CSE508 Information Retrieval
Winter 2024
Assignment - 3

Due Date: April 3, 2024 ; 23:59 **Max. Marks:** 100

Instructions:

1. The assignment is to be attempted individually.
2. The proposed solutions will be evaluated during your code demo and viva.
3. Institute plagiarism policy will be strictly followed.
4. The programming language allowed is Python. You are allowed to use Google Colab.
5. Ensure that your code is thoroughly documented for clarity and understanding.
6. You can utilize libraries such as NLTK and BeautifulSoup for data preprocessing in your Python code.
7. You are required to use version control via GitHub:
    a. Make a GitHub repository with the name:
    CSE508_Winter2024_A3_<Roll_No.>.
    b. Add your assignment TA as a contributor. The TA assigned (along with their GitHub handle) to you for this assignment will be released on the classroom.
8. You must make a detailed report with the name **CSE508_Winter2024_A3_<Roll_No>_Report.pdf** with a brief overview of your approach, methodologies, assumptions, and results for each problem.
9. Steps for submission:
    a. A zipped folder **CSE508_Winter2024_A3_<Roll_No.>** consisting of all your code files, dumped files and **Report.pdf**
    b. A text file **CSE508_Winter2024_A3_<Roll_No.>.txt** consisting of the link to your GitHub repository.
10. If it has been mentioned to code a solution from scratch, using any library is strictly not allowed.

Product Recommendation System based on Amazon Review

The goal is to create a predictive model that accurately forecasts user ratings and evaluates the usefulness of reviews. Leveraging collaborative filtering techniques, we seek to recommend the most relevant items to users by analyzing their past interactions and preferences. Through this approach, we aim to personalize the user experience and enhance engagement by offering tailored recommendations that closely match individual interests and preferences.

1. The dataset for Amazon Review Electronics Product is available at Amazon Reviews Dataset . Download the 5-core dataset for Electronics Category, under the heading of ***Small subset for experimentation***. Read the file to a dataframe. Remember to keep the product metadata in a distinct dataframe as well.
2. Choose a product of your choice. Let's say 'Headphones'.
3. Report the total number of rows for the product. Perform appropriate pre-processing as handling missing values, duplicates and other.
4. Obtain the Descriptive Statistics of the product as : -
    a. Number of Reviews.
    b. Average Rating Score.
    c. Number of Unique Products.
    d. Number of Good Rating.
    e. Number of Bad Ratings (Set a threshold of >=3 as 'Good' and rest as 'Bad'), and
    f. Number of Reviews corresponding to each Rating.
5. Preprocess the Text
    a. Removing the HTML Tags.
    b. Removing accented characters.
    c. Expanding Acronyms.
    d. Removing Special Characters
    e. Lemmatization
    f. Text Normalizer
6. To extract relevant statistics, perform the following EDA -
    a. Top 20 most reviewed brands in the category that you have chosen.
    b. Top 20 least reviewed brands in the category you have chosen.
    c. Which is the most positively reviewed 'Headphone' (Or for any other electronic product you have selected)
    d. Show the count of ratings for the product over 5 consecutive years.
    e. Form a Word Cloud for 'Good' and 'Bad' ratings. Report the most

commonly used words for positive and negative reviews by observing the good and bad word clouds.

f. Plot a pie chart for Distribution of Ratings vs. the No. of Reviews.
g. Report in which year the product got maximum reviews.
h. Which year has the highest number of Customers?

7. Use a relevant feature engineering technique to model review text as Bag of Words model, TF-IDF, Hashing Vectorizer or Word2Vec.
8. The Rating Class is divided into three categories
   - > 3 as Good
   - =3 as Average
   - <3 as Bad.
9. From the dataset, take the Review Text as input feature and Rating Class as target variable. Divide the data into Train and Test Data in the ratio of 75:25.
10. Compare the performance of 5 Machine Learning based models on the basis of Precision, Recall, F-1 Score and Support for each of the 3 target classes distinctly.
11. **Collaborative Filtering** :
    a) Create a user-item rating matrix
    b) Normalize the ratings, by using min-max scaling on user's reviews
    c) Create a user-user recommender system - i.e,

    i) Find the top N similar users, by using cosine similarity. N = 10, 20, 30, 40, 50

    ii) Use K-folds validation. K = 5. Explanation: Create 5 subsets, and take 1 of them as the validation set. Take the rest 4 to be the training set.

    iii) Use the training set to predict the missing values, and use the validation set to calculate the error. (Error = |actual_rating - predicted_rating|)

    iv) Report the MAE (Mean Absolute Error) for taking K = 10, 20, 30, 40, 50 similar users.

    d) Create an item-item recommender system. Use the same steps as above.
    e) Plot separate graphs for each of the two recommender systems, plotting MAE against K

12. Also, report the TOP 10 products by User Sum Ratings.