CMPE493 – Information Retrieval – Assignment2

Meltem Arslan
2016400117

(i) Preprocessing & getting wordsDict

wordsDict is the name of dictionary to hold inverted index. Its structure will be explained in the second part.

First, I read all files in the current directory ending with '.sgm' and collected into 'files' set.
According to the NEWIDs in each .sgm file, 'fileIDs' list is created which consists of all NEWIDs in the whole dataset.
With fields, I got the articles in 'articles' list with 'getArticles()' function.
With the 'getCleanedArticle()' function called in 'getArticles()' function, all articles cleaned by sending 'clean_data()' function one by one.
In 'clean_data()' function:

1. Text characters converted into lowercase characters.
2. Punctuation removed.
3. Stop words removed.

Since I am visiting each word in the document one by one and know its index and documents while doing this, I have filled the dictionary by cleaning the data to save time.

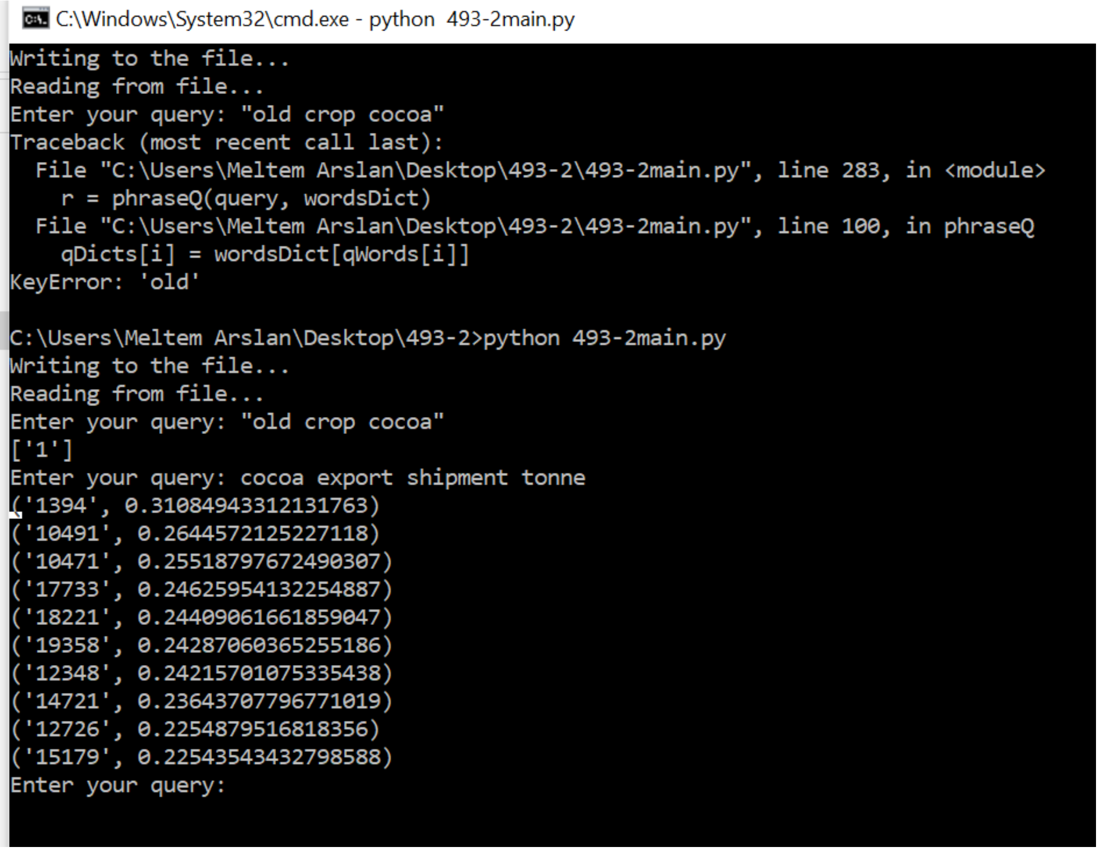
(ii) Data structure used for representing the inverted index

'wordsDict' dictionary is used to store inverted index. The structure is:

```
{'word1':
        {'doc1':[index1, index2, …],
         'doc2':[index1, index2, …],
          …
        },
 'word2':
        {'doc5':[index1, index2, …],
         'doc7':[index1, index2, …],
          …
        },
 'word3':
        {'doc5':[index1, index2, …],
         'doc7':[index1, index2, …],
          …
```

