**Spark Lab Assignment – 2**

**Name: Aravind Sheri, Class Id – 18, Team – 06**

**Name: Durgam Dileep, Class Id – 03, Team – 06**

**Task- 1:**

**Classification Algorithms in Spark:**

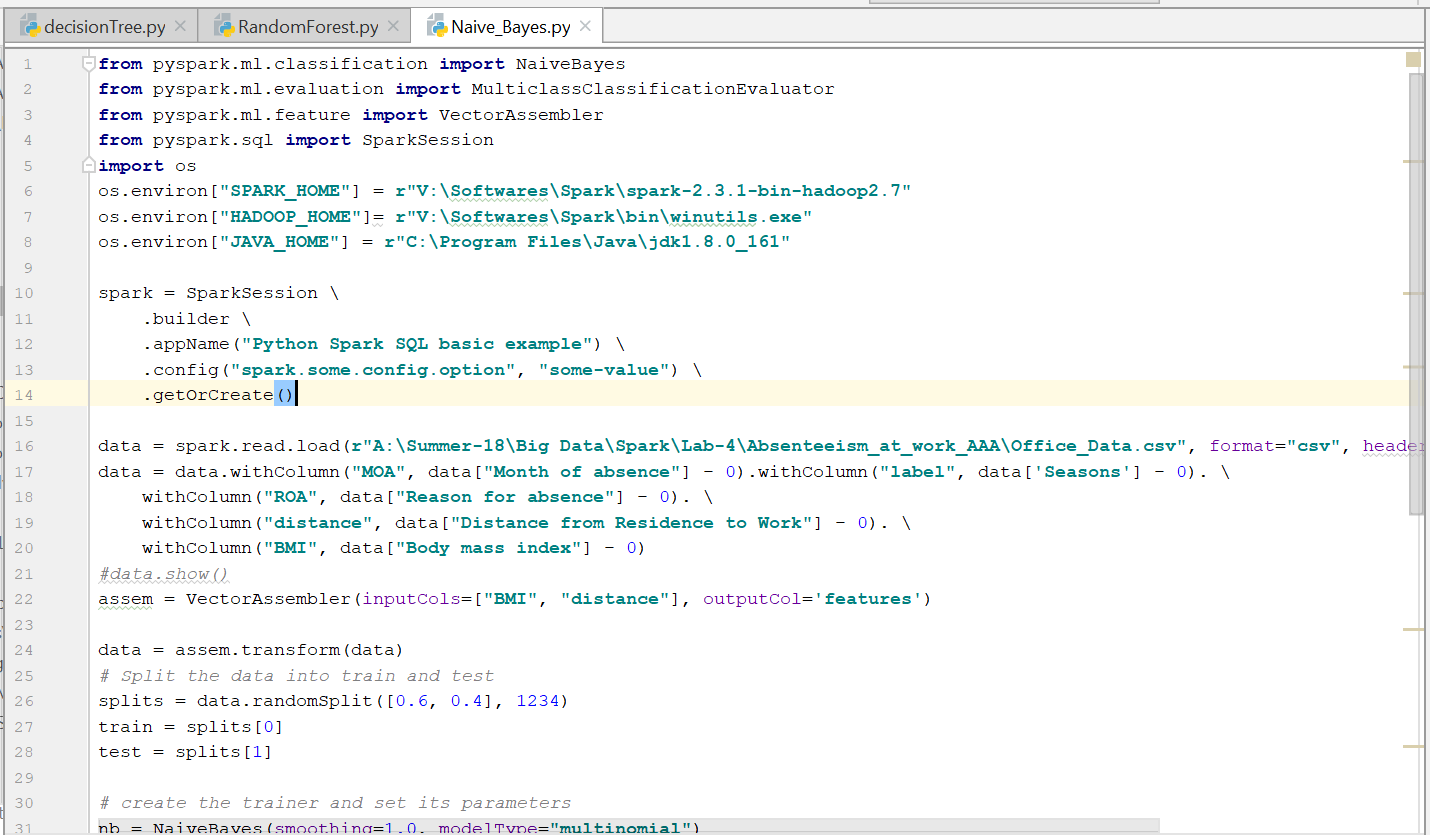
**Objective:** Apply Classification Algorithms on the data sets that are provided in the lab and predict the values based on the values that are given in the dataset.

**Approach:** Picked the Absenteeism as the data and loaded the data into the spark classification algorithm and fit the data into the model and predict the value based on the previous data. Different Classification algorithms are used to predict the data and calculate the loss and the accuracy to find the model accuracy and the various factors that affect the algorithm.

