PRICIPLES OF BIG DATA MANAGEMENT
COMP-SCI 5540 (GROUP-19)
WARM-UP PROJECT – Fall 2021

TEAM MEMBERS

RAVILLA HARISH	AVINASH KONGARA	PRANEETH CHEEKATI
SAI KIRAN MERUGU	RAVI CHANDRA THOTA	VARUNDEV NUTALAPATI

Word Clouds: Word clouds are visual representations of word frequency that give greater prominence to words that appear more frequently in a source text. The larger the word in the visual the more common the word was in the documents.

Data used for the project:

- Kaggle Source
- **File Name:** nytimes_news_articles.txt (Used Databricks File System as file storage).

Environment: Databricks included Apache Spark.

Find Spark Installation: Command used - ! pip install findspark.

Word Cloud Installation: Command used - pip install wordcloud.

TASK 1: Build a word cloud for NY Times articles

- Listed top 100 words used in all articles.
- Excluded stop words.
- Extracted top 100 words in dictionary.
- Plotted the word cloud using matplotlib.

TASK 2: Build a word cloud for top 5 news category

- Listed the top 5 Categories based on the URL's listed in the source file.
  - Top 5 Categories are sports, world, us, business, nyregion.
- Listed the top 100 used based on the categorized articles.
- Plotted the word cloud.

TASK 3: List the top 10 words that are shared among the highest number of news articles in the same category

- Listed all the categories from the data file.
- For each category, listed the words which are commonly shared among articles.
- Then listed the top 10 frequency words for each category.

GITHUB URL: PROJECT LINK

DATABRICKS URL: WORKSPACE URL