Information Retrieval and Web Search

Course overview

Instructor: Rada Mihalcea
The mobile web receives 217 new users.

YouTube users upload 48 hours of new video.

Email users send 204,166,667 messages.

Google receives over 2,000,000 search queries.

Facebook users share 684,478 pieces of content.

571 new websites are created.

Twitter users send over 100,000 tweets.

Foursquare users perform 2,083 check-ins.

Flickr users add 3,125 new photos.

Instagram users share 3,600 new photos.

Tumblr blog owners publish 27,778 new posts.

Brands & organizations on Facebook receive 34,722 “likes.”

Apple receives about 47,000 app downloads.

571 new websites are created.

With no signs of slowing, the data keeps growing. These are just some of the more common ways that Internet users add to the big data post. In truth, depending on the size of business you're in, there are virtually countless other sources of relevant data to pay attention to. Consider the following:

The global internet population grew 10% percent from 2010 to 2011 and now represents 2.1 billion people.
What is this course about?

- Processing
- Indexing
- Retrieving
- ... textual data
- (or audio, video, geo-spatial, ..., data)

- Fits in four lines, but much more complex and interesting than that
Need for Information Retrieval

- With the advance of WWW - more than 20 Billion documents indexed on Yahoo, Google, Bing

- Various needs for information:
  - Search for documents that fall under a given topic
  - Search for an answer to a question
  - Search for information in a different language
  - Search for emails
  - Search for patents
  - ...
  - Search for images
  - Search for music
  - Search for a (candidate) friend
Definition of IR

Salton (1989): “Information-retrieval systems process files of records and requests for information, and identify and retrieve from the files certain records in response to the information requests. The retrieval of particular records depends on the similarity between the records and the queries, which in turn is measured by comparing the values of certain attributes to records and information requests.”
Restated...

- Information Retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).

- These days we often think of Web search, but there are also other types of searches, e.g.:
  - Search your own computer
  - Search knowledge bases
  - Search the library catalogue
  - Search the deep Web (e.g., search for a certain car on a rental agency web page)
Examples of IR systems

- Conventional (library catalog)
  Search by keyword, title, author, etc. E.g.: You are probably familiar with mirlyn.lib.umich.edu

- Text-based (Lexis-Nexis, Google, Bing).
  Search by keywords. Some may use queries in natural language.

- Multimedia (YouTube, Flickr, Tineye)
  Search for/by visual appearance (shapes, colors,...).

- Question answering systems (Ask, Start)
  Search in (restricted) natural language

- Other:
  cross language information retrieval, music retrieval
IR systems on the Web

- Search for Web pages http://www.google.com
- Search for answers to questions http://www.ask.com
- Search for tweets https://twitter.com/search-home
- Search for images http://www.picsearch.com
- Search using image queries http://images.google.com
- Search for similar images http://www.incogna.com
- Search for (image) colors http://labs.tineye.com/multicolr
- Music retrieval http://www.peachnote.com, Shazam app
Course information

• Instructor: Rada Mihalcea
  – Besyter 3769, mihalcea@umich.edu

• GSI: Shibamouli Lahiri
  – Beyster 1695, lahiri@umich.edu

• Class meets MW, 12:00-1:30pm

• Office hours
  – Instructor: W 2:00-3:00pm
  – GSI: T 11:30-1:30pm, Th 11:30-1:30pm, F 12:30-2:30pm
  – Any time electronically
Course resources

• Class webpage:
  – http://web.eecs.umich.edu/~mihalcea/courses/498IR
  – check periodically for updates, announcements, etc.

• Textbook:
  – Introduction to Information Retrieval
    Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze

• Recommended:
  – Readings in Information Retrieval
    K.Sparck Jones and P. Willett
  – Modern Information Retrieval
    Ricardo Baeza-Yates and Berthier Ribeiro-Neto

• Papers:
  – Several papers will be assigned throughout the semester
Course communication

• Use the Piazza forum for any technical communication related to the class
  – Likely to get a faster answer than if you email the instructor or GSI individually
  – We will try to answer any question sent on the forum within 24 hours (but your peers may answer even faster!)
Grading (tentative)

- Four programming assignments: 35%
  - Start early! Some may be time consuming
  - 3 days late policy

- Exam I: 20%

- Exam II: 20%

- Project: 25%

- No final – final is replaced by the project
Programming language

• All assignments / project will be in Python
  • Makes life much much easier for text processing problems and for Web based applications
  • Information Retrieval involves a lot of text processing, and often involves Web access
    – Code reusability

• Code must run on CAEN

• Do not use libraries that directly solve the assignment/project
  – If in doubt, ask the instructor/GSI
Tentative schedule

- Course Overview
- Introduction to IR models and methods
- Web crawling
- Text analysis and text properties
- Boolean model
- Vector-based model
- Probabilistic model; other IR models
- IR evaluation and IR test collections
- Relevance feedback, query expansion
- Web search: link based and content based
- Query-based and content sensitive link analysis
Tentative schedule

• Text classification and text clustering
• Question answering and information extraction
• Text summarization and keyword extraction
• Cross Language IR
• Social media, crowdsourcing
• Image retrieval
• Music retrieval
• Geospatial search
• Two guest lectures - TBA