Lecture 5.1: Document Preprocessing

01204453 Web Information Retrieval and Mining

Department of Computer Engineering
Faculty of Engineering, Kasetsart University
Bangkok, Thailand.

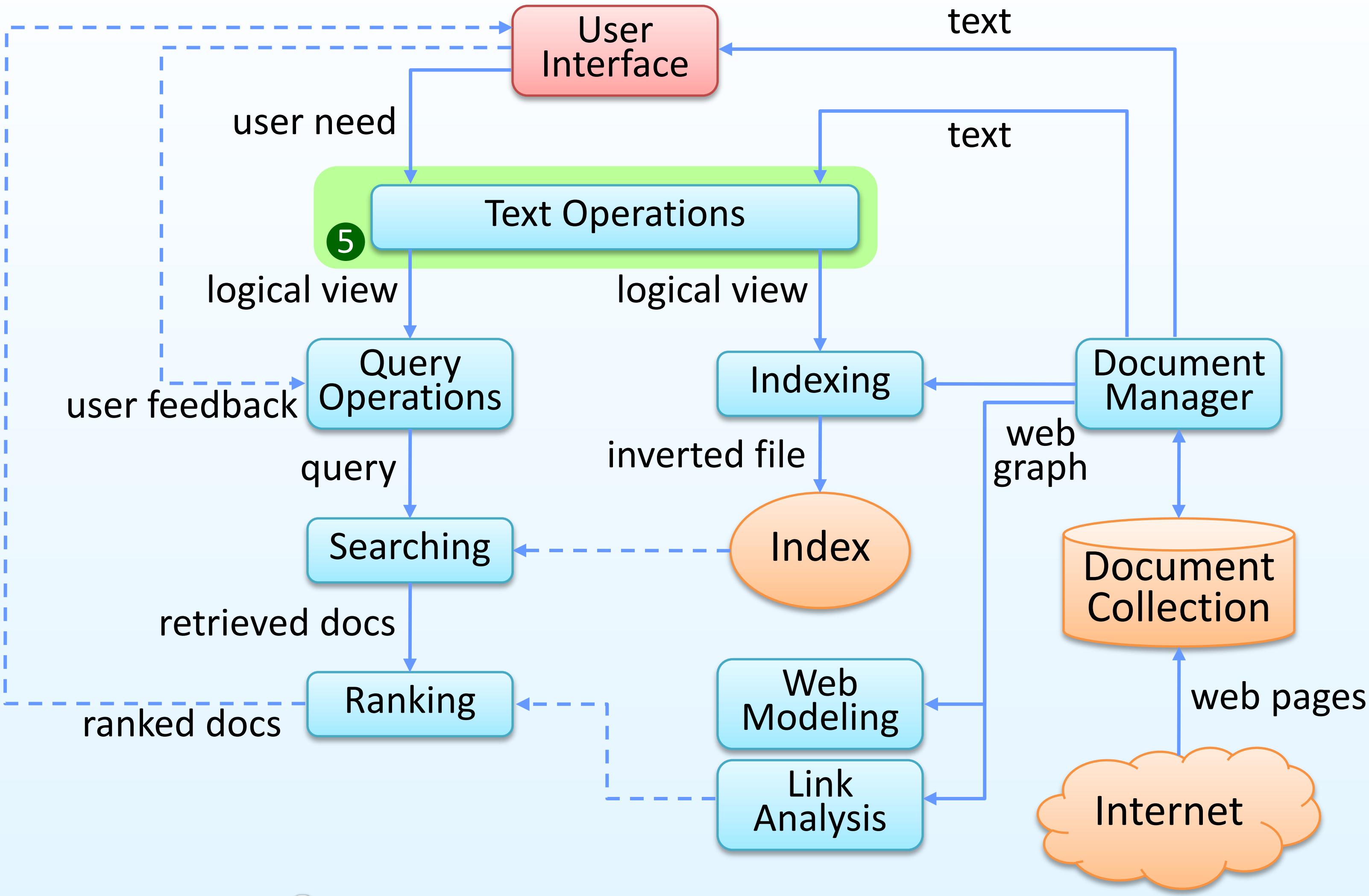
Department of Computer Engineering Kasetsart University

