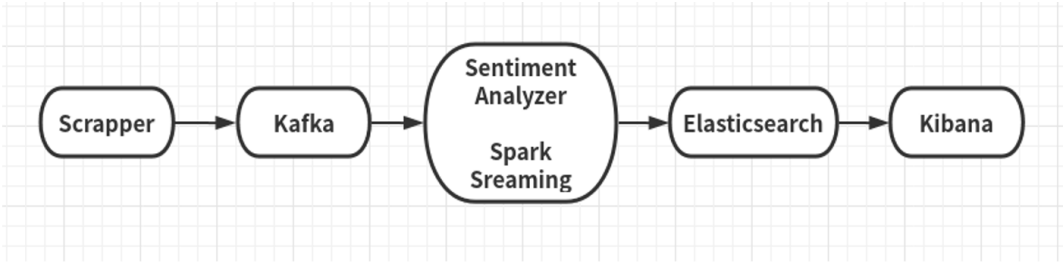
CS6350 Big data Management Analytics and Management Spring 2020
Homework 3
Submitted By Keerti Keerti – KXK190012

Spark Streaming of the Tweets with #coronavirus and #trump



1. Scrapper (python)

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

a. Collecting tweets in real-time with particular hash tags #trump, #coronavirus.

b. After filtering, sent them to Kafka.

c. Used Kafka API (producer) in my program

d. Scrapper program will run infinitely and takes hash tag as input parameter while running.

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.