CS6350

Big data Management Analytics and Management

Spring 2022

Homework 4

Submission Deadline: April 29th, 11:59 p.m.

Part 1: Spark Streaming and Visualization (100 points)

Q1.

You are required to implement the following framework using Apache Spark Streaming, Kafka (optional), Elastic, and Kibana. The framework performs SENTIMENT analysis of particular hash tags in twitter data in real-time. For example, we want to do the sentiment analysis for all the tweets for #trump, #coronavirus. Note that if you implement this framework with Scala, there is no need for Kafka and you can connect to twitter via the internal API. But if you want to implement it with Python, Kafka is required. Be careful about the Scala version compatibility.

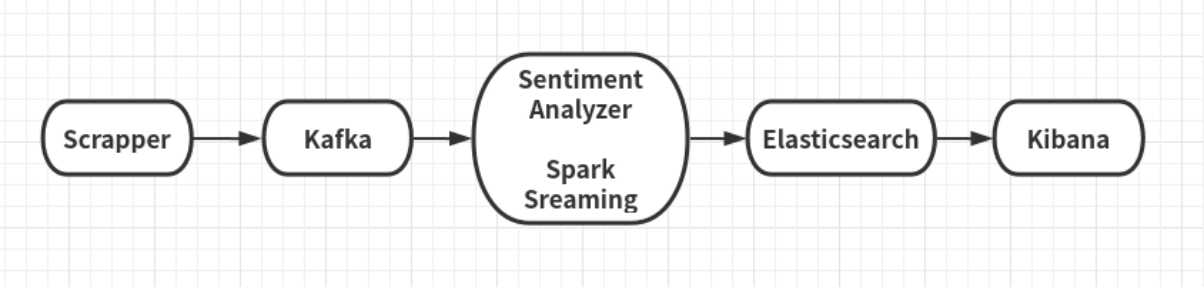

Figure: Sentiment analysis framework

The above framework has the following components:

1. Scrapper (for python, but Scala needs to produce same result)

The scrapper will collect all tweets and sends them to Kafka for analytics. The scraper will be a standalone program written in PYTHON and should perform the followings:

a. Collecting tweets in real-time with particular hash tags. For example, we will collect all tweets with #blacklivesmatter.

b. After filtering, we will send them to Kafka in case if you use Python.
c. You should use Kafka API (producer) in your program (https://kafka.apache.org/090/documentation.html#producerapi)
d. Your scrapper program will run infinitely and should take hash tag as input parameter while running.

2. Kafka (for Python)

You need to install Kafka and run Kafka Server with Zookeeper. You should create a dedicated channel/topic for data transport

3. Spark Streaming

In Spark Streaming, you need to create a Kafka consumer (for python, shown in the class for streaming) and periodically collect filtered tweets (required for both Scala and python) from scrapper. For each hash tag, perform sentiment analysis using Sentiment Analyzing tool (discussed below).

3. Sentiment Analyzer

Sentiment Analysis is the process of determining whether a piece of writing is positive, negative or neutral. It's also known as opinion mining, deriving the opinion or attitude of a speaker. For example, the following tweets taken from Twitter are shown along with their sentiment.

"RT @jeremycorbyn: It is shameful the UK Government won't condemn Trump. Now is the time to speak up for justice and equality. #BlackLives"- has positive sentiment.

"RT @larryelder: How many unarmed blacks were killed by cops last year? 9. How many unarmed whites were killed by cops last year? 19. More" - has negative sentiment.

You can use any third-party sentiment analyzer like Stanford CoreNLP (Scala), NLTK(python) for sentiment analyzing. For example, you can add Stanford CoreNLP as an external library using SBT/Maven in your Scala project. In python you can import NLTK by installing it using pip.

4. Elasticsearch

You need to install the Elasticsearch and run it to store the tweets and their sentiment information for further visualization purpose.

You can point http://localhost:9200 to check if it's running.

For further information, you can refer:

https://www.elastic.co/guide/en/elasticsearch/reference/current/getting-started.html

5. Kibana

Kibana is a visualization tool that can explore the data stored in Elasticsearch. In this assignment, instead of directly output the result, you are supposed to use the visualization tool to show your tweets sentiment classification result in a real-time manner.

