Cluster and Cloud Computing COMP90024

Project 2 Report

Team 33

Ke Yang, 1219623 - city: Anhui

Keang Xu, 1008807 - city: Hubei

Jintong Liu, 1074498 - city: Hebei

Xinwei Qian, 1068271 - city: Jiangsu

Yimeng Liu, 1074206 - city: Guangdong

May 16, 2022

Github https://github.com/KeYANG813/COMP90024Group33.git

YouTube https://www.youtube.com/playlist?list=PLlzQWmBYCR2yelx5vJ9zaeI22Yf3cuMq-

Abstract

This is the report for COMP90024 Group33 Assignment 2. Nowadays Twitter is one of the most popular online social networking tools. Therefore, the research of Twitter tends to reveal people's lifestyles and hot topics they are interested in. The purpose of our project is to explore the liveability of Melbourne through collecting tweets and analyzing some social media scenarios based on different indicators of liveability. The system and application architecture is built on the Melbourne Research Cloud (MRC) using OpenStack API. The tweet data is crawled by Twitter API and stored in the different databases through CouchDB. For the automatic deployment, Ansible is used to automate the four virtual machines on MRC, set up software environments and deploy crawlers, databases and web services. The docker is utilized for packing the code and libraries into containers, which can quickly deploy and scale our application into any environment. Flask is chosen for the backend framework of the web application and transfers the processed results from back-end to front-end by RESTful API, which accesses data from HTTP requests. The application renders a static HTML page on the front-end to visualize the analyzing results.

Contents

1 Introduction

The Melbourne Research Cloud (MRC) is largely Infrastructure-as-a-Service (IaaS), supplying a virtual server instance which could let users store data as well as deploy applications on cloud. In this project, a cloud system will be developed to collect and analyze the tweets from different cities in Australia using Twitter API, then achieve the data visualization on a web application. Twitter is a very popular and commonly used social media in Australia, there is a lot of information that can be extracted from tweets, for example, a variety of emotions shown based on sentiment analysis of sentences, and hot topics discussed in public could be summaries based on hashtags. The analysis of live tweets in Australia which compared with the official data from AURIN could provide some insights into the relationship between people's activities on social media and some factors(e.g. multiculture or income levels) that might impact the livability of a city. Moreover, a horizontal comparison between five cities in Australia: Melbourne, Sydney, Adelaide, Darwin, Brisbane is also discussed for liveability. The related information about total tweets and tweets for each city, language distribution, Top-10 hashtags, as well as the sentiment analysis for Top-10 hashtags are visualized and analyzed.

In section 2, we talk about System Architecture and Design in detail. In section 3, the introduction of the main deployment tools (MRC, Ansible and Docker) and the benefits of choosing them will be discussed. In section 4, tools used for data collection and some error handling is stated. In section 5, the tool for data storage and an overview of the database is demonstrated. In section 6 we introduce the tools for deploying the web application, how the backend and frontend were developed, and also include a detailed description of charts chosen for data visualization. In section 7 a deployment guide is presented which could be followed to rebuild our cloud system. In section 8, a detailed analysis of 4 scenarios will be displayed. In section 9, there is a brief summary of the overall project and what we have learned throughout this semester about this project.

