

**NLP**

# Introduction to NLP

*Information Retrieval*

# Introduction

- People search the Web daily
- Search engines
  - Google
  - Bing
  - Baidu
  - Yandex
- Information Retrieval is about search engines

# Yahoo Search

The screenshot shows a web browser window with the address bar displaying the URL: `https://search.yahoo.com/yhs/search?p=ebola&ei=UTF-8&hspart=mozilla&hsimp=yhs-001`. The search bar contains the text "ebola" and a "Search" button. The page header includes the "YAHOO!" logo, navigation links for "Web", "Images", "Video", "Local", and "Maps", and a "Anytime" filter dropdown. The main content area is titled "Ebola Virus News" and features several news snippets:

- Liberia sees surge in new Ebola cases in border county**  
Reuters via Yahoo! News 2 hours ago  
MONROVIA (Reuters) - An outbreak of Ebola cases in a western Liberia county threatens the country's goal of recording no new cases of the disease by the end of the year. From Dec. 1 to 25, some 49...
- Malaria killing thousands more than Ebola in West Africa**  
Associated Press via Yahoo! News 19 hours ago
- The forgotten epidemic? Malaria kills thousands more than Ebola**  
CBS News 2 hours ago
- Ebola virus disease - Wikipedia, the free encyclopedia**  
[en.wikipedia.org/wiki/Ebola\\_virus\\_disease](http://en.wikipedia.org/wiki/Ebola_virus_disease) Cached  
Ebola virus disease (EVD; also Ebola hemorrhagic fever, or EHF), or simply Ebola, is a disease of humans and other primates caused by ebolaviruses.
- #5 of 10 Most Popular Galleries of 2014: Ebola Outbreak in West Africa - Yahoo**  
[news.yahoo.com](http://news.yahoo.com)

On the right side, there are several "Ads" (sponsored links):

- Ebola Virus Pictures**  
[Lifescript.com/Health](http://Lifescript.com/Health)  
Find Facts, Symptoms & Treatments. Trusted By 50 Million Visitors.
- The Ebola Outbreak**  
[news.yahoo.com/ebola](http://news.yahoo.com/ebola)  
Stay informed with Katie Couric Yahoo News reveals the real world.
- Ebola Virus Treatment**  
[bit.ly/EbolaTreatment](http://bit.ly/EbolaTreatment)  
Find out How to Prevent and Treatment for Ebola Virus, Read it Here!
- About The Disease Ebola**  
[buyerpricer.com/aboutdiseaseebola](http://buyerpricer.com/aboutdiseaseebola)  
Searching for About The Disease Ebola? Find Info & Browse Results Now.

At the bottom, there is a small image of a man being carried away in a white van, with the caption: "A man is carried away to be tested for Ebola after collapsing on a street in Monrovia December 9, 2014. The death toll from the Ebola outbreak in West". Below this, a red banner displays a warning icon and the text: "no bucket Y! Confidential".

# Amazon Search

The screenshot shows the Amazon website interface with a search for "samsung galaxy". The browser address bar shows the URL: `www.amazon.com/?ref=UTR&field-keywords=samsung+galaxy&index=blended&link_code=qs&sourceid=Mc`. The search bar contains "samsung galaxy". The page displays 1-16 of 9,697,207 results. The left sidebar shows navigation links for "Cell Phones & Accessories" and "Computers & Accessories", along with a "Refine by" section listing brands like Samsung, Hongyada, Ancerson, shopping\_shop2000, ACEFAST INC, EVTECH, Gravydeals, and Intaure. The main content area features "Related Searches" for "samsung galaxy s3", "samsung galaxy s4", and "samsung galaxy s5". Below this is "Amazon's Samsung Store" with a grid of product categories: Samsung, Unlocked Cell Phones, Cell Phones & Accessories, Computer Tablets, and Computers & Accessories. The first product listing is the "Samsung Galaxy Tab 4 (7-Inch, White)" by Samsung, dated May 1, 2014. It shows a price of \$131.49 (reduced from \$199.99) and a 4.5-star rating with 1,120 reviews. The second product listing is the "Samsung Galaxy S Duos II S7582 White DUAL SIM Factory Unlocked International Ver" by Samsung, with a 4.5-star rating and 214 reviews. A "Sign in" pop-up is visible over the account link in the top right.

