CS 510 Pre-project Report: Development Track

By: Keval Morabia, Malcolm Tivelius, Tara Vijaykumar, Shubham Singhal
morabia2 mct6 tgv2 ss77

Literature Search Engine System

Data

The data used for the development and testing of the system is the ACL Anthology UIUC Corpus that has been pre-extracted using the Grobid toolkit. This dataset is a collection of scientific reports and we have chosen to focus on the title, abstract and introduction of said reports for this pre-project assignment.

Pre-processing

For pre-processing we rely heavily on the NLTK library. Using this library we process the titles, abstracts and introductions by first converting them to lowercase, tokenizing them into individual words, and then remove stop words using NLTKs english stop word corpus. Finally we utilize NLTK to perform stemming (Porter stemmer) and lemmatization (WordNet lemmatizer). The final output is a list of tokenized words for each document in the collection. We process the user input query the same way.

Search Algorithm

The search algorithm used in this pre-project exploration is Rank-BM25. The algorithm is imported from https://pypi.org/project/rank-bm25/, which is implemented according to the scientific paper: http://www.cs.otago.ac.nz/homepages/andrew/papers/2014-2.pdf
Specifically, we use the BM25+ variant proposed by Lv & Zhai (Lv, Y., C. Zhai, Lower-bounding term frequency normaliza- tion, *CIKM 2011*, p. 7-16.)

User Interface

The backend of the user interface is built in Python3 using the Flask library and the frontend is built using HTML and JavaScript. It allows the user to type in a query and receive matching documents sorted in descending order of relevance. The user can give feedback as to which documents were relevant to the query, and which were not. These relevance judgements are stored in a file on the server in the form of a tuple *(query, document ID, Relevant/Non-Relevant)*.

Demo

The working demo of this model is hosted at: ss77.web.illinois.edu/cs510_pre_project

Link to Github Repository: github.com/kevalmorabia97/ACL-Search-Engine-CS510