**Inputs:** CSV data is used to train the model by loading the data into it.

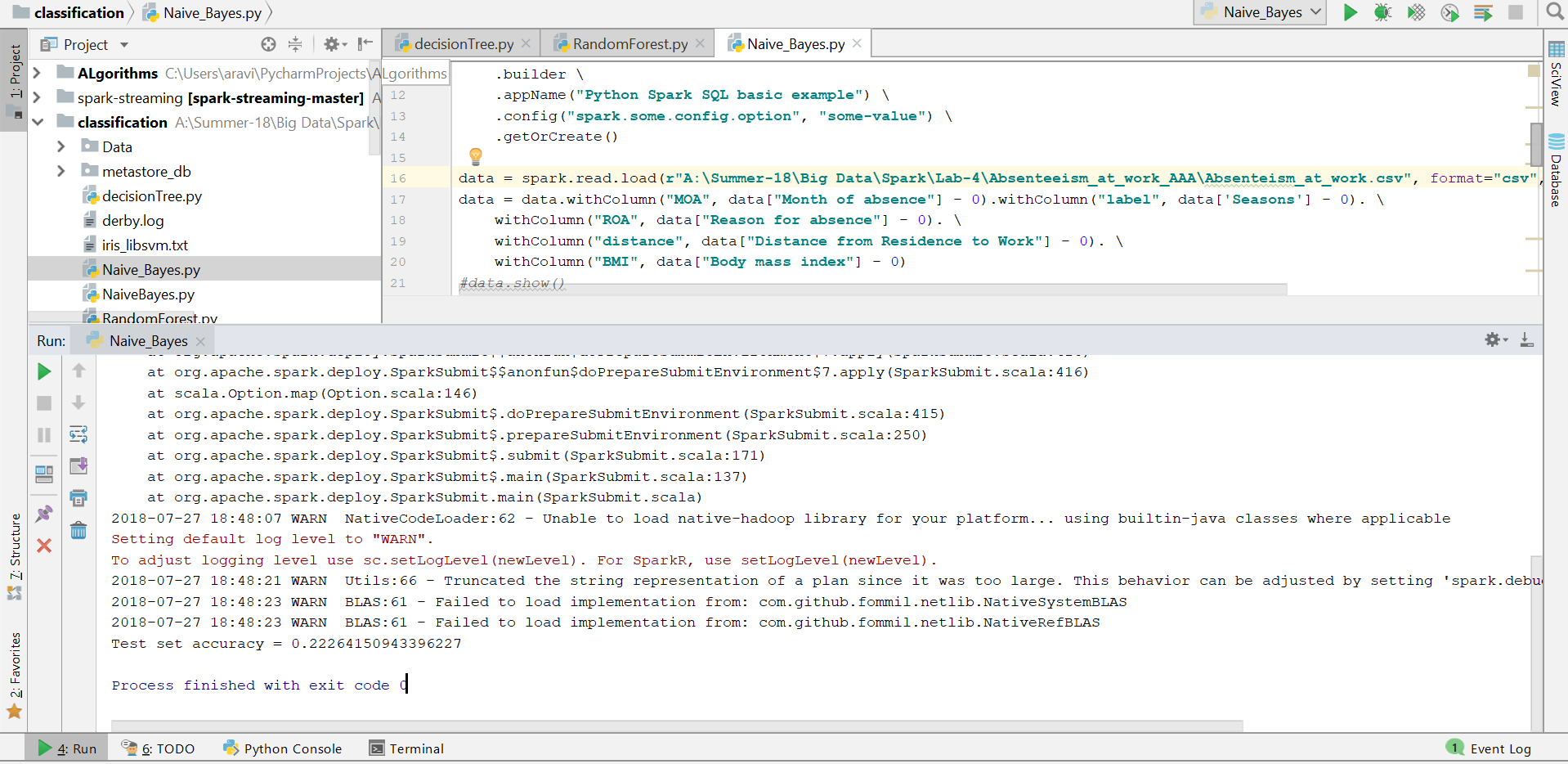
1. **Naïve Bayes Algorithm:** Naïve Bayes classification algorithm is an algorithm for binary and multi class classification problems.

