# Crawler Side

## 1 Description

The crawler part in this project is in charge of obtaining data from various data sources—to be specific—Twitter and AURIN APIs. In order to support the research topic sound and well, the crawler itself should be able to retrieve data within a certain range of specified keywords and location coordinates, which guarantee a higher effectiveness of data usage and sentimental analysis afterwards. In addition, it is critical to accumulate a large amount of raw data, which lessens or eliminates the possibilities towards wrong conclusions during analysing process, and contributes the final goal of this project.

## 2 Crawling Tweets

Tweeter is researcher-friendly and provides a number of APIs for retrieving tweets from its database. But still there are lots of problems and challenges having been met during the crawler’s development.

### 2.1 Challenges & Solutions

#### 2.1.1 Outdated Documents

Although the Twitter gives out a range of APIs to researchers, the Documents of them, largely, do not contain the newest features about. Thus for the usage of Twitter APIs in the crawler side in the beginning, only the most basic and simple functions are put into, which only provides poor efficiency in both data retrieving and analysis afterwards.

To solve the challenge above, a great range of projects on GitHub and information in various forum have been viewed, for extra information of Twitter APIs new features. And finally, one of the API: *tweepy.api.cursor()* has been found, which covers the most of the functions of previous attempts, enhancing the fetching performance to a great extent at the same time.

#### 2.1.2 Limitation of Data Requests

To constraint the total data throughputs, Twitter limits the rate of requests invoked by API users—180 times per 15 minutes (1 request every 5s) by default, which makes the process of accumulating tweets problematic.

Solution to this problem should lie in not only breaking the limitation of requests rate, but also improving the percentage of valid tweets retrieved from Twitter.

Possible solutions to these issues are:

* 1. Get tweets as many as possible within one request invoked in crawler and sent to Twitter.
  2. Describe the request to Twitter as explicit as possible, to ensure the majority of tweets in each request are valid.

#### 2.1.3 Duplicated tweets

#### 2.1.4 Search Exactly

#### 2.1.5 Scalability

### 2.2 Trim Raw Data

## 3 Crawling AURIN

## 4 Write to CouchDB