**COMP90024**

**Cluster and Cloud Computing Assignment 2**

**Australian Social Media Analytics**

**Team No 49.**

|  |  |  |  |
| --- | --- | --- | --- |
| City | Name | Student ID | Email Id |
| Melbourne |  |  |  |
| Melbourne |  |  |  |
| Melbourne |  |  |  |
| Melbourne |  |  |  |
| Melbourne |  |  |  |

**Abstract**

Twitter is a popular social media platform containing large amount of texture data. Aurin(Australian Urban Research Infrastructure Network) 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. Using this environment we will be mining interesting geoinformation by summarizing tweets from eight cities around Australia and combine them with city based information we accessed from 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

[1. Introduction 3](#_Toc513676193)

[2. System Design 3](#_Toc513676194)

[2.1 Cluster architecture 3](#_Toc513676195)

[2.2 System architecture 3](#_Toc513676196)

[3. Data Collector and Processor 4](#_Toc513676197)

[3.1 Hybrid crawler for tweets 4](#_Toc513676198)

[3.2 Embedded Sentiment Analyser 5](#_Toc513676199)

[3.3 Topic parsing and Hashtag Parsing 6](#_Toc513676200)

[3.4 Tweet Timepoint Partition 7](#_Toc513676201)

[3.5 Aurin collector and parser 7](#_Toc513676202)

[3.6 Format of processed data 7](#_Toc513676203)

[<add aurin format> 8](#_Toc513676204)

[4. Ansible and Boto 8](#_Toc513676205)

[5. CouchDB as Database 8](#_Toc513676206)

[4.1 CouchDB in cluster 8](#_Toc513676207)

[4.2 Duplication prevention 8](#_Toc513676208)

[6. Scenario Study 8](#_Toc513676209)

[5.1 I love tweet in the Morning 8](#_Toc513676210)

[5.2 Passion for Sports or Gambling? 10](#_Toc513676211)

[5.3 Marriage is a Disaster? 11](#_Toc513676212)

[5.4 A Rich Man’s Game? 12](#_Toc513676213)

[7. System UI and User Guide 13](#_Toc513676214)

[Reference 13](#_Toc513676215)

# Introduction

Nectar [1] stands for National eResearch Collaboration Tools and Resources project. It gives us fixed number of computational resources, thus allowing researchers to create a cluster according to 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. For getting the required data from Aurin we implemented a Aurin parser which parses the required data from Aurin which is in Json format and populate it into CouchDb. 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

Before beginning the implementation of the project, we sat down to decide which whether we will be collecting data suburb wise or city wise. During this discussion, we decided to find some information among eight cities around Australia.We choosed city wise as large amounts of data were available . We choose 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.

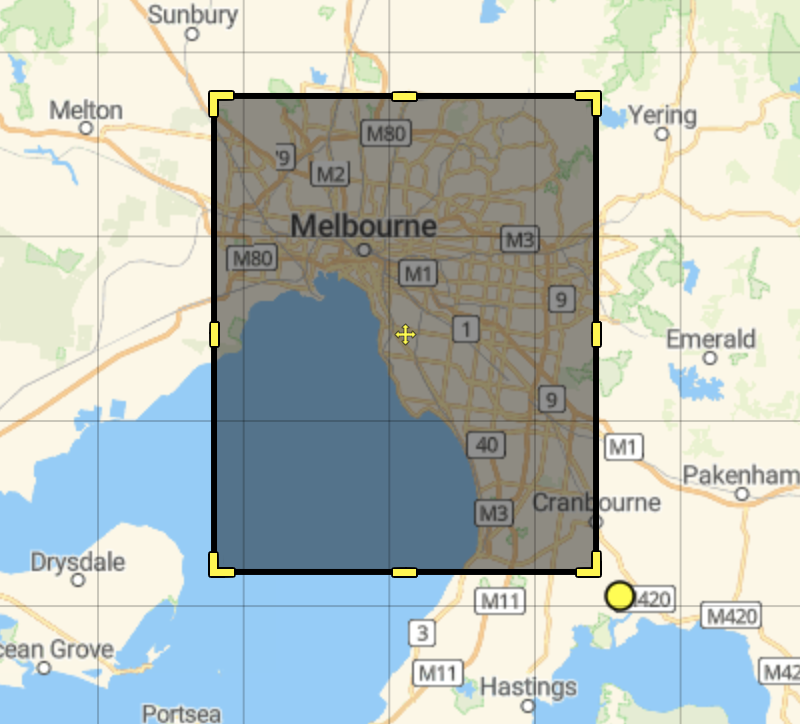


Figure 1.1 Bounding box for geo-filtering

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

# Ansible and Boto

4.1 Ansible

4.2 Boto

# CouchDB as Database

## 5.1 CouchDB

CouchDB is one of a new breed of database management system called NoSQL. Specifically CouchDB is a document oriented database where data is stored in JSON format and each document field is stored in key-value pair, map or list.