**Code Screenshots:**



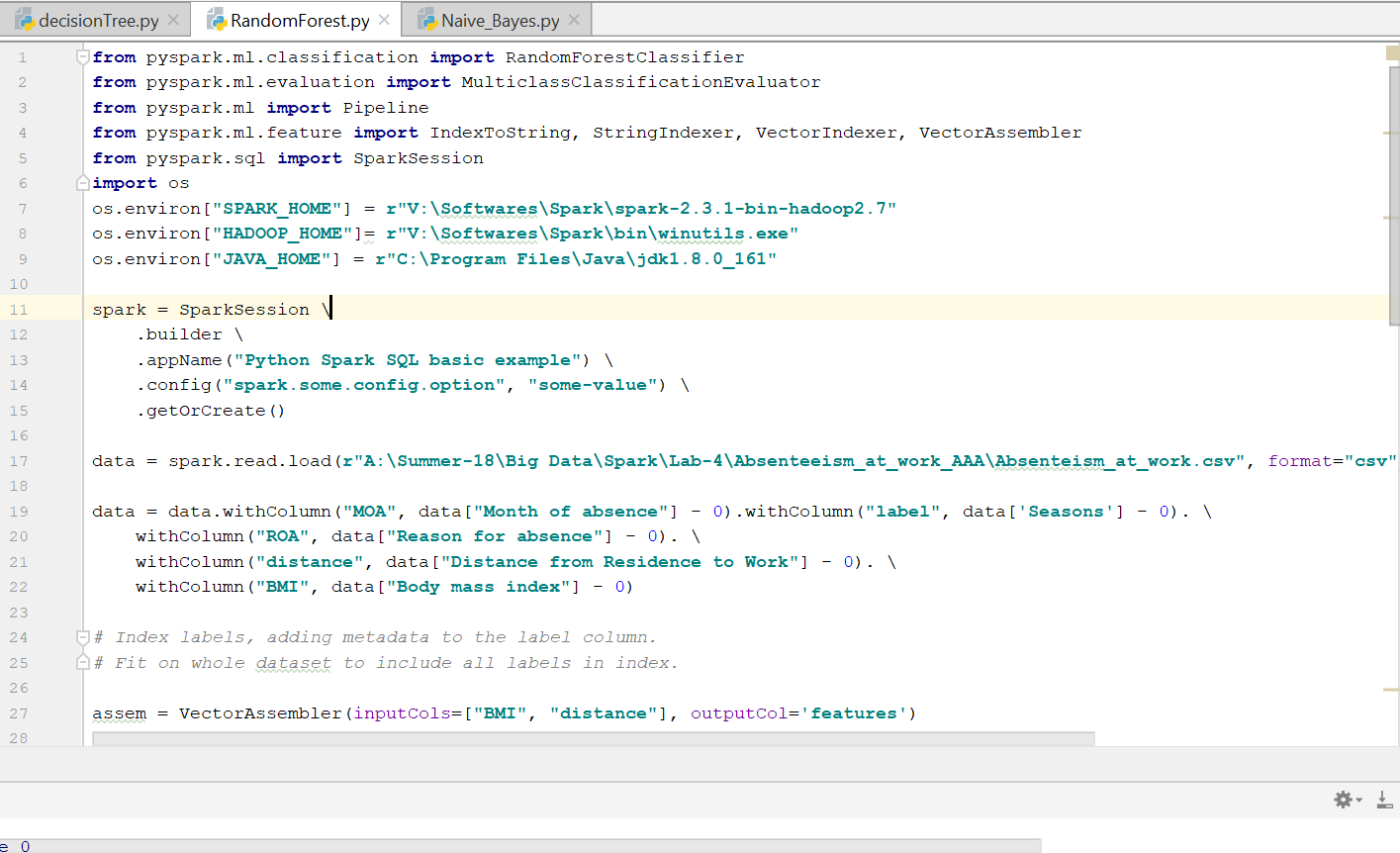


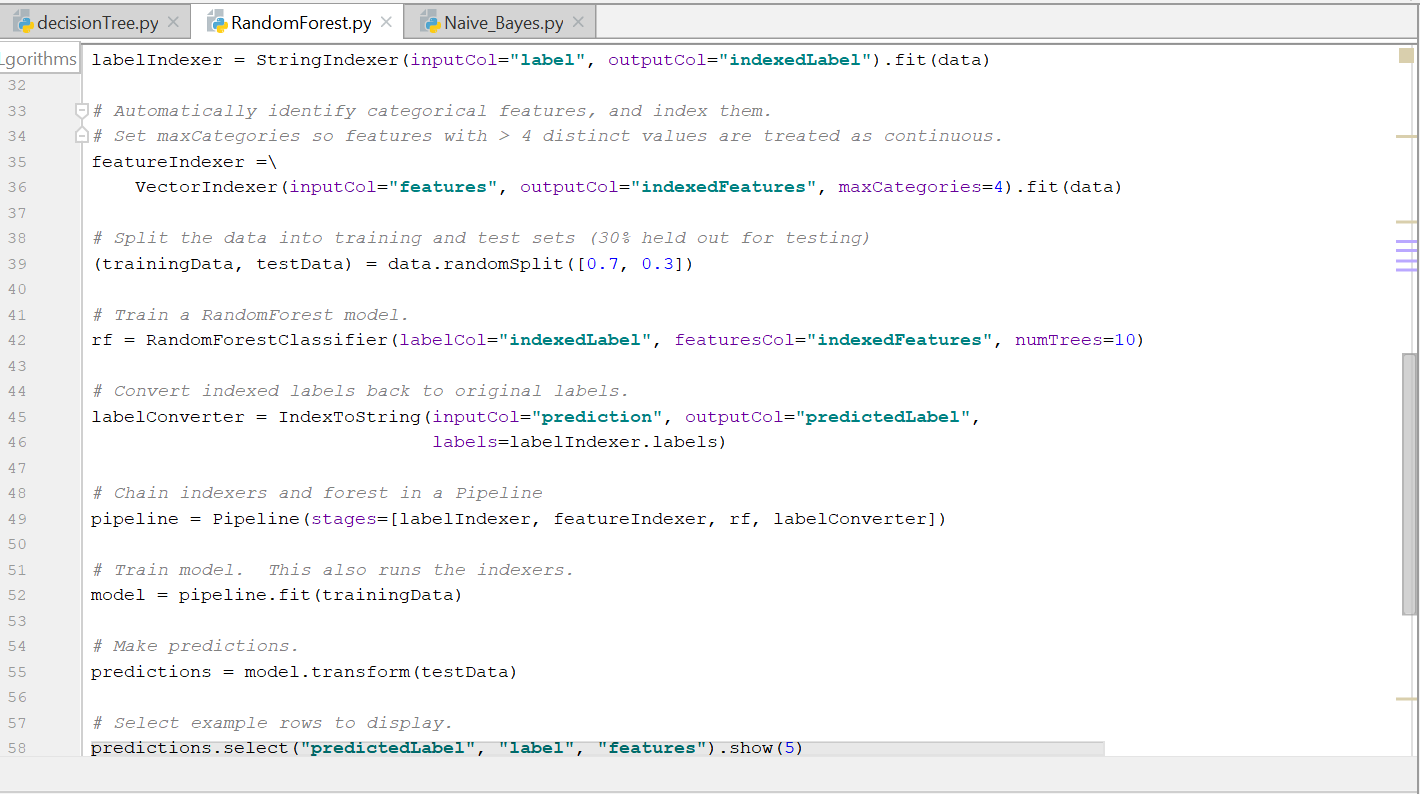
**Naïve Bayes Algorithm Output:**

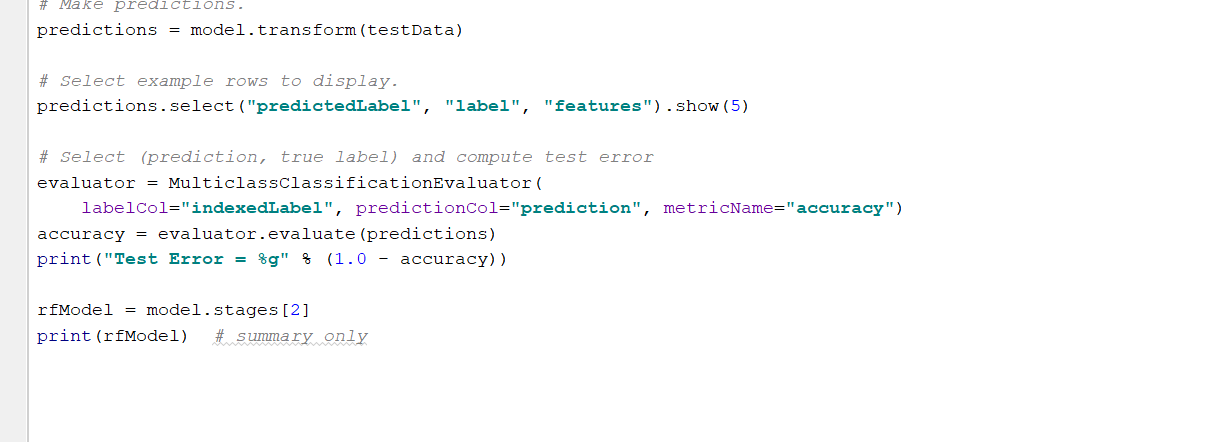


1. **Random Forest Algorithm:** Random forest classifier creates a set of decision trees from the input provided dataset. It then aggregates the results and various votes from different decision trees and decide the final class of the test object.