MIKE LABORATORY

Outline

- Text Properties
- Document Preprocessing

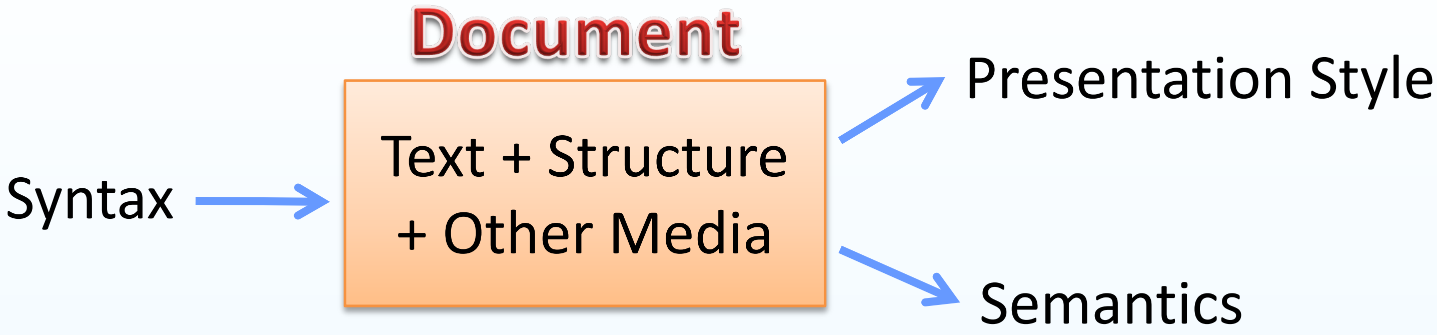
Review: Search Engine Architecture



The Document

- Denote a single unit of information
- Have a syntax and structure
- Have a semantics, specified by the author
- May have a presentation style
  - Given by its syntax and structure
  - Specify how to display or print
  - Related to a specific application

The Document



- Document syntax
  - Express structure, presentation style, and semantics
  - One or more of these elements might be implicit in the text or given together.
    - For example, structural element (e.g., a section) can have fixed formatting style.

Queries in search engines

- Differ from normal text
- Can be considered as short pieces of text
- Semantics often ambiguous due to polysemy
- Not simple to infer user intent behind a query
- Understanding them is very important

Text Properties

Modeling Natural Language

- Several issues were considered:
  - How the symbols are distributed over text
  - How the different words are distributed inside each document
  - How many the number of distinct words in a document is
  - How many the average length of words is

Distribution of Symbols

- Text is composed of symbols from a finite alphabet Σ.
- The symbols can be divided in two disjoint subsets:
  - Symbols that separated words (separators)
  - Symbols that belong to words

- Obviously, symbols are not uniformly distributed in a text
  - e.g., in English, the vowels are usually more frequent than most consonants.

Distribution of Symbols

- A simple model to generate text is the Binomial model.

$$F(\sigma,k) = \binom{\sigma}{k} p^{k} (1-p)^{\sigma-k}$$

  - However, probability of a symbol depends on previous symbols
    - e.g., in English, the letter ‘f’ cannot appear after the letter ‘c’.

- We can use a finite-context or Markov model to reflect this dependency.
  - The model can consider one or more letters to generate the next symbol.

- More complex models include finite-state models, and grammar models.
  - However, finding the right grammar is still a difficult problem.

