



THE UNIVERSITY
of EDINBURGH

Text Technologies for Data Science

INFR11145

Definitions

Instructor:
Walid Magdy

19-Sep-2017

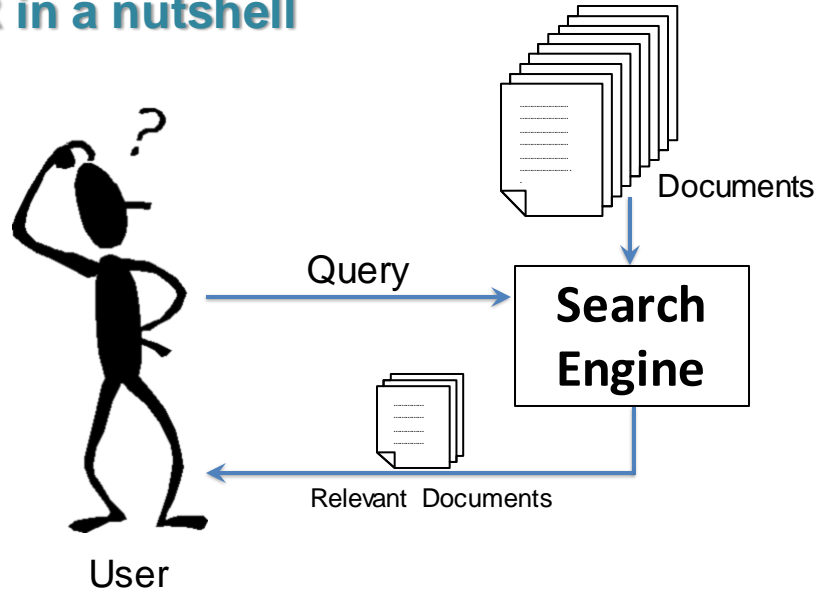
Lecture Objectives

- Learn about main concepts in IR
 - Document
 - Information need
 - Query
 - Index
 - BOW



THE UNIVERSITY
of EDINBURGH

IR in a nutshell



Walid Magdy, TTDS 2017/2018



IR, basic form

- Given Query **Q**, find relevant documents **D**

The screenshot shows a Google search for "donald trump". Red arrows and brackets highlight the query "donald trump" and the relevant documents returned, including top stories, a biography, and a Twitter profile.

Top stories:

- Trump on Irma: "We've never seen anything like this" - CNN.com - 1 hour ago
- Bound to No Party, Trump Opens 150 Years of Two-Party Rule - The New York Times - 19 h...
- Hurricane Irma: Florida Keys hit by 'most catastrophic storm ever' - latest news - The Telegraph - 20 mins ago

Donald Trump Biography: Trump Organization Hotel's Real Estate Golf ...
www.trump.com/biography/
 Donald J. Trump is the very definition of the American success story, continually setting the standards of excellence while expanding his interests in real estate, ...

Donald J. Trump (@realDonaldTrump) · Twitter
<https://twitter.com/realDonaldTrump>

Donald Trump
 45th U.S. President
donalddump.com

Donald John Trump is the 45th and current President of the United States, in office since January 20, 2017. Before entering politics, he was a businessman and television personality. Wikipedia

Born: 14 June 1946 (age 71), Jamaica Hospital Medical Center
Height: 1.88 m
Net worth: 3.5 billion USD (2017) Forbes
Spouse: Melania Trump (m. 2005), Maria Maples (m. 1993–1999), Ivana Trump (m. 1977–1992)
Education: Wharton School of the University of Pennsylvania (1968), MORE ...

Walid Magdy, TTDS 2017/2018



Two main Issues in IR

About 293,000,000 results (0.79 seconds)

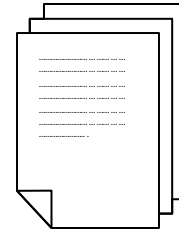
- Effectiveness
 - need to find **relevant** documents
 - needle in a haystack
 - very different from relational DBs (SQL)
- Efficiency
 - need to find them quickly
 - vast quantities of data (10's billions pages)
 - thousands queries per second (Google, 40,000)
 - data constantly changes, need to keep up
 - compared with other NLP areas IR is **very** fast

IR main components

- Documents
- Queries
- Relevant documents

Documents

- The element to be retrieved
 - Unstructured nature
 - Unique ID
 - N documents \rightarrow Collection
- web-pages, emails, book, page, sentence, tweets
- 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)



Walid Magdy, TTDS 2017/2018



Queries

- Free text to express user's information need
- Same information need can be described by multiple queries
 - Latest news on the hurricane in the US
 - Florida storm
 - Irma
- Same query can represent multiple information needs
 - Apple
 - Jaguar



Walid Magdy, TTDS 2017/2018



Queries – different forms

- Web search → keywords, narrative ...
- Image search → keywords, sample image
- QA → question
- Music search → humming a tune
- Filtering/recommendation → user's interest/history
- Scholar search → structured (author, title ..)
- Advanced search
`#wsyn(0.9 #field (title, #phrase (homer,simpson)) 0.7 #and (#>
 (pagerank,3), #ow3 (homer,simpson)) 0.4 #passage (homer, simpson,
 dan, castellaneta))`

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?
(satisfy information need)
 - is it novel (not redundant)?
- *Relevance = what is the topic about?*
 - i.e. **D**, **Q** share similar “meaning”
 - about the same topic / subject / issue

What is the challenge in relevance?