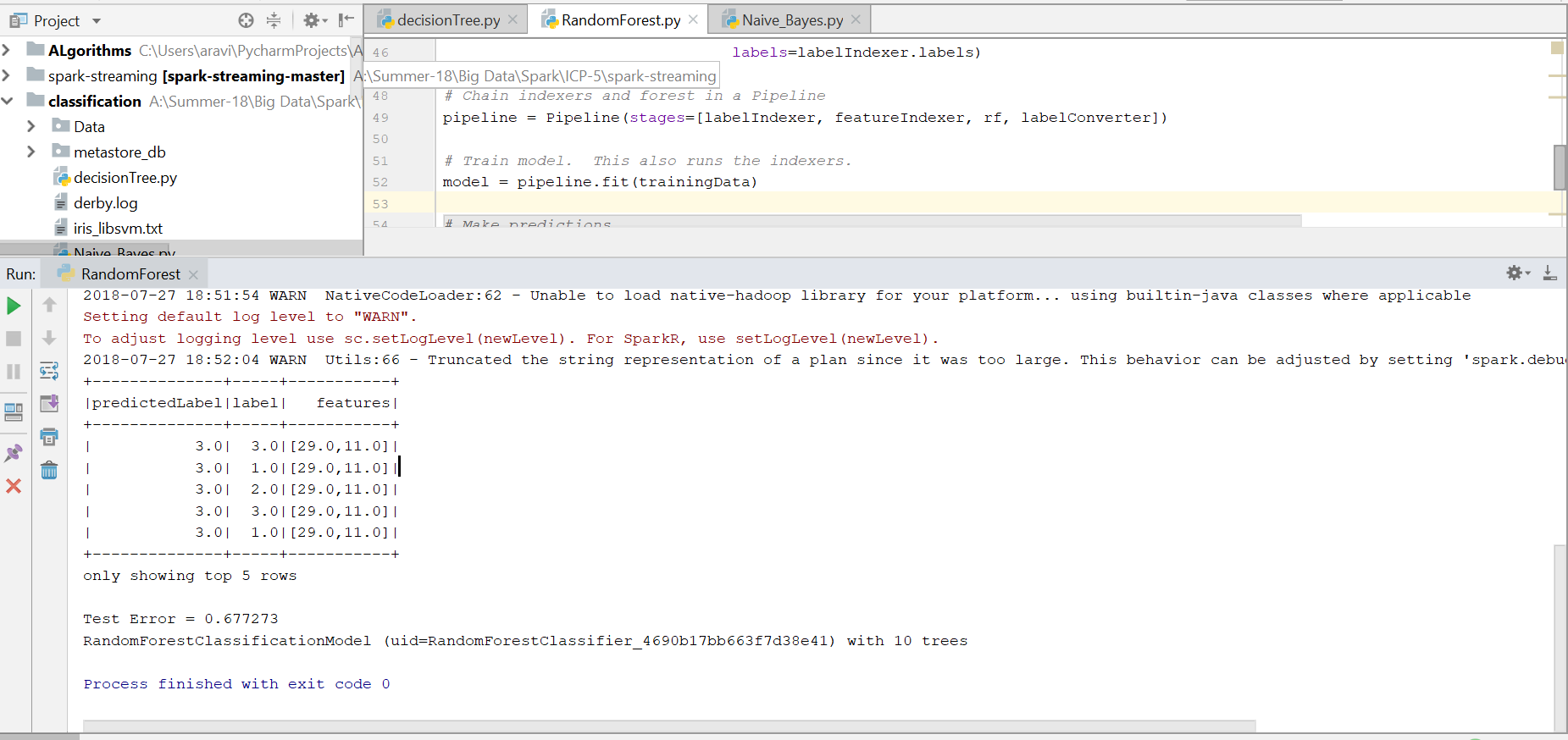
**Code Screenshots:**





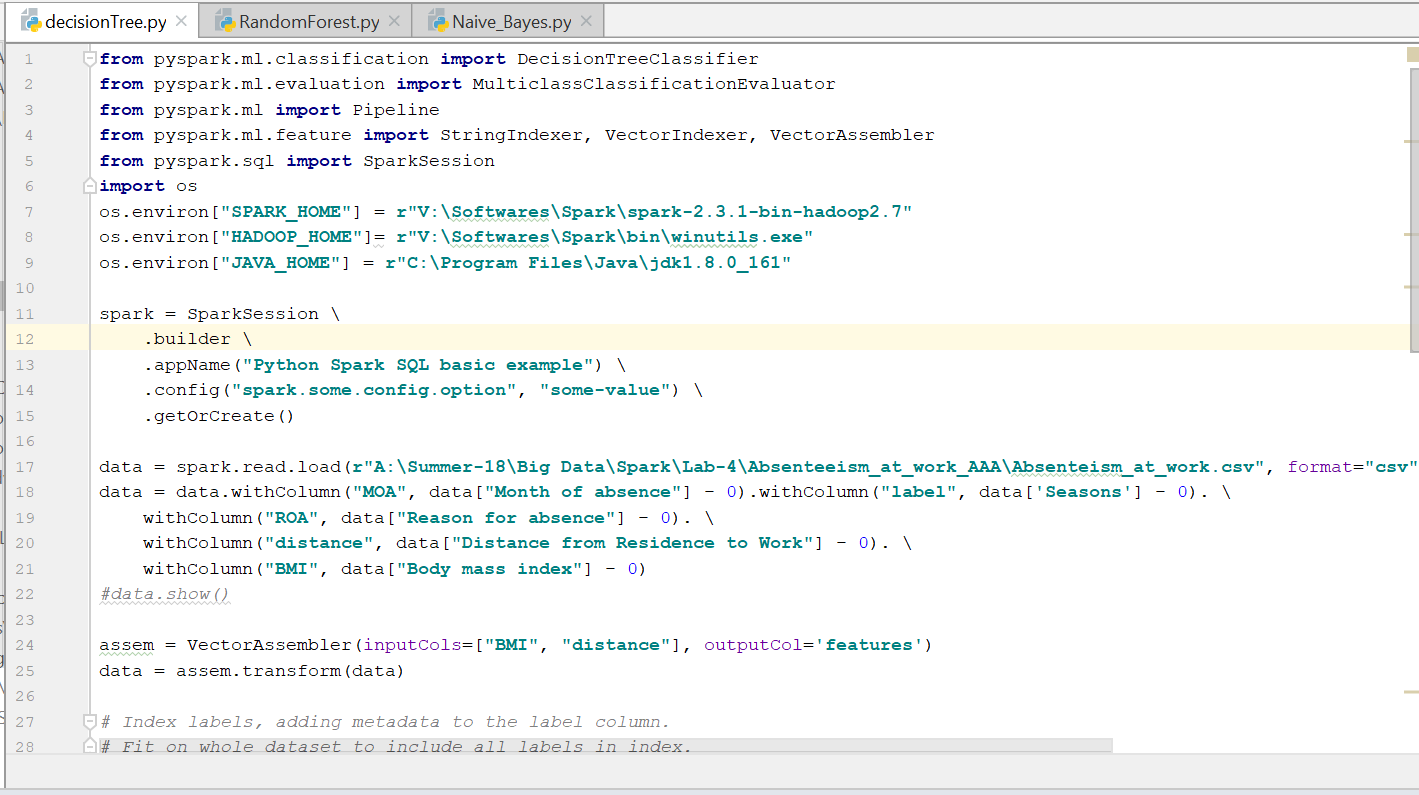


**Random Forest Output:**

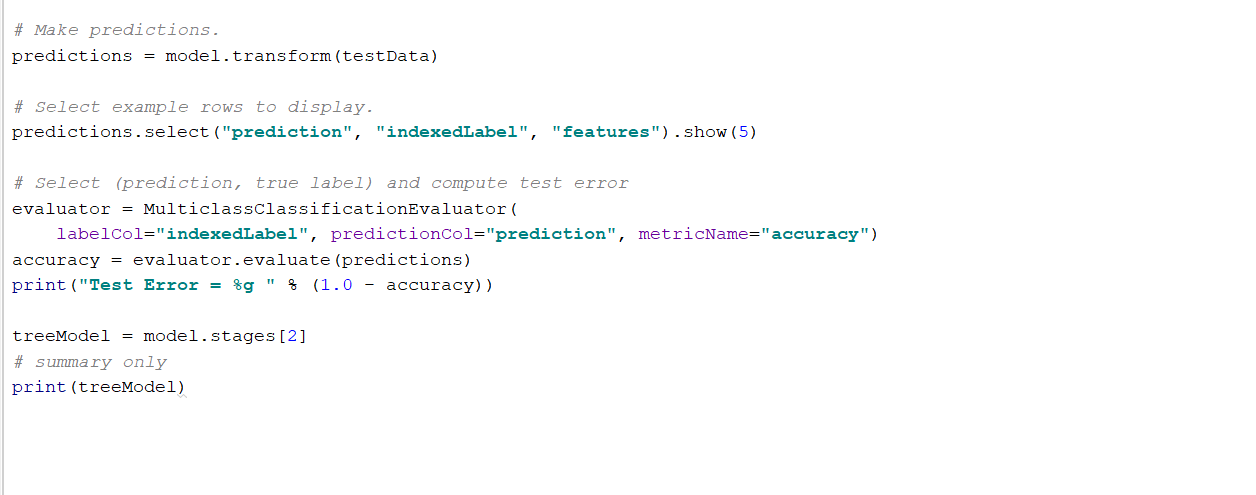


1. **Decision Tree Algorithm:** Decision Tree Algorithm is a Supervised Learning Algorithm; the main motive of the algorithm is to create a training model which can use to predict class or value of the target variables by making decisions from the previously trained data.

