

Course project

[Start Assignment](#)

Due Dec 20 by 11:59pm **Points** 100 **Submitting** a file upload

Available Nov 10 at 12am - Dec 20 at 11:59pm about 1 month

Due 12/20/2021

Form a group of no more than 4 students.

Tasks:

1. Do twitter streaming analysis using Spark streaming
2. Train machine learning algorithms to perform sentiment analysis based on the training data and classify stream of tweets using Spark MLlib
3. Store the fetched tweets data and their classification results in a database
4. Submit your codes and report to Canvas
5. Submit the screenshot of the code to do all the tasks (the running code) to Canvas as well

You will need to:

1. Register on Twitter Apps
2. Go to <https://developer.twitter.com/en/apply-for-access.html> and apply for a twitter developer account.
3. Login at <https://apps.twitter.com/>
4. Click "Create New App", fill out the form, and click "Create your Twitter application". For the Website URL, you can put any URL here. You only need to fill in all the required places.
5. On the next page, click on the "API keys" tab, and copy your "API key" and "API secret".
6. Scroll down and click "Create my access token", copy your "Access token" and "Access token secret".

Training data could be obtained from here:

<https://docs.google.com/file/d/0B04GJPshIjmPRnZManQwWEdTZjg/edit>

With the following format:

The data is a CSV with emoticons removed. Data file format has 6 fields:

0 - the polarity of the tweet (0 = negative, 2 = neutral, 4 = positive)

1 - the id of the tweet

2 - the date of the tweet

3 - the query. If there is no query, then this value is NO_QUERY.

4 - the user that tweeted

5 - the text of the tweet

Useful links:

<https://spark.apache.org/docs/latest/streaming-programming-guide.html>

[\(https://spark.apache.org/docs/latest/streaming-programming-guide.html\)](https://spark.apache.org/docs/latest/streaming-programming-guide.html)

<https://docs.python.org/3/library/re.html> (<https://docs.python.org/3/library/re.html>)