- No clear semantics, contrast:
 - “William Shakespeare”
 - Author history’s? list of plays? a play by him?
- Inherent ambiguity of language:
 - synonymy: “Florida storm” = “Irma hurricane”
 - polysemy: “Apple”, “Jaguar”
- Relevance highly subjective
 - Rel: yes/no, Rel: perfect/excellent/good/fair/bad
- On the web: counter SEOs / spam

Relevant Items are Similar

- Key idea:
 - Use similar vocabulary → similar meaning
 - Similar documents relevant to same queries
- Similarity
 - String match
 - Word overlap
 - $P(D|Q)$ → retrieval model

IR vs. DB

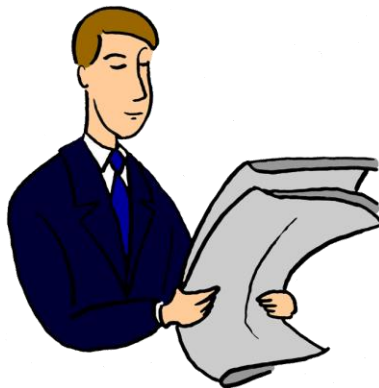
	Databases	IR
What we're retrieving	Structured data. Clear semantics based on a formal model.	Mostly unstructured. Free text with some metadata.
Queries we're posing	Formally-defined (relational algebra, SQL). Unambiguous.	Free text ("natural language"), Boolean
Results we get	Exact (always "correct")	Imprecise (need to measure relevance)
Interaction with system	One-shot queries.	Interaction is important.

Tamer Elsayed, QU

Walid Magdy, TTDS 2017/2018

THE UNIVERSITY
of EDINBURGH

How IR sees documents?



Walid Magdy, TTDS 2017/2018

THE UNIVERSITY
of EDINBURGH

Bag-of-words trick

- Can you guess what this is about:
 - per is salary hour £5,594 Neymar
 - obesity French is of full cause and fat fries
- Re-ordering doesn't destroy the topic
 - individual words – “building blocks”
 - “bag” of words: a “composition” of “meanings”

Walid Magdy, TTDS 2017/2018



Bag-of-words trick

- Most search engines use BOW
 - treat documents, queries as bags of words
- A “bag” is a set with repetitions
 - match = “degree of overlap” between **D**, **Q**
- Retrieval models
 - statistical models (function) that use words as features
 - decide which documents most likely to be relevant
- What should be the top results for **Q**?
 - BOW makes these models tractable

Walid Magdy, TTDS 2017/2018



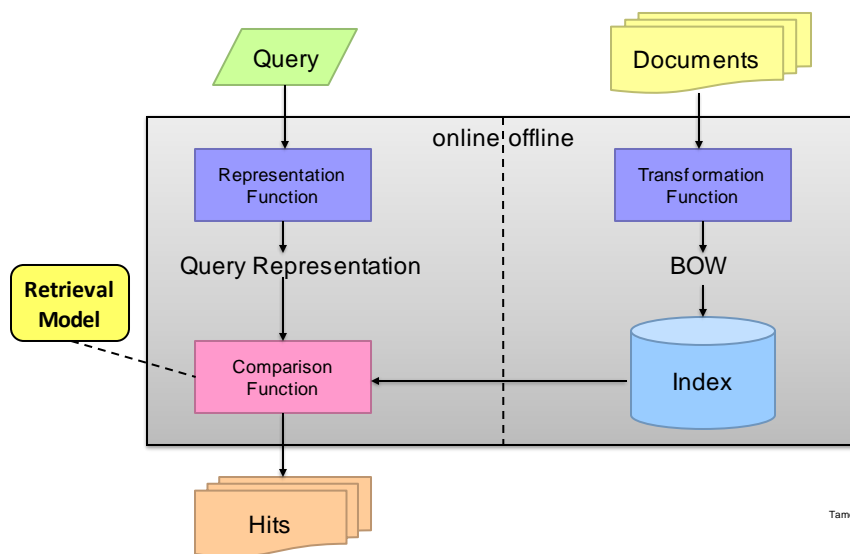
Bag-of-words: Criticism

- word meaning lost without context
 - True, but BOW doesn't really discard context
- what about negations, etc.?
 - {not, climate change is real} vs. {climate change is not real}
- does not work for all languages
 - No natural "word" unit for Chinese, images, music
 - Solve by "segmentation" or "feature induction"

