# Hw-6

JUNAID ALI SAYYED

Q1:

I have used cosine distance method. When the vectors' magnitude is irrelevant, cosine similarity is typically utilized as a distance metric.

For instance, when working with text data represented by word counts, this occurs. The most frequent application of this measure is with text data.

However, you might also wish to use cosine similarity in other situations when the instances' characteristics cause the weights to be higher without necessarily implying anything different.

Such an example could be sensor values that were recorded throughout a range of times (between instances).

I would assign topic name as: Cluster1: EARTH and space

cluster2: novel

cluster3: religion

cluster4: Internet

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 1 | 0.83 | 0.70 | 0.76 | 332 |
| 2 | 0.90 | 0.68 | 0.77 | 314 |
| 3 | 0.68 | 0.88 | 0.77 | 355 |
| 7 | 0.69 | 0.77 | 0.73 | 273 |
| accuracy | | | 0.76 | 1274 |
| macro avg | 0.78 | 0.75 | 0.76 | 1274 |
| weighted avg | 0.78 | 0.76 | 0.76 | 1274 |

Q2:

I would use a grid method to pick the parameters like clusters and variance type or I would use bic to find the optimum parameters

I achieved better performance in K-means clusters compared to GMM

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 1 | 0.73 | 0.69 | 0.71 | 332 |
| 2 | 0.56 | 0.86 | 0.68 | 314 |
| 3 | 0.79 | 0.73 | 0.76 | 355 |
| 7 | 0.83 | 0.48 | 0.61 | 273 |
| accuracy | | | 0.70 | 1274 |
| macro avg | 0.73 | 0.69 | 0.69 | 1274 |
| weighted avg | 0.73 | 0.70 | 0.69 | 1274 |

Q3.

Based on analysis I would assign topic name as: Cluster1: EARTH and space

cluster2: novel

cluster3: religion

cluster4: Internet.

LDA basically produced probability distribution for grouping and uses frequency counts. So I would you parameters grid ranging from 3,6 for this problem to select optimum number of clusters.

My LDA achieved the same performance as K-means and better than GMM

```
iteration: 1 of max_iter: 40, perplexity: 3747.0206
iteration: 2 of max_iter: 40, perplexity: 3492.8285
iteration: 3 of max_iter: 40, perplexity: 3294.2049
iteration: 4 of max_iter: 40, perplexity: 3142.3757
iteration: 5 of max_iter: 40, perplexity: 3035.2926
iteration: 6 of max_iter: 40, perplexity: 2960.7164
iteration: 7 of max_iter: 40, perplexity: 2905.9848
iteration: 8 of max_iter: 40, perplexity: 2861.7826
iteration: 9 of max_iter: 40, perplexity: 2826.3852
iteration: 10 of max_iter: 40, perplexity: 2798.3178
iteration: 11 of max_iter: 40, perplexity: 2776.2680
iteration: 12 of max_iter: 40, perplexity: 2758.8788
iteration: 13 of max_iter: 40, perplexity: 2745.1875
iteration: 14 of max_iter: 40, perplexity: 2734.3637
iteration: 15 of max_iter: 40, perplexity: 2725.2595
iteration: 16 of max_iter: 40, perplexity: 2717.2603
iteration: 17 of max_iter: 40, perplexity: 2710.9113
iteration: 18 of max_iter: 40, perplexity: 2705.7915
iteration: 19 of max_iter: 40, perplexity: 2701.2028
iteration: 20 of max_iter: 40, perplexity: 2696.9479
iteration: 21 of max_iter: 40, perplexity: 2692.9061
iteration: 22 of max_iter: 40, perplexity: 2689.6689
iteration: 23 of max_iter: 40, perplexity: 2686.9185
iteration: 24 of max_iter: 40, perplexity: 2684.5347
iteration: 25 of max_iter: 40, perplexity: 2682.5008
iteration: 26 of max_iter: 40, perplexity: 2680.6478
iteration: 27 of max_iter: 40, perplexity: 2679.1157
iteration: 28 of max_iter: 40, perplexity: 2677.8414
iteration: 29 of max_iter: 40, perplexity: 2676.5962
iteration: 30 of max_iter: 40, perplexity: 2675.4058
iteration: 31 of max_iter: 40, perplexity: 2674.3191
iteration: 32 of max_iter: 40, perplexity: 2673.2329
iteration: 33 of max_iter: 40, perplexity: 2672.1683
iteration: 34 of max_iter: 40, perplexity: 2671.0624
iteration: 35 of max_iter: 40, perplexity: 2670.0382
iteration: 36 of max_iter: 40, perplexity: 2669.1880
iteration: 37 of max_iter: 40, perplexity: 2668.4208
iteration: 38 of max_iter: 40, perplexity: 2667.5857
iteration: 39 of max_iter: 40, perplexity: 2666.6989
iteration: 40 of max_iter: 40, perplexity: 2665.9193
Topic 0:
[('water', '461.94'), ('nthe', '305.13'), ('energy', '289.43'), ('light',
'272.52'), ('earth', '260.50'), ('air', '247.68'), ('10', '232.56'),
('used', '218.01'), ('number', '200.51'), ('does', '198.05'), ('time',
'168.76'), ('mass', '165.80'), ('gas', '162.85'), ('like', '158.07'),
('speed', '151.22'), ('force', '147.22'), ('sun', '143.98'), ('heat',
'136.61'), ('space', '134.39'), ('answer', '134.01')]
```

Topic 1:
[('like', '651.71'), ('help', '615.40'), ('just', '597.85'), ('body', '519.62'), ('weight', '474.70'), ('don', '454.14'), ('good', '437.74'), ('know', '411.42'), ('need', '407.74'), ('day', '387.72'), ('eat', '386.03'), ('time', '381.43'), ('doctor', '361.23'), ('blood', '351.22'), ('really', '333.62'), ('make', '331.50'), ('want', '314.45'), ('does', '287.30'), ('use', '286.97'), ('way', '283.35')]

Topic 2:
[('people', '1311.95'), ('god', '1021.88'), ('just', '825.18'), ('like', '764.06'), ('think', '750.31'), ('know', '686.30'), ('don', '608.54'), ('life', '563.32'), ('time', '454.74'), ('say', '433.68'), ('believe', '401.57'), ('way', '378.23'), ('want', '378.08'), ('does', '365.99'), ('good', '363.12'), ('really', '361.68'), ('person', '357.69'), ('jesus', '350.25'), ('world', '347.32'), ('did', '339.70')]

Topic 3:
[('com', '605.23'), ('www', '484.23'), ('nhttp', '420.22'), ('business', '370.24'), ('need', '357.11'), ('work', '355.34'), ('want', '352.55'), ('good', '348.54'), ('money', '344.76'), ('help', '338.73'), ('job', '335.66'), ('know', '295.42'), ('credit', '265.25'), ('pay', '251.57'), ('http', '249.22'), ('like', '228.16'), ('company', '224.93'), ('don', '222.37'), ('make', '220.87'), ('question', '217.78')]

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 1 | 0.71 | 0.89 | 0.79 | 332 |
| 2 | 0.94 | 0.64 | 0.76 | 314 |
| 3 | 0.84 | 0.83 | 0.84 | 355 |
| 7 | 0.70 | 0.74 | 0.72 | 273 |
| accuracy | | | 0.78 | 1274 |
| macro avg | 0.80 | 0.78 | 0.78 | 1274 |
| weighted avg | 0.80 | 0.78 | 0.78 | 1274 |