Introduction

Text mining is the process of deriving high quality information from text. More specifically, large amounts of documents are to be automatically parsed and analyzed in order to extract the content. With an emphasis on patterns and “machine learning” for the bulk of the work, text mining saves time and resources when dealing with large amounts of data.

The aim of this collaborative research project is to develop sensible 2-3 word topic labels from a corpus. The program extracts common doubles and triples, which will be referred to as a “combo” or as a “candidate topic label”. A scoring system is developed based on the combo’s frequency and the relevancy of each word in the combo. More specifically, the scoring system takes into account part of speech, stemmed words, discrete statistics and predefined topics. Utilizing a specified technique, the scoring system has been implemented using the Perl, Shell Scripting, and Matlab(R) programming languages. The expected result is an effective summary of each topic chosen from the list candidate topic label. The results are then compared to mesh labels produced by PubMed® (http://www.ncbi.nlm.nih.gov/pubmed/).

Overview

PubMed® publishes a large collection of biomedical articles. Articles from this database are used in this research project in an attempt to create a working algorithm for discerning 2-3 word topic labels.

I: Creating the Topic Model

Each document is assigned a PubMed® ID (pmid.txt). For each collection of documents, the articles were previously analyzed and compiled to produce:

- sumwp.txt: count of the frequency of each word in the topic
- topic.$i.txt: list of top 200 documents most related to topic $i
- titles.txt: a list of the titles of each document
- vocab.txt: list of the words in the topic
- stopwords.txt: list of common stopwords

II: The Algorithm

Using the topic model information, the scoring algorithm picks the best candidate topic labels based on the $Pr(word | topic)$ and the frequency of the combo in the top 200 titles of the topic. The program go.sh scores all the topics for a given category.