What you will learn

• How to build a search engine
  – which search results to rank at the top
  – how to do it fast and on a massive scale
• How to evaluate a search algorithm
  – is system A really better than system B
• How to work with text
  – two tweets talk about the same topic?
  – handle misspellings, morphology, synonyms
  – build algorithms for languages you don’t know

Overview

• Information Retrieval
  – Two main issues in IR: speed and accuracy
  – Documents, queries, relevance
  – Bag-of-words trick
• Overview of Search System Architectures
• Other IR tasks
• Course logistics

Information Retrieval (IR)

“Information retrieval is a field concerned with the structure, analysis, organization, storage, searching, and retrieval of information.”
(Salton, 1968)

• IR – core technology for text processing
• widely used in NLP/DBMS applications
• driving force behind web technologies

IR in a nutshell

Two main issues in IR

• Effectiveness
  – need to find relevant documents
  – needle in a haystack:
  – very different from relational DBs (SQL)
• Efficiency
  – need to find them quickly:
  – vast quantities of data (100b pages)
  – thousands queries per second
  – data constantly changes, need to keep up
  – compared with other NLP areas IR is very fast

Documents

• “documents” has a very wide meaning:
• "documents" has a very wide meaning:
  – web-pages, emails, word/pdf/excel, news
  – photos, videos, musical pieces, code
  – answers to questions
  – product descriptions, advertisements
  – may be in a different language
  – may not have words at all (e.g. DNA)
• IR: match A against a large set of Bs
  – problem arises in many different domains

8 Queries
• web search:
  – query = a few keywords ("homer simpson")
• query = expression of information need
  – describes what you want to find
  – can have many forms:
    • keywords, narrative, example “document”
    • question, photo, scribble, humming a tune
    • \#wsum(0.9 \#field (title, \#phrase (homer,simpson))
      0.7 \#and (\#> (pagerank,3), \#owl (homer,simpson))
      0.4 \#passage (homer, simpson, dan, castellaneta))

9 Relevance
• at an abstract level, IR is about:
  does item $D$ match item $Q$? …or…
  is item $D$ relevant to item $Q$?
• relevance a tricky notion
  – will the user like it / click on it?
  – will it help the user achieve a task?
  – is it novel (not redundant)?
• common take: relevance = topicality / aboutness
  – i.e. $D,Q$ share similar “meaning”
  – about the same topic / subject / issue

10 Why is matching a challenge?
• no clear semantics, contrast:
  – author = X123456 ("Shakespeare, William") vs.
  – "play, frequently attributed to Shakespeare, is in fact"
• inherent ambiguity of language:
  – synonymy: “banking crisis” = “financial meltdown”
  – polysemy: “Homer” can be “Simpson” or “Greek author”
• relevance highly subjective
  – Anomalous State of Knowledge (Belkin)
• relevance not observable (when we need it)
• on the web: counter SEOs / spam

11 How do search engines do it?
• not with relational DBs
  – ok in niche domains (libraries)
  – “tagging” works for multi-media
    • spammers, loses “clarity” with scale
  – “semantic web” \( \Rightarrow \) inconsistent ontologies
• not by “understanding” the language
  – NLP brittle in unrestricted domains
    • good w. fixed structure/vocabulary (e.g. takeovers)
      – computationally expensive

12 Relevant Items are Similar
• Key idea:
  – use similar vocabulary \( \Rightarrow \) similar meaning
    – similar documents relevant to same queries
• Similarity
  – string match
  – word overlap
  – P (same model)

13 Bag-of-words trick
• Can you guess what this is about:
  – beating falls 355 Dow another takes points
    – said fat fries McDonalds French obesity
• Re-ordering doesn’t destroy meaning
  – individual words – “building blocks”
  – “bag” of words: a “composition” of “meanings”

14 Bag-of-words trick (2)
• Most search engines use BOW
  – treat documents, queries as bags of words
    • a “bag” is a set with repetitions (multi-set, urn)
  – match = “degree of overlap” between \( D, Q \)
• Retrieval models
  – statistical models that use words as features
    – decide which docs most likely to be relevant
      • what should be the top 10 for “homer simpson”? 
    – BOW makes these models tractable

15 Bag-of-words: criticisms
• word meaning lost without context
  – true, but BOW doesn’t really discard context
    • it discards surface form / well-formedness of text
• what about negations, etc.?
• “not, but he loves me” vs. “but he loves me not”
  • still discusses the same subject (him, me, love)
  • propagate negations to words: “but he not_loves me”
• does not work for all languages
  – no natural “word” unit in Chinese, images, music
  – circumvent by “segmentation” or “feature induction”
  • break/aggregate until units reflect “aboutness”

16 Systems perspective on IR
• get the data into the system
  – acquire the data from crawling, feeds, etc.
  – store the originals (if needed)
  – transform to BOW and “index”
• satisfy users’ requests
  – assist user in formulating query
  – retrieve a set of results
  – help user browse / re-formulate
  – log user’s actions, adjust retrieval model

17 Indexing Process
18 Search Process