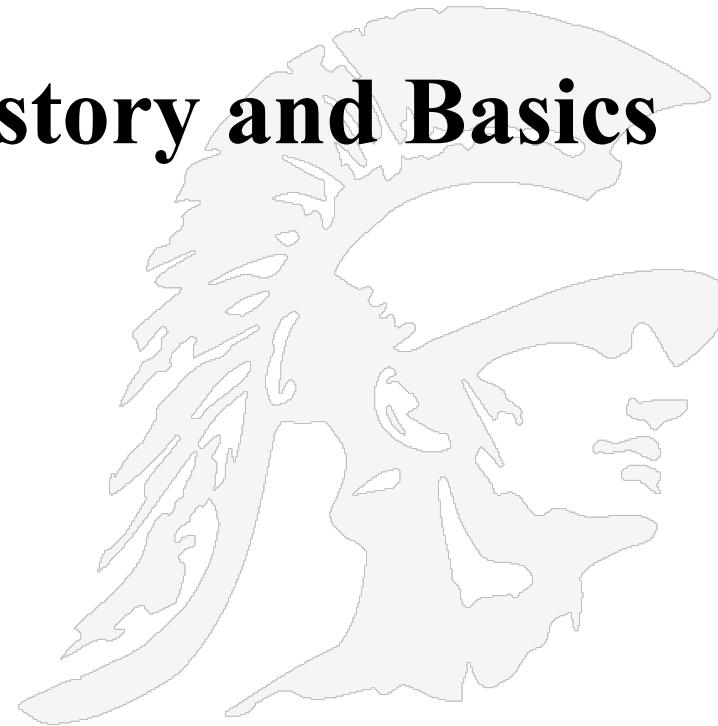


Search Engine History and Basics



A Brief Chronology of Search Engines

- 1991
 - Gopher, Archie, Veronica
- 1993
 - Wanderer,
 - ALIWeb
 - Excite

early search engines, non-web

- 1994
 - Galaxy
 - Yahoo
 - Lycos
 - WebCrawler
 - Alta Vista
- 1995
 - Infoseek
 - Metacrawler
 - SavvySearch
 - LookSmart

<http://www.excite.com/>

powerful indexing

<http://www.galaxy.com/>
<http://www.yahoo.com/>
<http://www.lycos.com/>
<http://www.webcrawler.com/>
<http://www.altavista.com/>

Early searchable directory
 Sophisticated searchable directory
 Improved query matching
 Includes full text of pages
 a large index

<http://www.infoseek.com/>
<http://www.metacrawler.com/>
<http://www.savvysearch.com/>
<http://www.looksmart.com>

included in Netscape Navigator
 combines results from other engines
 combines results from other engines
 convenient organization

<http://www.inktomi.com>
<http://www.hotbot.com/>

a large index using commodity hardware
 a large index

<http://www.askjeeves.com/>

fancy query processing

<http://www.goto.com/>
<http://www.google.com>

introduces auctioning of positions
 ranking using content and links

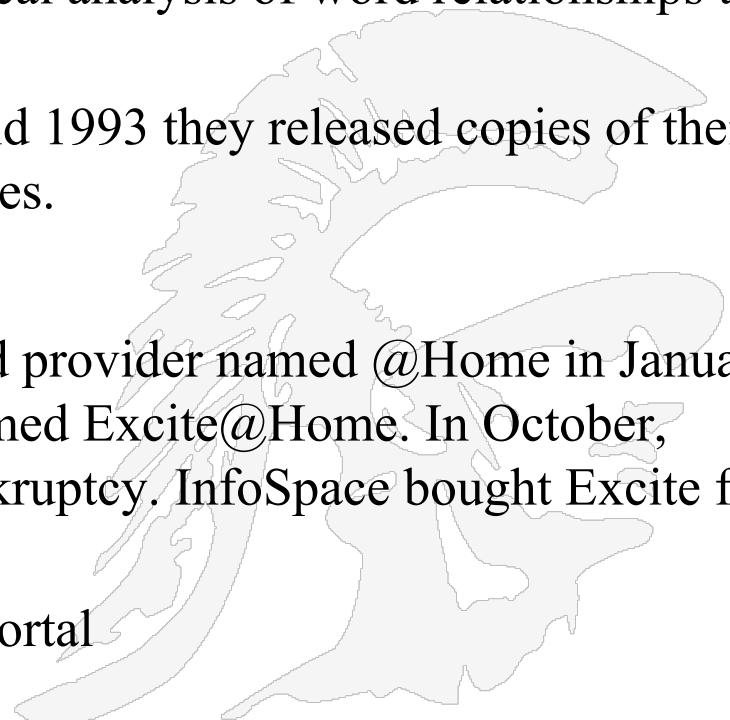
- Today there are hundreds of search engines, many are specialized
- See Search Engine History

- A very long web page describing the history of search Copyright © Ellits Idf goudt in 2011-2022

Archie, Veronica, Gopher

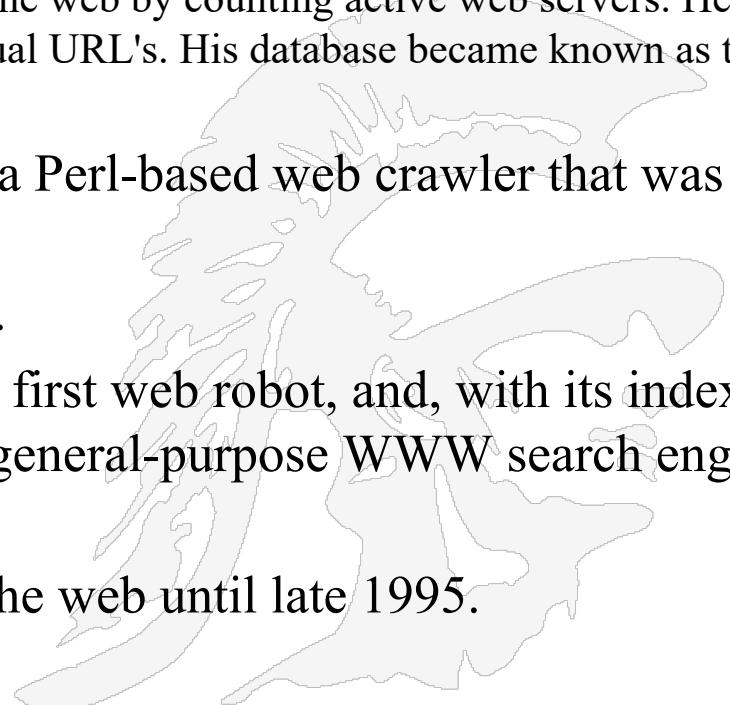
- By late 1980's many files were available by anonymous FTP.
- In 1990, Alan Emtage, P. Deutsch, et al of McGill Univ. developed Archie (short for “archives”)
 - Assembled lists of files available on many FTP servers.
 - Allowed regex search of these file names.
- In 1993, Veronica and Jughead were developed to search names of text files available through Gopher servers
 - The **Gopher protocol** is a TCP/IP application layer protocol designed for distributing, searching, and retrieving documents over the Internet. Strongly oriented towards a menu-document design
 - The Gopher ecosystem is often regarded as the effective predecessor of the World Wide Web

- Excite came from the project Architext, which was started in February, 1993 by six Stanford undergrad students.
 - They had the idea of using statistical analysis of word relationships to make searching more efficient.
 - They were soon funded, and in mid 1993 they released copies of their search software for use on web sites.
- Later developments
 - Excite was bought by a broadband provider named @Home in January, 1999 for \$6.5 billion, and was named Excite@Home. In October, 2001 Excite@Home filed for bankruptcy. InfoSpace bought Excite from bankruptcy court for \$10 million
 - www.excite.com still exists as a portal

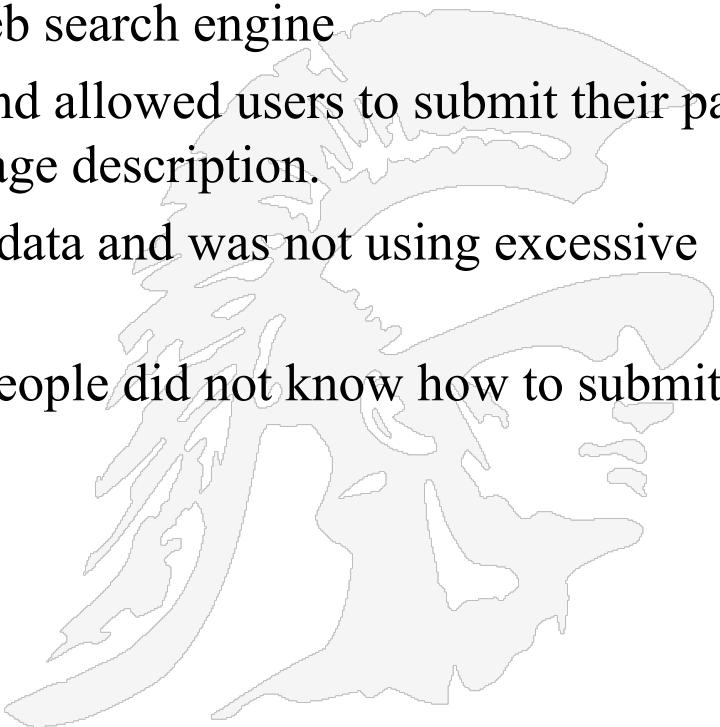


World Wide Web Wanderer

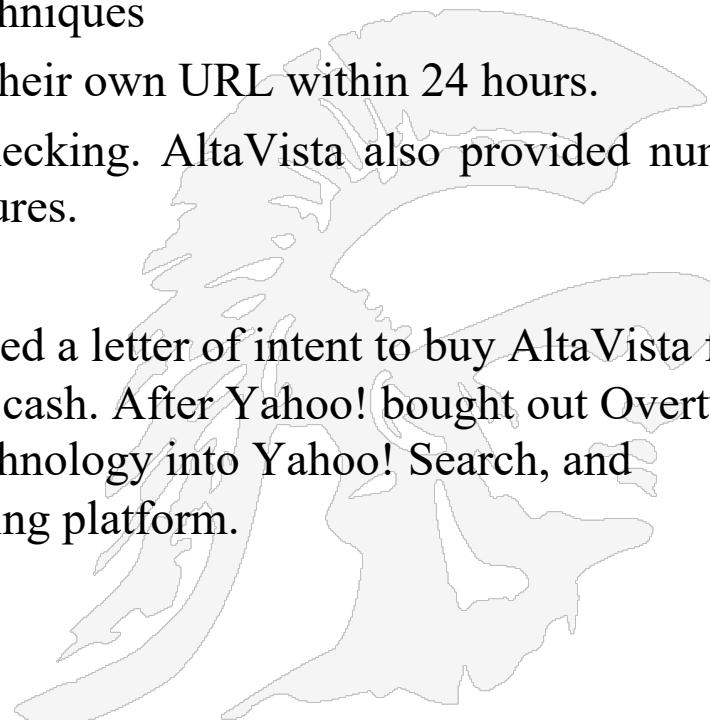
- In June 1993 Matthew Gray while at MIT introduced the World Wide Web Wanderer.
 - Initial goal was to measure the growth of the web by counting active web servers. He soon upgraded the software to capture actual URL's. His database became known as the Wandex.
- The World Wide Web Wanderer was a Perl-based web crawler that was first deployed in June 1993
- Matthew Gray now works for Google.
- While the Wanderer was probably the first web robot, and, with its index, clearly had the potential to become a general-purpose WWW search engine it never went that far
- The Wanderer charted the growth of the web until late 1995.



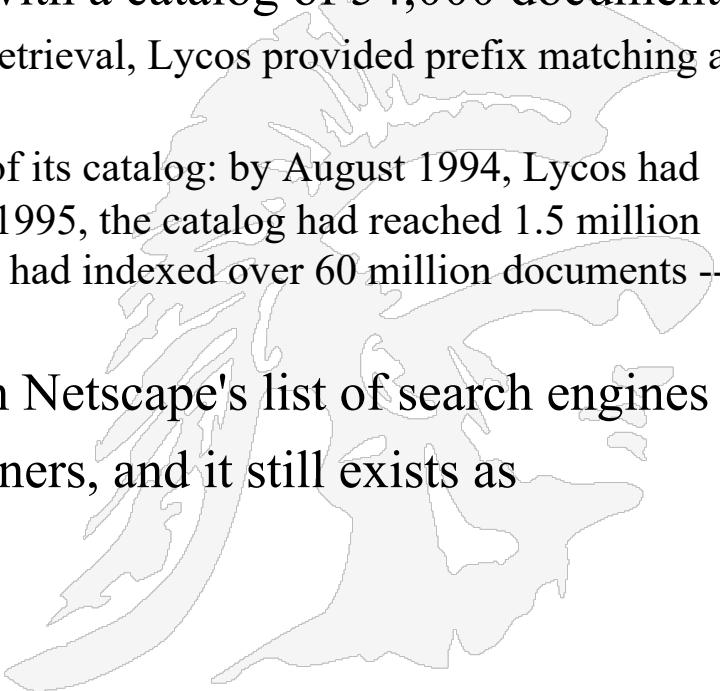
- In November of 1993 Martijn Koster created “Archie-Like Indexing of the Web”, or ALIWEB in response to the Wanderer.
 - Some consider it to be the first Web search engine
- ALIWEB crawled meta information and allowed users to submit their pages they wanted indexed with their own page description.
- This meant it needed no bot to collect data and was not using excessive bandwidth.
- One downside of ALIWEB was that people did not know how to submit their site



- AltaVista debut online came during December, 1995. AltaVista brought many important features to the web scene.
 - They were the first to allow natural language queries
 - They offered advanced searching techniques
 - They allowed users to add or delete their own URL within 24 hours.
 - They even allowed inbound link checking. AltaVista also provided numerous search tips and advanced search features.
- Later developments
 - On February 18, 2003, Overture signed a letter of intent to buy AltaVista for \$80 million in stock and \$60 million cash. After Yahoo! bought out Overture they rolled some of the AltaVista technology into Yahoo! Search, and occasionally used AltaVista as a testing platform.

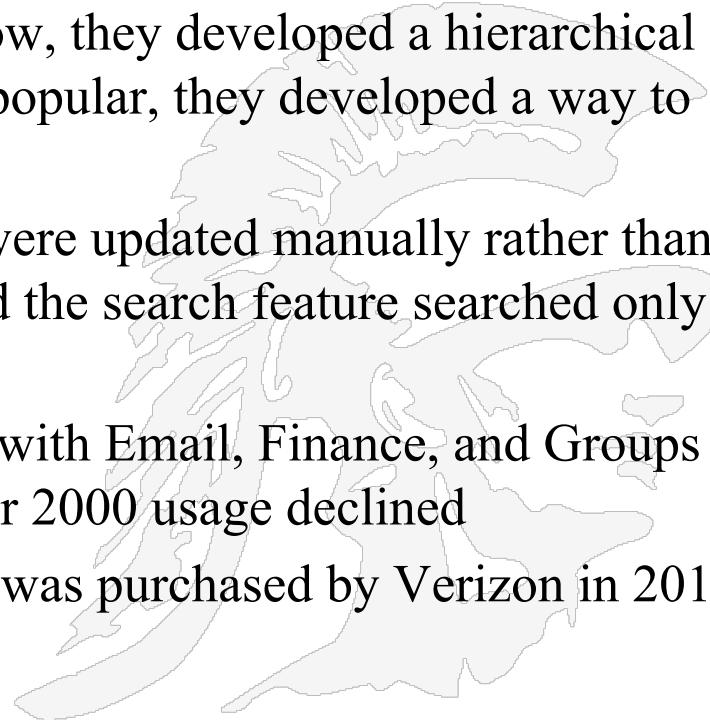


- Lycos was designed at Carnegie Mellon University around July of 1994. Michael Loren Mauldin was responsible for this search engine and was the chief scientist at Lycos Inc in the early years.
- On July 20, 1994, Lycos went public with a catalog of 54,000 documents.
 - In addition to providing ranked relevance retrieval, Lycos provided prefix matching and word proximity bonuses.
 - Lycos' main difference was the sheer size of its catalog: by August 1994, Lycos had identified 394,000 documents; by January 1995, the catalog had reached 1.5 million documents; and by November 1996, Lycos had indexed over 60 million documents -- more than any other Web search engine.
- In October 1994, Lycos ranked first on Netscape's list of search engines
- Lycos has gone through a series of owners, and it still exists as www.lycos.com

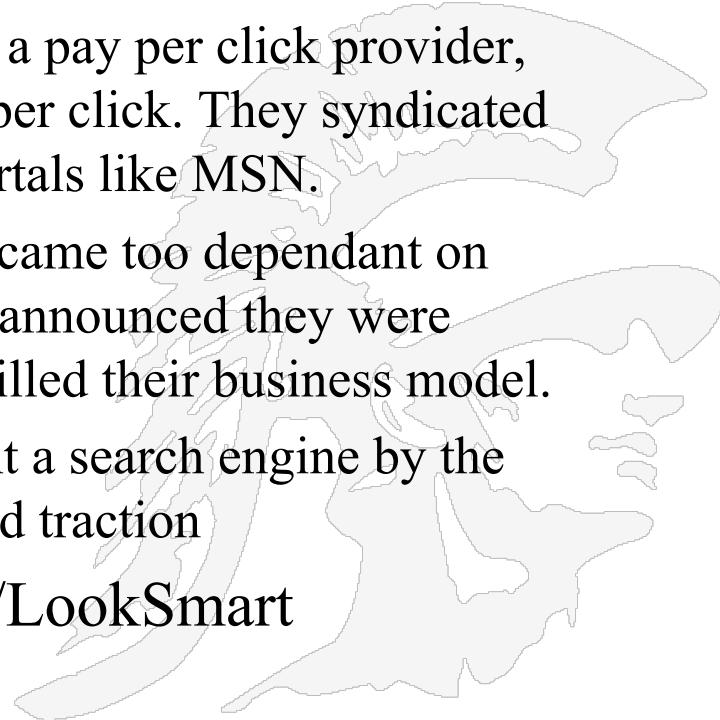


- Infoseek also started out in 1994, founded by Steve Kirsch
- In December 1995 they convinced Netscape to use them as their default search engine, which gave them major exposure.
- One popular feature of Infoseek was allowing webmasters to submit a page to the search index in real time, which was a search spammer's paradise
- They were the first search engine to sell advertising on a CPM (Cost per Thousand) impressions basis
- Infoseek was bought by Walt Disney Company in 1998

- In 1994, two Stanford Ph.D. students David Filo and Jerry Yang posted web pages with links on them, organized into a topical hierarchy.
- As the number of links began to grow, they developed a hierarchical listing. As the pages become more popular, they developed a way to search through all of the links.
- Early on all the links on the pages were updated manually rather than automatically by spider or robot and the search feature searched only those links
- Yahoo home page acted as a portal with Email, Finance, and Groups being very successful; however after 2000 usage declined
- After many years of decline Yahoo was purchased by Verizon in 2017 for \$4.48 billion, and it lives on



- Looksmart was founded in 1995 in Australia. They competed with the Yahoo! Directory by frequently increasing their inclusion rates
- Later developments
 - In 2002 Looksmart transitioned into a pay per click provider, which charged listed sites a flat fee per click. They syndicated those paid listings to some major portals like MSN.
 - The problem was that Looksmart became too dependant on MSN, and in 2003, when Microsoft announced they were dumping Looksmart that basically killed their business model.
 - In March of 2002, Looksmart bought a search engine by the name of WiseNut, but it never gained traction
- See <https://en.wikipedia.org/wiki/LookSmart>

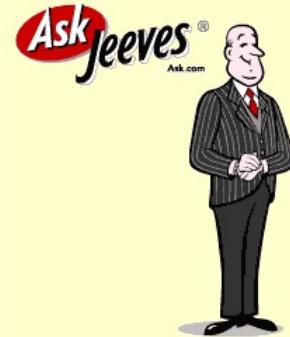




Inktomi

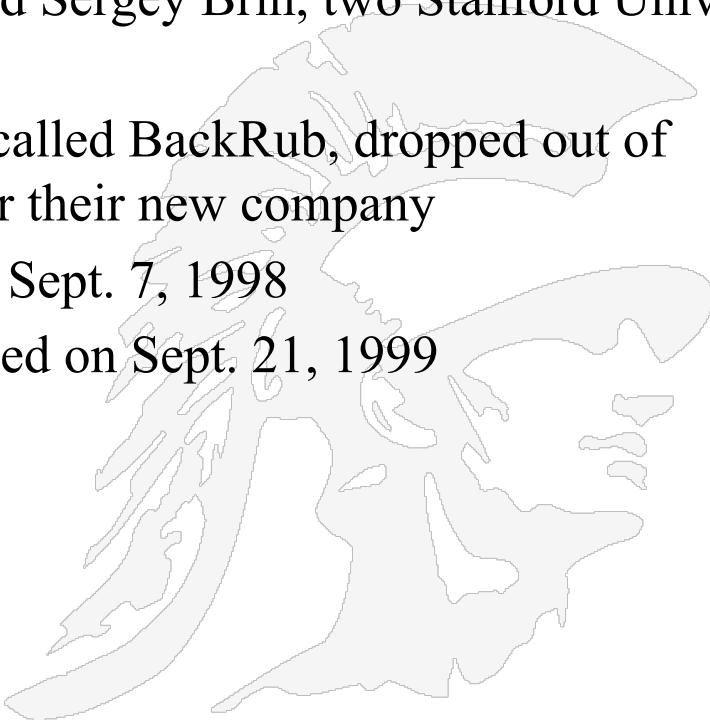
- The Inktomi Corporation came about on May 20, 1996 with its search engine Hotbot. Two Cal Berkeley cohorts created Inktomi from the improved technology gained from their research
- Later developments
 - In October of 2001 Inktomi accidentally allowed the public to access their database of spam sites, which listed over 1 million URLs at that time.
 - Inktomi pioneered ***the paid inclusion model*** in which a website pays a fee to the search engine that guarantees the site will be displayed when certain search terms are entered
 - The model was nowhere near as efficient as the pay-per-click auction model developed by Overture. Licensing their search results also was not profitable enough to pay for their scaling costs. They failed to develop a profitable business model, and sold out to Yahoo! for approximately \$235 million, or \$1.65 a share, in December of 2003.

*<http://searchenginewatch.com/article/2066745/Inktomi-Spam-Database-Left-Open-To-Public>

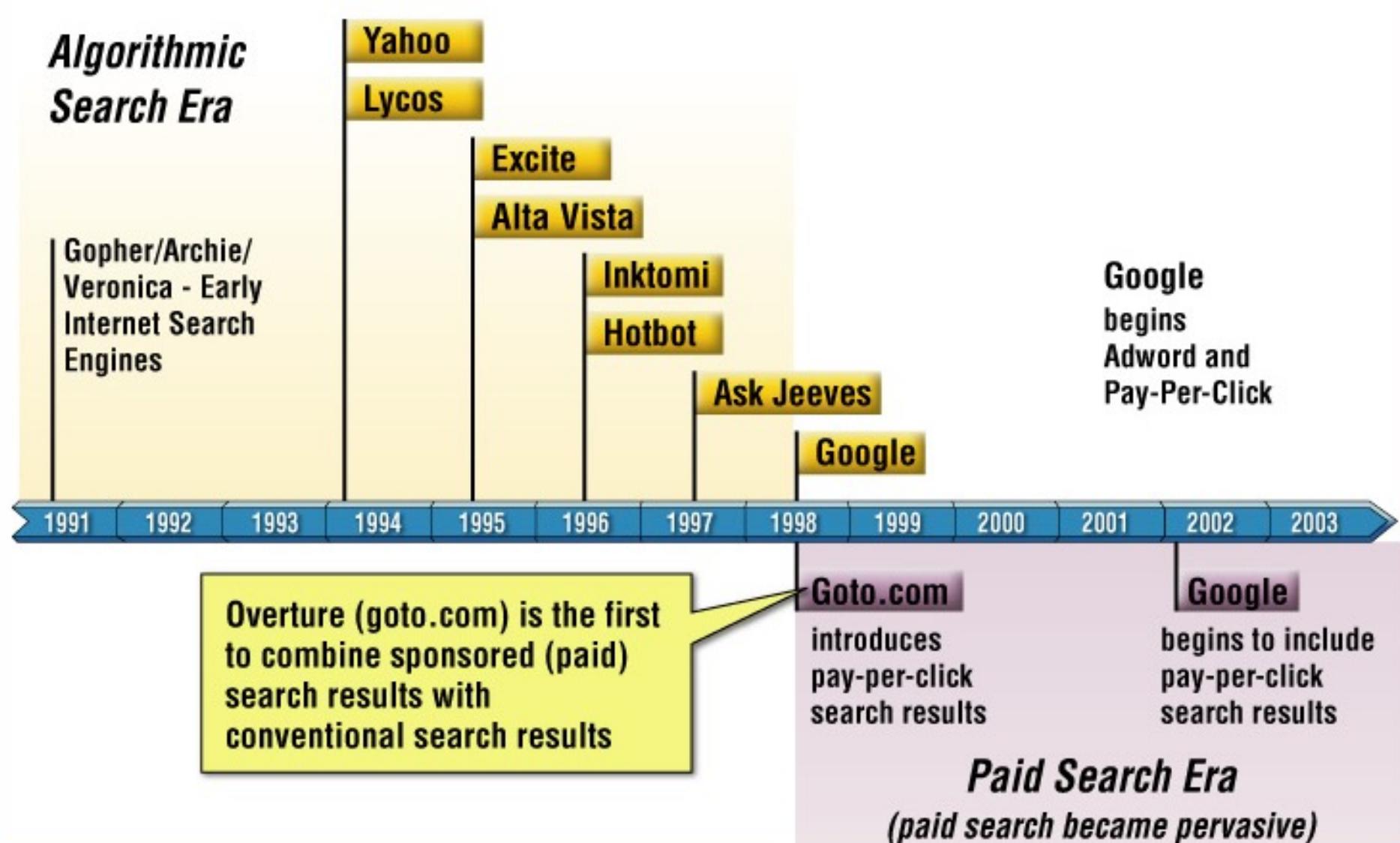


- In April of 1997 Ask Jeeves was launched as a natural language search engine.
 - Ask Jeeves used human editors to try to match search queries.
 - Ask was powered by DirectHit for a while, which aimed to rank results based on their popularity, but that technology proved too easy to spam.
 - In 2000 the Teoma search engine was released, which uses clustering to organize sites by Subject Specific Popularity, which is another way of saying they tried to find local web communities. In 2001 Ask Jeeves bought Teoma to replace the DirectHit search technology.
 - On March 21, 2005 Barry Diller's IAC agreed to acquire Ask Jeeves for 1.85 billion dollars. IAC owns many popular websites like Match.com, Ticketmaster.com, and Citysearch.com, and is promoting Ask across their other properties.
 - In 2006 Ask Jeeves was renamed to Ask.

- Google is a play on the word Googol, coined by Milton Sirotta; it refers to a 1 followed by 100 zeros, 10000000.....0
- A googol is bigger than the number of atoms in the universe
- Google was founded by Larry Page and Sergey Brin, two Stanford Univ. Computer Science graduate students
- In 1998 they built a prototype system called BackRub, dropped out of school, and tried to attract investors for their new company
- Google Inc. released a beta version on Sept. 7, 1998
- www.google.com was officially released on Sept. 21, 1999



A Brief Chronology of Search Engines

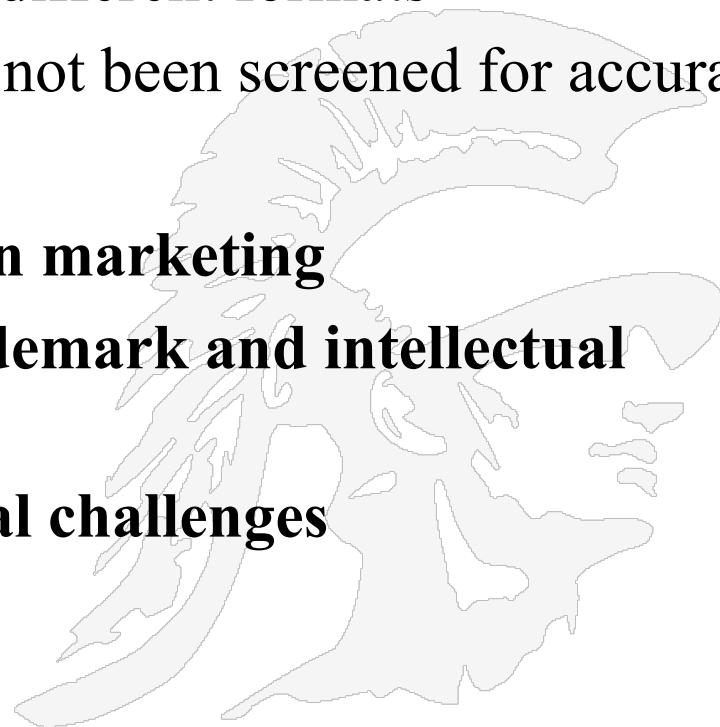


Search Engine Basic Behavior



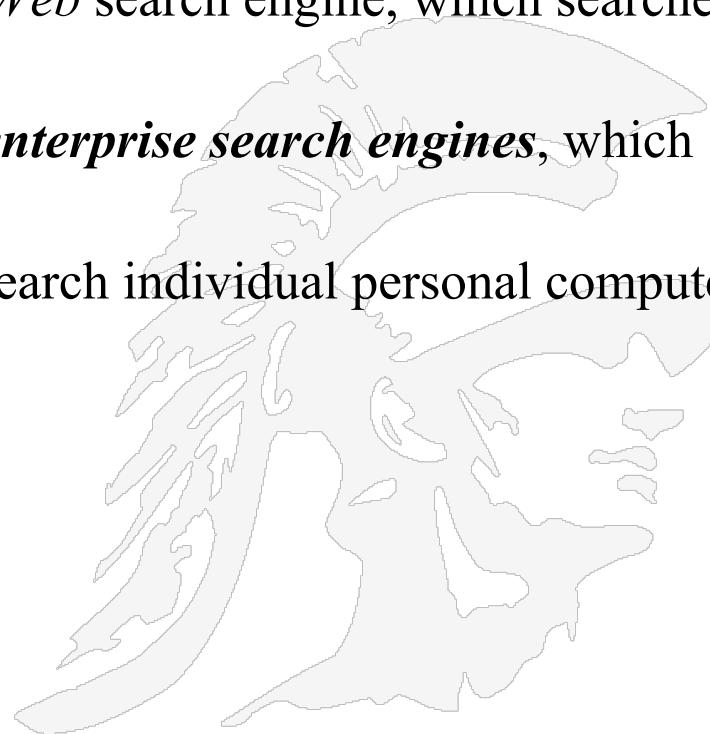
What is Web Search?

- **Providing access to heterogeneous, distributed information that is publicly available on the World Wide Web**
 - Information comes in many different formats
 - Most of the information has not been screened for accuracy
- **Multi-billion dollar business**
- **Source of new opportunities in marketing**
- **Strains the boundaries of trademark and intellectual property laws**
- **A source of unending technical challenges**

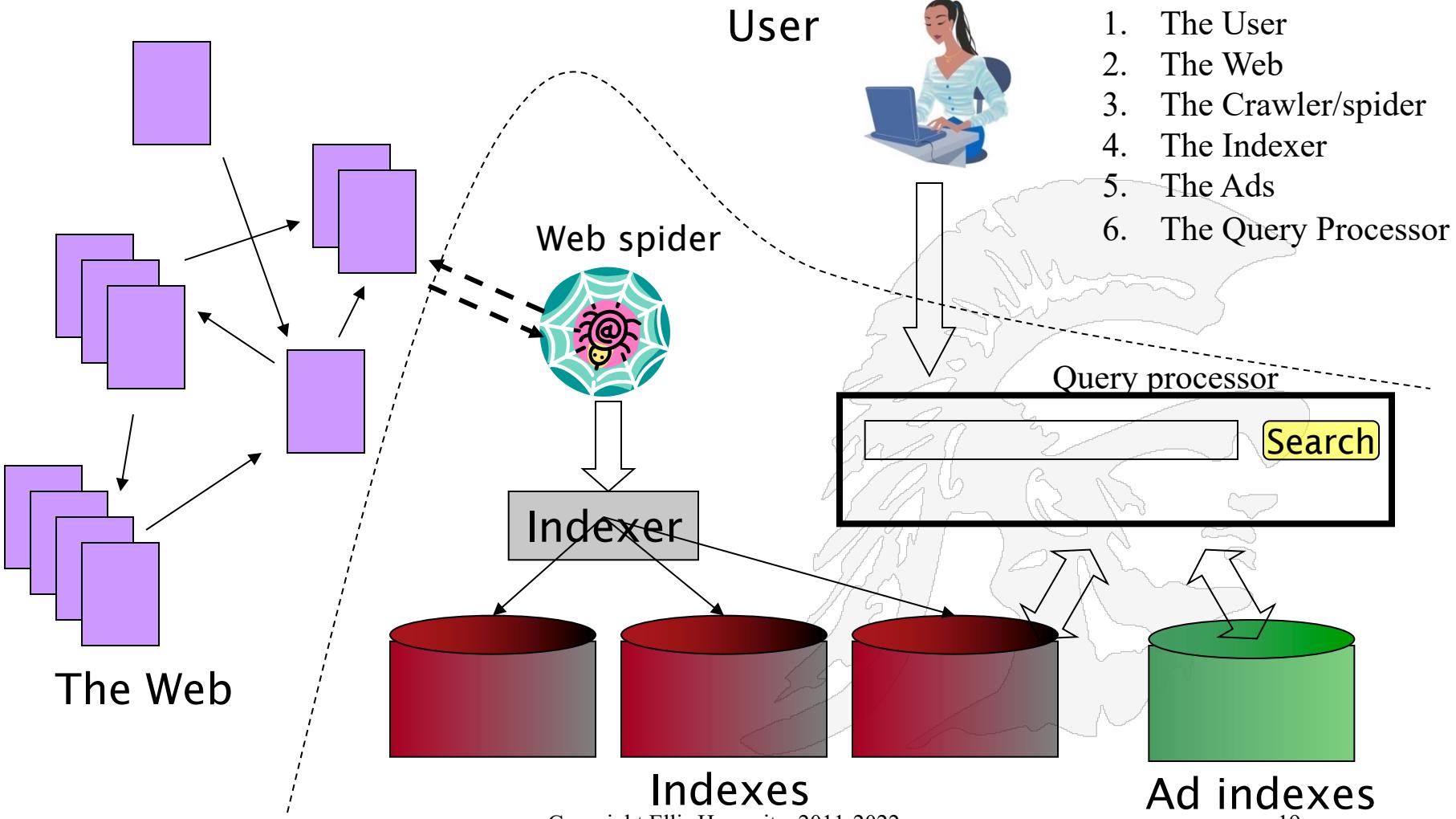


Web Search Engine Definitions

- “A search engine is a program designed to help find information stored on a computer system such as the World Wide Web, inside a corporate or proprietary network or a personal computer” *wikipedia*
 - *search engine* usually refers to a *Web search engine*, which searches for information on the public Web.
 - Other kinds of search engine are *enterprise search engines*, which search on intranets,
 - *personal search engines*, which search individual personal computers

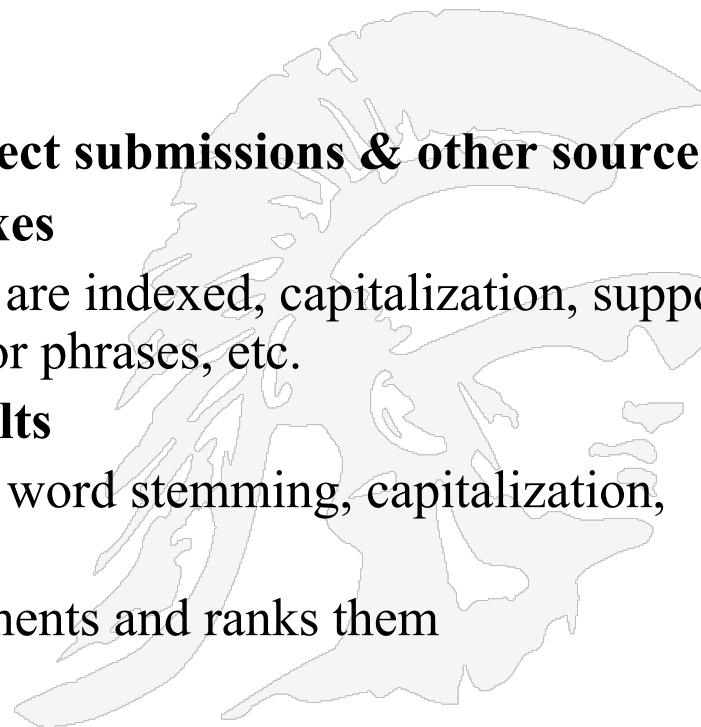


Basic Web Search Internals

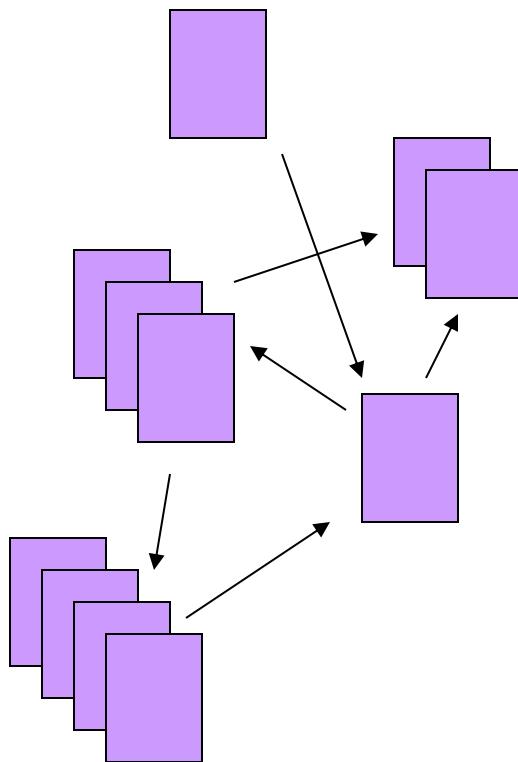


Web Search Engine Elements

- ***Spider* (a.k.a. crawler/robot) – builds **corpus****
 - **Collects web pages recursively**
 - For each known URL, fetch the page, parse it, and extract new URLs
 - Repeat
 - **Additional pages come from direct submissions & other sources**
- **The *indexer* – creates inverted indexes**
 - Various policies wrt which words are indexed, capitalization, support for Unicode, stemming, support for phrases, etc.
- ***Query processor* – serves query results**
 - **Front end** – query reformulation, word stemming, capitalization, optimization of Booleans, etc.
 - **Back end** – finds matching documents and ranks them

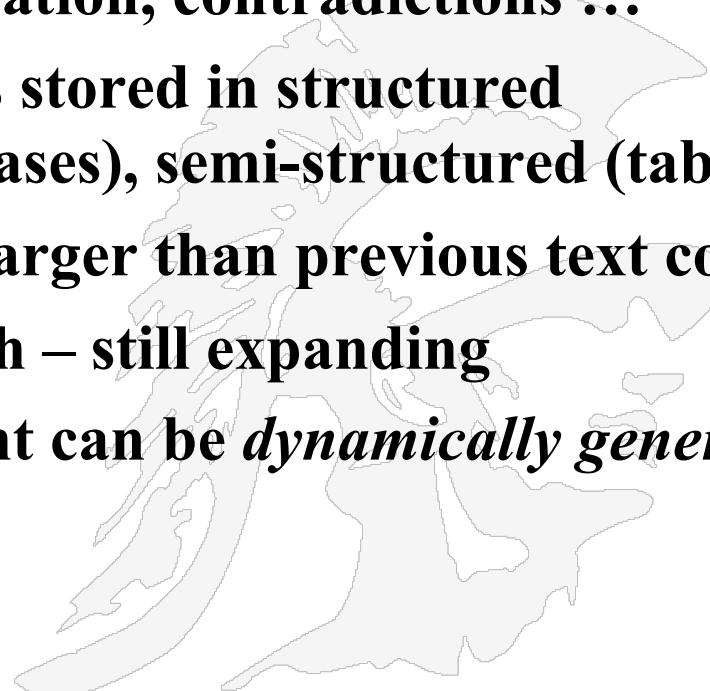


The Web



The Web

- No design/co-ordination
- Distributed content creation, linking
- Content includes truth, lies, obsolete information, contradictions ...
- Data is stored in structured (databases), semi-structured (tables)...
- Scale larger than previous text corpora
- Growth – still expanding
- Content can be *dynamically generated*



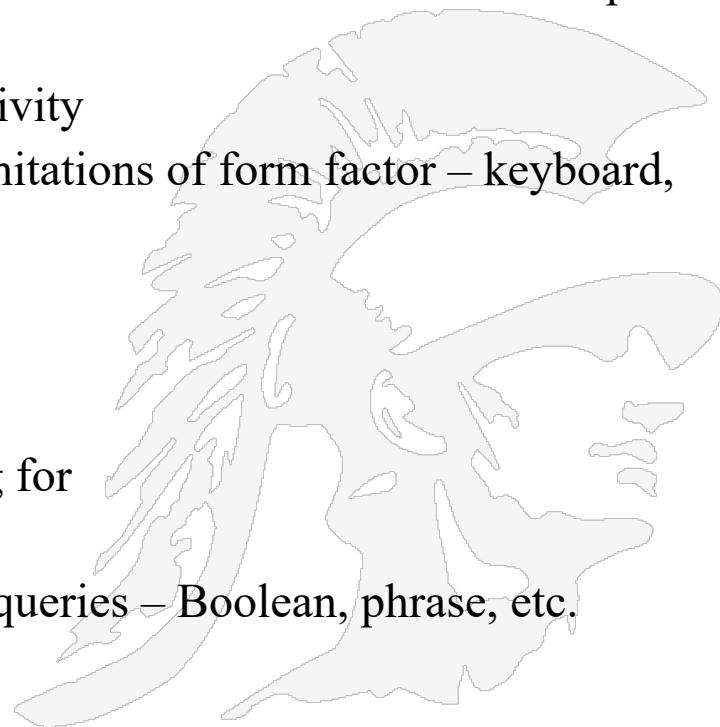
Most Dynamic Content is Missed

- Typically dynamic content is not indexed
- A page without a static html version, e.g.
 - current status of flight AA129
 - current availability of rooms at a hotel
- Dynamic content is usually assembled at the time of a request from a browser
 - To spot dynamic content typically a URL has a ‘?’ character in it
 - Some dynamic content includes malicious spider traps (infinite loops)
- The term *deep web* refers to content missed by search engine crawlers



The User

- **Diverse in background/training**
 - Users sometimes cannot tell the difference between a search bar from the URL address field (**Chrome conflates the two**)
 - Users rarely use the scroll bar, so key results must be at or near the top
- **Diverse in access methodology**
 - Increasingly, high bandwidth connectivity
 - Growing segment of mobile users: limitations of form factor – keyboard, display
- **Diverse in search methodology**
 - Search, search + browse,
 - Average query length ~ 2.5 terms
 - Has to do with what they're searching for
- **Poor comprehension of syntax**
 - Early engines offered rich syntax for queries – Boolean, phrase, etc.
 - Current engines hide these



User's Information Needs Are Diverse

- **Informational** – want to learn about something (~40%)

e.g. Low hemoglobin

- **Navigational** – want to go to that page (~25%)

e.g. United Airlines

- **Transactional** – want to do something (web-mediated) (~35%)

- Access a service

Los Angeles weather

- Downloads

Mars surface images

- Shop

Nikon CoolPix Camera

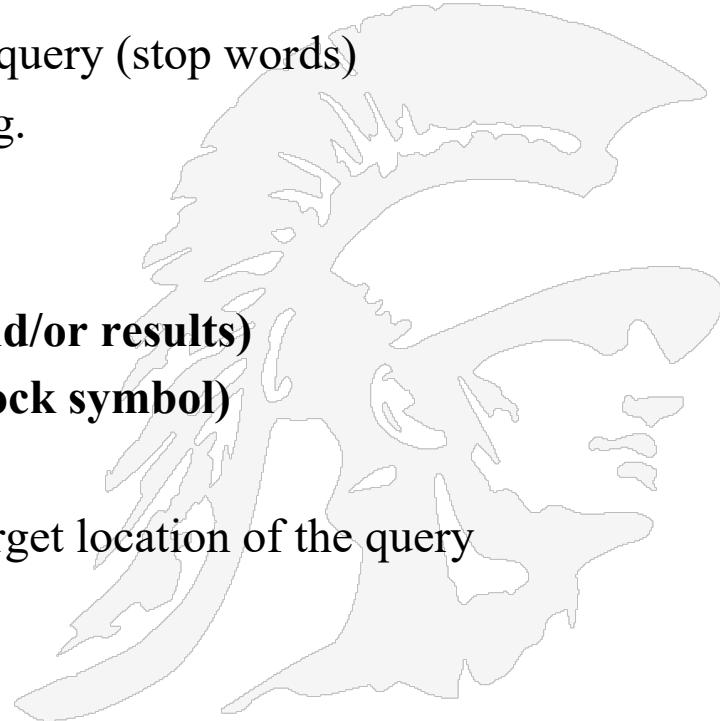
- **Gray areas**

- Find a good hub

Car rental in Finland

- Exploratory search “see what’s there”

- Query processing involves much more than just matching query terms with document terms
- Semantic analysis of the query includes:
 1. Determining the language of the query
 2. Filtering of unnecessary words from the query (stop words)
 3. Looking for specific types of queries, e.g.
 - **Personalities (triggered on names)**
 - **Cities (travel info, maps)**
 - **Medical info (triggered on names and/or results)**
 - **Stock quotes, news (triggered on stock symbol)**
 - **Company info ...**
 4. Determining the user's location or the target location of the query
 5. Remembering previous queries
 6. Maintaining a user profile



A Person Query

File Edit View History Bookmarks Tools Help

george clooney - Google S... +

https://www.google.com/#q=george+clooney

Google george clooney

Web News Images Videos Shopping More Search tools

About 37,700,000 results (0.23 seconds)

In the news

 [George Clooney, Amal Alamuddin Honeymoon in New British Home](#)
Us Magazine - 4 hours ago
Newlyweds George Clooney and Amal Alamuddin have skipped the traditional far-flung ...

People: [George Clooney's Wedding Cost About \\$1.6 Million](#)
Yahoo! Voices - 2 days ago

George Clooney & Amal Alamuddin Could Nab His & Hers Nobel Peace Prizes, Friend Predicts
People Magazine - 23 hours ago

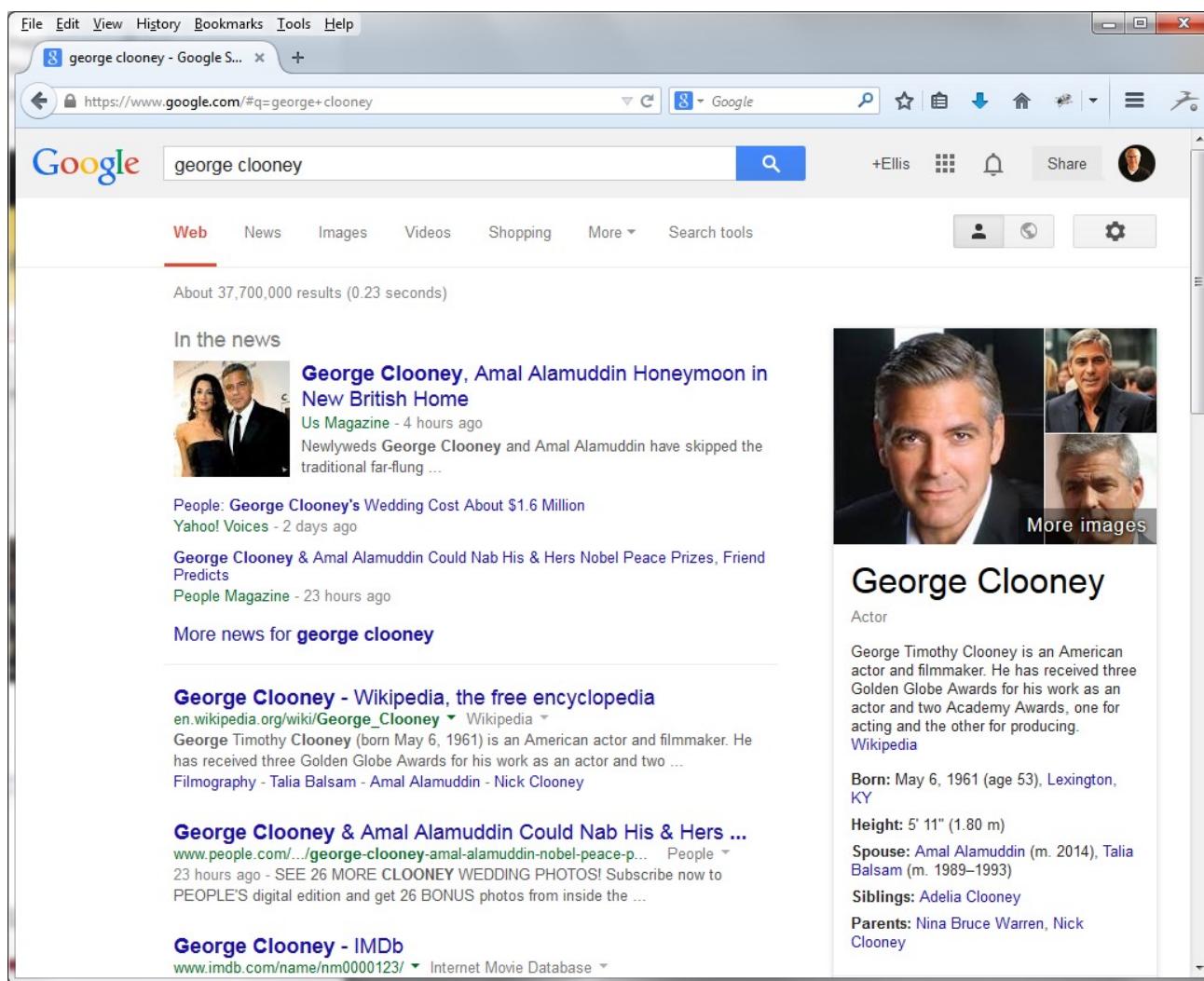
[More news for george clooney](#)

[George Clooney - Wikipedia, the free encyclopedia](#)
en.wikipedia.org/wiki/George_Clooney ▾ Wikipedia

George Timothy Clooney (born May 6, 1961) is an American actor and filmmaker. He has received three Golden Globe Awards for his work as an actor and two ...
Filmography - Talia Balsam - Amal Alamuddin - Nick Clooney

[George Clooney & Amal Alamuddin Could Nab His & Hers ...](#)
www.people.com/.../george-clooney-amal-alamuddin-nobel-peace-p... People ▾
23 hours ago - SEE 26 MORE CLOONEY WEDDING PHOTOS! Subscribe now to PEOPLE'S digital edition and get 26 BONUS photos from inside the ...

[George Clooney - IMDb](#)
www.imdb.com/name/nm0000123/ ▾ Internet Movie Database ▾



The right side of the screen displays a detailed profile of George Clooney. It includes a large portrait photo, a smaller inset photo, and a "More images" link. Below the photo, the name "George Clooney" is displayed in a large font, followed by the title "Actor". A biography states: "George Timothy Clooney is an American actor and filmmaker. He has received three Golden Globe Awards for his work as an actor and two Academy Awards, one for acting and the other for producing." A "Wikipedia" link is provided. Key facts listed are: Born: May 6, 1961 (age 53), Lexington, KY; Height: 5' 11" (1.80 m); Spouse: Amal Alamuddin (m. 2014), Talia Balsam (m. 1989–1993); Siblings: Adelia Clooney; Parents: Nina Bruce Warren, Nick Clooney.

Includes the following:

Latest news

Biography

Photos

Basic facts

born

married

parents

career

A Place Query

las vegas - Google Search

<https://www.google.com/search?q=las+vegas&oq=las+vegas&aqs=chrome..69i57j0l5.1920j0j8&sourceid=chrome&ie=UTF-8>

Apps CSCI152 Home Page CSCI571 Home Page CSCI351 Home Page Ellis Horowitz' Home... Computer Science D... Other bookmarks

Google las vegas

Web News Images Maps Videos More Search tools

About 168,000,000 results (0.41 seconds)

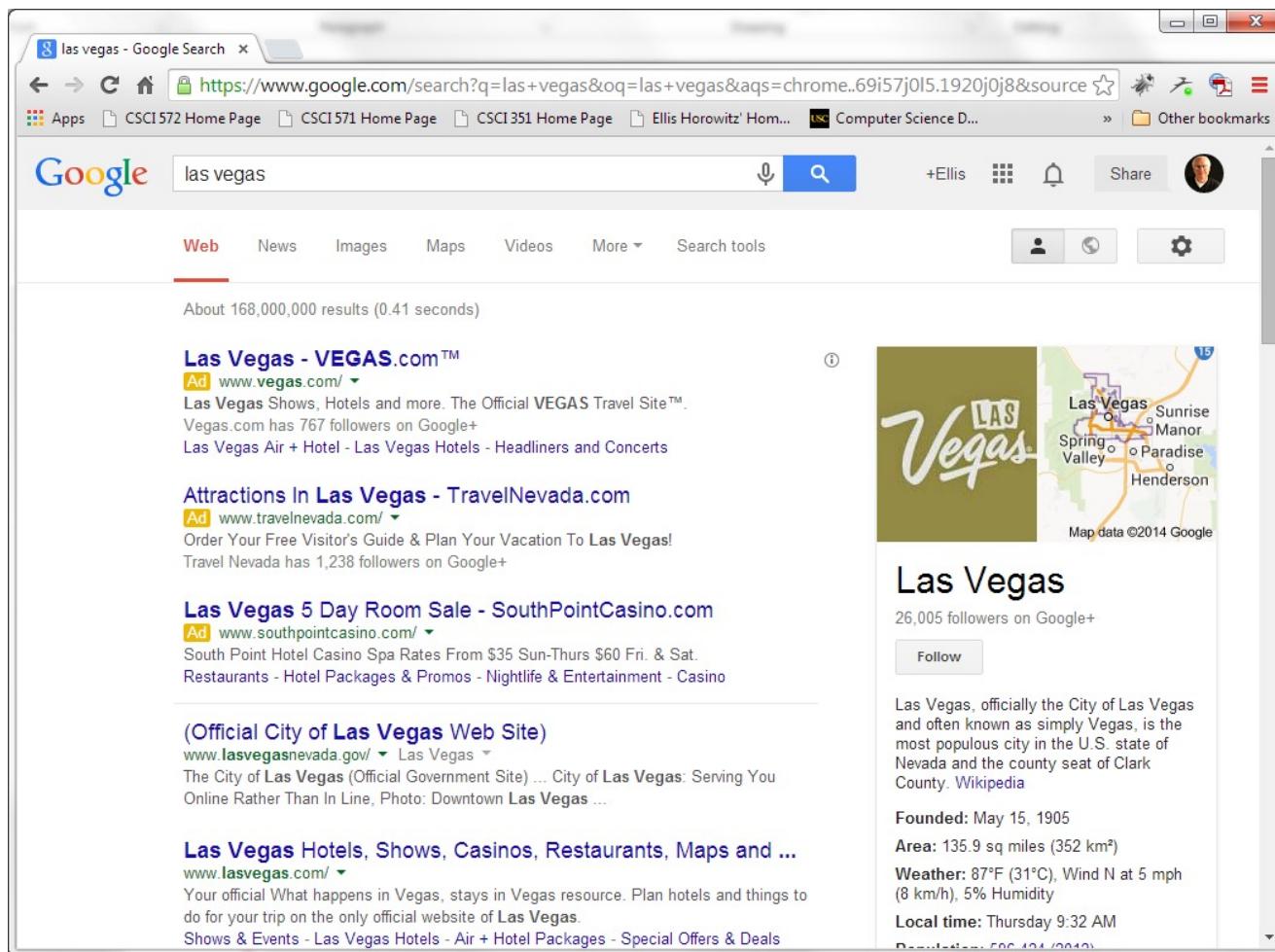
Las Vegas - VEGAS.com™
Ad www.vegas.com/ Las Vegas Shows, Hotels and more. The Official VEGAS Travel Site™. Vegas.com has 767 followers on Google+. Las Vegas Air + Hotel - Las Vegas Hotels - Headliners and Concerts

Attractions In Las Vegas - TravelNevada.com
Ad www.travelnevada.com/ Order Your Free Visitor's Guide & Plan Your Vacation To Las Vegas! Travel Nevada has 1,238 followers on Google+

Las Vegas 5 Day Room Sale - SouthPointCasino.com
Ad www.southpointcasino.com/ South Point Hotel Casino Spa Rates From \$35 Sun-Thurs \$60 Fri. & Sat. Restaurants - Hotel Packages & Promos - Nightlife & Entertainment - Casino

(Official City of Las Vegas Web Site)
www.lasvegasnevada.gov/ Las Vegas The City of Las Vegas (Official Government Site) ... City of Las Vegas: Serving You Online Rather Than In Line, Photo: Downtown Las Vegas ...

Las Vegas Hotels, Shows, Casinos, Restaurants, Maps and ...
www.lasvegas.com/ Your official What happens in Vegas, stays in Vegas resource. Plan hotels and things to do for your trip on the only official website of Las Vegas. Shows & Events - Las Vegas Hotels - Air + Hotel Packages - Special Offers & Deals



Includes the following:

Official site
Map
Essential facts
founded
area
weather
time
population

An Hotel Query

sheraton times square hotel x

<https://www.google.com/search?q=las+vegas&oq=las+vegas&aqs=chrome..69i57j0l5.1920j0j8&sourceid=chrome&ie=UTF-8>

Apps CSCI 572 Home Page CSCI 571 Home Page CSCI 351 Home Page Ellis Horowitz' Hom... Computer Science D... Other bookmarks

Google sheraton times square hotel nyc

Web Maps Images News Shopping More Search tools

About 24,000,000 results (0.55 seconds)

Sheraton™ New York Hotel - Official Site - Our Best Rates
Ad www.sheraton.com/TimesSquare
 Guaranteed. Book Now!
 Sheraton Hotels & Resorts has 1,593 followers on Google+
 811 7th Avenue, New York, NY
 Photos Make a Reservation Features & Amenities Special Offers

NYC Times Square Hotel - sofitel-new-york.com
Ad www.sofitel-new-york.com/
 Sofitel New York Hotel in Manhattan near Times Square. Book Direct Now!
 Make Your Reservation - Enjoy Business @ Sofitel - Sumptuous Terrace Suites

Sheraton New York Times Square Hotel: Hotel Near Time...
www.sheratonnewyork.com/
 The Sheraton New York Times Square Hotel is located between Central Park and Times Square, and offers newly renovated spaces and sophisticated ...
 3.2 ★★★★☆ 141 Google reviews - Write a review - \$239▼
 811 7th Ave, New York, NY 10019 (212) 581-1000 Photos & Videos - Sheraton New York Times ...

Sheraton New York Times Square Hotel - Starwood Hotel...
www.starwoodhotels.com/sherato... Starwood Hotels & Resorts Worldwide ▾
 ★★★★☆ Rating: 3.7 - 1,851 reviews

Map data ©2014 Google

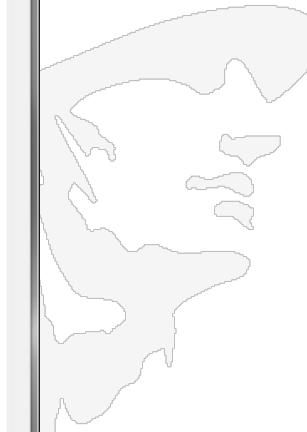
Sheraton New York Times Square Hotel
 \$239 Book Directions

Are you the business owner? Feedback

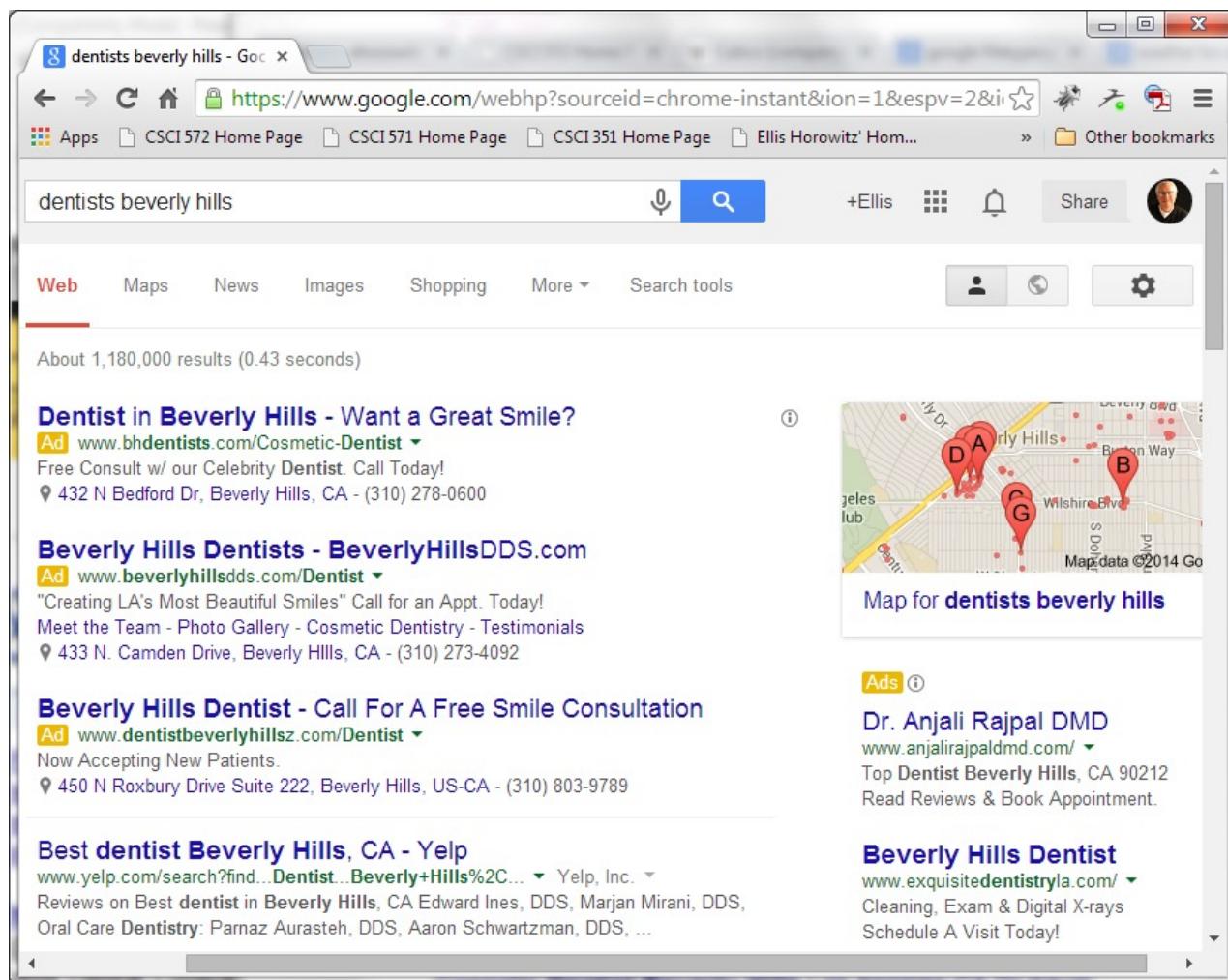
Ads Sheraton Times Square Hotel sheratonnewyork.reservations.com/
 Book Sheraton Times Square For Limited Offers - Reserve Today!

Includes the following:

- Main hotel website
- Map
- Address
- Phone number
- Price of a room
- Directions



Find a Professional Query



A screenshot of a Google search results page for "dentists beverly hills". The search bar at the top shows the query. Below it, there are several search result snippets. The first snippet is for "Dentist in Beverly Hills - Want a Great Smile?", which is an ad from www.bhdentists.com/Cosmetic-Dentist. It includes a phone number and address. The second snippet is for "Beverly Hills Dentists - BeverlyHillsDDS.com", also an ad, with a phone number and address. The third snippet is for "Beverly Hills Dentist - Call For A Free Smile Consultation", another ad, with a phone number and address. The fourth snippet is for "Best dentist Beverly Hills, CA - Yelp", linking to www.yelp.com/search?find...Dentist...Beverly+Hills%2C.... To the right of the search results, there is a map of Beverly Hills with four red pins labeled A, B, D, and G. Below the map is a link to "Map for dentists beverly hills". Further down the page, there are two more ads: one for "Dr. Anjali Rajpal DMD" with a link to www.anjalirajpalmd.com/, and another for "Beverly Hills Dentist" with a link to www.exquisitedentistryla.com/.

Includes the following:

Ads at top and side

Map pointing to specific dentists

Reviews of dentists (Yelp)

Query Expansion for “Dentists”

8 dentists beverly hills - Goo x

https://www.google.com/webhp?sourceid=chrome-instant&ion=1&espv=2&i...

Apps CSCI 572 Home Page CSCI 571 Home Page CSCI 351 Home Page Ellis Horowitz' Hom... Other bookmarks

Beverly Hills Pediatric Dentist - Beverly Hills Pediatric Dental Care. Welcome to our practice! Drs. Gross, Lempert, and associates have been providing the ...

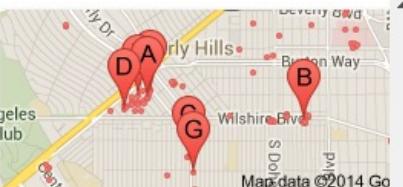
Beverly Hills Dentist | Top Los Angeles Cosmetic Dentistry
www.drmaddahi.com/
by Kourosh Maddahi - in 52 Google+ circles
As a top Beverly Hills cosmetic dentist, Dr. Maddahi is known throughout the world by his patients as the smile transformation expert. Our Los Angeles Dental ...

Beverly Hills Dentist, Prosthodontics, Cosmetic Dentistry
www.beverlyhillsdds.com/
Beverly Hills CA Prosthodontists provide Dental Implants, Porcelain Veneers, Teeth Whitening, Restorative Dentistry, Full Mouth Reconstruction. 310-273-4092.

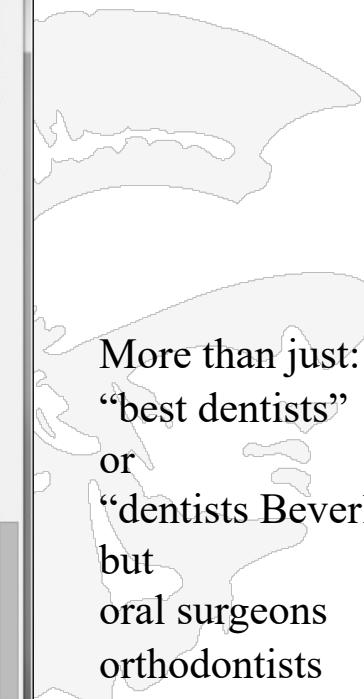
Searches related to dentists beverly hills

oral surgeons beverly hills best dentists beverly hills
yellow pages beverly hills dentists beverly hills saturday hours
orthodontists beverly hills beverly hills dental
dentists hollywood beverly hills dentist reviews

Goooooooooooooogle >
1 2 3 4 5 6 7 8 9 10 Next

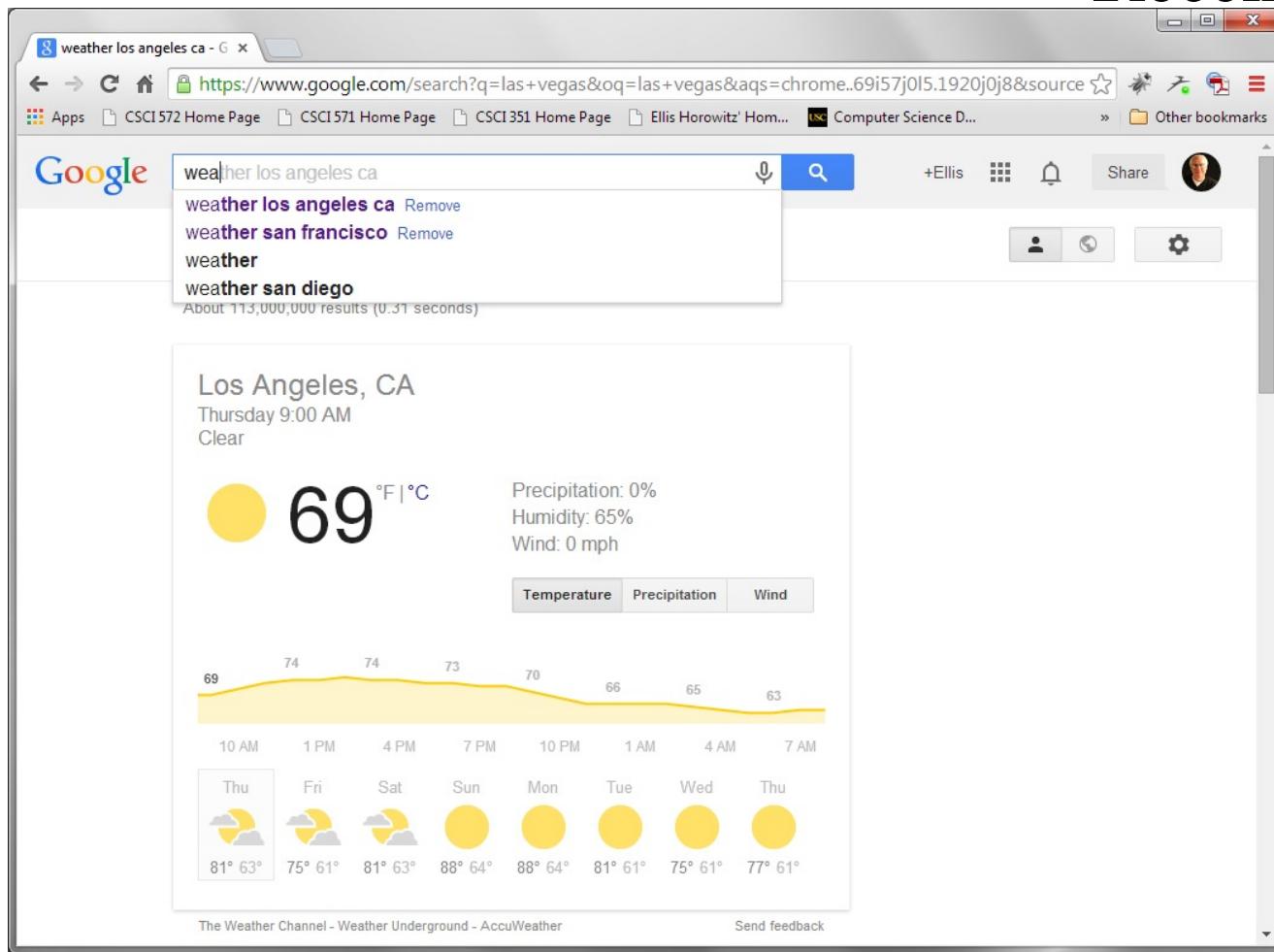


Map for dentists beverly hills



More than just:
“best dentists”
or
“dentists Beverly Hills”
but
oral surgeons
orthodontists

Google Maintains Your Recent Query History



A screenshot of a Google search results page. The search query "weather los angeles ca" is entered in the search bar. Below the search bar, a dropdown menu shows recent queries: "weather los angeles ca", "weather san francisco", "weather", and "weather san diego". The main search result is for Los Angeles, CA, showing the current temperature of 69°F (21°C), clear skies, and a 7-day forecast.

Recent Searches:

- weather los angeles ca
- weather san francisco
- weather
- weather san diego

Search Results:

Los Angeles, CA
Thursday 9:00 AM
Clear

69 °F | °C

Precipitation: 0%
Humidity: 65%
Wind: 0 mph

Temperature Precipitation Wind

Time	Temp (°F)	Temp (°C)
10 AM	69	21
1 PM	74	23
4 PM	74	23
7 PM	73	22
10 PM	70	21
1 AM	66	19
4 AM	65	18
7 AM	63	17

Thu Fri Sat Sun Mon Tue Wed Thu

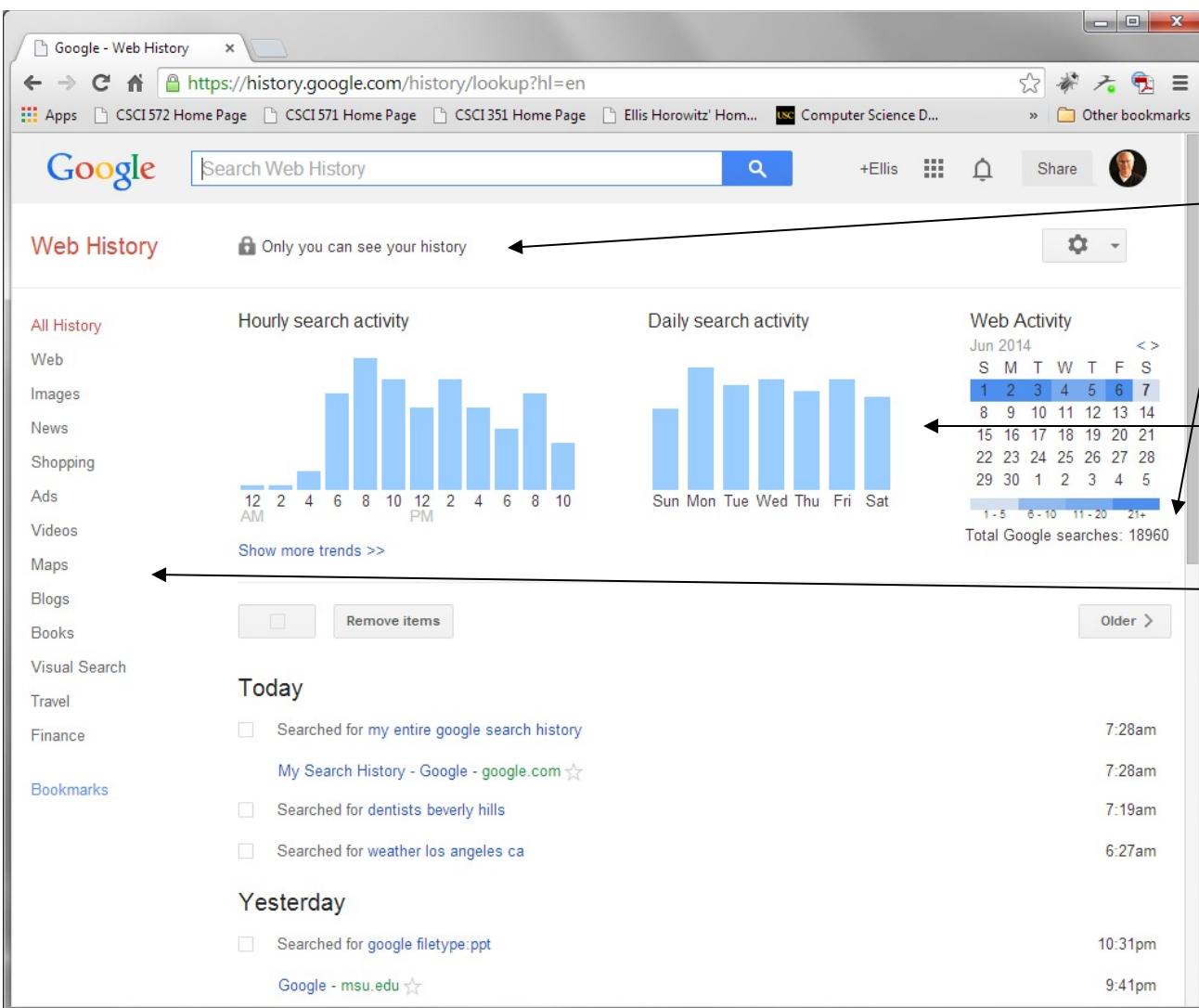
Day	Icon	Temp (°F)	Temp (°C)
Thu	Sun	81° 63°	27° 17°
Fri	Sun	75° 61°	24° 16°
Sat	Sun	81° 63°	27° 17°
Sun	Sun	88° 64°	31° 18°
Mon	Sun	88° 64°	31° 18°
Tue	Sun	81° 61°	27° 16°
Wed	Sun	75° 61°	24° 16°
Thu	Sun	77° 61°	25° 16°

The Weather Channel - Weather Underground - AccuWeather

Send feedback

Maintain previous queries
Helps to minimize typing
Allow users to remove old ones

Google Retains a User's Entire Query History!



They claim that only I can see my history;
 I have issued a total of 18,960 queries;

Graphs show my queries by hour and by week;

I can view my Web queries as distinct from my Image queries or my News queries, etc

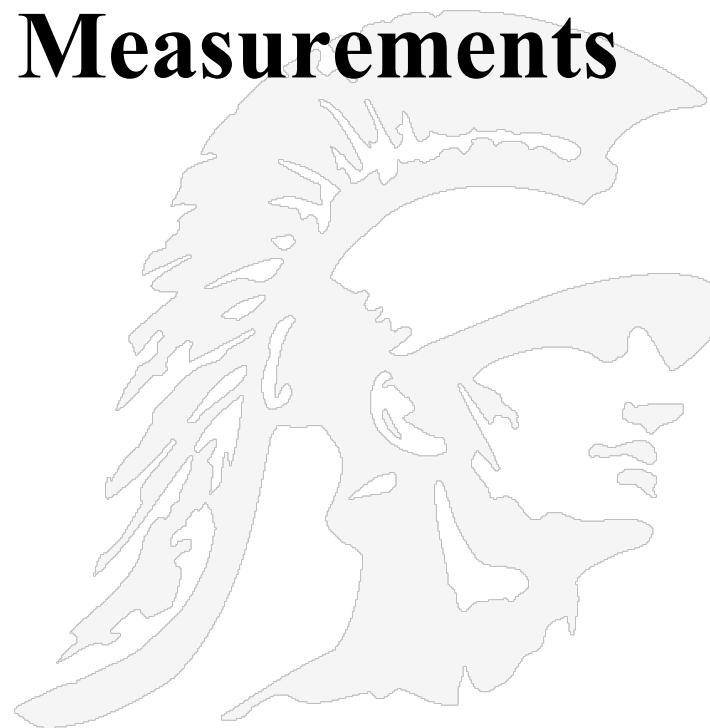
As a result, Google now knows a great deal about us!

Search Engines are an Industry

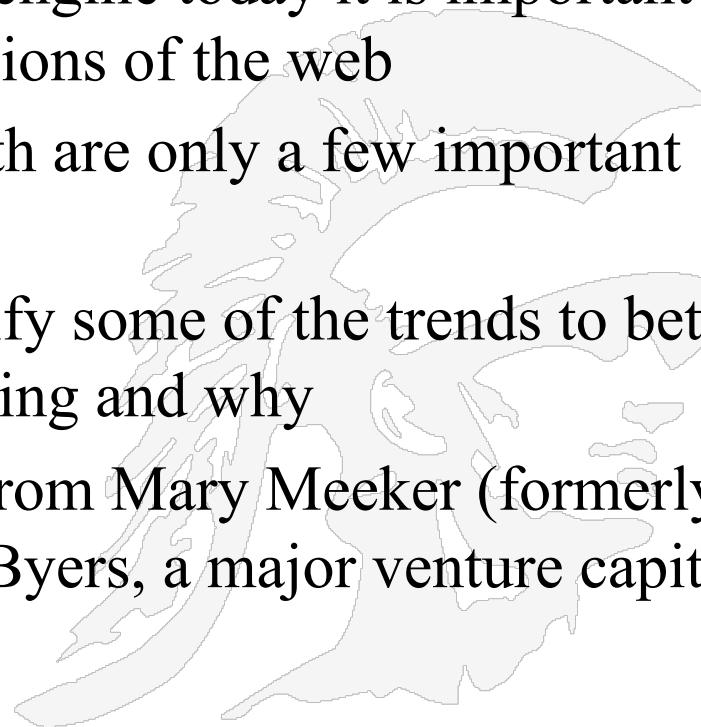
- The search engine industry is 20+ years old, having started with WebCrawler and Lycos in 1994 who sold banner ads as their business model
- Search engine revenue today
 - **Google:** 2020: \$181 billion; 2019: \$162 billion; 2018: \$116 Billion; 2017: \$109 Billion; 2016: \$90 Billion; 2015: \$74.5 Billion; 2014: \$66 Billion; 2013: \$37 Billion
 - **Baidu:** 2020: \$16.4 billion; 2019: \$15 billion; : \$11.3 Billion; 2017: \$13 Billion; 2016: \$10.1 Billion; 2015: \$10.2 Billion; 2014: 8.0 Billion
 - **Yahoo:** 2019: 6.97Billion; 2018: 3.03 Billion; 2017: 3.0 Billion; 2016: 2.98 billion; 2015: \$4.9 Billion; 2014: 4.6 Billion; 2013: 4.6Billion
 - **Bing:** 2020 \$1.6 Billion
 - Microsoft says that in Q1 2016 Bing became profitable



Web Trends and Measurements



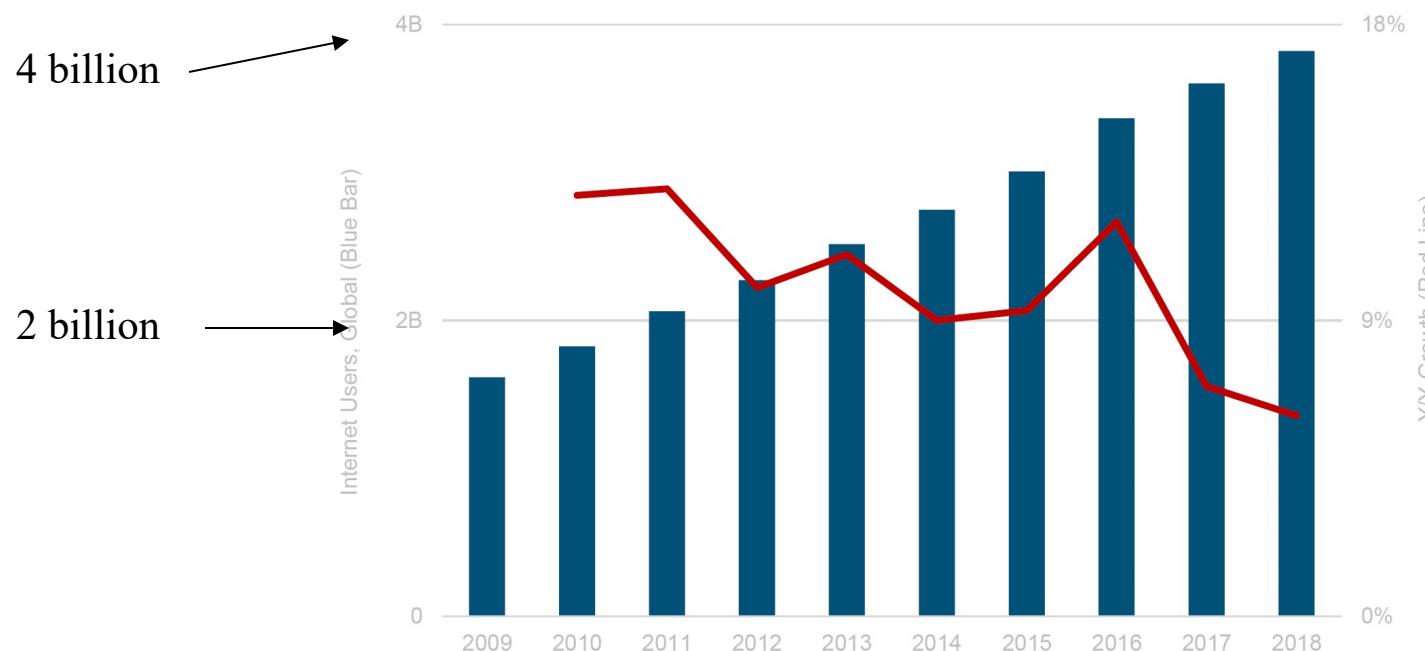
- Web has changed dramatically over the last 30+ years
- If one is building a web search engine today it is important to understand the different dimensions of the web
 - Scale, complexity and growth are only a few important factors
- In today's lecture I try to quantify some of the trends to better understand where the web is going and why
- many of the early slides come from Mary Meeker (formerly of) Kleiner, Perkins, Caufield and Byers, a major venture capital firm, <http://www.kpcb.com/>



Global Internet Users = 3.8B @ 46% Penetration

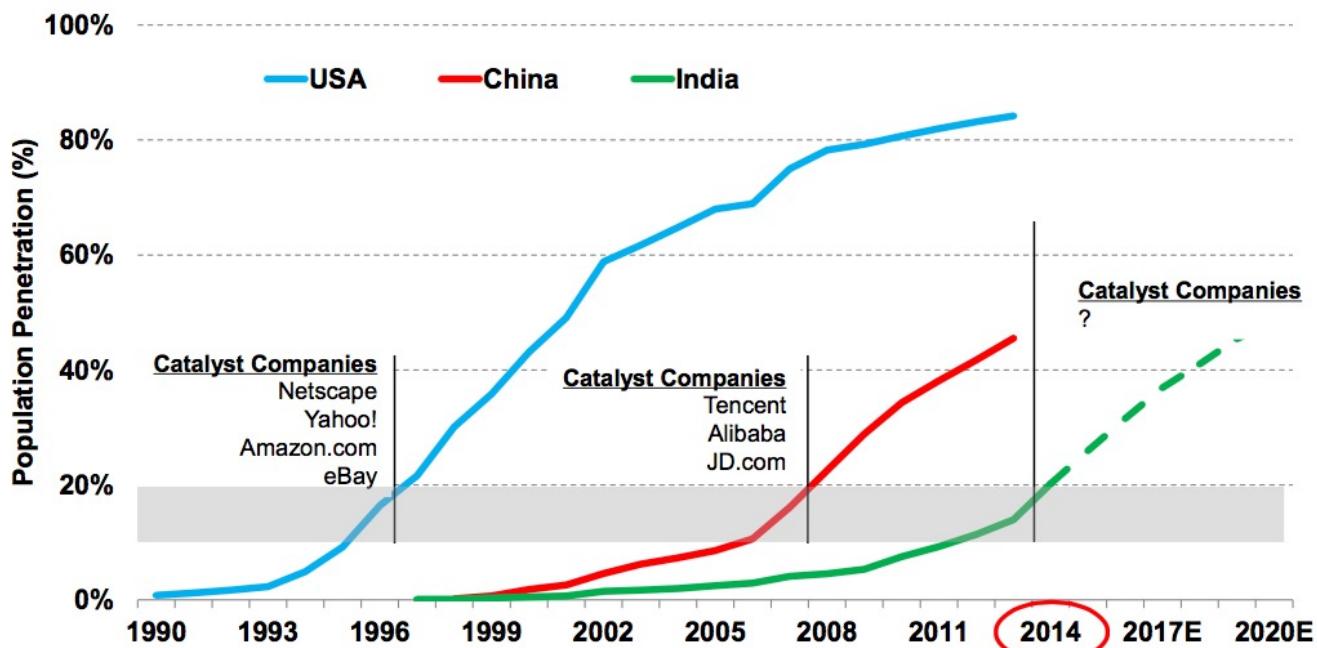
Global Internet User Growth =
Solid But Slowing +6% vs. +7% Y/Y

Internet Users vs. Y/Y Growth



India = Appears to Be @ Internet Penetration Growth Inflection

Internet User Penetration Curve, USA / China / India, 1990 – 2020E



@KPCB

Source: World Bank, Hillhouse Capital forecast for India beyond 2014.

Hillhouse Capital

165

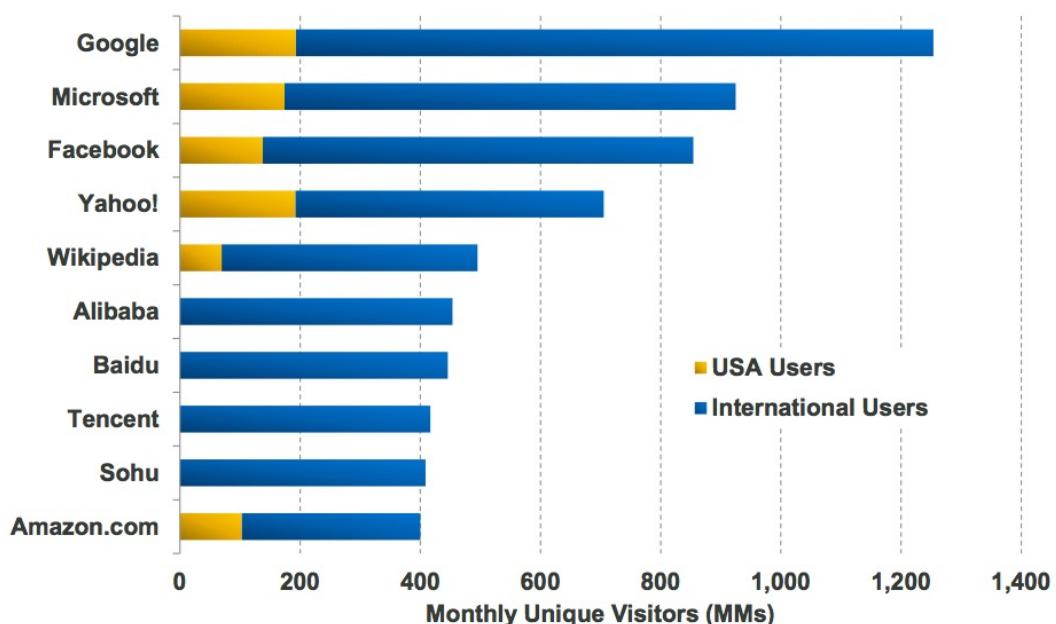
The US leads in the development of highly popular Internet websites;

Baidu is a Chinese search engine

*Tencent is a Chinese holding company of Internet properties, among the most popular being, QQ, for chatting;
 Sohu.com Inc. is a Chinese online media, search, gaming, community and mobile service group.*

**3/14 – 6 of Top 10 Global Internet Properties ‘Made in USA’...
 >86% of Their Users Outside America...China Rising Fast**

Top 10 Internet Properties by Global Monthly Unique Visitors, 3/14



@KPCB Source: comScore, 3/14.

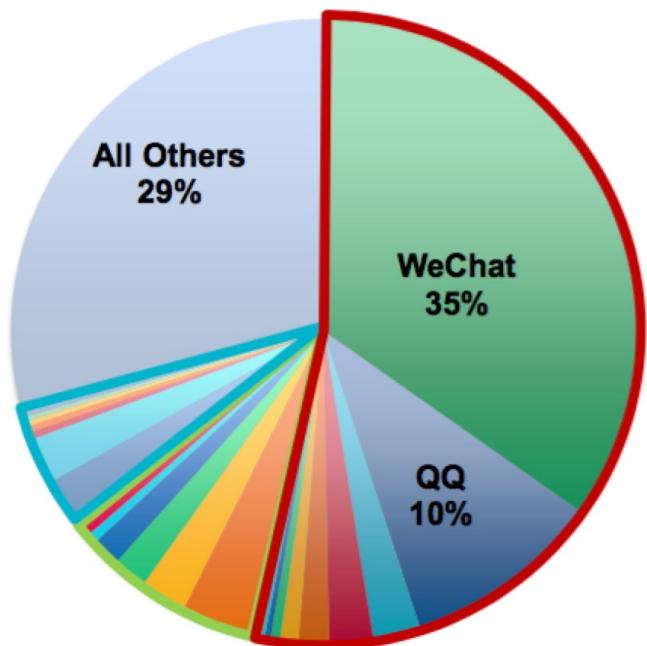
131



China Mobile Internet Usage Leaders...

Tencent + Alibaba + Baidu = 71% of Mobile Time Spent

Share of Mobile Time Spent, April 2016
Daily Mobile Time Spent = ~200 Minutes per User, Average



Tencent

Alibaba

Baidu

- WeChat
- QQ
- QQ Browser
- Tencent Video
- Tencent News
- Tencent Games
- QQ Music
- JD.com
- QQ Reading

- UCWeb Browser
- Taobao
- Weibo
- YouKu Video
- Momo
- Shuqi Novel
- AliPay
- AutoNavi

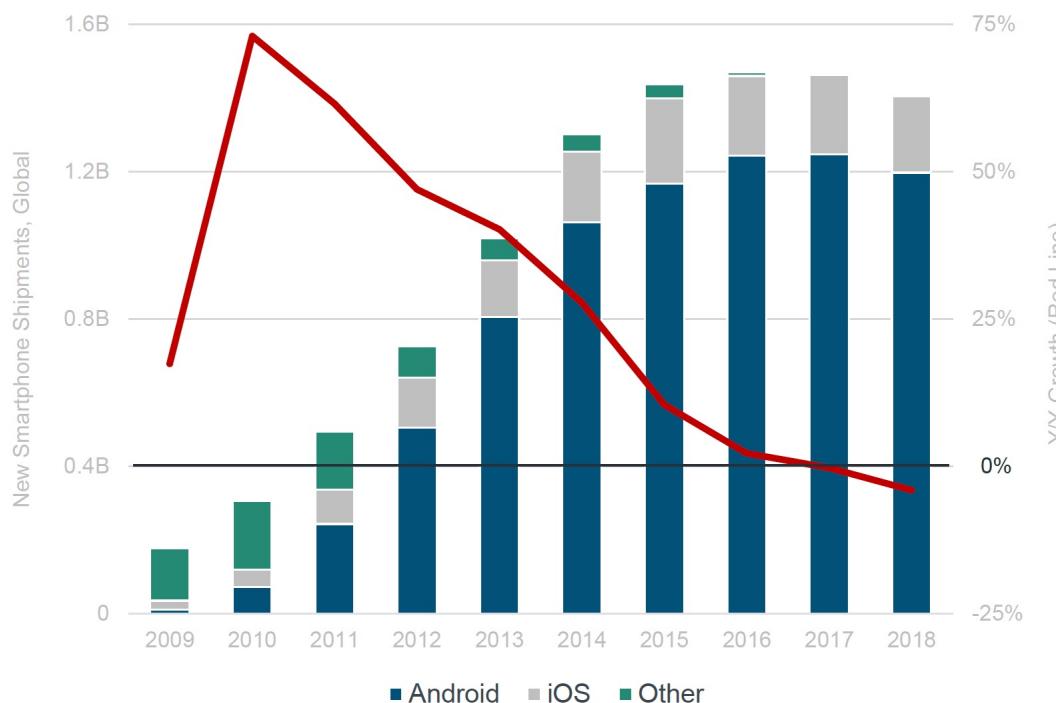
- Mobile Baidu
- iQiyi / PPS Video
- Baidu Browser
- Baidu Tieba
- 91 Desktop
- Baidu Maps
- All Other

Shipments by Operating System

Android and iOS

Global New Smartphone Unit Shipments =
Declining -4% vs. 0% Y/Y

New Smartphone Unit Shipments vs. Y/Y Growth



World's Content is Increasingly Findable + Shared + Tagged - Digital Info Created + Shared up 9x in Five Years

There has been exponential growth in online information;

1 Zettabyte = 1,024 Exabytes

1 Exabyte = 1,024 Petabytes

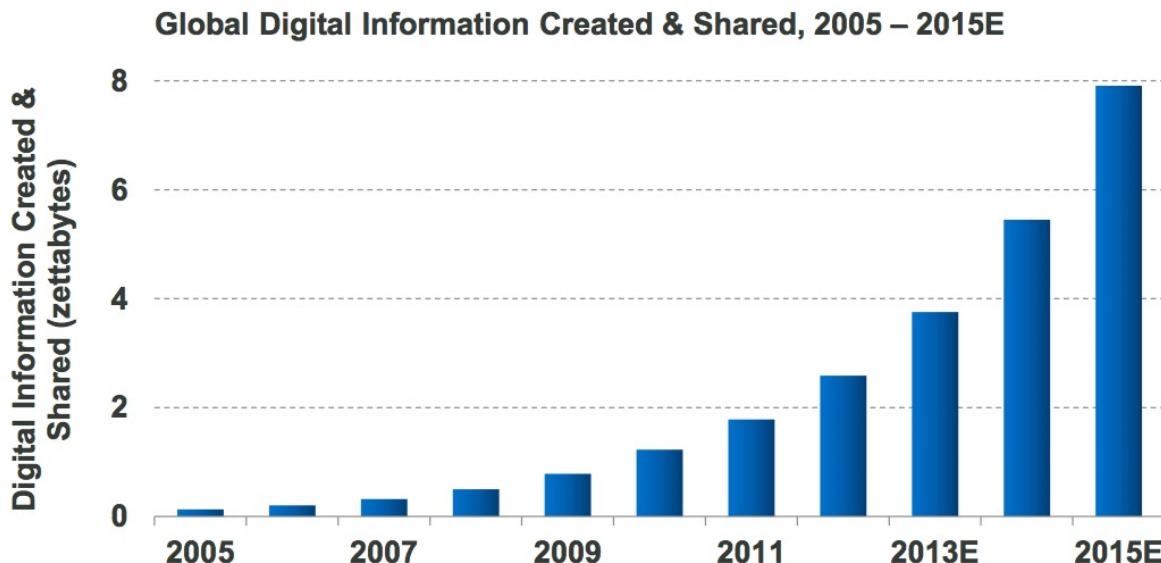
1 Petabyte = 1,024 Terabytes

1 Terabyte = 1,024 Gigabytes

or

1 Zettabyte = 1,000,000,000,000
gigabytes

*Amount of global digital information created & shared
– from documents to pictures to tweets –
grew 9x in five years to nearly 2 zettabytes* in 2011, per IDC.*



KPCB

Note: * 1 zettabyte = 1 trillion gigabytes. Source: IDC report "Extracting Value from Chaos" 6/11. 11



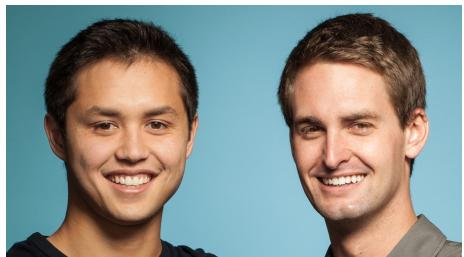
Photos Alone = 1.8B+ Uploaded & Shared Per Day... Growth Remains Robust as New Real-Time Platforms Emerge

500 million photos are uploaded every day and that number is doubling every year

Yahoo has recently made a major upgrade to **Flickr**

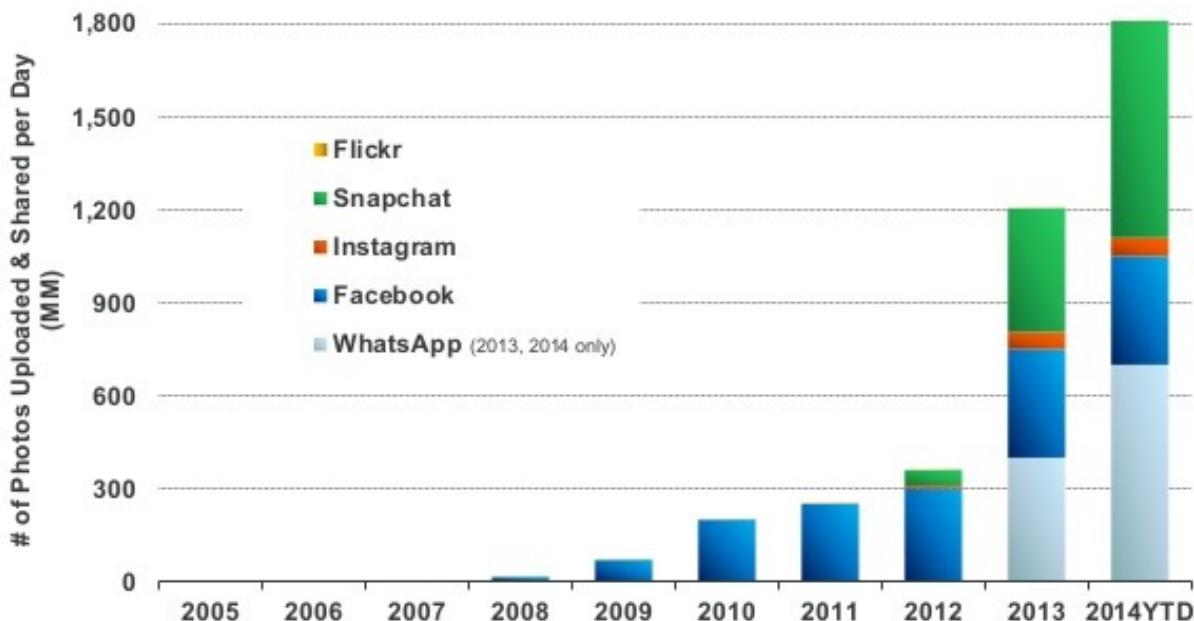
Instagram was in 2010 purchased by Facebook for \$1 billion

Snapchat is a photo messaging application developed by two Stanford students (\$9B valuation);



bobby Murphy - Evan Spiegel

Daily Number of Photos Uploaded & Shared on Select Platforms,
2005 – 2014YTD



@KPCB

Source: KPCB estimates based on publicly disclosed company data. 2014 YTD data per latest as of 5/14.

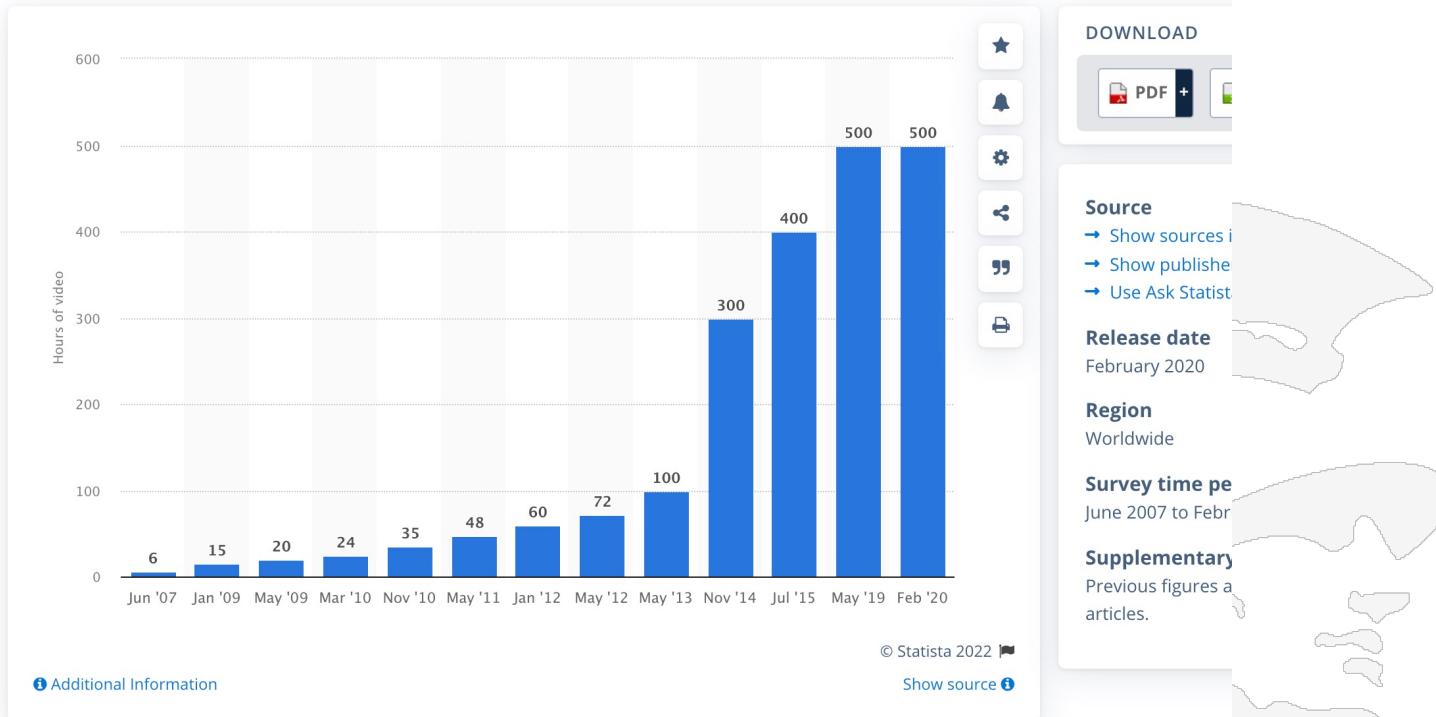
62



Content Uploaded to YouTube

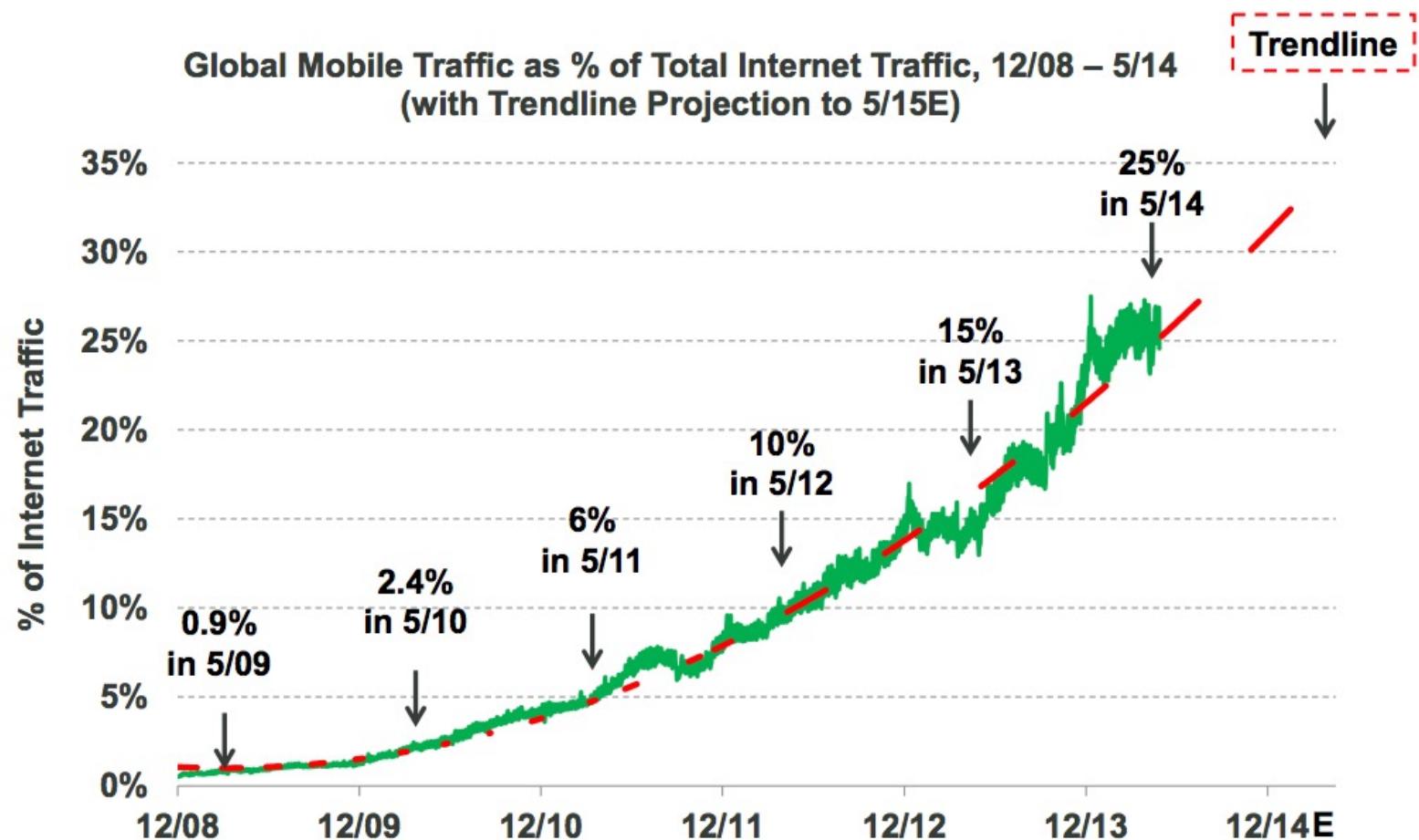
Internet > Online Video & Entertainment

Hours of video uploaded to YouTube every minute as of February 2020

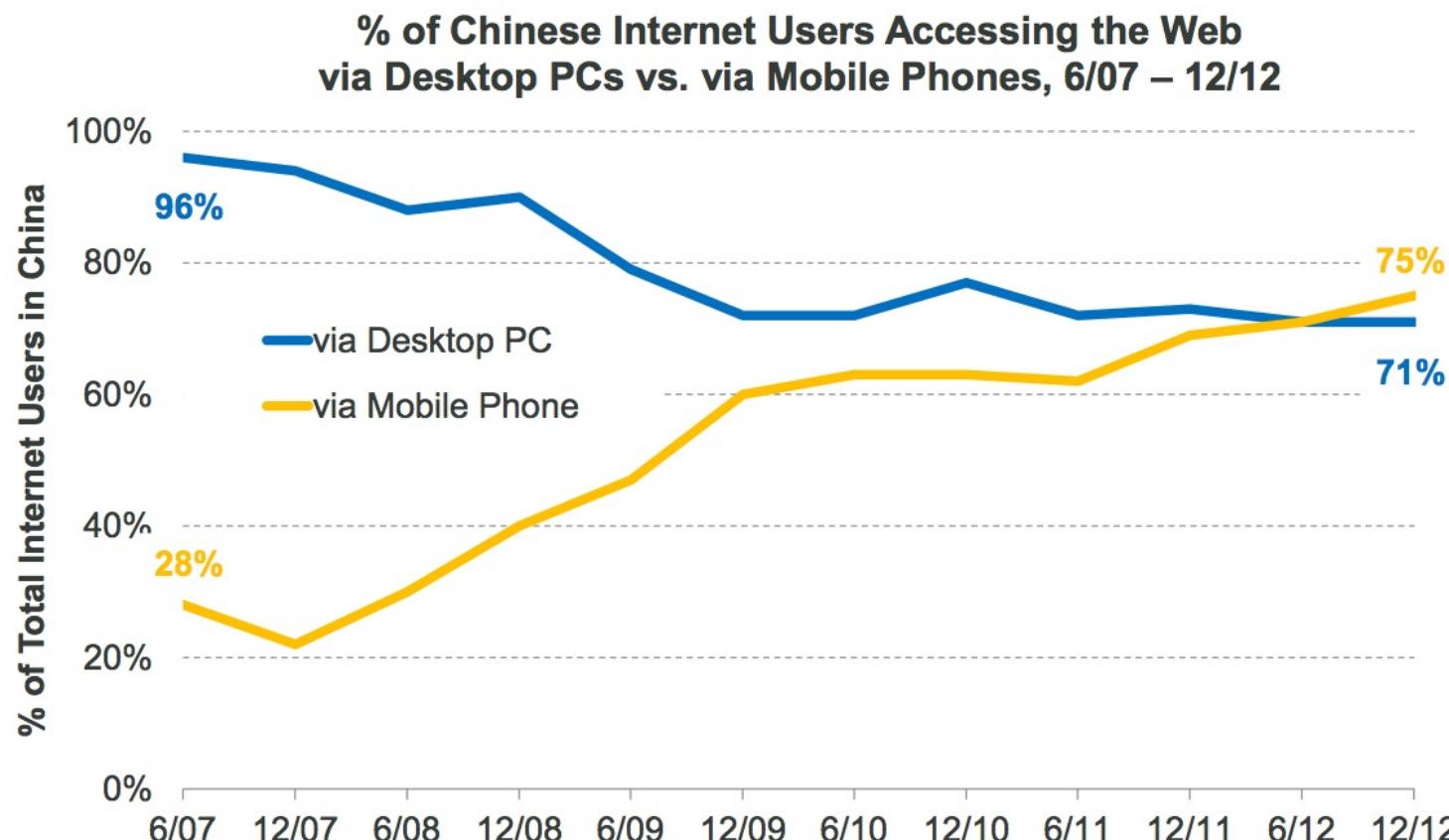


- As of February 2020, more than 500 hours of video were uploaded to YouTube every minute.
- This equates to approximately 30,000 hours of newly uploaded content per hour.
- The number of video content hours uploaded every 60 seconds grew by around 40 percent between 2014 and 2020.

Mobile Traffic as % of Global Internet Traffic = Growing >1.5x per Year & Likely to Maintain Trajectory or Accelerate



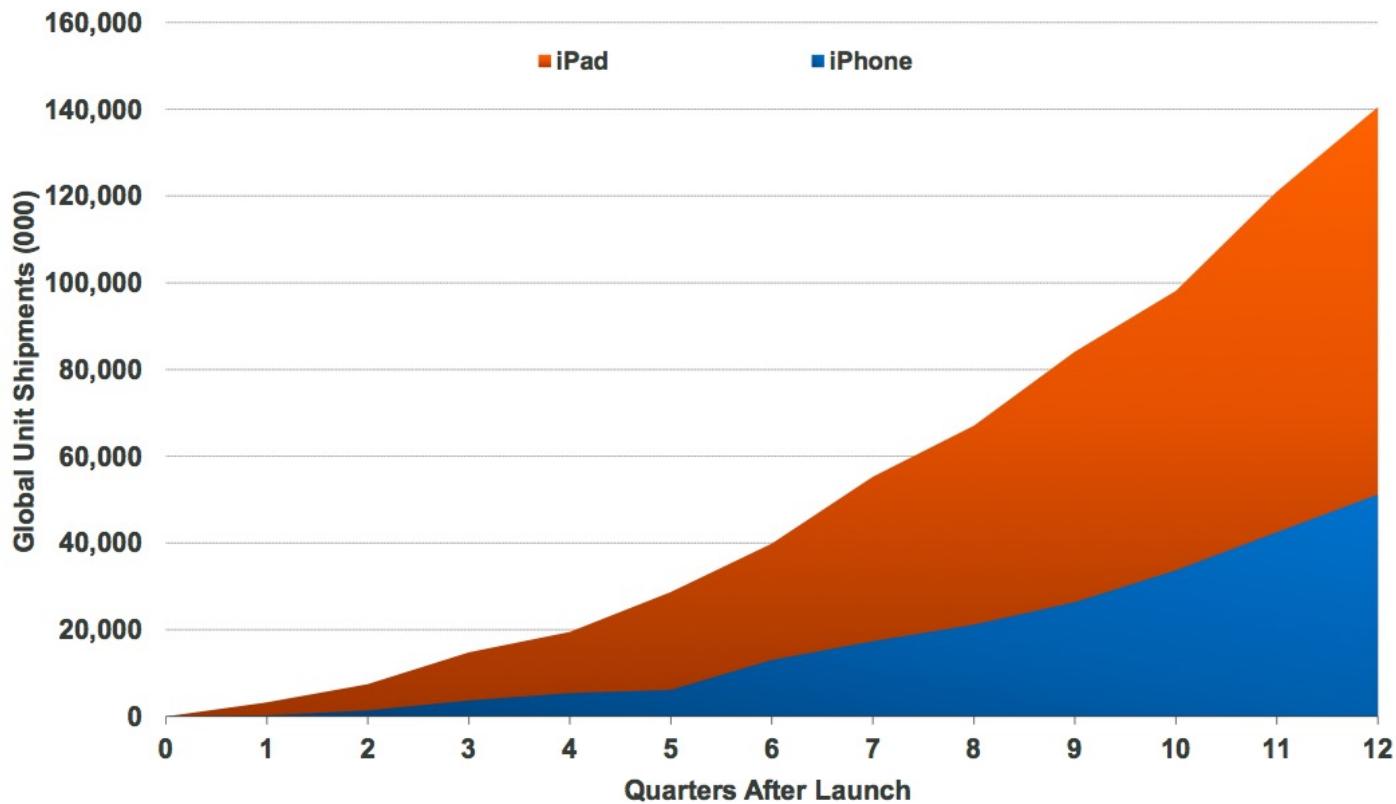
China – Mobile Internet Access Surpassed PC, Q2:12

**KPCB**

Source: CNNIC, 1/13. 33

Tablet Growth = More Rapid than Smartphones, iPad = ~3x iPhone Growth

First 12 Quarters Cumulative Unit Shipments, iPhone vs. iPad



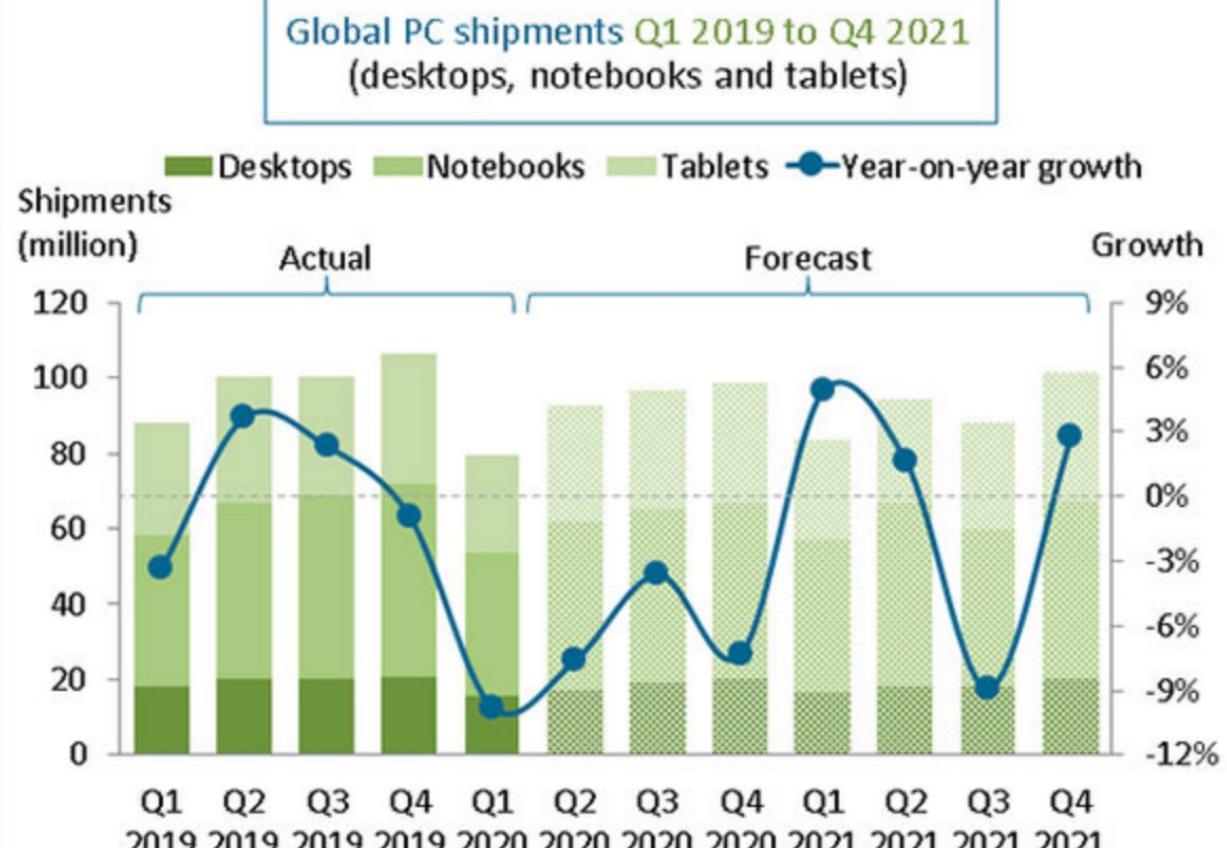
Source: Apple, as of CQ1:13 (12 quarters post iPad launch).
Launch Dates: iPhone (6/29/07), iPad (4/3/10).

44

KPCB

Worldwide PC and tablet shipments will fall 7% in 2020, due to a recession caused by the coronavirus.

A return to stability is expected in 2021.



Technology Cycles – Still Early Cycle on Smartphones + Tablets, Now Wearables Coming on Strong, Faster than Typical 10-Year Cycle

Technology Cycles Have Tended to Last Ten Years

Mainframe Computing
1960s



Mini Computing
1970s



Personal Computing
1980s



Desktop Internet Computing
1990s



Mobile Internet Computing
2000s



Wearable / Everywhere Computing
2014+



Others?

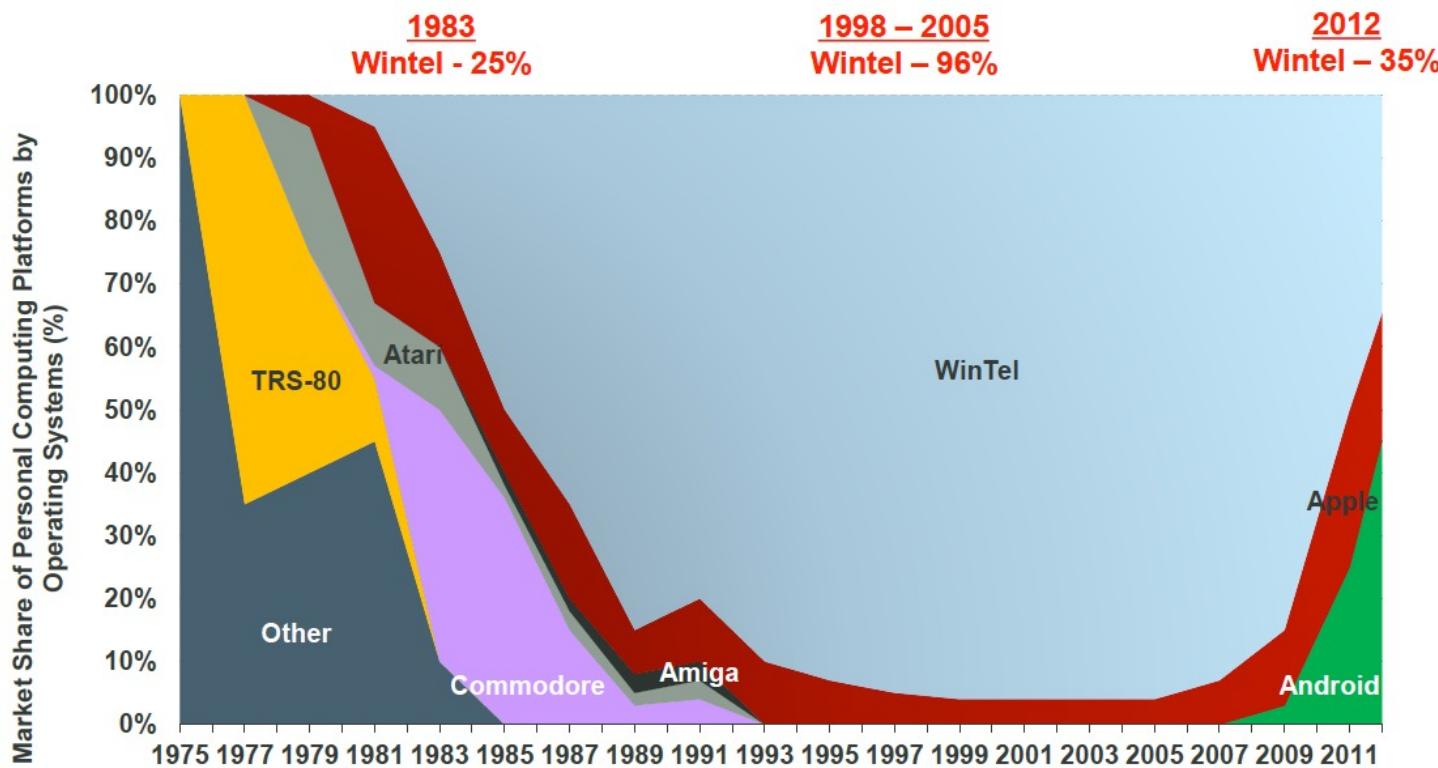
KPCB

49

Image Source: Computersciencelab.com, Wikipedia, IBM, Apple, Google, NTT docomo, Google, Jawbone, Pebble.

Re-Imagination of Computing Operating Systems - iOS + Android = 60% Share vs. 35% for Windows

Global Market Share of Personal Computing Platforms by Operating System Shipments, 1975 – 2012

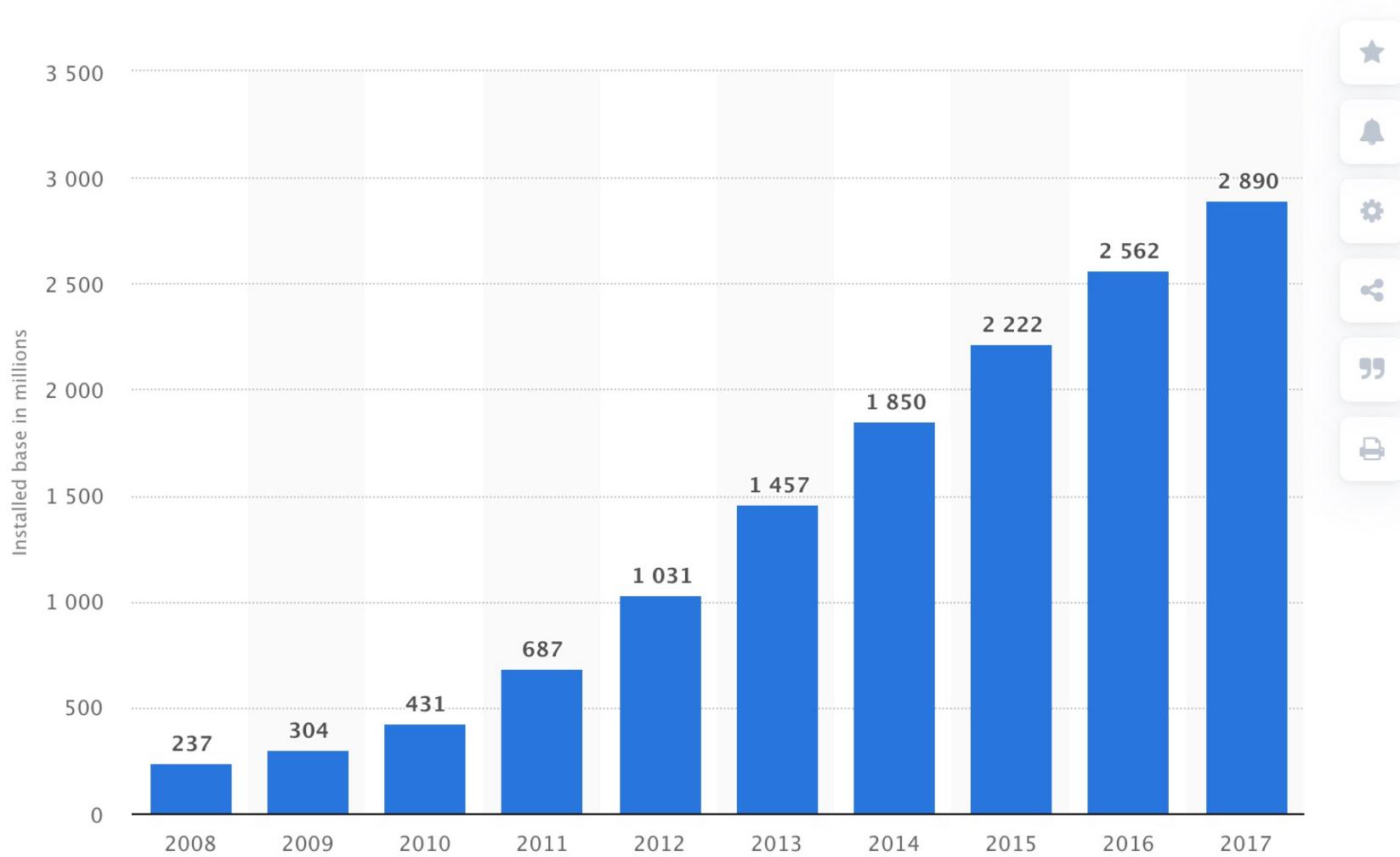


KPCB

Source: Asymco.com (as of 2011), Public Filings, Morgan Stanley Research, Gartner for 2012 data.

109

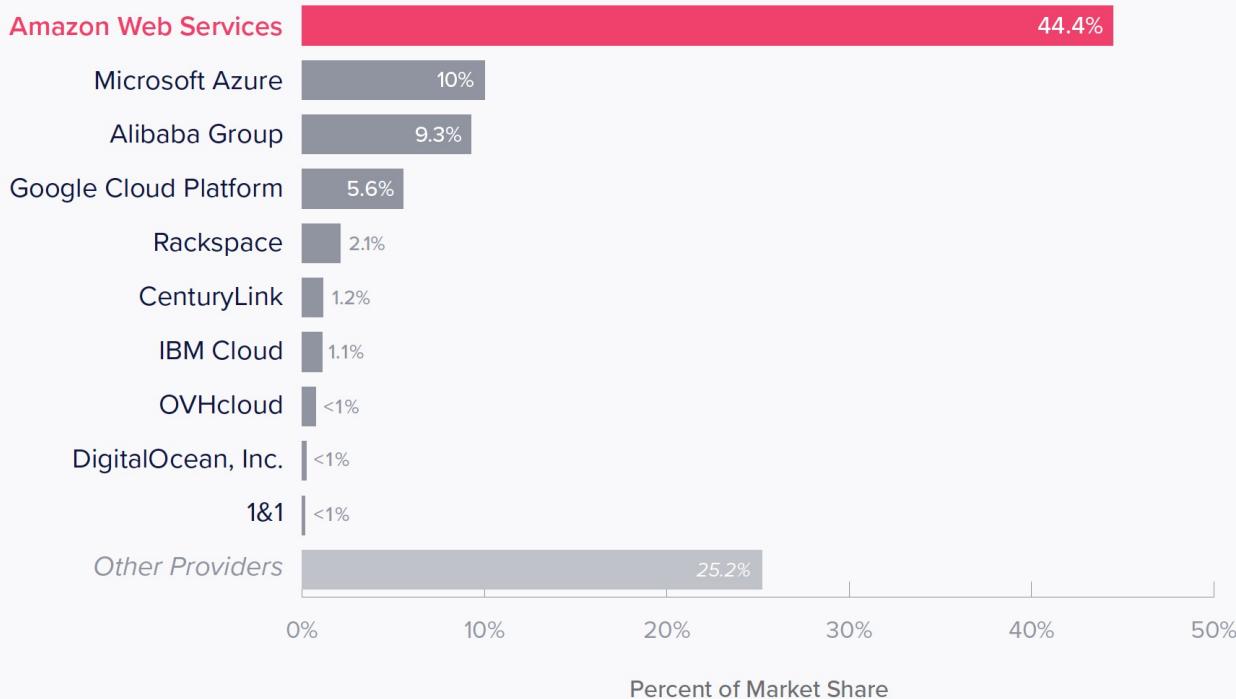
Smartphones worldwide installed base from 2008 to 2017 *(in millions)*



Major Cloud Providers

Market Share Of Leading Cloud Hosting Providers

Top 10 Providers by Total 2020 Market Share

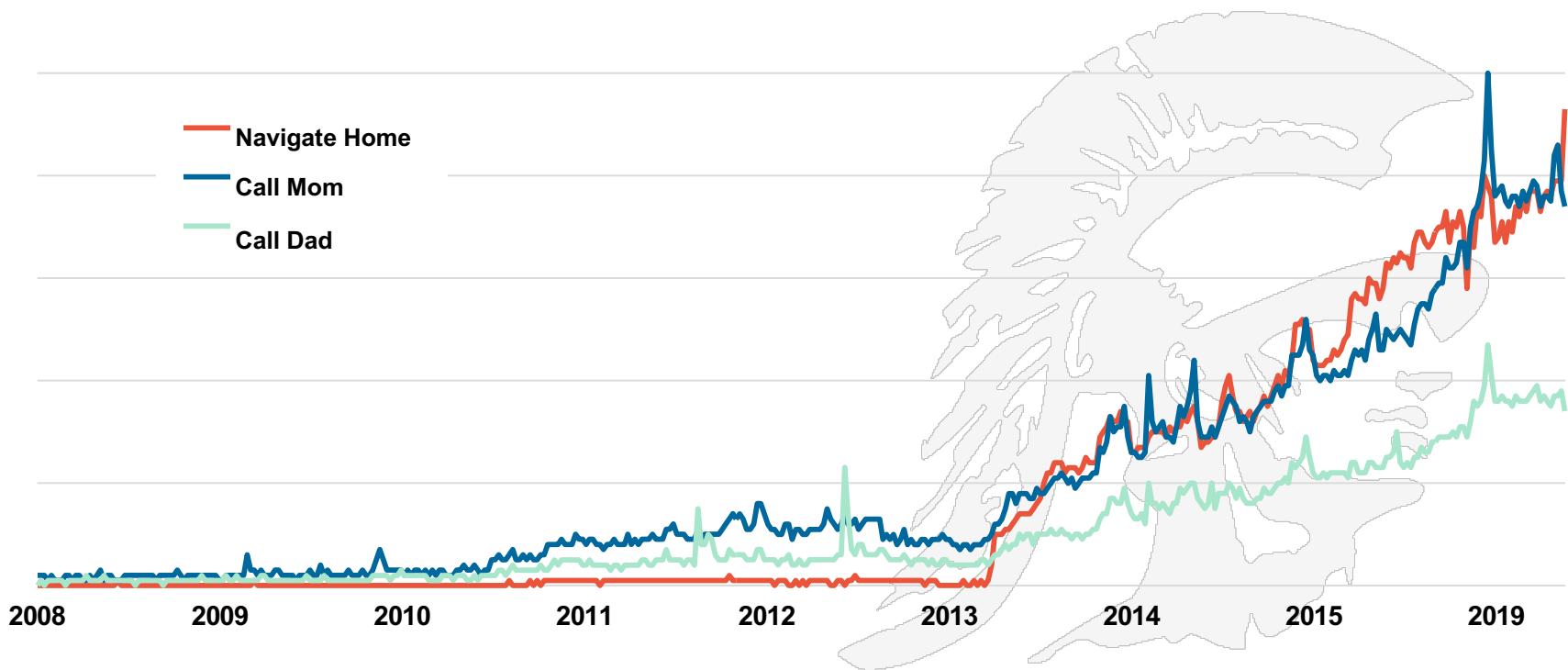


Source: Intricately data, April 2021

Google Voice Search Queries

Google Trends imply queries associated with voice-related commands have risen >35x since 2008 after launch of iPhone & Google Voice Search

Google Trends, Worldwide, 2008 – 2019



Summary of Recent Trends in Web/Internet Development

- Growth in number of users connected
- Growth in Smartphone use
- Growth in digital data, especially photos and video
- Growth in Social Media as an advertising platform
- Transition from desktop/laptop use to mobile
- Growth in tablet usage over desktops/laptops
- Decreased dominance of Microsoft Windows
- Move away from server farms to cloud computing
- Growth in voice communication with devices

Measuring the Web

- The World Wide Web (the Web, the publicly accessible web) is so dynamic it is hard to describe it and have the description be valid for very long
- In this lecture we look at what is known,
 - Measuring the Web by number of web sites
 - Measuring the Web by the Languages of Web Pages
 - Measuring the Web by Rate of Change of Pages
 - Measuring the Web by Document Content Type
 - Measuring the Web by linkage
 - Measuring the Web as a Graph
 - Measuring the Web by Content
 - (using the best statistics we can find)



Number of Websites

Jan. 2020:

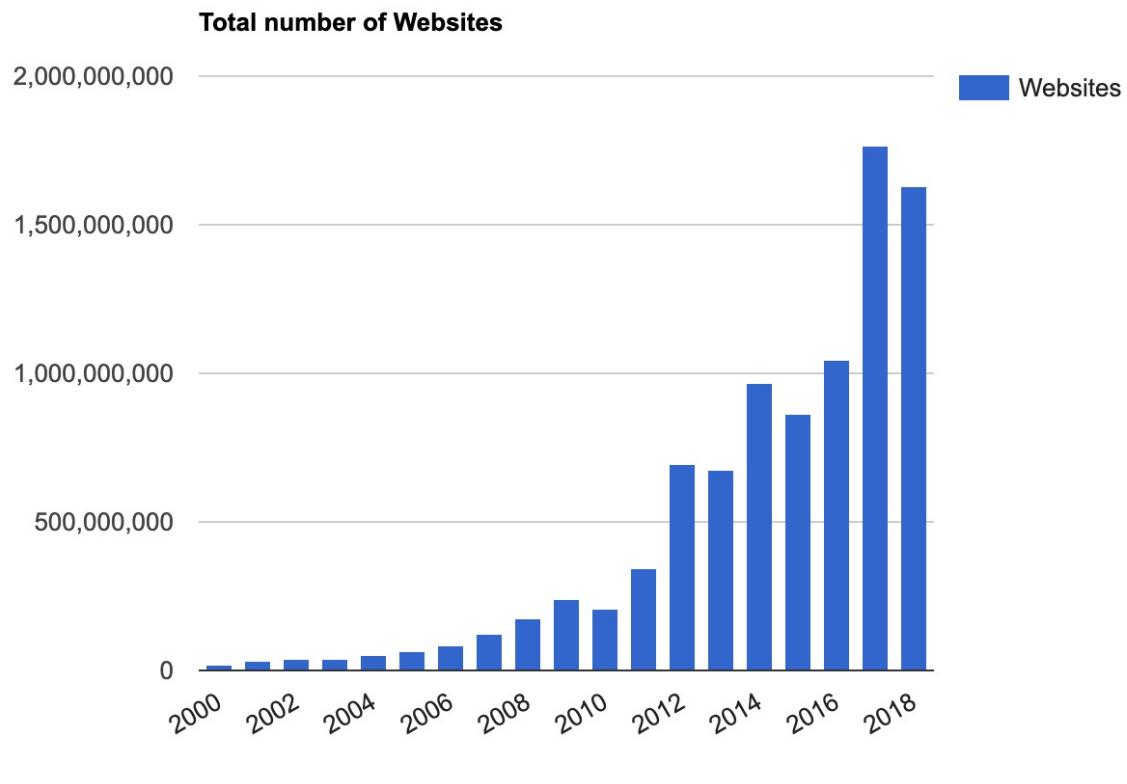
~1.7 Billion sites

nginx web server
had the largest
growth;

Over 50% of websites
Are hosted either by
Apache or nginx;

But Microsoft web
servers still power
43.2% of all sites

Around 75% of websites
are not active, but parked

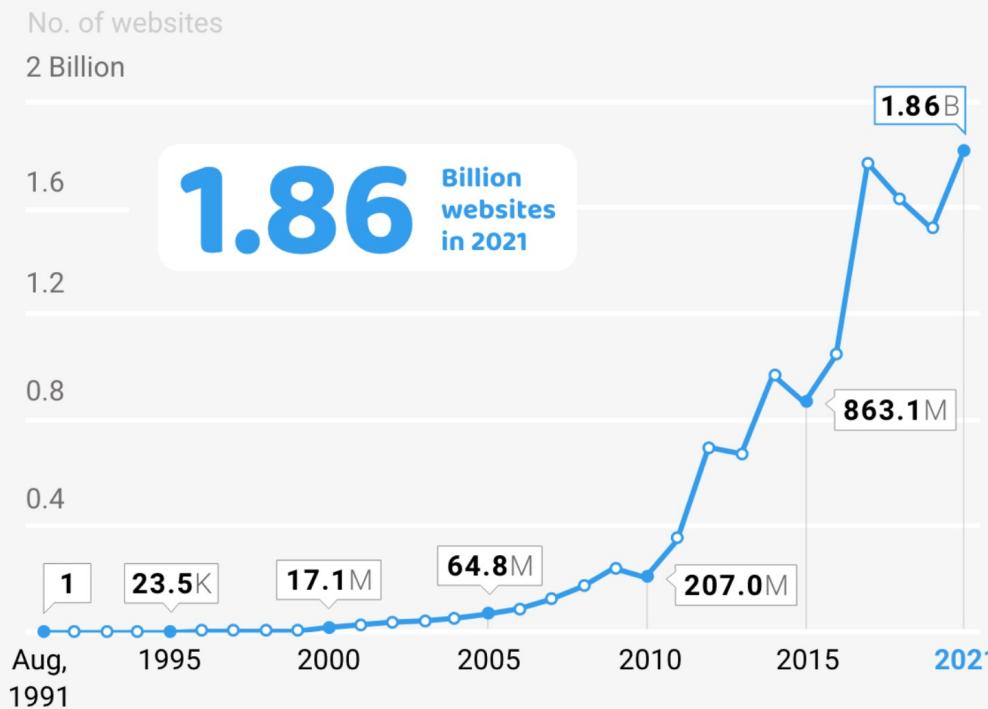


<http://www.internetlivestats.com/total-number-of-websites/>

Number of websites in the world



The global number of websites has more than doubled from 2015 to 2021. Websites growth rate from 1991 to 2021



Distribution Across TLDs

136 million in .com,
 21million in .tk, 14
 million in .de, etc

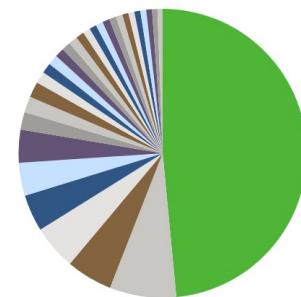
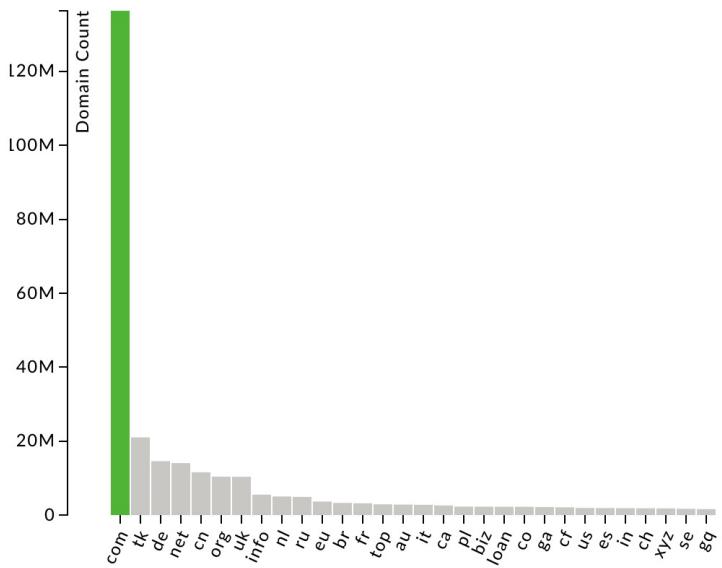
what is .tk and why is
 it so large? (Tokelau)

Domain Count Statistics for TLDs

This page displays the count of all Domains in each TLD. For Registry's publishing a domain count, "Our Count" should closely match their published record. For registry's that don't provide a zone file or publish an up-to-date record, Our Count represents all domains we know about, which is usually more accurate.

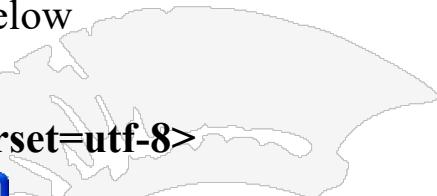
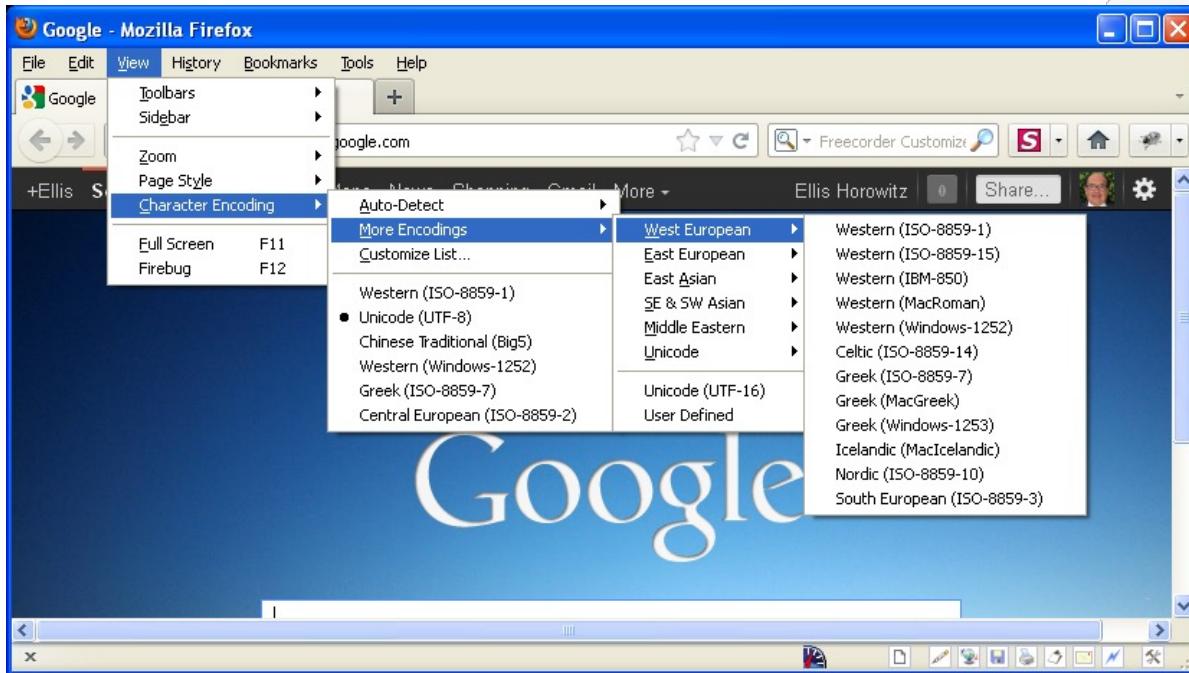
TLD	Our Count
.com	136,346,743
.tk	21,014,704
.de	14,586,920
.net	14,061,971
.cn	11,579,296
.org	10,398,501
.uk	10,361,314
.info	5,528,696
.nl	5,042,167
.ru	4,928,746
.eu	3,673,059
.br	3,282,608
.fr	3,195,248
.top	2,921,081
.au	2,845,979
.it	2,786,780
.ca	2,611,854
.pl	2,302,199
.biz	2,269,454
.loan	2,253,686

Display: Top25 - Default - None



Web Page Language Diversity

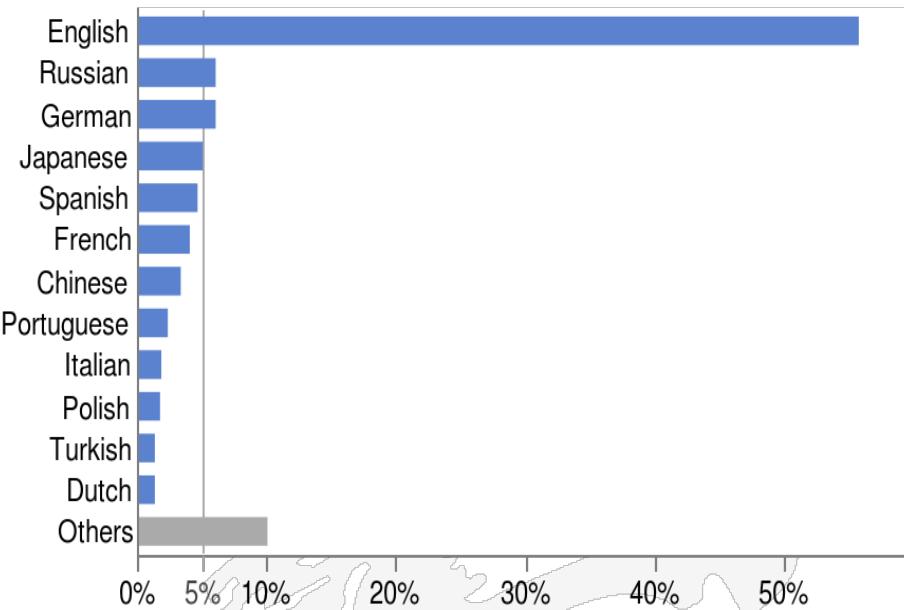
- The Web contains pages in many different languages
- Characters in a language are encoded such that each character is paired with a number
- Unicode and its parallel standard, the ISO/IEC 10646 Universal Character Set, together constitute a modern, unified character encoding.
- Most modern web browsers feature automatic character encoding detection. In Firefox, for example, see the View/Character Encoding submenu, shown below
- In HTML one can specify the character encoding using
- `<meta http-equiv="Content-Type content="text/html" charset=utf-8>`



- If charset is missing ISO-8859-1 is taken as the default unless there is a browser setting;
- Websites in non-western languages typically use UTF-8

Measuring Language Diversity

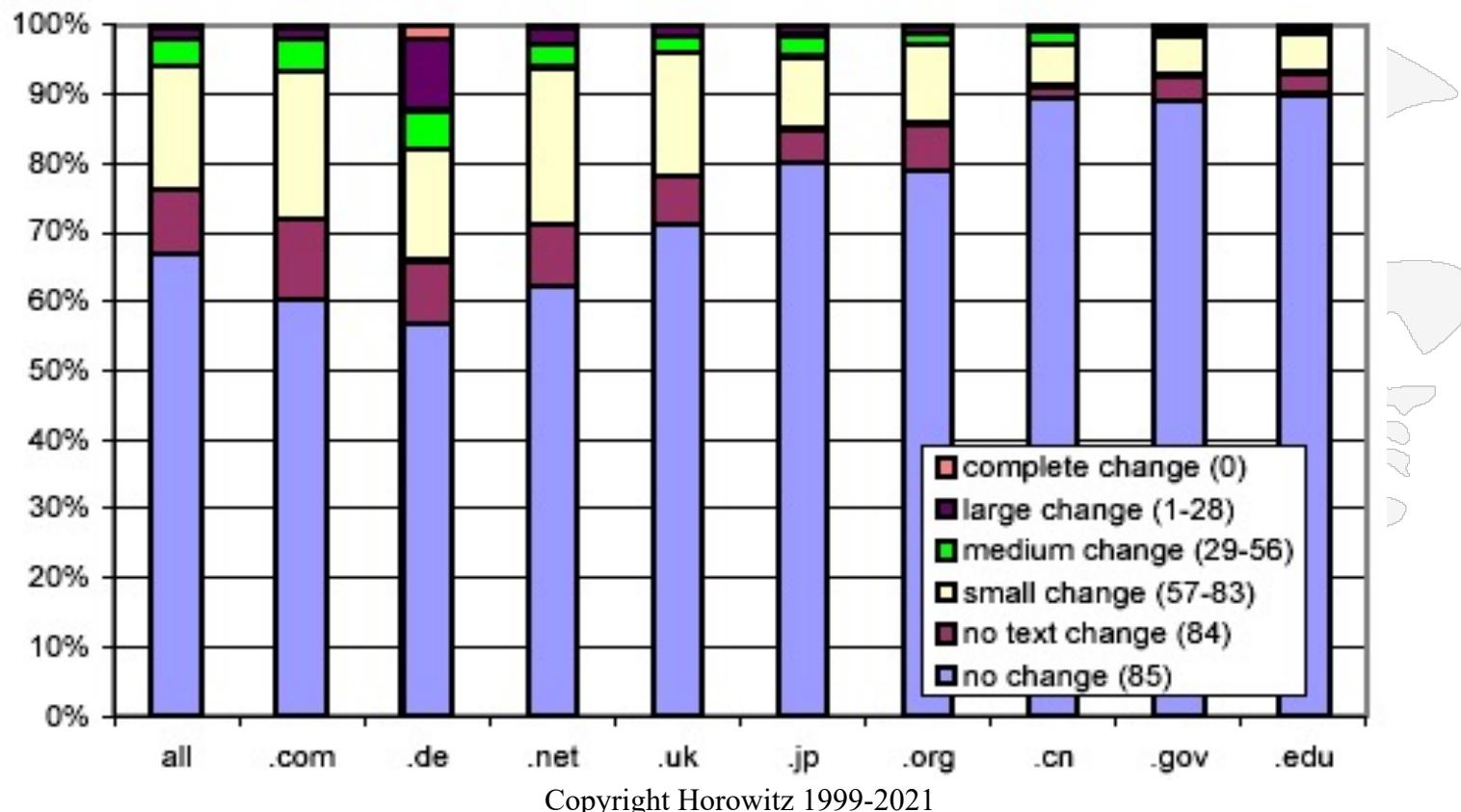
- It is estimated that about 40,000 different languages have been created by human beings
- Only between 6,000-9,000 are still in use
- Study done by the United Nations
 - <http://unesdoc.unesco.org/images/0018/001870/187016e.pdf>
 - The methodology was to examine the pages on a search engine and attempt to identify the primary language in which the page is written
 - Conclusions
 - From 1996 – 2008 English was predominant, occupying roughly 80% of web pages
 - At the same time the number of Internet users who had English as their primary language dropped from 80% to 40%



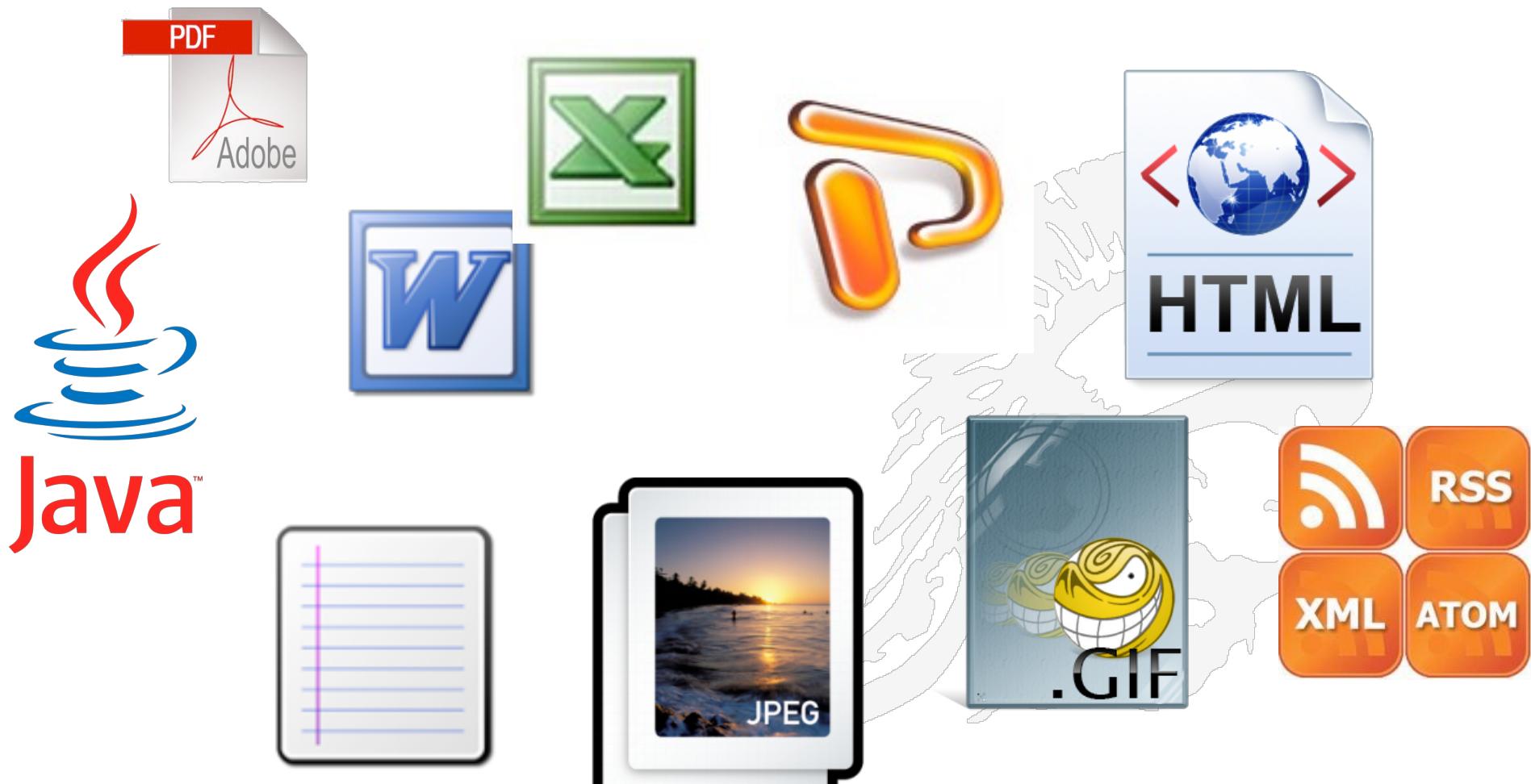
Content languages for websites as of 12 March 2014 [\[2\]](#)

Static Pages: Rate of Change

- Fetterly et al. study several views of data, 150 million pages over 11 weekly crawls
 - Divided into 6 groups by extent of change
- Changes in *.com are more frequent than .gov or .edu

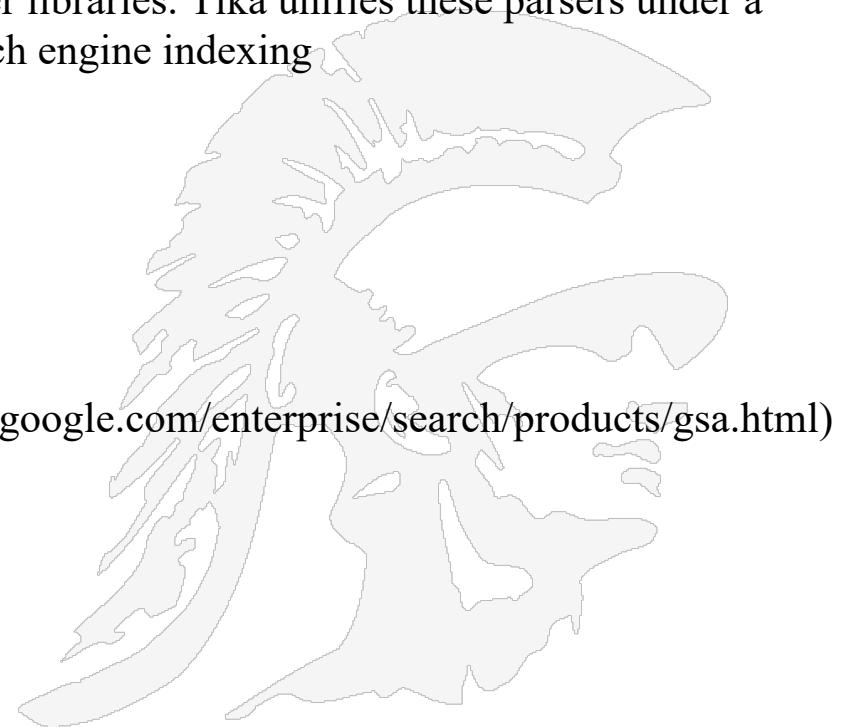


Complexity of Data Types



Proliferation of Content Types Available

- By some accounts, there are 16,000 to 51,000 content types*
- What to do with content types?
 - Parse them
 - How? The Apache Tika™ toolkit detects and extracts metadata and text content from various documents using existing parser libraries. Tika unifies these parsers under a single interface. Tika is useful for search engine indexing
 - Extract their text and structure
 - Index their metadata
 - Use an indexing technology like
 - Lucene, <http://lucene.apache.org/>
 - Solr, <http://lucene.apache.org/solr/>
 - Google Search Appliance (<http://www.google.com/enterprise/search/products/gsa.html>)
 - Identify what language they belong to
 - N-grams



*<http://fileext.com/> (see if you can name the top 20 file extensions)

Content Types Indexed by Google

What file types can Google index

support.google.com/webmasters/bin/answer.py?hl=en&answer=35287

CSCI 572 Home Page | CSCI 571 Home Page | CSCI 351 Home Page | Ellis Horowitz' Home Page

+You Gmail Calendar Documents Photos Sites Search More Sign in

Google Search Webmaster Tools Help home

What file types can Google index?