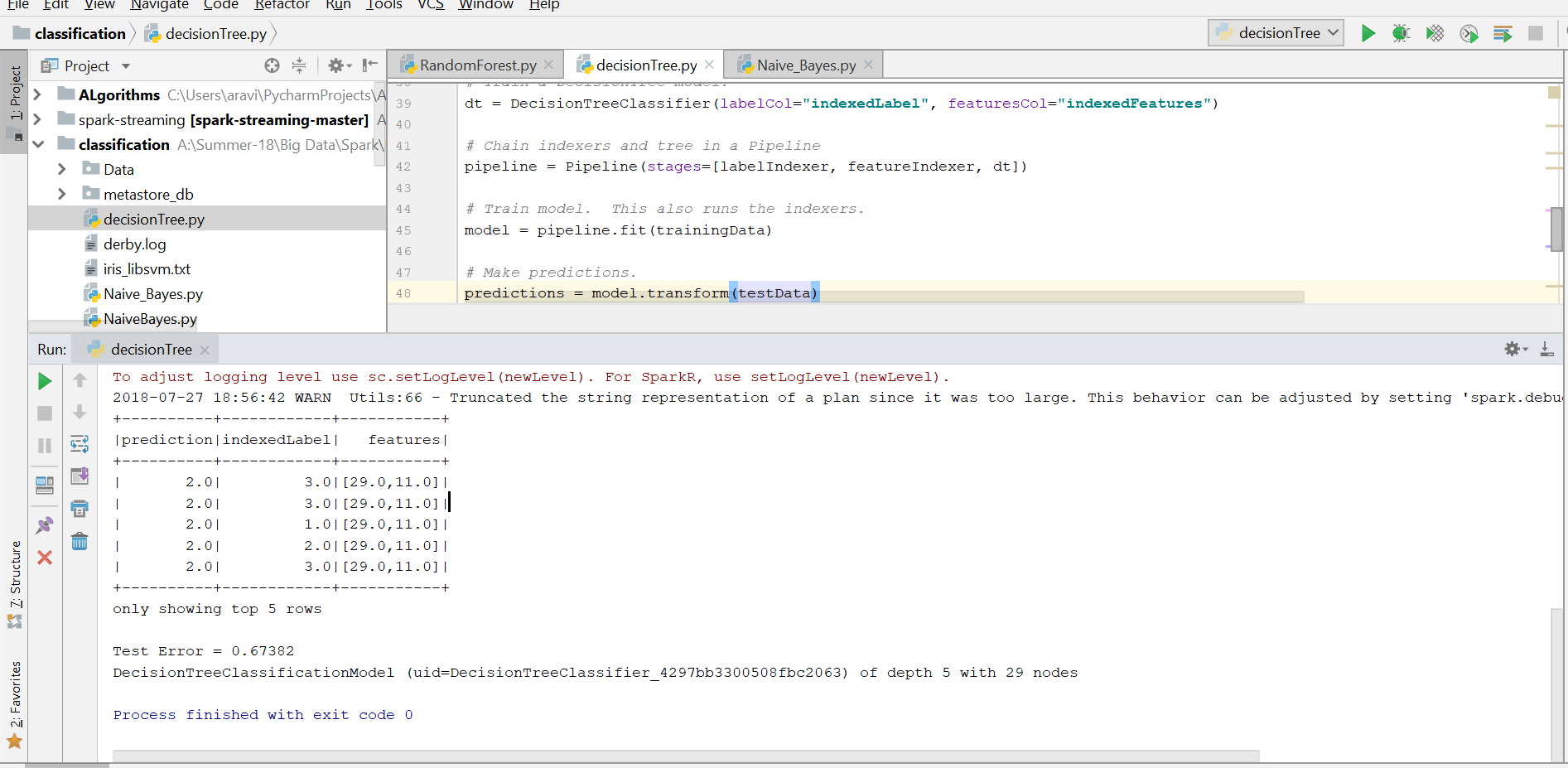
**Code Screenshots:**







**Decision Tree Output:**



**Task-2: Twitter Streaming Data Word Count.**

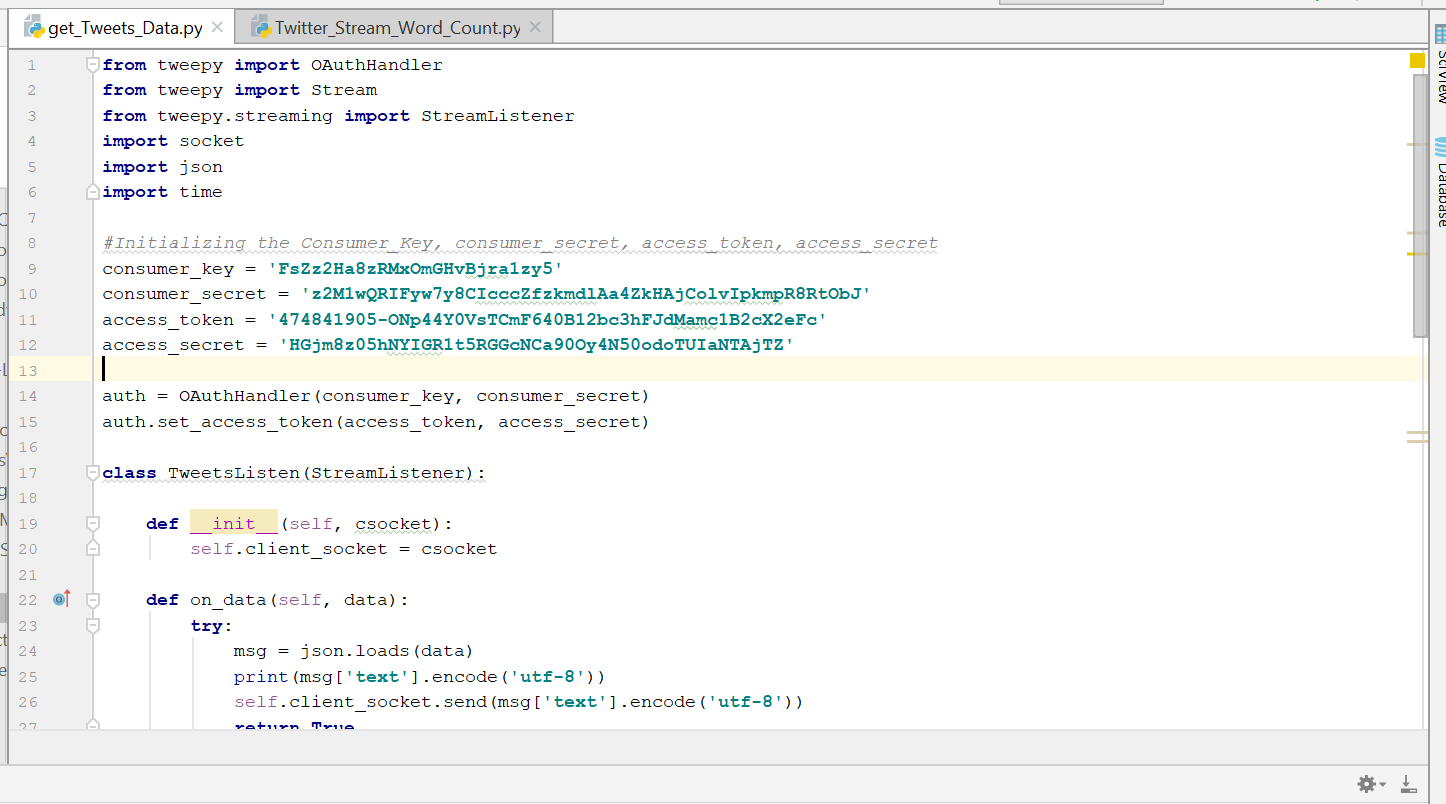
**Objective:** Perform the word count program on the live twitter data streaming.

