|  |
| --- |
|  |
| Master Project |
| Twitter Event Detection |
|  |
| **Decheng Yang** |
| **Student ID: 817090**  **dechengy@student.unimelb.edu.au** |

# Abstract

Table of Contents

[Abstract 2](#_Toc525321922)

[1. Introduction 4](#_Toc525321923)

[1.1 Background 4](#_Toc525321924)

[1.2 Scope and Structure 5](#_Toc525321925)

[Development Environment 6](#_Toc525321926)

[2. Design 7](#_Toc525321927)

[3. Implementation 8](#_Toc525321928)

[3.1 Data Structure 8](#_Toc525321929)

[4. Conclusion 10](#_Toc525321930)

[Reference 11](#_Toc525321931)

[APPENDIX 12](#_Toc525321932)

# Introduction

## 1.1 Background

With the development of Network and increasing of consumption level, social media has become a part of life. According to the (1) Sensis Social Media Report 2017, the mean number of users that using social media has increased 10 precents last year and 8 out of 10 people are on social media in 2017. Above all the social media platforms, Twitter is one of the most popular social media and has more than 3,000,000 active users every month. (2) The context that people posted in Twitter is called tweet, a tweet can contain text, picture or videos. And the most important characteristic is that tweets are posted in real-time. (9)

Those characteristics make Twitter a valuable resource that provides data for researchers in different fields like data mining, social science and others. (4) For social science those data can be used to detect events (5), analyse rumour (6) and even predict crime (7).

**Twitter** has its special features: mentions and **hashtags** which engage people to share events in a minute and get instant feedback (3). A hashtag that begins with a “#” signal and follows a word or phrase (8) usually indicates the subject of the tweet. Thus, it contains useful information and can be used to detect events. For example, “#ausopen” and “#australianopen” both refer to event The Australian Open tennis tournament. Meanwhile, some people may post geo-hashtags like “#Melbourne” along with the tweets, which may also point out the place of event. (4)

Research on using online resource to detect event has been studied since 1998(10). Recent years, analysing social media to detect event become much popular. (11, 12). Twitter, as its context contains amount of information and posted in real-time, can be used to detect event efficiently and instantly. But the its microtext characteristic also brings some challenges. For instance, some people only post hashtags with photos, and others posts text with abbreviated words and grammar and spelling errors. (4) So, to detect event though Twitter, it is important to extract useful information from hashtags and text.

**Point-of-interest**, or POI, is a terminology in Cartography which indicates a special point location that people may think useful or interesting. (13) POI is widely used in GPS navigation software and maps. With the help of POI, it is possible to locate the location of event when tweets do not contain specific location information.

## 1.2 Scope

The project is aimed to develop a software to detect events happened in Melbourne in real-time based on the real-time stream of tweet posted in Melbourne, the data of previous 3 days tweets and the data of POI information of Melbourne.

The paper is structured as follows. In next section I will list the requirements for setting up the environment. Then a design of detecting and recognizing events will be described. In implement section introduces the third-party libraries and use of those libraries. Finally, the results will be tested and evaluated.

## 1.3 Development Environment

OS: macOS High Sierra 10.13.4

Project language: Java

Project SDK: Java 10

Project IDE: IntelliJ IDEA Community Edition 2018.1

Project management tool: Apache Maven 4.0.0

Third-party dependencies:

1. Joinery Dataframe

<https://github.com/cardillo/joinery>

1. Apache Commons – Math

<http://commons.apache.org/proper/commons-math/>

1. Apache Commons – Text

<https://commons.apache.org/proper/commons-text/>

1. Twitter4J

<http://twitter4j.org/en/>

# Design

The data resources used in project are:

1. Tweets
2. POIs of Melbourne

To detect event, we need to link tweets with POIs. Meanwhile, it’s impossible to detect event by a single tweet. Due to the real-time characteristic of tweets, a time block of 10 mins is set to detect event. That is, the system will check every 10 mins to figure out whether there are events or not based on the tweets posted within 10 mins. An event means something special happened, so we can compare current 10 mins tweets with previous tweets to select out candidates of events, then apply a further algorithm to recognize true events.

The whole processing graph of project should be:

Figure 1 Processing graph of project

All stages will be introduced in the following sections.

# Implementation

In design section the work flow of system has been showed. In this section, the system will be implemented step by step.

