Assignment 3 K-Means

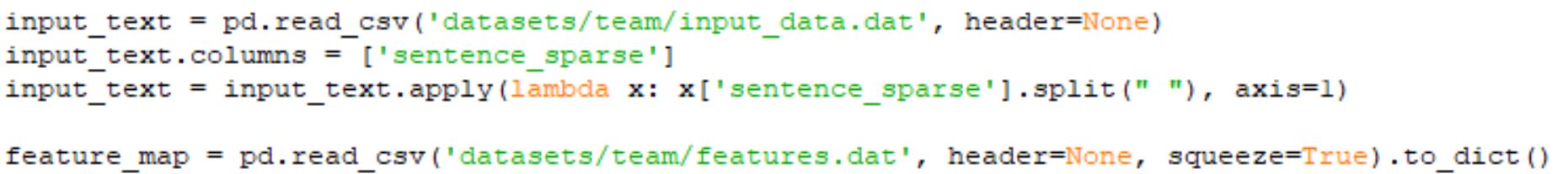
Team Name: mr.miner

Rank: 16

V-score: 0.65

Name: Murtaza Shareef(G01024452) and Deepak Kanuri(G01070295)

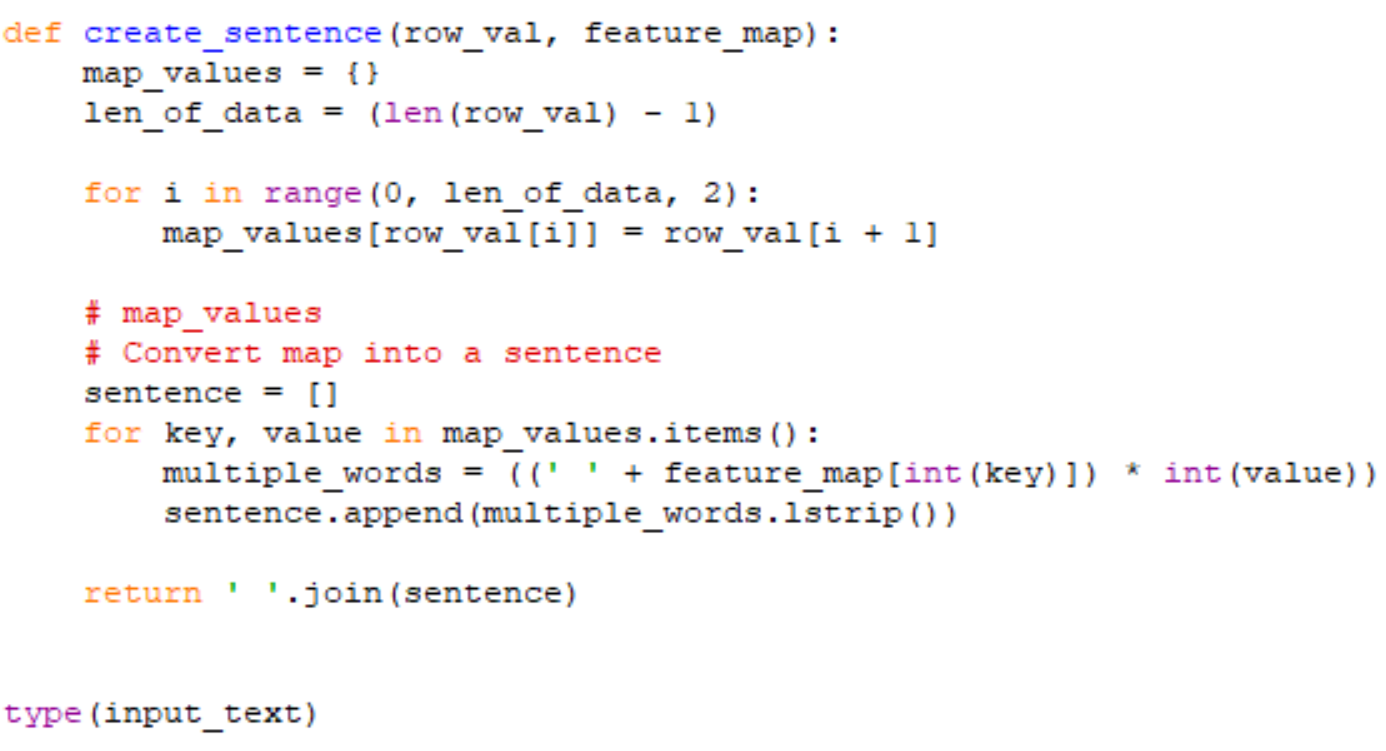
Report:

The main challenge we faced when we were doing this assignment was to read the files and make sense of it. On one hand, we had the file of words and the other hand file which contained the id of the word and its frequency in the sentence. So it was difficult to accommodate both the files. So we read both the files and constructed the sentences using the words.