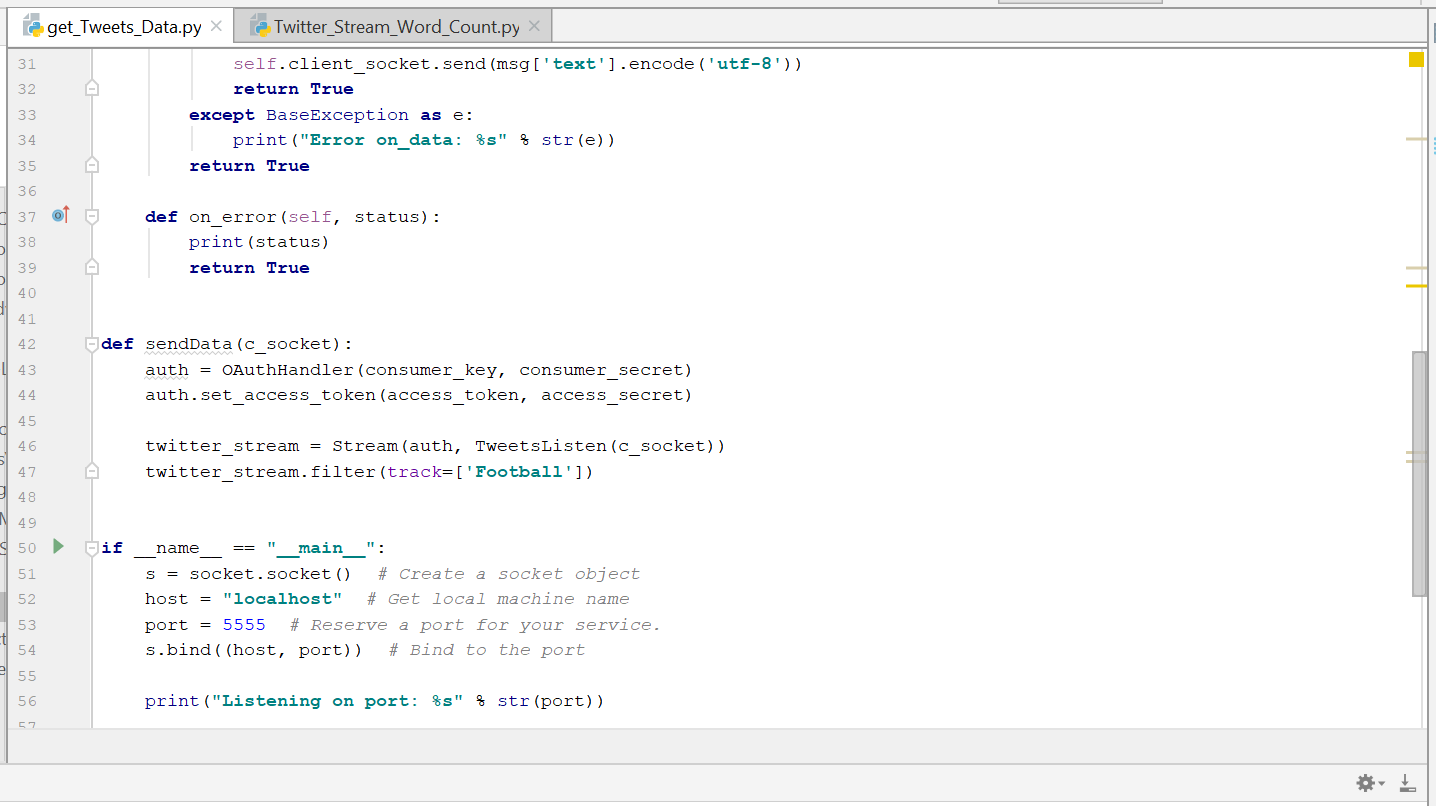
**Approach:** Used Tweepy module to get the tweets from twitter to perform the word count program. Used the socket to perform the listening of the live data and perform the word count on the tweets.