Learn more About Google Search

Google+ Sitelinks Site title and description Instant Preview Google Basics How are videos ranked? Special Google searches Adding a site to Google Ranking

What file types can Google index?

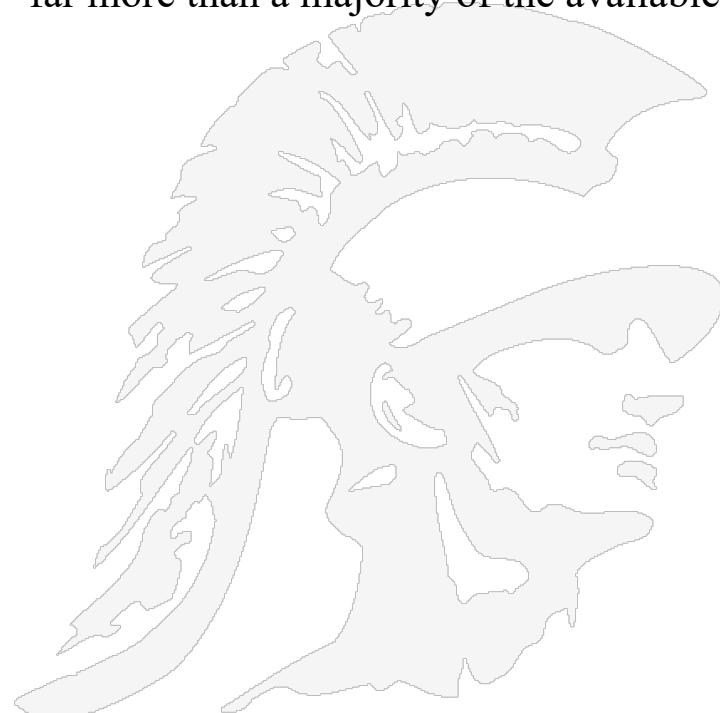
Google can index the content of most types of pages and files. The most common file types we index include:

- Adobe Flash (.swf)
- Adobe Portable Document Format (.pdf)
- Adobe PostScript (.ps)
- Autodesk Design Web Format (.dwf)
- Google Earth (.kmz, .kmv)
- GPS eXchange Format (.gpx)
- Hancom Hanword (.hwp)
- HTML (.htm, .html, other file extensions)
- Microsoft Excel (.xls, .xlsx)
- Microsoft PowerPoint (.ppt, .pptx)
- Microsoft Word (.doc, .docx)
- OpenOffice presentation (.odp)
- OpenOffice spreadsheet (.ods)
- OpenOffice text (.odt)
- Rich Text Format (.rtf, .wri)
- Scalable Vector Graphics (.svg)
- TeX/LaTeX (.tex)
- Text (.txt, .text, other file extensions), including source code in common programming languages:
 - Basic source code (.bas)
 - C/C++ source code (.c, .cc, .cpp, .cxx, .h, .hpp)
 - C# source code (.cs)
 - Java source code (.java)
 - Perl source code (.pl)
 - Python source code (.py)
- Wireless Markup Language (.wml, .wap)
- XML (.xml)

Related

- Flash and other rich media files Content guidelines > Images and video
- Google+ Webmaster FAQ About Google Search > Google+
- Search Engine Optimization (SEO) Google-friendly sites > General guidelines
- Webmaster FAQ Get Started
- Meta tags Google-friendly sites > Content guidelines
- Image Sitemaps Sitemaps > Specialized Sitemaps (Video, images, geo, News, mobile ...)

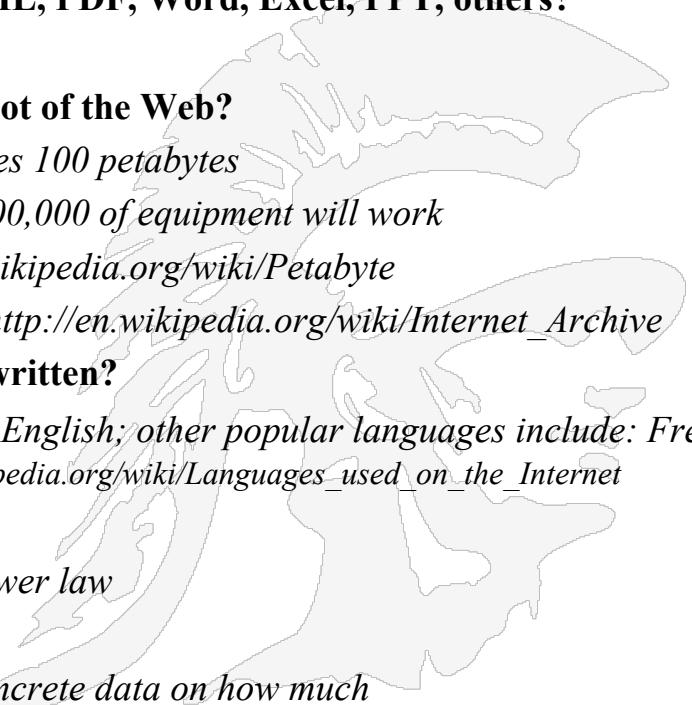
Considering the fact that there are thousands of file types of content stored on the web, Google actually indexes only a small number, less than 3 dozen, but they may well constitute far more than a majority of the available content



<http://support.google.com/webmasters/bin/answer.py?hl=en&answer=35287>

A Summary of Some Web Facts

- **How many websites?** $\sim 1.86 \text{ billion}$
- **How are they distributed across TLDs or across countries?**
 - $112 \text{ million out of } 148 \text{ million belong to .com or about 72\%}$
- **How many web pages are there?** $30 \text{ trillion unique URLs from Google found in 2012,}$
see <http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html>
- **Which content types hold the most information:** HTML, PDF, Word, Excel, PPT, others?
 - *There are thousands of different content types*
- **How much storage is required to hold a single snapshot of the Web?**
 - *1 trillion web pages at 100K bytes per page requires 100 petabytes*
 - *1 petabyte storage costs under \$1,000, so \$100,000 of equipment will work*
 - *Google processes 24 petabytes per day, <http://en.wikipedia.org/wiki/Petabyte>*
 - *The Internet Archive has more than 10 petabytes, http://en.wikipedia.org/wiki/Internet_Archive*
- **What are the languages in which the documents are written?**
 - *According to the Internet Archive, about 55% is in English; other popular languages include: French, German, Spanish and Chinese, also see http://en.wikipedia.org/wiki/Languages_used_on_the_Internet*
- **General properties of the Web graph**
 - *In-degree and out-degree distribution follows a power law*
- **Categories of Content: pornography, spam, mirrors**
 - *Presumably there is a lot of the above, but little concrete data on how much*





Manual Hierarchical Web Taxonomies

The screenshot shows a Mozilla Firefox browser window with the title bar "european cars - Yahoo! Directory Search Results - Mozilla Firefox". The menu bar includes File, Edit, View, History, Bookmarks, Tools, and Help. The toolbar features standard icons for Back, Forward, Stop, Refresh, and Home, along with a search button and a Freecorder Custom icon. The address bar displays the URL "dir.search.yahoo.com/search;_ylt=A0oGdWcnZA5PSgEAu3hXNyo". The main content area shows the "YAHOO! DIRECTORY" logo and a search bar with the query "european cars". Below the search bar, it says "190 results". On the left, there's a "FILTER" section with "Show All" selected, followed by categories: Regional (93), Business and Economy (79), Recreation (11), and Arts (3). Under "FILTER BY TIME", "Any time" is selected, with options for Last 3 months, Last 6 months, and Last year. The main results list includes:

- Also try:** [european car parts](#), [european cars for sale](#), [More...](#)
- European Car Sharing**
Umbrella organization for car sharing companies in Europe.
Category: [Business and Economy](#)>[Shopping and Services](#)>[Automotive](#)>[Car Sharing](#)
www.carsharing.org
- European New Car Assessment Programme (EuroNCAP)**
Aims to provide motoring consumers with a realistic and independent assessment of the safety performance of cars sold in Europe.
Category: [Recreation](#)>[Automotive](#)>[Driving](#)>[Safety](#)>[Organizations](#)
www.euroncap.com
- European Car Free Day**
Take place September 22, 2000, to protest problems of urban mobility, air pollution, and noise.
Category: [Recreation](#)>[Travel](#)>[Transportation](#)>[Auto-Free Transportation](#)>[Organizations](#)
www.22september.org
- ClassicDriver.com - The European Car Webzine**
Focuses on prestige marques, includes articles, web broadcasting, screen savers, dealer guide, and more.
Category: [Recreation](#)>[Automotive](#)>[News and Media](#)>[Magazines](#)
www.classicdriver.com

- **Yahoo** originally used human editors to assemble a large hierarchically structured directory of web pages.

http://www.yahoo.com/

Yahoo still retains the hierarchy as seen to the left; under european cars we see categories: regional with 93 matches, business & economy with 79 matches, etc

Open Directory Project

ODP - Open Directory Project - Mozilla Firefox

File Edit View History Bookmarks Tools Help

ODP - Open Directory Project +

www.dmoz.org  Freecorder Custom 

d m o z open directory project In partnership with AOL Search.

[about dmoz](#) [dmoz blog](#) [suggest URL](#) [help](#) [link](#) [editor login](#)

[advanced](#)

Arts	Business	Computers
Movies, Television, Music...	Jobs, Real Estate, Investing...	Internet, Software, Hardware...
Games	Health	Home
Video Games, RPGs, Gambling...	Fitness, Medicine, Alternative...	Family, Consumers, Cooking...
Kids and Teens	News	Recreation
Arts, School Time, Teen Life...	Media, Newspapers, Weather...	Travel, Food, Outdoors, Humor...
Reference	Regional	Science
Maps, Education, Libraries...	US, Canada, UK, Europe...	Biology, Psychology, Physics...
Shopping	Society	Sports
Clothing, Food, Gifts...	People, Religion, Issues...	Baseball, Soccer, Basketball...
World		
Català, Dansk, Deutsch, Español, Français, Italiano, 日本語, Nederlands, Polski, Pycckий, Svenska...		

[Become an Editor](#) | Help build the largest human-edited directory of the web

Copyright © 2012 Netscape

4,976,587 sites - 93,429 editors - over 1,009,376 categories

- **Open Directory Project, known as DMoZ, is an effort to organize the web according to an ontology;**
- **An approach similar to Yahoo's;**
- **Based on the distributed labor of volunteer editors (“net-citizens provide the collective brain”).**
- **Used by most other search engines.**
- **Started by Netscape.**
 - <http://www.dmoz.org/>
- **Distributes its data using RDF format**
- **DMOZ shut down in 2016**

Drilling Down By Category

Open Directory - Science - Mozilla Firefox

File Edit View History Bookmarks Tools Help

Open Directory - Science +

www.dmoz.org/Science/ 

In partnership with AOL Search.

[about dmoz](#) | [dmoz blog](#) | [report abuse/spam](#) | [help](#)

dmoz open directory project

Search the entire directory ▾

Top: Science (104,480) [Description](#)

[A|B|C|D|E|F|G|H|I|J|K|L|M|N|O|P|Q|R|S|T|U|V|W|X|Y|Z]

- [Agriculture \(3,670\)](#)
- [Anomalies and Alternative Science \(480\)](#)
- [Astronomy \(3,595\)](#)
- [Biology \(31,012\)](#)
- [Chemistry \(4,171\)](#)
- [Computer Science \(1,971\)](#)
- [Earth Sciences \(6,098\)](#)
- [Academic Departments \(9\)](#)
- [By Region \(0\)](#)
- [Chats and Forums \(16\)](#)
- [Directories \(27\)](#)
- [Educational Resources \(352\)](#)
- [Employment \(69\)](#)
- [Events \(54\)](#)
- [History \(346\)](#)
- [Instruments and Supplies \(2,370\)](#)
- [Libraries \(88\)](#)
- [Methods and Techniques \(100\)](#)
- [Environment \(6,572\)](#)
- [Math \(9,541\)](#)
- [Physics \(4,281\)](#)
- [Science in Society \(680\)](#)
- [Social Sciences \(19,489\)](#)
- [Technology \(10,560\)](#)
- [Women \(153\)](#)
- [Museums \(479\)](#)
- [News and Media \(234\)](#)
- [Organizations \(131\)](#)
- [People \(0\)](#)
- [Publications \(247\)](#)
- [Reference \(389\)](#)
- [Research Groups and Centers \(57\)](#)
- [Search Engines \(9\)](#)
- [Software \(783\)](#)
- [Weblogs \(117\)](#)

Selecting Category “Science”

Open Directory - Computers: Computer Science - Mozilla Firefox

File Edit View History Bookmarks Tools Help

Open Directory - Computers: Computer Sci... +

www.dmoz.org/Computers/Computer_Science/ 

In partnership with AOL Search.

[about dmoz](#) | [dmoz blog](#) | [report abuse/spam](#) | [help](#)

dmoz open directory project

Search the entire directory ▾

Top: Computers: Computer Science (1,971) [Description](#)

- [Academic Departments \(553\)](#)
- [Conferences \(203\)](#)
- [Directories \(8\)](#)
- [Organizations \(71\)](#)
- [People \(271\)](#)
- [Publications \(80\)](#)
- [Reference \(4\)](#)
- [Research Institutes \(74\)](#)
- [Artificial Intelligence \(1,294\)](#)
- [Artificial Life \(230\)](#)
- [Computational Geometry \(60\)](#)
- [Computer Graphics \(39\)](#)
- [Database Theory \(82\)](#)
- [Distributed Computing \(225\)](#)
- [Parallel Computing \(367\)](#)
- [Software Engineering \(114\)](#)
- [Theoretical \(361\)](#)

See also:

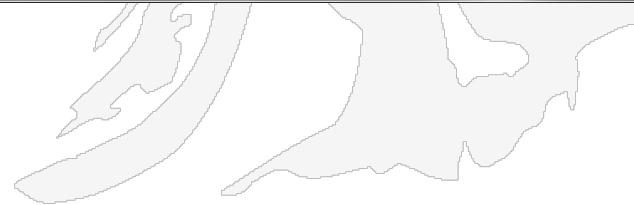
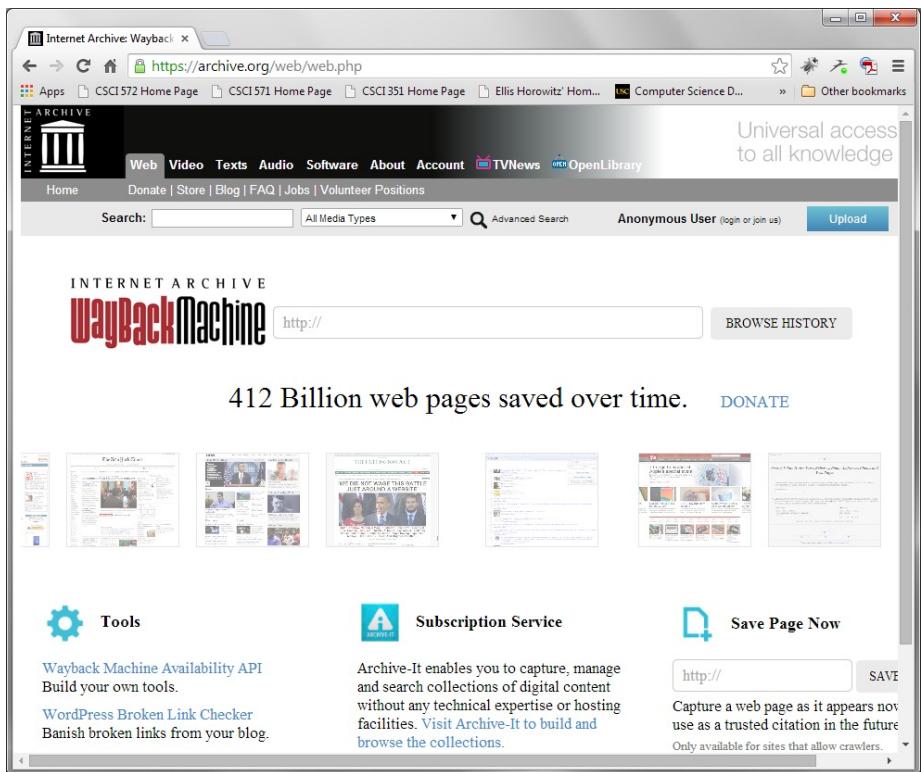
- [Computers: Algorithms \(284\)](#)
- [Computers: Programming \(15,619\)](#)
- [Science: Math \(9,541\)](#)
- [Science: Technology: Electrical Engineering \(237\)](#)

This category in other languages:

Selecting Category “Computer Science”

- The Internet Archive has been taking a snapshot of the World Wide Web every two months since 1997 – has used **Apache Nutch**
- The results are made available through [the Wayback Machine](#),
- Its database is approximately 4.5 petabytes
- The founder is Brewster Kahle
- For the past 13 years, the Internet Archive has been growing rapidly, most recently by about 100TB of data per month.
- Their crawler surveys the web every two months. The algorithm first performs a broad crawl that starts with a few "seed sites," such as Yahoo's directory. After snapping a shot of the home page, it then moves to any referable pages within the site until there are no more pages to capture. If there are any links on those pages, the algorithm automatically opens them and archives that content as well.

Internet Archive



Surface Web

SURFACE WEB



A diagram showing an iceberg floating in blue water. The visible part above the surface is labeled "SURFACE WEB". The submerged part below the surface is labeled "DEEP WEB". The hidden part completely underwater is labeled "(DARK WEB)".

Google

Bing

Wikipedia

Academic Information

Medical Records

Legal Documents

Scientific Reports

Subscription Information

DEEP WEB

*Contains 90% of the information on
the Internet, but is not accessible
by Surface Web crawlers.*

Social Media

Multilingual Databases

Financial Records

Government Resources

Competitor Websites

Organization-specific
Repositories

(DARK WEB)

A part of the Deep Web accessible only through certain browsers such as Tor designed to ensure anonymity. Deep Web Technologies has zero involvement with the Dark Web.

Illegal Information

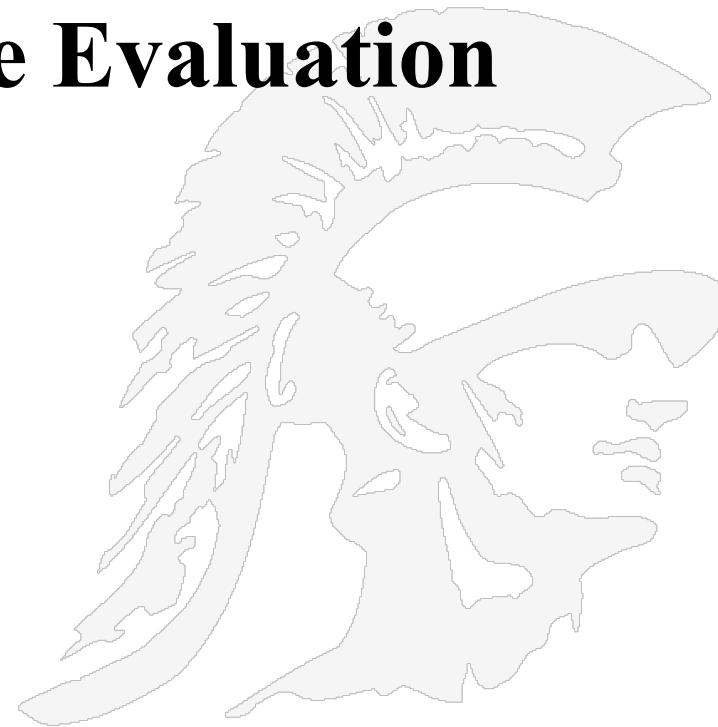
TOR-Encrypted sites

Drug Trafficking sites

Political Protests

Private Communications

Search Engine Evaluation

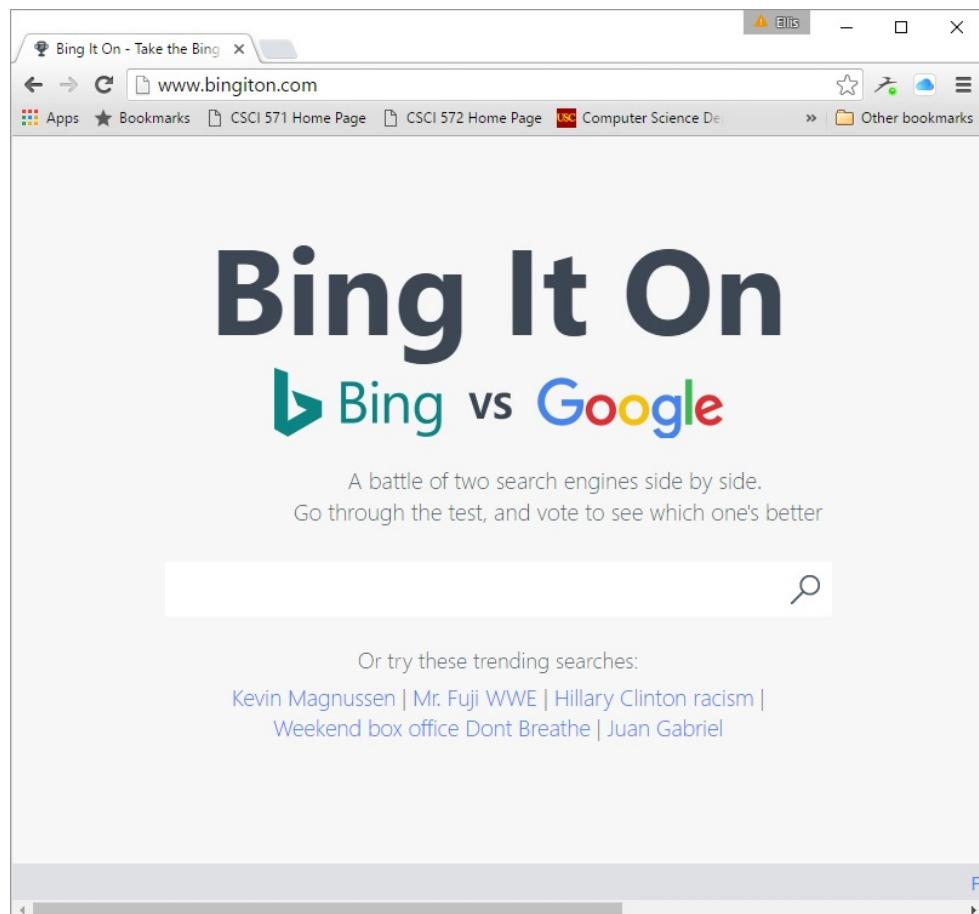


Outline

- **Defining precision/recall**
- **Mean Average Precision**
- **Harmonic Mean and F Measure**
- **Discounted Cumulative Gain**
- **Elements of Good Search Results**
- **Google's Search Quality Guidelines**
- **Using log files for evaluation**
- **A/B Testing**



Comparing Bing and Google



A screenshot of a web browser window titled "Bing It On - Take the Bing". The address bar shows "www.bington.com". The page content features a large title "Bing It On" with a subtitle "Bing vs Google". Below this, a subtext reads "A battle of two search engines side by side. Go through the test, and vote to see which one's better". There is a search bar with a magnifying glass icon. Underneath the search bar, text says "Or try these trending searches:" followed by a list of queries: "Kevin Magnussen | Mr. Fuji WWE | Hillary Clinton racism | Weekend box office Dont Breathe | Juan Gabriel".

- This site is no longer active, but we can simulate the experiment

Try it yourself!

here are some queries:

- ac versus dc current
- best bottled water
- worst hotel in Santa Monica
- how many gears do I need on a bicycle
- Clint Eastwood's best movie

- How do we measure the quality of search engines?
- Precision = #(relevant items retrieved)
divided by
#(all retrieved items)
- Recall = #(relevant items retrieved)
divided by
#(all relevant items)



Formalizing Precision/Recall

	Relevant	Nonrelevant
Retrieved	True positive (tp)	False positive (fp)
Not retrieved	False negative (fn)	True negative (tn)

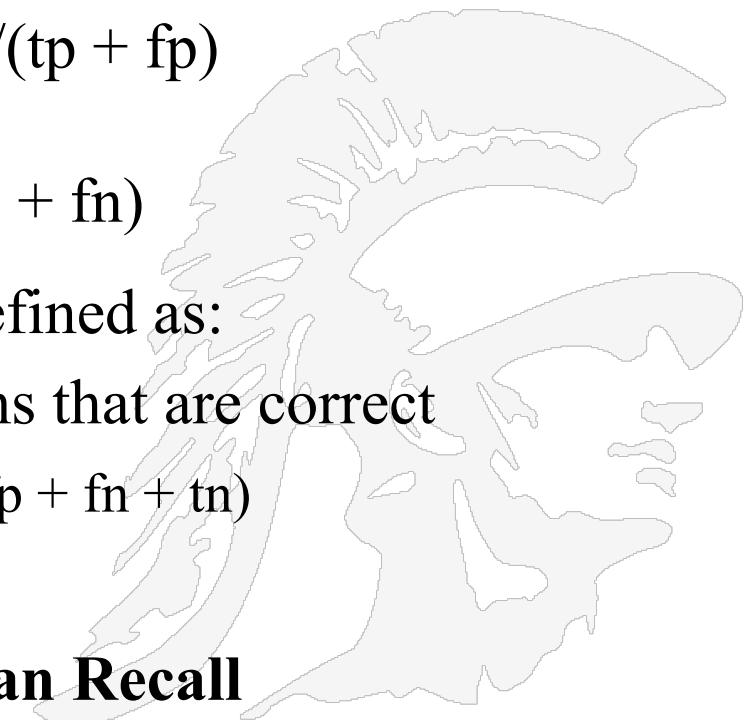
$$\text{Precision} = \text{tp}/(\text{tp} + \text{fp})$$

$$\text{Recall} = \text{tp}/(\text{tp} + \text{fn})$$

- The accuracy of an engine is defined as:
the fraction of these classifications that are correct

$$(\text{tp} + \text{tn}) / (\text{tp} + \text{fp} + \text{fn} + \text{tn})$$

**For web applications,
Precision is more important than Recall**



Precision/Recall Using Set Notation

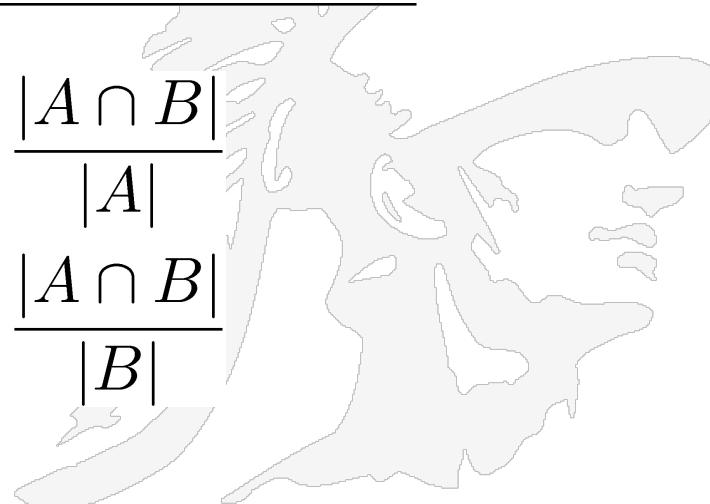
A is set of relevant documents,
 B is set of retrieved documents

	Relevant	Non-Relevant
Retrieved	$A \cap B$	$\overline{A} \cap B$
Not Retrieved	$A \cap \overline{B}$	$\overline{A} \cap \overline{B}$

you may not be able
 to see them, but
 A and B have a bar
 over them and it
 denotes the
 complement set

$$\text{Recall} = \frac{|A \cap B|}{|A|}$$

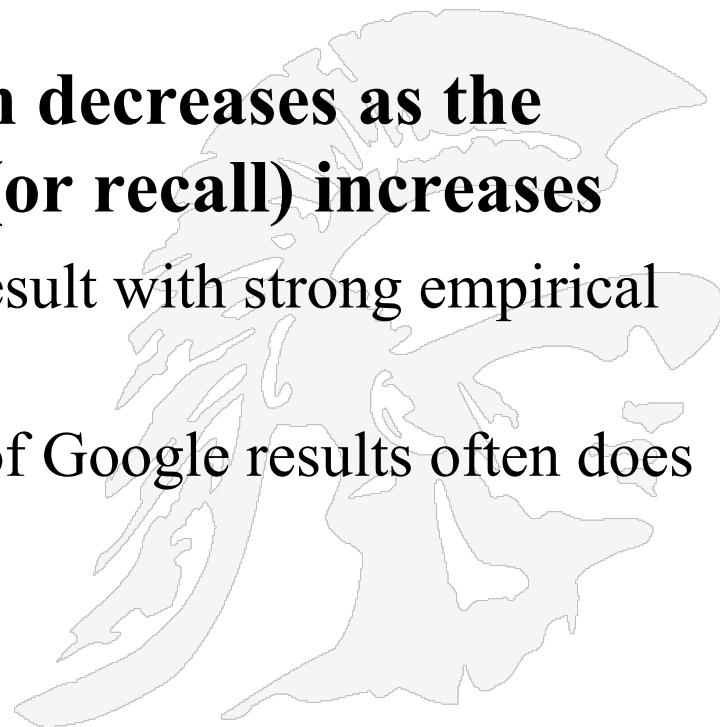
$$\text{Precision} = \frac{|A \cap B|}{|B|}$$



Precision/Recall

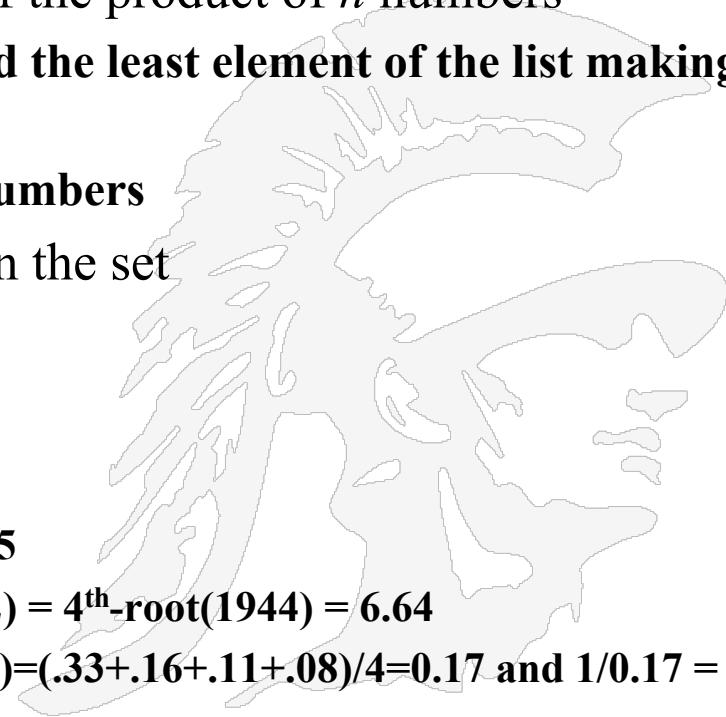
Two Observations

- You can get high recall (but low precision) by retrieving all docs for all queries!
 - a rather foolish strategy
- In a good system, precision decreases as the number of docs retrieved (or recall) increases
 - This is not a theorem, but a result with strong empirical confirmation
 - E.g. viewing multiple pages of Google results often does not improve precision at all



Harmonic Mean

- **There are three Pythagorean means**
 - 1. *arithmetic mean*, 2. *geometric mean*, 3. *harmonic mean*
 - of course we all know how to compute the arithmetic mean
 - the geometric mean is the n th root of the product of n numbers
- **The harmonic mean tends strongly toward the least element of the list making it useful in analyzing search engine results**
- **To find the harmonic mean of a set of n numbers**
 1. add the reciprocals of the numbers in the set
 2. divide the sum by n
 3. take the reciprocal of the result
- **e.g. for the numbers 3, 6, 9, and 12**
 - The arithmetic mean is: $(3+6+9+12)/4 = 7.5$
 - The geometric mean is: $\text{nth-root}(3*6*9*12) = 4^{\text{th}}\text{-root}(1944) = 6.64$
 - The harmonic mean is: $(1/3+1/6+1/9+1/12)=(.33+.16+.11+.08)/4=0.17 \text{ and } 1/0.17 = 5.88$



F Measure

- The harmonic mean of the precision and the recall is often used as an aggregated performance score for the evaluation of algorithms and systems: called the **F-score (or F-measure)**.
- *Harmonic mean of recall and precision is defined as*

$$F = \frac{1}{\frac{1}{2}(\frac{1}{R} + \frac{1}{P})} = \frac{2RP}{(R+P)}$$

- harmonic mean emphasizes the importance of small values, whereas the arithmetic mean is affected more by outliers that are unusually large
- More general form of F-Measure
 - β is a parameter that controls the relative importance of recall and precision

$$F_\beta = (\beta^2 + 1)RP / (R + \beta^2 P)$$

Calculating Recall/Precision at Fixed Positions



= the relevant documents

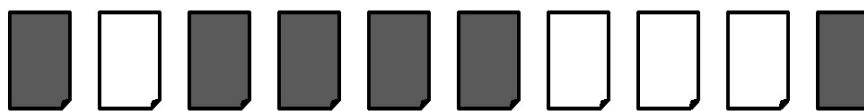
Steps

Recall:

$$1/6 = 0.17$$

Precision:

Ranking #1



e.g.
Google
Result

Recall:

$$1/6 = 0.17$$

Precision:

$$1/2 = 0.5$$

Recall:

$$2/6 = 0.33$$

Precision:

$$2/3 = 0.67$$

Recall:

$$3/6 = 0.5$$

Precision:

$$3/4 = 0.75$$

	Recall	0.17	0.17	0.33	0.5	0.67	0.83	0.83	0.83	1.0
--	--------	------	------	------	-----	------	------	------	------	-----

	Precision	1.0	0.5	0.67	0.75	0.8	0.83	0.71	0.63	0.56	0.6
--	-----------	-----	-----	------	------	-----	------	------	------	------	-----



e.g.
Bing
Result

	Recall	0.0	0.17	0.17	0.17	0.33	0.5	0.67	0.67	0.83	1.0
--	--------	-----	------	------	------	------	-----	------	------	------	-----

	Precision	0.0	0.5	0.33	0.25	0.4	0.5	0.57	0.5	0.56	0.6
--	-----------	-----	-----	------	------	-----	-----	------	-----	------	-----



$$\text{Recall} = \#RelevItemsRetr / \text{allRelevItems}$$

$$\text{Prec} = \#RelevItemsRetr / \text{allItemsRetr}$$

Average Precision of the Relevant Documents



= the relevant documents

computes the sum of the precisions of the relevant documents

Ranking #1



	Recall	0.17	0.17	0.33	0.5	0.67	0.83	0.83	0.83	1.0
	Precision	1.0	0.5	0.67	0.75	0.8	0.83	0.71	0.63	0.56



Ranking #2



	Recall	0.0	0.17	0.17	0.17	0.33	0.5	0.67	0.67	0.83	1.0
	Precision	0.0	0.5	0.33	0.25	0.4	0.5	0.57	0.5	0.56	0.6

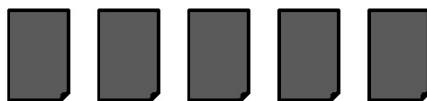


→ Ranking #1: $(1.0 + 0.67 + 0.75 + 0.8 + 0.83 + 0.6)/6 = 0.78$

→ Ranking #2: $(0.5 + 0.4 + 0.5 + 0.57 + 0.56 + 0.6)/6 = 0.52$

Conclusion: Ranking #1 for this query is best

Averaging Across Queries

 = relevant documents for query 1

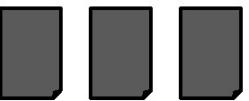
Ranking #1



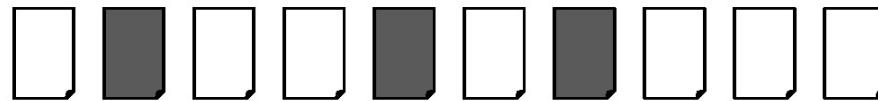
	Recall	0.2	0.2	0.4	0.4	0.4	0.6	0.6	0.6	0.8	1.0
Precision	1.0	0.5	0.67	0.5	0.4	0.5	0.43	0.38	0.44	0.5	

Average precision across the two queries for relevant docs is:

$$(1 + .67 + .5 + .44 + .5 + .5 + .4 + .43)/8 = 0.55$$

 = relevant documents for query 2

Ranking #2



	Recall	0.0	0.33	0.33	0.33	0.67	0.67	1.0	1.0	1.0	1.0
Precision	0.0	0.5	0.33	0.25	0.4	0.33	0.43	0.38	0.33	0.3	

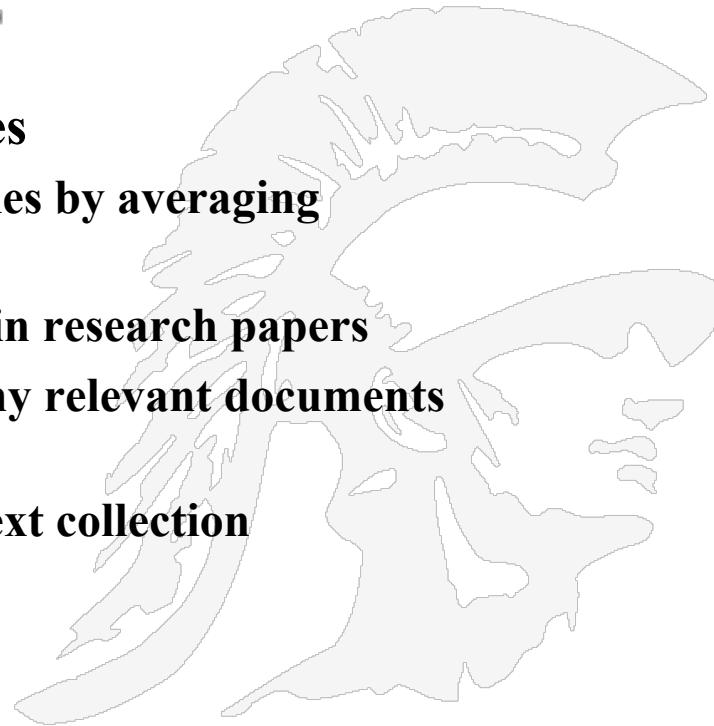
Averaging Across Queries

- ***Mean average precision (MAP)*** for a set of queries is the mean of the average precision scores for each query.

$$\text{MAP} = \frac{\sum_{q=1}^Q \text{AveP}(q)}{Q}$$

where Q is the number of queries

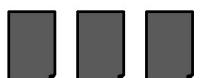
- Summarize rankings from multiple queries by averaging average precision
- This is the ***most commonly used*** measure in research papers
- Assumes user is interested in finding many relevant documents for each query
- Requires many relevance judgments in text collection



Mean Average Precision Example

 = relevant documents for query 1

Ranking #1									
Recall	0.2	0.2	0.4	0.4	0.4	0.6	0.6	0.8	1.0
Precision	1.0	0.5	0.67	0.5	0.4	0.5	0.43	0.38	0.44

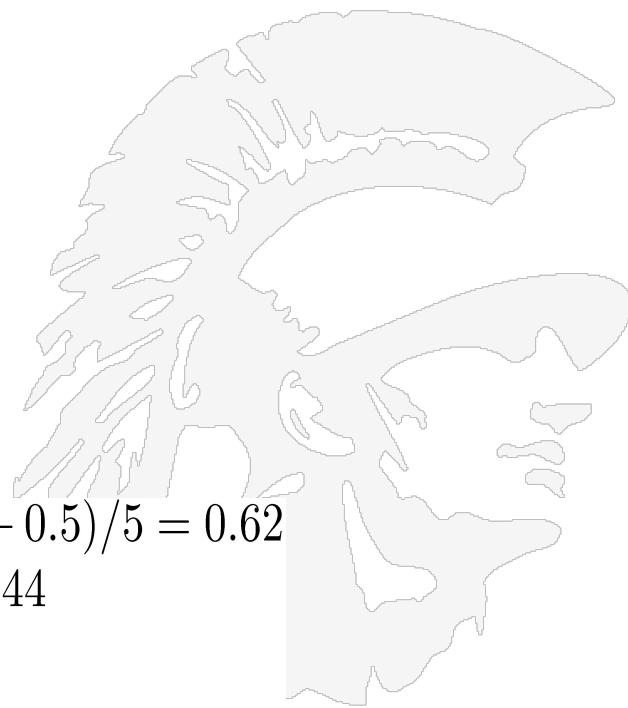
 = relevant documents for query 2

Ranking #2									
Recall	0.0	0.33	0.33	0.33	0.67	0.67	1.0	1.0	1.0
Precision	0.0	0.5	0.33	0.25	0.4	0.33	0.43	0.38	0.33

$$\text{average precision query 1} = (1.0 + 0.67 + 0.5 + 0.44 + 0.5) / 5 = 0.62$$

$$\text{average precision query 2} = (0.5 + 0.4 + 0.43) / 3 = 0.44$$

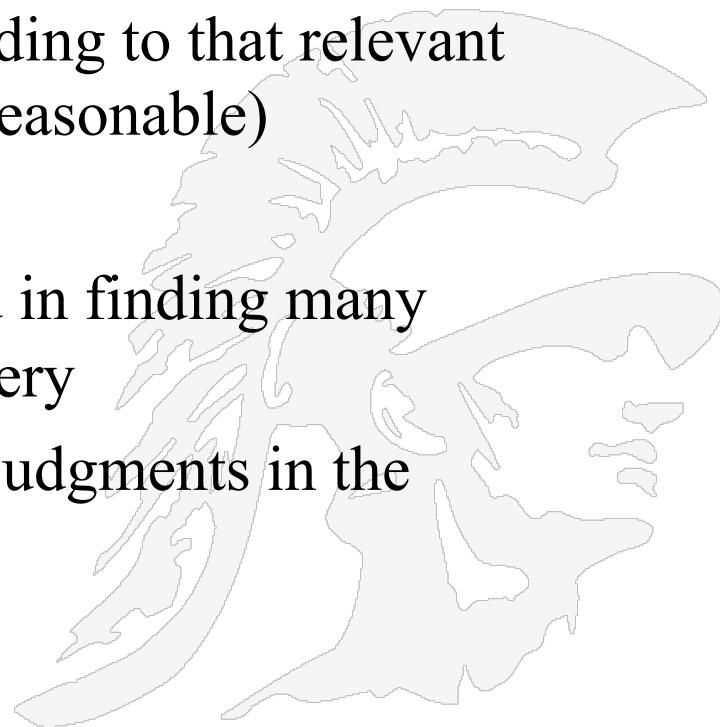
$$\text{mean average precision} = (0.62 + 0.44) / 2 = 0.53$$



More on Mean Average Precision Calculation

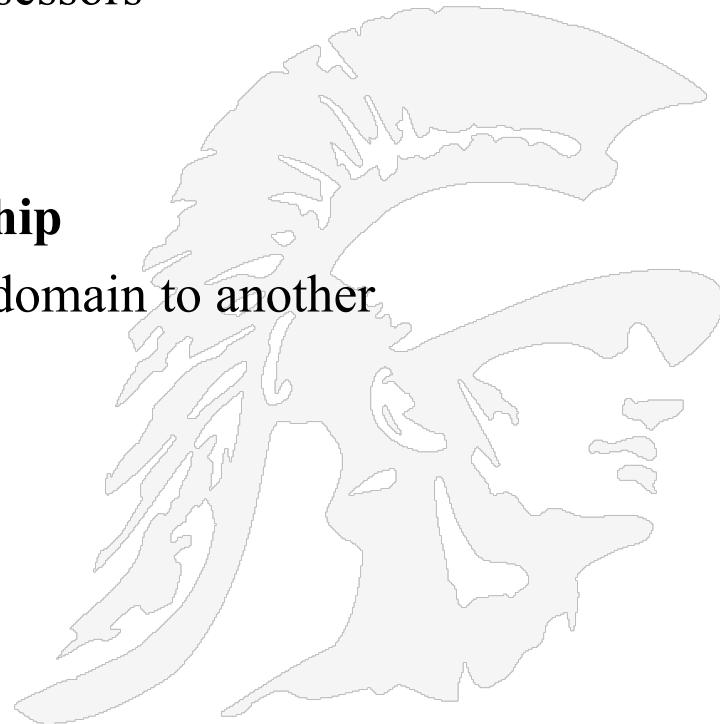
■ Mean Average Precision (MAP)

- Some negative aspects
 - If a relevant document never gets retrieved, we assume the precision corresponding to that relevant doc to be zero (this is actually reasonable)
 - Each query counts equally
 - MAP assumes user is interested in finding many relevant documents for each query
 - MAP requires many relevance judgments in the document collection



Difficulties in Using Precision/Recall

- Should average over large document collection and query ensembles
- Need human relevance assessments
 - But people aren't always reliable assessors
- Assessments have to be binary
 - Nuanced assessments?
- Heavily skewed by collection/authorship
 - Results may not translate from one domain to another



A Final Evaluation Measure: Discounted Cumulative Gain

- The premise of DCG is that highly relevant documents appearing lower in a search result list should be penalized as the graded relevance value is reduced logarithmically proportional to the position of the result.
- The discounted CG accumulated at a particular rank position p is defined as

$$\text{DCG}_p = \sum_{i=1}^p \frac{\text{rel}_i}{\log_2(i+1)} = \text{rel}_1 + \sum_{i=2}^p \frac{\text{rel}_i}{\log_2(i+1)}$$

where rel_i is the graded relevance of the result at position i

- Gain is accumulated starting at the top of the ranking and may be reduced, or *discounted*, at lower ranks
- Typical discount is $1/\log(\text{rank})$
 - With base 2, the discount at rank 4 is $1/2$, and at rank 8 it is $1/3$
- An alternative formulation of DCG places stronger emphasis on retrieving relevant documents:

$$\text{DCG}_p = \sum_{i=1}^p \frac{2^{\text{rel}_i} - 1}{\log_2(i+1)}$$

Discounted Cumulative Gain Example

we want high weights for high rank documents, because searchers are likely to inspect them, and low weights for low rank documents that searchers are unlikely to ever see.

The discount factor is commonly chosen as $\log_2(\text{rank} + 1)$ and is used to divide the relevance grade.

Using a logarithm for the position penalty makes the decay effect more gradual compared to using the position itself.

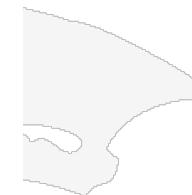
Discount examples

Rank	Grade	Discount1 [1/rank]	Discount2 [log2(rank + 1)]	Discount1 Grade	Discount2 Grade
1	4	1.000	1.000	4.000	4.000
2	3	0.500	0.631	1.500	1.893
3	2	0.333	0.500	0.667	1.000
4	1	0.250	0.431	0.250	0.431
5	1	0.200	0.387	0.200	0.387

Search Engine Evaluation Metrics

Metrics table

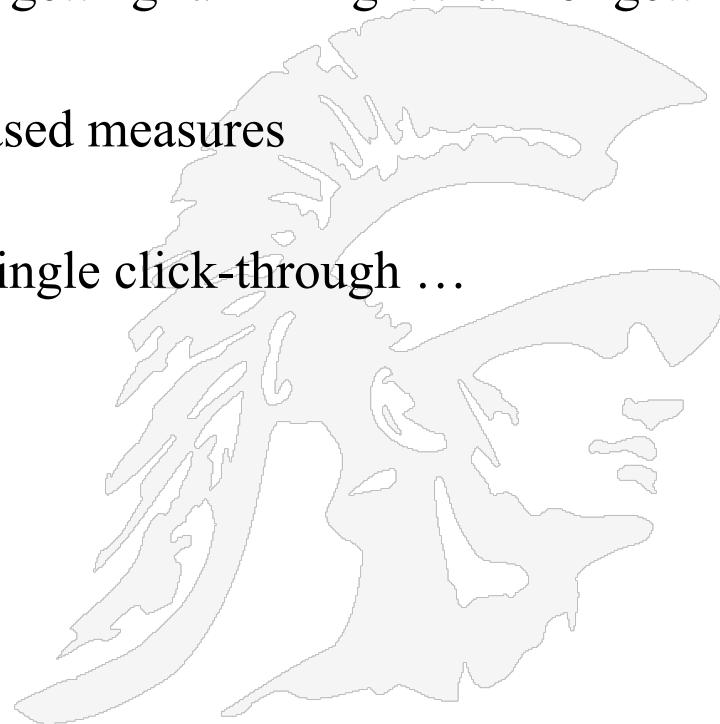
Scale	Metric	Measures	Drawbacks
Binary	Precision (P)	The relevance of the entire results set (gridded results display)	Doesn't account for position
Binary	Average Precision (AP)	Relevance to a user scanning results sequentially	Large impact of low-rank results
Graded	Cumulative Gain (CG)	Information gain from a results set	Same as <i>Precision</i> doesn't factor in position
Graded	Discount Cumulative Gain (DCG)	Information gain with positional weighting	Difficult to compare across queries
Graded	normalized DCG (nDCG)	How close the results are to the best possible	No longer shows information gain



Finally see [https://en.wikipedia.org/wiki/Evaluation_measures_\(information_retrieval\)](https://en.wikipedia.org/wiki/Evaluation_measures_(information_retrieval))
 Copyright Ellis Horowitz 2011-2021

How Evaluation is Done at Web Search Engines

- Search engines have test collections of queries and hand-ranked results
- Recall is difficult to measure on the web
- Search engines often use precision at top k positions, e.g., $k = 10$
- . . . or measures that reward you more for getting rank 1 right than for getting rank 10 right.
- Search engines also use non-relevance-based measures
 - Click-through on first result
 - Not very reliable if you look at a single click-through . . .
but pretty reliable in the aggregate.
 - Studies of user behavior in the lab
 - A/B testing



Google's Search Quality Rating Guidelines Document

- Google relies on raters, working in many countries and languages around the world
- The data they generate is rolled up statistically to give
 - a view of the quality of search results and search experience over time, and
 - an ability to measure the effect of proposed changes to Google's search algorithms

General Guidelines

October 14, 2020

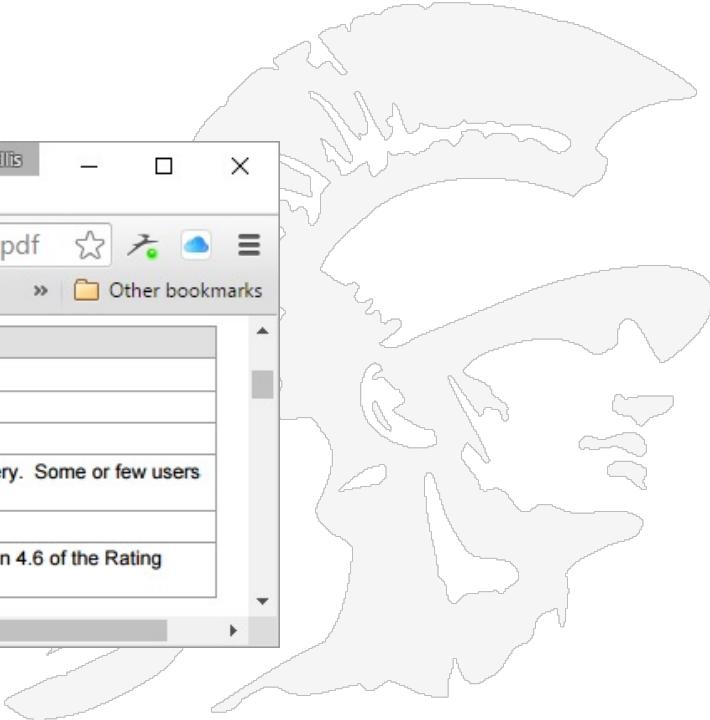
General Guidelines Overview	5
Introduction to Search Quality Rating	6
0.0 The Search Experience	6
0.1 The Purpose of Search Quality Rating	6
0.2 Raters Must Represent People in their Rating Locale	6
0.3 Browser Requirements	7
0.4 Ad Blocking Extensions	7
0.5 Internet Safety Information	7
0.6 The Role of Examples in these Guidelines	7
Part 1: Page Quality Rating Guideline	8
1.0 Introduction to Page Quality Rating	8
2.0 Understanding Webpages and Websites	8
2.1 Important Definitions	8
2.2 What is the Purpose of a Webpage?	9
2.3 Your Money or Your Life (YMYL) Pages	10
2.4 Understanding Webpage Content	10
2.4.1 Identifying the Main Content (MC)	10
2.4.2 Identifying the Supplementary Content (SC)	11
2.4.3 Identifying Advertisements/Monetization (Ads)	11
2.4.4 Summary of the Parts of the Page	12
2.5 Understanding the Website	12
2.5.1 Finding the Homepage	12
2.5.2 Finding Who is Responsible for the Website and Who Created the Content on the Page	14
2.5.3 Finding About Us, Contact Information, and Customer Service Information	14
2.6 Reputation of the Website or Creator of the Main Content	15
2.6.1 Research on the Reputation of the Website or Creator of the Main Content	16
2.6.2 Sources of Reputation Information	16
2.6.3 Customer Reviews of Stores/Businesses	16
2.6.4 How to Search for Reputation Information	16
2.6.5 What to Do When You Find No Reputation Information	18
3.0 Overall Page Quality Rating	19
3.1 Page Quality Rating: Most Important Factors	19
3.2 Expertise, Authoritativeness, and Trustworthiness (E-A-T)	19
4.0 High Quality Pages	20
4.1 Characteristics of High Quality Pages	20
4.2 A Satisfying Amount of High Quality Main Content	21
4.3 Clear and Satisfying Website Information: Who is Responsible and Customer Service	21
4.4 Positive Reputation	21
4.5 A High Level of Expertise/Authoritativeness/Trustworthiness (E-A-T)	22
4.6 Examples of High Quality Pages	22
5.0 Highest Quality Pages	26

http://csci572.com/papers/2020_10searchqualityevaluatorguidelines.pdf

Google's Search Quality Ratings Guidelines Document

- This document gives evaluators examples and guidelines for appropriate ratings.
- the evaluator looks at a search query and a result that could be returned. They rate the relevance of the result for that query on a scale described within the document.

The six rating scale categories



Rating Scale	Description
Vital	A special rating category. See Section 4.1 of the Rating Guidelines.
Useful	A page that is very helpful for most users.
Relevant	A page that is helpful for many or some users.
Slightly Relevant	A page that is not very helpful for most users, but is somewhat related to the query. Some or few users would find this page helpful.
Off-Topic or Useless	A page that is helpful for very few or no users.
Unratable	A page that cannot be evaluated. A complete description can be found in Section 4.6 of the Rating Guidelines.

1. Precision Evaluations

People use the Guidelines to rate search results

2. Side-by-Side Experiments

people are shown two different sets of search results and asked which they prefer

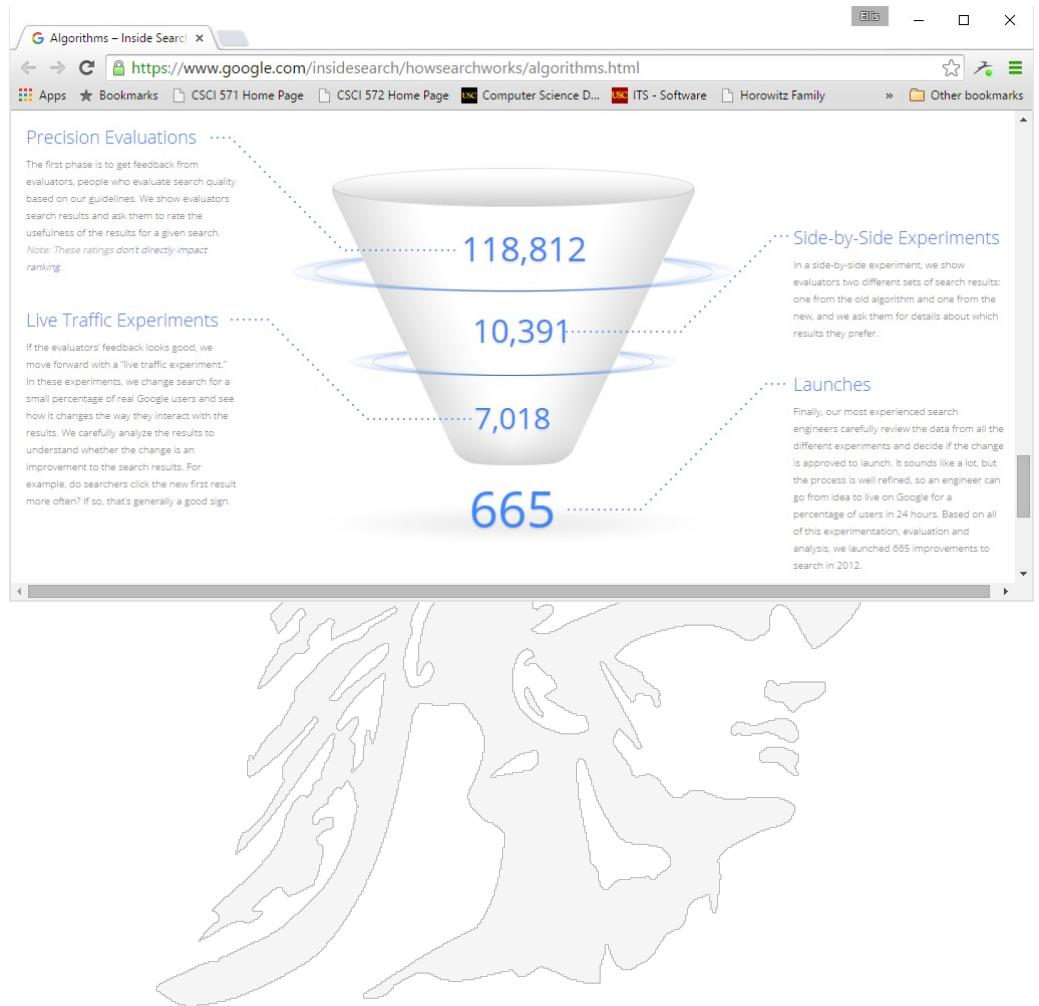
3. Live Traffic Experiments

the search algorithm is altered for a small number of actual users

4. Full Launch

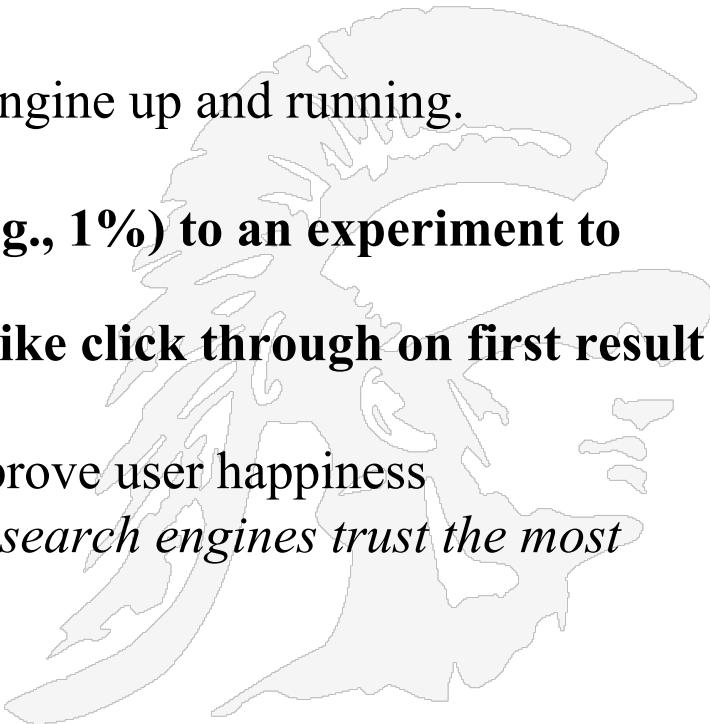
A final analysis by Google engineers and the improvement is released

Google's 4-Step Process for Changing Their Search Algorithm



A/B Testing at Web Search Engines

- **A/B testing** is comparing two versions of a web page to see which one performs better. You compare two web pages by showing the two variants (let's call them **A** and **B**) to similar visitors at the same time. The one that gives a better conversion rate, wins!
- 1. **Purpose:** Test a single innovation
- 2. **Prerequisite:** You have a large search engine up and running.
- 3. **Have most users use old system**
- 4. **Divert a small proportion of traffic (e.g., 1%) to an experiment to evaluate an innovation**
- 5. **Evaluate with an automatic measure like click through on first result**
- we directly see if the innovation does improve user happiness
- *This is the evaluation methodology large search engines trust the most*



USING USER CLICKS FOR EVALUATION



What Do Clicks Tell Us?

ALL RESULTS

RELATED SEARCHES
[CIKM 2008](#)

SEARCH HISTORY
Turn on search history to start remembering your searches.
Turn history on

ALL RESULTS

[CIKM 2008 | Home](#)
Napa Valley Marriott Hotel & Spa: Napa Valley, California October 26-30, 2008
[cikm2008.org](#) · [Cached page](#)

Papers Program Committee
Themes News
Important Dates Napa Valley
Banquet Posters
[Show more results from cikm2008.org](#)

[Conference on Information and Knowledge Management \(CIKM\)](#)
Provides an international forum for presentation and discussion of research on information and knowledge management, as well as recent advances on data and knowledge bases ...
[www.cikm.org](#) · [Cached page](#)

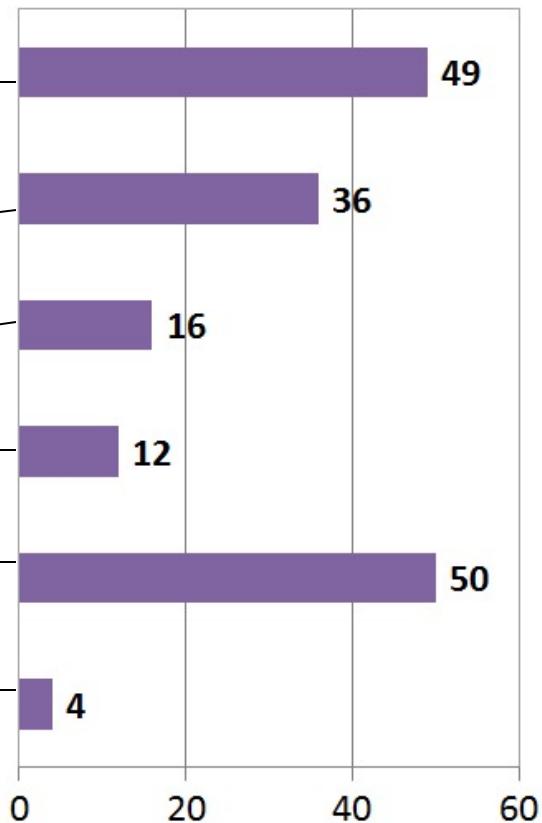
[Conference on Information and Knowledge Management \(CIKM'02\)](#)
SAIC Headquarters, McLean, Virginia, USA, 4-9 November 2002.
[www.cikm.org/2002](#) · [Cached page](#)

[ACM CIKM 2007 - Lisbon, Portugal](#)
News and announcements: 12/02 - Best interdisciplinary paper award at CIKM 2007 went to Fei Wu and Daniel Weld for Autonomously Semantifying Wikipedia.
[www.fc.ul.pt/cikm2007](#) · [Cached page](#)

[CIKM 2009 | Home](#)
CIKM 2009 (The 18th ACM Conference on Information and Knowledge Management) will be held on November 2-6, 2009, Hong Kong. Since 1992, CIKM has successfully brought together ...
[www.comp.polyu.edu.hk/conference/cikm2009](#) · [Cached page](#)

[Conference on Information and Knowledge Management \(CIKM\)](#)
CIKM Conference on Information and Knowledge Management The Conference on Information and Knowledge Management (CIKM) provides an international forum for presentation and ...
[cikmconference.org](#) · [Cached page](#)

of clicks received



There is strong position bias, so absolute click rates unreliable

Relative vs Absolute Ratings

ALL RESULTS

ALL RESULTS 1-10 of 131,000 results · Advanced

[CIKM 2008 | Home](#)
 Napa Valley Marriott Hotel & Spa: Napa Valley, California October 26-30, 2008
[cikm2008.org](#) · [Cached page](#)

Papers	Program Committee
Themes	News
Important Dates	Napa Valley
Banquet	Posters

Show more results from cikm2008.org

[Conference on Information and Knowledge Management \(CIKM\)](#)
 Provides an international forum for presentation and discussion of research on information and knowledge management, as well as recent advances on data and knowledge bases ...
[www.cikm.org](#) · [Cached page](#)

[Conference on Information and Knowledge Management \(CIKM'02\)](#)
 SAIC Headquarters, McLean, Virginia, USA, 4-9 November 2002.
[www.cikm.org/2002](#) · [Cached page](#)

[ACM CIKM 2007 - Lisbon, Portugal](#)
 News and announcements: 12/02 - Best interdisciplinary paper award at CIKM 2007 went to Fei Wu and Daniel Weld for Autonomously Semantifying Wikipedia.
[www.fc.ul.pt/cikm2007](#) · [Cached page](#)

[CIKM 2009 | Home](#)
 CIKM 2009 (The 18th ACM Conference on Information and Knowledge Management) will be held on November 2-6, 2009, Hong Kong. Since 1992, CIKM has successfully brought together ...
[www.comp.polyu.edu.hk/conference/cikm2009](#) · [Cached page](#)

[Conference on Information and Knowledge Management \(CIKM\)](#)
 CIKM Conference on Information and Knowledge Management The Conference on Information and Knowledge Management (CIKM) provides an international forum for presentation and ...
[cikmconference.org](#) · [Cached page](#)

User's click sequence

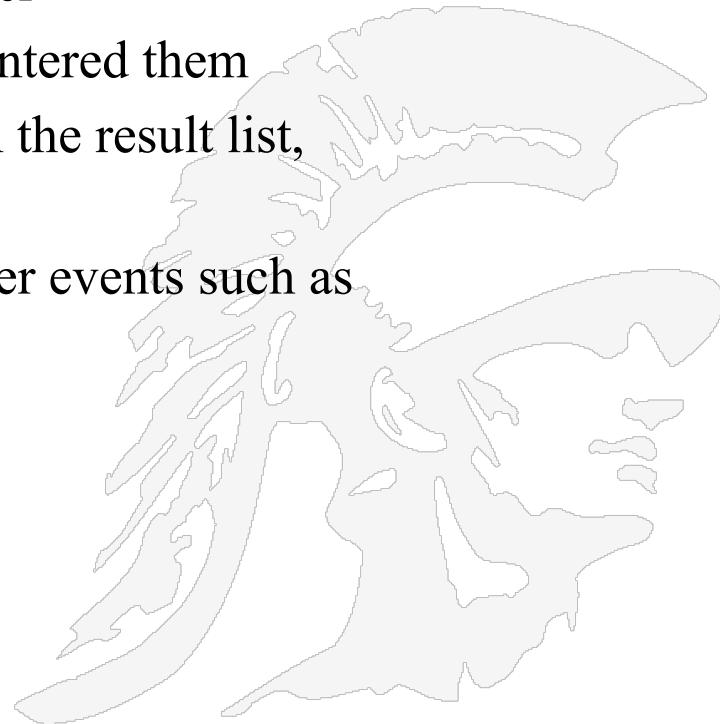


Hard to conclude Result1 > Result3

Probably can conclude Result3 > Result2

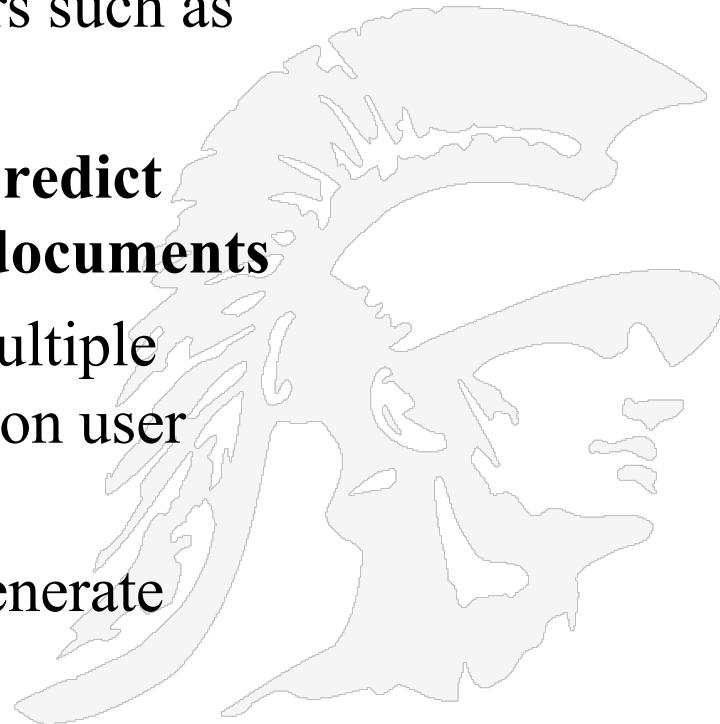
Query Logs

- **Used for both tuning and evaluating search engines**
 - also for various techniques such as query suggestion
- **Typical contents of the query log files**
 - User identifier or user session identifier
 - Query terms - stored exactly as user entered them
 - List of URLs of results, their ranks on the result list, and whether they were clicked on
 - Timestamp(s) - records the time of user events such as query submission, clicks



How Query Logs Can Be Used

- **Clicks are not relevance judgments**
 - although they are correlated
 - biased by a number of factors such as rank on result list
- **Can use clickthrough data to predict *preferences* between pairs of documents**
 - appropriate for tasks with multiple levels of relevance, focused on user relevance
 - various “policies” used to generate preferences



A Final Thought

Google's Enhancements of Search Results

Display improvements

- immediate answers
- autocomplete anticipations

Extensions to More Data

- results from books
- results from news
- results from images
- results from patents
- results from air schedules

New Input forms

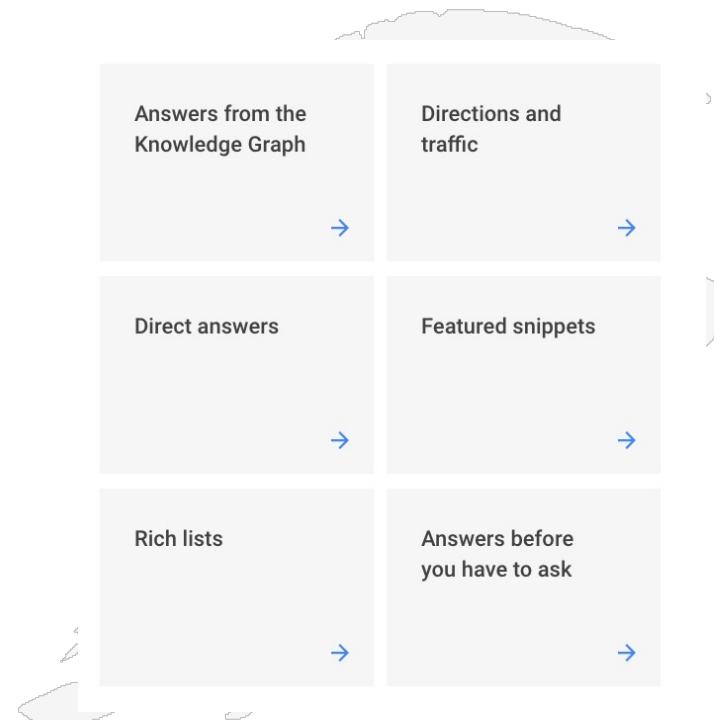
- search by voice
- search by image

information retrieval improvements

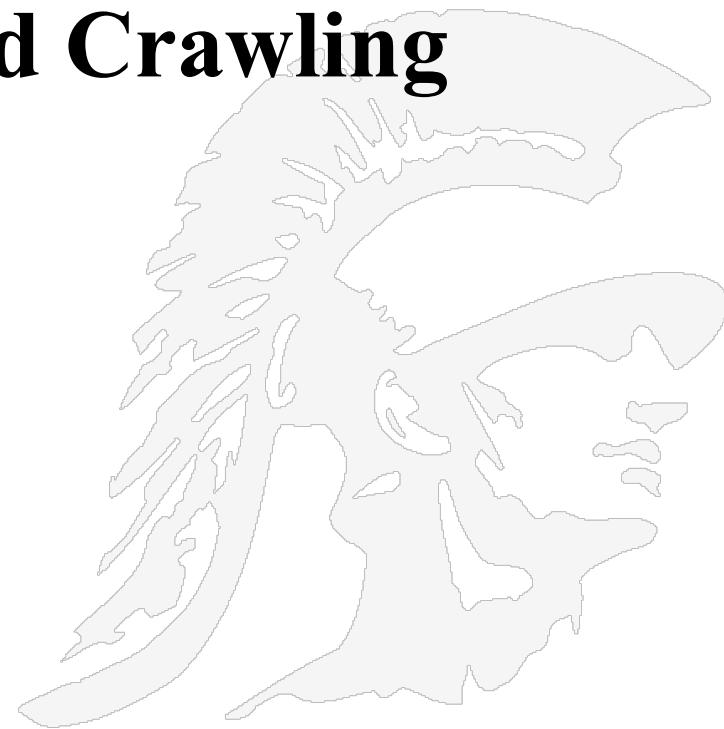
- snippets
- spelling correction
- translations
- People Also Ask boxes
- use of synonyms
- use of knowledge graph

The page below discusses the many aspects that go into producing search results at Google

<https://www.google.com/search/howsearchworks>



Crawlers and Crawling



There are Many Crawlers

- A web crawler is a computer program that visits web pages in an organized way

- Sometimes called a spider or robot

- A list of web crawlers can be found at

http://en.wikipedia.org/wiki/Web_crawler

Google's crawler is called googlebot, see

<http://support.google.com/webmasters/bin/answer.py?hl=en&answer=182072>

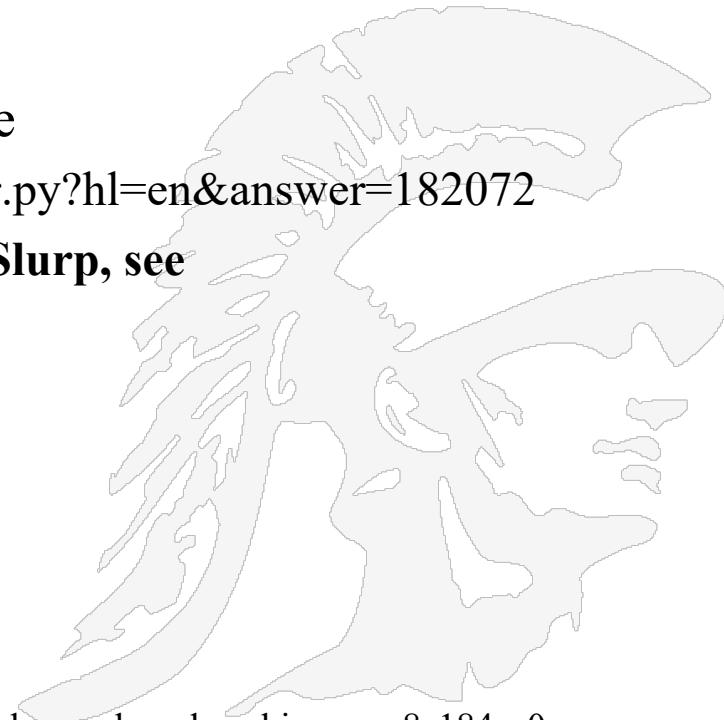
- Yahoo's web crawler is/was called Yahoo! Slurp, see

http://en.wikipedia.org/wiki/Yahoo!_Search

- Bing uses five crawlers

- Bingbot, standard crawler
 - Adidxbot, used by Bing Ads
 - MSNbot, remnant from MSN, but still in use
 - MSNBotMedia, crawls images and video
 - BingPreview, generates page snapshots

- For details see: <http://www.bing.com/webmaster/help/which-crawlers-does-bing-use-8c184ec0>



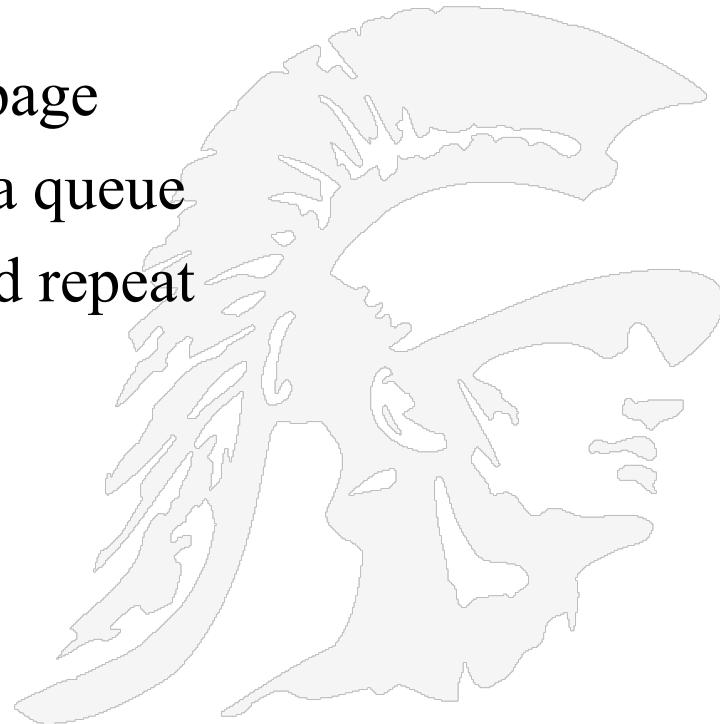
Web Crawling Issues

- **How to crawl?**
 - *Quality*: how to find the “Best” pages first
 - *Efficiency*: how to avoid duplication (or near duplication)
 - *Etiquette*: behave politely by not disturbing a website’s performance
- **How much to crawl? How much to index?**
 - *Coverage*: What percentage of the web should be covered?
 - *Relative Coverage*: How much do competitors have?
- **How often to crawl?**
 - *Freshness*: How much has changed?
 - How much has really changed?



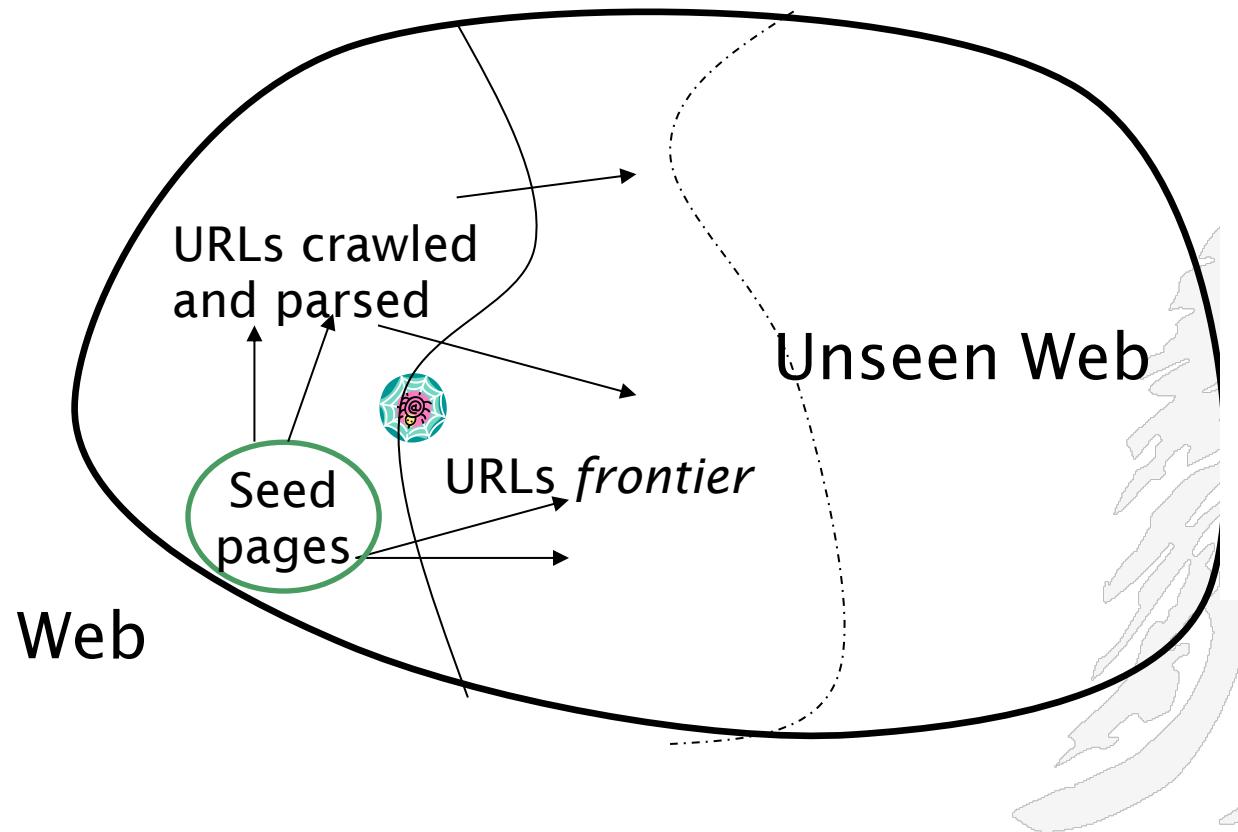
Simplest Crawler Operation

- Initialize (begin with known “seed” pages)
- Loop: Fetch and parse a page
 - Place the page in a database
 - Extract the URLs within the page
 - Place the extracted URLs on a queue
 - Fetch a URL on the queue and repeat



Crawling Picture

20 Web crawling and indexes



20.1 Overview

Web crawling is the process by which we gather pages from the Web, in order to index them and support a search engine. The objective of crawling is to quickly and efficiently gather as many useful web pages as possible, together with the link structure that interconnects them. In Chapter 19 we studied the complexities of the Web stemming from its creation by millions of uncoordinated individuals. In this chapter we study the resulting difficulties for crawling the Web. The focus of this chapter is the component shown in Figure 19.7 as *web crawler*; it is sometimes referred to as a *spider*.

The goal of this chapter is not to describe how to build the crawler for a full-scale commercial web search engine. We focus instead on a range of issues that are generic to crawling from the student project scale to substantial research projects. We begin (Section 20.1.1) by listing desiderata for web crawlers, and then discuss in Section 20.2 how each of these issues is addressed. The remainder of this chapter describes the architecture and some implementation details for a distributed web crawler that satisfies these features. Section 20.3 discusses distributing indexes across many machines for a web-scale implementation.

20.1.1 Features a crawler *must* provide

We list the desiderata for web crawlers in two categories: features that web crawlers *must* provide, followed by features they *should* provide.

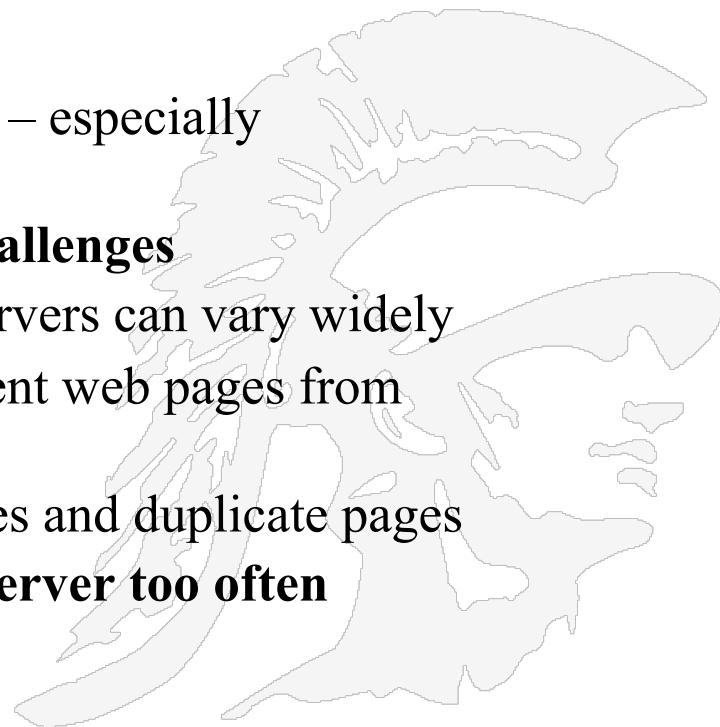
Robustness: The Web contains servers that create *spider traps*, which are generators of web pages that mislead crawlers into getting stuck fetching an infinite number of pages in a particular domain. Crawlers must be designed to be resilient to such traps. Not all such traps are malicious; some are the inadvertent side-effect of faulty website development.

Online edition (c) 2009 Cambridge UP

Our textbook

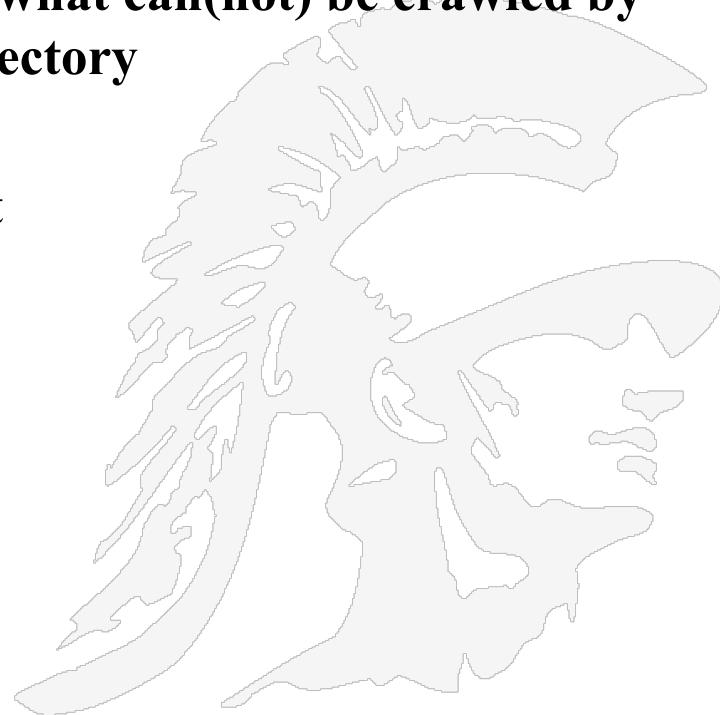
Simple Picture – Complications

- Crawling the entire web isn't feasible with one machine
 - But all of the above steps can be distributed
- Challenges
 - Handling/Avoiding malicious pages
 - Some pages contain spam
 - Some pages contain spider traps – especially dynamically generated pages
 - Even non-malicious pages pose challenges
 - Latency/bandwidth to remote servers can vary widely
 - Robots.txt stipulations can prevent web pages from being visited
 - How can one avoid mirrored sites and duplicate pages
 - Maintain politeness – don't hit a server too often



Robots.txt

- There is a protocol that defines the limitations for a web crawler as it visits a website; its definition is here
 - <http://www.robotstxt.org/orig.html>
- The website announces its request on what can(not) be crawled by placing a robots.txt file in the root directory
 - e.g. see
<http://www.ticketmaster.com/robots.txt>



Robots.txt Example

- **No robot visiting this domain should visit any URL starting with "/yoursite/temp/":**

User-agent: *

Disallow: /yoursite/temp/

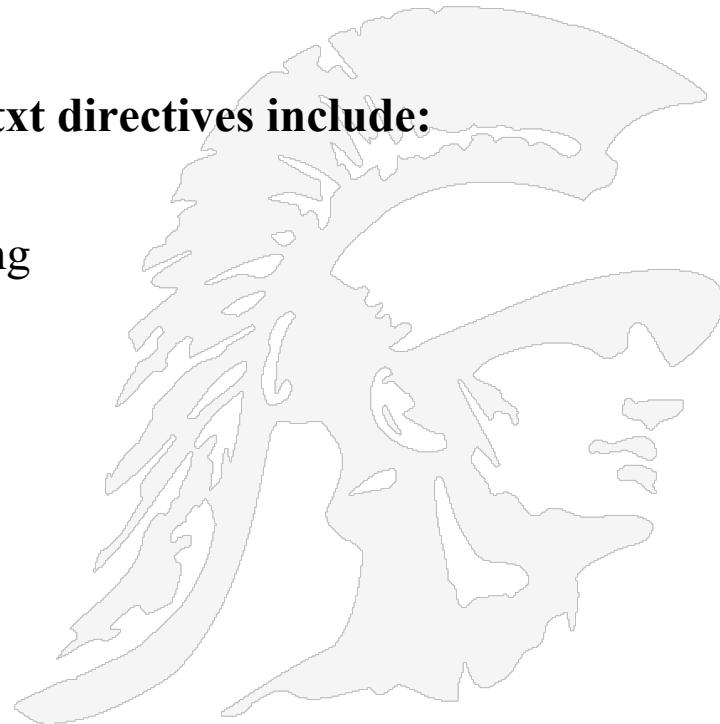
- **Directives are case sensitive**
- **Additional symbols allowed in the robots.txt directives include:**
 - '*' - matches a sequence of characters
 - '\$' - anchors at the end of the URL string
- **Example of '*':**

User-agent: Slurp

Allow: /public*/

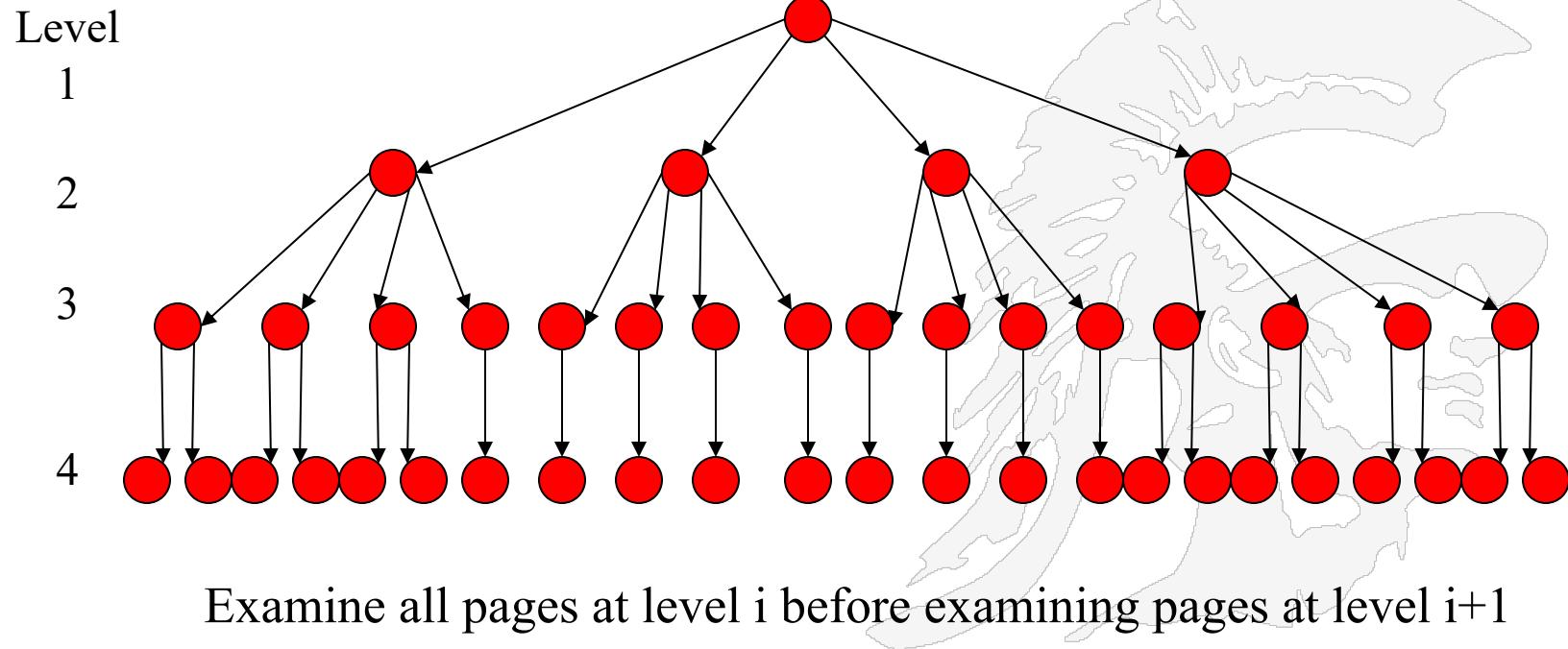
Disallow: /*_print*.html

Disallow: /*?sessionid



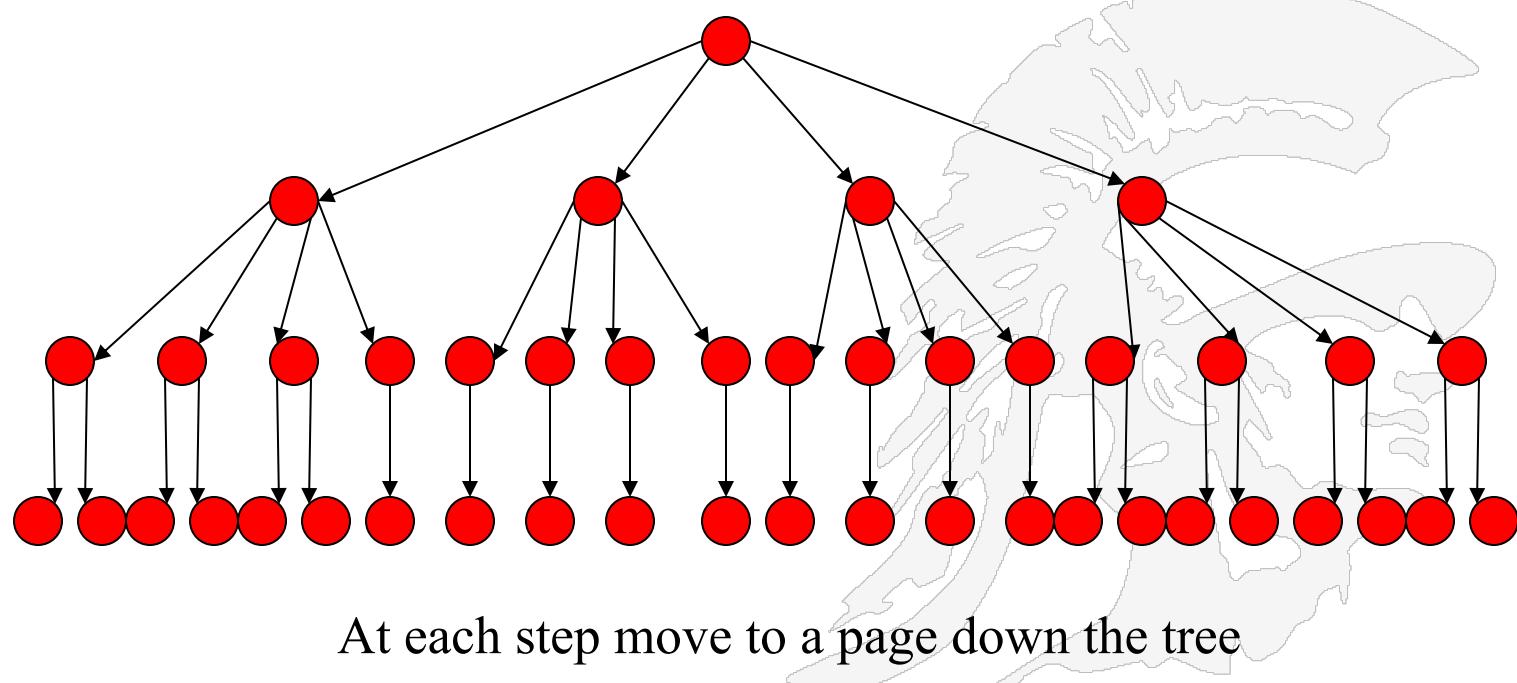
Basic Search Strategies

Breadth-first Search

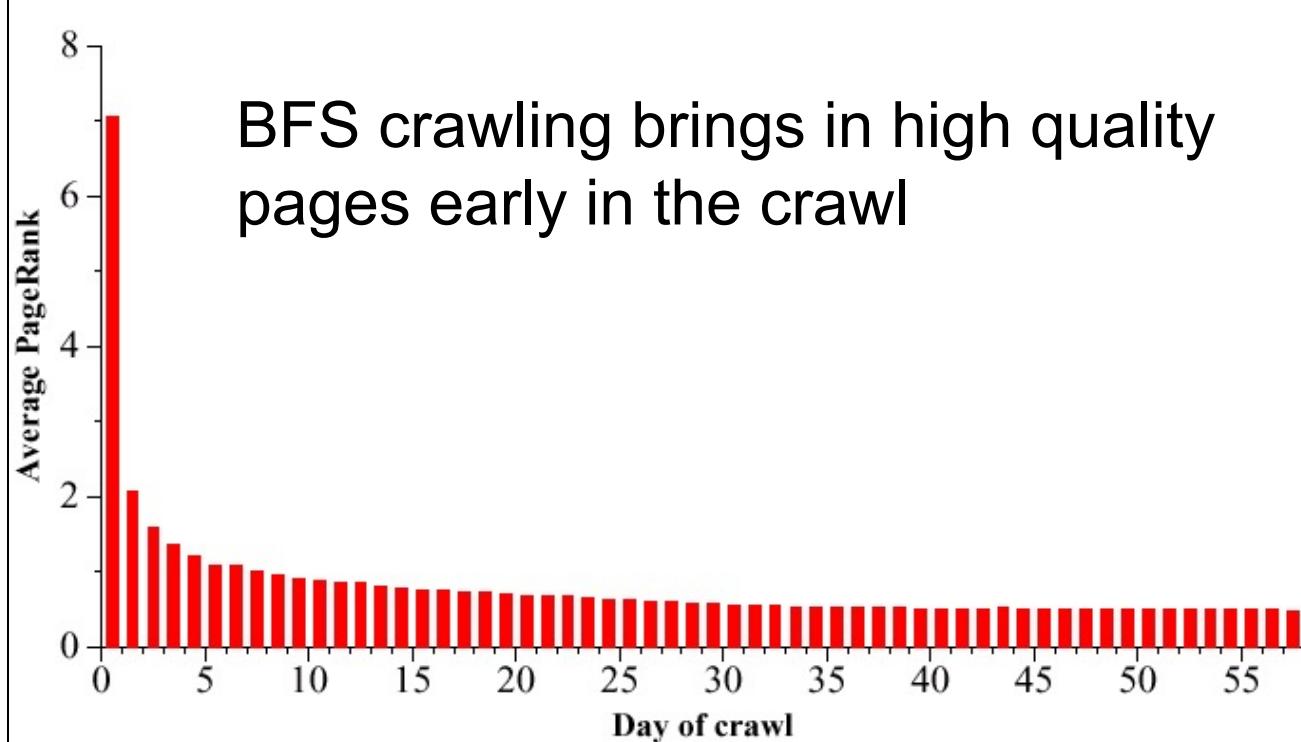


Basic Search Strategies (cont)

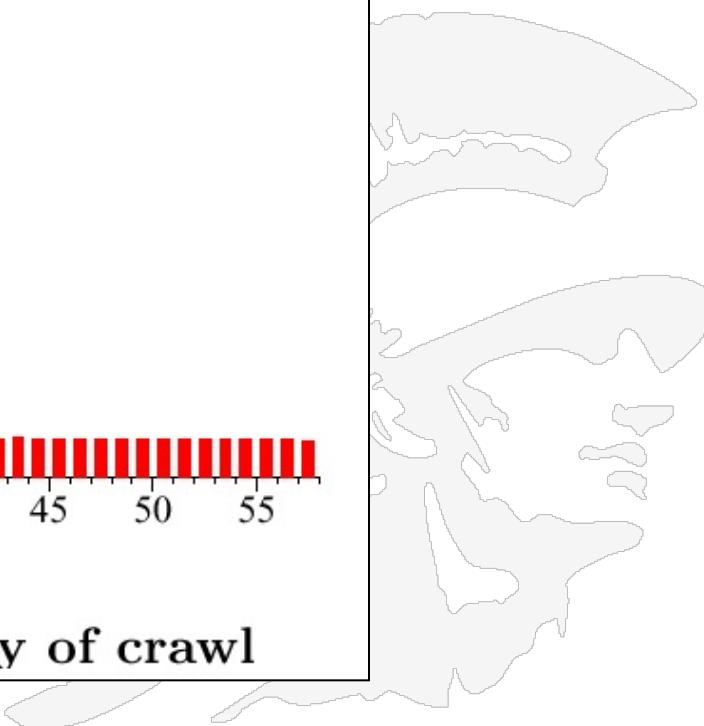
Depth-first Search



Web Wide Crawl (328M pages) [Najo01]



Average PageRank score by day of crawl



Page Rank is an algorithm developed by Google for determining the value of a page

Crawling Algorithm – Version 2

Initialize queue (Q) with initial set of known URL's.

Loop until Q empty or page or time limit exhausted:

Pop a URL, call it L, from the front of Q.

If L is not an HTML page (e.g. .gif, .jpeg,)

continue the loop

If L has already been visited, continue the loop.

Download page, P, for L

If cannot download P (e.g. 404 error, robot excluded)

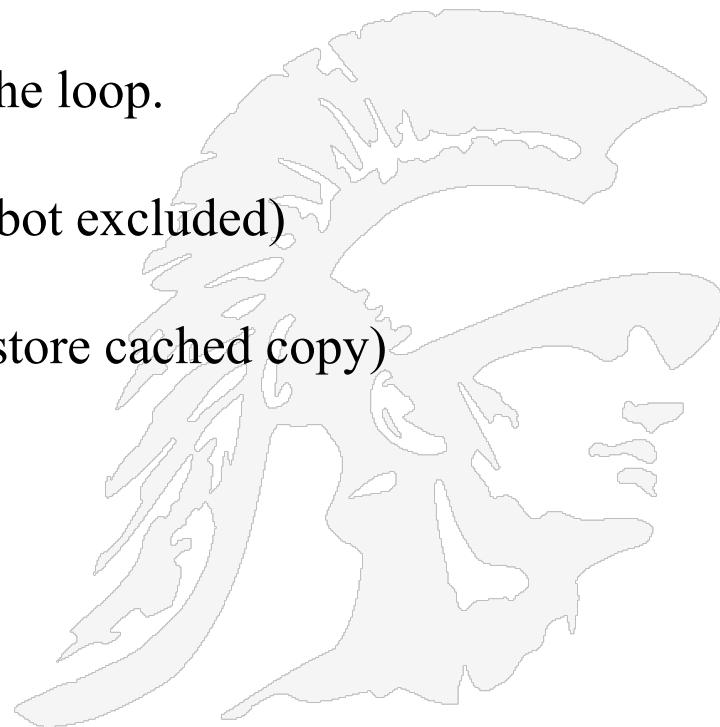
continue loop

Index P (e.g. add to inverted index and store cached copy)

Parse P to obtain list of new links N.

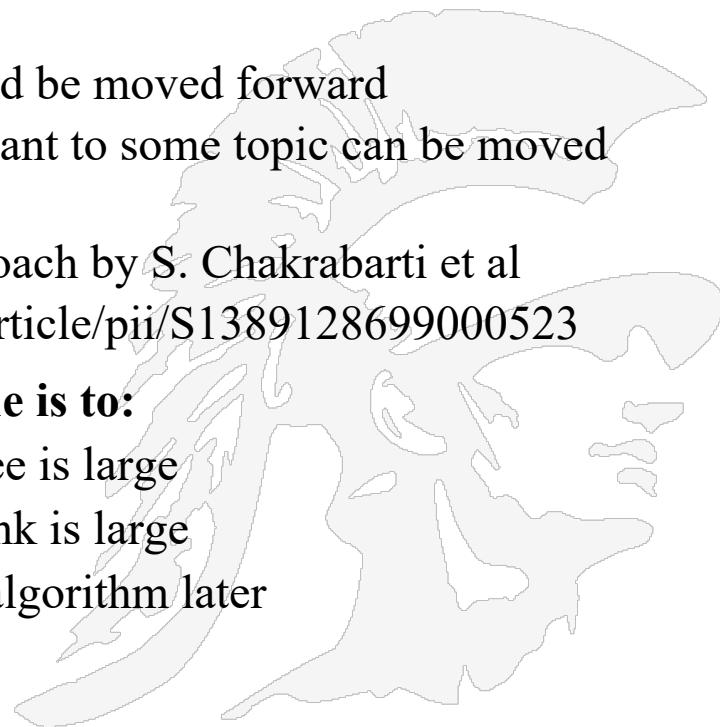
Append N to the end of Q

End loop



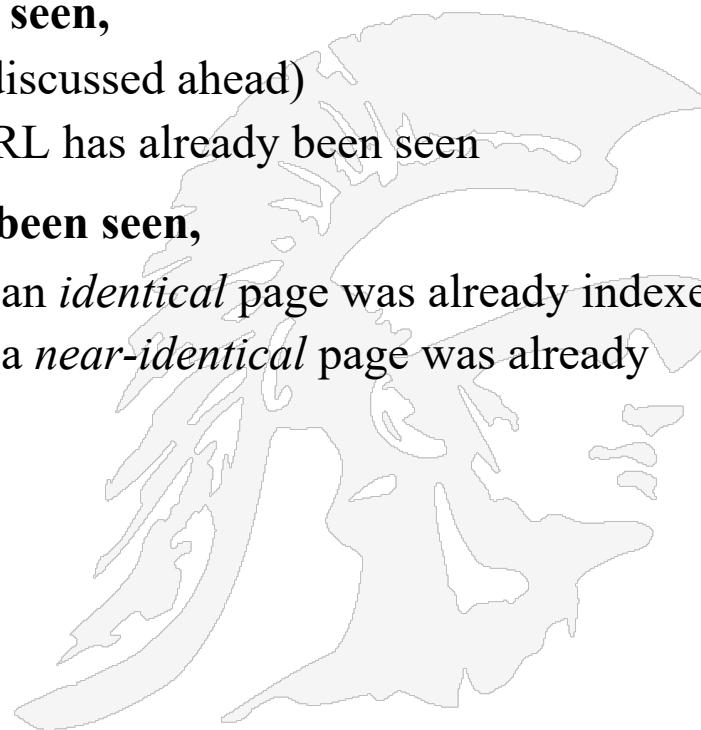
Queueing Strategy

- How new links are added to the queue determines the search strategy.
- FIFO (append to end of Q) gives breadth-first search.
- LIFO (add to front of Q) gives depth-first search.
- Heuristically ordering the Q gives a “focused crawler” that directs its search towards “interesting” pages; e.g.
 - A document that changes frequently could be moved forward
 - A document whose content appears relevant to some topic can be moved forward
 - e.g. see Focused Crawling: A New Approach by S. Chakrabarti et al
 - <https://www.sciencedirect.com/science/article/pii/S1389128699000523>
- One way to re-order the URLs on the queue is to:
 - Move forward URLs whose In-degree is large
 - Move forward URLs whose PageRank is large
 - We will discuss the PageRank algorithm later



Avoiding Page Duplication

- A crawler must detect when revisiting a page that has already been crawled (Remember: the web is a graph not a tree).
- Therefore, a crawler must efficiently index URLs as well as already visited pages
- To determine if a URL has already been seen,
 - Must store URLs in a standard format (discussed ahead)
 - Must develop a fast way to check if a URL has already been seen
- To determine if a new page has already been seen,
 - Must develop a fast way to determine if an *identical* page was already indexed
 - Must develop a fast way to determine if a *near-identical* page was already indexed



Link Extraction

- **Must find all links in a page and extract URLs;**

```
var links = document.querySelectorAll("a");
for (var i = 0; i < links.length; i++) {
    var link = links[i].getAttribute("href");
    console.log(link); }
```

- But URLs occur in tags other than <a>, e.g.
- <frame src="site-index.html">, <area href="...">, <meta>, <link>, <script>

- **Relative URL's must be completed, e.g. using current page URL or <base> tag**

- to http://www.myco.com/special/tools/proj.html
- to http://www.myco.com/special/outline/syllabus.html

- **Two Anomalies**

1. Some anchors don't have links, e.g.
2. Some anchors produce dynamic pages which can lead to looping

Representing URLs

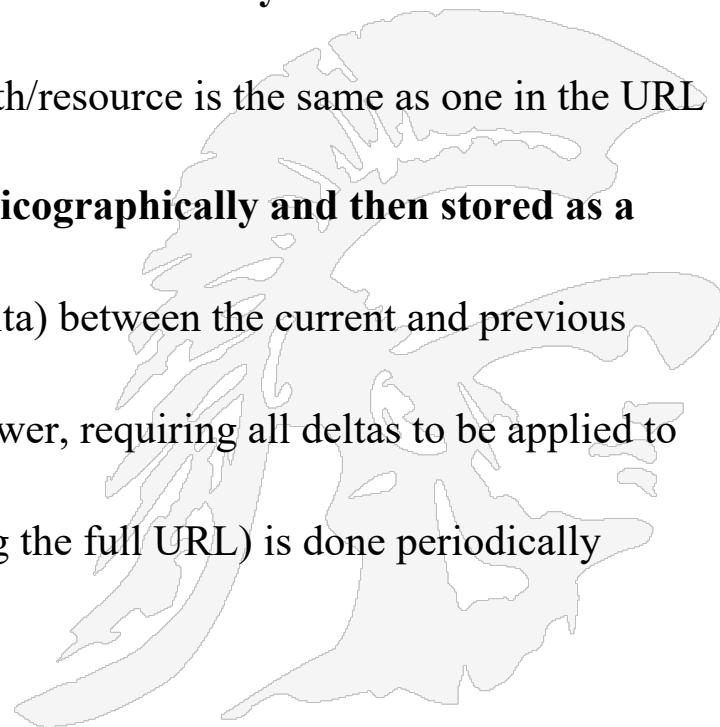
- URLs are rather long, 80 bytes on the average, implying 1 trillion URLs will require 80 Terabytes
 - Recently Google reported finding 30 trillion unique URLs, which by the above would require 2400 terabytes (or 2.4 petabytes) to store

1. One Proposed Method: To determine if a new URL has already been seen

- First hash on host/domain name, then
- Use a trie data structure to determine if the path/resource is the same as one in the URL database

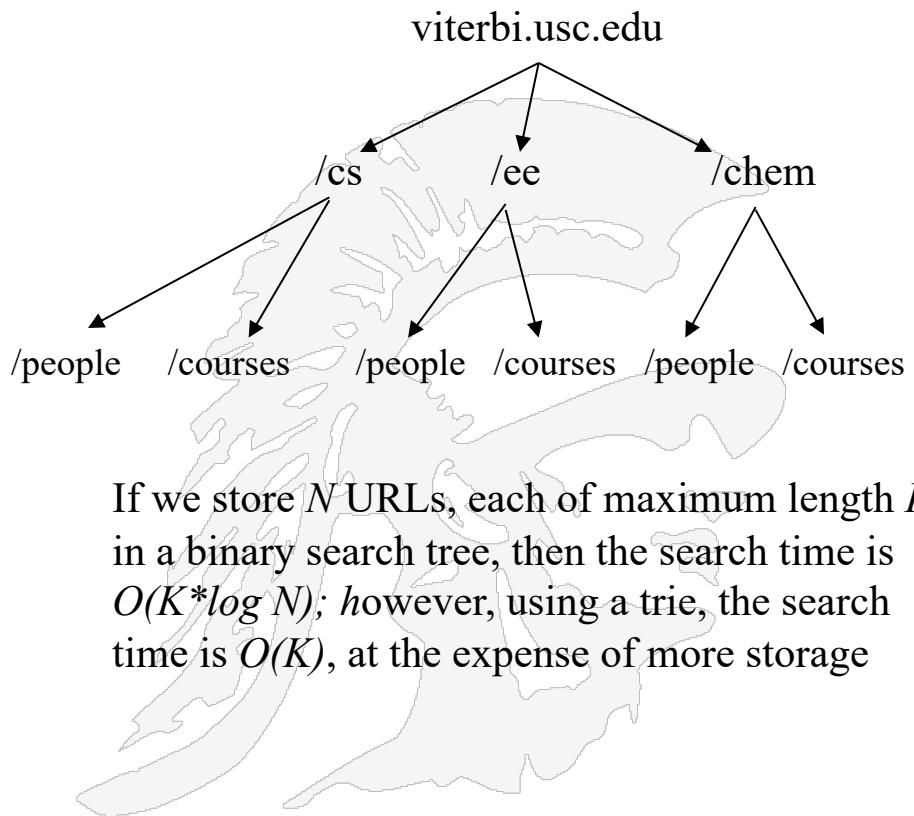
2. Another Proposed Method: URLs are sorted lexicographically and then stored as a delta-encoded text file

- Each entry is stored as the difference (delta) between the current and previous URL; this substantially reduces storage
- However, restoring the actual URL is slower, requiring all deltas to be applied to the initial URL
- To improve speed, checkpointing (storing the full URL) is done periodically



Trie for URL Exact Matching

- Simplest (and worst) algorithm to determine if a new URL is in your set
 - `grep -i <search_url> <url_file>`
 - For N URLs and maximum length K , time is $O(NK)$
- Characteristics of tries
 - They share the same prefix among multiple “words”
 - Each path from the root to a leaf corresponds to one “word”
 - *Endmarker symbol, \$, at the ends of all words*
 - To avoid confusion between words with almost identical elements
 - Assume all words are \$ terminated



Why Normalizing URLs is Important

- For example, all the following URLs have the same meaning (return the same web page), but different hashes:
 - `http://www.google.com`
 - `http://www.google.com/`
 - `https://www.google.com`
 - `www.google.com`
 - `google.com`
 - `google.com/`
 - `google.com.`



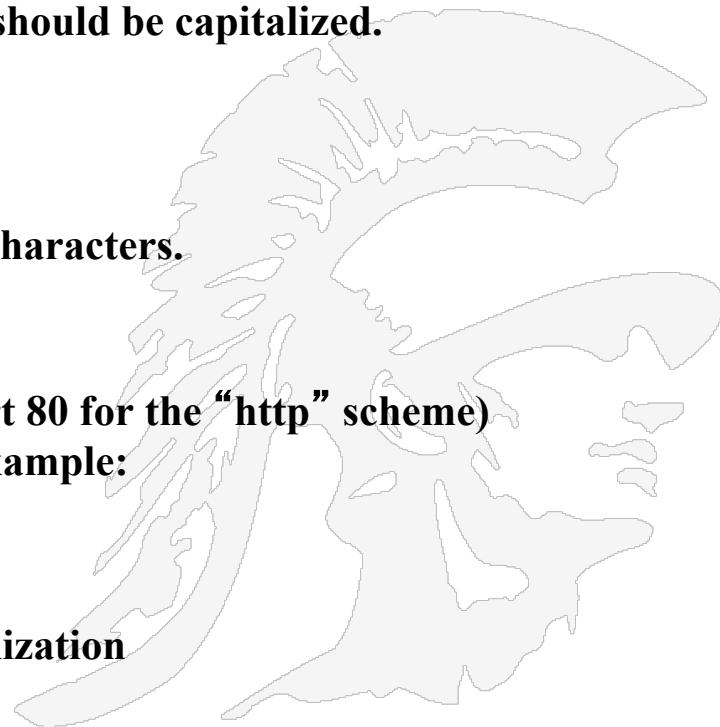
Normalizing URLs (4 rules)

1. Convert the scheme and host to lower case. The scheme and host components of the URL are case-insensitive.
 - HTTP://www.Example.com/ → http://www.example.com/
2. Capitalize letters in escape sequences. All letters within a percent-encoding triplet (e.g., "%3A") are case-insensitive, and should be capitalized.

Example:

 - http://www.example.com/a%c2%b1b →
http://www.example.com/a%C2%B1b
3. Decode percent-encoded octets of unreserved characters.

http://www.example.com/%7Eusername/ →
http://www.example.com/~username/
4. Remove the default port. The default port (port 80 for the “http” scheme) may be removed from (or added to) a URL. Example:
 - http://www.example.com:80/bar.html →
http://www.example.com/bar.html
 - See https://en.wikipedia.org/wiki/URL_normalization



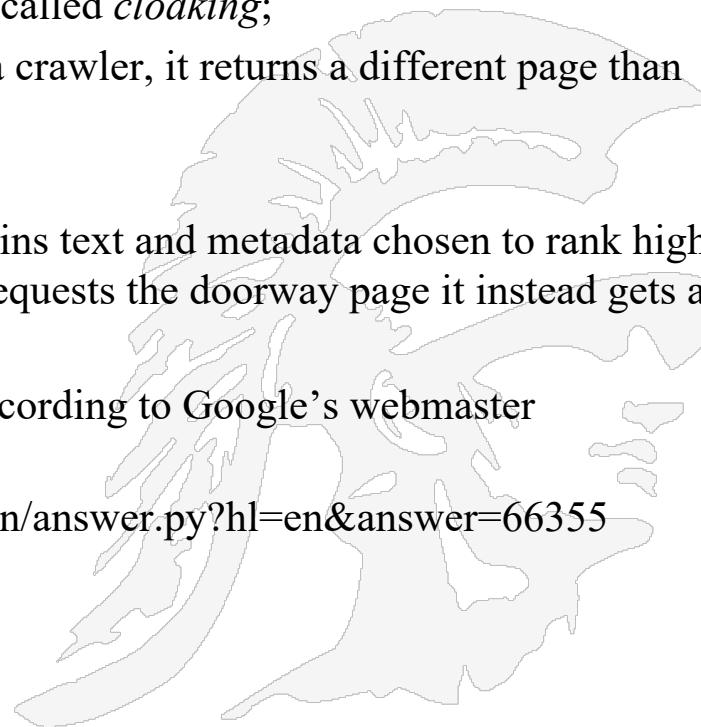
Avoiding Spider Traps

- A spider trap is when a crawler re-visits the same page over and over again
- The most well-known spider trap is the one created by the use of Session ID's
 - J2EE, ASP, .NET, and PHP all provide session ID management
- A Session ID is often used to keep track of visitors, and some sites puts a unique ID in the URL:
 - An example is www.webmasterworld.com/page.php?id=264684413484654 (**Note** this URL doesn't exist).
Each user gets a unique ID and it's often requested from each page.
The problem here is when Googlebot comes to the page, it spiders the page and then leaves, it goes to another page and it finds a link to the previous page, but since it has been given a different session id now, the link shows up as another URL.

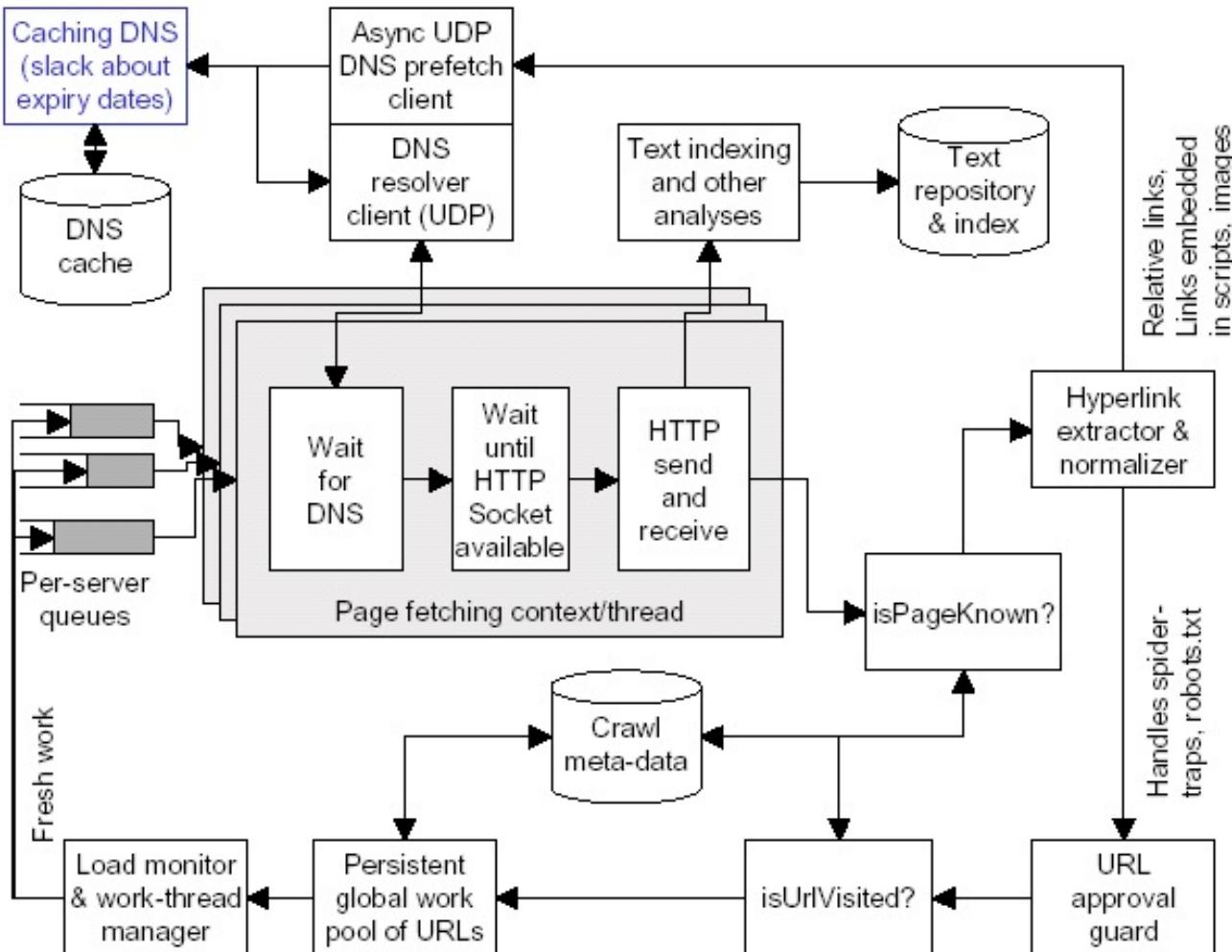
- One way to avoid such traps is for the crawler to be careful when the querystring "ID=" is present in the URL
- Another technique is to monitor the length of the URL, and stop if the length gets "too long"

Handling Spam Web Pages

- The **first generation** of spam web pages consisted of pages with a high number of repeated terms, so as to score high on search engines that ranked by word frequency
 - Words were typically rendered in the same color as the background, so as to not be visible, but still count
- The **second generation** of spam used a technique called *cloaking*:
 - When the web server detects a request from a crawler, it returns a different page than the page it returns from a user request
 - The page is mistakenly indexed
- A **third generation**, called a doorway page, contains text and metadata chosen to rank highly on certain search keywords, but when a browser requests the doorway page it instead gets a more “commercially oriented” (more ads) page
- **Cloaking** and **doorway pages** are not permitted according to Google’s webmaster suggestions
 - See <http://support.google.com/webmasters/bin/answer.py?hl=en&answer=66355>



The Mercator Web Crawler

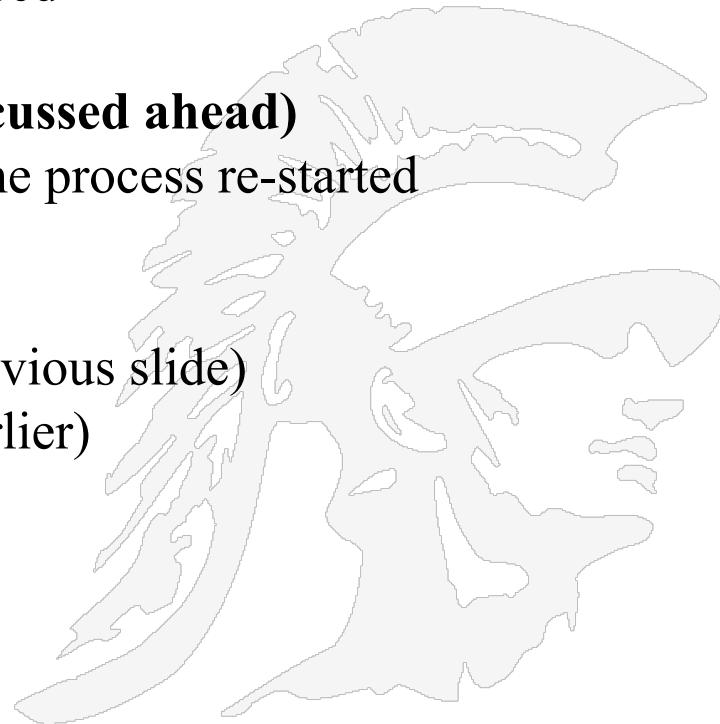


The diagram points out all of the key elements of a crawler;
Notice

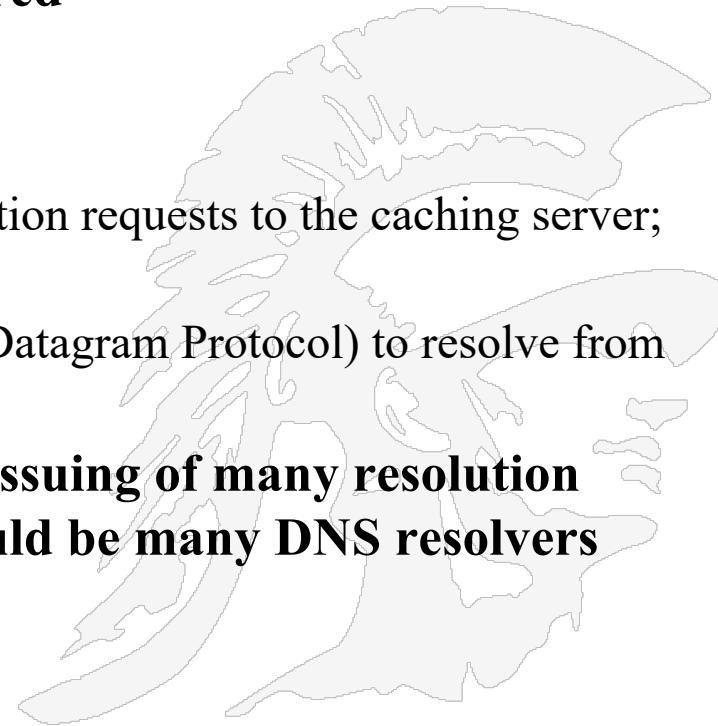
1. The DNS caching server
2. Use of UDP for DNS
3. Load and thread monitor
4. Parallel threads waiting for a page to download

Measuring and Tuning a Crawler

- **Measuring and tuning a crawler for peak performance eventually reduces to**
 - Improving URL parsing speed
 - Improving network bandwidth speed
 - Improving fault tolerance
- **More Issues (some of which are discussed ahead)**
 - Refresh Strategies: how often is the process re-started
 - Detecting duplicate pages
 - Detecting mirror sites
 - Speeding up DNS lookup (see previous slide)
 - URL normalization (discussed earlier)
 - Handling malformed HTML

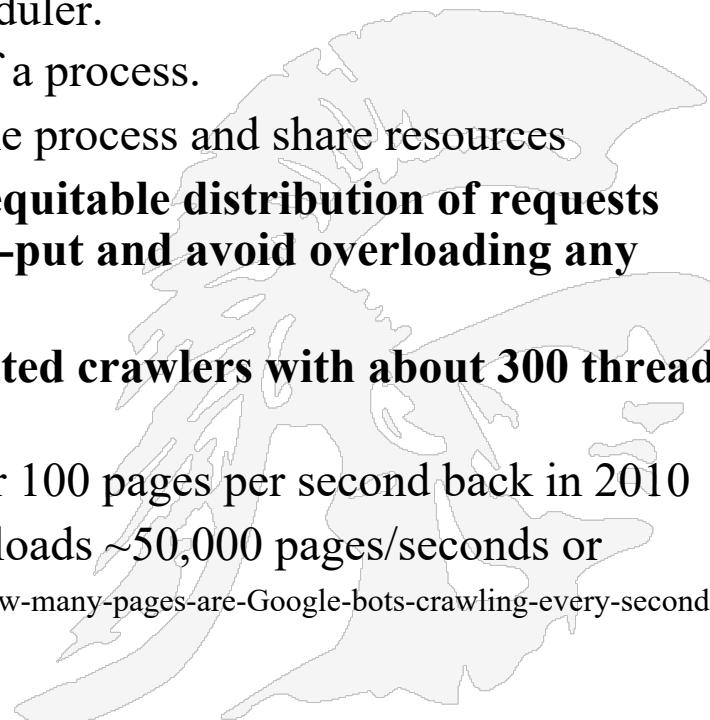


- *A common operating system's implementation of DNS lookup is blocking: only one outstanding request at a time; so*
- 1. **DNS caching:** build a caching server that retains IP-domain name mappings previously discovered
- 2. **Pre-fetching client**
 - once a page is parsed,
 - immediately make DNS resolution requests to the caching server; and
 - if unresolved, use UDP (User Datagram Protocol) to resolve from the DNS server
- 3. **Customize the crawler so it allows issuing of many resolution requests simultaneously; there should be many DNS resolvers**



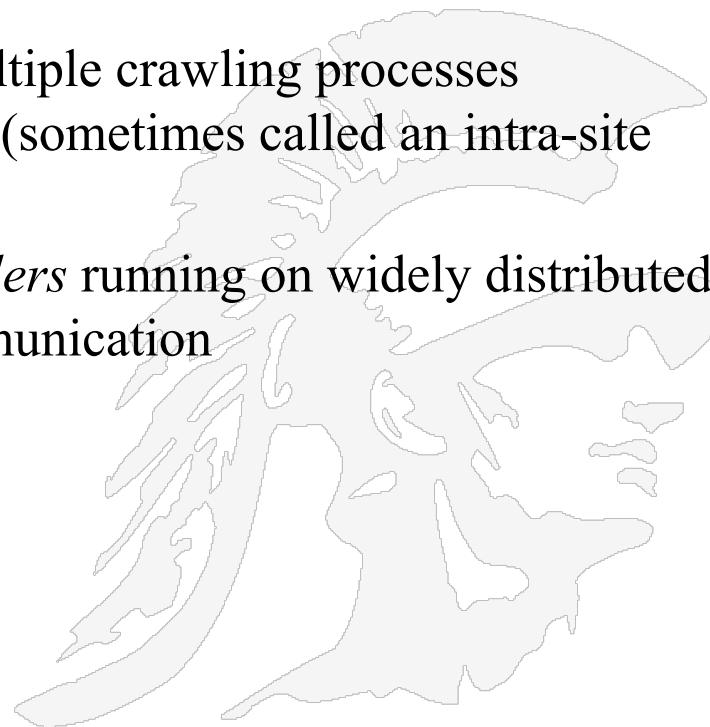
Multi-Threaded Crawling

- One bottleneck is network delay in downloading individual pages.
- It is best to have multiple threads running in parallel each requesting a page from a different host.
 - a **thread** of execution is the smallest sequence of programmed instructions that can be managed independently by a scheduler.
 - In most cases, a thread is a component of a process.
 - Multiple threads can exist within the same process and share resources
- Distribute URL's to threads to guarantee equitable distribution of requests across different hosts to maximize through-put and avoid overloading any single server.
- Early Google spider had multiple coordinated crawlers with about 300 threads each,
 - together they were able to download over 100 pages per second back in 2010
 - It is estimated that in 2021 Google downloads ~50,000 pages/seconds or 4billion+ in a day, see <https://www.quora.com/How-many-pages-are-Google-bots-crawling-every-second>



Distributed Crawling Approaches

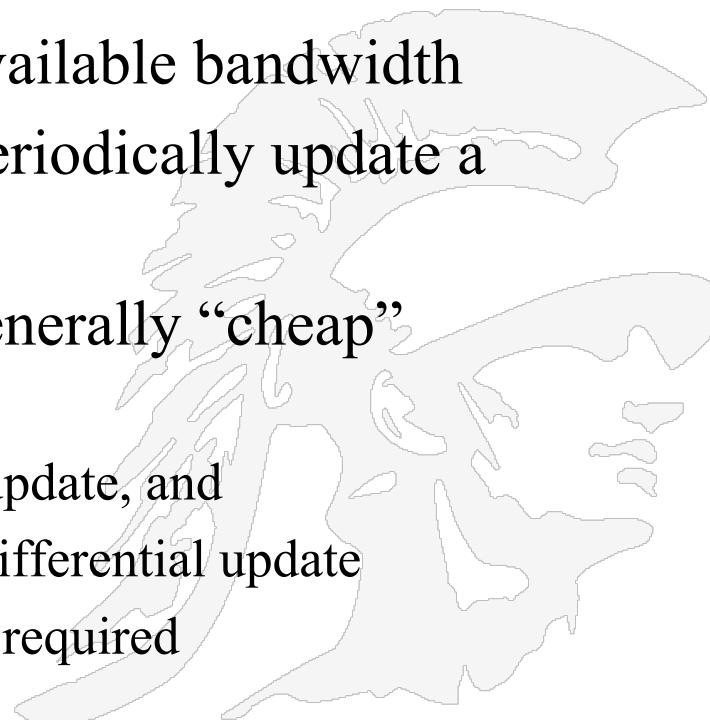
- Once the crawler program itself has been optimized, the next issue to decide is how many crawlers will be running at any time
- **Scenario 1:** A *centralized crawler* controlling a set of parallel crawlers all running on a LAN
 - A *parallel crawler* consists of multiple crawling processes communicating via local network (sometimes called an intra-site parallel crawler)
- **Scenario 2:** A *distributed set of crawlers* running on widely distributed machines, with or without cross communication



Distributed Model

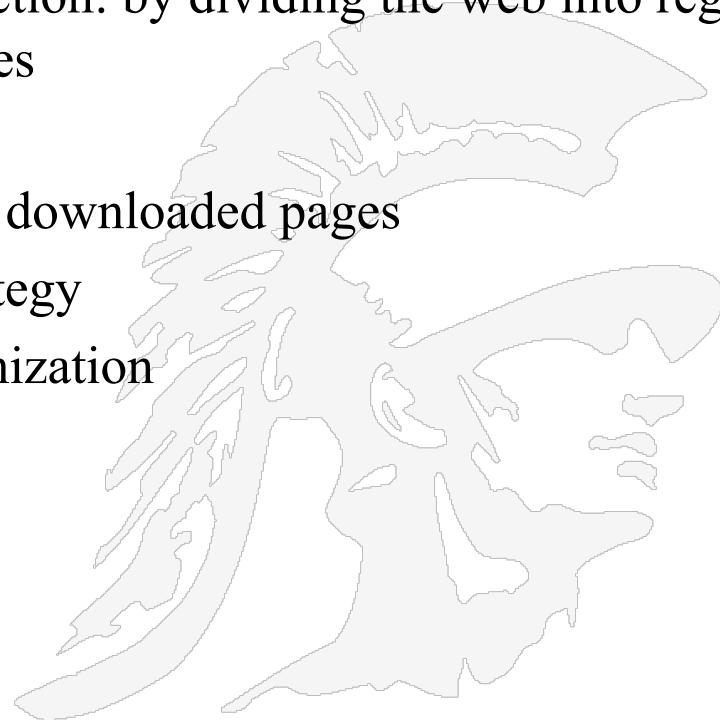
- If crawlers are running in diverse geographic locations, how do we organize them

- By country, by region, by available bandwidth
- Distributed crawlers must periodically update a master index
- But incremental update is generally “cheap”
 - Why? Because
 - a. you can compress the update, and
 - b. you need only send a differential update both of which will limit the required communication



Issues and Benefits of Distributed Crawling

- **Benefits:**
 - scalability: for large-scale web-crawls
 - costs: use of cheaper machines
 - network-load dispersion and reduction: by dividing the web into regions and crawling only the nearest pages
- **Issues:**
 - overlap: minimization of multiple downloaded pages
 - quality: depends on the crawl strategy
 - communication bandwidth: minimization



Coordination of Distributed Crawling

– Three strategies

1. Independent:

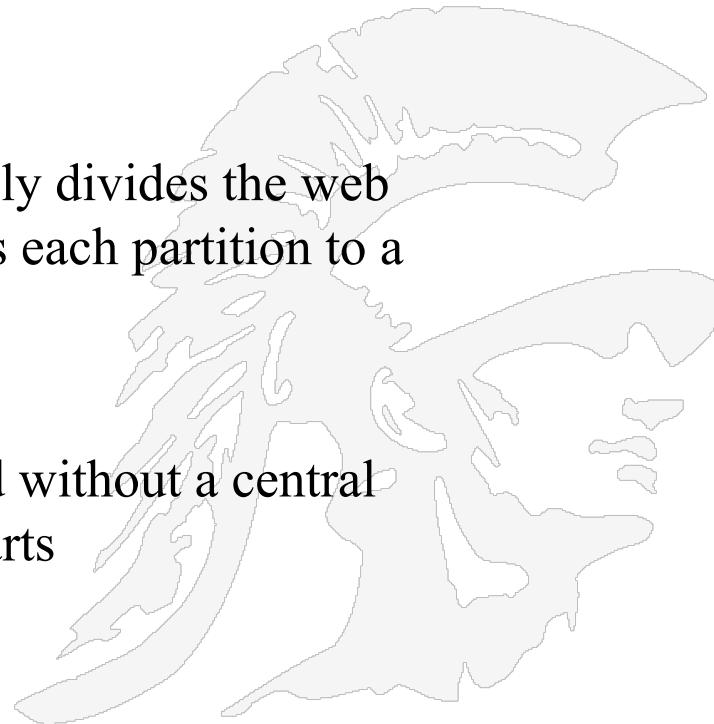
- ▶ no coordination, every process follows its extracted links

2. Dynamic assignment:

- ▶ a central coordinator dynamically divides the web into small partitions and assigns each partition to a process

3. Static assignment:

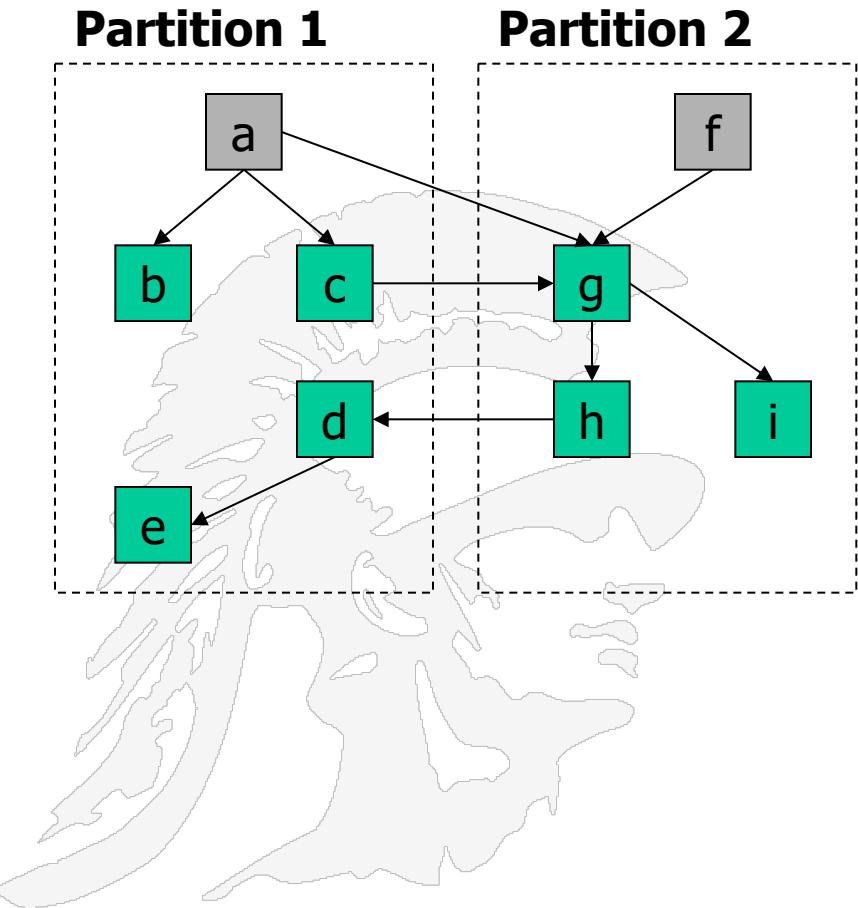
- ▶ Web is partitioned and assigned without a central coordinator before the crawl starts



Static Assignment

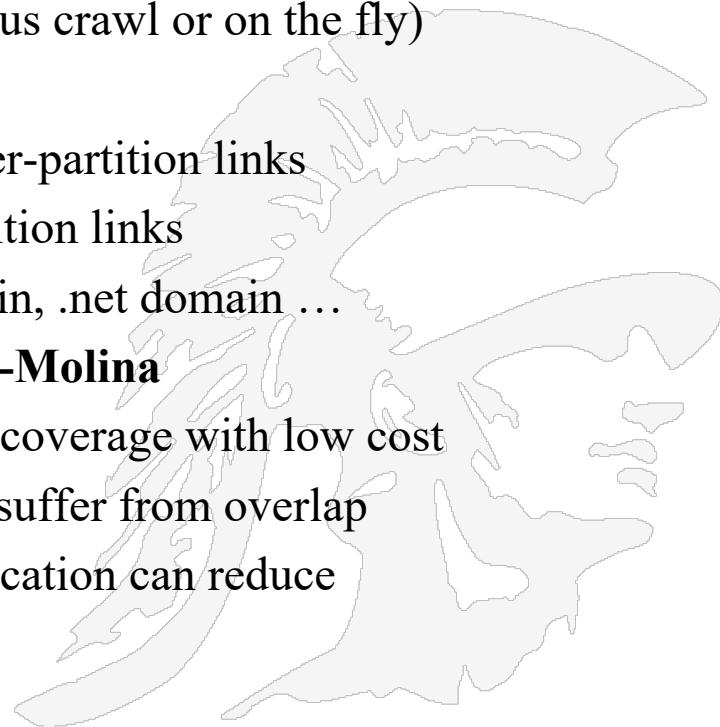
Links from one partition to another (inter-partition links) can be handled in one of three ways:

1. *Firewall mode:*
a process does not follow any inter-partition link
2. *Cross-over mode:*
a process also follows inter-partition links and possibly discovers also more pages in its partition
3. *Exchange mode:*
processes exchange inter-partition URLs; this mode requires communication

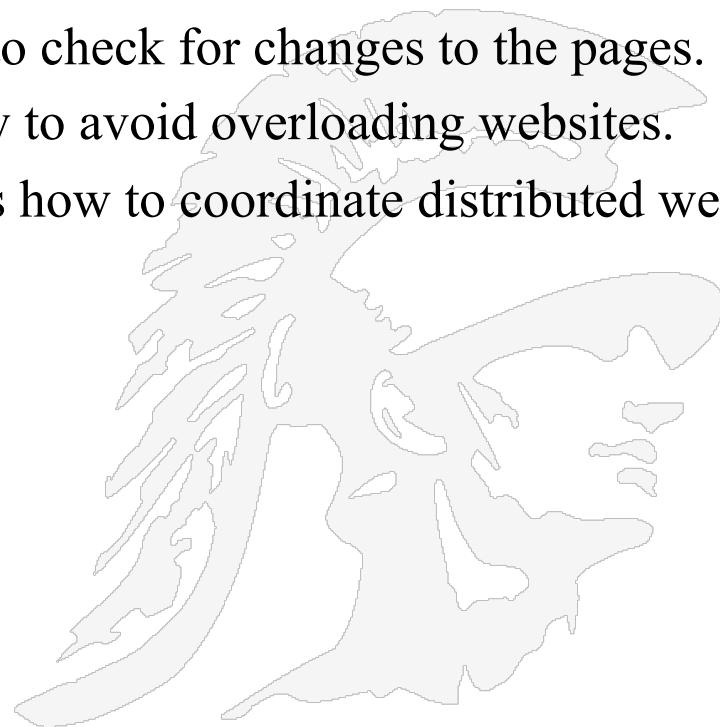


Classification of Parallel Crawlers

- If exchange mode is used, communication can be limited by:
 - Batch communication: every process collects some URLs and sends them in a batch
 - Replication: the k most popular URLs are replicated at each process and are not exchanged (previous crawl or on the fly)
- Some ways to partition the Web:
 - URL-hash based: this yields many inter-partition links
 - Site-hash based: reduces the inter partition links
 - Hierarchical: by TLD, e.g. .com domain, .net domain ...
- General Conclusions of Cho and Garcia-Molina
 - Firewall crawlers attain good, general coverage with low cost
 - Cross-over ensures 100% quality, but suffer from overlap
 - Replicating URLs and batch communication can reduce overhead

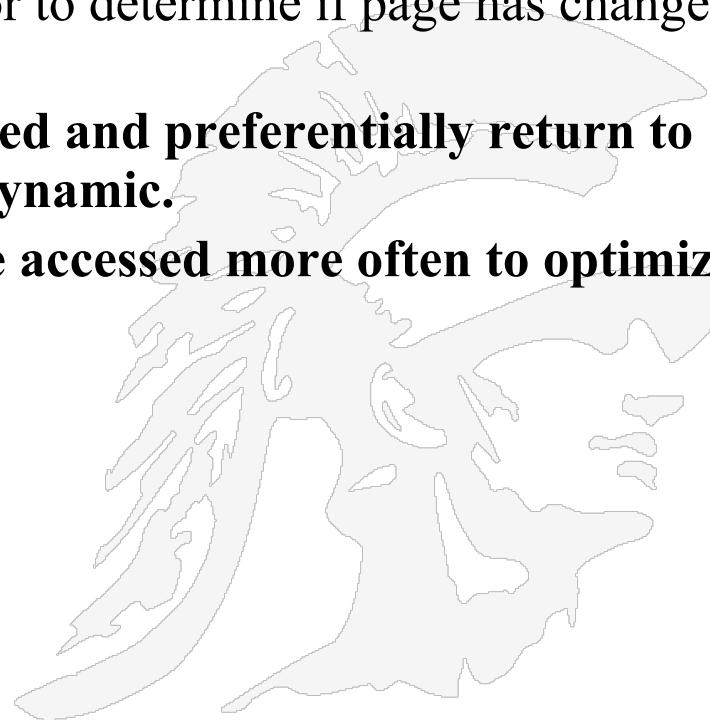


- **The behavior of a Web crawler is the outcome of a combination of policies:**
 - A *selection policy* that states which pages to download.
 - A *re-visit policy* that states when to check for changes to the pages.
 - A *politeness policy* that states how to avoid overloading websites.
 - A *parallelization policy* that states how to coordinate distributed web crawlers.