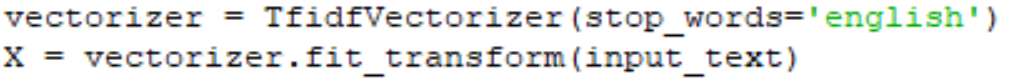


Here for each row, the sentence is created.

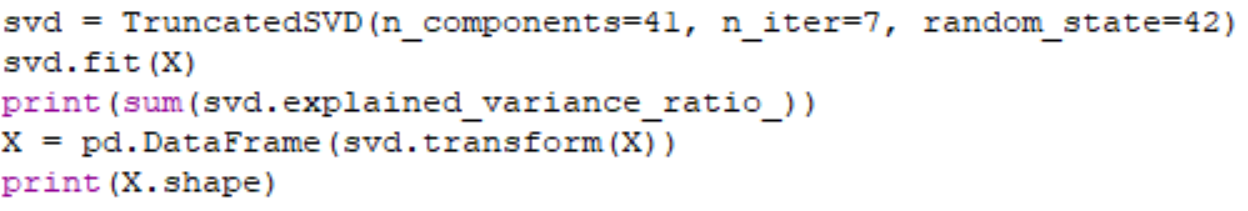
The code for mapping the words to its sentence is



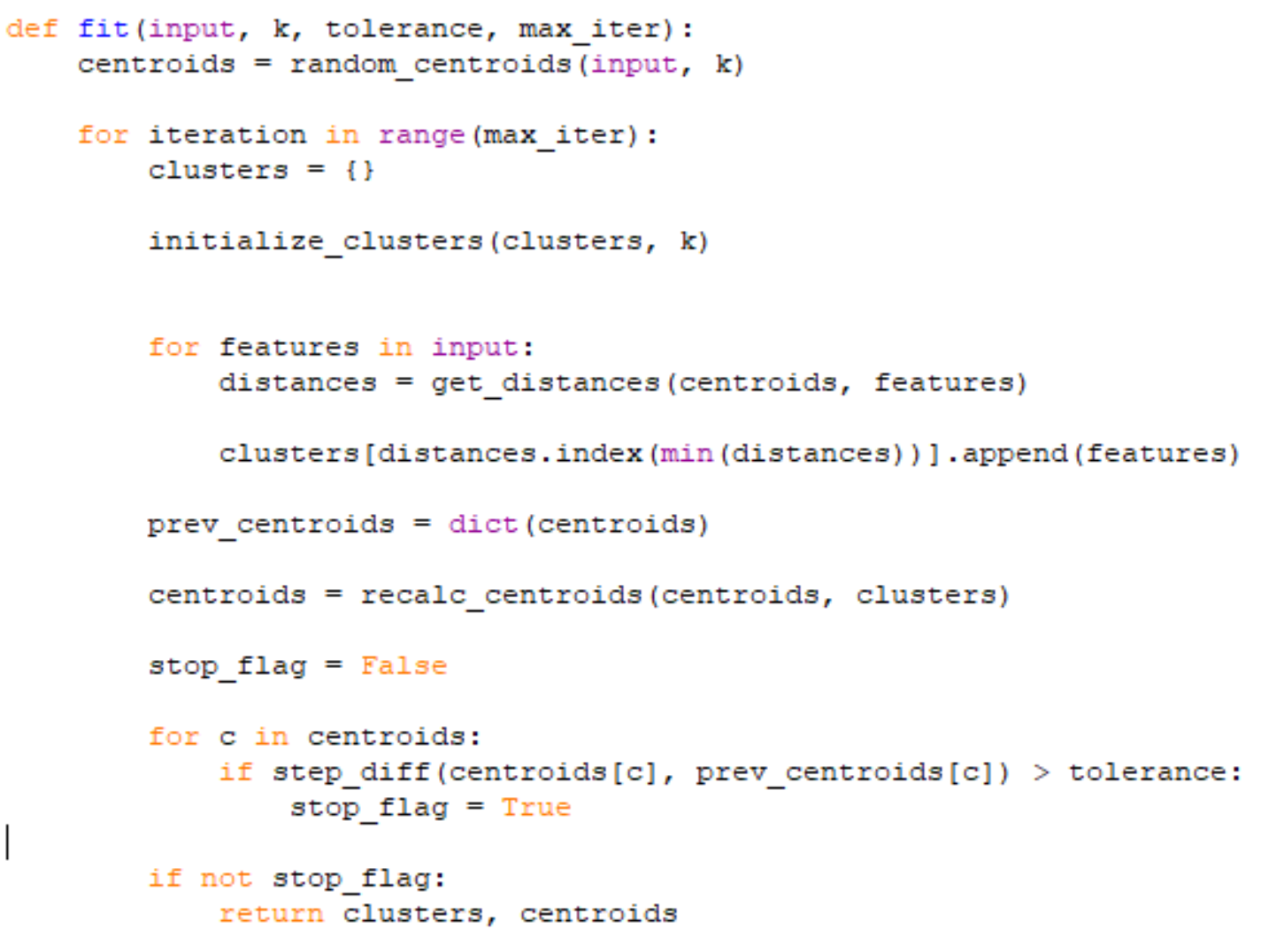
Our next step was to vectorize the document, as it is a text document.