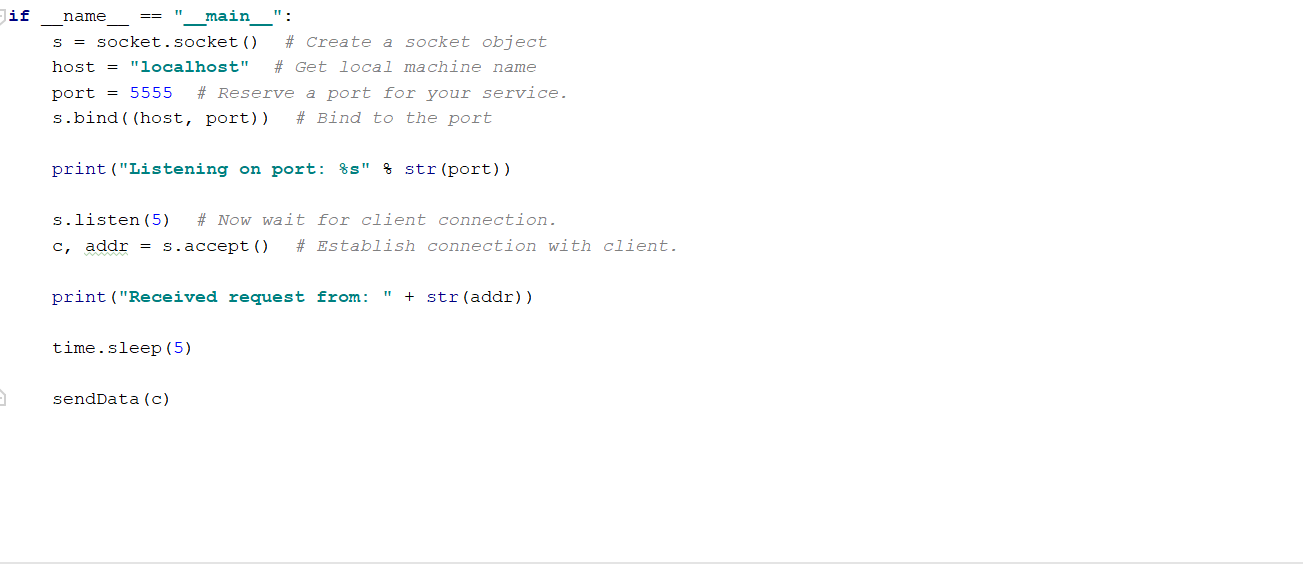
**Inputs:** Live Twitter data for the word count program on Spark.

**Code Screenshots:**

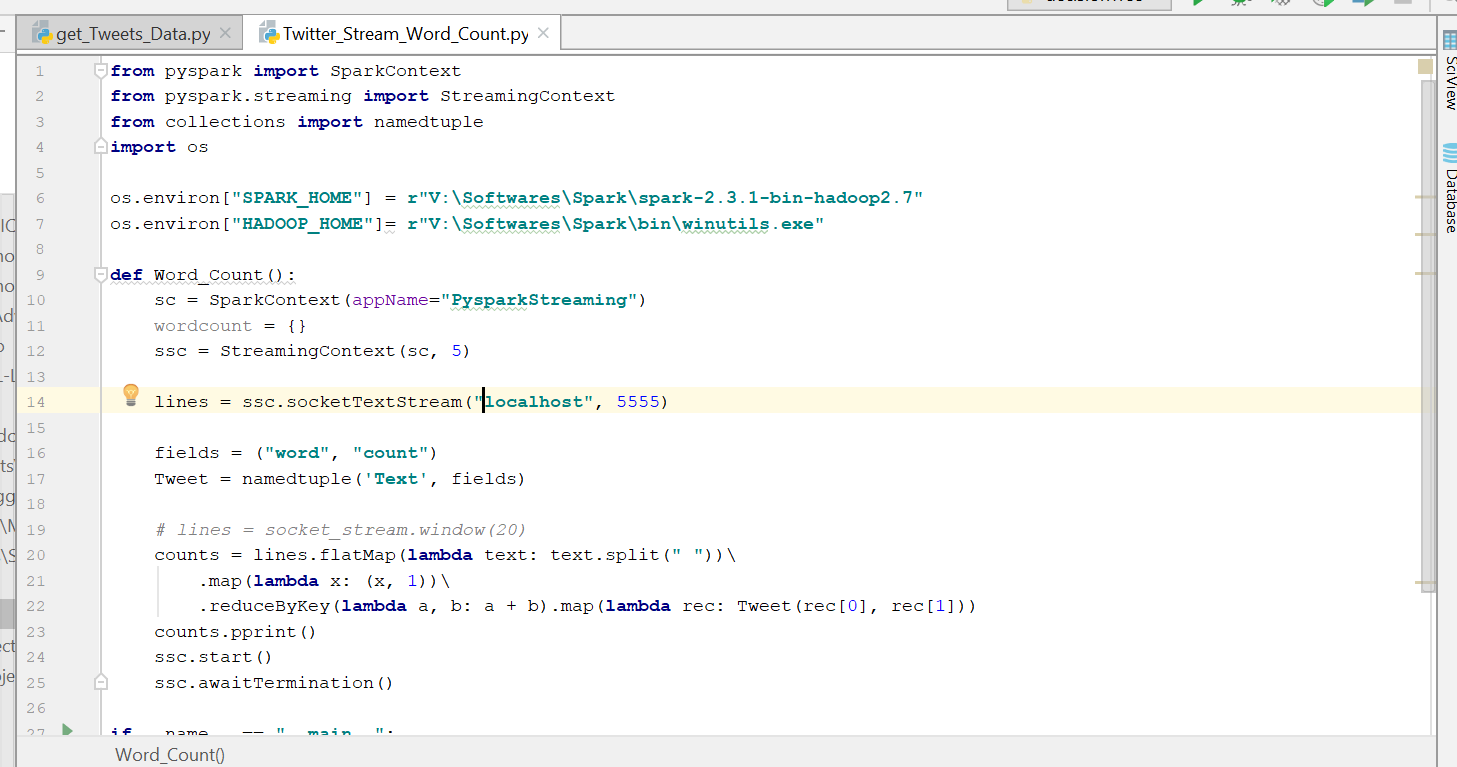
**Code for collecting the tweets from twitter:**

