NLP

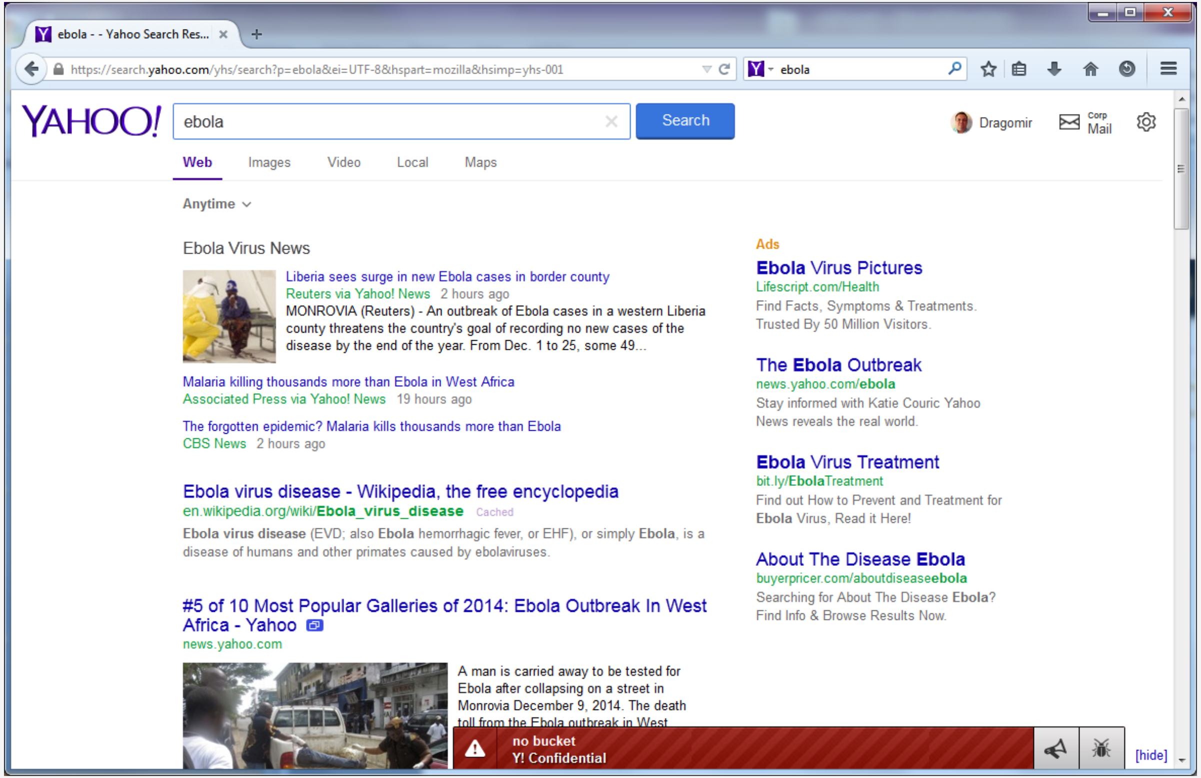
Introduction to NLP

Information Retrieval

Introduction

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

Yahoo Search



Amazon Search

Library of Congress Search

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
- 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