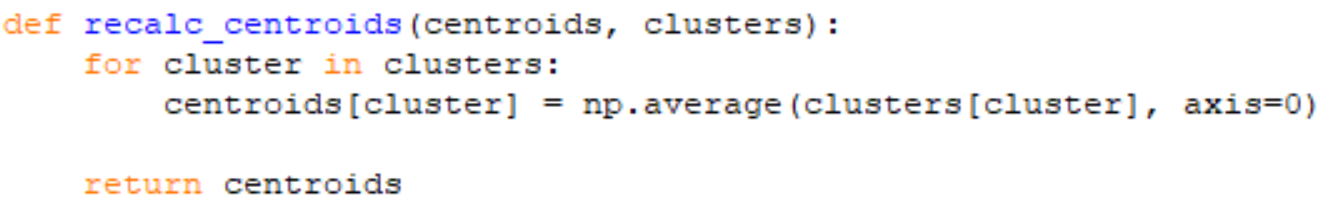
As the dataset was large we tried to reduce the dimensionality using PCA, but that did not work as it was a sparse matrix. So we went forward with TruncatedSVD().



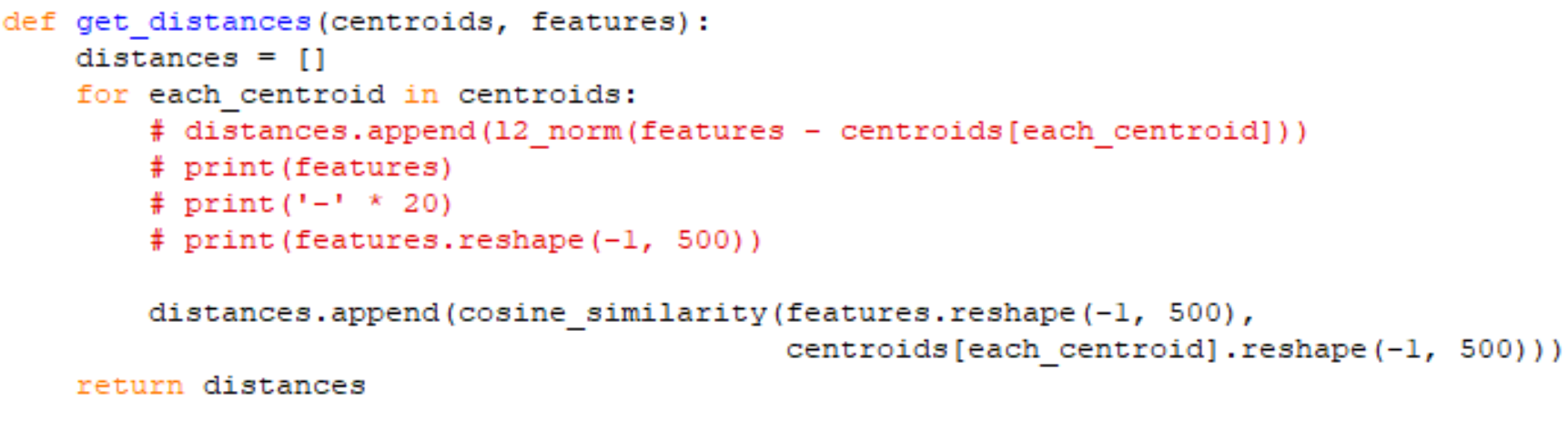
Now our next step was to define our own K-Means code. We first defined a fit() method which takes in the K, max-iterations and tolerance values. This method tries to find the clusters and centroid and does the same function over the max-iterations.



