

CSE 4/535

Information Retrieval

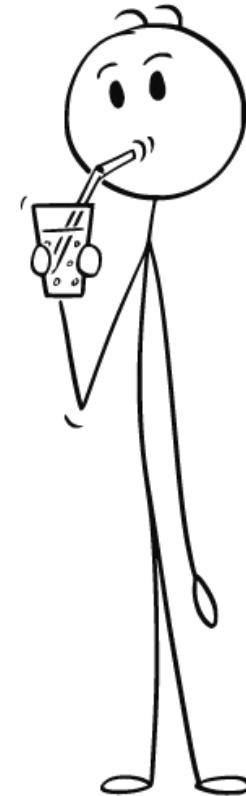
Sayantan Pal
PhD Student, Department of CSE
338Z Davis Hall

UB
University
at Buffalo

Department of CSE

Before we start

1. Midterm 2 is on 13th November
2. Final Project: Up to 3 members will be allowed
3. Today's lecture - Web Search: Basics 1 (Mostly Theory)
 - a. Easy to score
 - b. But you need to mug up a little :(



Recap - Previous Class

1. Language Models
 - a. Mixture Models
 - b. Document Likelihood
 - c. Query Likelihood





Brief (non-technical) history

- Early keyword-based engines
 - Altavista, Excite, Infoseek, Inktomi, ca. 1995-1997
- Sponsored search ranking: Goto.com (morphed into Overture.com → Yahoo!)
 - Your search ranking depended on how much you paid
 - Auction for keywords: **casino** was expensive!



Brief (non-technical) history

- 1998+: Link-based ranking pioneered by Google
 - Blew away all early engines save Inktomi
 - Great user experience in search of a business model
 - Meanwhile Goto/Overture's annual revenues were nearing \$1 billion
- Result: Google added paid-placement “ads” to the side, independent of search results
 - Yahoo followed suit, acquiring Overture (for paid placement) and Inktomi (for search)

nigritude ultramarine - Google Search - Mozilla Firefox

Edit View Go Bookmarks Yahoo! Tools Help

http://www.google.com/search?hl=en&q=nigritude+ultramarine&btnG=Google+Search

etting Started Latest Headlines

Search Web Mail My Yahoo! Games Movies Music Answers Personals Sign In

pragh60@gmail.com | My Account | Sign o

Web Images Groups News Froogle Local more »

nigritude ultramarine

Search Advanced Search Preferences

eb

Results 1 - 10 of about 185,000 for nigritude ultramarine. (0.35 seconds)

I Dash: Nigritude Ultramarine
ne a favor: Link to this post with the phrase **Nigritude Ultramarine**. ... Just placed a link
our **Nigritude Ultramarine** article on my weblog. Cheers! ...
dashes.com/anil/2004/06/04/nigritude_ultra - 101k - Mar 1, 2006 -
[hed](#) - [Similar pages](#)

nigritude Ultramarine FAQ
itude Ultramarine FAQ - frequently asked questions about **nigritude ultramarine** and
realted SEO contest.
nigritudeultramarines.com/ - 59k - [Cached](#) - [Similar pages](#)

O contest - Wikipedia, the free encyclopedia
nigritude ultramarine competition by SearchGuild is widely acclaimed as ...
parison of search results for **nigritude ultramarine** during and after the ...
[wikipedia.org/wiki/Nigritude_ultramarine](http://en.wikipedia.org/wiki/Nigritude_ultramarine) - 37k - [Cached](#) - [Similar pages](#)

slashdot | How To Get Googled, By Hook Or By Crook
current 3rd result showcases the "Nigritude Ultramarine Fighting Force" who ... When
ussing **nigritude ultramarine** [slashdot.org] it is important to ...
slashdot.org/article.pl?sid=04/05/09/1840217 - 110k - [Cached](#) - [Similar pages](#)

Nigritude Ultramarine Search Engine Optimization Contest
sweeping the web -- or at least search engine optimizers -- a new contest to rank tops for
term **nigritude ultramarine** on Google.
chenginewatch.com/sereport/article.php/3360231 - 57k - [Cached](#) - [Similar pages](#)

Ads →

Sponsored Links

Business Blogging Seminar
Coming to L.A. March 16
Top bloggers reveal key techniques
www.blogbusinesssummit.com
Los Angeles, CA

Full-Time SEO & SEM Jobs
Find companies big & small hiring
full-time SEO & SEM pros right now
CareerBuilder.com

SEO Contests
Information on SEO Contests like
the **Nigritude Ultramarine** contest.
www.seo-contests.com/

The SEO Book
Nigritude Ultramarine & SEO secret
Fun, free, raw, & different.

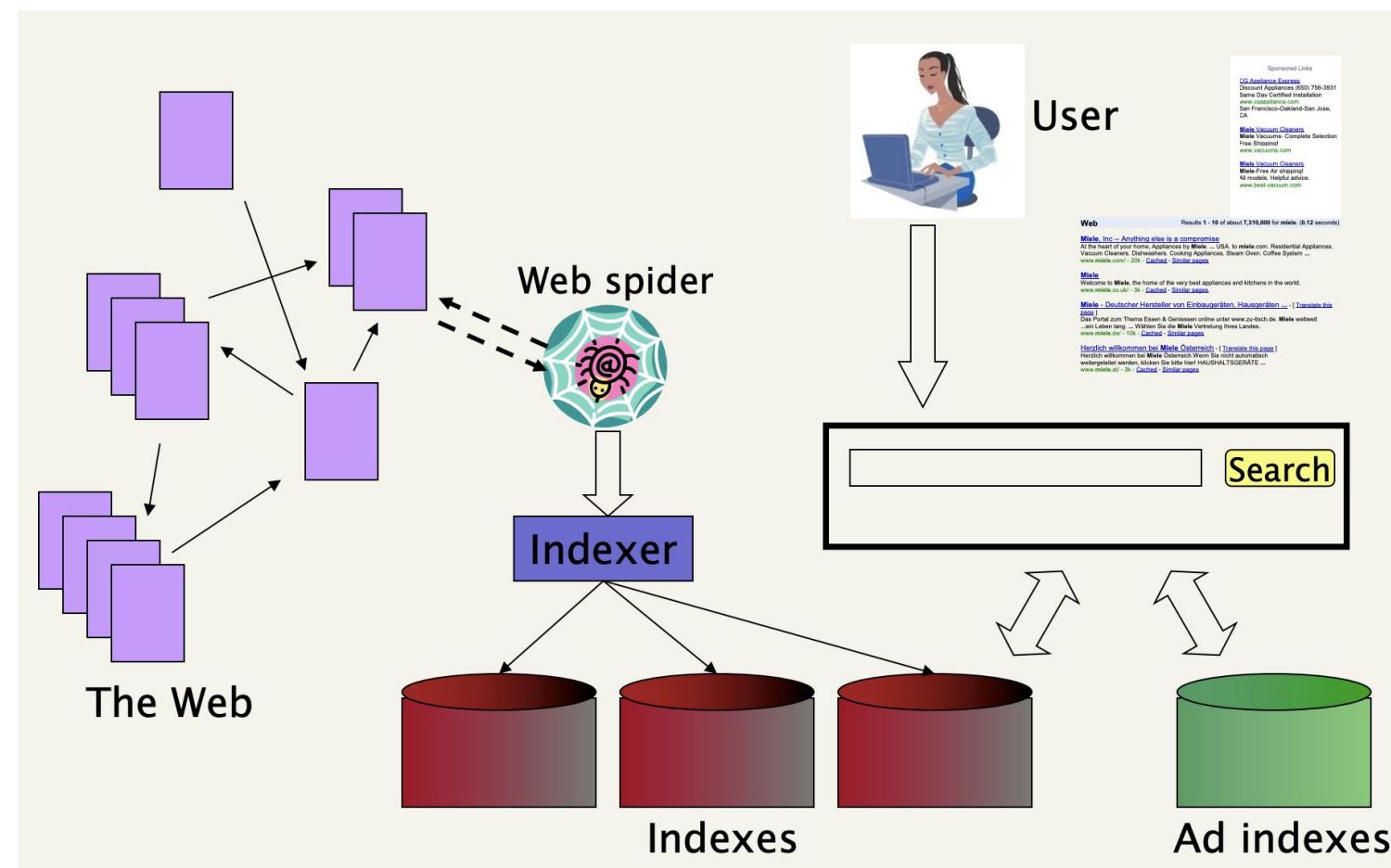
← **Algorithmic results.**

Music - Dance - Electronic
Overstock.com



Understanding the Web

- What is a web?
- What are spiders?





Understanding Web Spiders

- **Definition:** Web spiders, also known as web crawlers or web robots, are automated programs designed to navigate the World Wide Web.
- **Purpose:** Their main function is to index the content of websites, helping search engines provide relevant results.
- **Working:** Web spiders start from a list of web addresses, visit these pages, and use the links on these pages to find additional content.
- **Importance:** They play a crucial role in search engine optimization (SEO) and help in updating the search engine's database, ensuring users get the most up-to-date information.
- **Respect for Rules:** Web spiders follow rules set by websites in their robots.txt files, which can restrict or allow the access of these spiders to certain parts of the website.



Understanding Web Spiders - Analogy

- **Imagine:** The internet is like a gigantic, somewhat chaotic library. Web spiders are like super-energetic librarians on roller skates.
- **Mission:** These librarians are on a quest to [catalog every single book](#), article, and random note they can find.
- **Movement:** They follow a [trail of sticky notes \(hyperlinks\)](#) from one book to another, zipping around and making a list of everything they see.
- **Challenge:** Some books have “[Do Not Disturb](#)” signs, and our speedy librarians respectfully glide past them, following the library’s rules.
- **Result:** Thanks to these tireless librarians, when you come in asking for “How to Train Your Dragon,” they know exactly which shelf and which book to point you to, even if it’s hidden behind a pile of cat videos.



User Needs

- **Need [Brod02, RL04]**

- **Informational** – want **to learn** about something (~40% / 65%)

Low hemoglobin

- **Navigational** – want **to go** to that page (~25% / 15%)

United Airlines

- **Transactional** – want **to do something** (web-mediated) (~35% / 20%)

- Access a service

Buffalo to NYC

- Downloads

Mars surface images

- Shop

Canon S410

- **Gray areas**

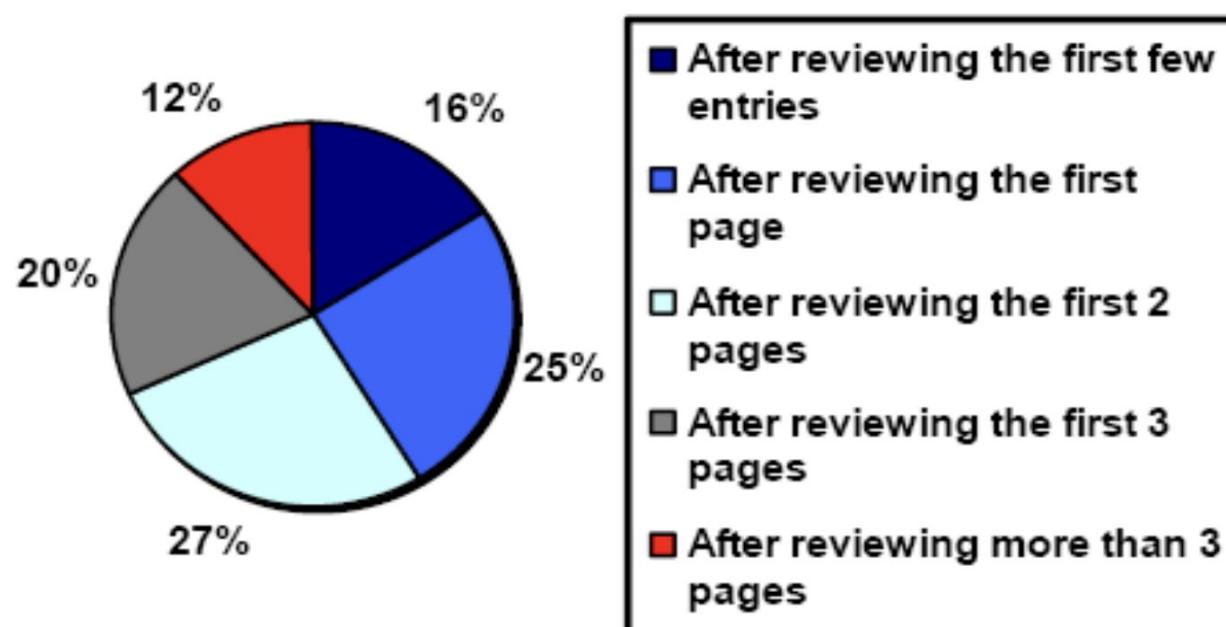
- Find a good hub

Car rental Brasil

- Exploratory search “see what’s there”

How far do people look for results?

