Programming Assignment - 1 [18/01/2022]
CSL7340

Instructions:

1. **Total marks - 15 = 10 (code) + 5 (video)**
2. **Assignment firm deadline - 5 PM 01-02-2022.**
3. **All libraries, tools are allowed.**
4. Submit a **.zip** file containing all the working codes (.py files and .pdf file). The zip file should be named in the format <RollNo1_RollNo2_RollNo3_NLU_A1>.zip.
5. Submit a **.pdf** report which should contain:
    a. A detailed description of what all you have done,
    b. A description of what else could have been done - what could be the challenges,
    c. Links to the Google-Colab files (if any),
    d. Clearly mention the contribution of each group member.
6. Create one video presentation of **max 5 min** explaining the highlights of your results, observation about the corpus that you gained from this exploration **(faces should be clearly visible)**. [**Video upload assignment will be shared - don't add the video to .zip**]
7. **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.**

Datasets:

1. IMDB dataset - https://www.kaggle.com/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews
2. Rotten Tomatoes dataset - https://www.kaggle.com/c/sentiment-analysis-on-movie-reviews/data?select=train.tsv.zip
3. **(Choose any one of ~24 categories)** from Amazon Product Review Dataset - http://jmcauley.ucsd.edu/data/amazon/

Data Rules:

1. **Choose any one of the following:**
    a. **Any of 1,2,3.**
    b. **If choosing 3, you can choose any one category out of ~24 categories.**
2. **The chosen dataset should contain at least 25k samples [choose at most 25k samples if no. of samples > 25k]**

3. ***A single dataset can be chosen by a max of two groups. Please refer to the following sheet for datasets chosen by other teams. It's a first come first serve basis.***
4. **Update dataset choices here.**

Module - 1 (Statistics):

Tasks:-

1. Explain the text processing pipeline adopted by you.
2. Generate term statistics: [You can use nltk, Stanford parsers, spacy etc.]
    a. Vocabulary size with word frequencies
    b. N-grams
    c. POS collections (Like Nouns - frequency, Verbs - frequency, Adverbs - frequency etc.)
    d. Most Frequent Noun Phrases
    e. Most Frequent Verb Phrases.
    f. NERs with their frequencies and types.
3. Which set of terms best describe your corpus? How did you arrive at it?
4. Create a visualization and justify your answer for 3.
5. Plot a graph of the frequency of word vs rank of the word. How would you characterize the relationship? For your reference - consult "Zipf's law" – and determine the best fit for your corpus?

Module - 2 (Topic analysis):

Tasks:-

1. Extract the topics from the corpus using LDA and present a visualization of them in your report. [You can use any ML library e.g Mallet]
2. For each topic: Present 20 most significant sentences.
3. Analyse whether the topics extracted make sense. Do you feel the topics cover the entire dataset correctly? Justify your claim. If you find anything is missing, please state what is it and why do you think LDA might have missed them. Do you have any suggestions for improvement?