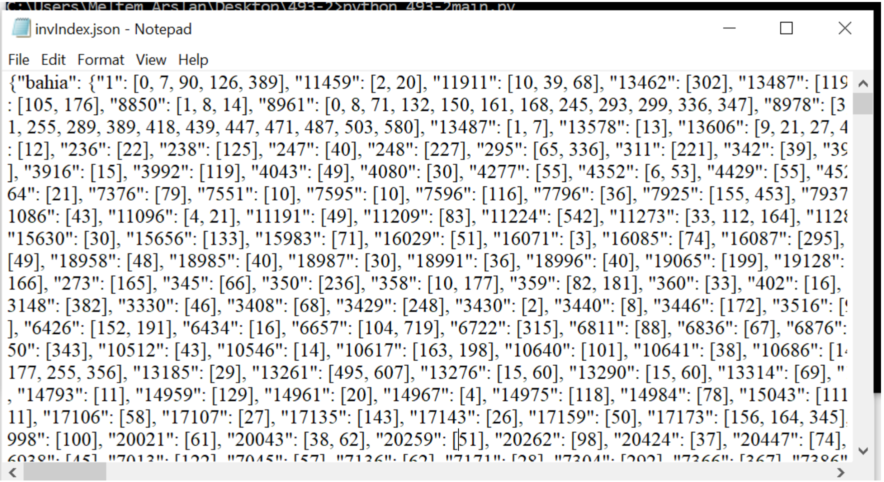
}

(iii) Provide a screenshot of running the indexing module of your system.



C:\Windows\System32\cmd.exe - python 493-2main.py

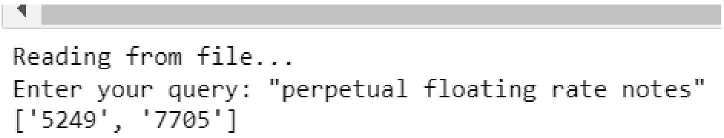
```
Writing to the file...
Reading from file...
Enter your query: "old crop cocoa"
Traceback (most recent call last):
  File "C:\Users\Meltem Arslan\Desktop\493-2\493-2main.py", line 283, in <module>
    r = phraseQ(query, wordsDict)
  File "C:\Users\Meltem Arslan\Desktop\493-2\493-2main.py", line 100, in phraseQ
    qDicts[i] = wordsDict[qWords[i]]
KeyError: 'old'

C:\Users\Meltem Arslan\Desktop\493-2>python 493-2main.py
Writing to the file...
Reading from file...
Enter your query: "old crop cocoa"
['1']
Enter your query: cocoa export shipment tonne
('1394', 0.31084943312131763)
('10491', 0.2644572125227118)
('10471', 0.25518797672490307)
('17733', 0.24625954132254887)
('18221', 0.24409061661859047)
('19358', 0.24287060365255186)
('12348', 0.24215701075335438)
('14721', 0.23643707796771019)
('12726', 0.2254879516818356)
('15179', 0.22543543432798588)
Enter your query:

C:\Users\Meltem Arslan\Desktop\493-2>python 493-2main.py
```



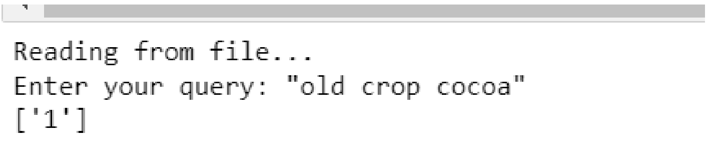
invIndex.json - Notepad

```
{"bahia": {"1": [0, 7, 90, 126, 389], "11459": [2, 20], "11911": [10, 39, 68], "13462": [302], "13487": [119
: [105, 176], "8850": [1, 8, 14], "8961": [0, 8, 71, 132, 150, 161, 168, 245, 293, 299, 336, 347], "8978": [3
1, 255, 289, 389, 418, 439, 447, 471, 487, 503, 580], "13487": [1, 7], "13578": [13], "13606": [9, 21, 27, 4
: [12], "236": [22], "238": [125], "247": [40], "248": [227], "295": [65, 336], "311": [221], "342": [39], "39
], "3916": [15], "3992": [119], "4043": [49], "4080": [30], "4277": [55], "4352": [6, 53], "4429": [55], "45
64": [21], "7376": [79], "7551": [10], "7595": [10], "7596": [116], "7796": [36], "7925": [155, 453], "7937
1086": [43], "11096": [4, 21], "11191": [49], "11209": [83], "11224": [542], "11273": [33, 112, 164], "1128
"15630": [30], "15656": [133], "15983": [71], "16029": [51], "16071": [3], "16085": [74], "16087": [295],
[49], "18958": [48], "18985": [40], "18987": [30], "18991": [36], "18996": [40], "19065": [199], "19128":
166], "273": [165], "345": [66], "350": [236], "358": [10, 177], "359": [82, 181], "360": [33], "402": [16],
3148": [382], "3330": [46], "3408": [68], "3429": [248], "3430": [2], "3440": [8], "3446": [172], "3516": [
], "6426": [152, 191], "6434": [16], "6657": [104, 719], "6722": [315], "6811": [88], "6836": [67], "6876":
50": [343], "10512": [43], "10546": [14], "10617": [163, 198], "10640": [101], "10641": [38], "10686": [1
177, 255, 356], "13185": [29], "13261": [495, 607], "13276": [15, 60], "13290": [15, 60], "13314": [69], "
, "14793": [11], "14959": [129], "14961": [20], "14967": [4], "14975": [118], "14984": [78], "15043": [111
11], "17106": [58], "17107": [27], "17135": [143], "17143": [26], "17159": [50], "17173": [156, 164, 345]
998": [100], "20021": [61], "20043": [38, 62], "20259": [51], "20262": [98], "20424": [37], "20447": [74],
6938": [45], "7013": [122], "7045": [57], "7136": [62], "7171": [28], "7304": [292], "7366": [367], "7386"
```

