**HW - Q3**

**In the file PROBLEM-3. ipynb**

**Explanation:**

1. First, all the **review/text** context from **finefood.txt** is collected in a list called **reviews**
2. All the **stop words** present in the file **LongStop.txt** are collected in a list called **stop\_words,** and are converted to lowercase
3. All the words from **reviews** list are stored in a **words** list
4. **Set** operation is performed on this **words** list to remove any duplicates and this new list is stored in **L**
5. **W** contains all the words from **L** without the stop words and punctuations (done by comparing if the word is **not present** in the **stop\_word** list and if **isalpha()** then appended to the list)
6. Using **collections.counter()** the frequency for all words present in the **words** lists is obtained and stored in a list called **countofwords** (This is done as we need the count of a particular word in the whole **review/text** field, operations like set removes any duplicates hence the **collections.counter()** is performed on **words** list)
7. Using the function **most\_common()**, we get the top 500 words and store them in a list called **data**
8. To just get the words present in data and not their count so it can be used as vocabulary for vectorizing reviews, we take the first value of the tuple (that is the string) and append it to the **top\_500\_words** list
9. Using **k-means** from sklearn, for **10 clusters** obtain the **10 centroid vectors** in **sortingCenter**
10. Then, from each centroid, top 5 words are selected, and the **word\_list** is appended to **output\_words** list, and the **feature\_list** value is appended to the **feature\_values** list
11. Then the words and feature values are displayed cluster wise

**Execution:** I’ve attached the LongStop.txt (contains all the stop words), dataset file (finefoods.txt) and PROBLEM-3. ipynb in the zip folder

**Step 1:** Extract the Zip at the desired location

**Step 2:** Make sure you’re in the same directory as the contents of the zip after extraction

**Step 3:** Open the file by Launching Jupyter notebook making sure the dataset and the stopword file is present

**Step 4:** Sequentially run the code using Shift + Enter to get the output of that block

**(!!! Some cells might take 15-20 minutes max)**

**Step 5:** Extra cells can be created for you to check the intermediate results, by selecting the cell and then click on the **‘+’** icon beside the save icon to create a new cell.