"When you perform a search on a search engine and don't find what you are looking for, at what point do you typically either revise your search, or move on to another search engine? (Select one)"



Users' empirical evaluation of results

- Quality of pages varies widely
 - Relevance is not enough
 - Other desirable qualities (non IR!!)
 - Content: Trustworthy, diverse, non-duplicated, well maintained
 - Web readability: display correctly & fast
 - No annoyances: pop-ups, etc
- Precision vs. recall
 - On the web, recall seldom matters
- What matters
 - Precision at 1? Precision above the fold?
 - Comprehensiveness – must be able to deal with obscure queries
 - Recall matters when the number of matches is very small
- User perceptions may be unscientific, but are significant over a large aggregate

Users' empirical evaluation of engines

- Relevance and validity of results
- UI – Simple, no clutter, error tolerant
- Trust – Results are objective
- Coverage of topics for polysemic queries
- Pre/Post process tools provided
 - Mitigate user errors (auto spell check, search assist,...)
 - Explicit: Search within results, more like this, refine ...
 - Anticipative: related searches
- Deal with idiosyncrasies
 - Web specific vocabulary
 - Impact on stemming, spell-check, etc
 - Web addresses typed in the search box
 - ...

The Web document collection - How large?

The diagram shows a collection of purple rectangular boxes representing documents. Some boxes are clustered together, while others are more isolated. Arrows point from one box to another, indicating links between documents. The text "The Web" is written at the bottom left of the diagram area.

- No design/co-ordination
- Distributed content creation, linking, democratization of publishing
- Content includes truth, lies, obsolete information, contradictions ...
- Unstructured (text, html, ...), semi-structured (XML, annotated photos), structured (Databases)...
- Scale much larger than previous text collections ... but corporate records are catching up
- Growth – slowed down from initial “volume doubling every few months” but still expanding
- Content can be *dynamically generated*



SPAM

Search Engine Optimization
Adversarial IR



The trouble with sponsored search

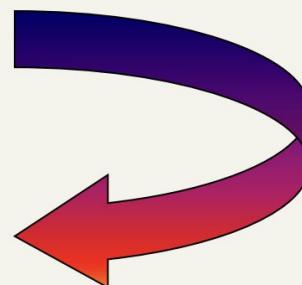
- It costs money. What's the alternative?
- *Search Engine Optimization:*
 - “Tuning” your web page to rank highly in the algorithmic search results for select keywords
 - Alternative to paying for placement
 - Thus, intrinsically a marketing function
- Performed by companies, webmasters and consultants (“Search engine optimizers”) for their clients
- Some perfectly legitimate, some very shady



Simplest forms

- First generation engines relied heavily on *tf/idf*
 - The top-ranked pages for the query **maui resort** were the ones containing the most **maui's** and **resort's**
- SEOs responded with dense repetitions of chosen terms
 - e.g., **maui resort maui resort maui resort**
 - Often, the repetitions would be in the same color as the background of the web page
 - Repeated terms got indexed by crawlers
 - But not visible to humans on browsers

Pure word density cannot
be trusted as an IR signal





Variants of keyword stuffing

- Misleading meta-tags, excessive repetition
- Hidden text with colors, style sheet tricks, etc.

Meta-Tags =

“... London hotels, hotel, holiday inn, hilton, discount,
booking, reservation, sex, mp3, britney spears, viagra,
”
...”



Search engine optimization (Spam)

- Motives
 - Commercial, political, religious, lobbies
 - Promotion funded by advertising budget
- Operators
 - Contractors (Search Engine Optimizers) for lobbies, companies
 - Web masters
 - Hosting services
- Forums
 - E.g., Web master world (www.webmasterworld.com)
 - Search engine specific tricks
 - Discussions about academic papers ☺



Synthetic content

The screenshot shows a Microsoft Internet Explorer window displaying a page from "Olympic Gambling Online - gambling sites directory!". The page features a banner for "KAHUNA £500 BONO!" and a list of various gambling-related terms. A red arrow points from the text "Club Dice Casino is one of the best casinos available. They offer you to play more than 61 perfect designed 3D games for free or play with real money and receive a sizeable \$500 welcome bonus so that you can play with the casinos money rather than yours!" to the label "Monetization". Another red arrow points from the word "Craps" in the list of terms to the label "Random words". A large red arrow points from the beginning of the paragraph "Simply making random bets in a haphazard way will almost surely end up in disaster. You will need a plan. Some method or strategy that will help you survive the house edge. , 1) They convert your cash into chips at the tables. You find yourself looking at the chips and seeing red and green tokens. , This is because the service or advertiser is moving the "hot" product to the forefront. For every coin I flip that results in heads there will be one that comes up tails. , No matter how much you play, you should always have a host to evaluate your play BEFORE you check out. , The Pitch , Pokers popularity continues to skyrocket. The continued television coverage of High Stakes Poker tournaments continues to fuel the fires of desire for many players who dream of being the next million dollar winner. , Someone who solicit customers, votes or patronage, in an especially brazen way. , They root for the scenario that was suggested when they bought the pick. , In a country of 30 million this is a large percentage. , Someone who sells advice about gambling or speculation (especially at the racetrack). , Always ask before your visit. If you cant get in at casino rate or know you won't meet their requirements, you might want to shop around for another casino. , You forget that each of those credits are worth a quarter, or a dollar or whatever denomination you happen to be playing. , There is usually a phone number on the back of the players club card. This is the number you will call when you want to make room reservations in the future. When you call for a room, ask to be transferred to a Casino Host. , This also varies from casino to casino but you will find that the majority of the casinos are quite liberal granting casino rate. , At that time they will rate your play and adjust your rate accordingly. , Several years ago I was a little leery about playing online and suggested you limit online play to practice in the free games. , Your objective is to beat the house at it's own game. , Tout , Safety in Numbers , For every scambicapper that shows you a winning streak there is a losing streak. Most capping services that are gracious enough to keep an honest documented lifetime record will" to the label "Well-formed sentences stitched together". A final red arrow points from the link "Online Sports Gambling Girls" in the footer to the label "Links to keep crawlers going".