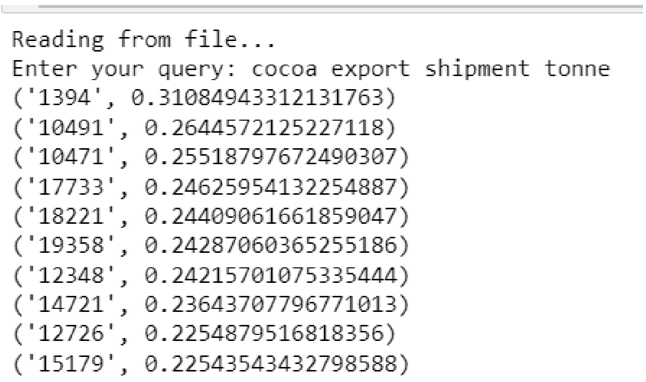
(iv) Provide four screenshots of running your system for each of the two types of queries



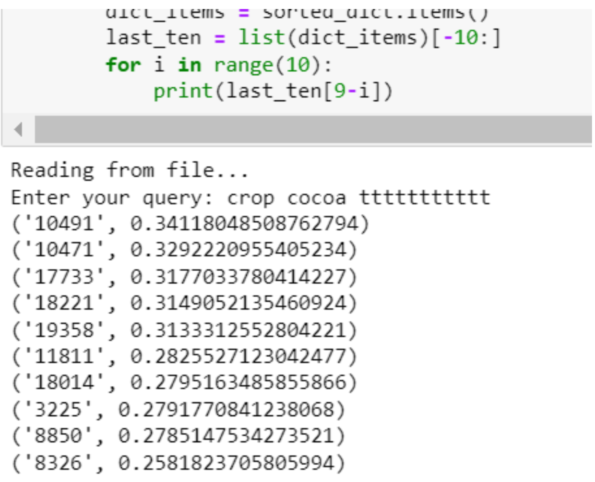
```
Reading from file...
Enter your query: "perpetual floating rate notes"
['5249', '7705']
```



```
Reading from file...
Enter your query: "old crop cocoa"
['1']
```



```
Reading from file...
Enter your query: cocoa export shipment tonne
('1394', 0.31084943312131763)
('10491', 0.2644572125227118)
('10471', 0.25518797672490307)
('17733', 0.24625954132254887)
('18221', 0.24409061661859047)
('19358', 0.24287060365255186)
('12348', 0.24215701075335444)
('14721', 0.23643707796771013)
('12726', 0.2254879516818356)
('15179', 0.22543543432798588)
```



```
    dict_items = sorted_dict.items()
    last_ten = list(dict_items)[-10:]
    for i in range(10):
        print(last_ten[9-i])
```

```
Reading from file...
Enter your query: crop cocoa tttttttttt
('10491', 0.34118048508762794)
('10471', 0.3292220955405234)
('17733', 0.3177033780414227)
('18221', 0.3149052135460924)
('19358', 0.3133312552804221)
('11811', 0.2825527123042477)
('18014', 0.2795163485855866)
('3225', 0.2791770841238068)
('8850', 0.2785147534273521)
('8326', 0.2581823705805994)
```