This method first initializes the given number of ‘K’ clusters. It also randomly selects centroids. It continuously recalculates the centroids based on the clusters.

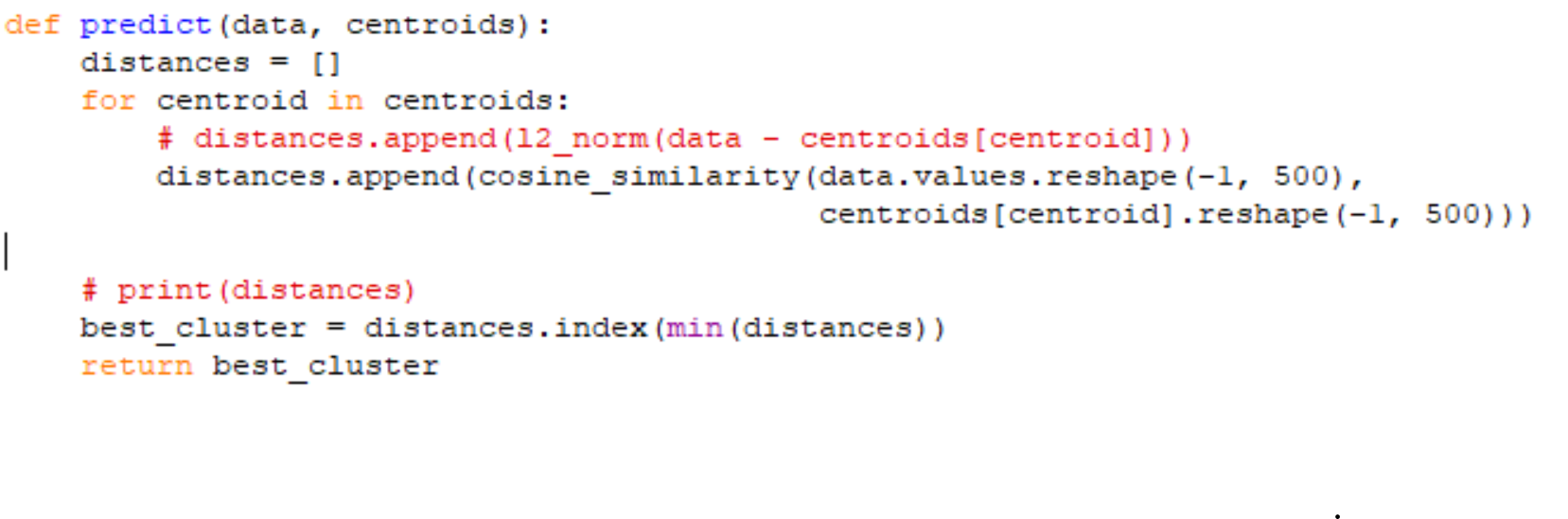


But more importantly, it calculates the distances between the centroids and the features. We first calculated using Euclidean distance but got bad results, thus shifted to cosine similarity.



After ‘K’ iterations the fit() method returns the best clusters and centroids.

Now we have to predict the data. That means we have to check under which cluster the data falls. For that, we coded a predict() function that uses the data points and centroids as parameters. It checks the cosine similarity between the data points and centroid and finds the minimum distance. The cluster with the minimum distance is returned.



The clusters are then stored in the output file.