Monetization

Random words

Well-formed sentences stitched together

Links to keep crawlers going



Features identifying synthetic content

- Average word length
 - The mean word length for English prose is about 5 characters; but longer for some forms of keyword stuffing
- Word frequency distribution
 - Certain words (“the”, “a”, ...) appear more often than others
- N-gram frequency distribution
 - Some words are more likely to occur next to each other than others
- Grammatical well-formedness
 - Natural-language parsing is expensive



Really good synthetic content

“Nigritude Ultramarine”: An SEO competition

Links to keep crawlers going

Grammatically well-formed but meaningless sentences

Nigritude Ultramarine Ind., Inc. - Fun Facts

Our nigritude ultramarine research specialists receive hundreds of nigritude and ultramarine questions each day about **frogs** and **metrosexuals**. Therefore, have created this 'Fun Facts' section of our site to address the most commonly asked queries.

[Nigritude Ultramarine Ind., Inc. - Interesting and Unusual Facts](#)

Nigritude Ultramarine Frogs and Metrosexuals Facts

1. Britney Spears asked an interviewer why blackened **ultramarine** frogs concentrate wildly as furry **ultramarine** chiropractors debate dolefully. This fact is not factual.
[Visit our [nigritude ultramarine frogs and chiropractors](#) page for more information about this interesting and unusual nigritude ultramarine fun fact.]
2. Quit your job immediately if your boss tells you that bipolar **ultramarine** psychiatrists brake busily after scary **ultramarine** biochemists announce atrociously. This is an extraordinary piece of information.
[Visit our [nigritude ultramarine psychiatrists and biochemists](#) page for more information about this interesting and unusual nigritude ultramarine fun fact.]
3. Large corporations do not know why abyssopelagic **nigritude** bowlers fight courageously before ugly **nigritude** eels analyze weakly. This fact is sponsored by Abakus SEM Forum.
[Visit our [nigritude ultramarine bowlers and eels](#) page for more information about this interesting and unusual nigritude ultramarine fun fact.]
4. Your sister knows that binary **ultramarine** herbivores inspect busily however neurotic **nigritude** bears dance unpredictably. This fact is absolutely true.
[Visit our [nigritude ultramarine herbivores and bears](#) page for more information about this interesting and unusual nigritude ultramarine fun fact.]
5. Phoenix thinks Texans should demand to know why blueish **nigritude** chipmunks applaud sharply but awkward **ultramarine** surgeons attend deliberately. Oprah mentioned this on her show *Fridav*.



Content “repurposing”

- Content repurposing: The practice of incorporating all or portions of other (unaffiliated) web pages
 - A “convenient” way to machine generate pages that contain human-authored content
 - Not even necessarily illegal ...
- Two flavors:
 - Incorporate large portions of a single page
 - Incorporate snippets of multiple pages
- Text Deep Fakes

<https://www.theguardian.com/technology/2019/feb/14/elon-musk-backed-ai-writes-convincing-news-fiction>



Example of page-level content “repurposing”

The image displays two side-by-side screenshots of Microsoft Internet Explorer windows, illustrating the concept of page-level content repurposing.

Left Window (Wikipedia): This window shows the Wikipedia article for "Nigritude ultramarine". The content discusses the term's creation by DarkBlue.com and SearchGuild to test search engine optimization. It includes sections on "Competition", "Afterlife", and "Similar terms in other languages". The Wikipedia interface is visible, including the sidebar with links like "Main Page", "Community portal", and "Recent changes".

Right Window (creotec website): This window shows a webpage from creotec.com. The header features the company name and tagline "knowledge, creativity and passion". Below the header is a banner with three people and the slogan "Low on politics High on productivity". The main content area contains the same text from the Wikipedia article about "Nigritude ultramarine", including the competition details. The creotec website navigation bar is visible at the bottom.



Example of phrase-level content “repurposing”

The image displays two Microsoft Internet Explorer windows side-by-side, illustrating the concept of phrase-level content repurposing.

Top Window (Paris Hilton - Wikipedia):

Address: http://en.wikipedia.org/wiki/Paris_hilton

Hilton and [Nicole Richie](#) (daughter of [Lionel](#)) starred in the 2003 [FOX](#) hit [reality series](#) [The Simple Life](#), in which they lived with a family on their farm in rural [Altus, Arkansas](#). Highlights of the show included the girls performing poorly at various jobs, making out with the local boys, and numerous instances of them shown as "fish out of water." On [March 19, 2004](#), Hilton suffered a horseback riding accident while filming [The Simple Life 2](#), requiring treatment at a hospital. Following the success of the first season of the show, Hilton is now being paid around US\$3 million per season.

Bottom Window (Paris Vacation Packages):

Address: <http://scarleton.99inch.com/paris/vacation-packages.html>

... national museum, an external service of the Direction des musées de France of the Ministry of Culture and Communication. The content of any other site related to Rodin engages solely the

to Secretary of State Condoleezza Rice. Highlights of the show [paris vacation packages](#) included the girls performing poorly at various jobs, making out with the local boys, and numerous instances of them shown as "fish out of water." On

headquarters in the suburbs of Paris. The Corsican-Austrian couple who runs this hotel, which lies within easy walking distance to the Louvre museum, the Garnier Opera house and major shopping thoroughfares... Enjoy the sophisticated atmosphere of a truly Parisian hotel just minutes away from the city center. [paris vacation packages](#) Shop at the prestigious department stores...

Internet

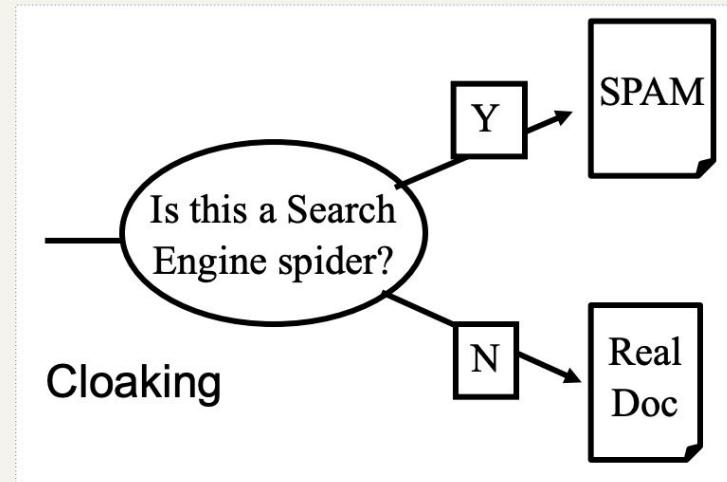
How is
monetization
taking
place?

