Assignment 4, CSE 474/574

The number of points per question are in parentheses here (but not in the jupyter notebook).

Notes on grading:

- We reserve the right to spot check any code to ensure that the answers provided are based on the code that you wrote in your assignment. Corect answers on the report that are not clearly based on analysis in the notebook are subject to receiving zero points, and depending on what we believe transpired, an AI violation. So, please show your work!

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`

Part 2.3 - Vectorizing documents (12 points)

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

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.
- **3.1.2** What is the average number non zero elements in each row?
- **3.1.3** On average what percentage of elements in each row have non-zero elements?

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?

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

Part 4.1 - Clustering and evaluation (16 points)

- **4.1.1** What is the range of possible values of silhouette coefficients?
- **4.1.2** Describe what a silhouette score of -1 and 1 mean?
- **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.
- **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.

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?
- **4.3.2.** Write the function bellow that returns nearest samples to a cluster center. Use this function and explain why there are overlaps in your labels?
- **4.3.3.** Can you infere the overlapping label(s) by checking out most central samples? check with original labels.

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

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`.
- **5.1.2.** Evaluate your model on test datas (`test_umap` and `test_svd`). Which model performs better? Why?
- **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?
- **5.1.4.** Shortly describe why the two sets of values (macro and micro) are so similar in this case.

Contribution Statement (Minus 15 points if you do not submit this)

Please describe what each group member contributed to this project. Note that the professor and TA reserve the right to challenge this statement, and **falsification of effort will be considered a violation of Academic Integrity** . That is, if we find reason to believe that a specific group member's claim about the work they contributed is not valid, **then we reserve the right to take steps to ensure that the group member did, in fact, contribute as stated**. We also reserve the right to adjust grades based on extreme differences in effort put into the assignment.