Readme File

Steps to reproduce the results:

1. Load the ‘fine foods review’ file in the csv format.
2. Define the ‘cleaning’ function that takes in every individual review and converts into lower case and stems the word to its root origin.
3. Pass every review to the above function and append every review to the ‘clean_Text_sw’ list
4. Eliminate duplicate and select the unique words in the ‘clean_Text_sw’ and assign to a new list L.
5. Load the ‘stopwords’ from a text file and create a global list of stopwords.
6. Define a ‘remove_stopword’ function that agains cleans the data and remove the stopwords.
7. From the unique list we eliminate the stop words.
8. Calculate the word frequency for all the unique words and select the top 500 words.
9. Create an excel file to store the 500 words
10. Run the ‘tfidfVectorizer’ to vectore all reviews using the top 500 words
11. Fit the vectorizer on our dataset
12. Run K means algorithm with k=10 and run it for 1000 iterations
13. Select from each centroid, select top 5 words that represents the centroid along with the feature value