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
A: `['comp.graphics', 'rec.autos', 'rec.sport.hockey', 'sci.med', 'soc.religion.christian', 'talk.politics.guns', 'talk.politics.mideast']`

Part 2.3 - Vectorizing documents (12 points)
• 2.3.1. What does TF-IDF stand for?
A: 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?
A: If we only use the term frequency, we cannot measure the importance of words accurately because a word that is measured a high importance in a document can has a high frequency in other documents, which means it is not that much important.

A: The benefit of adding inverse document frequency is that the importance of the word can be measured more accurately by determining the importance of the word by considering the number of documents including the word.

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

	Document1	Document2
this	0	0
is	0	0
a	0.12	0
sample	0.06	0
another	0	0.10
example	0	0.15

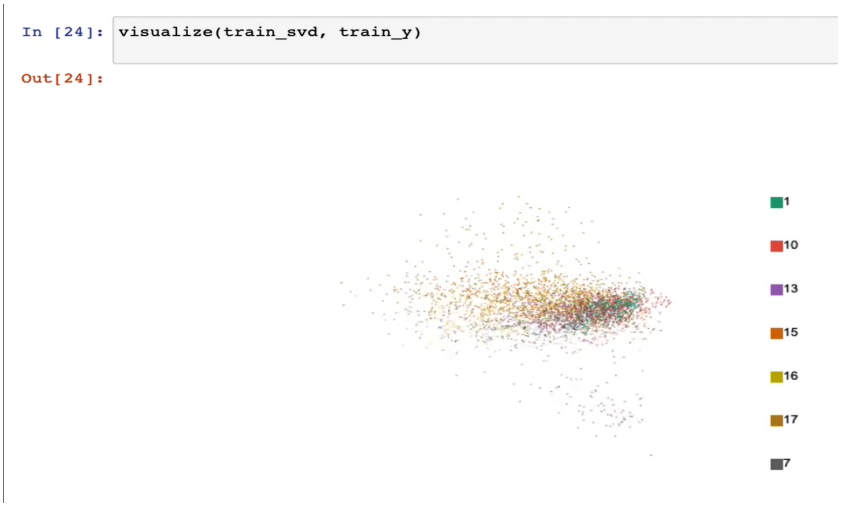
Part 3.1 - Sparsity (12 points) In this section we will interpret the coefficients from the final model you trained on all of the training data.
• 3.1.1 Count the number of non-zeros in each row of the train_vec matrix.
A: `696063` (it is the total number of non-zeros that obtained from each row in the matrix)
If the number of non-zeros for each row is needed, please see the PA_4.ipynb
(it is too long to paste here)

• 3.1.2 What is the average number non zero elements in each row?
A: 170 - Average number of non zero elements
If this question asks the average value, then please see the code

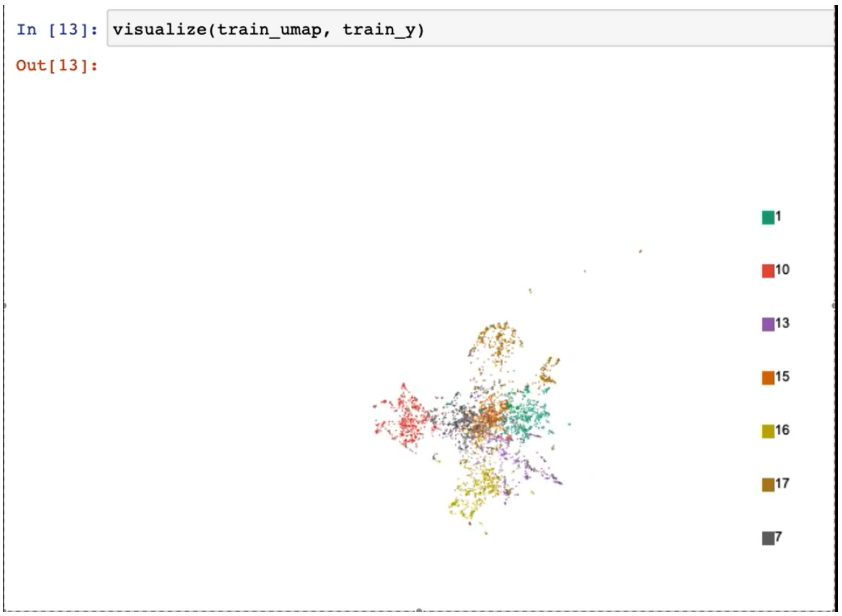
• 3.1.3 On average what percentage of elements in each row have non-zero elements?
A: 100%

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?
A: News Topics (that contain the most targets in the selected target list)

Part 3.4 - Visualization (8 points)
SVD



UMAP



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

A: SVD is linear dimensionality reduction method, while the UMAP is nonlinear dimensionality reduction method.