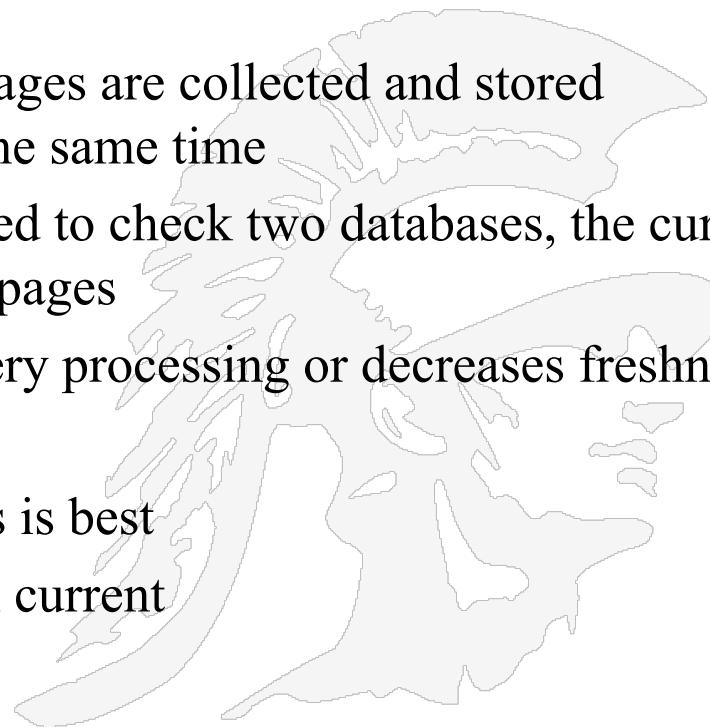
Keeping Spidered Pages Up to Date

- Web is very dynamic: many new pages, updated pages, deleted pages, etc.
- Periodically check crawled pages for updates and deletions:
 - Just look at LastModified indicator to determine if page has changed, only reload entire page if needed
- Track how often each page is updated and preferentially return to pages which are historically more dynamic.
- Preferentially update pages that are accessed more often to optimize freshness of more popular pages.



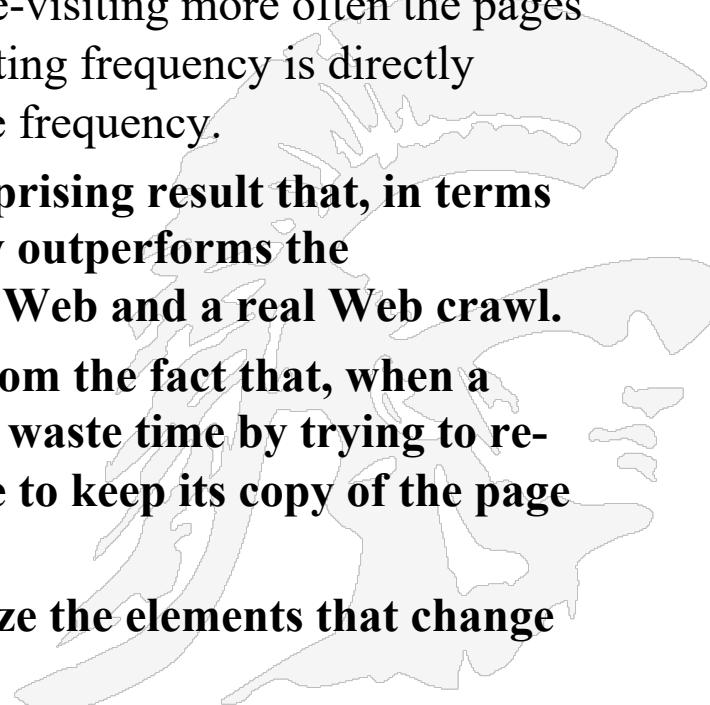
Implications for a Web Crawler

- **A *steady crawler* runs continuously without pause**
 - Typically search engines use multiple crawlers
- **When a crawler replaces an old version by a new page, does it do it “in-place” or “shadowing”**
 - Shadowing implies a new set of pages are collected and stored separately and all are updated at the same time
 - The above implies that queries need to check two databases, the current database and the database of new pages
 - Shadowing either slows down query processing or decreases freshness
- **Conclusions:**
 - running multiple types of crawlers is best
 - Updating in-place keeps the index current



Cho and Garcia-Molina, 2000

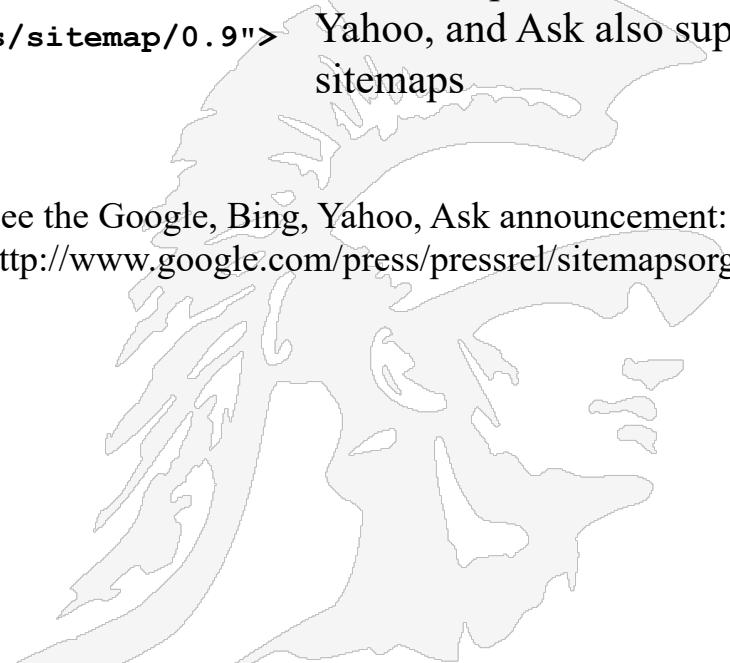
- Two simple re-visiting policies
 - Uniform policy: This involves re-visiting all pages in the collection with the same frequency, regardless of their rates of change.
 - Proportional policy: This involves re-visiting more often the pages that change more frequently. The visiting frequency is directly proportional to the (estimated) change frequency.
- Cho and Garcia-Molina proved the surprising result that, in terms of average freshness, the uniform policy outperforms the proportional policy in both a simulated Web and a real Web crawl.
- The explanation for this result comes from the fact that, when a page changes too often, the crawler will waste time by trying to re-crawl it too fast and still will not be able to keep its copy of the page fresh.
- To improve freshness, we should penalize the elements that change too often



Help the Search Engine Crawler Creating a SiteMap

- A sitemap is a list of pages of a web site accessible to crawlers
- This helps search engine crawlers find pages on the site
- XML is used as the standard for representing sitemaps
- Here is an example of an XML sitemap for a three page website

```
<?xml version="1.0" encoding="UTF-8"?>
<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">
<url>
  <loc>http://www.example.com/?id=who</loc>
  <lastmod>2009-09-22</lastmod>
  <changefreq>monthly</changefreq>
  <priority>0.8</priority> </url>
<url>
  <loc>http://www.example.com/?id=what</loc>
  <lastmod>2009-09-22</lastmod>
  <changefreq>monthly</changefreq>
  <priority>0.5</priority> </url>
<url>
  <loc>http://www.example.com/?id=how</loc>
  <lastmod>2009-09-22</lastmod>
  <changefreq>monthly</changefreq>
  <priority>0.5</priority> </url>
</urlset>
```



Back in 2006 Google introduced the sitemap format; now Bing, Yahoo, and Ask also support sitemaps

See the Google, Bing, Yahoo, Ask announcement:
<http://www.google.com/press/pressrel/sitemapsorg.html>

Google Crawlers

- Google now uses multiple crawlers
 - APIs-Google
 - AdSense
 - AdsBot Mobile Web Android
 - AdsBot Mobile Web
 - AdsBot
 - Googlebot Images
 - Googlebot News
 - Googlebot Video
 - Googlebot (desktop)
 - Googlebot (smartphone)
 - Mobile AdSense
 - Mobile Apps Android
 - Feedfetcher
 - Google Read Aloud

Crawler	User agent token (product token)	Full user agent string
APIs-Google	APIs-Google	APIs-Google (+https://developers.google.com/webmasters/APIs-Google.html)
AdSense	Mediapartners-Google	Mediapartners-Google
AdsBot Mobile Web Android	AdsBot-Google-Mobile	Mozilla/5.0 (Linux; Android 5.0; SM-G920A) AppleWebKit (KHTML, like Gecko) Chrome Mobile Safari (compatible; AdsBot-Google-Mobile; +http://www.google.com/mobile/adsbot.html)
AdsBot Mobile Web	AdsBot-Google-Mobile	Mozilla/5.0 (iPhone; CPU iPhone OS 9_1 like Mac OS X AppleWebKit/601.1.46 (KHTML, like Gecko) Version/9.1.1 Safari/601.1 (compatible; AdsBot-Google-Mobile; +http://www.google.com/mobile/adsbot.html)
AdsBot	AdsBot-Google	AdsBot-Google (+http://www.google.com/adsbot.html)
Googlebot Images	<ul style="list-style-type: none"> • Googlebot-Image • Googlebot 	Googlebot-Image/1.0
Googlebot News	<ul style="list-style-type: none"> • Googlebot-News • Googlebot 	Googlebot-News
Googlebot Video	<ul style="list-style-type: none"> • Googlebot-Video • Googlebot 	Googlebot-Video/1.0
Googlebot (Desktop)	Googlebot	<ul style="list-style-type: none"> • Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)

For details see

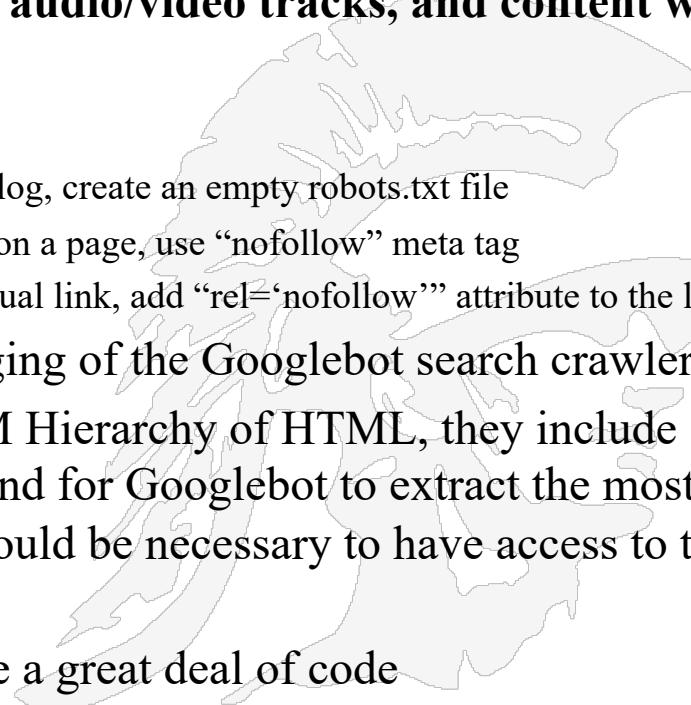
<https://support.google.com/webmasters/answer/1061943?hl=en>

see also Google's tool for checking how Googlebot sees your website

<https://support.google.com/webmasters/answer/6066468?rd=2>

Google's Googlebot

- Begins with a list of webpage URLs generated from previous crawls
- Uses Sitemap data provided by webmasters
- Many versions of Googlebot are run on multiple machines located near the site they are indexing
- Googlebot cannot see within Flash files, audio/video tracks, and content within programs
- Advice
 - To prevent “File not found” in a website’s error log, create an empty robots.txt file
 - To prevent Googlebot from following any links on a page, use “nofollow” meta tag
 - To prevent Googlebot from following an individual link, add “rel=“nofollow”” attribute to the link
- Assertion: Chrome was in fact a repackaging of the Googlebot search crawler;
- Why: browsers don’t just render the DOM Hierarchy of HTML, they include transformations via CSS and JavaScript, and for Googlebot to extract the most meaningful features from a web page it would be necessary to have access to these transformations
- Conclusion: Googlebot and Chrome share a great deal of code

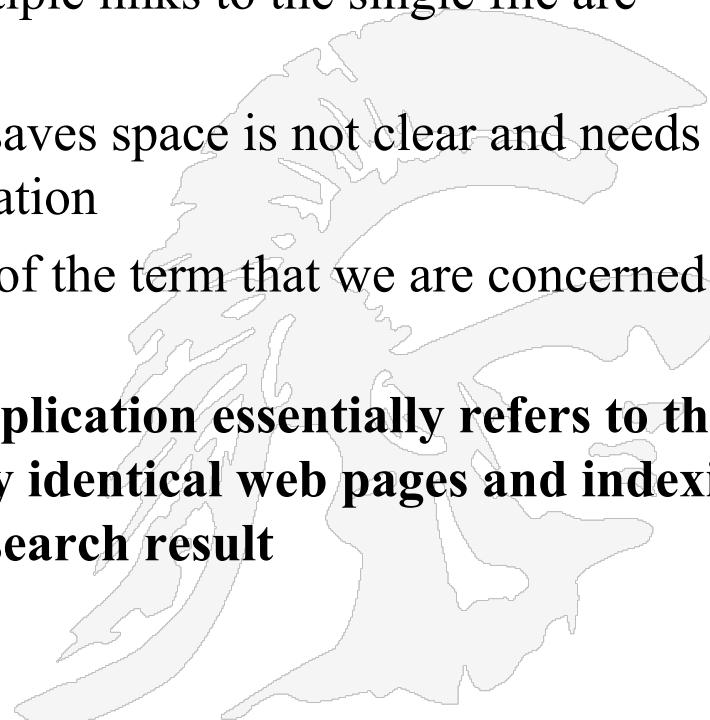


De-Duplication



The Definition of De-Duplication

- *De-Duplication* – the process of identifying and avoiding essentially identical web pages
- The term is often used in connection with *locker storage* where only a single copy of a file is stored and multiple links to the single file are managed
 - Whether this strategy effectively saves space is not clear and needs analysis for each particular application
 - However, this is *not* the meaning of the term that we are concerned about in this class
- **With respect to *web crawling*, de-duplication essentially refers to the identification of identical and nearly identical web pages and indexing only a single version to return as a search result**

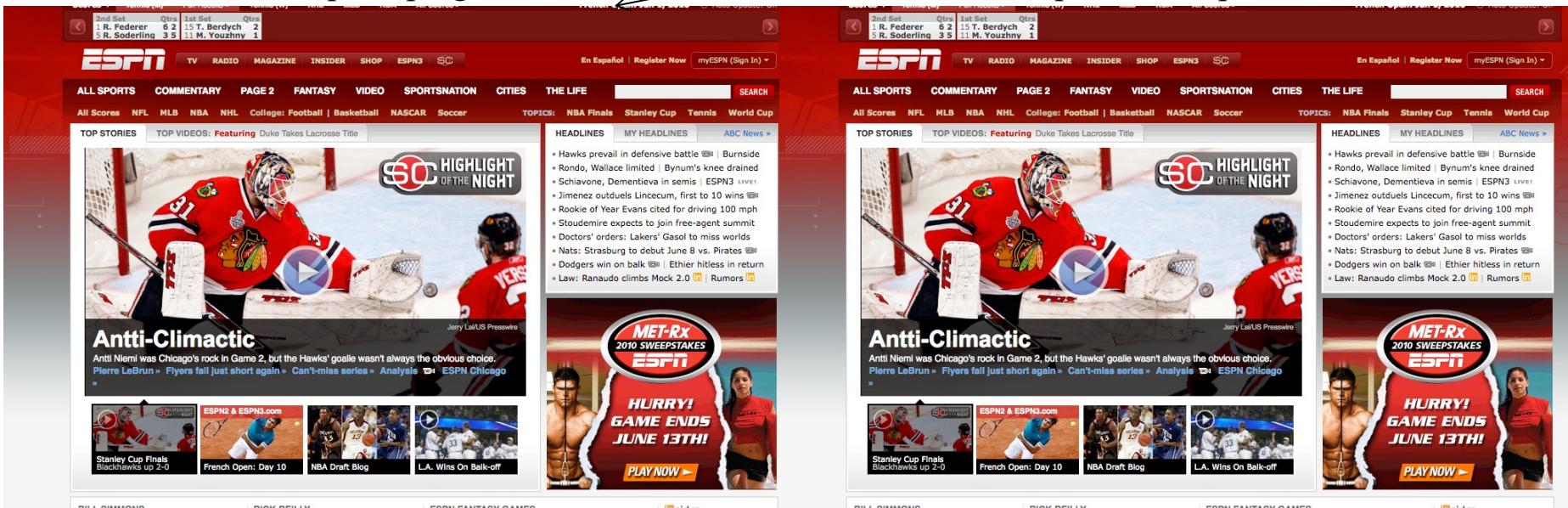


What are Web Duplicates?

- One example is the same page, referenced by different URLs

<http://espn.go.com>

<http://www.espn.com>



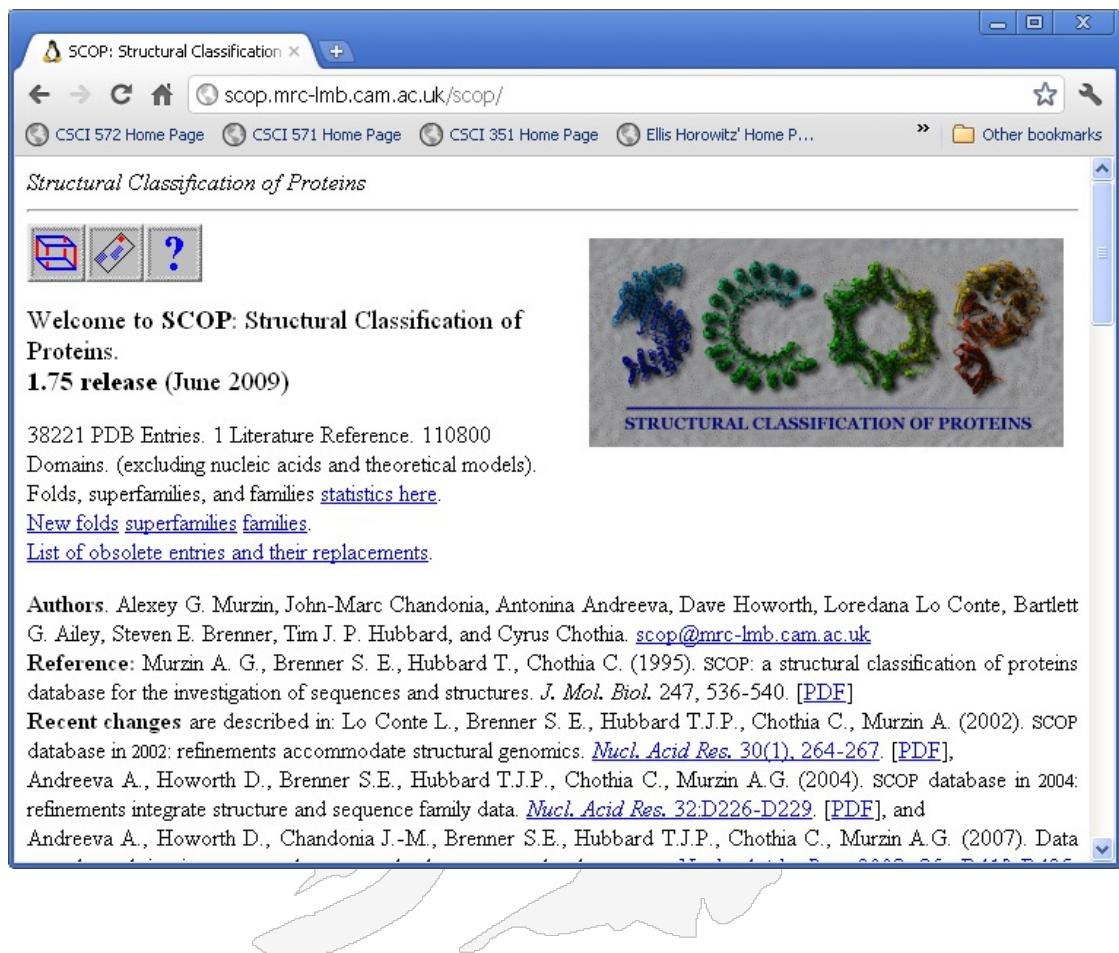
- How can two URLs differ yet still point to the same page?

- the URL's host name can be distinct (virtual hosts) sharing the same document folder,
- the URL's protocol can be distinct (http, https), but still deliver the same document
- the URL's path and/or page name can be distinct

Distinct URLs May Deliver the Same Page

- At one time* all 3 URLs below pointed to the identical page
- Structural Classification of Proteins
 - <http://scop.mrc-lmb.cam.ac.uk/scop>
 - <http://scop.berkeley.edu/>
 - <http://scop.protres.ru/>
- **The three URLs have distinct domain names, but all redirect to the same page**

* At least they did when I took this snapshot, no longer



Near Duplicates-An Example

Another example is two web pages whose content differs slightly



The New York Times - Breaking News | nytimes.com

brother at your side Printing Solutions for Offices & Small Workgroups Brother Monochrome Laser Printer, HL-L...

Monday, September 7, 2020

World U.S. Politics N.Y. Business Opinion Tech Science Health Sports Arts Books Style Food Travel Magazine T Magazine Real Estate Video

The Book Review Podcast Jeffrey Toobin on President Trump; Elena Ferrante's new novel.

Coronavirus Schools Briefing How is the pandemic reshaping education? Get the latest news.

Listen to New Songs Hear tracks by SZA, Ava Max, Tricky, Bill Callahan and others.

New Delhi  Danish Siddiqui/Reuters

With Washington Deadlocked on Aid, States Reckon With Dire Fiscal Crises

- With federal lawmakers at odds over sending more aid, local officials are slashing funding for everything from orchestra subsidies to composting to education.
- But economists warn that further state spending reductions could prolong the downturn by shaking the confidence of residents, who depend on local services.

For Long-Haulers, Covid-19 Takes a Toll on Mind as Well as Body

Long-haulers, or those with symptoms that persist for months, are turning to support groups where many have shared how their mental health has suffered.

In Recessions, Used Cars Turn Into Hot Commodity

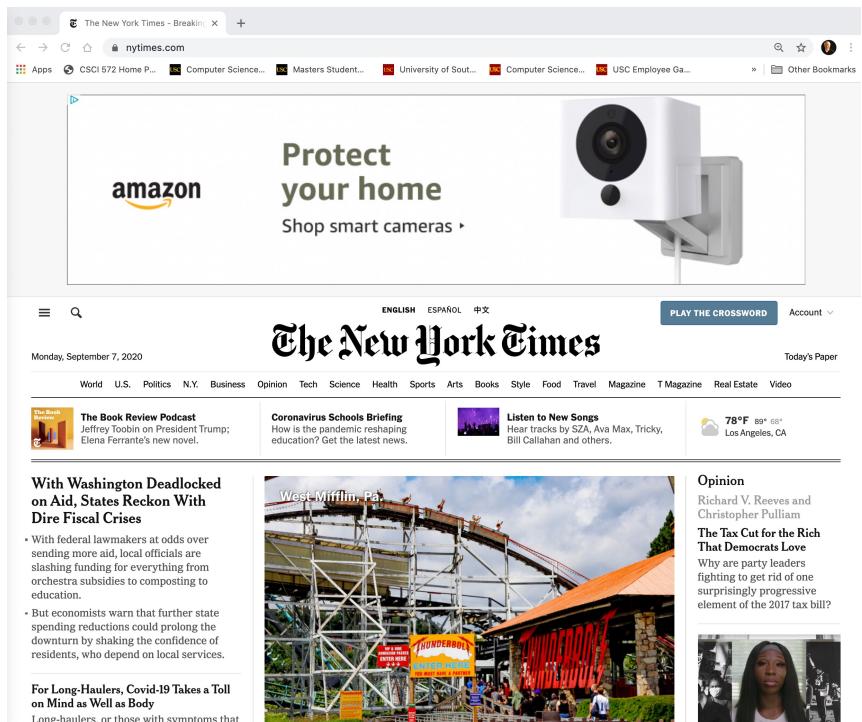
Consumers are buying used cars to avoid public transportation during the pandemic. India has now reached more than 4.2 million cases. Here's the latest.

Live

Photos World Cases U.S. Curve

S&P 500 -0.81% Dow -0.56% Nasdaq -1.27% 78°F 89° 65° Los Angeles, CA

PLAY THE CROSSWORD Account



The New York Times - Breaking News | nytimes.com

amazon Protect your home Shop smart cameras >

Monday, September 7, 2020

World U.S. Politics N.Y. Business Opinion Tech Science Health Sports Arts Books Style Food Travel Magazine T Magazine Real Estate Video

The Book Review Podcast Jeffrey Toobin on President Trump; Elena Ferrante's new novel.

Coronavirus Schools Briefing How is the pandemic reshaping education? Get the latest news.

Listen to New Songs Hear tracks by SZA, Ava Max, Tricky, Bill Callahan and others.

78°F 89° 65° Los Angeles, CA

With Washington Deadlocked on Aid, States Reckon With Dire Fiscal Crises

- With federal lawmakers at odds over sending more aid, local officials are slashing funding for everything from orchestra subsidies to composting to education.
- But economists warn that further state spending reductions could prolong the downturn by shaking the confidence of residents, who depend on local services.

For Long-Haulers, Covid-19 Takes a Toll on Mind as Well as Body

Long-haulers, or those with symptoms that

West Mifflin, Pa. 

Opinion Richard V. Reeves and Christopher Pulliam

The Tax Cut for the Rich That Democrats Love

Why are party leaders fighting to get rid of one surprisingly progressive element of the 2017 tax bill?

Pico Iyer The Best Reason to Go to College

Adam Grant and Allison Sweet Grant Kids Can Learn to Love Learning, Even Over

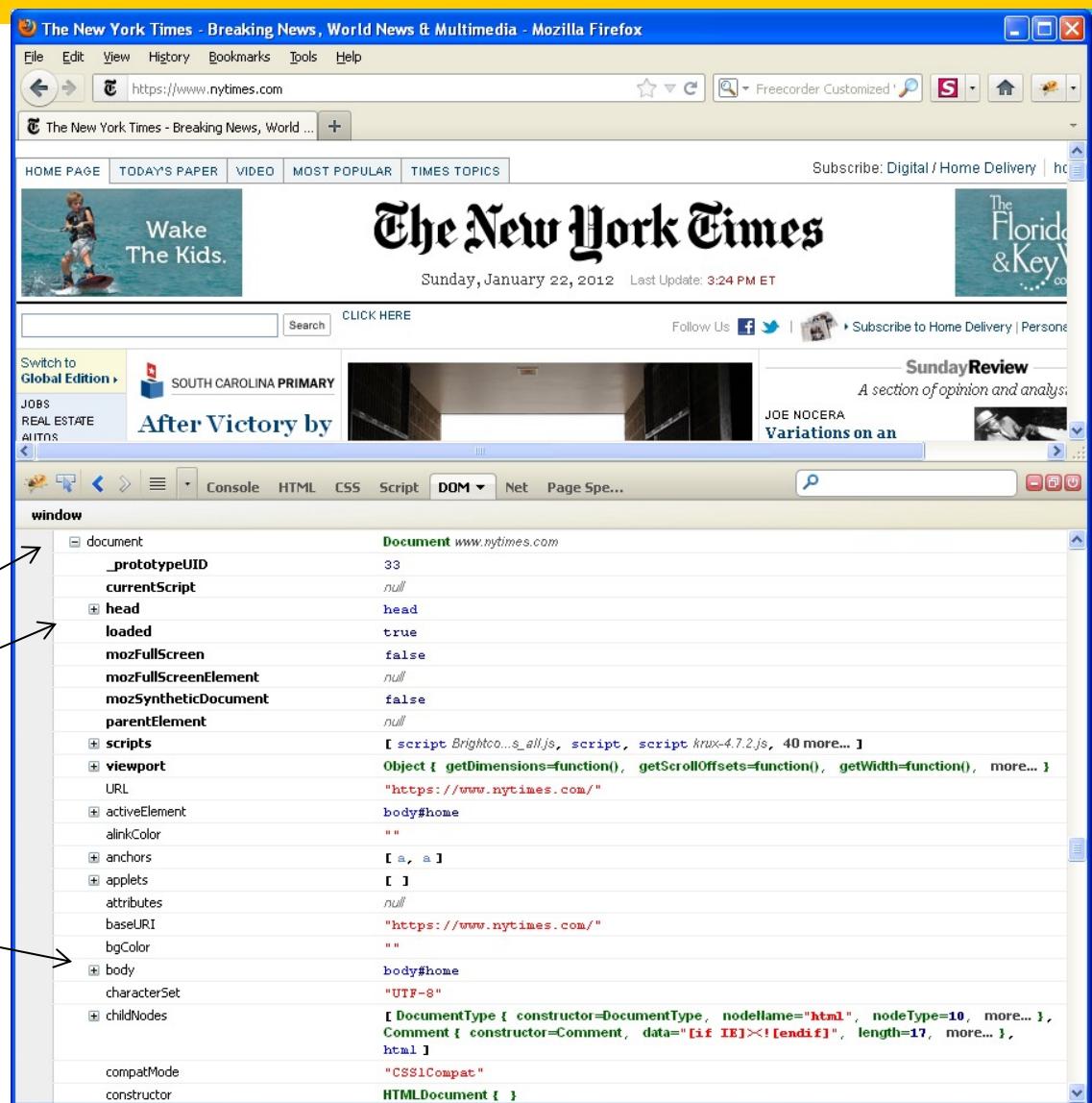
The Editorial Board How to Fix New York's \$5 Billion Budget Crisis

PLAY THE CROSSWORD Account

Two copies of www.nytimes.com snapshot within a few seconds of each other;
The pages are essentially identical except for the ads at the top and the photo in the middle

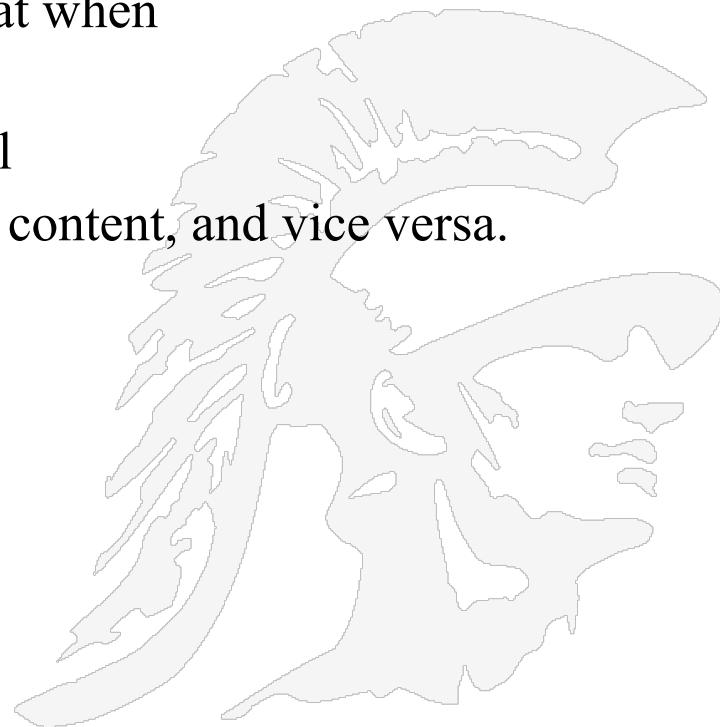
Spotting Near Duplicates

- In examining a web page a search engine may want to ignore ads, navigation links and other elements that do not specifically relate to the contents of the web page
- One way to do this is to delve into the structure of a web page and focus on content blocks
- E.g. the Document Object Model for HTML displays a web page as a tree hierarchy
 - Document
 - Head
 - Body
- However this is time consuming



Web Duplicates-Mirroring

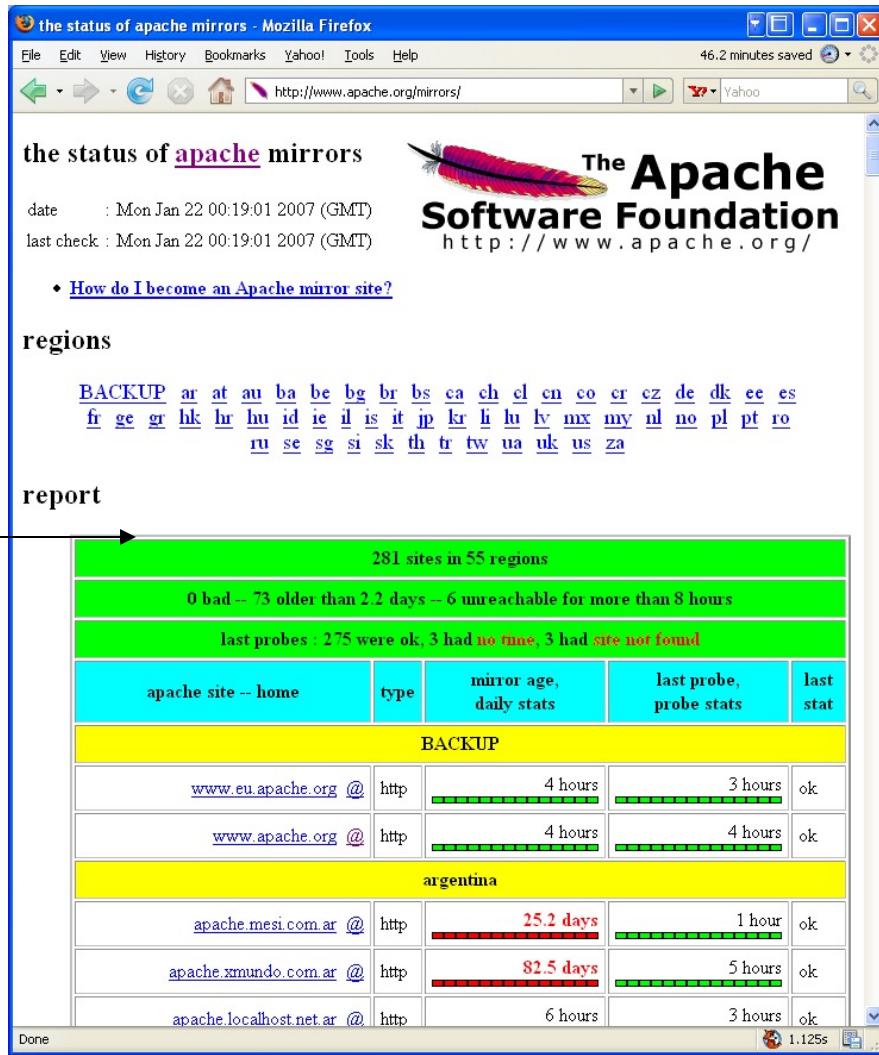
- Mirroring is the systematic replication of web pages across hosts.
 - Mirroring is the **single largest cause** of duplication on the web
- Host1/ α and Host2/ β are mirrors iff
 - For all (or most) paths p such that when
 $\text{http://Host1/ } \alpha / p$ exists
 - $\text{http://Host2/ } \beta / p$ exists as well
 - with identical (or near identical) content, and vice versa.



List of Apache Mirror Sites

List of countries

281 sites in 55 regions



the status of [apache](#) mirrors

date : Mon Jan 22 00:19:01 2007 (GMT)
 last check : Mon Jan 22 00:19:01 2007 (GMT)

♦ [How do I become an Apache mirror site?](#)

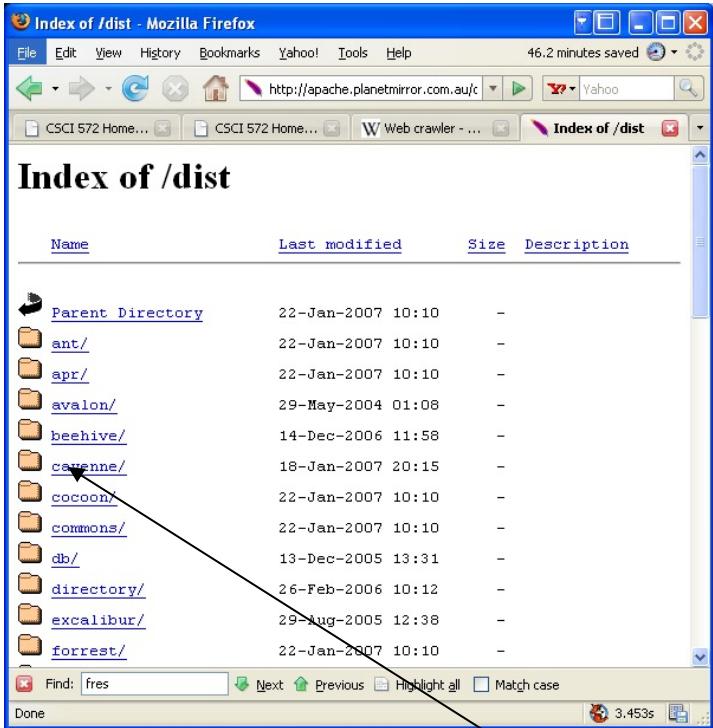
regions

BACKUP	ar	at	au	ba	be	bg	br	bs	ca	ch	cl	cn	co	cr	cz	de	dk	ee	es				
	fr	ge	gr	hk	hr	hu	id	ie	il	is	it	jp	kr	li	lu	lv	mx	my	nl	no	pl	pt	ro
	ru	se	sg	si	sk	th	tr	tw	ua	uk	us	za											

report

281 sites in 55 regions				
0 bad -- 73 older than 2.2 days -- 6 unreachable for more than 8 hours				
last probes : 275 were ok, 3 had no time, 3 had site not found				
apache site -- home	type	mirror age, daily stats	last probe, probe stats	last stat
BACKUP				
www.eu.apache.org @	http	4 hours	3 hours	ok
www.apache.org @	http	4 hours	4 hours	ok
argentina				
apache.mesi.com.ar @	http	25.2 days	1 hour	ok
apache.xmundo.com.ar @	http	82.5 days	5 hours	ok
apache.localhost.net.ar @	http	6 hours	3 hours	ok

Two Sample Mirror Sites



Index of /dist

Name	Last modified	Size	Description
Parent Directory	22-Jan-2007 10:10	-	
ant/	22-Jan-2007 10:10	-	
apr/	22-Jan-2007 10:10	-	
avalon/	29-May-2004 01:08	-	
beehive/	14-Dec-2006 11:58	-	
cavenne/	18-Jan-2007 20:15	-	
cocoon/	22-Jan-2007 10:10	-	
commons/	22-Jan-2007 10:10	-	
db/	13-Dec-2005 13:31	-	
directory/	26-Feb-2006 10:12	-	
excalibur/	29-Aug-2005 12:38	-	
forrest/	22-Jan-2007 10:10	-	

Done

Find: fres Next & Previous Highlight all Match case

3.453s

Site in Australia

Note identical
directories



Apache Software Foundation Distribution Directory

The directories linked below contain current software releases from the Apache Software Foundation projects. Older non-recommended releases can be found on our [archive site](#).

To find the right download for a particular project, you should start at the project's own webpage or on our [project resource listing](#) rather than browsing the links below.

Please do not download from apache.org! If you are currently at apache.org and would like to browse, please instead visit a [nearby mirror site](#).

Projects

Name	Last modified	Size	Description
Parent Directory	04-Dec-2006 14:44	-	
ant/	19-Dec-2006 23:24	-	
apr/	06-Dec-2006 22:56	-	
avalon/	28-May-2004 12:08	-	
beehive/	13-Dec-2006 22:58	-	
cocoon/	22-Dec-2006 04:52	-	
commons/	04-Nov-2003 06:08	-	
db/	13-Dec-2005 00:31	-	
directory/	25-Feb-2006 21:12	-	
excalibur/	28-Aug-2005 23:38	-	
forrest/	23-Jun-2005 22:51	-	

Done

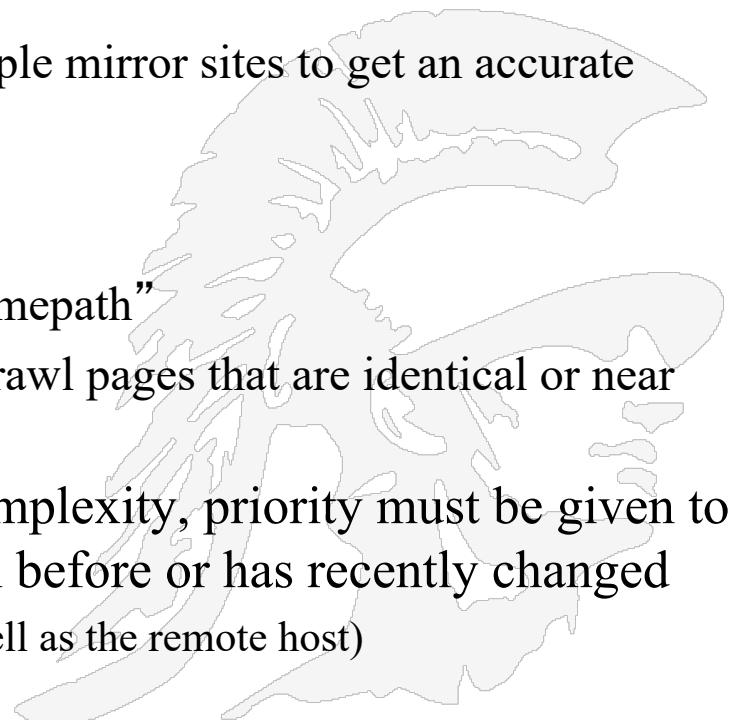
Find: fres Next & Previous Highlight all Match case

0.750s

Site in Argentina

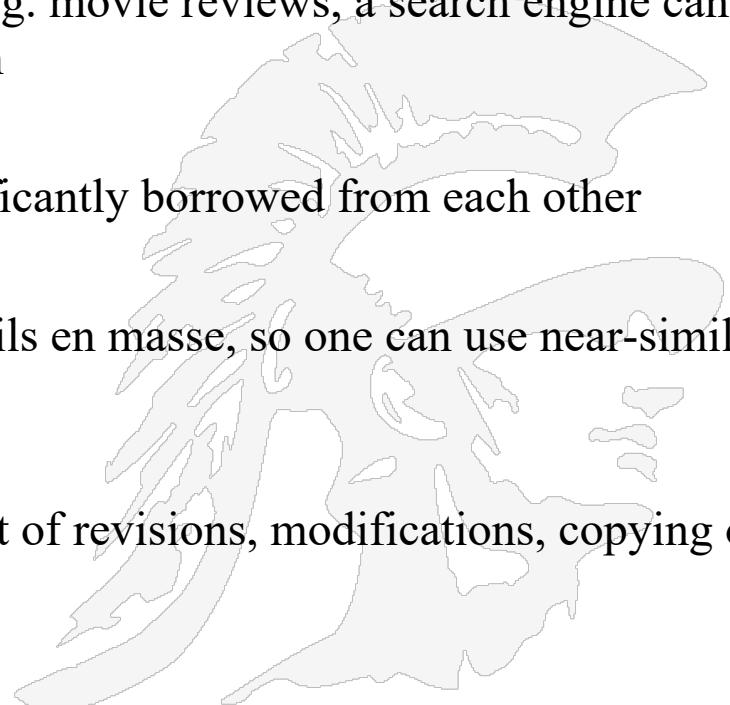
Why Detect Exact Duplicates

- ***Smarter crawling***
 - Avoid returning many duplicate results to a query
 - Allow fetching from the fastest or freshest server
- ***Better connectivity analysis***
 - By combining in-links from the multiple mirror sites to get an accurate PageRank (measure of importance)
 - Avoid double counting out-links
- ***Add redundancy in result listings***
 - “If that fails you can try: <mirror>/samepath”
- ***Reduce Crawl Time***: Crawlers need not crawl pages that are identical or near identical
- ***Ideally***: given the web’s scale and complexity, priority must be given to content that has **not** already been seen before or has recently changed
 - Saves resources (on the crawler end, as well as the remote host)
 - Increases crawler politeness
 - Reduces the analysis that a crawler will have to do later



Why Detect Near Duplicates

- **Clustering**
 - Given a news article some people might wish to see “related articles” describing the same event
- **Data extraction**
 - Given a collection of similar pages, e.g. movie reviews, a search engine can extract and categorize the information
- **Plagiarism**
 - Identify pairs that seem to have significantly borrowed from each other
- **Spam detection**
 - Spammers typically send similar emails en masse, so one can use near-similarity techniques to identify the spam
- **Duplicates within a domain**
 - To identify near-duplicates arising out of revisions, modifications, copying or merging of documents



1. *Duplicate Problem: Exact match;*

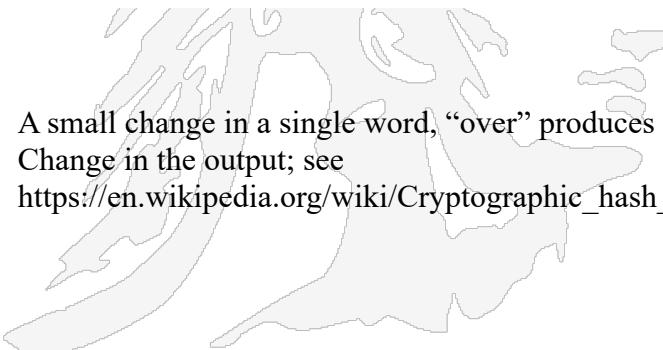
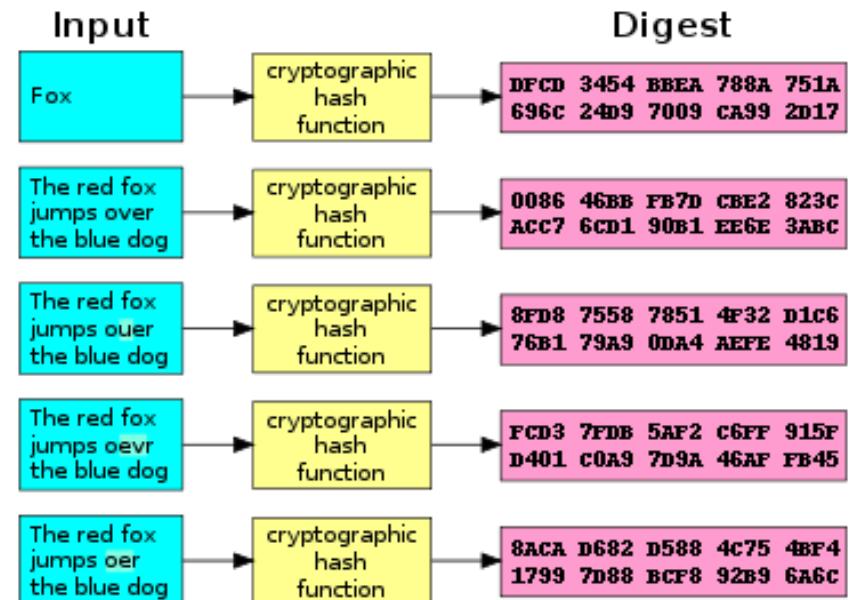
- Solution: compute fingerprints using cryptographic hashing
- SHA-1 and MD5 are the two most popular cryptographic hashing methods
- Most useful for URL matching (see crawling slides), but also works for detecting identical web pages

2. *Near-Duplicate Problem: Approximate match*

- Solution: compute the syntactic similarity with an edit-distance measure, and
- Use a similarity threshold to detect near-duplicates
 - e.g., Similarity > 80% => Documents are “near duplicates”
- The remaining slides are devoted to specific methods for duplicate and near duplicate detection

Using a Cryptographic Hash Function to Convert a Web Page to a Number

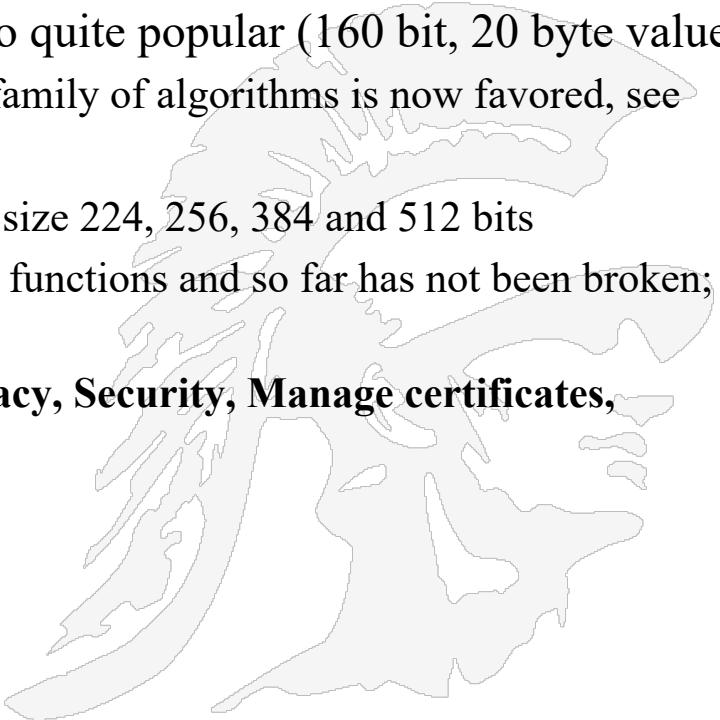
- A **cryptographic hash function** is a hash function which takes an input (or 'message') and returns a fixed-size alphanumeric string, which is called the **hash value** (sometimes called a **message digest**, **digital fingerprint**, **digest** or a **checksum**).
- The cryptographic hash function has four main properties:
 1. It is extremely easy (i.e. fast) to calculate a hash for any given data.
 2. It is extremely computationally difficult to calculate an alphanumeric text that has a given hash.
 3. A small change to the text yields a totally different hash value.
 4. It is extremely unlikely that two slightly different messages will have the same hash.



A small change in a single word, "over" produces a major Change in the output; see
https://en.wikipedia.org/wiki/Cryptographic_hash_function

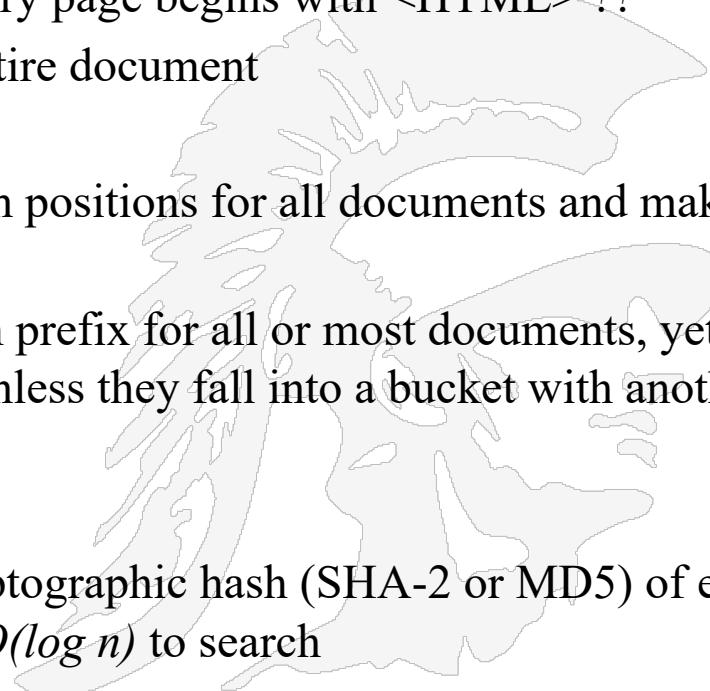
Some Popular Cryptographic Hash Functions

- The **MD5** (message-digest) hash function is a widely used cryptographic hash function producing a 128-bit (16-byte) hash value, typically expressed in text format as a 32 digit hexadecimal number.
 - Invented by Ron Rivest of MIT in 1991; replaced MD4
- The **SHA-1, SHA-2** hash functions are also quite popular (160 bit, 20 byte value)
 - SHA-1 was broken in 2005; using SHA-2 family of algorithms is now favored, see
 - <https://en.wikipedia.org/wiki/SHA-2>
- **SHA-3**, released in 2015; it produces digests of size 224, 256, 384 and 512 bits
- **RIPEMD-160** – a family of cryptographic hash functions and so far has not been broken; produces a 160 bit (20 byte) digest
- **E.g. See Chrome, Settings, Security and Privacy, Security, Manage certificates, certificates, Verisign**

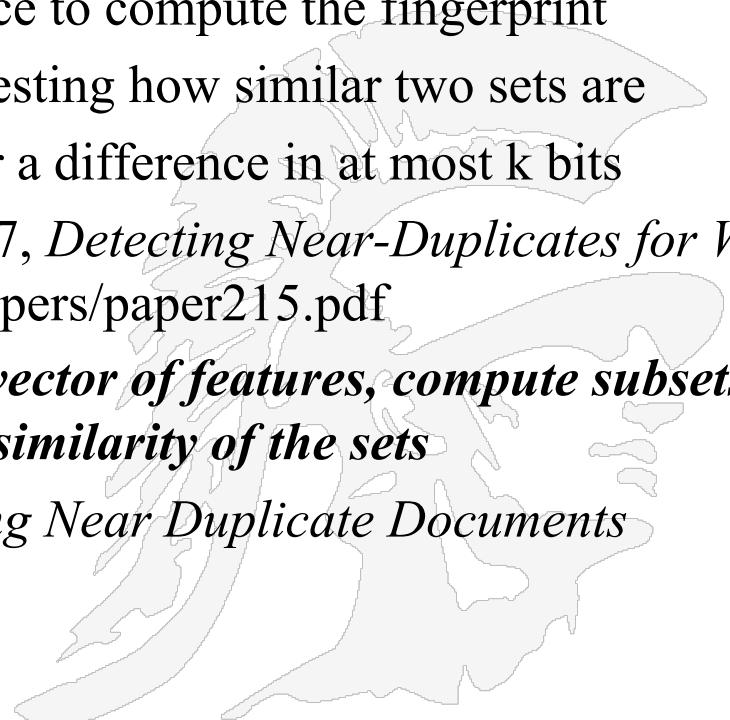


Identifying Identical Web Pages – Four Approaches

1. *Compare character by character* two documents to see if they are identical
 - very time consuming !!
2. *Hash just the first few characters and compare* only those documents that hash to the same bucket
 - But what about web pages where every page begins with <HTML> ??
3. *Use a hash function* that examines the entire document
 - But this requires lots of buckets
4. *Better approach* - pick some fixed random positions for all documents and make the hash function depend only on these;
 - This avoids the problem of a common prefix for all or most documents, yet we need not examine entire documents unless they fall into a bucket with another document
 - But we still need a lot of buckets
5. *Even better approach:* Compute the cryptographic hash (SHA-2 or MD5) of each web page and maintain in sorted order, $O(\log n)$ to search

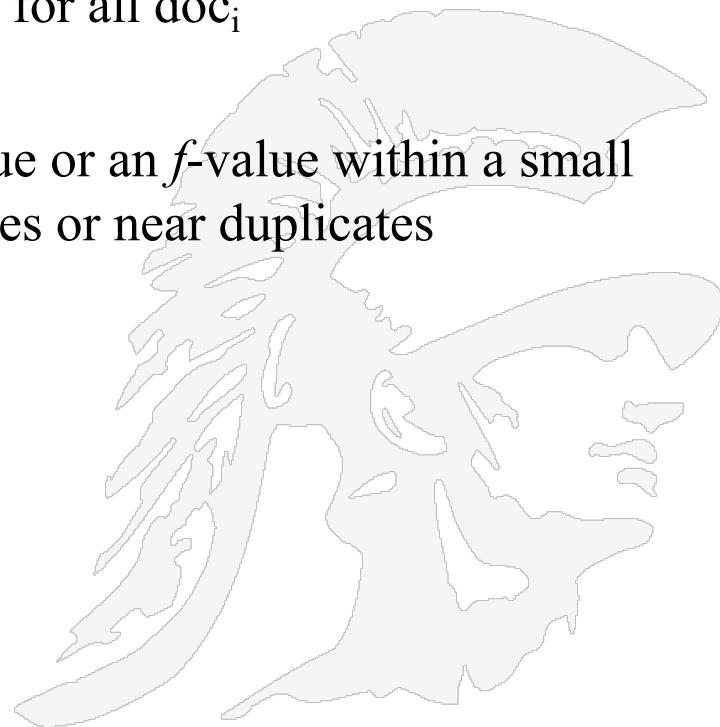


Identifying Near Identical Web Pages - Two Approaches

- 
1. ***Produce fingerprints and test for similarity*** - Treat web documents as defined by a set of features, constituting an n -dimensional vector, and transform this vector into an f -bit fingerprint of a small size
 - Use Simhash or Hamming Distance to compute the fingerprint
 - SimHash is an algorithm for testing how similar two sets are
 - Compare fingerprints and look for a difference in at most k bits
 - E.g. see Manku et al., WWW 2007, *Detecting Near-Duplicates for Web Crawling*, <http://www2007.org/papers/paper215.pdf>
 2. ***Instead of documents defined by n -vector of features, compute subsets of words (called shingles) and test for similarity of the sets***
 - Broder et al., WWW 1997, *Finding Near Duplicate Documents*

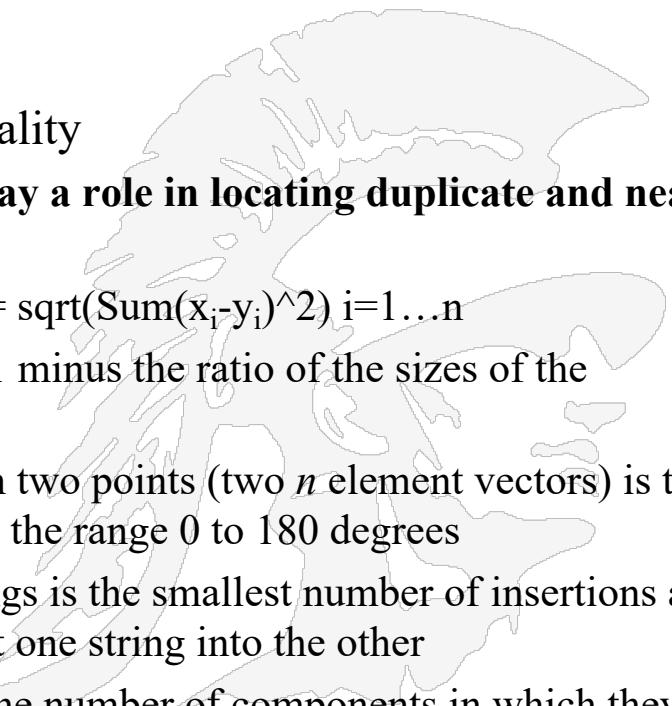
General Paradigm

1. Define a function f that captures the contents of each document in a number
 - E.g. hash function, signature, or a fingerprint
2. Create the pair $\langle f(doc_i), ID \text{ of } doc_i \rangle$ for all doc_i
3. Sort the pairs
4. Documents that have the same f -value or an f -value within a small threshold are believed to be duplicates or near duplicates



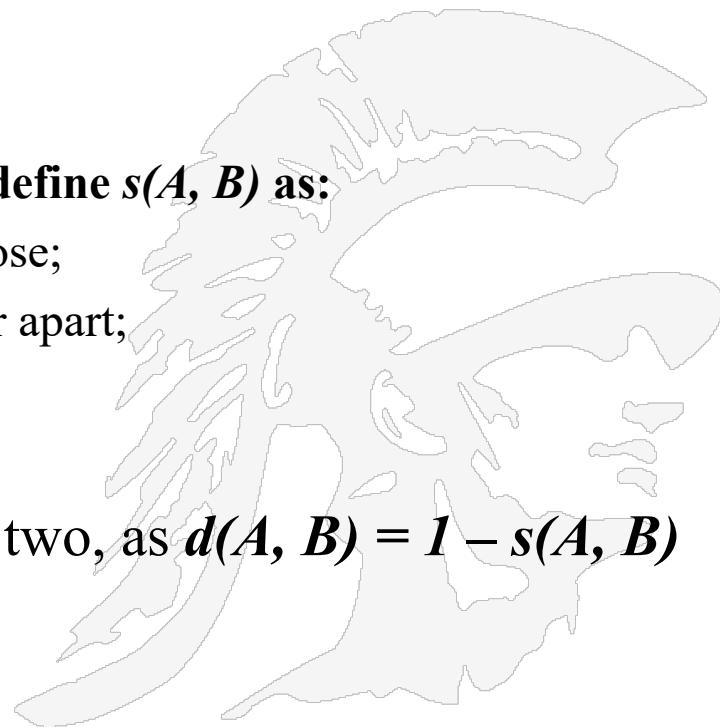
General Properties of Distance Measures

- To compute similarity, we need a distance measure
- A distance measure must satisfy 4 properties
 1. No negative distances
 2. $D(x,y) = 0$ iff $x=y$
 3. $D(x,y) = D(y,x)$ symmetric
 4. $D(x,y) \leq D(x,z) + D(z,y)$ triangle inequality
- There are several distance measures that can play a role in locating duplicate and near-duplicate documents
 - Euclidean distance – $D([x_1 \dots x_n], [y_1, \dots, y_n]) = \sqrt{\sum (x_i - y_i)^2}$ $i=1 \dots n$
 - Jaccard distance – $D(x,y) = 1 - \text{SIM}(x,y)$ or 1 minus the ratio of the sizes of the intersection and union of sets x and y
 - Cosine distance – the cosine distance between two points (two n element vectors) is the angle that the vectors to those points make; in the range 0 to 180 degrees
 - Edit distance – the distance between two strings is the smallest number of insertions and deletions of single characters that will convert one string into the other
 - Hamming distance – between two vectors is the number of components in which they differ (usually used on Boolean vectors)



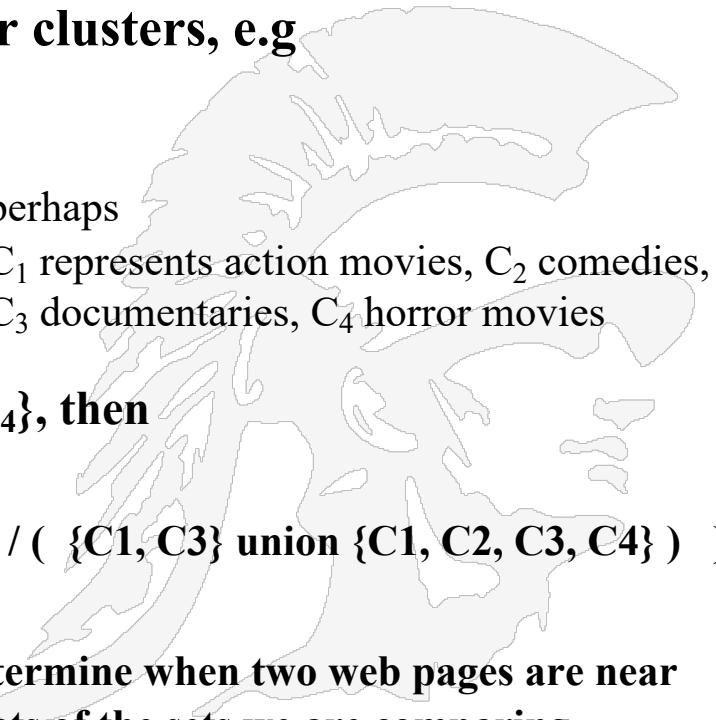
Set Distance and Set Similarity

- A set is an unordered collection of objects, e.g. $\{a, b, c\}$
- Focusing on the notion of *distance* of two sets we define a distance $d(A, B)$ as
 - *small*, if objects in A and B are close;
 - *large*, if objects in A and B are far apart;
 - *0*, if they are the same, and finally
 - $d(A, B)$ is in the range $[0, \text{infinity}]$
- Focusing on the notion of *similarity* we define $s(A, B)$ as:
 - *large*, if the objects in A and B are close;
 - *small*, if the objects in A and B are far apart;
 - *1*, if they are the same, and finally
 - $s(A, B)$ is in the range $[0, 1]$
- Often we can convert between the two, as $d(A, B) = 1 - s(A, B)$



Jaccard Similarity

- Consider $A = \{0, 1, 2, 5, 6\}$ and $B = \{0, 2, 3, 5, 7, 9\}$
- $JS(A, B) = \text{size}(A \text{ intersection } B) / \text{size}(A \text{ union } B)$
 $= \text{size}(\{0, 2, 5\}) / \text{size}(\{0, 1, 2, 3, 5, 6, 7, 9\})$
 $= 3 / 8 = 0.375$
- Suppose we divide our items into four clusters, e.g
 - $C_1 = \{0, 1, 2\}$
 - $C_2 = \{3, 4\}$
 - $C_3 = \{5, 6\}$
 - $C_4 = \{7, 8, 9\}$
- If $A_{\text{clu}} = \{C_1, C_3\}$ and $B_{\text{clu}} = \{C_1, C_2, C_3, C_4\}$, then
- $JS_{\text{clu}}(A, B) = JS(A_{\text{clu}}, B_{\text{clu}}) =$
 $\text{size}(\{\{C_1, C_3\} \text{ intersect } \{C_1, C_2, C_3, C_4\}\}) / (\{\{C_1, C_3\} \text{ union } \{C_1, C_2, C_3, C_4\}\})$
 $= 5 / 10 = 0.5$
- If we are going to use Jaccard similarity to determine when two web pages are near duplicates; we need to say what are the elements of the sets we are comparing



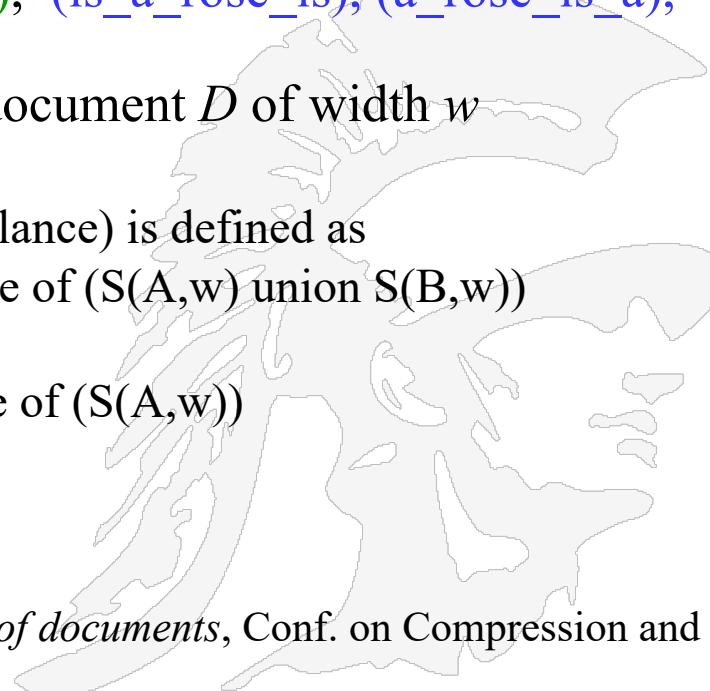
Computing Jaccard Similarity from Sets Containing Shingles

- **Definition of Shingle:**

- a contiguous subsequence of words in a document is called a *shingle*;
The 4-shingling of the phrase below produces a bag of 5 items:
“a rose is a rose is a rose” => a set $S(D,w)$ is defined as
 - $\{ (a_rose_is_a), (rose_is_a_rose), (is_a_rose_is), (a_rose_is_a), (rose_is_a_rose) \}$
- $S(D,w)$ is the set of shingles of a document D of width w

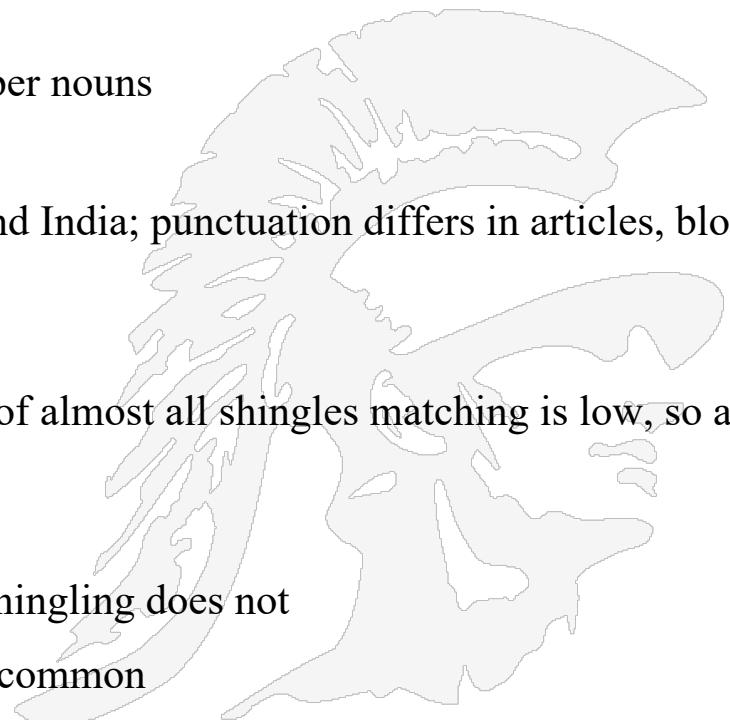
- **Similarity Measures**

- $Jaccard(A,B)$ (also known as Resemblance) is defined as
size of $(S(A,w) \cap S(B,w)) / \text{size of } (S(A,w) \cup S(B,w))$
 - $Containment(A,B)$ is defined as
size of $(S(A,w) \cap S(B,w)) / \text{size of } (S(A,w))$
 - $0 \leq \text{Resemblance} \leq 1$
 - $0 \leq \text{Containment} \leq 1$
- See *On the resemblance and containment of documents*, Conf. on Compression and Complexity, DEC Research Center, 1997



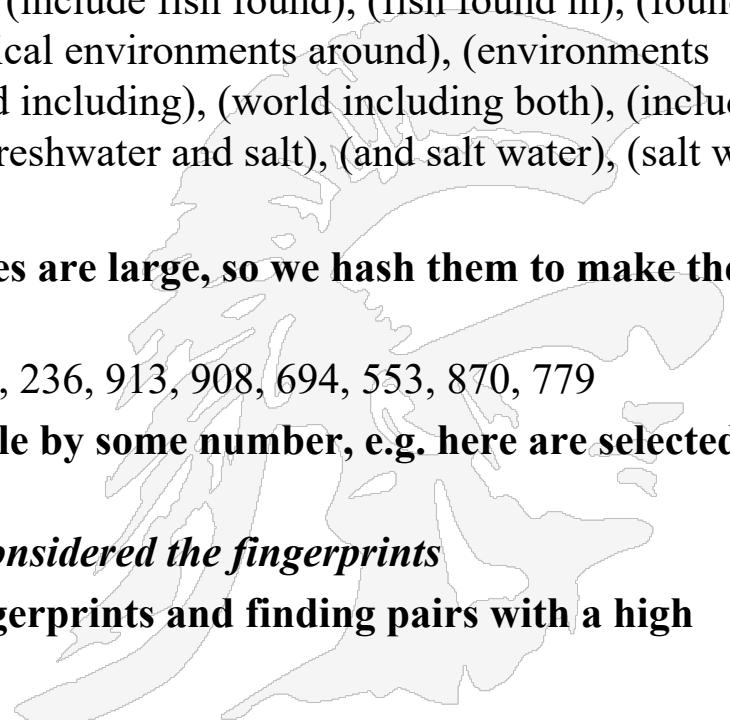
Shingling Modeling Choices

- **White space?**
 - Should we include spaces and returns? Sometimes it makes sense, e.g.
“plane has touch down” versus “threw a touchdown”
(the space between “touch” and “down” is significant)
- **Capitalization?**
 - Sam versus sam. Can help to distinguish proper nouns
- **Punctuation?**
 - English is punctuated differently in the US and India; punctuation differs in articles, blogs, and tweets
- **How large should k be?**
 - General rule: high enough so the probability of almost all shingles matching is low, so a collision is meaningful;
- **Count replicas?**
 - Typically bag of words counts replicas, but shingling does not
- **Stop words?** Typically omitted as they are so common



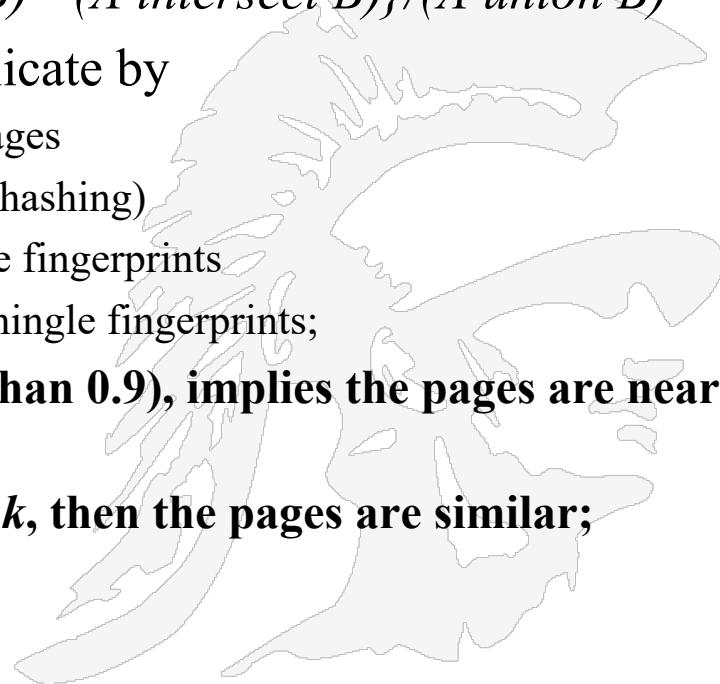
Mapping Shingles to Numbers

- **Original text**
 - “Tropical fish include fish found in tropical environments around the world, including both freshwater and salt water species”
- **All 3-shingles (there are 16 of them)**
 - (Tropical fish include), (fish include fish), (include fish found), (fish found in), (found in tropical), (in tropical environments), (tropical environments around), (environments around the), (around the world), (the world including), (world including both), (including both freshwater), (both freshwater and), (freshwater and salt), (and salt water), (salt water species)
- **Hash values for the 3-shingles (sets of shingles are large, so we hash them to make them more manageable, and we select a subset)**
 - 938, 664, 463, 822, 492, 798, 78, 969, 143, 236, 913, 908, 694, 553, 870, 779
- **Select only those hash values that are divisible by some number, e.g. here are selected hash values using $0 \bmod 4$**
 - 664, 492, 236, 908; *these are considered the fingerprints*
- **Near duplicates are found by comparing fingerprints and finding pairs with a high overlap**



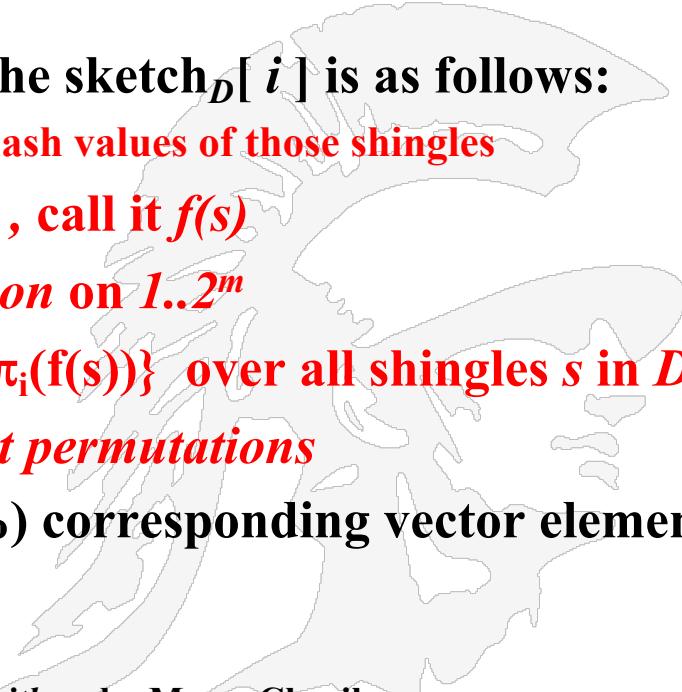
Jaccard Similarity and Shingling

- Recall the Jaccard similarity of sets A and B, $J(A,B)$, is defined as
$$|A \text{ intersect } B| / |A \cup B|$$
- The Jaccard distance of sets A and B, measuring *dissimilarity* is defined as
$$1 - J(A,B), \text{ or equivalently } \{(A \cup B) - (A \cap B)\} / (A \cup B)$$
- We can test if two pages are near duplicate by
 1. First compute the k -shingles of the two pages
 2. Map the k -shingles into numbers (e.g. by hashing)
 3. Select a subset of the shingles to act as the fingerprints
 4. Compute the Jaccard similarity of the k -shingle fingerprints;
- **A high Jaccard similarity (e.g. greater than 0.9), implies the pages are near duplicate; or**
- **$\text{if } (J(\text{fingerprint}(A), \text{fingerprint}(B))) > k, \text{ then the pages are similar;}$**



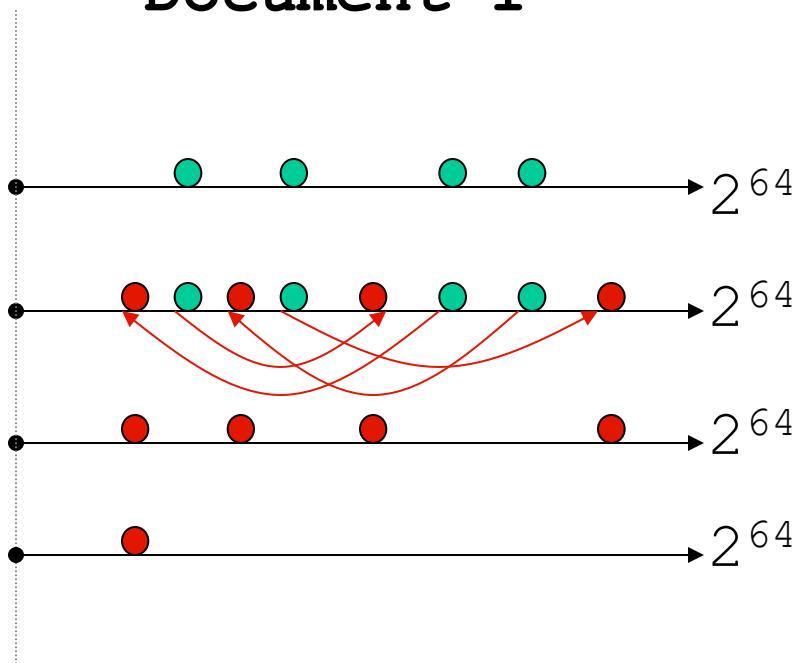
Solution to Speed Up the Use of Shingles for Near Document Similarity

- If we have to compare the shingles of a newly found document with the fingerprints of all documents found so far, this process is still too time consuming
 - So we adopt a probabilistic approach to reduce the number of comparisons we must make
 - For each document D compute the sketch $_D[i]$ is as follows:
 - Compute its shingles and then the hash values of those shingles
 - Map the hash values to $1..2^m$, call it $f(s)$
 - Let π_i be a *random permutation* on $1..2^m$
 - Pick the lowest bit, ie $\text{MIN } \{\pi_i(f(s))\}$ over all shingles s in D
 - *Do the above for 200 different permutations*
- Documents that share $\geq t$ (say 80%) corresponding vector elements are deemed to be near duplicates
- For more details see
- *Similarity estimation techniques from rounding algorithms* by Moses Charikar
- <http://www.cs.princeton.edu/courses/archive/spring04/cos598B/bib/CharikarEstim.pdf>



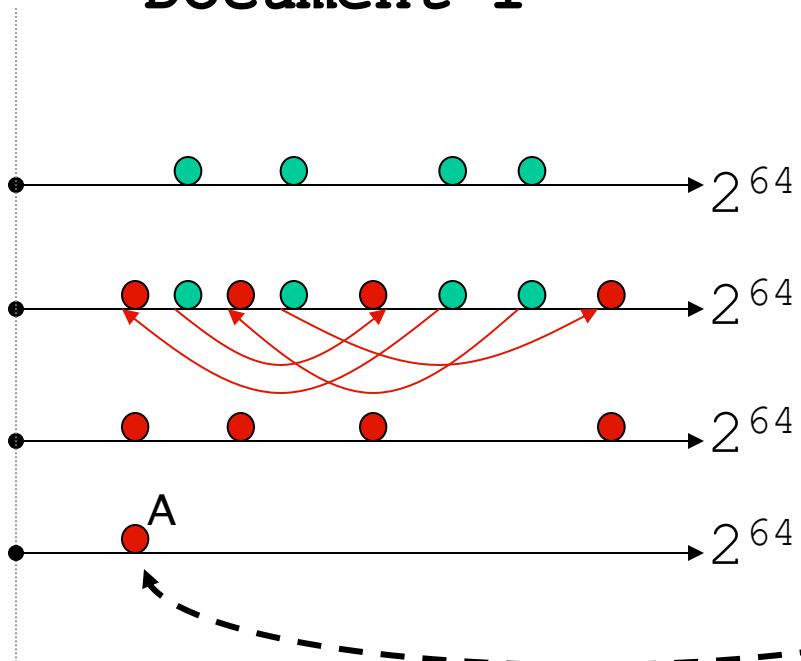
Computing Sketch[i] for Doc1

Document 1

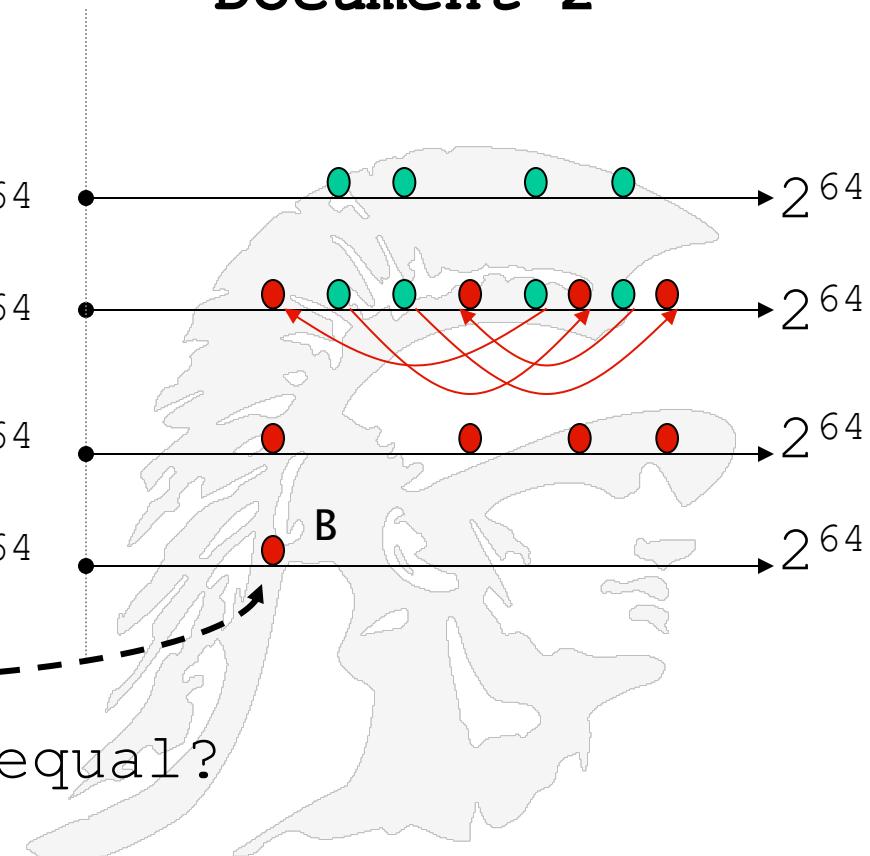


Test if Doc1.Sketch[i] = Doc2.Sketch[i]

Document 1

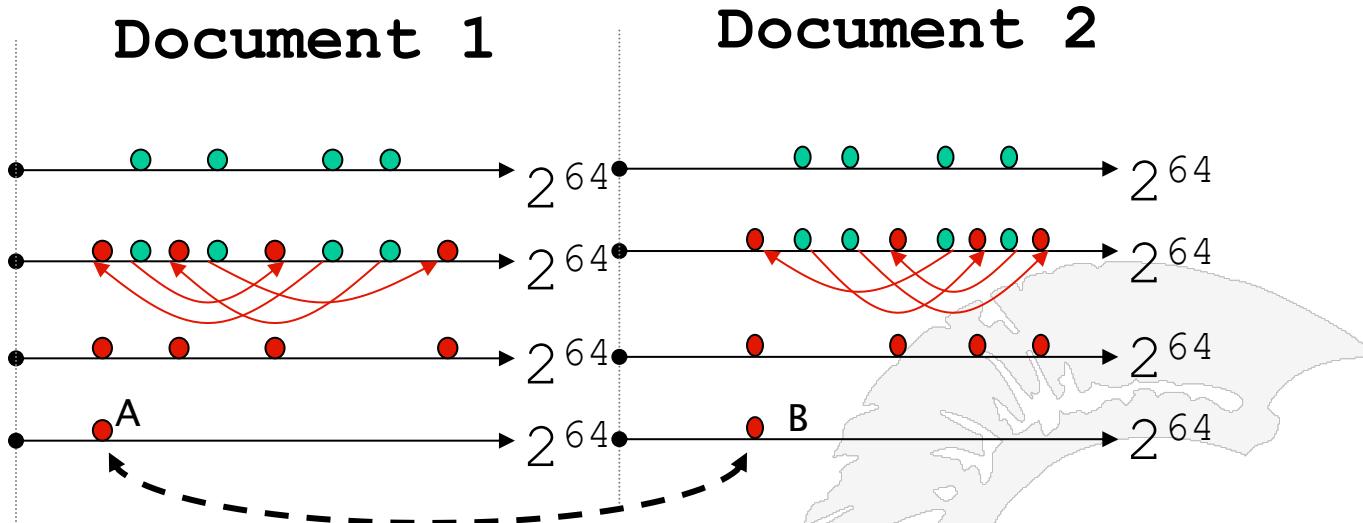


Document 2



Are these equal?

Test for 200 random permutations: $\pi_1, \pi_2, \dots, \pi_{200}$



Doc1 (is near similar to) Doc2 iff the shingle with the MIN value in the union of Doc1 and Doc2 also lies in the intersection

Key Property: This happens with probability

Size_of_intersection / Size_of_union

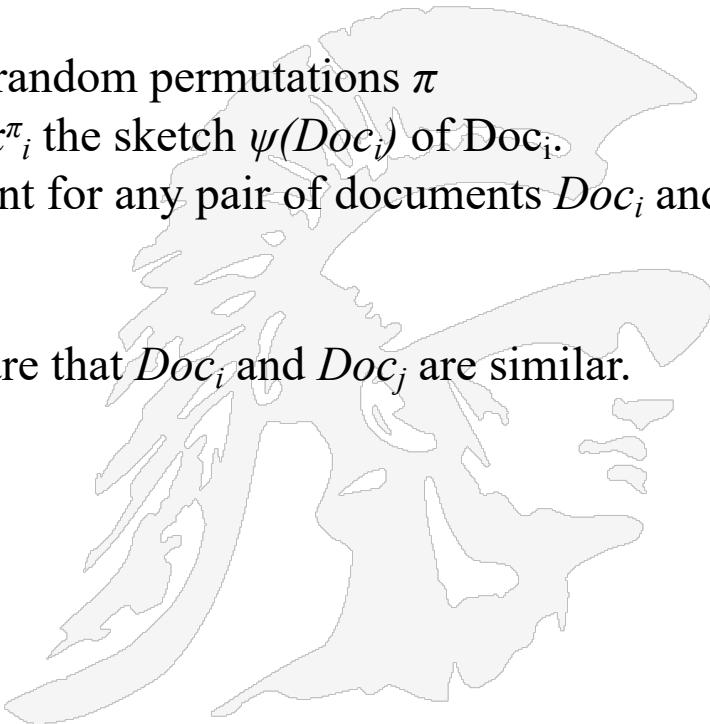
which means this method is just as good as computing the complete Jaccard Similarity

Computing Near Duplicates

Using Jaccard Similarity and Probability

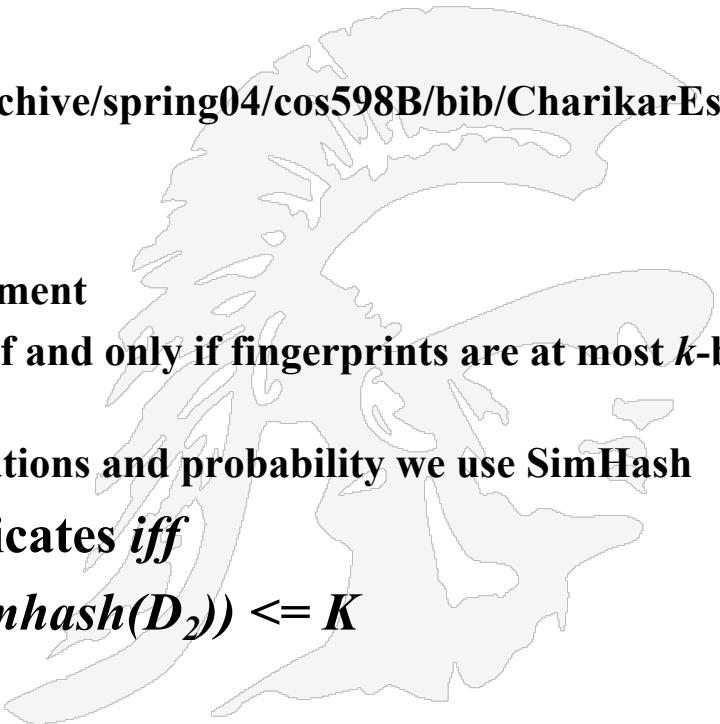
- *A surprising result:* $J(S(Doc_1), S(Doc_2)) = Prob(x^{\pi}_1 = x^{\pi}_2)$
- *Said another way: the Jaccard similarity of the shingle sets of Doc1 and Doc2 is the same as the probability that the permutation of x1 equals the permutation of x2*

- Repeat the process independently for 200 random permutations π
- Call the set of the 200 resulting values of x^{π}_i the sketch $\psi(Doc_i)$ of Doc_i .
- We can then estimate the Jaccard coefficient for any pair of documents Doc_i and Doc_j to be
- $|\psi i \cap \psi j|/200;$
- if this exceeds a preset threshold, we declare that Doc_i and Doc_j are similar.
- See our textbook, section 19.6 for a proof



Using SimHash to Detect Near Duplicates

- There is another way to determine if two web pages are near duplicates
- The method is called SimHash
- It was developed by Moses Charikar and is described in his paper *Similarity Estimation Techniques from Rounding Algorithms*, STOC May 2002
 - <https://www.cs.princeton.edu/courses/archive/spring04/cos598B/bib/CharikarEstim.pdf>
- The basic idea is the same as before
 - obtain an f -bit fingerprint for each document
 - A pair of documents are near duplicate if and only if fingerprints are at most k -bits apart
 - But in this case instead of using permutations and probability we use SimHash
- Documents D_1 and D_2 are near duplicates iff
$$\text{Hamming-Distance}(\text{Simhash}(D_1), \text{Simhash}(D_2)) \leq K$$
- Typically $f = 64$ and $k = 3$



Simhash by Moses Charikar

A Locally Sensitive Hash Function

- A hash function usually hashes different values to totally different hash values; here is an example

p1 = 'the cat sat on the mat'

p2 = 'the cat sat on a mat'

p3 = 'we all scream for ice cream'

p1.hash => 415542861

p2.hash => 668720516

p3.hash => 767429688

- Simhash is one where similar items are hashed to similar hash values
(by similar we mean the bitwise Hamming distance between hash values is small)

p1.simhash => 851459198

0011001011000000001111000111110

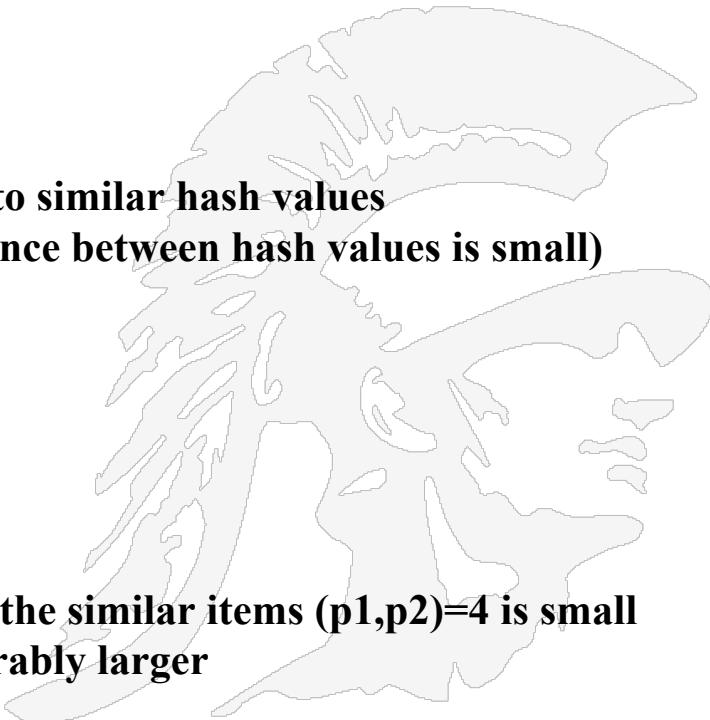
p2.simhash => 847263864

00110010100000000011100001111000

p3.simhash => 984968088

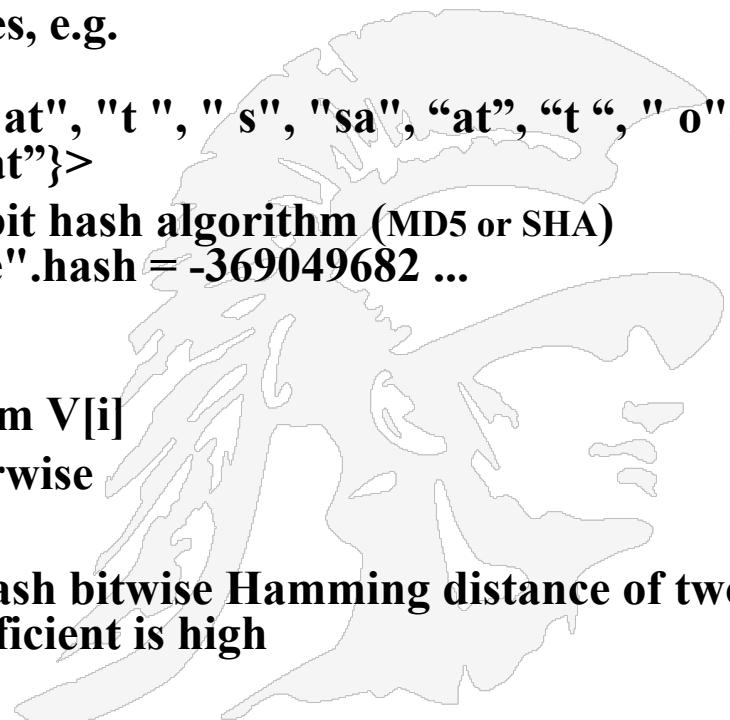
00111010101101010110101110011000

- in this case we can see the hamming distance of the similar items $(p1, p2) = 4$ is small whereas $(p1, p3) = 16$ and $(p2, p3) = 12$ are considerably larger



The Simhash Algorithm

- The simhash of a phrase is calculated as follows:
 1. pick a hashsize, lets say 32 bits
 2. let $V = [0] * 32$ # (ie a vector of 32 zeros)
 3. break the input phrase up into shingles, e.g.
'the cat sat on a mat'.shingles(2) =>
#<Set: {"th", "he", "e ", " c", "ca", "at", "t ", "s", "sa", "at", "t ", " o",
"on", "n ", " a", "a ", " m", "ma", "at"}>
 4. hash each feature using a normal 32-bit hash algorithm (MD5 or SHA)
"th".hash = -502157718 "he".hash = -369049682 ...
 5. for each hash
if bit_i of hash is set then add 1 to $V[i]$
if bit_i of hash is not set then take 1 from $V[i]$
 6. simhash bit_i is 1 if $V[i] > 0$ and 0 otherwise
- Simhash is useful because if the Simhash bitwise Hamming distance of two phrases is low then their Jaccard coefficient is high



Simhash Continued

- In the case that two numbers have a low bitwise Hamming distance and the difference in their bits are in the lower order bits then it turns out that they will end up close to each other if the list is sorted.
- consider numbers
- 1 37586 1001001011010010
- 2 50086 1100001110100110 7 <--(this column lists hamming
- 3 2648 0000101001011000 11 distance to previous entry)
- 4 934 0000001110100110 9
- 5 40957 100111111111101 9
- 6 2650 0000101001011010 9
- 7 64475 111101111011011 7
- 8 40955 100111111111011 4

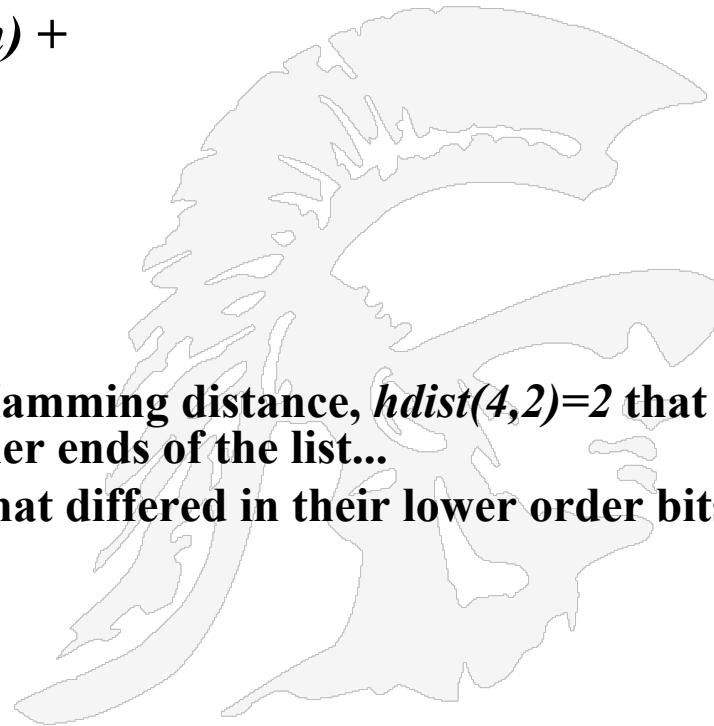
if we sort them

4	934	0000001110100110
3	2648	0000101001011000 9
6	2650	0000101001011010 1
1	37586	1001001011010010 5
8	40955	100111111111011 6
5	40957	100111111111101 2
2	50086	1100001110100110 9
7	64475	111101111011011 9

notice that two pairs with very smallest hamming distance
 $\text{hdist}(3,6)=1$ and $\text{hdist}(8,5)=2$ have ended up adjacent to each other.

Simhash Continued

- Rather than check every combo we could just check the adjacent pairs of the list, each is a good candidate.
- This reduces the runtime from $n^*(n-1)/2$ coefficient calculations, $O(n^2)$ to
 - n fingerprints calculations $O(n)$ +
 - a sort $O(n \log n)$ +
 - n coefficient calculations $O(n)$,
- which is $O(n \log n)$ overall;
- A problem:
 - there is another pair with a low Hamming distance, $hdist(4,2)=2$ that have ended up totally apart at other ends of the list...
 - sorting only picked up the pairs that differed in their lower order bits.



Simhash Continued

- To get around this consider another convenient property of bitwise Hamming distance, *a permutation of the bits of two numbers preserves Hamming distance*
- If we permute by 'rotating' the bits, i.e. bit shift left and replace lowest order bit with the 'lost' highest order bit we get 'new' fingerprints that have the same Hamming distances

'rotate' bits left twice

4	3736	0000111010011000
3	10592	0010100101100000 9
6	10600	0010100101101000 1
1	19274	0100101101001010 5
8	32750	011111111101110 6
5	32758	011111111110110 2
2	3739	0000111010011011 9
7	61295	1110111101101111 9

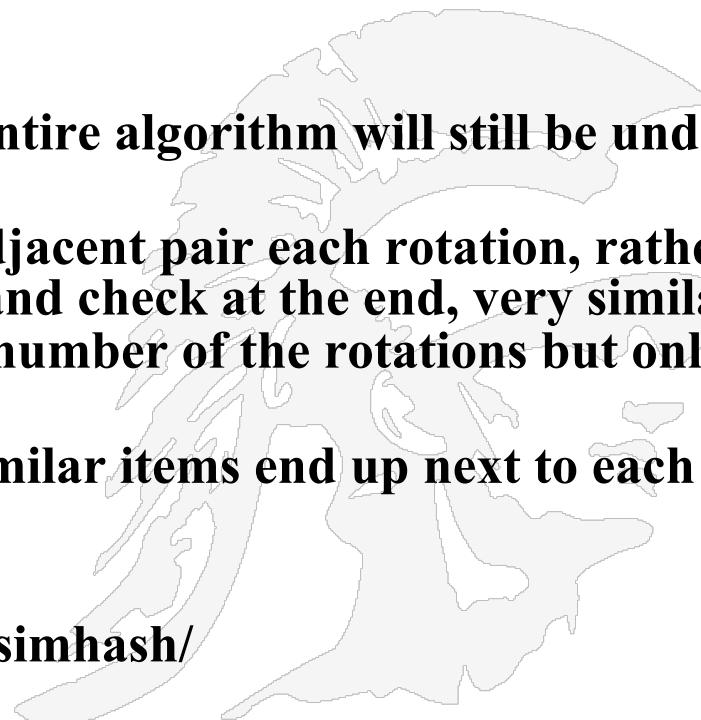
if we sort again by fingerprint

4	3736	0000111010011000
2	3739	0000111010011011 2
3	10592	0010100101100000 11
6	10600	0010100101101000 1
1	19274	0100101101001010 5
8	32750	011111111101110 6
5	32758	011111111110110 2
7	61295	1110111101101111 6

this time the (2,4) pair ended up adjacent
 we also identified the (3,6) and (5,8) pairs as candidates again

Simhash Continued

- So to avoid whether the items differ in the higher or lower order bits we can just do the following B times (where B is the bit length of the fingerprint)
 1. rotate the bits
 2. sort
 3. check adjacent
- in the likely case that $B \ll n$ the entire algorithm will still be under $O(n^2)$
- (in fact we wouldn't check each adjacent pair each rotation, rather collect the adjacent pairs in a set and check at the end, very similar items would end up adjacent in a number of the rotations but only need to be checked once)
- but it all hinges on the fact that similar items end up next to each other in the sorted lists
- <http://matpalm.com/resemblance/simhash/>



HW2 Assignment

The screenshot shows a web browser window with the title bar "CSCI 572 Home Page". The address bar says "Not Secure | csci572.com". The page content is titled "CS572 Course Assignments" with a subtitle "Last Modified: Jan 24, 2022". On the left is a sidebar with a logo and links: "Home Page", "Recent Search Engine Articles", "Schedule of Lectures", "Assignments", "Special Resources", "Course Grading", "Course Materials", and "Class News Group Piazza". A large arrow points from the "Course Grading" link in the sidebar to the "Homework 2: Web Crawling" section on the right. This section contains a list of items: "[Instructions for Installing Eclipse and crawler4j]", "[Flowchart for Crawler4j]", "[Web Crawler Exercise]", "[Grading Guidelines]", and "Homework #2 Due Feb 10".

- Involves
 1. Java programming
 - I assume all of you know how to program in Java!
 2. Eclipse Software Development Environment
 3. crawler4j, an open source java web crawler
 4. a crawl and analysis of a web site and an analysis of the crawl

What is Eclipse?

- Eclipse started as a proprietary IBM product (IBM Visual age for Smalltalk/Java)
 - Embracing the open source model IBM opened the product up
- Open Source
 - It is a general purpose open platform that facilitates and encourages the development of third party plug-ins
- Best known as an Integrated Development Environment (IDE)
 - Provides tools for coding, building, running and debugging applications
- Originally designed for Java, now supports many other languages
 - Good support for C, C++
 - Python, PHP, Ruby, etc...

Prerequisites for Running Eclipse

- Eclipse is written in Java and will thus need an installed JRE (Java Runtime Environment) or JDK (Java Development Kit) in which to execute
 - JDK recommended

Obtaining Eclipse

- Eclipse can be downloaded from...
<https://www.eclipse.org/downloads/packages/>
- Eclipse comes bundled as a zip file (Windows) or a tarball (all other operating systems)
- Eclipse version to Install - Eclipse IDE for Java Developers

The Eclipse Installer 2021-06 R now includes a JRE for macOS, Windows and Linux.

Try the Eclipse **Installer** 2021-06 R

The easiest way to install and update your Eclipse Development Environment.

[Find out more](#)

[2,031,205 Installer Downloads](#)

[2,337,429 Package Downloads and Updates](#)

Download

macOS [x86_64](#)
Windows [x86_64](#)
Linux [x86_64 | AArch64](#)

Eclipse IDE 2021-06 R Packages



The Eclipse
Installer 2021-06
R now includes a
JRE for macOS,
Windows and
Linux.

Get **Eclipse IDE 2021-06**

Installing Eclipse

- Simply unwrap the zip file to some directory where you want to store the executables
- The document

“Instructions for Installing Eclipse and crawler4j”

- located at

<http://csci572.com/2022Spring/hw2/Crawler4jinstallation.pdf>

describes the installation for both Windows and Macs

Launching Eclipse

- Once you have the environment setup, go ahead and launch eclipse
- You should see a splash screen such as the one below



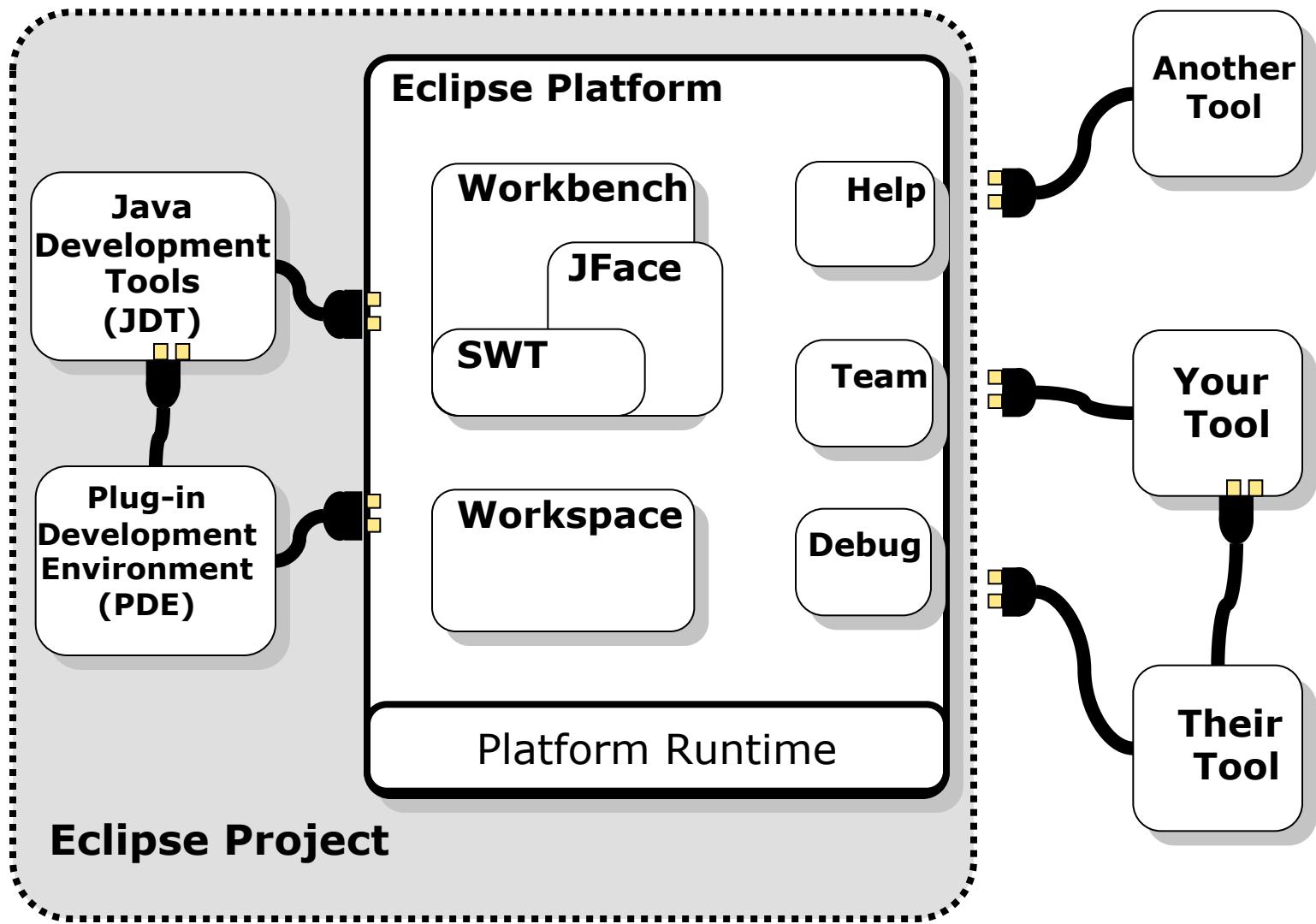
Eclipse Installer

Eclipse Structure

- » The following components constitute the rich client platform:
 - Core platform - boot Eclipse, run plug-ins
 - OSGi - a standard bundling framework
 - the Standard Widget Toolkit (SWT) - a portable widget toolkit
 - JFace - file buffers, text handling, text editors
 - The Eclipse Workbench - views, editors, perspectives, wizards

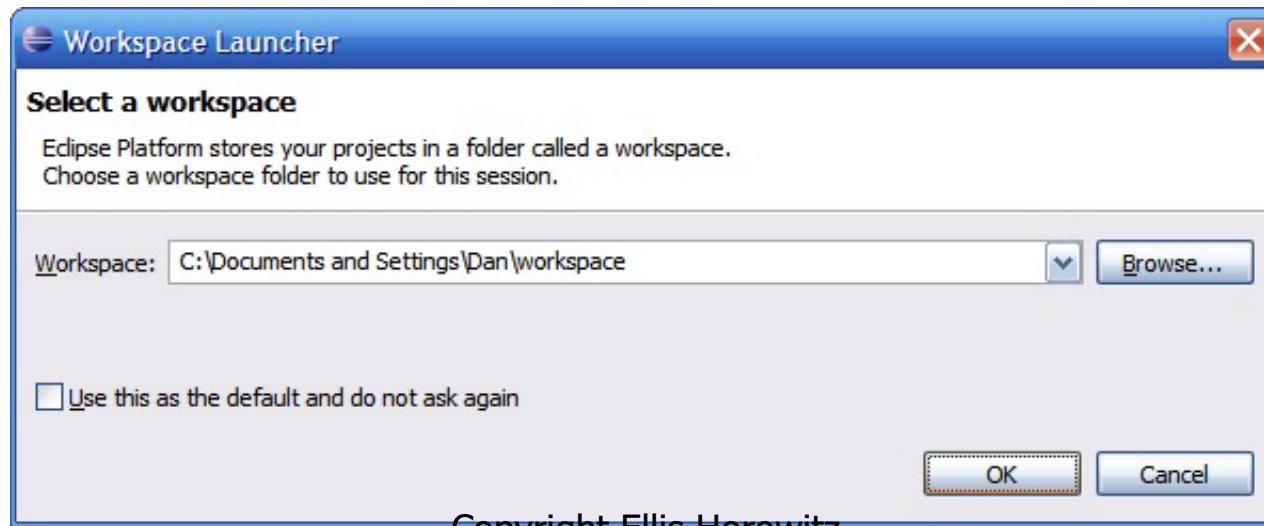
Eclipse Platform is the common base
Consists of several key components

Eclipse Overview



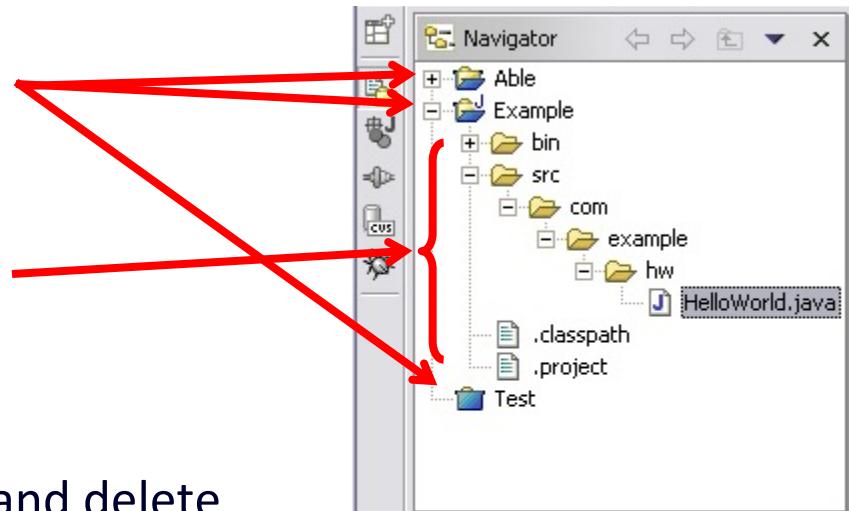
Selecting a Workspace

- In Eclipse, all of your code will live under a *workspace*
- A *workspace* is nothing more than a location where we will store the source code and where Eclipse will write out preferences
- Eclipse allows you to have multiple workspaces – each tailored in its own way
- Choose a location where you want to store your files, then click OK

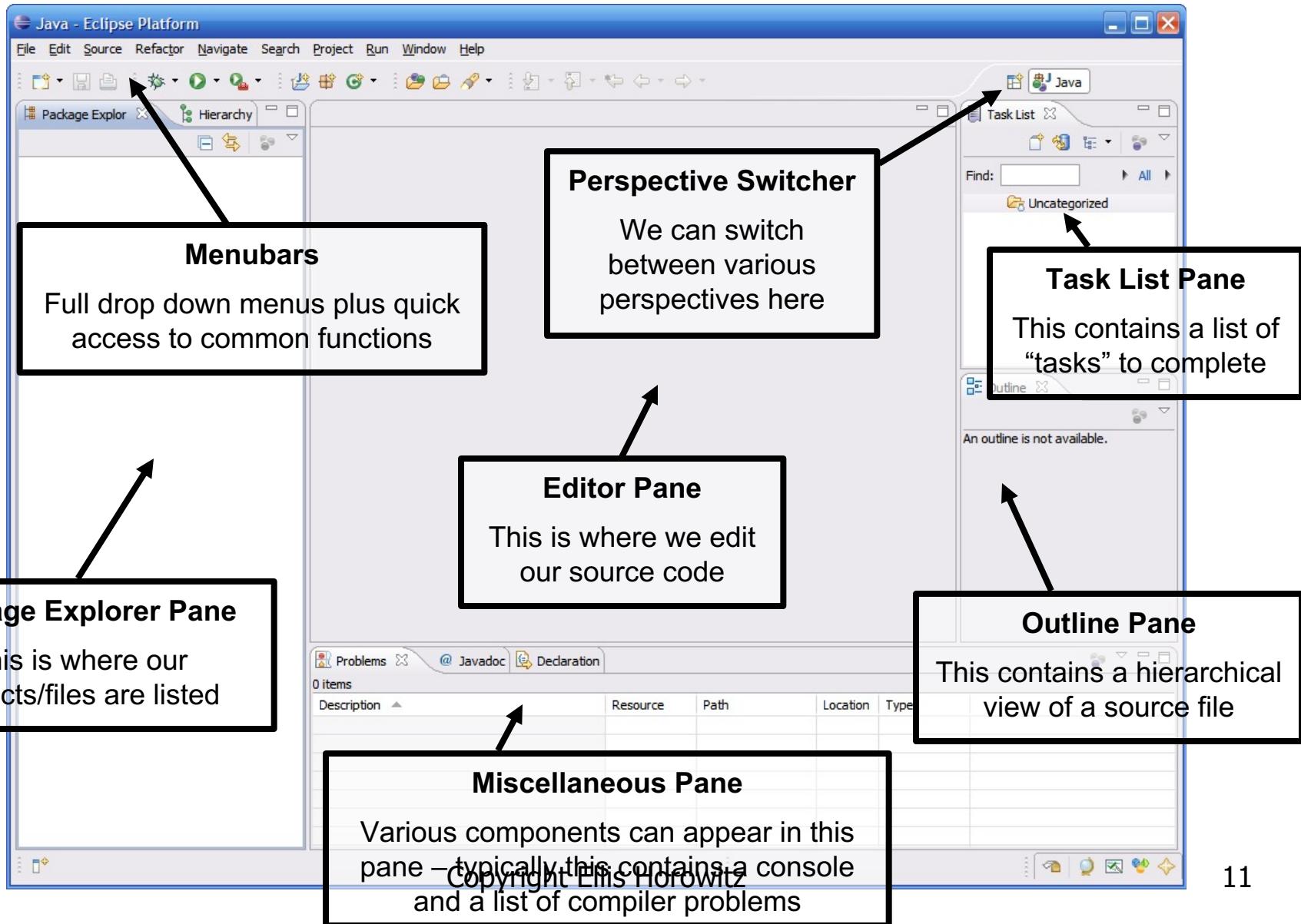


Workspace Component

- Tools operate on files in user's **workspace**
- Workspace holds 1 or more top-level **projects**
- Projects map to directories in file system
- Tree of **folders** and **files**
- {Files, Folders, Projects} termed **resources**
 - Tools read, create, modify, and delete resources in workspace
 - Plug-ins access via workspace and resource APIs

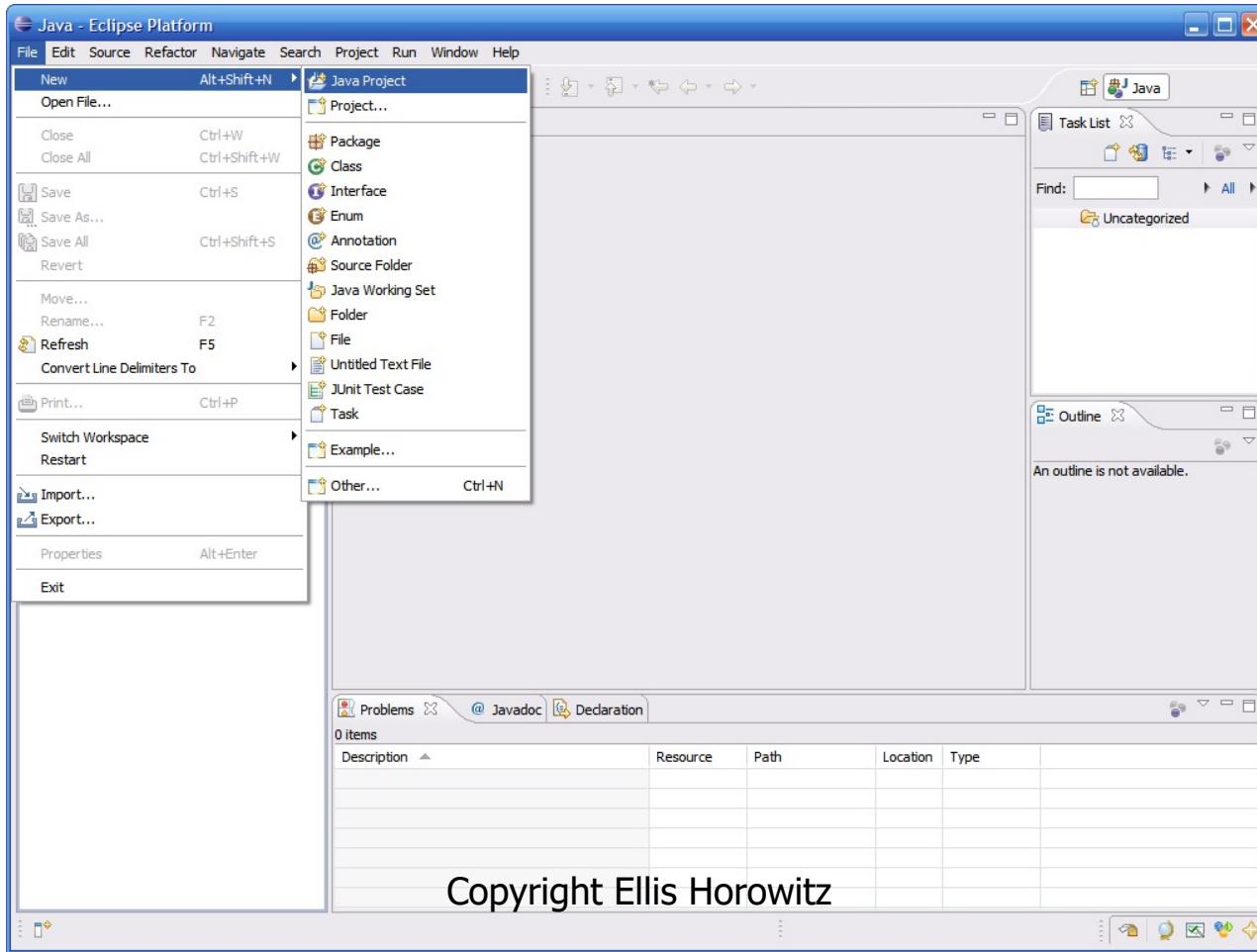


Eclipse IDE Components



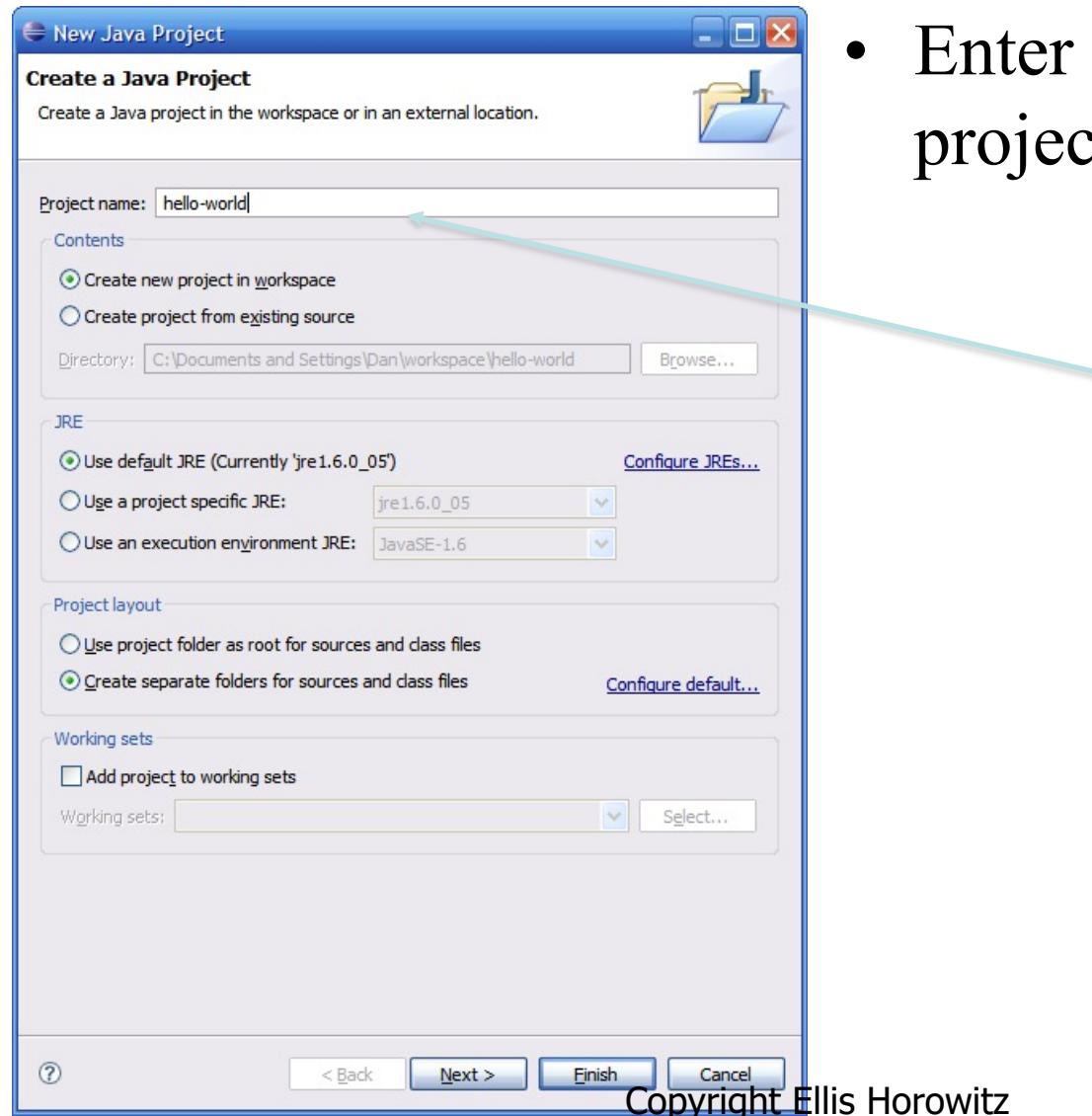
Creating a New Project

- All code in Eclipse needs to live under a project
- To create a project: File → New → Java Project



Copyright Ellis Horowitz

Creating a New Project (continued)

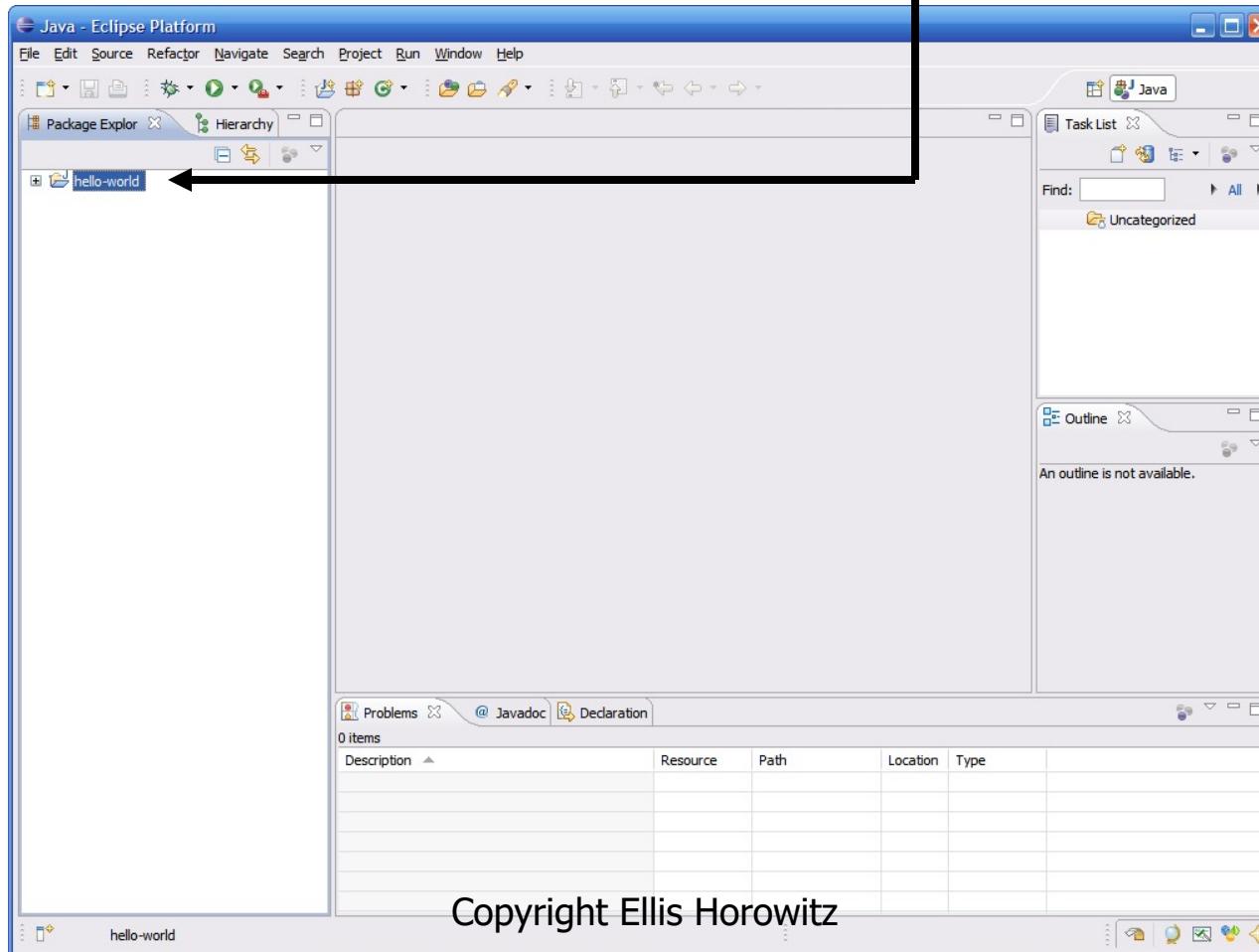


- Enter a name for the project, then click Finish

Hello-world Project

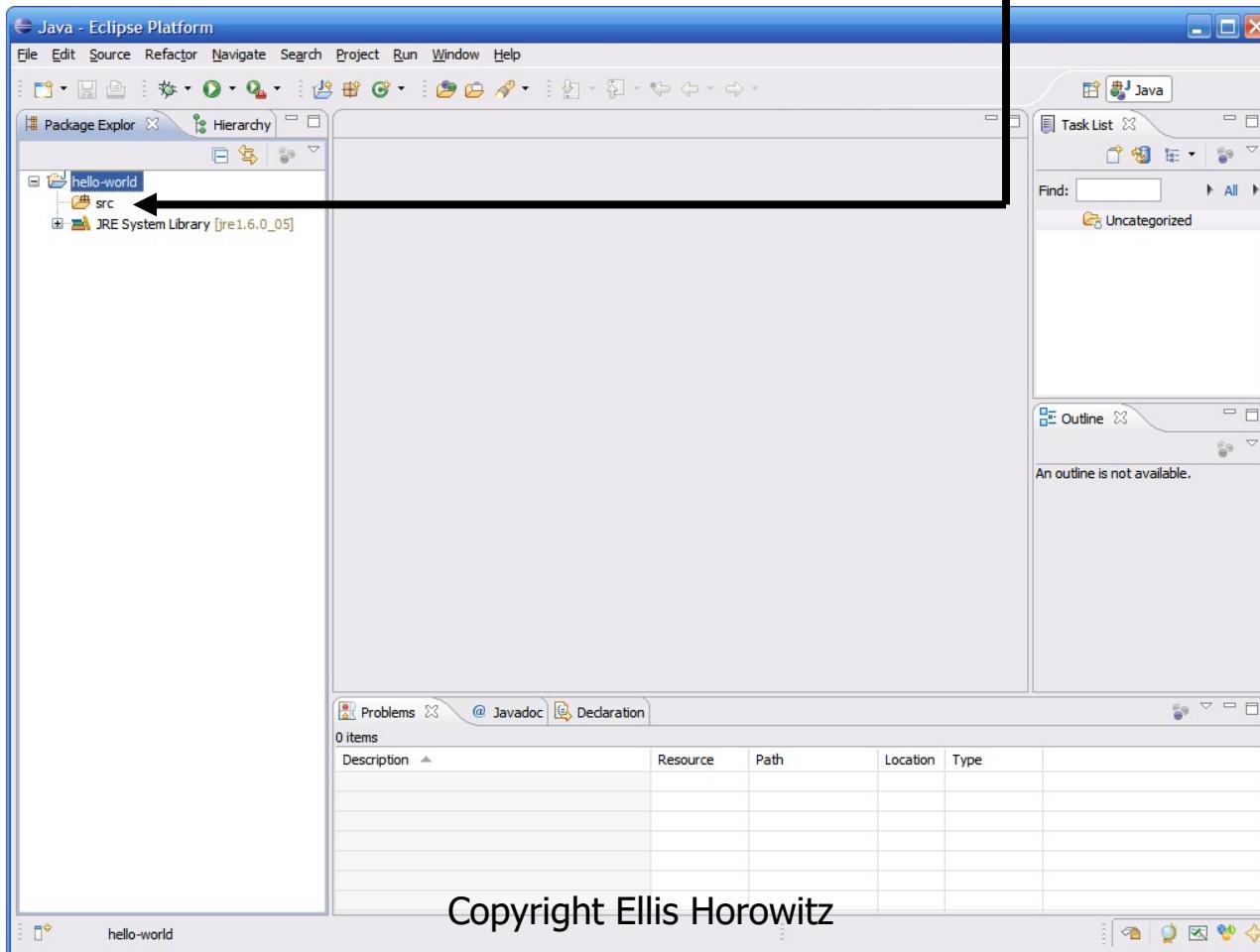
Creating a New Project (continued)

- The newly created project should then appear under the Package Explorer



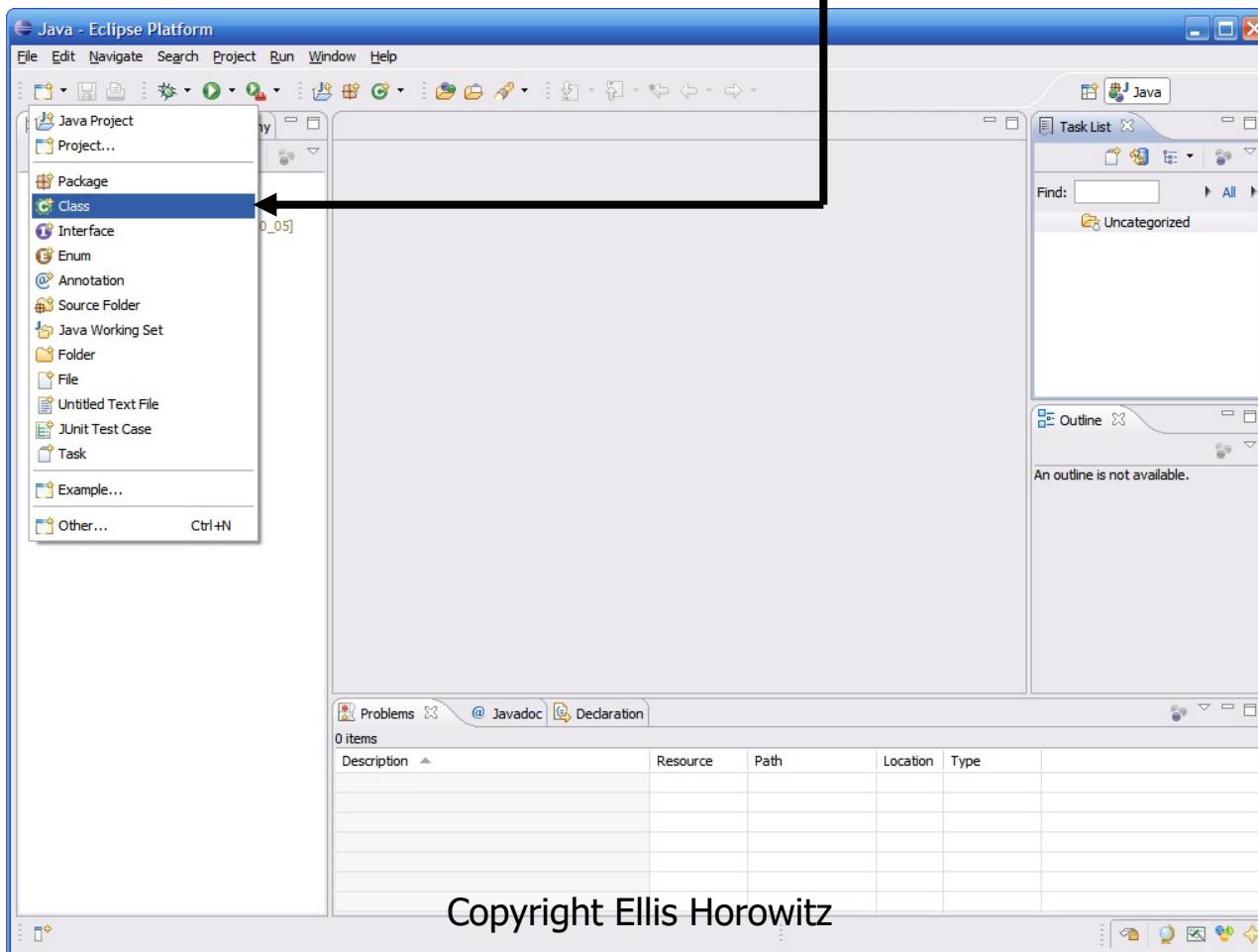
The src folder

- Eclipse automatically creates a folder to store your source code in called src

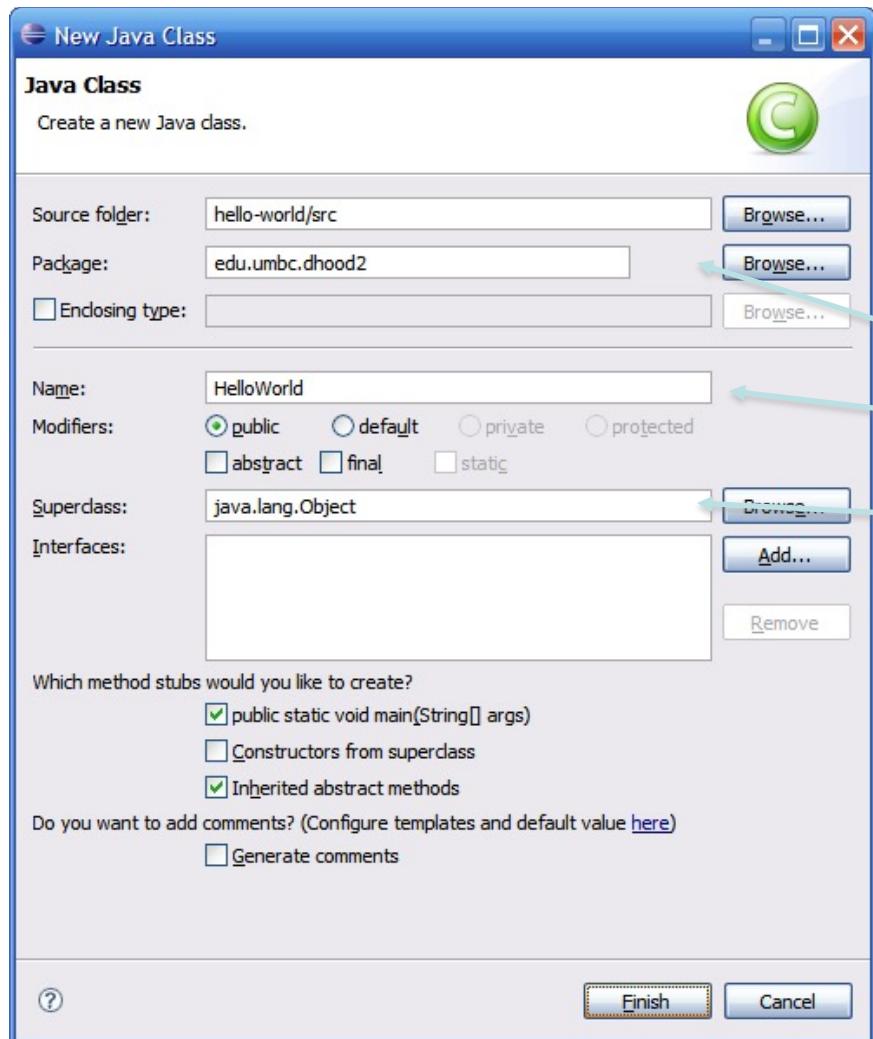


Creating a Class

- To create a class, simply click on the New button, then select Class



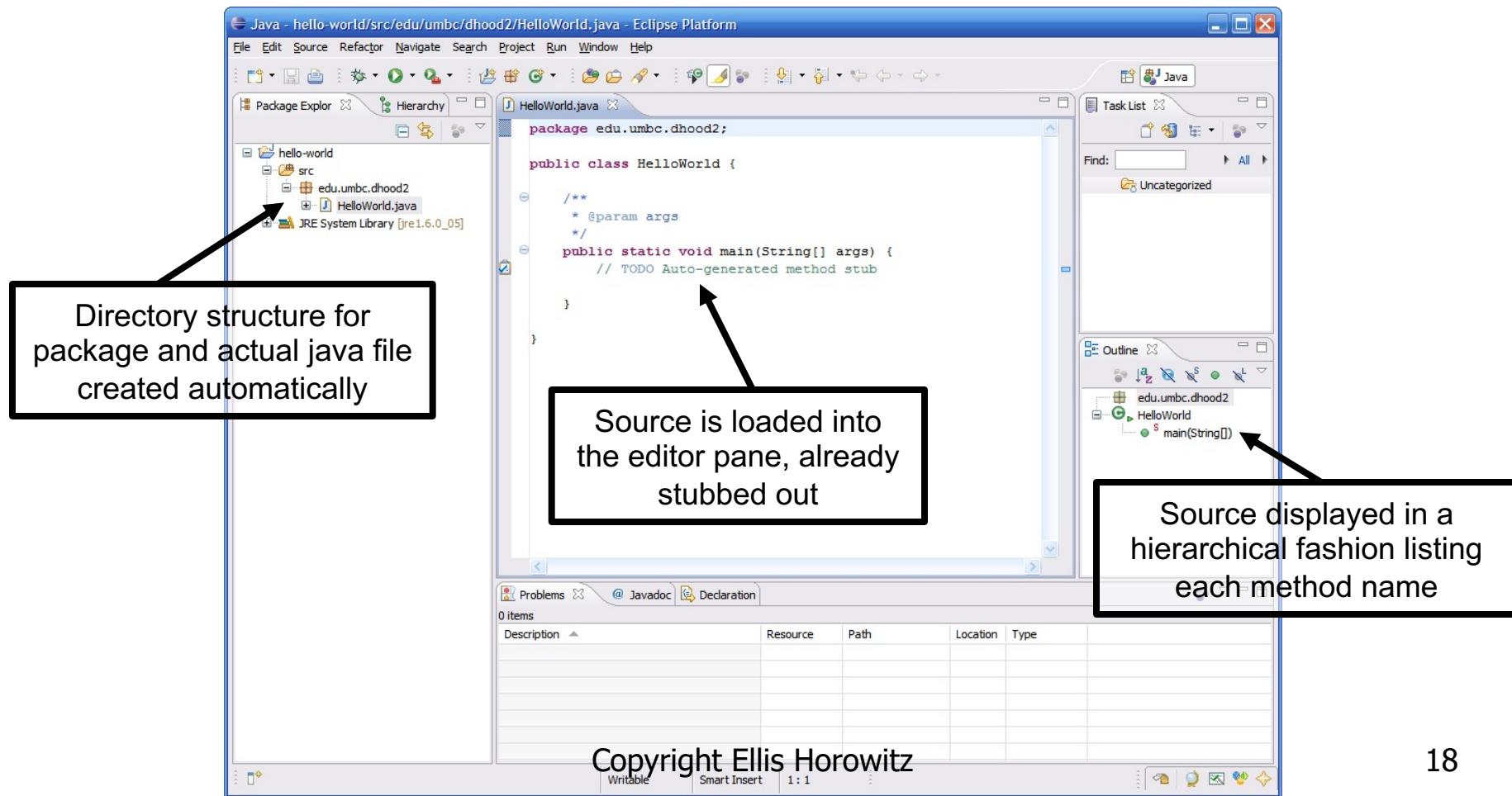
Creating a Class (continued)



- This brings up the new class wizard
- From here you can specify the following...
 - Package
 - Class name
 - Superclass
 - Whether or not to include a main
 - Etc...
- Fill in necessary information then click Finish to continue

The Created Class

- As you can see a number of things have now happened...

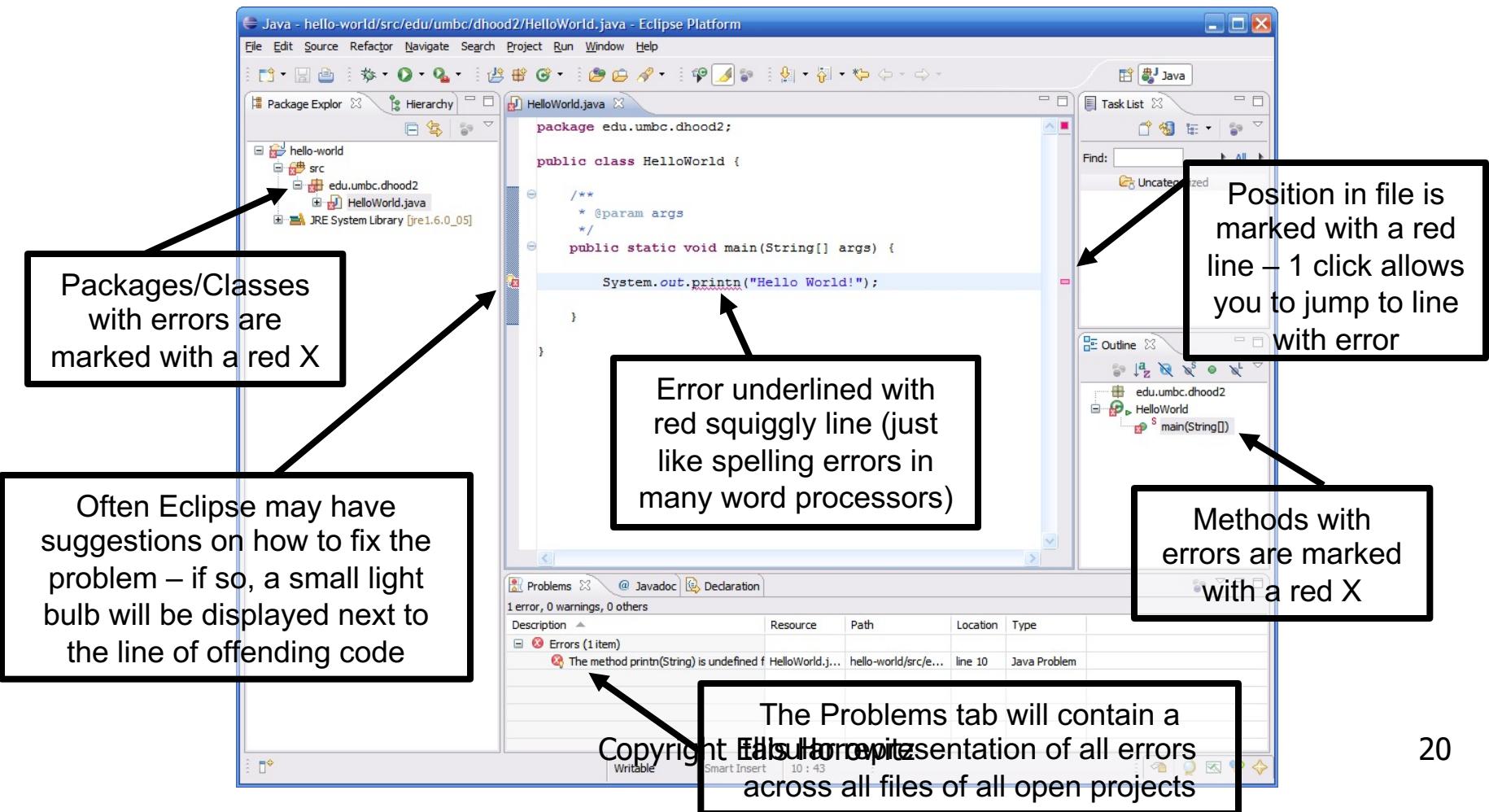


Compiling Source Code

- One important feature of Eclipse is that it automatically compiles your code in the background
- This means that errors can be corrected when made
 - We all know that iterative development is an excellent approach to developing code, but going to shell to do a compile can interrupt the normal course of development
 - You no longer need to go to the command prompt and compile code directly

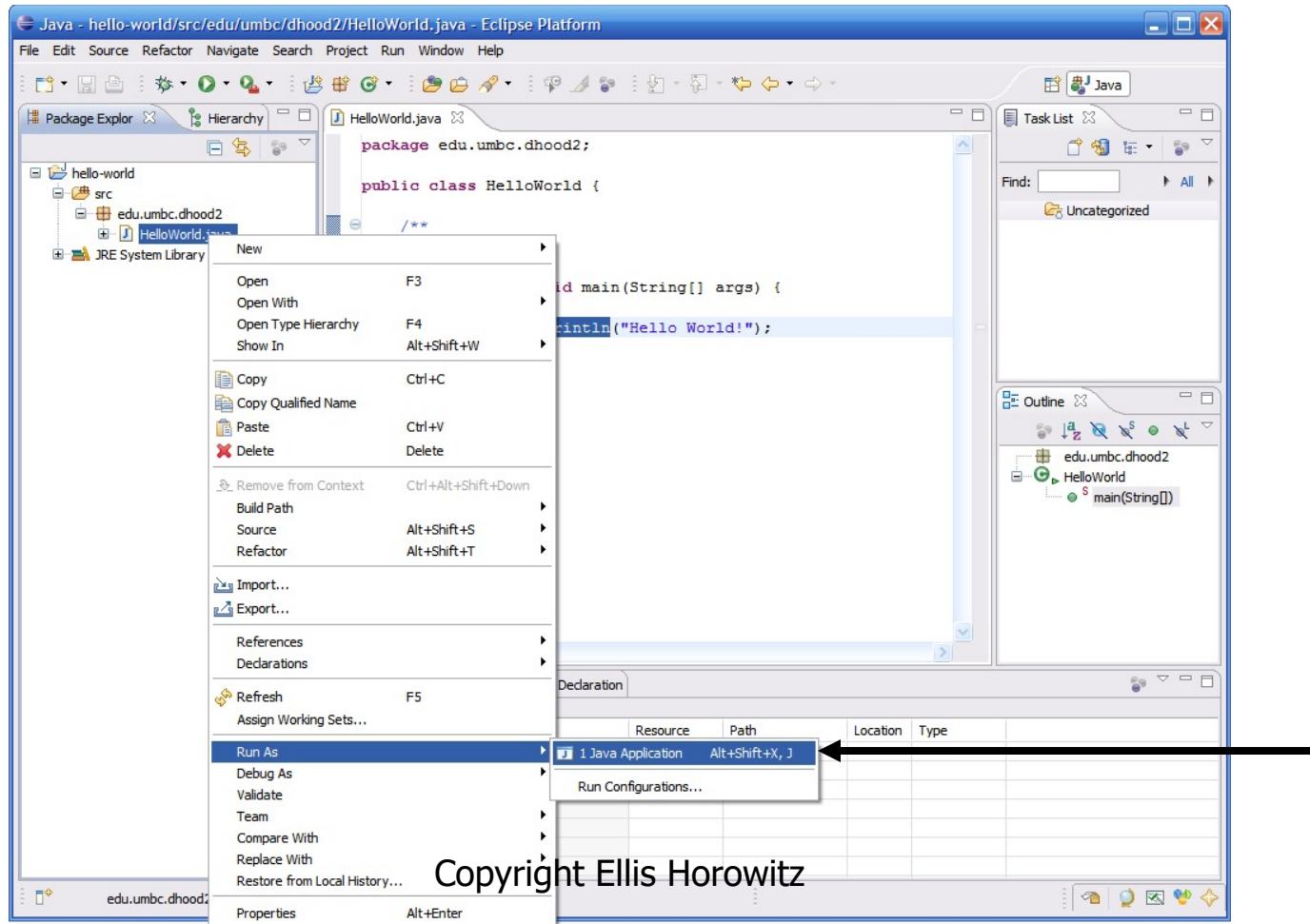
Example Compilation Error

- This code contains a typo in the `println` statement...



