

Outline

- View search engine interface homework
- Document preparation
 - Analysis, normalization
 - Term extraction
- Manual indexing

Document Preparation

- Automatic indexing requires that the documents be in a standard form
- Steps to building an index:
 - Break up corpus into individual documents that can be accessed separately
 - how large can a document be?
 - is it a paragraph, a page, a file, or a directory?
 - Document analysis and purification
 - Token analysis / term extraction

Doc. Purification and Analysis

- Examine how each doc is organized (recognizing the subparts, such as title, author, body, etc.).
- Determine what information or parts of the document will be indexed.
 - In the WWW, search engines often ignore comments, ALT text, META tags, image maps, frames, tiny or 'invisible' text.
- Convert the document to a standard format (e.g., from PostScript to ASCII text, or from HTML to plain text).

Term Extraction

- Determine which words or phrases should be used to represent the semantic content.
 - How can you recognize a word?
 - What about numbers, punctuation, capitalization?
- Possibly remove ubiquitous and/or singleton terms.
- Eliminate (non-content) terms in 'stop lists'

Stopwords

- Function words and connectives (non-content terms)
 - Appear in a large number of documents and are little use in pinpointing documents
 - e.g., a, about, above, after, an, and, another, any
- Indexing stopwords
 - Stopwords not indexed
 - For reducing index space and improving performance
 - Replace stopwords with a placeholder (to record the offset)
- Issues
 - Queries containing only stopwords ruled out
 - Words that are stopwords in one sense but not in others
 - E.g.; *can* as a verb vs. *can* as a noun

Term Extraction, cont.

- Apply stemming to the term.
 - Replaces a term with its root (dropping prefixes and suffixes).
 - Removes inflections that convey parts of speech, tense and number
 - Reduces the vocabulary.
 - run, runs, running -> run
 - university, universal -> univers
 - reformation, reformative, reformatory, reformed -> reform
 - Conflates words to help match a query term with a morphological variant in the corpus.
 - May combine distinct terms.

Stemming, cont.

– Techniques

- Difficult to do accurately
- Apply rules from morphological analysis
 - (e.g., Porter's algorithm)
- Dictionary lookup (e.g., WordNet)

– Stemming may increase recall but at the price of precision

- Abbreviations, polysemy and names coined in the technical and commercial sectors (more common in unedited WWW)
- E.g.: Stemming “ides” to “IDE”, “SOCKS” to “sock”, “gated” to “gate”, may be bad !

Manual Indexing

- Manual Indexing (e.g., Yahoo, DMOZ, Libraries)
 - Assign keywords, categories
 - Finds high-quality connections between documents
 - Time-consuming, expensive
 - Inconsistent from person to person
 - May miss connections (esp. if categories change over time)