CMPE 493 IR ASSIGNMENT 1

Student : Salih Sevgican
Student No: 2013400219

(i) Describe the steps you have performed for data preprocessing and provide answers for the following questions.
(a) How many tokens does the corpus contain before stopword removal and stemming?
2902784
(b) How many tokens does the corpus contain after stopword removal and stemming?
2147367
(c) How many terms (unique tokens) are there before stopword removal, stemming, and case-folding?
129173
(d) How many terms (unique tokens) are there after stopword removal, stemming, and case-folding?
45278

(e) List the top 20 most frequent terms before stopword removal, stemming, and case-folding?

1. the 119584
2. of 72137
3. to 68413
4. and 53275
5. in 49814
6. a 48193
7. said 35721
8. for 25120
9. mln 24805
10. The 22828
11. &#3; 19043
12. on 17696
13. it 17497
14. is 16645
15. said. 15784
16. dlrs 15434
17. from 14919
18. that 14837
19. its 14715
20. vs 14599

(f) List the top 20 most frequent terms after stopword removal, stemming, and case-folding?

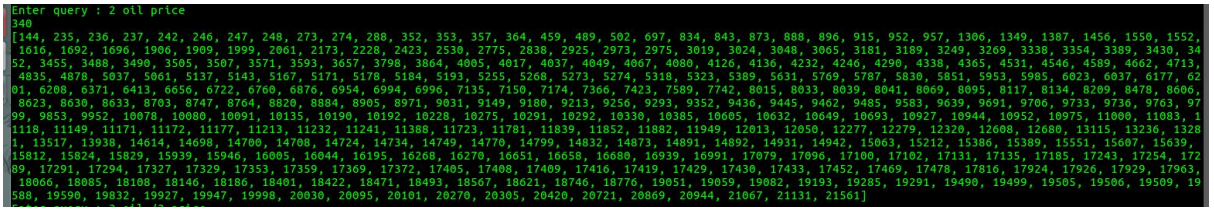
1. to 73074
2. said 53096
3. 3 26814
4. mln 26732
5. dlrs 21273
6. reuter 18964
7. pct 18046
8. lt 16680
9. 1 15895
10. from 15277
11. vs 14836
12. 000 13448
13. year 13109
14. billion 10726
15. has 10185
16. 2 9996
17. company 9699
18. cts 9219
19. would 9200
20. not 8308

(ii) Describe the data structures that you used for representing the positional index (dictionary and postings).
Dictionary is a list which contains dictionaries(Python data structure). List index represents docID, and dictionary holds tokens belong to the specified docID.Values of tokens are incremented one by one starting from zero.
Python represantation:
dictionary = [{ token0 : 0 , token1 : 1 , token2 : 2 } , { token3 : 3 , token4 : 4 , token5 : 5 }]

Positional Index is a dictionary(Python data structure) , keys are tokens and values are another dictionary, which holds docID as a key and positions of token in that docID as a list.
Python represantation:
posIndex = { token0 : {docID1 : [positions] , docID2 : [positions2] } , token1 : { docID1 : [positions] , docID5 : [positions] } }

(iii) Provide a screenshot of running your system for a conjunctive query.

(iv) Provide a screenshot of running your system for a phrase query.



(iv) Provide a screenshot of running your system for a proximity query.