2 System Architecture and Design

2.1 Twitter Data Management

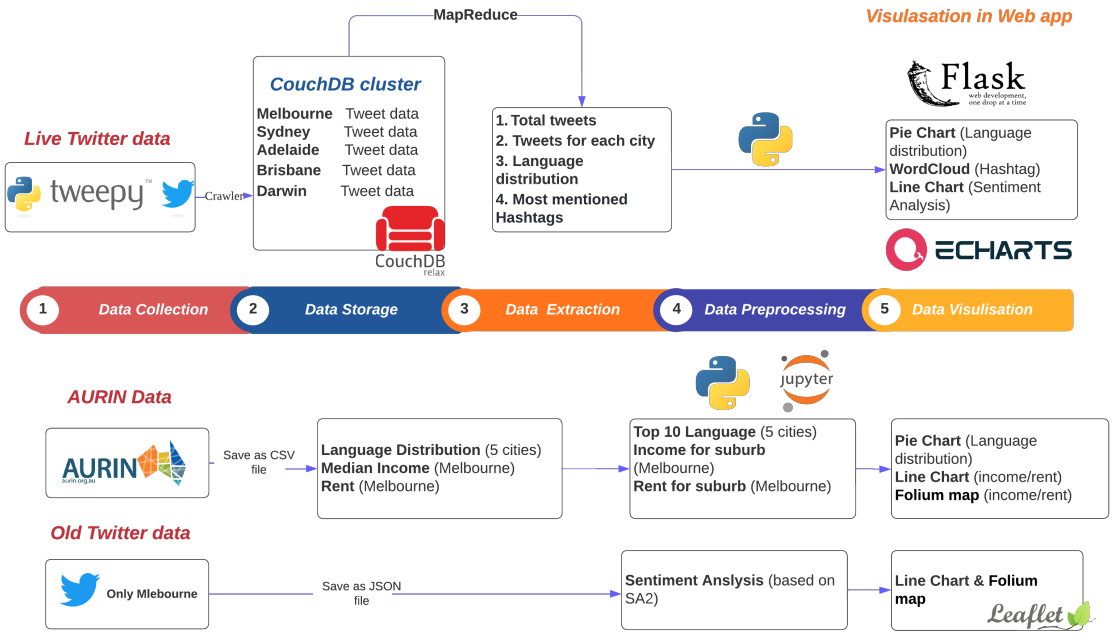

Figure 1: Pipeline of Data Management System

The overall workflow of data management is shown in Figure 1.
For the live tweets:

- The live tweet data of different cities are crawled from the **Twitter API**.
- The live tweets are stored in the **CouchDB** cluster, there are 5 databases for all cities and tweets were stored in the corresponding city based on geo-location.
- The desired results are queried through **MapReduce** from CouchDB, saved in the specific route in the **Flask** backend and ready for the visualization in web application.
- The web application is deployed to visualize different scenarios which are analyzed from the live tweets.

For the AURIN data:
The official data for different cities are collected from AURIN, by comparing with the live tweets/old tweets data, we can have a more clear understanding of the most liveable cities in Australia/most liveable suburb in Melbourne.

