Cmpe 493 Introduction to Information Retrieval, Fall 2021
Assignment 2 - A Simple Search System for Phrase and Free Text Queries
Zuhal Didem Aytaç – 2018400045

1. Data Preprocessing

I created a file preprocess.py. This code handles the initial processing of data. It reads the reutx-xxx.sgm files and the stopwords.txt file. It extracts the necessary content from the sgm files. Then with the helper functions, it performs case-fold, punctuation removal and stop-word removal operations on the extracted texts. It keeps the normalized, tokenized text using a dictionary, with new_id as the key.

This preprocess module is not run by itself, it is called by the indexing module.

2. The Inverted Index

The indexize module calls the preprocess module described in part 1. The preprocess module returns a dictionary (*normalized*) with new_id as key and list of tokens as value. The indexize module traverses that dictionary (*normalized*) and builds two new dictionaries. One is *inverted_intex* with token as keys and list of new_ids as value. The other is *document_index* with new_id as keys and {token:frequency} dictionary as value.

After that, the *build_frequency_index* function is called with *inverted_intex* dictionary. This function builds the dictionary *index_with_frequencies*. It stores for each token, the document frequency and the term frequencies and positions for each news article containing the token. The dictionary is as follows:

```
{ token:
   {
   'df': document frequency,
   'documents': { new_id: { 'tf': term frequency, 'positions': list of positions } }
   }
}
```

Finally, the *index_with_frequencies* and *document_index* dictionares are dumped to json files (*index.json* and *documents.json*).

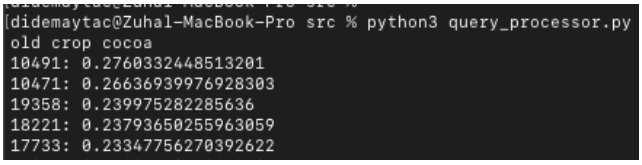
3. The Indexing Module

```
didemaytac@Zuhal-MacBook-Pro src % python3 indexize.py
```

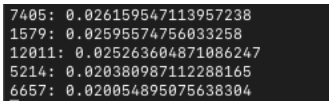
4. Query Processing

Type python3 query_processor.py, then type the query and press enter. The output will be shown at the command line. If no results are found, nothing is printed.

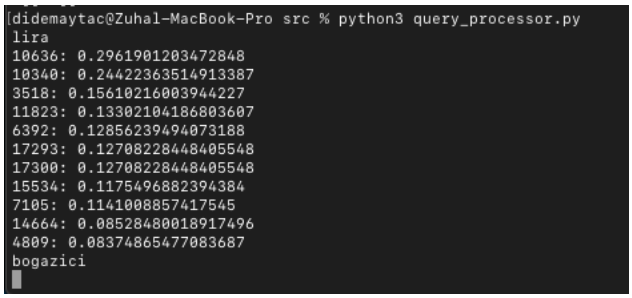
Free text query old crop cocoa (only first and last 5 results are screenshotted):



…



Other free text query examples (lira) (bogazici) :



Phrase query examples (“old crop cocoa”, “bogazici university”, “turkish lira”, “lira”):