

## 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 stopwords removal and stemming?  
2902784

(b) How many tokens does the corpus contain after stopwords removal and stemming?  
2147367

(c) How many terms (unique tokens) are there before stopwords removal, stemming, and case-folding?  
129173

(d) How many terms (unique tokens) are there after stopwords removal, stemming, and case-folding?  
45278

(e) List the top 20 most frequent terms before stopwords 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. mIn 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. It 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 representation:

```
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 representation:

```
posIndex = { token0 : {docID1 : [positions] , docID2 : [positions2] } , token1 : { docID1 : [positions] , docID5 : [positions] } }
```

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

```
Enter query : 1 oil AND price
1127, 144, 145, 191, 194, 208, 213, 235, 236, 237, 242, 246, 247, 248, 263, 273, 274, 288, 349, 352, 353, 357, 364, 370, 459, 471, 489, 502, 543, 597, 668, 697, 704, 829,
833, 834, 843, 873, 885, 888, 896, 915, 952, 957, 1026, 1211, 1306, 1349, 1370, 1379, 1387, 1456, 1558, 1552, 1616, 1692, 1696, 1711, 1799, 1848, 1875, 1906, 1909, 1990, 1
999, 2045, 2046, 2061, 2068, 2074, 2121, 2132, 2173, 2228, 2231, 2251, 2383, 2423, 2449, 2522, 2538, 2585, 2696, 2775, 2828, 2833, 2838, 2925, 2973, 2975, 2998, 3019, 3024
, 3048, 3065, 3174, 3181, 3189, 3249, 3269, 3303, 3338, 3342, 3354, 3364, 3389, 3430, 3452, 3455, 3466, 3488, 3490, 3505, 3507, 3535, 3540, 3563, 3571, 3593, 3597, 3657, 3
798, 3818, 3846, 3855, 3864, 3869, 3888, 3929, 3985, 4005, 4017, 4035, 4037, 4049, 4061, 4067, 4080, 4126, 4136, 4174, 4214, 4232, 4246, 4290, 4338, 4365, 4367, 4453, 4474
, 4481, 4531, 4546, 4547, 4564, 4569, 4576, 4584, 4589, 4590, 4593, 4634, 4662, 4679, 4713, 4744, 4835, 4870, 4963, 4981, 5037, 5061, 5137, 5143, 5145, 5156, 5167, 5171, 5
178, 5179, 5184, 5193, 5238, 5244, 5255, 5268, 5270, 5273, 5274, 5295, 5318, 5323, 5376, 5389, 5391, 5559, 5561, 5631, 5684, 5692, 5712, 5761, 5769, 5787, 5793, 5830, 5851
, 5936, 5953, 5985, 6023, 6037, 6054, 6086, 6119, 6121, 6169, 6177, 6201, 6208, 6253, 6264, 6348, 6371, 6413, 6432, 6606, 6656, 6722, 6740, 6742, 6760, 6876, 6922, 6954, 6
994, 6996, 7135, 7150, 7174, 7200, 7366, 7397, 7408, 7423, 7548, 7589, 7639, 7643, 7731, 7742, 7854, 7937, 8015, 8033, 8039, 8041, 8050, 8051, 8069, 8086, 8095, 8117, 8119
, 8131, 8134, 8149, 8156, 8173, 8209, 8210, 8478, 8596, 8598, 8606, 8610, 8615, 8623, 8630, 8633, 8672, 8703, 8747, 8764, 8820, 8882, 8884, 8905, 8959, 8960, 8964, 8971, 8
988, 9031, 9077, 9149, 9156, 9180, 9193, 9206, 9213, 9251, 9256, 9293, 9352, 9392, 9402, 9436, 9445, 9462, 9485, 9583, 9639, 9658, 9691, 9706, 9722, 9732, 9733, 9734, 9736
, 9742, 9761, 9763, 9769, 9799, 9853, 9933, 9947, 9952, 10015, 10078, 10080, 10091, 10133, 10135, 10168, 10190, 10192, 10228, 10261, 10268, 10272, 10275, 10291, 10292, 103
96, 10330, 10348, 10375, 10385, 10567, 10605, 10632, 10649, 10693, 10703, 10720, 10845, 10873, 10908, 10927, 10944, 10947, 10952, 10975, 11080, 11083, 11118, 11149, 11171,
11172, 11177, 11213, 11224, 11231, 11232, 11236, 11241, 11273, 11350, 11388, 11444, 11455, 11541, 11559, 11711, 11723, 11752, 11753, 11768, 11778, 11781, 11839, 11852, 11
880, 11882, 11949, 12013, 12050, 12111, 12277, 12279, 12281, 12320, 12361, 12608, 12647, 12670, 12680, 12791, 12799, 13115, 13236, 13256, 13265, 13266, 13276, 13281, 13290
, 13291, 13350, 13420, 13501, 13517, 13576, 13653, 13861, 13938, 13993, 14107, 14183, 14279, 14558, 14611, 14614, 14649, 14690, 14698, 14700, 14708, 14724, 14734, 14749, 1
4778, 14799, 14832, 14833, 14853, 14873, 14891, 14892, 14931, 14942, 15038, 15063, 15084, 15203, 15212, 15238, 15322, 15386, 15389, 15500, 15551, 15675, 15687, 15635, 1563
9, 15812, 15814, 15824, 15829, 15875, 15939, 15946, 16005, 16044, 16086, 16093, 16116, 16130, 16195, 16215, 16268, 16270, 16358, 16403, 16577, 16589, 16604, 16607, 16649,
16651, 16658, 16680, 16772, 16939, 16948, 16956, 16991, 17015, 17018, 17079, 17096, 17100, 17101, 17102, 17131, 17135, 17161, 17173, 17177, 17185, 17199, 17243, 17254, 172
89, 17291, 17294, 17327, 17329, 17353, 17359, 17369, 17372, 17385, 17386, 17389, 17405, 17408, 17409, 17416, 17419, 17429, 17430, 17433, 17441, 17446, 17452, 17463, 17469,
17477, 17478, 17519, 17708, 17759, 17771, 17812, 17816, 17875, 17878, 17883, 17892, 17913, 17949, 17926, 17929, 17963, 18066, 18085, 18108, 18146, 18150, 18186, 18193, 18
280, 18367, 18378, 18393, 18401, 18403, 18420, 18422, 18432, 18447, 18448, 18464, 18471, 18480, 18493, 18512, 18523, 18567, 18621, 18680, 18689, 18701, 18728, 18736, 18738
18744, 18746, 18754, 18765, 18773, 18776, 18795, 18810, 18846, 18996, 19017, 19038, 19051, 19059, 19069, 19082, 19083, 19091, 19110, 19128, 19151, 19192, 19285, 19291, 1
9397, 19478, 19490, 19497, 19499, 19505, 19506, 19509, 19534, 19551, 19556, 19559, 19588, 19590, 19662, 19832, 19867, 19927, 19947, 19998, 20008, 20030, 20095, 20101, 2027
9, 20385, 20352, 20389, 20406, 20420, 20461, 20462, 20526, 20666, 20709, 20721, 20869, 20919, 20936, 20944, 21067, 21076, 21131, 21152, 21267, 21363, 21486, 21525, 21561]
Enter query : 1
```

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