## 3.1 Data Structure

Twitter development platform(<https://developer.twitter.com>) provides variety of API for the developers. To make use of Twitter APIs in Java, the third-party library Twitter4J(http://twitter4j.org/en/) is used in this project. Twitter4J is an unofficial library that widely used by the Java developers.

At the beginning stage of project, we need harvest public tweets though Twitter4J; in the following stage, we also need Twitter4J to collect real-time tweets through Twitter stream API.

Twitter Filter Real-time API enable developers to get real-time public tweets with filter function applied. The raw tweet is stored in Json structure and contains a lot of information that not used in this project. To make the data more readable and clear, as well as easier to edit data in the following steps, it is necessary to change raw tweet data into a suitable data structure like DataFrame in Pandas (a data analysis tolls for Python).

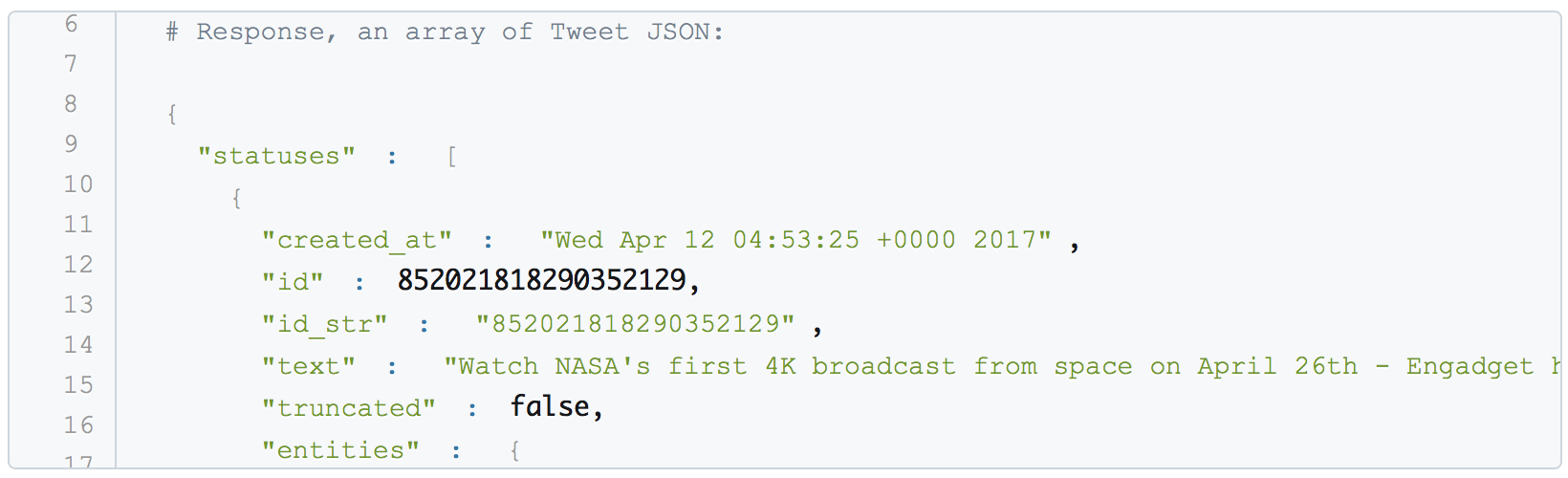


Figure1 Example of raw tweet stored in Json (from <https://developer.twitter.com> 9/18)

There are some third-party libraries such as Apache Spark DataFrame and others. After comparisons, Joinery DataFrame is chosen because it is light and more dependency free.

Joinery DataFrame is similar to Pandas DataFrame and has several easy-to-use functions that can edit, select and show DataFrames.

