

Text Processing Overview

Lecture 2
 CS 410/510
 Information Retrieval on the Internet

Lecture Outline

- Text processing goals
- Accessing content
 - Text extraction
- Feature extraction
 - Structure
 - Terms
- Feature selection/generation
 - Text operations
 - Selection of index terms

Text Processing Goals

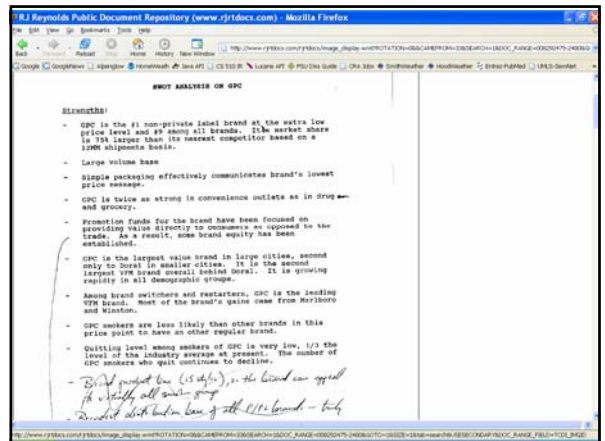
- Enable high quality retrieval results
- Represent document content and characteristics efficiently
 - Storage space
 - Efficient matching of queries to documents

Accessing Document Content

- Identify document types
- Handle each appropriately

Accessing Document Content

Challenge	Possible approach
Scanned documents	OCR
Compressed, encrypted, zipped files	Uncompress, decrypt, unzip
Word processed and other application-specific documents	Application-specific handler
Complex HTML pages: <ul style="list-style-type: none"> •Frames •Pages from CMS •Dynamic content •Scripts embedded in HTML 	Parse HTML to extract content Site-specific wrapper?



Deciding What and How to Index

- HTML tags?
- URL? Page title? Javascript?
- Ads?
- Navigation bars?
- Text in tables?
- Text associated with images?
- Maintain information about document structure?
 - Index document parts as separate fields?



What should you index on this page?

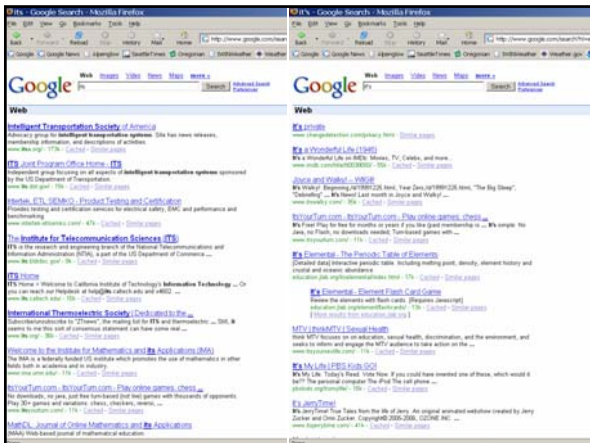
How would you extract it from all the other stuff?

Lexical analysis

- Parse the text into words
 - What is a word?
- Challenges:
 - Punctuation
 - Special characters
 - Hyphens
 - Case
 - Digits

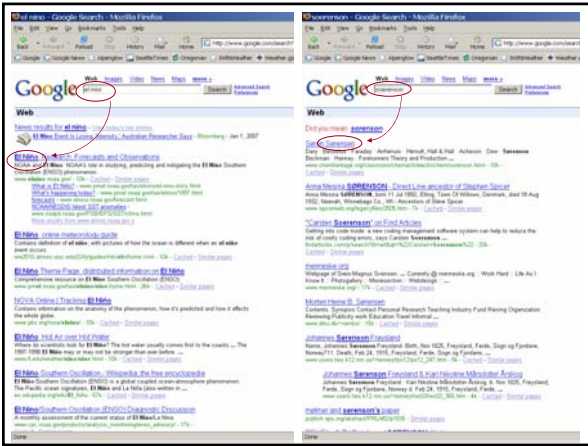
Lexical analysis

- Punctuation
 - Most can just be eliminated
 - Periods
 - End of sentence
 - Abbreviations and acronyms
 - "Dots" in URLs?
 - Apostrophes
 - How much do they change the meaning?
 - Grants versus Grant's
 - Exact match only?
 - Eliminate all in both query and index?
 - Eliminate some and not others?
 - Retain and use some algorithm for matching?
- What do current search engines do?



Lexical analysis

- Special characters
 - What to do with &, \$, !?
 - Are they different if they occur alone vs. part of a word?
- Characters used in other languages
 - Exact match only?
 - Automatically match common translations?
 - el nino and el niño?
- What do current web search engines do?