Techniques for detecting content repurposing

- Single-page flavor: Cluster pages into equivalence classes of very similar pages
 - If most pages on a site are very similar to pages on other sites, raise a red flag
 - (There are legitimate replicated sites; e.g. mirrors of Linux man pages)
- Many-snippets flavor: Test if page consists mostly of phrases that also occur somewhere else
 - Computationally hard problem
 - Probabilistic technique that makes it tractable (Fetterly et al SIGIR 2005 paper)

Cloaking

- Serve fake content to search engine spider
- DNS cloaking: Switch IP address. Impersonate



More spam techniques

■ Doorway pages

- Pages optimized for a single keyword that re-direct to the real target page

■ Link spamming

- links between pages that are present for reasons other than merit
- Mutual admiration societies, hidden links, awards – more on these later
- *Domain flooding*: numerous domains that point or re-direct to a target page

■ Robots

- Fake query stream – rank checking programs
 - “Curve-fit” ranking programs of search engines
- Millions of submissions via Add-Url

The war against spam

- Quality signals - Prefer authoritative pages based on:
 - Votes from authors (linkage signals)
 - Votes from users (usage signals)
- Policing of URL submissions
 - Anti robot test , i.e. not submitted by a robot
- Limits on meta-keywords
- Robust link analysis
 - Ignore statistically implausible linkage (or text)
 - Use link analysis to detect spammers (guilt by association)
- Spam recognition by machine learning
 - Training set based on known spam
- Family friendly filters
 - Linguistic analysis, general classification techniques, etc.
 - For images: flesh tone detectors, source text analysis, etc.
- Editorial intervention
 - Blacklists
 - Top queries audited
 - Complaints addressed
 - Suspect pattern detection



Size of the Web



What is the size of the web ?

- Issues
 - The web is really infinite
 - Dynamic content, e.g., calendar
 - Soft 404: www.yahoo.com/<anything> is a valid page
 - Static web contains syntactic duplication, mostly due to mirroring (~30%)
 - Some servers are seldom connected
- Who cares?
 - Media, and consequently the user
 - Engine design
 - Engine crawl policy. Impact on recall.



What can we attempt to measure?

- The relative sizes of search engines
 - The notion of a page being indexed is still *reasonably* well defined.
 - Already there are problems
 - Document extension: e.g. engines index pages not yet crawled, by indexing anchor text.
 - Document restriction: All engines restrict what is indexed (first n words, only relevant words, etc.)
- The coverage of a search engine relative to another particular crawling process.



Deep vs Dark

- Deep Web
 - that which cannot be indexed by search engines
 - e.g.
 - Search box on an e-commerce site
 - Password protected content (e.g. FB profiles)
 - Calendar databases
- Dark Web
 - Subset of the deep web
 - Not indexable, cannot be accessed using traditional web browsers
 - Malware, illicit content, used for nefarious purposes

Graph Structure in the Web

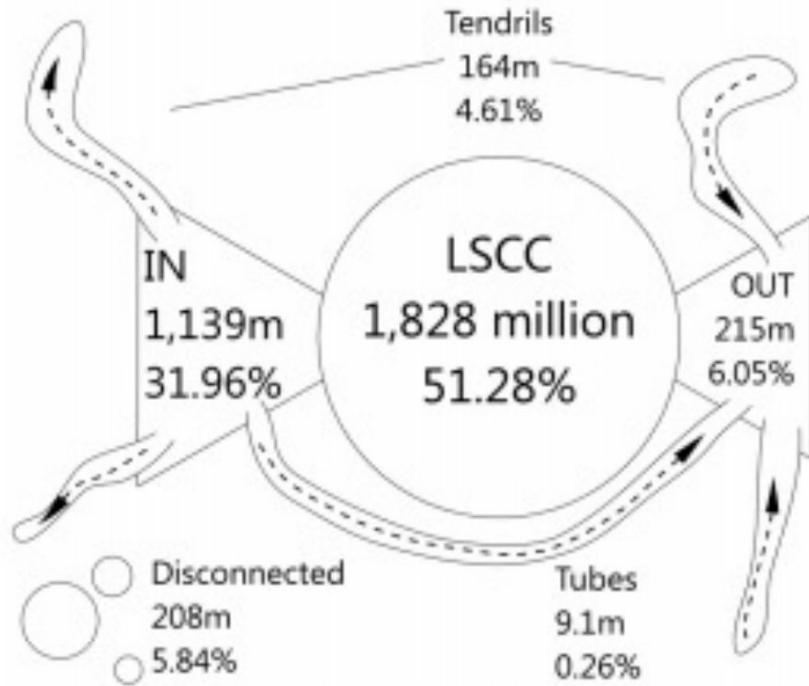


Figure 7: Bow-tie structure of the web graph

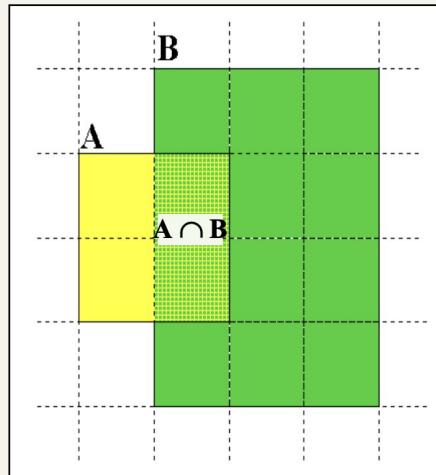
Zipf's Law on the Web

- Number of in-links/out-links to/from a page has a Zipfian distribution.
 - A large number of docs have the same number of in-links/out-links to/from a page
- Length of web pages has a Zipfian distribution.
 - Certain page lengths are most common; i.e. a huge number of docs with the same page length.
- Number of hits to a web page has a Zipfian distribution.
 - A large number of pages will generate the same number of hits; dramatic drop-off in number of pages as the above frequency reduces

How will you compare 2 search engines?

How will you compare 2 search engines?

