Task: Popular User Identification and Word Frequency Analysis Using PySpark

In this task, you will carry out a complete MapReduce-style data analysis workflow using PySpark, focusing on processing and analyzing tweet data from a CSV file.

Dataset:

You are provided with: Amazon_Responded_Oct05.csv, containing approximately 400,000 tweets. You will work with the following three columns:
- user_id_str: unique identifier of the user
- user_followers_count: number of followers the user has
- text_: the text content of the tweet

Objectives:

1. Identify popular users who have more than 5,000 followers.
2. Analyze tweet content from these users to extract the Top 10 most frequently used words.

Workflow Steps:

1. Load and Clean the Data
- Load the CSV into a Pandas DataFrame, clean missing values and special characters.
- Convert it into a Spark DataFrame for distributed processing.

2. Select Relevant Columns
- Extract the columns user_id_str, user_followers_count, and text_.

3. Handle Duplicate Users
- Some users appear multiple times with different follower counts.
- For each user, keep only the record with the maximum follower count.

4. Filter for Popular Users
- Filter the dataset to keep only users with more than 5,000 followers.

5. Word Frequency Analysis (MapReduce Style)
- Collect tweets from popular users and clean the text:
* Remove URLs, punctuation, and numbers
* Convert all text to lowercase
- Use RDDs to:

* Split tweets into words (Map)
* Count word frequencies (Reduce)
- Extract the Top 10 most frequent words used by popular users.

6. Save Output

- Save the Top 10 words and their frequencies into a new text file (output.txt) in your Colab environment.

Expected Output:

- A printed list of the top 10 most common words.
- A saved file /content/output.txt containing those words and their counts.