CSCE 670

Implement the Boolean Retrieval engine

1. Tokenize entities and definitions using whitespaces and punctuations as delimiters
2. Remove stop words and stemming with **nltk** packages
3. Build an inverted index to support Boolean Retrieval
4. Rank the documents with sum of **TF-IDF** scores, vector space model with TF-IDF, and BM25 respectively

Find significant Twitter users

1. Build a re-tweet graph by parsing the tweets in the dataset
2. Implement **PageRank** algorithm to find the top ten "impactful" users with highest scores

Recommender System

1. Implement **Matrix Factorization** to predict ratings on MovieLens dataset, evaluate the model by computing the MAE and RMSE value on the testing dataset
2. Use a **BPR** package to experiment with **top-K** item recommendation on a Spotify playlist recommendation dataset, evaluate the results with **NDCG**

Word Embeddings for Information Retrieval and Query Expansion

1. Use the **Word2Vec** algorithm to generate **word embeddings** for tokens in the dataset
2. Match the query and the document via the cosine similarity between the embeddings of them
3. Expand the original query and redo the vector space model via word embeddings