#### Instructions

Attached is a dataset with Movie and TV reviews. It has 4 columns. The first represent the domain which is Movies\_TV, the second represent the label as positive (POS) or negative (NEG). Third column is rating from 0 - 5 with 0 being the least likable and 5 as the most likeable. The fourth column is the content of the review. Your tasks for the first assinment are;

1) Read data from the dataset and separate it into 4 columns with each row as a review submitted.

2) Preprocess the reviews content by;

a) Removing unwanted whitespaces

b) Normalizing case

c) Removinig stopwords

d) Removing punctuations

e) Stemming words

f) Lemmatizing words

3) Find all possible unigrams, bigrams and trigrams for the first review (its computationally expensive)

4) Find the probabilities of unigrams, bigrams and trigrams

5) Provide the following information about reivews;

a) total tokens

b) Vocabulary (unique tokens) before and after preprocessing

c) Average length of a review

d) Average length of tokens within a review