Walid Magdy, TTDS 2017/2018



IR Black Box



Tamer Elsayed, QU

Walid Magdy, TTDS 2017/2018



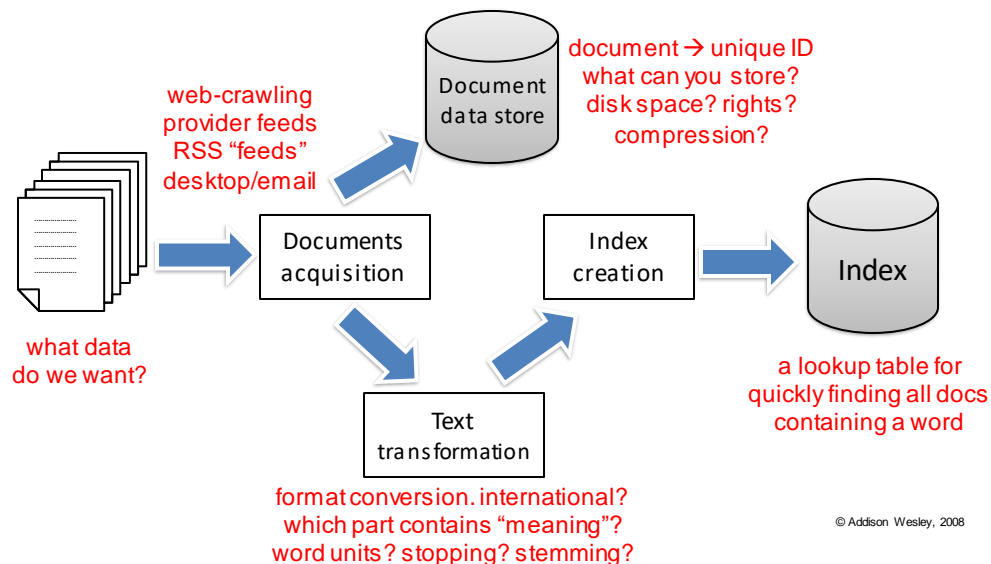
Systems perspective on IR

- Indexing Process: (*offline*)
 - get the data into the system
 - acquire the data from crawling, feeds, etc.
 - store the originals (if needed)
 - transform to BOW and “index”
- Search (retrieval) Process: (*online*)
 - 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

Walid Magdy, TTDS 2017/2018



Indexing Process

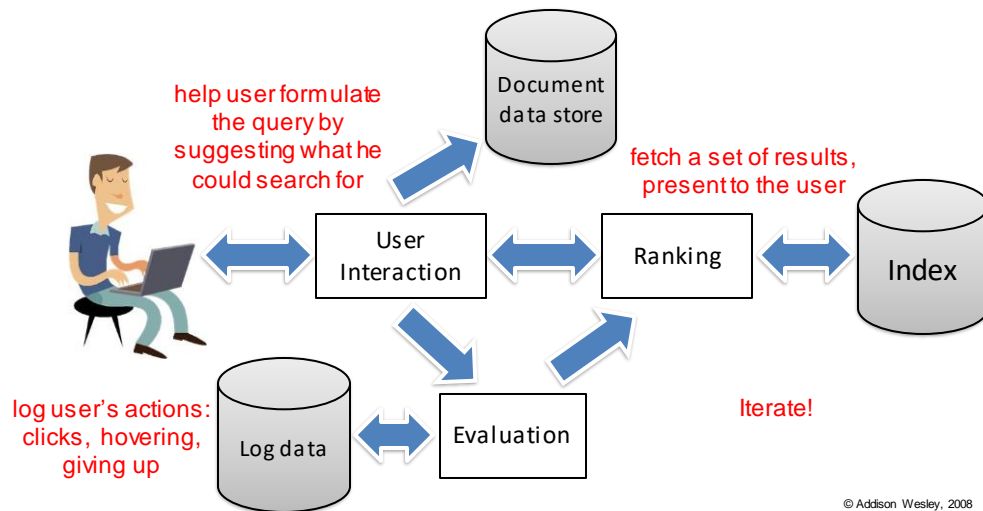


© Addison Wesley, 2008

Walid Magdy, TTDS 2017/2018



Search Process



Walid Magdy, TTDS 2017/2018



Summary

- Information Retrieval (IR): core technology
 - selling point: IR is very fast, provides context
- Main issues: effectiveness and efficiency
- Documents, queries, relevance
- Bag-of-words trick
- Search system architecture:
 - indexing: get data into the system
 - searching: help users find relevant data

Walid Magdy, TTDS 2017/2018



Questions

- Next time:
 - Laws of text (Zipf)
 - Vector space models
- Reading:
 - Search Engines: Information Retrieval in Practice, chapter 1 & 2
- Videos:
 - The Zipf Mystery, Vsauce
- Tools:
 - Perl regular expressions: <https://perldoc.perl.org/perire.html>

