Instructions

1. Follow the instructions in each question carefully.
2. Only two files should be uploaded in canvas without zipping them. One is ipynb file and other one html output of the ipynb file. No other files should be uploaded
3. Any assignment submitted using other python IDEs are not considered for grading.
4. If there are any issues in accessing the links to datasets, you can search for the same dataset from any repositories and use them.
5. Incorrect Assignment Set submitted will not be considered.

NLP Assignment 2

Set A

Link to the Dataset:

https://drive.google.com/file/d/1x0oiWyLUns9002jTDj2CzIE6yqbglN__/view?usp=sharing

Note: Use first 10000 rows of dataset from the original dataset given

Description of Data:

This is the Amazon Fine food review dataset.Each record consists of the following attributes:
The column or features in the dataset:

- Id
- ProductId — unique identifier for the product
- UserId — unqiue identifier for the user
- ProfileName
- HelpfulnessNumerator — number of users who found the review helpful
- HelpfulnessDenominator — number of users who indicated whether they found the review helpful or not
- Score — rating between 1 and 5
- Time — timestamp for the review
- Summary — brief summary of the review
- Text — text of the review

1. Perform EDA and necessary pre-processing steps in dataset. (2 Mark)

2. Using the LDA algorithm create the Topics (Min 10) for the Corpus (2 Mark)
   NOTE: Use Text Column

3. Compute the coherence score and print Topics Extracted. (2 Mark)

4. Visualize the topics (1 Mark)

5. Plot the dependency parser for any **two random sentences** from the entire corpus/dataset that has at least 10 words in the sentence. Make sure that dependency parser looks good and should visually understandable. (3 Mark)