Assignment 4, CSE 474/574

Part 2.2 - Filtering target classes (4 points)

• 2.2.1. Print the name of classes in your training set along with selected_targets you can use target_names attribute of newsgroups_train

comp.graphics : 1
rec.autos : 7
rec.sport.hockey : 10
sci.med : 13
soc.religion.christian : 15
talk.politics.guns : 16
talk.politics.mideast : 17

Part 2.3 - Vectorizing documents (12 points)

- **2.3.1. What does TF-IDF stand for?**

  The full form for TF-IDF is Term Frequency- Inverse Document Frequency

- **2.3.2. Why don't we only use term frequency of the words in a document as its feature vector? what is the benefit of adding inverse document frequency?**

  Term Frequency is the ratio of a word occurrence to the total number of words in the document. On the other hand, inverse document frequency is used to calculate the importance of words. Sometimes we have words like of, and, which are used mostly but they have very little significance and these are given equal attention if we only use term frequency. So using inverse document frequency tends to give more weightage to significant words.

- **2.3.3. Calculate the tf-idf vectors of the following two documents, assuming this is the entire corpus:**

Document 1

Term	Term Count
this	1
is	1
a	2
sample	1

Document 2

Term	Term Count
this	1
is	1
another	2
example	3

Document 1 -> [½, ½, 2, 1]
Document 2 -> [½, ½, 2, 3]

Part 3.1 - Sparsity (12 points)

In this section, we will interpret the coefficients from the final model you trained on all of the training data.

- **3.1.1 Count the number of non-zeros in each row of the train_vec matrix.**

  [89, 94, 217, 70.....258, 342, 205, 124]

  Note - Complete answer is in the notebook.

- **3.1.2 What is the average number non zero elements in each row?**

  170.56187209017398

- **3.1.3 On average what percentage of elements in each row have non-zero elements?**

  0.3037448971384858

Part 3.2 - SVD (4 points)

- **3.2.1. What portion of the variance in your dataset is explained by each of the SVD dimensions?**

  There are 3 dimensions. The percentage of variance explained by each of the selected components is -

  [0.01618638 0.00617073 0.00540306]

Part 3.4 - Visualization (8 points)

- **3.4.1. Based on your observation, what is the difference between SVD and UMAP embeddings? 1-2 sentences should suffice.**

  UMAP looks way neater. The clustering has less spread (variance) so the clusters are easily distinguishable, i.e. the intra-cluster distance also seems to be more. SVD graph has a lot of overlap.

- **3.4.2. Which one do you prefer to use for a classification task? why? 1-2 sentences should suffice**

  We would prefer UMAP since the clustering is well represented spatially and has low variance.

Part 4.1 - Clustering and evaluation (16 points)

- **4.1.1 What is the range of possible values of silhouette coefficients?**

  (-1,1)

- **4.1.2 Describe what a silhouette score of -1 and 1 mean?**

  The Silhouette Coefficient for a sample is (b - a) / max(a, b), where a = mean intra-cluster distance & b = the distance between a sample and the nearest cluster that the sample is not a part of.

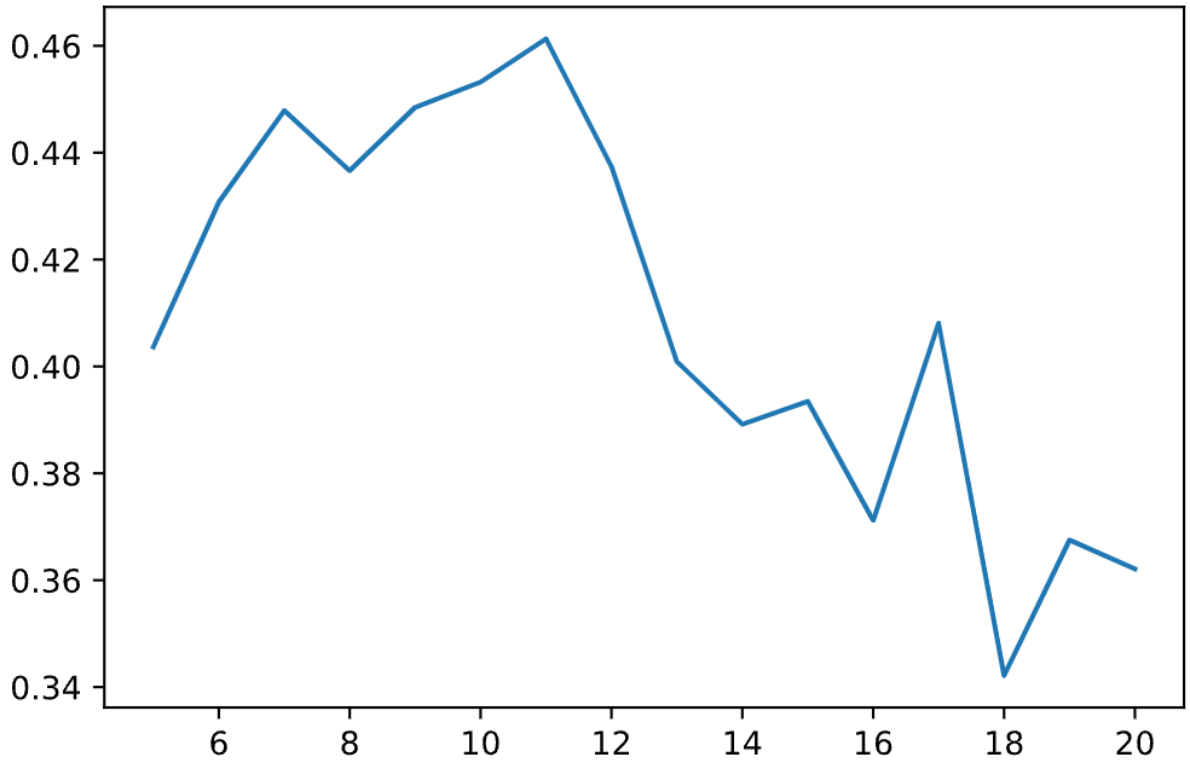
A silhouette score of -1 is the worst possible score we can get which means the distance between a sample and the nearest cluster that the sample is not a part of is 0 so the clustering is wrong.

A silhouette score of 1 (b > a & a=0) is the best possible score we can get which means that clusters are well apart from each other and clearly distinguished.

- **4.1.3. Use silhouette score and KMeans from sklearn library to find the optimum number of clusters in your train_umap. Don't forget to use SEED as your kmeans random_seed. In order to do this try different values of cluster numbers from 5 to 20. Choose the one that results in the best score.**

  The best number of clusters = 11

- **4.1.4. Plot silhouette score for different values of n_clusters (a plot with n_clusters on the x-axis and silhouette score on the y-axis). Don't forget to put the plot in your report.**



Part 4.2 - Making a Kmeans classifier (4 points)

- **4.2.1 show your mapping (resulted dictionary) inside your project report.**

  {5: 13, 10: 1, 2: 16, 9: 7, 0: 10, 4: 17, 3: 15, 8: 13, 7: 17, 1: 13, 6: 15}

Part 4.3 - Analyzing clusters (12 points)

- **4.3.1. Are there any two clusters in your clustering output with the same original label (for example, are there two clusters which both have same training label)? Use your visualizations and describe why?**

  Yes, there are.
  {1, 5, 8} -> 13, {3, 6} -> 15, {4, 7} -> 17

  It's mostly because of the intracluster distance and overlap between them. For e.g. in the baby plot generated using train_umap, clustering. labels_ (Kmeans cluster labels) in the notebook clusters 1, 5 & 8 have some overlap and the data points are harder to distinguish.

- **4.3.2. Write the function below that returns nearest samples to a cluster center. Use this function and explain why there are overlaps in your labels?**

  For cluster = 1

```
(4081, 11)
<class 'numpy.ndarray'>
4081

array([ 0.07115079,  0.11022696,  0.14070763, ..., 10.057724  ,
       10.069924  , 10.07467   ], dtype=float32)
```

  Here, the closest value to the cluster center are 0.07115079, 0.11022696, 0.14070763

For cluster = 5

```
(4081, 11)
<class 'numpy.ndarray'>
4081

array([ 0.07115079,  0.11022696,  0.14070763, ..., 10.057724  ,
       10.069924  , 10.07467   ], dtype=float32)
```

Here, the closest value to the cluster center are 0.07115079, 0.11022696, 0.14070763

For cluster = 8

```
(4081, 11)
<class 'numpy.ndarray'>
4081

array([ 0.07115079,  0.11022696,  0.14070763, ..., 10.057724  ,
       10.069924  , 10.07467   ], dtype=float32)
```

Here, the closest value to the cluster center are 0.07115079, 0.11022696, 0.14070763

- **4.3.3. Can you infere the overlapping label(s) by checking out most central samples? check with original labels.**

  Looking at the top matches from clusters 1, 5 & 8, we can see that most of the data is about related to the medical field. This makes sense since all 3 of these clusters later map to cluster 13 in the dataset. We know that sci. med: 13

Part 4.4 - Evaluate your Kmeans model on test dataset (12 points)

- **4.4.1. Using the generated mapping, and your clustering model, predict the labels of test dataset (you can use the embeddings of test data that you generated by umap test_umap)**

  {0: 1, 1: 15, 2: 1, 3: 10, 4: 10, 5: 10, 6: 13, 7: 7, 8: 1}

- **4.4.2. Calculate the accuracy of model**

  The accuracy of our model is: 0.2119205298013245

- **4.4.3. Calculate both micro and macro values of precision, recall, and F1 score**

  Precision micro = 0.2119205298013245, Recall micro = 0.2119205298013245,

  F1 micro = 0.2119205298013245

  Precision macro = 0.16438320249305233, Recall macro = 0.20699630799012975,

  F1 macro = 0.17517260997281656

574 ONLY Part 5.1 - KNN classification (16 points)

- **5.1.1. Train two seperate KNN models on both SVD and UMAP embeddings. Use n_neighbors=100.**

  Two separate KNN models on both SVD and UMAP are created in the code.

- **5.1.2. Evaluate your model on test datas (test_umap and test_svd). Which model performs better? Why?**

  UMAP has performed better as compared to SVD as the precision, recall, and F1 values for UMAP were higher than SVD. This is again probably due to UMAP generating a better clustering of the data and being able to work directly on a sparse matrix.

- **5.1.3. Calculate macro and micro precision recall and fscore for test_umap. Which one of the two do you prefer for evaluating your model? Why?**

  For micro:

  Precision micro = 0.7711552612214864, Recall micro = 0.7711552612214864,
  F1 micro= 0.7711552612214864

  For macro:

  Precision macro = 0.7789546890411838, Recall macro = 0.771345552517061, F1 macro = 0.7710583416123292

  Accuracy 0.7711552612214864

  We would prefer the UMAP micro since the F1 micro is slightly better than the F1 macro.

- **5.1.4. Shortly describe why the two sets of values (macro and micro) are so similar in this case.**

  When the class imbalance is higher than we could observe a large difference between macro and micro, but in our case, the class imbalance is not that much great hence we could say that the two sets of values (macro and micro).

Contribution Statement

All of us individually completed all parts and then we matched our results to iron out any issues. We worked on the report together and everyone contributed equally.