Connecting...  
www.amazon.com/?ref=UTR&field-keywords=samsung galaxy&index=blended&link\_code=qs&sourceid=Mc  
samsung galaxy

Firefox has prevented the outdated plugin "Adobe Flash" from running on www.amazon.com. Continue Blocking Allow...

amazon  
Try Prime  
Your Amazon.com Today's Deals Gift Cards Sell Help

Shop by Department Search All samsung galaxy Go

Hello, Sign in Your Account Try Prime Cart Wish List

1-16 of 9,697,207 results for "samsung galaxy"

Show results for

Cell Phones & Accessories >  
Unlocked Cell Phones  
Cell Phones  
+ See more

Computers & Accessories >  
Computer Tablets  
+ See All 32 Departments

Refine by

Eligible for Free Shipping  
Free Shipping by Amazon

Brand

- ☐ Samsung
- ☐ Hongyada
- ☐ Ancerson
- ☐ shopping\_shop2000
- ☐ ACEFAST INC
- ☐ EVTECH
- ☐ Gravydeals
- ☐ Intaure

Related Searches: samsung galaxy s3, samsung galaxy s4, samsung galaxy s5.

Amazon's Samsung Store

Samsung Unlocked Cell Phones Cell Phones & Accessories Computer Tablets Computers & Accessories

Samsung Galaxy Tab 4 (7-Inch, White) May 1, 2014  
by Samsung  
\$199.99 Click to see price Prime  
Get it by Tuesday, Dec 30  
More Buying Choices  
\$131.49 used & new (40 offers)

★★★★★ 1,120  
FREE Shipping and 1 more promotion  
Product Description  
... web, or watching movies, the Samsung Galaxy Tab 4 features a 7.0-inch ...  
Electronics: See all 4,851,153 items

Samsung Galaxy S Duos II S7582 White DUAL SIM Factory Unlocked International Ver  
by Samsung  
\$199.99 214

# Library of Congress Search

The screenshot shows a web browser window displaying the Library of Congress search results for the term "bulgaria". The browser's address bar shows the URL `www.loc.gov/search/?in=&q=bulgaria&new=true&st=`. The Library of Congress logo and navigation menu are visible at the top. The search results page shows 1 to 25 of 1,942 results. The results are refined by "Available Online" (1,942) and "All Items" (40,021). The "Original Formats" section lists various document types and their counts. The first result is a photograph titled "Bulgaria" with a description of an arch built for the 25th anniversary of the accession to the Bulgarian throne of Ferdinand I (1861-1948). The second result is a map titled "Bulgaria" with a description of its base and availability.

Search Results for "bulgari..."

www.loc.gov/search/?in=&q=bulgaria&new=true&st=

LOC.GOV CONGRESS.GOV COPYRIGHT.GOV

LIBRARY OF CONGRESS

Discover Services Visit Education Connect About

All Formats bulgaria GO

Library of Congress > Search

Print Subscribe Share/Save Give Feedback

Results for "bulgaria" 1 - 25 of 1,942

Refined by:

Refine your search

Sort By Relevance Go

View List Go

Available Online 1,942

All Items 40,021

Original Formats

Web Pages	1,041
Books	600
Manuscripts/Mixed Material	327
Photos, Prints, Drawings	200
Legislation	164
Archived Web Sites	55

**Bulgaria**

1 negative : glass ; 5 x 7 in. or smaller. | Photograph possibly shows an arch built for the celebration of the 25th anniversary of the accession to the Bulgarian throne of Ferdinand I (1861-1948). Arch has dates 1887-1912. (Source: Flickr Commons project, 2011)

Contributor: Bain News Service  
Original Format: Photos, Prints, Drawings  
Date: 1912

**Bulgaria.**

"Base 802231 (R01234) 6-94". Also issued with shaded relief. Includes note. Available also through the Library of Congress Web site as a raster image.