Lexical analysis

- Hyphens
 - Usage often varies
 - May or may not change meaning
 - *world wide, world-wide, worldwide*
 - *client server, client-server, client/server, ClientServer*
 - *A-Boy, a boy*
 - Generally want to conflate different forms that refer to the same concept
 - What do current search engines do?
 - Are results for *Baeza-Yates* == *Baeza Yates*?

20

Lexical analysis

- Case (upper case, lower case, mixed)
 - Case can be useful for distinguishing proper nouns and acronyms
 - Most search engines appear to convert all text to lower or upper case
 - *Grant* == *grant* on Google, Yahoo, MSN
 - *Bush* == *bush* on Google, Yahoo, MSN
 - Or do they?
 - look further down in the Google results; order differs slightly
 - Ultraseek (commercial software for website/intranet use)
 - if query is lower case, matching is case-insensitive
 - If query has any upper case characters, matching is exact only

21

Lexical analysis

- Digits
 - Baeza-Yates text suggests that digits generally not very useful
 - What about addresses, phone numbers, numbers with significant meaning? (e.g. 911)
 - Major search engines do index digits
 - Perhaps related to ability to handle larger vocabularies than when textbook written?

22

Stemming

- Purpose is to conflate inflexional variations on a word that refer to the same concept
 - singular and plural forms of nouns
 - different tenses of the same verb
 - query *rain cat dog* should match *"raining cats and dogs"*
- Okay if stemmed form is not meaningful
 - User doesn't see or care that *computing* in query and *computed* in text were both stemmed to *comput*
- Obviously is language dependent
 - Inflexions in English are different from those in French

23

Stemming strategies (1)

- Algorithmic
 - Affix removal
 - Usually only remove suffixes since prefixes often change meaning of the word (e.g. do, undo)
 - Successor variety
 - Finds boundaries between morphemes (word segments)
 - Morphemes are the smallest linguistic units with semantic meaning (e.g. unreadable *un-read-able*)
 - Successor variety of a string is the number of different characters that follow it in words in some body of text.
 - Text: read, readable, reading, reads, red, rope, ripe
 - R→3, RE→2, REA→1, READ→3, READA→1, READAB→1, READABL→1, READABLE→1
 - Look for peak and plateau to break words

24

Stemming strategies (2)

- Algorithmic, cont.
 - N-grams
 - An N-gram is a sub-sequence of length n from a sequence
 - Uses character sequences within words
 - Clusters words based on number of shared N-grams
 - Language-independent
- Dictionary
 - Table lookup
- Mixed
 - Use algorithm with table lookup for exceptions

25

Stemming

Porter Stemmer Output	
a	a
abase	abas
abate	abat
abated	abat
abatement	abat
abess	abess
abbey	abbei
abide	abid
abides	abid
abjectly	abjectli

26

Stemming Errors

- Understemming
 - Words that should be conflated are not *Examples?*
- Overstemming
 - Words are conflated that do not share meaning *Examples?*
- Mis-stemming
 - Removing an ending that is part of the stem: *reply*
- What words should be conflated may depend on context
- See discussion on web-site for the Lancaster (Paice/Husk) stemming algorithm, especially <http://www.comp.lancs.ac.uk/computing/research/stemming/Links/error.htm>

27

Stemming

- Effect on retrieval is controversial
- Some claim it improves recall but may degrade precision
- Probably depends on many factors:
 - Language
 - Stemming method
 - Document collection
 - Queries
 - Nature of user task

28

Stemming summary

Advantages	Disadvantages
Searcher doesn't have to anticipate author's exact usage <i>tsunami</i> can match: •What happens to a tsunami as it approaches land? • Tsunamis have been historically referred to as tidal waves because ...	May decrease precision •Might want to match specific word •Might be okay for query <i>tired</i> to match other forms of to tire , but not <i>automobile tire</i>
May increase recall	Stemming can change meaning • <i>bushing</i> – an insulating liner in an opening through which conductors pass • <i>bush</i> - A low shrub with many branches
Results in more compact indexing vocabulary	Stemming algorithms imperfect – occasionally lead to some odd results

29

Stemming

- Do the major search engines stem?
- If so, how?

