Data-Intensive Computing, ID2221

Project Proposal

Oliver Chin Wee Gui, Erik Kongpachith, Timothée Ly
Group 7

October 6, 2022

1 Problem description

Covid-19 has been a major upset worldwide for the past 3 years. We are interested in studying people's reactions about the pandemic by computing statistics on their tweets. For instance, how many tweets with the hashtag `#covid19` were published each month for the past 3 years, what proportion of tweets does it make? Do users posting about Covid once often post again about it? - or is it more of a one time thing? We would also analyze the other hashtags related with Covid to get a better overview of what topics are meddled with the Covid pandemic. (vaccine, politics, etc?)

2 Data

The data will be extracted from Twitter in the form of Tweets. Our team will query the Twitter Filtered Stream Application Programming Interface (API) to extract tweets that fulfil a certain set of rules, such as tweets that contain the hashtag `#covid19`. Our team aims to process these Tweets live and update descriptive statistics calculated from these Tweets and data visualizations in real time.

3 Tools

Stream processing and data visualizations will be applied in our project. Therefore, the following tools will be used:
Spark SQL
Twitter API
Kafka/Spark Streaming

4 Methodology

1. Use Twitter API to get data of corona related tweets from the past 3 years.
2. Use Spark SQL for processing and perform an analysis of the data.
3. Use Spark streaming/Kafka for an interface which displays the most trending corona related tweets in real time.