Problem Statement

Dataset 1:

The dataset shared represents a subset of stack overflow dataset. The dataset consists of (textId, Text) where text represents the questions asked in stackoverflow and textId is an identifier for the text. There are 32,577 questions in this file.

Dataset 2:

The dataset shared represents abstract of various research papers. The dataset consists of (Id, Abstract) for different technical papers.

Task 1:

- This task corresponds to identification of significant bigrams in each of the collections (Dataset-1 and Dataset-2).

- Assign score to each bigram $w_1 w_2$ .Mentioned below are two scoring methods:
  - Based on collection frequency
    - $score_{Freq}(w_1 w_2) = Number\ of\ occurrences\ of\ the\ bigram\ in\ the\ collection$
  - Based on collocation score
    - $score_{PMI} = \log_{10} \frac{P(w_1 w_2)}{P(w_1)p(w_2)}$

- For each of the scoring methods, for each dataset, do the following
  - Output the top 30 bigrams
  - Sort the bigrams in decreasing order of their scores. Plot these sorted scores.

- Analyze the plots and include your observations in the report. If you have any observation form the list of bigrams, include that as well.

Task 2:

Create a plot where x-axis represents bigrams based on collocation score and y-axis represents frequency score. Hence each bigram can be

considered as a point in this 2D space. Plot the points and report your observations from this plot.

Briefly mention (in 2-3 points) how this bigram analysis can be useful.

Task 3:

Taking clue from your observations, can you think of other scoring methods identifying the significant bigrams? If yes, then briefly mention about that, and output the top-30 bigrams according to that scoring method.

Give this method an appropriate name.