Distribution of Word Frequencies

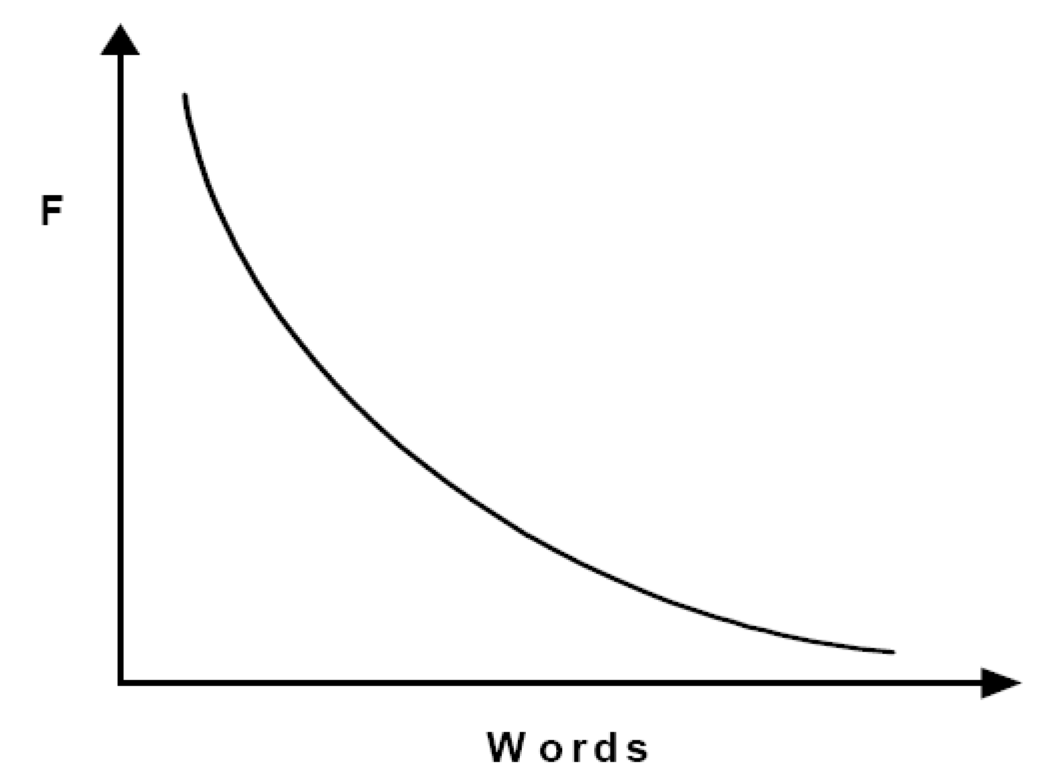
- An approximate model is the Zipf's Law.
  - This law states that the frequency f_i of the i-th most frequent word is given by

$$f_i = \frac{f_1}{i^\alpha}$$

where f_1 is the frequency of the most frequent word and α is a text dependent parameter.

Distribution of Word Frequencies

- Figure below illustrates the distribution of frequencies of the terms in a text.
  - Word arranged in decreasing order of their frequencies



Distribution of Word Frequencies

- For a text of n words with a vocabulary of V words, we have

$$n = \Sigma_{i=1}^{V} \frac{1}{i^{\alpha}} f_1 = f_1 \times \left(\Sigma_{i=1}^{V} \frac{1}{i^{\alpha}} \right)$$

- The factor enclosed in brackets depends only on the text parameters α and V.

- Let $H_V(\alpha)$ be the harmonic number of order α of V

$$H_V(\alpha) = \Sigma_{i=1}^{V} \frac{1}{i^{\alpha}}$$

- Then,

$$f_1 = \frac{n}{H_V(\alpha)}$$

Distribution of Word Frequencies

- Since the distribution of words is very skewed, words that are too frequent (called stopwords) can be disregarded.

- A stopword is a word which does not carry meaning in natural language (or low discrimination power).
  - e.g., “a”, “the”, “by”, “and”, etc.
  - Fortunately, the most frequent words are stopwords.
    - Therefore, half of the words appearing in a text do not need to be considered.

The Vocabulary Size

- Finding the number of distinct words in a document

- To predict the growth of the vocabulary size, we use the Heaps' Law.
  - The vocabulary of a text of n words is of size