```
Enter query : 2 oil price
140
1144, 235, 236, 237, 242, 246, 247, 248, 273, 274, 288, 352, 353, 357, 364, 459, 489, 502, 697, 834, 843, 873, 888, 896, 915, 952, 957, 1306, 1349, 1387, 1456, 1550, 1552,
1616, 1692, 1696, 1906, 1909, 1999, 2061, 2173, 2228, 2423, 2530, 2775, 2838, 2925, 2973, 2975, 3019, 3024, 3048, 3065, 3181, 3189, 3249, 3269, 3338, 3354, 3389, 3430, 34
52, 3455, 3488, 3490, 3505, 3507, 3571, 3593, 3657, 3798, 3864, 4005, 4017, 4037, 4049, 4067, 4080, 4126, 4136, 4232, 4246, 4290, 4338, 4365, 4531, 4546, 4589, 4662, 4713,
4835, 4870, 5037, 5061, 5137, 5143, 5167, 5178, 5184, 5193, 5255, 5268, 5273, 5274, 5318, 5323, 5389, 5631, 5769, 5787, 5838, 5851, 5953, 5985, 6023, 6037, 6177, 62
61, 6280, 6371, 6413, 6656, 6722, 6760, 6876, 6954, 6994, 6996, 7135, 7150, 7174, 7366, 7423, 7589, 7742, 8015, 8033, 8039, 8041, 8069, 8095, 8117, 8134, 8209, 8478, 8606,
8623, 8630, 8633, 8703, 8747, 8764, 8820, 8884, 8905, 8971, 9031, 9149, 9180, 9213, 9256, 9293, 9352, 9436, 9445, 9462, 9485, 9583, 9639, 9691, 9706, 9732, 9736, 9763, 97
99, 9853, 9952, 10078, 10080, 10091, 10135, 10190, 10192, 10228, 10275, 10291, 10292, 10330, 10385, 10605, 10632, 10649, 10693, 10927, 10944, 10952, 10975, 11000, 11083, 11
118, 11149, 11171, 11172, 11213, 11232, 11241, 11388, 11723, 11781, 11839, 11852, 11882, 11949, 12013, 12050, 12277, 12279, 12320, 12608, 12680, 13115, 13236, 1328
1, 13517, 13938, 14614, 14698, 14700, 14708, 14724, 14734, 14749, 14770, 14799, 14832, 14873, 14891, 14892, 14931, 14942, 15063, 15212, 15386, 15389, 15551, 15607, 15639,
15812, 15824, 15829, 15939, 15946, 16005, 16044, 16195, 16268, 16270, 16651, 16658, 16680, 16939, 16991, 17079, 17096, 17100, 17102, 17131, 17135, 17185, 17243, 17254, 172
89, 17291, 17294, 17327, 17329, 17353, 17359, 17369, 17372, 17405, 17408, 17409, 17416, 17419, 17429, 17430, 17433, 17452, 17469, 17478, 17816, 17924, 17926, 17929, 17963,
18066, 18085, 18108, 18146, 18186, 18401, 18422, 18471, 18493, 18567, 18621, 18746, 18776, 19051, 19059, 19082, 19193, 19285, 19291, 19490, 19499, 19505, 19506, 19509, 19
588, 19590, 19832, 19927, 19947, 19998, 20030, 20095, 20101, 20270, 20305, 20420, 20721, 20869, 20944, 21067, 21131, 21561]
```

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