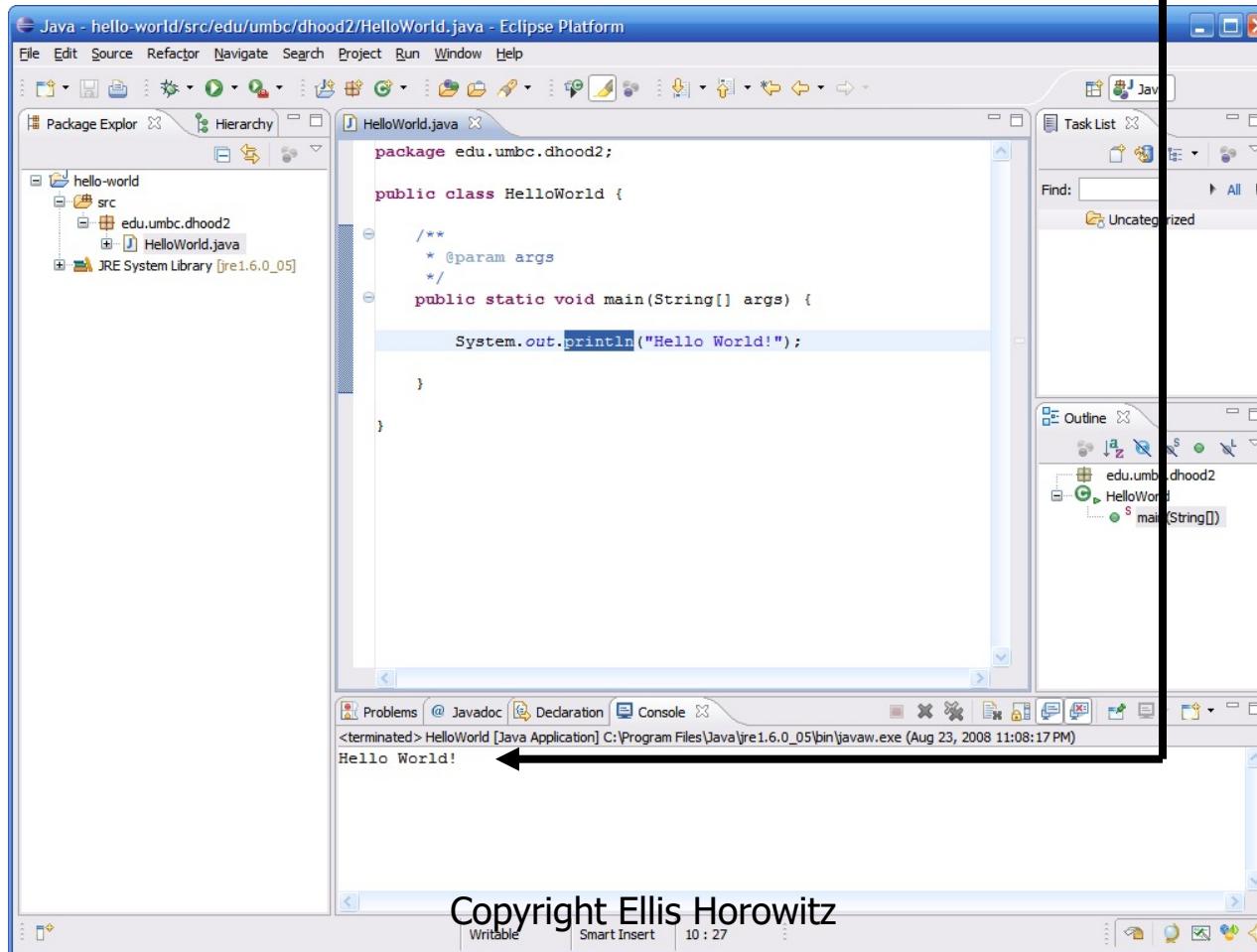
Running Code

- An easy way to run code is to right click on the class and select Run As → Java Application



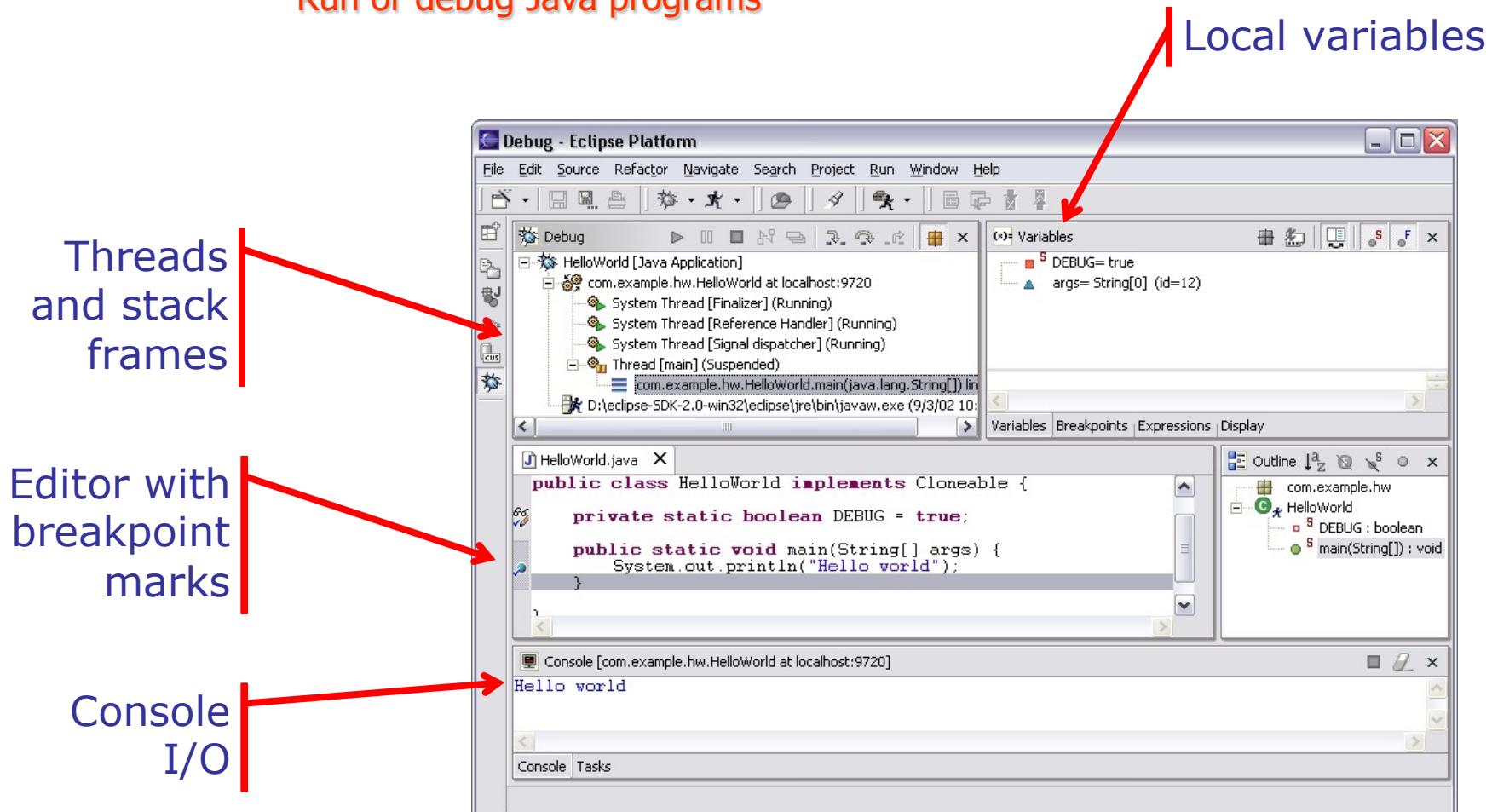
Running Code (continued)

- The output of running the code can be seen in the Console tab in the bottom pane



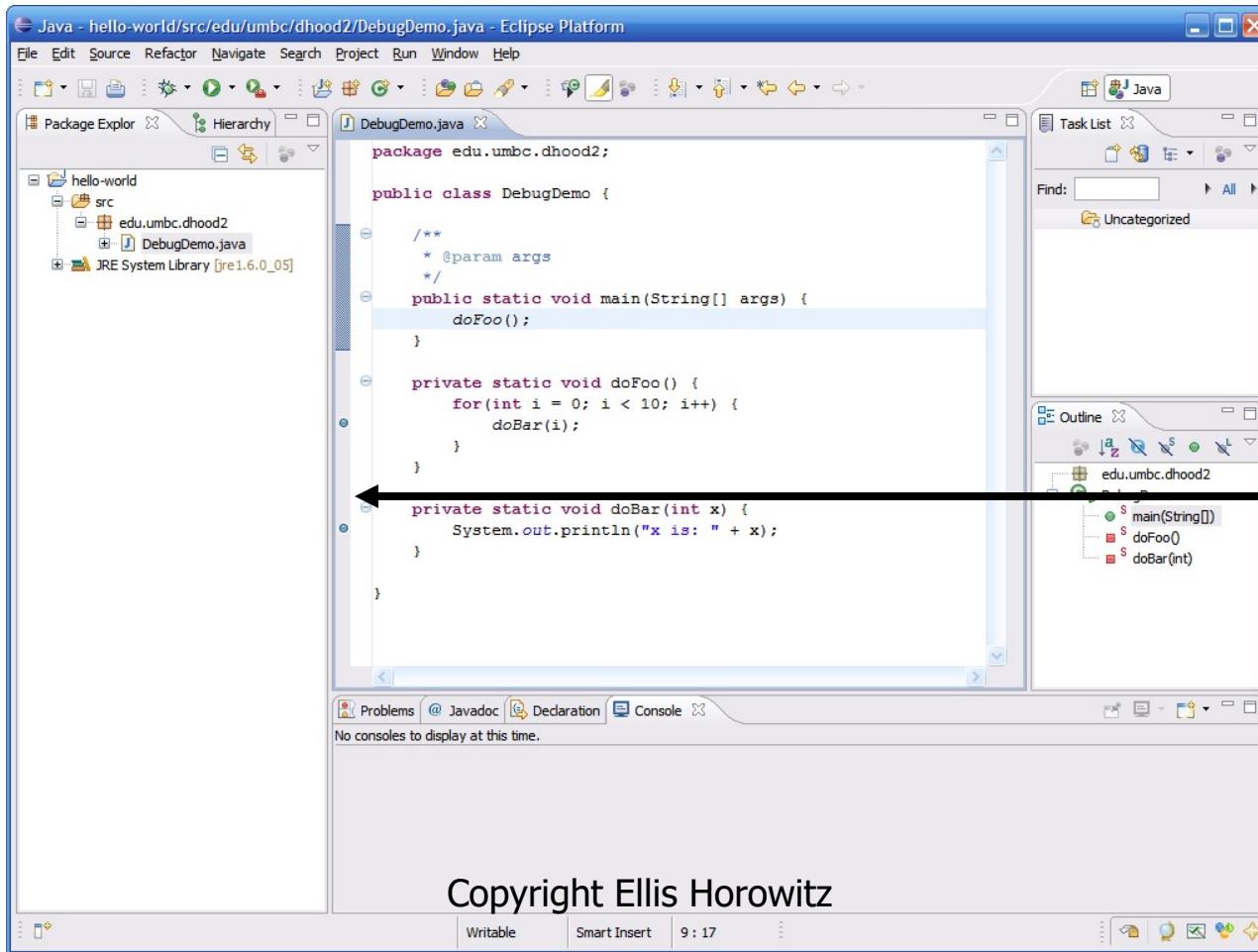
Eclipse Java Debugger

Run or debug Java programs



Debugging Code

- Eclipse comes with a pretty good built-in debugger
- You can set breakpoints in your code by double clicking in the left hand margin – break points are represented by these blue bubbles



End of Eclipse Tutorial

Tools for Surface Web Crawling

- Command line for issuing http requests
 - wget, pre-installed in Ubuntu
 - get a single page
 - wget http://www.example.com/index.html
 - support http, ftp etc., e.g.
 - wget ftp://ftp.gnu.org/pub/gnu/wget/wget-latest.tar.gz
 - curl, OSX pre-installed also supports http requests
- Simple crawling programs
 - Crawler4j, written in Java
 - Scrapy: <http://scrapy.org>, written in Python
- Large-scale crawling programs
 - Heritrix, crawler for archive.org
 - Nutch, Apache Software Foundation

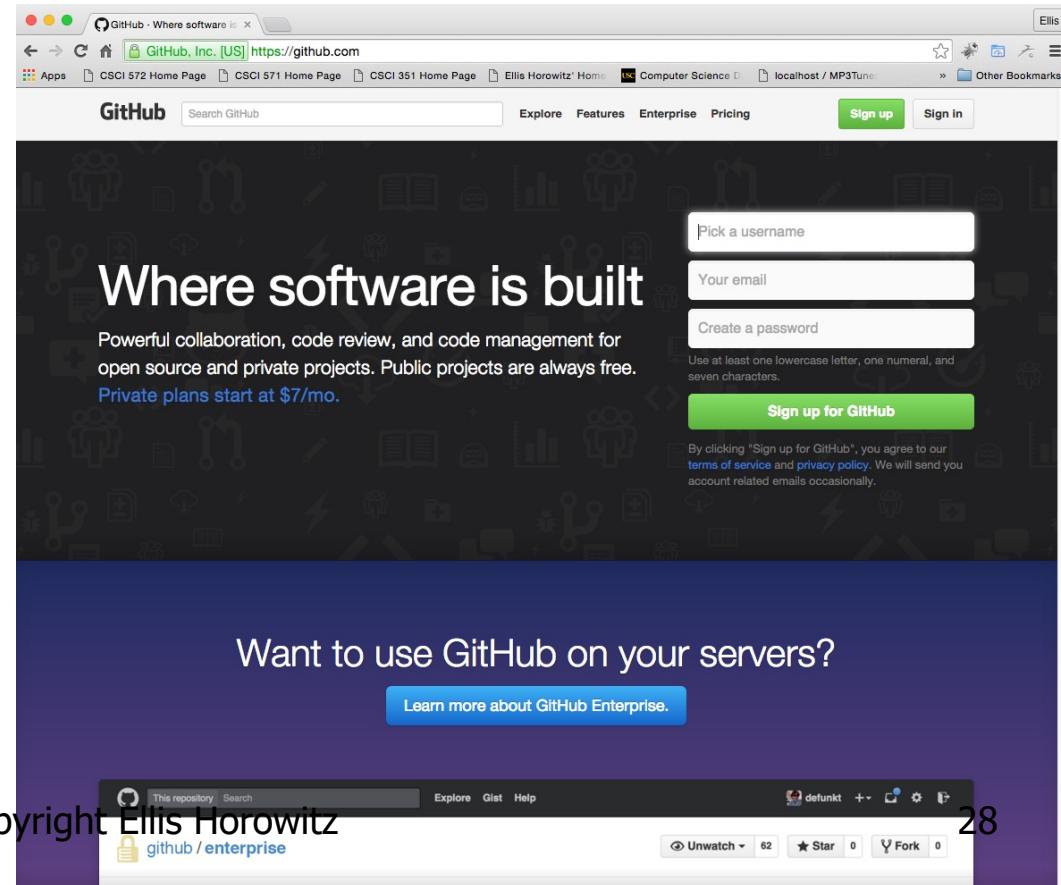
How To Get a Web Page in Java

```
import java . net .*;
import java . io .*;
public class URLReader {
public static void main(String [] args) throws Exception { } }
URL oracle = new URL("http://www.oracle.com/");
BufferedReader in = new BufferedReader (
new InputStreamReader(oracle.openStream()));
String inputLine ;
while (( inputLine = in . readLine ()) != null)
    System . out . println ( inputLine );
    in . close ();
}
}
```

- After you create a URL, you can call the URL's openStream() method to get a stream from which you can read the contents of the URL.
- The openStream() method returns a [java.io.InputStream](#) object

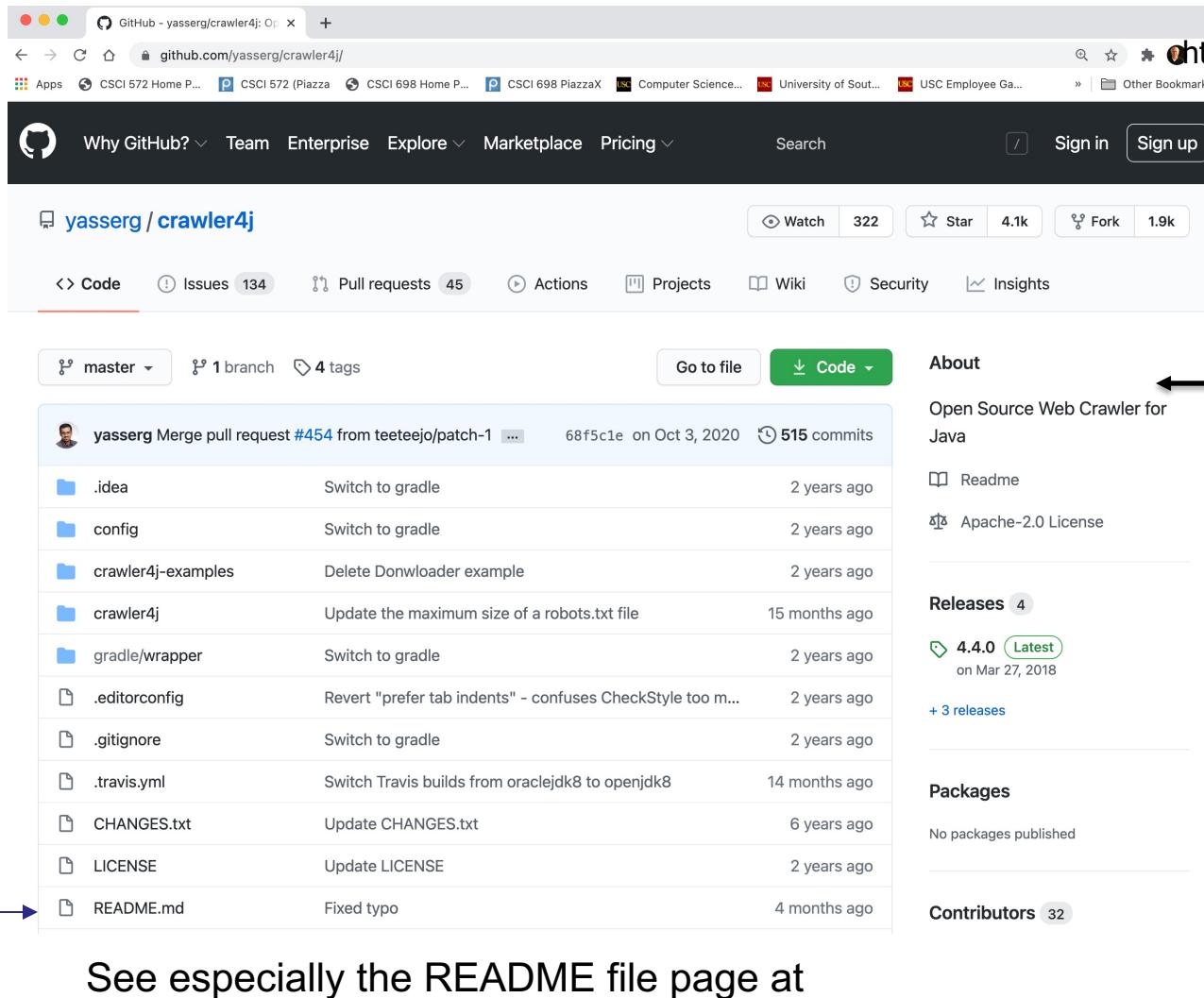
Instructions for Installing Crawler4j

- download crawler4j from github
 - GitHub is a web-based repository hosting service for software. Originally the Git system offered distributed revision control and source code management (SCM) functionality, but on the command line; GitHub offers a web interface and some additional features.
 - As of January 2020, GitHub reports having over 56 million users and over 100 million repositories
 - Microsoft purchased GitHub



Copyright Ellis Horowitz

Downloading Crawler4j from GitHub



The screenshot shows a Mac OS X desktop with a browser window open to the GitHub repository page for `yasserg/crawler4j`. The URL in the address bar is <https://github.com/yasserg/crawler4j>. The repository has 322 stars and 1.9k forks. The repository page includes sections for Code, Issues (134), Pull requests (45), Actions, Projects, Wiki, Security, and Insights. A large list of commits is shown, with the most recent being a merge pull request from `teeteejo/patch-1`. The right side of the page features sections for About, Releases (4), Packages, and Contributors.

GitHub - yasserg/crawler4j: Open

https://github.com/yasserg/crawler4j/

Why GitHub? Team Enterprise Explore Marketplace Pricing

Search Sign in Sign up

yasserg / crawler4j

Code Issues 134 Pull requests 45 Actions Projects Wiki Security Insights

master 1 branch 4 tags Go to file Code

yasserg Merge pull request #454 from teeteejo/patch-1 ... 68f5c1e on Oct 3, 2020 515 commits

File	Description	Time Ago
.idea	Switch to gradle	2 years ago
config	Switch to gradle	2 years ago
crawler4j-examples	Delete Downloader example	2 years ago
crawler4j	Update the maximum size of a robots.txt file	15 months ago
gradle/wrapper	Switch to gradle	2 years ago
.editorconfig	Revert "prefer tab indents" – confuses CheckStyle too m...	2 years ago
.gitignore	Switch to gradle	2 years ago
.travis.yml	Switch Travis builds from oraclejdk8 to openjdk8	14 months ago
CHANGES.txt	Update CHANGES.txt	6 years ago
LICENSE	Update LICENSE	2 years ago
README.md	Fixed typo	4 months ago

About

Open Source Web Crawler for Java

Readme Apache-2.0 License

Releases 4

4.4.0 Latest on Mar 27, 2018 + 3 releases

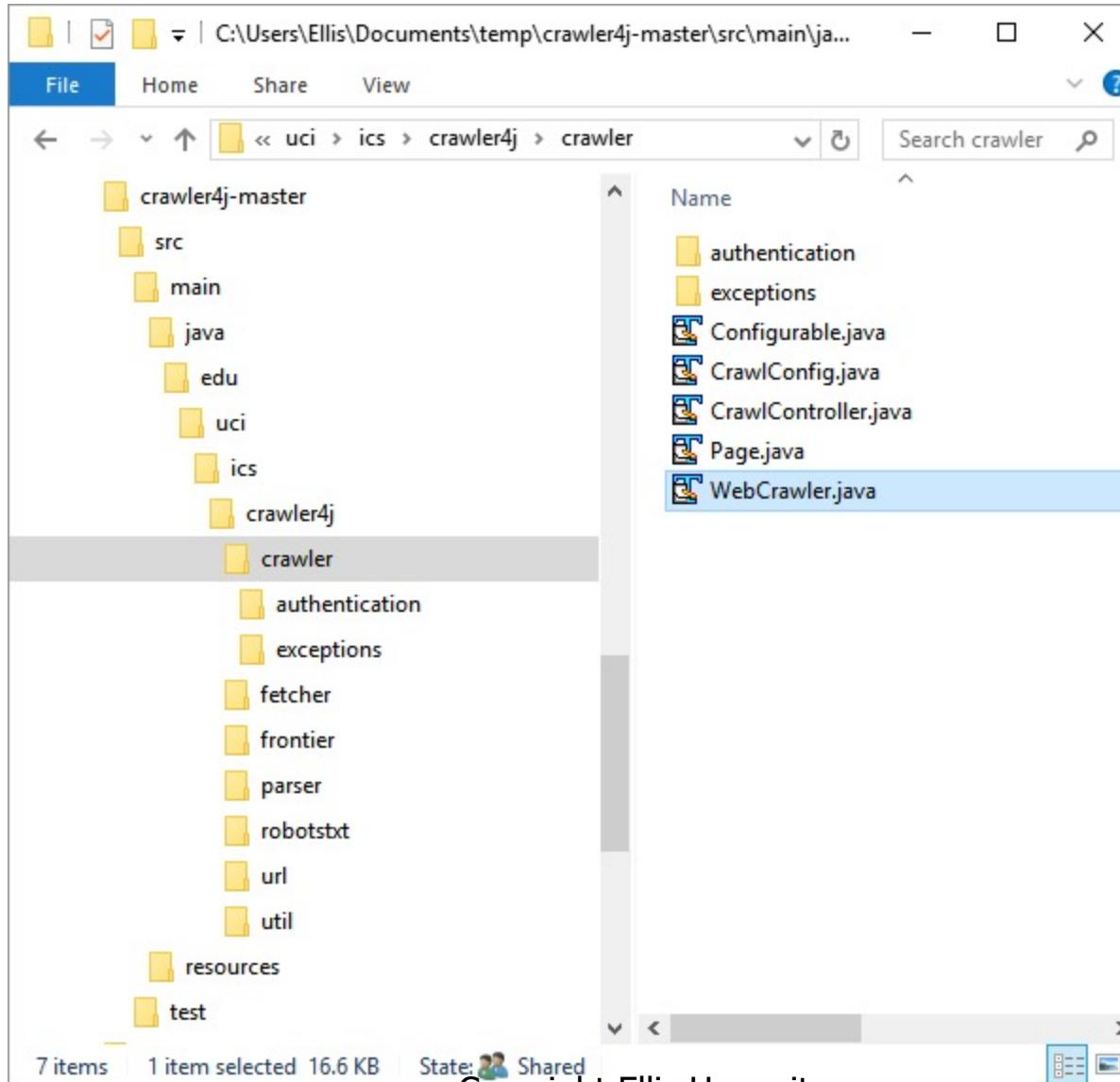
Packages

No packages published

Contributors 32

See especially the README file page at
<https://github.com/yasserg/crawler4j/blob/master/README.md>

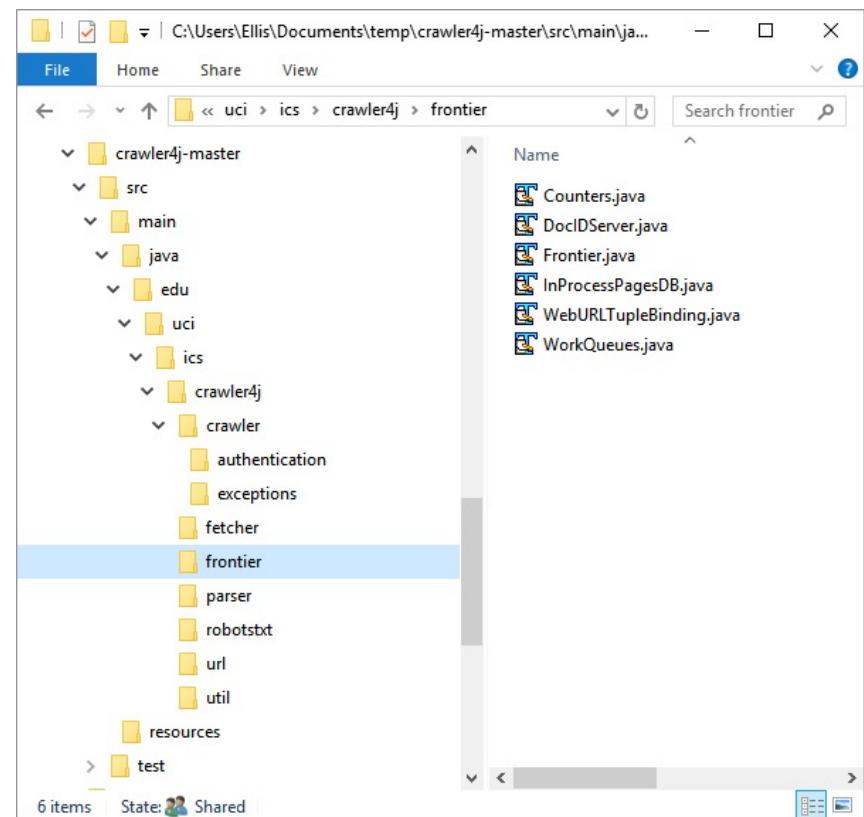
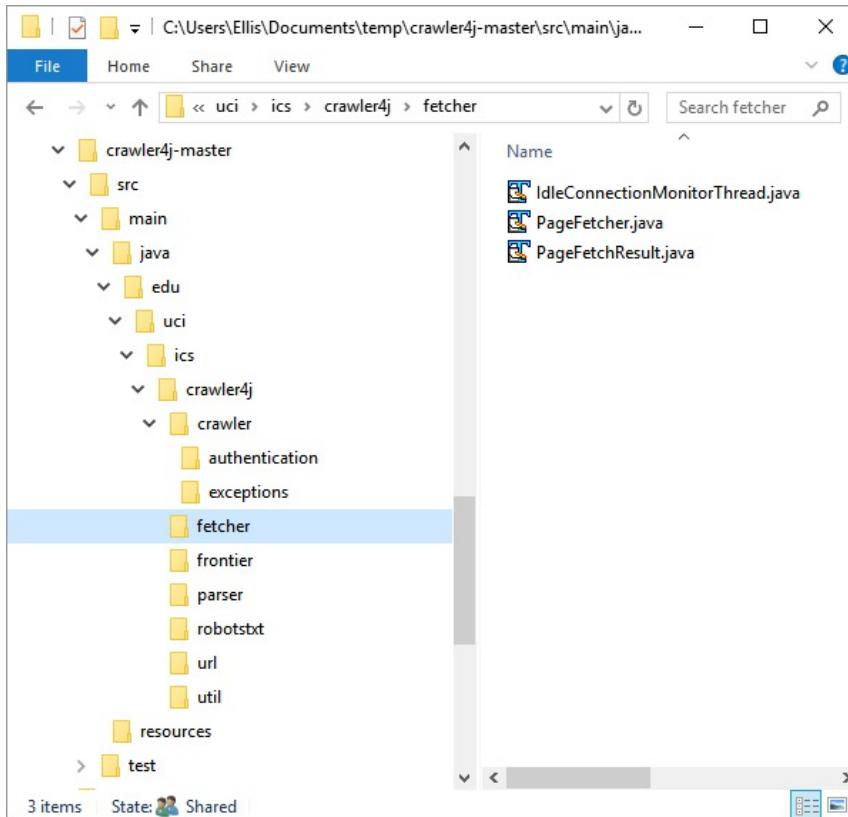
Crawler4j Source Code



Copyright Ellis Horowitz

Crawler folder, a good place to start; look especially at WebCrawler.java

Crawler4j Source code is Logically Organized into folders



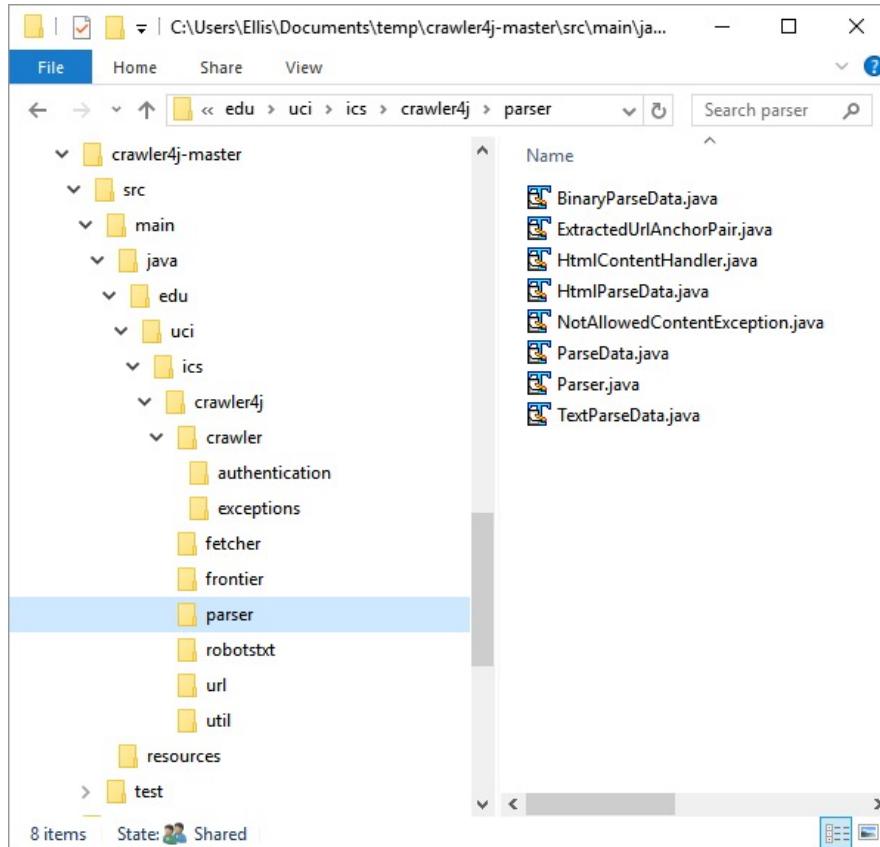
Fetcher Code handles:

- schemes: http, https
- politeness delay;
- redirects;
- max-size settings;
- expired connections

Frontier Code handles:

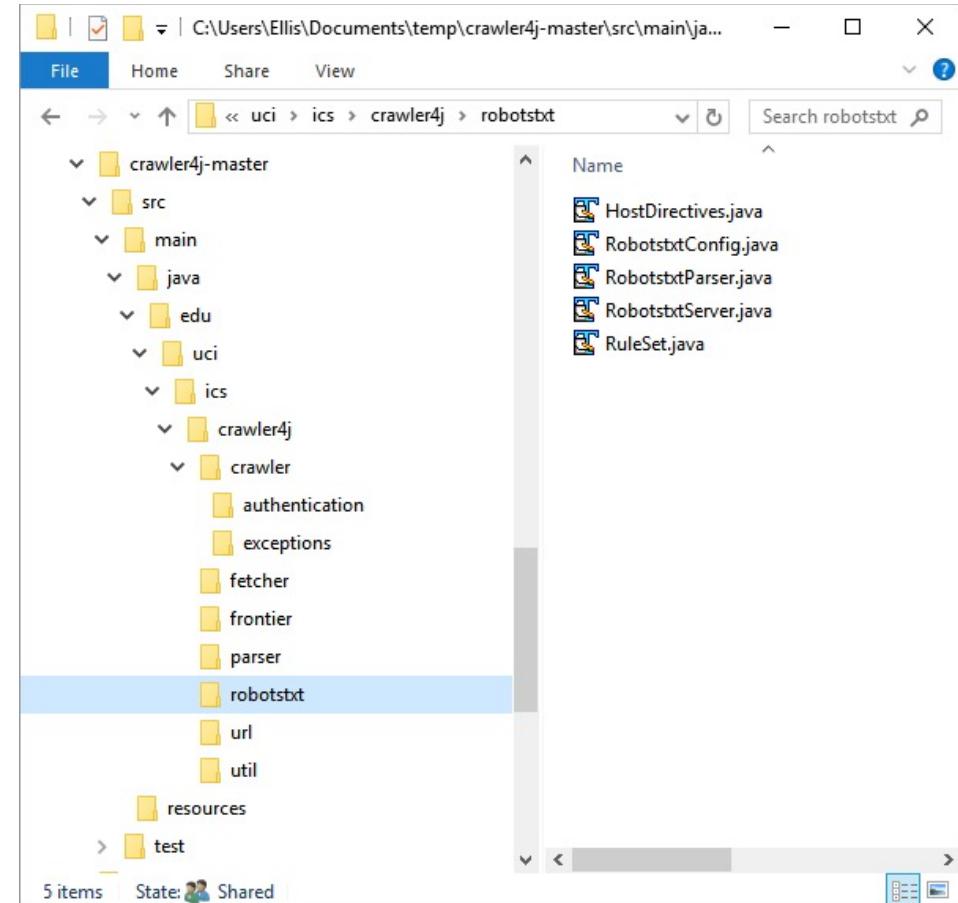
- statistics database;
- previously seen URLs
- queue of pending URLs

Crawler4j Routines are Named According to their Function



Parser Code handles:

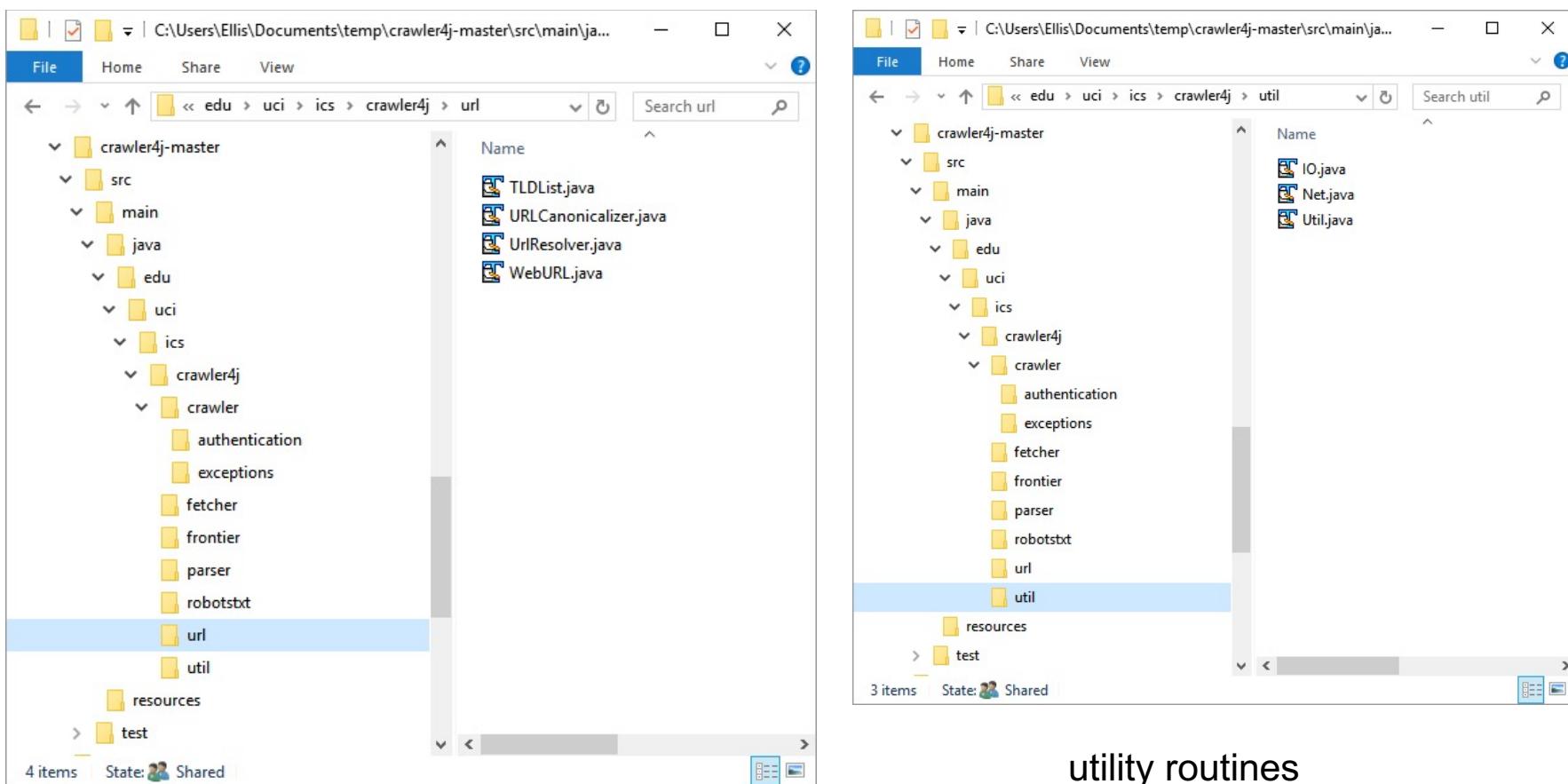
- binary data
- html pages
- extracting links



Robots.txt Code handles:

- fetching and re-fetching robots.txt
- caching robots.txt files
- interpreting commands
- working with Page Fetcher

More crawler4j Source code



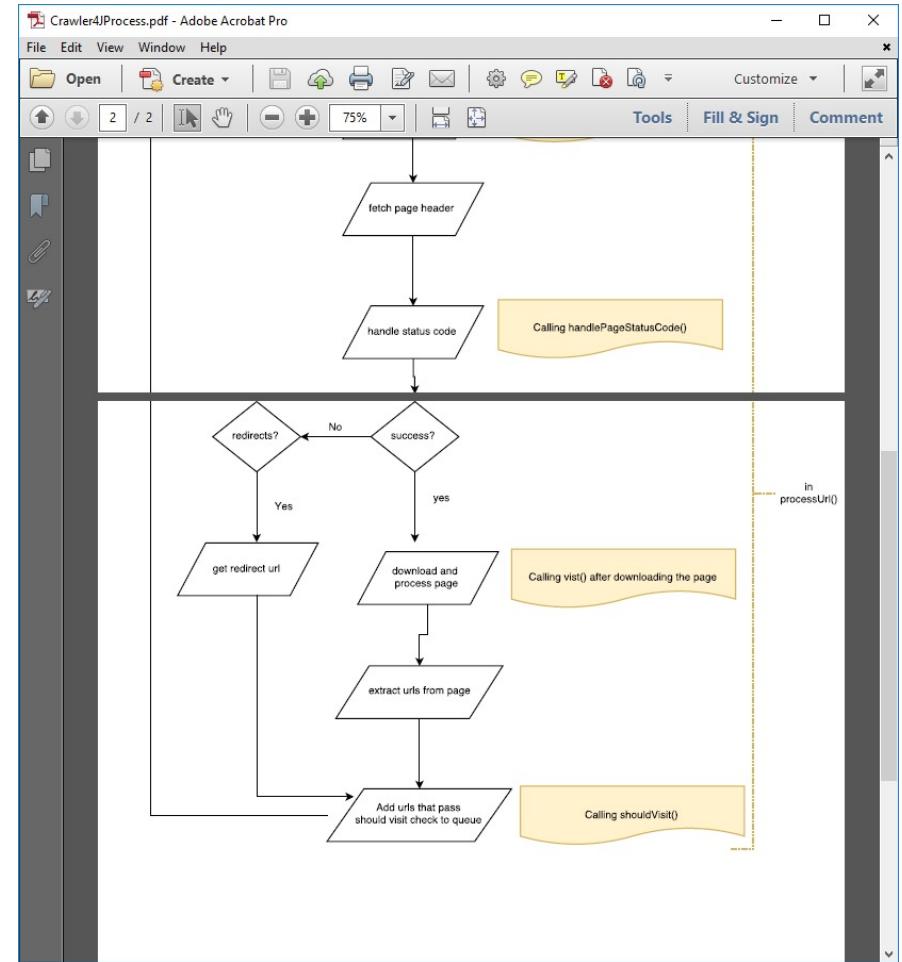
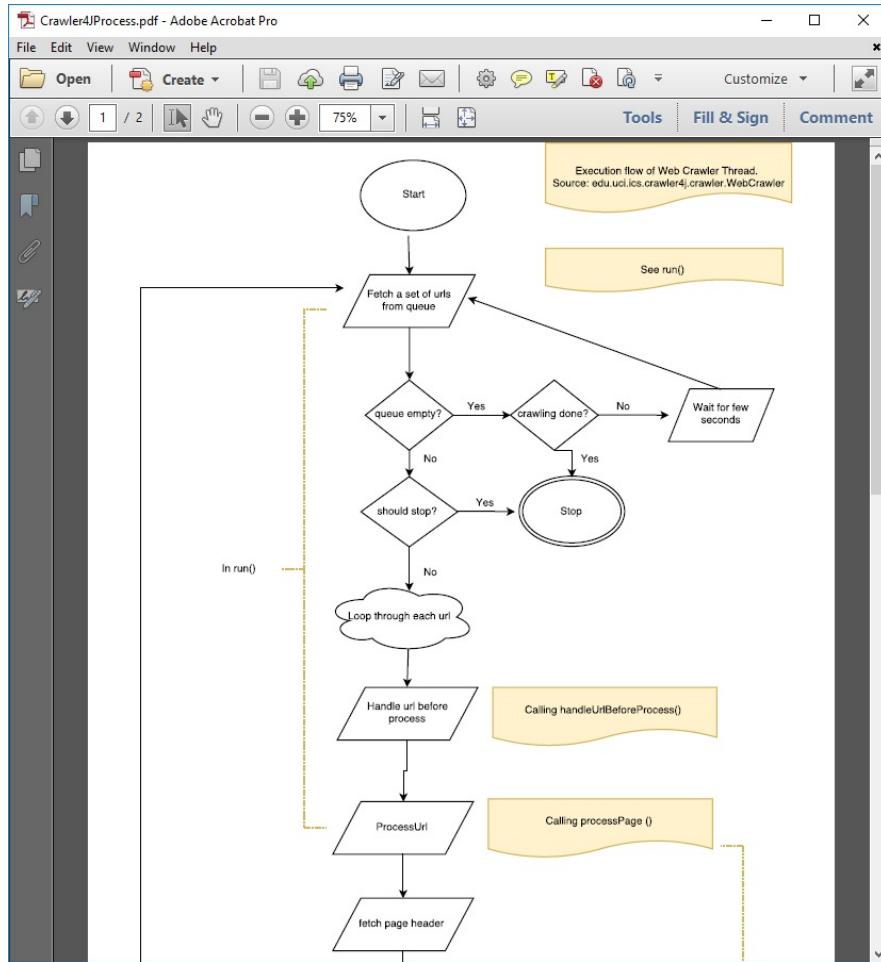
utility routines

URL resolver and canonicalizer handles:

- checking against list of TLDs
- normalizes URL, removes . or .., etc
- alters name/value pairs
- converts #nn values
- evaluates <base>

Copyright Ellis Horowitz

Logic Flowchart



<http://csci572.com/2022Spring/hw2/Crawler4JProcess.pdf>

Configuring the Crawler and Seeding it

```
public class Controller {  
    public static void main(String[] args) throws Exception {  
        String crawlStorageFolder = "/data/crawl"; ← folder to store  
        int numberOfCrawlers = 7; ← #crawlers  
        CrawlConfig config = new CrawlConfig();  
        config.setCrawlStorageFolder(crawlStorageFolder);  
        /* Instantiate the controller for this crawl.*/  
        PageFetcher pageFetcher = new PageFetcher(config); ← set up pagefetcher  
        RobotstxtConfig robotstxtConfig = new RobotstxtConfig(); ← and robots.txt  
        RobotstxtServer robotstxtServer = new RobotstxtServer(robotstxtConfig, pageFetcher);  
        CrawlController controller = new CrawlController(config, pageFetcher, robotstxtServer);  
        /* For each crawl, you need to add some seed urls. These are the first  
         * URLs that are fetched and then the crawler starts following links  
         * which are found in these pages */  
        controller.addSeed("https://www.nytimes.com/"); ← crawling  
        /* Start the crawl. This is a blocking operation, meaning that your code  
         * will reach the line after this only when crawling is finished. */  
        controller.start(MyCrawler.class, numberOfCrawlers);  
    }  
}
```

Crawling Pages Other Than HTML

- To make sure you are not just crawling HTML, and missing pdf and doc files, you need to set BinaryContent to true
- Examine the routine BasicCrawlController and see line 31
- Turn config.setIncludeBinaryContentInCrawling(false);

<https://github.com/yasserg/crawler4j/blob/master/crawler4j-examples/crawler4j-examples-base/src/test/java/edu/uci/ics/crawler4j/examples/basic/BasicCrawlController.java>

Defining Which Pages to Crawl

```
public class MyCrawler extends WebCrawler {  
    private final static Pattern FILTERS =  
        Pattern.compile(".*(\\".(css|js|gif|jpg" + "|png|mp3|mp3|zip|gz))$"); see next slide  
    /** This method receives two parameters. The first parameter is the page  
     * in which we have discovered this new url and the second parameter is  
     * the new url. You should implement this function to specify whether  
     * the given url should be crawled or not (based on your crawling logic).  
     * In this example, we are instructing the crawler to ignore urls that  
     * have css, js, git, ... extensions and to only accept urls that start  
     * with "http://www.latimes.com/". In this case, we didn't need the  
     * referring Page parameter to make the decision. */  
    @Override  
    public boolean shouldVisit(Page referringPage, WebURL url) {  
        String href = url.getURL().toLowerCase();  
        return !FILTERS.matcher(href).matches()  
            && href.startsWith("http://www.nytimes.com/");  
    }  
}
```

Matching URLs

- `".*(\\.\\.(css|js|gif|jpg" + "|png|mp3|mp4|zip|gz))$"`
- A regular expression, specified as a string, must first be compiled into an instance of this class.
- a Matcher object that can match arbitrary character sequences against the regular expression
- See <https://docs.oracle.com/javase/7/docs/api/java/util/regex/Pattern.html>
- In the above there are two strings concatenated by plus; consider the simpler form:
- `".*(\\.\\.(css|js|zip|gz))$"`
 - . matches any character
 - * matches zero or more of preceding character
 - \\. matches a literal dot
 - \$ anchors the pattern at the end of the string

Parsing the Downloaded Page

```
/** This function is called when a page is fetched and ready
 * to be processed by your program. */
@Override
public void visit(Page page) {
    String url = page.getWebURL().getURL();
    System.out.println("URL: " + url);
    if (page.getParseData() instanceof HtmlParseData) {
        HtmlParseData htmlParseData = (HtmlParseData) page.getParseData();
        String text = htmlParseData.getText();
        String html = htmlParseData.getHtml();
        Set<WebURL> links = htmlParseData.getOutgoingUrls();
        System.out.println("Text length: " + text.length());
        System.out.println("Html length: " + html.length());
        System.out.println("Number of outgoing links: " + links.size());
    }
}
```

The Actual Exercise

- the URLs it attempts to fetch, `fetch.csv`. The number of rows should be no more than 20,000 as that is our pre-set limit.
- the files it successfully downloads, `visit.csv`; clearly the number of rows will be less than the number of rows in `fetch.csv`
- all of the URLs that were discovered and processed in some way; `urls.csv`. This file could be much larger than 20,000 rows as it will have numerous repeated URLs

Things to Save

- Fetch statistics:
 - # fetches attempted:

The total number of URLs that the crawler attempted to fetch. This is usually equal to the MAXPAGES setting if the crawler reached that limit; less if the website is smaller than that.
 - # fetches succeeded:

The number of URLs that were successfully downloaded in their entirety, i.e. returning a HTTP status code of 2XX.
 - # fetches failed or aborted:

The number of fetches that failed for whatever reason, including, but not limited to: HTTP redirections (3XX), client errors (4XX), server errors (5XX) and other network-related errors.
-

Outgoing URLs

- Outgoing URLs: statistics about URLs extracted from visited HTML pages
 - Total URLs extracted:
The grand total number of URLs extracted from all visited pages
 - # unique URLs extracted:
The number of unique URLs encountered by the crawler
 - # unique URLs within the news web site:
The number of unique URLs encountered that are associated with the news website,
i.e. the URL begins with the given root URL of the news website.
 - # unique URLs outside the news website:
The number of unique URLs encountered that were not from the website.

Sample Crawl Report for NY Times

Using 20,000 as the Download Limit

```
News site crawled: https://www.nytimes.com/
Fetch Statistics
=====
# fetches attempted:19750
# fetches succeeded:19067
# fetches aborted or failed:683
Outgoing URLs
=====
Total URLs extracted:964461
# unique URLs extracted:73450
# unique URLs within nytimes:31115
# unique URLs outside nytimes:42335
Status Codes
=====
200 OK:19067
301 Moved Permanently:140
302 Moved Temporarily:21
503 Service Unavailable:366
401 Unauthorized Error:1
403 Forbidden:8
404 Not Found:143
500 Internal Server Error:1
400 Bad Request:3
File Sizes
=====
<1KB:14
1KB~<10KB:1052
10KB~<100KB:694
100KB~<1MB:8602
>=1MB:1
Content Type
=====
text/html:18349
image/jpeg=5
image/png=3
```

Sample Fetch File for NY Times

	A	B	C	D	E	F	G	H	I	J	K	L
2	https://www.nytimes.com/	200										
3	https://www.nytimes.com/video	200										
4	https://www.nytimes.com/section/arts/dance	200										
5	https://www.nytimes.com/section/us	200										
6	https://www.nytimes.com/svc/collections/v1/publish/https://www.nytimes.com/se	200										
7	https://www.nytimes.com/section/technology	200										
8	https://www.nytimes.com/svc/collections/v1/publish/https://www.nytimes.com/se	200										
9	https://www.nytimes.com/column/speakingindance	200										
10	https://www.nytimes.com/live/2020/impeachment-trial-live-01-27/the-house-want	200										
11	https://www.nytimes.com/svc/collections/v1/publish/https://www.nytimes.com/se	200										
12	https://www.nytimes.com/video/investigations	200										
13	https://www.nytimes.com/subscription?campaignId=37WXW	200										
14	https://www.nytimes.com/section/science	200										
15	https://www.nytimes.com/section/business/smallbusiness	200										
16	https://www.nytimes.com/svc/collections/v1/publish/https://www.nytimes.com/co	200										
17	https://www.nytimes.com/svc/collections/v1/publish/https://www.nytimes.com/se	200										
18	https://www.nytimes.com/section/food	200										
19	https://www.nytimes.com/news-event/2020-election	200										
20	https://www.nytimes.com/live/2020/impeachment-trial-live-01-27/starr-manages-t	200										
21	https://www.nytimes.com/subscription	200										
22	https://www.nytimes.com/live/2020/impeachment-trial-live-01-27/rudy-giuliani-is-a	200										

Sample Visit File for NY Times

The screenshot shows a Microsoft Excel spreadsheet titled "Sample Visit File for NY Times". The table consists of 21 rows and 5 columns. The columns are labeled A through F, and the rows are numbered 1 through 21. The data includes URLs from the NY Times website, their file sizes, the number of outgoing links, and their content type.

	A	B	C	D	E	F	G	H	I	J	K	L
1	URL	Size	Outgoing Links	Content Type								
2	https://www.nytimes.com/	1451085	139	text/html								
3	https://www.nytimes.com/video	429553	109	text/html								
4	https://www.nytimes.com/section/arts/dance	736631	157	text/html								
5	https://www.nytimes.com/section/us	946441	120	text/html								
6	https://www.nytimes.com/section/technology	1112691	161	text/html								
7	https://www.nytimes.com/column/speakingindance	459835	141	text/html								
8	https://www.nytimes.com/live/2020/impeachment-trial-live-01-27/the-k	314892	149	text/html								
9	https://www.nytimes.com/subscription?campaignId=37WXW	72384	18	text/html								
10	https://www.nytimes.com/video/investigations	1247877	166	text/html								
11	https://www.nytimes.com/section/science	1012715	182	text/html								
12	https://www.nytimes.com/section/business/smallbusiness	817605	141	text/html								
13	https://www.nytimes.com/section/food	1092598	173	text/html								
14	https://www.nytimes.com/news-event/2020-election	1338179	173	text/html								
15	https://www.nytimes.com/subscription	72384	18	text/html								
16	https://www.nytimes.com/live/2020/impeachment-trial-live-01-27/starr	256830	136	text/html								
17	https://www.nytimes.com/live/2020/impeachment-trial-live-01-27/rudy	309304	148	text/html								
18	https://www.nytimes.com/newsletters/signup/CN	107087	92	text/html								
19	https://www.nytimes.com/subscription/education?campaignId=7KL9U	93183	19	text/html								
20	https://www.nytimes.com/live/2020/impeachment-trial-live-01-27/trum	339037	153	text/html								
21	https://www.nytimes.com/section/food/drinks	811026	166	text/html								

What to Submit

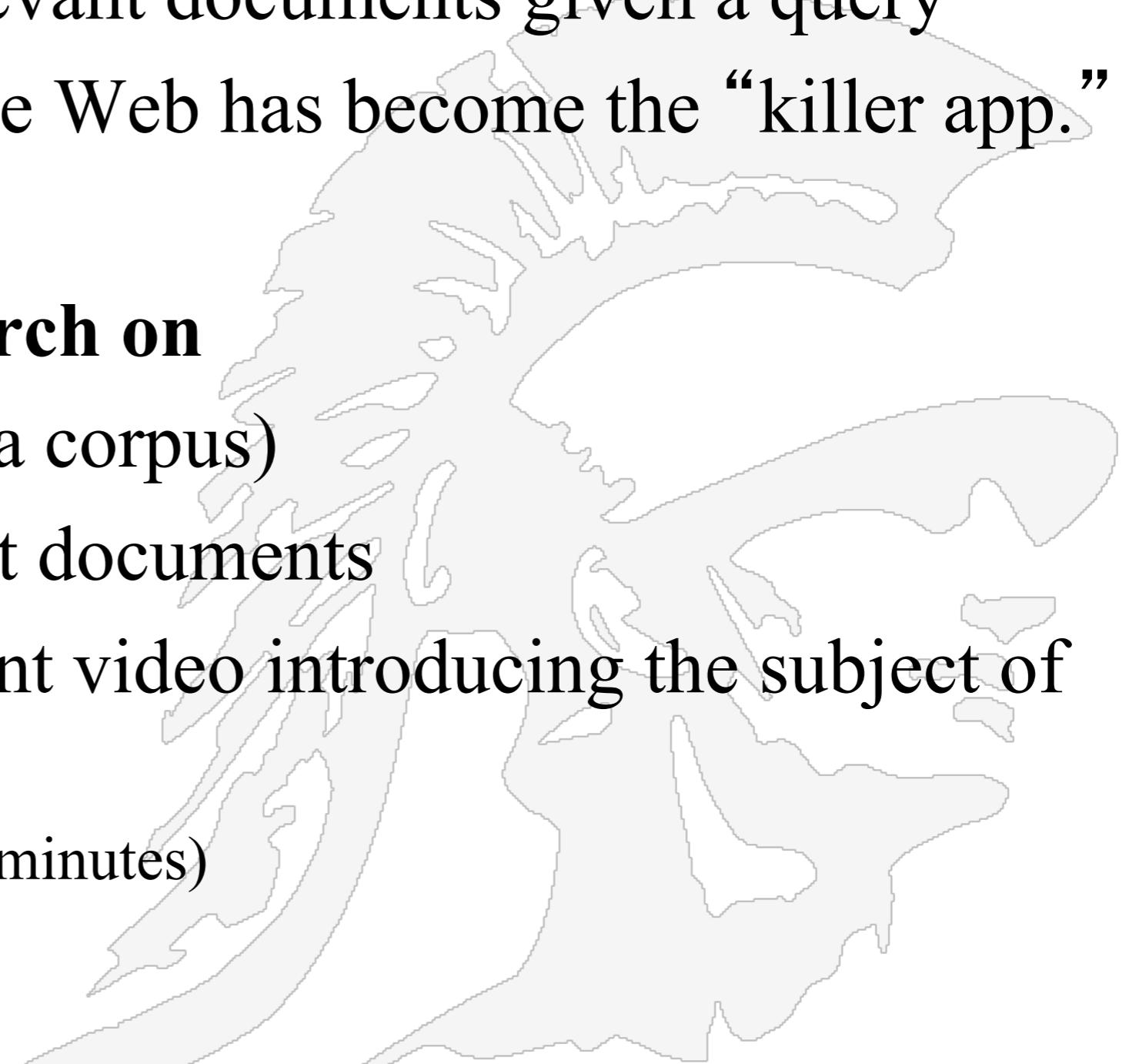
- Compress all of the above into a single zip archive and name it:
crawl.zip
- Use only standard zip format. Do NOT use other formats such as zipx, rar, ace, etc. For example the zip file might contain the following three files:
 1. CrawlReport_nytimes.txt,
 2. fetch_nytimes.csv
 3. visit_nytimes.csv
- Place your crawl.zip file in your csci572/hw2 folder

Introduction to Information Retrieval

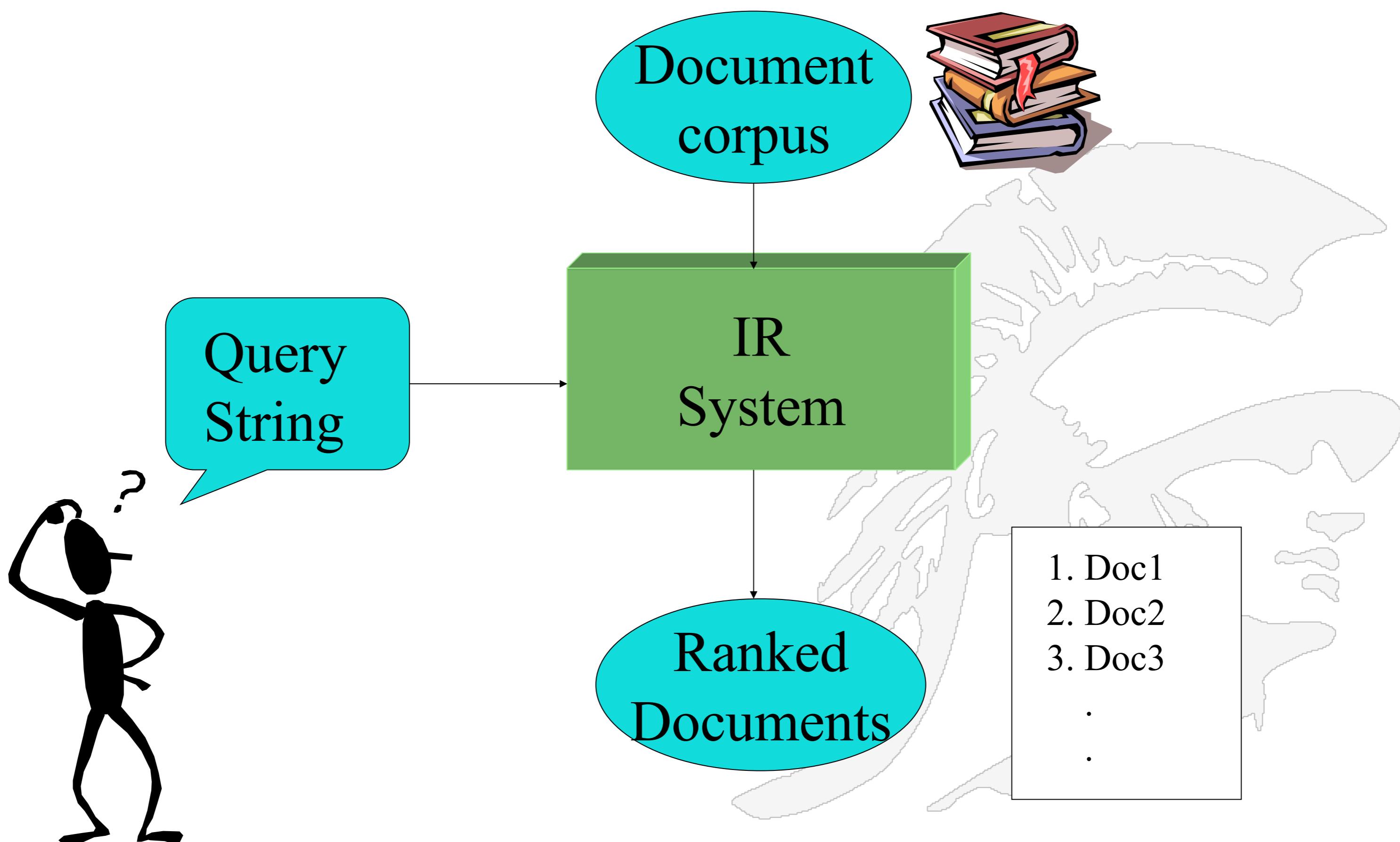


Information Retrieval

- ***Information retrieval (IR) has been a computer science subject for many decades***
 - Traditionally it deals with the *indexing* of a given set of textual documents and the *retrieval* of relevant documents given a query
- Searching for pages on the World Wide Web has become the “killer app.”
- **There has been a great deal of research on**
 - How to index a set of documents (a corpus)
 - How to efficiently retrieve relevant documents
- Jurafsky and Manning have an excellent video introducing the subject of Information Retrieval;
- http://csci572.com/movies/01_IntroIR.mp4 (9 minutes)
then jump to slide 16

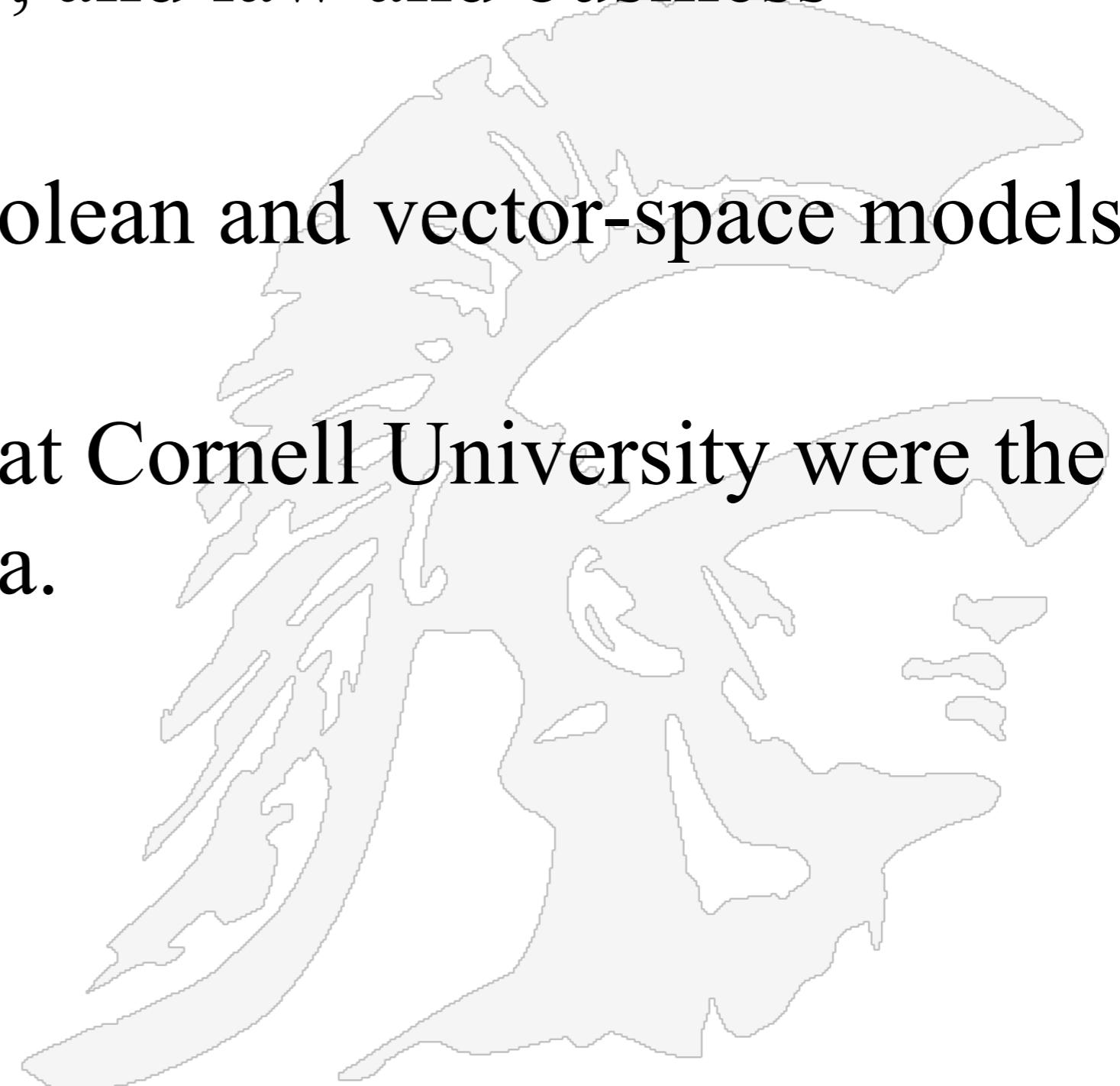


The Traditional IR System



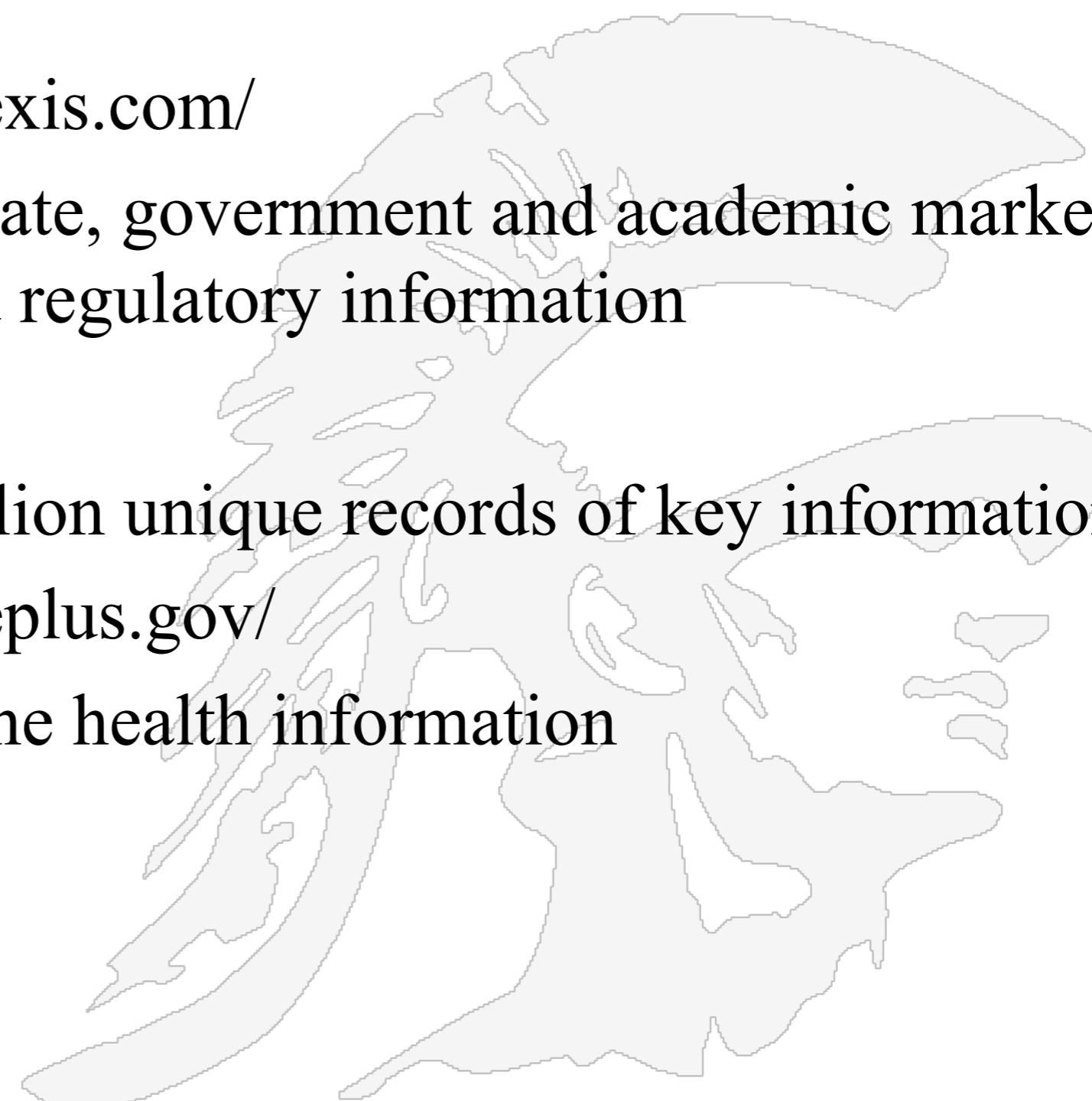
History of IR

- **1960-70's:**
 - Initial exploration of text retrieval systems for “small” corpora of scientific abstracts, and law and business documents.
 - Development of the basic Boolean and vector-space models of retrieval.
 - Prof. Salton and his students at Cornell University were the leading researchers in the area.



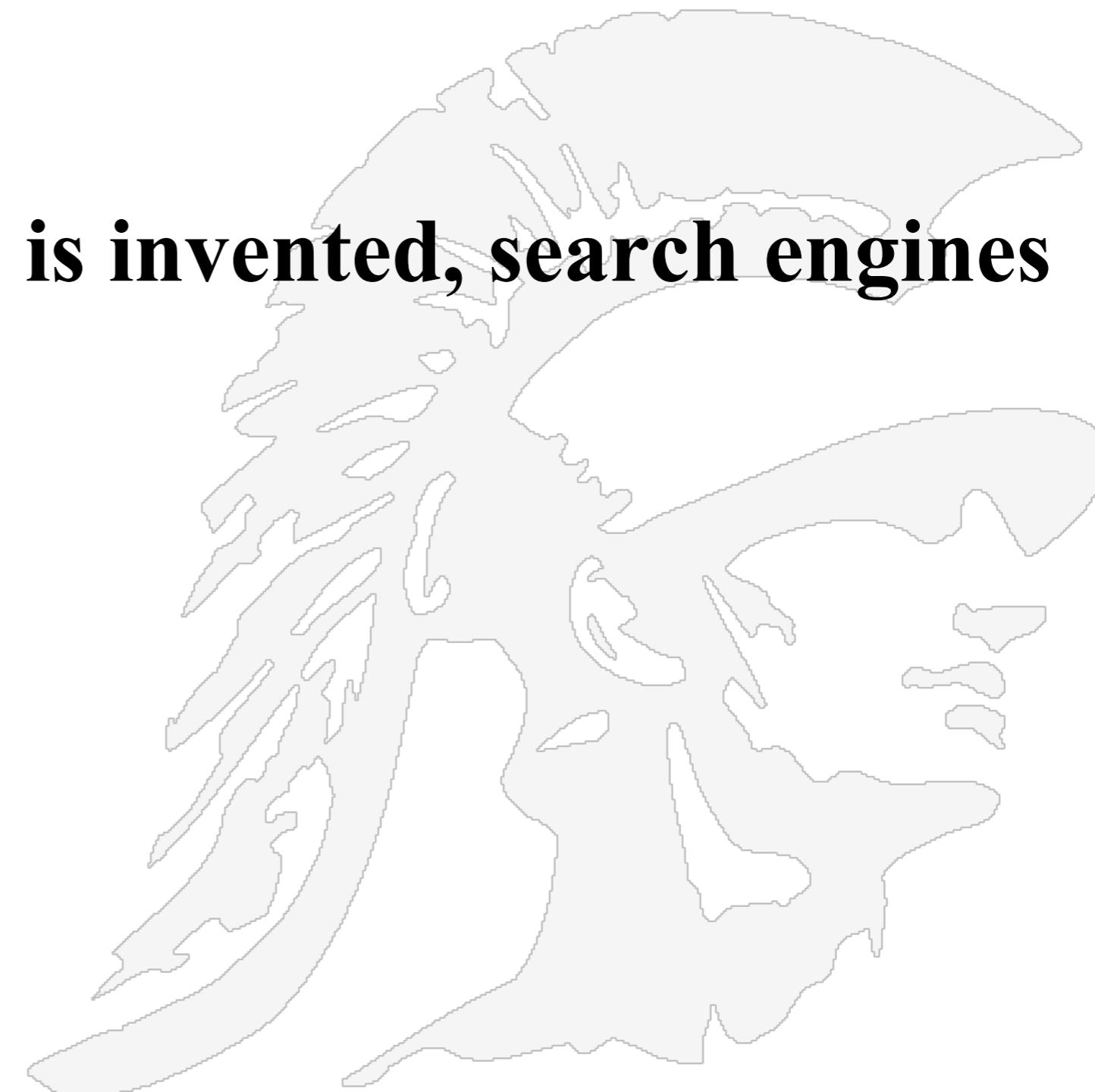
IR History Continued

- **1980's:**
 - **Creation of large document database systems, many run by companies:**
 - Lexis-Nexis, <http://www.lexisnexis.com/>
 - information to legal, corporate, government and academic markets, and publishes legal, tax and regulatory information
 - Dialog, <http://www.dialog.com/>
 - data from more than 1.4 billion unique records of key information.
 - MEDLINE, <http://www.medlineplus.gov/>
 - National Library of Medicine health information



IR History Continued

- 1990's:
 - Searching FTP'able documents on the Internet
 - Archie
 - WAIS
 - After the World Wide Web is invented, search engines appear
 - Lycos
 - Yahoo
 - Altavista



IR History Continued

- 1990's continued:
 - Organized Competitions
 - NIST TREC (Text REtrieval Conferences, <http://trec.nist.gov/>)
 - Sponsored by National Institute of Standards and Technology, NIST
 - Several New Types of IR Systems are Developed

1. *Recommender Systems*: computer programs which attempt to predict items (movies, music, books, news, web pages) that a user may be interested in, given some information about the user's profile.

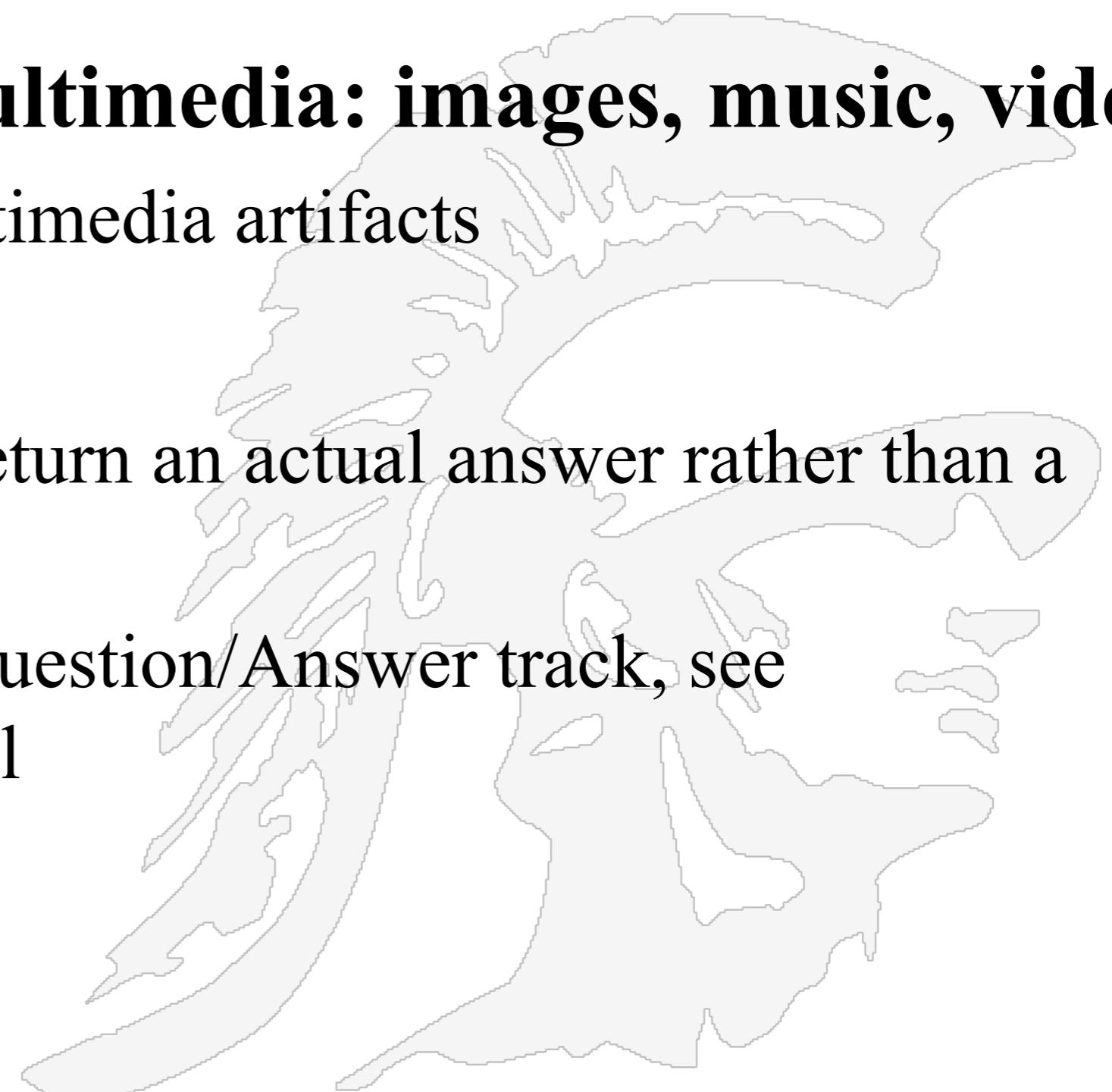
- Often implemented as a collaborative filtering algorithm, examples include:
 - » YouTube, perhaps the largest scale such system in existence
 - » <https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/45530.pdf>
 - » Amazon's recommendation system, see
<https://stackoverflow.com/questions/2323768/how-does-the-amazon-recommendation-feature-work>
 - » <http://rejoiner.com/resources/amazon-recommendations-secret-selling-online/>

2. *Automated Text Categorization & Clustering Systems*

- Useful for grouping news articles

Recent IR History Moves to the Web

- **2000's**
 - **Link analysis for Web Search**
 - Google started this
 - **Extension to retrieval of multimedia: images, music, video**
 - It is much harder to index multimedia artifacts
 - **Question Answering**
 - Question answering systems return an actual answer rather than a ranked list of documents
 - Since 1999 TREC has had a Question/Answer track, see <http://trec.nist.gov/data/qa.html>





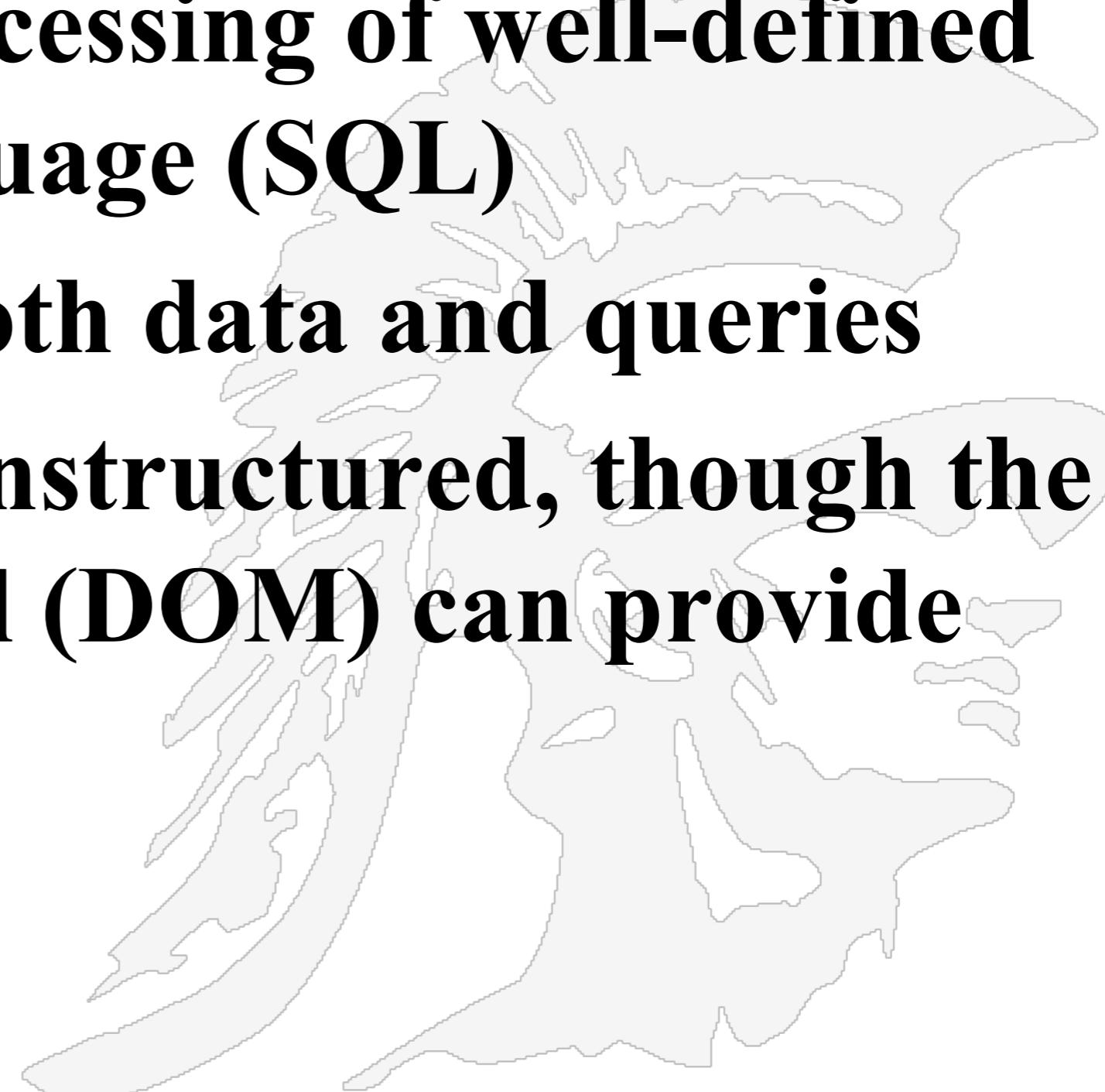
Areas Related To, But Different Than, Information Retrieval

- Database Management
- Library and Information Science
- Artificial Intelligence
- Natural Language Processing
- Machine Learning
- Data Science

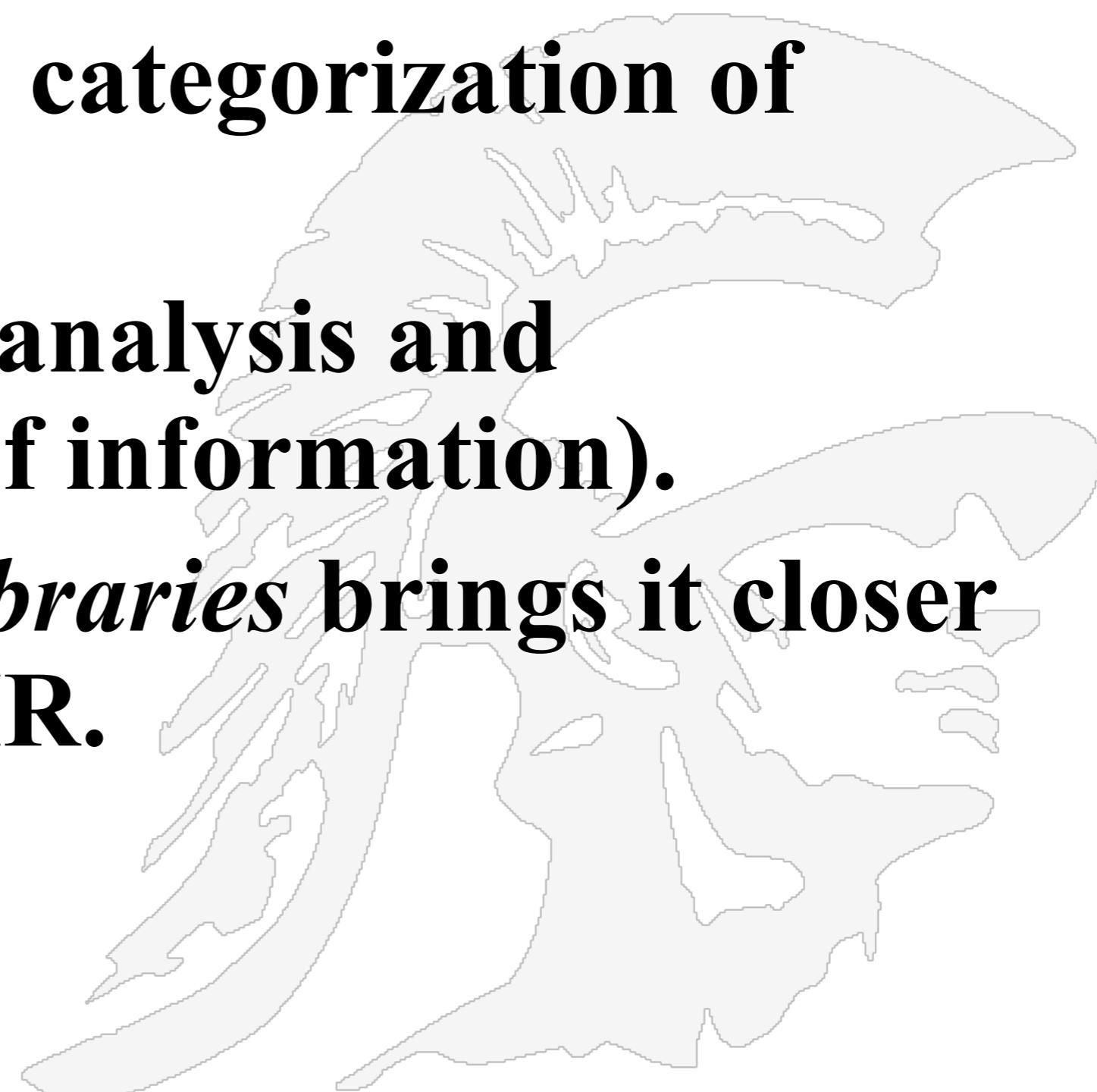


Database Management is Different from IR

- Focused on *structured* data stored in relational tables rather than free-form text
- Focused on efficient processing of well-defined queries in a formal language (SQL)
- Clearer semantics for both data and queries
- Web pages are mostly unstructured, though the Document Object Model (DOM) can provide some clues

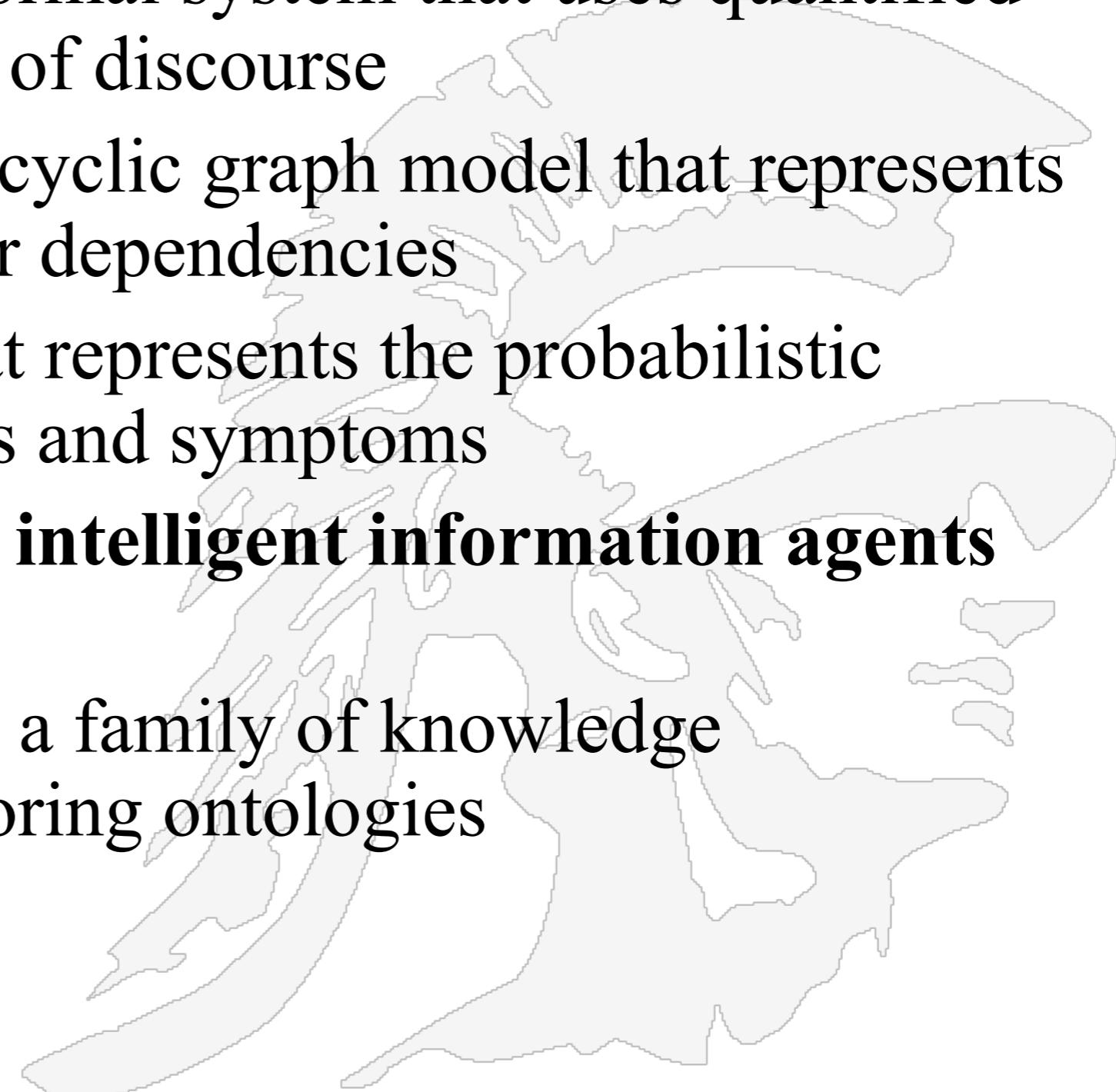


- Focused on the human user aspects of information retrieval (human-computer interaction, user interface, visualization).
- Concerned with effective categorization of human knowledge.
- Concerned with citation analysis and *bibliometrics* (structure of information).
- Recent work on *digital libraries* brings it closer to Computer Science & IR.



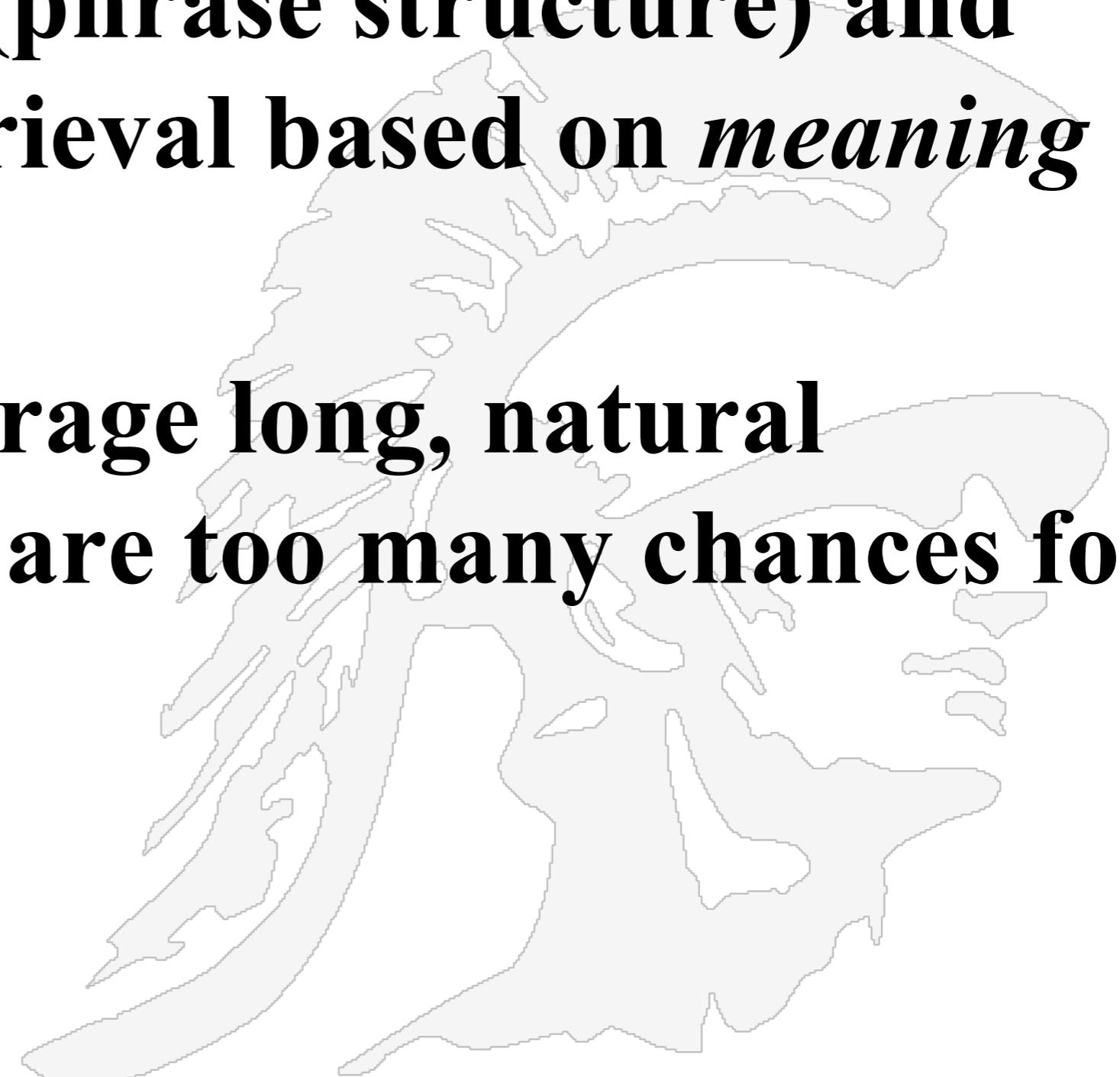
Artificial Intelligence

- Focused on the representation of knowledge, reasoning, and intelligent action.
- Formalisms for representing knowledge and queries:
 - *First-order Predicate Logic* – a formal system that uses quantified variables over a specified domain of discourse
 - *Bayesian Networks* – a directed acyclic graph model that represents a set of random variables and their dependencies
 - E.g. A Bayesian Network that represents the probabilistic relationships between diseases and symptoms
- Recent work on web ontologies and intelligent information agents brings it closer to IR
 - Web Ontology Language OWL is a family of knowledge representation languages for authoring ontologies
 - See <https://www.w3.org/OWL/>



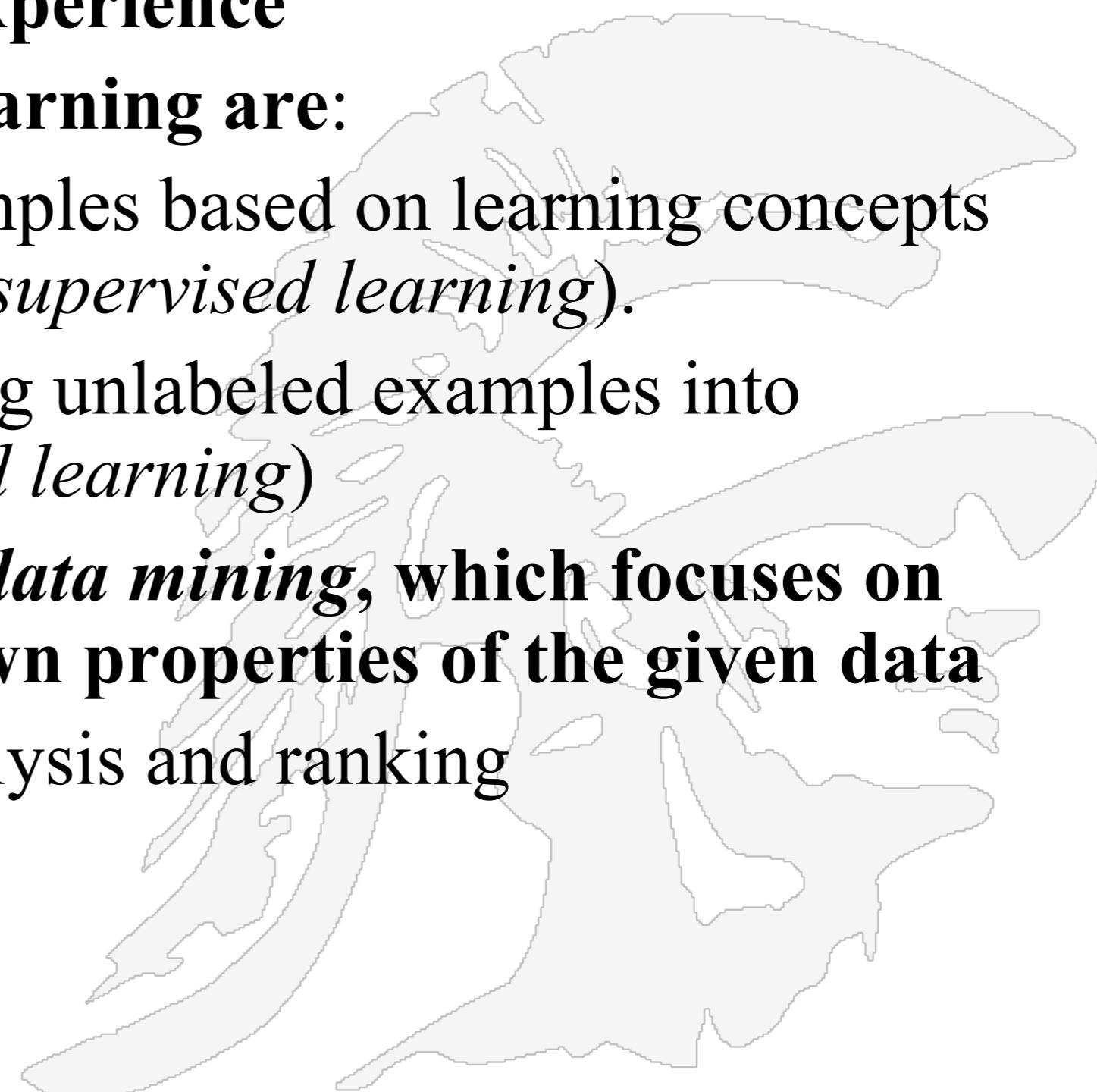
Natural Language Processing

- Focused on the syntactic, semantic, and pragmatic analysis of natural language text and discourse.
- Ability to analyze syntax (phrase structure) and semantics could allow retrieval based on *meaning* rather than keywords
- But search engines discourage long, natural language queries as there are too many chances for ambiguity of meaning



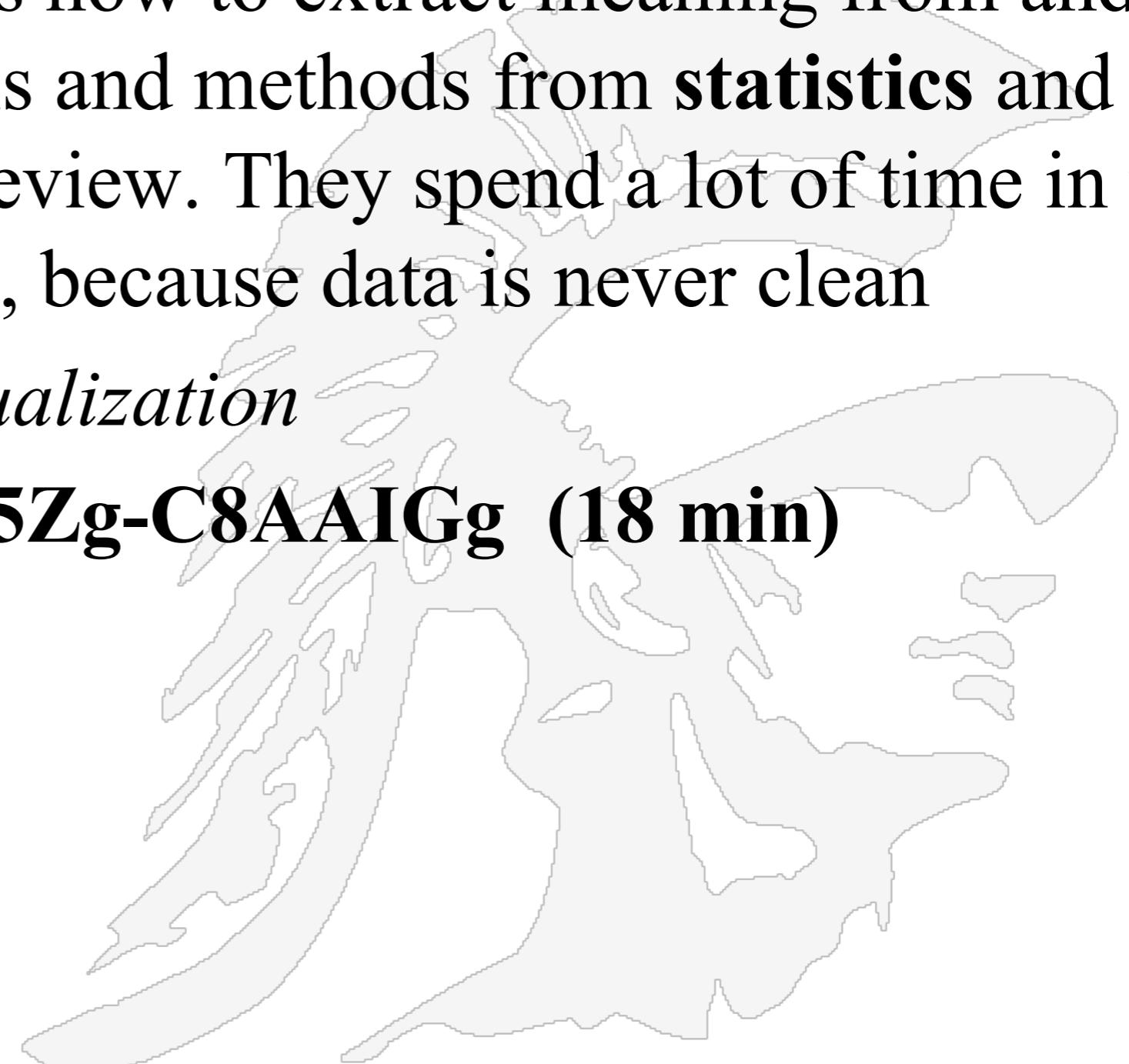
Machine Learning

- A branch of Artificial Intelligence concerned with algorithms that allow computers to evolve their behavior based on empirical data
- Focused on the development of computational systems that improve their performance with experience
- Two major subtypes of machine learning are:
 - Automated classification of examples based on learning concepts from labeled training examples (*supervised learning*).
 - Automated methods for clustering unlabeled examples into meaningful groups (*unsupervised learning*)
- Machine learning is distinct from *data mining*, which focuses on the discovery of previously unknown properties of the given data
 - Data mining is akin to query analysis and ranking



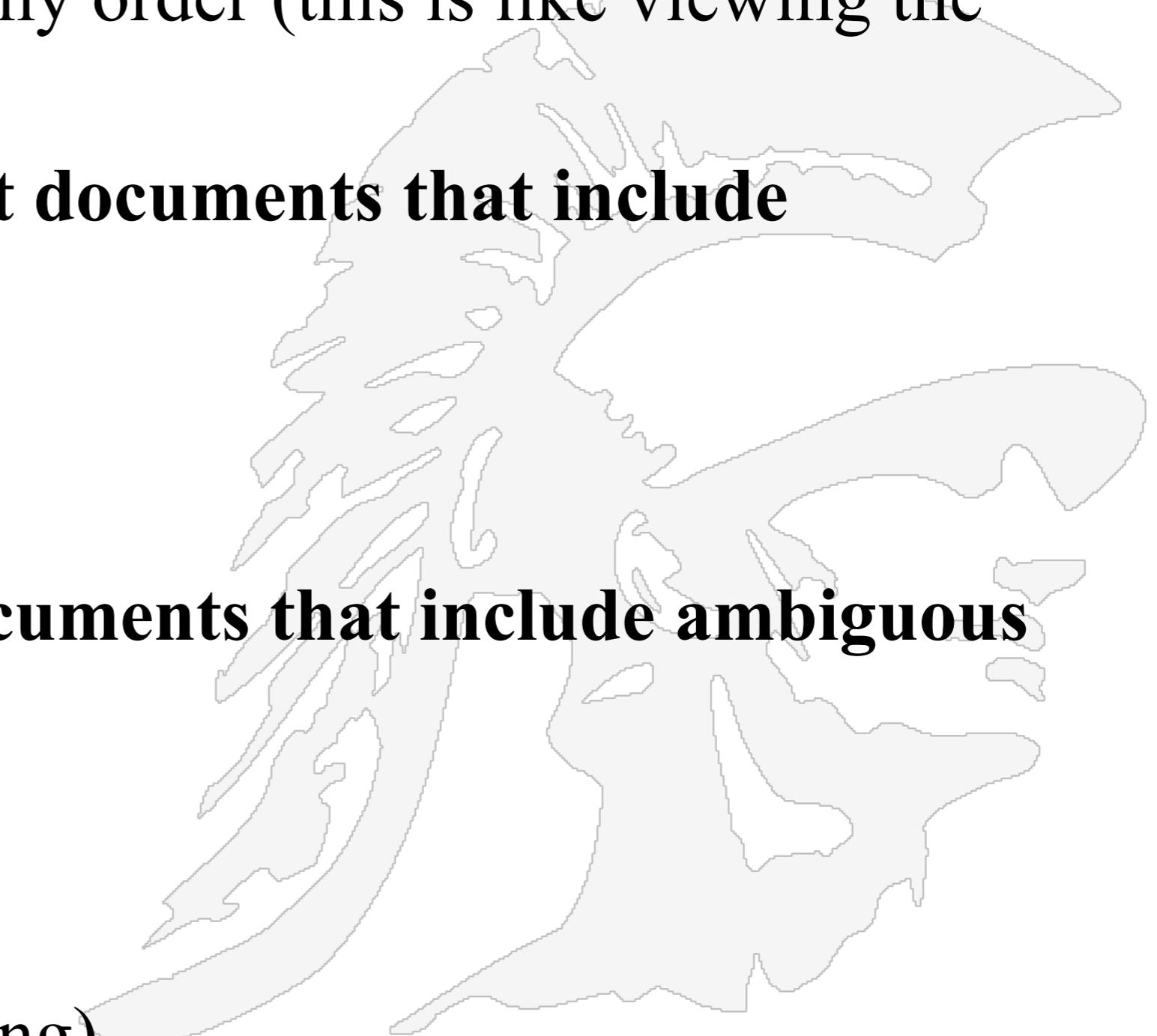
Data Science

- “*Data science* is an inter-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from many structural and unstructured data. Data science is related to data mining, machine learning and big data.” wikipedia
- A *data scientist* is someone who knows how to extract meaning from and interpret data, which requires both tools and methods from **statistics** and **machine learning**, as well as human review. They spend a lot of time in the process of collecting and cleaning data, because data is never clean
- For fun watch *The Beauty of Data Visualization*
- <https://www.youtube.com/watch?v=5Zg-C8AAIGg> (18 min)

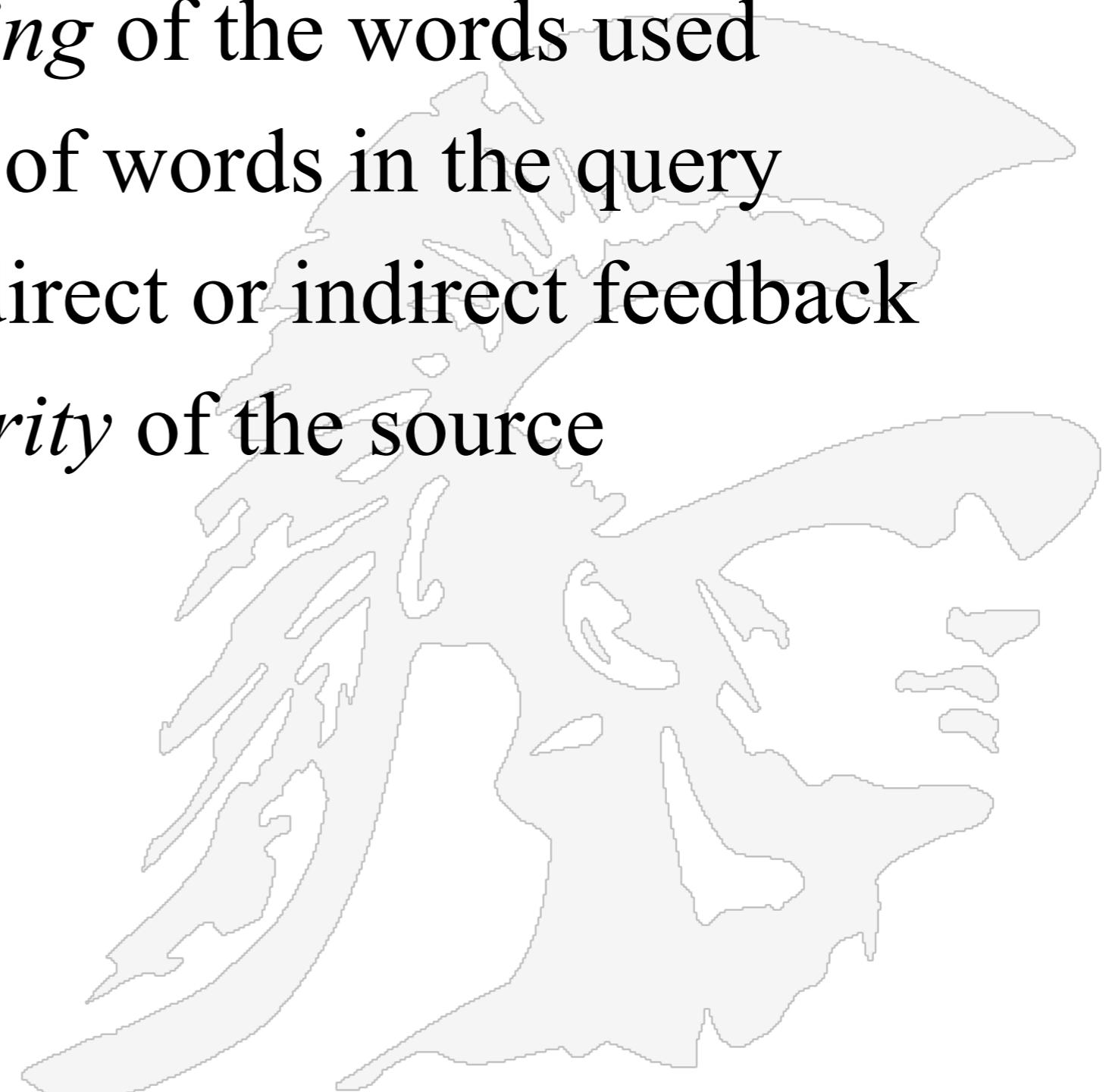


Basic Information Retrieval Begins with Keyword Matching

- **Simplest notion of relevance is that the query string appears verbatim in the document.**
 - Slightly less strict notion is that the words in the query appear frequently in the document, in any order (this is like viewing the document as a *bag of words*).
- **But that may not retrieve relevant documents that include synonymous terms.**
 - “restaurant” vs. “café”
 - “PRC” vs. “China”
- **And it may retrieve irrelevant documents that include ambiguous terms.**
 - “bat” (baseball vs. mammal)
 - “Apple” (company vs. fruit)
 - “bit” (unit of data vs. act of eating)



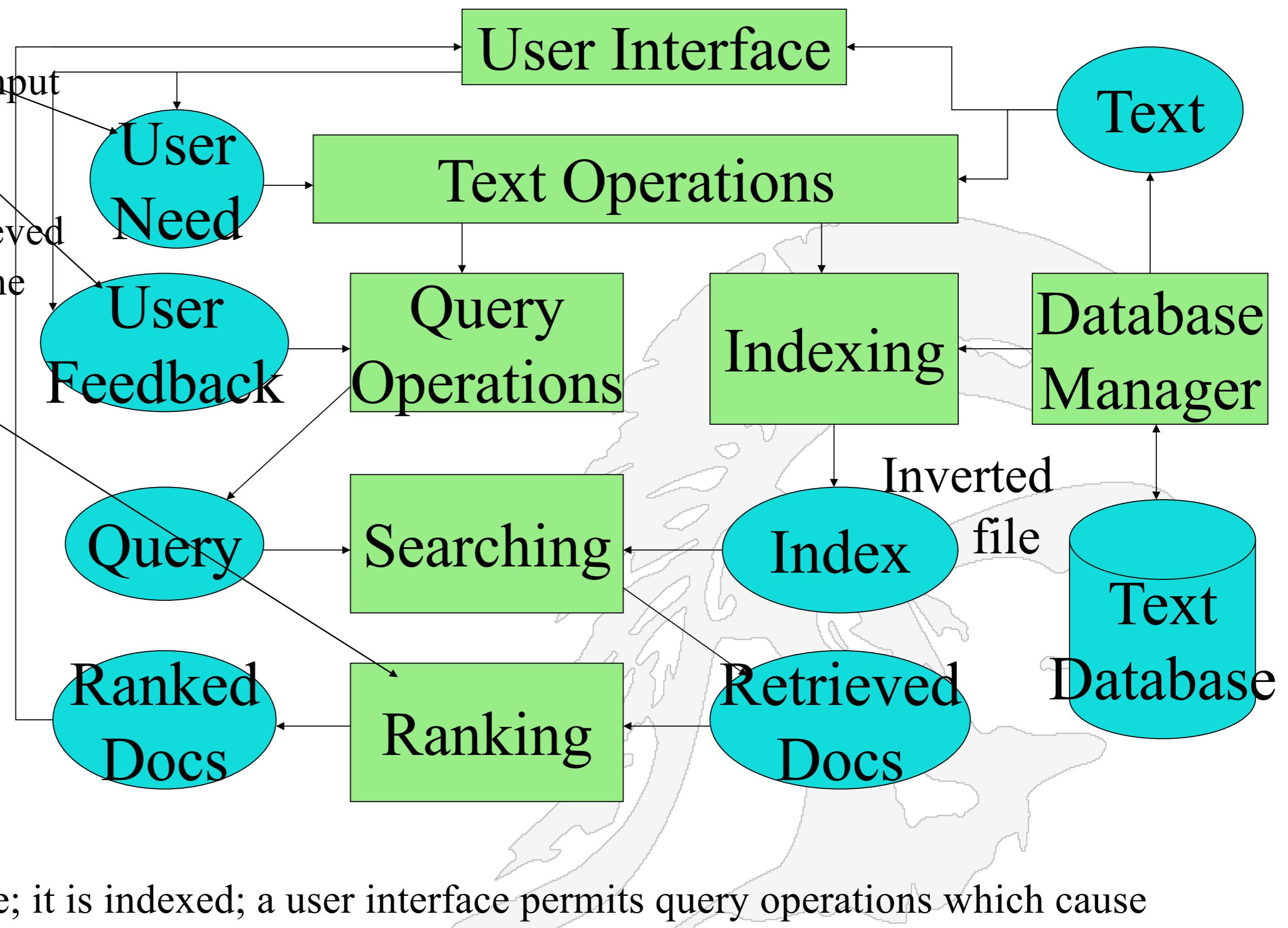
- **Goes beyond using just keyword matching, instead it**
 - Takes into account the *meaning* of the words used
 - Takes into account the *order* of words in the query
 - Adapts to the user based on direct or indirect feedback
 - Takes into account the *authority* of the source



A More Detailed IR Architecture

Logical View

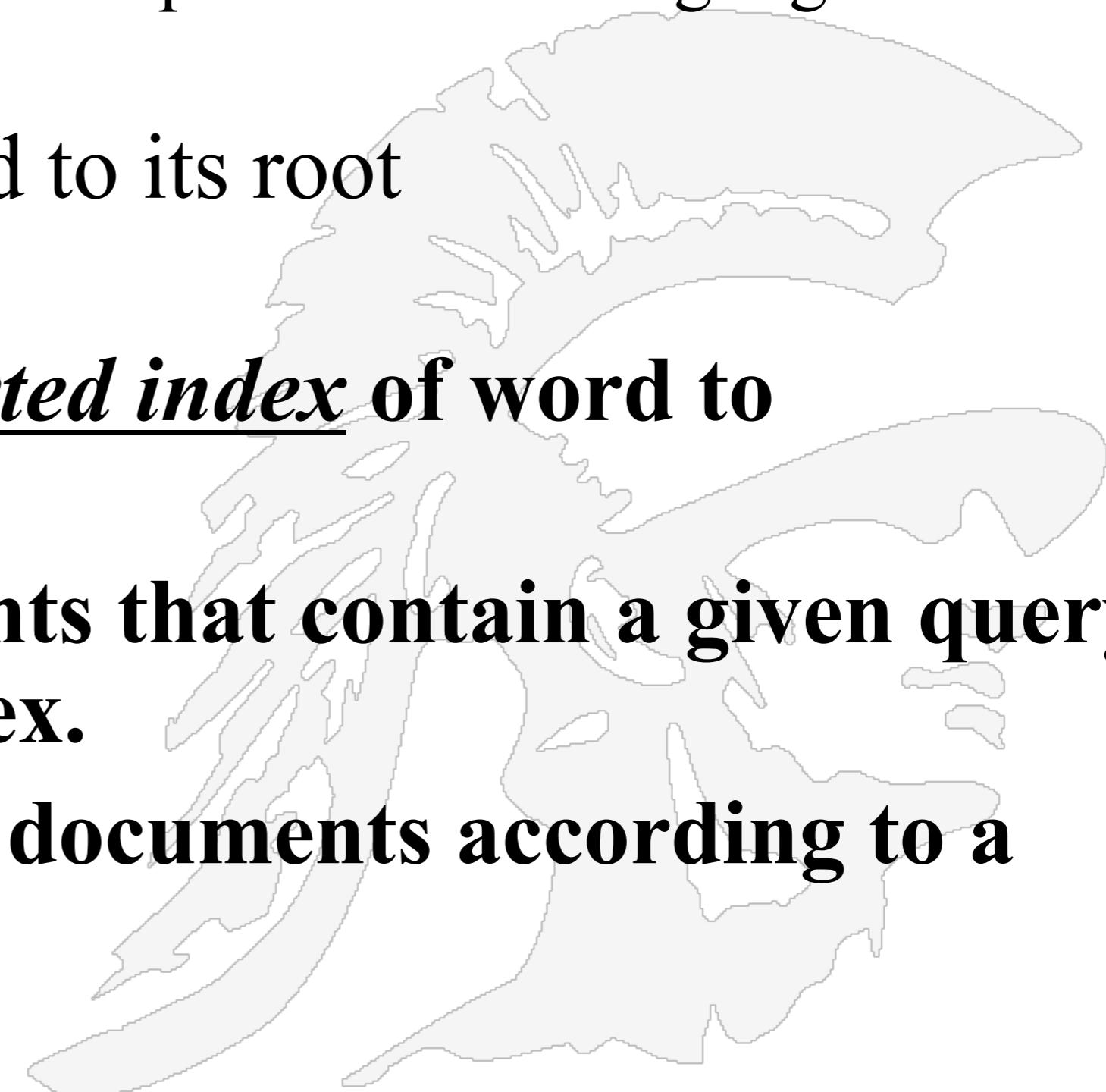
- User needs are part of the input
- User feedback is provided
- Queries initiate a search of the index and docs are retrieved
- A ranking function orders the results



Start with a text database; it is indexed; a user interface permits query operations which cause a search on the Index; matched documents are retrieved and ranked

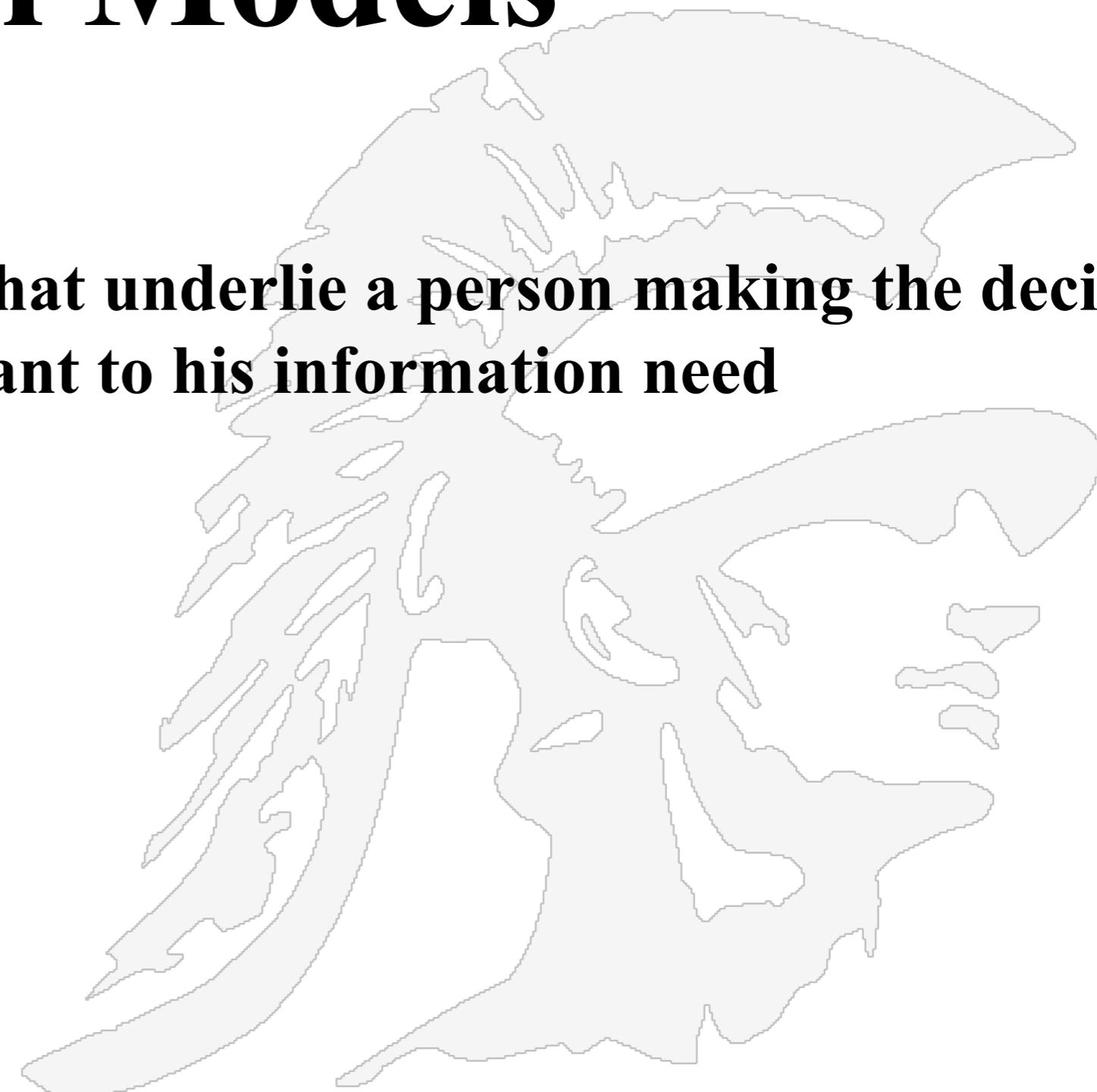
Defining Terms

- **Parsing** forms index words (**tokens**) and includes:
 - *Stopword* removal
 - See <http://www.ranks.nl/tools/stopwords.html> for google stopwords
 - *Stemming*: reducing a word to its root
 - More about this later
- **Indexing** constructs an *inverted index* of word to document pointers.
- **Searching** retrieves documents that contain a given query token from the inverted index.
- **Ranking** scores all retrieved documents according to a relevance metric.



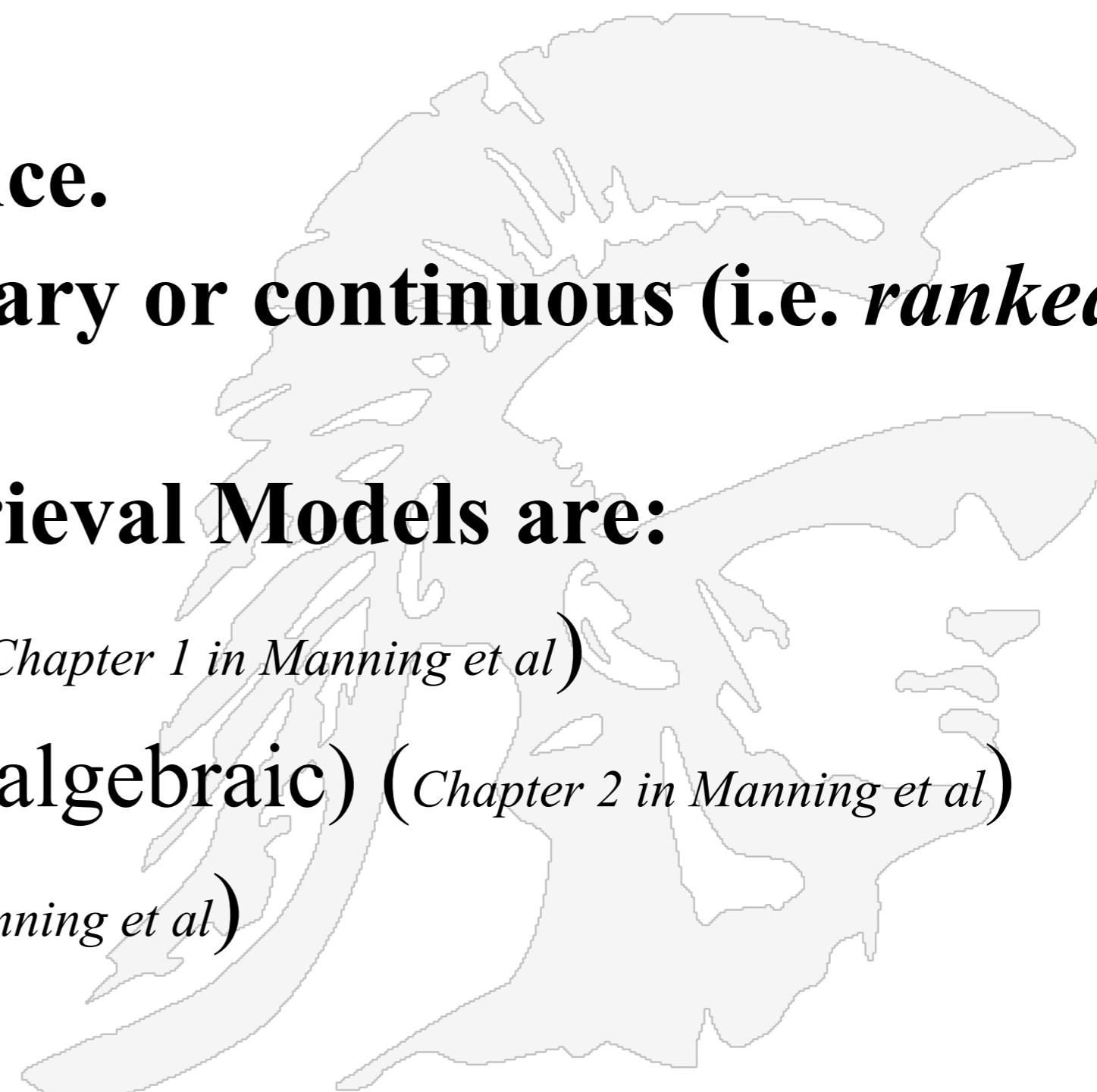
Boolean and Vector Space Retrieval Models

Primary goal: to formalize the processes that underlie a person making the decision
that a piece of text is relevant to his information need

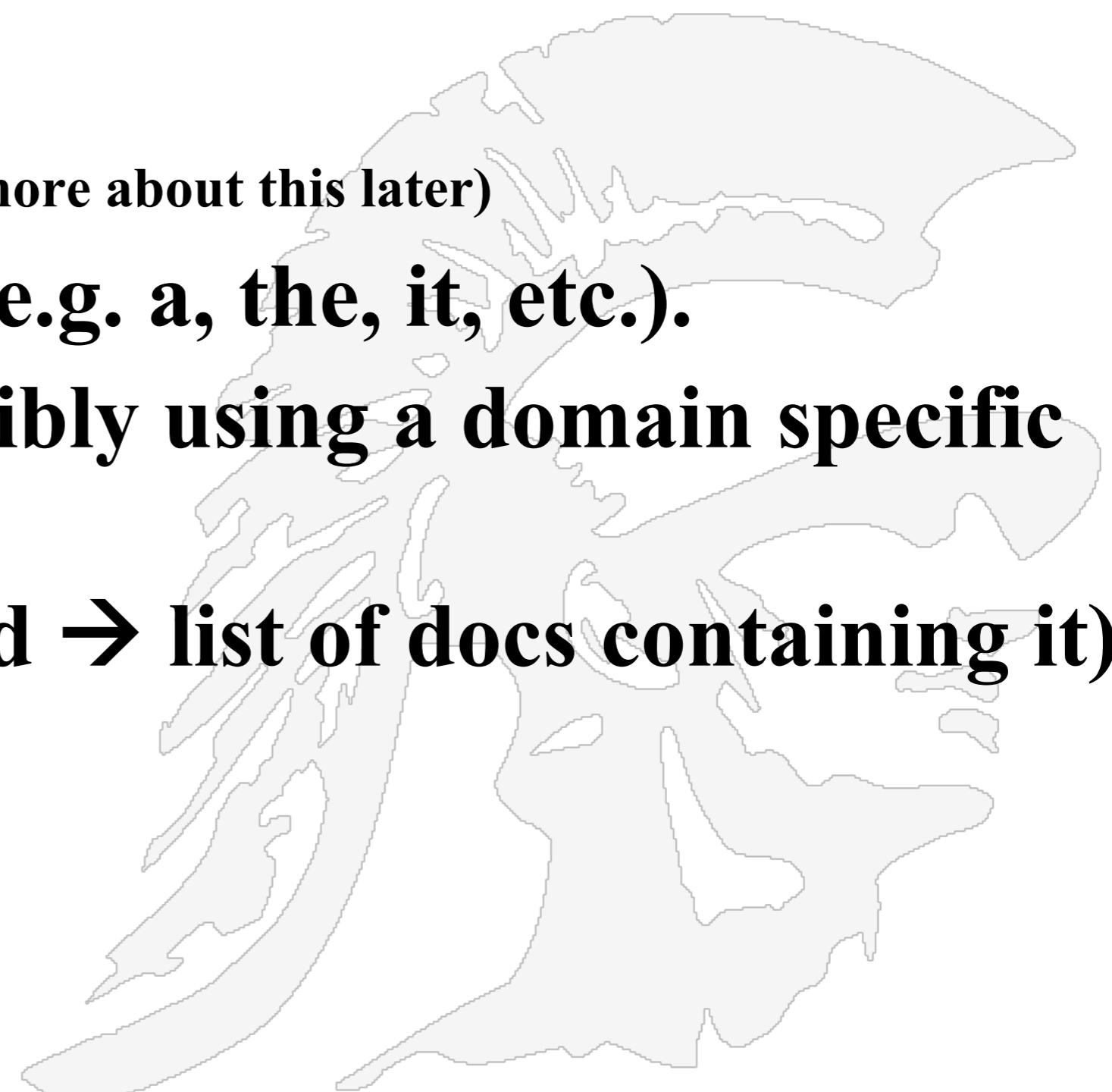


Retrieval Models

- **A retrieval model specifies the details of:**
 - Document representation
 - Query representation
 - Retrieval function
- **Determines a notion of relevance.**
- **Notion of relevance can be binary or continuous (i.e. *ranked retrieval*)**
- **Three major Information Retrieval Models are:**
 1. Boolean models (set theoretic) (*Chapter 1 in Manning et al*)
 2. Vector space models (statistical/algebraic) (*Chapter 2 in Manning et al*)
 3. Probabilistic models (*Chapter 11 in Manning et al*)



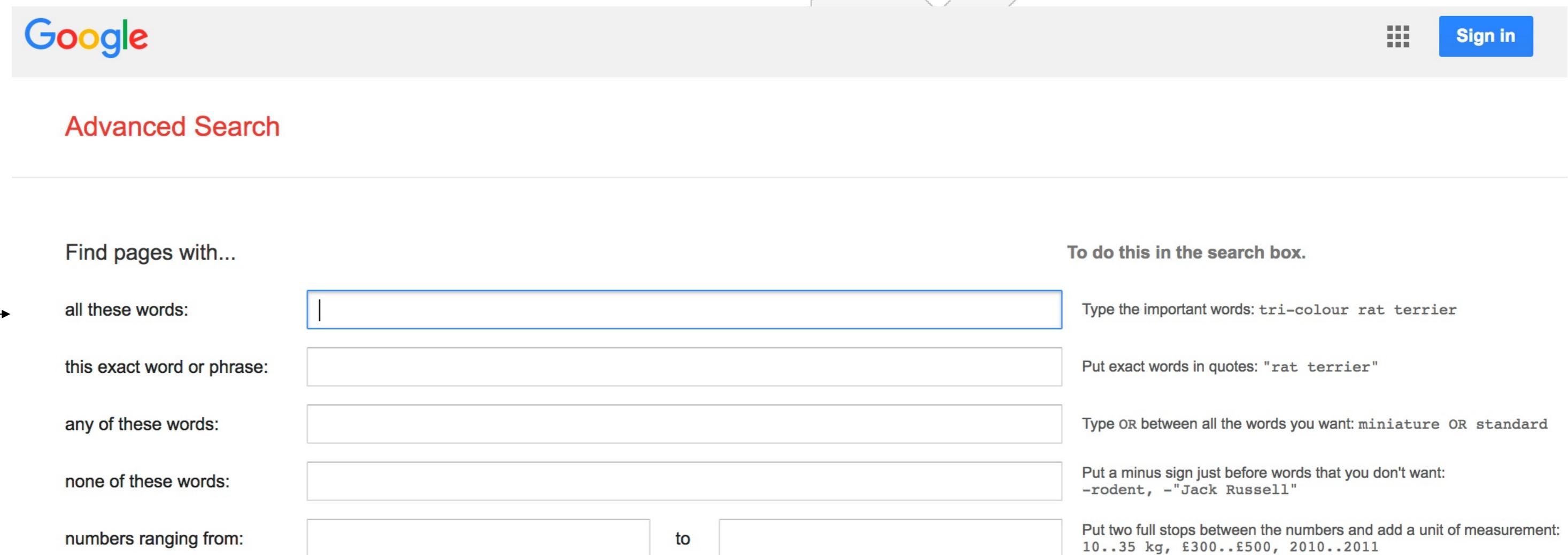
Common Pre-Processing Steps

- 
1. Strip unwanted characters/markup (e.g. HTML tags, punctuation, page numbers, etc.).
 2. Break into tokens (keywords) separating out whitespace.
 3. Stem tokens to “root” words
 - computational → comput (more about this later)
 4. Remove common stopwords (e.g. a, the, it, etc.).
 5. Detect common phrases (possibly using a domain specific dictionary).
 6. Build inverted index (keyword → list of docs containing it).

Boolean Model

- A document is represented as a **set** of keywords.
- Queries are Boolean expressions of keywords, connected by AND, OR, and NOT, including the use of brackets to indicate scope
- Here is a sample Boolean query with explicit AND, OR, NOT operators
 - [[Rio & Brazil] | [Hilo & Hawaii]] & hotel & !Hilton]

Google Advanced Search;
Note inclusion of AND, OR, NOT operators



The screenshot shows the Google Advanced Search interface. At the top, there's a navigation bar with the Google logo, a "Sign in" button, and a "Advanced Search" link. Below the navigation, there's a section titled "Find pages with..." with various search operators:

all these words: this exact word or phrase: any of these words: none of these words: numbers ranging from:	<input type="text"/> Type the important words: tri-colour rat terrier <input type="text"/> Put exact words in quotes: "rat terrier" <input type="text"/> Type OR between all the words you want: miniature OR standard <input type="text"/> Put a minus sign just before words that you don't want: -rodent, -"Jack Russell" <input type="text"/> to <input type="text"/> Put two full stops between the numbers and add a unit of measurement: 10..35 kg, £300..£500, 2010..2011
--	---

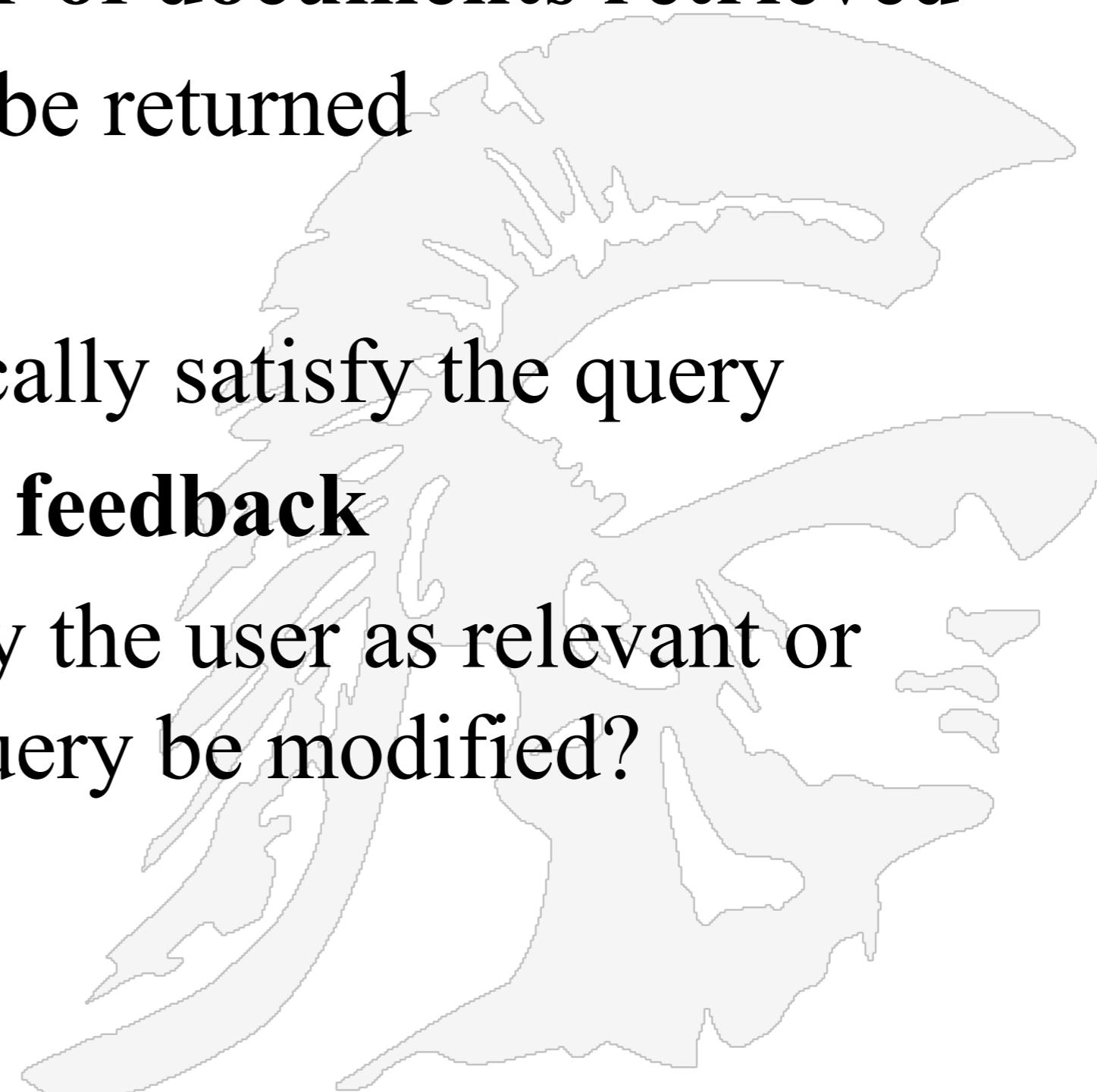
Boolean Retrieval Model

- Popular retrieval model because:
 - Easy to understand for simple queries.
 - Clean formalism.
- Boolean models can be extended to include ranking
- Reasonably efficient implementations possible for normal queries.



Boolean Models – Problems

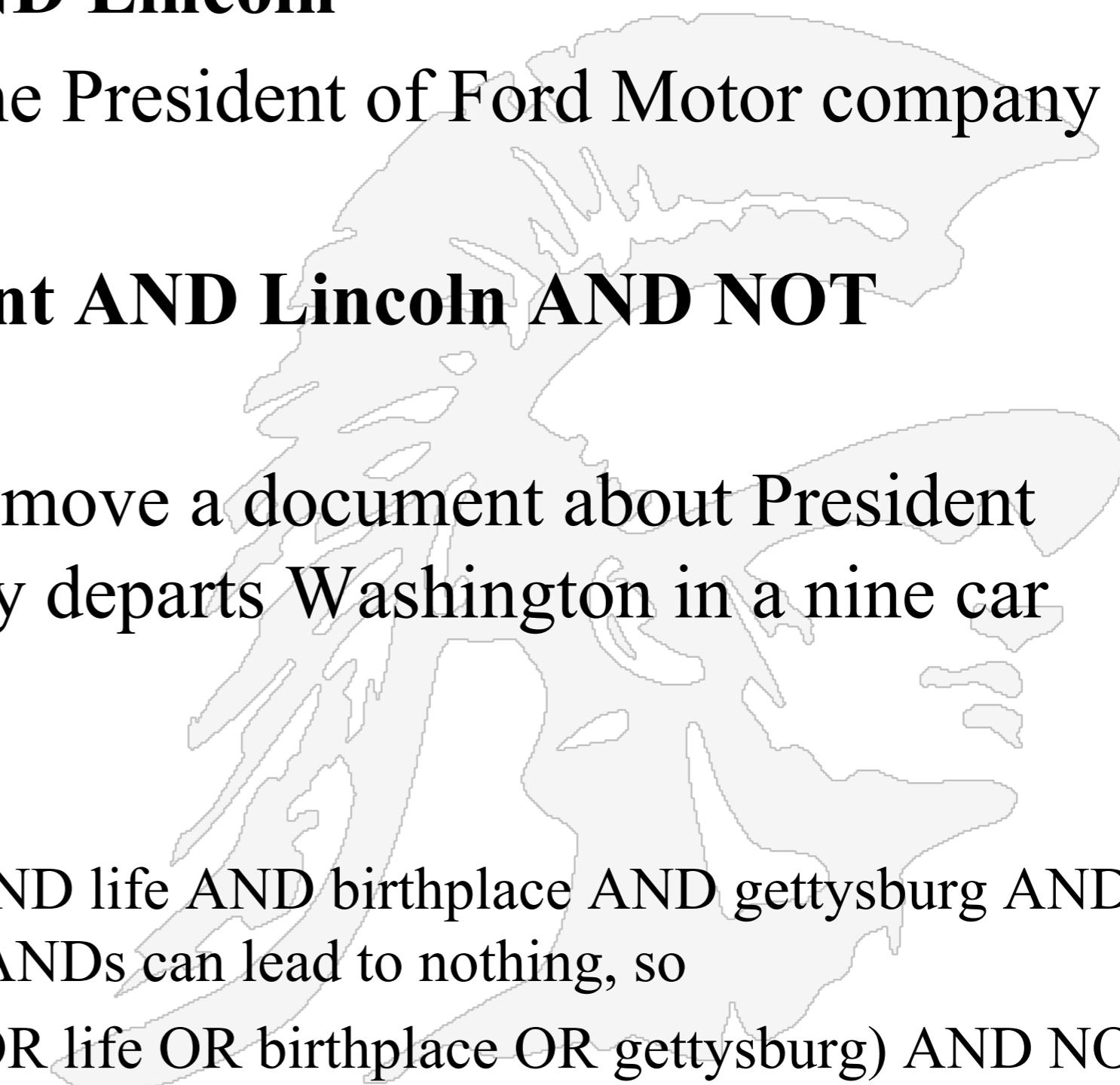
- **Very rigid:** AND means all; OR means any
- **Difficult to express complex user requests**
- **Difficult to control the number of documents retrieved**
 - *All* matched documents will be returned
- **Difficult to rank output**
 - *All* matched documents logically satisfy the query
- **Difficult to perform relevance feedback**
 - If a document is identified by the user as relevant or irrelevant, how should the query be modified?





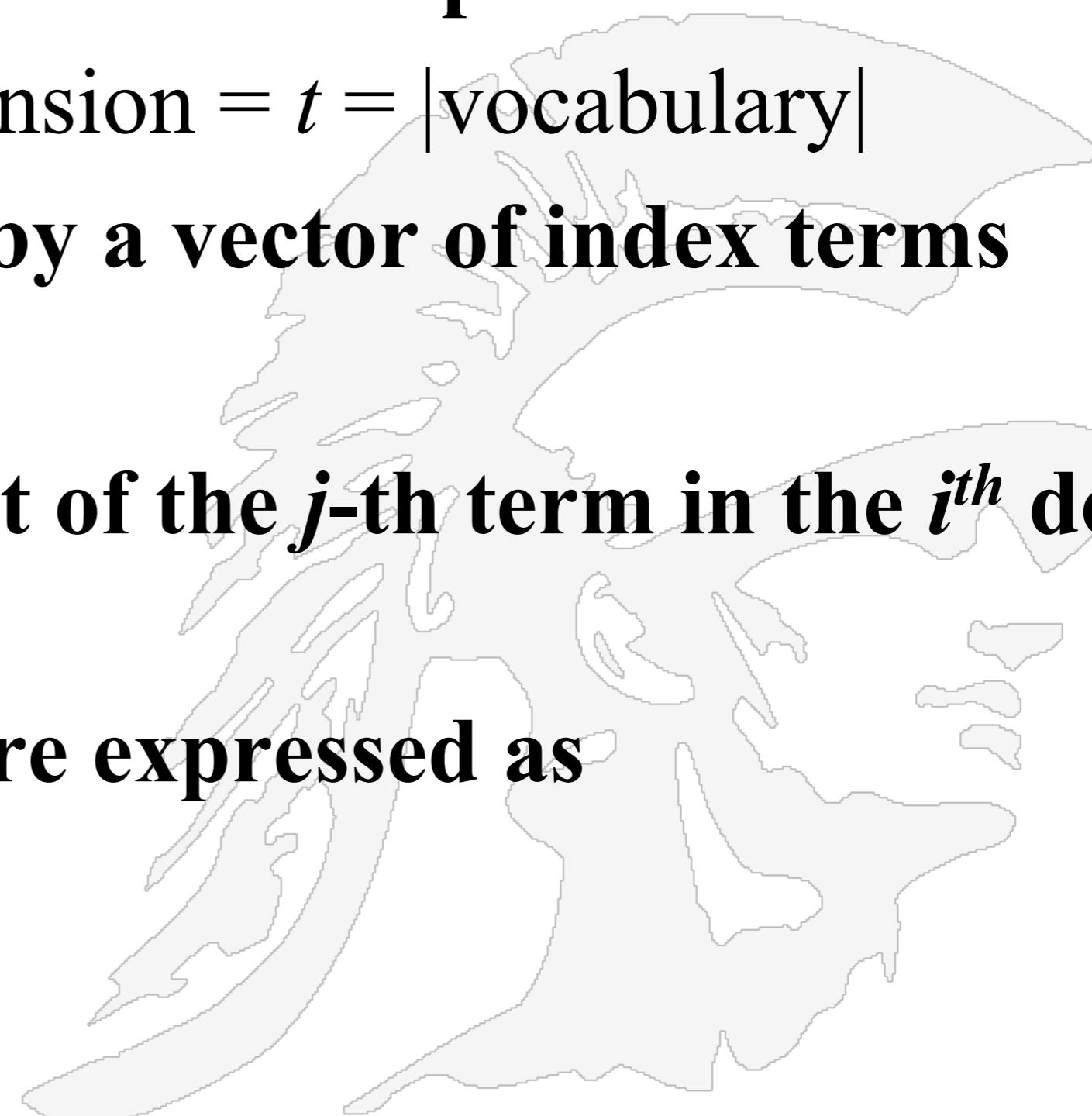
Problem Example For the Boolean Model

- **The simple query “Lincoln”**
 - Too many matches including Lincoln cars and places named Lincoln as well as Abraham Lincoln
- **More detailed query “President AND Lincoln”**
 - Returns documents that discuss the President of Ford Motor company that makes the Lincoln car
- **Even more detailed query “president AND Lincoln AND NOT (automobile OR car)”**
 - Better, but the use of NOT will remove a document about President Lincoln that says “Lincoln’ s body departs Washington in a nine car funeral train”
- **Perhaps try**
 - President AND lincoln AND biography AND life AND birthplace AND gettysburg AND NOT (automobile OR car), but too many ANDs can lead to nothing, so
 - President AND lincoln AND (biography OR life OR birthplace OR gettysburg) AND NOT (automobile OR car)



The Vector-Space Model

- Assume t distinct terms remain after preprocessing; call them **index terms** or the **vocabulary**
- These “orthogonal” terms form a **vector space**
size of the vocabulary = Dimension = $t = |\text{vocabulary}|$
- A document D_i is represented by a **vector of index terms**
$$D_i = (d_{i1}, d_{i2}, \dots, d_{it})$$
- Where d_{ij} represents the **weight** of the j -th term in the i^{th} doc
 - *but how is the weight computed?*
- Both documents and queries are expressed as **t -dimensional vectors**



Graphic Representation Example

Example:

$$D_1 = 2T_1 + 3T_2 + 5T_3$$

$$D_2 = 3T_1 + 7T_2 + T_3$$

$$Q = 0T_1 + 0T_2 + 2T_3$$

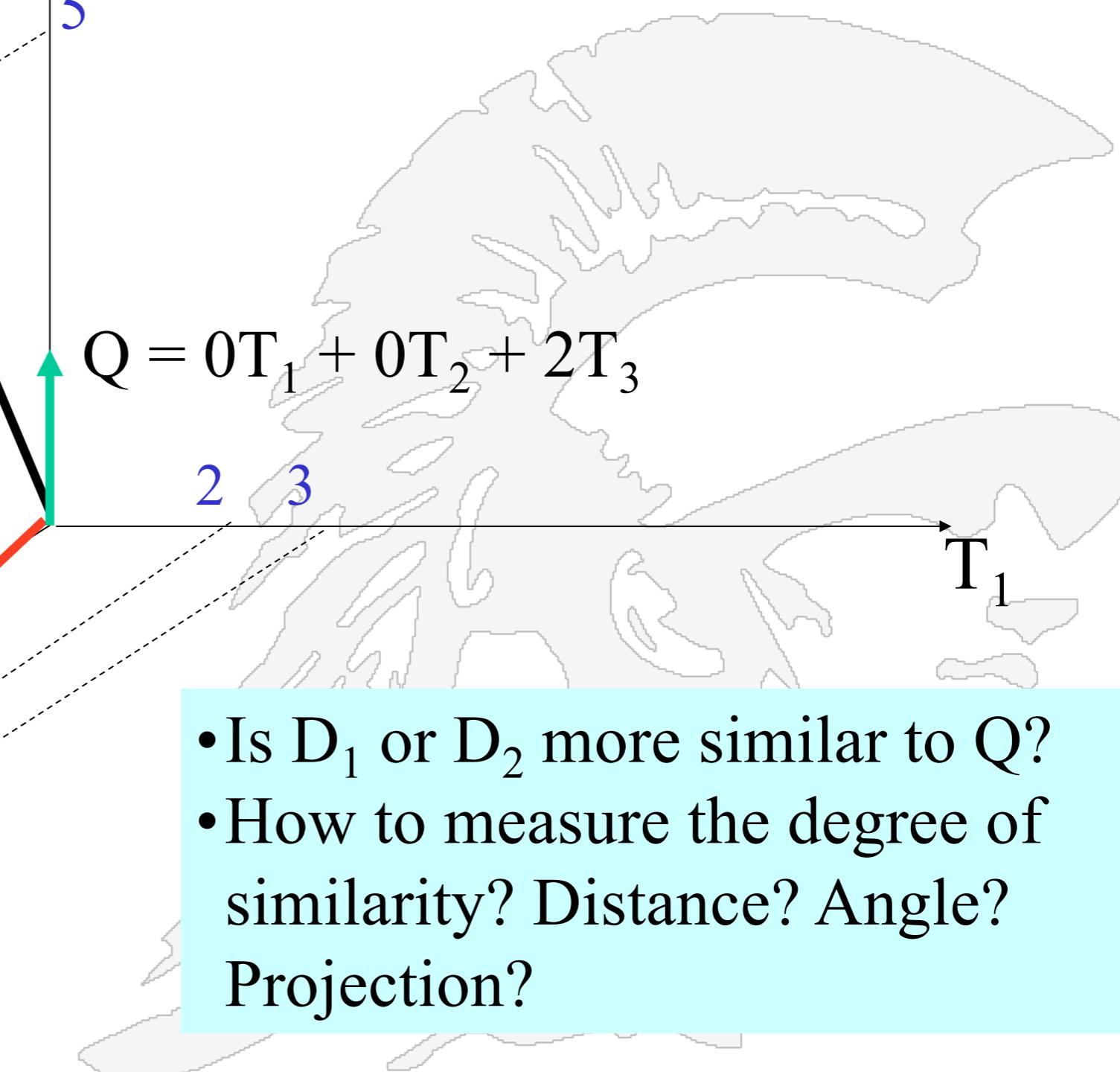
$$D_1 = 2T_1 + 3T_2 + 5T_3$$

$$D_2 = 3T_1 + 7T_2 + T_3$$

$$T_2$$

$$7$$

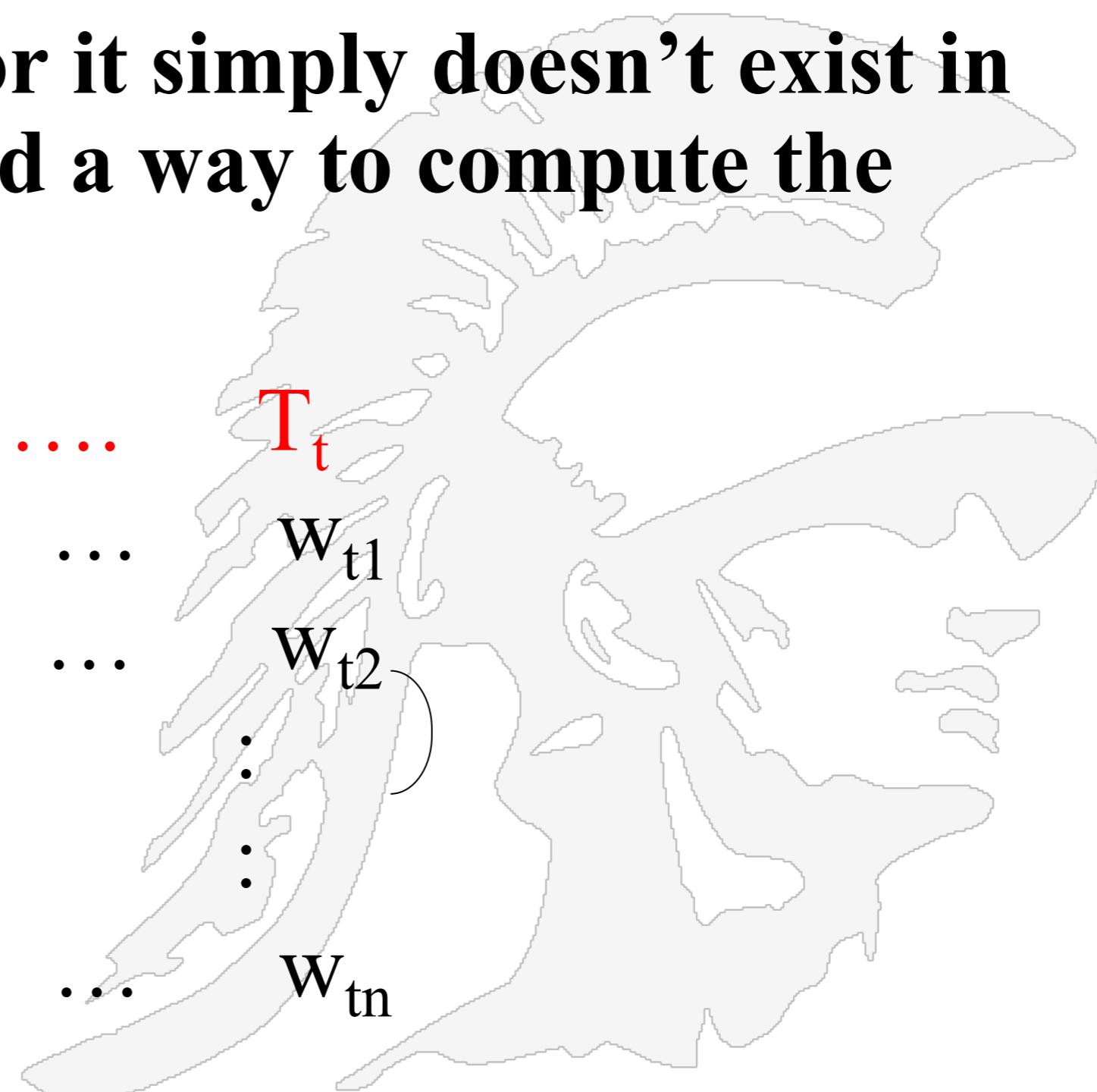
Vocabulary consists of 3 terms
with weights the coefficients
There are two documents, D_1 and
 D_2 ; there is one query, Q



- Is D_1 or D_2 more similar to Q ?
- How to measure the degree of similarity? Distance? Angle? Projection?