# Examples Of Search Engines

- Conventional (library catalog)
  - Search by keyword, title, author, etc.
- Text-based (Lexis-Nexis, Google, Yahoo!)
  - Search by keywords. Limited search using queries in natural language.
- Image-based
  - shapes, colors, keywords
- Question answering systems (ask.com)
  - Search in (restricted) natural language
- Clustering systems (Vivísimo, Clusty)
- Research systems (Lemur, Nutch)

## Sample Queries

- How to get rid of stretch marks
- Dodge
- Kourtney Kardashian
- How many calories are in pumpkn pie
- Angelina Jolie and Brad Pitt
- How to vote
- Derek Jeter
- Interstellar trailer
- What is Ebola?

<https://www.google.com/trends/topcharts>



# The Size Of The World Wide Web

- The size of the indexed world wide web pages (by 2014)
  - Indexed by Google: about 45 Billion pages
  - Indexed by Bing: about 25 Billion pages

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

## Web Statistics

- Twitter hits 400 million tweets per day
  - June, 2012. Dick Costolo, CEO at Twitter
- Over 2.5 billion photos uploaded to Facebook each month (2010)
  - [blog.facebook.com](http://blog.facebook.com)
- Google's clusters process a total of more than 20 petabytes of data per day.
  - 2008. Jeffrey Dean from Google

# Challenges

- Dynamically generated content
- New pages get added all the time
- The size of the blogosphere doubles every 6 months

# Characteristics Of User Queries

- Sessions
  - users revisit their queries
- Very short queries
  - typically 2 words long
- A large number of typos
- A small number of popular queries
  - A long tail of infrequent ones
- Almost no use of advanced query operators
  - with the exception of double quotes

# Information Retrieval

- **Baseline Process**
  - Given a collection of documents
  - And a user's query
  - Find the most relevant documents

## Key Terms Used in IR

- **Query**
  - a representation of what the user is looking for – can be a list of words or a phrase.
- **Document**
  - an information entity that the user wants to retrieve
- **Collection**
  - a set of documents
- **Index**
  - a representation of information that makes querying easier
- **Term**
  - word or concept that appears in a document or a query

# Documents

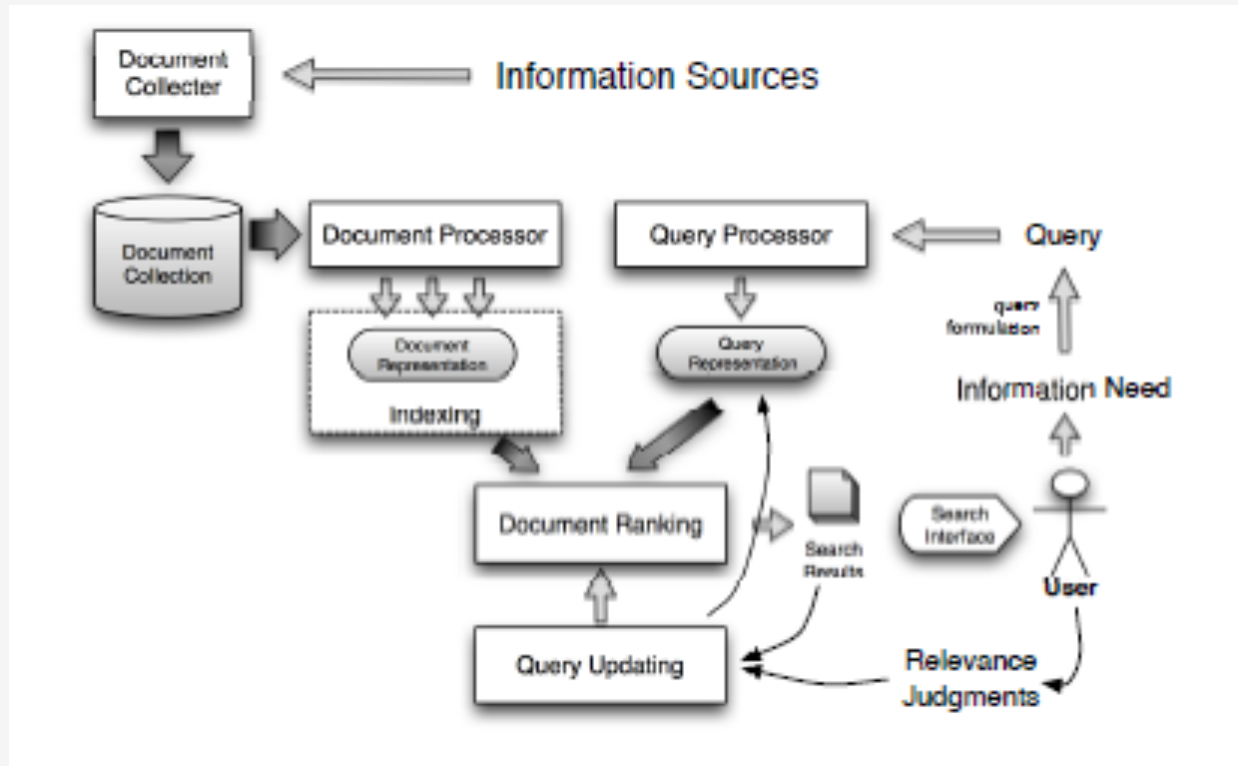
- Not just printed paper
- Can be records, pages, sites, images, people, movies
- Document encoding (Unicode)
- Document representation
- Document preprocessing (e.g., removing metadata)
- Words, terms, types, tokens

