Lab 7 – SECTION A , BATCH 1 Date: 29th Dec 2021

Exer 1: Clustering

Download the data set "*Online Retail.xlsx"* from *https://archive.ics.uci.edu/ml/datasets/online+retail*

11. Read and write a summary of the metadata .
12. Select only the transactions that have occurred from 01/04/ 2011 and 09/12/2011 and create a dataset.
13. Calculate the RFM values for each customer (by customer id). RFM represents:
    - R (Recency) − Recency should be calculated as the number of months before he or she has made a purchase from the online store. If he/she made a purchase in the month of December 2011, then the Recency should be 0. If purchase is made in November 2011 then Recency should be 1 and so on and so forth.
    - F (Frequency) − Number of invoices by the customer from 01/04/ 2011 and 09/12/2011.
    - M (Monetary Value) − Total spend by the customer from 01/04/ 2011 and 09/12/2011.
14. Use the elbow method to identify how many customer segments exist, using the RFM values for each customer.
15. Create the customer segments with K-means algorithm by using number of clusters is suggested by elbow method.

    ```
    from sklearn.cluster import KMeans
    ```
16. Plot the clusters in a scatter plot and mark each segment differently using lmplot.
17. Print the cluster centers of each customer segment and explain them intuitively.
18. Create the customer segments with Agglomerative algorithm by using number of clusters is suggested by elbow method.

    ```
    from sklearn.cluster import AgglomerativeClustering
    ```
19. Visualize the clusters using the dendrogram.
20. Compare the clusters obtained using KMeans vs. Agglomeration.

Exer 2: Text Analysis

Download the amazon_baby.zip file and answer the following:

12. Check the number of the reviews received for each product.
13. Check the products that have more than 15 reviews.
14. Find any missing review are present or not, If present remove those data.
15. Clean the data and remove the special characters and replace the contractions with its expansion by converting the uppercase character to lower case. Also, remove the punctuations.
16. Add the Polarity, length of the review, the word count and average word length of each review.
17. Visualize the distribution of the word count, review length, and polarity.
18. Visualize polarity considering the rating.
19. Visualize the count of the reviews of each rating available in the dataset.
20. List the Top 20 products based on the polarity.
21. Visualize to check whether the review length changes with rating.
22. Visualize the distribution of Top 25 Unigram, Bigram and Trigram.