- A collection of n documents can be represented in the vector space model by a term-document matrix.
- An entry in the matrix corresponds to the “**weight**” of a **term in the document**; zero means the term has no significance in the document or it simply doesn’t exist in the document; but we still need a way to compute the weight

	T ₁	T ₂	T _t
D ₁	W ₁₁	W ₂₁	...	W _{t1}
D ₂	W ₁₂	W ₂₂	...	W _{t2}
:	:	:	...	⋮
⋮	⋮	⋮	...	⋮
D _n	W _{1n}	W _{2n}	...	W _{tn}



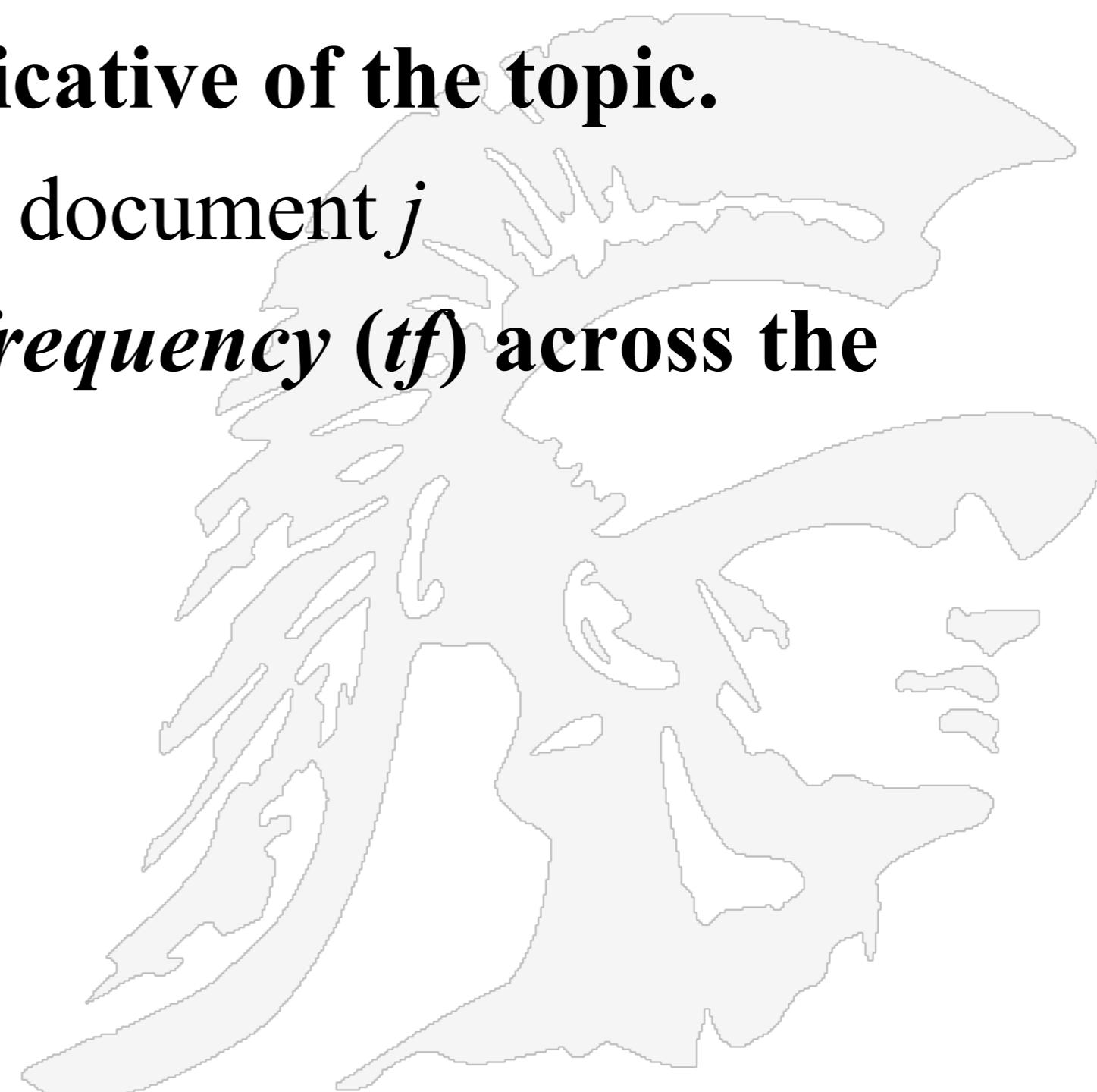
Term Weights: Term Frequency

- One way to compute the weight is to use the term's frequency in the document
- Assumption: the more frequent terms in a document are more important, i.e. more indicative of the topic.

f_{ij} = frequency of term i in document j

- May want to normalize *term frequency (tf)* across the entire corpus:

$$tf_{ij} = f_{ij} / \max\{f_{ij}\}$$



Term Weights:
Inverse Document Frequency

- Terms that appear in many *different* documents are *less* indicative of overall topic

df_i = document frequency of term i

= number of documents containing term i

of course df_i is always $\leq N$ (total number of documents)

idf_i = inverse document frequency of term i ,

= $\log_2 (N / df_i)$

(N : total number of documents)

- An indication of a term's *discrimination power*
- Log is used to dampen the effect relative to tf

An Example of Inverse Document Frequency

- | <i>term</i> | df_i | idf_i |
|-------------|-----------|-----------------------|
| Calpurnia | 1 | $\log(1,000,000/1)=6$ |
| animal | 100 | 4 |
| Sunday | 1,000 | 3 |
| fly | 10,000 | 2 |
| under | 100,000 | 1 |
| the | 1,000,000 | 0 |
-
- $idf_i = \log_{10}(N/df_i)$, $N = 1,000,000$
 - there is one idf value for each term t in a collection



TF.IDF Weighting

- A typical combined term importance indicator is ***tf-idf weighting*** (*note: it is often written with a hyphen, but the hyphen is NOT a minus sign; some people replace the hyphen with a dot*):

$$w_{ij} = tf_{ij} \cdot idf_i = (1 + \log tf_{ij}) * \log_2 (N/df_i)$$

- A term occurring frequently in the document but rarely in the rest of the collection is given high weight.
- Many other ways of determining term weights have been proposed.
- Experimentally, *tf.idf* has been found to work well
- Given a query q , then we score the query against a document d using the formula
- $Score(q, d) = \sum (tf.idf_{t,d})$ where t is in $q \cap d$

Computing TF.IDF -- An Example

Given a document containing 3 terms with given frequencies:

A(3), B(2), C(1)

Assume collection contains 10,000 documents and document frequencies of these 3 terms are:

A(50), B(1300), C(250)

Then:

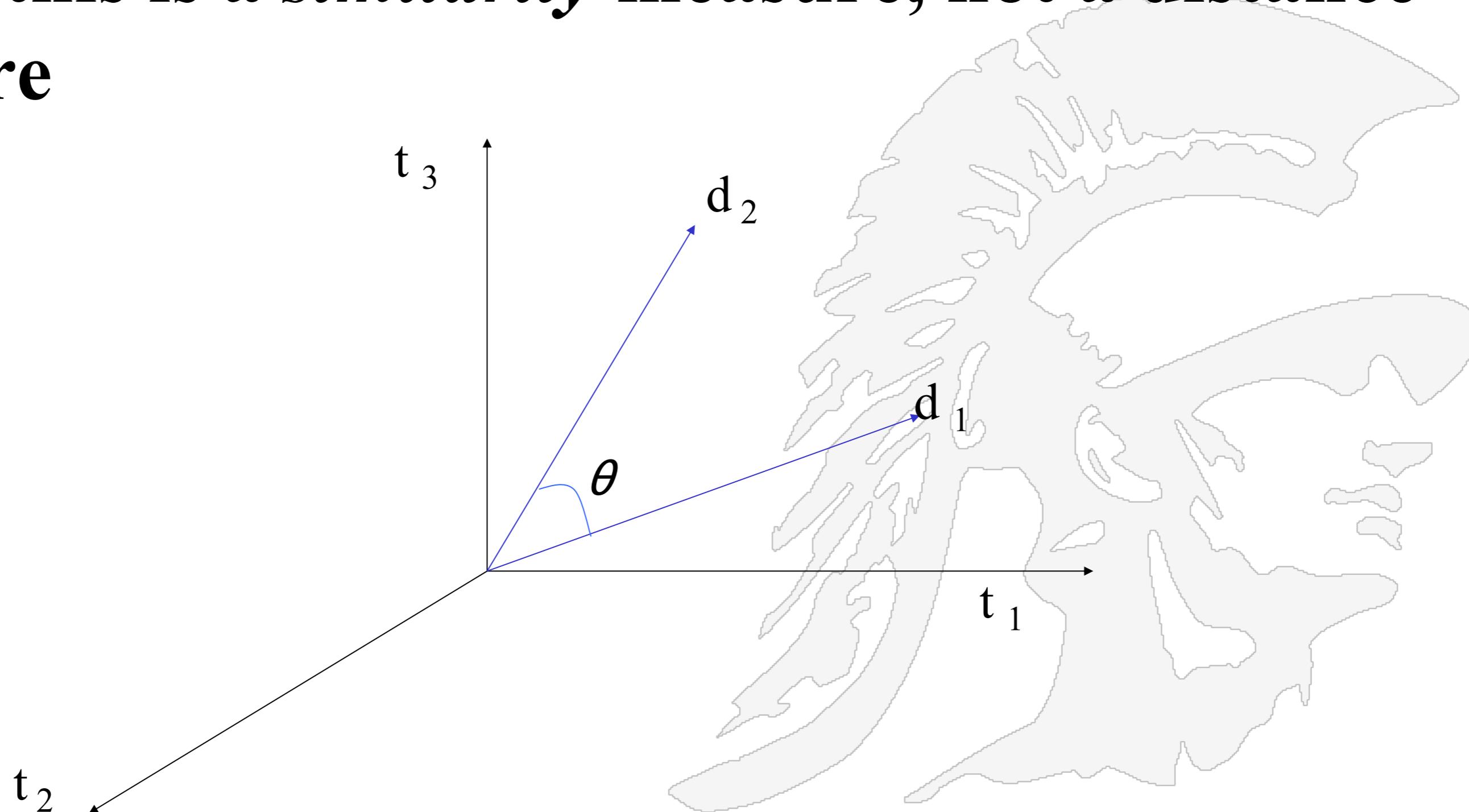
A: $tf = 3/3$; $idf = \log(10000/50) = 5.3$; $tf.idf = 5.3$

B: $tf = 2/3$; $idf = \log(10000/1300) = 2.0$; $tf.idf = 1.3$

C: $tf = 1/3$; $idf = \log(10000/250) = 3.7$; $tf.idf = 1.2$

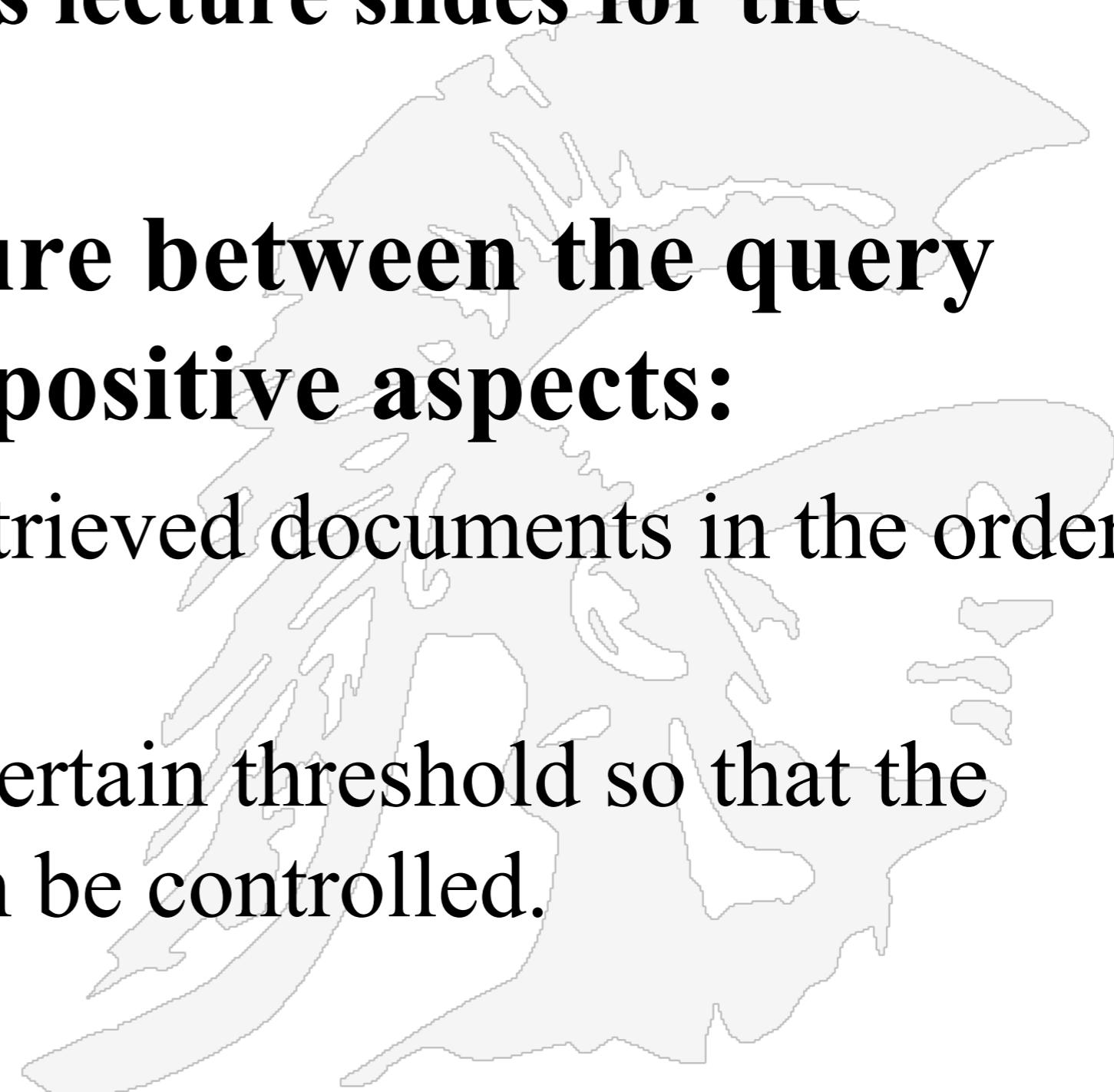
Cosine Similarity

- Distance between vectors d_1 and d_2 is *captured* by the cosine of the angle x between them.
- Note – this is a *similarity* measure, not a distance measure



Similarity Measure

- A **similarity measure** is a function that computes the *degree of similarity* between two vectors
 - Look back at the previous lecture slides for the definition of similarity
- Using a similarity measure between the query and each document has positive aspects:
 - It is possible to rank the retrieved documents in the order of presumed relevance.
 - It is possible to enforce a certain threshold so that the size of the retrieved set can be controlled.



Normalized Vectors

- A vector can be normalized (given a length of 1) by dividing each of its components by the vector's length
- This maps vectors onto the unit circle:
- Then, $|\vec{d}_j| = \sqrt{\sum_{i=1}^n w_{i,j}^2} = 1$
- Longer documents don't get more weight
- For normalized vectors, the cosine is simply the dot product:

$$\cos(\vec{d}_j, \vec{d}_k) = \vec{d}_j \cdot \vec{d}_k$$

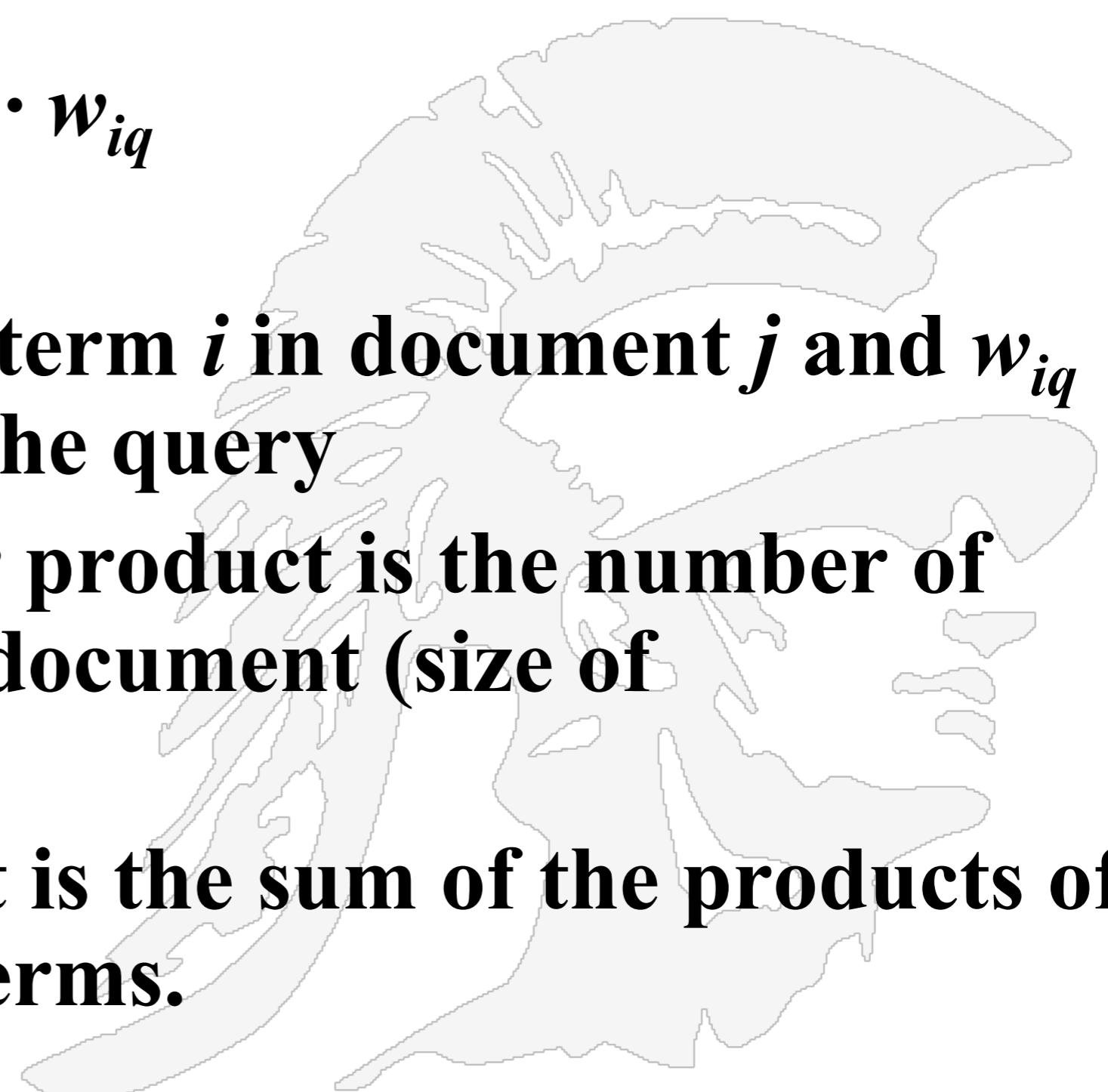
Similarity Measure for Document and Query

- Similarity between vectors for the document d_j and query q can be computed as the vector inner product:

$$\text{sim}(d_j, q) = d_j \cdot q = \sum_{i=1}^t w_{ij} \cdot w_{iq}$$

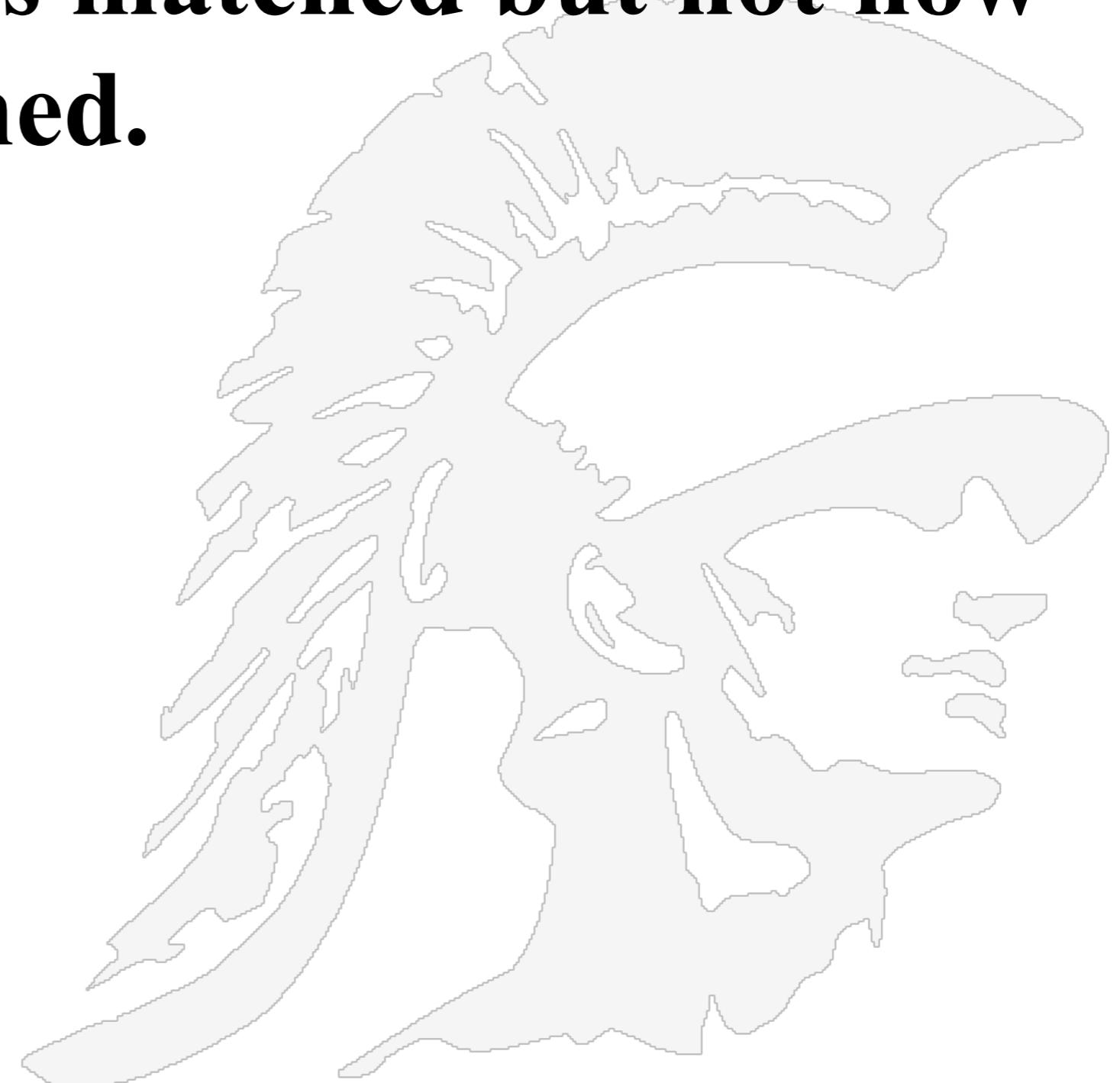
where w_{ij} is the weight of term i in document j and w_{iq} is the weight of term i in the query

- For binary vectors, the inner product is the number of matched query terms in the document (size of intersection).
- For weighted term vectors, it is the sum of the products of the weights of the matched terms.



Limitations of the Inner Product

- Favors long documents with a large number of unique terms.
- Measures how many terms matched but not how many terms are *not* matched.



Cosine Similarity Using Inner Product -- Examples

Binary:

- $D = [1, 1, 1, 0, 1, 1, 0]$ retrieval database architecture computer text management information
- $Q = [1, 0, 1, 0, 0, 1, 1]$ Size of vector = size of vocabulary = 7
0 means corresponding term not found in document or query

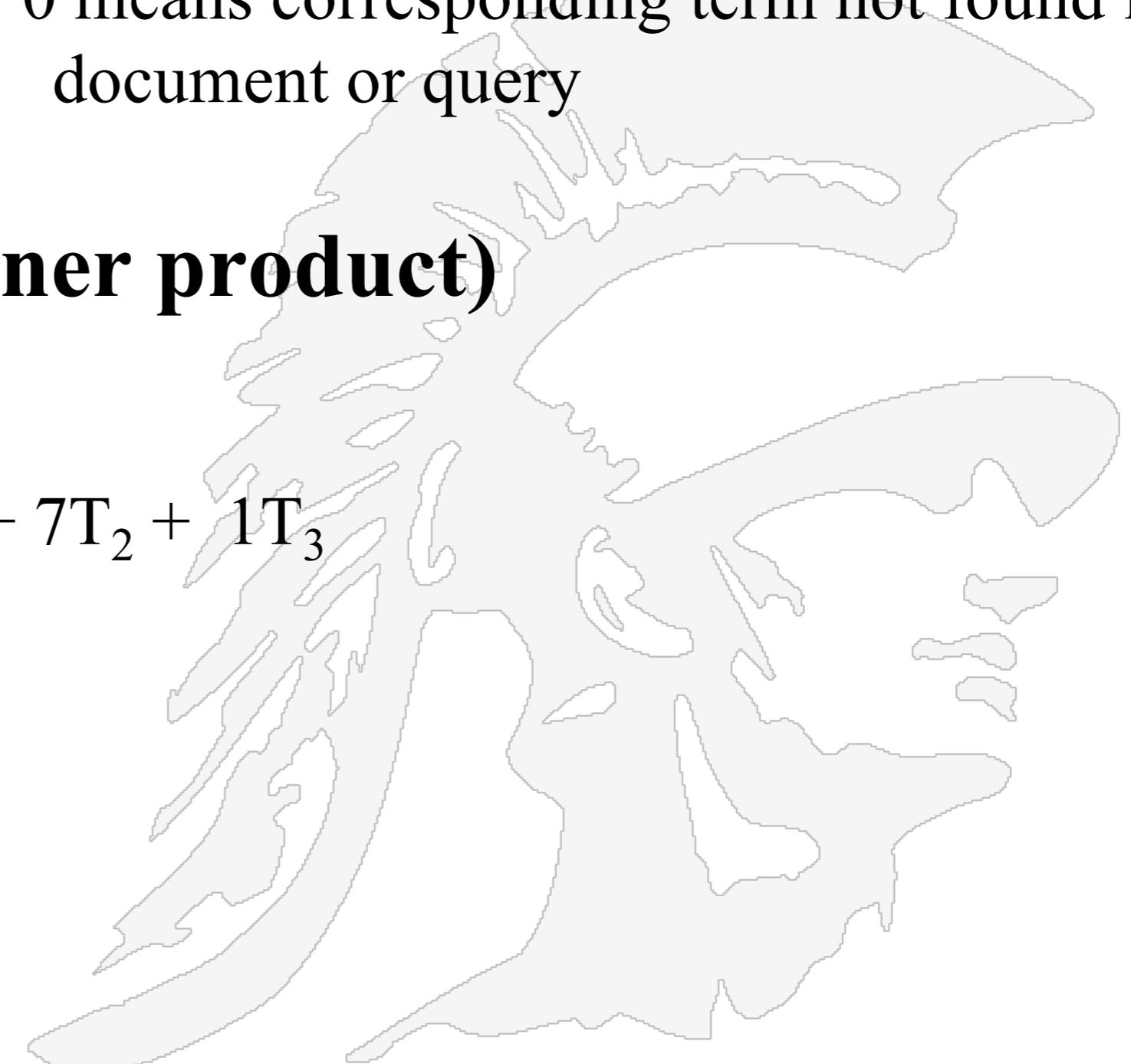
$$\text{similarity}(D, Q) = 3 \quad (\text{the inner product})$$

Weighted:

$$\begin{aligned} D_1 &= 2T_1 + 3T_2 + 5T_3 & D_2 &= 3T_1 + 7T_2 + 1T_3 \\ Q &= 0T_1 + 0T_2 + 2T_3 \end{aligned}$$

$$\text{sim}(D_1, Q) = 2*0 + 3*0 + 5*2 = 10$$

$$\text{sim}(D_2, Q) = 3*0 + 7*0 + 1*2 = 2$$

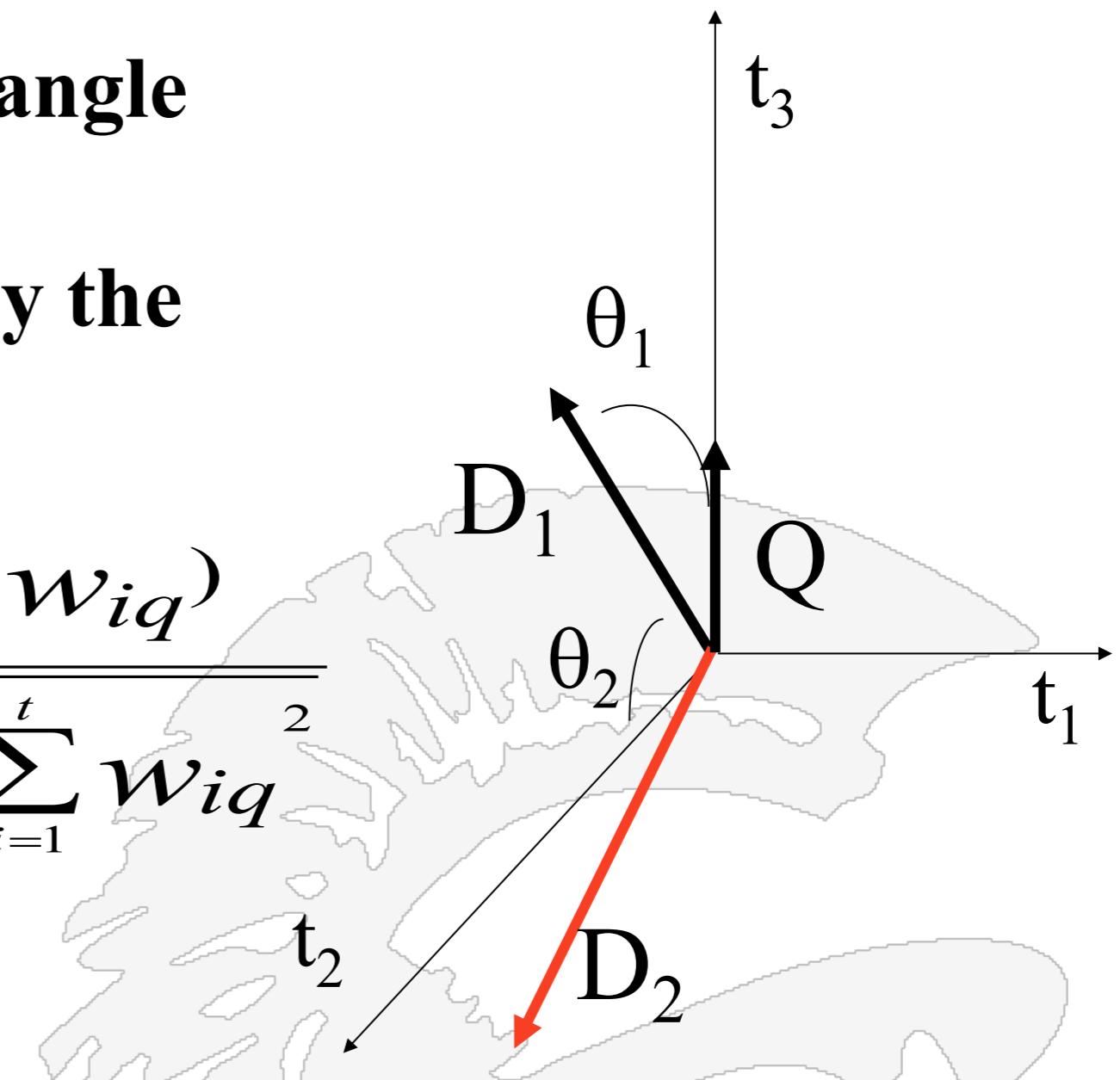




Cosine Similarity Measure Normalized

- Cosine similarity measures the cosine of the angle between two vectors
- We compute the inner product normalized by the vector lengths

$$\text{CosSim}(d_j, q) = \frac{\vec{d}_j \cdot \vec{q}}{|\vec{d}_j| \cdot |\vec{q}|} = \frac{\sum_{i=1}^t (w_{ij} \cdot w_{iq})}{\sqrt{\sum_{i=1}^t w_{ij}^2 \cdot \sum_{i=1}^t w_{iq}^2}}$$



$$D_1 = 2T_1 + 3T_2 + 5T_3$$

$$D_2 = 3T_1 + 7T_2 + 1T_3$$

$$Q = 0T_1 + 0T_2 + 2T_3$$

$$\text{CosSim}(D_1, Q) = 10 / \sqrt{(4+9+25)(0+0+4)} = 0.81$$

$$\text{CosSim}(D_2, Q) = 2 / \sqrt{(9+49+1)(0+0+4)} = 0.13$$

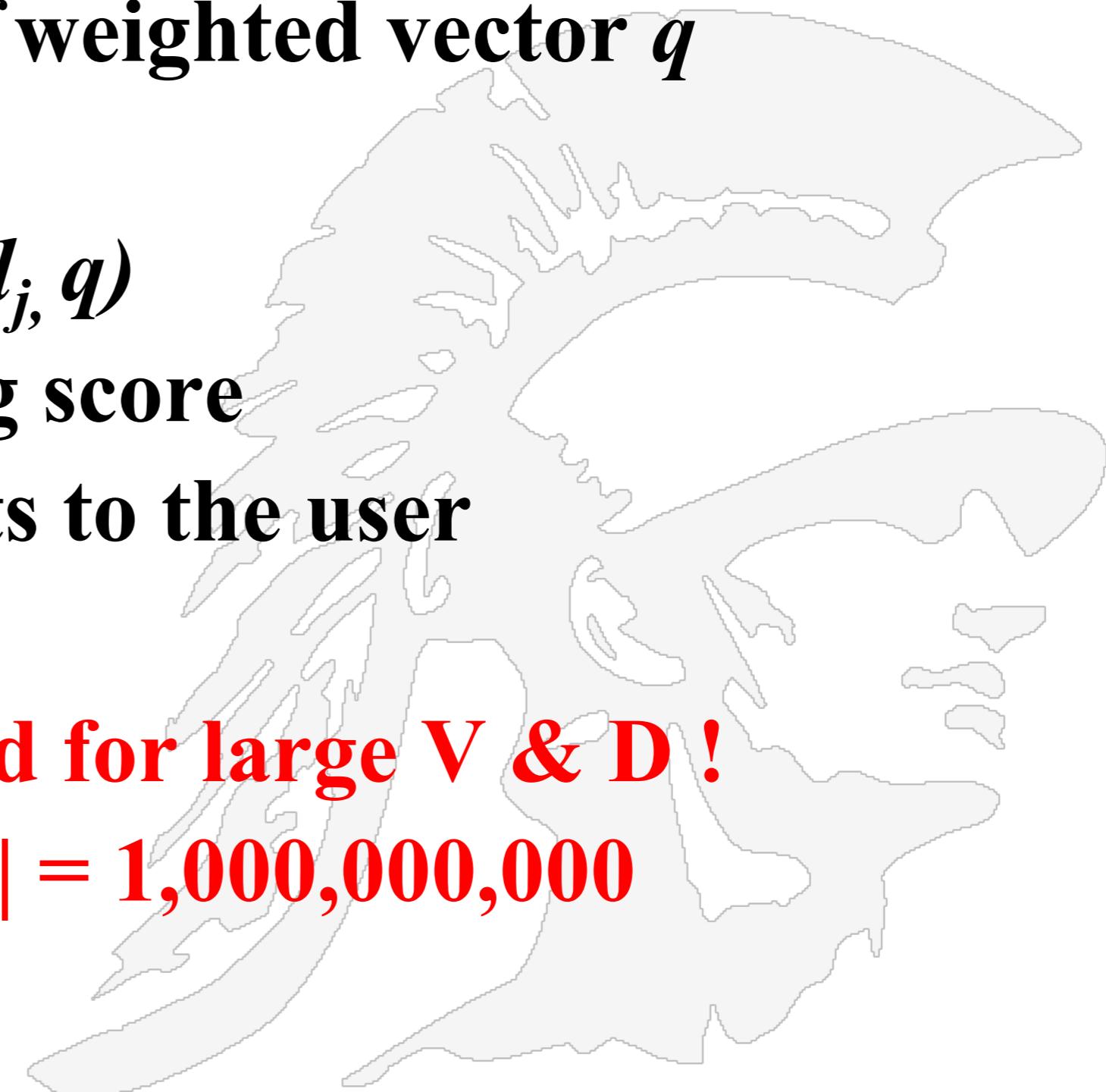
D₁ is 6 times better than D₂ using cosine similarity but only 5 times better using inner product.

Naïve Implementation

1. Convert all documents in collection D to $tf.idf$ weighted vectors, the j^{th} document denoted by d_j , for keywords in vocabulary V
2. Convert each query to a $tf.idf$ weighted vector q
3. For each d_j in D do
 - Compute score $s_j = \text{cosSim}(d_j, q)$
4. Sort documents by decreasing score
5. Present top ranked documents to the user

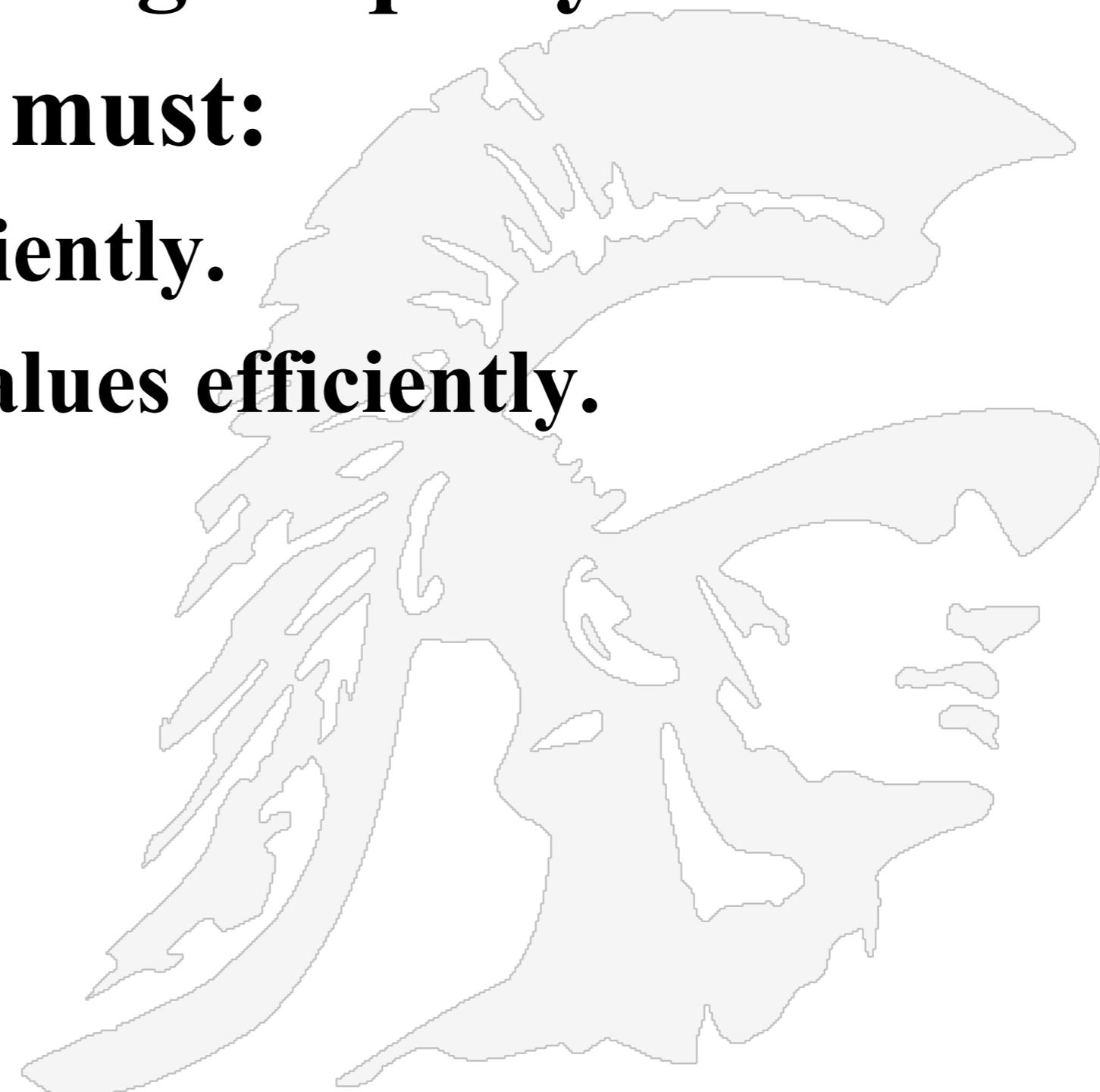
Time complexity: $O(|V| \cdot |D|)$ Bad for large V & D !

$|V| = 10,000$; $|D| = 100,000$; $|V| \cdot |D| = 1,000,000,000$



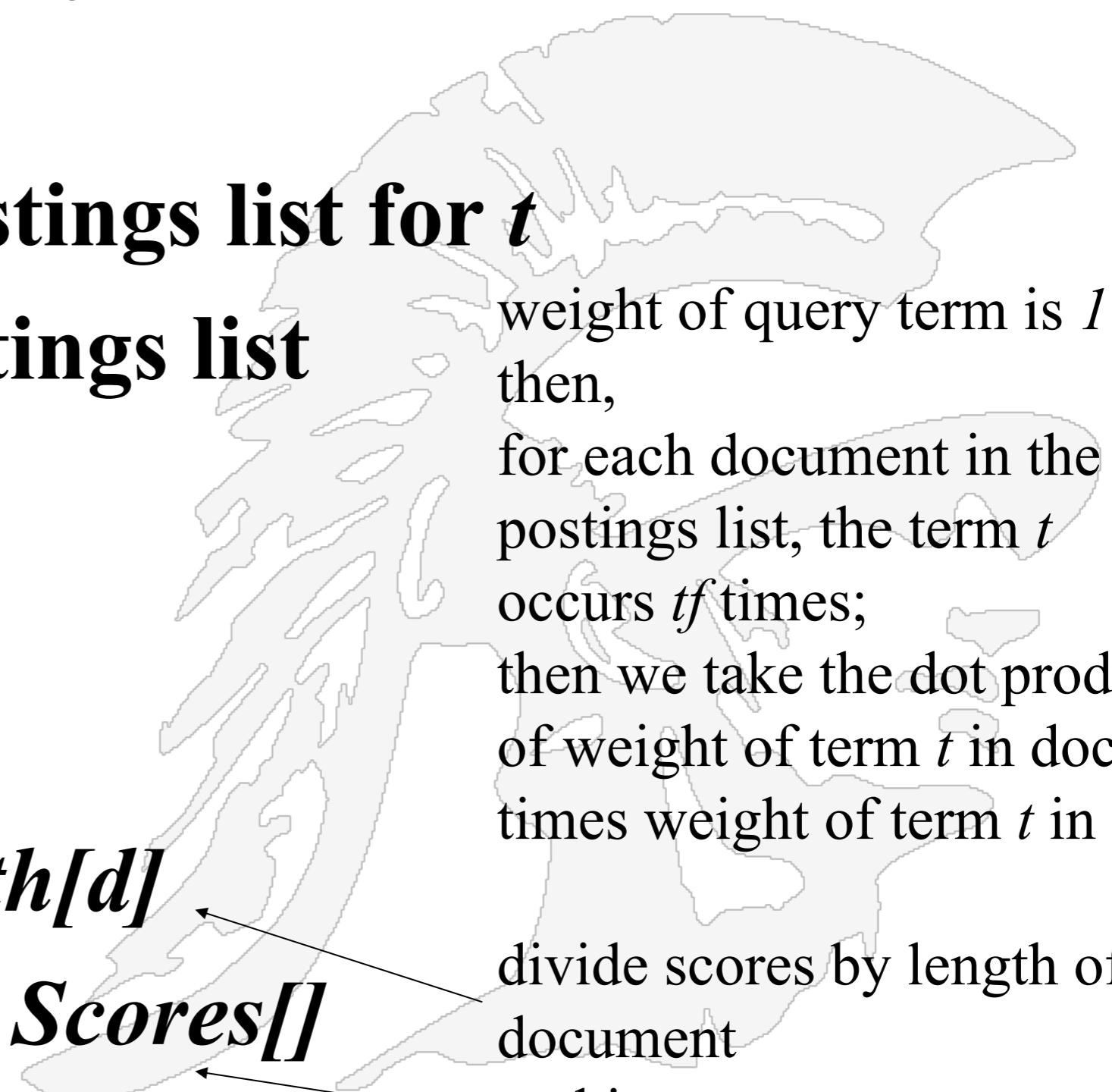
Efficient Cosine Ranking

- Ranking consists of computing the k docs in the corpus “nearest” to the query $\Rightarrow k$ largest query-doc cosines.
- To do efficient ranking one must:
 - Compute a single cosine efficiently.
 - Choose the k largest cosine values efficiently.



Computing Cosine Scores

cosineScore(q)

1. float *Scores[N]* = 0; //Scores array for all documents
 2. float *Length[N]* //lengths of all documents
 3. for each query term t
 4. do calculate $w_{t,q}$ and fetch postings list for t
 5. for each pair $(d, tf_{t,d})$ in postings list
 6. do $Scores[d] += w_{t,d} \times w_{t,q}$
 7. Read the array Length
 8. for each d do
 9. $Scores[d] = Scores[d]/Length[d]$
 10. return Top K components of *Scores[]*
- 
- weight of query term is 1;
then,
for each document in the
postings list, the term t
occurs tf times;
then we take the dot product
of weight of term t in document
times weight of term t in query;
- divide scores by length of each
document
- ranking

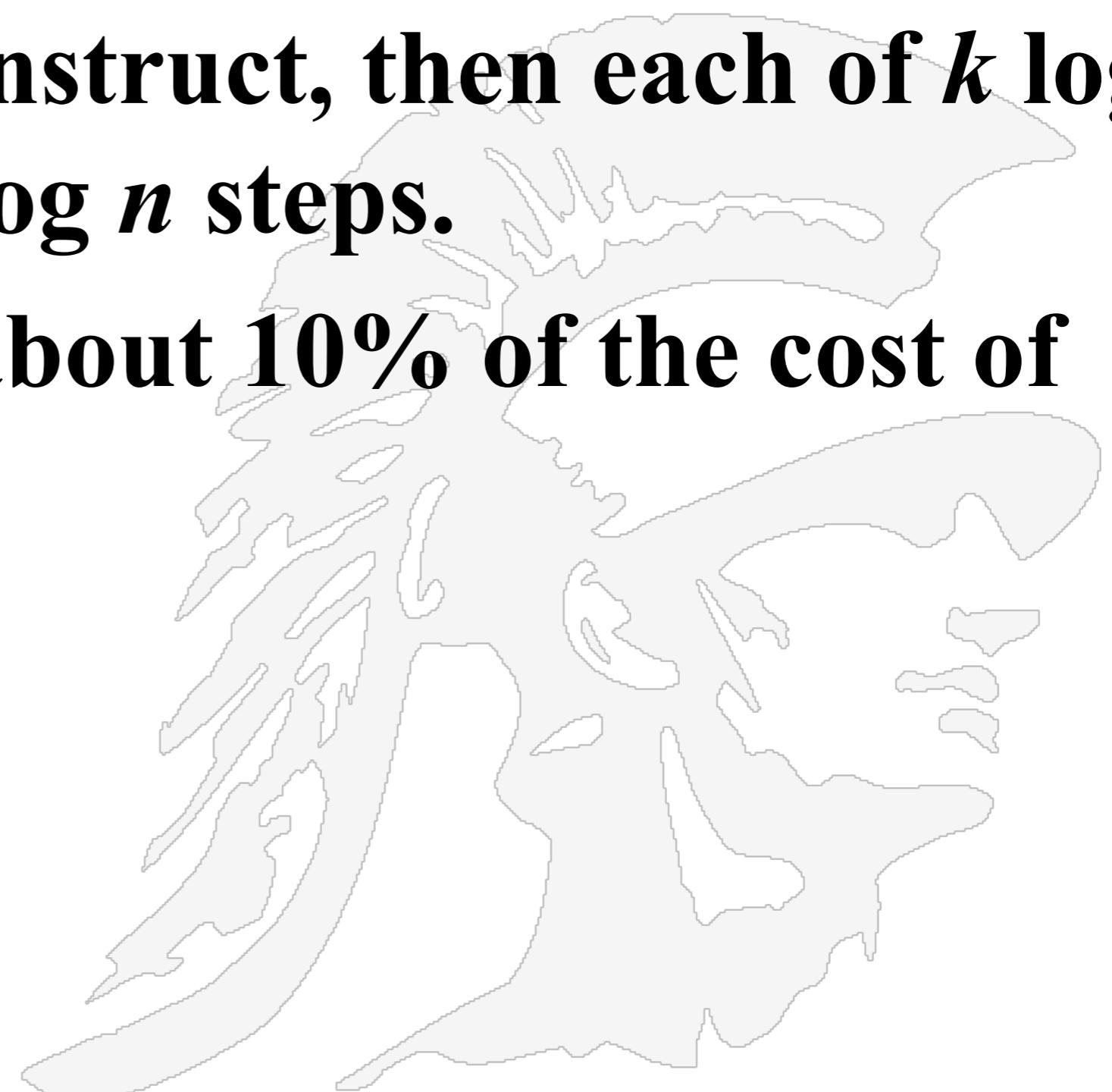
in practice we only work with a subset of the documents

Summary of Algorithm for Vector Space Ranking

- Represent the query as a weighted $tf.idf$ vector
- represent each document as a weighted $tf.idf$ vector
- compute the cosine similarity score for the query vector and each document vector that contains the query term
- Rank documents with respect to the query by score
- Return the top k (e.g. $k=10$) to the user

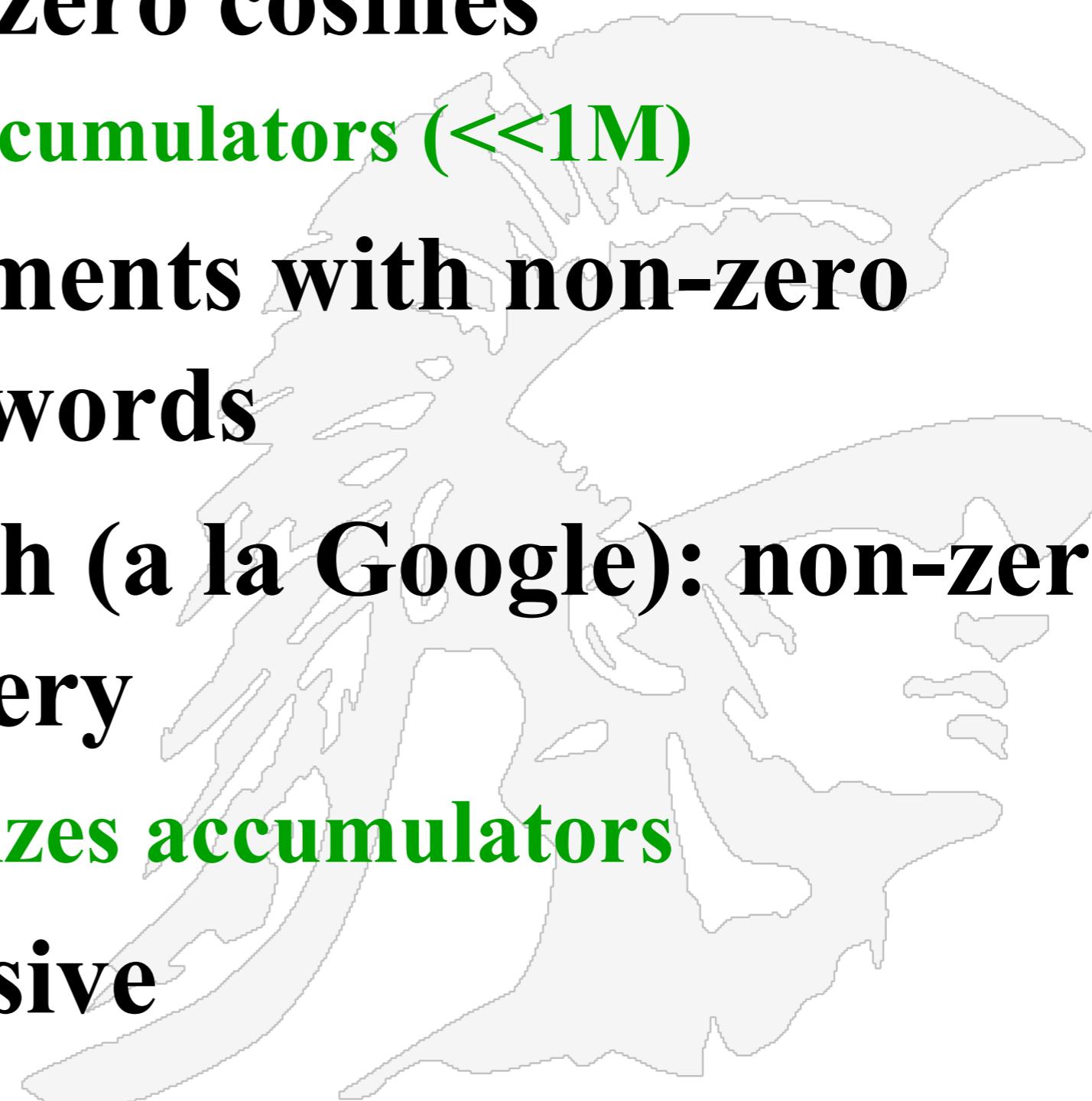
Use Heap for Selecting Top k

- Binary tree in which each node's value > values of children
- Takes $2n$ operations to construct, then each of $k \log n$ “winners” read off in $2\log n$ steps.
- For $n=1M, k=100$, this is about 10% of the cost of sorting.



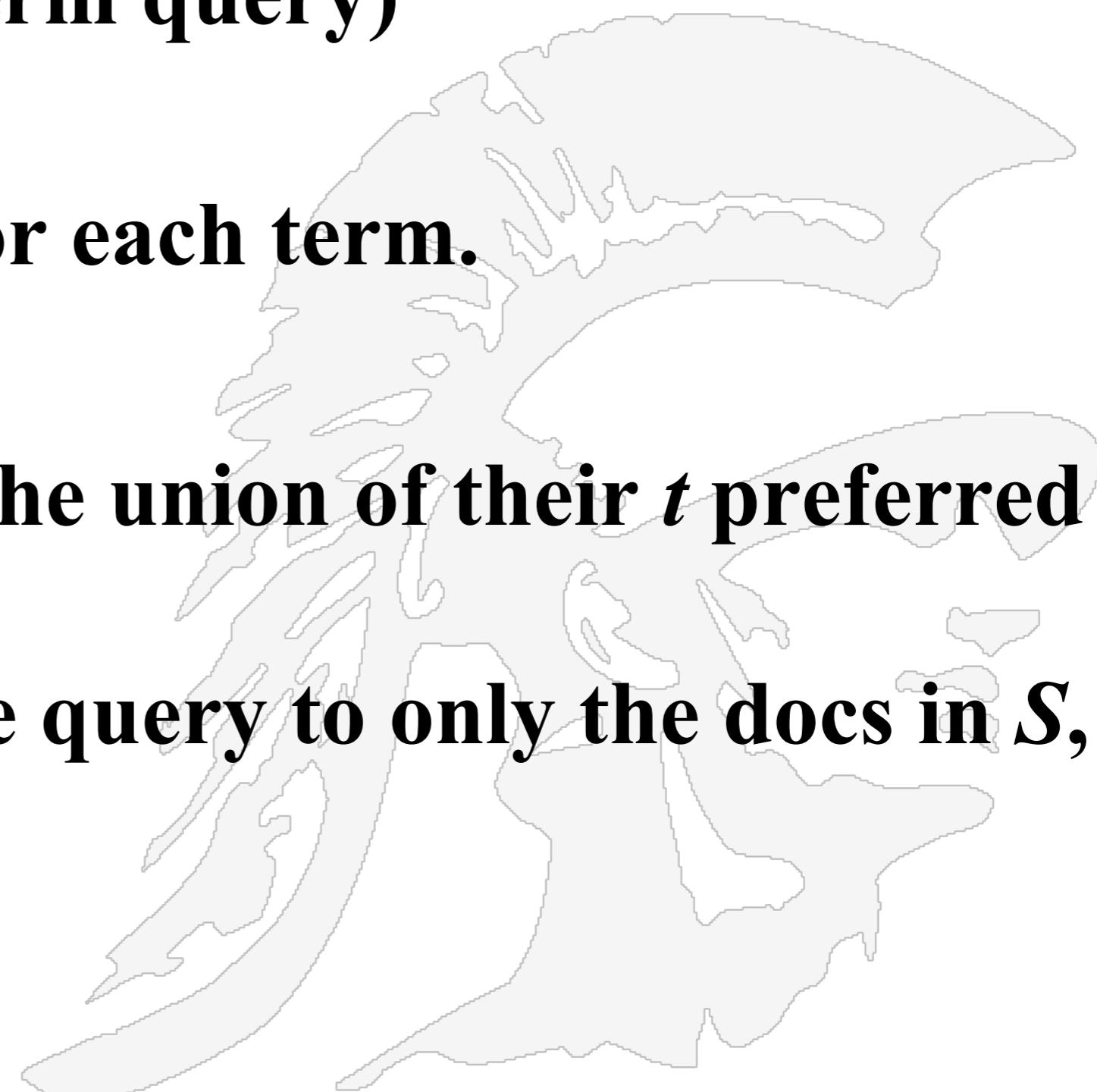
Bottleneck

- Still need to first compute cosines from query to each of n docs → several seconds for $n = 1M$
- Can select from only non-zero cosines
 - Need union of postings lists accumulators (<<1M)
- Can further limit to documents with non-zero cosines on rare (high idf) words
- Enforce conjunctive search (a la Google): non-zero cosines on *all* words in query
 - Need min of postings lists sizes accumulators
- But still potentially expensive



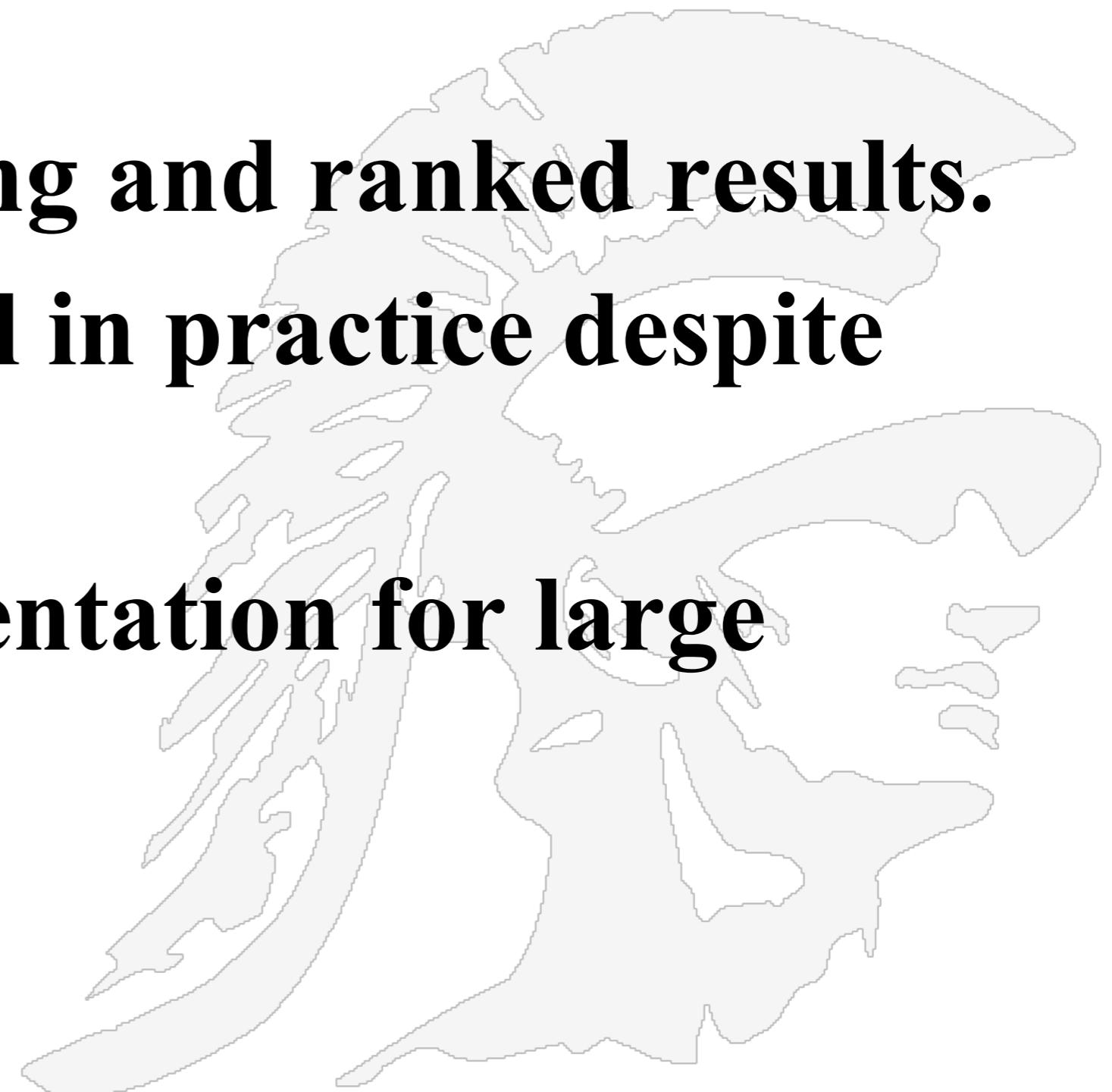
A Pre-Processing Strategy

- **Preprocess**: Pre-compute, for each term, its k nearest docs.
 - (Treat each term as a 1-term query)
 - lots of preprocessing.
 - Result: “preferred list” for each term.
- **Search**:
 - For a t -term query, take the union of their t preferred lists – call this set S .
 - Compute cosines from the query to only the docs in S , and choose top k .



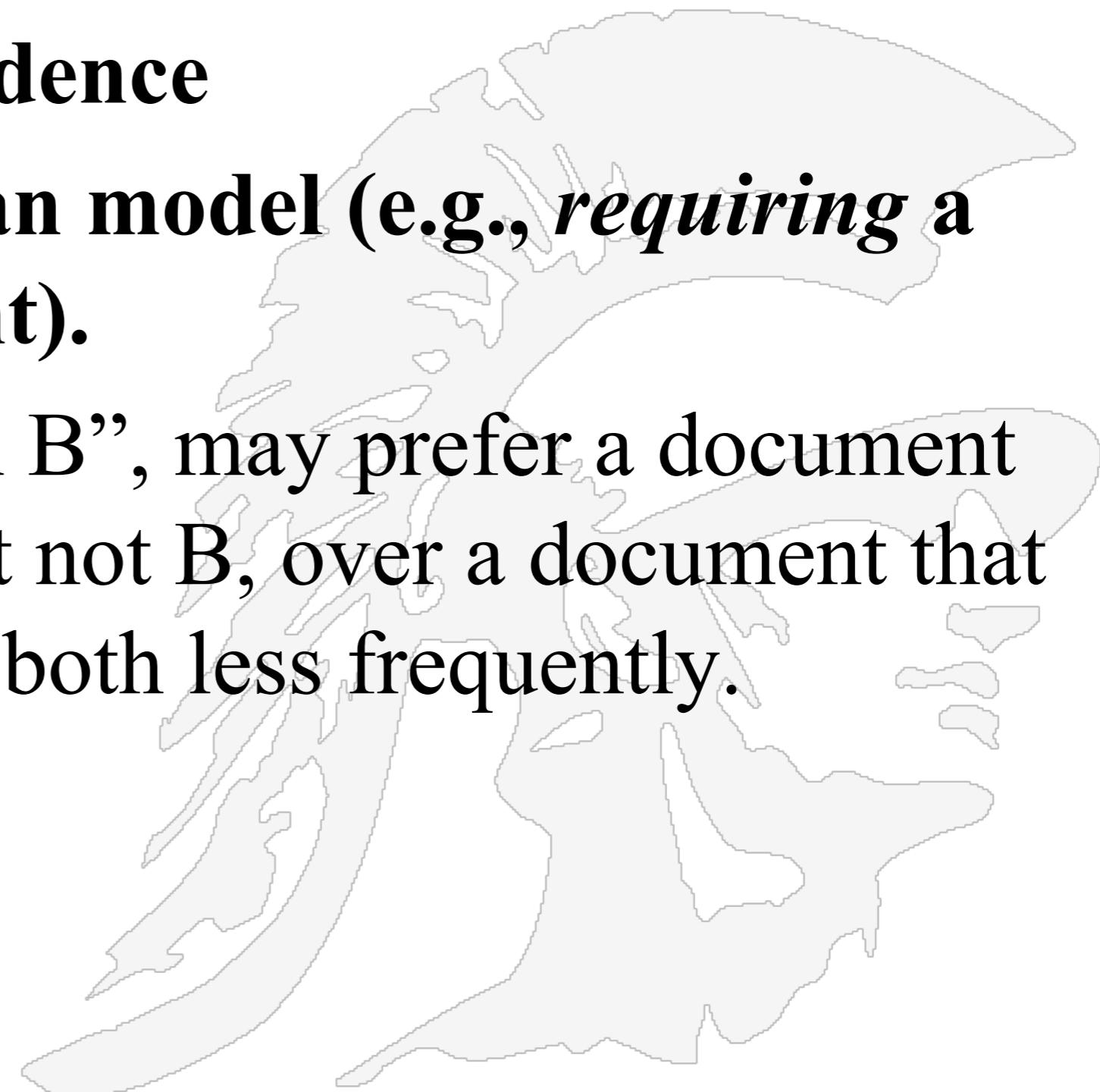
Comments on Vector Space Model

- Simple, mathematically based approach.
- Considers both local (tf) and global (idf) word occurrence frequencies.
- Provides partial matching and ranked results.
- Tends to work quite well in practice despite obvious weaknesses.
- Allows efficient implementation for large document collections.



Problems with Vector Space Model

- Missing semantic information (e.g. word sense).
- Missing syntactic information (e.g. phrase structure, word order, proximity information).
- Assumption of term independence
- Lacks the control of a Boolean model (e.g., *requiring a term to appear in a document*).
 - Given a two-term query “A B”, may prefer a document containing A frequently but not B, over a document that contains both A and B, but both less frequently.

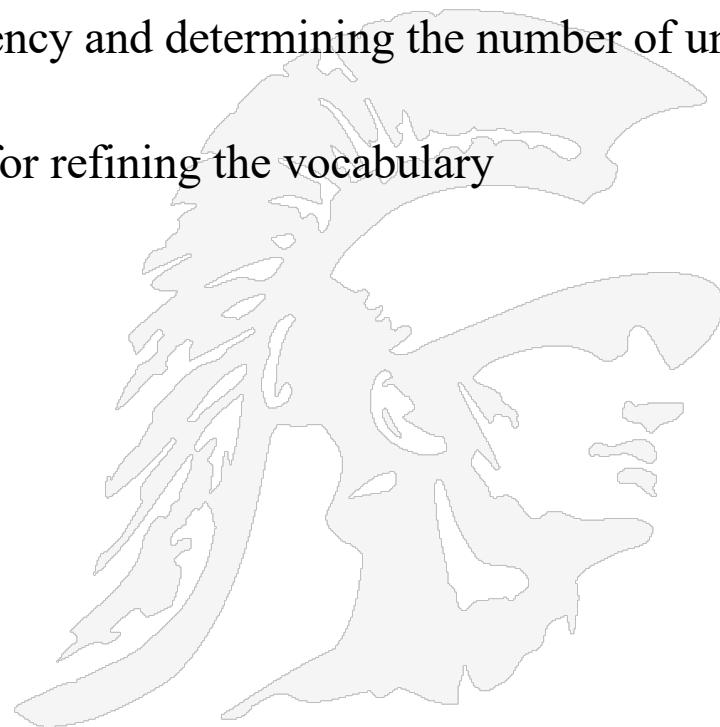


Lexicon & Text Normalization



Overview of these Slides

- These slides are focused on the issue of how a search engine creates its database of documents.
- In particular how the search engine determines what words go into its inverted index, the **lexicon**
- we begin with a discussion of word frequency and determining the number of unique words (the **vocabulary**)
- then we look at some specific operations for refining the vocabulary
 - tokenization
 - stop words
 - capitalization
 - case folding
 - synonyms (thesaurus)
 - similar sounding words
 - stemming



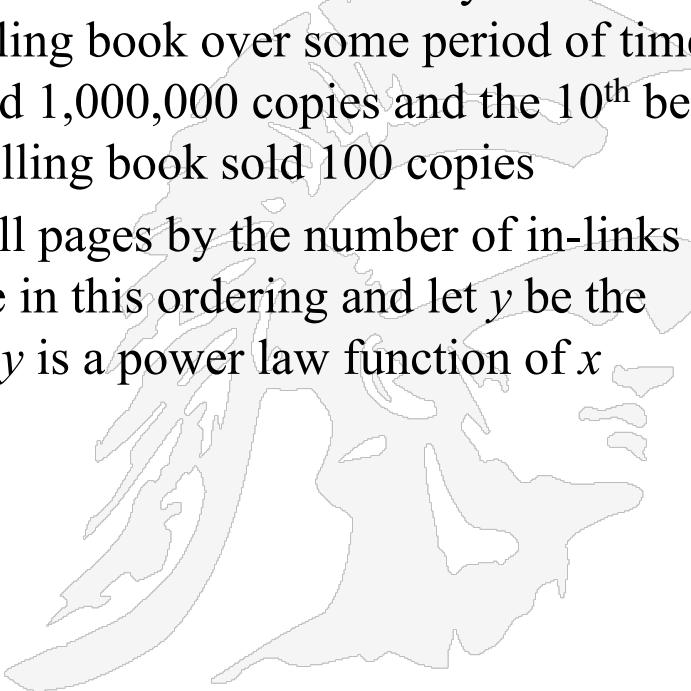
Power Laws and Zipf's Law

- An equation of the form $y = kx^c$ is called a power law
 - k and c are constants
- Zipf's law is a power law with $c = -1$
 - George Zipf noticed the rule while examining word frequencies in text
 - His rule states that the frequency of any word is inversely proportional to its rank in the frequency table
 - Thus the most frequent word will occur approximately twice as often as the second most frequent word, three times as often as the third most frequent word, etc
- On a log-log plot, power laws give a straight line with slope c .

$$\log(y) = \log(kx^c) = \log k + c \log(x)$$

Some Power Law Examples

- ***Population of U.S. states:*** if we order states in the US by population and let y be the population of the x -th most populous state, then x and y obey Zipf's law approximately.
- ***Book sales at Amazon.com:*** let x represent the rank of books by sales and let y be the number of sales of the x^{th} best selling book over some period of time. According to Amazon the best seller sold 1,000,000 copies and the 10th best sold 10,000 copies, and the 100th best selling book sold 100 copies
- ***Node degrees in the web graph:*** order all pages by the number of in-links to that page. Let x be the position of a page in this ordering and let y be the number of in-links to the x^{th} page. Then y is a power law function of x

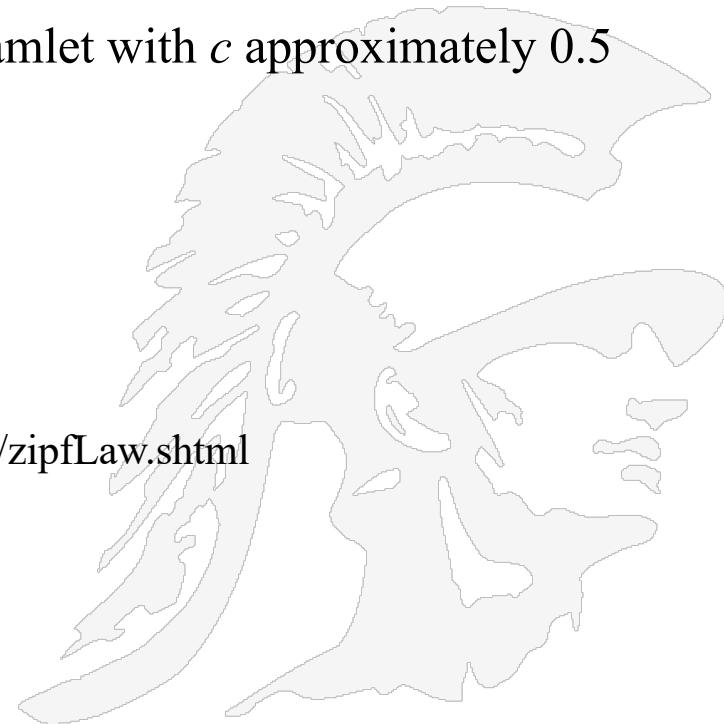


Other Power Law Examples

- Frequency of accesses to web pages
 - For example: the access counts on Wikipedia pages
- Words in the English language
 - for instance, in Shakespeare's play Hamlet with c approximately 0.5
- Sizes of settlements
- Income distributions among individuals
- Size of earthquakes
- Notes in musical performances

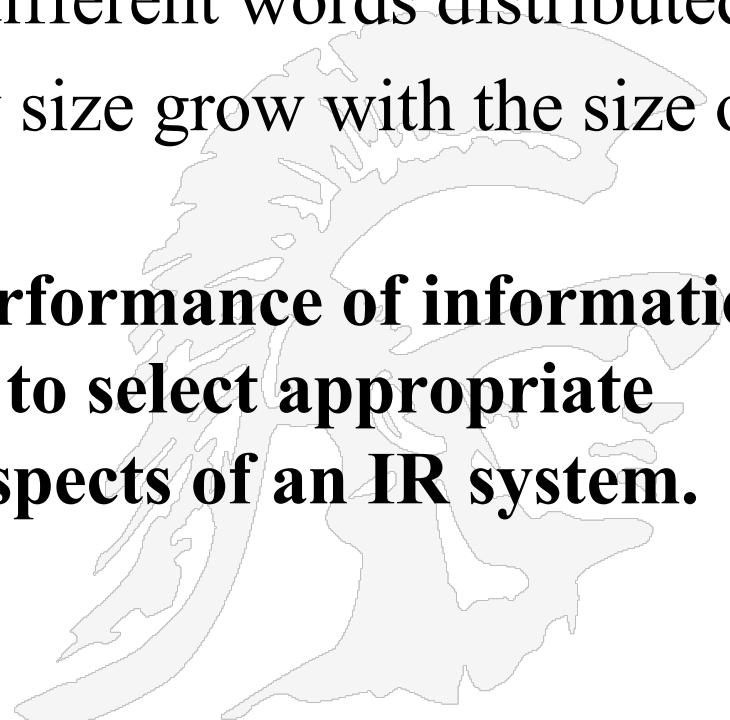
http://en.wikipedia.org/wiki/Zipf's_law

http://www.cut-the-knot.org/do_you_know/zipfLaw.shtml



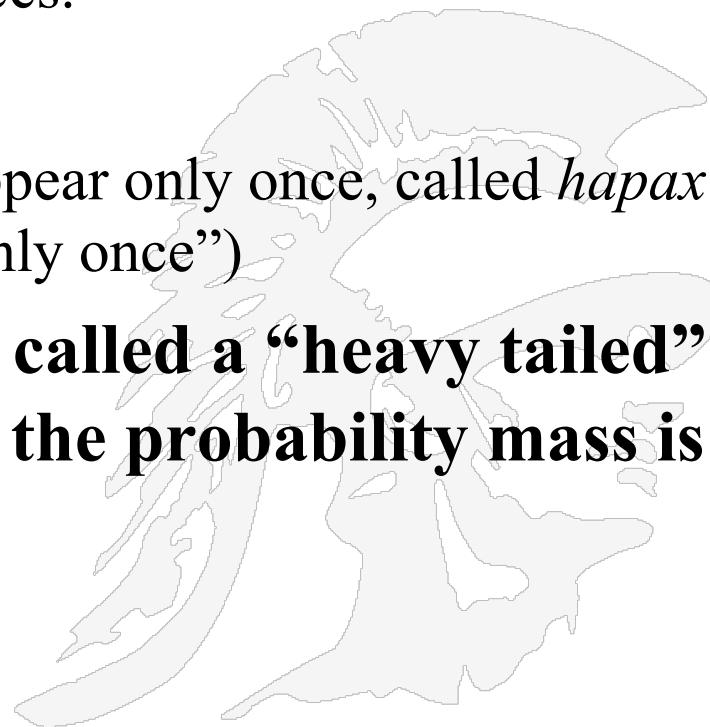
Statistical Properties of Text

- **We can ask two questions about words in a set of documents**
 1. How is the frequency of different words distributed?
 2. How fast does vocabulary size grow with the size of a set of documents?
- **Such factors affect the performance of information retrieval and can be used to select appropriate term weights and other aspects of an IR system.**



Word Frequency

- **A few words are very common.**
 - Two very frequent words (e.g. “the”, “of”) can account for about 10% of word occurrences.
- **Most words are very rare.**
 - Half the words in a corpus appear only once, called *hapax legomena* (Greek for “read only once”)
- **The above phenomenon is called a “heavy tailed” distribution, since most of the probability mass is in the “tail”**



There are Many Lists of Stop Words

A stop list of 25 semantically non-selective words

a	an	and	are	as	at	be	by	for	from	
Has	he	in	is	it	its	of	on	that	the	
to	was	were	will	with						

- The MOST COMMONLY used list of stop words
- <https://nlp.stanford.edu/IR-book/html/htmledition/dropping-common-terms-stop-words-1.html>



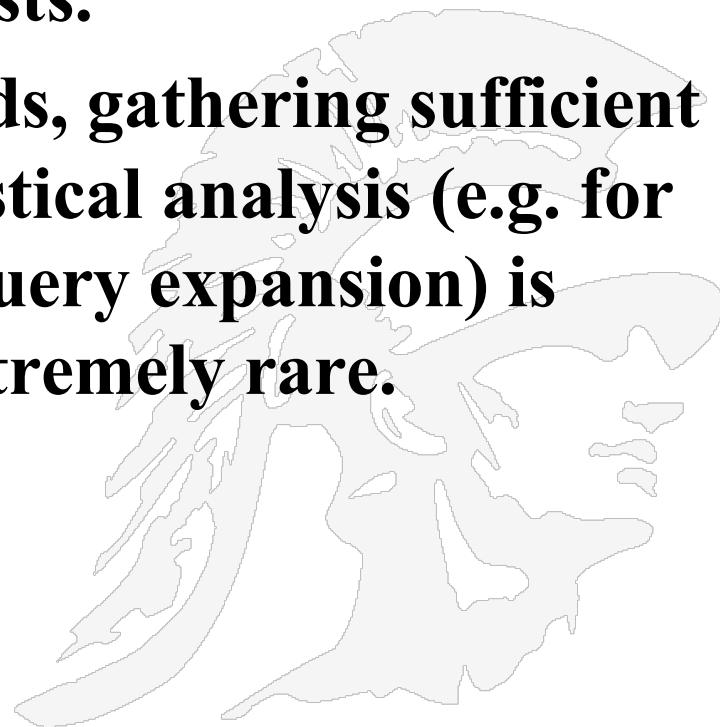
Sample Word Frequency Data (from B. Croft, UMass)

Frequent Word	Number of Occurrences	Percentage of Total
the	7,398,934	5.9
of	3,893,790	3.1
to	3,364,653	2.7
and	3,320,687	2.6
in	2,311,785	1.8
is	1,559,147	1.2
for	1,313,561	1.0
The	1,144,860	0.9
that	1,066,503	0.8
said	1,027,713	0.8

Frequencies from 336,310 documents in the 1GB TREC Volume 3 Corpus
125,720,891 total word occurrences; 508,209 unique words

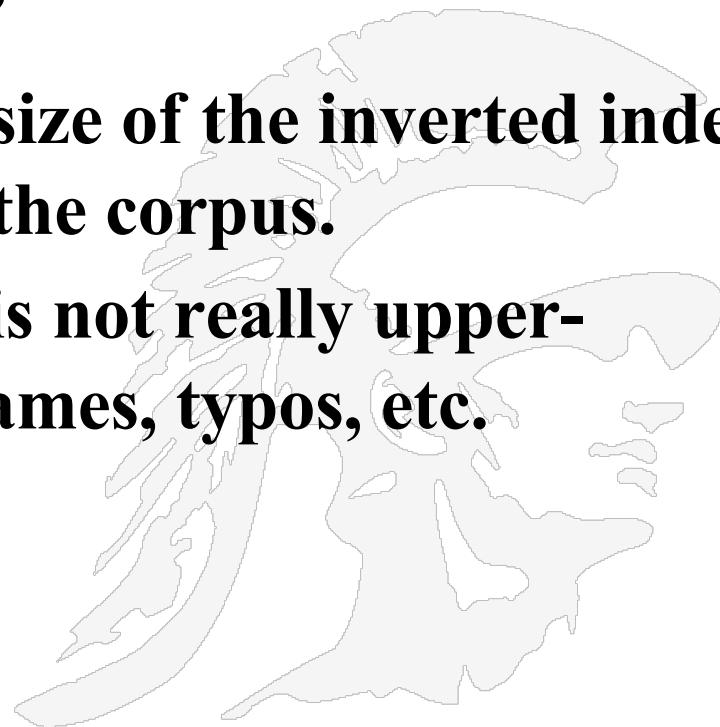
Zipf's Law Impact on IR

- **Good News:** Stopwords will account for a large fraction of text so eliminating them greatly reduces inverted-index storage costs.
- **Bad News:** For most words, gathering sufficient data for meaningful statistical analysis (e.g. for correlation analysis for query expansion) is difficult since they are extremely rare.



Vocabulary Growth

- **Our second question:** How does the size of the overall vocabulary (number of unique words) grow with the size of the corpus?
- **This determines how the size of the inverted index will scale with the size of the corpus.**
- **In practice a vocabulary is not really upper-bounded due to proper names, typos, etc.**

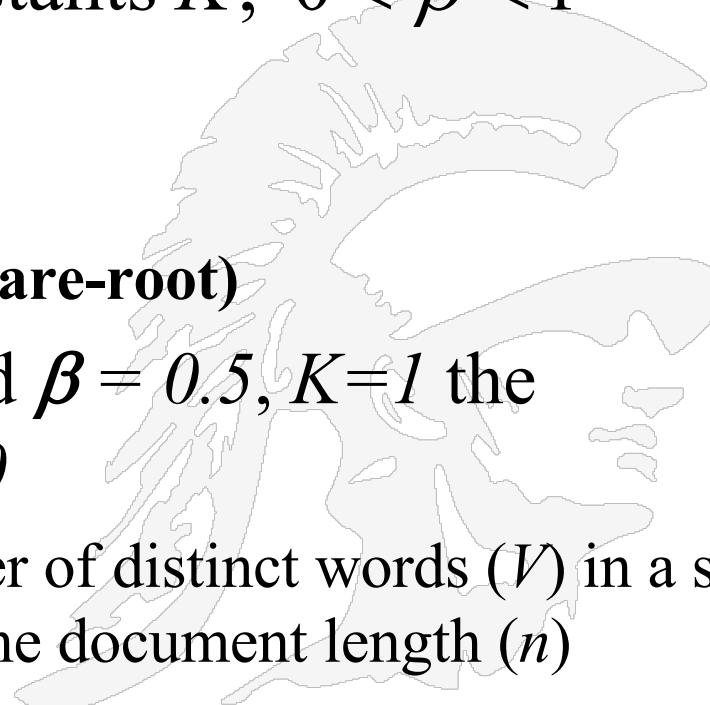


Heaps' Law

- If V is the size of the vocabulary and n is the number of words:

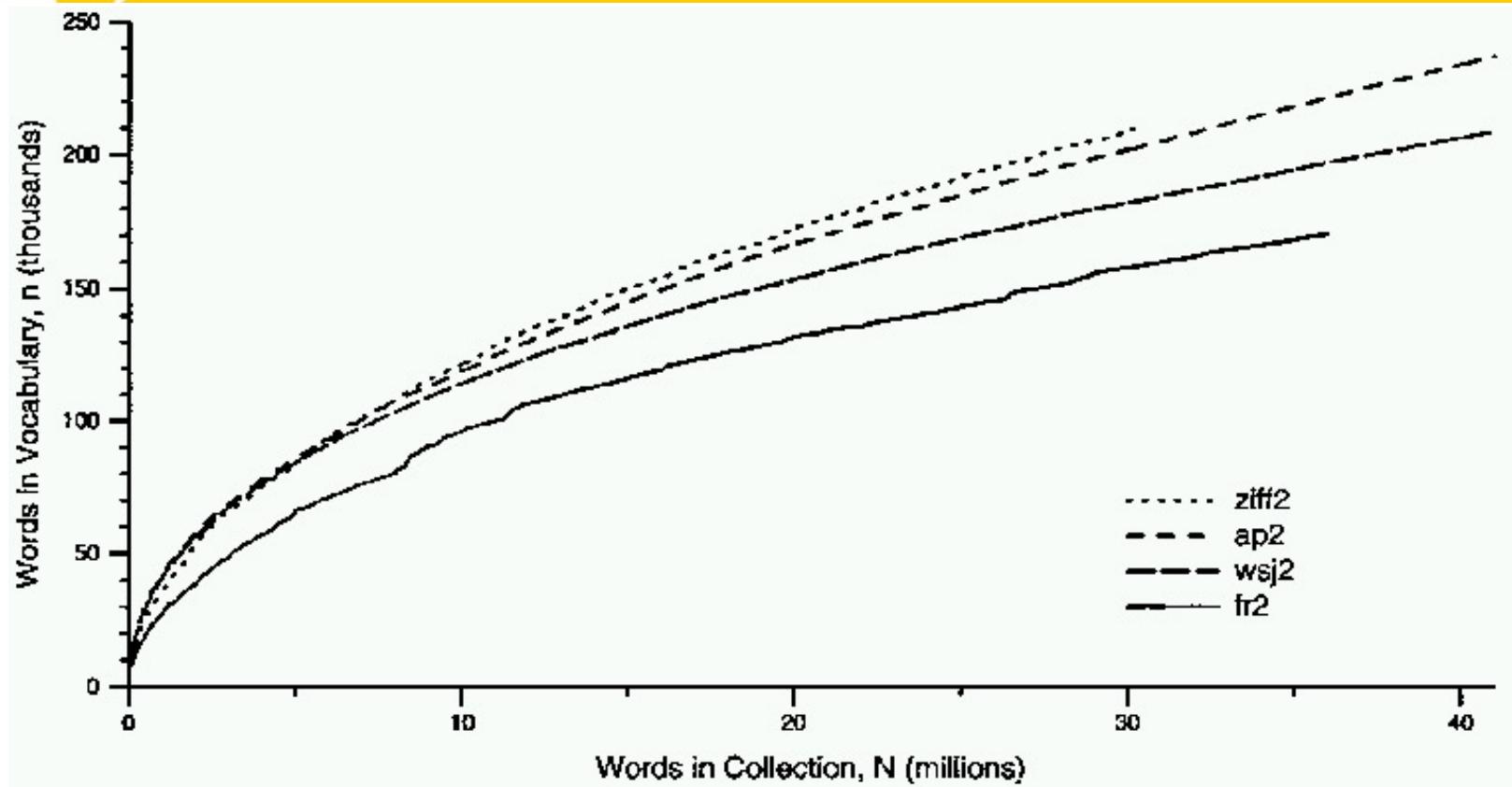
$$V = Kn^\beta \quad \text{with constants } K, 0 < \beta < 1$$

- Typical constants:
 - $K \approx 10 - 100$
 - $\beta \approx 0.4 - 0.6$ (approx. square-root)
- So for $n = 100,000,000$ and $\beta = 0.5$, $K=1$ the vocabulary size is $\approx 10,000$
- Heap's law describes the number of distinct words (V) in a set of documents as a function of the document length (n)



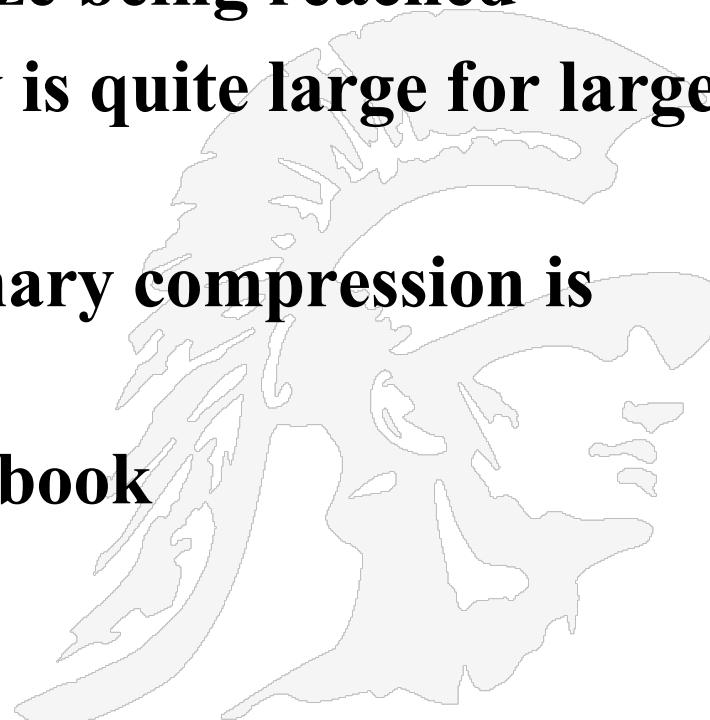
The law is named after Harold Stanley Heaps, but was originally discovered by Gustav Herdan

Heaps' Law Data: An Example

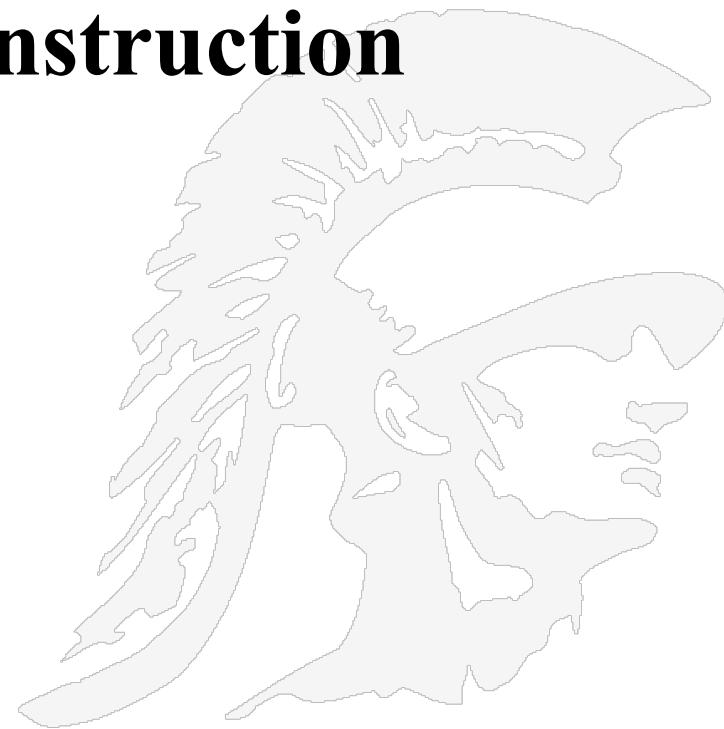


Here 40,000,000 words implies a 250,000 vocabulary further implying $k = 33$ and $b=0.5$

Heap's Law Conclusions

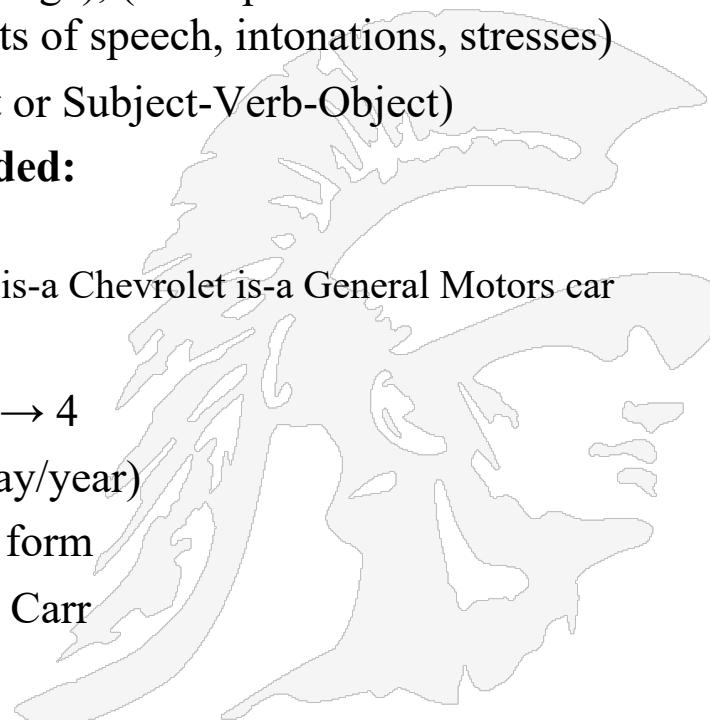
1. the dictionary size continues to increase with more documents in the collection, rather than a maximum vocabulary size being reached
 2. the size of the dictionary is quite large for large collections
- 1 and 2 imply that dictionary compression is important to do
 - See section 5.2 of our textbook
- 

Lexicon Construction



What is a Lexicon?

- **Definition:** A database of the vocabulary of a particular domain (or a language)
- It is more than a list of words/phrases
- Usually it includes some linguistic information
 - *Morphology* (manag- e/es/ing/ed → manage), (description of the structure of a language's units such as root words, parts of speech, intonations, stresses)
 - *Syntactic patterns* (such as Verb-Object or Subject-Verb-Object)
- Often some semantic information is included:
 - *Is-a hierarchy*, e.g.
 - lion is-a primate is-a mammal; corvette is-a Chevrolet is-a General Motors car
 - *Synonyms* (e.g. restaurant, café)
 - Numbers convert to normal form: Four → 4
 - Dates convert to normal form (month/day/year)
 - Alternative names converted to explicit form
 - Mr. Carr, Tyler, Presenter → Tyler Carr



Determining the Characters in a Document

- Digital documents are typically bytes in a file. The first step of processing is to convert this byte sequence into a linear sequence of characters.
 - Plain English text in ASCII encoding poses no problem
 - However, multi-byte encoding schemes, such as Unicode are more difficult to identify and process.
 - Determining the correct encoding is often handled by heuristic methods or by using provided document metadata such as:
 - `<html lang="en-US">, <html lang="fr">, <Q lang="he" dir="rtl">`
- **Unicode** (Worldwide Character Standard) is a **character** encoding standard that stores letters and other **characters** by assigning a number for each one.
- Unicode allows for **17** planes, each of **65,536** possible characters (or 'code points'). This gives a total of **1,114,112** possible characters. At present, only about 10% of this space has been allocated
- Unicode may be 8-bit, 16-bit or 32-bit
 - ASCII uses only one byte to represent each character, Unicode up to 4 bytes
- UTF-8 is a popular version of Unicode that uses 8-bit encoding
- There are 3,178 emojis in the Unicode standard as of 10/2019, see
- <https://unicode.org/emoji/charts/full-emoji-list.html>

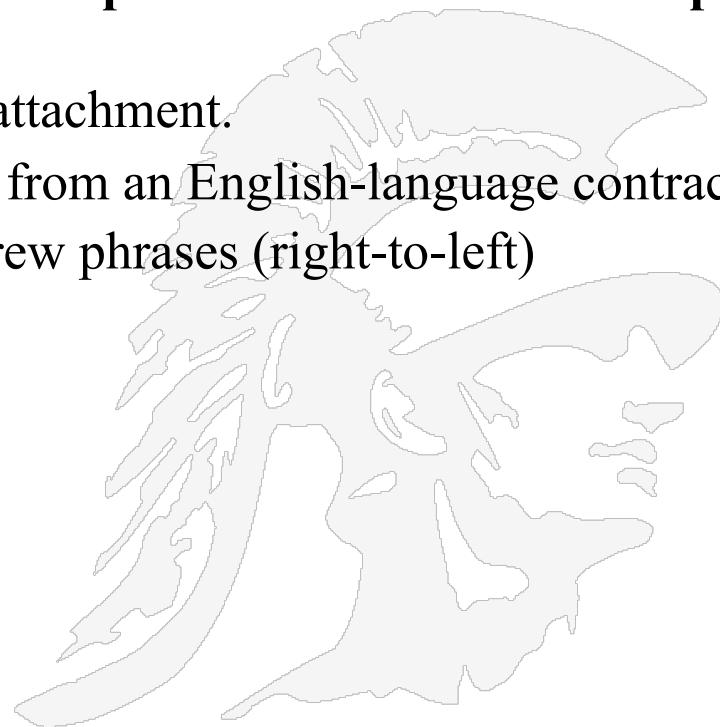
Determining the Characters

- Alternatively, the characters may have to be decoded out of some binary representation like Microsoft Word DOC files and/or a compressed format such as zip files
- Even for plain text documents, additional decoding may need to be done. In XML document's character entities, such as &, need to be decoded to give the correct character
- Finally, the textual part of the document may need to be extracted out of other material, e.g. when handling postscript (PS) or PDF files
- **Solution:** *This problem is usually solved by licensing a software library that handles decoding document formats and character encodings.*



More Complications: Format/Language

- Documents being indexed can include items from **many different languages**
 - A single index may contain terms from many languages.
- **Sometimes a single document or its components can contain multiple language formats, e.g.**
 - French email with a German pdf attachment.
 - French email with quoted clauses from an English-language contract
 - French email with Arabic or Hebrew phrases (right-to-left)

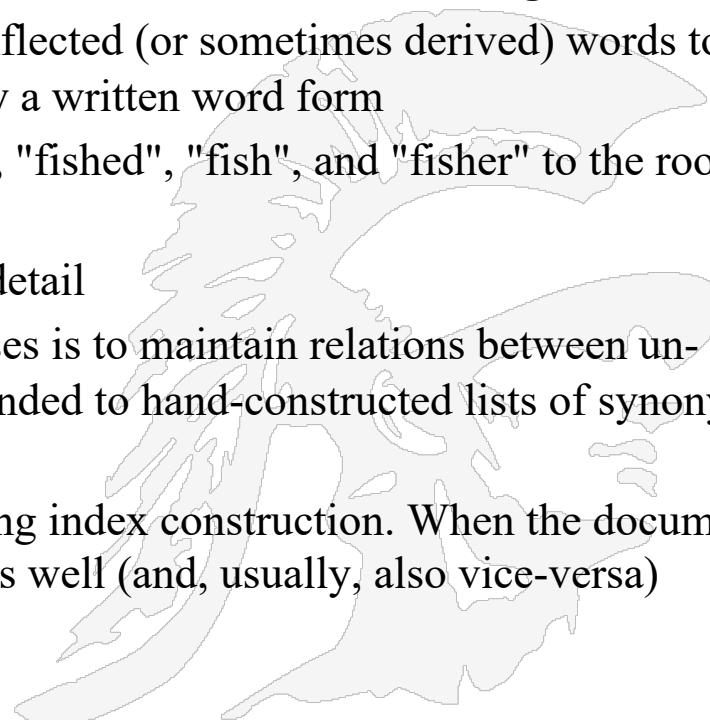


Tokenization

- **Definition:** *tokenization*: The task of chopping a document unit into pieces, called tokens, and possibly throwing away certain characters
- A *token* is an instance of a sequence of characters in some particular document that are grouped together as a useful semantic unit
- A *term* is a (possibly normalized) type that is included in the lexicon
- A **tokenizer** breaks a stream of text into tokens,
- A **lexer** is basically a **tokenizer**, but it usually attaches extra context to the tokens – e.g. this token is a number, that token is a string literal, this other token is an equality operator.
- **Simple tokenization algorithm:** “chop on whitespace and throw away punctuation characters”, but consider these examples:
 - “Mr. O’Neill thinks that the boys’ stories about Chile’s capital aren’t amusing”
 - O’Neill: is the proper token, but according to our rule above it might be treated as one of these five possibilities: neill, oneill, o’neill, o’ neill, o neill?
 - aren’t: is the proper token, but according to our rule above it might be treated as one of these four possibilities: aren’t, arent, are n’t, aren t?
- So the simple strategy of splitting on all non-alphanumeric characters can and does fail and one needs a more powerful rule

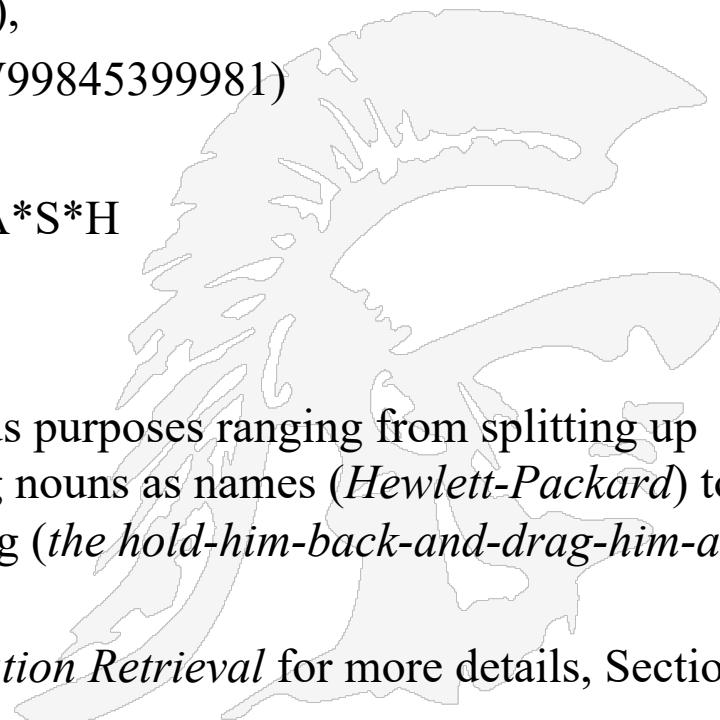
Token Normalization

- *Token normalization* is the process of canonicalizing tokens so that matches occur despite superficial differences in the character sequences of the tokens.
- The **standard** way to normalize is to implicitly create *equivalence classes*, which are normally named after one member of the set; this is often called **stemming**;
 - *stemming* is the process for reducing inflected (or sometimes derived) words to their stem, base or root form – generally a written word form
 - A stemmer reduces the words "fishing", "fished", "fish", and "fisher" to the root word, "fish".
 - Later slides discuss stemming in more detail
- An **alternative** to creating equivalence classes is to maintain relations between un-normalized tokens. This method can be extended to hand-constructed lists of synonyms such as *car* and *automobile*.
- One way is to perform the expansion is during index construction. When the document contains *automobile*, we index it under *car* as well (and, usually, also vice-versa)



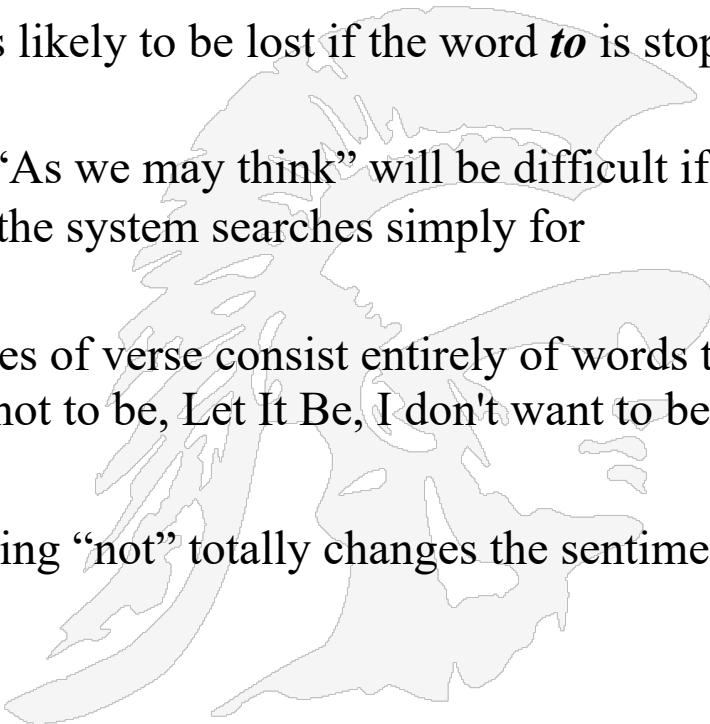
Handling Unusual Specific Tokens

- A tokenizer should recognize as a single token:
 - email addresses (jblack@mail.yahoo.com),
 - web URLs (http://stuff.big.com/new/specials.html),
 - numeric IP addresses (142.32.48.231),
 - package tracking numbers (1Z9999W99845399981)
 - E.g. aircraft names like B-52,
 - Unusual TV show name such as M*A*S*H
 - phone numbers (800) 234-2333
 - dates (Mar 11, 1983).
- In English, *hyphenation* is used for various purposes ranging from splitting up vowels in words (*co-education*) to joining nouns as names (*Hewlett-Packard*) to a copyediting device to show word grouping (*the hold-him-back-and-drag-him-away maneuver*).
- See our textbook, *Introduction to Information Retrieval* for more details, Section 2.2



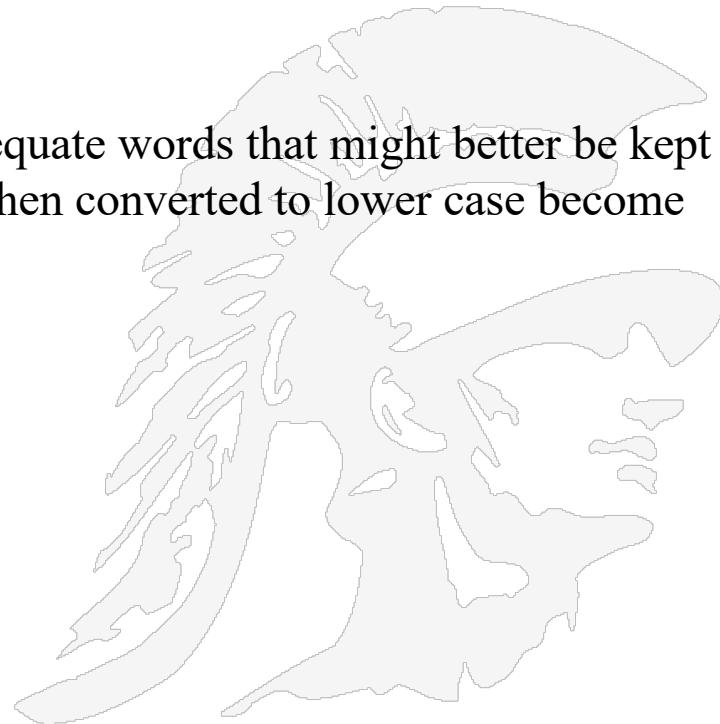
Stop Words Can be Critical to Meaning

- **Removal:** for the query: “how to implement the BM25 retrieval formula” it is best to remove stop words “how”, “to” and “the” and to focus on “implement BM25 retrieval formula” see https://en.wikipedia.org/wiki/Okapi_BM25
- **Exceptions: phrase searches**
 - The meaning of “flights **to** London” is likely to be lost if the word **to** is stopped out.
 - A search for Vannevar Bush's article “As we may think” will be difficult if the first three words are stopped out, and the system searches simply for documents containing the word think.
 - Some song titles and well known pieces of verse consist entirely of words that are commonly on stop lists (To be or not to be, Let It Be, I don't want to be, ...).
- **Exceptions: sentiment determination**
 - ”I told you she was not happy” removing “not” totally changes the sentiment of the sentence



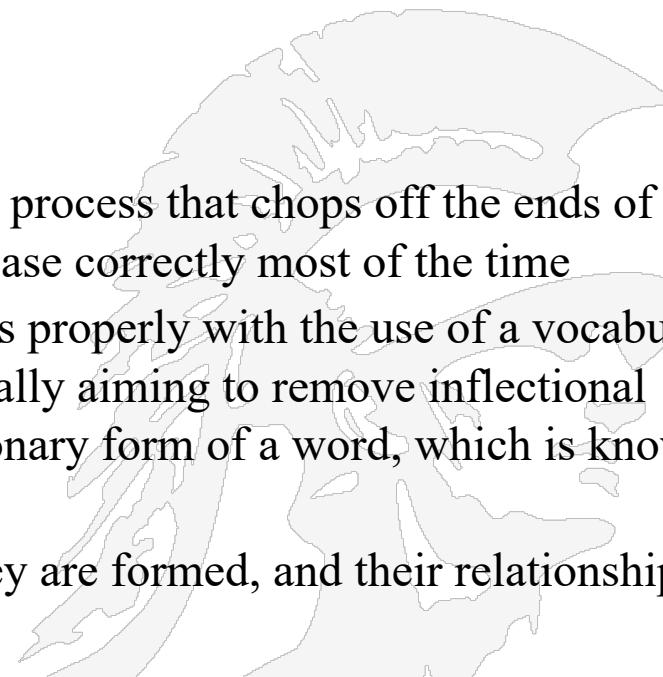
Capitalization/Case-Folding

- *case-folding* is reducing all letters to lower case.
- Often this is a good idea:
 - *automobile* will match *Automobile*
 - *ferrari* will match *Ferrari*
- On the other hand, such case folding can equate words that might better be kept apart. Here are some Proper Nouns that when converted to lower case become something different:
 - *General Motors*,
 - *The Associated Press*
 - *the Fed* vs. *fed*
 - person names such as *Bush*, *Black*



Stemming and Lemmatization

- **Goal:** to reduce multiple forms of a word to a common base form, e.g.
 - am, are, is => be
 - car, cars, car's, car' => car
- **So for example**
 - the boy's cars are different colors =>
 - the boy car be differ color
- *Stemming* usually refers to a crude heuristic process that chops off the ends of words in the hope of achieving a common base correctly most of the time
- *Lemmatization* usually refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the *lemma*
- **morphology** is the study of **words**, how they are formed, and their relationship to other **words**
 - It analyzes the structure of **words** and parts of **words**, such as stems, root **words**, prefixes, and suffixes.



Stemming vs Lemmatization

- Lemmatization and stemming are special cases of normalization. They identify a canonical representative for a set of related word forms.
- The goal of both stemming and lemmatization is to reduce inflectional forms of a word to a common base form.
- *Stemming* usually refers to a crude heuristic process that chops off the ends of words
- *Lemmatization* usually refers to the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma
- <https://stackoverflow.com/questions/1787110/what-is-the-difference-between-lemmatization-vs-stemming/1787121#1787121>
- Example:
 - The word "better" has "good" as its lemma. This link is missed by stemming, as it requires a dictionary look-up.
 - The word "walk" is the base form for word "walking", and hence this is matched in both stemming and lemmatization.
 - The word "meeting" can be either the base form of a noun or a form of a verb ("to meet") depending on the context, e.g., "in our last meeting" or "We are meeting again tomorrow". Unlike stemming, lemmatization can in principle select the appropriate lemma depending on the context.

CSCI 571 - Home USC Computer Science... ITS - Software Masters Student ... University of Sout...

Javascript Porter Stemmer Online

[View the source \(minified\)](#)

Find out more about the Porter Stemming algorithm at the [official site](#).

Example:
 Do you really think it is weakness that yields to temptation? I tell you that there are terrible temptations which it requires strength, strength and courage to yield to ~ Oscar Wilde

Stem the above content

Overlay

Example Do you really think it is weakness that yields to temptation I tell you that there are terrible temptations which it requires strength strength and courage to yield to Oscar Wilde

Stemmed

Examp Do you realli think it is weak that yield to temptat I tell you that there ar terribl temptat which it requir strength strength and courag to yield to Oscar Wild

JavaScript Porter Stemmer Online

- A stemmer written by Martin Porter has become the de-facto standard algorithm used for English stemming

- The website to the left implements Porter's stemmer algorithm in JavaScript

- Example: “to be or not to be” is not changed at all

Example: Since the department's founding in 1968, our faculty have made pioneering contributions to fundamental and interdisciplinary fields of computing.”

Since -> Sinc
 department -> depart
 founding -> found
 pioneering -> pioneer
 contributions -> contribut
 fundamental -> fundament

The Porter Stemming Algorithm

- The official site is located at <http://tartarus.org/~martin/PorterStemmer/>
- The site offers versions of the algorithm in multiple languages including C, Perl, Java, JavaScript, PHP

Original paper
written in
1979;

JavaScript version can
be found at
<http://tartarus.org/~martin/PorterStemmer/js.txt>

The JavaScript is very easy
to follow



This page was completely revised Jan 2006. The earlier edition is [here](#).

This is the 'official' home page for distribution of the Porter Stemming Algorithm, written and maintained by its author, [Martin Porter](#).

The Porter stemming algorithm (or 'Porter stemmer') is a process for removing the commoner morphological and inflectional endings from words in English. Its main use is as part of a term normalisation process that is usually done when setting up Information Retrieval systems.

History

The original [stemming algorithm paper](#) was written in 1979 in the Computer Laboratory, Cambridge (England), as part of a larger IR project, and appeared as Chapter 6 of the final project report,

C.J. van Rijsbergen, S.E. Robertson and M.F. Porter, 1980. *New models in probabilistic information retrieval*. London: British Library. (British Library Research and Development Report, no. 5587).

With van Rijsbergen's encouragement, it was also published in,

M.F. Porter, 1980, An algorithm for suffix stripping, *Program*, 14(3) pp 130–137.

And since then it has been reprinted in

Karen Sparck Jones and Peter Willett, 1997, *Readings in Information Retrieval*, San Francisco: Morgan Kaufmann, ISBN 1-55860-454-4.

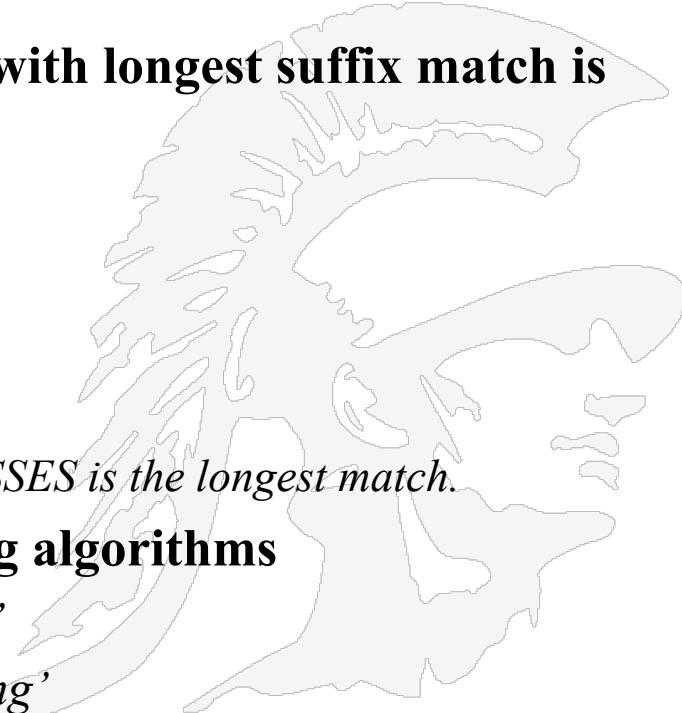
The original stemmer was written in [BCPL](#), a language once popular, but now defunct. For the first few years after 1980 it was distributed in its BCPL form, via the medium of punched paper tape. Versions in other languages soon began to appear, and by 1999 it was being widely used, quoted and adapted. Unfortunately there were numerous variations in functionality among these versions, and this web page was set up primarily to 'put the record straight' and establish a definitive version for distribution.

- Porter's algorithm has five phases of word reductions, applied sequentially
- Each phase has a set of conventions for how to select the transformation
- For a given set of rules only the rule with longest suffix match is applied; for example

- *sses* → *ss*
- *ies* → *i*
- *ational* → *ate*
- *tional* → *tion*

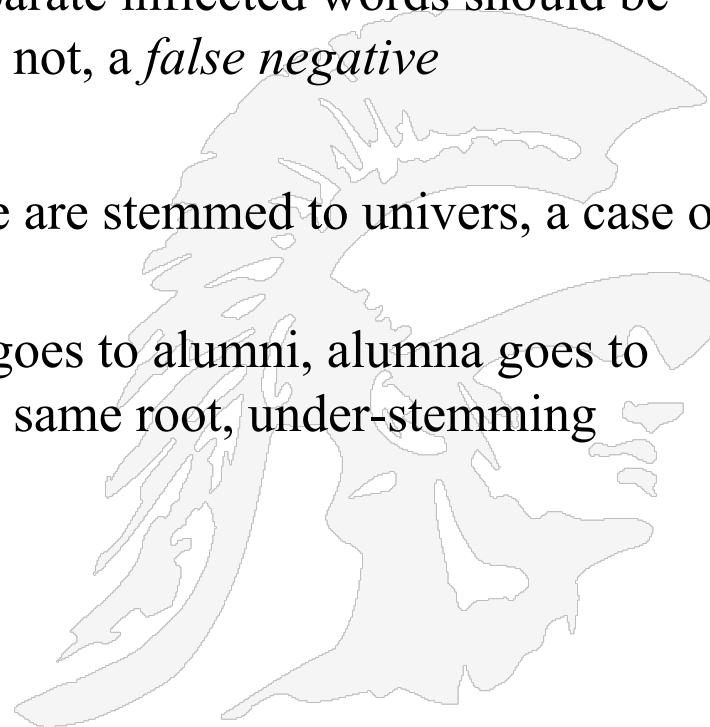
For Ex. CARESSES maps to CARESS since SSES is the longest match.

- Other possibilities are suffix stripping algorithms
 - If the word ends in 'ed' remove the 'ed'
 - If the word ends in 'ing', remove the 'ing'
 - If the word ends in 'ly', remove the 'ly'



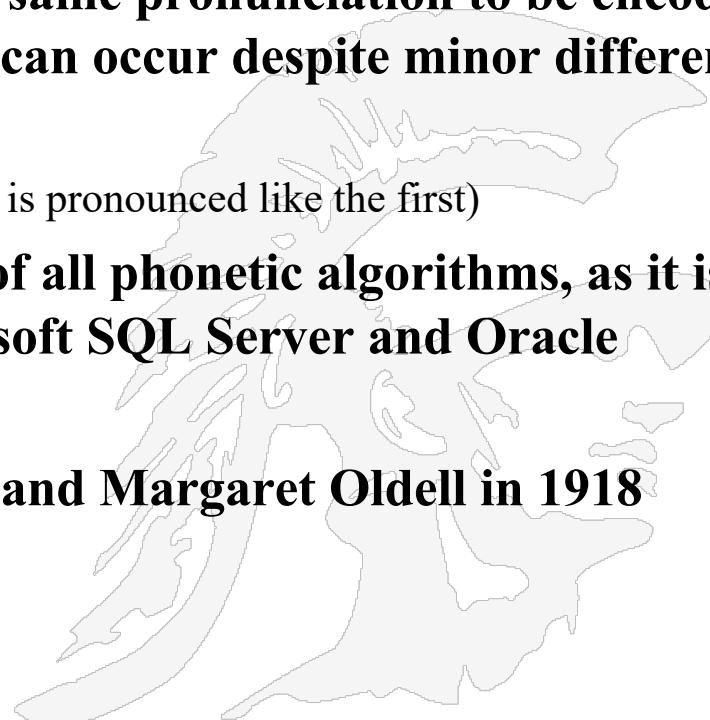
Error Metrics

- **Two error measurements**
 - **Over-stemming:** two separate inflected words are stemmed to the same root, but should not have been, a *false positive*
 - **Under-stemming:** where two separate inflected words should be stemmed to the same root, but are not, a *false negative*
- **In the Porter stemmer**
 - Universal, university and universe are stemmed to univers, a case of over-stemming
 - alumnus goes to alumnu, alumni goes to alumni, alumna goes to alumna, but all should stem to the same root, under-stemming



One Last Algorithm Soundex Algorithm

- Soundex is a phonetic algorithm *for indexing names by their sound when pronounced in English.*
- The basic aim is for names with the same pronunciation to be encoded to the same string so that matching can occur despite minor differences in spelling, e.g.
 - SMITH and SMYTH (the second option is pronounced like the first)
- Soundex is the most widely known of all phonetic algorithms, as it is a standard feature of MySQL, Microsoft SQL Server and Oracle
- E.g., *chebyshev* → *tchebycheff*
- It was developed by Robert Russell and Margaret Oldell in 1918



Soundex – Typical Algorithm

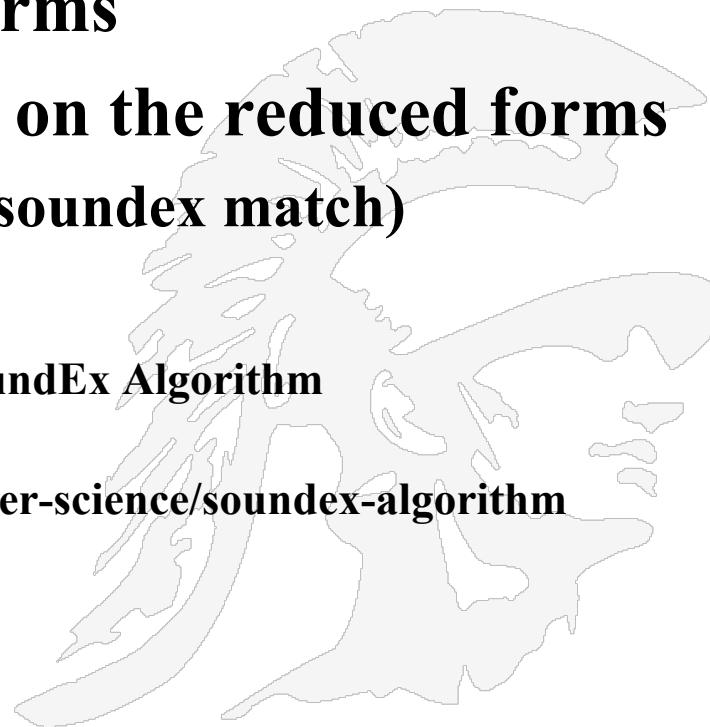
- Turn every token to be indexed into a 4-character reduced form
- Do the same with query terms
- Build and search an index on the reduced forms
 - (when the query calls for a soundex match)

Here are some good links that discuss the SoundEx Algorithm

<https://en.wikipedia.org/wiki/Soundex>

<https://www.sciencedirect.com/topics/computer-science/soundex-algorithm>

<http://www.blackwasp.co.uk/soundex.aspx>



Soundex – Algorithm

- The algorithm mainly encodes consonants; a vowel will not be encoded unless it is the first letter.
- Every Soundex code consists of a letter and three numbers, e.g. W-252;
 - the letter is the first letter of the surname
 - The numbers are assigned to the remaining letters of the surname according to the Soundex guide
- Washington is coded W-252 (W, 2 for the S, 5 for the N, 2 for the G, and remaining letters disregarded)
- Lee is coded L-000 (L the leading letter and since the ee are dropped the zeros are used as padding)

A web page containing versions of Soundex in 36 different programming languages is here

<http://rosettacode.org/wiki/Soundex>

<http://www.archives.gov/research/census/soundex.html>

Soundex – Typical Algorithm

1. Retain the first letter of the word.
2. Change all occurrences of the following letters to '0' (zero):
'A', 'E', 'I', 'O', 'U', 'H', 'W', 'Y'.
3. Change letters to digits as follows:
 - B, F, P, V → 1
 - C, G, J, K, Q, S, X, Z → 2
 - D, T → 3
 - L → 4
 - M, N → 5
 - R → 6

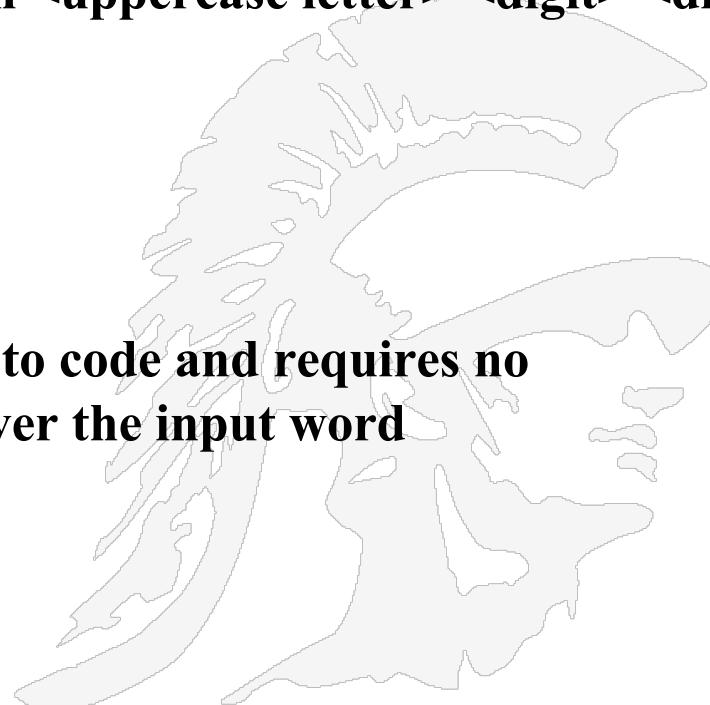


Soundex Continued

4. Remove all pairs of consecutive digits.
5. Remove all zeros from the resulting string.
6. Pad the resulting string with trailing zeros and return the first four positions, which will be of the form <uppercase letter> <digit> <digit> <digit>.

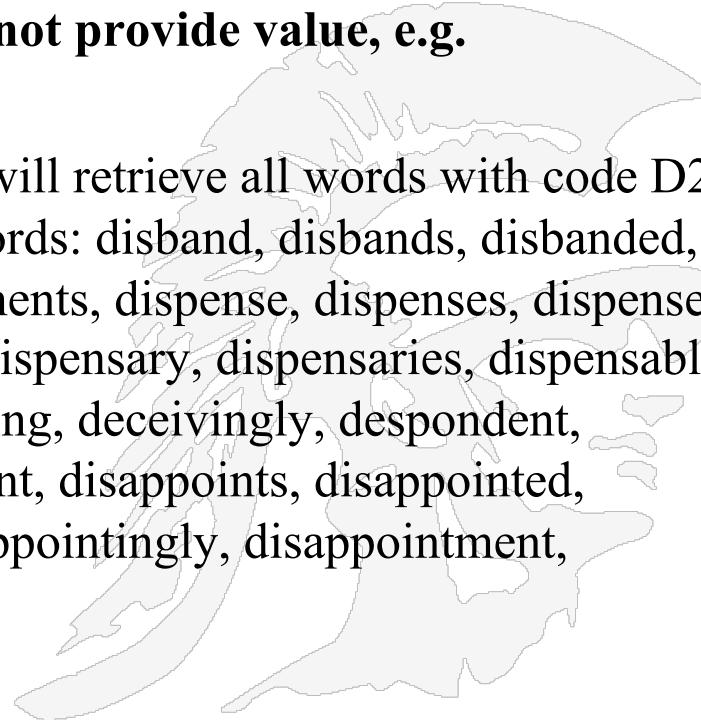
E.g., *Herman* becomes H655

- The algorithm is straight forward to code and requires no backtracking or multiple passes over the input word



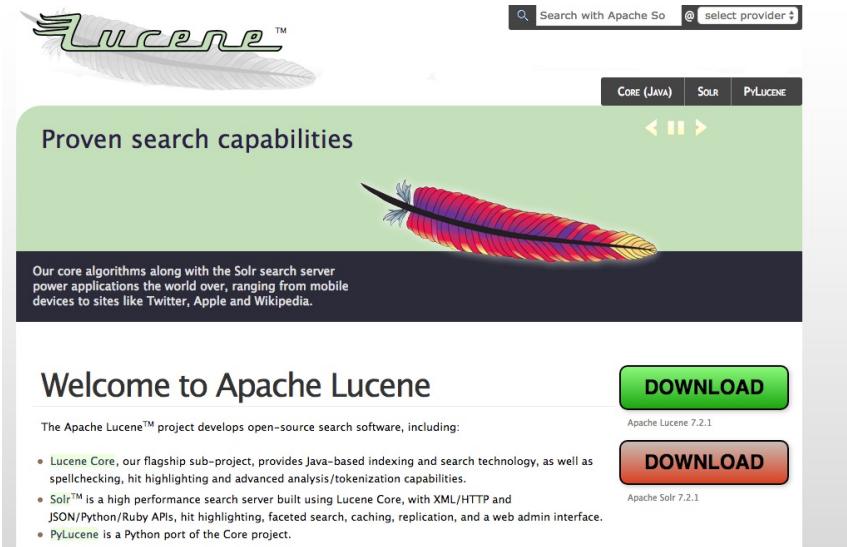
Soundex Example

- **Example 1: Zbygniewski has code Z125**
 - starts with Z, consonant in group 1 (the b), one in group 2 (the g), one in group 5 (the n), remainder is ignored
 - if Zbignyefsky is entered, the same code is produced
- **For some words using Soundex does not provide value, e.g.**
- **Example 2: disapont has code D215**
 - a spelling corrector in its database will retrieve all words with code D215 which includes the following 31 words: disband, disbands, disbanded, disbanding, disbandment, disbandments, dispense, dispenses, dispensed, dispensing, dispenser, dispensers, dispensary, dispensaries, dispensable, dispensation, dispensations, deceiving, deceivingly, despondent, despondently, disobeying, disappoint, disappoints, disappointed, disappointing, disappointedly, disappointingly, disappointment, disappointments, disavowing



Postscript: Lucene and Solr

- **Lucene**
 - a high-performance, full-featured text search engine library written entirely in Java.
 - suitable for any application that requires full-text search
 - an open source project available for free download
- **Solr**
 - HTTP-based Search Server
 - Uses XML for configuration
 - Many, many nice features that Lucene users need
 - Faceting, spell checking, highlighting
 - Caching, Replication, Distributed
- Download both from <http://lucene.apache.org>



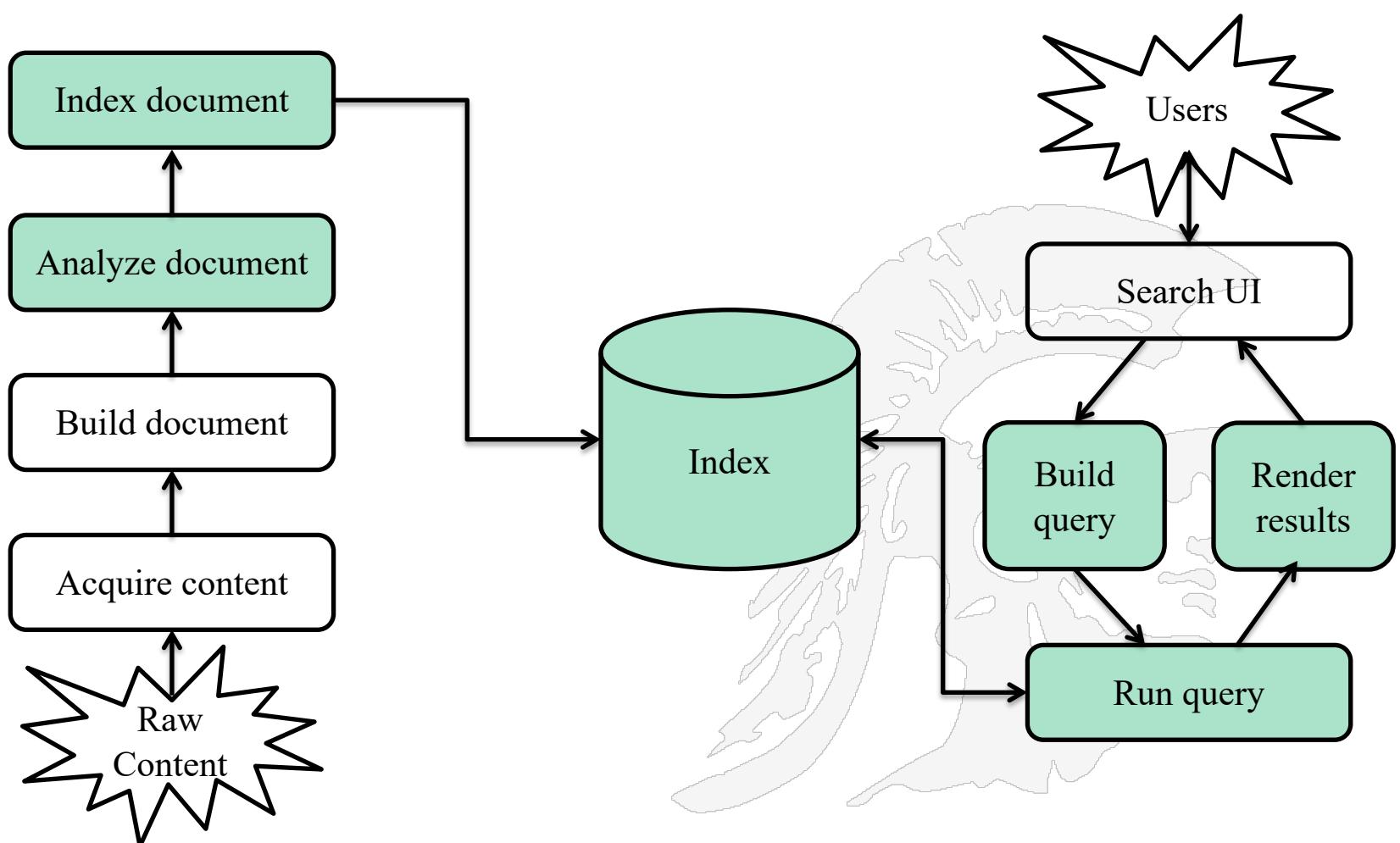
The screenshot shows the Apache Lucene homepage. At the top right is a search bar with a magnifying glass icon and the placeholder "Search with Apache Solr". Below it are three buttons: "select provider", "CORE (JAVA)", "SOLR", and "PYLUCENE". A large feather graphic is on the left. The main content area has a green header "Proven search capabilities" and a black footer with text about core algorithms and Solr's applications. Below the green header is another feather graphic. The central text area says "Welcome to Apache Lucene" and describes the project's purpose. It includes two "DOWNLOAD" buttons: one for "Apache Lucene 7.2.1" in green and one for "Apache Solr 7.2.1" in orange. The bottom right features a large, stylized feather graphic.

Main Lucene Modules

- **Lucene is the underlying software that imports documents and creates an inverted index**
 - **Document Parsing and Analysis**
 - This section is in charge of tokenization
 - **Document Identification**
 - Where the Document ID is created
 - Date of Document is extracted
 - Title of document is extracted
 - **Indexing**
 - Provides access to indexes
 - Maintains indexes

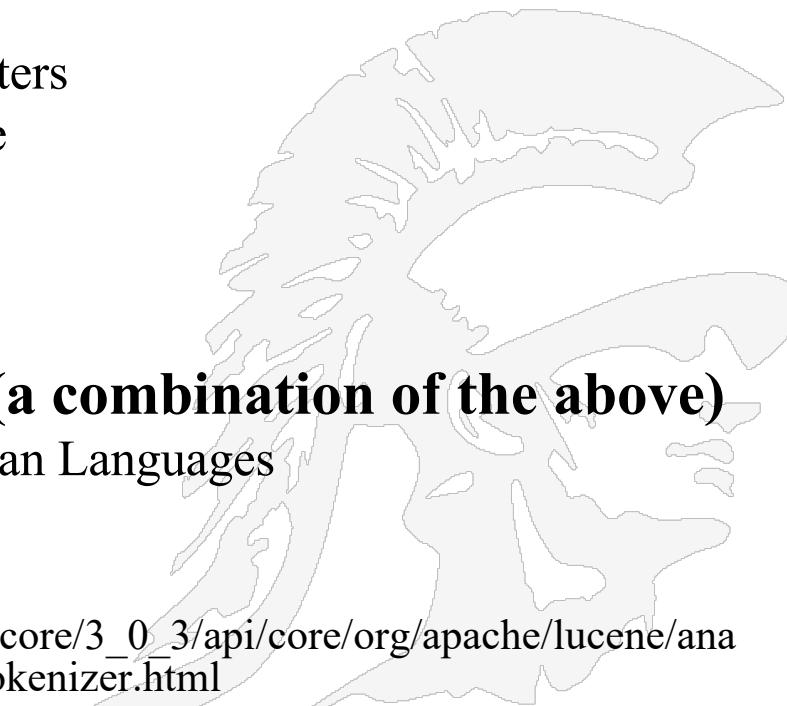


Lucene in a search system



Lucene Tokenizers Specify How the Text in a Field is to be Indexed

- **Tokenizers in Lucene**
 - **WhitespaceTokenizer**
 - divides text at whitespace
 - **SimpleTokenizer**
 - divides text at non-letters
 - converts to lower case
 - **StopTokenizer**
 - SimpleAnalyzer
 - removes stop words
 - **StandardTokenizer (a combination of the above)**
 - good for most European Languages
 - removes stop words
 - convert to lower case
 - https://lucene.apache.org/core/3_0_3/api/core/org/apache/lucene/analysis/standard/StandardTokenizer.html
- **In Lucene you can create your own Tokenizers**



Example 1

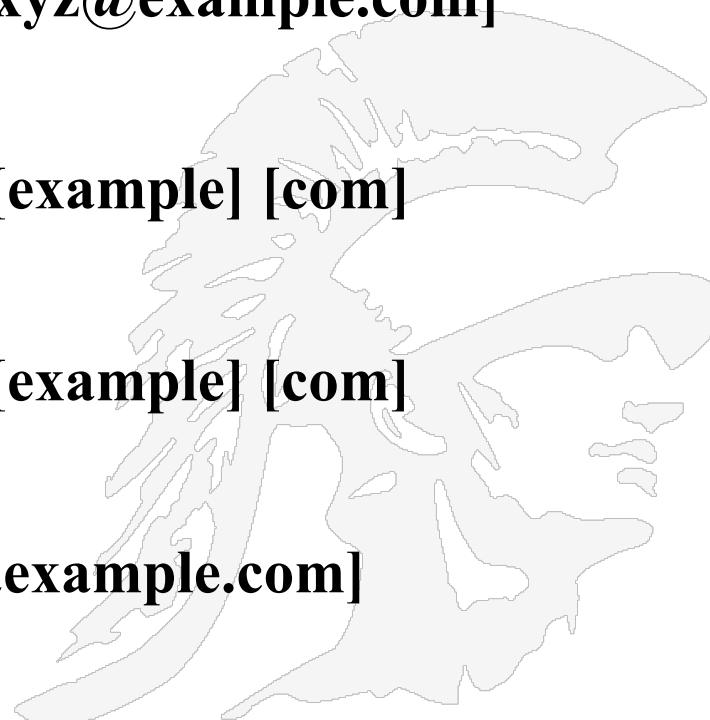
Analysis example

- “The quick brown fox jumped over the lazy dog”
- **WhitespaceAnalyzer**
 - [The] [quick] [brown] [fox] [jumped] [over] [the] [lazy] [dog]
- **SimpleAnalyzer**
 - [the] [quick] [brown] [fox] [jumped] [over] [the] [lazy] [dog]
- **StopAnalyzer**
 - [quick] [brown] [fox] [jumped] [over] [lazy] [dog]
- **StandardAnalyzer**
 - [quick] [brown] [fox] [jumped] [over] [lazy] [dog]



Example 2

- “**XY&Z Corporation – xyz@example.com**”
- **WhitespaceAnalyzer**
 - [XY&Z] [Corporation] [-] [xyz@example.com]
- **SimpleAnalyzer**
 - [xy] [z] [corporation] [xyz] [example] [com]
- **StopAnalyzer**
 - [xy] [z] [corporation] [xyz] [example] [com]
- **StandardAnalyzer**
 - [xy&z] [corporation] [xyz@example.com]



the LexCorp BFG-900 is a printer

Lex corp bfg900 printers

TEXT**QUERY***WhitespaceTokenizer*

the LexCorp BFG-900 is a printer

Lex corp bfg900 printers

WordDelimiterFilter

the Lex Corp BFG 900 is a printer

Lex corp bfg 900 printers

LowerCaseFilter

the lex corp bfg 900 is a printer

lex corp bfg 900 printers

StopwordFilter

lex corp bfg 900 printer

lex corp bfg 900 printers

StemmerFilter

lex corp bfg 900 print-

lex corp bfg 900 print-

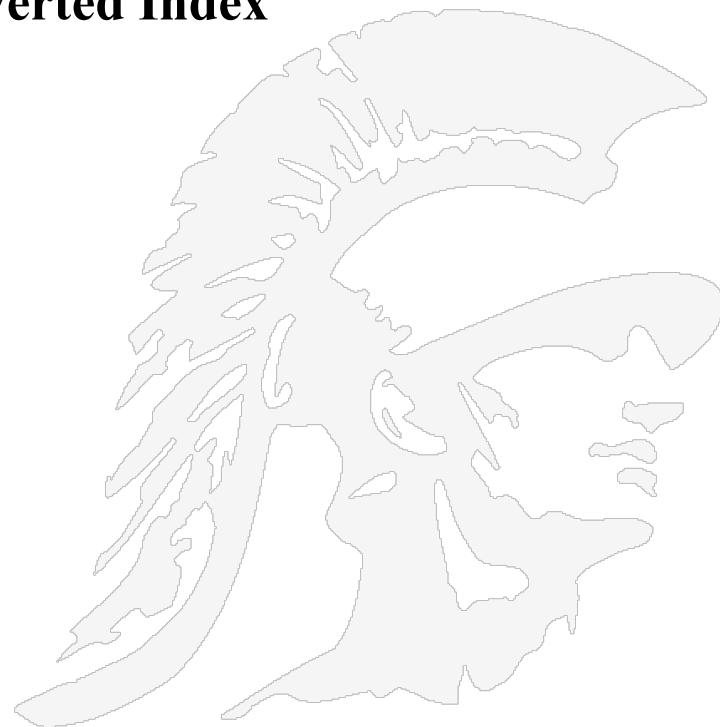
Match

Inverted Indexing



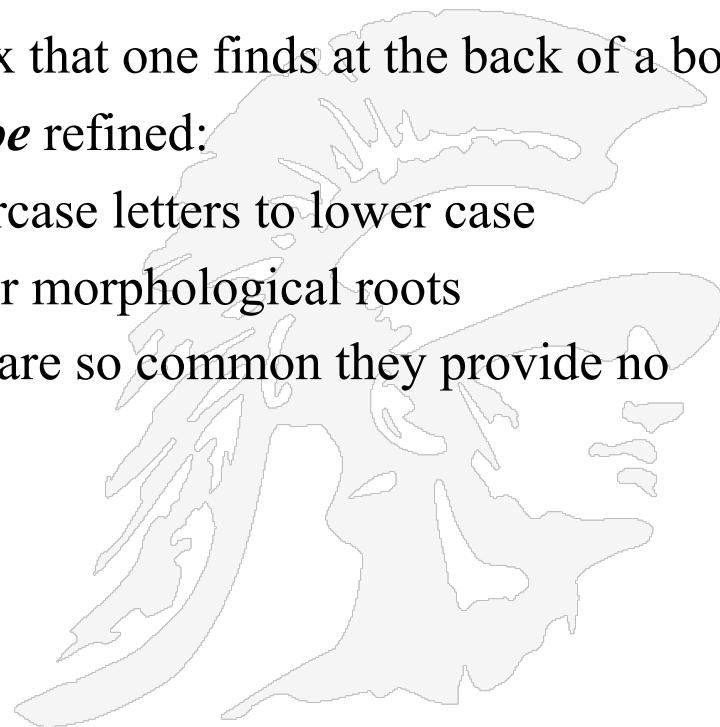
Outline

- **Definition of an Inverted index**
- **Examples of Inverted Indices**
- **Representing an Inverted Index**
- **Processing a Query on a Linked Inverted Index**
- **Skip Pointers to Improve Merging**
- **Phrase Queries**
- **biwords**
- **Grammatical Tagging**
- **N-Grams**
- **Distributed Indexing**



Creating an Inverted Index

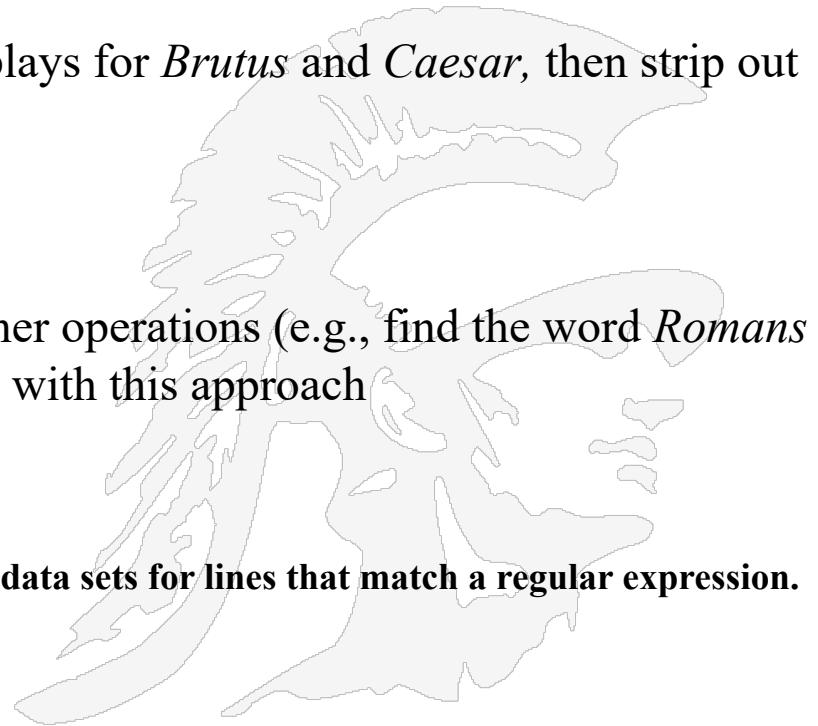
- An inverted index is typically composed of a vector containing all distinct words of the text collection in lexicographical order (which is called the **vocabulary**) and for each word in the vocabulary, a list of all documents (and text positions) in which that word occurs
 - This is nothing more than an index that one finds at the back of a book
- Terms in the inverted file index *may be* refined:
 - **Case folding:** converting all uppercase letters to lower case
 - **Stemming:** reducing words to their morphological roots
 - **Stop words:** removing words that are so common they provide no information



Processing a Query

An Example

- The Query
 - Which plays of Shakespeare contain the words *Brutus* AND *Caesar* but NOT *Calpurnia*?
- One Possible Solution
 - One could grep all of Shakespeare's plays for *Brutus* and *Caesar*, then strip out lines containing *Calpurnia*?
 - Too Slow (for large corpora)
 - Requires lots of space
 - This method doesn't allow for other operations (e.g., find the word *Romans* near *countrymen*) are not feasible with this approach



grep is a command-line utility for searching plain-text data sets for lines that match a regular expression.

