# Assignment 2

## 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 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 of dictionary named index\_dict. I store newId’s of each word and their corresponding positions in documents 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. After these steps I added these words to dictionary and create a dictionary for each word. For these words I added their news ids and their positions in the document. 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 dictionary of dictionary to store words, news ids and word positions in documents. Since .sgm files read in ascending order of news ids it stores new ids in ascending order. Also, positions are in ascending order in document. I used AND operator like algorithm while implementing phrase queries. In text free queries I used Lec03 as my reference.

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

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

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

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.

“old crop cocoa”

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

"sugar price"

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

"leverage position"

metin, iç mekan, ekran görüntüsü, kapat içeren bir resim

Açıklama otomatik olarak oluşturuldu

"United States District"

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

cocoa export shipment tonne

metin, açık hava içeren bir resim

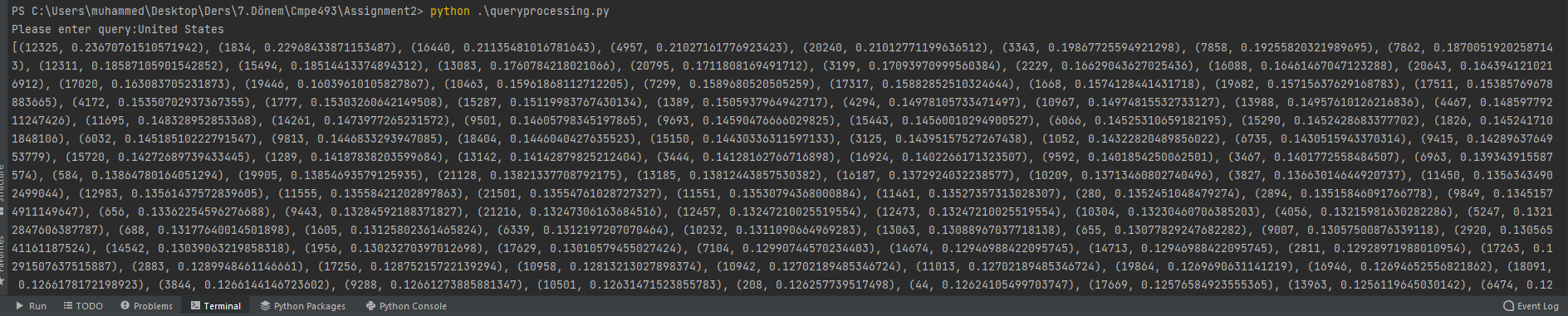
Açıklama otomatik olarak oluşturuldu

Oil price

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

United States



payments market

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu