CS4225 Assignment 2 Task 1

Final Output

Using prefined parameters, my final k-means clusters looked like -

Centroid	Cluster Size	Median Score	Average Score
(50000,0)	124563	0	0
(50000,2)	236993	2	2
(50000,61)	3798	52	62
(50000,255)	586	233	257
(50000,726)	130	680	726
(50000,1683)	38	1595	1683
(50000,5007)	5	4441	5007
(100000,0)	214232	0	0
(100000,4)	165156	3	4
(100000,64)	4103	53	64
(100000,282)	431	242	282
(100000,1077)	34	961	1077
(100000,10271)	2	10271	10271
(150000,1)	314828	1	1
(150000,84)	1345	67	84
(150000,446)	82	383	446
(150000,2131)	4	1777	2131
(200000,2)	172020	2	2
(200000,64)	2682	52	64
(200000,307)	259	270	308
(200000,922)	41	799	922
(200000,3770)	3	3335	3770
(250000,0)	98797	0	0
(250000,2)	249119	2	2
(250000,23)	14657	20	23
(250000,104)	1707	92	105
(250000,333)	257	300	336
(250000,923)	30	787	933
(300000,1)	150853	1	1
(300000,10)	28726	9	10
(300000,64)	1729	55	64
(300000,259)	186	229	259
(300000,772)	32	639	772
(300000,3636)	2	3636	3636
(350000,4)	55409	1	4
(400000,3)	113982	1	3
(450000,2)	94617	1	2
(450000,110)	903	89	110
(450000,557)	82	473	557
(500000,3)	24001	2	3
(572231,6)	23764	3	6
(673571,3)	21634	2	3

Insights from Results

- There seems to be almost an inverse relationship between cluster size and average/median score of the cluster. Bigger clusters tend to have lower average scores of around 0-5. However, a few small clusters (size < 10), have average scores which are very high (3000 < score < 11000). This seems reasonable since most posts on Q/A forums are expected to have a low score and some popular ones have vastly higher scores.

- When sorted by the centroid vectors (as is shown in the output above), we can notice a zig-zag pattern of cluster sizes. In every range of domain indices (i.e in 50000s, 100000s, 150000s, 200000s ...), the biggest cluster is seen where the max score is smaller. This also aligns with our expectations of having the majority of the posts with low scores.

- From the data, we can also find out the most prominent topics. The largest clusters and their corresponding topics are -

- (150000,1)	- Algorithm	(Size - 314828)
- (100000,0)	- Compute-Science	(Size - 214232)
- (200000,2)	- Big-Data	(Size - 172020)
- (300000,1)	- Security	(Size - 150853)
- (50000,0)	- Machine-Learning	(Size - 124563)

Further discussion on the system performance

- In the current implementation, we are using k random vectors as our initial centers. This is not the optimal selection and would lead to subpar results and performance. Results can be bad because the clusters discovered are not the best ones. Performance can also be bad because it will take more iterations to reach convergence. To improve this, we could use K-means++ or other sampling methods.

- Secondly, the proper use of .persist() and .cache() with Spark can lead to some boosts in performance as well. For RDDs that need constant access, this could give us significant performance boosts.

CS4225 Assignment 2 Task 2

Final Output

Since the output is really long, I've put it in the Appendix at the end of this document.

Description and analysis of the results

There are a lot of insights that we can derive from the results -

- Grammatically correct and incorrect use of English is classified into mostly different clusters. For eg, Cluster 3's top words are - "u", "lol", "know", "im" and "dont".

- Clusters also capture themes and emotions. For example, Cluster 2's top words are - "thanks", "thank", "good", "love" and "hope". Similarly, Cluster 13's top words are - "hurts", "sore", "throat", "tummy" and "headache".

- Most clusters are or comparable sizes (1000s) except for clusters that capture a very specific theme such as Cluster 14 (Size 273) which seems to be about "#seb-day". Or it could be due to grammatical anomalies like Cluster 16 (Size 466) which captures repeated question marks.

Analysis of the parameters in k-means

The three main hyperparameters that required tuning in this assignment were K (i.e number of clusters), Vector Size (i.e number of features extracted by Word2Vec), and Word2Vec Minimum Count.

In the parts from here on, I shall refer to these configs as a tuple. So, a (10, 5, 2) config means that K = 10, Vector Size = 5 and Word2Vec Minimum Count = 2.

Word2Vec Minimum Count
This specifies what is the minimum times a word must appear in a tweet to be considered significant for Word2Vec. Since the documents we are dealing with tweets, the chances of a word re-occurring are low even if it is a significant word. Thus the highest value I considered for the Minimum Count was 2. Comparing results for (10,5,1) and (10,5,2), their silhouette scores were 0.216 and 0.220. Thus, Min Count of 2 seems to lead to more tightly coupled clusters but it is a minute difference.

As for performance, using 2 instead of 1 for Minimum Count allowed for slightly faster processing but it was not a substantial difference.

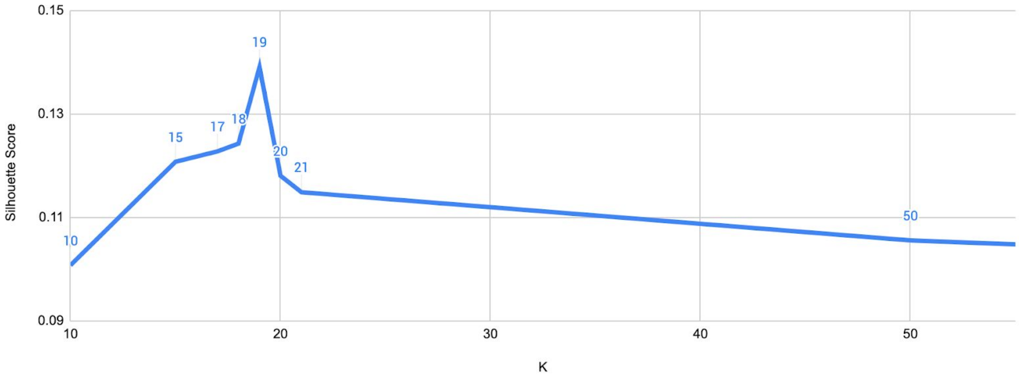
Vector Size (Number of features extracted by Word2Vec)
This was hard to measure. Lower vector size meant fewer dimensions. This made the computations significantly faster. I tried a range of vector sizes. What was interesting to note was its impact on the silhouette score -

K	Vector Size	Minimum Count	Silhouette Score
10	2	2	0.4798
10	3	2	0.3453
10	4	2	0.2811
10	5	2	0.2195

If a higher silhouette score is seen as better, then it feels as though fewer features give better results. But that seems counter-intuitive. My understanding of what's happening here is that as the number of dimensions increases, the vectors are more sparse, spread out. This results in weaker clusters being formed as many points lie closer to the boundaries of these clusters.

Some research papers online show that the expressibility of Word2Vec steadily increases until a vector size of 350. After that returns are diminishing. So, it would make sense to use a large vector size < 350. But since the computing power of the SoC clusters and of my local machine are limited, I have decided to use a vector size 10.

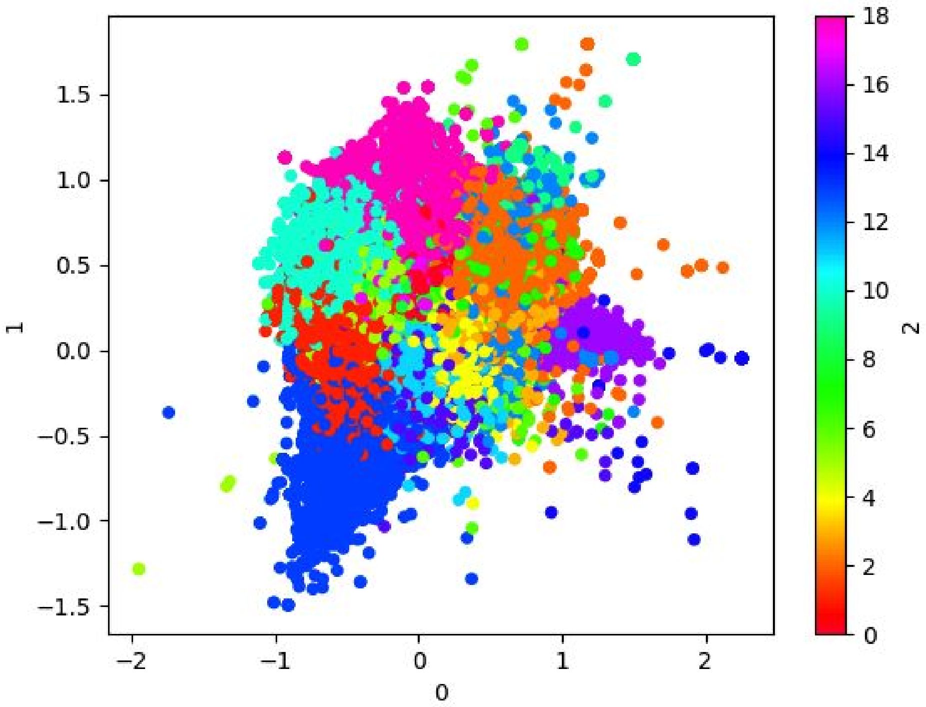
K: Number of Clusters
With Minimum Count set to 2 and Vector Size set to 10, I could run multiple experiments with different K values and use [Silhouette Score Analysis]() to find the best K.



From the graph above, we can see that the best value for K (if we keep other hyperparameters constant) is 19.

Visualization of the result (Bonus point)

For visualization, I had to use Principal Component Analysis to reduce my vectors' dimensions from 19 to 2. I then exported these 2-D vectors and their corresponding cluster labels predicted by K-Means to plot the following graph -



The two axes of the graph are the Principal Component A and Principal Component B. Each color represents one cluster. Although it is not visible, there are 19 clusters shown here.

As we can see, PCA has been able to capture a lot of cluster information. However, it is not perfect as blue points can be seen on the far right where they are closer to the centroid of the purple cluster.

It is hard to gauge the size of each cluster because big patches don't necessarily show size but perhaps spread of the points. This is in the case of Cluster 14, which is one of the smallest clusters but on the graph, it looks like a giant blue patch.

Appendix

The output of Task 2

```
Silhouette = 0.13918721997700298

====== Results for Cluster 0 ======
Cluster Center: (-0.06011799969291251, -0.023177028831928598,
-0.014996501480494709, -0.1049467915040036, -0.10673278589327158,
-0.2537027074044517, -0.010950834291087286, -0.013231797552051797,
0.22052066087870317, -0.19750055519177376)
Cluster Size: 152595
Top 5 tokens:
Token          Frequency
i'm            13838
going          11959
miss           11260
can't          10063
im             8356
======================

====== Results for Cluster 1 ======
Cluster Center: (0.04678089533740316, 0.036702254558853946, -0.07285620237884785,
-0.037099531695305075, 0.27455587502694667, -0.23638620112187914,
-0.2449231022928867, -0.22855999561851817, 0.31194675621241513,
-0.04644831316430648)
Cluster Size: 58272
Top 5 tokens:
Token          Frequency
just           5146
eating         4258
like           3326
&amp;          3099
ice            3042
======================

====== Results for Cluster 2 ======
Cluster Center: (-0.15475690709481502, -0.024568396498337888,
-0.038902296436861585, -0.07992899289305529, -0.19787052259082386,
-0.07460542541550333, -0.03209284738860794, -0.12314013384121451,
0.20435407013952325, -0.48871306265773534)
Cluster Size: 102578
Top 5 tokens:
Token          Frequency
thanks         15959
thank          10991
good           8996
love           7426
hope           5858
======================

====== Results for Cluster 3 ======
Cluster Center: (-0.1834427330443112, -0.021654699163230835, -0.13869451490988158,
-0.058684922548964534, -0.030748304332417663, -0.0484308236336129,
0.09840458547273775, 0.027387152665385252, 0.29492612544481234,
-0.27444902490564005)
Cluster Size: 99746
Top 5 tokens:
Token          Frequency
u              29541
lol            11039
know           10171
im             8448
dont           8055
======================

====== Results for Cluster 4 ======
Cluster Center: (-0.1863434904382457, 0.024416346857956638, -0.1938033235882132,
0.026041122848280855, 0.011883065387359917, -0.11078237003966497,
-0.0469078175527965, -0.04293378107587743, 0.10233059740868433,
-0.26045379338300856)
Cluster Size: 211884
Top 5 tokens:
Token          Frequency
i'm            26832
like           24833
just           23343
don't          22876
know           17812
======================

====== Results for Cluster 5 ======
Cluster Center: (-0.06712496228187131, 0.05712609158609434, -0.007798553647965843,
-0.08733912817944588, -0.09419893775282664, -0.34689199391183334,
0.06124548263699945, -0.11461384847553184, 0.1755741001447111, 0.08224342578658361)
Cluster Size: 96224
Top 5 tokens:
Token          Frequency
2              11024
just           9082
got            7269
days           6904
3              6332
======================

====== Results for Cluster 6 ======
Cluster Center: (0.023259647852385677, -0.355798026507883, 0.06552471869172441,
0.08152799190812682, 0.038075287949532625, -0.45743263055727085,
0.045933780620450594, -0.02942036131069606, 0.19788568300019635,
-0.17265014048821295)
Cluster Size: 43648
Top 5 tokens:
Token          Frequency
watching       9090
new            5238
movie          4203
can't          3873
wait           3447
======================
====== Results for Cluster 7 ======
Cluster Center: (-0.13396287851475358, -0.21942914716167225, -0.06979419035606664,
-0.015838628147850732, -0.023153950781512587, -0.18967752822313813,
-0.0381411636672848, -0.07704457959897788, 0.20152775879998733,
-0.3113095768297426)
Cluster Size: 103555
Top 5 tokens:
Token          Frequency
love           20751
-              9580
just           9262
like           6974
new            5268
======================

====== Results for Cluster 8 ======
Cluster Center: (0.0041197477712397994, 0.14571125301350124, -0.14445668865086503,
-0.0762476314447993, 0.07199524503185936, -0.2945368713369767, 0.08418110684405716,
-0.043557598020296955, 0.1684126576677986, -0.19137873302784214)
Cluster Size: 105759
Top 5 tokens:
Token          Frequency
i'm            27516
just           15540
im             13322
going          11838
sleep          8700
======================

====== Results for Cluster 9 ======
Cluster Center: (0.003584128540830505, 0.41846937305263393, -0.762854894643l339,
1.6855293247103136, -2.2433465370942747, -0.239077277719541, -0.45033052784922617,