A DataFrame looks like:

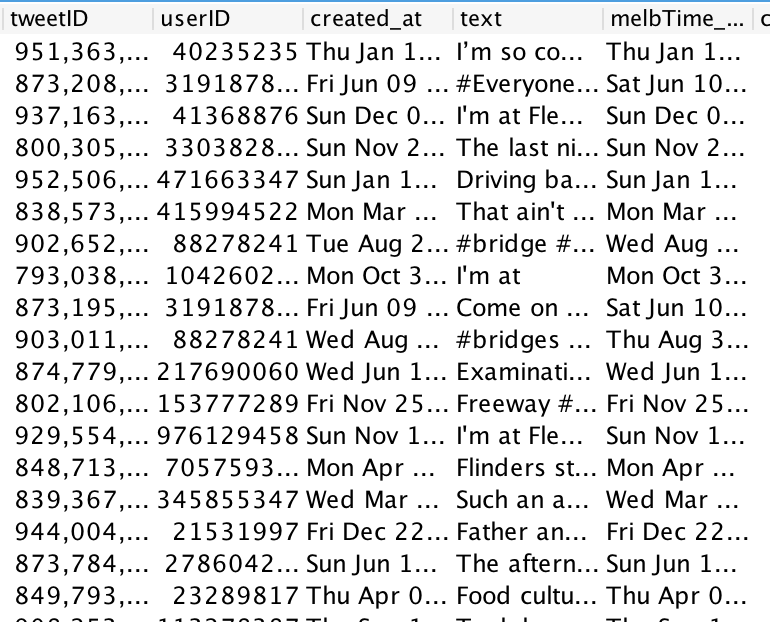
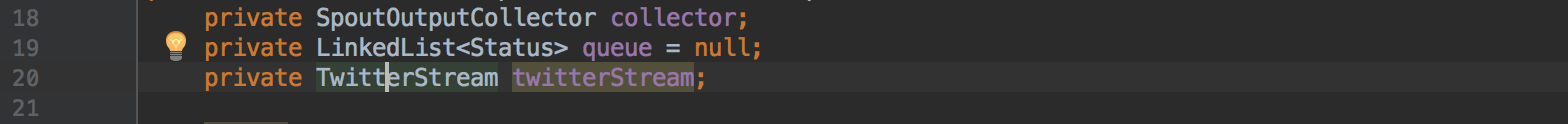


Figure 2 Example of tweets in DataFrame (not completed)

Operations on DataFrame is similar as the operations on tables of relational database. In this project, select data, set data, add new column, add new row, outer join two DataFrames, right join two DataFrames will be used.

## 3.2 Data Collection

We can simply implement tweet API by ‘tweet4j' library. The example codes are here:



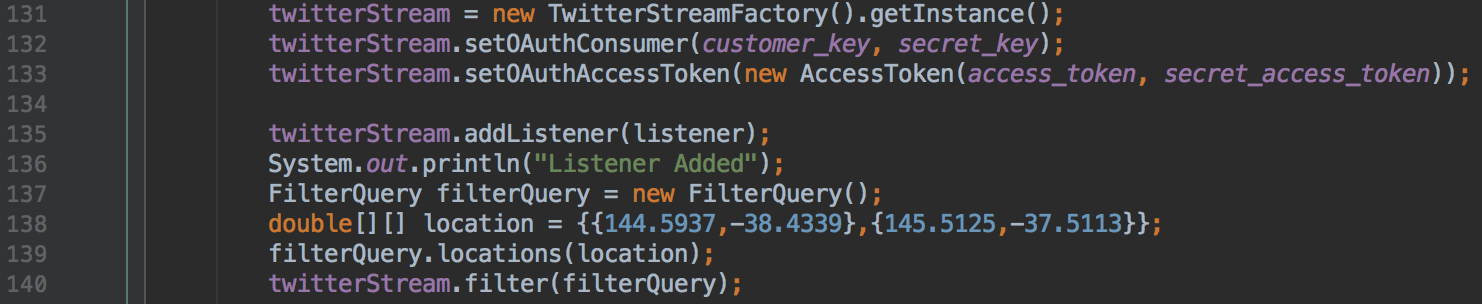


Figure 3 Codes to collect tweet by tweet4j

Class TwitterStream is used to create an instance of real-time twitter listen stream, and FilterQuery can help us to filter the tweets we will collect. Because we only need tweets from Melbourne, so the location filter is added into filterQuery.

The raw data collected contains full information of a tweet, and only some of them will be used in this project. To clarify, those data will be used later:

1. Tweet ID – unique number to identify a tweet
2. User ID – unique number to identify a Twitter User
3. Created\_at – indicates the create date of tweet
4. Lat – indicates the latitude of tweet
5. Long – indicates the longitude of tweet
6. Text – the full text of tweet

After extracting useful data from a raw tweet, POI information need to be linked to this tweet.

