Capstone Project 2, Milestone Report:

Academic Paper Recommender

Overview:

The aim of this project is to create a recommender system to help researchers and scientists find related articles to a specific paper. There are thousands of papers published everyday and it is not feasible for a researcher to browse all these papers in order to find a related paper to a desired subject. This recommender system helps scientists with their search and saves them a lot of time looking into article resources.

Dataset:

The data for this project is collected using API for PubMed, which is a repository for biomedical data. The data is extracted in raw XML in a full text format and information such as paper's ID, title, authors last name, year, journal, abstract, tags, and citations is collected from each paper. This is programmatically queried via the NCBI Entrez E-utilities interface.

Data Scraping and Data Cleaning:

The code for data scraping can be found in getData.ipynb. 10000 papers up to the PMID of 27000000 is extracted and data for PMID, title, authors, publication year, name of journal, abstract, tags, and citations is used in a dataframe. Since the search is going to be based on title and/or abstract and also tags, the papers with missing title and tags were removed from the data. The data is saved as a .csv file and is uploaded in Databricks, which is a web-based platform for working with Spark, and provides automated cluster management and IPython-style notebooks.

Modelling Approach:

Different methods and information is used to come up with a recommender for a paper. The first approach is text similarity-based recommender. In this method, title and abstract of each paper is used and a TF-IDF based similarity is calculated in order to recommend the n number of related papers. The other method is based on the tags used for each paper. Papers which are sharing the most number of tags with the reference paper would be recommended.