Apache Spark Assignment
Twitter Sentiment Analysis

Introduction to Big Data

Farid Kopzhassarov, DS1
f.kopzhassarov@innopolis.ru

Contents

1 Introduction

In this assignment we were asked to apply Apache Spark Streaming to stream tweets from Twitter and classify their sentiment according to their content. So, I have chosen to use three distinct groups: Negative, Neutral, and Positive.

2 Streaming

Due to the strict policy of getting access to Twitter API, it was a problem to use it. So, QueryFeed was suggested instead.

To use the supplied RSS feed from QueryFeed service, a separate library was used. It helped to parse and convert the RSS feed's XML file to Apache Spark compatible RDD stream.

3 Preprocessing

Before the analysis task, the incoming data needs to be preprocessed, i.e. drop useless or empty columns, filter the values in the remaining columns, thus the noisy data will be removed and won't affect the classifier results.

After processing the raw data, only 2 columns left: Tweet and Sentiment value.

So, we can proceed to the tweet preprocessing. In the provided program, it consists of tokenizing, lowercasing, lemmatizing, and stop words removal. The result of tweet preprocessing is stored in a new column. An example is shown in the figure below.

tweet	**sentiment**
What a WONDERFUL day!	NULL
I am going home.	NULL

Table 1: Before preprocessing

tweet	**preprocessed**	**sentiment**
What a WONDERFUL day!	[wonderful, day]	NULL
I am going home.	[go, home]	NULL

Table 2: After preprocessing

4 Sentiment Analysis

Sentiment analysis is done with the help of Stanford CoreNLP which provides a set of natural language processing tools. It initially classifies the text into 5 groups: very negative, negative, neutral, positive, very positive. But to make the output simpler, I decided to limit it to 3 groups (negative, neutral, positive) by merging very negative with negative, and very positive with positive.

The topic chosen was *#clevelandcavaliers* related to one of the NBA teams known as Cleveland Cavaliers.

tweet	**sentiment**
Cavs win back to back vs Contenders! Were still Eastern Conference Champions!	Positive
Grandpas and baby boys first game!	Neutral
Thunder star Russell Westbrook to demolish Cavaliers? Heres his sporty wife	Negative
Magic is getting nervous that his team might be using LeBron like Cleveland did	Negative

Table 3: Correctly classified tweets

tweet	**classifed**	**correct**
Cavs just beat Philly and Houston. Tristan put up 20 and 16 .. is he cheating on Khloe again ? Because those ain't Kardashian numbers	Negative	Positive
Cleveland Cavaliers 2016 NBA Finals Champions Land of Champions T-Shirt - Navy	Negative	Neutral
Game time!! Lets Go CAVS!!!	Neutral	Positive
#ClevelandCavaliers #Suck	Neutral	Negative

Table 4: Incorrectly classified tweets

5 Conclusion

Actually, the precision was not as high as expected. I can surely state that it is all because of misusing the hashtags, or some neutral posts like team's merchandise. So, mostly, the results of the classifier strongly depend on the context of tweet.

Appendix I. Launch the Spark application

After deploying jar file, to launch the Spark application:

"jarfile topic outputfile"