**IR ASSIGNMENT-2**

**DOCUMENTATION**

Functions Used:

* clean\_sentence: This function removes all the characters that are not letters or numbers, converts the sentence to lowercase and removes the stop-words occuring in the document.
* clean\_dataframe: This function removes all the documents that does not have plots and applies the clean\_sentence function to the documents that have plots.
* random\_coefficients: Generate random coefficients that are used to build hash functions for minhashing.
* getTriangleIndex: This function is used to take all the combinations of columns to calculate similarities