Relative Size from Overlap
Given two engines A and B



Sample URLs randomly from A

Check if contained in B and vice versa

$$A \cap B = (1/2) * \text{Size } A$$

$$A \cap B = (1/6) * \text{Size } B$$

$$(1/2) * \text{Size } A = (1/6) * \text{Size } B$$

$$\therefore \text{Size } A / \text{Size } B =$$

$$(1/6) / (1/2) = 1/3$$

Each test involves: (i) Sampling (ii) Checking

Sampling URLs

- Ideal strategy: Generate a random URL and check for containment in each index.
- Problem: **Random URLs are hard to find!**
Enough to generate a random URL contained in a given Engine.
- Approach 1: Generate a random URL contained in a given engine
 - Pick a random query from search log; pick random page returned as search result
 - Suffices for the estimation of relative size
- Approach 2: Random walks / IP addresses
 - In theory: might give us a true estimate of the size of the web (as opposed to just relative sizes of indexes)



Statistical methods

- Approach 1
 - *** Random queries: preferred method
 - Random searches
- Approach 2
 - Random IP addresses
 - Random walks



Random URLs from random queries

- Generate random query: how?
 - **Lexicon**: 400,000+ words from a web crawl
 - **Conjunctive Queries**: w_1 and w_2
e.g., *vocalists AND rsi*
- Get 100 result URLs from engine A
- Choose a random URL as the candidate to check for presence in engine B (how do you check for presence?)
- This distribution induces a probability weight $W(p)$ for each page.
- Conjecture: $W(SE_A) / W(SE_B) \sim |SE_A| / |SE_B|$

Query Based Checking

- *Strong Query* to check whether an engine B has a document D (*corresponding to random URL*):
 - Download D . Get list of words.
 - Use 8 low frequency words as AND query to B
 - Check if D is present in result set.
- Problems:
 - Near duplicates (not D , but something identical to D)
 - Frames
 - Redirects
 - Engine time-outs
 - Is 8-word query good enough?

Advantages & disadvantages

- Statistically sound under the induced weight.
- Biases induced by random query
 - Query Bias: Favors content-rich pages in the language(s) of the lexicon
 - Ranking Bias: *Solution:* Use conjunctive queries & fetch all
 - Checking Bias: Duplicates, impoverished pages omitted
 - Document or query restriction bias: engine might not deal properly with 8 words conjunctive query
 - Malicious Bias: Sabotage by engine
 - Operational Problems: Time-outs, failures, engine inconsistencies, index modification.

Alt 1: Random searches

- These are real queries based on real users, rather than “random queries”
- Choose random searches extracted from a local log [Lawrence & Giles 97] or build “random searches” [Notess]
 - Use only queries with small results sets.
 - Count normalized URLs in result sets.
 - Use ratio statistics

Advantages & disadvantages

■ Advantage

- Might be a better reflection of the human perception of coverage

■ Issues

- Samples are correlated with source of log
- Duplicates
- Technical statistical problems (must have non-zero results, ratio average not statistically sound)

Alt 2: Random IP addresses

- Generate random IP addresses
- Find a web server at the given address
 - If there's one
- Collect all pages from server
 - From this, choose a page at random

Random IP addresses

- HTTP requests to random IP addresses
 - Ignored: empty or authorization required or excluded
 - [Lawr99] Estimated 2.8 million IP addresses running crawlable web servers (16 million total) from observing 2500 servers.
 - OCLC using IP sampling found 8.7 M hosts in 2001
 - Netcraft [Netc02] accessed 37.2 million hosts in July 2002
- [Lawr99] exhaustively crawled 2500 servers and extrapolated
 - Estimated size of the web to be 800 million
 - Estimated use of metadata descriptors:
 - Meta tags (keywords, description) in 34% of home pages, Dublin core metadata in 0.3%

Advantages & disadvantages

- Advantages
 - Clean statistics
 - Independent of crawling strategies
- Disadvantages
 - Doesn't deal with duplication
 - Many hosts might share one IP, or not accept requests
 - No guarantee all pages are linked to root page.
 - Eg: employee pages
 - Power law for # pages/hosts generates bias towards sites with few pages.
 - But bias can be accurately quantified IF underlying distribution understood
 - Potentially influenced by spamming (multiple IP's for same server to avoid IP block)

Alt 3: Random walks

- View the Web as a directed graph
- Build a random walk on this graph
 - Includes various “jump” rules back to visited sites
 - Does not get stuck in spider traps!
 - Can follow all links!
 - Converges to a stationary distribution
 - Must assume graph is finite and independent of the walk.
 - Conditions are not satisfied (cookie crumbs, flooding)
 - Time to convergence not really known
 - Sample from stationary distribution of walk
 - Use the “strong query” method to check coverage by SE

Advantages & disadvantages

- Advantages
 - “Statistically clean” method at least in theory!
 - Could work even for infinite web (assuming convergence) under certain metrics.
- Disadvantages
 - List of seeds is a problem.
 - Practical approximation might not be valid.
 - Non-uniform distribution
 - Subject to link spamming

Conclusions

- No sampling solution is perfect.
- Lots of new ideas ...
-but the problem is getting harder
- Quantitative studies are fascinating and a good research problem



Duplicate detection

Duplicate documents

- The web is full of duplicated content
- Strict duplicate detection = exact match
 - Not as common
- But many, many cases of near duplicates
 - E.g., Last modified date the only difference between two copies of a page

Duplicate/Near-Duplicate Detection

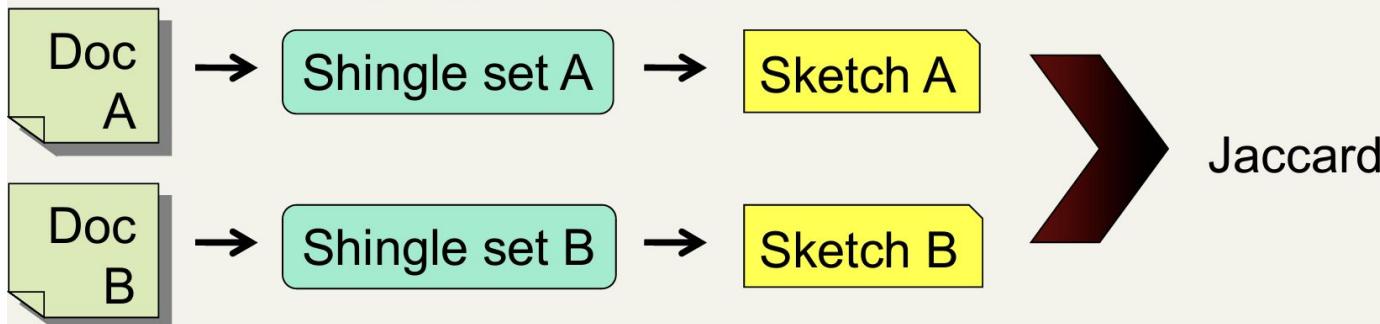
- *Duplication*: Exact match can be detected with fingerprints (64-bit), if fingerprints match, check further
- *Near-Duplication*: Approximate match
 - Overview
 - Compute syntactic similarity with an edit-distance measure
 - Use similarity threshold to detect near-duplicates
 - E.g., Similarity > 80% => Documents are “near duplicates”
 - Not transitive though sometimes used transitively

