**Analyzing Twitter Dataset using**

**Apache Spark**

**Scenario:**

Nowadays, data is growing and accumulating faster than ever before.

One of the main sources of data today are social networks. This Capstone gives a real-life example: dealing, analysing, and extracting insights from social network data using one of the most important big data solutions out there—Apache Spark.

**Brief introduction:**

**Twitter Dataset**:

This Twitter Dataset is a txt file which has a high volume of tweets on particular subjects in specific locations (England & India).

**Output:**

Read the tweets from the twitter dataset and display the top 10 trending hashtags with their counts.

**Submission:**

[**Here**](https://forms.gle/T3Zpu7Fyq7eN9Bnb8) **is the Submission form where you have to upload these two things**.

* Word file with the code snippet explaining your approach for each code snippet.
* Upload to the Google Form.

**Evaluation Rubric:**

| **Criteria** | **Meets expectation** | **Does not meet expectation** |
| --- | --- | --- |
| Data loading, understanding and preparation(25%) | Understand the problem statement and objective. | Importing the libraries which are not used in code. |
| Load the dataset and Extract the features what is required what is not required | Proper data cleaning not done |
| Neatness of Coding (30%) | The code is well commented with meaningful variable names. It is well structured with the use of data structures/dataframes. | The code is difficult to read and understand |
| Analysis of results (25%) | Proper display of the top 10 Trending hashtags.  Proper display of all tweets for the top 5 trending hashtags.  Proper display of tweets related to Olympics. | Not able to display the display the Trending Hashtags |
| Doc file (30%) | Proper documentation with the code snippertand explaining your approach for each code snippet | Proper explanation is not done |