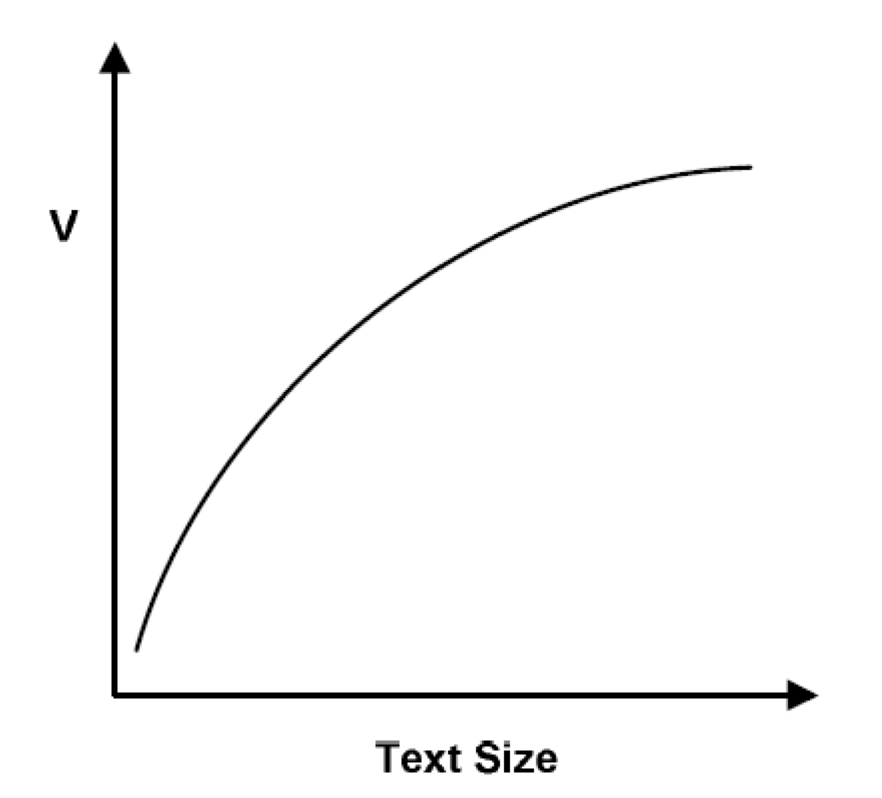
$$V = Kn^{\beta}$$

    where K and β depend on the text.

    - Usually, $10 \leq K \leq 100$ and $0 < \beta < 1$
    - In the TREC-2 collection, commonly $0.4 \leq \beta \leq 0.6$

The Vocabulary Size

- The figure below illustrates that vocabulary size grows sub-linearly with text size.



The Average Length of Words

- Relate with the text size of bytes

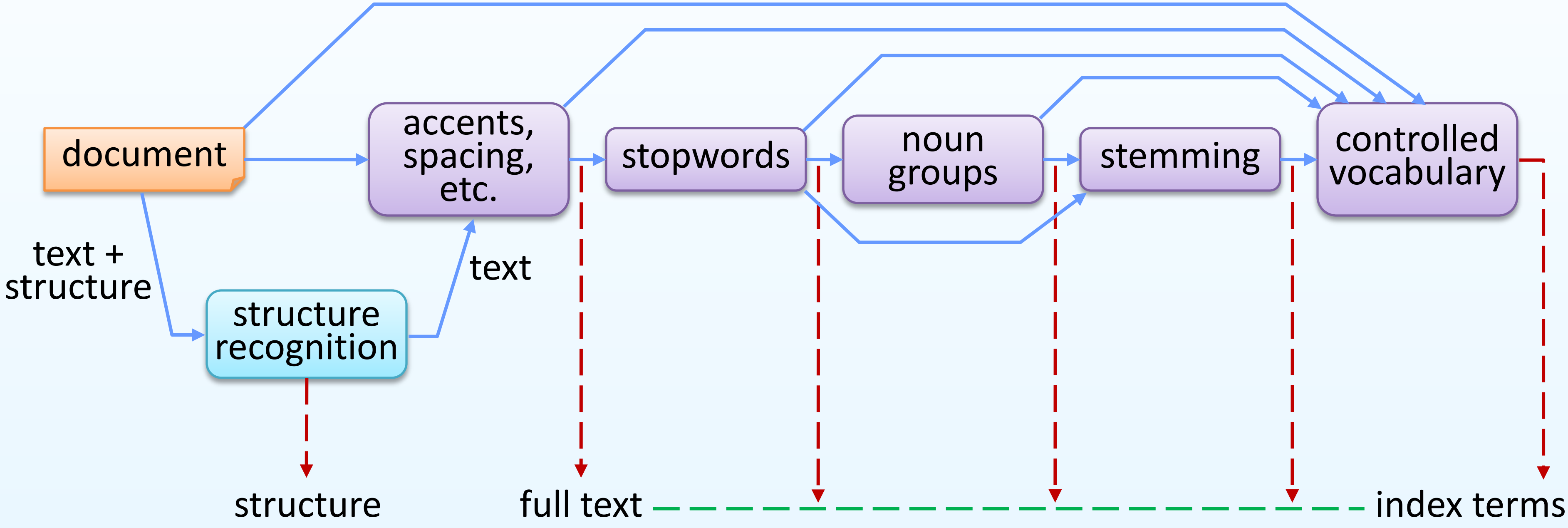
- In sub-collections of TREC-2, average length of words is very close to 5 letters.
- By removing the stopwords, the average length of words increase to a number between 6 and 7 letters.

