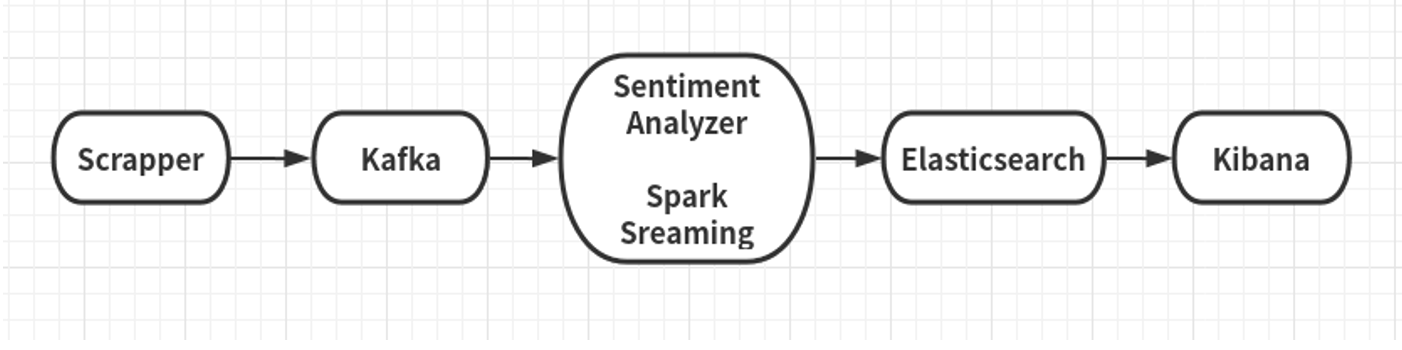
**CS6350 Big data Management Analytics and Management Spring 2020**

**Homework 3**

**Submitted By Keerti Keerti – KXK190012**

**Spark Streaming of the Tweets with #coronavirus and #trump**



**1. Scrapper (python)**

The scrapper collects all tweets and sends them to Kafka for analytics.

1. Collecting tweets in real-time with particular hash tags #trump, #coronavirus.
   1. Create a twitter developer acct and create an api to get access token and end points.
   2. Reference: <https://medium.com/@leowgriffin/scraping-tweets-with-tweepy-python-59413046e788> : Using tweepy library, fetch all the tweets using the access keys.
   3. Collected tweets are filtered for #trump and #coronavirus and using producer api are sent to the Kafka topic

**2. Kafka (Python)**

Installed Kafka and ran Kafka Server with Zookeeper with a dedicated channel/topic for data transport.

**3. Spark Streaming**

In Spark Streaming, created a Kafka consumer and periodically collected filtered tweets from scrapper. For each hash tag, performs sentiment analysis using Sentiment Analyzing tool.

**4. 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.

Used third party sentiment analyzer nltk(python) for sentiment analyzing.

**5. Elasticsearch**

Installed the Elasticsearch and ran it to store the tweets and their sentiment information for further visualization purpose.

**5. Kibana**

Kibana is a visualization tool that can explore the data stored in elasticsearch. Used the visualization tool to show the tweets sentiment classification result in a real-time manner.

A close up of a logo

Description automatically generated

A picture containing screenshot

Description automatically generated

Twitter Sentiment Ananlyis

1. Created a twitter Development Account to get access key, access secret key and consumer key and consumer secret key.

2. After which, using the keys and tweepy library, scraped the tweets and filtered the scraped tweets using #trump, #coronavirus

3. All the filtered tweets are then sent Via Producer API to the mentioned kafka topic

4. Consumer subscribes to the mentioned kafka topic and fetches all the tweets sent to the topic.

5. Using nltk, the tweets are classified into negative and positive tweets.

Have used naive bayes classification method.

Started,

Spark Streaming - pyspark, Pycharm started : Goto Documents/Pycharm\*/bin

## ./pycharm.sh

a demo project is created and add kafka and analyzer python files downloaded from elearning.

Zookeeper - started : Goto Documents/Kafka-2.12-2.0.5/bin -- open a new terminal

## ./zookeeper-server-start.sh /home/keerti/Documents/kafka-2.12-2.0.5/config/zookeeper.poperties

Kafka-started : Goto Documents/Kafka-2.12-2.0.5/bin -- open a new terminal

## ./kafka-server-start.sh /home/keerti/Documents/kafka-2.12-2.0.5/config/server.poperties

Elastic Search started: Goto Documents/Elasticsearch/bin -- open a new terminal

## ./elasticsearch

verify : browser : type localhost:9200 and localhost:9200/<indexname>/\_doc/1

Kibana started :sudo systemctl start kibana.service

verify: command line : type journalctl -u kibana.service

browser : localhost:5601

Index linking and the visualize