Term-Document Incidence Matrix

One way to think about an inverted index is to consider it as a sparse matrix where rows represent terms and columns represent documents

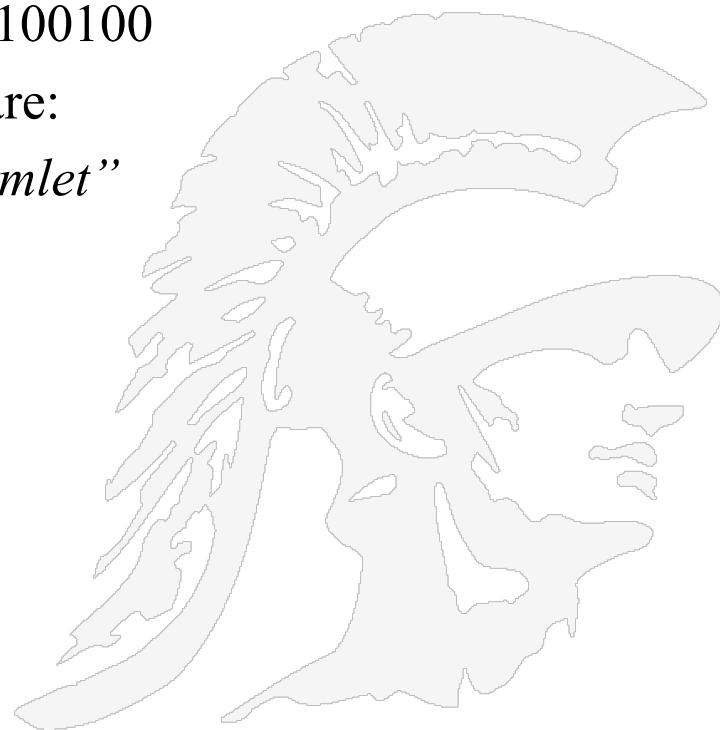
documents	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
terms	Antony	1	1	0	0	0
	Brutus	1	1	0	1	0
	Caesar	1	1	0	1	1
	Calpurnia	0	1	0	0	0
	Cleopatra	1	0	0	0	0
	mercy	1	0	1	1	1
	worser	1	0	1	1	0

The Query:
Brutus AND Caesar but NOT Calpurnia

1 if the play contains word, 0 otherwise

Incidence Vectors

- So we have a 0/1 vector for each term.
- To answer the previous query: take the vectors for *Brutus*, *Caesar* and *Calpurnia* (complemented) and do a bitwise *AND*.
- $110100 \text{ AND } 110111 \text{ AND } 101111 = 100100$
- So the two plays matching the query are:
“Anthony and Cleopatra”, *“Hamlet”*

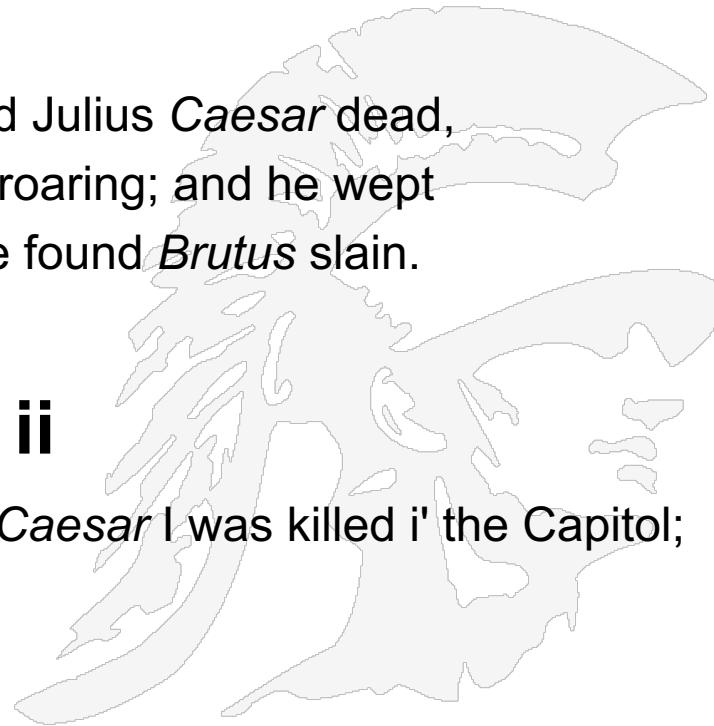


Actual Answers to the Query

- **Antony and Cleopatra, Act III, Scene ii**
- ***Agrippa [Aside to DOMITIUS ENOBARBUS]:***

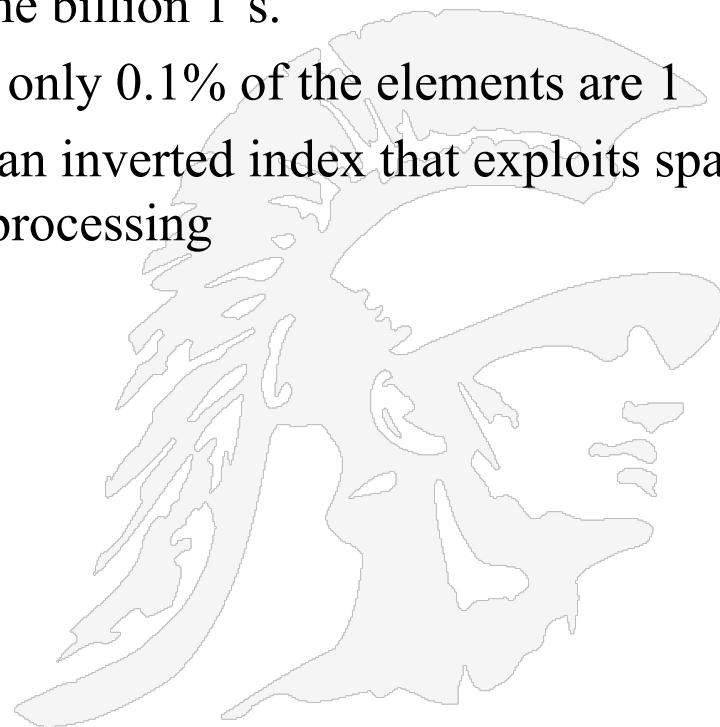
- Why, Enobarbus,
- When Antony found Julius Caesar dead,
- He cried almost to roaring; and he wept
- When at Philippi he found Brutus slain.

- **Hamlet, Act III, Scene ii**
- ***Lord Polonius:*** I did enact Julius Caesar
I was killed i' the Capitol;
Brutus killed me.

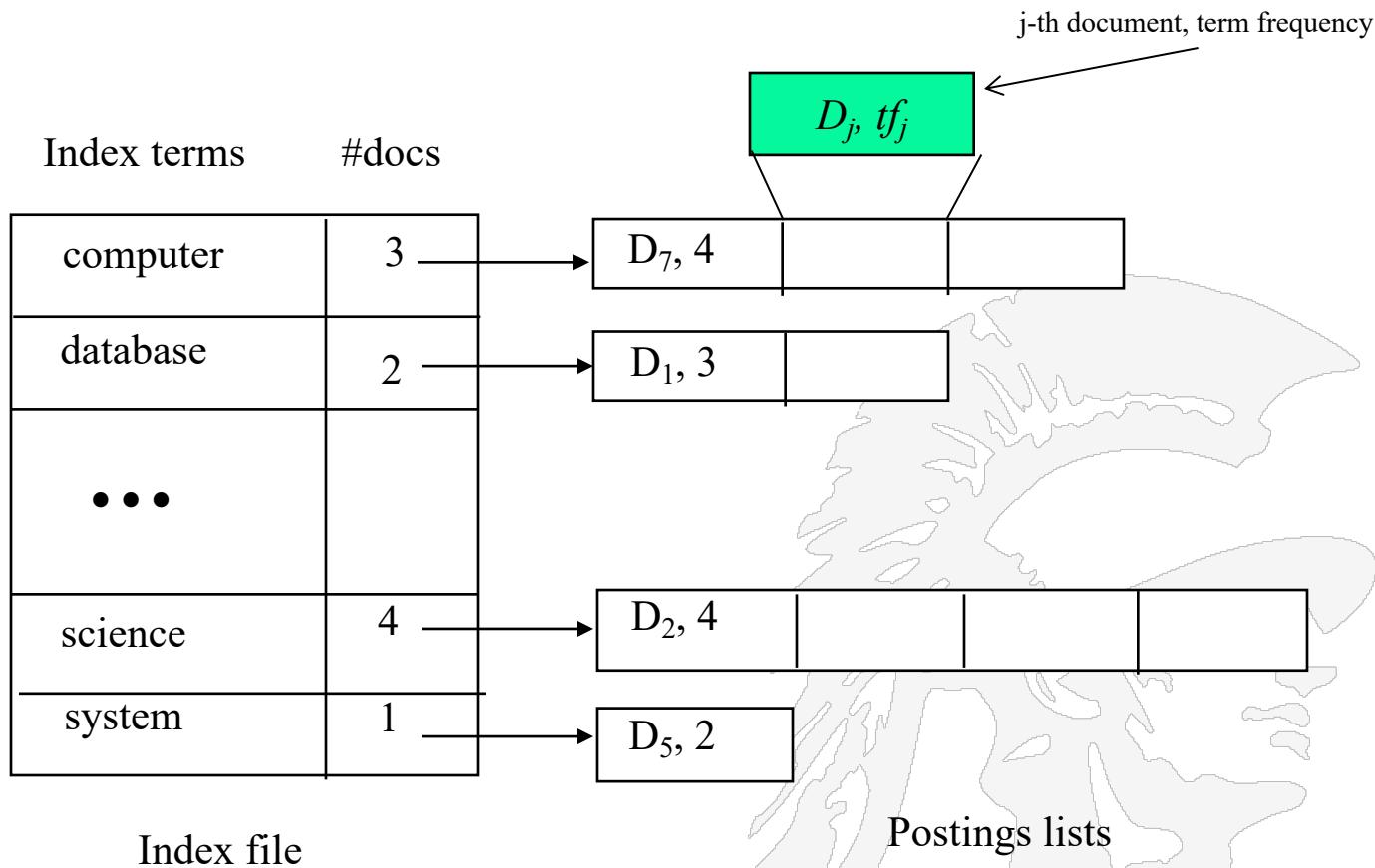


Term-Document Incident Matrices are Naturally Sparse

- Given 1 million documents and 500,000 terms
- The (term x document) matrix in this case will have size 500K x 1M or half-a-trillion 0's and 1's.
- But it will likely have no more than one billion 1's.
 - So the matrix is extremely sparse, only 0.1% of the elements are 1
- So instead we use a data structure for an inverted index that exploits sparsity and then devise algorithms for query processing



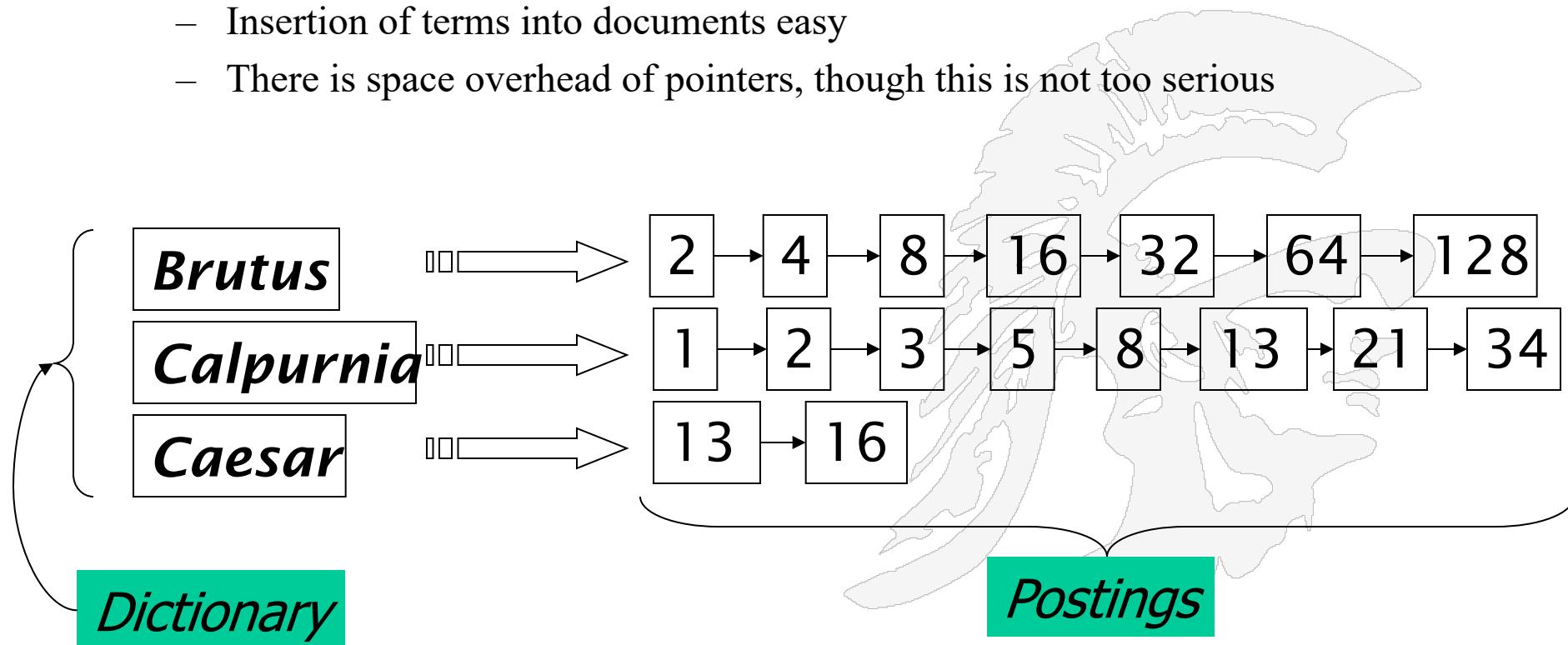
Inverted Index Example



The two parts of an inverted index. The **dictionary** (Index file) is usually kept in memory, with pointers to each **postings list**, which is stored on disk. The dictionary has been sorted alphabetically and the postings list is sorted by document ID

Inverted Index Stored In Two Parts

- For each term T , we must store a list of all documents that contain T .
- Linked lists are generally preferred to arrays, why . . .
 - Dynamic space allocation
 - Insertion of terms into documents easy
 - There is space overhead of pointers, though this is not too serious



Parsing Documents To Create an Inverted Index

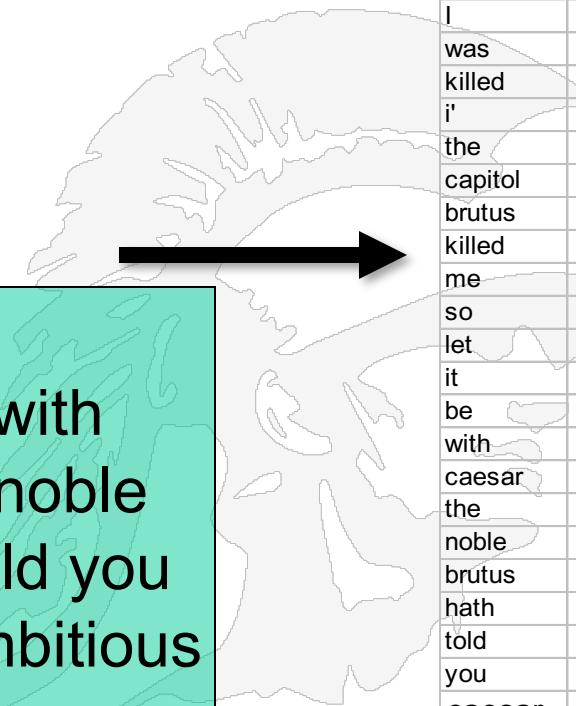
- Documents are parsed to extract words and these are saved with the document ID i.e a sequence of (possibly modified token, Document ID) pairs

Doc 1

I did enact Julius
Caesar I was killed
i' the Capitol;
Brutus killed me.

Doc 2

So let it be with
Caesar. The noble
Brutus hath told you
Caesar was ambitious

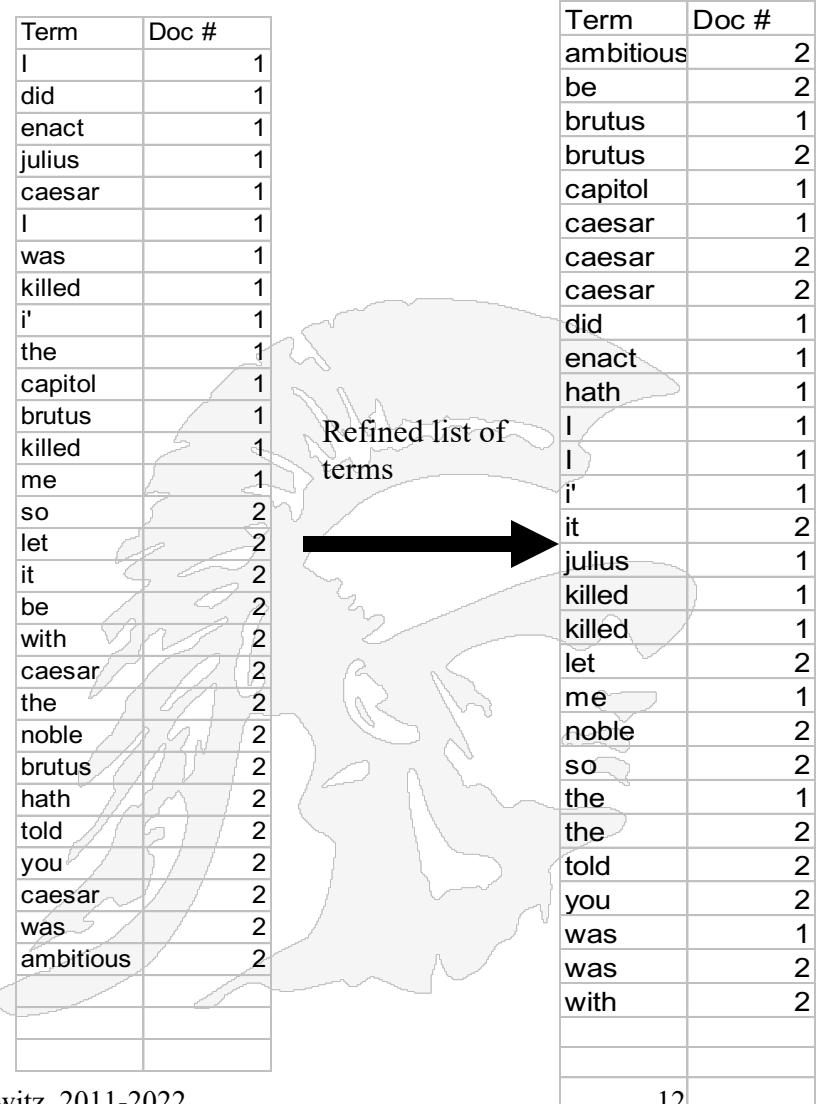


Term	Doc #
I	1
did	1
enact	1
julius	1
caesar	1
I	1
was	1
killed	1
i'	1
the	1
capitol	1
brutus	1
killed	1
me	1
so	2
let	2
it	2
be	2
with	2
caesar	2
the	2
noble	2
brutus	2
hath	2
told	2
you	2
caesar	2
was	2
ambitious	2

- If the corpus is known in advance, then after all documents have been parsed the inverted file is sorted by terms

Initial capture of terms →

- However, on the Web, documents are constantly being added and the terms are constantly increasing

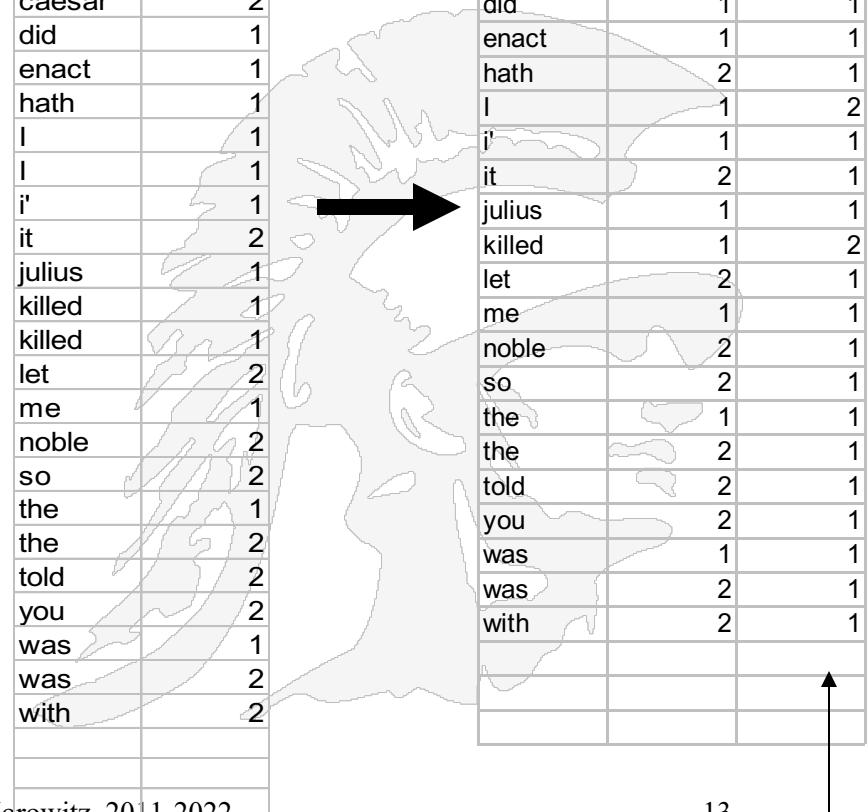


The diagram illustrates the refinement of an inverted index. On the left, an arrow points from a table labeled "Initial capture of terms" to a larger table labeled "Refined list of terms". The "Initial capture of terms" table contains all unique words from the corpus. The "Refined list of terms" table contains only the words that appear in at least two documents, as indicated by the "Doc #" column.

Term	Doc #
I	1
did	1
enact	1
julius	1
caesar	1
I	1
was	1
killed	1
i'	1
the	1
capitol	1
brutus	1
killed	1
me	1
so	2
let	2
it	2
be	2
with	2
caesar	2
the	2
noble	2
brutus	2
hath	2
told	2
you	2
caesar	2
was	2
ambitious	2

Term	Doc #
ambitious	2
be	2
brutus	1
brutus	2
capitol	1
caesar	1
caesar	2
caesar	2
did	1
enact	1
hath	1
I	1
I	1
i'	1
it	2
julius	1
killed	1
killed	1
let	2
me	1
noble	2
so	2
the	1
the	2
told	2
you	2
was	1
was	2
with	2

- Multiple term entries in a single document are merged.
- Frequency information is added.

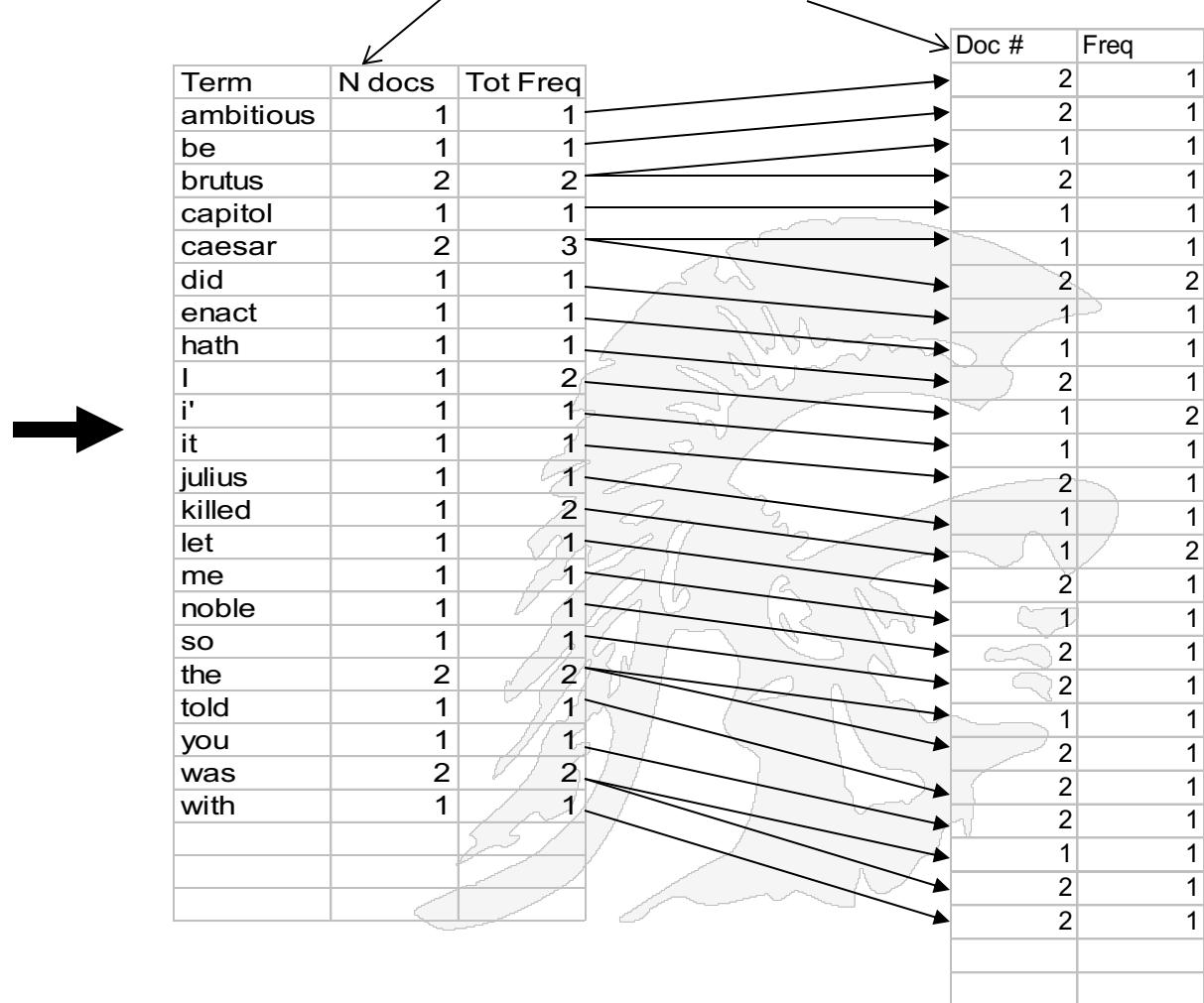


A large black arrow points from the text "Frequency information is added." in the list above to the rightmost column of the table below, which is labeled "Freq".

Term	Doc #	Freq
ambitious	2	1
be	2	1
brutus	1	1
brutus	2	1
capitol	1	1
caesar	1	1
caesar	2	2
caesar	2	2
did	1	1
enact	1	1
hath	1	1
I	1	2
I	1	1
i'	1	1
it	2	1
julius	1	1
killed	1	2
killed	1	1
let	2	1
me	1	1
noble	2	1
so	2	1
the	1	1
the	2	1
told	2	1
you	2	1
was	1	1
was	2	1
with	2	1

e.g. Caesar Occurs in Documents 1 and 2, With Total Frequency 3

- The file is commonly split into a *Dictionary* and a *Postings* file

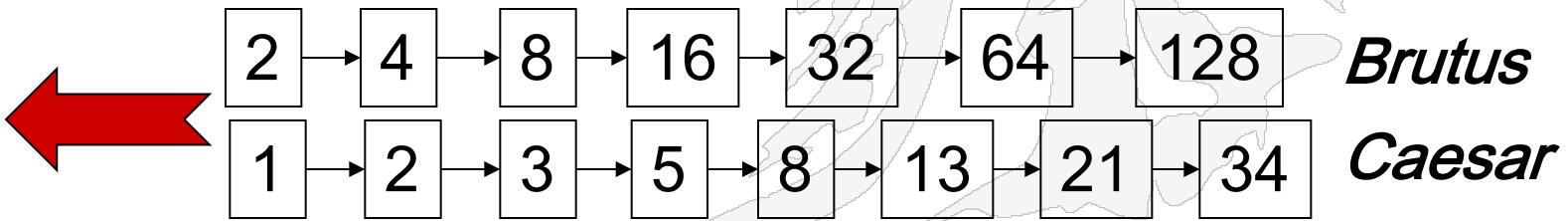


Query Processing Across the Postings List

- Consider processing the query:

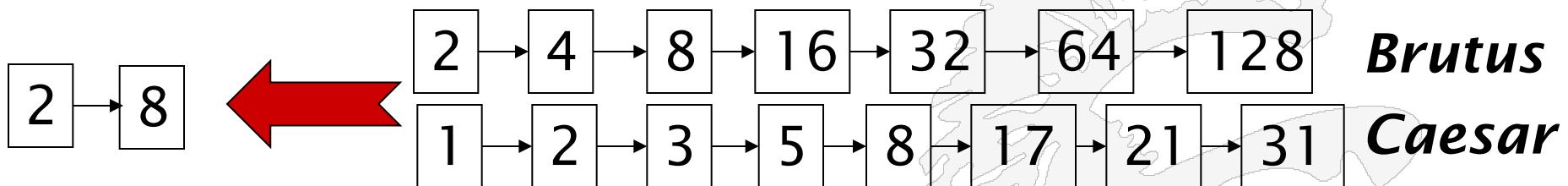
Brutus AND Caesar

- Locate *Brutus* in the Dictionary;
 - Retrieve its postings.
- Locate *Caesar* in the Dictionary;
 - Retrieve its postings.
- “Merge” the two postings and select the ones in common (postings are document ids):



Basic Merge

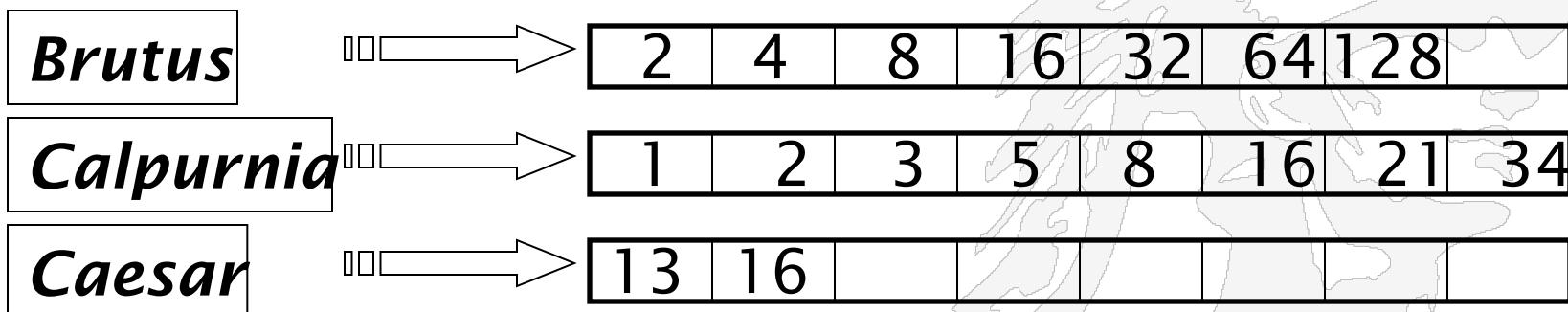
- Walk through the two postings simultaneously, in time linear in the total number of postings entries



If the list lengths are m and n , the merge takes $O(m+n)$ operations.

Query Optimization

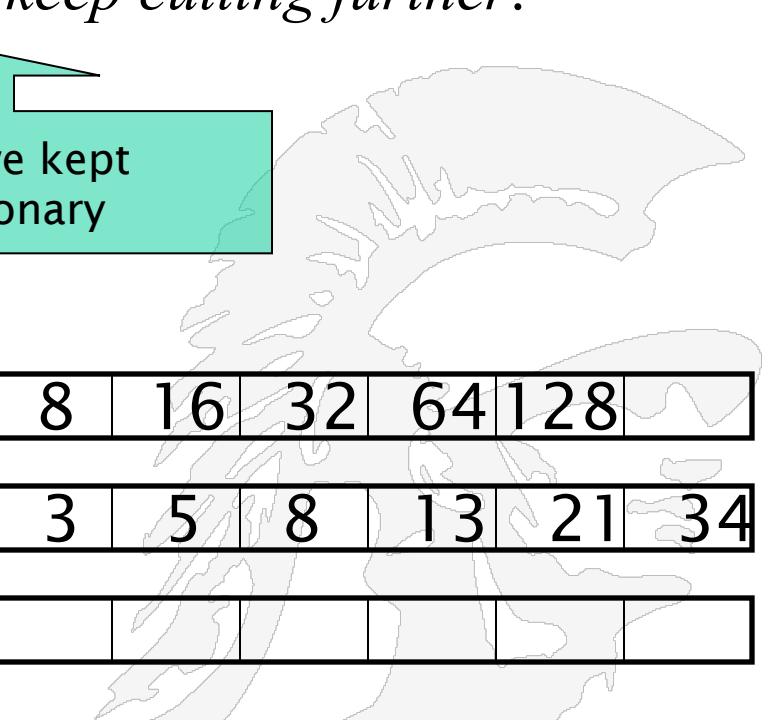
- What is the best order for query processing?
- Consider a query that is an *AND* of t terms.
- For each of the t terms, get its postings, then *AND* together.



Query: *Brutus AND Calpurnia AND Caesar*

Query Optimization Example

- Process in order of increasing frequency of occurrence:
 - *start with smallest set, then keep cutting further.*



This is why we kept
freq in dictionary

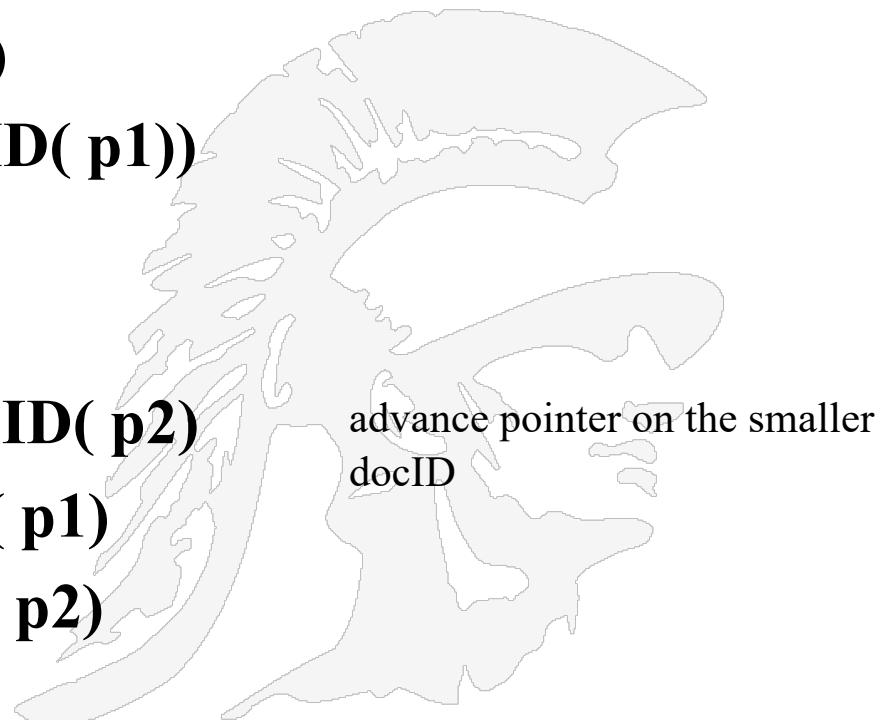
Brutus	↔	2 4 8 16 32 64 128
Calpurnia	↔	1 2 3 5 8 13 21 34
Caesar	↔	13 16

Execute the query as (*Caesar AND Brutus*) AND *Calpurnia*.

Algorithm for the intersection of two postings lists p1 and p2

INTERSECT(p1, p2)

```
1 answer ← ()  
2 while p1 ≠ NIL and p2 ≠ NIL  
3 do if docID( p1) = docID( p2)  
4     then ADD(answer, docID( p1))  
5             p1 ← next( p1)  
6             p2 ← next( p2)  
7     else if docID( p1) < docID( p2)  
8         then p1 ← next( p1)  
9     else p2 ← next( p2)  
10    return answer
```

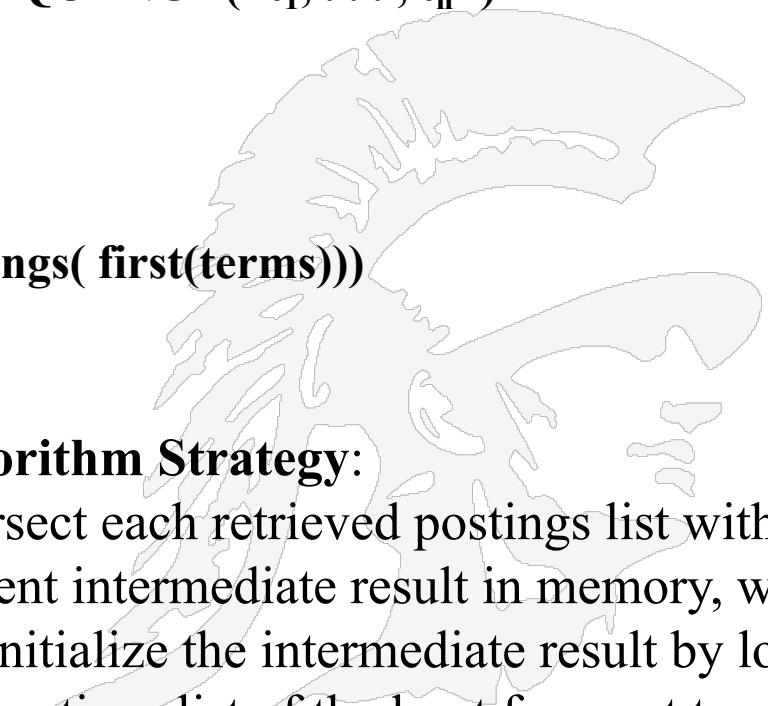


Algorithm for conjunctive queries that returns the set of documents containing each term in the input list of terms

INTERSECT(< t_1, \dots, t_n >)

- 1 terms \leftarrow **SORTBYINCREASINGFREQUENCY(< t_1, \dots, t_n >)**
- 2 result \leftarrow **postings(first(terms))**
- 3 terms \leftarrow **rest(terms)**
- 4 while terms \neq **NIL** and result \neq **NIL**
- 5 do result \leftarrow **INTERSECT(result, postings(first(terms)))**
- 6 terms \leftarrow **rest(terms)**
- 7 return result

Algorithm Strategy:



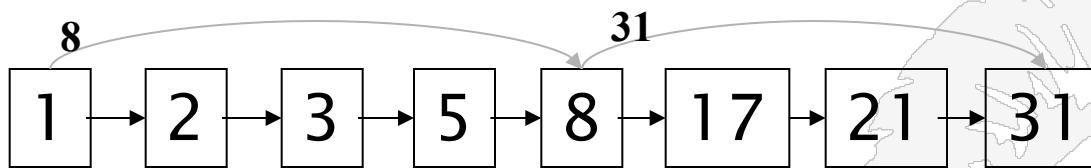
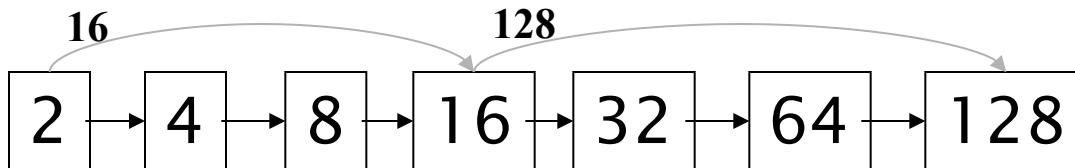
intersect each retrieved postings list with the current intermediate result in memory, where we initialize the intermediate result by loading the postings list of the least frequent term

To speed up the merging of postings we
use the technique of *Skip Pointers*

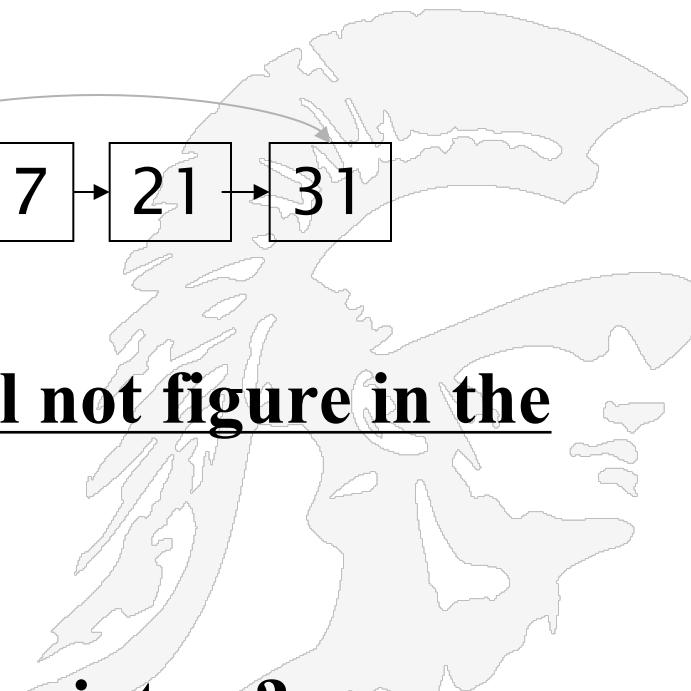


The Technique of Skip Pointers

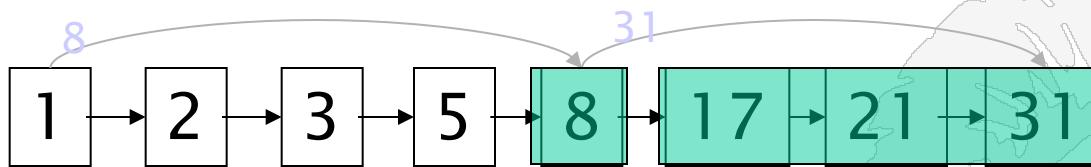
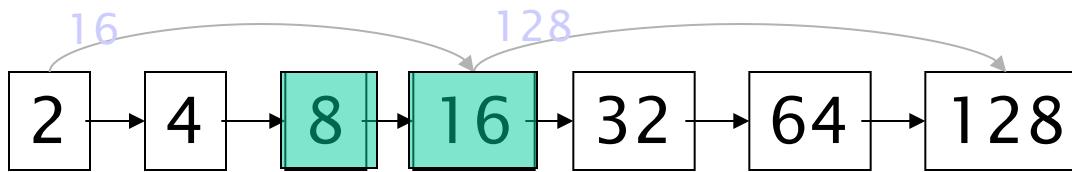
Augment postings with skip pointers (at indexing time)



- Why?
- To skip postings that will not figure in the search results.
- How?
- Where do we place skip pointers?



Query Processing With Skip Pointers



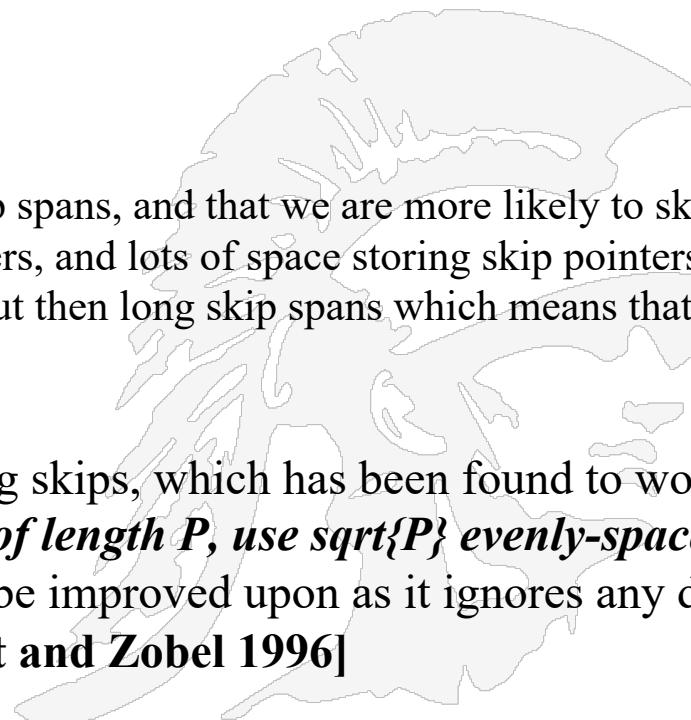
Suppose we've stepped through the lists until we process 8 on each list.

When we get to 16 on the top list, we see that its successor is 32.

But the skip successor of 8 on the lower list is 31, so we can skip ahead past the intervening postings 17 and 21.

Facts on Skip Pointers

- **Skip pointers are added at indexing time**; they are shortcuts, and they only help for AND queries and they are useful when the corpus is relatively static
- there are two questions that must be answered:
 - 1. where should they be placed?
 - 2. how do the algorithms change?
- **The Argument:** More skips means shorter skip spans, and that we are more likely to skip. But it also means lots of comparisons to skip pointers, and lots of space storing skip pointers. Fewer skips means few pointer comparisons, but then long skip spans which means that there will be fewer opportunities to skip.
- **The Solution:** A simple heuristic for placing skips, which has been found to work well in practice, is that *for a postings list of length P , use \sqrt{P} evenly-spaced skip pointers*. This heuristic possibly can be improved upon as it ignores any details of the distribution of query terms. **[Moffat and Zobel 1996]**
- See the YouTube video <http://www.youtube.com/watch?v=tPsCQOsA7j0> (15 min)



```

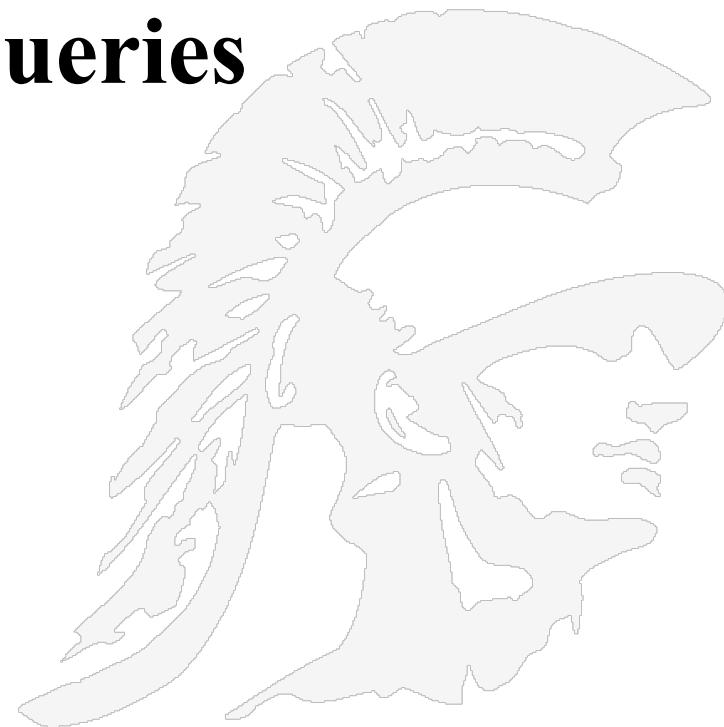
INTERSECTWITHSKIP(p1, p2)
1  answer ← ⟨ ⟩
2  while p1 ≠ NIL and p2 ≠ NIL
3  do if docID(p1) = docID(p2)
4    then ADD(answer, docID(p1))
5    p1 ← next(p1)
6    p2 ← next(p2)
7  else if docID(p1) < docID(p2)
8    then if hasSkip(p1) and (docID(skip(p1)) ≤ docID(p2))
9      then while hasSkip(p1) and (docID(skip(p1)) ≤ docID(p2))
10     do p1 ← skip(p1)
11     else p1 ← next(p1)
12   else if hasSkip(p2) and (docID(skip(p2)) ≤ docID(p1))
13     then while hasSkip(p2) and (docID(skip(p2)) ≤ docID(p1))
14     do p2 ← skip(p2)
15     else p2 ← next(p2)
16 return answer

```

Faster Postings List Intersection via Skip Pointers

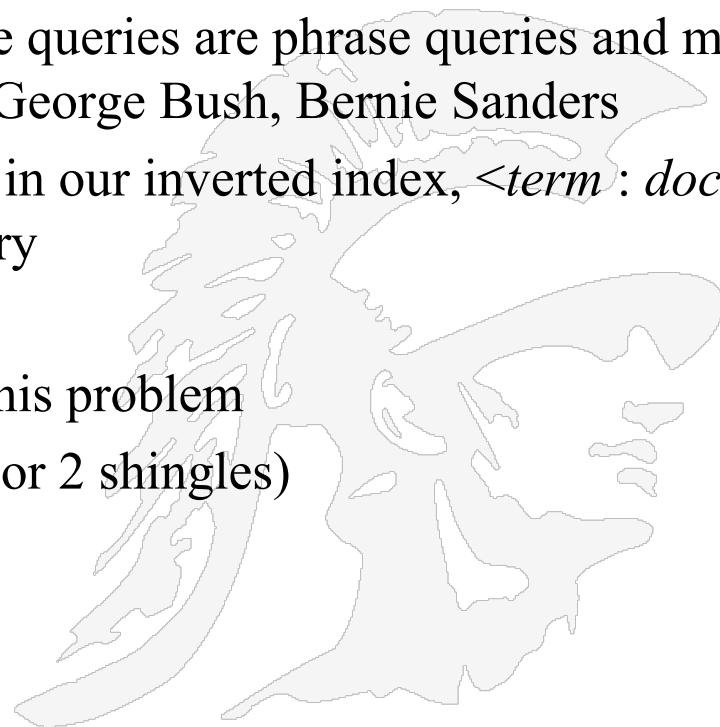
- Skip pointers will only be available for the original postings lists.
- For an intermediate result in a complex query, the call hasSkip(p) will always return false.
- Finally, note that the presence of skip pointers only helps for AND queries, not for OR queries.

Phrase Queries



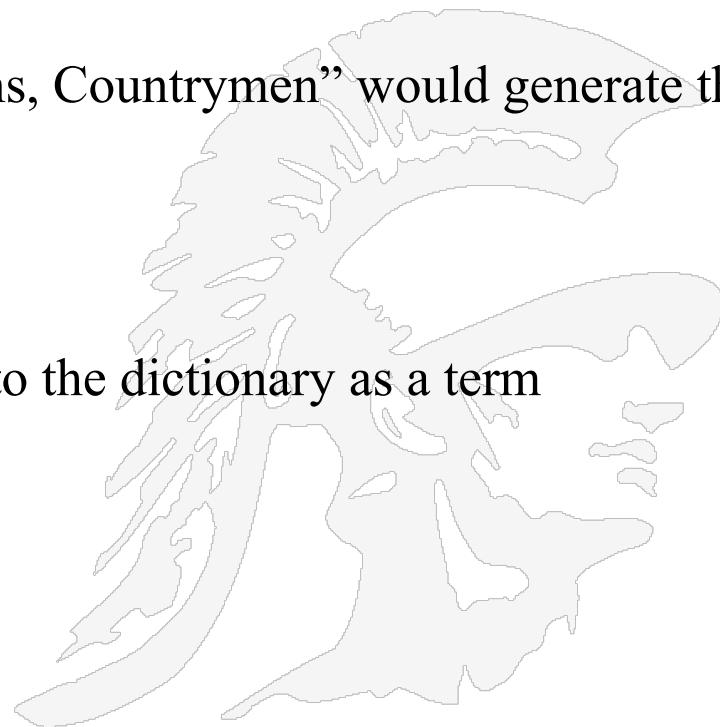
Phrase queries

- We want to answer queries such as *stanford university* – as a phrase
 - Many search engines allow users to specify a phrase using double quotes, "Stanford University", which most people find easy to understand
 - As many as 10% of web search engine queries are phrase queries and many more are implied phrase queries, e.g. George Bush, Bernie Sanders
 - If we only store terms and documents in our inverted index, $\langle term : docs \rangle$ then how we can answer a phrase query
-
- There are two approaches to solving this problem
 1. Bi-word indexes (also called 2-grams or 2 shingles)
 2. Positional indexes



Using Biword Indexes for Phrase Searching

- Definition: A *bi-word* (or a 2-gram) is a consecutive pair of terms in some text
- To improve phrase searching one approach is to index every bi-word in the text
- For example the text “Friends, Romans, Countrymen” would generate the bi-words
 - *friends romans*
 - *romans countrymen*
- Each of these bi-words is now added to the dictionary as a term



Handling Longer Phrase Queries

- Consequences
 - Bi-words will cause an explosion in the vocabulary database
 - Queries longer than 2 words will have to be broken into bi-word segments
- Example: suppose the query is the 4 word phrase

stanford university palo alto

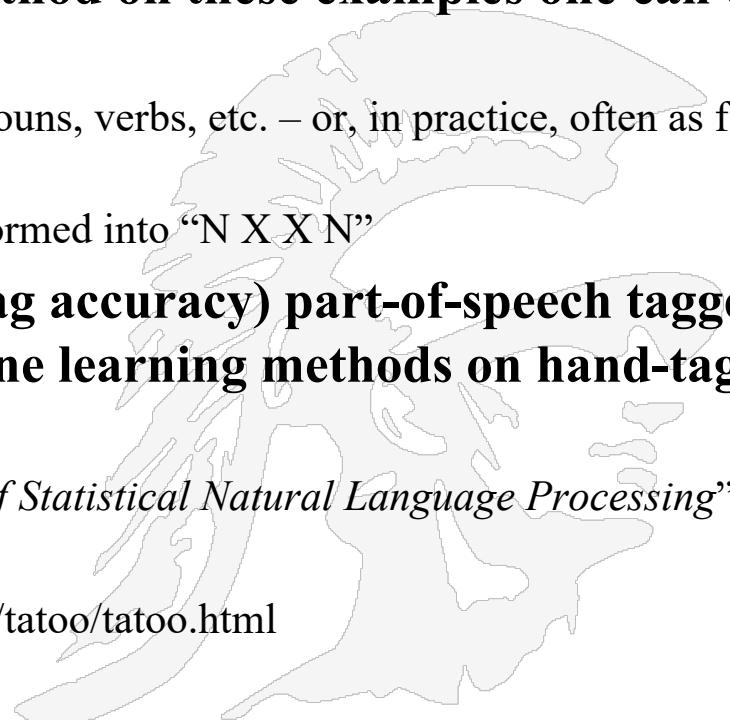
The query can be broken into the Boolean query on bi-words:

stanford university AND university palo AND palo alto

- Matching the query to terms in the index will work, but may also produce false positives (i.e. occurrences of the bi-words, but not the full 4 word query)
- A Bi-word index that is extended to longer sequences, or even variable length sequences is called a *phrase index*

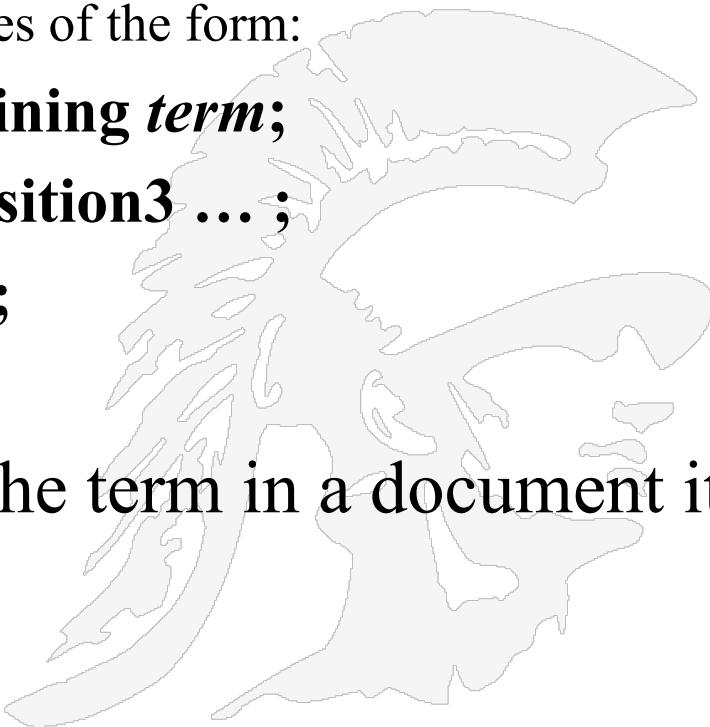
Part-of-Speech Tagging

- **Many two word phrases have embedded stop words, e.g.**
 - the abolition of slavery
 - negotiation of the constitution
- **To salvage the bi-word indexing method on these examples one can use part-of-speech tagging**
 - Part-of-speech taggers classify words as nouns, verbs, etc. – or, in practice, often as finer grained classes like “plural proper noun”.
 - “Negotiation of the constitution” is transformed into “N X X N”
- **Many fairly accurate (c. 96% per-tag accuracy) part-of-speech taggers now exist, usually trained by machine learning methods on hand-tagged text**
 - See Manning and Schutze “*Foundations of Statistical Natural Language Processing*”
 - <https://nlp.stanford.edu/fsnlp/>
 - <https://www.issco.unige.ch/en/staff/robert/tatoo/tatoo.html>



Alternate Solution - Using Positional Indexes

- Given the limitations of a bi-word index, (i.e. the enormous growth in the vocabulary) the alternate solution is most commonly used, called a **Positional Index**
- Store, for each *term* in the index, entries of the form:
<term, number of docs containing term;
***doc1:* position1, position2, position3 ... ;**
***doc2:* position1, position2 ... ;**
etc.>
- i.e. for each occurrence of the term in a document its position is stored



Positional Index Example

for each term in the vocabulary, we store postings of the form

docID: position1, position2, ...,

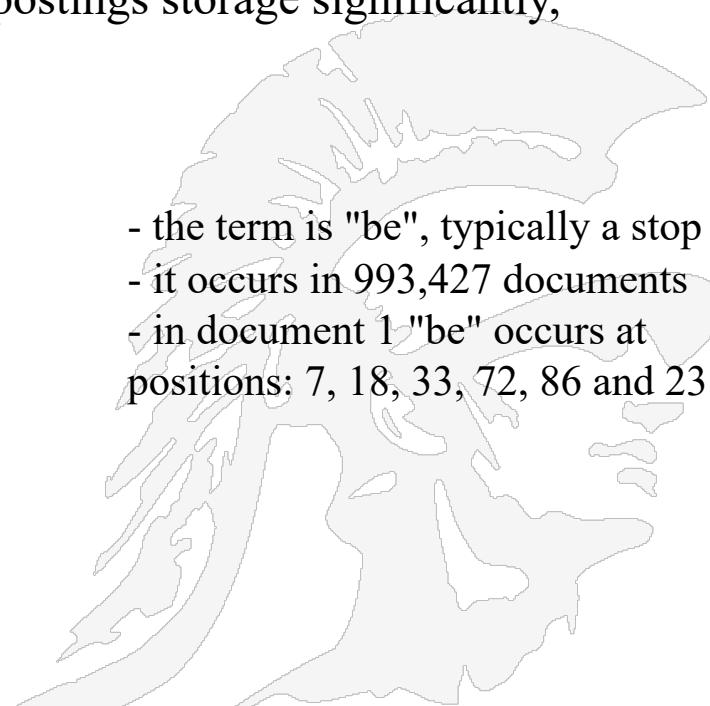
where each position is a token index in the document.

Each posting will also usually record the term frequency

Adopting a positional index expands required postings storage significantly,
even if we compress position values/offsets

<**be**: 993427;
1: 7, 18, 33, 72, 86, 231;
2: 3, 149;
4: 17, 191, 291, 430, 434;
5: 363, 367, ...>

Lots of documents
Lots of occurrences

- 
- the term is "be", typically a stop word
 - it occurs in 993,427 documents
 - in document 1 "be" occurs at positions: 7, 18, 33, 72, 86 and 231

- this scheme expands postings storage *substantially (rather than the vocabulary)*

Processing a Phrase Query

- Algorithm for matching a phrase query:
 1. Extract inverted index entries for each distinct term: e.g. *to*, *be*, *or*, *not*, *to*, *be*
 2. Merge their *doc:position* lists to enumerate all positions with “*to be or not to be*”.
 3. Match those documents that contain the terms in the adjacent positions
 - ***to*:**
 - 2:1,17,74,222,551; 4:8,16,190,429,433; 7:13,23,191; ...
 - ***be*:**
 - 1:17,19; 4:17,191,291,430,434; 5:14,19,101; ...
 - Same general method for proximity searches
 - In document 4 the word “*to*” appears in position 16 and the word “*be*” appears in position 17, so they are adjacent

Algorithm for Proximity Queries with k words

2 The term vocabulary and postings lists

```

POSITIONALINTERSECT( $p_1, p_2, k$ )
1 answer  $\leftarrow \langle \rangle$ 
2 while  $p_1 \neq \text{NIL}$  and  $p_2 \neq \text{NIL}$ 
3 do if  $\text{docID}(p_1) = \text{docID}(p_2)$ 
4   then  $l \leftarrow \langle \rangle$ 
5      $pp_1 \leftarrow \text{positions}(p_1)$ 
6      $pp_2 \leftarrow \text{positions}(p_2)$ 
7     while  $pp_1 \neq \text{NIL}$ 
8       do while  $pp_2 \neq \text{NIL}$ 
9         do if  $|\text{pos}(pp_1) - \text{pos}(pp_2)| \leq k$ 
10        then ADD( $l, \text{pos}(pp_2)$ )
11        else if  $\text{pos}(pp_2) > \text{pos}(pp_1)$ 
12          then break
13           $pp_2 \leftarrow \text{next}(pp_2)$ 
14        while  $l \neq \langle \rangle$  and  $|l[0] - \text{pos}(pp_1)| > k$ 
15          do DELETE( $l[0]$ )
16          for each  $ps \in l$ 
17            do ADD(answer,  $\langle \text{docID}(p_1), \text{pos}(pp_1), ps \rangle$ )
18             $pp_1 \leftarrow \text{next}(pp_1)$ 
19             $p_1 \leftarrow \text{next}(p_1)$ 
20             $p_2 \leftarrow \text{next}(p_2)$ 
21        else if  $\text{docID}(p_1) < \text{docID}(p_2)$ 
22          then  $p_1 \leftarrow \text{next}(p_1)$ 
23        else  $p_2 \leftarrow \text{next}(p_2)$ 
24 return answer

```

► **Figure 2.12** An algorithm for proximity intersection of postings lists p_1 and p_2 . The algorithm finds places where the two terms appear within k words of each other and returns a list of triples giving docID and the term position in p_1 and p_2 .

The algorithm finds places where the two terms appear within k words of each other and returns a list of triples giving docID and the term position in p_1 and p_2 .



Some High Frequency Noun Phrases from TREC and Patent DataSets

TREC

<i>Frequency</i>	<i>Phrase</i>
65824	United States
61327	article type
33864	Los Angeles
18062	Hong Kong
17788	North Korea
17308	New York
15513	San Diego
15009	Orange county

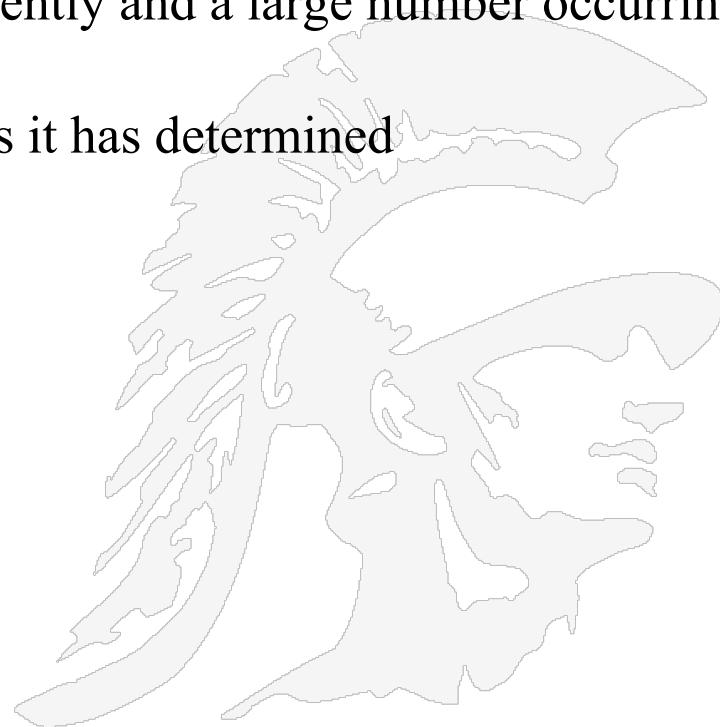
Patent

<i>Frequency</i>	<i>Phrase</i>
975362	present invention
191625	u.s. pat
147352	preferred embodiment
95097	carbon atoms
87903	group consisting
81809	room temperature
78458	seq id
75850	brief description

The phrases above were identified by POS tagging; The data above shows that common phrases are used more frequently in patent data as patents have a very formal style; many of the TREC phrases are proper nouns, whereas patent phrases are those that occur in all patents

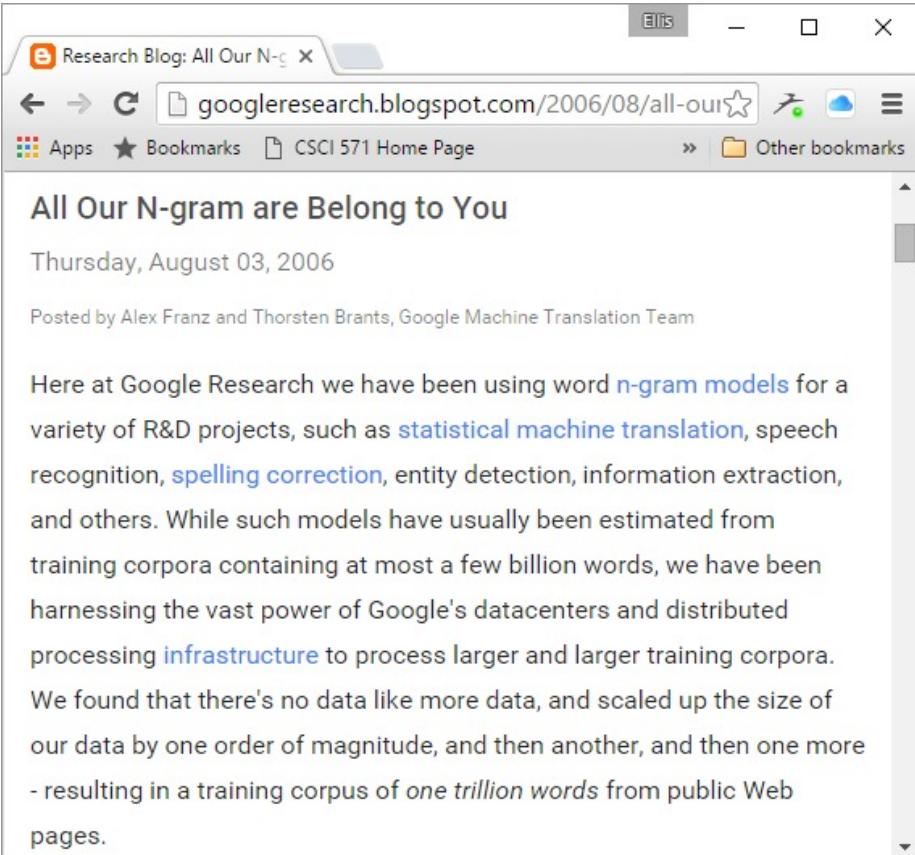
Google Relies Upon *n*-gram Indexes

- Google has investigated the use of n-grams stored in its index, $n \geq 2$;
- N-grams of all lengths form a **Zipf distribution (power law)** with a few common phrases occurring very frequently and a large number occurring with frequency 1
- Google has released the set of n-grams it has determined



Google's N-Gram Database Facts

- Google made available a file of n-grams derived from the web pages it indexed
- <http://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html>
- Statistics for the Google n-gram sample
- Number of tokens 1,024,908,267,229
(1 trillion, 24 billion, 908 million, . . .)
- Number of sentences 95,119,665,584
- Number of unigrams 13,588,391
- Number of bigrams 314,843,401
- Number of trigrams 977,069,902
- Number of four grams 1,313,818,354
- Number of five grams 1,176,470,663



The screenshot shows a web browser window with the title bar "Research Blog: All Our N-gram are Belong to You". The address bar contains the URL "googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html". The page content starts with the title "All Our N-gram are Belong to You" and the date "Thursday, August 03, 2006". It is posted by "Alex Franz and Thorsten Brants, Google Machine Translation Team". The text discusses the use of n-gram models for various R&D projects like machine translation, speech recognition, spelling correction, entity detection, and information extraction. It mentions scaling up the data size by orders of magnitude and processing infrastructure to handle one trillion words from the Web.

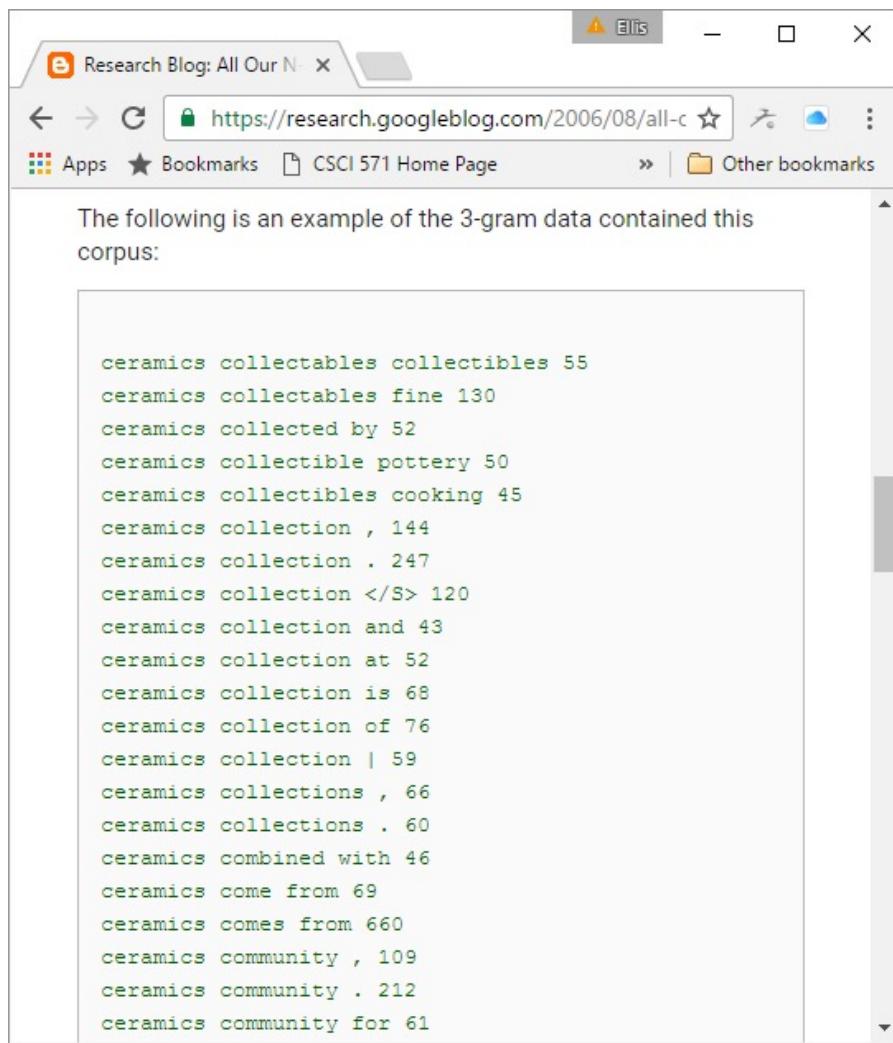
All Our N-gram are Belong to You

Thursday, August 03, 2006

Posted by Alex Franz and Thorsten Brants, Google Machine Translation Team

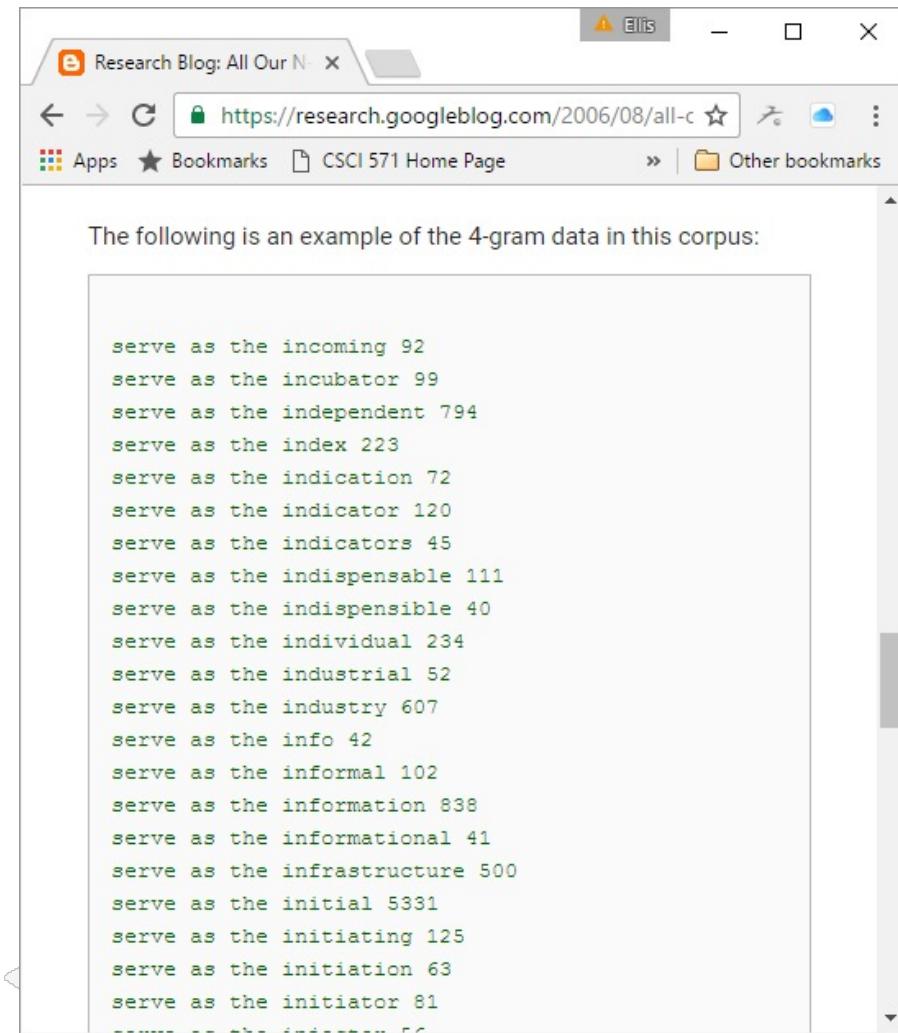
Here at Google Research we have been using word [n-gram models](#) for a variety of R&D projects, such as [statistical machine translation](#), speech recognition, [spelling correction](#), entity detection, information extraction, and others. While such models have usually been estimated from training corpora containing at most a few billion words, we have been harnessing the vast power of Google's datacenters and distributed processing [infrastructure](#) to process larger and larger training corpora. We found that there's no data like more data, and scaled up the size of our data by one order of magnitude, and then another, and then one more - resulting in a training corpus of [one trillion words](#) from public Web pages.

Examples of 3-Gram, 4-Gram Data



The following is an example of the 3-gram data contained this corpus:

```
ceramics collectables collectibles 55
ceramics collectables fine 130
ceramics collected by 52
ceramics collectible pottery 50
ceramics collectibles cooking 45
ceramics collection , 144
ceramics collection . 247
ceramics collection </S> 120
ceramics collection and 43
ceramics collection at 52
ceramics collection is 68
ceramics collection of 76
ceramics collection | 59
ceramics collections , 66
ceramics collections . 60
ceramics combined with 46
ceramics come from 69
ceramics comes from 660
ceramics community , 109
ceramics community . 212
ceramics community for 61
```



The following is an example of the 4-gram data in this corpus:

```
serve as the incoming 92
serve as the incubator 99
serve as the independent 794
serve as the index 223
serve as the indication 72
serve as the indicator 120
serve as the indicators 45
serve as the indispensable 111
serve as the indispensable 40
serve as the individual 234
serve as the industrial 52
serve as the industry 607
serve as the info 42
serve as the informal 102
serve as the information 838
serve as the informational 41
serve as the infrastructure 500
serve as the initial 5331
serve as the initiating 125
serve as the initiation 63
serve as the initiator 81
----- -- --- ----- --
```

N-Grams Used To Find Best and Worst Search Queries

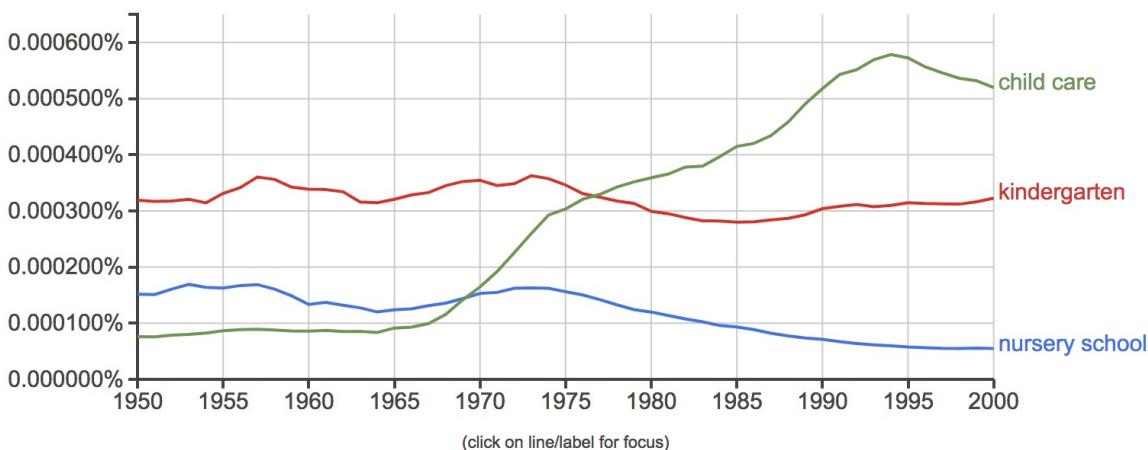
- Lacrosse occurs in 46,579 search terms on Google;**
- Overall it produced \$843,708 revenue by spending \$152,575**
- “Lacrosse equipment” was the most profitable, while “girls lacrosse cleats the least**

Unigram	Count	Clicks	Cost	Transactions	Revenue	ROAS
lacrosse	46,579	200,727	\$152,575	6,614	\$843,708	553%

Search Queries Containing "Lacrosse" (Sample)					
Search Query	Impressions	Clicks	Costs	Order	Revenue
lacrosse equipment	13548	637	\$493.04	31	\$5,306.48
lacrosse gloves	28552	631	\$319.14	24	\$2,182.20
lacrosse gear	18234	80	\$686.99	23	\$3,529.44
lacrosse cleats	21965	958	\$682.38	18	\$2,200.82
box lacrosse bicep pads	275	27	\$17.82	10	\$521.96
lacrosse sticks	46274	553	\$422.03	7	\$890.46
lacrosse elbow pads	10299	124	\$111.54	6	\$303.60
youth lacrosse gloves	2330	105	\$73.57	6	\$308.53
womens lacrosse cleats	4982	184	\$148.76	6	\$1,066.30
women's lacrosse cleats	4251	178	\$138.19	5	\$563.62
lacrosse heads	7334	151	\$111.79	5	\$550.05
lacrosse bags	8584	157	\$80.80	5	\$337.56
youth lacrosse gear	1628	122	\$109.82	5	\$1,369.11
discount lacrosse gear	191	39	\$18.57	5	\$1,078.36
lacrosse socks	3190	89	\$48.09	5	\$506.72
lacrosse gear bag	1179	40	\$23.13	4	\$410.71
discount lacrosse heads	19	6	\$3.77	4	\$272.78
girls lacrosse cleats	4564	114	\$77.77	4	\$354.86

Google's N-Gram Viewer

- The **Google N-Gram Viewer** or **Google Books N-Gram Viewer** is an online search engine that charts frequencies of any set of comma-delimited search strings using a yearly count of n-grams found in sources printed between 1500 and 2008 in **Google's** text corpora in English, Chinese (simplified), French, German, Hebrew, Italian, ...
 - <https://books.google.com/ngrams/>, or for examples see books.google.com/ngrams/info
- The program can search for a single word or a phrase, including misspellings
- The n-grams are matched with the text within the selected corpus, and, if found in 40 or more books, are then plotted on a graph
- The data used for the search are composed of total _ counts, 1-grams, 2-grams, 3-grams, 4-grams, and 5-grams files for each language



This shows trends in three n-grams from 1950 to 2000:
 "nursery school" (a 2-gram or bigram),
 "kindergarten" (a 1-gram or unigram),
 and
 "child care" (another bigram)

Comparing N-Grams Across Languages

- S. Yang et al, N-gram statistics in English and Chinese: Similarities and differences, ICSC, 2007, Int'l Conf. on semantic computing, 454-460
- http://static.googleusercontent.com/external_content/untrusted_dlcp/research.google.com/en/us/pubs/archive/33035.pdf
- The authors above analyzed 200 million randomly sampled English and Chinese Web pages and concluded:
 1. The distribution of the unique number of n-grams is similar between English and Chinese, though the Chinese distribution is shifted to larger N
 2. The distribution indicates that on average 1.5 Chinese characters correspond to 1 English word
 3. While frequency distributions of uni-grams and bi-grams are very different, the frequency distribution for 3-grams and 4-grams are strikingly similar

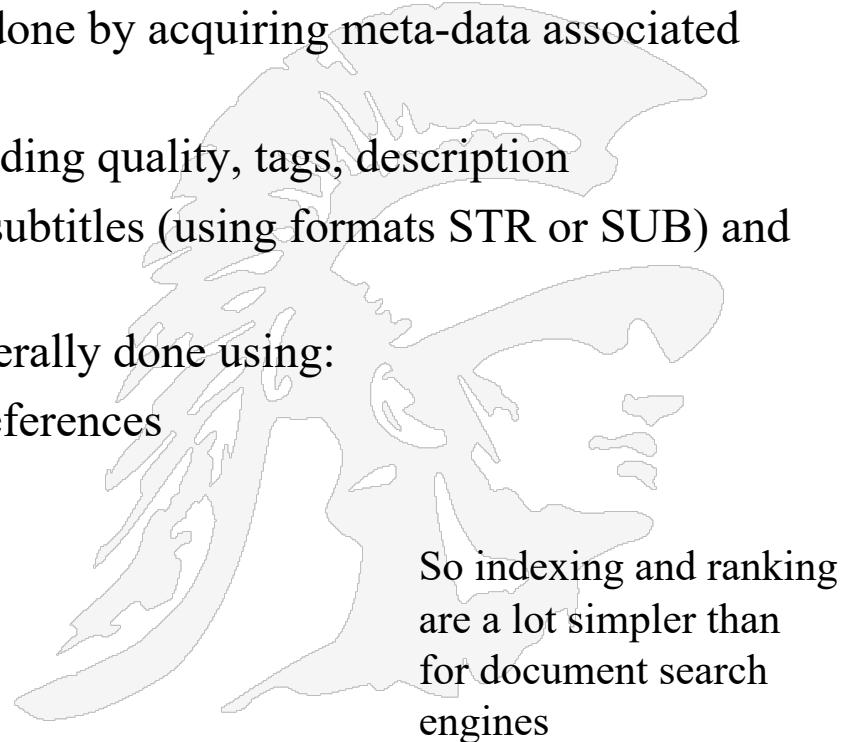


Video Search Engines YouTube et al



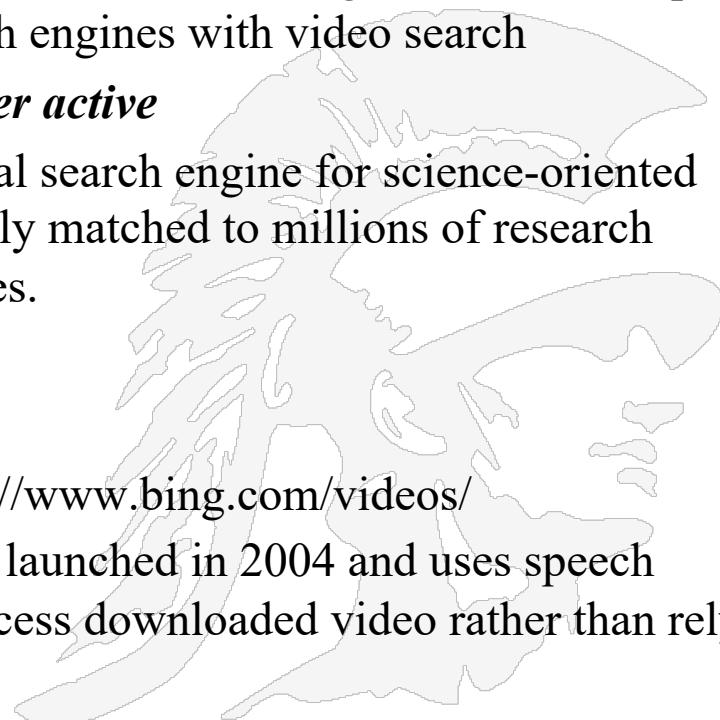
Video Search Engines – Quick Summary

- A **video search engine** is a web-based search engine which crawls the web primarily for video content.
 - YouTube is not strictly a video search engine as it does not crawl the web looking for video content
- The **indexing** of video content is normally done by acquiring meta-data associated with the video, e.g.
 - Author, title, creation date, duration, coding quality, tags, description
 - Other aspects of video recognition are subtitles (using formats STR or SUB) and transcription (using format TTXT)
- The **ranking** of videos under a query is generally done using:
 - Relevance: using metadata and user preferences
 - Ordered by date of upload
 - Ordered by number of views
 - Ordered by duration
 - Ordered by user rating



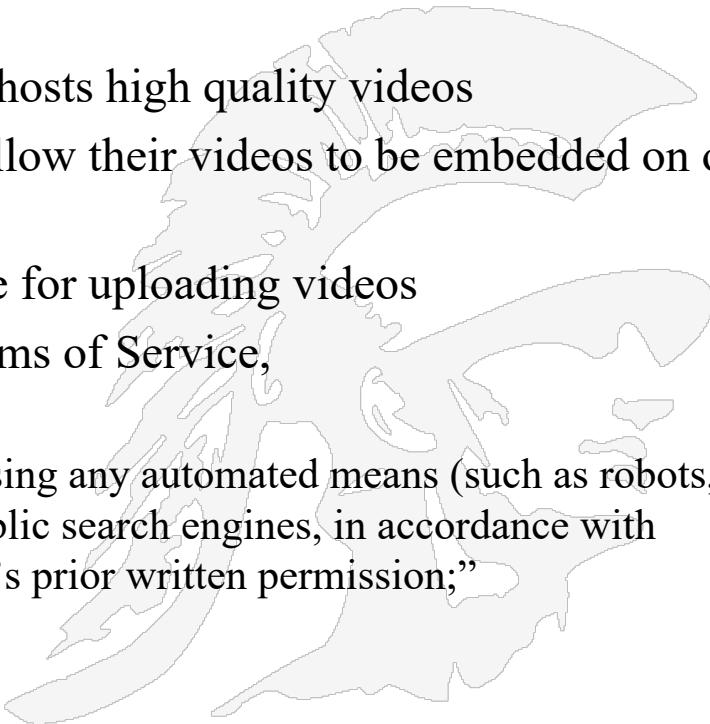
Video Search Engines That *Crawl* for Content

- *Those no longer existing*
 - *CastTV* was a Web-wide video search engine that was founded in 2006
 - *No longer active*
 - *Munax* released their first version all-content search engine in 2005 and powers both nationwide and worldwide search engines with video search
 - <http://www.munax.com/> *no longer active*
 - *ScienceStage* is an integrated universal search engine for science-oriented videos. All videos are also semantically matched to millions of research documents from open-access databases.
 - *No longer active*
- *A few remain*
 - *Bing* does crawl for videos, see <https://www.bing.com/videos/>
 - *blinkx* (renamed as RhythmOne) was launched in 2004 and uses speech recognition and visual analysis to process downloaded video rather than rely on metadata alone
 - <http://www.blinkxtv.com/> *now redirects to 360Daily.com*



Video Search Engines That Host

- Largely because of the large file sizes involved, video hosting is highly concentrated on a fairly small number of websites
 - **vimeo.com**, first to support HD video, focuses on short, arty, films
 - **vevo.com**, a joint venture of Universal Music Group, Sony Music Entertainment and Warner Music Group
 - **dailymotion.com**, owned by Vivendi, hosts high quality videos
- Most of these websites which host video allow their videos to be embedded on other websites
- **YouTube.com** has become the defacto site for uploading videos
- It is legal to crawl YouTube, see their Terms of Service,
www.youtube.com/static?template=terms
- “3. You are not allowed to access the Service using any automated means (such as robots, botnets or scrapers) except (a) in the case of public search engines, in accordance with YouTube’s robots.txt file; or (b) with YouTube’s prior written permission;”



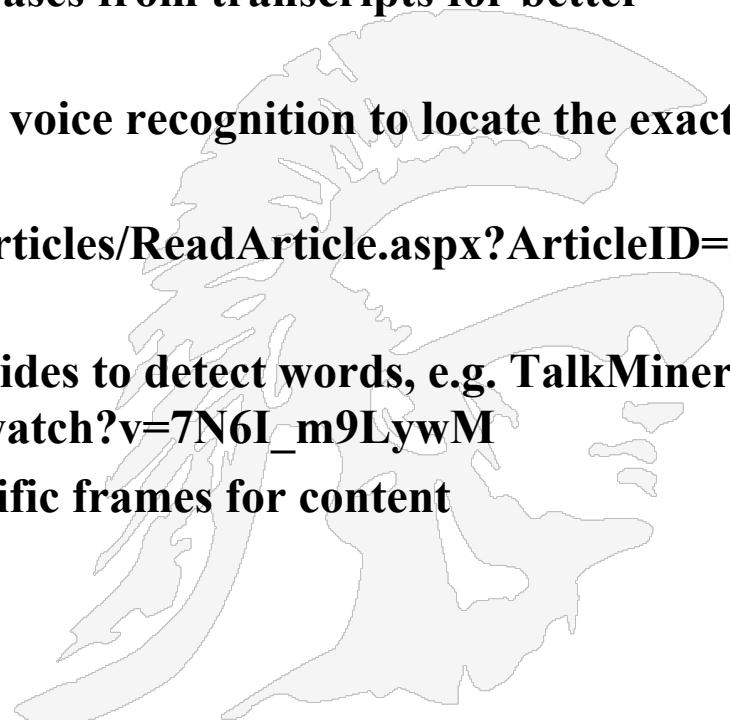
Video Search Engines That Stream Entertainment

- **Hulu** is an America subscription video on demand service jointly owned by Walt Disney, 21st Century Fox, Comcast, and Time Warner
 - In December 2017, Disney acquired Fox's partial ownership, giving it a majority stake; other owners include Comcast
 - It is primarily oriented towards instant streaming of television series', carrying current and past episodes of many series from its owners' respective television networks and other content partners
- **Netflix** is an American subscription video on demand service, that originally delivered DVDs;
 - They develop their own content as well as offering content from major film distributors
- **Amazon Prime** is an American subscription video on demand service offering television and file shows for rent or purchase
- **Disney+** a recent entry

- **Entertainment**

Some Video Technologies

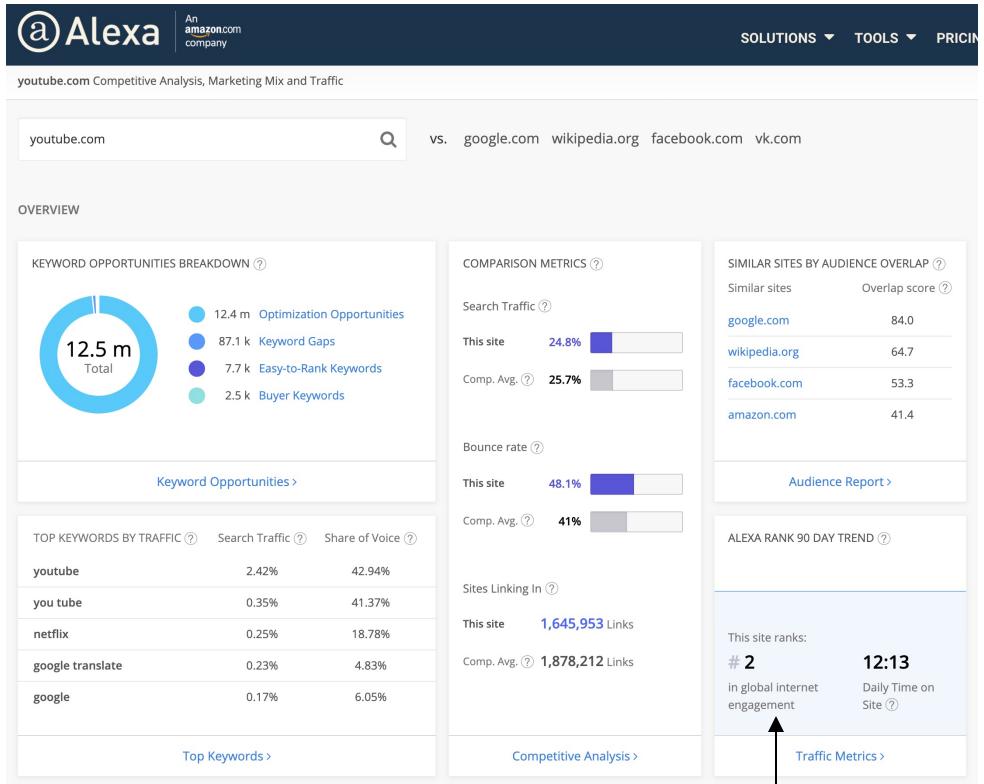
- **Subtitles:** there are two formats, one for subtitles and one for transcripts
 - SRT or SUB for subtitles
 - TTXT for transcripts
- **Speech Recognition,** used to extract phrases from transcripts for better indexing
 - Gaudi, Google Audio Indexing uses voice recognition to locate the exact spot where words are spoken
 - <https://www.speechtechmag.com/Articles/ReadArticle.aspx?ArticleID=51280>
- **Text Recognition:** uses OCR on video slides to detect words, e.g. TalkMiner System, see https://www.youtube.com/watch?v=7N6I_m9LywM
- **Frame Analysis:** tools that analyze specific frames for content



YouTube Background

- YouTube is an American video hosting website headquartered in San Bruno, California, created by three former PayPal employees: Chad Hurley, Steve Chen Jawed Karim in February 2005.
- In November 2006, it was bought by Google for US\$1.65 billion
- In 2020 Google announced that YouTube generated revenue of \$19.8 billion
- The site allows users to upload, view, rate, share, add to favorites, report and comment on videos
- In January 2022, the website was ranked as the second most popular site by Alexa Internet, a web traffic analysis company (now owned by Amazon)

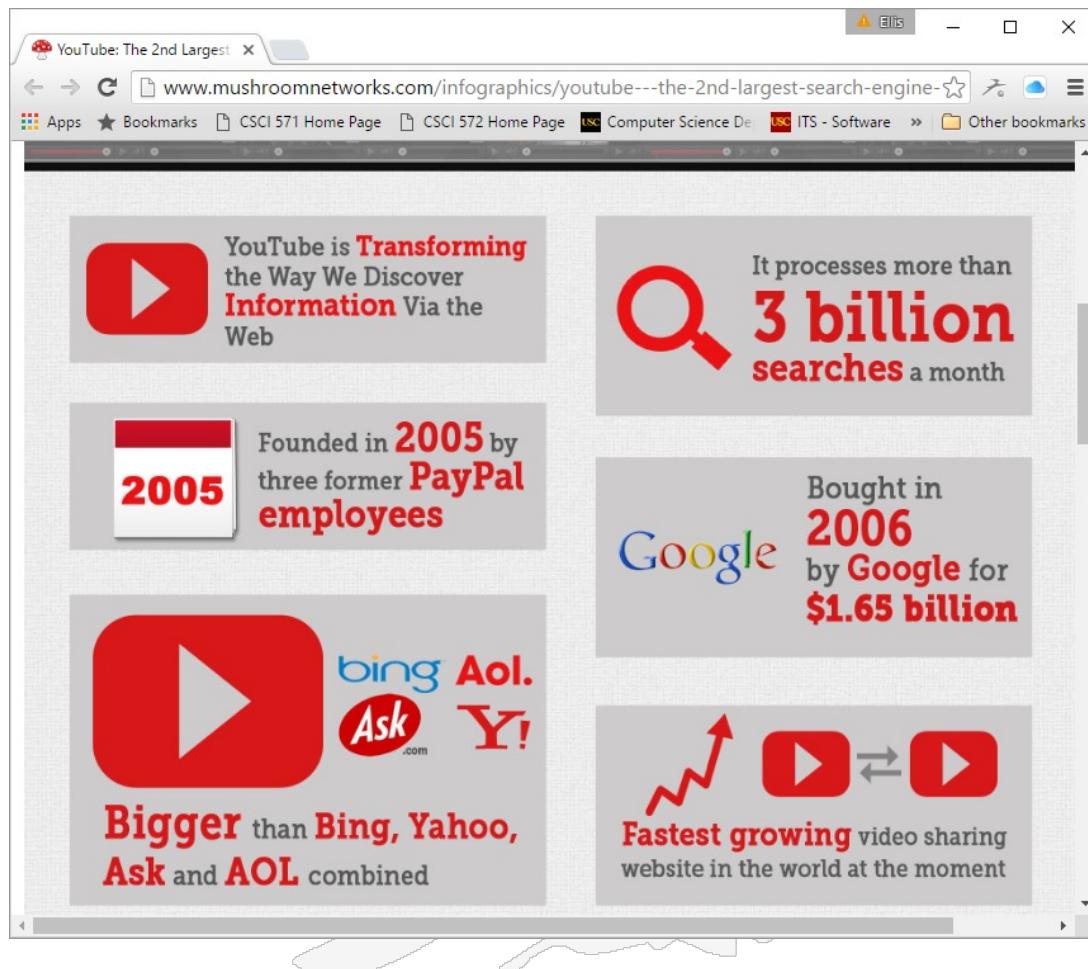
– See also
https://en.wikipedia.org/wiki/List_of_most_popular_websites



For details see Related Articles page, Mar 2020

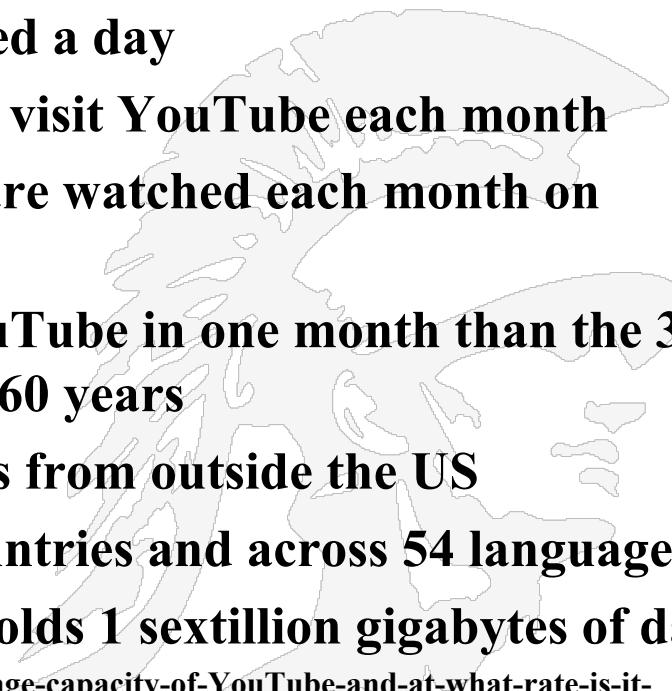
YouTube as a Search Engine

- YouTube - The 2nd Largest Search Engine (cite:Infographic)
- YouTube processes more than 3 billion searches a month.
- It's bigger than Bing, Yahoo!, Ask and AOL combined!
- <http://www.mushroomnetworks.com/infographics/youtube---the-2nd-largest-search-engine-infographic>



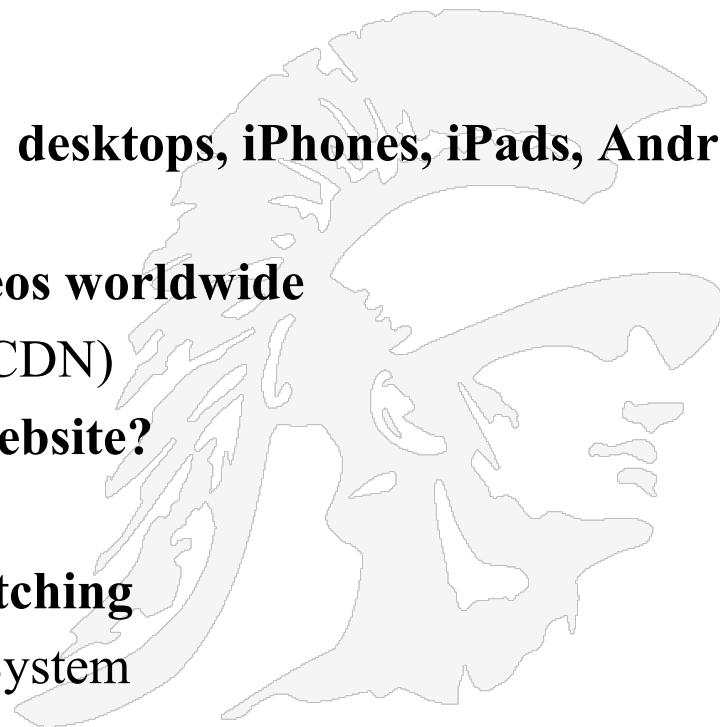
YouTube Traffic - Some Facts

- **As of 2019:**
 - **60 hours of video are uploaded every minute, or one hour of video is uploaded to YouTube every second.**
 - **Over 4 billion videos are viewed a day**
 - **Over 800 million unique users visit YouTube each month**
 - **Over 3 billion hours of video are watched each month on YouTube**
 - **More video is uploaded to YouTube in one month than the 3 major US networks created in 60 years**
 - **70% of YouTube traffic comes from outside the US**
 - **YouTube is localized in 39 countries and across 54 languages**
 - **It is estimated that YouTube holds 1 sextillion gigabytes of data**
 - <https://www.quora.com/What-is-the-total-size-storage-capacity-of-YouTube-and-at-what-rate-is-it-increasing-How-is-Google-keeping-up-with-the-increasing-demands-of-Youtube%E2%80%99s-capacity-given-that-thousands-of-videos-are-uploaded-every-day>



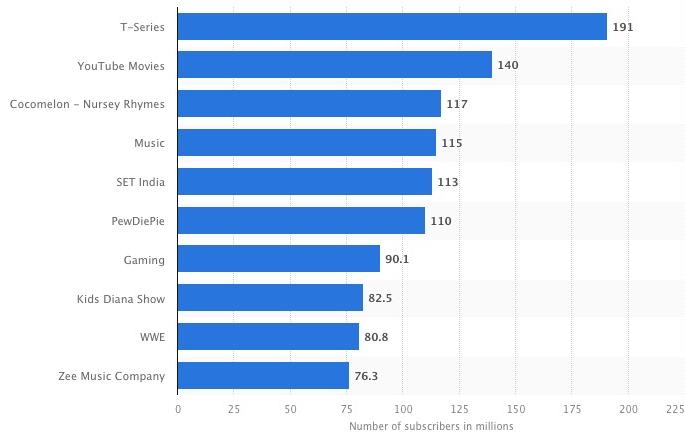
YouTube Search Engine Issues to Consider

- Since crawling, indexing and ranking are not big challenges for YouTube, what are the major hurdles
 - 1. **What video formats are acceptable**
 - For uploading
 - For downloading
 - 2. **How are videos to be displayed on: desktops, iPhones, iPads, Android devices, etc**
 - 3. **How does YouTube distribute videos worldwide**
 - A content distribution network (CDN)
 - 4. **How does YouTube monetize its website?**
 - YouTube's ContentID system
 - 5. **How does YouTube keep users watching**
 - The YouTube Recommendation System



YouTube Channels

- In order to upload a video you must be a registered user
- In addition YouTube offers a special type of account called a *channel*; channels include
 - thumbnails of videos you've uploaded,
 - members to whom you've subscribed,
 - videos from other members you've picked as favorites,
 - lists of members who are your friends,
 - your subscribers, and
- Biggest YouTube Channels as of 2021



With 1 million subscribers, a YouTuber will make between \$300,000 – \$2 million To be in the top 1000 YouTubers you must have ~1.8 million subscribers As of 09/2020, there are more than 2000 YouTubers with over a million subscribers

<https://www.statista.com/statistics/277758/most-popular-youtube-channels-ranked-by-subscribers/>

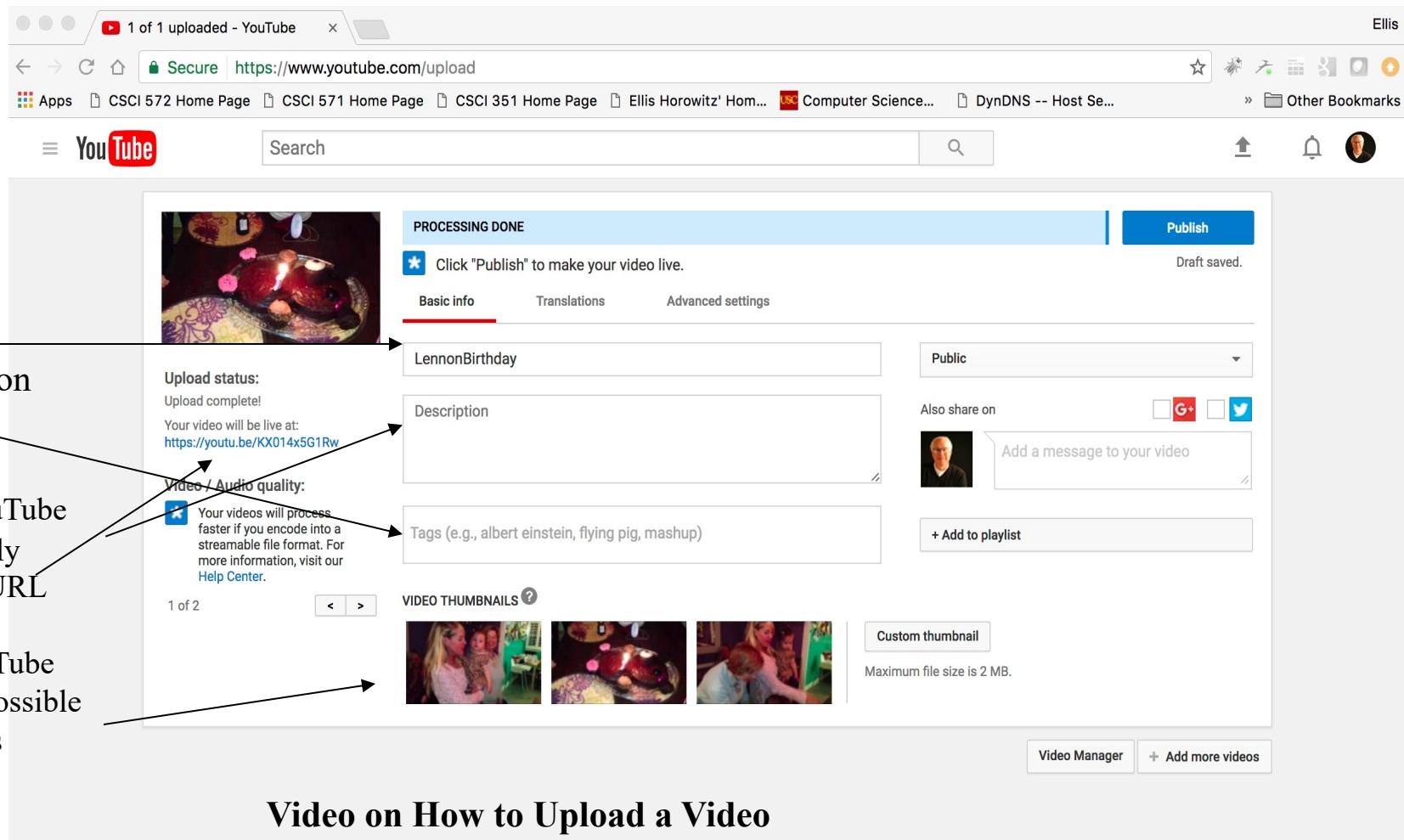
YouTube Gathers Information When Videos are Uploaded

YouTube captures:

Name _____
Description _____
Tags _____

Note: YouTube immediately assigns a URL

Note: YouTube suggests possible thumbnails



The screenshot shows the YouTube upload interface. At the top, it says "1 of 1 uploaded - YouTube". The URL in the address bar is <https://www.youtube.com/upload>. Below the address bar are various browser tabs and icons.

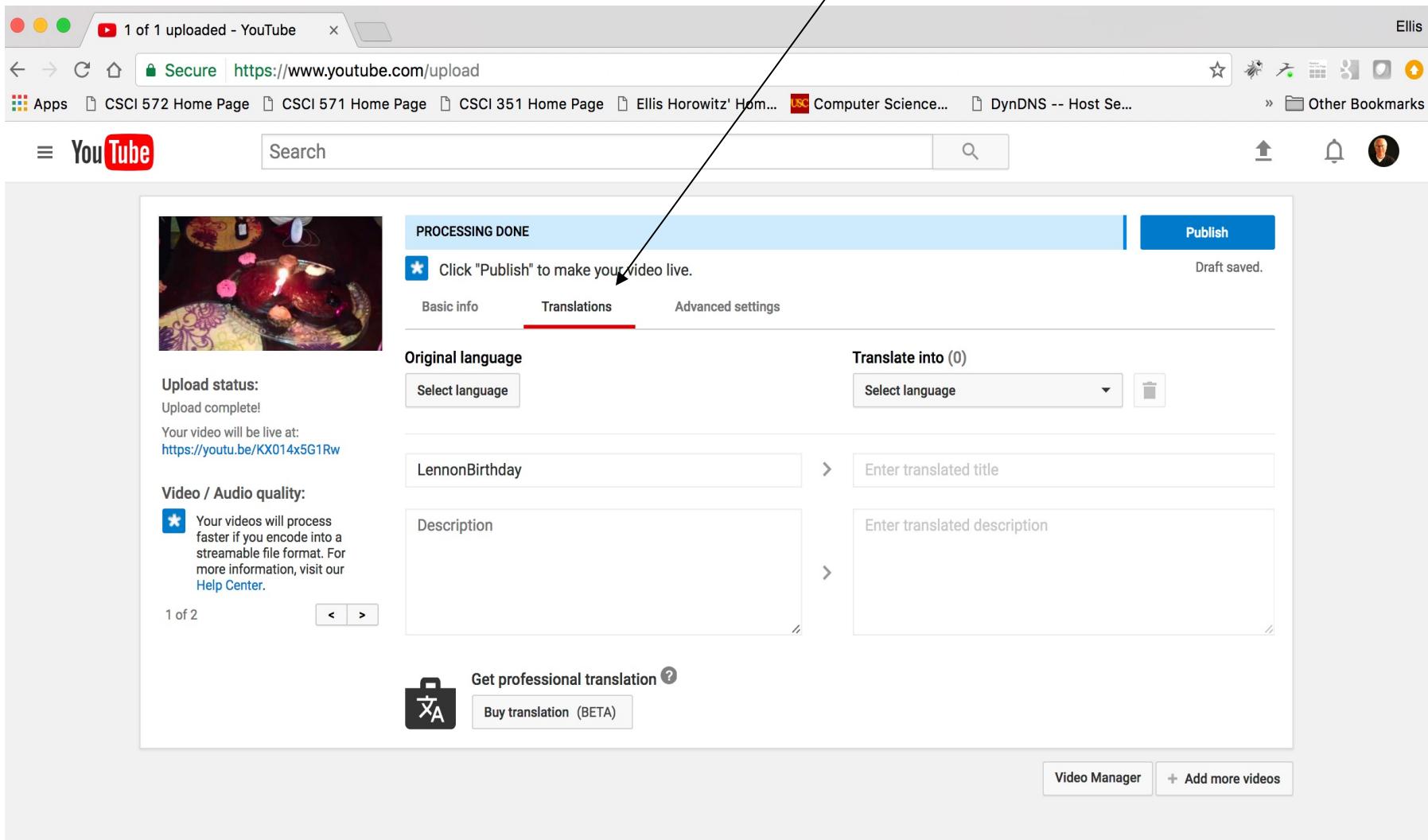
The main area shows a video thumbnail of a birthday cake with lit candles. To the right of the thumbnail, the status is "PROCESSING DONE" with a note to "Click 'Publish' to make your video live." Below this, there are tabs for "Basic info", "Translations", and "Advanced settings", with "Basic info" being the active tab. The "Basic info" section contains fields for "Name" (LennonBirthday), "Description", and "Tags (e.g., albert einstein, flying pig, mashup)". To the right of these fields are dropdown menus for "Visibility" (set to "Public") and "Share on" (with options for Google+ and Twitter). There is also a "Custom thumbnail" section with a "Custom thumbnail" button and a note about file size (maximum 2 MB).

On the left side of the interface, there are annotations with arrows pointing to specific elements:

- An arrow points from the "Name" label to the "Name" input field.
- An arrow points from the "Description" label to the "Description" input field.
- An arrow points from the "Tags" label to the "Tags" input field.
- An arrow points from the "Note: YouTube immediately assigns a URL" text to the URL displayed below the "Upload status" message: <https://youtu.be/KX014x5G1Rw>.
- An arrow points from the "Note: YouTube suggests possible thumbnails" text to the "VIDEO THUMBNAILS" section, which shows three suggested thumbnail images and a "Custom thumbnail" button.

Video on How to Upload a Video
<https://support.google.com/youtube/answer/57407>

Uploading to YouTube Second Input Screen



The screenshot shows the YouTube upload interface. At the top, a banner indicates "1 of 1 uploaded - YouTube". The URL in the address bar is <https://www.youtube.com/upload>. On the right side of the header, there are links for "Ellis", "Apps", "CSCI 572 Home Page", "CSCI 571 Home Page", "CSCI 351 Home Page", "Ellis Horowitz' Hom...", "USC Computer Science...", "DynDNS -- Host Se...", and "Other Bookmarks". Below the header, the YouTube logo and search bar are visible.

The main content area displays a thumbnail image of a video showing a lit candle on a decorated plate. A blue banner at the top says "PROCESSING DONE" and contains the instruction "Click "Publish" to make your video live." Below this, tabs for "Basic info", "Translations", and "Advanced settings" are shown, with "Translations" being the active tab. A message "Draft saved." is displayed next to the "Publish" button.

On the left, under "Upload status:", it says "Upload complete!" and provides the video link <https://youtu.be/KX014x5G1Rw>. Under "Video / Audio quality:", a note says "Your videos will process faster if you encode into a streamable file format. For more information, visit our Help Center." Below this, it shows "1 of 2" with navigation arrows.

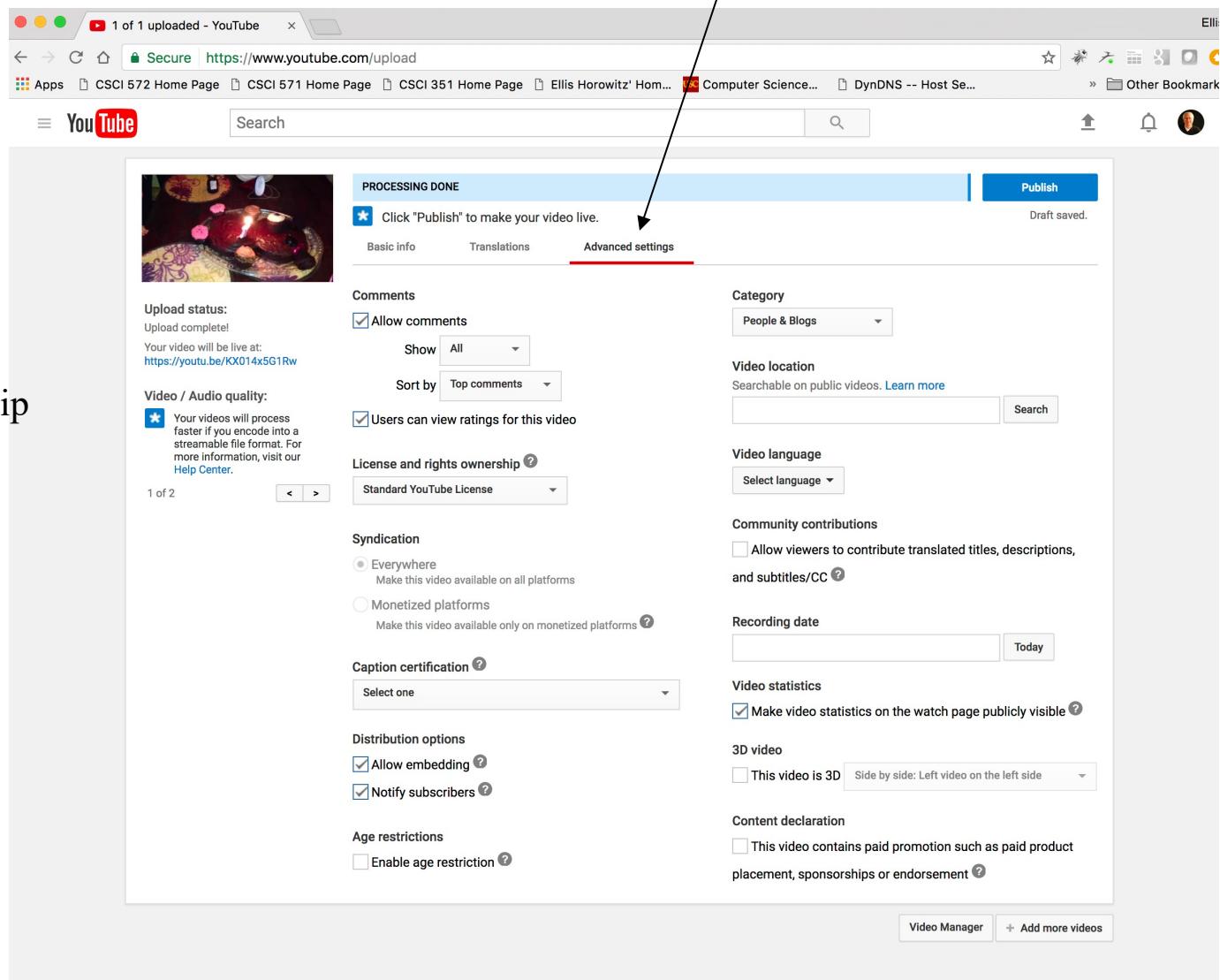
The "Translations" section allows users to translate the video's title and description into other languages. It includes fields for "Original language" (Select language), "Translate into (0)" (Select language), and two sets of "Enter translated title" and "Enter translated description" fields for "LennonBirthday".

At the bottom, there is a "Get professional translation" button with a briefcase icon and a "Buy translation (BETA)" button. Navigation buttons for "Video Manager" and "+ Add more videos" are also present.

Uploading to YouTube Third Input Screen

YouTube allows
the creator to
specify:

License and ownership
Syndication
Caption
Embedding
Age restrictions
Categories
• • •



The screenshot shows the YouTube upload interface after processing is done. A large blue banner at the top says "PROCESSING DONE" with a star icon and the text "Click 'Publish' to make your video live." Below this, the "Advanced settings" tab is selected. The page is filled with various configuration options:

- Comments:** Includes checkboxes for "Allow comments" (checked) and "Users can view ratings for this video".
- Category:** Set to "People & Blogs".
- Video location:** Set to "Searchable on public videos".
- Video language:** "Select language".
- Community contributions:** "Allow viewers to contribute translated titles, descriptions, and subtitles/CC".
- Recording date:** "Today".
- Video statistics:** "Make video statistics on the watch page publicly visible".
- 3D video:** "This video is 3D" with a dropdown for "Side by side: Left video on the left side".
- Content declaration:** "This video contains paid promotion such as paid product placement, sponsorships or endorsement".
- Distribution options:** "Allow embedding" (checked) and "Notify subscribers" (checked).
- Age restrictions:** "Enable age restriction".

At the bottom right are buttons for "Video Manager" and "+ Add more videos".

Business Model: Ads, Ads, Ads

Sample YouTube Search Results for Katy Perry

First result is an Ad

2nd and 4th results are stored at Vevo

3rd and 5th results are links to a Katy Perry channel with 106 videos

To the right is a mix of Katy Perry songs and some “related” artists

YouTube katy perry

About 17,000,000 results

Lipstick By Katy Perry
Ad COVERGIRL
 3,094,672 views
 Get Exclusive Katy Kat Lipstick For Pillow Soft, Rich Matte Lips Now!

Katy Perry - Chained To The Rhythm (Lyric Video) ft. Skip Marley
 KatyPerryVEVO
 2 days ago • 10,827,898 views
 Get "Chained to the Rhythm": <http://katy.to/cttr> Directed by: Aya Tanimura Executive Producer: Danny Lockwood Produced by: ...
 NEW

KatyPerryVEVO
 106 videos
 CHANNEL Subscribe 20,804,904

Katy Perry - Roar (Official)
 KatyPerryVEVO
 3 years ago • 1,790,071,602 views
 Get "Roar" from Katy Perry's 'PRISM': <http://smarturl.it/PRISM> Official music video for Katy Perry's "Roar" brought to you in ...
 CC

Katy Perry - Topic
 433 videos
 Katheryn Elizabeth Hudson, known professionally as Katy Perry, is an American singer and songwriter. After singing in church ...
 CHANNEL Subscribe 297,603

Katy Perry
 Music

Top Tracks Albums

- Roar 4:30
- Dark Horse 3:46
- Last Friday Night (T.G.I.F.) 8:11
- Firework 3:54
- Hot n Cold 4:44
- Part of Me 4:12
- Wide Awake 4:37
- The One That Got Away 4:51
- Unconditionally 3:57
- This Is How We Do 3:30
- [View all](#)

Related Artists

-  Taylor Swift
-  Lady Gaga
-  Russell Brand

Ranking: Ads, Views, Age YouTube Search Results

Begins with an ad

The next 4 results are ordered by the number of views: 420,004, 369,979, 228,004

Subsequent listings are a mixture of highly viewed videos, but older, e.g. Lec 1 MIT has 3 million+ views but is 7 years old

It is not obvious how the ranking was determined

YouTube computer science

About 18,600,000 results

Technology For Students
Ad Best Buy
41,606 views
Check Out Best Buy's Student Device Management Programs For Schools!

Lecture 0 - Introduction to Computer Science I
Asim Ali
2 years ago • 420,004 views
This is first lecture from the series of course "Introduction to Computer Science I", Harvard OpenCourseWare with Instructor David ...
CC

Computer Science a good major?
ENGINEERED TRUTH
3 years ago • 369,979 views
You should ask a lot of people for advice. In my opinion, most people in the world should get their bachelors in CS before working ...

Computer science is for everyone | Hadi Partovi | TEDxRainier
TEDx Talks
2 years ago • 228,044 views
This talk was given at a local TEDx event, produced independently of the TED Conferences. This persuasive talk shows how ...

Computer Science vs Self-Taught vs Coding Bootcamp (ft. Quincy Larson)
ENGINEERED TRUTH
4 months ago • 155,464 views
Quincy Larson is the creator of FreeCodeCamp.com, the #1 way to learn code for free. FreeCodeCamp is also the most starred ...

Computer science education: why does it suck so much and what if it didn't? | Ashley Gavin |...
TEDx Talks
1 year ago • 220,105 views
Ashley's talk shines a light on the major problem that is American Computer Science education. In 2020, 1.4 million new jobs will ...

Computer science education: why does it suck so much and what if it didn't? | Ashley Gavin |...
TEDx Talks
1 year ago • 220,105 views
Ashley's talk shines a light on the major problem that is American Computer Science education. In 2020, 1.4 million new jobs will ...

Lec 1 | MIT 6.00 Introduction to Computer Science and Programming, Fall 2008
MIT OpenCourseWare
7 years ago • 3,423,564 views
Lecture 1: Goals of the course; what is computation; introduction to data types, operators, and variables Instructors: Prof. CC

Question: How Important is Math in a Computer Science Degree?
Eli the Computer Guy Live
1 year ago • 119,331 views
I would like to know how hard it is the mathematics part in the computer science undergraduate course. I love computers and ...

Computer Science Explained in less than 3 minutes
shaun diem-lane
2 years ago • 257,833 views
Computer Programming is an amazing field of complication, amazement, difficulty, but above all, fun. Computer Programming ...

Computer Science Tutor
77 videos
CHANNEL Subscribe 6,009

Vlog: What to expect in a Computer Science course
lcc0612
1 year ago • 25,738 views
Being pretty near graduation now, I decide that, by reflecting upon my own experience, answer some of the most commonly asked ...

- During a search YouTube provides filters for users to refine their search:
 - UPLOAD DATE
 - TYPE
 - DURATION
 - FEATURES
 - SORT BY

YouTube Advanced Search Ranking Filters

YouTube search results for "algorithms".

About 1,690,000 results

FILTER

UPLOAD DATE	TYPE	DURATION	FEATURES	SORT BY
Last hour	Video	Short (< 4 minutes)	4K	Relevance
Today	Channel	Long (> 20 minutes)	HD	Upload date
This week	Playlist		Subtitles/CC	View count
This month	Movie		Creative Commons	Rating
This year	Show		3D	
			Live	
			Purchased	
			360°	

Intro to Algorithms: Crash Course Computer Science #13
 CrashCourse • 173K views • 4 months ago
 Algorithms are the sets of steps necessary to complete computation - they are at the heart of what our devices actually do. And this ...
 CC

MIT 6.006 Introduction to Algorithms, Fall 2011
 MIT OpenCourseWare
 1. Algorithmic Thinking, Peak Finding • 53:22
 2. Models of Computation, Document Distance • 48:52
[VIEW FULL PLAYLIST \(47 VIDEOS\)](#)

John MacCormick's Nine Algorithms That Changed the Future
 Princeton University Press • 5.4K views • 5 years ago
 Every day, we use our computers to perform remarkable feats. A simple web search picks out a handful of relevant needles from ...

YouTube Ranking Factors

- YouTube uses the following metrics for ranking search results:

1. Meta Data

- video titles, descriptions and tags are core ranking factors
- include links to a website and social profiles

2. Video Quality

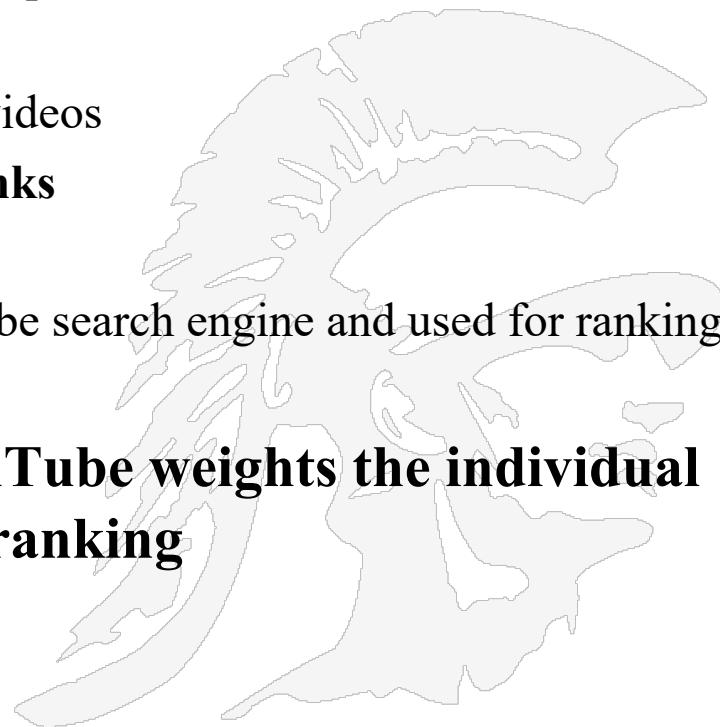
- HD ranks higher than low quality videos

3. Number of views, likes, shares and links

4. Subtitles and Closed Captions

- captions are crawled by the YouTube search engine and used for ranking

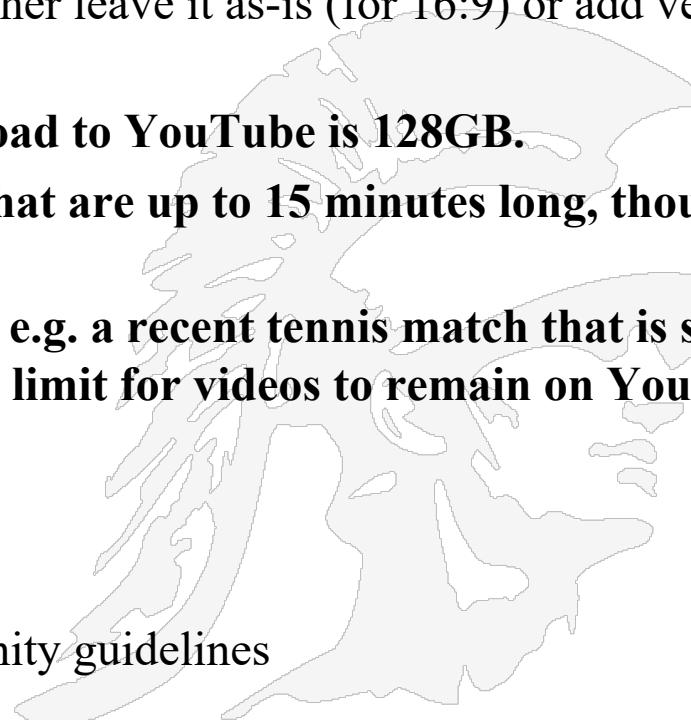
- **What is not known is how YouTube weights the individual factors to make up their final ranking**



YouTube Upload Characteristics

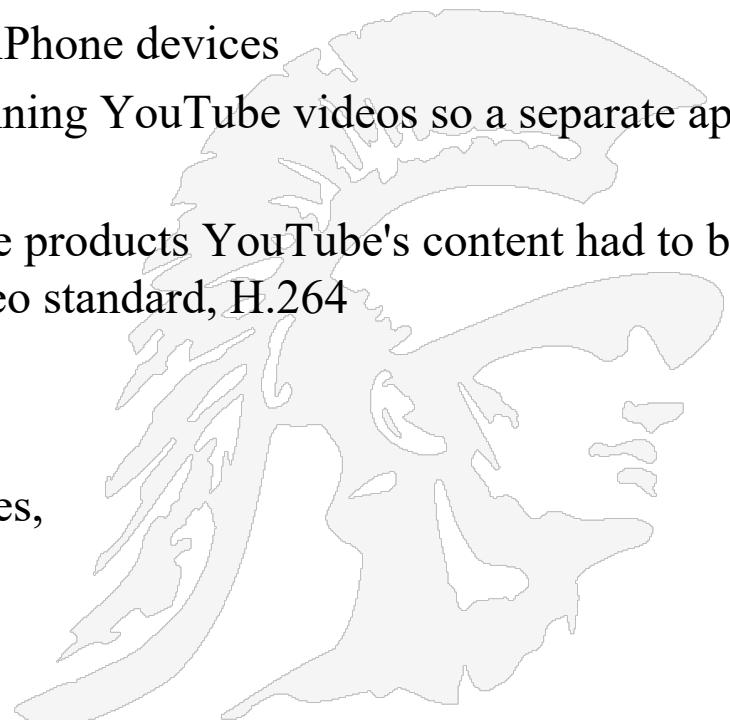
- **YouTube Upload Characteristics**

- **YouTube** supports 8 video formats for uploading: MOV, MP4 (MPEG4), AVI, WMV, FLV, 3GP, MPEGPS, WebM
- **Aspect Ratio:** the standard aspect ratios are: 4:3 or 16:9. When the video is uploaded to the site, YouTube will either leave it as-is (for 16:9) or add vertical black bars (for 4:3)
- **The maximum file size you can upload to YouTube is 128GB.**
- **By default, you can upload videos that are up to 15 minutes long, though that can be extended**
- **Many videos have a short life cycle, e.g. a recent tennis match that is soon forgotten, however, there is no time limit for videos to remain on YouTube, unless**
 - You delete the video.
 - You delete your account.
 - You violate copyright or community guidelines



YouTube Videos Run On Multiple Platforms

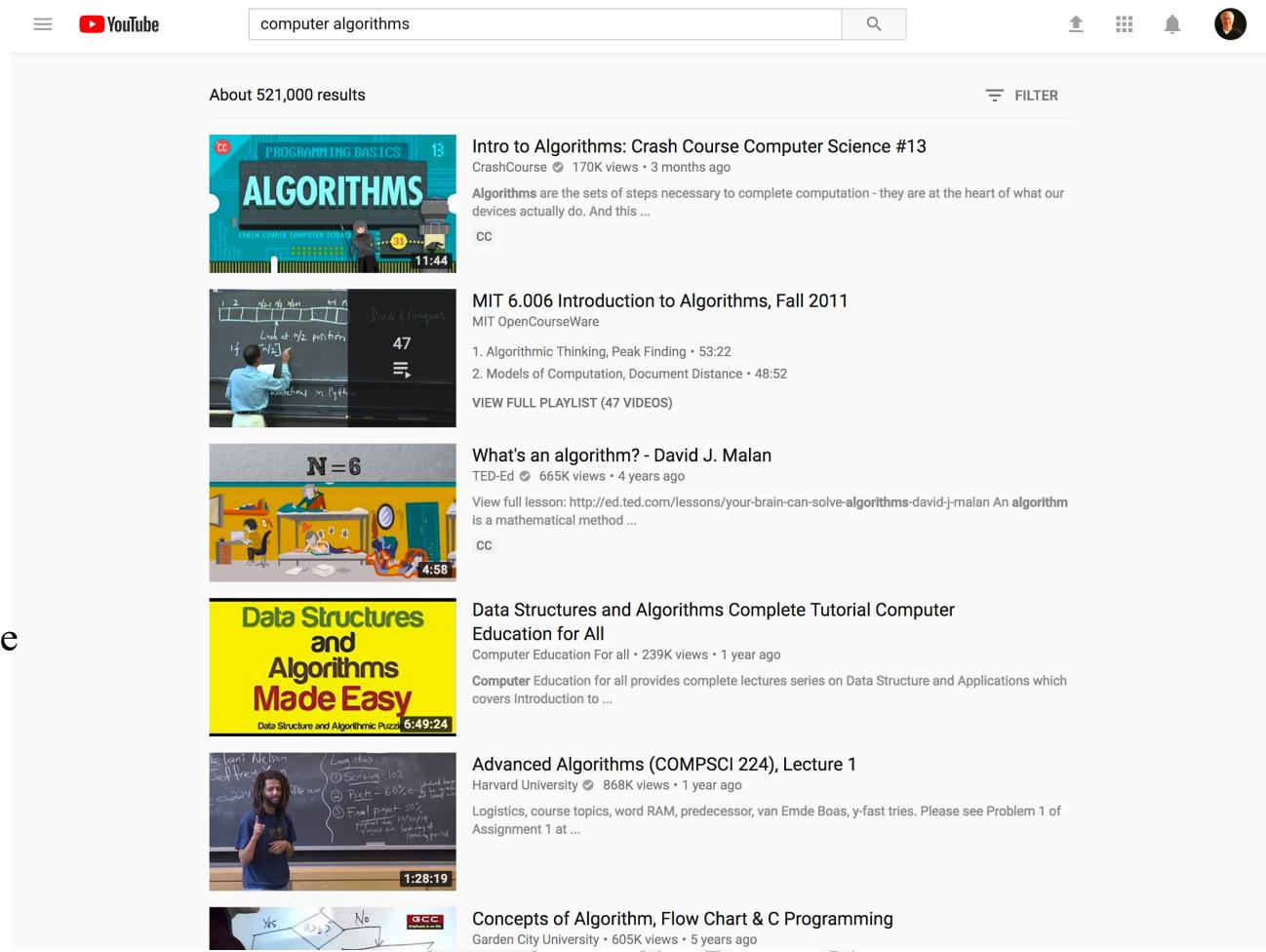
- **Desktops/laptops**
 - Videos are played in your browser assuming it supports HTML5
 - This avoided the need to use Adobe Flash Player
- **Smartphones**
 - YouTube apps exist for Android and iPhone devices
 - There is no native support for running YouTube videos so a separate app is required
 - For YouTube's videos to run on Apple products YouTube's content had to be transcoded into Apple's preferred video standard, H.264
- **Other Devices**
 - Apple TV, Fire TV, iPod Touch,
 - TiVo, PlayStation, Wii Game consoles,
 - Xbox Live, Roku Players
 - Google Chromecast



YouTube Makes Recommendations to Retain Viewers

- YouTube Search Results Example for query “computer algorithms”
- Assume we choose the first result

Recommendations are made to maximize watch time



The screenshot shows a YouTube search results page for the query "computer algorithms". The search bar at the top contains the text "computer algorithms". Below the search bar, there is a button labeled "FILTER". The results section displays five video thumbnails:

- Intro to Algorithms: Crash Course Computer Science #13** by CrashCourse (170K views, 3 months ago). Description: "Algorithms are the sets of steps necessary to complete computation - they are at the heart of what our devices actually do. And this ...".
- MIT 6.006 Introduction to Algorithms, Fall 2011** by MIT OpenCourseWare (1. Algorithmic Thinking, Peak Finding • 53:22, 2. Models of Computation, Document Distance • 48:52). Description: "VIEW FULL PLAYLIST (47 VIDEOS)".
- What's an algorithm? - David J. Malan** by TED-Ed (665K views, 4 years ago). Description: "View full lesson: <http://ed.ted.com/lessons/your-brain-can-solve-algorithms-david-j-malan> An algorithm is a mathematical method ...".
- Data Structures and Algorithms Complete Tutorial Computer Education for All** by Computer Education for all (239K views, 1 year ago). Description: "Computer Education for all provides complete lectures series on Data Structure and Applications which covers Introduction to ...".
- Advanced Algorithms (COMPSCI 224), Lecture 1** by Harvard University (868K views, 1 year ago). Description: "Logistics, course topics, word RAM, predecessor, van Emde Boas, y-fast tries. Please see Problem 1 of Assignment 1 at ...".
- Concepts of Algorithm, Flow Chart & C Programming** by Garden City University (605K views, 5 years ago). Description: "GCC".

<https://www.nbcnews.com/tech/social-media/algorithms-take-over-youtube-s-recommendations-highlight-human-problem-n867596>

YouTube Recommendation Algorithm

- Given the query “computer algorithms” followed by a selection, YouTube makes recommendations for subsequent videos
- Recommendations account for 60% of all video clicks



YouTube Recommendation System

- **Association Rule Mining**

- For each pair of videos v_i, v_j compute co-visitation counts, i.e. they count how often they were co-watched; if $c_{i,j}$ is the co-visitation count, then relatedness is defined as

$$r(v_i, v_j) = \frac{c_{ij}}{f(v_i, v_j)}$$

where c_i and c_j are the total occurrence counts across all sessions for videos v_i and v_j . $f(v_i, v_j)$ is a normalization function that takes the global popularity of both the seed video and the candidate video into account; e.g. $f(v_i, v_j) = c_i * c_j$

The set of related videos, R_i for a given seed video v_i is determined by taking the top N candidate videos ranked by their scores $r(v_i, v_j)$

Related videos induce a directed graph over the set of videos, namely:

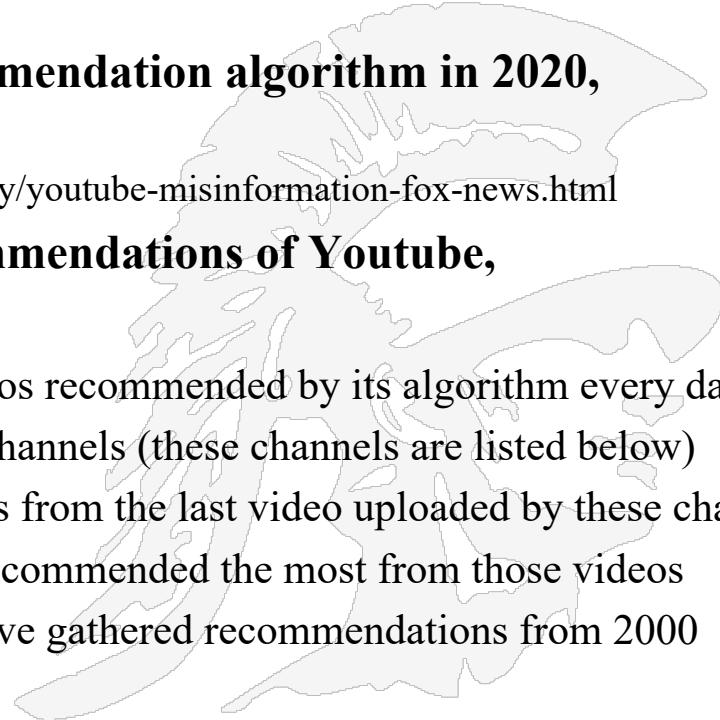
For each pair of videos (v_i, v_j) , there is an edge e_{ij} from v_i to v_j iff v_j is in R_i

For details see: *The YouTube Recommendation System*

<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.434.9301&rep=rep1&type=pdf>

Media Sites (including YouTube) Move Away from False Information

- **YouTube's recommendation algorithm used to send people to misinformation, e.g. see**
 - <https://www.youtube.com/watch?v=FI8tFmBIPak> (3 min)
 - <https://www.wsj.com/articles/how-youtube-drives-viewers-to-the-internets-darkest-corners-1518020478>
- **As a result YouTube changed its recommendation algorithm in 2020, eliminating so-called “fringe” sites**
 - <https://www.nytimes.com/2020/11/03/technology/youtube-misinformation-fox-news.html>
- **There is a website that tracks the recommendations of Youtube,**
- **<https://algotransparency.org/>**
- “We used a multi-step program to analyze videos recommended by its algorithm every day”
 - Step 1: We start from a list of 1000+ US channels (these channels are listed below)
 - Step 2: We gather all recommended videos from the last video uploaded by these channels
 - Step 3: We compute which channel was recommended the most from those videos
 - Step 4: We repeat step 2 and 3 until we have gathered recommendations from 2000 channels
 - Step 5: For each video that was observed, we count and display from how many channels it was recommended



A **video rich snippet** means that when someone searches for something on Google, you can have a small tiny **video** show up next to your result to let the user know that particular result (yours) has a **video** to help

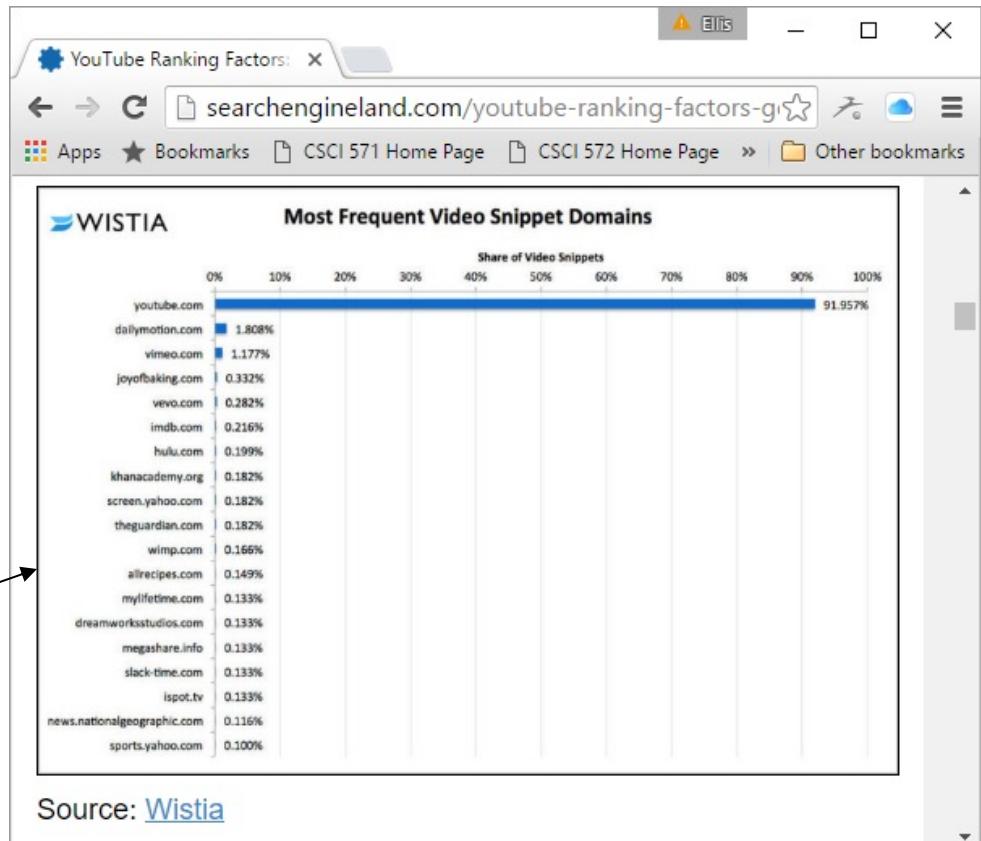
Google weeded out the video competition in Web search by predominantly displaying **only video-rich snippets** for YouTube videos back in 2014.

Here is a graph outlining the percentage share of video-rich snippets in Google; 91% are from YouTube

see

<https://wistia.com/blog/where-did-my-video-snippets-go>

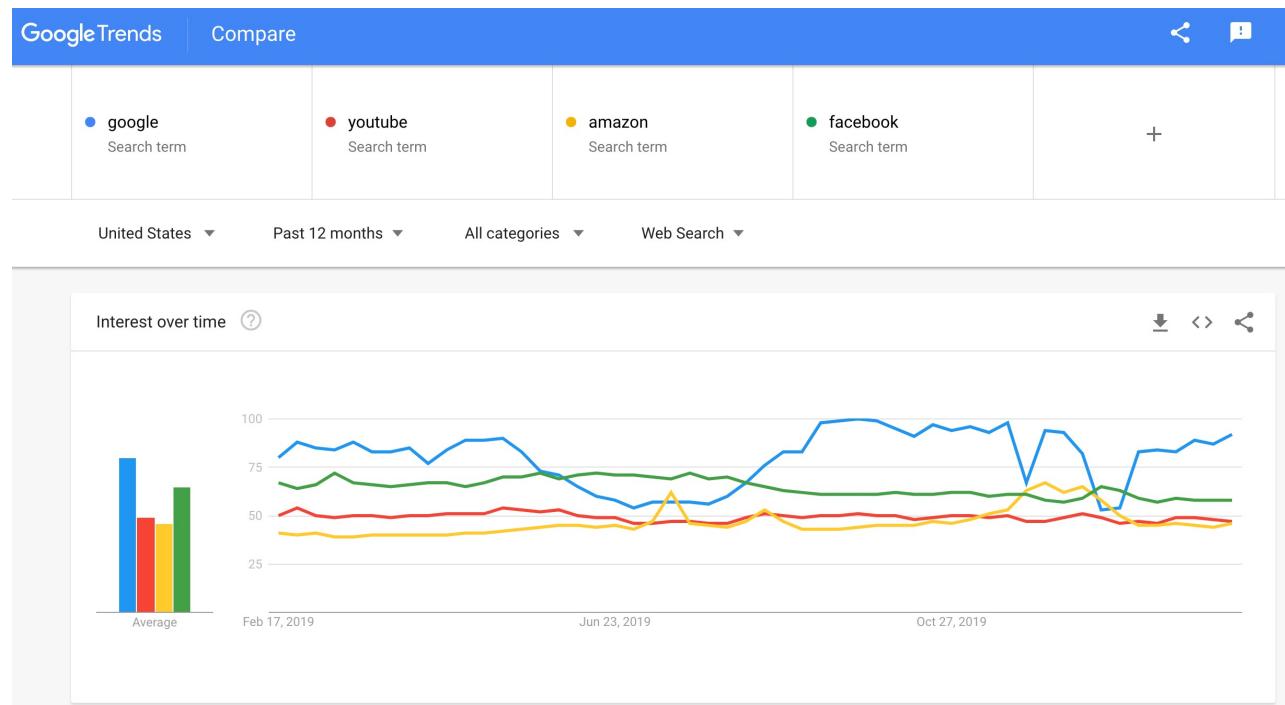
Google Search is Biased Towards YouTube Videos



e.g. try “tutorial on bitcoin”

Google is Biased Towards YouTube Videos

- Google trends analyzes the popularity of top search queries, see trends.google.com
- Google also made an update to Google trends recently by including YouTube trending topics in the tool.
- This graph shows how interest in Google compares with YouTube, Amazon and Facebook

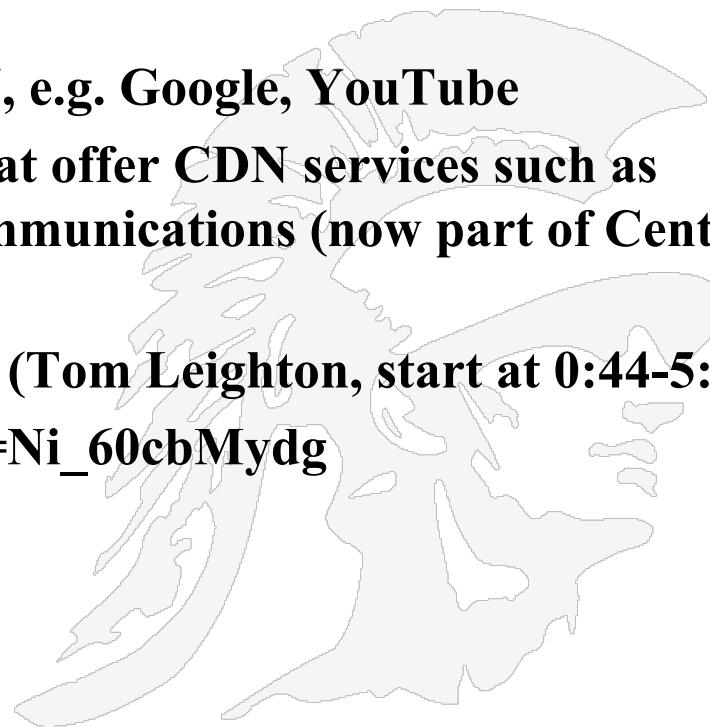


<https://searchengineland.com/youtube-ranking-factors-getting-ranked-second-largest-search-engine-225533>



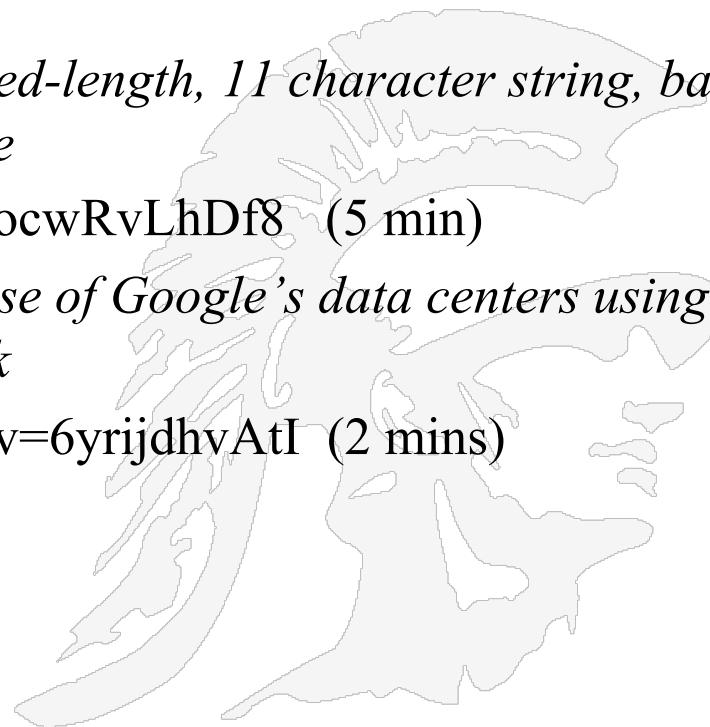
Content Delivery Networks

- A content distribution network (CDN) consists of a large set of content servers and a means for dynamically selecting servers based on knowledge of the location of the user and possibly the content being requested
- Some sights operate their own CDN, e.g. Google, YouTube
- There are third party companies that offer CDN services such as Akamai, Limelight and Level 3 Communications (now part of Century Link)
- See the Akamai video for 5 minutes (Tom Leighton, start at 0:44-5:00),
- https://www.youtube.com/watch?v=Ni_60cbMydg



YouTube Video Delivery System

- **Two Critical Technology Challenges for YouTube:**
 - *how to identify billions of videos*
 - *How to efficiently deliver the video to the desktop/mobile device*
- **The Solutions:**
- **Identification:** *YouTube assigns a fixed-length, 11 character string, base 64, unique identifier to each video, see*
- <https://www.youtube.com/watch?v=gocwRvLhDf8> (5 min)
- **Efficient Delivery:** *YouTube makes use of Google's data centers using them as a content distribution network*
 - <https://www.youtube.com/watch?v=6yrijdhvAtI> (2 mins)



YouTube (Google's) Content Delivery Datacenters

- A map of Google's data centers, see
- <https://www.google.com/about/datacenters/inside/locations/index.html>

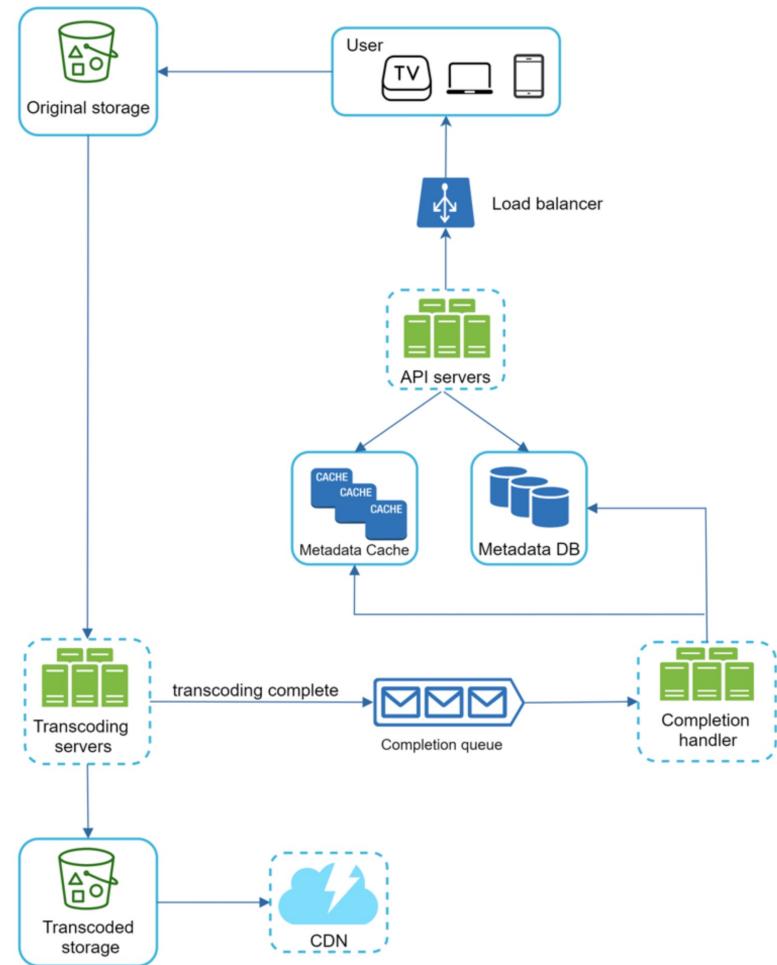


Figure 4: Geographical distribution of YouTube Video Cache Locations.