Some of the key features of CouchDB are:

1. HTTP­based RESTful APIs

CouchDB is easily accessed via port 5984 once installed. We were able to access the database and perform CRUD operations on data using simple HTTP requests. In this project, We load the JSON through the curl –X put command and access the view from curl –X get command.

Futon

CouchDB provides its users with a web­based administration console called Futon. It gives us a clear visualization of data and design views. Futon presents the JSON document in a more readable format so humans can easily understand the underlying structure. Another important advantage of futon is it allows us to create temporary view in the database which helps us to test and debug our Map Reduce function without wasting space for storing it.

1. Document oriented storage

CouchDB uses schema free JSON documents for storing data, this means the documents are stored like the real world document, which are self-contained without any data model structure. This makes the designing more scalable and flexible as we can store the document directly rather than dividing and inserting into separate table.

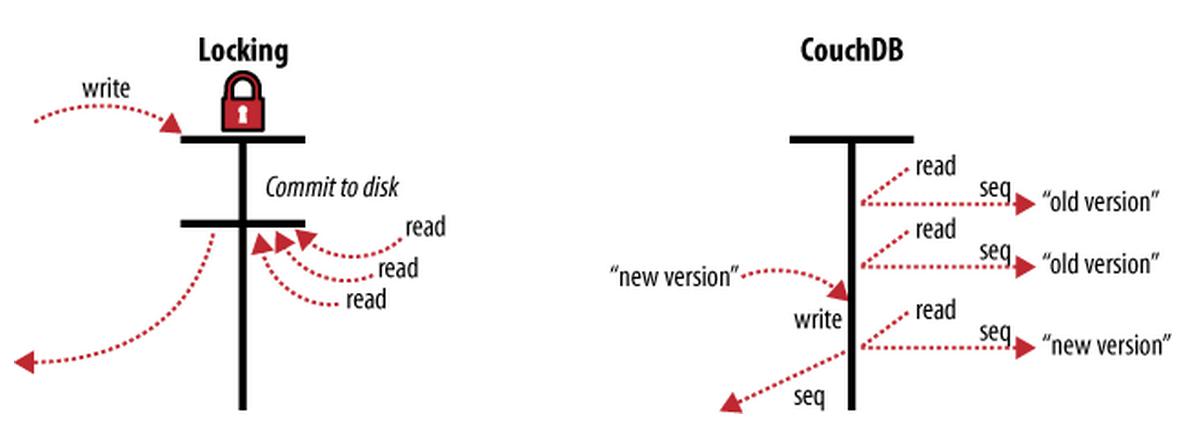
1. MapReduce

MapReduce in CouchDB is used to create views on the database. In Map() function we filter and sort each document in database and in Reduce() function we summarize that. This Technique is really powerful and efficient as it process large amount of data in less time. Although it takes considerable time at first to build B-tree for the view, but once views are created the query time is less. Another advantage of having the B-tree structure is that it is scalable, when new data is added it will perform the required calculation on the new data rather than building the entire tree unless the MapReduce function remains same.

In our project we use the Map function to create a particular view which filters the required data and in reduce function we count the occurrence of the targeted topic.

1. No Locking

Another advantage of using CouchDB is it does not have any locking mechanism and uses Multi-Version Concurrency Control (MVCC) to organize concurrent access to database from multiple user. Thus CouchDB does not block user reading the data when update is happening.



Couchdb(http://docs.couchdb.org/en/2.0.0/intro/consistency.html)

CouchDB generates different version of data when updates occur. This feature helps to effectively utilize the processing power as clients won’t have to wait while the system is processing other requests. The system is able to perform analysis on existing data while the harvester is still adding new data to the database

## 5.2 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.

5.3 CouchDB Design

The Twitter API procedure and the CouchDB expects the data in JSON format which is the standard format in web data interchanges. As a pre-existing twitter data was already available to use as a preliminary source, the final structure for tweets in CouchDB is mirrored to some extent.

CouchDB’s *\_id* and *\_rev* tags are top level tags provided by CouchDB. In our case, *\_id* has been manually adjusted by the harvester to mirror that of the tweet id, specifically *id\_str.* This technique is used to ease the handling of duplicated tweets in the database, it offloads any duplicate checking directly to CouchDB, if the \_id that exist is attempted to be the CouchDB will return conflict error.

Additionally, the Twitter Harvester provides *sentiment*, *AFL, cricket, timestamp* tag to each document.

The *sentiment* tag can have two possible values, either 0 to -1 or 0 to 1. These values relate to either a negative sentiment (>0), or positive sentiment (<0).

The *AFL* tag captures the all the tweets that have mentioned topic of AFL. The Topic contains keywords related to AFL. For Example the topic of AFL have keyword like footy ladder, arena etc.

The *cricket* tag captures the all the tweets that have mentioned topic of cricket. The Topic contains keywords related to Cricket. For Example the topic of cricket have keyword like bowling wicket, sledging etc.

The Timestamp tag captures the tweets for time period. The time period is divided into three period morning, afternoon and night. Morning time period being 4am to 12pm, afternoon time period being 12 pm to 8 pm and the night period being 8 pm to 4 am.

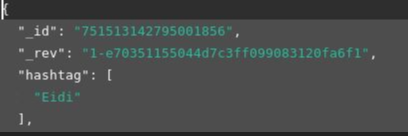


Figure - Top level document Json data structure

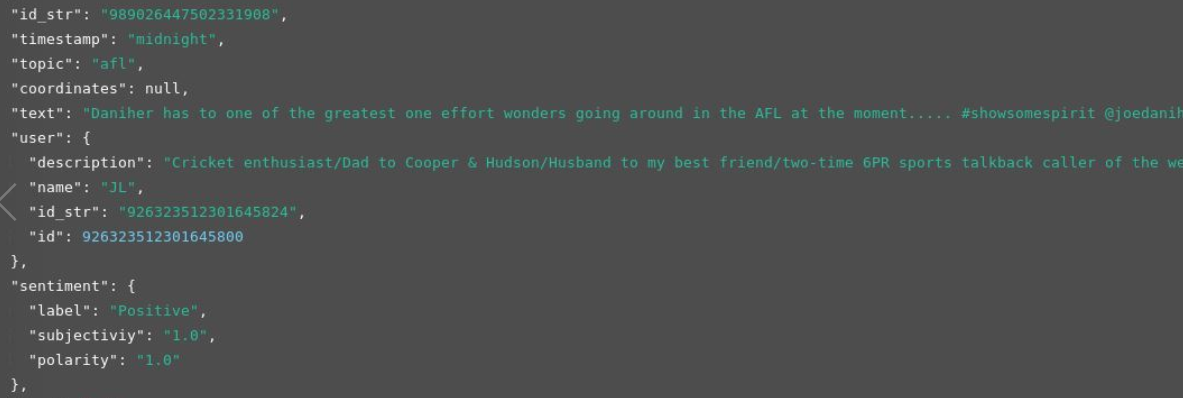


Figure - Tweet Json data structure for sentiments and AFL



Figure - Tweet Json data structure for timestamp and cricket

CouchDB’s design document and views are created by uploading a Json file containing the MapReduce functions and relevant design document.

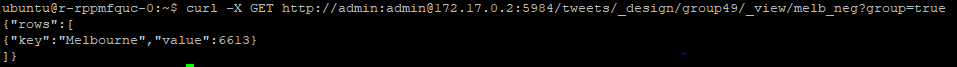
curl -X PUT 'http://localhost:5984/twitter-data/tweets/\_design/group49/\_view' -d @melb\_neg.json

Each view contains a Map function and optionally a reduce function. The structure is that of a nested Json document, which each view being a child of the views field. Map and Reduce functions are string representations of function calls written in Javascript



Figure: example of CouchDB views in Json format, showing Map and Reduce functions

Internally, the Node.js web server returns relevant data to the user, as Json, via requesting the respective views from CouchDB. Additional parameters may be imposed on the returned data



# Scenario Study

## 6.1 I love tweet in the Morning

**Description:**

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 labelled this time intervals as midnight,morning,afternoon and evening respectively. 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.

**Visualization:**

**Observation:**

As we can observe from the graph, the rate of sad tweets in morning in almost all the cities is more as compared to any other time of the day. So this statistics supports our belief that morning is the time where people are usually in a cosy mood and are reluctant to start their day early. This can also imply that the people in Australia usually likes to sleep late and due to incomplete sleep, the mood is normally bad in the morning which results in more sad tweets during this time of the day. City wise, Darwin has been the city with highest rate of negative tweets. This maybe also because Darwin has the more youth population. Hobart has the lowest rate of sad tweets among all the cities. The rate of sad tweets can be seen decreasing as the day progresses. This maybe indication that the people gets settled in their lifestyle as the day progresses and starts to tweet more positively. City wise, Darwin again has the highest rate of sad tweets in afternoon while Canberra, the capital city of Australia has the lowest rate. That is the good news for Australia as the administrative and political functioning of whole of Australia is done in Canberra as it is the site of Parliament House and numerous government departments and agencies. So we can say that the people in the most important city strategically has the less number of people who are sad at the most important time of the day work wise. Then again we can observe that the rate of sad tweets goes increases slightly during midnight. Of course the rate don’t go up as it was in morning. This maybe an indication of the pressure of the uncertainty about the next day. As we know the human mind tends to think too much about the future. During midnight, we normally thinks about the things we have to do tomorrow, deadlines we have to meet in the coming days, commitments we have to fulfil. So that anxiety may lead the people to tweet negatively during midnight. So that can be the analysis made of the sad tweets rate increase during midnight. City wise, Canberra has the highest sad tweets rate while Perth has the lowest rate. Do the load of being part of the administrative capital of Australia taking on for Canberra people during their thoughts on Midnight? Probably yes. Darwin has the 2nd lowest rate of sad tweets during midnight. That’s significant improvement from morning’s rate for Darwin. So interestingly, we can analyse that Darwin people are more happy during the later part of the day than initial part of the day. We expected the rate of sad tweets to be lowest for Melbourne for at least one part of the day as it is considered to be the most liveable city of the World but sadly, it was not able to satisfy the expectations from it.

## 6.2 Passion for Sports or Gambling?

**Description:**

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.

**Visualization:**

**Observation:**

OnY axis, we are capturing the weighted average of the people involved in gambling in different cities of Australia and weighted average of the sports related tweets coming from these cities. We have taken the weighted average of both these components to reduce the bias towards the number of people in the city and number of tweets coming in from these cities. As wen can observe the weighted average of the people involved in gambling is highest in Brisbane while it is lowest in Sydney. The number of sports tweets tell a different story as more number of Sports tweets are coming from Melbourne and those coming from Brisbane are less than those compare to the other cities. So we can observe that the number of sports tweets does not correlate with the number of people involved in gambling for most of the cities in Australia. So from the data we can analyse that it is not necessary that the people involved in gambling are sports lovers. But we cant make concrete inferences from this data as number of sports tweets coming in from the cities of Australia in recent days is not that high. These maybe because no major sporting event is going on over whole of Australia currently. We could have got more number of sports tweets during Commonwealth Games time. So based on available data we can observe that there is not much correlation between number of gambling activities in the city and number of sports tweets coming in from these cities.

## 6.3 Marriage is a Disaster?

**Description:**

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.

**Visualization:**

**Observation:**

One graph shows the weighted average of the married people in different cities of Australia while other graph shows average sentiment score for the tweets coming from these cities. For finding the weighted average of married people, we have taken into consideration the population of the city to ensure that there is no bias towards the cities which have high populations. More the average sentiment score than one indicates more rate of positive tweets coming from the city. As we can see from the graph, the city with highest weighted average of marriage people is Sydney while Brisbane has lowest weighted average of marriage people. If we have a look at the sentiment score of these cities, we can see that Sydney has pretty good Sentiment Score and that of Brisbane is less than that of Sydney. The city with the highest Sentiment score is Hobart and it was 6th highest weighted average of the married people. So the observations contradicts the belief that the happiness of the people is inversely proportional to the marital status. We are able to witness from the data of the most of the cities that if the city has more number of married people than other city, then the average sentiment score of that city is also higher than the city with less number of married people. So we can analyse that more number of positive tweets are coming from the cities having more number of married people. This scenario captures the interesting marital relationships analysis about the people in different cities of Australia.

## 6.4 A Rich Man’s Game?

**Description:**

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.

**Visualization:**

**Observation:**

One graph shows the number of sports tweets coming from different cities of Australia, while the other graph shows the average monthly income of these cities. As we can observe that more number of sports tweets are coming from Melbourne while the least number of tweets are coming from Darwin. The most popular sports followed in Australia are Cricket and AFL and its not surprising that number of sports tweets are coming more from Melbourne. The sports facilities in Melbourne are world famous with Melbourne Cricket Ground being one of the most renowned Cricket Stadium in the world which also has a highest seating capacity among all the cricket grounds in the world. The craze of AFL is also at peak in Melbourne currently. On the other hand, average monthly income of the people in Melbourne is less than that of the people in other cities. Darwin has the highest average monthly income among the Australian cities, but the number of sports tweets coming in from Darwin is substantially less. The observation can be made that more tweets are coming from cities which are known globally for sports facilities. For example, Sydney has a good cricket facilities as well as have hosted Olympics previously and has fair bit amount of sports tweets coming in. So that’s good news that the income in the city is not directly proportional to the number of sports enthusiasts in the city. We also managed to make an extra analysis that number of sports lovers in the city correlates with the sports facilities in that city while making an observation for this scenario.

# 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/