Document Preprocessing

Document Preprocessing

- Can be divided into 5 text operations
  - Lexical analysis of the text
  - Elimination of stopwords
  - Stemming words
  - Selection of index terms or keywords
  - Construction of term categorization structures (thesaurus)

Logical View of a Document



Text operations reduce the complexity of documents

Lexical Analysis of the Text

- Convert stream of characters into stream of words
  - Major objective: indentify words in the text
- Word separators
  - Space: most common separator
  - Number: inherently vague, need context for disambiguation
    - e.g., "2012", "Euro2012", citizen ID, phone number, date, etc.
  - Hyphens: breaking up hyphens can cause inconsistent semantics
    - e.g., "relevant" and "non-relevant"
  - Punctuation marks: normally, not have an impact in performance
    - e.g., "510B.C.", but not for a case of "x.id" program variable
  - Case of the letters: usually, not important
    - e.g., "Thailand", but not for distinguishing "Lotus" and "lotus"

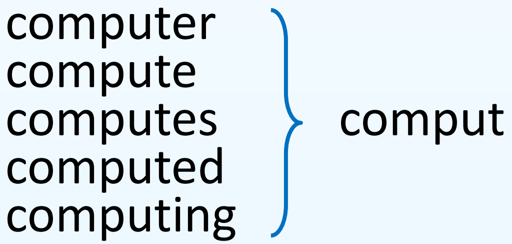
Elimination of Stopwords

- Stopwords
  - Words appearing too frequently
  - Words having very low discrimination power
  - Natural candidates: articles, prepositions, conjunctions

- Elimination of stopwords
  - Normally, filtered out as potential index terms
  - Effectively reduce size of index terms (by 40% or more)
  - Expense of reducing recall
    - e.g., not able to retrieve documents that contain “believe it or not”

Stemming

- In English, words can be in plural, gerund, or past tense suffix forms.
- Stemming is a process for reducing inflected words into their stem (or root form).
- Stemming reduces size of the index terms.
- Affix removal strategy, e.g., the Porter algorithm



computer
compute
computes
computed
computing

comput

  - http://tartarus.org/martin/PorterStemmer/
  - http://9ol.es/porter_js_demo.html
- There is controversy about benefits for retrieval.
- Many search engines do not adopt any stemming.

Keyword Selection

- Not all words in text used as index terms

- But, use
  - nouns—most concrete part of speech
  - noun groups—2 or 3 nouns in a single component
    - e.g., "computer science", "Tesco Lotus"

Thesaurus (Thesauri in plural)

- A treasury of words for reference
  - Precompiled list of important words in a knowledge domain
  - A set of related words derived from a synonymy relationship

- In general, a thesaurus includes a complex structure.
  - e.g., Peter Roget’s (general domain) thesaurus

cowardly *adjective*
Ignobly lacking in courage: *cowardly turncoats*.
Syns: chicken (slang), chicken-hearted, craven, dastardly, faint-hearted, gutless, lily-livered, pusillanimous, unmanly, yellow (slang), yellow-bellied (slang).

Thesaurus

- The main purposes of a thesaurus are to provide:
  - A standard vocabulary for indexing and searching
  - A mean to find terms for proper query (re)formulation
  - Classified hierarchies to allow broadening/narrowing queries

- Thesaurus as a controlled vocabulary for indexing and searching
  - Normalization of indexing concepts
  - Reduction of noise
  - Identification of indexing terms with a clear semantic meaning
  - Retrieval based on concepts rather than on words
  - However, only in specific domains of knowledge

Any Question?