

# **Report on Warm-up project PDM -Spring 2021**

**By:** Sai Teja Chennu (16314417)

Anusha Yanamala (16317767)

Aravind Yanamala (16315604)

Chaitanya Tummala (16315816)

Amarnath Reddy Tatireddy (16314179)

## **Introduction:**

A histogram is basically used to represent data provided in a form of some groups. It is an accurate method for the graphical representation of numerical data distribution. It is a type of bar plot where the X-axis represents the bin ranges while the Y-axis gives information about frequency. It can be almost any type of data. The written data is transposed onto a chart that has vertical blocks; the number of blocks depends on the categories of data collected.

## **Project Description:**

Generating a Histogram of the size of news articles under the news category 'US politics'.

## **Installation Steps:**

1. Importing required libraries like pandas, nltk.
2. Download stopwords using nltk library i.e., `nltk.download('stopwords')`.

### 3. Mounting google drive.

#### Task1:

Steps to read the text file for extracting words and storing the text in list as shown below:

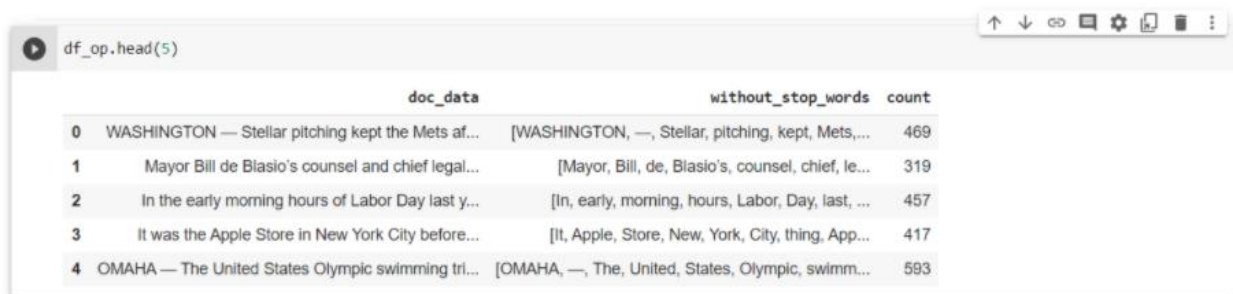
- Reading the data in the NY times text file which is in the path  
/content/drive/MyDrive/Warm-up Project\_Prin of Big Data  
Mgmt/nytimes\_news\_articles.txt
- Adding the text to CSV file to perform operations.

Create column doc\_data for storing the text in the form of list.

And Stopwords downloaded from the nltk package and stored in stop(list) variable.

#### Task2:

Removing stop words from the list in the article file.



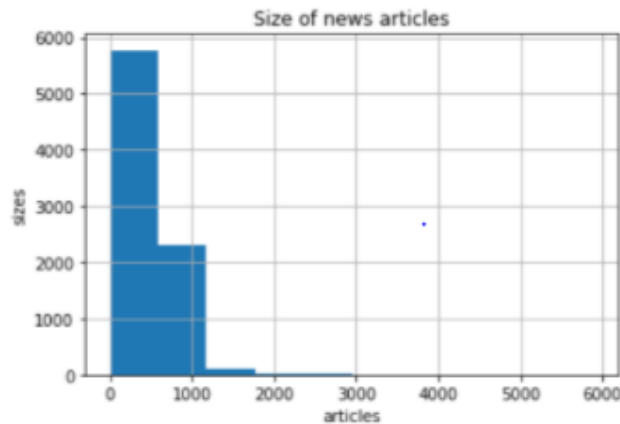
The screenshot shows a Jupyter Notebook interface with a code cell containing `df_op.head(5)`. Below the code cell, a table displays the first five rows of the DataFrame. The table has three columns: `doc_data`, `without_stop_words`, and `count`. The rows are indexed from 0 to 4.

	doc_data	without_stop_words	count
0	WASHINGTON — Stellar pitching kept the Mets af...	[WASHINGTON, —, Stellar, pitching, kept, Mets,...	469
1	Mayor Bill de Blasio's counsel and chief legal...	[Mayor, Bill, de, Blasio's, counsel, chief, le...	319
2	In the early morning hours of Labor Day last y...	[In, early, morning, hours, Labor, Day, last, ...	457
3	It was the Apple Store in New York City before...	[It, Apple, Store, New, York, City, thing, App...	417
4	OMAHA — The United States Olympic swimming tri...	[OMAHA, —, The, United, States, Olympic, swimm...	593

The above diagram displaying the first five rows with out stop words and word count.

### Task3:

Plotting the Histogram size of news articles with a number of words in each article belongs to the US Politics using pandas object. DataFrame The pandas object holding the data.



The above figure shows that X-axis represent the No.of articles and Y-axis represent the size of the article after removing the stopwords.

### GitHub Url:

[https://github.com/chennusaiteja/Computer-Science-5540-0001-Prin-of-Big-Data-Mgmt/tree/main/Warm-up%20Project\\_histogram](https://github.com/chennusaiteja/Computer-Science-5540-0001-Prin-of-Big-Data-Mgmt/tree/main/Warm-up%20Project_histogram)

### Wiki Url:

<https://github.com/chennusaiteja/Computer-Science-5540-0001-Prin-of-Big-Data-Mgmt/wiki/Warmup-Project-Histogram>