30

Stopwords

- Very common words are often thought not to be very good discriminators between documents
 - Common words often convey little of what a document is about
 - Articles, conjunctions, prepositions
 - Does *the* or *of* tell you what a document is about?
- Eliminating common words (stopwords) reduces the size of the index

31

Stopwords

- Typically implemented as a set of words for filtering both document text and queries
- Size of list may vary, e.g.
 - MEDLARS stoplist only 7 words
 - *and, an, by, from, of, the, with*
 - Other published lists have 250, 471 words
- May want a collection-specific stopword list

32

Example stopword list applied to a document

a, about, an, and, are, as, by, for, from, had, have, he, his, him, in, into, of, on, or, that, the, this, to, was, with, were

Here ~~were~~ ~~the~~ servants ~~of~~ your adversary,
~~And~~ yours, close fighting ere I did approach:
I drew ~~to~~ part them: ~~in~~ ~~the~~ instant came
~~The~~ fiery Tybalt, ~~with~~ ~~his~~ sword prepared,
Which, ~~as~~ ~~he~~ breathed defiance ~~to~~ my ears,
~~He~~ swung about ~~his~~ head ~~and~~ cut ~~the~~ winds,
Who nothing hurt withal hiss'd ~~him~~ ~~in~~ scorn:
While we were interchanging thrusts ~~and~~ blows,
Came more ~~and~~ more ~~and~~ fought ~~on~~ part ~~and~~ part,
Till ~~the~~ prince came, who parted either part.

33

Stopwords

- Common words may be important
 - Phrases
 - Query for “to be or not to be”
 - Special meaning in context
 - Vitamin A
 - Abbreviations and acronyms
 - OR as abbreviation (ORegon) or acronym (Operating Room)

34

Phrases containing common words

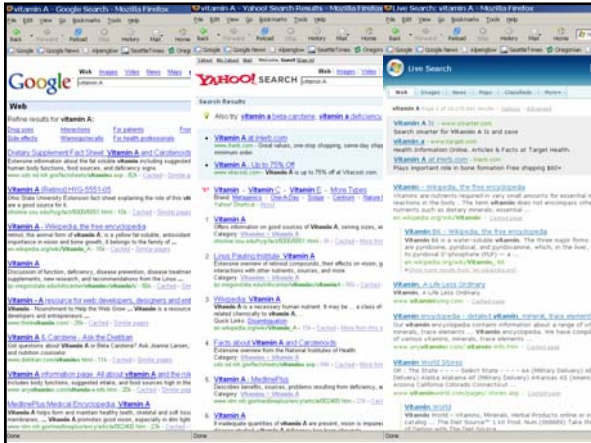
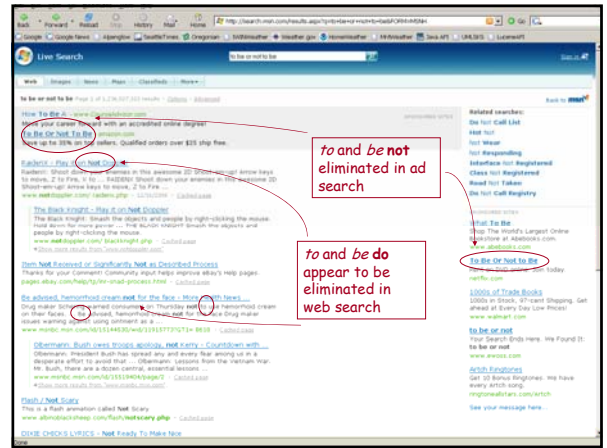
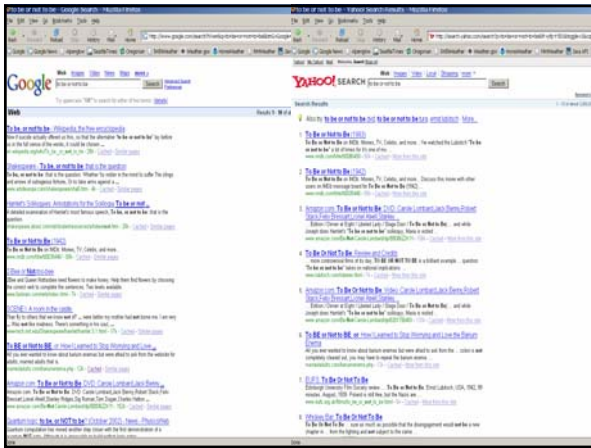
- Index all words
- Ignore stopwords; use position information for “significant” words “the cat in the hat”
 - Accept risk of false matches
 - Filter matching docs by examining full text
- Two indexes
 - Document-level index
 - Eliminate stopwords
 - Use for fast evaluation of standard ranked queries
 - Word-level index with all words
 - Use for evaluating phrase queries

35

Stopwords

- When might stopword elimination hurt your search?
- How would you decide whether to use stopwords or not?
- What do the major web search engines do about stopwords?

36



Index term selection

- Automated
 - All words in text (+/- stopwords)
 - Nouns only
 - Noun groups
 - Statistically common phrases
- Manual
 - Terms assigned from a controlled vocabulary
 - Freely-selected indexing terms

Tradeoffs

- What are some of the tradeoffs involved in pre-processing documents before indexing?