YouTube Upload Architecture

1. videos are uploaded from a desktop to a central Data Center
2. the video is then transcoded into multiple formats
3. transcoded copies are sent to the Content Distribution Network

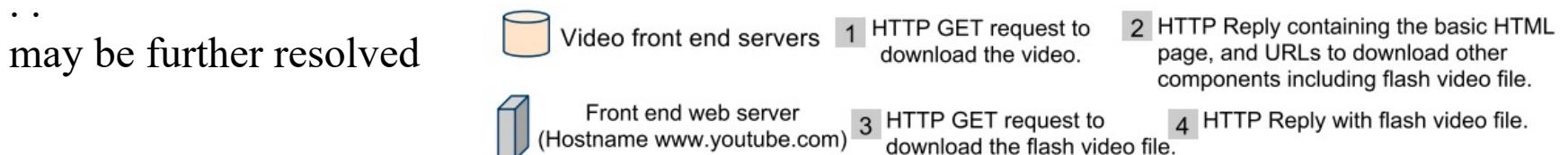
Video transcoding is a technique of converting a video into multiple different formats and resolutions to make it playable across different devices and bandwidths. The technique is also known as *video encoding*. This enables YouTube to stream videos in different resolutions such as *144p, 240p, 360p, 480p, 720p, 1080p & 4K*.



YouTube's Content Distribution Network Downloading a YouTube Video

A local DNS server resolves www.youtube.com and is redirected to a YouTube server which downloads the page information and a pointer to a YouTube server that can deliver the video, e.g.
v23.lscache5.c.youtube.com
The request to v23.lscache5

...
may be further resolved



4 steps describing the delivery of a YouTube video

<http://www-users.cs.umn.edu/~zhzhang/Papers/youtube-tech-report.pdf>

YouTube Delivery System

- The design of the YouTube video delivery system consists of three components:
 - a “flat” video id space,
 - a multi-layered logical server organization consisting of five anycast namespaces (and two unicast namespaces), and
 - a 3-tiered physical cache hierarchy with (at least) 38 primary locations, 8 secondary and 5 tertiary locations.

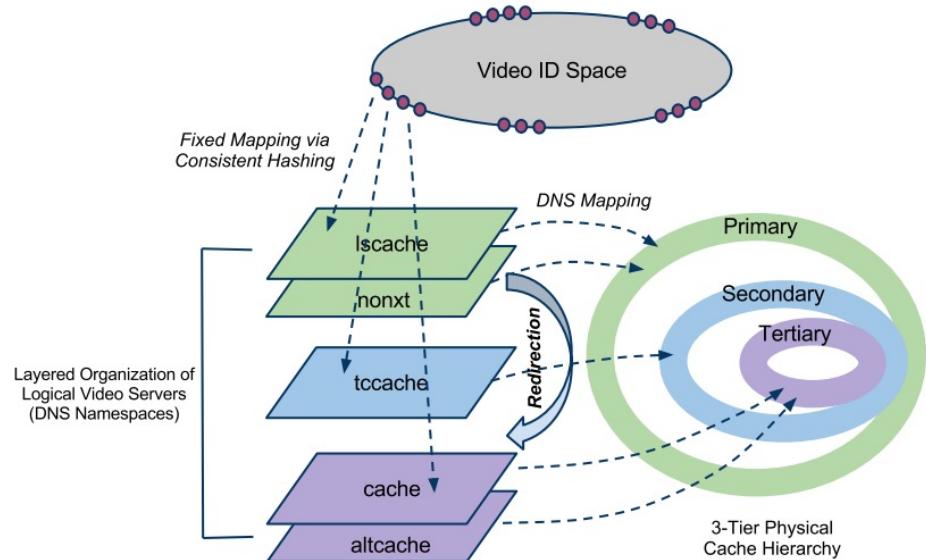
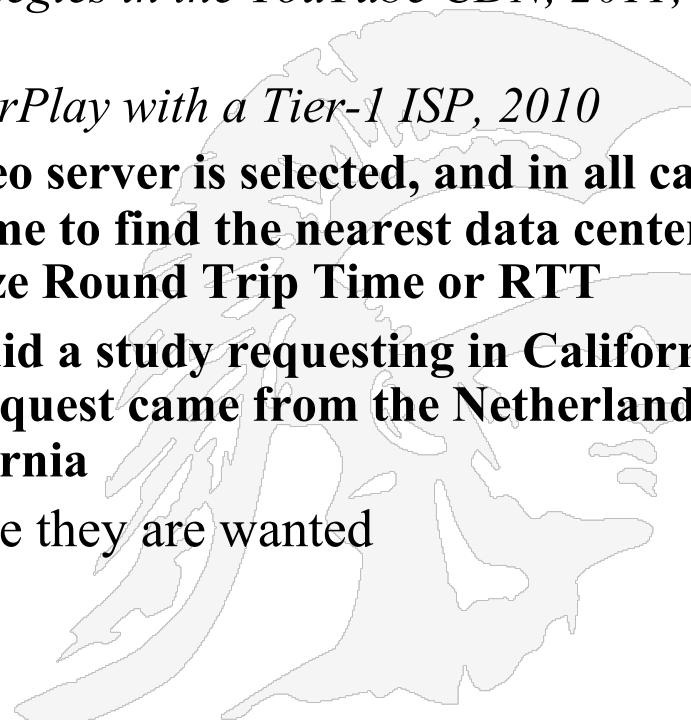


Figure 3: YouTube Architectural Design.

<https://www-users.cse.umn.edu/~zhang089/Papers/youtube-tech-report.pdf>

References to YouTube's CDN

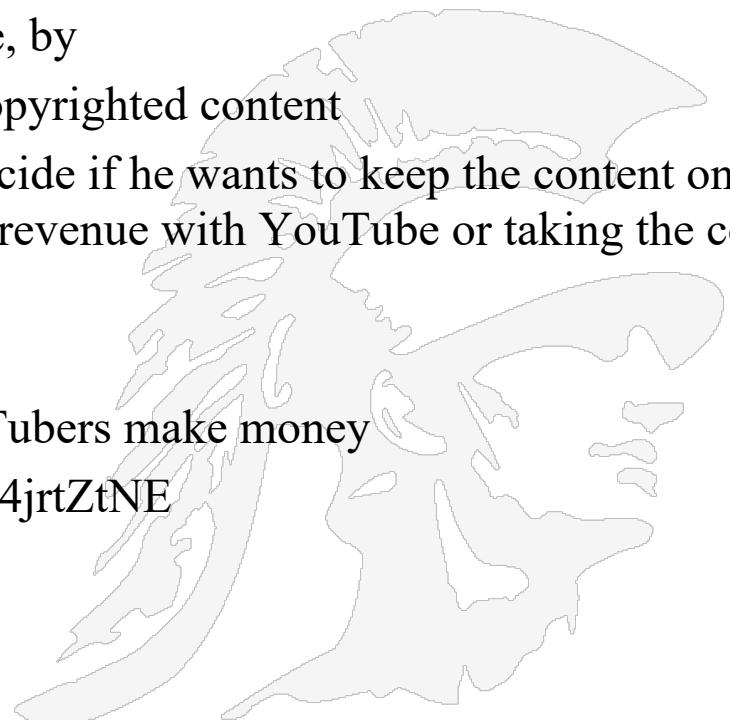
- There are three research papers that investigated and discussed the YouTube CDN, they are:
 1. *Vivisecting YouTube: An Active Measurement Study*, 2012, cited by Jefay
 2. *Dissecting Video Server Selection Strategies in the YouTube CDN*, 2011, cited by Jefay
 3. *YouTube Traffic Dynamics and Its InterPlay with a Tier-1 ISP*, 2010
- Both papers try to determine how a video server is selected, and in all cases there is a complicated re-direction scheme to find the nearest data center to serve the video; they attempt to minimize Round Trip Time or RTT
- For rarely-called-for videos *Dissecting* did a study requesting in California a rare video and observed that the first request came from the Netherlands, but future requests were served from California
 - Conclusion: videos are moved where they are wanted



Monetizing YouTube

- **YouTube challenges in the early days**

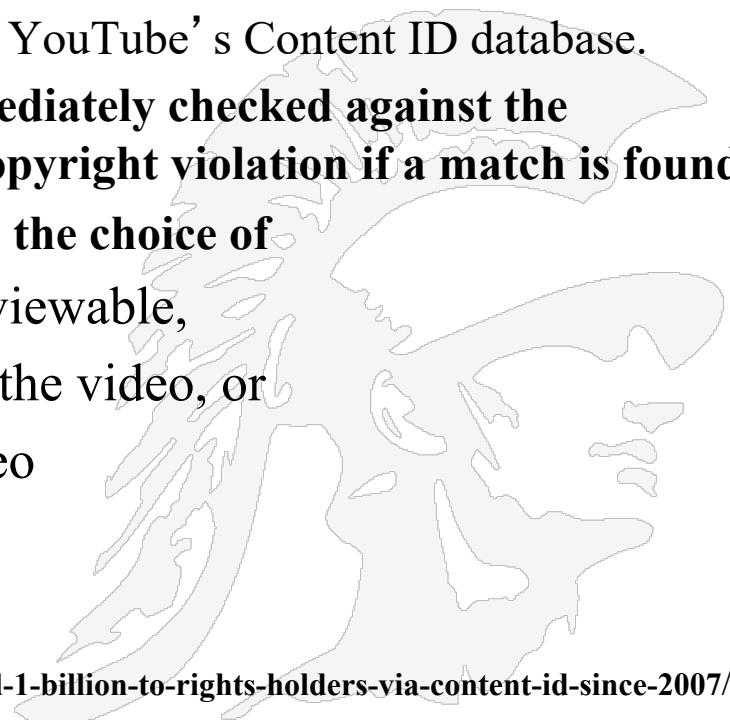
- YouTube had no way of making money and its infrastructure is very expensive
- YouTube was being sued by content creators as many of YouTube's videos were uploaded illegally
- YouTube **solved both problems** at once, by
 - Developing a system for spotting copyrighted content
 - Allowing the copyright owner to decide if he wants to keep the content on the site and let ads appear, splitting the revenue with YouTube or taking the content down



- Here is a video that describes how YouTubers make money
- <https://www.youtube.com/watch?v=v8F4jrtZtNE>
- (8 min)

ContentID

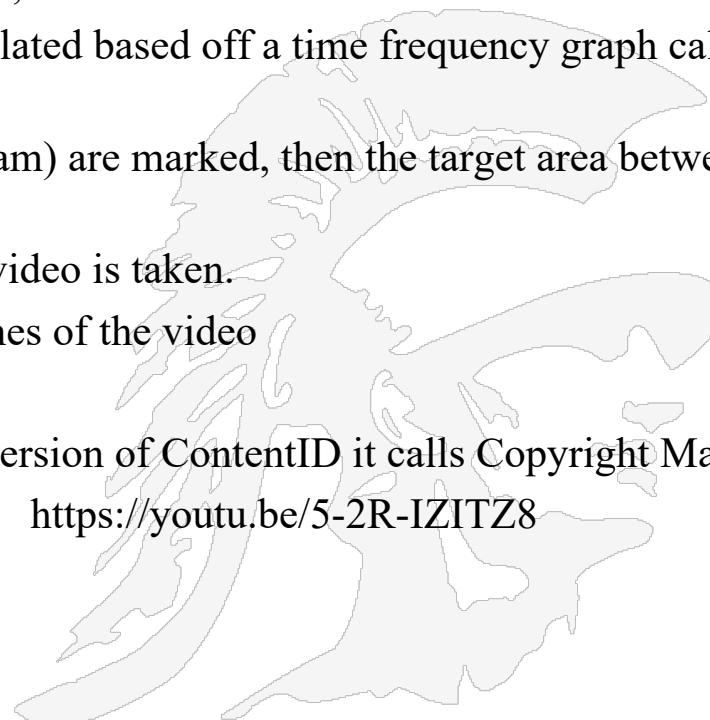
- YouTube's solution was to create a fingerprint database of copyrighted content, called Content ID
- YouTube solicited cooperation from content owners asking them to submit copies of their content so YouTube could fingerprint them
 - There are millions of reference files in YouTube's Content ID database.
- When a new video is uploaded, it is immediately checked against the database, and the video is flagged as a copyright violation if a match is found.
- When this occurs, the content owner has the choice of
 1. blocking the video to make it unviewable,
 2. tracking the viewing statistics of the video, or
 3. adding advertisements to the video



<https://arstechnica.com/tech-policy-policy/2014/10/youtube-has-paid-1-billion-to-rights-holders-via-content-id-since-2007/>

More Details on ContentID

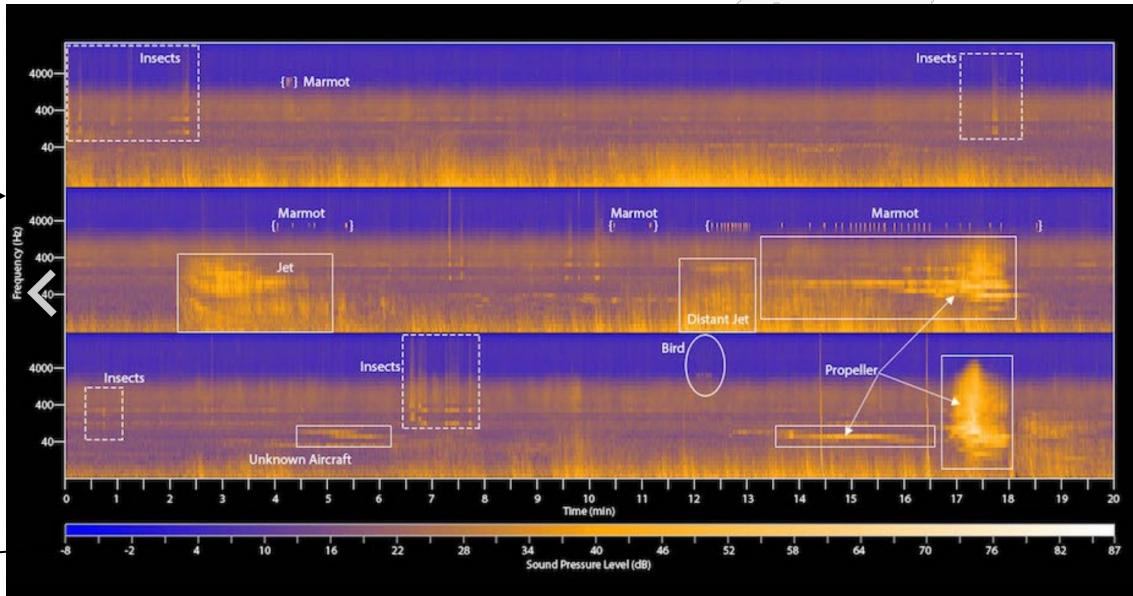
1. Content ID is based off audio and video samples that rights holders have uploaded to YouTube
 2. User uploads a video.
 3. YouTube then queues up the video to be processed i.e. it is transcoded into multiple formats including:
 - HTML5, H.264, WebM VP8, HD, non-HD, and others
 4. *If the video contains audio*, a hash is then calculated based off a time frequency graph called a spectrogram.
 - Target zones (peak points in the spectrogram) are marked, then the target area between them is also taken and hashed
 5. *For a video*, a sample section of frames of the video is taken.
 - A hash is created from those sampled frames of the video
-
- **Note** recently YouTube has introduced a new version of ContentID it calls Copyright Match
 - See the following videos for details, (2 min). <https://youtu.be/5-2R-IZITZ8>



Creating an Acoustic Fingerprint

- The audio signal is digitized and converted to a spectrogram – a time-frequency graph
 - The graph below plots three dimensions of audio: frequency versus amplitude versus time
 - A common format is a graph with two dimensions: one axis represents time, and the other axis represents frequency; a third dimension indicating the amplitude of a particular frequency at a particular time is represented by the intensity or color of each point in the image.

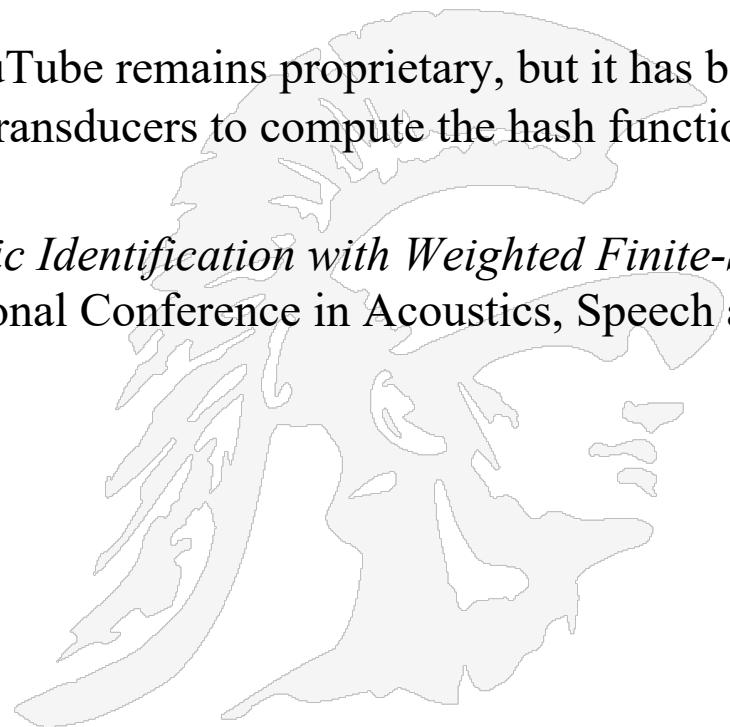
frequency →



Time axis →

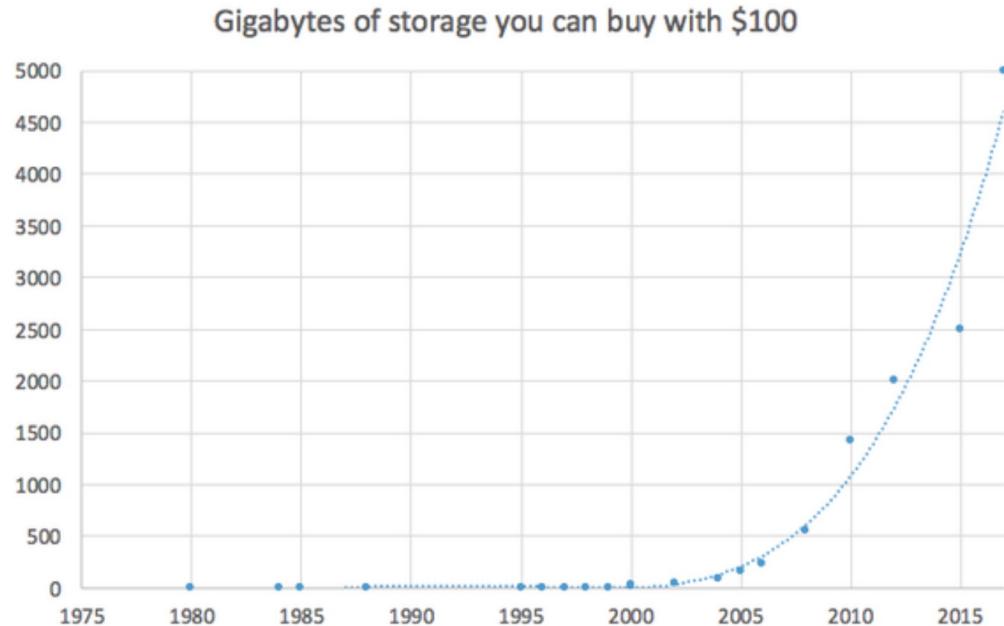
How Good is Content ID

- According to stats released by YouTube **99.5 percent** of all copyright issues specifically related to sound recordings are automatically resolved by Content ID
- In addition to music, Content ID also identifies 98% of copyright claims, including those tied to film, TV, gaming
- The actual hashing algorithm used by YouTube remains proprietary, but it has been suggested that YouTube uses finite-state transducers to compute the hash function, e.g. see
- Eugene Weinstein, Pedro J. Moreno; *Music Identification with Weighted Finite-State Transducers*, Proceedings of the International Conference in Acoustics, Speech and Signal Processing (ICASSP), 2007



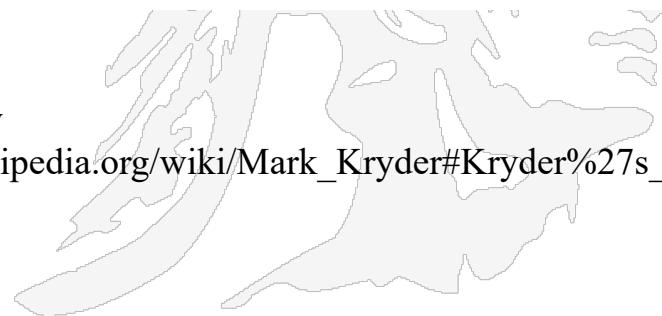
- The storage you can buy with \$100 has grown exponentially — or equivalently, the cost of storing 1GB of videos has decreased exponentially

Will YouTube Ever Run Out of Storage



Kryder's Law

https://en.wikipedia.org/wiki/Mark_Kryder#Kryder%27s_law_projection



- An answer by Rasty Turek from Quora
- There is roughly 24TB of new videos uploaded daily
- Each video is re-encoded based on pre-selected profiles and each is stored as a separate file
- Here is his computation:

 1280x720	mp4	 download - 59.01 MB
 640x360	mp4	 download - 15.34 MB
 640x360	webm	 download - 19.07 MB
 400x240	flv	 download - 8.51 MB
 320x240	3gp	 download - 5.94 MB
 176x144	3gp	 download - 2.12 MB
 4k (no audio)	mp4	 download - 297.69 MB

$$24\text{TB} * 4 \text{ (for profiles)} * 365 \text{ days} = 35\text{PB/year}$$

So YouTube needs to store roughly 35PB of new data every year.

From multiple sources we know that there is roughly 1B videos uploaded to this day to YouTube. Each video

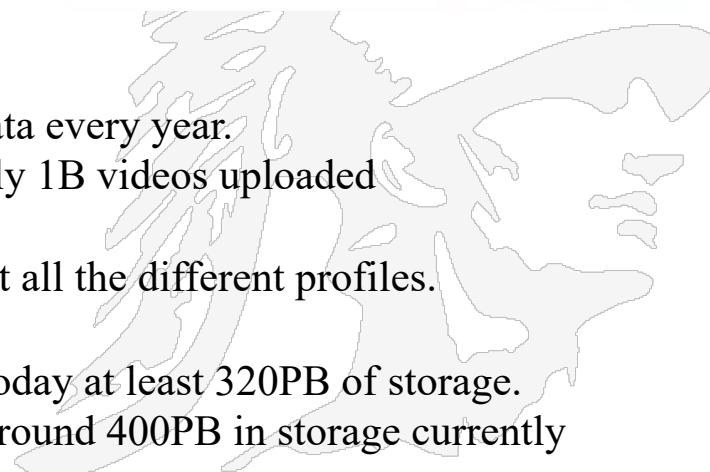
has on average size of 86MB. We still have to count all the different profiles.

$$86\text{MB} * 4 \text{ (for profiles)} * 1,000,000,000 = 320\text{PB}$$

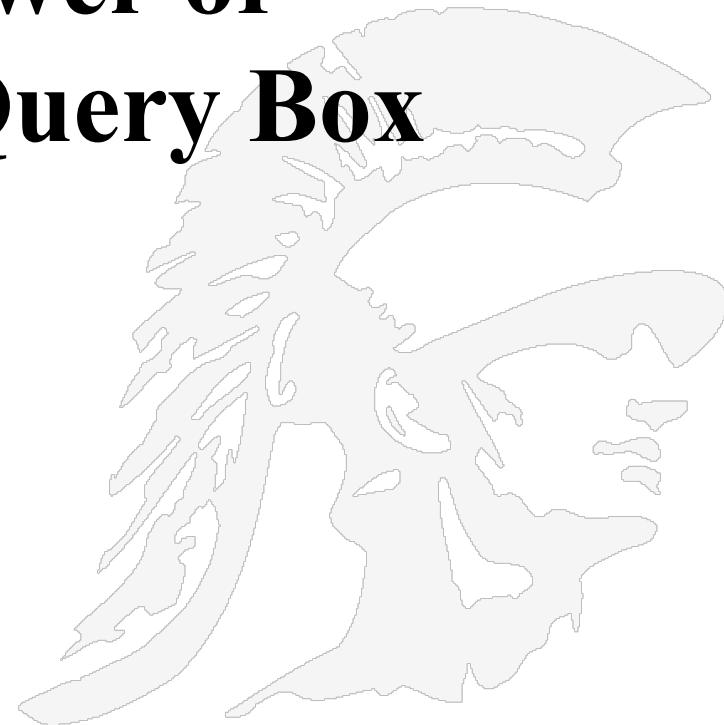
To store all this data YouTube needs to have as of today at least 320PB of storage.

From that we can estimate that they have roughly around 400PB in storage currently allocated for storing YouTube videos.

YouTube currently owns around 400,000TB of storage.

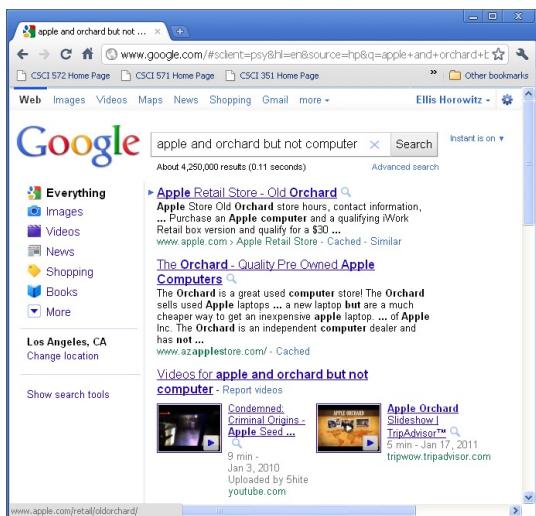


The Power of Google's Query Box



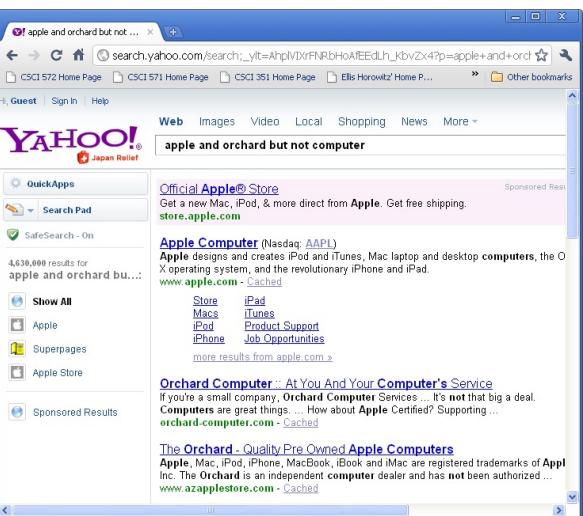
Traditional Boolean Queries in a Search Box Don't Work As Expected

Query: "apple AND orchard BUT NOT computer" still returns Apple Computer results



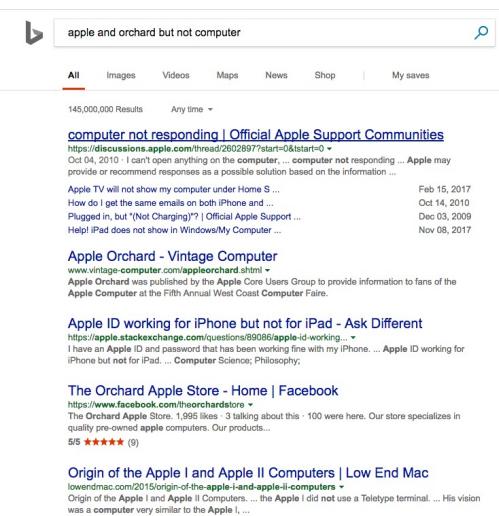
Google search results for "apple and orchard but not computer":

- Apple Retail Store - Old Orchard** (www.apple.com/retail/oldorchard) - 4,250,000 results (0.11 seconds)
- The Orchard - Quality Pre Owned Apple Computers** (www.azapplestore.com) - 4,630,000 results for apple and orchard bu...
- Videos for apple and orchard but not computer**: Report videos (Condemed, Criminal Origins - Apple Seed ...)



Yahoo search results for "apple and orchard but not computer":

- Official Apple® Store** (store.apple.com) - Sponsored Results
- Apple Computer (Nasdaq: AAPL)** (www.apple.com) - 4,630,000 results for apple and orchard bu...
- Orchard Computer :: At You And Your Computer's Service** (orchardcomputer.com) - If you're a small company, Orchard Computer Services... It's not that big a deal.
- The Orchard - Quality Pre Owned Apple Computers** (www.azapplestore.com) - 4,630,000 results for apple and orchard bu...

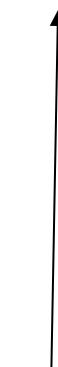


Bing search results for "apple and orchard but not computer":

- computer not responding | Official Apple Support Communities** (apple.com/discussions/apple/computer-not-responding) - Oct 04, 2010 - I can't open anything on the computer, ... computer not responding ... Apple may provide or recommend responses as a possible solution based on the information ...
- Apple Orchard - Vintage Computer** (www.vintage-computer.com/appleorchard.shtml) - Oct 14, 2010 - Apple Orchard was published by the Apple Computer Users Group to provide information to fans of the Apple Computer at the Fifth Annual West Coast Computer Fair.
- Apple ID working for iPhone but not for iPad - Ask Different** (apple.stackexchange.com/questions/69008/apple-id-working...) - Oct 14, 2010 - I have an Apple ID and password that has been working fine with my iPhone... Apple ID working for iPhone but not for iPad... Computer Science; Philosophy
- The Orchard Apple Store - Home | Facebook** (https://www.facebook.com/theorchardstore) - Oct 14, 2010 - The Orchard Apple Store, 1,995 likes · 3 talking about this · 100 were here. Our store specializes in quality pre-owned apple computers. Our products... \$5 ***** (9)
- Origin of the Apple I and Apple II Computers | Low End Mac** (lowendmac.com/2010/09/origin-of-the-apple-i-and-apple-ii-computers) - Oct 14, 2010 - Origin of the Apple I and Apple II Computers... the Apple I did not use a Teletype terminal... His vision was a computer very similar to the Apple I...

Related searches

- apple computer not working
- apple computer not responding
- apple computer not turning on
- apple computer not charging
- apple computer not starting
- apple computer not loading



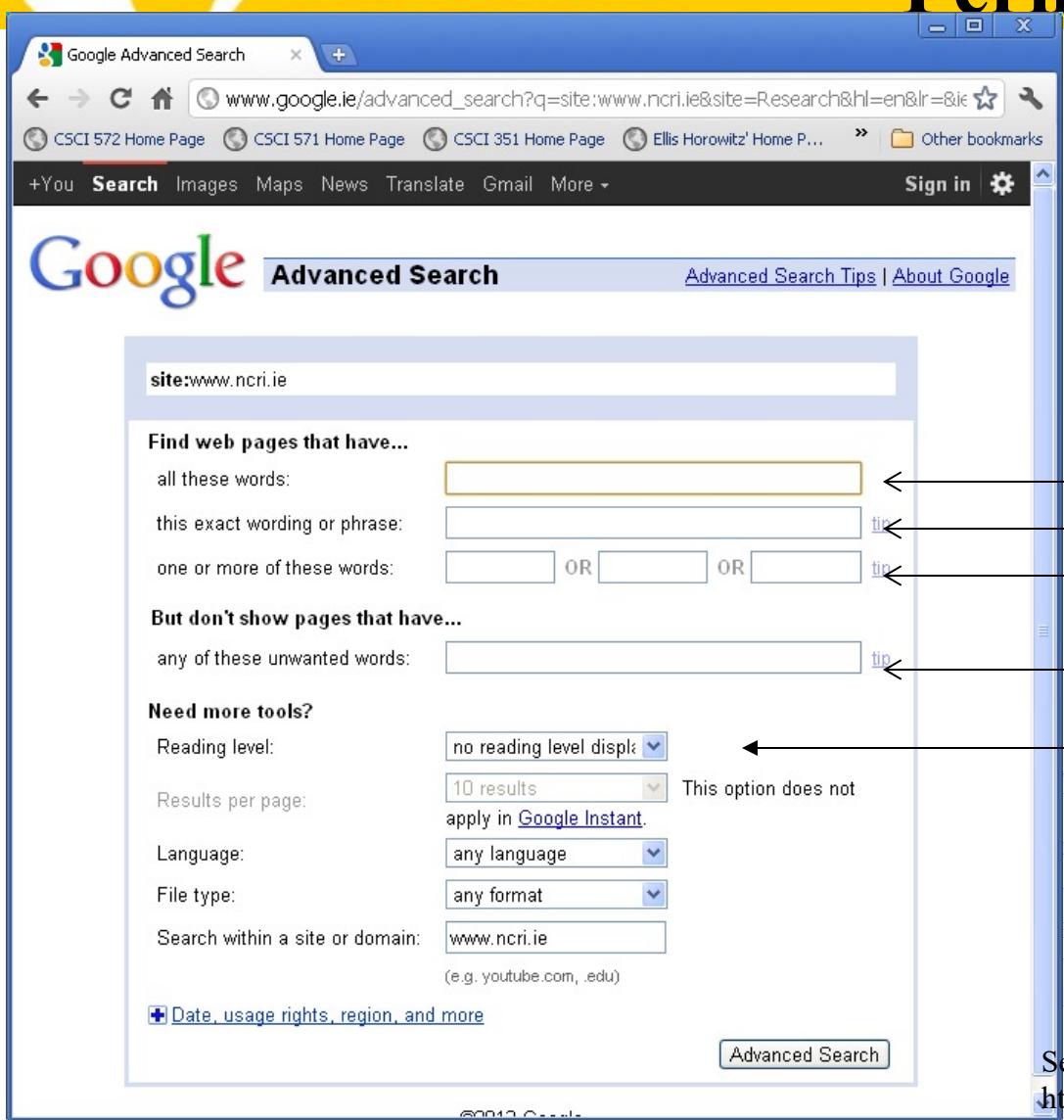
Notice
Related
searches

Google

Yahoo

Bing

But Google Advanced Search Permits Boolean Queries



The screenshot shows the Google Advanced Search interface. At the top, there's a search bar with the query "site:www.ncri.ie". Below it, under "Find web pages that have...", there are three input fields: "all these words:" (containing "tip"), "this exact wording or phrase:" (containing "tip"), and "one or more of these words:" (containing "tip"). To the right of these fields, arrows point from the text "ANDing a set of words", "Exact phrase", and "ORing a set of words" respectively. Further down, under "But don't show pages that have...", there's a field containing "tip", with an arrow pointing to the text "NOT words". On the left, under "Need more tools?", there are dropdown menus for "Reading level" (set to "no reading level displayed"), "Results per page" (set to "10 results"), "Language" (set to "any language"), "File type" (set to "any format"), and "Search within a site or domain" (set to "www.ncri.ie"). A note next to the results per page dropdown says "This option does not apply in Google Instant." At the bottom, there's a link "Date, usage rights, region, and more" and a "Advanced Search" button.

Search engines typically offer an “Advanced Search” page where users **can enter**, in effect, a Boolean query; e.g. Here is Google’s “advanced search” screen

ANDing a set of words

Exact phrase

ORing a set of words

NOT words

reading level

language
file type

etc

See also

<https://help.bing.microsoft.com/#apex/bing/en-US/10002/-1>

Advanced Search Works Properly

Google Advanced Search

Find pages with...

all these words:

this exact word or phrase:

any of these words:

none of these words:

numbers ranging from: to

Then narrow your results by...

language:

region:

last update:

site or domain:

terms appearing:

SafeSearch:

file type:

usage rights:

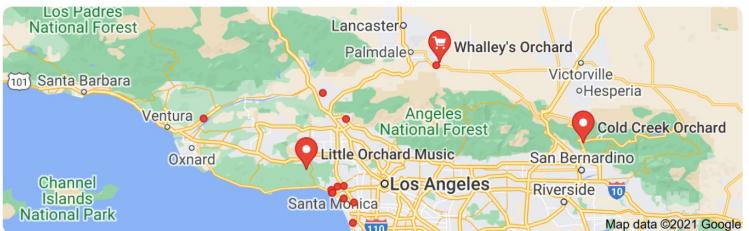
[Advanced Search](#)

apple orchard -computer

All Maps Images Videos News More Tools

About 66,700,000 results (0.76 seconds)

Apple Orchard -Computer



Rating Hours

Little Orchard Music
 5.0 ★★★★★ (3) · Orchard
 Topanga, CA · (310) 455-2950
 "Great place.I enjoyed the Little Orchard Music!!!"

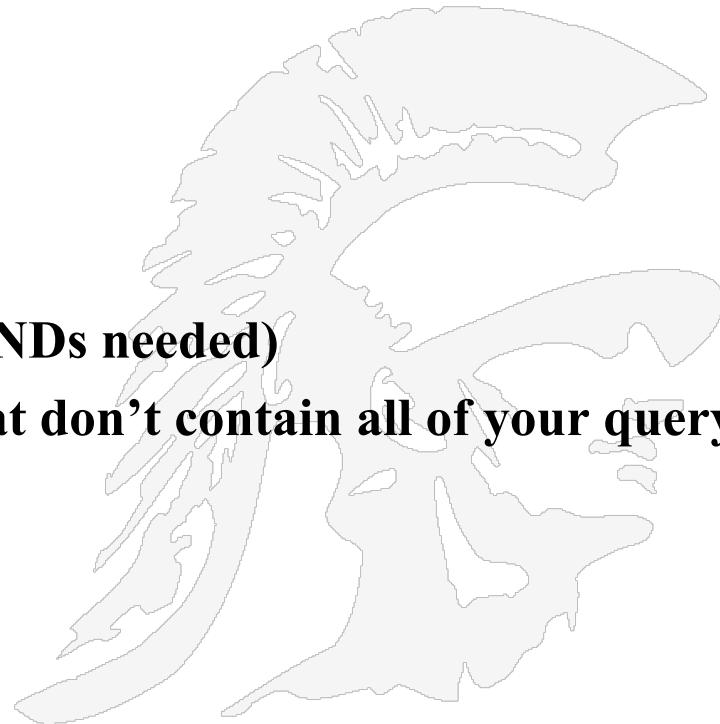
Apple Farm Collections
 No reviews · Farm
 Sherman Oaks, CA · (818) 788-1529

Whalley's Orchard
 4.4 ★★★★★ (50) · Produce market
 Littlerock, CA · (661) 480-6454
 Open · Closes 5PM
 "They also have the best Pluots, grapes, apples, and peaches."

Directions Directions

Query Box Default is AND

- If you search for more than one keyword at a time, Google/Bing will automatically search for pages that contain **ALL** of your keywords
 - This is called “implicit AND”
- A search for
disney disneyland pirates
is the same as searching for
disney AND disneyland AND pirates
(without the ANDs needed)
- Google sometimes returns pages that don't contain all of your query terms

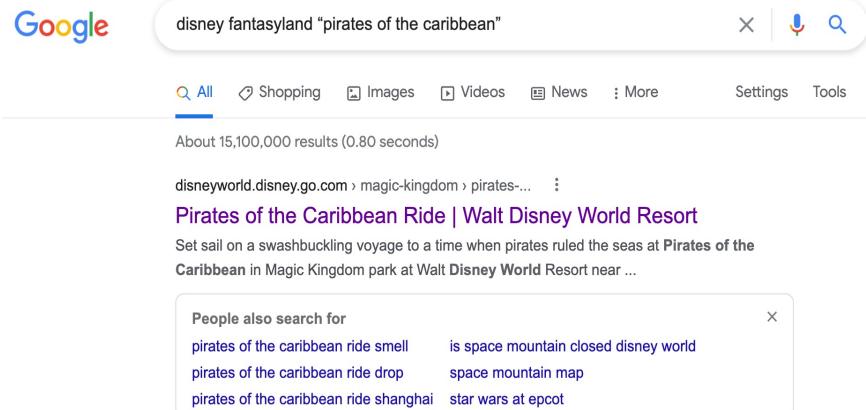


Put Exact Phrases in Quotes

- To search for phrases, just put your phrase in quotes.
- For example,

disney disneyland “pirates of the caribbean”

- This would show you all the pages in Google's index that contain the word **disney** AND the word **disneyland** AND the exact phrase **“pirates of the Caribbean”** (of course, without the quotes)



Google

disney fantasyland “pirates of the caribbean”

All Shopping Images Videos News More Settings Tools

About 15,100,000 results (0.80 seconds)

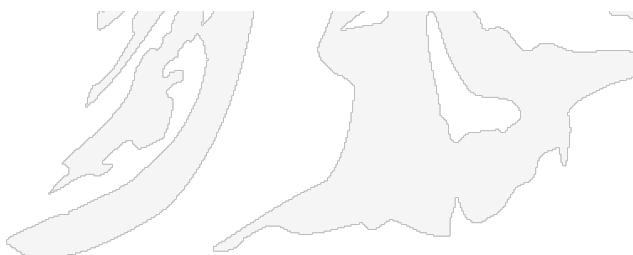
disneyworld.disney.go.com › magic-kingdom › pirates-... ::
Pirates of the Caribbean Ride | Walt Disney World Resort
Set sail on a swashbuckling voyage to a time when pirates ruled the seas at Pirates of the Caribbean in Magic Kingdom park at Walt Disney World Resort near ...

People also search for

pirates of the caribbean ride smell	is space mountain closed disney world
pirates of the caribbean ride drop	space mountain map
pirates of the caribbean ride shanghai	star wars at epcot



en.wikipedia.org › wiki › Pirates_of_the_Caribbean_(at... ::
Pirates of the Caribbean (attraction) - Wikipedia
Pirates of the Caribbean is a pirate-themed Shoot the Chute attraction at Disneyland, Walt Disney World's Magic Kingdom, Tokyo Disneyland, Disneyland Park at Disneyland Paris, and Shanghai Disneyland.
Number of drops: Disneyland and Disneylan... Drop: 14 ft (4.3 m)
Theme: Pirates of the Caribbean movie serie... Audio-animatronics: 119
History · Modifications · Adaptations · Soundtrack



Differences With/Without Quotes

Query: general principles of electricity

 "general principles of electricity"

All Videos News Images Shopping More Settings Tools

About 189,000 results (0.55 seconds)

[PDF] [General principles of electricity supply systems](#)

www.idc-online.com/.../pdfs/.../General_principles_of_electricity_supply_systems.pdf ▾
The generator consists of a prime mover and a magnetic field excitor. The magnetic field is produced electrically by passing a direct current (d.c.) through a winding on an iron core, which rotates inside three-phase windings on the stator of the machine. The magnetic field is rotated by means of a prime mover which may be ...

[General principles of electricity supply systems | EEP | Transmission ...](#)

<https://www.pinterest.com/pin/35888128254768417/>
Electric Power Substation ... Volts, alternating, amperage, blackout, built, cable, communications, connection, construction, current, danger, distribution, electrical, electricity, energies, equipment, factory, fuel, fuse, generation, generator, high, in, industry, insulation, insulators, iron, lines, link, metal, multi-generation, nobody, ...

[articles - Mentor EBS](#)

www.mentor-ebs.si/ARTICLES ▾
To give the answer to this question, we need knowledge of the **general principles of electricity** pricing in the electricity market. What we have in mind here is the market price of electricity, i.e. the price at which electricity is sold to suppliers by producers, not the rate in the end-user market. The latter, naturally, indirectly has to ...

[Electricity - Guides.turnitin.com](#)

https://guides.turnitin.com/.../Prompt_Library/Secondary_Education ▾
Dec 19, 2017 - Prompt: Today you will research electricity and consider some of the methods used in science texts to support different purposes. First, you will read a passage that explains some **general principles of electricity**. Then you will read an article about what causes a short circuit. Finally, you will read an article ...

[PDF] [C:\Physics Dept History\1965Curriculum.wpd](#)

blogs.mtu.edu/physics/files/2000/01/Curriculum1965.pdf ▾
PH336. Electric and Magnetic Measurements. (0-3-3) s (alternate years) 4. A continuation of work begun in PH316-317, intended for those who wish to pursue further the theory and practice of precise electric and magnetic measurements. Types of all the principal instruments used in modern electrical methods are ...

With quotes

Copyright Ellis Horowitz 2011-2022

 general principles of electricity

All Videos News Images Shopping More Settings Tools

About 2,120,000 results (0.60 seconds)

Even materials that conduct **electricity** resist the flow of electrons. The unit of **electrical** resistance is an ohm. The pressure needed to make one coulomb per second (one ampere) flow through a conductor having a resistance of one ohm is one volt. ... The quantity of **electric** charge is measured in coulombs.

[The Principles of Electricity - Energizer](#)

www.energizer.com/science-center/the-principles-of-electricity

 About this result  Feedback

People also ask

[What is the law of electricity?](#)



[What is the basic principle of the electric generator?](#)



[What is the basic principle of magnetism?](#)



[What is the theory of electricity?](#)



[The Basic Principles of Electricity | Anixter](#)

https://www.anixter.com/en_us/resources/.../the-basic-principles-of-electricity.html ▾
The Volt. The pressure that is put on free electrons that causes them to flow is known as electromotive force (EMF). The volt is the unit of pressure, i.e., the volt is the amount of electromotive force required to push a current of one ampere through a conductor with a resistance of one ohm.

[The Principles of Electricity - Energizer](#)

www.energizer.com/science-center/the-principles-of-electricity ▾

Even materials that conduct **electricity** resist the flow of electrons. The unit of **electrical** resistance is an ohm. The pressure needed to make one coulomb per second (one ampere) flow through a conductor having a resistance of one ohm is one volt. ... The quantity of **electric** charge is measured in coulombs.



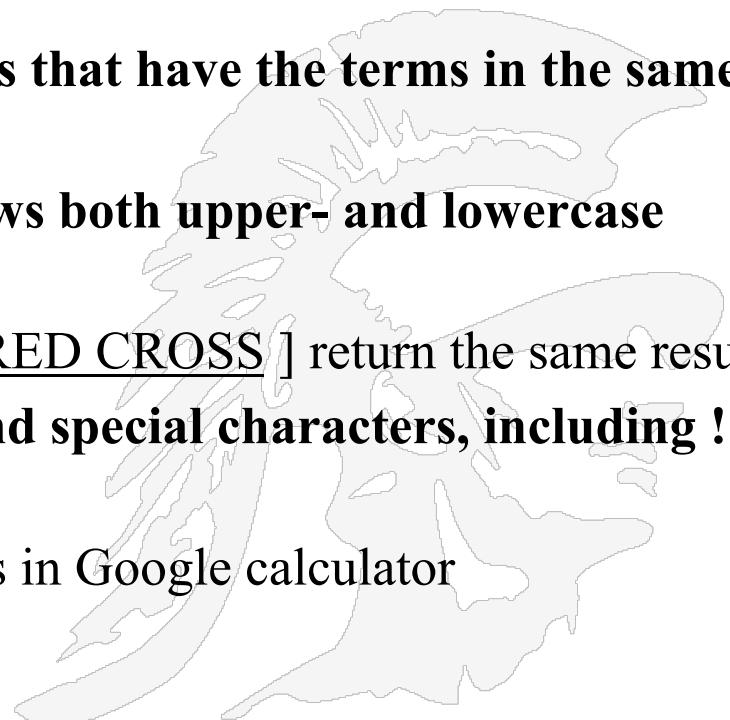
Without quotes

Google Stemming and Stop Words

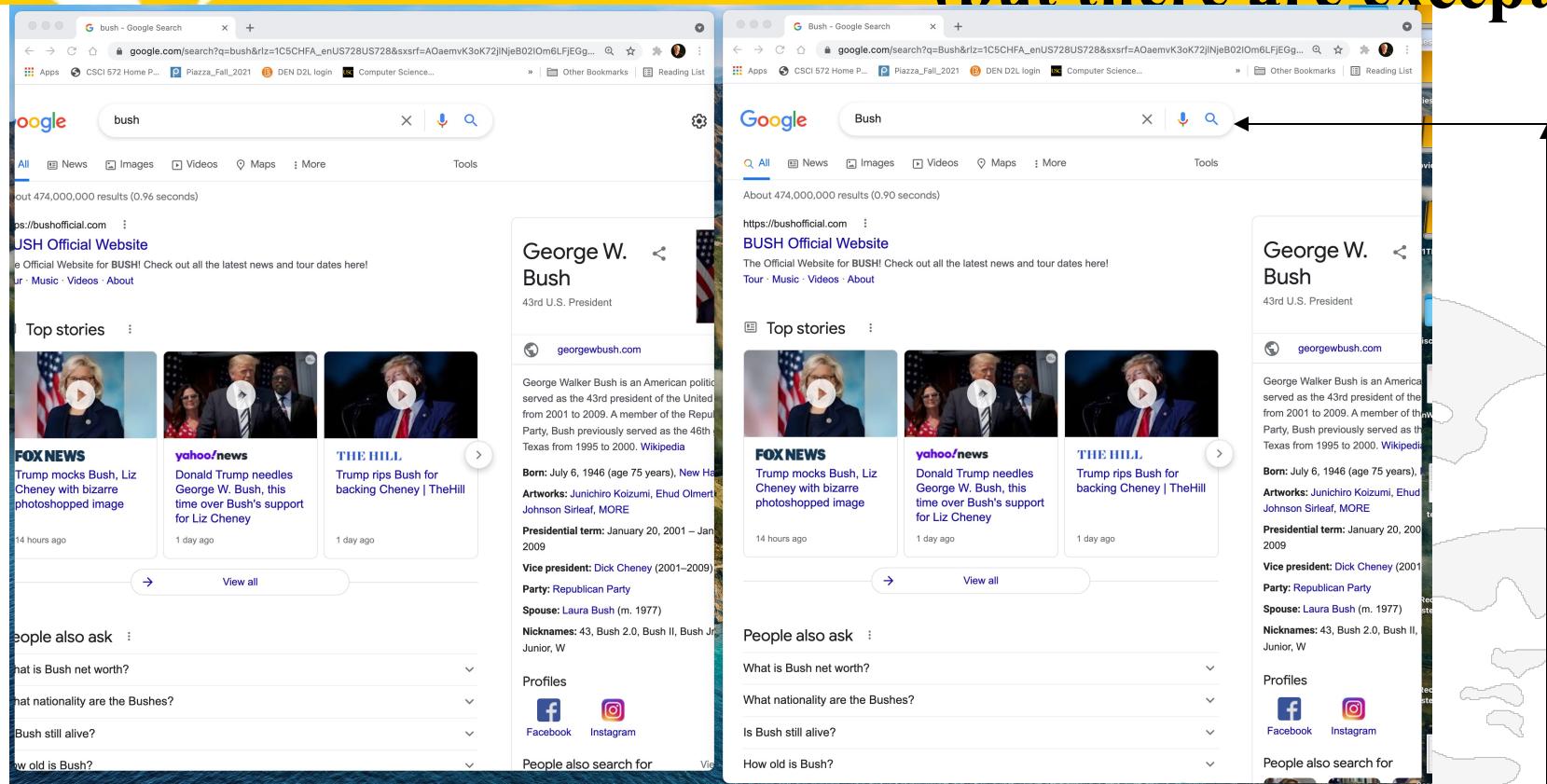
- The query [**child bicycle helmet**] finds pages that contain words that are similar to some or all of your search terms,
 - e.g., “child,” “children,” or “children's,” “bicycle,” “bicycles,” “bicycle's,” “bicycling,” or “bicyclists,” and “helmet” or “helmets.”
 - Google calls this feature *word variations* or *automatic stemming*
- Google will often ignore Stop Words
 - However, a query with only Stop Words, e.g. [the who] gets treated as significant, returning pages for the Rock Group, the Who
- Google limits queries to 32 words
 - Google will choke on the query
 - aardvark aback abacus abalone abandon abashed abbey abbreviate abdicate abdomen abduct aberration abhor abide ability abject able abnormal aboard abode abolish abolitionist abort about above abrade abridge abroad abrupt abscond absent absinthe

Other Google Query Rules

- **Google favors results that have your search terms near each other**
 - The query [snake grass] finds pages about plants; [snake in the grass] finds pages about sneaky people
- **Google gives higher priority to pages that have the terms in the same order as in your query**
- **Google is NOT case sensitive; it shows both upper- and lowercase results**
 - [Red Cross], [red cross], and [RED CROSS] return the same results.
- **Google ignores some punctuation and special characters, including ! ? , . ; [] @ / # < >**
 - Exceptions: C++, or math symbols in Google calculator



Capitalization Does NOT Matter (but there are exceptions)



The image displays three side-by-side Google search results for the query "bush". Each result shows the same search interface with a search bar containing "bush", a results count of "About 474,000,000 results (0.96 seconds)", and a snippet for the "BUSH Official Website". Below the snippet are "Top stories" from FOX NEWS, yahoo/news, and THE HILL, and a "People also ask" section. The results are identical for both "bush" and "Bush".

bush and Bush yield the same results
as does “apple” and “Apple”

More Exceptions

Two screenshots of Google search results for "disney" and "Disney".

Left Screenshot (Google Search for "disney"):

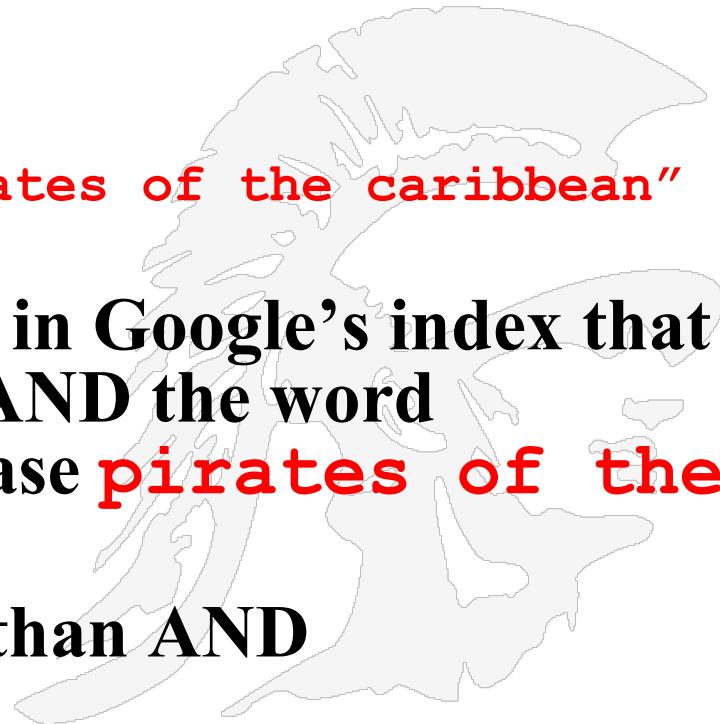
- Search bar: disney
- Results count: About 1,620,000,000 results (1.06 seconds)
- Ad: Disney+ - Start Streaming
- Text: Classic favorites. New releases. Disney Channel throwbacks. Exclusive Original series. From classics to new releases, get access to stories you can't stream anywhere else.
- Link: Disney+
- Text: Entretenimiento Sin Fin Suscibete Ahora
- Link: Disney+, Hulu, and ESPN+
- Text: The Ultimate Streaming Trio There's something for everyone.
- Link: Marvel Movie Home
- Text: Captain America, Thor, and More! Stream Your Favorites Today!
- Link: Stream Disney Throwbacks
- Text: The Vault is Wide Open. Your Disney Favorites Await.
- Link: https://www.disney.com
- Text: Disney.com | The official home for all things Disney
- Text: The official website for all things Disney: theme parks, resorts, movies, tv programs, characters, games, videos, music, shopping, and more!
- Link: Video · Princess · Disney News · The Cast of Disney's Jungle...
- Link: https://disneyworld.disney.go.com
- Text: Walt Disney World Resort in Orlando, Florida
- Text: Welcome to Walt Disney World. Come and enjoy the magic of Walt Disney World Resort in Orlando, FL. Plan your family vacation and create memories for a ...

Right Screenshot (Google Search for "Disney"):

- Search bar: Disney
- Results count: About 1,520,000,000 results (1.16 seconds)
- Link: https://www.disney.com
- Text: Disney.com | The official home for all things Disney
- Text: The official website for all things Disney: theme parks, resorts, movies, tv programs, characters, games, videos, music, shopping, and more!
- Link: Video · Princess · Disney News · The Cast of Disney's Jungle...
- Link: https://disneyworld.disney.go.com
- Text: Walt Disney World Resort in Orlando, Florida
- Text: Come and enjoy the magic of Walt Disney World Resort in Orlando, FL. Plan your family vacation and create memories for a lifetime.
- Section: Top stories
- Card 1: THE VERGE · Disney sues for control of key Marvel characters · 2 hours ago
- Card 2: The New York Times · Disney Sues to Keep Complete Rights to Marvel Characters · 2 hours ago
- Card 3: THE WALL STREET JOURNAL · Disney Isn't Investigating Handling of Sexual-Assault Allegations, After ABC News Boss Urge... · 1 day ago
- Section: Films produced

Boolean OR

- The Boolean **OR** operator is acceptable in Google queries, placed between keywords, and **OR** is *always* in all caps
- For example,
disney disneyland OR “pirates of the caribbean”
- This would show you pages in Google’s index that contain the word **disney** AND the word (**disneyland** OR the phrase **pirates of the caribbean**)
- OR has higher precedence than AND

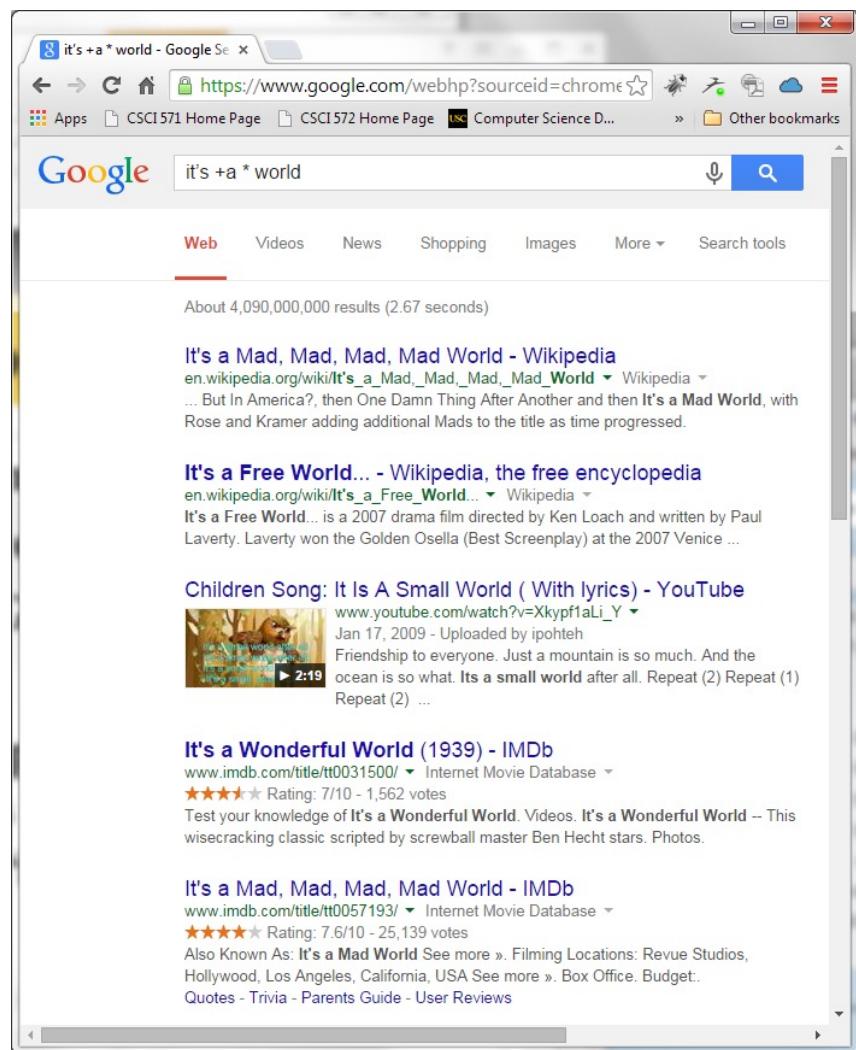


How Google/Bing Treat Queries with Boolean Operators

- All query terms are implicitly ANDed
- OR has higher precedence than AND
- Three examples (a, b, c stand for query terms):
 1. a b OR c d is treated as a AND (b OR c) AND d
 2. a OR b c OR d is treated as (a OR b) AND (c OR d)
 3. a OR “b c” d is treated as (a OR (“b c”)) AND d
- and see
 - <https://support.google.com/websearch/answer/2466433?hl=en>
 - <https://help.bing.microsoft.com/#apex/bing/en-US/10002/-1>

- Google offers full-word wildcard queries
- For example, if you search Google for **it's +a * world**,
- Google shows you all of the pages in its database that contain the phrase, e.g. “it's a small world” ... and “it's a nano world” ... and “it's a Linux world” ... and so on
- The + before **a** is required because it is a stop word and would otherwise be ignored
- This query works the same in Bing
- There must be a space after the +a

Google and Wildcards



Google Advanced Operators

Query modifiers

- **daterange:**
- **filetype:**
- **inanchor:**
- **intext:**
- **intitle:**
- **inurl:**
- **site:**

Alternative query types

- **cache:**
- **link:**
- **related:**
- **info:**

Other information needs

- **stocks:**

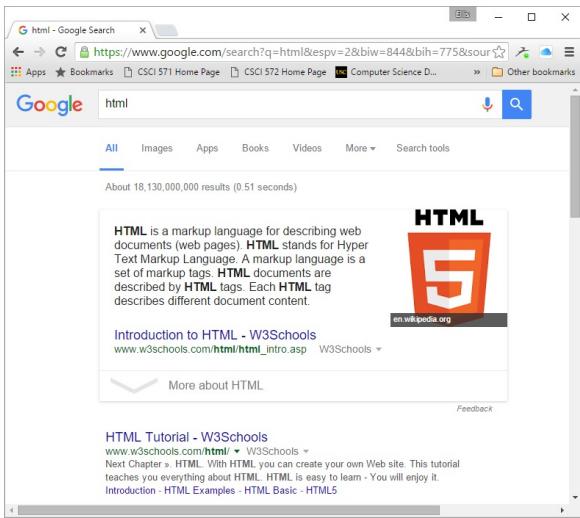
Search Service	Search Operators
Web Search	<u>allinanchor:</u> , <u>allintext:</u> , <u>allintitle:</u> , <u>allinurl:</u> , <u>cache:</u> , <u>define:</u> , <u>filetype:</u> , <u>id:</u> , <u>inanchor:</u> , <u>info:</u> , <u>intext:</u> , <u>intitle:</u> , <u>inurl:</u> , <u>link:</u> , <u>related:</u> , <u>site:</u>
Image Search	<u>allintitle:</u> , <u>allinurl:</u> , <u>filetype:</u> , <u>inurl:</u> , <u>intitle:</u> , <u>site:</u>
Groups	<u>allintext:</u> , <u>allintitle:</u> , <u>author:</u> , <u>group:</u> , <u>insubject:</u> , <u>intext:</u> , <u>intitle:</u>
Directory	<u>allintext:</u> , <u>allintitle:</u> , <u>allinurl:</u> , <u>ext:</u> , <u>filetype:</u> , <u>intext:</u> , <u>intitle:</u> , <u>inurl:</u>
News	<u>allintext:</u> , <u>allintitle:</u> , <u>allinurl:</u> , <u>intext:</u> , <u>intitle:</u> , <u>inurl:</u> , <u>location:</u> , <u>source:</u>
Product Search	<u>allintext:</u> , <u>allintitle:</u>

http://www.googleguide.com/advanced_operators_reference.html

See also

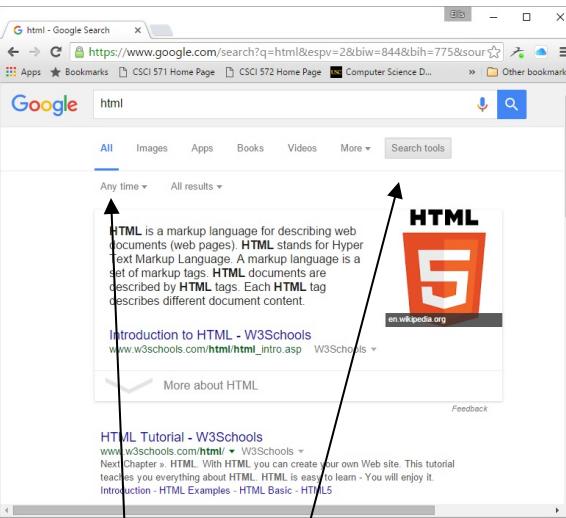
<https://help.bing.microsoft.com/#apex/bing/en-US/10001/-1>

Searching by Date



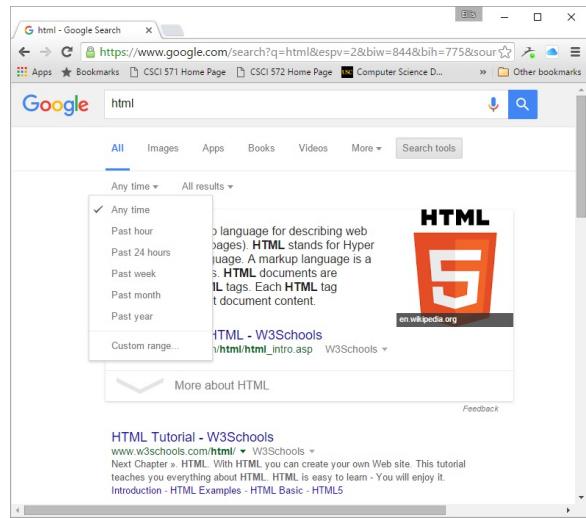
Google search results for "html". The first result is "Introduction to HTML - W3Schools" from www.w3schools.com/html/html_intro.asp. The page content describes HTML as a markup language for web documents.

1. Google search for "html"



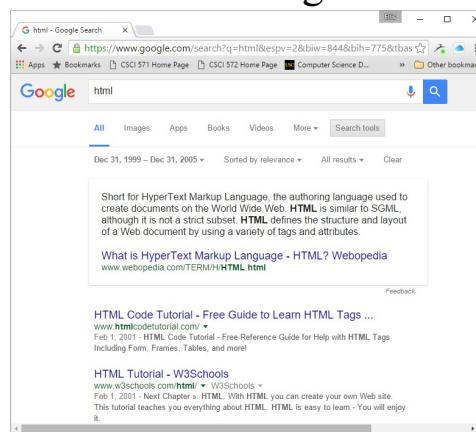
Google search results for "html". The "Search tools" dropdown is open, and the "Any time" option is selected. The results remain the same as in the previous screenshot.

2. click on "Search Tools" and "Any Time" appears



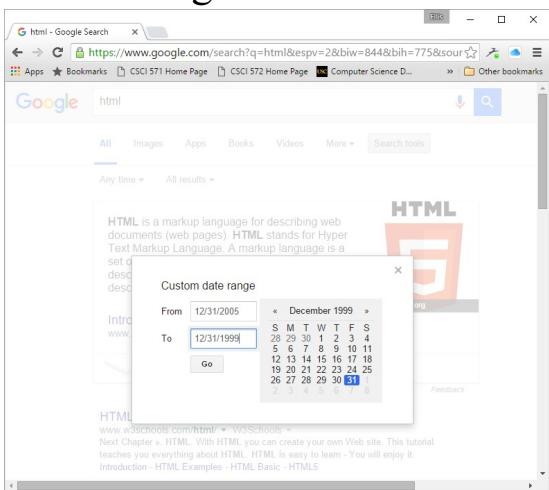
Google search results for "html". The "Search tools" dropdown is open, and the "Custom range" option is selected. A date range calendar is displayed, showing the month of December 1999. The "From" field is set to 12/31/2005 and the "To" field is set to 12/31/1999.

3. click on Any Time and select Custom Range



Google search results for "html" filtered by the date range 12/31, 1999 – Dec 31, 2005. The results show a single entry: "What is HyperText Markup Language - HTML? Webopedia" from www.webopedia.com/TERM/H/HTML.html. The page content describes HTML as the authoring language used to create documents on the World Wide Web.

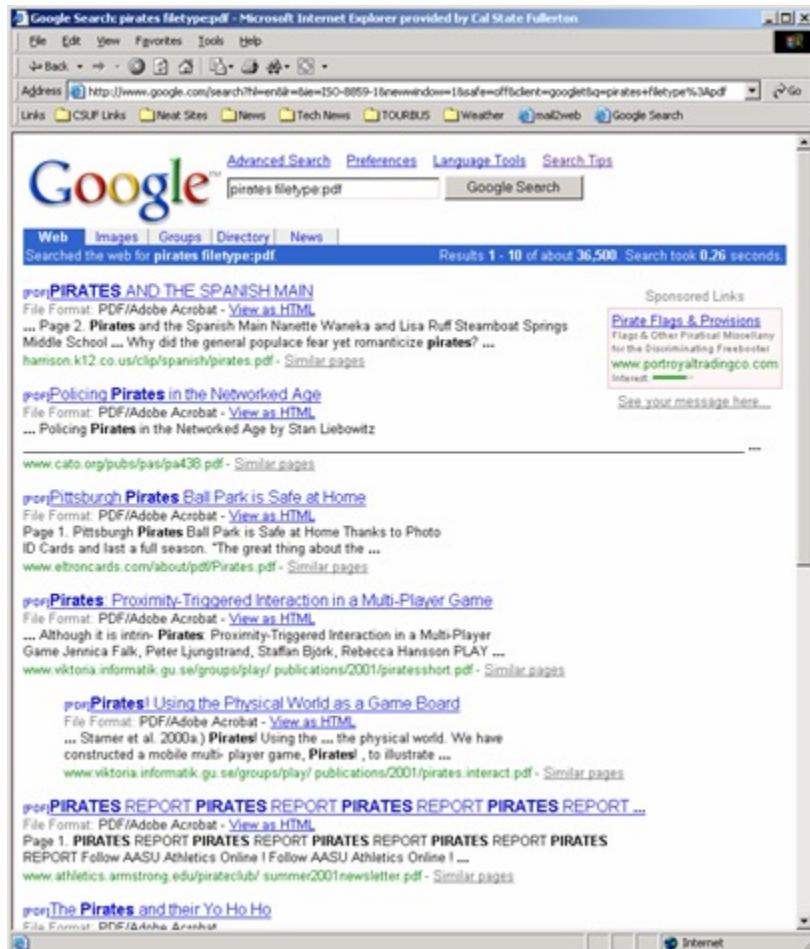
4. calendar appears, enter dates



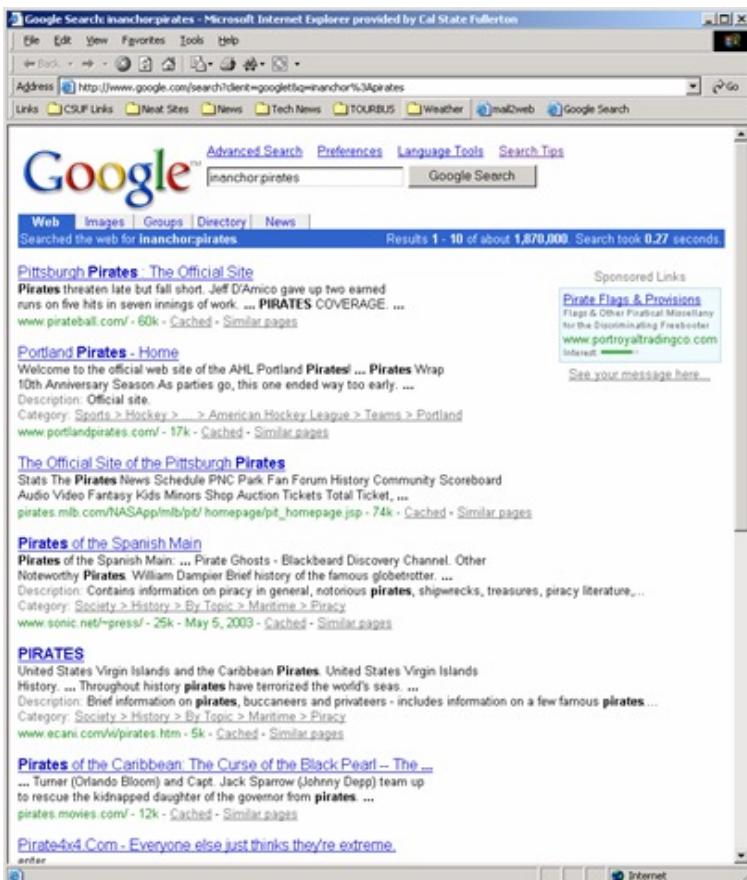
Google search results for "html". The "Search tools" dropdown is open, and the "Custom date range" option is selected. A date range dialog is displayed, showing "From 12/31/2005" and "To 12/31/1999".

5. Final Result

filetype:



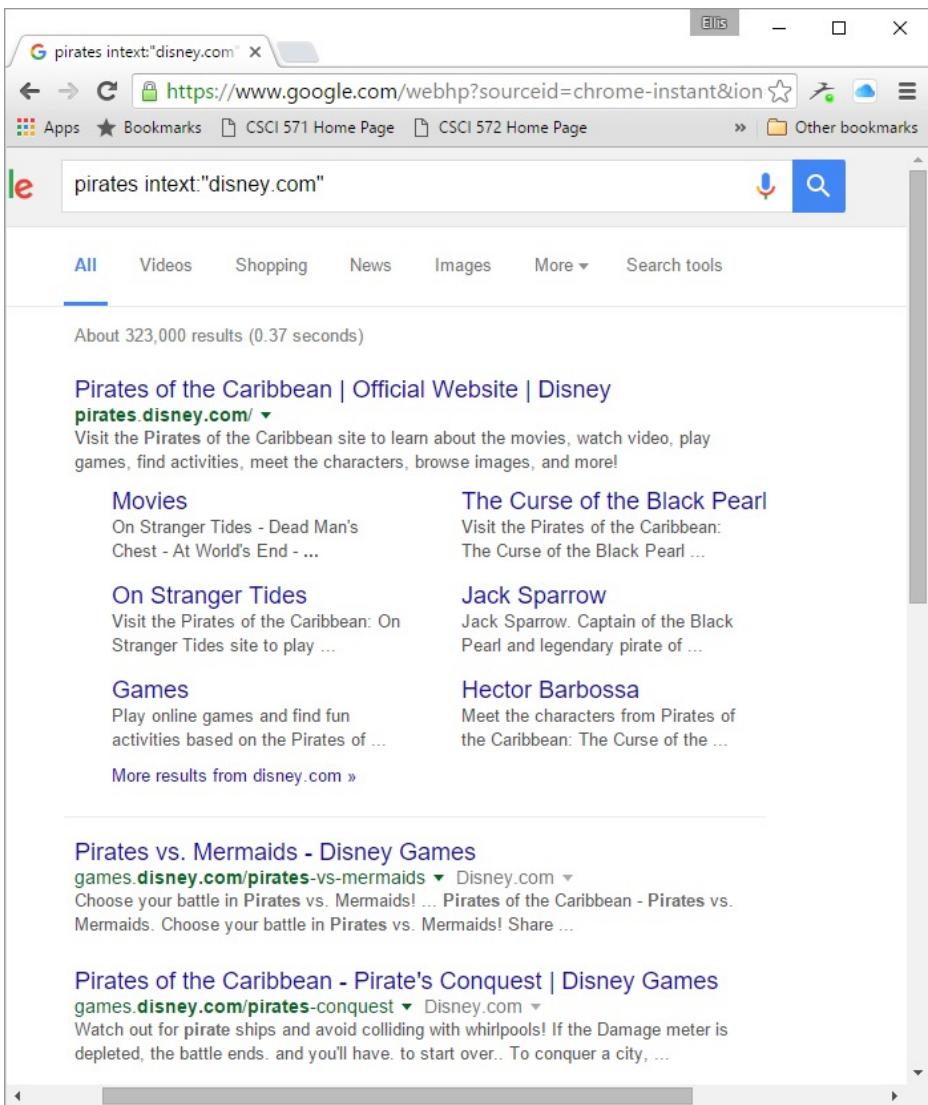
- **filetype:** restricts your results to files ending in a specific file suffix, e.g ".doc" (or .xls, .ppt. etc.), and shows you only files created with the corresponding program
- There can be no space between **filetype:** and the file extension
- The “dot” in the file extension – .doc – is optional
- filetype:doc will NOT return docx



inanchor:

- **inanchor:** will restrict the results to pages containing the query terms you specify in the anchor text or links to the page.
- For example,
[restaurants inanchor:gourmet]
- will return pages in which the anchor text on links to the pages contain the word “gourmet” and the page contains the word “restaurants.”
- **allinanchor:** restricts results to pages containing all query terms you specify in the anchor text on links to the page.

For example, [allinanchor: best museums sydney] will return only pages in which the anchor text on links to the pages contain the words “best,” “museums,” and “sydney.”



A screenshot of a web browser window titled "Ellis". The address bar shows the query "pirates intext:"disney.com"". The search results page displays the following content:

About 323,000 results (0.37 seconds)

Pirates of the Caribbean | Official Website | Disney
pirates.disney.com/ ▾
Visit the Pirates of the Caribbean site to learn about the movies, watch video, play games, find activities, meet the characters, browse images, and more!

Movies
On Stranger Tides - Dead Man's Chest - At World's End - ...

On Stranger Tides
Visit the Pirates of the Caribbean: On Stranger Tides site to play ...

Games
Play online games and find fun activities based on the Pirates of ...
More results from disney.com »

Pirates vs. Mermaids - Disney Games
games.disney.com/pirates-vs-mermaids ▾ Disney.com ▾
Choose your battle in Pirates vs. Mermaids! ... Pirates of the Caribbean - Pirates vs. Mermaids. Choose your battle in Pirates vs. Mermaids! Share ...

Pirates of the Caribbean - Pirate's Conquest | Disney Games
games.disney.com/pirates-conquest ▾ Disney.com ▾
Watch out for pirate ships and avoid colliding with whirlpools! If the Damage meter is depleted, the battle ends. and you'll have. to start over.. To conquer a city, ...

intext:

- **intext:** ignores link text, URLs, and titles, and only searches body text.
- **intext:** helps you avoid query words that are too common in URLs and links.
- **pirates intext:"Disney.com"** requires Disney.com to be within the body of the web page

intitle:

Google search results for "intitle:'pirates of the caribbean'"

About 74,200 results (0.82 seconds)

Videos

- Pirates of the Caribbean: Dead Men Tell No Tales - Official ...
YouTube · Walt Disney Studios Mar 2, 2017
- Pirates of the Caribbean: Dead Men Tell No Tales - "Legacy ..."
YouTube · Walt Disney Studios Apr 26, 2017
- 4 key moments in this video
- Pirates of the Caribbean: Dead Men Tell No Tales: Extended ...
YouTube · Walt Disney Studios Feb 5, 2017

<https://www.imdb.com/title/tt4154694/>

Pirates of the Caribbean: Dead Men Tell No Tales (2017) - IMDb
 Captain Jack Sparrow is pursued by old rival Captain Salazar and a crew of deadly ghosts who have escaped from the Devil's Triangle.
 ★★★★☆ Rating: 6.5/10 · 280,585 votes

<https://www.imdb.com/title/tt2423976/>

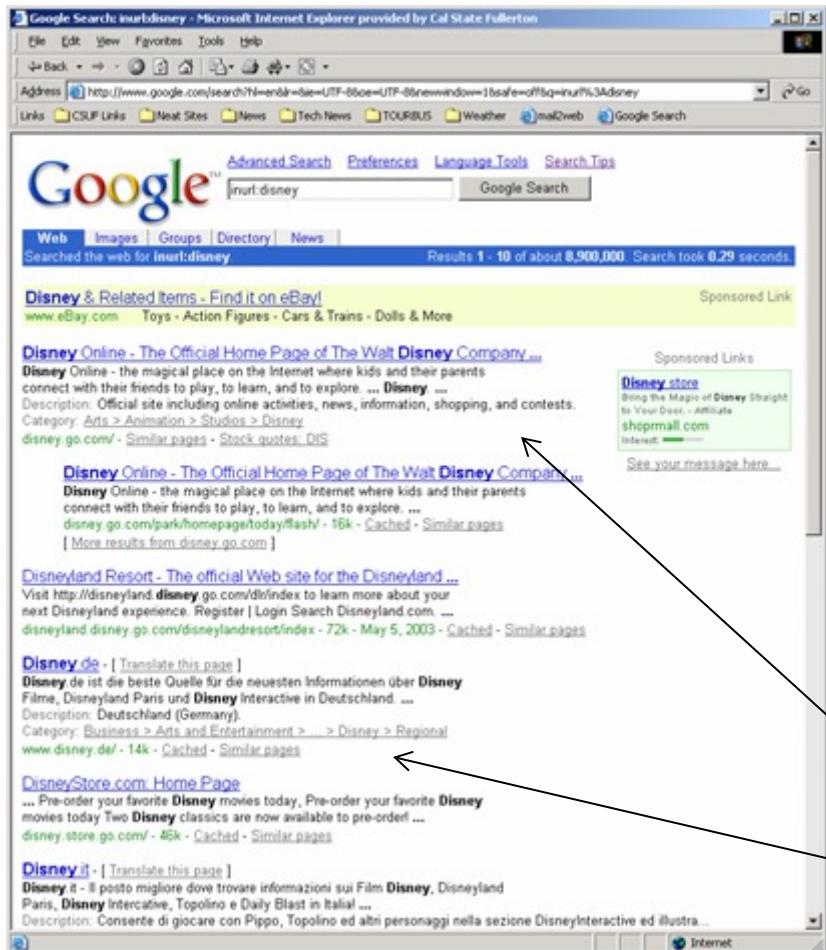
Pirates of the Caribbean: On Stranger Tides (2011) - IMDb
 Jack Sparrow and Barbossa embark on a quest to find the elusive fountain of youth, only to discover that Blackbeard and his daughter are after it too.
 ★★★★☆ Rating: 6.6/10 · 497,456 votes

Missing: intitle:piratesofthecaribbean" | Must include: intitle:piratesofthecaribbean"

- **intitle:** restricts the results to documents containing a particular word in its title.
- You can also search for phrases. Just put your phrase in quotes
- **intitle:pirates**
- **Intitle:"pirates of the caribbean"**

inurl:

- **inurl:** restricts the results to documents containing a particular word in its URL.
- **inurl:disney**
- Results include
 - Disney.go.com
 - www.disney.de



site:

Google

masters site:cs.usc.edu

[All](#)[News](#)[Images](#)[Videos](#)[Maps](#)[More](#)[Settings](#)[Tools](#)

About 668 results (0.50 seconds)

[www.cs.usc.edu](#) › Academic Programs**M.S. Program - USC Viterbi | Department of Computer Science**

The Master of Science in Computer Science provides intensive preparation in the concepts and techniques related to the design, programming, and application ...

[www.cs.usc.edu](#) › Academic Programs › M.S. Program**Computer Science (General) - USC Viterbi | Department of ...**

The Master of Science in Computer Science provides intensive preparation in the basic concepts and techniques related to the design, programming and ...

[www.cs.usc.edu](#) › Academic Programs › M.S. Program**Intelligent Robotics - USC Viterbi | Department of Computer ...**

The Master of Science in Computer Science (Intelligent Robotics) educates students on the design, construction, operation, and application of robots, as well as ...

[www.cs.usc.edu](#) › Academic Programs › M.S. Program**Data Science - USC Viterbi | Department of Computer Science**

The Master of Science in Computer Science (Data Science) provides students with a core background in Computer Science and specialized algorithmic, ...

You've visited this page 2 times. Last visit: 9/9/20

[www.cs.usc.edu](#) › Academic Programs › M.S. Program**Software Engineering - USC Viterbi | Department of Computer ...**

The Master of Science in Computer Science (Software Engineering) focuses on providing its graduates not only software development skills, but also systems ...

[www.cs.usc.edu](#) › Academic Programs › M.S. Program**Game Development - USC Viterbi | Department of Computer ...**

The Master of Science in Computer Science (Game Development) program graduates students with a strong foundation in computer science, ...

You've visited this page 2 times. Last visit: 8/12/19

- **site:** restricts the results to those websites in a domain.
- There can be no space between **site:** and the domain.
- Query is:
- **masters site:cs.usc.edu**

Using site:

- You can use **site:** in conjunction with another search term or phrase.
pirates site:disney.com
- You can also use **site:** and negation to exclude sites.
pirates -site:disney.com
- You can use **site:** to exclude or include entire top level domains (and, like with filetype, the dot is optional).
pirates -site:com
pirates site:edu



G USC Viterbi | Department of C x +
Not Secure | webcache.googleusercontent.com/search?q=cache%3Awww.cs.usc.edu&q=cache%3Awww.cs.usc.edu&aqs=chrome.....

Apps CSCI 572 Home P... Piazza Spring2022 CSC1672_Spring2... DEN D2L Page USC Schedule of... Computer Science... Other Bookmarks Reading List

This is Google's cache of <https://www.cs.usc.edu/>. It is a snapshot of the page as it appeared on Feb 15, 2022 08:53:41 GMT. The [current page](#) could have changed in the meantime. [Learn more](#).

[Full version](#) [Text-only version](#) [View source](#)

Tip: To quickly find your search term on this page, press **Ctrl+F** or **⌘-F** (Mac) and use the find bar.

≡ USC Viterbi



Department of
Computer Science



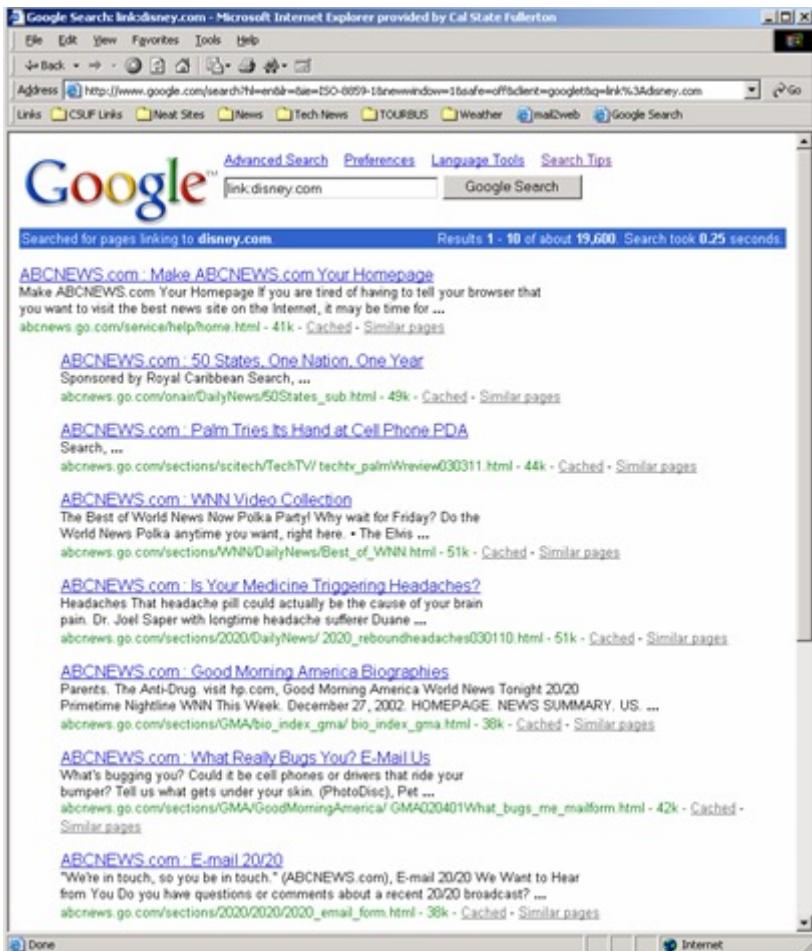
"Deepfaking the Mind" Could Improve Brain-Computer Interfaces for People with Disabilities

cache:

- **cache:url** shows the version of a web page that Google has in its cache.
- Getting to the cached page has changed multiple times



link:



- **link:** restricts the results to those web pages that have links to the specified URL.
 - **link:Disney.com**
 - Note: apparently the link operator only returns a sample of web pages pointing to the link

related:

Google related:www.cs.usc.edu

All Images Maps Shopping More Search tools

About 45 results (0.21 seconds)

- CS | Computer Science**
www.cs.ucla.edu/ ▾
General information about the department and contacts.
- Department of Computer Science, Columbia University | Home**
www.cs.columbia.edu/ ▾
Established in 1979, the Department of Computer Science is located within the tree-lined Morningside campus on the Upper West Side of Manhattan.

- Donald Bren School of Information and Computer Sciences ...**
www.ics.uci.edu/ ▾
Department of Information and Computer Science. Teaching and research information.
- Computer Science - Computer Science**
www.cs.ucdavis.edu/ ▾
Course descriptions, course schedules, job positions, events, newsletter, and contact information.

- Stanford Computer Science**
www-cs.stanford.edu/ ▾
Stanford, California.

- Computer Science Department at Princeton University**
https://www.cs.princeton.edu/ ▾
The official website for Princeton University's Computer Science Department. Information on programs offered, news, events, and more.

- Computer Science and Engineering: Welcome to CSE @ UCR**
www.cs.ucr.edu/ ▾
Department of Computer Science and Engineering (CSE)

- UCSB Computer Science**
https://www.cs.ucsb.edu/ ▾
General information about the department and contacts.

- **related:** lists web pages that are "similar" to a specified web page.
- There can be no space between **related:** and the URL.

Google Search: related:disney.com - Microsoft Internet Explorer provided by Cal State Fullerton

File Edit View Favorites Tools Help

Address http://www.google.com/search?hl=en&rlz=1&sa=GO&tbo=q&client=google&q=related%3Adisney.com

Links CSUF Links Net Sites News Tech News TOURBUS Weather mail2web Google Search

Google Searched for pages similar to disney.com Results 1 - 10 of about 20 Search took 0.14 seconds

Disney Online - Where the Magic Comes to You!
Disney Online - the magical place on the Internet where kids and their parents connect with their friends to play, to learn, and to explore.
Description: Official site including online activities, news, information, shopping, and contests.
Category: Arts > Animation > Studios > Disney
disney.go.com/ ▾ Similar pages Stock quotes DIS

Yahooligans!
Yahooligans. Thursday March 20, 2003. Games. Animals. E-Cards. Movies. Jokes. Science. Reference. Ask Earl. News. Sports. Astrology. Cool Sites. Parents' Guide. ...
Description: Featuring comprehensive safe surfing, games, homework help, and many kid friendly activities.
Category: Kids & Teens > Directions
www.yahooligans.com/ ▾ 26k - Cached - Similar pages

Nickelodeon Online at Nick.com
What should Brent and Candace do while wearing HUGE underwear? Deliver the mail. Photocopy their butts Ride the elevator. Got NickPoints? Get e-Collectibles! ...
Description: The official site. Nickelodeon TV stuff, hot games, cool jokes and celebrity gossip. Check out favorite...
Category: Arts > Television > Networks > Cable > Nickelodeon
www.nick.com/ ▾ 43k - Cached - Similar pages

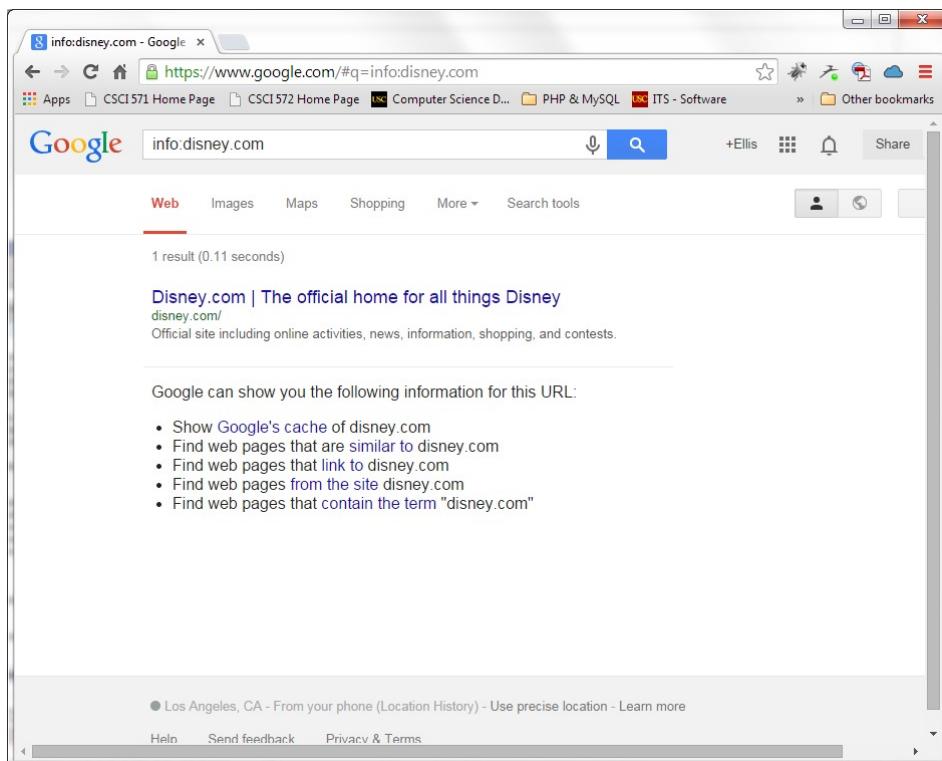
Paramount Pictures
-Pick a Show- ...
Description: Feature film production and distribution, video and DVD worldwide distribution, and production of...
Category: Arts > Movies > Studios
www.paramount.com/ ▾ 12k - Cached - Similar pages

Welcome to Discovery.com
Discovery Channel Store Bestseller! TV Schedules | HD Theater | Newsletter / My Discovery | Discovery School | International Sites ...
Description: Discovery Online
Category: Arts > Television > Networks > Cable > Discovery Channel
www.discovery.com/ ▾ 11k - Cached - Similar pages

MGM New and Upcoming Movies Trailers
... new DVD releases
Category: Business > Major Companies > Publicly Traded > M
www.mgm.com/ ▾ 41k - Cached - Similar pages Stock quotes: MGM

26

info:



- **info:** presents some information that Google has about a particular web page.

stocks:

Google search results for "stocks:aapl". The top result is for Apple Inc. (NASDAQ: AAPL). The page shows the market summary for Apple Inc., including the current price of 170.74 USD, a 1.10% increase today, and a 10:36 AM EST update. It also displays a 1-day chart from 10:00 AM to 4:00 PM, showing a price of 170.83 USD at 10:36 AM. Below the chart are key financial metrics: Open (170.97), High (171.45), Low (170.25), Mkt cap (2.79T), P/E ratio (28.36), Div yield (0.52%), GDP score (A-), and 52-wk high (182.94) and low (116.21). To the right, there is an "About" section with links to apple.com, information about the company's history, leadership, and financial details.

- If you begin a query with **stocks:** Google will treat the rest of the query terms as stock ticker symbols, and will link to a finance page showing stock information for those symbols.

Even More Special Features of the Google Query Box

- **Math expressions are evaluated:** $12 + 34 + 10 * (150 / 7) = 260.285714$
- **Dictionary definitions:** define:antidisestablishmentarianism
- Put @ in front of a word to search social media. For example: @twitter.
- Put \$ in front of a number and search for a price. For example: camera \$400.
- Put .. between two numbers and search in a range.
For example, camera \$50..\$100.
- Put a valid tracking number from FedEx or UPS and it will take you to the tracking site
- Put a valid airline and flight number and it will give you its status
- Put a tilde, ~car repair and it queries on ALL synonyms of car, like auto
 - See the following links to further discussion of Google's operators
 - <http://www.google.com/support/websearch/bin/answer.py?hl=en&answer=136861>
 - http://www.googleguide.com/advanced_operators_reference.html
 - <http://searchengineland.com/google-power-user-tips-query-operators-48126>

Even More Things One Can Do with Google

Google search results for the query "53493439531=english". The results page shows the number 53,493,439,531 followed by its verbal representation: "fifty-three billion four hundred ninety-three million four hundred thirty-nine thousand five hundred thirty-one".

Speaking/writing out a number

Google search results for the query "what's my ip". The results page shows the user's public IP address: 172.248.32.30.

What is my IP address

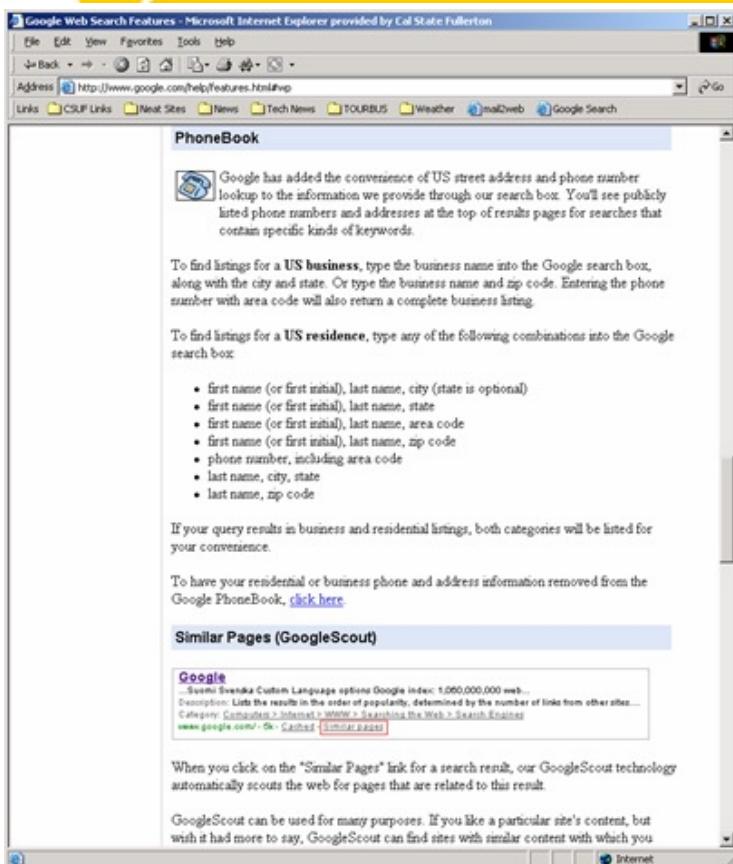
Google search results for the query "mortgage calculator". The results page displays a mortgage calculator tool. The input fields show a mortgage amount of \$100,000, an interest rate of 3.92%, and a mortgage period of 30 years. The output fields show a total cost of mortgage of \$170,213 and monthly payments of \$473.

Mortgage calculator

Google search results for the query "spinner". The results page displays a spinner fidget tool. The spinner has five segments labeled 1, 2, 3, 4, and 5. A yellow arrow points to the segment labeled 1. The tool includes a "Number" toggle switch and a "Fidget" toggle switch.

Spinners for Games

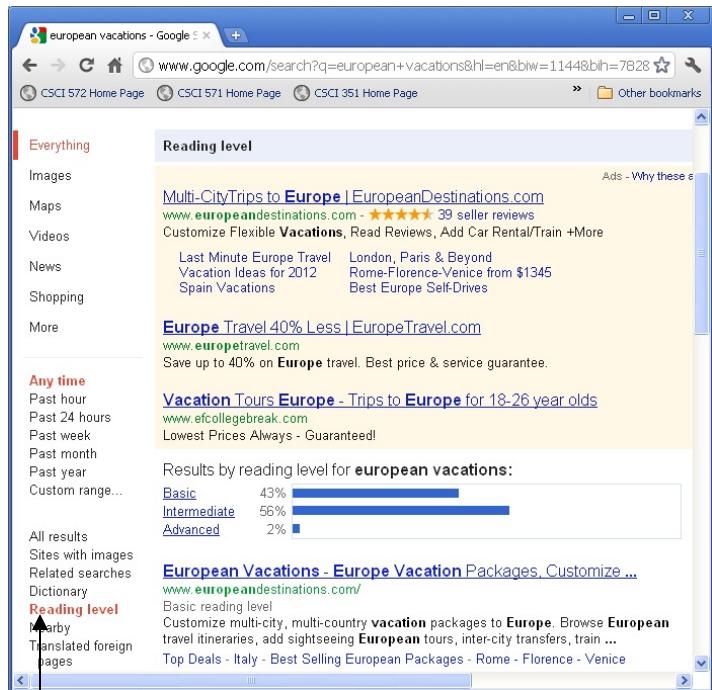
A Failed Google Experiment - phonebook: operator



- There were actually three different Google phonebook operators.
- **phonebook:** searches the entire Google phonebook.
- **rphonebook:** searches residential listings only.
- **bphonebook:** searches business listings only

As of 2010, Google's phone book feature has been officially retired. Both the phonebook: and the rphonebook: search operator have both been dropped due to many complaints about privacy violations

A Failed Google Experiment - Reading Level Examples

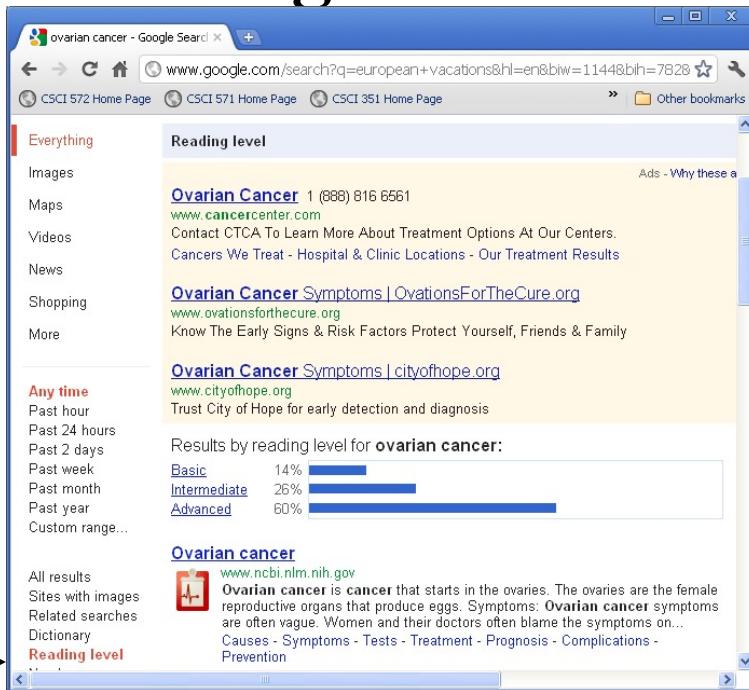


The screenshot shows a Google search results page for "european vacations". On the left, there's a sidebar with filters like "Everything", "Images", "Maps", "Videos", "News", "Shopping", and "More". Below these are time filters: "Any time", "Past hour", "Past 24 hours", "Past week", "Past month", "Past year", and "Custom range...". Under "Reading level", it shows three categories: Basic (43%), Intermediate (56%), and Advanced (2%). The main search results list includes links for "Multi-CityTrips to Europe | EuropeanDestinations.com" and "Europe Travel 40% Less | EuropeTravel.com". At the bottom, there's a snippet about European vacation packages.

Query: European vacations

The feature is based primarily on statistical models built with the help of teachers. Google paid teachers to classify pages for different reading levels, and then took their classifications to build a statistical model. With this model, they can compare the words on any webpage with the words in the model to classify reading levels.

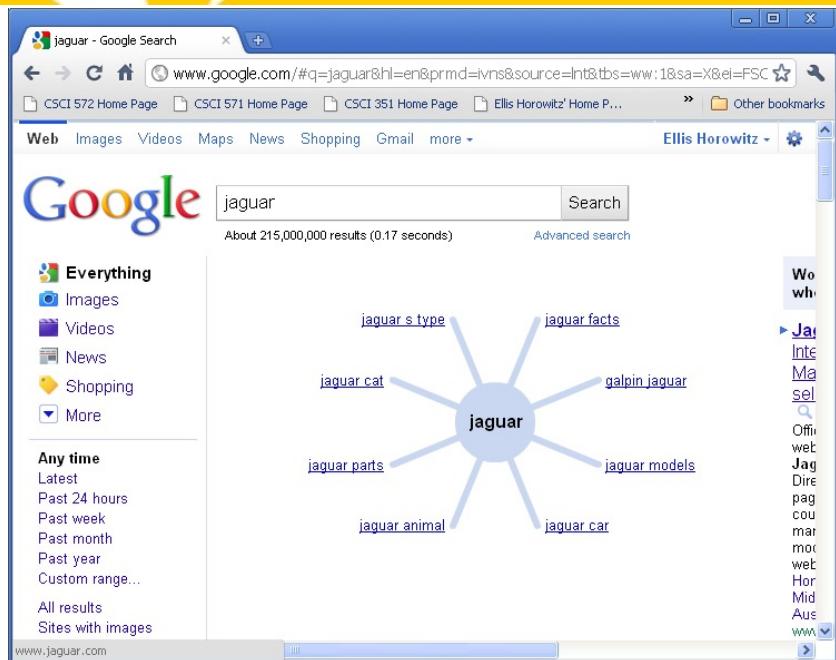
Google dropped this feature in 2015.



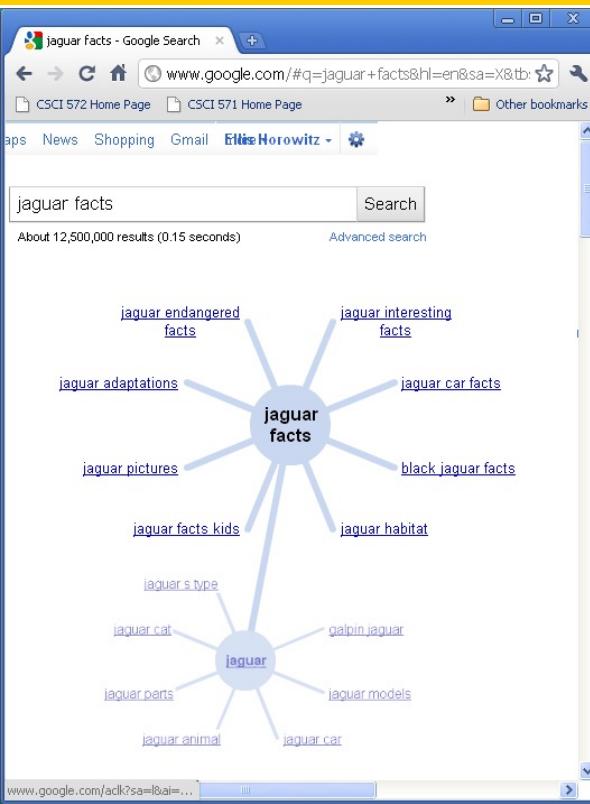
This screenshot shows a Google search results page for "ovarian cancer". The sidebar and filters are identical to the first screenshot. The "Reading level" section shows results for "ovarian cancer": Basic (14%), Intermediate (26%), and Advanced (60%). The main results list includes links for "Ovarian Cancer" on the cancercenter.org website and "Ovarian Cancer Symptoms" on ovationsforthecure.org. A snippet from cityofhope.org discusses early detection and diagnosis.

Query: ovarian cancer

A Failed Google Experiment - The Wonder Wheel



Google puts up a wheel of possible interpretations including the jaguar car and the animal; clicking on the spokes of the wheel bring up a refined wheel which eventually lead to a modified query



In 2011 Google removed the Wonder Wheel

In 2009 Google introduced the Wonder Wheel, a flash-based interface

In 2011 Google removed the Wonder Wheel but provided no concrete explanation for why it did so;

In 2012 Google restored the wonder wheel, renaming it the Contextual Targeting Tool

In 2014 it was re-focused to help advertisers chose their keywords

A Failed Google Experiment - Google Code Search

Google Code Search

From Wikipedia, the free encyclopedia

Not to be confused with [Google Code](#).

Google Code Search was a free [beta](#) product from [Google](#) which debuted in [Google Labs](#) on October 5, 2006, allowing web users to search for open-source code on the Internet. Features included the ability to search using operators, namely `lang:`, `package:`, `license:` and `file:`.

The code available for searching was in various formats including `.tar.gz`, `.tar.bz2`, `.tar`, and `.zip`, [CVS](#), [Subversion](#), [git](#) and [Mercurial](#) repositories.

Contents [hide]

- 1 Regular expression engine
- 2 Discontinuation
- 3 See also
- 4 References
- 5 External links

Google Code Search

	labs
Developer(s)	Google
Initial release	October 5, 2006
Development status	Discontinued
Operating system	Any (web-based application)
Type	Code search engine
Website	www.google.com/codesearch

Regular expression engine [edit]

The site allowed the use of [regular expressions](#) in queries, which at the time was not offered by any other search engine for code.^[citation needed] This makes it resemble [grep](#), but over the world's public code. The methodology employed combines a [trigram index](#) with a custom-built, [denial-of-service](#) resistant [regular expression engine](#).^[1]

In March 2010, the code of [RE2](#), the regular expression engine used in Google Code Search, was made open source.^[2]

Google Code Search supported POSIX extended regular expression syntax, excluding back-references, collating elements, and collation classes.

Languages not officially supported could be searched for using the `file:` operator to match the common file extensions for the language.

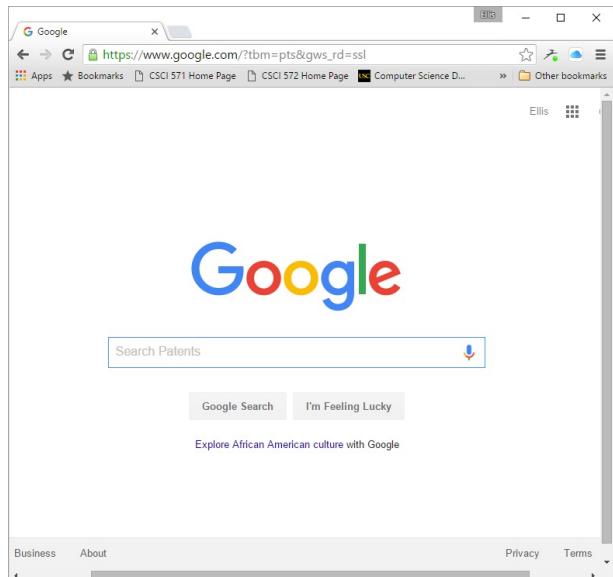
Discontinuation [edit]

In October 2011, Google announced that Code Search was to be shut down along with the Code Search API.^[3] The service remained online until March 2013,^[4] and it now returns a [404](#).

In January 2012, Russ Cox published an overview of history and the technical aspects of the tool, and open-sourced a basic implementation of a similar functionality as a set of standalone programs that can run fast indexed regular expression searches over local code.^[5]

Special Content Search Engines

Google Patents



Initial Google Patents Page

Document compression system and method for use with tokenspace repository
www.google.com/patents/US7917480
Grant - Filed Aug 13, 2004 - Issued Mar 29, 2011 - Jeffrey Dean - Google Inc.
Document compression system and method for use with tokenspace repository. US 7917480 B2. Abstract. The disclosed embodiments enable ...
[Overview](#) · [Related](#) · [Discuss](#)

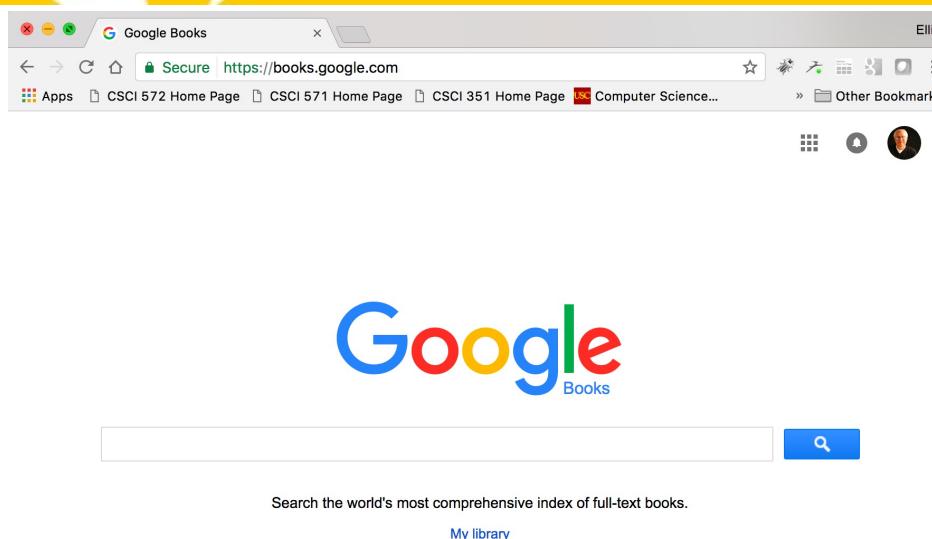
patent query
using patent
number

Publication number	US79
Publication type	Grant
Application number	US 10
Publication date	Mar 2
Filing date	Aug 1
Priority date	Aug 1
Fee status	Paid
Also published as	US83 US20
Inventors	Jeffre
Original Assignee	Googl
Export Citation	BIBTeX
Patent Citations	(11) Non-Patent

Sample patent results
page;
Note download

Special Content Search Engines

Google Books



Google Books is a service that searches the full text of books and magazines that Google has scanned, converted to text using optical character recognition (OCR), and stored in its digital database.

Books are provided either by

- publishers and authors, through the Google Books Partner Program, or by
- Google's library partners, through the Library Project.

Controversy:

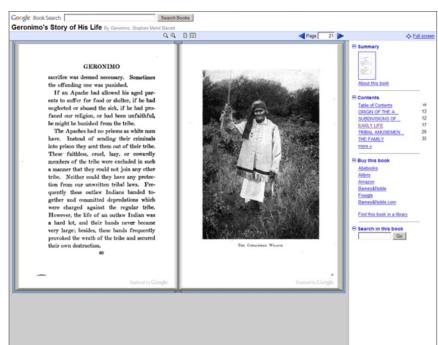
Google has been criticized for potential copyright violations, and lack of editing to correct the many errors introduced into the scanned texts by the OCR process.

As of October 2015, the number of scanned book titles was over 25 million.

Google estimated in 2010 that there were about 130 million distinct titles in the world, and stated that it intended to scan all of them.

Full View

You can see books in Full View if the book is out of copyright, or if the publisher or author has asked to make the book fully viewable. The Full View allows you to view any page from the book, and if the book is in the public domain, you can download, save and print a PDF version to read at your own pace.


Full View**Limited Preview**

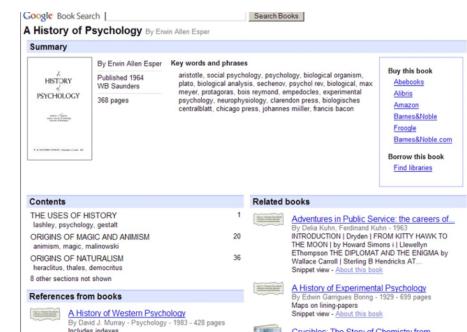
If the publisher or author has given us permission, you can see a limited number of pages from the book as a preview.


Limited View

What You See on Google Books

Snippet View

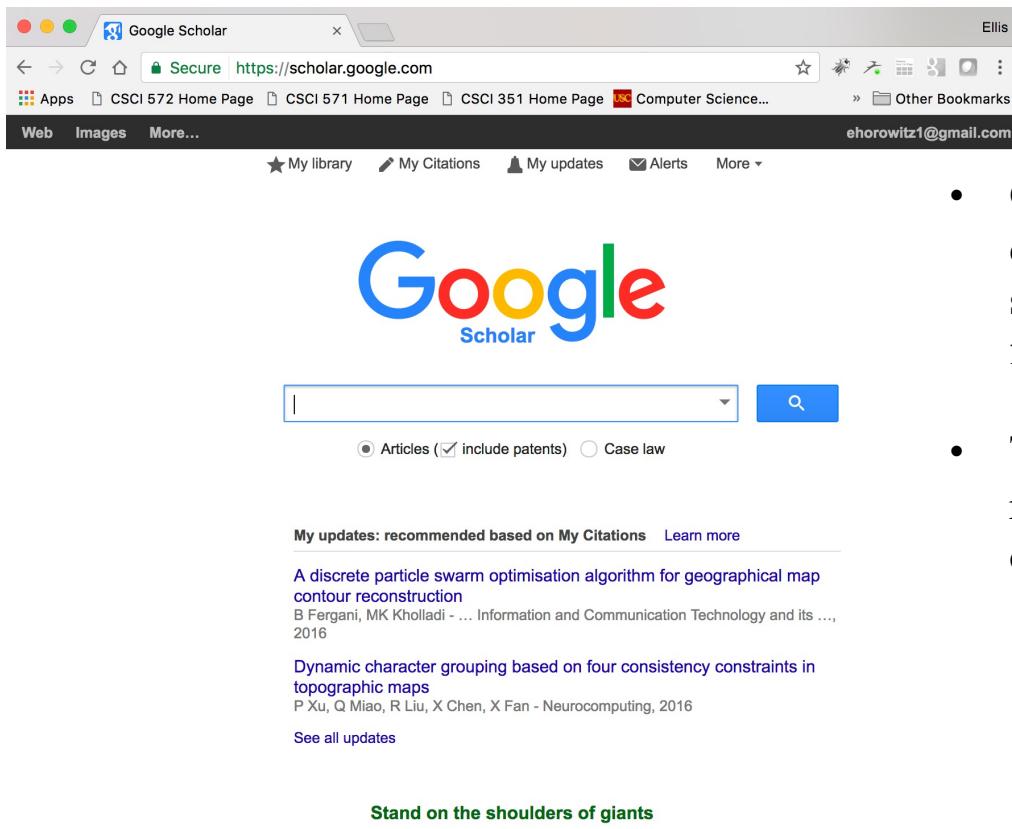
The Snippet View, like a card catalog, shows information about the book plus a few snippets – a few sentences to display your search term in context.


Snippet View

<https://www.google.com/googlebooks/library/screenshots.html#books-fullview>

Special Content Search Engines

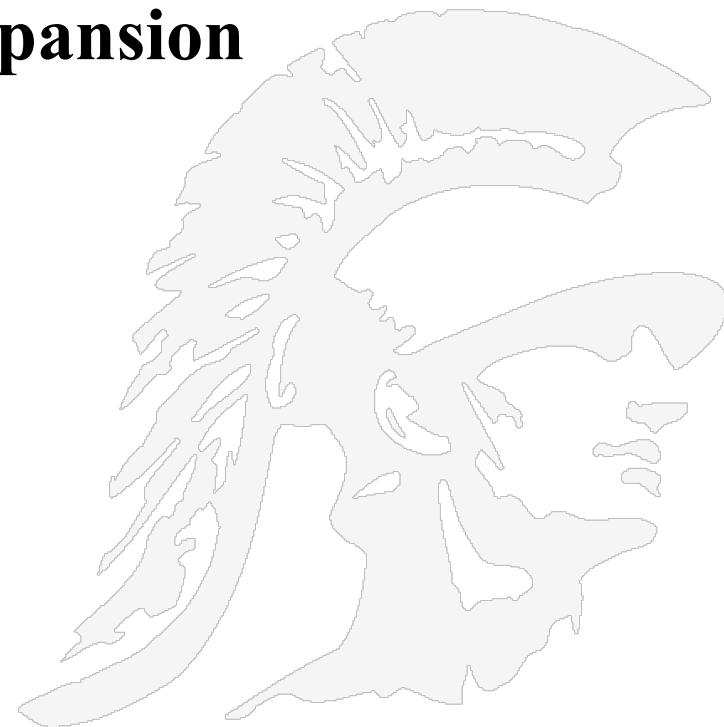
Google Scholar



The screenshot shows the Google Scholar homepage. At the top, there's a search bar with a dropdown arrow, a blue search button, and filter options for 'Articles' (checked) and 'Case law'. Below the search bar, a message says 'My updates: recommended based on My Citations' with a 'Learn more' link. Two academic papers are listed: one by B Fergani and MK Kholladi from 2016, and another by P Xu, Q Miao, R Liu, X Chen, and X Fan from 2016. At the bottom, there's a green footer with the text 'Stand on the shoulders of giants'.

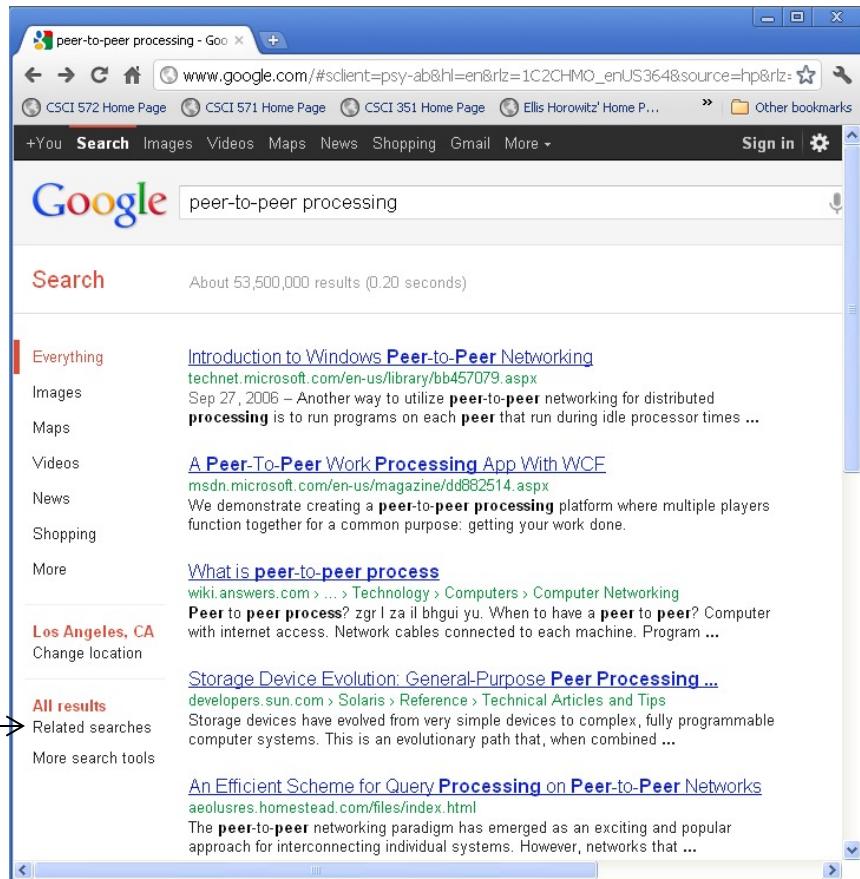
- **Google Scholar** is a freely accessible search engine that indexes the full text or metadata of scholarly literature across an array of publishing formats and disciplines.
- The Google Scholar index includes most peer-reviewed online academic journals and books, conference papers, theses, dissertations, etc

Relevance Feedback & Query Expansion



Relevance Feedback

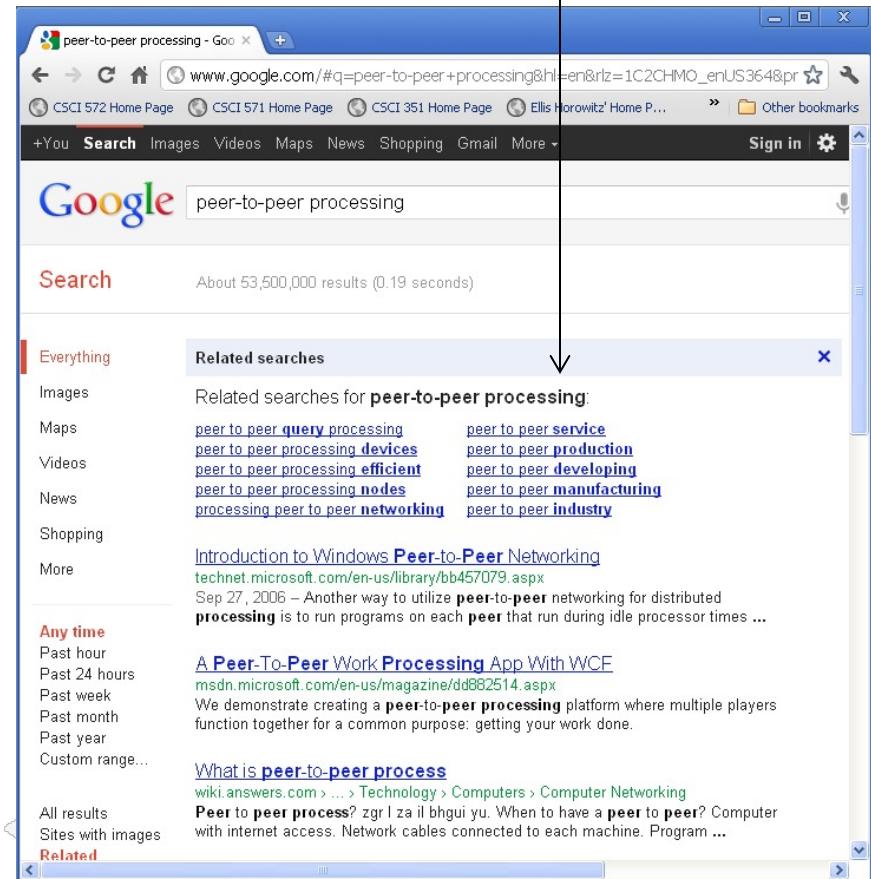
- After initial retrieval results are presented, allow the user to provide feedback on the relevance of one or more of the retrieved documents
 - Google offers this but not very prominently (see 10 related searches below)



Google search results for "peer-to-peer processing". The results page shows a list of links, including:

- [Introduction to Windows Peer-to-Peer Networking](http://technet.microsoft.com/en-us/library/bb457079.aspx)
- [A Peer-To-Peer Work Processing App With WCF](http://msdn.microsoft.com/en-us/magazine/dd882514.aspx)
- [What is peer-to-peer process](http://wiki.answers.com...>Technology>Computers>Computer Networking)
- [Storage Device Evolution: General-Purpose Peer Processing ...](http://developers.sun.com>Solaris>Reference>Technical Articles and Tips)
- [An Efficient Scheme for Query Processing on Peer-to-Peer Networks](http://aeolusres.homestead.com/files/index.html)

The sidebar on the left includes links for "Everything", "Images", "Maps", "Videos", "News", "Shopping", "More", and "Los Angeles, CA". At the bottom, there are links for "All results", "Related searches", and "More search tools".



Google search results for "peer-to-peer processing". The results page shows a list of links, including:

- [Introduction to Windows Peer-to-Peer Networking](http://technet.microsoft.com/en-us/library/bb457079.aspx)
- [A Peer-To-Peer Work Processing App With WCF](http://msdn.microsoft.com/en-us/magazine/dd882514.aspx)
- [What is peer-to-peer process](http://wiki.answers.com...>Technology>Computers>Computer Networking)

A "Related searches" section is highlighted in blue, containing a list of 10 related queries:

- peer to peer query processing
- peer to peer processing devices
- peer to peer processing efficient
- peer to peer processing nodes
- processing peer to peer networking
- peer to peer service
- peer to peer production
- peer to peer developing
- peer to peer manufacturing
- peer to peer industry

The sidebar on the left includes links for "Everything", "Images", "Maps", "Videos", "News", "Shopping", "More", and "Any time". At the bottom, there are links for "All results", "Sites with images", and "Related".

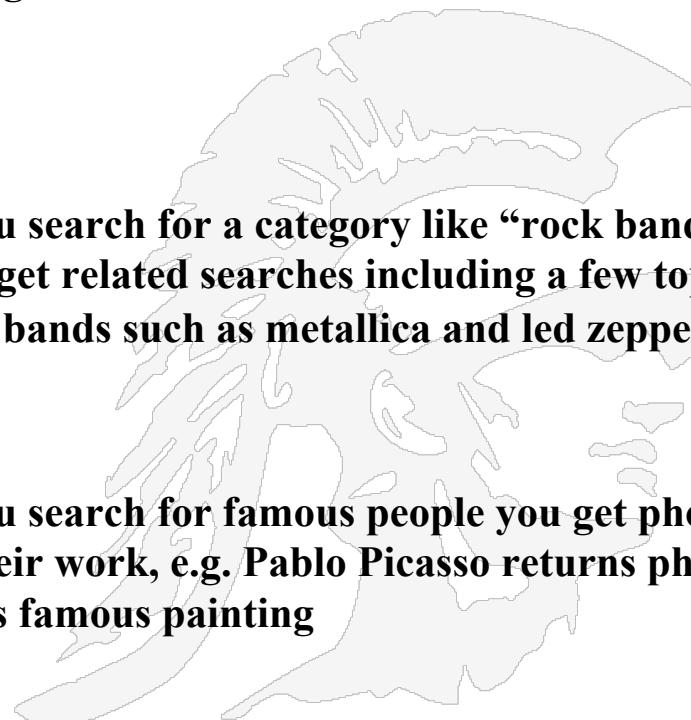
Google Related Searches



The figure consists of three vertically stacked screenshots of Google search results:

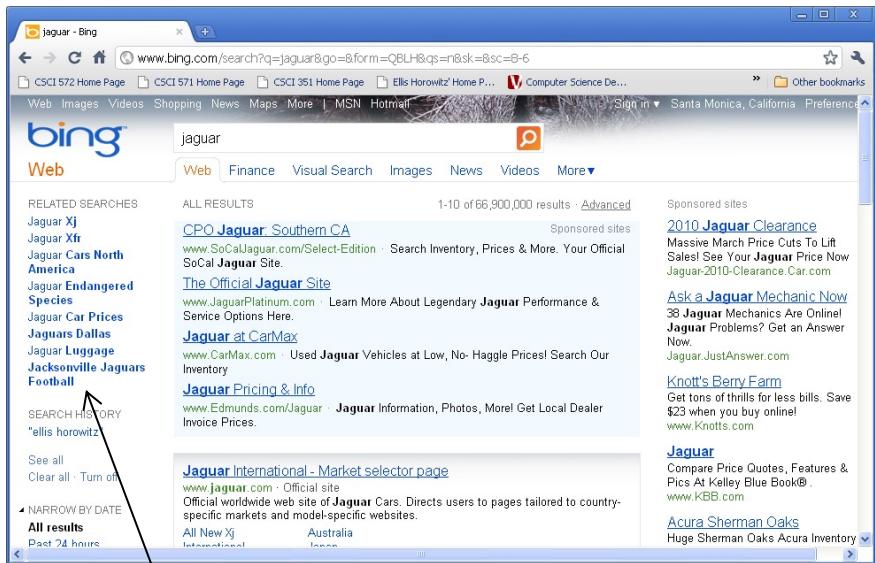
- Top Screenshot:** A screenshot of a browser window titled "german cars - Google Search". The URL is www.google.com/#hl=en&cp=1&q=german+cars. The page shows "Top references for german cars" with links to Audi, BMW, Mercedes Benz, Volkswagen, Porsche, Opel, Smart, Mini, Bentley, and "Feedback". An arrow points from the text "e.g. 'german cars'" to this screenshot.
- Middle Screenshot:** A screenshot of a browser window titled "rock bands - Google Search". The URL is www.google.com/#q=rock+bands&hl=en&prmd=1. The page shows "Related searches for rock bands:" with links to punk rock bands, list of rock bands, classic rock bands, new rock bands, 80's rock bands, metal bands, emo bands, heavy metal bands, screamo bands, pop bands, rock lyrics, rock music, rock songs, rock artists, rock groups, billboard, mtv, metallica, led zeppelin, and aerosmith. An arrow points from the text "e.g. 'rock bands'" to this screenshot.
- Bottom Screenshot:** A screenshot of a browser window titled "pablo picasso - Google Search". The URL is www.google.com/#hl=en&cp=4&gs_id=f&xhr=1. The page shows "Artwork searches for Pablo Picasso" with thumbnail images of Guernica, Les Demoiselles d'Avignon, The Old Guitarist, Three Musicians, and Dora Maar au Chat. Below the images, it says "Searches related to pablo picasso" with links to pablo picasso biography, facts, quotes, life, pablo picasso cubism, pictures, guernica, salvador dali, and pablo picasso's life. An arrow points from the text "e.g. Pablo Picasso returns photos of his famous painting" to this screenshot.

- Google has enhanced their related searches, e.g.
- If you search for the name of a category, Google will show the most popular members, e.g. “german cars”



- If you search for a category like “rock bands” You get related searches including a few top rock bands such as metallica and led zeppelin
- If you search for famous people you get photos of their work, e.g. Pablo Picasso returns photos of his famous painting

Search Engines and Relevance Feedback



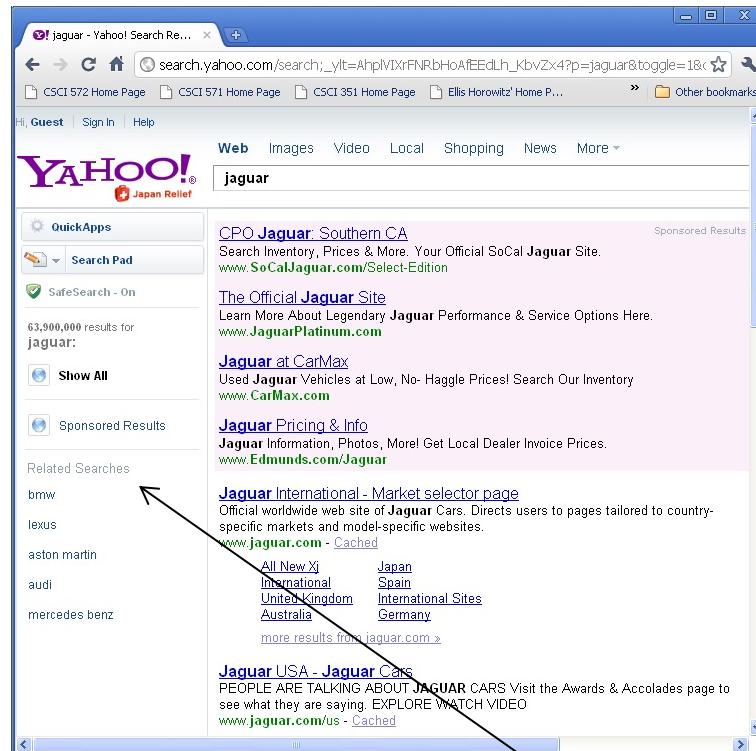
Bing search results for "jaguar":

- ALL RESULTS 1-10 of 66,900,000 results · Advanced
- CPO Jaguar** Southern CA Sponsored sites www.SoCalJaguar.com>Select-Edition · Search Inventory, Prices & More. Your Official SoCal Jaguar Site.
- The Official Jaguar Site** www.JaguarPlatinum.com · Learn More About Legendary Jaguar Performance & Service Options Here.
- Jaguar at CarMax** www.CarMax.com · Used Jaguar Vehicles at Low, No-Haggle Prices! Search Our Inventory
- Jaguar Pricing & Info** www.Edmunds.com/Jaguar · Jaguar Information, Photos, More! Get Local Dealer Invoice Prices.
- Jaguar International - Market selector page** www.jaguar.com · Official site Official worldwide web site of Jaguar Cars. Directs users to pages tailored to country-specific markets and model-specific websites.

SEARCH HISTORY
*ellis horowitz

See all
Clear all · Turn off

NARROW BY DATE
All results
Past 24 hours



Yahoo search results for "jaguar":

- CPO Jaguar** Southern CA Search Inventory, Prices & More. Your Official SoCal Jaguar Site. www.SoCalJaguar.com>Select-Edition
- The Official Jaguar Site** Learn More About Legendary Jaguar Performance & Service Options Here. www.JaguarPlatinum.com
- Jaguar at CarMax** Used Jaguar Vehicles at Low, No-Haggle Prices! Search Our Inventory www.CarMax.com
- Jaguar Pricing & Info** Jaguar Information, Photos, More! Get Local Dealer Invoice Prices. www.Edmunds.com/Jaguar
- Jaguar International - Market selector page** Official worldwide web site of Jaguar Cars. Directs users to pages tailored to country-specific markets and model-specific websites.

Related Searches

- bmw
- lexus
- aston martin
- audi
- mercedes benz

On the query “jaguar” yahoo’s 3rd result is the animal; Bing’s 2nd result is the animal; all have extensive ads for the car at the top and side indicating that the query for the animal is far rarer; Bing, on the left, puts up alternatives, e.g. Jaguar Luggage, Jacksonville Jaguars; Yahoo also provides related searches but only for the car: bmw, lexus, etc

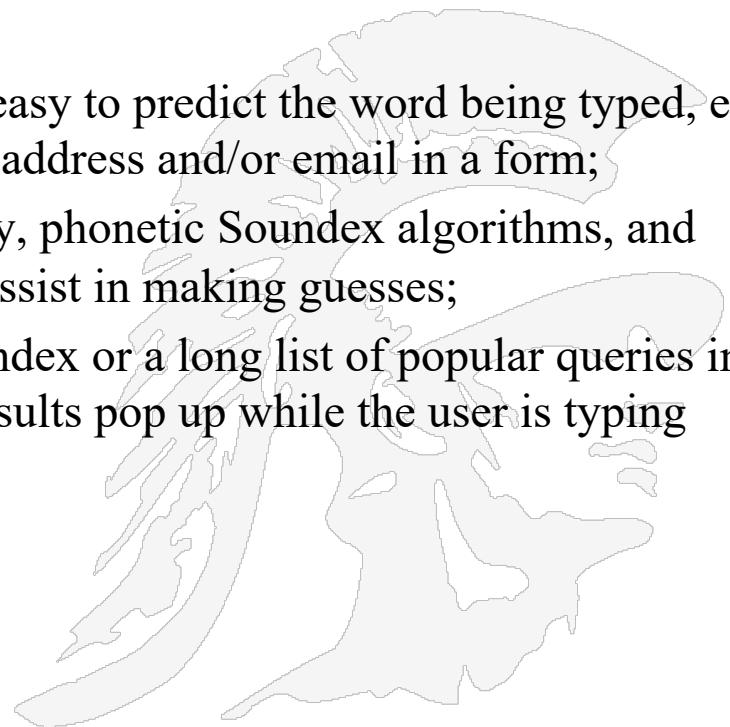
Bing and Yahoo clearly mark Related Searches

Auto-Completion



Auto-Completion

- ***Auto-completion* is the process of predicting a word or phrase that the user wants to type in without the user actually typing it in completely**
- **Auto-completion is a form of relevance feedback**
 - This feature is effective when it is easy to predict the word being typed, e.g. when a browser fills in your name, address and/or email in a form;
 - Search engines may use past history, phonetic Soundex algorithms, and spelling corrections algorithms to assist in making guesses;
 - The challenge is to search a large index or a long list of popular queries in a very short amount of time so the results pop up while the user is typing



Google Auto-Completion

Google Search

Everything Images Maps Videos News Shopping More

amazon
amazon
apple
aol
american airlines

[Amazon.com® Official Site](#)
www.amazon.com
amazon.com is rated ★★★★☆ 8,669 reviews
Huge Selection and Amazing Prices. Free Shipping on Orders Over \$25
12,679 people +1'd this page

[Amazon.com: Online Shopping for Electronics, Apparel, Computers ...](#)
www.amazon.com/
Online retailer of books, movies, music and games along with electronics, toys, apparel, sports, tools, groceries and general home and garden items. Region 1 ...
+ Show stock quote for AMZN
177 people in Los Angeles, CA, USA +1'd this

- Google has been offering auto-completion since 2008, though it was an experimental feature as far back as 2004.
- Google does automatic completion even after the user enters just the first character
- When the second character is entered a totally different set of possibilities may be offered

+Ellis Search Images Videos Maps News Shopping Gmail More Ellis Horowitz | Share...  

Google Search

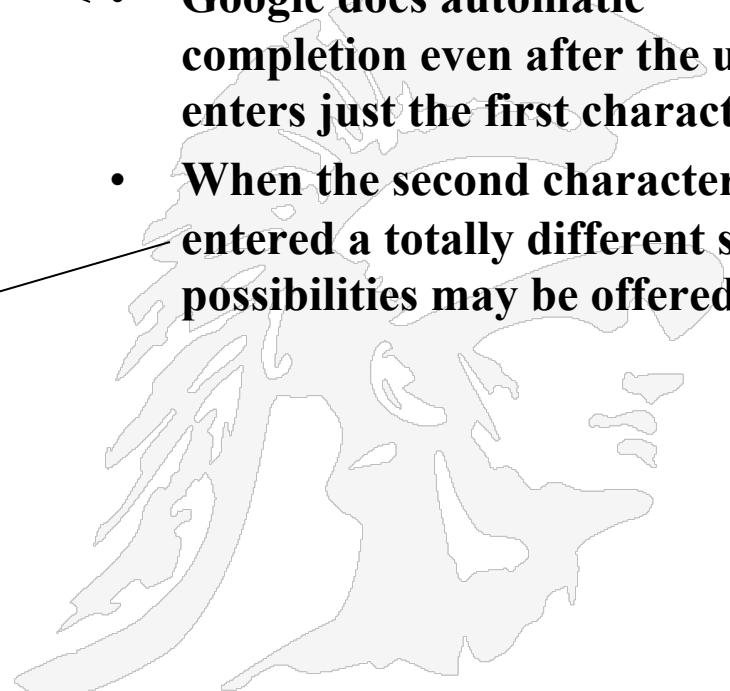
Everything Images Maps Videos News Shopping More

anniversary gifts
anniversary gifts Remove
anthropologie
angry birds
android market

[Anniversary Gifts](#) 1 (877) 530 9512
www.bluenile.com - ★★★★☆ 3,437 seller reviews
Find the perfect gift for her. Free FedEx & 30-Day Returns.

[Anniversary Gifts | PersonalizationMall.com](#)
www.personalizationmall.com - ★★★★☆ 2,157 seller reviews
Unique Romantic Anniversary Gifts Engraved Free & Ship in 1-2 Days!
Bridesmaid Gifts - Wedding Favors - Wedding Gifts - Groomsmen Gifts

[MY M&M'S® - Anniversary | MyMMs.com](#)
www.mydds.com
Personalized MY M&M'S® Candies Make the Perfect Anniversary Gift!



+Ellis Search Images Videos Maps News Shopping Gmail More Ellis Horowitz 1 Share...  

Google

Search anaheim ducks

Everything

- Images
- Maps
- Videos
- News
- Shopping
- More

anaheim ducks

anaheim

analytics

anagram

Anaheim Ducks Tickets - Buy Ducks Hockey Tickets Today.
www.ticketmaster.com/Ducks
Prices Starting at Only \$25!

714Tickets Ducks Tickets - Across from Honda Center
www.714tickets.com/
No hidden handling fees

Anaheim Ducks
ducks.nhl.com/
Official site. News, statistics, players, coaches, schedule, and arena information.
Regular season record: 19-24-7 Division standing: 5 Points: 45

+Ellis Search Images Videos Maps News Shopping Gmail More Ellis Horowitz 1 Share...  

Google

Search anaz

Everything

- Images
- Maps
- Videos
- News
- Shopping
- More

amazon

amazon promo code

amazon kindle

amazon prime

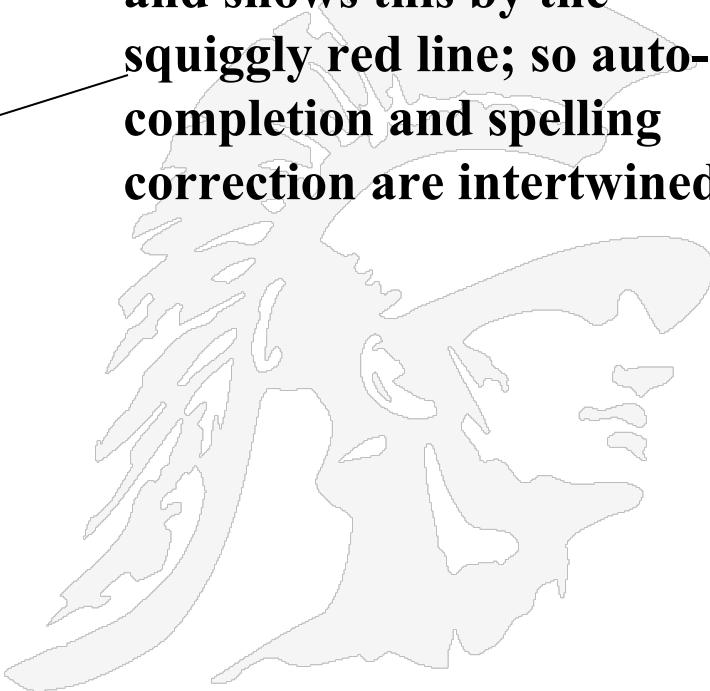
Amazon.com® Official Site
www.amazon.com
amazon.com is rated ★★★★☆ 8,669 reviews
Huge Selection and Amazing Prices. Free Shipping on Orders Over \$25
12,679 people +1'd this page

Showing results for **amazon**
Search instead for anaz

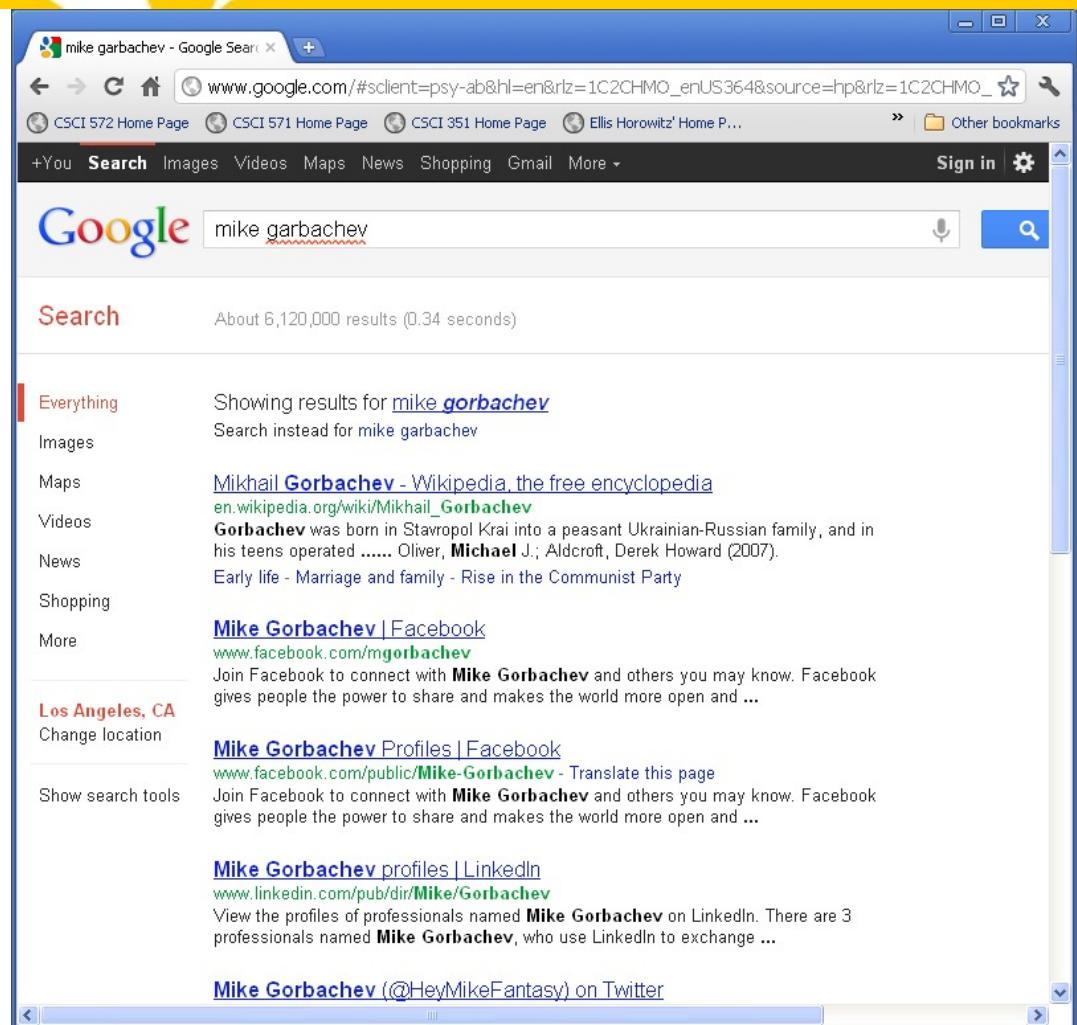
Amazon.com: Online Shopping for Electronics, Apparel, Computers ...
www.amazon.com/
Online retailer of books, movies, music and games along with electronics, toys, apparel.

Google Auto-Completion

- By the time the fourth character is entered Google has already guessed that the word has been misspelled, and shows this by the squiggly red line; so auto-completion and spelling correction are intertwined



Google Combines Spelling Correction and Auto Completion



mike garbachev - Google Search

www.google.com/#sclient=psy-ab&hl=en&rlz=1C2CHMO_enUS364&source=hp&rlz=1C2CHMO_...

CSCI 572 Home Page CSCI 571 Home Page CSCI 351 Home Page Ellis Horowitz' Home P... Other bookmarks

Sign in | More

Search

mike garbachev

Search

About 6,120,000 results (0.34 seconds)

Everything Showing results for [mike gorbachev](#)
Search instead for mike garbachev

[Mikhail Gorbachev - Wikipedia, the free encyclopedia](#)
en.wikipedia.org/wiki/Mikhail_Gorbachev
Gorbachev was born in Stavropol Krai into a peasant Ukrainian-Russian family, and in his teens operated Oliver, Michael J.; Aldcroft, Derek Howard (2007). Early life - Marriage and family - Rise in the Communist Party

[Mike Gorbachev | Facebook](#)
www.facebook.com/mgorbachev
Join Facebook to connect with **Mike Gorbachev** and others you may know. Facebook gives people the power to share and makes the world more open and ...

[Mike Gorbachev Profiles | Facebook](#)
www.facebook.com/public/Mike-Gorbachev - Translate this page
Join Facebook to connect with **Mike Gorbachev** and others you may know. Facebook gives people the power to share and makes the world more open and ...

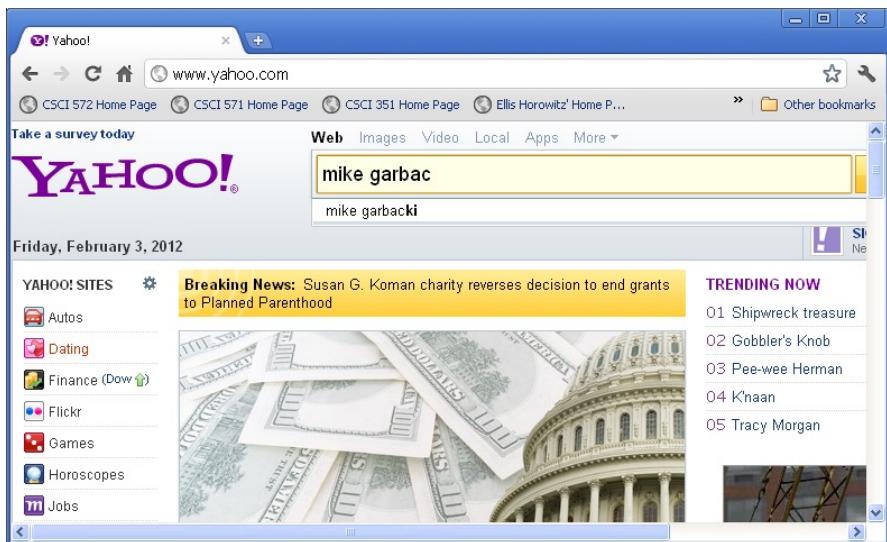
[Mike Gorbachev profiles | LinkedIn](#)
www.linkedin.com/pub/dir/Mike/Gorbachev
View the profiles of professionals named **Mike Gorbachev** on LinkedIn. There are 3 professionals named **Mike Gorbachev**, who use LinkedIn to exchange ...