# Search Engine Architecture

- Decide what to index
- Collect it
- Index it (efficiently)
- Keep the index up to date
- Provide user-friendly query facilities



# Search Engine Architecture



# Document Representations

- Term-document matrix ( $m \times n$ )
- Document-document matrix ( $n \times n$ )
- Typical example in a medium-sized collection
  - $n=3,000,000$  documents
  - $m=50,000$  terms
- Typical example on the Web
  - $n=30,000,000,000$
  - $m=1,000,000$
- Boolean vs. integer-valued matrices

# Storage Issues

- Imagine a medium-sized collection with  $n=3,000,000$  and  $m=50,000$
- How large a term-document matrix will be needed?
- Is there any way to do better? Any heuristic?

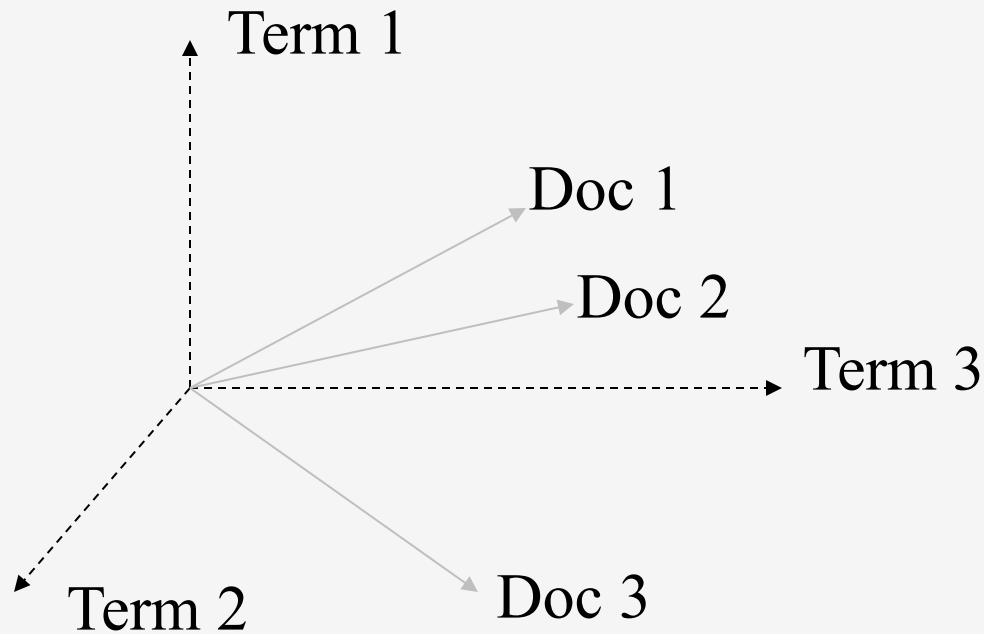
## Inverted Index

- Instead of an incidence vector, use a posting table
  - VERMONT: D1, D2, D6
  - MASSACHUSETTS: D1, D5, D6, D7
- Use linked lists to be able to insert new document postings in order and to remove existing postings.
- Can be used to compute document frequency
- Keep everything sorted! This gives you a logarithmic improvement in access.

# Basic Operations On Inverted Indexes

- **Conjunction (AND)**
  - iterative merge of the two postings:  $O(x+y)$
- **Disjunction (OR)**
  - very similar
- **Negation (NOT)**
  - can we still do it in  $O(x+y)$ ?
  - Example: VERMONT AND NOT MASSACHUSETTS
  - Example: MASSACHUSETTS OR NOT VERMONT
- **Recursive operations**
- **Optimization**
  - start with the smallest sets

# The Vector Model



# Queries as Documents

- **Advantages:**
  - Mathematically easier to manage
- **Problems:**
  - Different lengths
  - Syntactic differences
  - Repetitions of words (or lack thereof)

## Vector Queries

- Each document is represented as a vector
- Non-efficient representation
- Dimensional compatibility

