Spark Streaming - Twitter

Spark Streaming is an extension of the core Spark API that enables scalable, high-throughput, fault-tolerant stream processing of live data streams. Data can be ingested from many sources like Kafka, Flume, Kinesis, or Twitter, and can be processed using complex algorithms expressed with high-level functions like map, reduce, join and window. Finally, processed data can be pushed out to filesystems, databases, and live dashboards. In fact, you can apply Spark's machine learning and graph processing algorithms on data streams. To know more about Spark Streaming: https://spark.apache.org/docs/latest/streaming-programming-guide.html

In this project, I will build a simple application that reads online streams from Twitter, then processes the tweets using Apache Spark Streaming to identify hashtags and, finally, returns a specific trending hashtag and conducts simple sentiment analysis.

Step 1: First you need to create your own credentials for Twitter APIs. Here are some tutorials to do this:

Step 2: Build your Apache Spark Streaming Application on Twitter and collect **at least 1000** tweets with hashtag ***#hospital***.

Step 3: Conduct simple sentiment analysis by identifying positive or negative words using Bing Liu Opinion Lexicon.

For example, if there are three positive words and two negative words in a tweet, then the sentiment score is 3-2=1.

Step 4: Report the sentiment score of each tweet and save to a .csv file.

Output

The .CSV file will have the sentiment score of collected tweets in a format as:

Tweet_content	Sentiment score
…	…