IT585 Advanced Machine Learning
Lab8
LDA for Topic Modeling

Instructions:

1. You have to code the solution in Google colab
2. You can use inbuilt libraries from python
3. Your plots, code, any insights, observations written as text should be submitted as one ipynb file to google classroom
4. Deadline : Mar 20,2024 11:59 PM IST
5. Name of your file should be : yourrollno_lab8.ipynb

In this lab we will implement topic modeling using LDA on the Trip Advisor Hotel Reviews dataset which is available on kaggle [1]. We will focus on the text reviews. Each review can be considered as one document. Before application of LDA for topic modeling, some text-preprocessing is required.

Do the following:

1) Remove the stopwords
2) Apply tokenization, lemmatization and stemming to the text data.

Read about the above procedures. These are standard text preprocessing techniques for many NLP tasks. You can use the python nltk toolkit to perform the above tasks. You can also display a word cloud just to gain an idea of word distribution in your corpus.

Next create your dictionary and also convert the documents to the bag of words model. You can use the gensim module for the tasks. You can also use `sklearn.feature_extraction.text` count vectorizer to create document term matrix. Apply LDA on the data using either gensim model or sklearn.decomposition `LatentDirichletAllocation` module. Check the results for different values of K i.e. no. of topics that you assume. Looking at the high probability words in the obtained topics try to analyze the different reviews and explain your analysis and observations. You can use various plots or some print statements to support your conclusions.

[1] Alam, M. H., Ryu, W.-J., Lee, S., 2016. Joint multi-grain topic sentiment: modeling semantic aspects for online reviews. Information Sciences 339, 206–223.