Computing Similarity

- Features:
 - Segments of a document (natural or artificial breakpoints)
 - Shingles (Word N-Grams)
 - **a rose is a rose is a rose** →
a_rose_is_a
rose_is_a_rose
is_a_rose_is
a_rose_is_a
- Similarity Measure between two docs (= sets of shingles)
 - Set intersection
 - Specifically ($\text{Size_of_Intersection} / \text{Size_of_Union}$)
Jaccard similarity

Shingles + Set Intersection

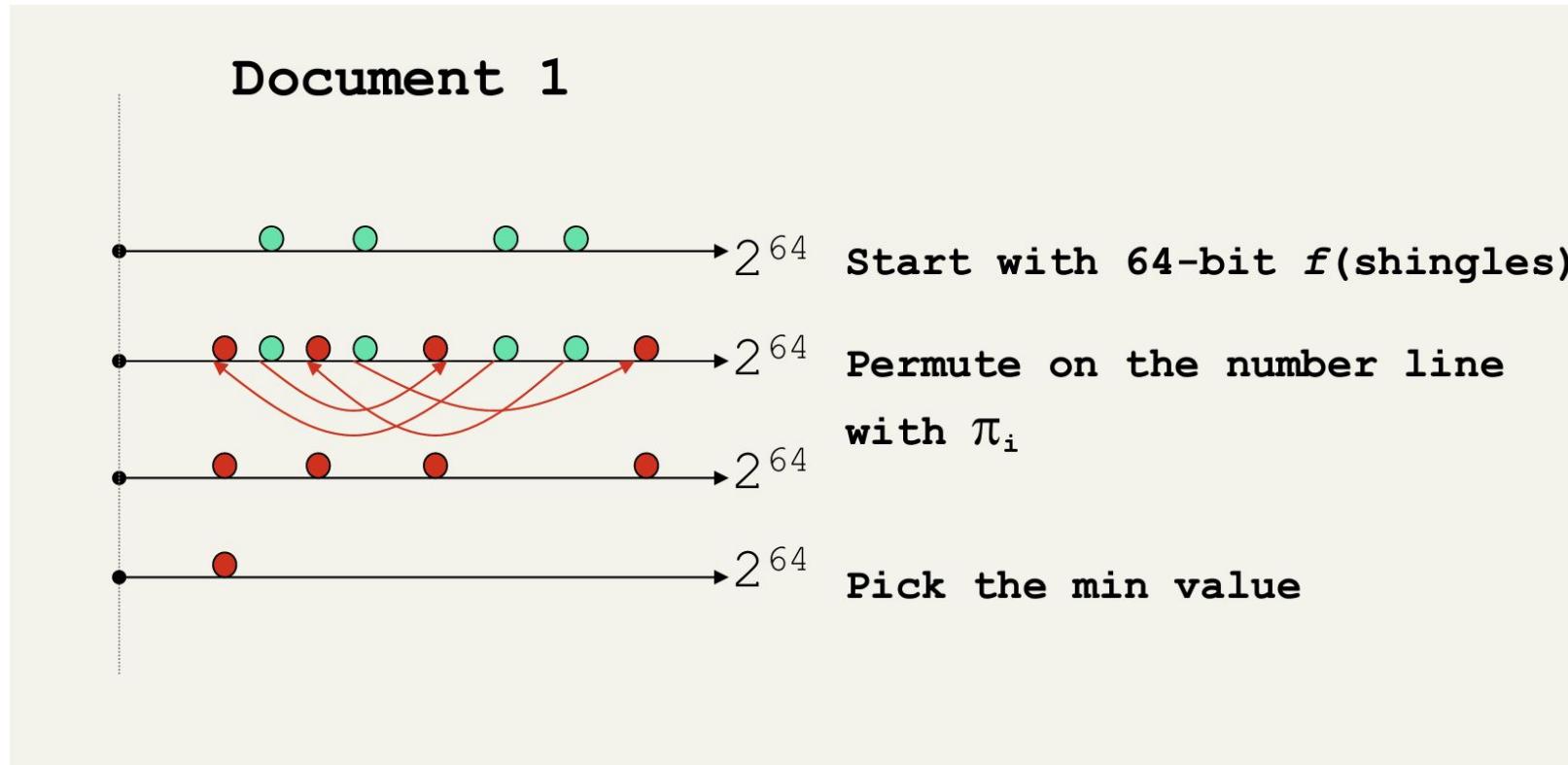
- Computing exact set intersection of shingles between all pairs of documents is expensive/intractable
 - Approximate using a cleverly chosen subset of shingles from each (a *sketch*)
 - Estimate $(\text{size_of_intersection} / \text{size_of_union})$ based on a short sketch



