STATISTICAL PROPERTIES OF TERMS IN IR

- The number of terms is the main factor in determining the size of the dictionary.
- Stemming reduces the number of distinct terms by 14%.
- How is the frequency of different words distributed?
- How fast does vocabulary size grow with the size of a corpus?
- Such factors affect the performance of IR.
- A few words are very common.
- Two most frequent words ("the", "of") can account for about 10% of words occurrences.

ZIPF'S LAW

- Zipf's law states that some corpus of natural language, the frequency of any word is inversely proportional to its rank.
- Named after the Harvard linguistic professor George Kingsely Zipf.
- Is used to understand how terms are distributed across documents.
- It states that if t_1 is the most common term in the collection, t_2 is the next most common and so on, and then the collection frequency cf_i of the i^{th} most common term is proportional to 1/i, (i.e. $cf_i \propto 1/i$).
- So the most frequent word, three times as often as the third frequent one.
- Example: in Brown corpus, "the" is the most frequently occurring word and by itself accounts for nearly 7% of all words occurrences (69971 out of 1 million).
- Second place word "of" accounts for slightly over 35% of word (36411).
- Third place is "and" (28852).
- The intuition is that frequency decreases very rapidly with rank.

DOCUMENT PROCESSING

- Can be divided into the following five operations:
  1. Lexical analysis (Morphological analysis)
  2. Elimination of stop words

3. Stemming
4. Selection of index term
5. Construction of term categorization structure such as Thesaurus

LEXICAL ANALYSIS OF THE TEXT

- Lexical analysis is the process of converting a stream of characters (the text of documents) into a stream of words (the candidate words to be adopted as index terms).
- Basically space is involved as word separator, but however the following cases also have to consider, (1) Digits, (2) Hyphens, (3) Punctuation marks and (4) Case of the letters.

1. Digits

- Numbers are usually not good index terms because without a surrounding context, they are inherently vague.
- For example: a user is interested in documents about the number of deaths due to car accidents between the year 1910 and 1989.
- Such a request could be specified as the set of index terms {deaths, car accidents, year, 1910, 1989}.
- However the presence of the numbers 1910 and 1989 in the query could lead to retrieval a variety of documents which refer to either o these two years.
- Thus in general, numbers are disregarded as index terms.
- But numbers like 510 B.C., sequence of 16 digits verifying a credit card number might be index term.

2. Hyphens

- Pose another difficult decision to the lexical analyzer.
- Breaking up hyphenated words might be useful due to inconsistency of usage.
- For example: "state-of-the-art" and "state of the art" are identical.
- But there are words which includes hyphen as an integral part.
- For example: co-education, B-49, etc.

3. Punctuation Marks
   - Normally, punctuation marks are removed entirely in the process of lexical analysis, while some punctuation marks are integral part of the world.
   - For example: Dr., B.C., etc.

4. Case of letters
   - The case of letters is usually not important for the identification of index terms.
   - As a result, the lexical analyzer normally converts all the text to either lower or upper case.
   - But, it may not work all the time. For example: the words "Bank" and "bank" have different meaning. UNIX commands are in uppercase.

INDEX TERM SELECTION

- If a full text representation of the text is adopted then all words in the text are used as index terms.
- The alternative is not all words are used as index terms.
- This implies that the set of terms used as indices must be selected.
- In the area of bibliographic sciences, such a selection of index terms is usually done by a specialist.
- A good approach is the identification of noun groups.
- A sentence in natural language text is usually composed of nouns, pronouns, articles, verbs, adjectives, adverbs and connectives.
- Most of the semantics is carried by the noun words.
- So it is like to use the noun as index terms.
- Also, the combination of noun ("Computer Science") can also be used as index.
- A noun group is a set of nouns whose syntactic distance in the text does not exceed a predefined threshold (for example: 3).

THESAURI

- A thesaurus is a collection of words with its synonyms and related words.

- It consists of:
  1. *A precompiled list of important words in a given domain knowledge.*
  2. *For each word in the list, a set of related words.*
- Thesaurus provides a standard vocabulary for indexing and searching.
- The terms are the indexing components of the thesaurus.
- Terms are basically noun.
- Thesaurus also contain phrase if a single word is unable to express semantic meaning. For example: "ballistic missiles".
- Basically, the terms are used in plural form, since the thesaurus represents class.
- Sometimes it is need to specify the precise meaning of a context in a particular thesaurus. For example: "seal" has different meaning in context of "documents" and "marine animals".

METADATA

- Information about a document that may not be a part of the document itself, i.e. data about data.
  1. *Descriptive metadata*
  2. *Semantic metadata*
- Descriptive metadata is external to the meaning of the document.
- For example: author, title, source, date, ISBN, length, etc.
- Semantic metadata concerns the content (semantic meaning).
- For example: abstract, keywords, etc.

WEB METADATA

- Meta tag in HTML.
- For example: <meta name = "keywords" content = "pets, cats, dogs">.