• 3.4.2. Which one do you prefer to use for a classification task? why? 1-2 sentences should suffice
A: UMAP. Although it is not an efficient method when applied to large datasets, the UMAP reduction method is good for visualizing clusters.

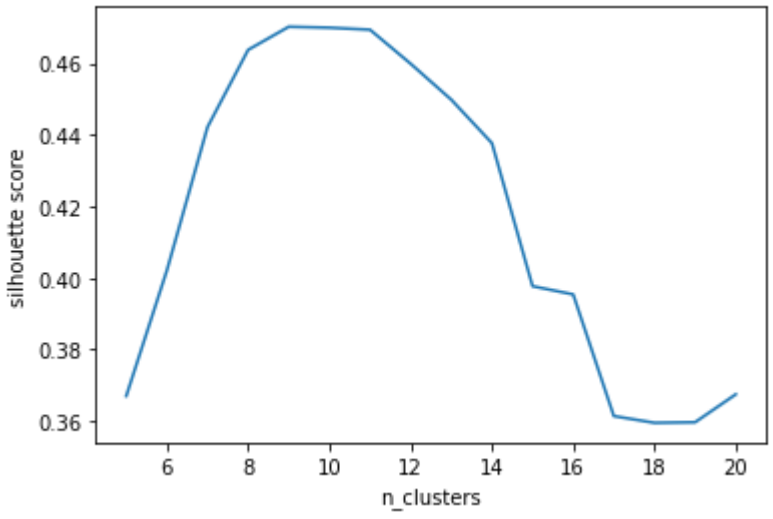
Part 4.1 - Clustering and evaluation (16 points)
• 4.1.1 What is the range of possible values of silhouette coefficients?
A: -1 to 1

• 4.1.2 Describe what a silhouette score of -1 and 1 mean?
A: 1 means clusters are well apart from each other and clearly distinguished, and -1 has opposite meaning.

• 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.
A: 9

• 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.

A:



Part 4.2 - Making a Kmeans classifier (4 points)
• 4.2.1 show your mapping (resulted dictionary) inside your project report.
A: `{6: 7, 1: 1, 2: 16, 3: 10, 4: 17, 8: 15, 0: 13, 7: 17, 5: 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?

{6: 7, 1: 1, 2: 16, 3: 10, 4: 17, 8: 15, 0: 13, 7: 17, 5: 15}
A: Yes, there are two pairs of clusters that have the same training label for each pair.

• 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?
A:
[13, 15, 7], [1, 15, 7], [16, 7, 13], [10, 15, 1], [17, 1, 7], [15, 13, 15], [7, 15, 16], [17, 1, 7], [15, 7, 1]

• 4.3.3. Can you infer the overlapping label(s) by checking out most central samples? check with original labels
A: [13, 15, 7], [1, 15, 7], [16, 7, 13], [10, 15, 1], [17, 1, 7], [15, 13, 15], [7, 15, 16], [17, 1, 7], [15, 7, 1]

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)

• 4.4.2. Calculate the accuracy of model
A: 0.71 (71%)

• 4.4.3. Calculate both micro and macro values of precision, recall and F1 score
A:

	precision	recall	f1-score	support
1	0.77	0.73	0.75	389
7	0.73	0.75	0.74	396
10	0.87	0.85	0.86	399
13	0.93	0.25	0.39	396
15	0.46	0.96	0.62	398
16	0.86	0.73	0.79	364
17	0.89	0.71	0.79	376
accuracy			0.71	2718
macro avg	0.79	0.71	0.71	2718
weighted avg	0.79	0.71	0.71	2718