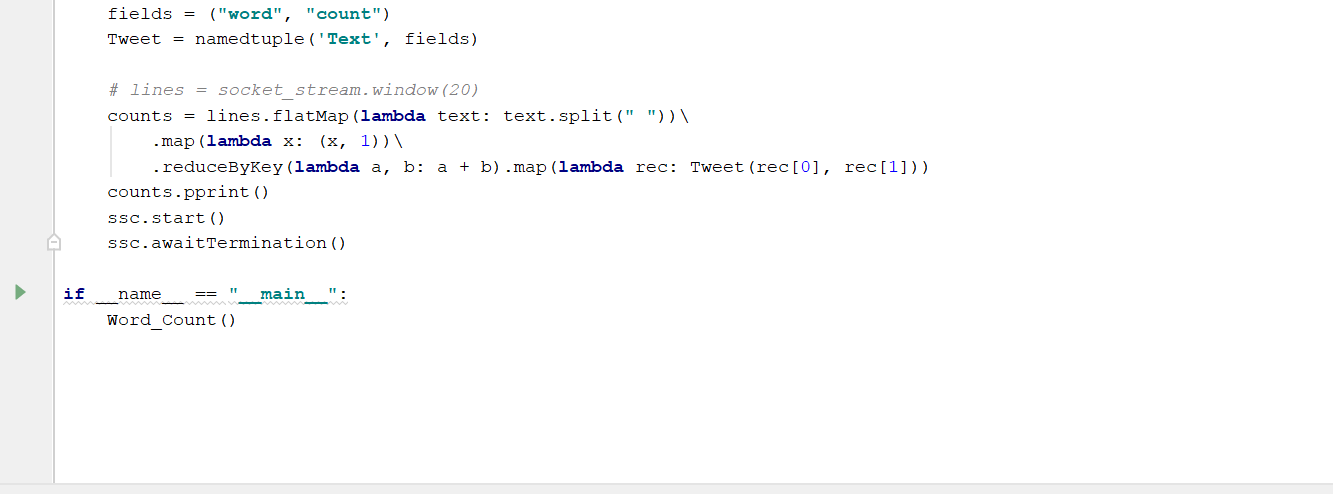






**Code to perform the word count on the tweets:**

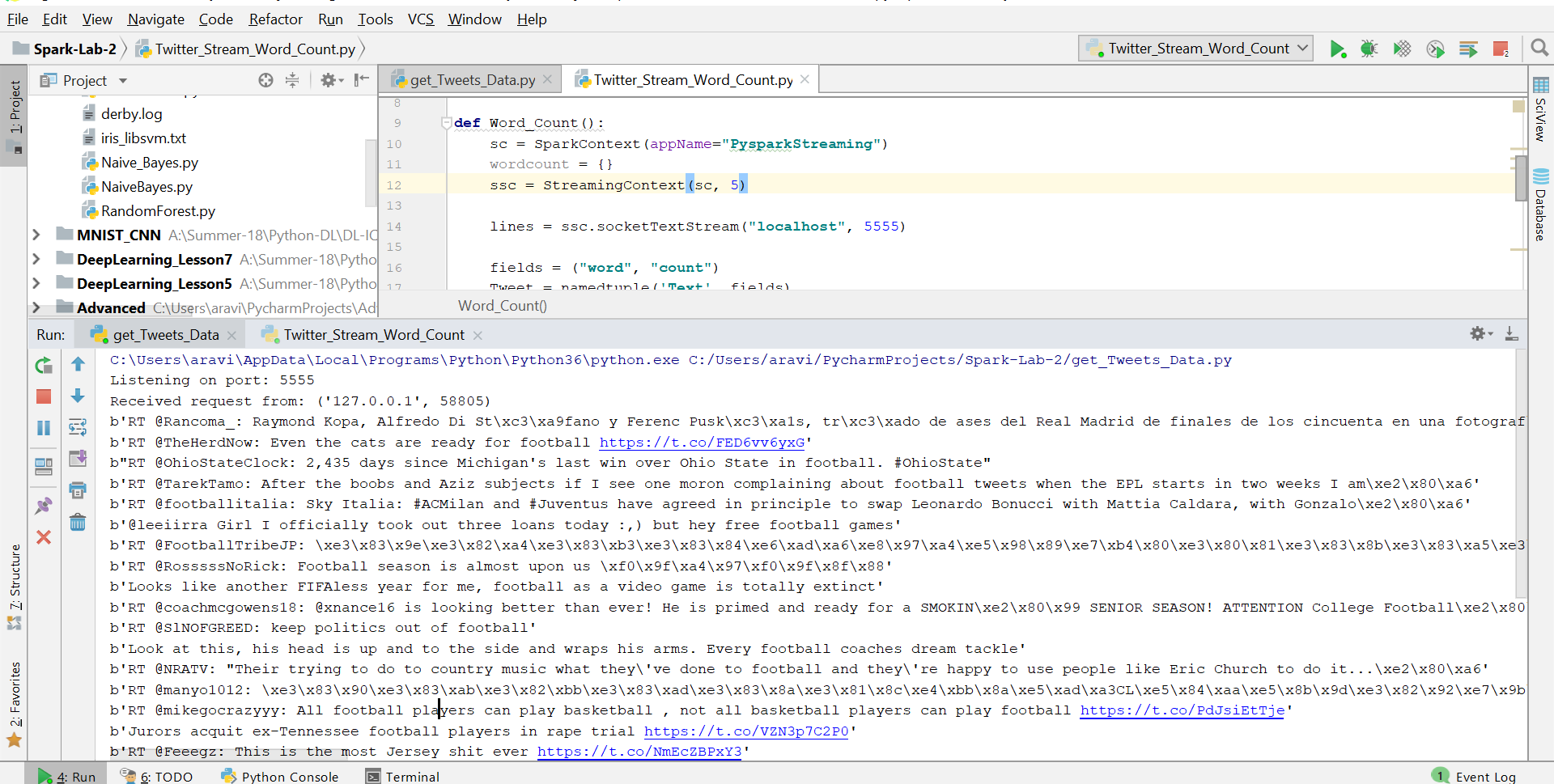




**Execution:** Executed the tweets collection file first and then executed the word count program to gather the tweets and perform the word count.

Output:

Tweets collection output:



Word Count Output:

