COMP90024 Cluster and Cloud Computing Project

* Group Members
* Jiyu Chen 908066

# Abstract

Twitter is a popular social media platform containing large amount of texture data. Aurin provide series of datasets developed and contributed by Australia’s leading researchers. In this project, we will leverage the NECTAR facility to create a four instances cluster environment and mining interesting geoinformation by summarizing tweets from eight cities around Australia and combine them with open sourced Aurin data. We will discuss the system structure, cluster design, tweet crawler, tweet data processor, sets of Aurin data we have leveraged, views of our data and guidance of system user interface.

Content

[Abstract 2](#_Toc513645210)

[1. Introduction 3](#_Toc513645211)

[2. System Design 4](#_Toc513645212)

[2.1 Cluster architecture 4](#_Toc513645213)

[2.2 System architecture 4](#_Toc513645214)

[3. Data Collector and Processor 4](#_Toc513645215)

[3.1 Hybrid crawler for tweets 4](#_Toc513645216)

[3.2 Embedded Sentiment Analyser 6](#_Toc513645217)

[3.3 Topic parsing and Hashtag Parsing 7](#_Toc513645218)

[3.4 Tweet Timepoint Partition 7](#_Toc513645219)

[3.5 Aurin collector and parser 7](#_Toc513645220)

[3.6 Format of processed data 8](#_Toc513645221)

[4. CouchDB as Database 8](#_Toc513645222)

[4.1 CouchDB in cluster 8](#_Toc513645223)

[4.2 Duplication prevention 9](#_Toc513645224)

[5. Scenario Study 9](#_Toc513645225)

[5.1 Happiest City Ranking 9](#_Toc513645226)

[5.2 I love tweet in the Morning 9](#_Toc513645227)

[5.3 Marriage is a Disaster? 10](#_Toc513645228)

[5.4 Passion for Sports or Gambling? 10](#_Toc513645229)

[5.5 A Rich Man’s Game? 11](#_Toc513645230)

[6. System UI and User Guide 12](#_Toc513645231)

[Reference 12](#_Toc513645232)

# Introduction

Nectar [1] stands for National eResearch Collaboration Tools and Resources project. It gives us fix number of computational resources and allowing researchers to create a cluster in their needs with high flexibility for system architecture design and management. We designed a <is there a standard definition of our structure> structure cluster by leveraging Nectar and using CouchDB to control passing messages between nodes include message storing, duplication prevention, resource backup, location transparency, communication synchronization. We also implement error handling mechanism and parallel computing to enhance the fault tolerance ability of our system. We tested the scalability of the system with different number of instances and result in good performances. Twitter data and AURIN is used in our study. For tweets, the twitter API allows us to harvest tweets in the past 7 days. The retrieved tweets are in twitter json format containing the user information, tweet content, timestamp, geo-tag and so on. We will discuss how we utilize the information in later section. We implemented a hybrid crawler leveraging both search and stream API and successfully perform harvesting large amounts of tweets without hitting the twitter API access time limit. <add intro of aurin n aurin harvester>. In tweet crawler, we designed an embedded machine learning sentiment analyser for classifying whether the tweet is sentimentally positive or negative. We also designed a baseline for comparison. We tested our sentiment analyser on NLTK twitter sample [2] which result in 98.41% average f1-score and on sent140 corpus [3] which result in 66.7% average f1-score. We also implement a basic pattern matching method in generalising tweets related to a certain topic including sports, crime, tobacco consumption and so on. Hashtags are extracted from each tweet for finding trending hashtags around Australia. We extract tweet timepoint and labelled with four per-defined slot tags including “morning”, “afternoon”, “evening” and “midnight”. <map/reduce> <web> <ansible>

# System Design

## 2.1 Cluster architecture

<add something>

Add description of scalability, fault tolerance, backup, resource consistency or more…

## 2.2 System architecture

<add something>

# Data Collector and Processor

## 3.1 Hybrid crawler for tweets

In the beginning of study, we decide to find some interesting information among eight cities around Australia, Melbourne, Sydney, Canberra, Brisbane, Perth, Adelaide, Darwin, and Hobart, based on their tweets. The data collection is designed by referring Twitter API which provides both standard Search API and Stream API for harvesting tweets in past 7 days. However, standard Search API keeps a 15 minutes access time limit and Stream API is also restrict to one connection each time with one developer access tokens which limiting our efforts in getting sufficient data because the quantity of data strongly influences the analysis result as more data coming in, more normal and general our conclusion will be. Hence, we created a hybrid crawler leveraging both search API and stream API for fast harvesting without touching the access limit.

Firstly, we created a geo-location filter box by getting marginal coordinates from klokantech[4]. We use eight squares (Figure 1.1) to crop out the area we are interested in so there might be some mis-crops at the edge as city areas are not squares. Since our studying granularity is on city level instead of suburbs, a few mis-crops on the edge are statistically tolerable. Secondly, some tweets filtered from stream API may not contain precise coordination as point longitude/latitude but only contain the city name and a bounding box. In our study, we will not leverage the precise coordinates of each tweets and only focus on the city they came from.

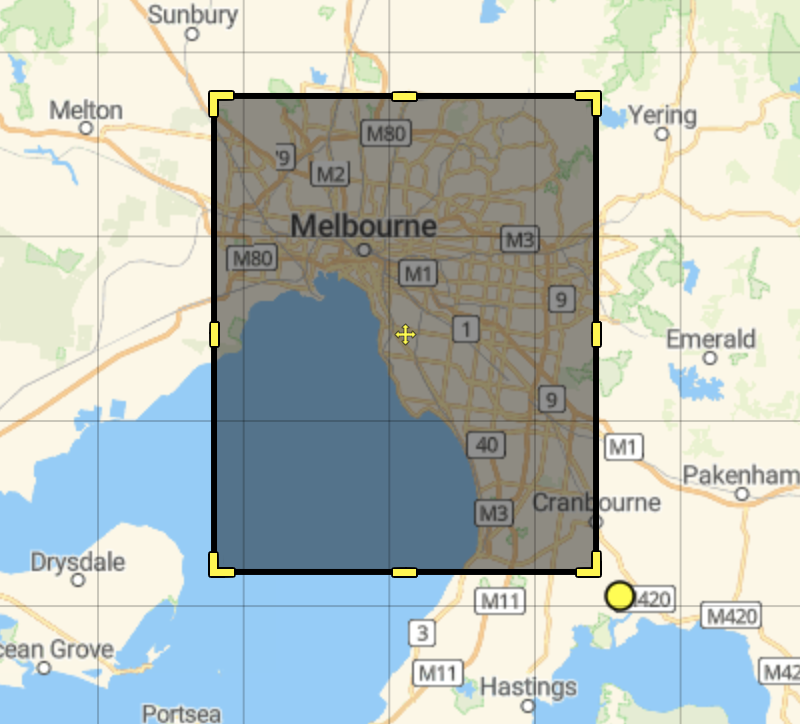


Figure 1.1 Bounding box for geo-filtering

Thirdly, we made an assumption that when a user posted a tweet from a typical city, other tweets from this user are likely to be posted from the same area. Therefore, we embedded a search API after harvesting one stream tweet and to make queries on that user’s timeline. In our test, the success search rate (number of useful tweets divided by number of total queries) increase rapidly. The query time decrease from 100 times per search to 15 times per search which successfully held up before accessing rate limit without slowing the search performance. In the meantime, we also implement a suspend/wake mechanism for search API to cease query after hitting rate limit and restart after a given time. During tweets harvesting, we implement an embedded sentiment analyser.

## 3.2 Embedded Sentiment Analyser

One of the most interesting information in our study is the polarity of people in different areas. We get this data by analysing tweets and classify them into two classes, positive and negative. We implemented a machine learning method leveraging textblob [5]. Textblob can score a sequence of texture symbols including English alphabets and some emojis ranging from -1(negative) to 1(positive). We firstly designed a baseline method to parse tweets without doing any pre-processing and test on NLTK twitter samples [2] and sent140 [3]. The baseline average f1-score on NLTK corpus is 96.32% and f1-score on sent140 is 42.24%. Then we implemented a pre-processor and tried different combination of pre-process methods including lemmatization, remove stop-words, lower alphabets and so on. We also created an extendable rule-based text parser to transform internet glossaries such as transforming OMG to oh my god. The improved average f1-score on NLTK corpus increased to 98.41% and on sent140 corpus increased to 66.7%. The reason that the performance on NLTK corpus is higher than sent140 is because all data in NLTK contain emojis while not in sent140 corpus. And performance of our sentiment classifier is influenced by the occurrence of emojis. (Compare show in Table 1.1).

Table 1.1 average f1-score compare of baseline and analyzer

Our analyser is not accurate enough compare to current benchmarks in sentiment analysis, but it will not cause bad influence on the final scenario study. We assume with sufficient number of tweets, the mis-labelled positive and negative tend to be normally distributed and mutual neutralized. Hence, when calculating the positive/negative rate, some numbers of mis-labelling are tolerable.

## 3.3 Topic parsing and Hashtag Parsing

Tweets may contain special topics that we are interested in. Are people more positive with AFL or cricket in the same city is an interesting study. Therefore, we implemented a basic pattern match topic tagger in the data processor. We constructed different extendable topic glossaries and make sure they will not overlap with each other. We created four topics including Tobacco, Crime, AFL, and Cricket. For tweets that exclude our defined topics will be tagged with null topic. Another interesting study is hashtag parsing. Each year, twitter will publish a summary on most popular hashtags people used around the world. We decided to summarize some popular hashtags around Australia in past 7 days and study people’s polarity trend on different hashtags. We use regular expression to extract all hashtags in each tweet and store them under hashtag keyword in a json format.

## 3.4 Tweet Timepoint Partition

In our study, timepoint of tweet stand for the point that one typical tweet was sent online is extracted and assigned with a timestamp defined by us. We partition 24 hours into four time slot and given each slot a timestamp name (Table 1.2).

|  |  |
| --- | --- |
| Morning | 07:00:00-12:59:59 |
| afternoon | 13:00:00-18:59:59 |
| Evening | 19:00:00-00:59:59 |
| midnight | 01:00:00-06:59:59 |

Table 1.2 Partition of 24 hours and correspond timestamp

## 3.5 Aurin collector and parser

<add something>

## 3.6 Format of processed data

|  |  |
| --- | --- |
| \_id | CouchDB unique document ID |
| \_rev | CoudhDB document rev |
| id\_str | Unique tweet id |
| coordinates | Twitter json coordinates |
| timestamp | morning/afternoon/evening/midnight |
| Place | Twitter place json |
| Place\_type | Granularity of the place, city in our study |
| Name | City name |
| Bounding box | Twitter json geo bounding box |
| Country\_code | AU |
| User | Information of user who posted tweet |
| Id | User id |
| Name | User name |
| description | User profile description |
| Lang | Language of tweet |
| Text | Text content of tweet |
| sentiment | Sentiment analysis information |
| Polarity | Range from -1 to 1 |
| subjectivity | Range from -1 to 1 |
| Label | Positive or Negative |
| Topic | Tobacco/Crime/AFL/Cricket/null |
| Hashtag | A list of hashtags or [] |

Table 1.3 Processed tweets

<add aurin format>

# CouchDB as Database

## 4.1 CouchDB in cluster

<add something related to sharding, structure, etc..>

## 4.2 Duplication prevention

In our system, three data processors in each slave work in parallel and save processed tweets into uniform CouchDB running on database instance. We leverage the automatic document duplication prevention mechanism in CouchDB to help us ignore harvesting redundant tweets. Each tweet was given unique id by twitter, and each document in CouchDB is given a unique id. Therefore, we use tweet ID as document ID and if there is a duplication exception from database, we will discard the tweets.

# Scenario Study

## 5.1 Happiest City Ranking

<add something>

## 5.2 I love tweet in the Morning

This scenario tries to find out during which part of the day people in different cities of Australia are sad. People are known to have different moods during different times of the day. This mood can be dependent on number of factors. For example, most people have early morning commitments such as work, meetings, lectures in University and does not usually enjoy getting up early. Then depending on how people enjoy their work and working with their colleagues, it can have an effect on their mood in the afternoon. Then in the evening, people may be relieved from their work, and would be excited to be back home mostly with their family and loved ones. Then their mood at night would depend on how the things are going on in their personal life. So, we thought that these different moods of the people at different times of the day can have significant impact on the number of sad tweets at different times of the day. So, in this scenario we tried to analyse whether the number of sad tweets differ during different times of the day. We divided the day into four time intervals namely 12.00 am to 6.00 am,6.00 am to 12 noon,12 noon to 6 pm and 6 pm to 12.00 am. We tried to find these observations city wise so that we can answer the questions like which city has more number of sad tweets during a particular time of the day or during which time of the day people of Sydney or Melbourne are saddest.

## 5.3 Marriage is a Disaster?

This scenario tries to compare the number of married and unmarried people in the city and tries to relate it with the number of happy/sad tweets in the city. There are many number of stories and regular discussions prevalent among people regarding the relation of marital status and happiness in life. So, we tried to analyse the same with the data available with us. We got the number of married and unmarried people from Aurin and populated into CouchDB. Then we already had polarity data on the tweets which we had got by performing sentiment analysis on twitter data. So, we tried to analyse whether the city which has maximum number of married people has more happy tweets or sad tweets. So, this scenario will let us answer the questions like which city has more number of married or unmarried people and whether this number has the significant impact on the number of happy or sad tweets coming in from this city.

## 5.4 Passion for Sports or Gambling?

This scenario tries to compare the number of people involved in gambling activities in a particular city with the number of sports tweets coming in from these cities. Gambling has been a very well-known platform for the sports lovers to use their knowledge of the game to predict the outcomes or the events in the game to earn money. It is said that along with the prediction skills, gambling requires the person to have a good knowledge about the sports on which that person is gambling. In recent years the social networking sites such as Twitter has been the medium for the sports lovers to express their opinions or emotions about the outcomes or the events in the sport that they love immensely. For example, there had been many tweets expressing opinions or we can say anger over the sandpaper incident which rocked the Australian cricket recently. There has also been lot of tweets about the AFL teams where AFL lovers expresses their opinion about their favourite teams or players. So, in this scenario we tried to analyse whether the number of sports tweets is higher in the city which have large number of people involved in the gambling activities. So, this scenario will let us answer the questions such has which city has the maximum number of sports related tweets, which city has maximum number of people involved in gambling activities and whether these factors correlate with each other for those cities.

## 5.5 A Rich Man’s Game?

This scenario tries to compare the number of sports tweets in the city with the financial status of the people in that city. It is a debatable topic that whether the rich people follow sports more or whether poor people follow it more. There can be factors such as access to the sport’s equipment’s the privilege of experiencing sports from the best possible place which can suggest that a rich person is more likely to be a sports enthusiast. But then there have been the stories of the famous sports persons who have shined from the bottom level financially. So that can suggest that poor people can also follow sports with the same enthusiasm as the rich people. Sports has always been the great source of entertainment from ages. It has also been credited with building the relations among the countries with varying cultures by bringing people together. People have emotions attached with the sports they follow, teams they support or the sportspersons they love or idolized. It provides them with great amount of refreshment from their busy schedule. So ideally both the rich and poor should have access to enjoy sports equally. So, by analysing this scenario we tried to study whether this is the case in Australia or not. Hence, this scenario will let us answer the questions such as which city has maximum median income for its people, which city has maximum number of sports tweets and whether these numbers correlate with each other for example if Sydney has more median income as compared to other cities then whether people who follow sports in Sydney or sports tweets coming from Sydney are more compared to those coming from other cities.

# System UI and User Guide

<add something>

# Reference

[1] Nectar Research Cloud, a collaborative Australian research platform supported by the National Collaborative Research Infrastructure Strategy (NCRIS).

[2] nltk twitter sample. copyright: Copyright (C) 2015 Twitter, Inc; license: Must be used subject to Twitter Developer Agreement (https://dev.twitter.com/overview/terms/agreement)

[3] sent140. [Twitter Sentiment Corpus](http://www.sananalytics.com/lab/twitter-sentiment/) by Niek Sanders

[4] http://boundingbox.klokantech.com.

[5] Textblob. http://textblob.readthedocs.io/en/dev/