-0.34030704946197465, 1.031639642331434, 0.5268851967942783)
Cluster Size: 1660
Top 5 tokens:
Token          Frequency
followers      1572
train          1521
add            1507
100            1506
pay            1500
======================

====== Results for Cluster 10 ======
Cluster Center: (0.13187107046424829, 0.23234899403026182, -0.018751806338054237,
-0.20149145951989647, -0.08535655000704044, -0.5170174654013264,
0.01960972960947412, -0.009001995571803348, 0.23580864761619574,
-0.0789962655043107)
Cluster Size: 62864
Top 5 tokens:
Token          Frequency
going          10974
work           9313
ready          6369
```

```
home           5843
tomorrow       5801
======================

====== Results for Cluster 11 ======
Cluster Center: (-0.11969194863207783, -0.00325881137058783, -0.008531369637356944,
0.002820023222158313, 0.05929167409100469, -0.10706034080383496,
-0.026347204022029332, -0.06387853574836205, 0.07505527488671257,
-0.04945928596674797)
Cluster Size: 208190
Top 5 tokens:
Token          Frequency
just           14302
-              10350
i'm            7729
got            7585
like           6954
======================

====== Results for Cluster 12 ======
Cluster Center: (-0.2253565069089552, 0.1032092409507347, -0.1760659126016635,
0.38217465139066126, -0.30938664932150756, -0.1909994442813527,
-0.009100875174109908, -0.11221574585112089, 0.2890189662422212,
-0.2325394423598657)
Cluster Size: 29797
Top 5 tokens:
Token          Frequency
twitter        4618
just           2953
followers      2750
new            2514
add            2312
======================

====== Results for Cluster 13 ======
Cluster Center: (-0.05603866137242635, 0.29478067619494636, -0.07674065689082486,
-0.12669185537063427, 0.42533299983658246, 0.0024232924676308716,
-0.017056588282757194, -0.23153850683553323, -0.10495843938088725,
0.07180102655685719)
Cluster Size: 23081
Top 5 tokens:
Token          Frequency
hurts          2637
sore           1557
throat         1462
headache       1454
tummy          1181
======================

====== Results for Cluster 14 ======
Cluster Center: (-0.8037943946759631, -0.6223498297791911, 0.9422515236720852,
-0.06586096449462055, 0.15406320094552933, 0.1569248983473181, 0.8665747202061064,
-1.0127143132431413, -1.4653509298307128, -1.9527203410600735)
Cluster Size: 273
Top 5 tokens:
```

```
Token          Frequency
#seb-day       482
died!          211
isplayer       210
sorry          210
#marsiscoming  59
======================

====== Results for Cluster 15 ======
Cluster Center: (-0.18856774575530813, 0.07174436485359573, -0.0512757554407361,
0.21712512730145656, -0.06051242688016094, -0.2279504118904251, -0.047182730321917,
-0.06191093010684792, 0.1258503082594285, -0.07516779934931224)
Cluster Size: 114162
Top 5 tokens:
Token          Frequency
just           12789
new            10579
-              8394
i'm            7387
got            6806
======================

====== Results for Cluster 16 ======
Cluster Center: (-1.0176477398370376, -0.02058763873479107, 1.0979897744620921,
0.1958536013389339, -1.1437660523907809, 1.1240543592662273, -0.5845239534596445,
-0.0797126723369229, 0.9923378976958799, 0.3347727224524171)
Cluster Size: 466
Top 5 tokens:
Token          Frequency
???            443
??             413
????           411
?????          368
?              306
======================

====== Results for Cluster 17 ======
Cluster Center: (-0.04559572431138217, 0.08215319515179872, -0.137727467756344,
-0.1108015961317697, -0.06178700401781939, -0.28494390274291387,
-0.16228422990920433, -0.12632159804275592, 0.13172575796657401,
-0.17594735405458464)
Cluster Size: 134148
Top 5 tokens:
Token          Frequency
good           20756
it's           16920
day            15195
i'm            12676
like           9918
======================

====== Results for Cluster 18 ======
Cluster Center: (0.07236703871699877, 0.10043365773671271, -0.1258537235326712,
-0.3105556360158873, -0.31066411812431616, -0.40275256594152237,
-0.22456325459306756, -0.24522093916811435, 0.18687107587406632,
-0.315686399825755)
Cluster Size: 51098
Top 5 tokens:
Token          Frequency
good           16144
day            15326
morning        9081
happy          5952
great          5224
======================
```