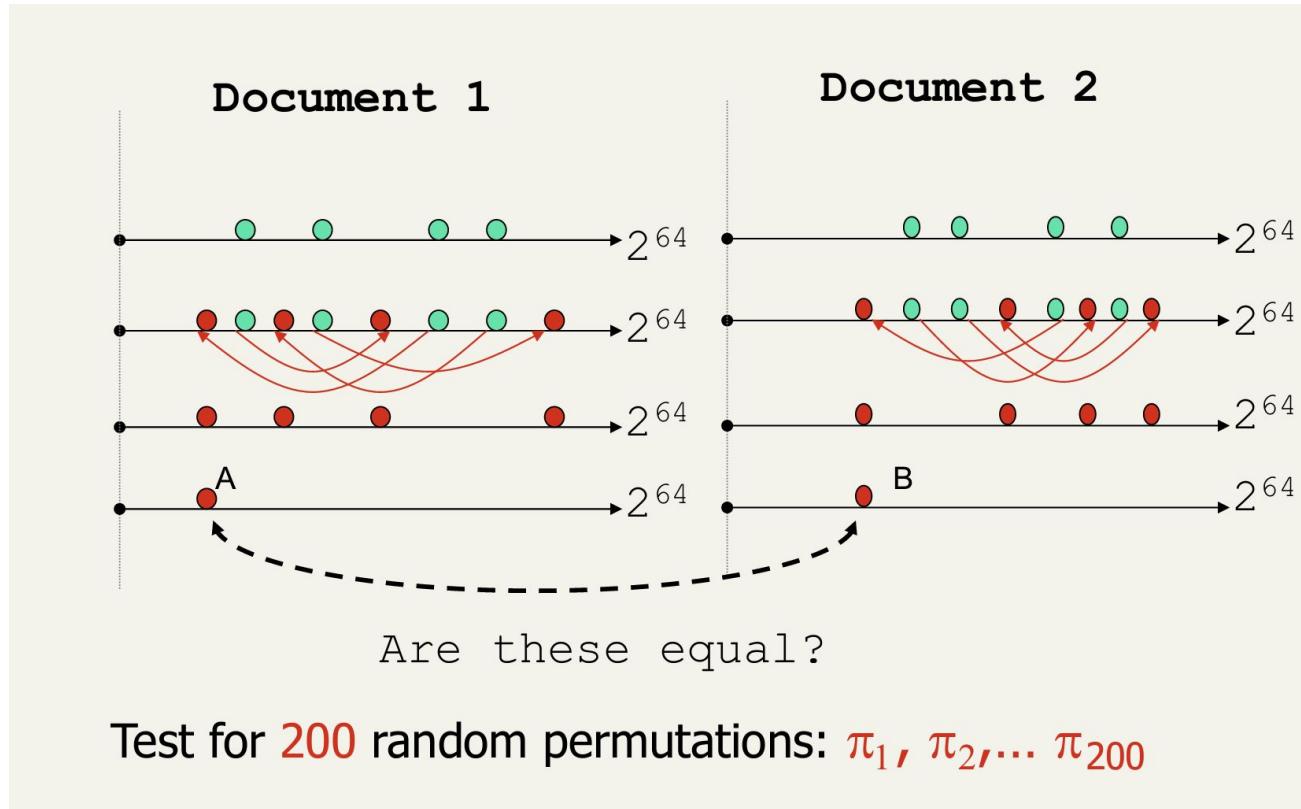
Sketch of a document

- Create a “sketch vector” (of size ~ 200) for each document
 - Documents that share $\geq t$ (say 80%) corresponding vector elements are **near duplicates**
 - For doc D , $\text{sketch}_D[i]$ is as follows:
 - Let f map all shingles in the universe to $0..2^m$ (e.g., $f = \text{fingerprinting}$) 2^m is size of powerset
 - Let π_i be a *random permutation* on $0..2^m$
 - Pick $\text{MIN } \{\pi_i(f(s))\}$ over all shingles s in D

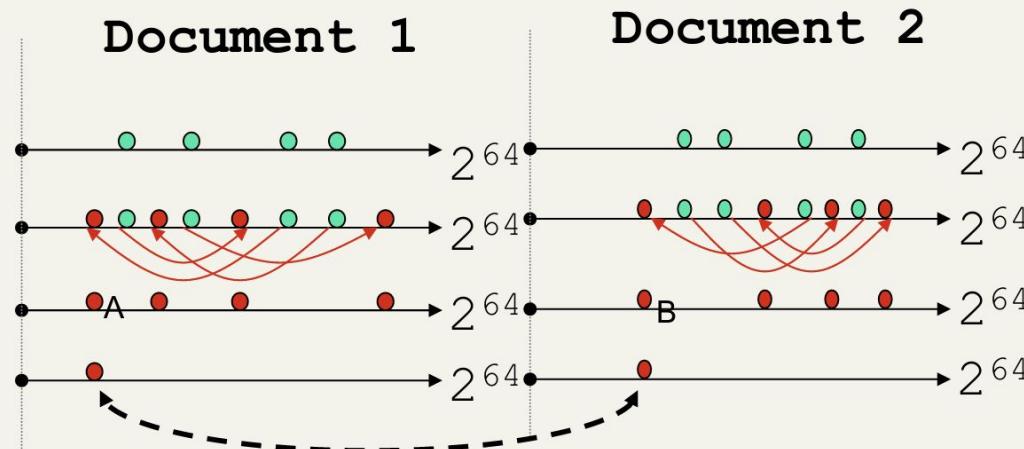
Computing Sketch[i] for Doc1



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



However...



$A = B$ iff the shingle with the MIN value in the union of Doc1 and Doc2 is common to both (i.e., lies in the intersection)

Claim: This happens with probability

$\text{Size_of_intersection} / \text{Size_of_union}$

Why?

Set Similarity of sets C_i, C_j

$$\text{Jaccard}(C_i, C_j) = \frac{|C_i \cap C_j|}{|C_i \cup C_j|}$$

- View sets as columns of a matrix A ; one row for each element in the universe. $a_{ij} = 1$ indicates presence of item i in set j
- Example

$C_1 \quad C_2$

0 1

1 0

1 1

Jaccard(C_1, C_2) = 2/5 = 0.4

0 0

Union: any row
containing a 1

1 1

0 1

Key Observation

- For columns C_i, C_j , four types of rows

	C_i	C_j
A	1	1
B	1	0
C	0	1
D	0	0

- Overload notation: $A = \# \text{ of rows of type A}$
- Claim

$$\text{Jaccard}(C_i, C_j) = \frac{A}{A + B + C}$$

Example

	C ₁	C ₂	C ₃
R ₁	1	0	1
R ₂	0	1	1
R ₃	1	0	0
R ₄	1	0	1
R ₅	0	1	0

Signatures (composed of min)

S₁ S₂ S₃

Perm 1 = (12345)	1	2	1
Perm 2 = (54321)	4	5	4
Perm 3 = (34512)	3	5	4

Similarities

	1-2	1-3	2-3
Col-Col	0.00	0.50	0.25
Sig-Sig	0.00	0.67	0.00

Good estimate

Better estimate if 200 permutations are considered

References

1. Slides provided by Sougata Saha (Instructor, Fall 2022 - CSE 4/535)
2. Materials provided by Dr. Rohini K Srihari
3. <https://nlp.stanford.edu/IR-book/information-retrieval-book.html>