[Mike Gorbachev \(@HeyMikeFantasy\) on Twitter](#)

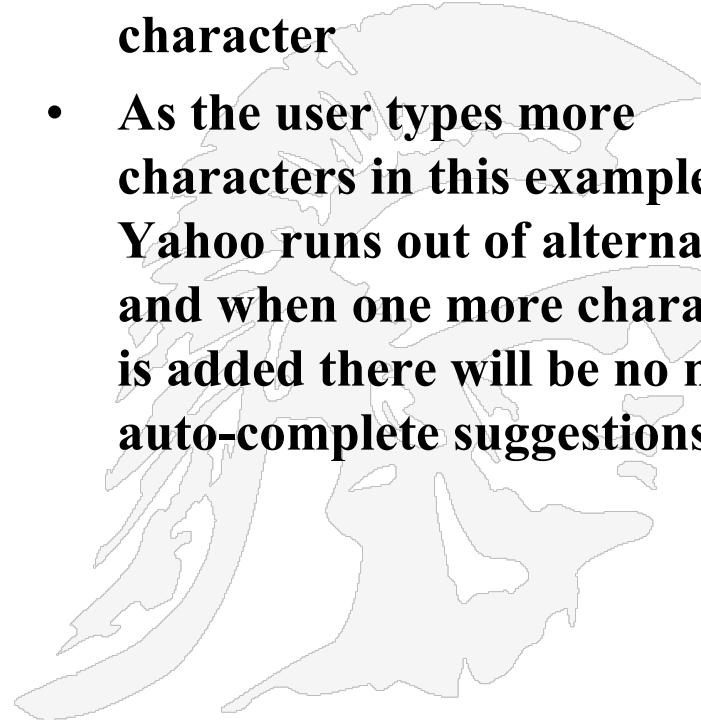
After hitting Enter, Google will display results for “**Mikhail Gorbachev**”, but also provide a link if instead the user actually wanted to search for “**mike garbachev**”



Yahoo Auto-Completion



- **Yahoo does auto-completion though it starts after typing the *third* character, whereas Google starts on the *first* character**
- **As the user types more characters in this example Yahoo runs out of alternatives and when one more character is added there will be no more auto-complete suggestions**



Yahoo Also Offers Spelling Correction

mike garbachev - Yahoo! Search Results

Hi, Guest | Sign In | Help

YAHOO!

mike garbachev

Search

WEB IMAGES VIDEO SHOPPING APPS BLOGS MORE

2,100,000 results

We have included [michael gorbachev](#) results - Show only [mike garbachev](#)

Mikhail Gorbachev - Wikipedia, the free encyclopedia
[Early life](#) | [Marriage and...](#) | [Rise in the...](#) | [General Secretary...](#)
 Archie Brown, in The Gorbachev Factor, uses the memoirs of many people around ...
 Retrieved 2 April 2009. ^ Oliver, Michael J.; Aldcroft, Derek Howard (2007).
[en.wikipedia.org/wiki/Mikhail_Gorbachev](#) - Cached
[More results from en.wikipedia.org »](#)

Mikhail Gorbachev: Biography from Answers.com
 McFaul, Michael. (2001). Russia's Unfinished Revolution: Political Change from Gorbachev to Putin. Ithaca, NY: Cornell University Press. Matlock, Jack F., Jr. (1995).
[www.answers.com/topic/mikhail-gorbachev](#) - Cached
[More results from answers.com »](#)

Mike Gorbachev - Image Results



Michael Gorbachev | Facebook
 Michael Gorbachev is on Facebook. Join Facebook to connect with Michael Gorbachev and others you may know. Facebook gives people the power to share and makes the ...
[www.facebook.com/people/Michael-Gorbachev/100002373955597](#) - Cached
[More results from facebook.com »](#)

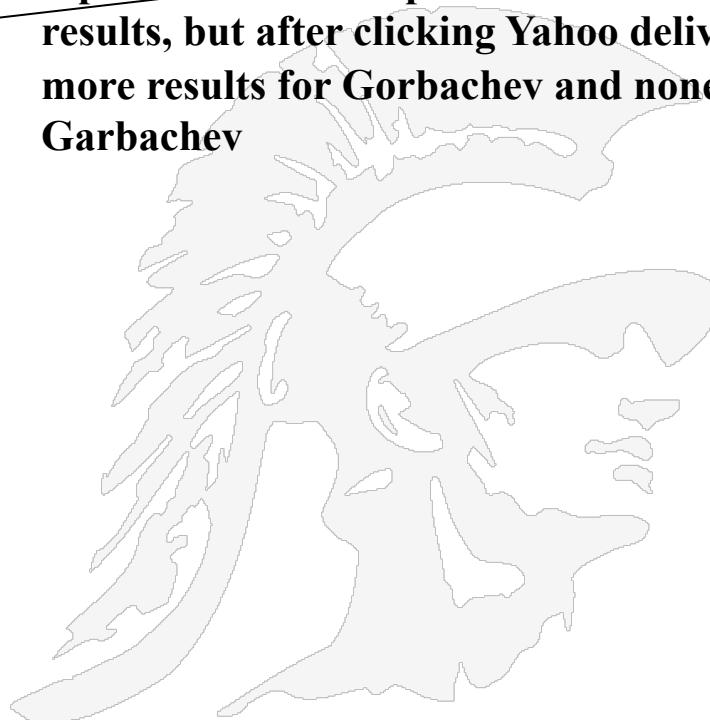
My Country and the World, by Michael Gorbachev
 A two-page summary of Gorbachev's evaluation of Russian and Soviet history dating back to 1917.
[www.fsmitha.com/review/r-gorb.htm](#) - Cached

Mikhail Gorbachev - Biography - Nobelprize.org
 Why Gorbachev Happened: His Triumphs and his Failures. New York: Simon & Schuster, 1991. Miller, John. Mikhail Gorbachev and the End of Soviet Power ...
[www.nobelprize.org/laureates/1990/gorbachev-bio.html](#) - Cached

michael gorbachev | crazymonk.org
 Part II of Errol Morris day. (Re-)watch this Morris short, which played during the 2002

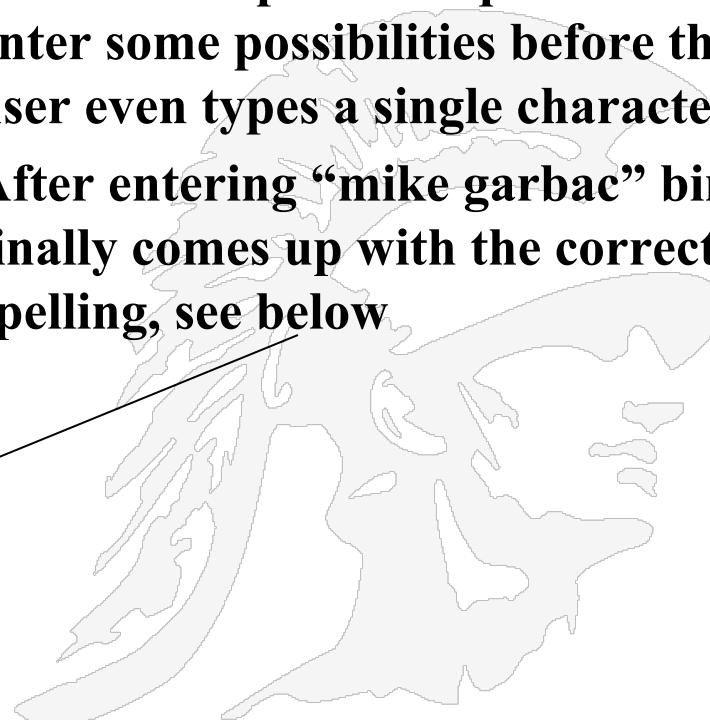
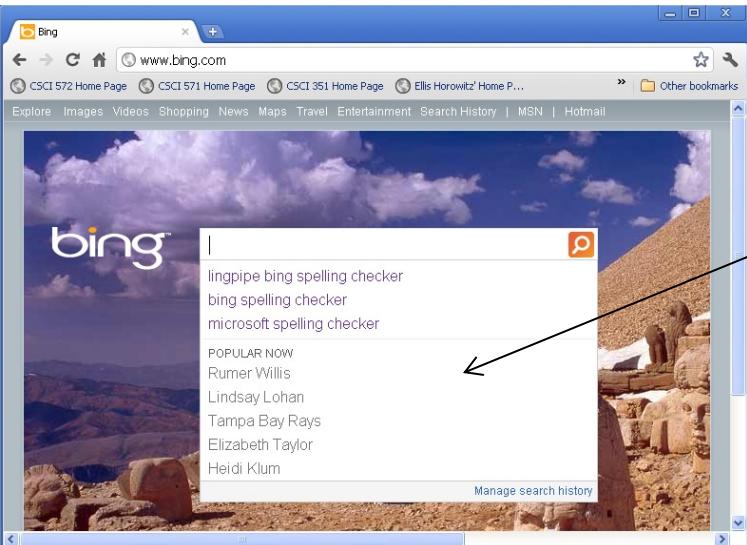
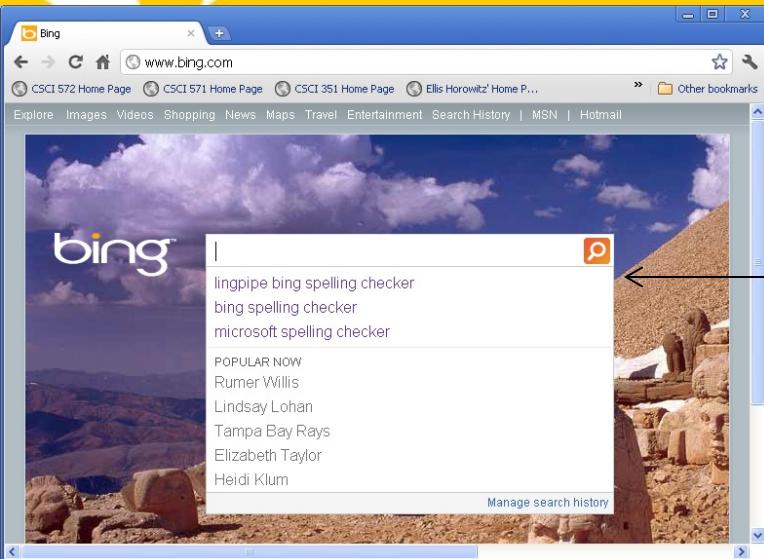
<http://shopping.yahoo.com/search;...rbachev&fr=yfp-t-701&fr2=piv-web>

After hitting Enter, Yahoo does offer the correct spelling of “Gorbachev” and returns many results; in case the user actually wanted to search for “garbachev” it provides a link to produce those results, but after clicking Yahoo delivers more results for Gorbachev and none for Garbachev



Bing Auto-Completion

- On the other hand, Bing does not even wait for the first character to be entered, as seen on the left they make use of previous queries and enter some possibilities before the user even types a single character
- After entering “mike garbac” bing finally comes up with the correct spelling, see below



Bing Final Result with Spelling Corrected

mike garbachev - Bing

www.bing.com/search?q=mike+garbachev&qs=n&form=QBL

Related Searches: Mikhail Gorbachev and Ronald Reagan, Mikhail Gorbachev Biography, Mikhail Gorbachev Policy, Mikhail Gorbachev Family, Mikhail Gorbachev Wife, Gorbachev Foundation, Reagan and Gorbachev, Mikhail Baryshnikov.

Search History: garbachi, brec baldwin, See all, Clear all - Turn off

Search Bar: mike garbachev

Results: ALL RESULTS 1-10 of 2,160,000 results · Advanced

Including results for [michael gorbachev](#). Do you want results for mike garbachev?

[Mikhail Gorbachev - Wikipedia, the free encyclopedia](#)
Early life · Marriage and family · Rise in the Communist ... Archie Brown, in The Gorbachev Factor, uses the memoirs of many people around ... Retrieved 2 April 2009. ^ Oliver, Michael J.; Aldcroft, Derek Howard (2007). en.wikipedia.org/wiki/Mikhail_Gorbachev

[Mikhail Gorbachev: Biography from Answers.com](#)
McFaul, Michael. (2001). Russia's Unfinished Revolution: Political Change from Gorbachev to Putin. Ithaca, NY: Cornell University Press. Matlock, Jack F., Jr. ... www.answers.com/topic/mikhail-gorbachev

[Images of mike garbachev](#)

[Michael Gorbachev | Facebook](#)
Lives in Tanjungbalai · From Medan · Worked at PT.KALAPAON Michael Gorbachev is on Facebook. Join Facebook to connect with Michael Gorbachev and others you may know. Facebook gives people ... www.facebook.com/people/Michael-Gorbachev/...

[My Country and the World, by Michael Gorbachev](#)
A two-page summary of Gorbachev's evaluation of Russian and Soviet history dating back to 1917. www.fsmitha.com/review/r-gorb.htm

[Mikhail Gorbachev - Biography - Nobelprize.org](#)
Why Gorbachev Happened: His Triumphs and his Failures. New York: Simon & Schuster, 1991. Miller, John. Mikhail Gorbachev and the End of Soviet Power ... www.nobelprize.org/nobel_prizes/peace/laurates/1990/gorbachev-bio...

[michael gorbachev | crazymonk.org](#)
Part II of Errol Morris day. (Re-)watch this Morris short, which played during the 2002 Oscar telecast, depicting a bunch of people talking about their favorite films. crazymonk.org/topics/michaelgorbachev

As with Google and Yahoo, Bing will offer results using the corrected spelling and include a link for the user spelling



Judging the Quality of Answers: Mean Reciprocal Rank (MRR) Scoring

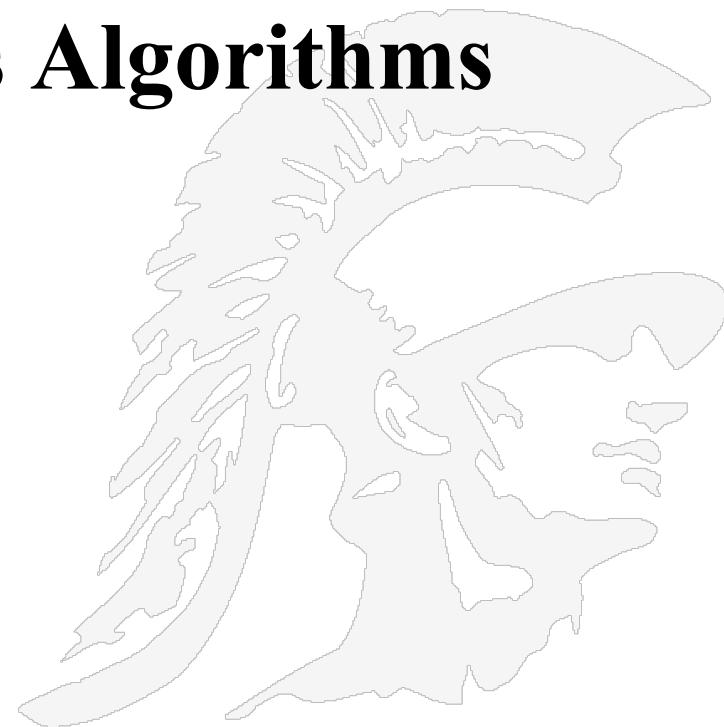
- The mean reciprocal rank is a statistical measure for evaluating any process that produces a list of possible responses to a sample of queries, ordered by probability of correctness.
The **reciprocal rank** of a query response is the multiplicative inverse of the **rank** of the first correct answer
- The mean reciprocal rank is the average of the reciprocal ranks of results for a sample of queries
- For example, suppose we have the following three sample queries for a system that tries to translate English words to their plurals. In each case, the system makes three guesses, with the first one being the one it thinks is most likely correct:

$$\text{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i}.$$

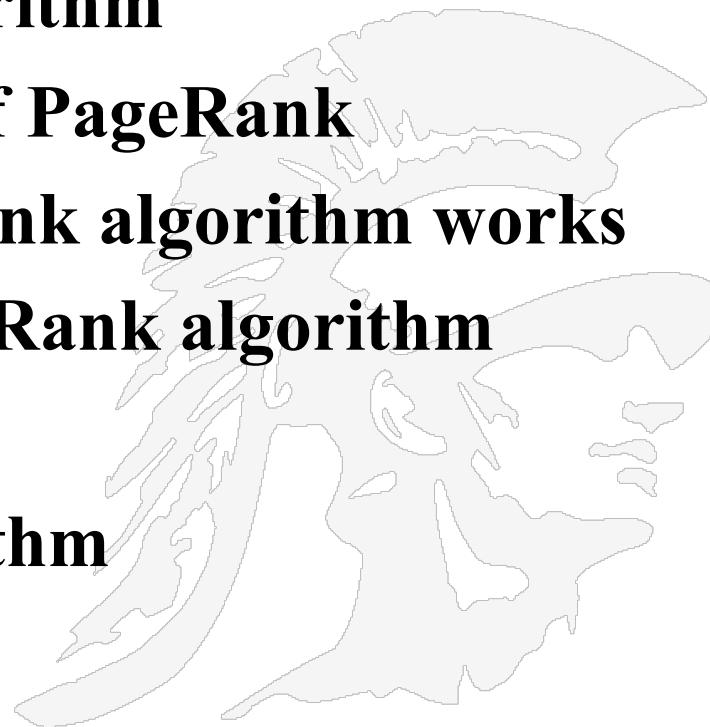
Query	Results	Correct response	Rank	Reciprocal rank
cat	catten, cati, cats	cats	3	1/3
tori	torii, tori , toruses	tori	2	1/2
virus	viruses , virii, viri	viruses	1	1

the mean reciprocal rank as $(1/3 + 1/2 + 1)/3 = 11/18$ or about 0.61.

Link Analysis Algorithms



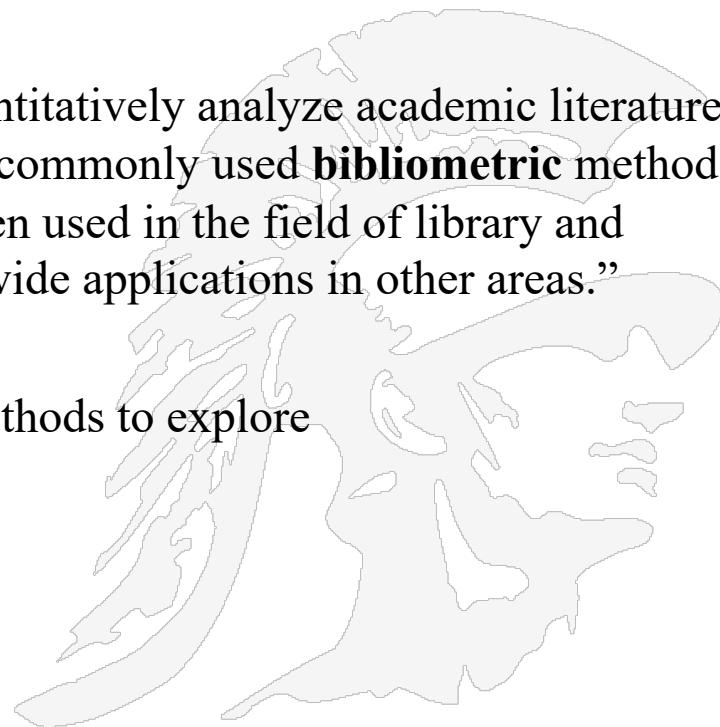
- **Background on Citation Analysis**
- **Google's PageRank Algorithm**
- **Simplified Explanation of PageRank**
- **Examples of how PageRank algorithm works**
- **Observations about PageRank algorithm**
- **Importance of PageRank**
- **Kleinberg's HITS Algorithm**



History of Link Analysis

- **Bibliometrics has been active since at least the 1960's**
- A definition from Wikipedia:

"**Bibliometrics** is a set of methods to quantitatively analyze academic literature. Citation analysis and content analysis are commonly used **bibliometric** methods. While **bibliometric** methods are most often used in the field of library and information science, **bibliometrics** have wide applications in other areas."
- Many research fields use **bibliometric** methods to explore
 - the impact of their field,
 - the impact of a set of researchers, or
 - the impact of a particular paper.



- **One common technique of Bibliometrics is *citation analysis***
- **Citation analysis** is the examination of the frequency, patterns, and graphs of citations in articles and books.
- citation analysis can observe links to other works or other researchers.
- **Bibliographic coupling:** two papers that cite many of the same papers
- **Co-citation:** two papers that were cited by many of the same papers
- **Impact factor (of a journal):** frequency with which the average article in a journal has been cited in a particular year or period

Bibliometrics

<http://citeseerx.ist.psu.edu/stats/citations>

Top Ten Most Cited Articles in CS Literature
CiteSeer^x is a search engine for academic papers



Most Cited Computer Science Citations

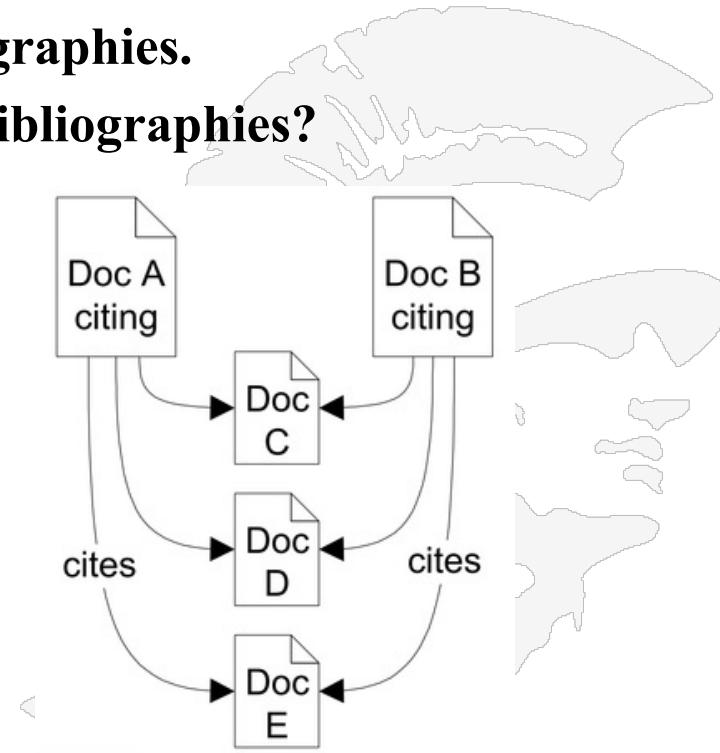
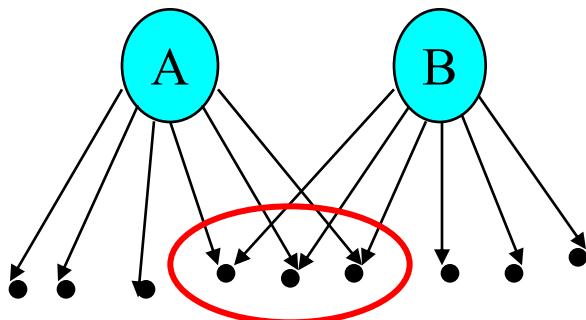
This list is generated from documents in the CiteSeer^x database as of March 19, 2015. This list is automatically generated and may contain errors. The list mode and citation counts may differ from those currently in the CiteSeer^x database, since the database is continuously updated.

All Years | 1990 | 1991 | 1992 | 1993 | 1994 | 1995 | 1996 | 1997 | 1998 | 1999 | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2015

1. M R Garey, D S Johnson
Computers and Intractability: A Guide to the Theory of NPCompleteness W.H. Freeman and 1979
11468
2. J Sambrook, E F Fritsch, T Maniatis
Molecular Cloning: A Laboratory Manual, Vol. 1, 2nd edn Nucleic Acids Research, 1989
10362
3. V Vapnik
Statistical Learning Theory. 1998
9698
4. T M Cover, J A Thomas
Elements of Information Theory Series in Telecommunications, 1991
9198
5. U K Laemmli
Cleavage of structural proteins during the assembly of the head of bacteriophage T4. Nature 227:680–685 1970
9092
6. T H Cormen, C E Leiserson, R L Rivest, C Stein
Introduction to Algorithms. 1990
9039
7. A P Dempster, N M Laird, D B Rubin
Maximum likelihood from incomplete data via the EM algorithm. 1977
8999
8. D E Goldberg
Genetic Algorithms in Search, Optimization and Machine Learning, 1989
8261
9. J Pearl
Probabilistic Reasoning in Intelligent Systems 1988
7473
10. C E Shannon, W Weaver
The Mathematical Theory of Communication 1949
7077

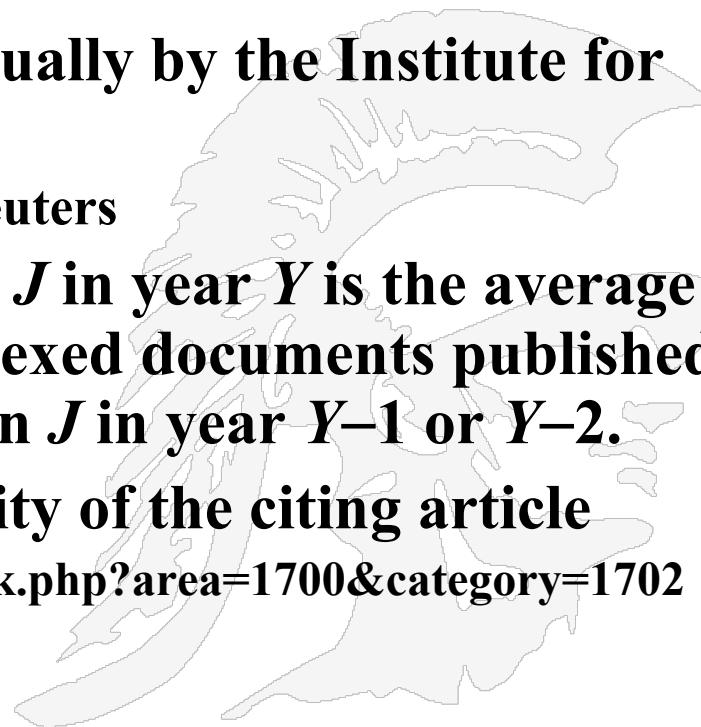
Bibliographic Coupling

- Measure of similarity of documents introduced by Kessler of MIT in 1963.
- The bibliographic coupling of two documents A and B is the number of documents cited by *both* A and B .
- Size of the intersection of their bibliographies.
- Maybe want to normalize by size of bibliographies?



Journal Impact Factor

- Developed by Garfield in 1972 to measure the importance (quality, influence) of scientific journals.
- Measure of how often papers in the journal are cited by other scientists.
- Computed and published annually by the Institute for Scientific Information (ISI).
 - It is now owned by Thomson Reuters
- The *impact factor* of a journal J in year Y is the average number of citations (from indexed documents published in year Y) to a paper published in J in year $Y-1$ or $Y-2$.
- Does not account for the quality of the citing article
- <https://www.scimagojr.com/journalrank.php?area=1700&category=1702>



Over all Computer Science

<http://www.guide2research.com/journals/>

Top Journals for Computer Science and Electronics

 Guide2Res... Like Page		
Ranking is based on Impact Factor. Vanity press and poor-quality journals are not listed <input type="button" value="All Categories"/> <input type="button" value="All Publishers"/> <input type="text" value="Search by keyword"/>		
Rank	Publisher	Journal Details
		Impact Factor
1	 IEEE	IEEE Communications Surveys and Tutorials ISSN:1553-877X , Quarterly
2	 IEEE	IEEE Transactions on Fuzzy Systems ISSN:1063-6706 , Bimonthly
3	 IEEE	IEEE Signal Processing Magazine ISSN:1053-5888 , Bimonthly
4	 IEEE	IEEE Transactions on Industrial Electronics ISSN:0278-0046 , Monthly
5	 Mary Ann Liebert	Soft Robotics ISSN:2169-5172 , Quarterly
6	 World Scientific www.worldscientific.com	International Journal of Neural Systems ISSN:0129-0657 , Bimonthly
7	 IEEE	IEEE Transactions on Pattern Analysis and Machine Intelligence ISSN:0162-8828 , Monthly
8	 IEEE	IEEE Transactions on Evolutionary Computation ISSN:1089-778X , Bimonthly
9	 ELSEVIER	Remote Sensing of Environment ISSN:0034-4257 , Monthly
10	 OXFORD UNIVERSITY PRESS	Bioinformatics ISSN:1367-4803 , Semimonthly
		9.220
		6.701
		6.671
		6.383
		6.130
		6.085
		6.077
		5.908
		5.881
		5.766

Top Journals for Computer Science

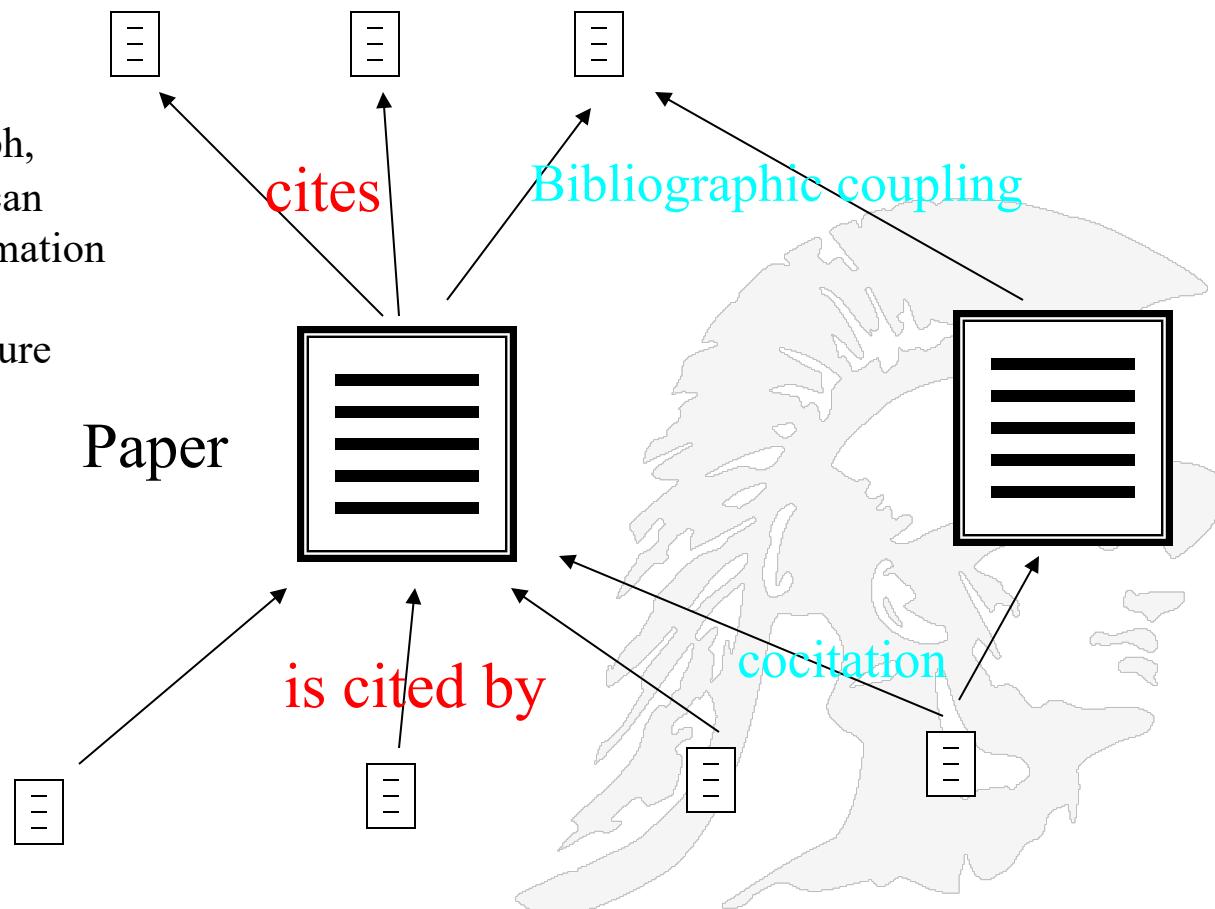
Software Engineering

Top Journals for Computer Science and Electronics

 Guide2Res... Like Page		
Ranking is based on Impact Factor. Vanity press and poor-quality journals are not listed <input type="button" value="Software Engineering"/> <input type="button" value="All Publishers"/> <input type="text" value="Search by keyword"/>		
Rank	Publisher	Journal Details
		Impact Factor
90	 WILEY <small>Publishers Since 1807</small>	INFORMATION SYSTEMS JOURNAL ISSN:1350-1917 , Bimonthly
113	 IEEE	IEEE Transactions on Reliability ISSN:0018-9529 , Quarterly
147	 Springer	Business and Information Systems Engineering ISSN:1867-0202 , Bimonthly
185	 ELSEVIER	Information Systems ISSN:0306-4379 , Bimonthly
215	 ELSEVIER	Advances in Engineering Software ISSN:0965-9978 , Monthly
235	 IEEE	IEEE Transactions on Dependable and Secure Computing ISSN:1545-5971 , Bimonthly
237	 ELSEVIER	Journal of Computer and System Sciences ISSN:0022-0000 , Bimonthly
241	 ELSEVIER	Information and Software Technology ISSN:0950-5849 , Monthly
254	 IEEE	IEEE Transactions on Software Engineering ISSN:0098-5589 , Monthly
256	 ACM <small>Association for Computing Machinery</small>	ACM Transactions on Software Engineering and Methodology ISSN:1049-331X , Quarterly
		2.522
		2.287
		2.059
		1.832
		1.673
		1.592
		1.583
		1.569
		1.516
		1.513

Citation Graph

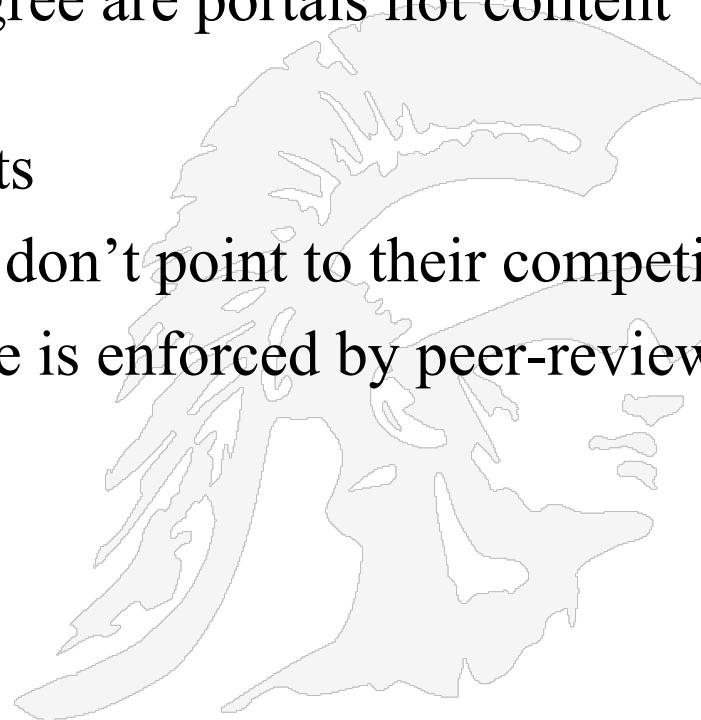
The structure of this graph, independent of content, can provide interesting information about the similarity of documents and the structure of information



Note that academic citations nearly always refer to the author's earlier work.

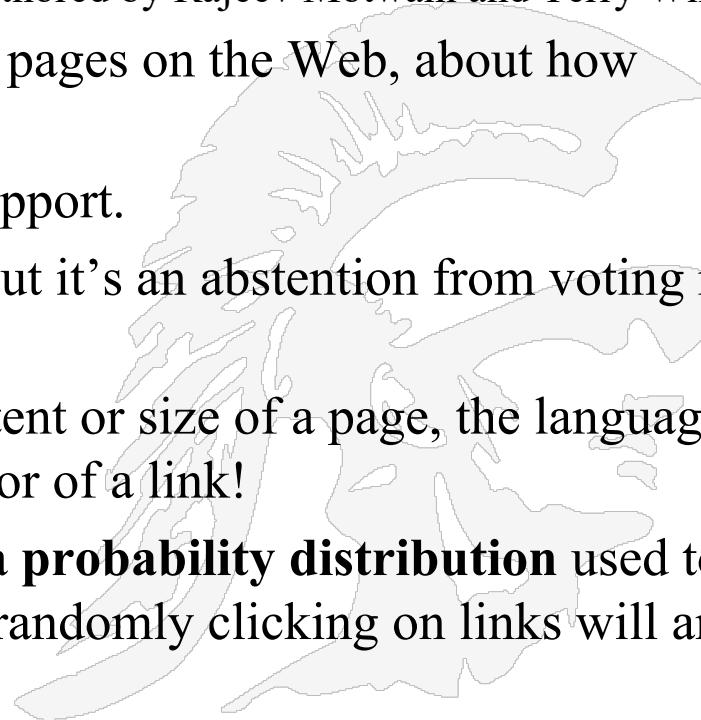
Citations vs. Web Links

- **Web links are a bit different than citations:**
 - Many links are navigational
 - Many pages with high in-degree are portals not content providers
 - Not all links are endorsements
 - Company websites normally don't point to their competitors
 - Citations to relevant literature is enforced by peer-review



What is PageRank?

- PageRank is a **web link analysis algorithm** introduced by Google
- PageRank was developed at Stanford University by Google founders **Larry Page and Sergey Brin**
 - The paper describing PageRank was co-authored by Rajeev Motwani and Terry Winograd
- PageRank is a “**vote**”, by all the other pages on the Web, about how important a page is.
- A link to a page counts as a vote of support.
- If there’s no link there’s no support (but it’s an abstention from voting rather than a vote against the page).
- PageRank says nothing about the content or size of a page, the language it’s written in, or the text used in the anchor of a link!
- Looked at another way, PageRank is a **probability distribution** used to represent the likelihood that a person randomly clicking on links will arrive at any particular page



www.freepatentsonline.com/6285999.pdf

USC Viterbi School • Main Page – Comput • Faculty Resources, C • Computer Science D • Google • Gmail – Inbox (3)


US006285999B1

(12) United States Patent
Page

(10) Patent No.: US 6,285,999 B1
(45) Date of Patent: Sep. 4, 2001

(54) METHOD FOR NODE RANKING IN A LINKED DATABASE

(75) Inventor: **Lawrence Page**, Stanford, CA (US)

(73) Assignee: **The Board of Trustees of the Leland Stanford Junior University**, Stanford, CA (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days.

(21) Appl. No.: **09/004,827**
(22) Filed: **Jan. 9, 1998**

Related U.S. Application Data
(60) Provisional application No. 60/035,205, filed on Jan. 10, 1997.

(51) Int. Cl.⁷ **G06F 17/30**
(52) U.S. Cl. **707/5; 707/7; 707/501**
(58) Field of Search **707/100, 5, 7, 707/513, 1-3, 10, 104, 501; 345/440; 382/226, 229, 230, 231**

(56) References Cited
U.S. PATENT DOCUMENTS

4,953,106 *	8/1990	Gansner et al.	345/440
5,450,535 *	9/1995	North	395/140
5,748,954	5/1998	Mauldin	395/610
5,752,241 *	5/1998	Cohen	707/3
5,832,494 *	11/1998	Egger et al.	707/102
5,848,407 *	12/1998	Ishikawa et al.	707/2
6,014,678 *	1/2000	Inoue et al.	707/501

OTHER PUBLICATIONS

S. Jeromy Carriere et al, "Web Query: Searching and Visualizing the Web through Connectivity", Computer Networks and ISDN Systems 29 (1997), pp. 1257-1267.*
Wang et al "Prefetching in Worl Wide Web", IEEE 1996, pp. 28-32.*
Ramer et al "Similarity, Probability and Database Organisation: Extended Abstract", 1996, pp. 272.276.*

(List continued on next page.)

Primary Examiner—Thomas Black
Assistant Examiner—Uyen Le
(74) Attorney, Agent, or Firm—Harrity & Snyder L.L.P.

(57) ABSTRACT

A method assigns importance ranks to nodes in a linked database, such as any database of documents containing citations, the world wide web or any other hypermedia database. The rank assigned to a document is calculated from the ranks of documents citing it. In addition, the rank of a document is calculated from a constant representing the probability that a browser through the database will randomly jump to the document. The method is particularly useful in enhancing the performance of search engine results for hypermedia databases, such as the world wide web, whose documents have a large variation in quality.

29 Claims, 3 Drawing Sheets

A
n 4

Page Rank Patented

A copy of the front page of the patent of the PageRank algorithm; Larry Page is credited as the inventor; the patent was awarded to Stanford University; the patent was filed January 1998

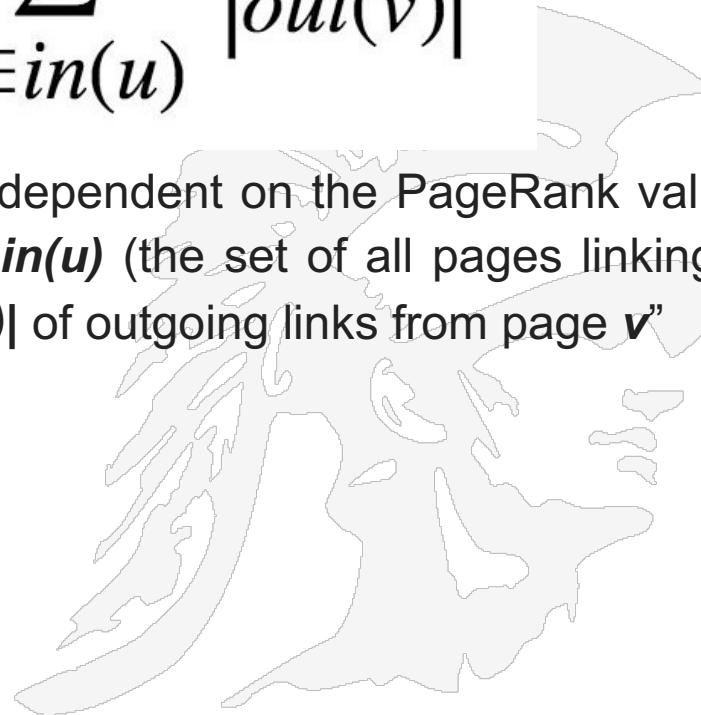
The PageRank patent expires in 2017. Google holds a perpetual license to the patent.

Google has never pursued other search engine companies for using the PageRank algorithm

Initial PageRank Formulation

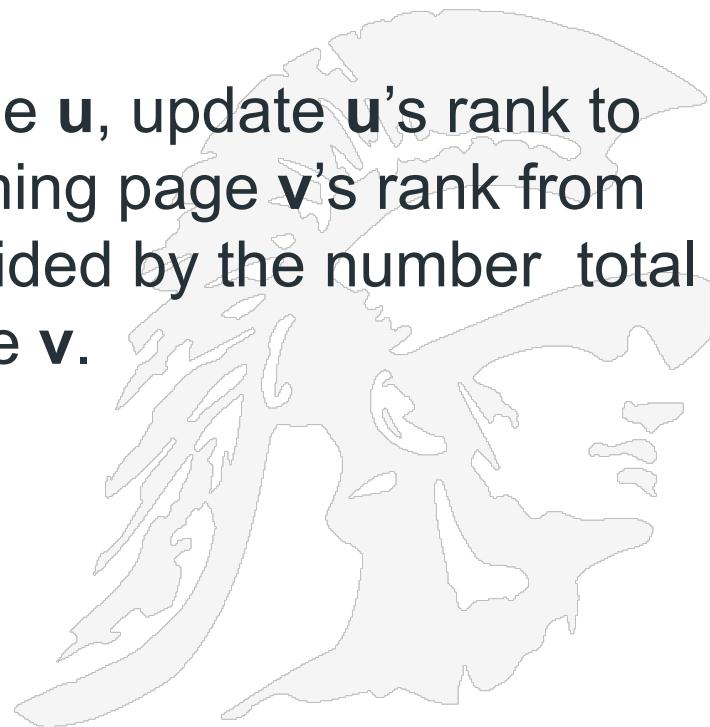
$$PR(u) = \sum_{v \in in(u)} \frac{PR(v)}{|out(v)|}$$

- “the PageRank value for a page u is dependent on the PageRank values for each page v contained in the set $in(u)$ (the set of all pages linking to page u), divided by the number $|out(v)|$ of outgoing links from page v ”



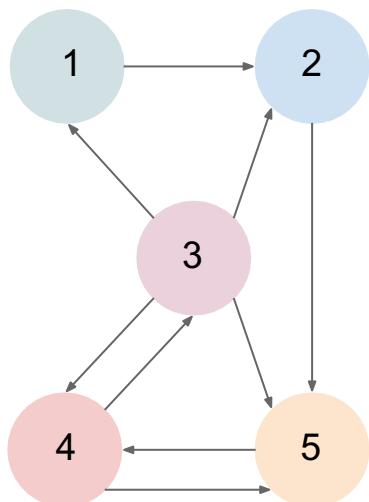
Steps for Simplified Algorithm

1. **Iteration 0:** Initialize all ranks to be $1/(\text{number of total pages})$.
2. **Iteration 1:** For each page u , update u 's rank to be the sum of each incoming page v 's rank from the previous iteration, divided by the number total number of links from page v .



The Simplified PageRank Algorithm

Example 1



	Iteration 0	Iteration 1
P ₁	1/5	1/20
P ₂	1/5	5/20
P ₃	1/5	1/10
P ₄	1/5	5/20
P ₅	1/5	7/20

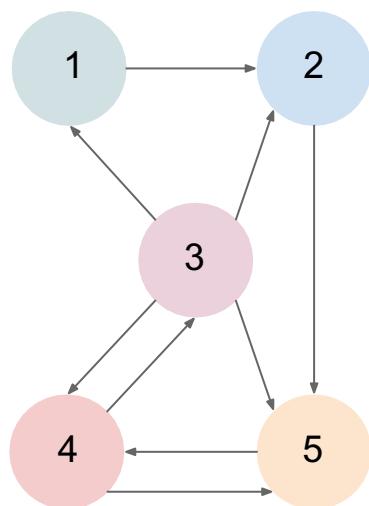
1. Iteration 0: Initialize all pages to have rank $\frac{1}{5}$.
2. Iteration 1:
3. P₁: has 1 link from P₃, and P₃ has 4 outbound links, so we take the rank of P₃ from iteration 0 and divide it by 4, which results in rank $(\frac{1}{5})/4 = 1/20$ for P₁
4. P₂: has 2 links from P₁ and P₃, P₁ has 1 outbound link and P₃ has 4 outbound links, so we take (the rank of P₁ from iteration 0 and divide it by 1) and add that to (the rank of P₃ from iteration 0 and divided that by 4) to get $\frac{1}{5} + \frac{1}{20} = 5/20$ for P₂

$$\text{PR}(P_1) = \frac{1}{5} + (\frac{1}{5})/4 = 1/20$$

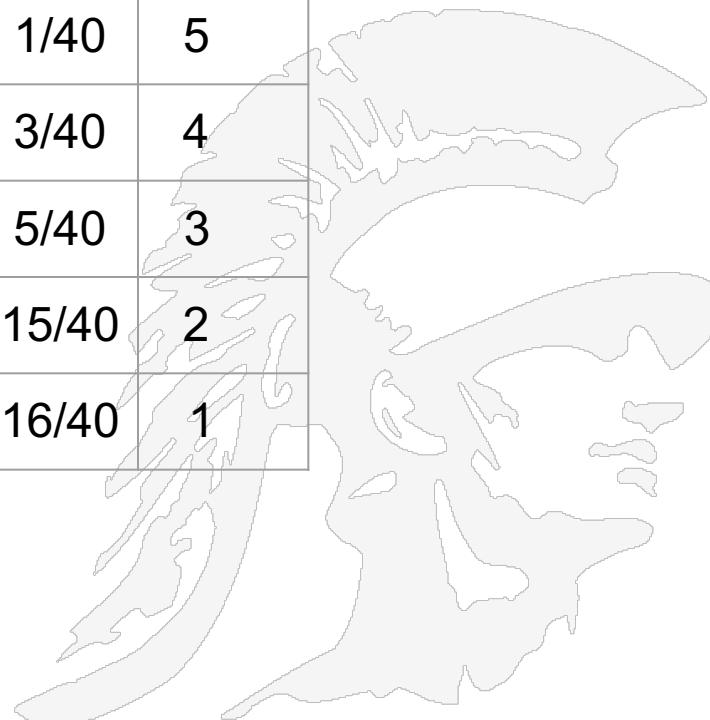
$$\text{PR}(P_2) = \frac{1}{5} + (\frac{1}{5})/4 = 5/20$$

The Simplified PageRank Algorithm

Example 1: After 2 iterations



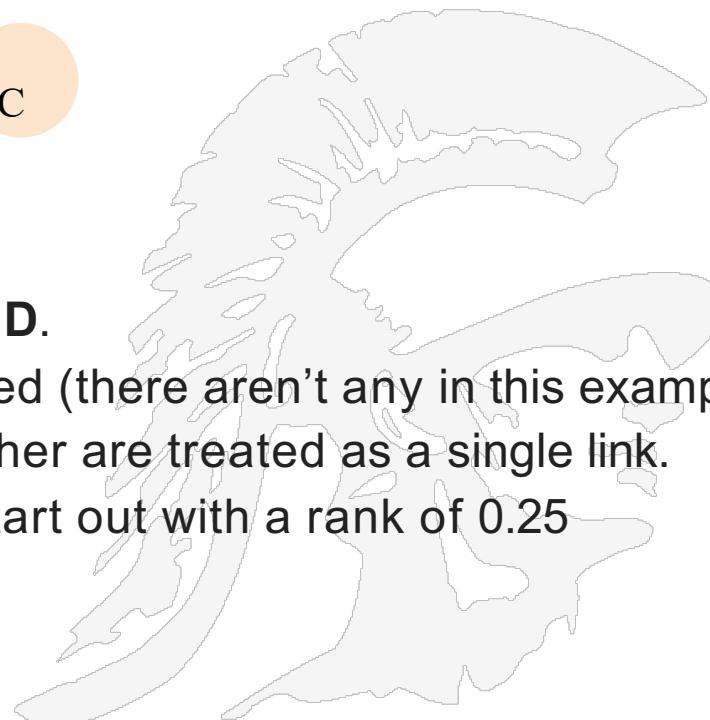
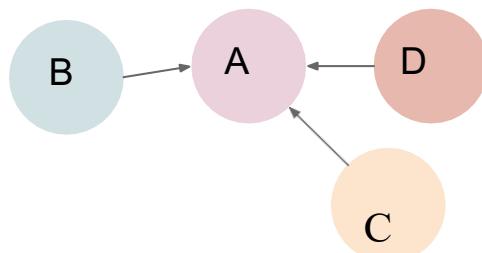
	Iteration 0	Iteration 1	Iteration 2	Final Ranking
P ₁	1/5	1/20	1/40	5
P ₂	1/5	5/20	3/40	4
P ₃	1/5	1/10	5/40	3
P ₄	1/5	5/20	15/40	2
P ₅	1/5	7/20	16/40	1



$$\text{PR}(P_5) = \frac{1}{5} + \frac{1}{5} * \frac{1}{4} + \frac{1}{5} * \frac{1}{2} = \\ \frac{7}{20}$$

Another Example

Example 2

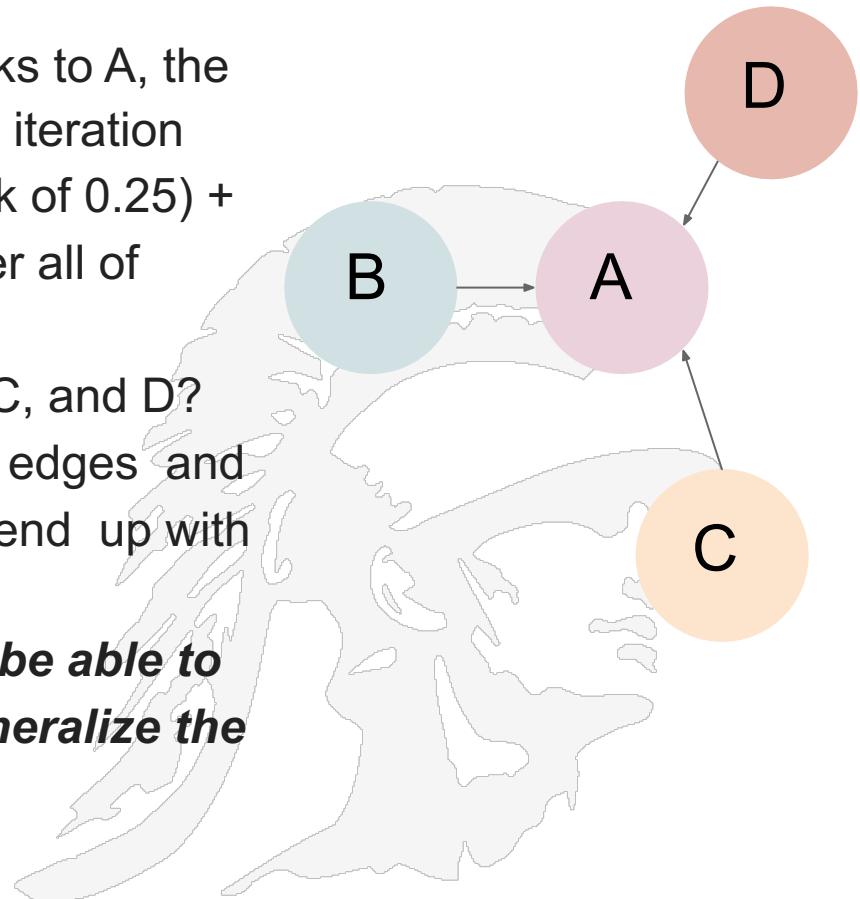


- Say we have four pages: **A, B, C and D**.
- Links from a page to itself are ignored (there aren't any in this example).
- Multiple links from one page to another are treated as a single link.
- In this example, every page would start out with a rank of 0.25

Another Example

Example 2

- Since B, C, and D all have outbound links to A, the Pagerank of A will be **0.75** upon the first iteration
 - ◆ (B with rank of 0.25) + (C with rank of 0.25) + (D with rank of 0.25) would transfer all of those ranks to A
- But wait! What about ranks of pages B,C, and D? Because B, C, and D have no incoming edges and they give all their rank to A, they will all end up with a rank of 0. This doesn't add up to 1 . . .
- ***So the simplified algorithm needs to be able to handle border cases, so we must generalize the PageRank algorithm!***



Complete PageRank Algorithm

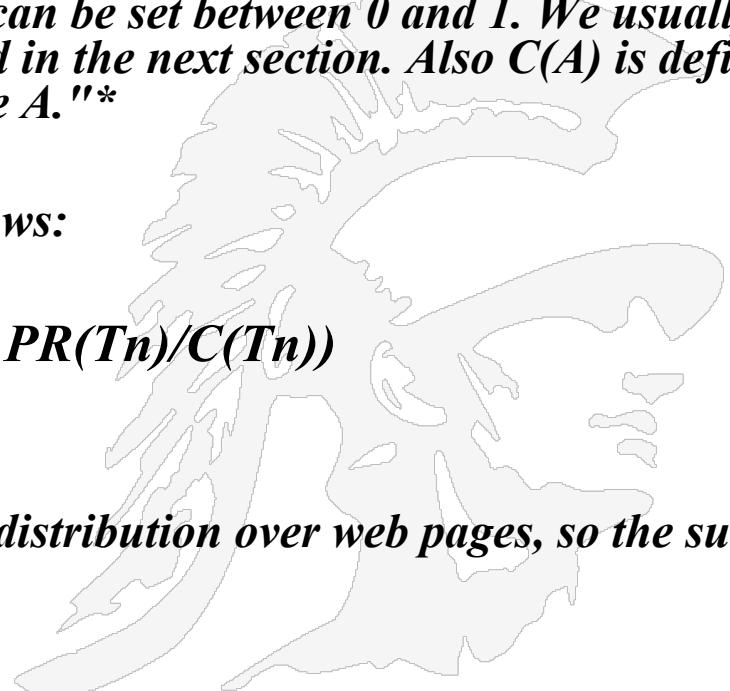
- Quoting from the original Google paper, PageRank is defined like this:

*"We assume page A has pages T1...Tn which point to it (i.e., are citations). The parameter d is a damping factor which can be set between 0 and 1. We usually set d to 0.85. There are more details about d in the next section. Also C(A) is defined as the number of links going out of page A."**

The PageRank of a page A is given as follows:

$$PR(A) = (1-d) + d \left(PR(T1)/C(T1) + \dots + PR(Tn)/C(Tn) \right)$$

- Note:
 - That the PageRanks form a probability distribution over web pages, so the sum of all web pages' PageRanks will be one.
- **The Anatomy of a Large-Scale Hypertextual Web Search Engine by Brin and Page,
<http://infolab.stanford.edu/pub/papers/google.pdf>*



Explanation

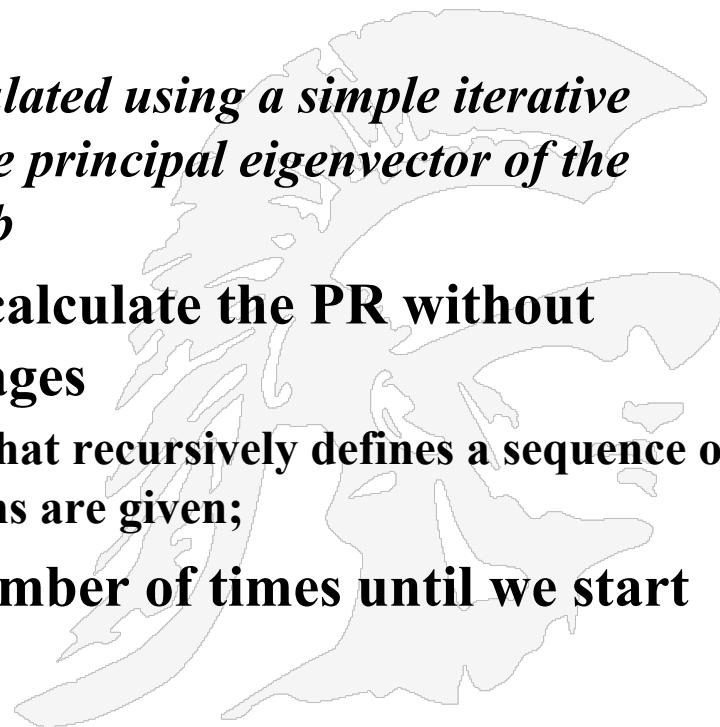
$$PR(A) = (1-d) + d \left(PR(T1)/C(T1) + \dots + PR(Tn)/C(Tn) \right)$$

- *PR(A) is PageRank of Page A (one we want to calculate)*
- *PR(T1) is the PageRank of Site T1 pointing to page A*
- *C(T1) is the number of outgoing links of page T1*
- *PR(Tn)/C(Tn) : If page A has a backlink from page “Tn” the share of the vote page A will get is “PR(Tn)/C(Tn)”*
- *d(...) : All these fractions of votes are added together but, to stop the other pages having too much influence, this total vote is “damped down” by multiplying it by the factor “d”, (0.85)*
- *(1-d) : Since “sum of all web pages’ PageRanks will be one”. Even if the d(...) is 0 then the page will get a small PR of 0.15*

How PageRank is Calculated

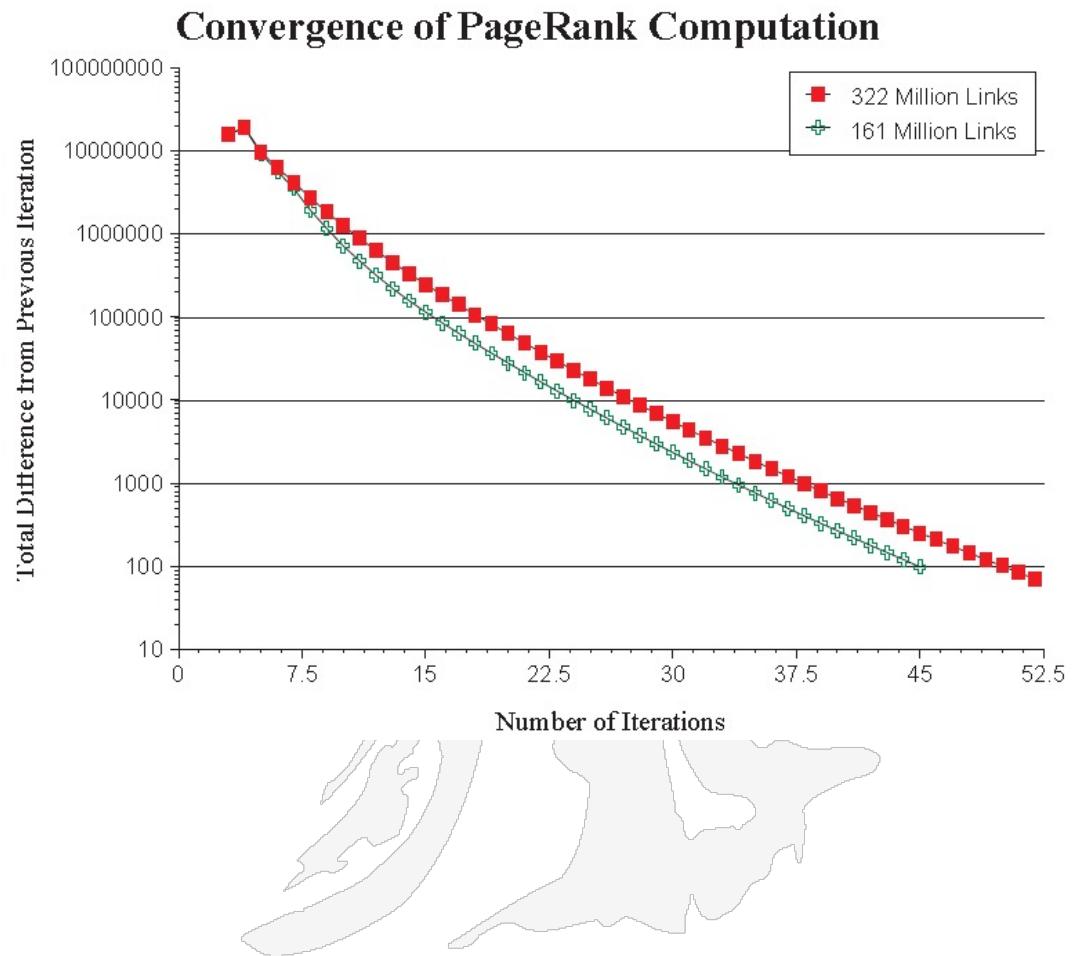
- PR of each page depends on PR of other pages which are pointing to it. But we don't know PR of a given page until the PR of other pages is calculated and so on...
- From the Google paper:

PageRank or $PR(A)$ can be calculated using a simple iterative algorithm, and corresponds to the principal eigenvector of the normalized link matrix of the web
- What this means is that we can calculate the PR without knowing the final PR of other pages
 - Recurrence Relation: an equation that recursively defines a sequence of values once one or more initial terms are given;
- We calculate PR iteratively a number of times until we start converging to the same value.



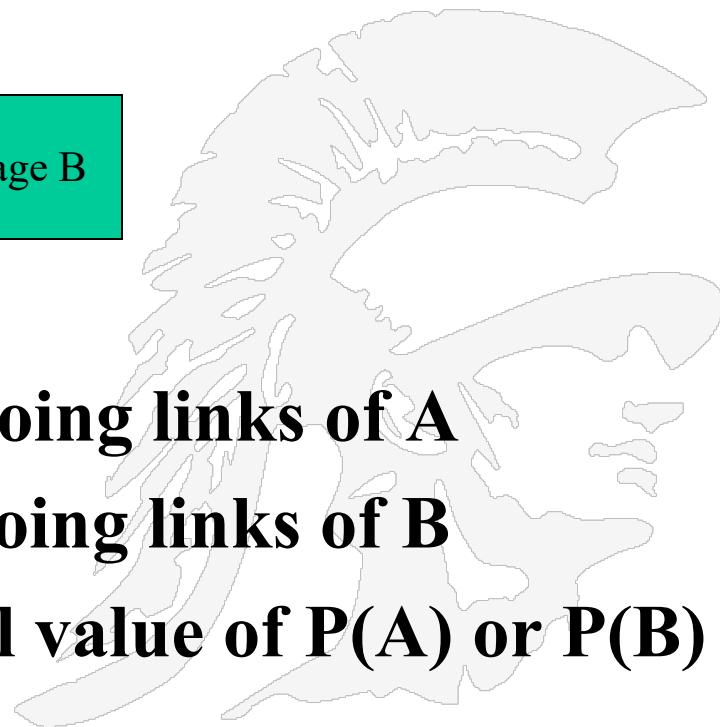
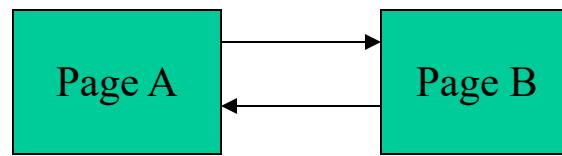
How Fast Does the PageRank Algorithm Converge

- Early experiments on Google used 322 million links
- PR (322 Million Links): 52 iterations
- PR (161 Million Links): 45 iterations
- Number of iterations required for convergence is empirically (but not formally derived) $O(\log n)$ (where n is the number of links)
- Hence the calculation is quite efficient



Computing PageRank by Iteration Example

- Consider 2 pages: Page A and Page B pointing to each other.



- $C(A) = 1$, number of outgoing links of A
- $C(B) = 1$, number of outgoing links of B
- What should be the initial value of $P(A)$ or $P(B)$?

Guess 1:

- Suppose the initial values are :
 - $P(A) = 1$ and $P(B) = 1$ and $d = 0.85$; then
$$PR(A) = (1 - d) + d * (PR(B)/1)$$
$$PR(B) = (1 - d) + d * (PR(A)/1)$$

i.e.

- $PR(A) = 0.15 + 0.85 * 1$
 $= 1$
- $PR(B) = 0.15 + 0.85 * 1$
 $= 1$
- In one iteration we are done
- Let's try another set of initial values.



Guess 2 With 3 Iterations

- Initial Values : $P(A) = 0$, $P(B) = 0$ and $d= 0.85$

$$PR(A) = (1 - d) + d(PR(B)/1)$$

$$PR(B) = (1 - d) + d(PR(A)/1)$$

- $PR(A) = 0.15 + 0.85 * 0 = 0.15$

$$PR(B) = 0.15 + 0.85 * 0.15 = 0.2775$$

Iterating again we get:

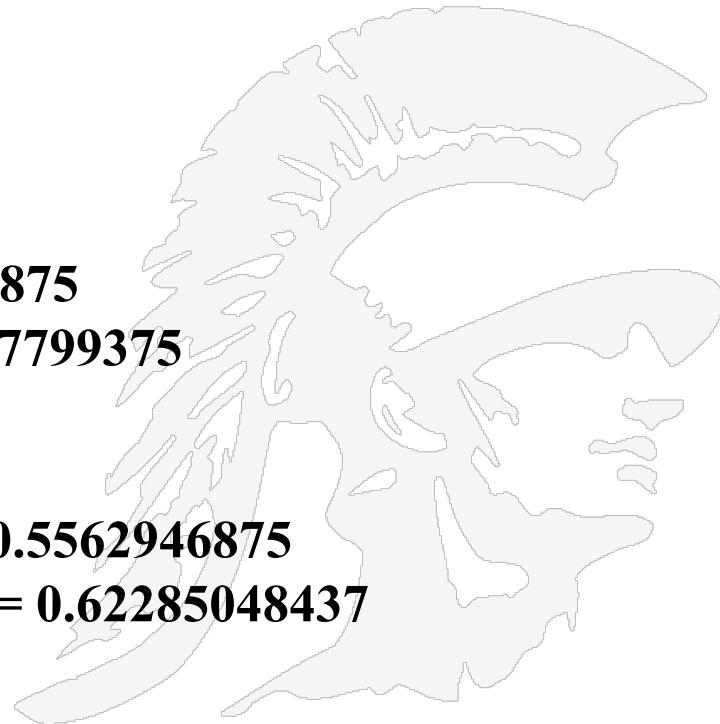
- $PR(A) = 0.15 + 0.85 * 0.2775 = 0.385875$

$$PR(B) = 0.15 + 0.85 * 0.385875 = 0.47799375$$

And iterating again

- $PR(A) = 0.15 + 0.85 * 0.47799375 = 0.5562946875$

$$PR(B) = 0.15 + 0.85 * 0.5562946875 = 0.62285048437$$



Guess 2: Continued...

- After 20 iterations...
- $\text{PR(A)} = 0.99$
- $\text{PR(B)} = 0.99$
- Both approaching to 1.

	A	B	C	D
1	C(A)		1	
2	C(B)		1	
3				
4	Iterations	PR(A)	PR(B)	
5				
6	0	0		0
7	1	0.15		0.2775
8	2	0.385875		0.47799375
9	3	0.556294688		0.622850484
10	4	0.679422912		0.727509475
11	5	0.768383054		0.803125596
12	6	0.832656756		0.857758243
13	7	0.879094506		0.89723033
14	8	0.912645781		0.925748914
15	9	0.936886577		0.94635359
16	10	0.954400552		0.961240469
17	11	0.967054399		0.971996239
18	12	0.976196803		0.979767283
19	13	0.98280219		0.985381862
20	14	0.987574582		0.989438395
21	15	0.991022636		0.99236924
22	16	0.993513854		0.994486776
23	17	0.99531376		0.996016696
24	18	0.996614191		0.997122063
25	19	0.997553753		0.99792069
26	20	0.998232587		0.998497699
27				

Guess 3:

- Initial Values : $P(A) = 40$ and $P(B) = 40$
 $d = 0.85$

$$PR(A) = (1 - d) + d(PR(B)/1)$$

$$PR(B) = (1 - d) + d(PR(A)/1)$$

- Notice decreasing value of PR
- The page rank is again approaching to 1.
- So it doesn't matter where you start your guess, once the PageRank calculations have settled down, the "*normalized probability distribution*" (the average PageRank for these two pages) will be 1.0

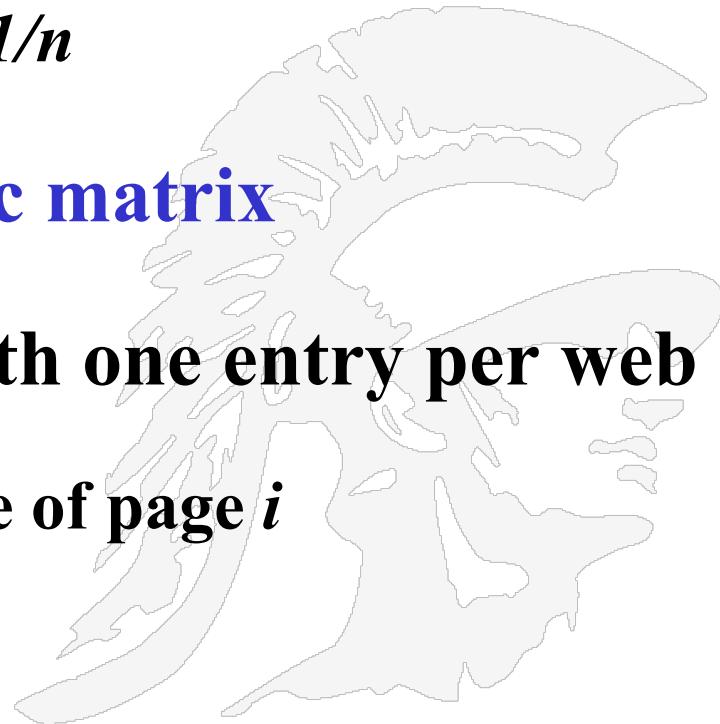
	A	B	C	D
1	C(A)		1	
2	C(B)		1	
3				
4	Iterations	PR(A)	PR(B)	
5				
6	0	0		40
7	1	34.15		29.1775
8	2	24.950875		21.35824375
9	3	18.30450719		15.70883111
10	4	13.50250644		11.62713048
11	5	10.03306091		8.678101769
12	6	7.526386504		6.547428528
13	7	5.715314249		5.008017112
14	8	4.406814545		3.895792363
15	9	3.461423509		3.092209982
16	10	2.778378485		2.511621712
17	11	2.284878455		2.092146687
18	12	1.928324684		1.789075981
19	13	1.670714584		1.570107397
20	14	1.484591287		1.411902594
21	15	1.350117205		1.297599624
22	16	1.252959681		1.215015728
23	17	1.182763369		1.155348864
24	18	1.132046534		1.112239554
25	19	1.095403621		1.081093078
26	20	1.068929116		1.058589749
27				

PageRank Convergence?

- Some observations about the damping factor
- The damping factor value and its effect:
 - For certain graphs the simple update rule can cause pagerank to accumulate and get stuck in certain parts of the graph
 - E.g. if a page has no outgoing links to other pages, it is called a sink
 - The simplified pagerank algorithm can get stuck in sinks
 - This is fixed by having each node on every round
 - Give a d fraction of its pagerank to its neighbors
 - Give a $(1-d)$ fraction of its pagerank to everyone in the graph
 - As a result, pages with no incoming links will get some pagerank
 - If too high, more iterations are required
 - If too low, you get repeated over-shoot,
 - Both above and below the average – the numbers just swing like pendulum and never settle down.

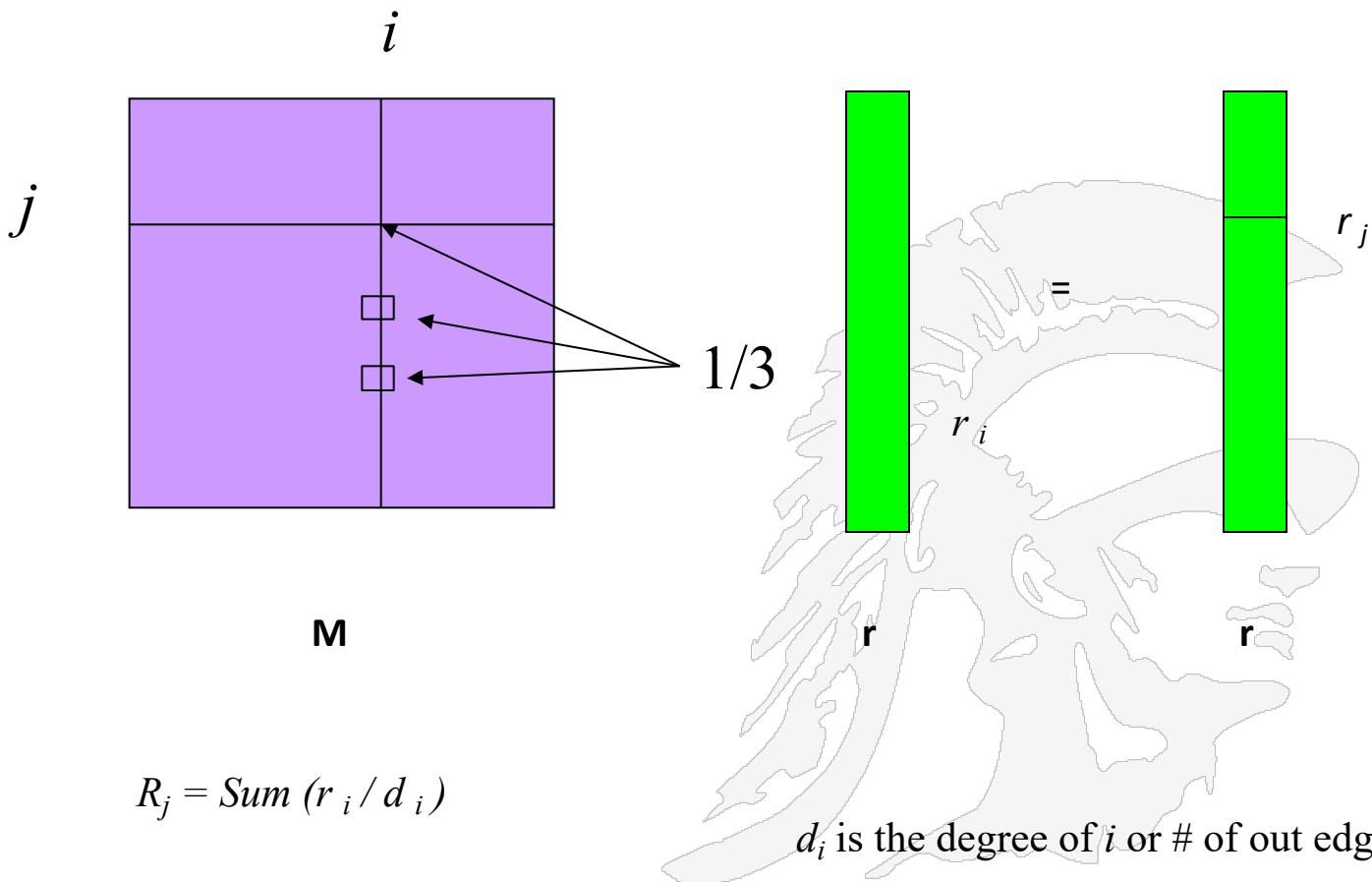
Matrix Formulation for Computing PageRank

- Suppose we define a matrix M to have one row and one column for each web page
- Suppose page j has n outlinks
 - If j points to i , then $M_{ij}=1/n$
 - Else $M_{ij}=0$
- M is a **column stochastic matrix**
 - i.e. its columns sum to 1
- Suppose r is a vector with one entry per web page
 - r_i is the importance score of page i
 - Call it the **rank vector**



Flow Equation in Matrix Form

Suppose page i links to 3 pages, $d_i = 3$, including j



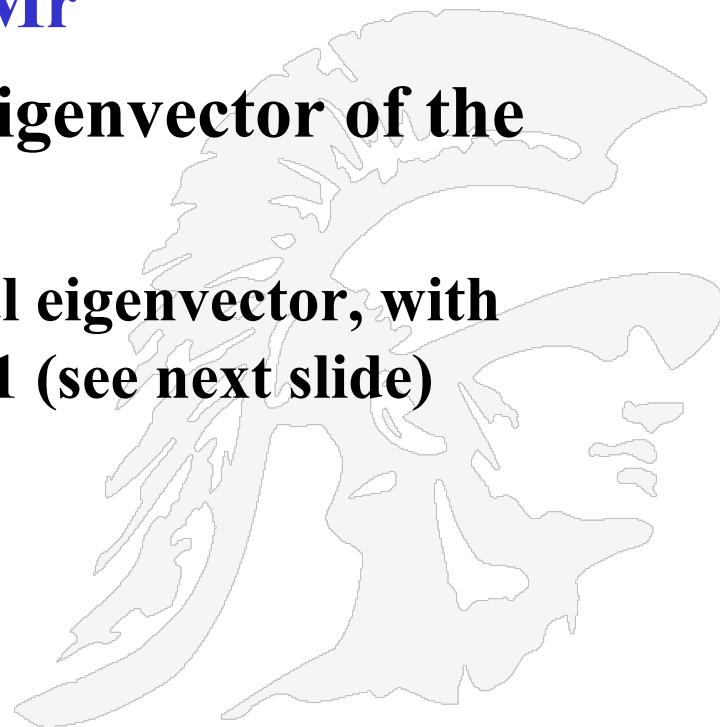
Eigenvector Formulation

- The flow equations can be written

$$\mathbf{r} = \mathbf{M}\mathbf{r}$$

- So the rank vector is an eigenvector of the stochastic web matrix

- In fact, its first or principal eigenvector, with corresponding eigenvalue 1 (see next slide)



Eigenvalue and Eigenvector

- Eigenvalues and Eigenvectors are properties of a matrix
- In general, a matrix acts on a vector by changing both its magnitude and direction
- However, a matrix may act on certain vectors by changing only their magnitude, and leaving their direction unchanged – *Eigenvector*
- A matrix acts on an eigenvector by multiplying its magnitude by a factor called the *Eigenvalue*

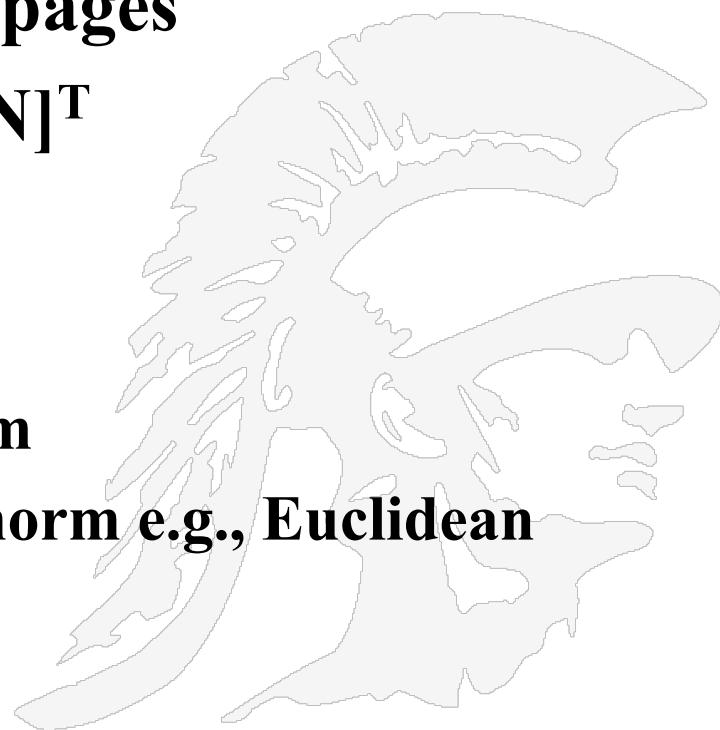
Given a linear transformation A , a non-zero vector x is defined to be an eigenvector of the transformation if it satisfies the eigenvalue equation

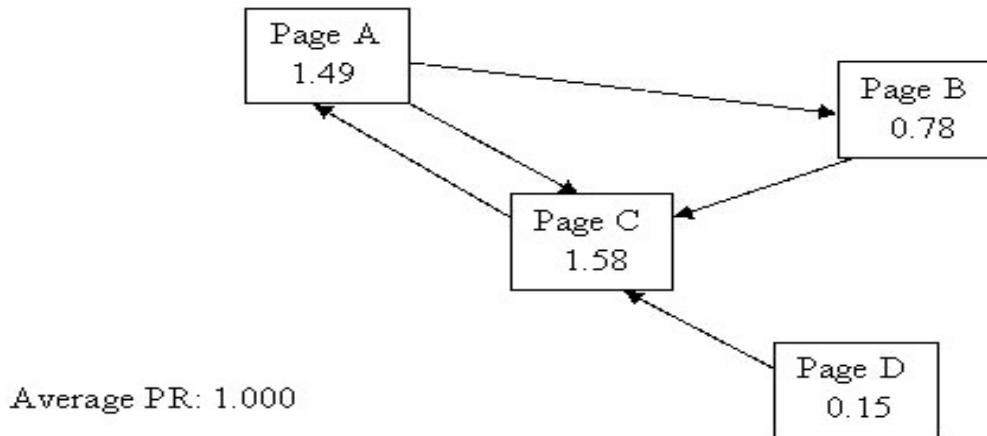
$$A \mathbf{x} = \lambda \mathbf{x}$$

In this situation, the scalar λ is called an *eigenvalue* of A corresponding to the eigenvector \mathbf{x}

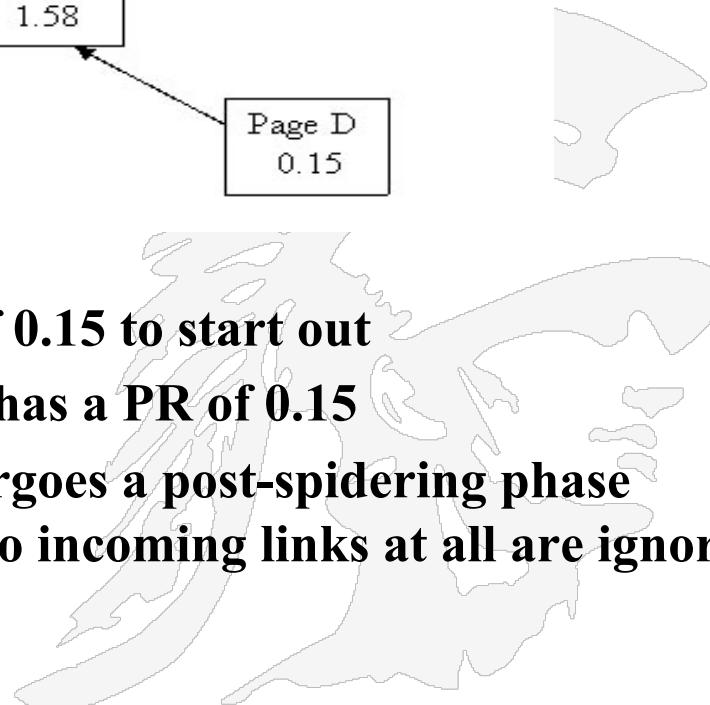
Power Iteration method

- Simple iterative scheme (aka **relaxation**)
- Suppose there are **N** web pages
- Initialize: $r^0 = [1/N, \dots, 1/N]^T$
- Iterate: $r^{k+1} = Mr^k$
- Stop when $|r^{k+1} - r^k|_1 < \epsilon$
 - $|x|_1 = \sum_{1 \leq i \leq N} |x_i|$ is the L₁ norm
 - Can use any other vector norm e.g., Euclidean

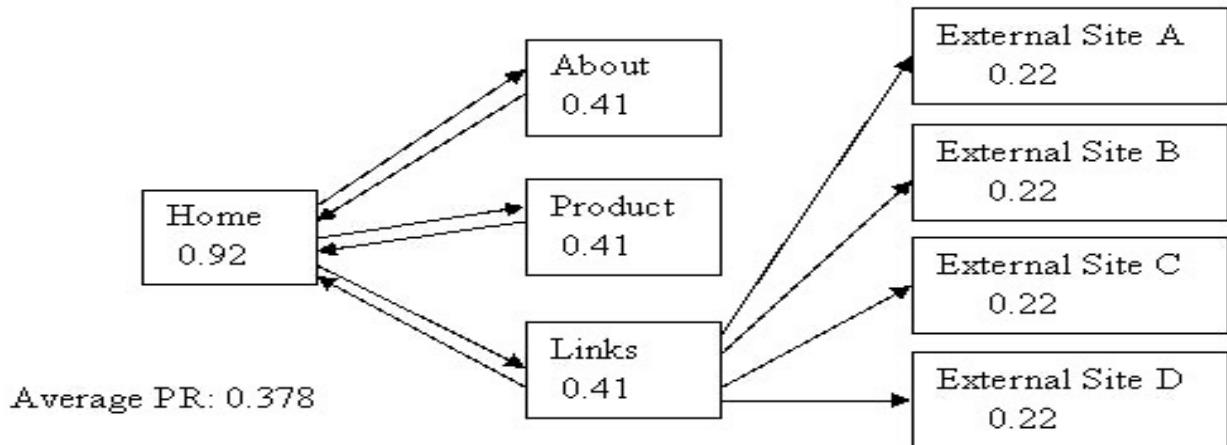


Examples of
Relative PageRanks 1:

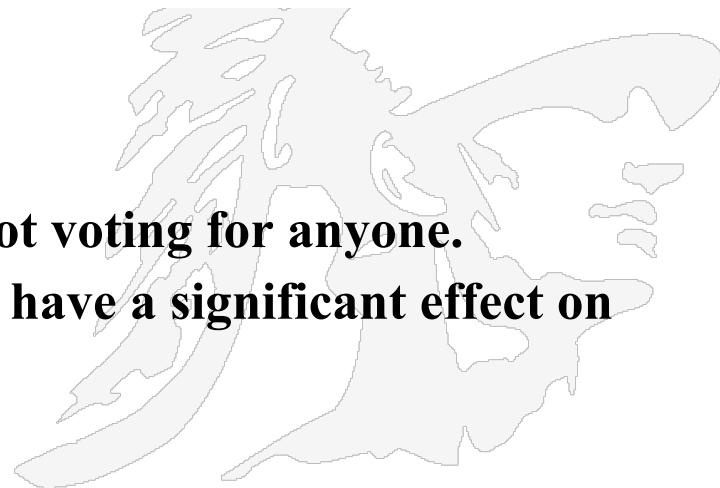
- **Observations:**
 - Every page has at least a PR of 0.15 to start out
 - Page D has no votes but still it has a PR of 0.15
 - It is believed that Google undergoes a post-spidering phase whereby any pages that have no incoming links at all are ignored wrt PageRank
 - Examples on the following pages are taken from <http://www.sirgroane.net/google-page-rank/>



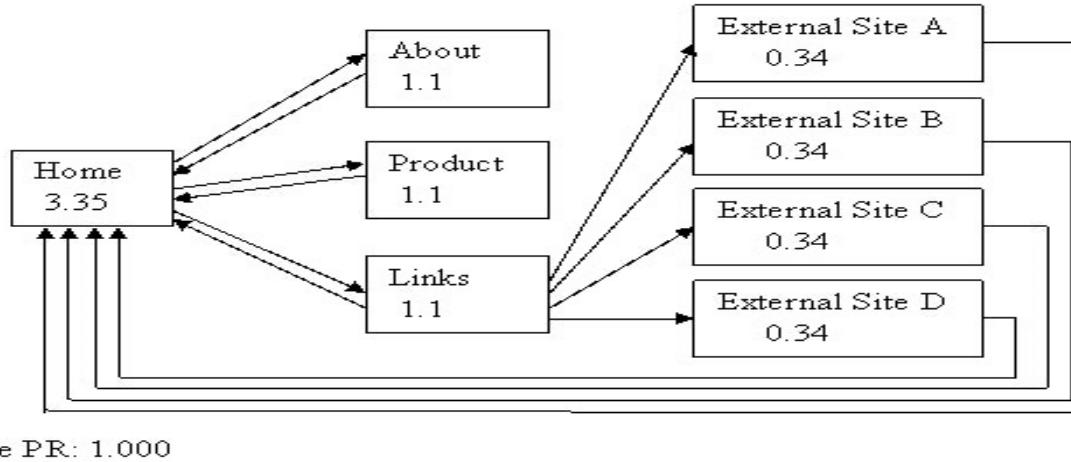
Example 2: Simple hierarchy with Some Outgoing Links



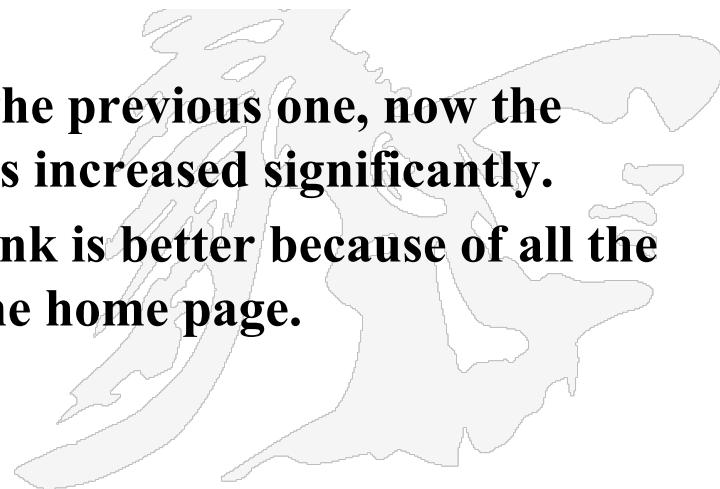
- **Observations:**
 - Home has the most PR
 - But average PR is 0.378
 - “External site” pages are not voting for anyone.
 - Links within your own site can have a significant effect on PageRank.



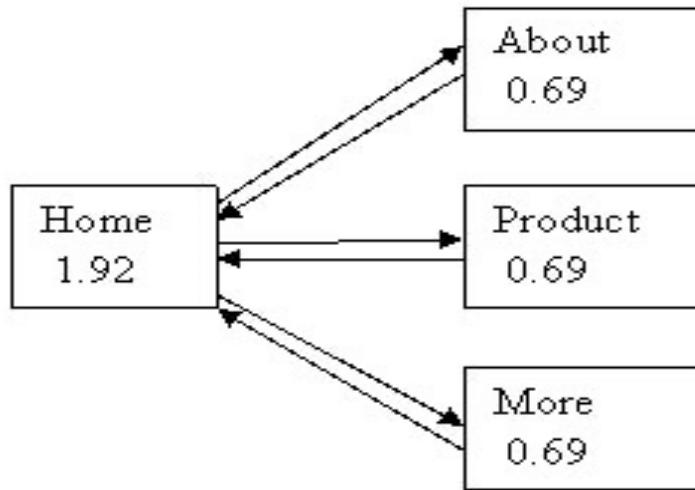
Example 3: Linking External Sites Back into our Home Page



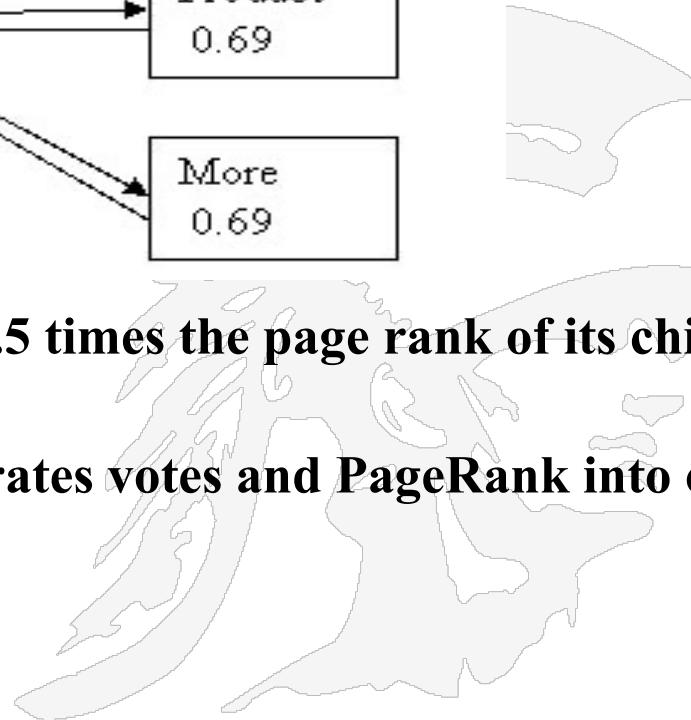
- **Observations:**
 - Comparing this example with the previous one, now the Pagerank of the Home Page has increased significantly.
 - Moreover, the average PageRank is better because of all the external sites linking back to the home page.



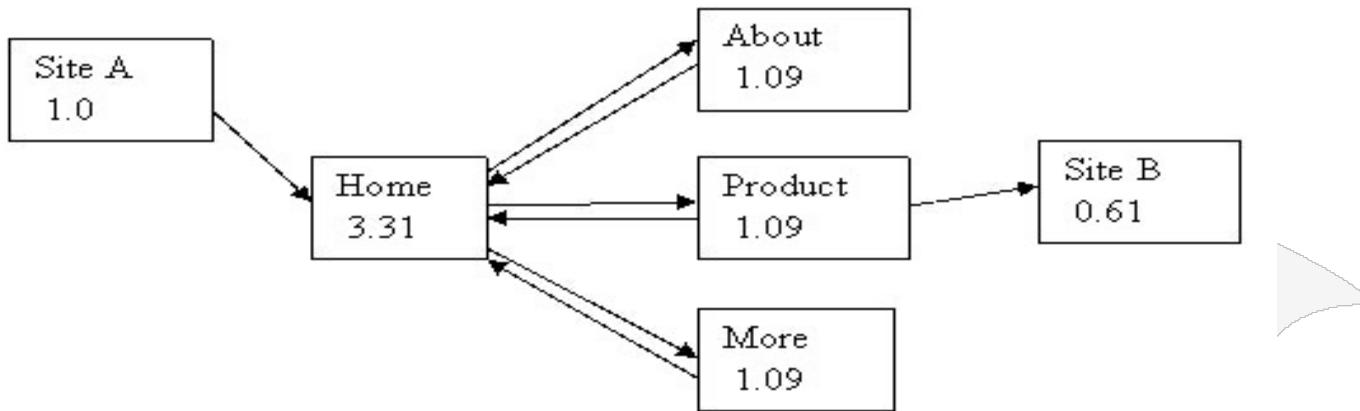
Example 4: Simple Hierarchy



- **Observations:**
 - Home Page has PageRank of 2.5 times the page rank of its child pages.
 - A hierarchy structure concentrates votes and PageRank into one page.

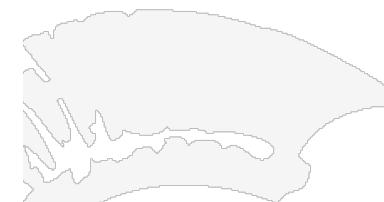
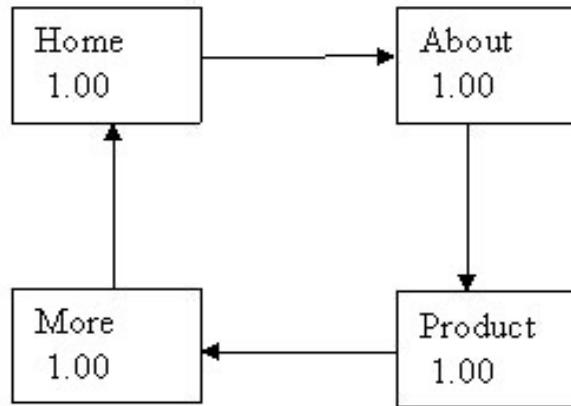


Example 5: Hierarchical – But with One Link In and One Out



- **Observations:**
 - The PageRank of Home page has increased from 1.92 (Previous Example) to 3.31
 - Site A contributed 0.85 PR to Home page and the raised PageRank in the “About”, “Product” and “More” pages has had a lovely “feedback” effect, pushing up the home page’s PageRank even further!
- A well structured site will amplify the effect of any contributed PR.

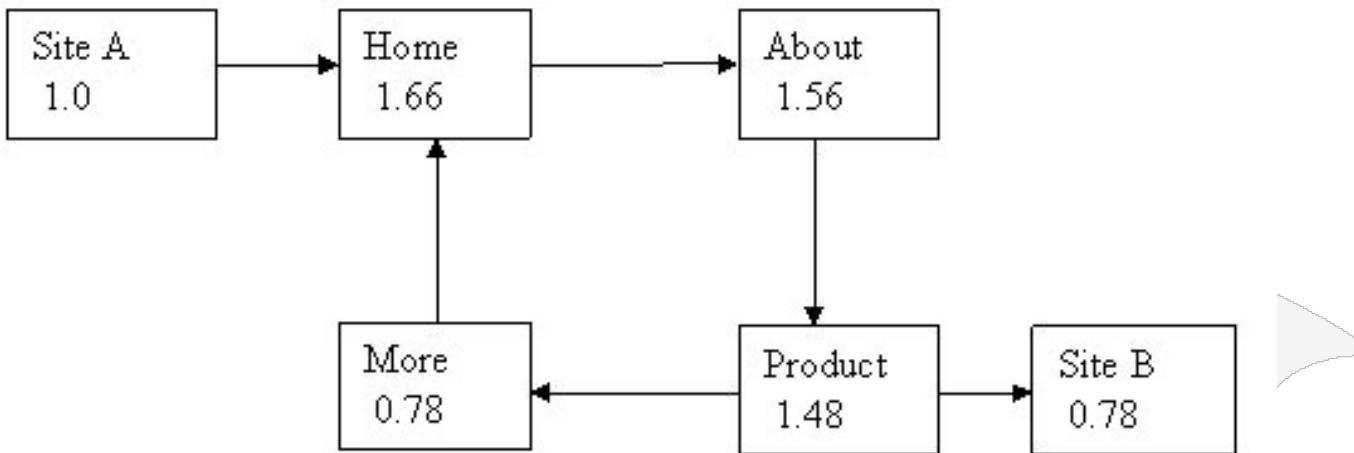
Example 6: Looping



- **Observations:**

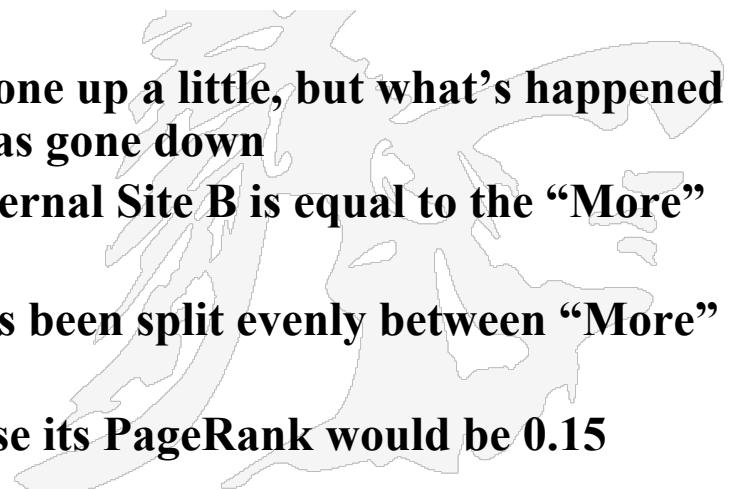
- All the pages have the same number of incoming links, i.e. all pages are of equal importance to each other.
- Each page has PR of 1.0
- Average PR is 1.0

Example 7: Looping – But with a Link in and a Link Out

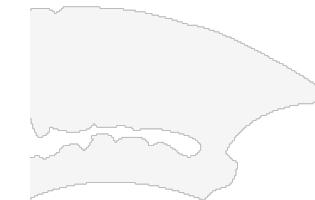
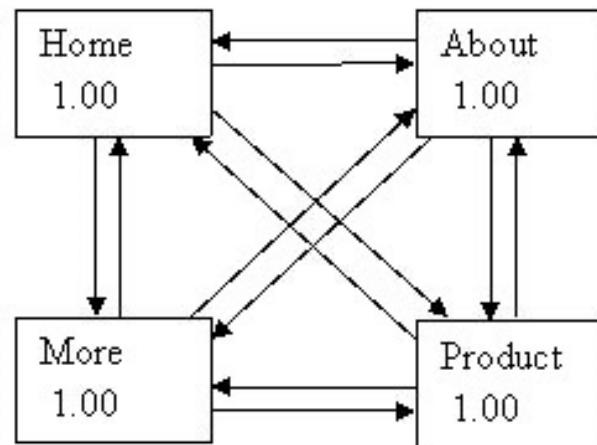


- **Observations:**

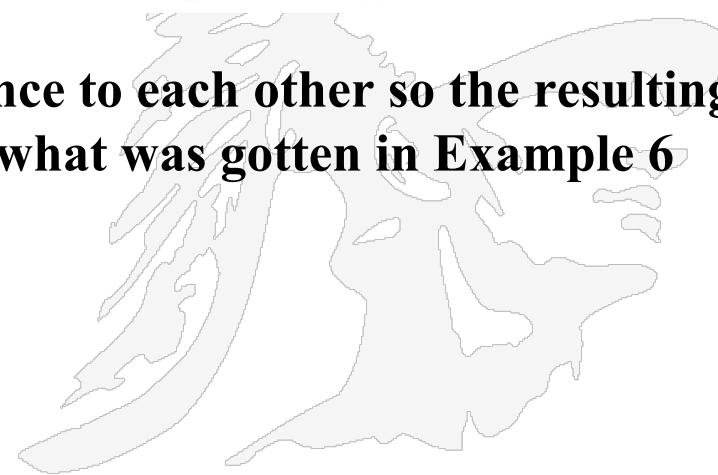
- PageRank of our home page has gone up a little, but what's happened to the “More” page? Its PageRank has gone down
- Now the Pagerank value of the external Site B is equal to the “More” page.
- The vote of the “Product” page has been split evenly between “More” page and the external site B.
- This is good for Site B for otherwise its PageRank would be 0.15



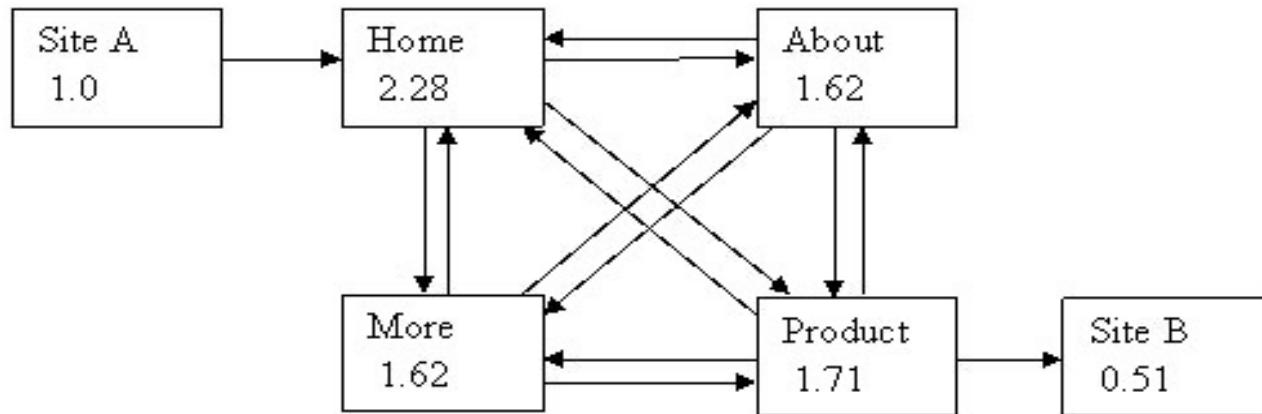
Example 8: Extensive Interlinking or Fully Meshed



- **Observations:**
 - All pages are of equal importance to each other so the resulting PageRank is no different than what was gotten in Example 6

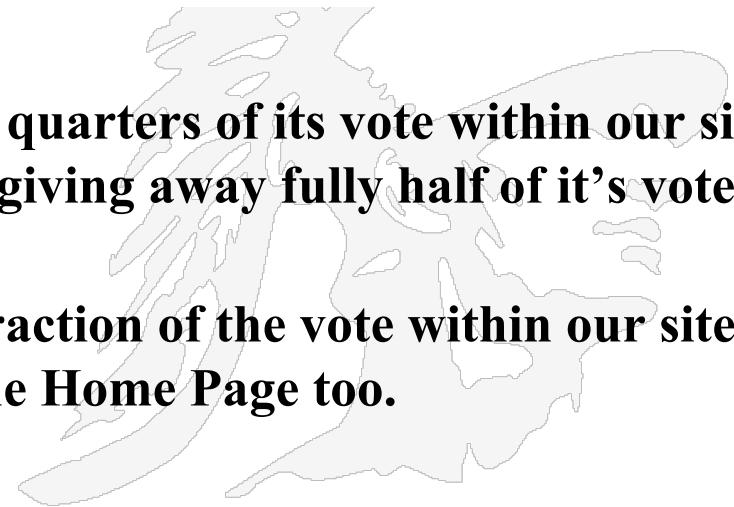


Example 9: Fully Meshed – But with One Vote in and One Vote Out

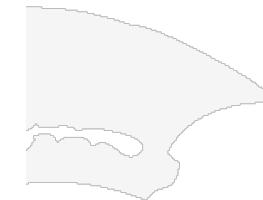
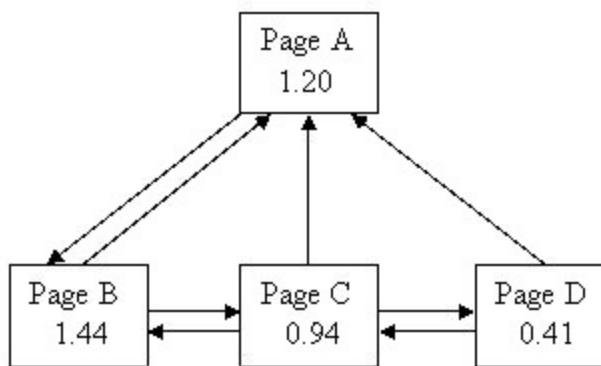


- **Observations:**

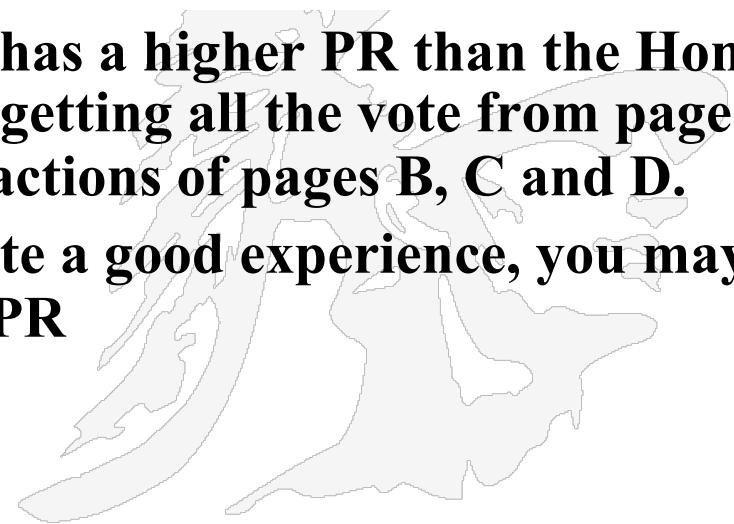
- “Product” page has kept three quarters of its vote within our site unlike example 9 where it was giving away fully half of it’s vote to the external site!
- Keeping just this small extra fraction of the vote within our site has had a very nice effect on the Home Page too.



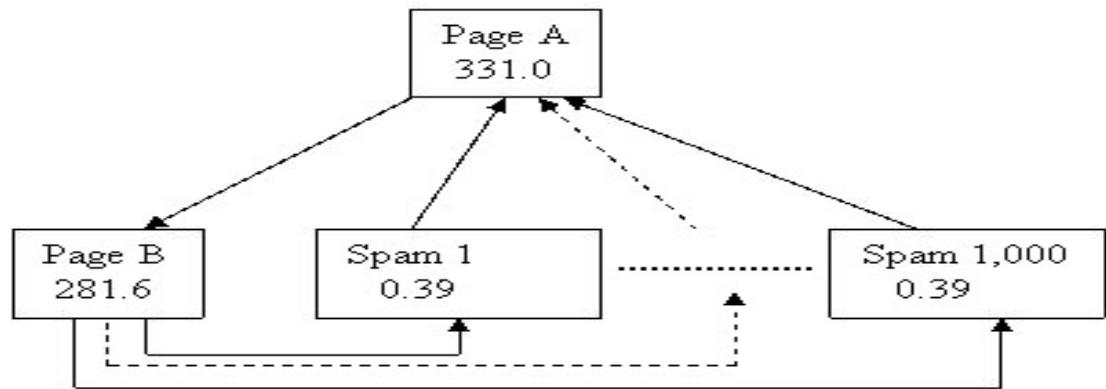
Example 10: Previous ... Next ... Documentation Page Layout



- The first page of the document has a higher PR than the Home Page! This is because page B is getting all the vote from page A, but page A is only getting fractions of pages B, C and D.
- In order to give users of your site a good experience, you may have to take a hit against your PR

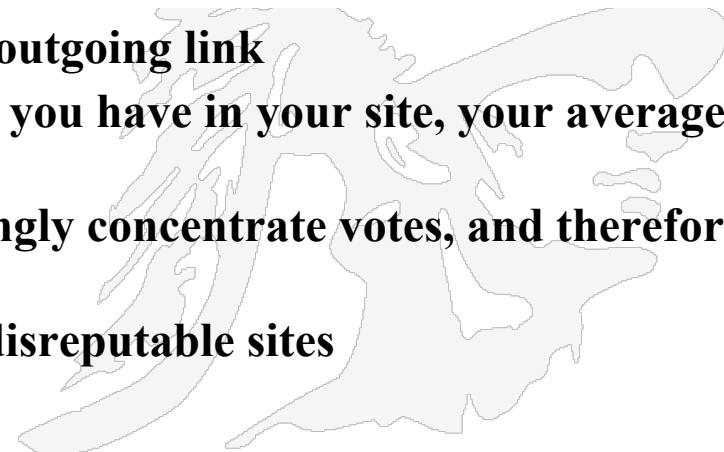


Example 11: Getting Higher PR the Wrong Way!



- **Observations:** Average PR: 1.000

- 1000 incoming links and only one outgoing link
- It doesn't matter how many pages you have in your site, your average PR will always be 1.0 at best.
- But a hierarchical layout can strongly concentrate votes, and therefore the PR, into the home page!
- This is a technique used by some disreputable sites

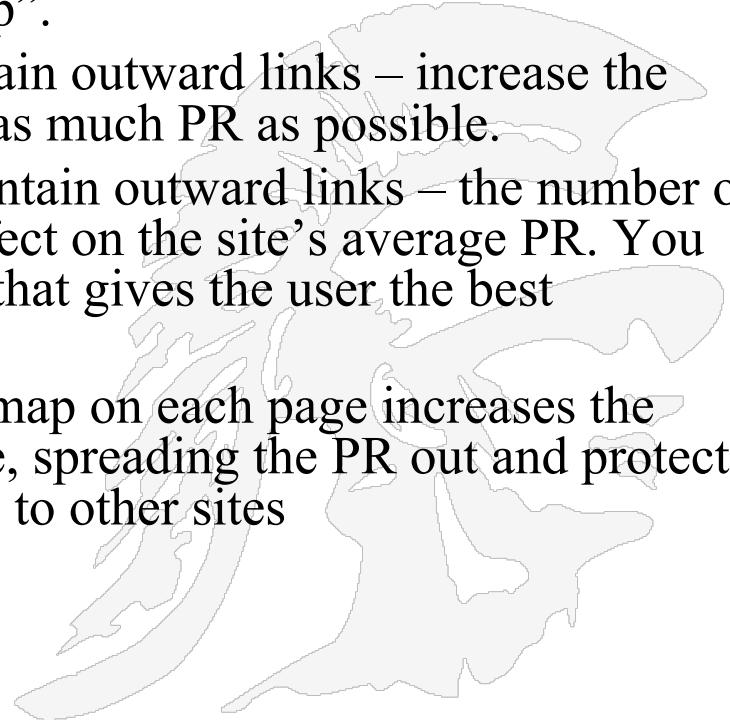


A link farm is set of web pages created with the sole aim of linking to a target page, in an attempt to improve that page's search engine ranking.

Some Suggestions Based on What We Have Seen in Examples.

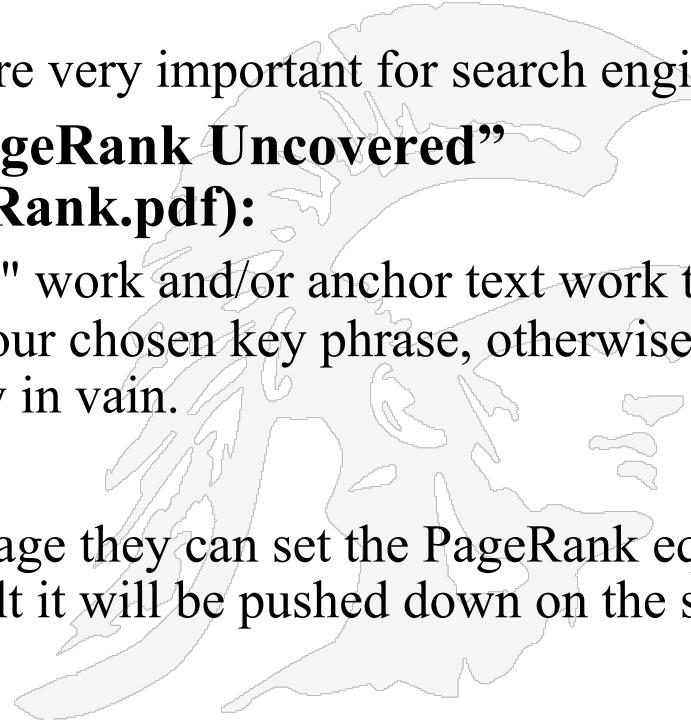
- **Suggestions for improving your page rank**

- Increasing the internal links in your site can minimize the damage to your PR when you give away votes by linking to external sites.
- If a particular page is highly important – use a hierarchical structure with the important page at the “top”.
- Where a group of pages may contain outward links – increase the number of internal links to retain as much PR as possible.
- Where a group of pages do not contain outward links – the number of internal links in the site has no effect on the site’s average PR. You might as well use a link structure that gives the user the best navigational experience.
- **Use Site Maps:** Linking to a site map on each page increases the number of internal links in the site, spreading the PR out and protecting you against your vote “donations” to other sites



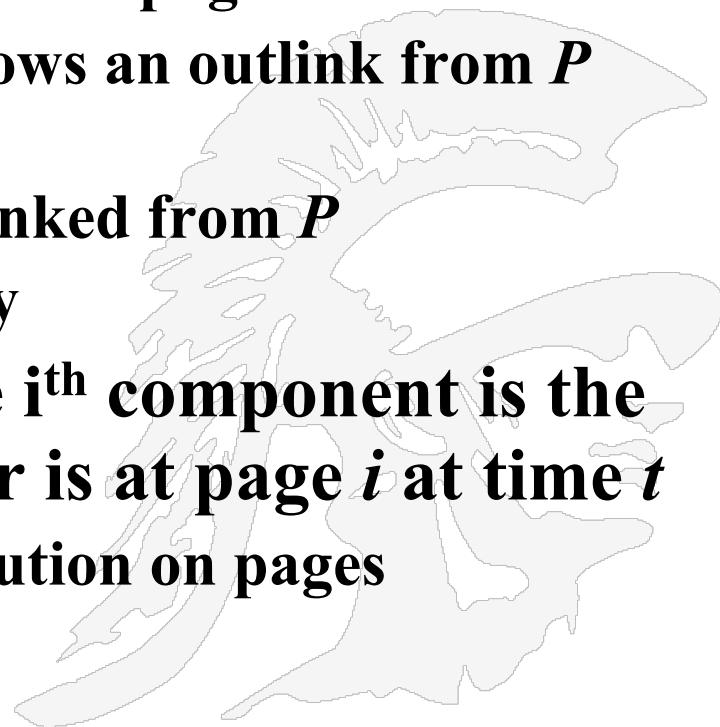
Importance of PageRank

- **PageRank is just one factor Google uses to determine a page's relevance. It assumes that people will link to your page only if they think the page is good. But this is not always true.**
- **Content is still the king!!!**
 - Anchor, body, title tags etc. still are very important for search engines
- **From Chris Ridings' Paper, “PageRank Uncovered” (<http://www.voelspriet2.nl/PageRank.pdf>):**
 - You must do enough "on the page" work and/or anchor text work to get into that subset of top pages for your chosen key phrase, otherwise your high PageRank will be completely in vain.
- **PageRank is a multiplier factor.**
 - If Google wants to penalize any page they can set the PageRank equal to a small number, even 0. As a result it will be pushed down on the search results page.



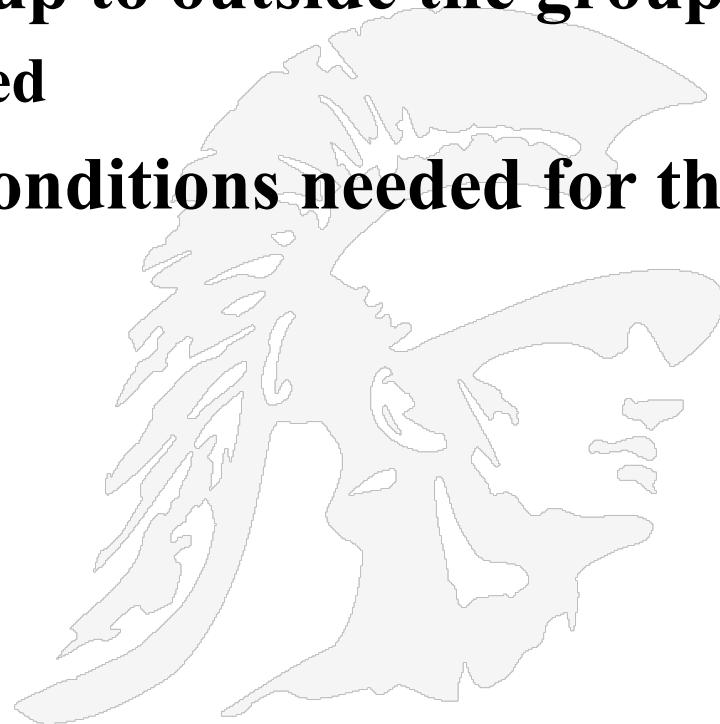
Random Walk Interpretation

- Imagine a **random web surfer**
 - At any time t , surfer is on some page P
 - At time $t+1$, the surfer follows an outlink from P uniformly at random
 - Ends up on some page Q linked from P
 - Process repeats indefinitely
- Let $p(t)$ be a vector whose i^{th} component is the probability that the surfer is at page i at time t
 - $p(t)$ is a probability distribution on pages



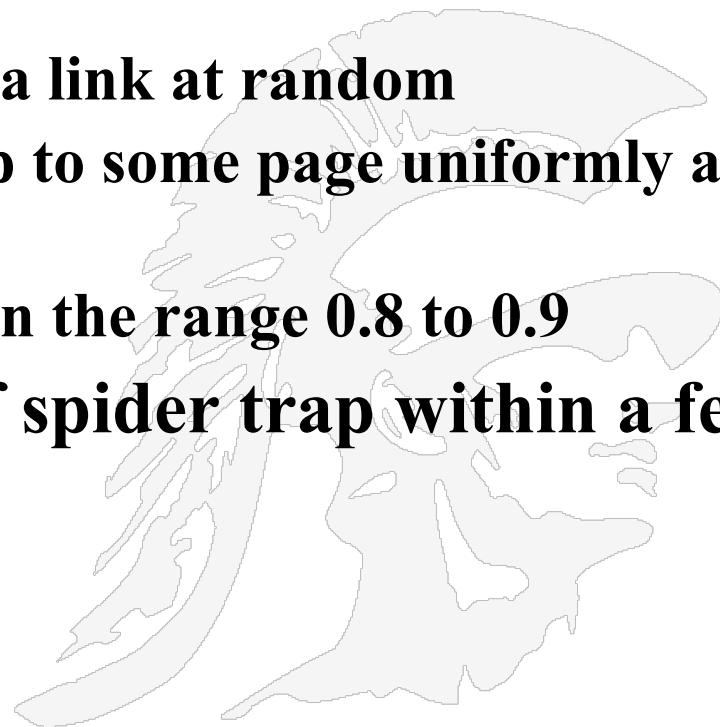
Spider traps

- A group of pages is a **spider trap** if there are no links from within the group to outside the group
 - Random surfer gets trapped
- Spider traps violate the conditions needed for the random walk theorem



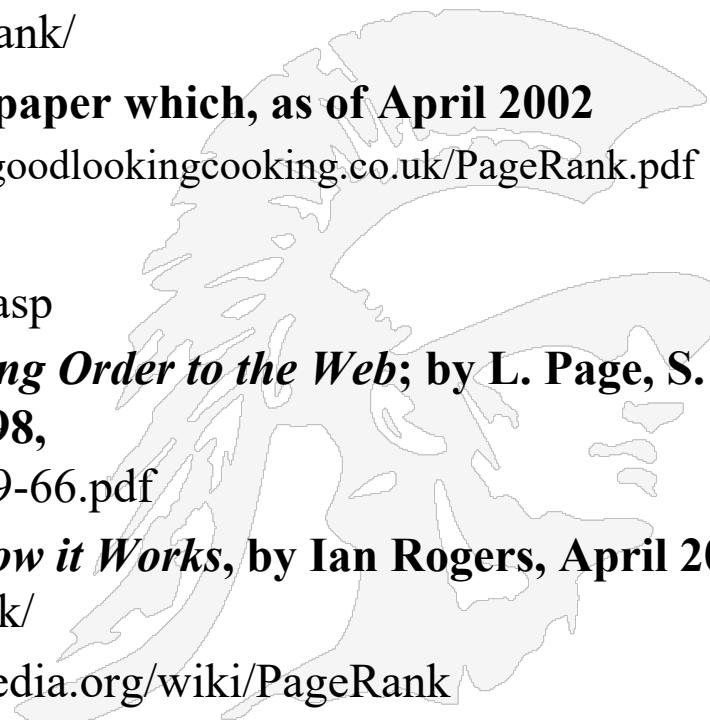
Random teleports

- The Google solution for spider traps
- At each time step, the random surfer has two options:
 - With probability β , follow a link at random
 - With probability $1-\beta$, jump to some page uniformly at random
 - Common values for β are in the range 0.8 to 0.9
- Surfer will teleport out of spider trap within a few time steps



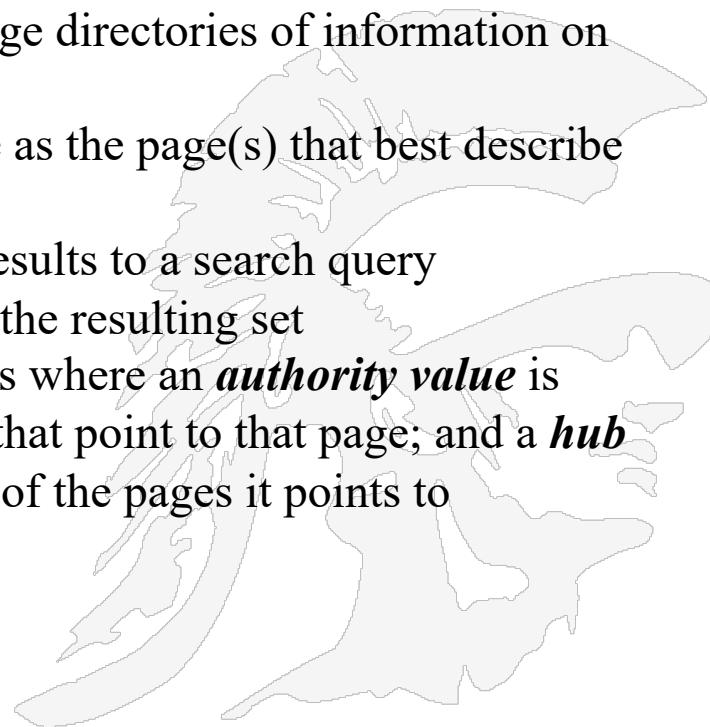
References

- The original PageRank paper by Google's founders Sergey Brin and Lawrence Page –
 - <http://www-db.stanford.edu/~backrub/google.html>
- PageRank explained correctly:
 - <http://www.iprcom.com/papers/pagerank/>
- Chris Ridings' "PageRank Explained" paper which, as of April 2002
 - http://web.archive.org/web/*/http://www.goodlookingcooking.co.uk/PageRank.pdf
- Tool to calculate PageRank
 - www.markhorrell.com/seo/pagerank.asp
- *The PageRank Citation Ranking: Bringing Order to the Web; by L. Page, S. Brin, R. Motwani, T. Winograd, January, 1998,*
<http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf>
- *The Google PageRank Algorithm and How it Works, by Ian Rogers, April 2002,*
<http://www.sirgroane.net/google-page-rank/>
- *PageRank on Wikipedia*, <http://en.wikipedia.org/wiki/PageRank>



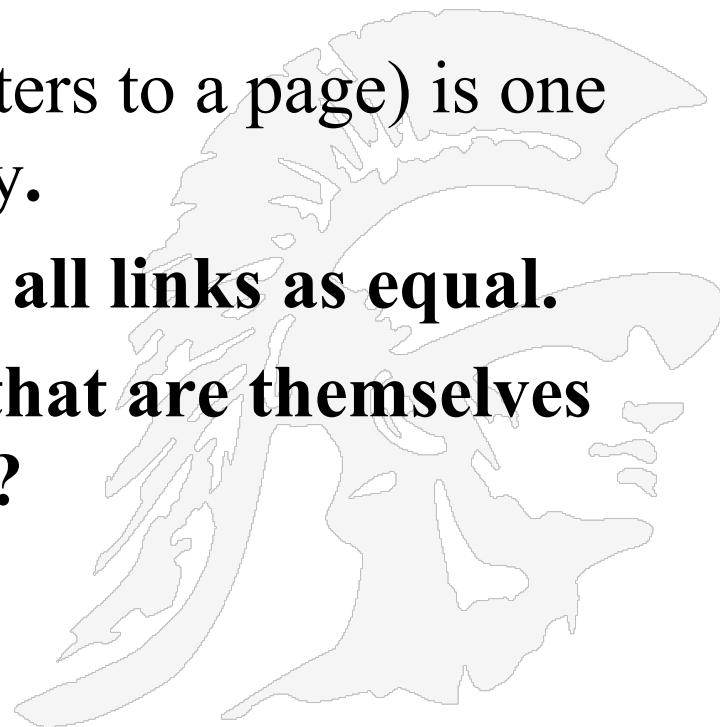
Another Link Analysis Algorithm – HITS by Kleinberg

- HITS stands for “Hyperlink-Induced Topic Search”
- HITS is a link analysis algorithm developed by Jon Kleinberg at Cornell
- HITS preceded PageRank
- The algorithm is based on the observations that
 - Some pages known as *hubs* serve as large directories of information on a given topic
 - Some pages known as *authorities* serve as the page(s) that best describe the information on a given topic
- The algorithm begins by retrieving a set of results to a search query
- The HITS algorithm is performed ONLY on the resulting set
- The algorithm goes through a set of iterations where an *authority value* is computed as the sum of the scaled hub values that point to that page; and a *hub value* is the sum of the scaled authority values of the pages it points to



Authorities

- *Authorities* are pages that are recognized as providing significant, trustworthy, and useful information on a topic.
- *In-degree* (number of pointers to a page) is one simple measure of authority.
- However in-degree treats all links as equal.
- Should links from pages that are themselves authoritative count more?

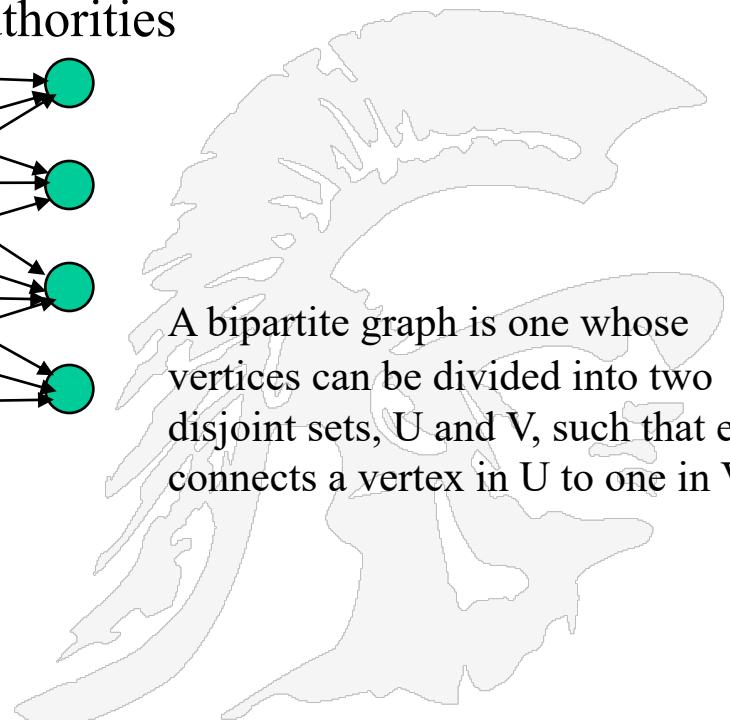
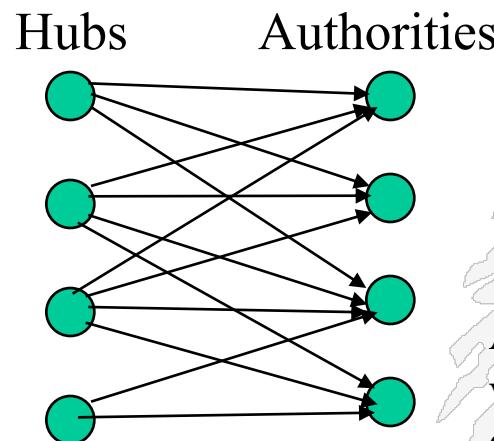


- ***Hubs*** are index pages that provide lots of useful links to relevant content pages (topic authorities).
- Ex: **pages are included in the course home page**



Hubs and Authorities

- Together they tend to form a bipartite graph:



A bipartite graph is one whose vertices can be divided into two disjoint sets, U and V, such that every edge connects a vertex in U to one in V

The General Idea

- **Authorities**- pages that are relevant and are linked to by many other pages
- **Hubs** – pages that have links to multiple relevant authorities

out-degree: # nodes

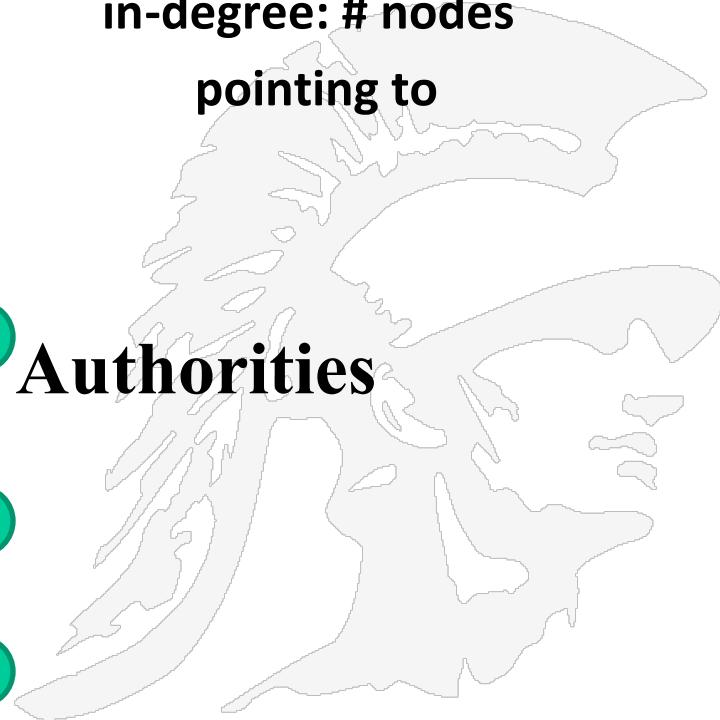
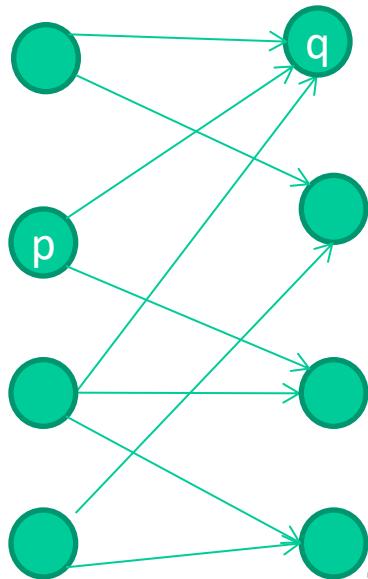
pointed by

in-degree: # nodes

pointing to

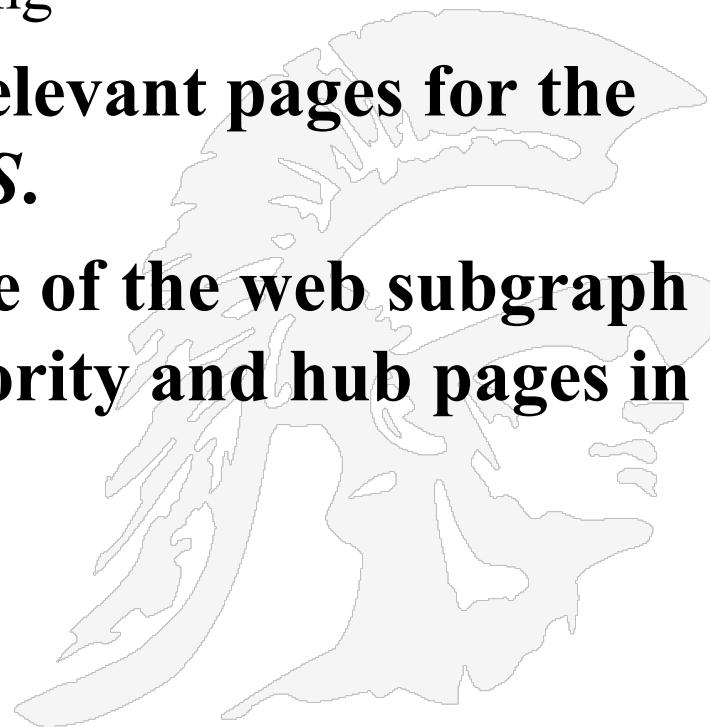
Hubs

Authorities



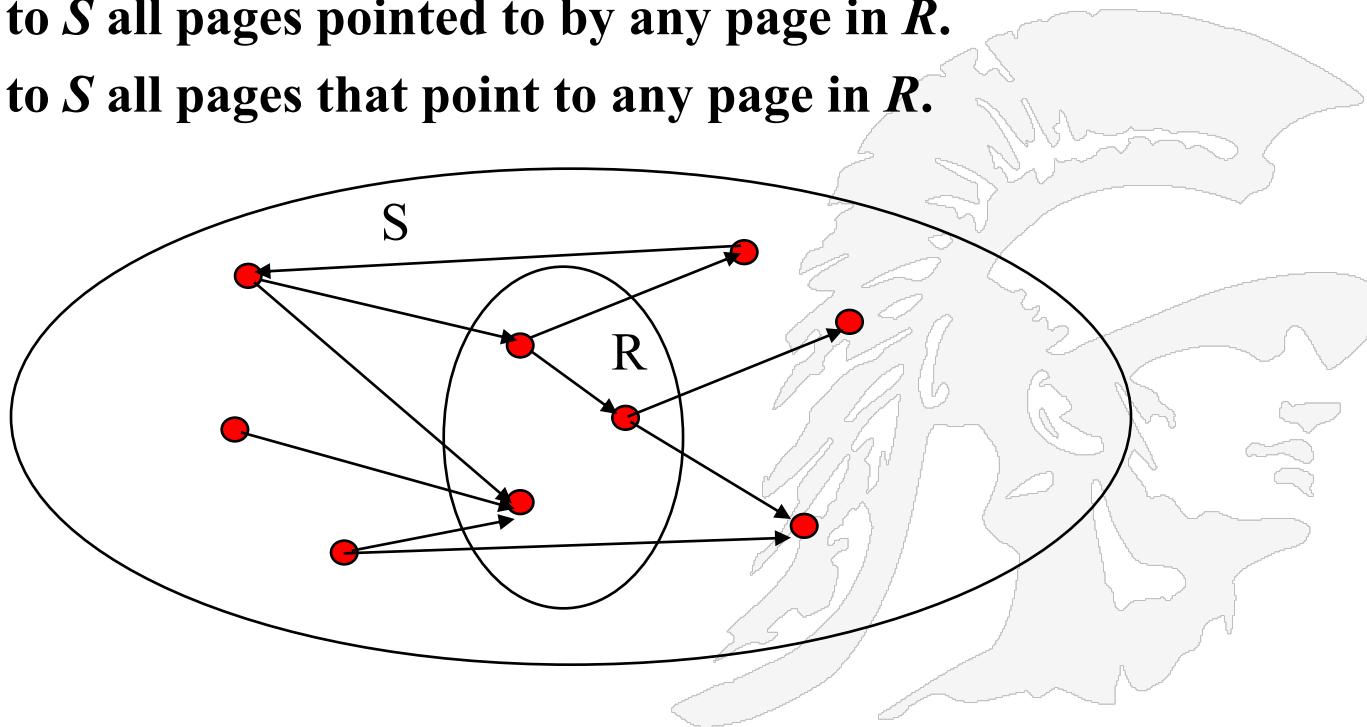
HITS Algorithm

- Computes hubs and authorities for a particular topic specified by a **normal query**.
 - Thus query dependent ranking
- First determine a set of relevant pages for the query called the *base* set S .
- Analyze the link structure of the web subgraph defined by S to find authority and hub pages in this set.



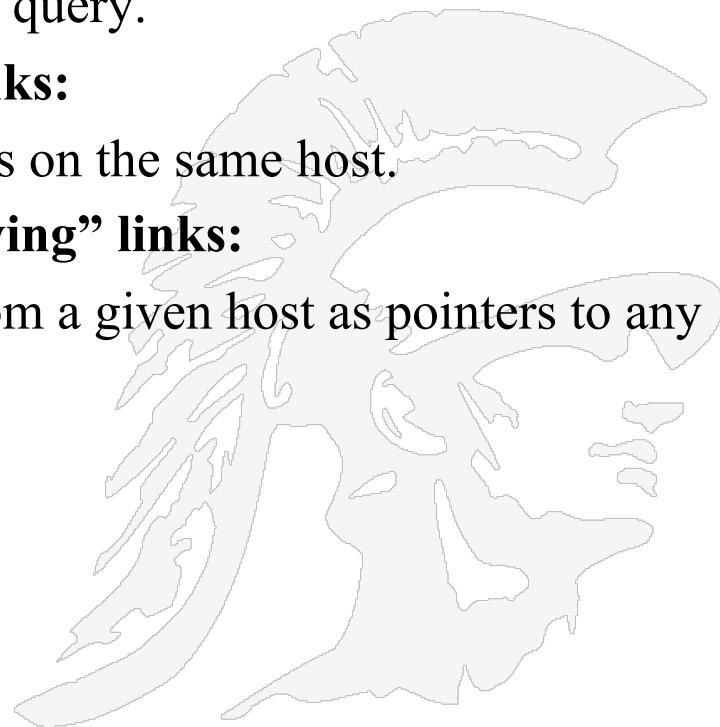
Constructing a Base Subgraph

- For a specific query Q , let the set of documents returned by a standard search engine be called the *root set* R .
- Initialize S to R .
- Add to S all pages pointed to by any page in R .
- Add to S all pages that point to any page in R .



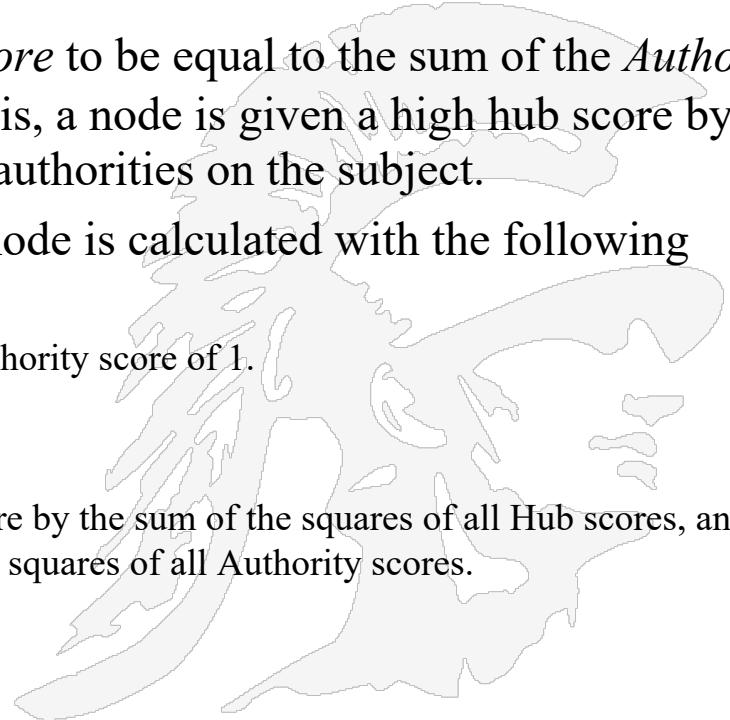
Base Limitations

- **To limit computational expense:**
 - Limit number of root pages to the top 200 pages retrieved for the query.
 - Limit number of “back-pointer” pages to a random set of at most 50 pages returned by a “reverse link” query.
- **To eliminate purely navigational links:**
 - Eliminate links between two pages on the same host.
- **To eliminate “non-authority-conveying” links:**
 - Allow only m ($m \approx 4-8$) pages from a given host as pointers to any individual page.



More Algorithm Detail

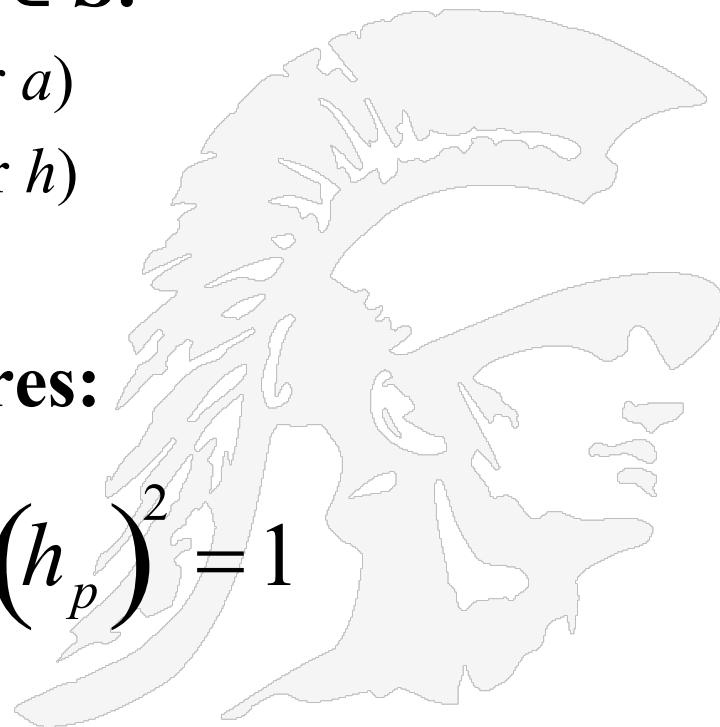
- The algorithm performs a series of iterations, each consisting of two basic steps:
- 1. **Authority Update:** Update each node's *Authority score* to be equal to the sum of the *Hub Scores* of each node that points to it. That is, a node is given a high authority score by being linked to by pages that are recognized as Hubs for information.
- 2. **Hub Update:** Update each node's *Hub Score* to be equal to the sum of the *Authority Scores* of each node that it points to. That is, a node is given a high hub score by linking to nodes that are considered to be authorities on the subject.
- The Hub score and Authority score for a node is calculated with the following algorithm:
 1. Start with each node having a hub score and authority score of 1.
 2. Run the Authority Update Rule
 3. Run the Hub Update Rule
 4. Normalize the values by dividing each Hub score by the sum of the squares of all Hub scores, and dividing each Authority score by the sum of the squares of all Authority scores.
 5. Repeat from the second step as necessary.



Iterative Algorithm

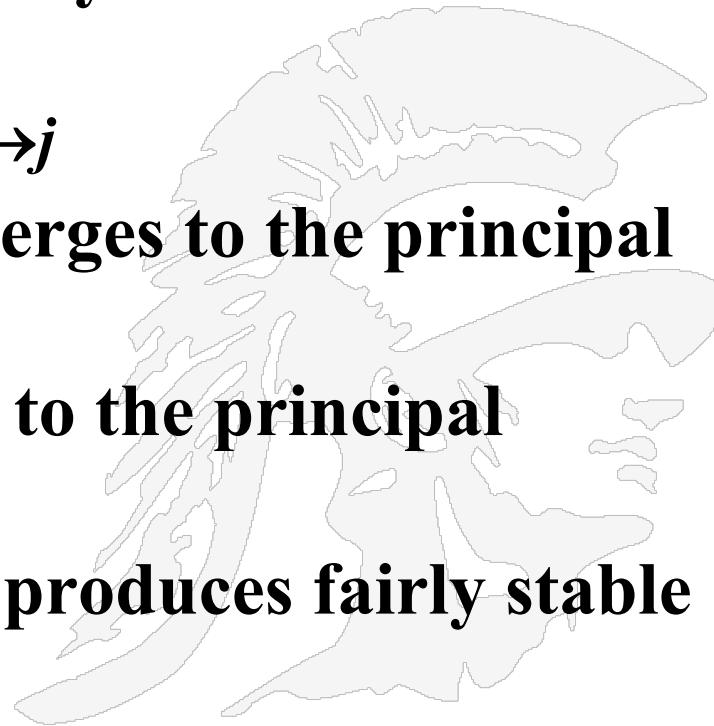
- Use an iterative algorithm to slowly converge on a mutually reinforcing set of hubs and authorities.
- Maintain for each page $p \in S$:
 - Authority score: a_p (vector a)
 - Hub score: h_p (vector h)
- Initialize all $a_p = h_p = 1$
- Maintain normalized scores:

$$\sum_{p \in S} (a_p)^2 = 1 \quad \sum_{p \in S} (h_p)^2 = 1$$



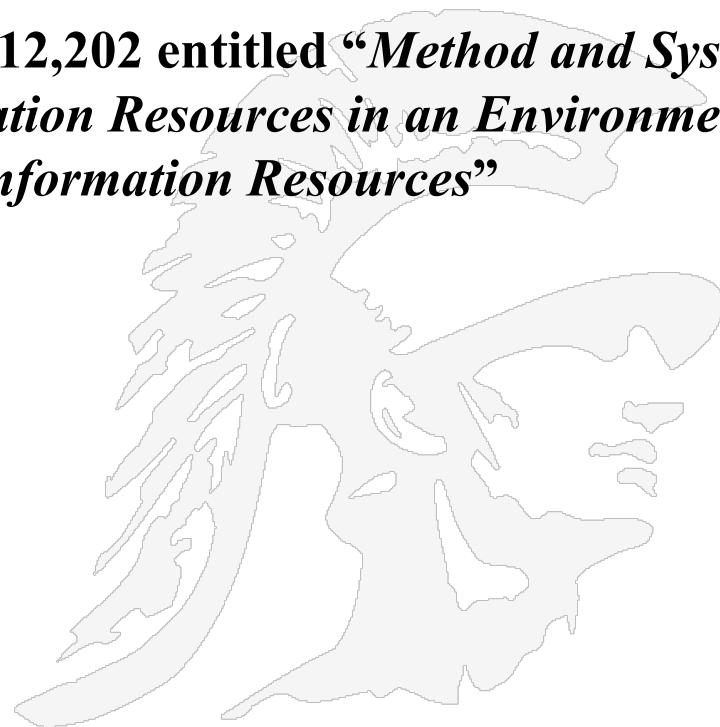
Convergence

- Algorithm converges to a *fix-point* if iterated indefinitely.
- Define A to be the adjacency matrix for the subgraph defined by S .
 - $A_{ij} = 1$ for $i \in S, j \in S$ iff $i \rightarrow j$
- Authority vector, a , converges to the principal eigenvector of $A^T A$
- Hub vector, h , converges to the principal eigenvector of $A A^T$
- In practice, 20 iterations produces fairly stable results.



Final Observations

- This algorithm is executed at query time, not at indexing time
 - So there is a penalty on performance
- It computes two scores for each page, a hub score and an authority score
- The method was patented as US 6,112,202 entitled *“Method and System for Identifying Authoritative Information Resources in an Environment with Content-based Links Between Information Resources”*



USC Viterbi

www.freepatentsonline.com/6112202.pdf

USC Viterbi School | Main Page - Comput. | Faculty Resources, C | Computer Science D | Google | Gmail - Inbox (3)

United States Patent [19]
Kleinberg


US006112202A

[11] Patent Number: **6,112,202**
[45] Date of Patent: **Aug. 29, 2000**

[54] **METHOD AND SYSTEM FOR IDENTIFYING AUTHORITATIVE INFORMATION RESOURCES IN AN ENVIRONMENT WITH CONTENT-BASED LINKS BETWEEN INFORMATION RESOURCES**

[75] Inventor: **Jon Michael Kleinberg**, Los Gatos, Calif.

[73] Assignee: **International Business Machines Corporation**, Armonk, N.Y.

[21] Appl. No.: **08/813,749**
[22] Filed: **Mar. 7, 1997**

[51] Int. Cl.⁷ **G06F 17/30**
[52] U.S. Cl. **707/5; 707/9; 707/101**
[58] Field of Search **707/1, 2, 4, 5, 707/10, 100, 101, 102, 501**

[56] **References Cited**

U.S. PATENT DOCUMENTS

5,257,185	10/1993	Farley et al.	707/100
5,446,891	8/1995	Kaplan et al.	707/2
5,778,363	7/1998	Light	707/5
5,826,031	10/1998	Nielsen	395/200.63
5,835,905	11/1998	Pirolli et al.	707/3

OTHER PUBLICATIONS

Savoy, J., Searching Information in Hypertext Systems using Multiple Sources of Evidence, International Journal of Man-Machine Studies, 1993, pp. 1017-1030.
 Steiger, H., Making Use of Hypertext Links when Retrieving Information, ACM, pp. 102-111, 1992.
 R.W. Schwanke et al., Cross References are Features, Sec. 10.1, Book/Machine Learning: From Theory to Applications, Cooperative Research at Siemens and MIT, Appeared in Proceedings of the 2nd International Workshop on Software Configuration Mng., Princeton, NJ, Oct. 1989, ACM SIGSoft, IEEE CS, and GI pp. 107-123.
 H.C. Arents et al., "Concept-Based Retrieval of Hypermedia Information: From Term Indexing to Semantic Hyperindexing," Information processing & Management vol. 29, No. 3, pp. 373-386, 1993.

R. Rada et al., "Retrieval Hierarchies in Hypertext," Information Processing & Mng., vol. 29, No. 3, (Printed in Great Britain) pp. 359-371, 1993.
 W.M. Shaw, Jr., "Subject and Citation Indexing, Part I: The Clustering Structure of Composite Representations in the Cystic Fibrosis Document Collection," JASIS-Journal of the American Society for Information Science, vol. 42, No. 9, Oct. 1991, pp. 669-675.
 W.M. Shaw, Jr., "Subject Indexing & Citation Indexing-Part II: A Evaluation and Comparison Information Processing & Management", vol. 26, No. 6, (printed in Great Britain) pp. 705-718, 1990.
 T.R. Kochtanek, "Brief Communication, Document Clustering, Using Macro Retrieval Techniques," Journal of the American Society for Information Science, vol. 34, No. 5, pp. 356-359, Sep. 1993.
 F. Narin et al., Chapter 2, "Bibliometrics," Pub. Annual Review of Information Sciences and Technology, pp. 35-58, 1977.

(List continued on next page.)

Primary Examiner—Thomas G. Black
 Assistant Examiner—John Loomis
 Attorney, Agent, or Firm—Khanh Q. Tran

[57] **ABSTRACT**

A system and method are provided for searching for desired items from a network of information resources. In particular, the system and method have advantageous applicability to searching for World Wide Web pages having desired content. An initial set of pages are selected, preferably by running a conventional keyword-based query, and then further selecting pages pointing to, or pointed to from, the pages found by the keyword-based query. Alternatively, the invention may be applied to a single page, where the initial set includes pages pointed to by the single page and pages which point to the single page. Then, iteratively, authoritativeness values are computed for the pages of the initial set, based on the number of links to and from the pages. One or more communities, or "neighborhoods", of related pages are defined based on the authoritativeness values thus produced. Such communities of pages are likely to be of particular interest and value to the user who is interested in the keyword-based query or the single page.

57 Claims, 5 Drawing Sheets

Choose print parameters:
 m = number of pages
 n = number of directions
 k = normal size

Copyright LIMS Totorowitz, 2011-2022

Patent invented by Jon Kleinberg
 Assigned to IBM; filed March 1997

