Muhammed Göktepe
2017400162

Assignment 1

Describe the steps you have performed for data preprocessing.

In this project we are assigned to implement a document retrieval system for simple boolean queries using the non-positional inverted indexing scheme. Reuters news is our dataset, and we have some stopwords to discard. First, I stored stopwords in a list for not to include these words while indexing texts. I have a dictionary named index_dict. I store newId's of each word in here. Then I started to traverse .sgm files in reuters21578 directory. I used BeautifulSoup library to ease .sgm file processing. Then, I store all articles in a file in news_list list. I iterate over each news and look for their titles and body's. If the article contains either of these, I combine the title and body and send it to the normalize function. In normalize function I made punctuation removal, case folding and finally stopwords removal. It also prevents duplicate entries in text. After these steps I added these words and their ids to dictionary and printed this dictionary to file.

Describe the data structures (hash, b-tree, linked list etc.) that you used for representing the inverted index (i.e., the dictionary and the postings lists).

I used list while storing posting lists of words. Since .sgm files read in ascending order of news ids it stores new ids in ascending order. While implementing AND operator I used pseudocode in Lec01. I traversed two list by comparing news ids and make necessary addition to result.

Provide a screenshot of running the indexing module of your system.

Since I am using BeautifulSoup library it must be installed on working environment.

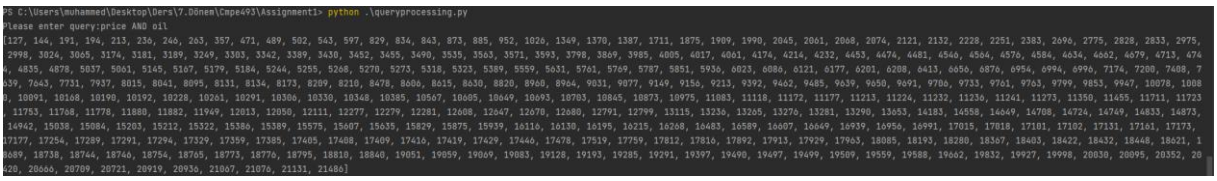
```
PS C:\Users\muhammed\Desktop\Ders\7.Dönem\Cmpe493\Assignment1> pip install bs4
WARNING: Ignoring invalid distribution -ip (c:\python310\lib\site-packages)
WARNING: Ignoring invalid distribution - (c:\python310\lib\site-packages)
WARNING: Ignoring invalid distribution -ip (c:\python310\lib\site-packages)
WARNING: Ignoring invalid distribution - (c:\python310\lib\site-packages)
Requirement already satisfied: bs4 in c:\python310\lib\site-packages (0.0.1)
Requirement already satisfied: beautifulsoup4 in c:\python310\lib\site-packages (from bs4) (4.10.0)
Requirement already satisfied: soupsieve>1.2 in c:\python310\lib\site-packages (from beautifulsoup4->bs4) (2.3.1)
WARNING: Ignoring invalid distribution -ip (c:\python310\lib\site-packages)
WARNING: Ignoring invalid distribution - (c:\python310\lib\site-packages)
WARNING: Ignoring invalid distribution -ip (c:\python310\lib\site-packages)
WARNING: Ignoring invalid distribution - (c:\python310\lib\site-packages)
WARNING: Ignoring invalid distribution -ip (c:\python310\lib\site-packages)
WARNING: Ignoring invalid distribution - (c:\python310\lib\site-packages)
WARNING: Ignoring invalid distribution -ip (c:\python310\lib\site-packages)
WARNING: Ignoring invalid distribution - (c:\python310\lib\site-packages)
WARNING: You are using pip version 21.2.3; however, version 21.3.1 is available.
You should consider upgrading via the 'C:\Python310\python.exe -m pip install --upgrade pip' command.
PS C:\Users\muhammed\Desktop\Ders\7.Dönem\Cmpe493\Assignment1> python .\preprocessing.py
```

Muhammed Göktepe
2017400162

You can install it by typing “pip install bs4”. Then you can run my preprocessing code by typing “python preprocessing.py”.

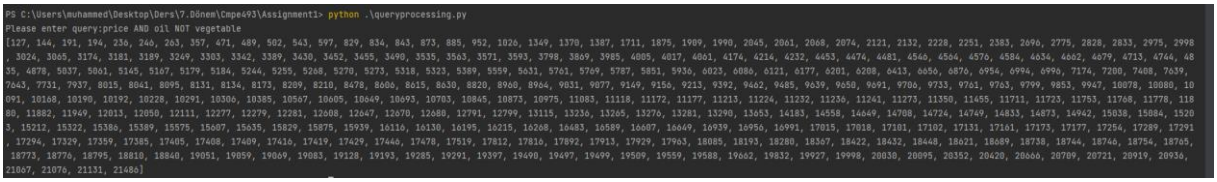
Provide four screenshots of running your system for each of the four types of queries.

price AND oil



petroleum OR oil OR gas

price AND oil NOT vegetable



Muhammed Göktepe
2017400162

petroleum OR oil NOT price