1. **Read the dataset using spark**
    a. using databricks.com by creating a cluster and loading the data directly
    b. Check the number of partitions for each part of data
    c. Repartition the data into 2 partitions because the cluster has one node 2 cores
    d. Store the data as parquet
    e. Reload the data again but from parquet files which save a lot of time
        i. JSON files take about 16.87 minutes
        ii. Parquet files take about 4.76 seconds

2. **Exploratory data analysis (EDA)**
    a. Print Schema to check the available columns
    b. Get most important columns (paper_id, title, abstract, body_text, back_matter)
    c. Add "source" column and set the value with "comm, uncomm, bio"
    d. Check about null and empty for all of these columns and this is the result
        - null titles: 0
        - empty titles: 617
        - null abstracts: 0
        - abstracts has less than 100 char: 1624
        - abstracts has less than 100 char and not empty: 107
        - null body_text: 0
        - body_text has less than 10000 char: 1188
        - body_text has less than 10000 char and not empty: 1188
        - null back_matter: 0
        - back_matter has less than 100 char: 3776
        - back_matter has less than 100 char and not empty: 974
    e. check about duplicate title and there are about 500 duplicate titles

f. check about language and there are about 100 documents non English (we filter step by step so these 5 after removing null duplicates and short body_text) using langdetect library

3. **Preparation and Cleaning the data**
    a. drop null and empty and short body text records
    b. drop duplicate titles
    c. keep only English document

4. **Preprocessing**
    a. Remove punctuations
    b. Remove stop words
    c. Remove custom stop words
    d. convert text to lower case

5. **Vectorization**

    We use TF-IDF.

    Here we try without using the feature_num parameters then it produce about 250k features, but unfortunately when apply PCA, it limited to only 16K features, so we limit the features num to 16K but then PCA has out of memery exception so then we decrease the features_num again and again with a lot of values (50000,10000,5000,1000,500)

6. **Clustering**
    a. Use Kmeans
    b. Use PCA to reduce dimensions