$\mathbf{W}_1$	$\mathbf{W}_2$	$\mathbf{W}_3$	$\mathbf{W}_4$	$\mathbf{W}_5$	$\mathbf{W}_6$	$\mathbf{W}_7$	$\mathbf{W}_8$	$\mathbf{W}_9$	$\mathbf{W}_{10}$
$C_1$	$C_2$	$C_3$	$C_4$	$C_5$	$C_6$	$C_7$	$C_8$	$C_9$	$C_{10}$



# The Matching Process

- Document space
- Matching is done between a document and a query (or between two documents)
- Distance vs. similarity measures.
  - Euclidean distance (define)
  - Manhattan distance (define)
  - Word overlap
  - Jaccard coefficient

# Similarity Measures

- The Cosine measure (normalized dot product)

$$\sigma(D, Q) = \frac{|D \cap Q|}{\sqrt{|D| \cdot |Q|}} = \frac{\sum (d_i \cdot q_i)}{\sqrt{\sum (d_i)^2} \cdot \sqrt{\sum (q_i)^2}}$$

- The Jaccard coefficient

$$\sigma(D, Q) = \frac{|D \cap Q|}{|D \cup Q|}$$

## Exercise

- Compute the cosine scores
  - $\sigma(D_1, D_2)$
  - $\sigma(D_1, D_3)$
- for the documents
  - $D_1 = \langle 1, 3 \rangle$
  - $D_2 = \langle 100, 300 \rangle$
  - $D_3 = \langle 3, 1 \rangle$
- Compute the corresponding Euclidean distances, Manhattan distances, and Jaccard coefficients.

## Phrase-based Queries

- Examples
  - “New York City”
  - “Ann Arbor”
  - “Barack Obama”
- We don’t want to match
  - York is a city in New Hampshire

# Positional Indexing

- Keep track of all words and their positions in the documents
- To find a multi-word phrase, look for the matching words appearing next to each other

# Document Ranking

- Compute the similarity between the query and each of the documents
- Use cosine similarity
- Use TF\*IDF weighting
- Return the top K matches to the user

# IDF: Inverse Document Frequency

- Motivation
- Example

$N$ : number of documents

$d_k$ : number of documents containing term  $k$

$f_{ik}$ : absolute frequency of term  $k$  in document  $i$

$w_{ik}$ : weight of term  $k$  in document  $i$

$$\text{idf}_k = \log_2(N/d_k) + 1 = \log_2 N - \log_2 d_k + 1$$

**NLP**