Information Retrieval

What is it?

Information retrieval system should satisfy users' information needs. The IR system seeks to locate documents relevant to this information.

A standard text IR system tends to do this based on relationship between the contents of the users search request and each potentially relevant document, which is available in the document collection.

Components of an IR System

- **Document Collection**
    - Published or professional documents - reports, articles
    - Web documents
    - Social media content

  Documents need to be processed into a standard format, remove or HTML or any other markup
- **Document Indexing**
    - Convert document into a file structure for rapid access - **inverted file**
- **Search Request**
    - Search request expressing their information need
- **Document Searching**
    - calculate set of potentially relevant documents and return to user
    - usually ranked by some score indicating likelihood of relevance
- **Relevance Feedback**
    - Modify search e.g. expand query using extra search items, based on relevance data

The Nature of Text

- Text is composed of word tokens taken from a surprisingly small vocabulary - 10GB of web documents typically has around 160k unique words
- Word morphology is changed as words are combined into phrases and further combined into sentences when used in natural language - **morphology** is the study of words, how they are formed
- There is no way of enforcing the use of the accepted grammar of the language by the authors documents
- There will be a number of spelling mistakes, even in published texts
- Large documents may be organised with structural elements
- Documents may be linked via hypertext

Variations in Text Style

- **Technical documentation -** complex sentences and phrases because topic is complex
- **Journalism -** newspapers articles, short and simple to read
- **Storybook prose -** can be complex, but makes it difficult to read
- **Email messages -** ungrammatical - c u l8r
- **Office memos -** grammatically correct, but not as complex as technical documentation
- **Formal language -** wills, deeds, legal documentation
- **Web pages -** variation of all of the above
- **Blogs, microblogs -** informal varied style, very terse short statements

Written vs Spoken Text

- Structure and content of written text is much different from manually written content
- Consider the difference between your own speaking and writing language

Text preprocessing methods

- The effectiveness and reliability of IR system can be improved by careful content preprocessing
- Process of bringing your text into a form that that is analyzable and predictable
- Pre processing usually should be automatic, as there are large amounts of data
    1. Lowercasing - most basic and often overlooked
    2. Stemming - process of reducing infection in words to their root word, e.g. troubled, troubles to trouble - **Porter's Algorithm**
    3. Lemmatization - similar to stemming, where the goal is to remove infectious words and map a word to its root form. Does it proper way, doesn't chop things off
    4. Stopword Removal - removal of 'is','are','the
    5. Normalization - transforming text into normal form - gud, goooood to good
    6. Noise removal - removal of characters and digits that can interfere with your text analysis, Hello to Hello
    7. Text enrichment / augmentation -

Automatic indexing

- Documents can only be accessed with information stored about them
- The most simple would be to store documents contents and search these for matches with words in the search request
- Identical words might instantiated differently in requests and documents
    - Synonyms
    - Different word form
    - Where this occurs, a simple match between words will fail
- Semantic analysis of the documents and requests to interpret their meaning and then matching them somehow could address this problem
- The search requirements of an IR system are typically underspecified - the users request does not give us a full piece description of what the user is looking for.
- IR system needs to be designed as far as possibly reliable to identify potentially relevant documents items within a broad scope of possible relevance

- Exception to this is when the search request is an actual question, i.e. looking for a specific piece of information - ‘What is the capital of Ireland?’

Tokenization

This is the process of chopping up and processing text strings into units that can form the basis of indexing units for retrieval.

- Token is a sequence of characters in a document that are grouped together into a meaningful semantic unit.
- Indexing unit may be a real ‘word’, but it may not, e.g. it may be a phrase or a part of a word.

Information Retrieval for Other Languages

Most IR techniques were developed for English, since the majority of electronic texts were in English.