Mid-Term Assignment

Deadline: 14th June 2024

Instructions

1. Solve the problems given in the Insights section using map-reduce steps only.
2. Using Pandas library or other Dataframe APIs like Spark Dataframe APIs to solve the problems will not be accepted as a solution.
3. Each step should have a heading. The steps should be explained in detail using comments or markdown text.
4. Databricks platform must be used for this assignment.
5. Upload the data file to the location: ***/FileStore/tables/jio2025/Sample_tweets.csv***
6. Use PEP8 standards for coding (Refer to link).
7. Submit the notebook as .ipynb file. It should be named as BDE_MidTerm_Groupn.ipynb, where n is the group number.
8. The notebook should have the Names and IDs of all members of your group at the beginning.
9. Notebooks that are not properly documented or commented will be penalized
10. Please do not copy code from your friends or generate code using ChatGPT. Plagiarism cases may be penalized.

Direct copying will result in zero marks for all the groups involved. Also, generating the code from GenAI tools will result in zero marks.

Dataset

This dataset comprises the Twitter discourse surrounding the Indian Premier League (IPL) from August 15 to November 11, 2020.

The data consists of 7 columns which are:

- **user_name** (The username of the Twitter account that posted the tweet)
- **text** (The actual tweet along with the hashtag)
- **user_followers** (The number of followers the user had at the time of posting the tweet)
- **date** (The date when the tweet was posted)
- **user_location** (location specified in the user's Twitter profile)
- **source** (The device/platform/application used to post the tweet)
- **user_description** (The bio or description provided by the user in their Twitter profile)

Download the dataset from the following link
Link to Data

Insights

1. Find top 5 hashtag trends for different months. The results should be in the following format:
    a. Month, rank, Hashtag, number of tweets
2. Find the trends of tweets (number of tweets) for each team over different months. The results should have
    a. Team name, month, number of tweets
    b. A list of abbreviated names of the teams can be used to filter tweets (e.g. CSK, MI, RCB etc.)
3. Identify the top 5 tweets for each hashtag, ordered by the number of followers the users have, who have tweeted. It should contain unique set of users. The result should contain:
    a. Hashtag, list of 5 tweets, the username who has tweeted and the number of followers.

4. Create an inverted index of hashtags and the tweets so that given a hashtag, you can print the list of tweets with the hashtags sorted by the months. The result should contain
    a. Hashtag, month, list of tweets (actual text)
5. Define 3 Insights of your own.

Guidelines:

- For each of the above problems, print a few records of the final result using take(5).
- Provide sufficient description or comments for each line of the code to make it more readable.
- Not following PEP 8 standard to write code will be penalized.
- There will be 30% of weightage given for documentation and clarity of code using the guidelines.