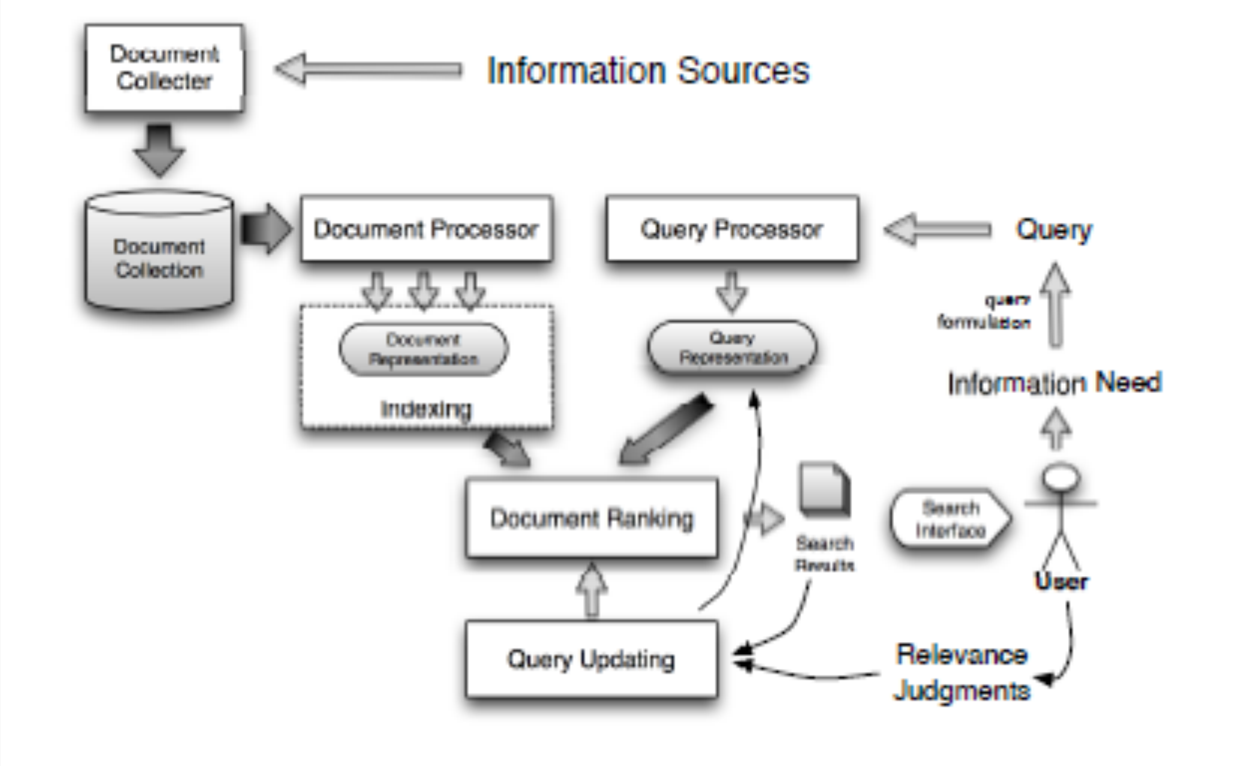
Documents

- 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 x n)
- Document-document matrix (n x 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?

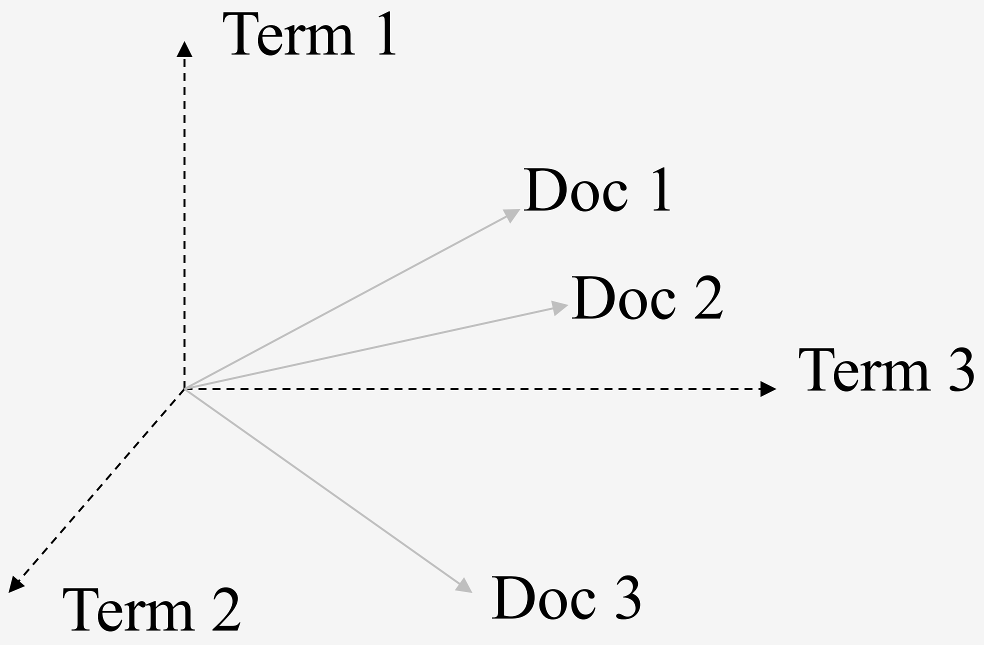
Inverted Index

- 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

W_1	W_2	W_3	W_4	W_5	W_6	W_7	W_8	W_9	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)

$$\sigma(D,Q) = \frac{|D \cap Q|}{\sqrt{|D| \, . \, |Q|}} = \frac{\Sigma\,(d_i \, . \, q_i)}{\sqrt{\Sigma\,(d_i)^2} \; . \; \sqrt{\Sigma\,(q_i)^2}}$$

- 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

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

NLP