Programming Assignment - 1
CSL 7340

Dataset:

1. Choose one category from http://jmcauley.ucsd.edu/data/amazon/ - amazon product review data.
2. Choose at least 25,000 (reviews). [if no. of reviews > 25k)
3. Review rule, for dataset:
   a. [overall > 3.0] - positive
   b. [overall <= 3.0] - negative

Module - 1 (Statistics):

Tasks:-

1. Explain the text processing pipeline adopted by you.
2. Generate term statistics:
   a. Vocabulary size with word frequencies
   b. N-grams
   c. POS collections
3. Verify Zipf’s law – what is the best fit for your corpus?
4. Which set of terms best describe your corpus? How did you arrive at it?

Module - 2 (Sentiment Analysis using statistical NLP):

Tasks:-

1. Use the following vector space models
   a. CountVectorizer.
   b. TF-IDF.
   c. Any external vectorizer (cite the original paper).
2. Do sentiment analysis using all (a,b,c) using classical ML techniques
   a. Naive Bayes Model.
   b. Decision Tree.

    c. Logistic Regression.

3. Report metrics [accuracy, f1 score, confusion matrix] for all the combinations in (1 and 2)
4. Analyse the results. [Report clearly which vector space model is giving better results on each model used]

Module - 3 (Topic analysis and topic (attribute) wise sentiment analysis):

Tasks:-

1. Extract the topics from the reviews using any topic extraction technique of your choice.
2. Report sentences under each topic.
3. Analyse whether the topics extracted make sense. Justify your claim with some examples.
4. Report topic wise sentiment distribution for the whole repository. Explain the method that you used. Give complete reference of any paper that you use for the purpose.

Instructions:

1. Submit a .zip file containing all the working codes (.py files). Zip file should be named in the format <RollNo1_RollNo2_RollNo3>.zip.
2. Submit a report which should contain:
    a. Detailed description of what all you have done,
    b. Links to the Google-Colab files,
    c. Clearly mention the contribution of each group member.
3. Copying from the Internet and/or your classmates is strictly prohibited. Any team found guilty will be awarded a suitable penalty as per IIT rules.