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

**By**: Sai Teja Chennu

(16314417)

Anusha Yanamala Aravind Yanamala

(16317767) (16315604)

Amarnath Reddy Tatireddy Chaitanya Tummala

(16314179) (16315816)

**Introduction:**

*Word Cloud is a data visualization technique used for representing text data in which the size of each word indicates its frequency or importance. Significant textual data points can be highlighted using a word cloud. Word clouds are widely used for analyzing data from social network websites.* Data for this warmup project is collected from New York Times articles.

**Spark Installation Steps:**

1. Installing spark using command !pip install findspark.
2. Importing required libraries for word cloud.
3. Creating spark session and sparkcontext.
4. Mounting google drive.
5. Uploading nytimes\_news\_articles.txt in the specified path to read the data from the file.

**Task1:**

To build a word cloud we used Apache Spark and Google Collab platform. Listed top 100 words from all the articles such as Sports, World, US, Business.

**Steps:**

1. Removing stopwords from the article.
2. Extracting top 100 in dictionary.
3. Plotting the wordcloud.

**Task2:**

Here we fabricated a word cloud top 5 new category depending upon the most news in the articles such as politics, sports, world, arts and films.

**Steps**:

* Plotting Word cloud for top 5 news category.

# Finding the top5 Categories that has the most of news articles.

* Finding the URL contains the category which the news belong.

**Task3:**

Here we gathered top 10 words that are divided between the highest number of news articles of a similar classification. For instance the word Country appeared 2191 times in the article.

**Steps**:

* Finding the top 10 words that are shared among the highest number of news articles of the same category.

1. sports
2. world
3. Us
4. business
5. nyregion

**GitHub Url:**

https://github.com/chennusaiteja/Computer-Science-5540-0001-Prin-of-Big-Data-Mgmt/tree/main/Warm-up%20Project/SOURCE%20CODE

**References:**

<https://www.geeksforgeeks.org/generating-word-cloud-python/>