Twitter Sentiment Analysis

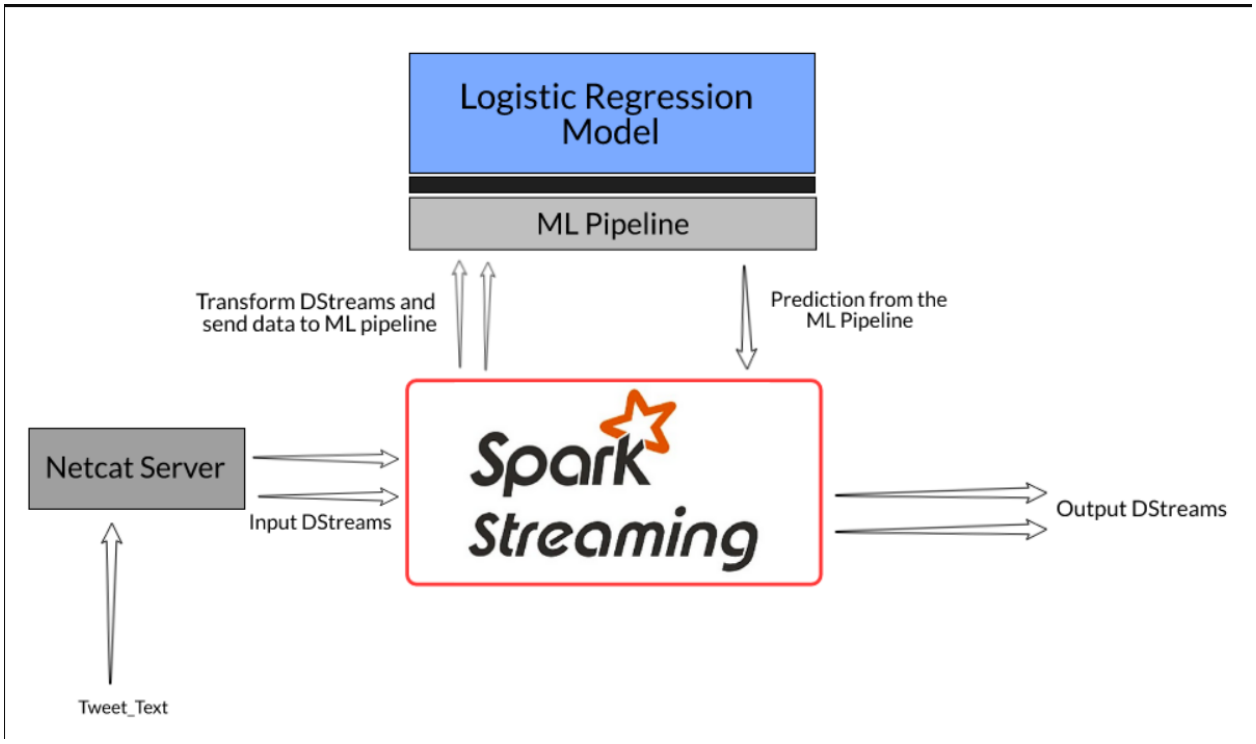
The objective of this task is to detect hate speech in tweets. For the sake of simplicity, we say a tweet contains hate speech if it has a racist or sexist sentiment associated with it. So, the task is to classify racist or sexist tweets from other tweets.

Formally, given a training sample of tweets and labels, where label '1' denotes the tweet is racist/sexist and label '0' denotes the tweet is not racist/sexist, your objective is to predict the labels on the test dataset.

Let's say we receive hundreds of comments per second and we want to keep the platform clean by blocking the users who post comments that contain hate speech. So, whenever we receive the new text, we will pass that into the pipeline and get the predicted sentiment.

You have to use itversity for this project.

The architecture of this project :

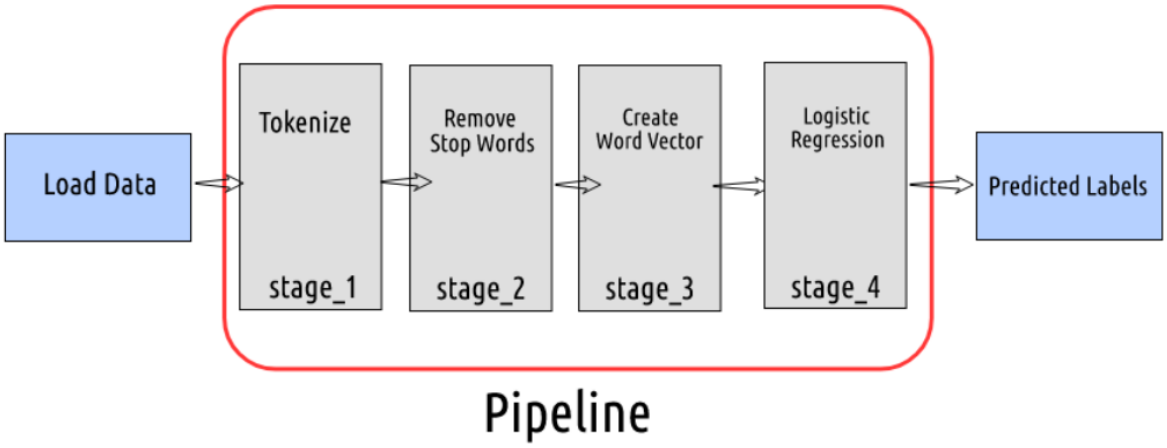


The Data:

1. train.csv - For training the models, we provide a labelled dataset of 31,962 tweets. The dataset is provided in the form of a csv file with each line storing a tweet id, its label and the tweet.

The Task:
You have to create a pyspark script (.py) in which you will do the following:

1. Perform basic analysis on the data (e.g. Number of rows, % of 1 and 0 in target variable, etc.)
2. You have to create a machine learning pipeline in spark. Here is a sample architecture for the pipeline, you are free to create your own as well.



3. Now, initialize the Spark Streaming context and define a batch duration of 3 seconds. This means that we will do predictions on data that we receive every 3 seconds.
4. Specify a port and hostname in the dstream that you created above.

Run the script in a terminal.

Create a netcat server that will be used to send text that you can enter.
`nc -lk port_number`
Use the above command in another terminal to then send strings to your spark script which will be listening to a port. This port number will be the one you specified for your dstream.

Submission Details:
You have to submit your script in a zip file. The evaluation will be done based on your script and understanding of how you . Do not worry if netcat does not work because marks will be given based on your script.