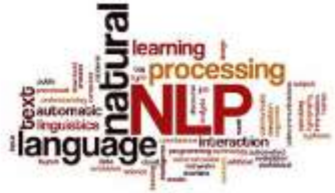
Natural Language Processing

Text Operations



Language Corpus

- A **corpus** is a body of text that has been collected for some purpose.
- A **balanced corpus** contains texts which represent different sources
- Examples
  - Brown corpus - written American English - compiled in the 60's
  - Lancaster-Oslo-Bergen (LOB) - written British English corpus - compiled in the 70’s
  - Both are about 1 million words
  - British National Corpus (BNC) contains approx 100 million words and includes 20 million words of spoken English.
  - The International Corpus of English (ICE) - a collection of 1,000,000 word corpora from each country or region where English is spoken as a first language. The corpus consists of a written and a spoken component.

Corpora

Corpora are

- important for many types of linguistic research
- essential for most modern NLP research

The compilation of text can be analysed in various ways to establish patterns of grammar and vocabulary usage.

Operations on Text

- look at the documents from a very high level
- Total Number of Words
- Total Number of Unique Words - Vocabulary count
- Total Number of Stop Words
- Total Number of Key Words

Tokenization

- Identify
- Paragraphs
- Sentences
- Words
- Frequency of words

Pattern Identification

Particular pattern that occurs a few times in this document

- How many times certain words occur together
- How many times two word that occur always together
- Three words that always occur together.
- To look at a particular pattern that occurs a few times in this document.