**Twitter Data Analysis Report**

1. **Description**

Twitter's stream live data can be retrieved through Twitter's API once you have a Twitter developer account. Hence, we can utilize it to do some interesting data analysis or make some useful applications to serve our own purpose. This is a mini experimental project which makes an application for showing frequency of tweet's hashtags with stream live data. Furthermore, some additional analysis can be implemented with stream live data.

1. **Design Methodology**

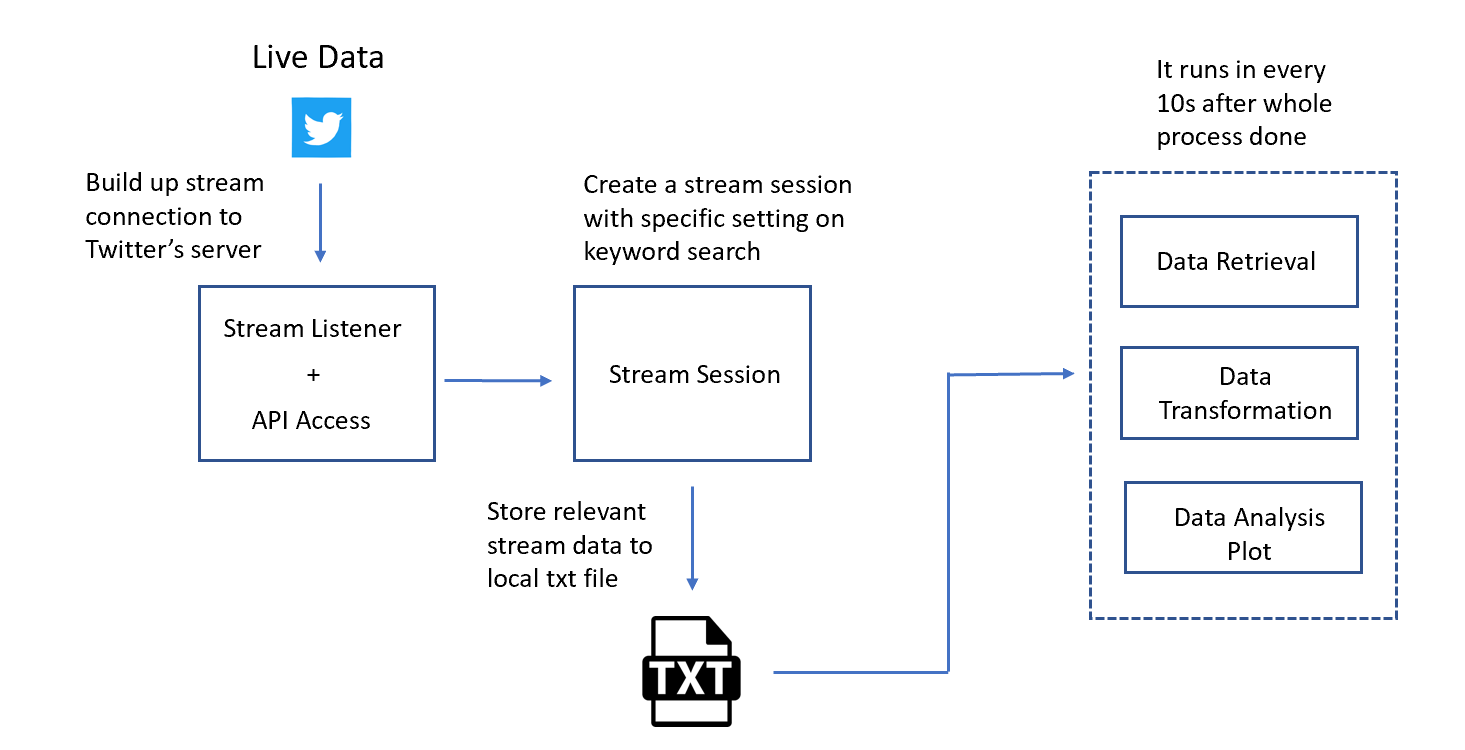


Fig 2.1 Application Design Architecture

My key design thinking is that make application to catch live data tightly and update the analysis plot. In my local pc, the easier way is that creating a stream session for retrieving the live data with Twitter API continuously. The stream data can be stored into a local file and be extracted for doing necessary data analysis.

In one round of operation, the data is extracted from txt file. Some ETL job is done with the extracted data, and the requested data analysis will be done with transformed data. The application can be set up to run in every 10 seconds after one round of operation is done. The data analysis plot will be kept updating.

However, there are some disadvantages in this design. The working efficiency will be impacted once the data size is built up into a huge amount. The analysis result shown in the plot may not be purely live, the result may only represent the situation at dozens of seconds before.

Besides above method, I also have tired some specific frameworks like Apache Kafka working with Pyspark for implementing data analysis without the steps which store and call the data in local. Unfortunately, it is not fully successful on setup of framework in local within limited time and make full process works. I believe that this shall be another working direction which is valuable to work on.

1. **Insights from Analysis**

From fig 3.1, except “python”, the most frequent appeared hashtag is “machine learning”, “AI”, “data science”, “analytics” in the tweets when we searched key word “python”. It means that python is quite relevant to these areas and it shall be the important development tool in these areas.