```
Enter query : 3 oil /3 price
400
1127, 144, 235, 236, 237, 242, 246, 247, 248, 273, 274, 288, 352, 353, 357, 364, 370, 459, 471, 489, 502, 543, 697, 834, 843, 873, 888, 896, 915, 952, 957, 1306, 1349, 137
9, 1387, 1456, 1550, 1552, 1616, 1692, 1696, 1799, 1906, 1909, 1999, 2046, 2061, 2173, 2228, 2383, 2423, 2530, 2696, 2775, 2838, 2925, 2973, 2975, 3019, 3024, 3048, 3065,
3181, 3189, 3249, 3269, 3303, 3338, 3354, 3389, 3430, 3452, 3455, 3488, 3490, 3505, 3507, 3571, 3593, 3657, 3798, 3818, 3864, 4005, 4017, 4035, 4037, 4049, 4067, 4080, 412
6, 4136, 4232, 4246, 4290, 4338, 4365, 4531, 4546, 4589, 4662, 4713, 4744, 4835, 4870, 5037, 5061, 5137, 5143, 5145, 5156, 5167, 5171, 5178, 5184, 5193, 5255, 5268, 5273,
5274, 5295, 5318, 5323, 5389, 5631, 5769, 5787, 5830, 5851, 5953, 5985, 6023, 6037, 6119, 6169, 6177, 6201, 6208, 6264, 6371, 6413, 6656, 6722, 6740, 6760, 6876, 6954, 699
4, 6996, 7135, 7150, 7174, 7366, 7397, 7408, 7423, 7589, 7742, 8015, 8033, 8039, 8041, 8050, 8051, 8069, 8095, 8117, 8134, 8173, 8209, 8478, 8606, 8623, 8630, 8633, 8703,
8747, 8764, 8820, 8884, 8905, 8971, 9031, 9149, 9180, 9193, 9213, 9256, 9293, 9352, 9436, 9445, 9462, 9485, 9583, 9639, 9691, 9706, 9722, 9732, 9736, 9761, 9763, 9769, 979
9, 9853, 9952, 10078, 10080, 10091, 10135, 10190, 10192, 10228, 10261, 10275, 10291, 10292, 10330, 10385, 10605, 10632, 10649, 10693, 10703, 10927, 10944, 10952, 10975, 11
000, 11083, 11118, 11149, 11171, 11172, 11213, 11232, 11236, 11241, 11388, 11559, 11711, 11723, 11752, 11753, 11768, 11778, 11781, 11839, 11852, 11882, 11949, 12013, 12050, 12277
12279, 12320, 12608, 12680, 12795, 13115, 13236, 13256, 13281, 13517, 13930, 14558, 14614, 14698, 14700, 14708, 14724, 14734, 14749, 14770, 14799, 14832, 14833, 14873, 1
4891, 14892, 14931, 14942, 15038, 15063, 15212, 15322, 15386, 15389, 15551, 15607, 15639, 15812, 15824, 15829, 15939, 15946, 16005, 16044, 16195, 16268, 16270, 16651, 1665
8, 16680, 16939, 16991, 17015, 17079, 17096, 17100, 17102, 17131, 17135, 17185, 17199, 17243, 17254, 17289, 17291, 17294, 17327, 17329, 17353, 17359, 17369, 17372, 17405,
17408, 17409, 17419, 17429, 17430, 17433, 17441, 17452, 17469, 17478, 17816, 17924, 17926, 17929, 17963, 18066, 18085, 18108, 18146, 18186, 18193, 18367, 18401, 184
22, 18471, 18480, 18493, 18567, 18621, 18680, 18701, 18736, 18738, 18744, 18746, 18773, 18776, 18810, 19017, 19051, 19059, 19082, 19193, 19285, 19291, 19478, 19490, 19499,
19505, 19506, 19509, 19588, 19590, 19832, 19927, 19947, 19998, 20030, 20095, 20101, 20270, 20305, 20420, 20721, 20869, 20919, 20944, 21067, 21131, 21267, 21363, 21561]
Enter query : 1
```