For example, there are the Melbourne POI data (showed in DataFrame):

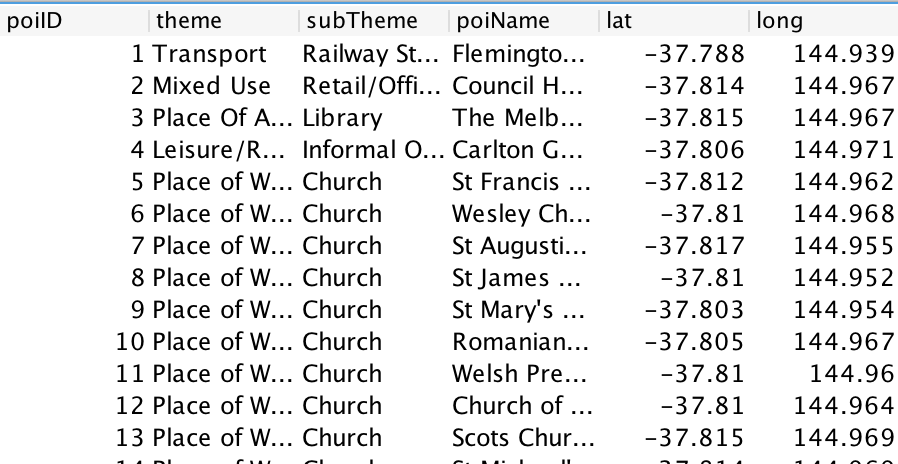


Figure 4 Example of Melbourne POI in DataFrame

And there are the extracted tweets:

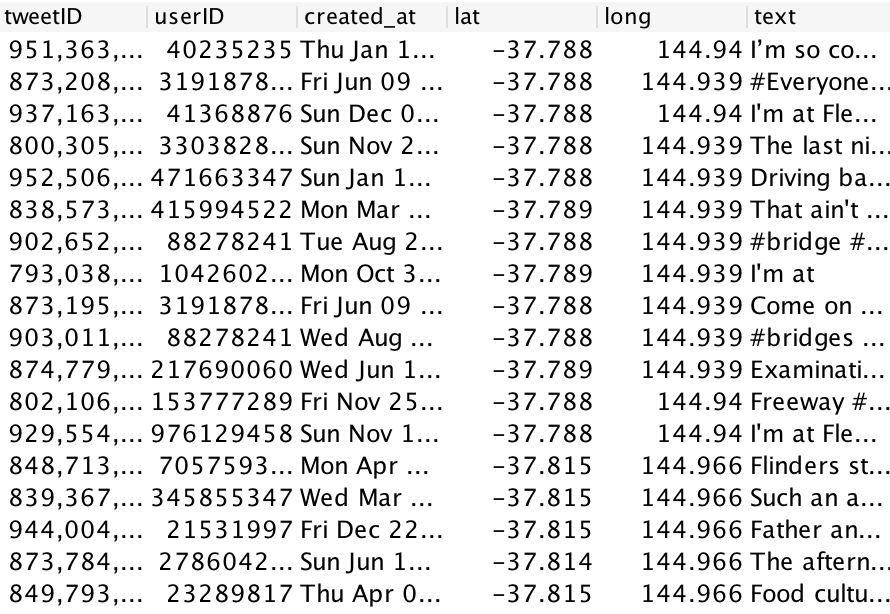


Figure 5 Example of extracted tweet in DataFrame

One specific tweet will be classified into one specific POI based on the geo-information of tweet. The information of POI will be ‘right join’ into the tweet DataFrame. In the end, we will get the DataFrame of tweets with POIs, as following figure shows:

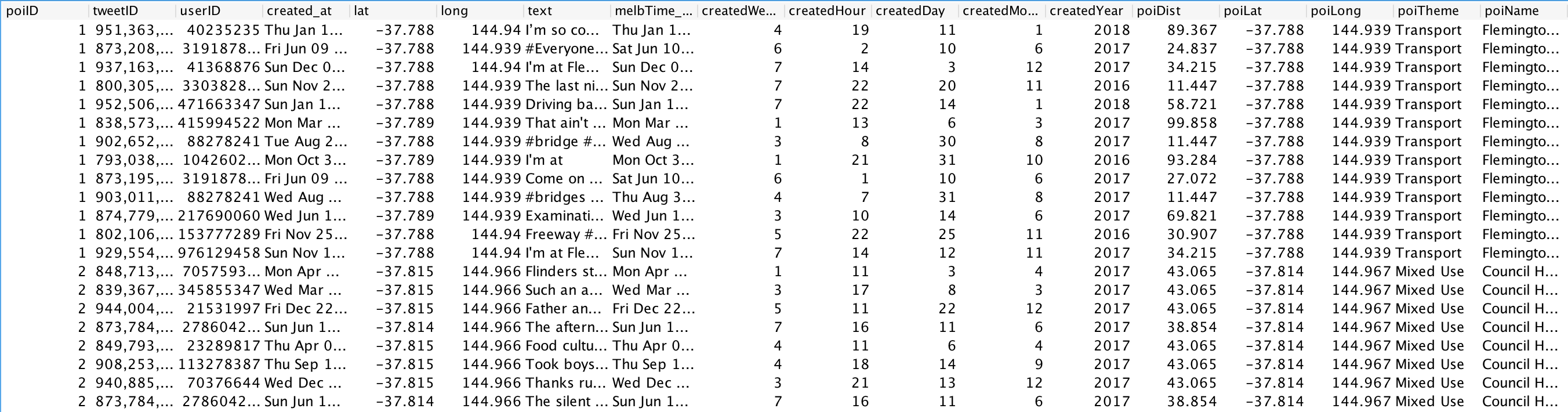


Figure 6 Example of tweets with POIs in DataFrame

There is the working flow graph of data collection stage:

Figure 7 working flow graph of data collection

The average speed of collecting tweets posted in Melbourne using single developer key through tweet stream API is 1 public tweet per minute. In this project, we use the tweets of 2017 as test data. (provided by xxx)

## 3.3 Candidate Selection

111

## 3.4 Event Recognition

111

## 3.5 Outcome Display

11

# Testing and Conclusion

# Reference

(1)(<https://irp-cdn.multiscreensite.com/535ef142/files/uploaded/Sensis-Social-Media-Report-2017.pdf>)

(2) (<https://www.civicwebmedia.com.au/australias-most-popular-social-media-sites-2018/>)

(3)(<https://devumi.com/2017/08/what-is-twitter-used-for-and-why-do-people-tweet/>).

(4) Multimodal Event Detection in Twitter Hashtag Networks

Yasin Yılmaz1 · Alfred O. Hero1

(5) Farzindar, A., & Khreich, W. (2015). A survey of techniques

for event detection in Twitter. Computational Intelligence, 31(1),

132–164

(6) Zhao, Z., Resnick, P., & Mei, Q. (2015). Enquiring minds: early

detection of rumors in social media from enquiry posts. In International

world wide web conference

(7) Wang, X., Gerber, M.S., & Brown, D.E. (2012). Automatic

crime prediction using events extracted from Twitter posts. In

International conference on social computing, behavioral-cultural

modeling and prediction.

(8) Feng, W., Han, J., Wang, J., Aggarwal, C., & Huang, J. (2015). STREAMCUBE:

Hierarchical spatio-temporal hashtag clustering for event exploration over the

Twitter stream. ICDE.

(9) Li, R., Lei, K. H., Khadiwala, R., & Chang, K. C. C. (2012, April). Tedas: A Twitterbased

event detection and analysis system. In Data engineering (icde), 2012 ieee

28th international conference on (pp. 1273-1276). IEEE.

(10) Yang, Y., Pierce, T., & Carbonell, J. (1998). A Study of retrospective

and on-line event detection. In ACM SIGIR conference on

research and development in information retrieval.

(11) Online city-scale hyper-local event detection via analysis of social media and human mobility

Jun Hu ; Yuxin Wang ; Ping Li

(12)

Real-Time Novel Event Detection from Social Media

Quanzhi Li ; Armineh Nourbakhsh ; Sameena Shah ; Xiaomo Liu

2017 IEEE 33rd International Conference on Data Engineering (ICDE)

(13) Yuan, Q., Cong, G., Ma, Z., Sun, A., & Thalmann, N. M. (2013, July). Time-aware point-of-interest recommendation. In Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval (pp. 363-372). ACM.

# APPENDIX