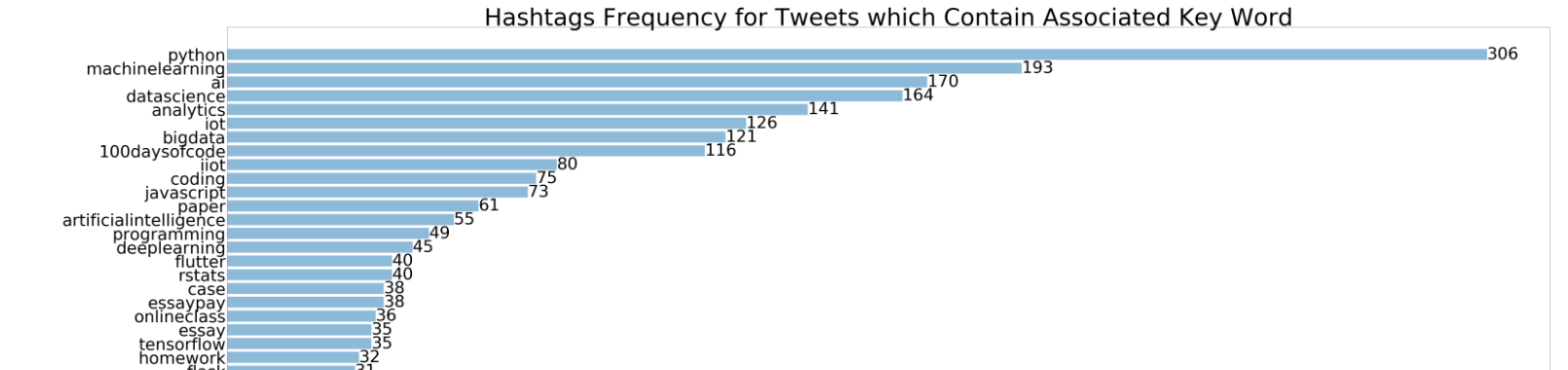
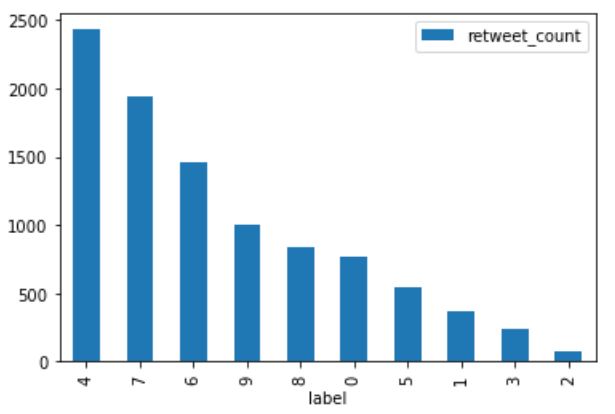
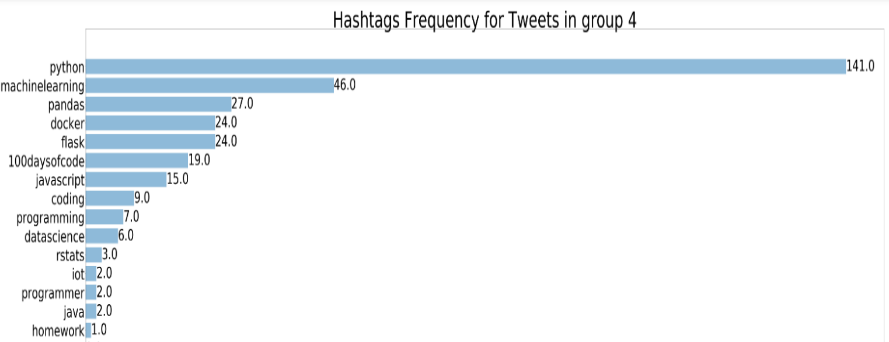


Fig 3.1 Hashtags Frequency

In further analysis session, by using hashtags info, I use clustering technic to find out clusters among all tweets we searched. From fig 3.2, we can link information from hashtags frequency and retweet count plot and read out some insights.

For example, the hashtags in cluster 4 which has highest retweet counts are mostly 'python', 'machine learning', 'pandas', 'docker', 'flask', '100 days of code', 'javascript'. The most tweets in this cluster seems relate to web development & machine learning and its deployment, it also actively links to the popular self-direct challenge "100-days of code". People are very responsive on this area's tweet. In cluster 7, the words are mostly 'ai', 'homework', 'essay', 'docker', we may conclude that the tweets in this cluster are mostly from students. They may be doing some homework related to AI area and publish tweets for getting some help, other people may like to retweet them for getting more people to help.



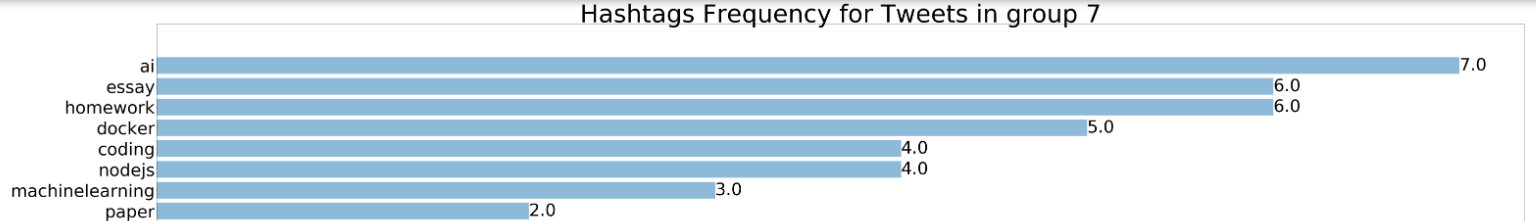


Fig 3.2 Comparison between retweets count and hashtags frequency in cluster 4 & 7.