Please see the documentation for more information:

https://www.elastic.co/guide/en/kibana/current/getting-started.html

Q2

Clustering

This is a two-part question where you are expected to perform incremental K-means clustering on Twitter data by collecting relevant tweets after every 30 second interval. You are also expected to show the changes in the clusters via a scatter plot that updates in real-time.

a) For the first part use the scrapper from the first question, extract tweets having the hashtag #BLM and perform K-Means clustering with K=3. Then plot the points and show which cluster they belong to.
b) Then after every 30 second interval, extract tweets for the same hashtag for an indefinite amount of time. For each interval, incrementally update the K-Means clusters and show how they change due to the addition of the new set of clusters.

What to submit:

1. Python/Scala code
2. Screenshots of your visualization charts

Part 2: Clustering and recommendation systems (100 points)

Section #1
Objective:

This assignment is for you to learn about clustering and recommendation system, particularly about different techniques of clustering.

Please solve the following problems. No computer programming is required to solve the problems.

Problem Statement:

1. K-Means algorithm:

Consider the following eight points in a 2-dimensional space: {(2, 10); (2, 5); (8, 4); (5, 8); (7, 5); (6, 4); (1, 2); (4, 9)}. Use the Euclidean distance metric to measure the separation of the points.

a. Plot the data points and group them into appropriate clusters. How many clusters are required and what are contents of each of these clusters?
b. Now consider we want to divide the points into 3 initial clusters (C1, C2, C3) with centers defined as {(2, 5), (5, 8), (4, 9)} respectively.
c. What's the center of the cluster after one iteration?
d. What's the center of the cluster after one 2nd iteration?
e. What's the center of the cluster after one 3rd iteration?
f. Compare the results of each iteration with your answers in part (a).
g. How many iterations are required for the clusters to converge?
h. What are the resulting centers and resulting clusters (K=3)? Plot the final data points.

2. Hierarchical algorithm:

Use the similarity matrix in the table below to perform single and complete link hierarchical clustering. Show your results by drawing a dendrogram. The dendrogram should clearly show the order in which the points are merged. (with enough explanation and calculation)

	P1	P2	P3	P4	P5
P1	1.00	0.10	0.41	0.55	0.35
P2	0.10	1.00	0.64	0.47	0.98
P3	0.41	0.64	1.00	0.44	0.85
P4	0.55	0.47	0.44	1.00	0.76
P5	0.35	0.98	0.85	0.76	1.00

3. DBSCAN algorithm

Consider the following eight point in a 2-dimensional space: {(2, 10); (2, 5); (8, 4); (5, 8); (7, 5); (6, 4); (1, 2); (4, 9)}. Suppose we use the Euclidean distance metric.

a. If Epsilon is 2 and min_samples is 2, what are the clusters that DBSCAN would discover. Plot the discovered clusters.
b. What if Epsilon is increased to $\sqrt{10}$?

4. Explain the shortcomings of BFR algorithm and describe how CURE algorithm overcomes the shortcomings.

Section #2: BigTable and Cassandra

1.There are many ways one might organize the distributed storage of data structured as rows and columns. What design decisions are shared by both Bigtable and Cassandra?
2. Bigtable uses a "master server" and "tablet servers" to manage the reading and writing of data; Cassandra does not have a distinguished master node. What are the pros and cons of each architecture?
3. How is replication handled in each of Bigtable and Cassandra?
4. What are the main steps in reading and writing data in Bigtable?
5. What are the main steps in reading and writing data in Cassandra?
6. What is the use of Bloom Filter in Cassandra?
7. How does Cassandra delete data?
8. What is SSTable? How is it different from other relational tables?
9. Explain CAP theorem.

Section #3: Recommendation Systems

Use Collaborative filtering to find the accuracy of ALS model. Use ratings.dat file. It contains: User id :: movie id :: ratings :: timestamp.
Your program should report the accuracy of the model. For details follow the link: https://spark.apache.org/docs/latest/mllib-collaborative-filtering.html. Please use 60% of the data for training and 40% for testing and report the MSE of the model. Submit the code along with the output of your code.