2.2 Overall System Architecture of Cloud System

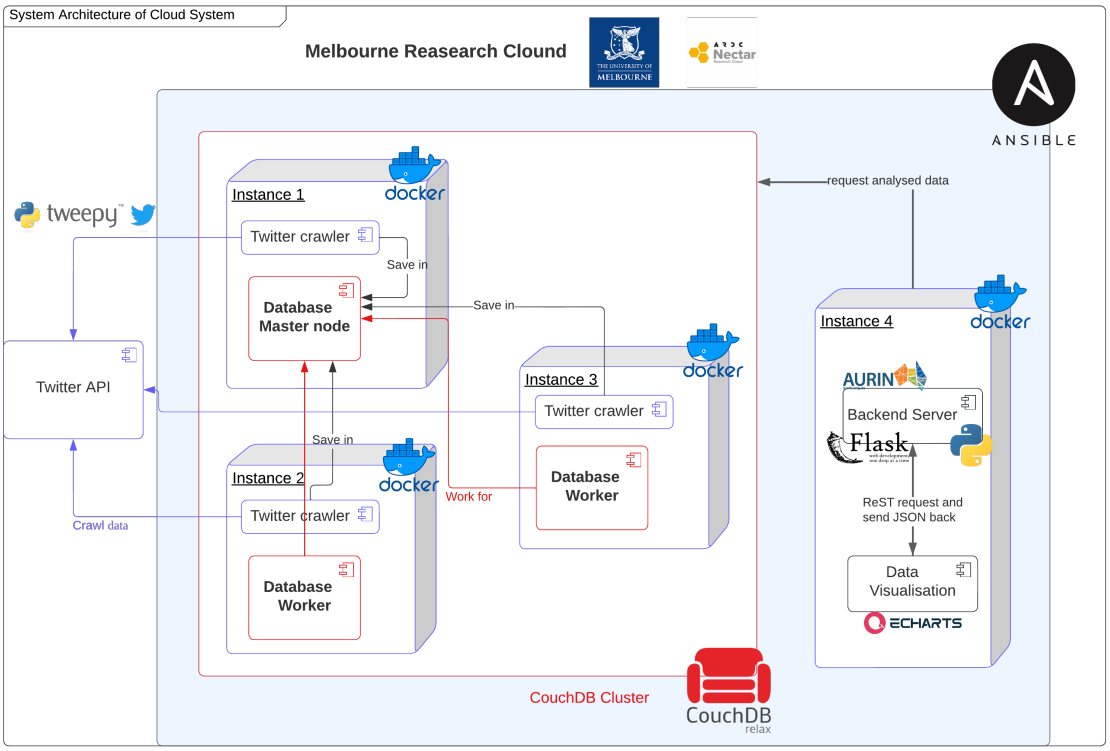

Figure 2: Cloud System Architecture

The above Figure 2 shows the overview of system architecture labeled with the corresponding deployment tools. There are totally 4 virtual machines running on the Melbourne Research Cloud, each instance has 2 virtual CPUs with 9GB memory in total, the provided cloud resource could support our system to achieve scenarios analyzing [Section 3.1].
Each instance (server) is design for specific use:

- Instance 1, instance 2 and instance 3 run crawlers using Tweepy for data collection [Section 4] and managing the CouchDB cluster for data storage [Section 5]. Instead of crawling tweets on master node (instance 1) only, we crawled Melbourne, Sydney tweets on instance 1 and instance 2 respectively (since they have greater data volume compared to other cities), and crawled the rest of 3 cities' tweets on instance 3 in order to balance the pressure among servers.

- Instance 4 is designed for both analyzing data and deploying web applications for data visualization [Section 6].

Docker is installed in 4 instances to support all application components (e.g.Twitter Crawler, CouchDB), not only benefit for the rapid deployments but also enable scalability and flexibility. [Section 3.3]

The overall cloud system including creation of instances and the following deployment of services are able to run automatically supported by Ansible [Section 3.2].

3 Deployment Tools

3.1 Melbourne Research Cloud

The Melbourne Research Cloud (MRC) offers free on-demand computing resources to researchers at the University of Melbourne. [1] In this project, we need a server to deploy our data analysis with huge data volume and host our web application, The University of Melbourne provides us a chance to deploy them on the Melbourne Research Cloud. We use the MRC with the interaction of OpenStack for creating the instance, which is an open source cloud computing infrastructure software project. We list and discuss the pros and cons of using Melbourne Research Cloud based on our experience during testing and developing systems.

3.1.1 Pros

- **Cost**
  Compared to the current popular commercial cloud providers such as Amazon Web Services (AWS) and Google Cloud Platform which are deployed on a pay-per-use basis, Melbourne Research Cloud is free for cloud services.

- **Availability**
  Firstly, The instance run on cloud 24/7 and we don't need to worry about the issue associated with stability. Secondly, a variety of official images which used to populate a blank instance can be directly accessed on Melbourne Research Cloud, we can develop instance on pre-made image: "NeCTAR Ubuntu 20.04 LTS (Focal) amd64 (with Docker)".

- **Security**
  We don't need to pay extra attention to encrypt accessing passwords or other security issues. Firstly, Key Pair authentication only allows our team members who have a private key to access each instance, and the API access password is automatically encrypted. Secondly, we can specify the security group when creating the instance, the state of the port on each instance could be adjusted based on our purpose.

- **Data Backup**
  Data snapshot technology allows us to spend little time backing up data without affecting the network bandwidth. Besides, we can use backup to quickly recover data. When an accident occurs, we can roll back to the data version at the specified point in time to recover data. In addition, we can use the data of the virtual hard disk generated by the snapshot to test the new application or the new operating system version, which can avoid damage to the production data and will not affect the currently running applications.

3.1.2 Cons

- **Accessibility**
  Melbourne Research Cloud is not convenient for students who try to connect to cloud outside of the university network, Cisco AnyConnect VPN is needed but it is always associated with slow-loading issues due to the network traffic.

3.2 Ansible

Ansible enables its users to "scale automation, manage complex deployments and speed productivity with enterprise automation platforms" (Red Hat Inc., 2020). We use Ansible to automate creation

of instances on MRC, set up software environments and deploy crawler, CouchDB and web services. We won't need to write custom code to automate your systems or configure the application manually. When we run a playbook from the control machine, Ansible uses SSH to communicate with the remote hosts and run all the tasks for parallel deployment, which is time-saving.The Ansible playbook that was created for this assignment enables the deployment and setup of the whole system from the bootstrapping process. Overall, a total of fifteen tasks contribute to the playbook of the system. Table shown below provides the detailed explanation of the Ansible playbook tasks in specific steps.

Hosts	**Roles**	**Description**
localhost	openstack-common	Install python3-pip, openstacksdk and update pip on localhost
localhost	openstack-images	Show all available Openstack images with names and ids
localhost	openstack-volume	Create volumes on MRC based on settings in host_vars
localhost	openstack-security-group	Create security group with rules based on settings in host_vars
localhost	openstack-instance	Create instances on MRC and add corresponding volume and security groups

Before installing the docker into each instance, we add proxy environment variables for each instance. HTTP/HTTPS proxy servers offer many uses, including a wide range of security benefits. They enable users to hide their IP addresses to anonymize browsing activity and can act as a tunnel or network link that provides devices with restricted access to networks. So, HTTP/HTTPS proxies are helpful for internet users who want to hide their identity online, access prohibited websites, or enforce access policies on specific websites.The URL, http://wwwproxy.unimelb.edu.au:8000/, is used in our project to provide HTTP/HTTPS service.

Hosts	**Roles**	**Description**
instances	add-proxy	Add proxy in /etc/environment, then allow internet connections
instances	docker-prereq-env	Install/update pip, install dependencies for docker
instances	docker-install	Install docker, docker-compose
instances	docker-mount-volume	Mount volumes created before. Each instance will allocate to a volume and mounted to file system created /dev/vdb

Hosts	**Roles**	**Description**
DataNodes	couchdb-install	Create a docker container and pull existed CouchDB image provided by IBM with version 3.0.0 into each node of DataNodes (Instance 1-3)
Masternode	couchdb-finish-setup	Create and set up CouchDB clusters on Masternode (Instance 1)

Hosts	**Roles**	**Description**
Masternode	twitter-crawler-mel	Deploy the crawler in docker container to collect tweets from Melbourne
instance2	twitter-crawler-syd	Deploy the crawler in docker container to collect tweets from Sydney
instance3	twitter-crawler-three	Deploy the crawlers in docker container to collect tweets from other 3 cities

Hosts	**Roles**	**Description**
ProcessingNodes	app-git-clone-repository	Git clone the source code from Github repo, then deploy our web application in the docker compose, including the data processing&analysis(backend) and visualisation (frontend)

3.3 Docker

3.3.1 Containerization using Docker

Docker is the containerization platform which packages our source codes and all its dependencies together, enabling them to run in different operating-system. In other words, it keeps consistency between the developer's laptop (local environment) and deployment environment (on Melbourne Research Cloud). The most obvious advantage of using Docker is that it could guarantee the same source code running correctly on different OS our members used, for example, our members work in Windows (Ubuntu installed) or Mac OS, if we try to run the same code there will be errors come up sometimes and it is hard to figure out compatibility issue, but the program could run anywhere within container.

In addition, Docker provides an isolated environment for each container, which makes sure different containers could run isolated and concurrently on each server/instance. For example, in instance 1, CouchDB running in one container and crawler running in another container, Docker provides a separate isolated environment configuration for each container so that they will not affect each other, which means a single service/container could be testing, updating and deploying separately.

3.3.2 Docker Compose

In our project, three instances are allocated to crawler tweets. We are interested in analyzing tweets in five cities(Melbourne, Sydney, Adelaide, Darwin, Brisbane). Since the volumes of tweets in Melbourne and Sydney are quite huge, we arrange instance 1 to crawl Melbourne's tweets, instance 2 to crawl Sydney's tweets and instance 3 to crawl the tweets from other 3 cities.

In this part, we use instance 3 (the instance to collect tweets from 3 cities) as an example, the directory tree shown in Figure 3 below.

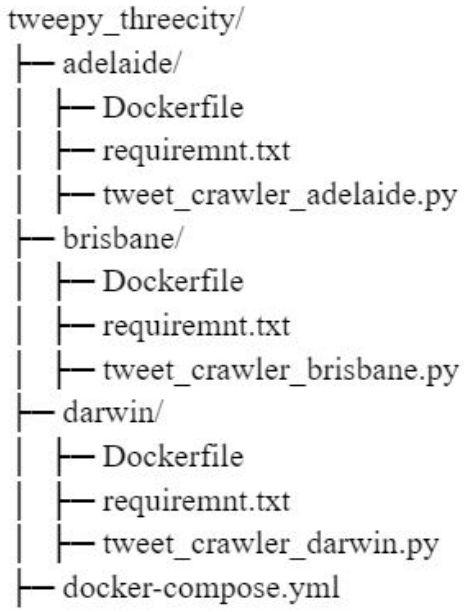

Figure 3: Directory tree for tweepy_threecity

Figure 4: docker-compose.yml file

Docker Compose is used for running multiple containers as a single service on a single host. For

example, we aim to crawl tweets from 3 cities together instead of starting each one repeatedly. A docker-compose file makes it possible to run them simultaneously which not only ensures easy configuration but also high productivity.

The other benefit is bind mounts which can avoid creating a new image if code was modified, As shown in Figure 4, `docker-compose.yml` file mount the local directory `./adelaide` to the shared folder `/usr/src/twpy1` in a running container. It could decrease the times of spending on rebuilding the Docker images which speed up the dockerized development workflow.

4 Data collection

We collect the data of our project through three channels, real -time Twitter collected through Twitter API, the old Twitter data provided by the university, and statistics from various Australian regions from AURIN.

We used Twitter API to collect Twitter information of 5 different cities (i.e. Adelaide, Brisbane, Darwin, Melbourne and Sydney), and use these data to compare the degree of liveability between different cities. We also analyzed the degree of liveability of each SA2 area in Melbourne in the old Twitter data. At the same time, we use data on AURIN to compare and verify our discovery.

4.1 Twitter API

4.1.1 Data selection

Our entire crawler is based on Twitter API V1. We wrote an automated deployment script based on Python (Tweepy), using the bounding-box (the boundaries of latitude and longitude) to automatically obtain real-time Twitter in different cities and store them in CouchDB after pre-processing. The border frame comes from the AURIN's large capital city statistics.

4.1.2 Crawler deployment

From 4.20 till the present, we crawled approximately 300,000 Twitter data from 5 cities in Australia. The rate of crawlers was about 10,000 per day. Five crawlers, corresponding to 5 cities, are deployed in three instances. We have applied for five API tokens of Twitter accounts that are assigned for each crawler. Five crawlers are executed in parallel. Finally, the crawled tweets will be stored in CouchDB by streaming.

4.1.3 Remove duplication processing

Removing duplicate tweets is a major problem. Our solution is to use tweet ID as the key of the document in the CouchDB database, which ensures that the tweets in the entire database are unique.

4.1.4 Problems encountered

Only some of the group members have successfully obtained Twitter's advanced account authorization. This limits the speed for crawling tweets. Because we found that only about 2% of tweets have the accurate geographical location.So our real-time tweet comparison is limited to the city level, we do not have enough tweets with accurate geographical location information, most of the tweet geographical information comes from the bounding box.

4.1.5 Error Handling

In the phase of crawling tweets, there are many exceptions that have to be handled to prevent the crawler from unexpectedly terminating.Those exceptions are shown in the following table.

Exception	**Handler**
Create Cloudant client fail	Pop up error message and exit
Create database fail	Pop up error message and exit
ProtocolError	Pop up timestamp and protocol error message
AttributeError	Pop up timestamp and attribute error message
Other Exception	Pop up timestamp and unknown exception message
420 status code	Sleep 10s
429 status code	Sleep 15 mins
Other unknown status code	Sleep 10s

4.2 Old Twitter Data

University also provides us with old Twitter data, exactly 2,500,000 Tweets. We finally obtained 434,847 rows of useful data by filtering tweets containing accurate geographical location and effective text. Time span is about the second half of 2014. Compared with real -time Twitter data, old Twitter data contains a higher proportion of geographical location, about 17%, and there are some data fields that we do not have permissions. Regions of Melbourne were divided into Statistical Areas Level 2 (SA2), and scenarios for these areas are analyzed through combining the AURIN data with the old Twitter data.

4.3 AURIN data

AURIN provides extensive data in Australia, such as Australia population distribution, per capita income, health care card and job seeker information.

In this project, we focus on the per capita income and rent in each SA2. Our analysis is mainly based on the census data of 2016 and 2021 (correspond to the old Twitter data in 2016). Simultaneously, we use AURIN to visualize data. Through comparison of AURIN's plots and results computed from old tweets, we can find some interesting conclusions.

5 Data Storage

5.1 CouchDB advantages

Compared to RDBM, No-SQL databases can handle large volumes of data at high speed with a scale-out architecture. Among many No-SQL databases, CouchBD is more outstanding with these following advantages:

- More flexible(JSON) document model.
- Simple HTTP-based REST API for document insertion,updates, retrieval and deletion.
- Fast indexing and retrieval.

- Easy replication of a database across multiple server instances.
- Outstanding cluster management(replication and sharding etc.) improves the availability and achieves the redundancy of the database.

5.2 CouchDB cluster settings

We select three instances to be the CouchDB servers, among which one instance as master node and two instances as child nodes. Then these three nodes are combined as a cluster. Thanks to the convenient feature of CouchDB, we only have to control and manage the master node because the CouchDB systems will automatically create several copies and shards to make the database fault-tolerant and distribute the computing load across each node. For our project, the default setting(3 copies and 2 shards) is used.

5.3 Overview of data storage

In this application, the CouchDB cluster consisting of three nodes stores the tweets from five different cities(Melbourne, Sydney, Brisbane, Adelaide and Darwin). Besides, three views of hashtags, languages and tweet count are created by the MapReduce technique.

The following table summarizes the format of docs store in the cities' database and an example is provided (only the attributes which we are interested in):

Field	**Description**
_id	The Tweet ID instead of the auto-generated ID by CouchDB. Before storing a new tweet into the database, if the ID of tweet already exists, skip it in order to avoid duplication.
create_at	The date that the tweet was posted.
user	The details of the author, including user ID, name and location of user etc.
place	The location that the tweet was posted, including some basic information of the place, such as geo (the latitude and longitude).
place.bounding_box	An area defined by two longitudes and two latitudes.
entities.hashtags	Constitutes several keywords that describe the category of the tweet.
lang	The language that the tweet uses.

5.4 Interaction of CouchDB

After setting up the nodes in each instance, we have to know how to utilize the database. This part will show the main operations of CouchDB in our application.

5.4.1 MapReduce

MapReduce is the key technique to solve problems in massive data. It applies a two-step process, map phase and reduce phase. The map phase looks at all documents in one specific database separately and creates a key-pairs result, which is called View. To extract the certain data field, we can design our own map function as needed. The reduce phase does some basic statistics in a certain View, such as count, sum and even custom reduce function. The process that the processed data in map step is distributed to different nodes for the next reduce step which is called shuffling. The map phase distributes data across machines, while the reduce phase hierarchically summarizes them until the result is obtained. Apart from parallelism, its advantage lies in moving the process to where

data is, greatly reducing network traffic. In this project, the python module Cloudant is a scalable and distributed cloud database based on Apache CouchDB which is used to connect the Couchdb databases, create views and execute the reduce function. The Cloudant API takes the MapReduce function as a parameter and returns a Python dictionary as the result. The meaning of statistics are elaborated below.

- Date: This records the total number of tweets in five different cities (Melbourne, Sydney, Darwin, Brisbane, Adelaide) and the year that these tweets were posted. The year, extracted from the created_at field of tweets, is emitted. And the reduce function is the built-in _count function. Hence, the total number and corresponding year of tweets can be obtained in a front-end application.

- Hash: This is the statistic of word frequency (Hashtag) of tweets similarly in five cities. Since the hashtags field is an embedded list, we should check if the entities field and hashtags (in the entitles) are not null and the hashtags list is not empty. Then we emit the element of the hashtags list. The reduce function is built-in _count function. Now we fetch a grouped word frequency JSON dictionary, which can be directly handled and visualized in front-end applications.

- Language: The distribution of language in these five cities are interesting. Just emit the lang field of tweets and utilize built-in _count reduce function. Front-end applications can get the language frequency used in tweets.

5.4.2 Interaction between CouchDB and back end

The interaction between CouchDB and the back-end is quite simple. Actually, when a new tweet comes in, the result of MapReduce is only updated for the new tweet, instead of going through all tweets in the database. The backend application just fetches the processed result, which is very efficient.

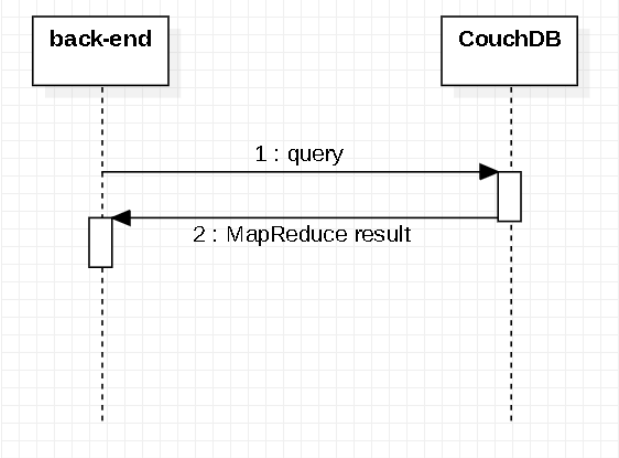

Figure 5: CouchDB Interaction with backend

6 Web Application Tools

6.1 RESTful API

6.1.1 Web application design

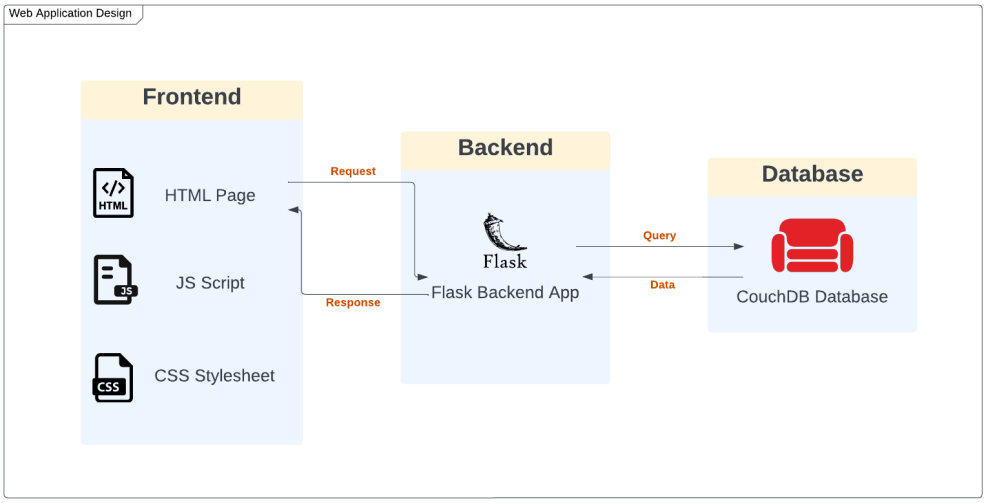

Figure 6: Web Application Design

In this section, we will talk about the system architecture of our Web Application in detail.

Firstly we built our backend, which is used for data analysis and processing. The processed data will be transferred to the frontend by RESTful API. Flask is employed on the backend of the web application, and it is a small and lightweight web application framework programmed in Python that is easy to use. Moreover, it provides a built-in development server and a quick debugger. Flask is not only ensuring the scalability of simple applications but is additionally highly flexible and easy to implement. For any extra functionality required, it can either import the corresponding third-party modules to satisfy the requirements, or directly call the functionality defined in other python files in the backend directory. However, flask is not suitable for large implementations due to the complexity of maintenance. Besides, flask can be quite slow when it handles a high number of requests as it needs to take turns to handle them and can only handle one at a time. As for this project, Flask is well enough to create a simple website to present the data visualization.

Then we built our frontend, which is used to display the results on the web application. A static HTML page is displayed on the frontend of the web application. The layout of the HTML page is formatted by CSS stylesheet. The charts and maps are presented through JS files that are intended to define functions to create charts. An HTML of the Melbounre's map is rendered and illustrated within the web application as well.

Simply running app.py in the backend can start the server. That is because app.py is the main code section, by defining routes and functions, each action corresponding to them can be executed.[2]

6.1.2 Communication between frontend and backend

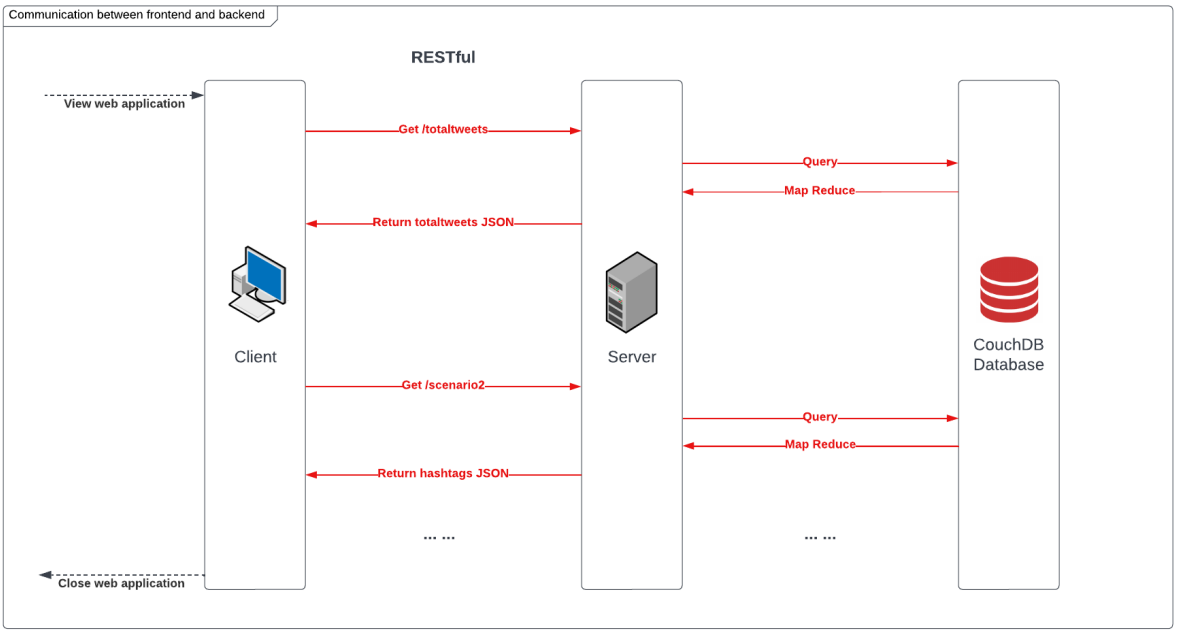

Figure 7: Commnication between frontend and backend

This application is a RESTful application that accesses and uses data with HTTP requests, in which each URL represents a resource. Axios, a "promise-based HTTP client",[3] is used to consume the desired data when the request is sent from the frontend.

For this project, when a user browses the website, the designed GET request will be sent from frontend to the backend in order to get the corresponding data. Then the HTTP response will be sent back to the user from the backend to the frontend. These will be displayed in the users' browsers, which again, are rendered by HTML file as different charts. The data used to implement different charts can be requested by various URLs using the Axios library.

For example, a GET request to the URL host:/senario3 will be sent if the frontend wants the top 10 hashtags for different cities. The backend will search for the relevant data in CouchDB and preprocess in the backend. The data finally will be returned to the frontend in JSON format.

6.2 Data visualization: Echarts and Folium map

Echarts, a declarative framework for web-based data visualization is employed to present charts from data in this project. As Echarts is an open-source JavaScript visualization library, it is simple to understand, straightforward and user-friendly. It offers not only a large library of a variety of intuitive, vivid, and interactive data visualization charts examples, but also many examples from the community with lots of creativity. Due to time constraints, it is suitable to use such an uncomplicated tool to visualize the processed data in a concise web application, which also helps save time and speed up the development of the whole project. Additionally, Folium is easy to visualize data manipulated in Python on an interactive leaflet map. It enables both the binding of data to a map for choropleth

visualizations as well as passing rich vector/raster/HTML visualizations as markers on the map.[4]

The main exploration throughout the project is whether Melbourne is a liveable city or not. First section is to investigate the factors which might potentially affect the liveability of each suburb in Melbourne. Four charts are utilised to display on the web application, a HTML of folium map of sentiment analysis for Melbourne, a line chart of monthly income levels and rent payments over suburbs, and its corresponding sentiment analysis by different kinds of line charts. The total number of tweets from the previous data is calculated and shown at the top of this section (Figure 9).

Through a primary analysis and visualisation of live tweets for five major Australian cities, Adelaide, Brisbane, Darwin, Melbourne and Sydney, a horizontal comparison is made to determine which city is the most liveable for the second section. There are a total of four charts to display via employing the data access from the GET request, which are geographical map of Australia with scatter points, pie charts of language distributions for the respective city, a live word cloud of top-10 hashtags for each city, as well as sentiment analysis displayed by various line charts. In addition, through the GET request, a total number of live tweets for all cities is accessed and displayed at the beginning of the live tweets section (Figure 8).

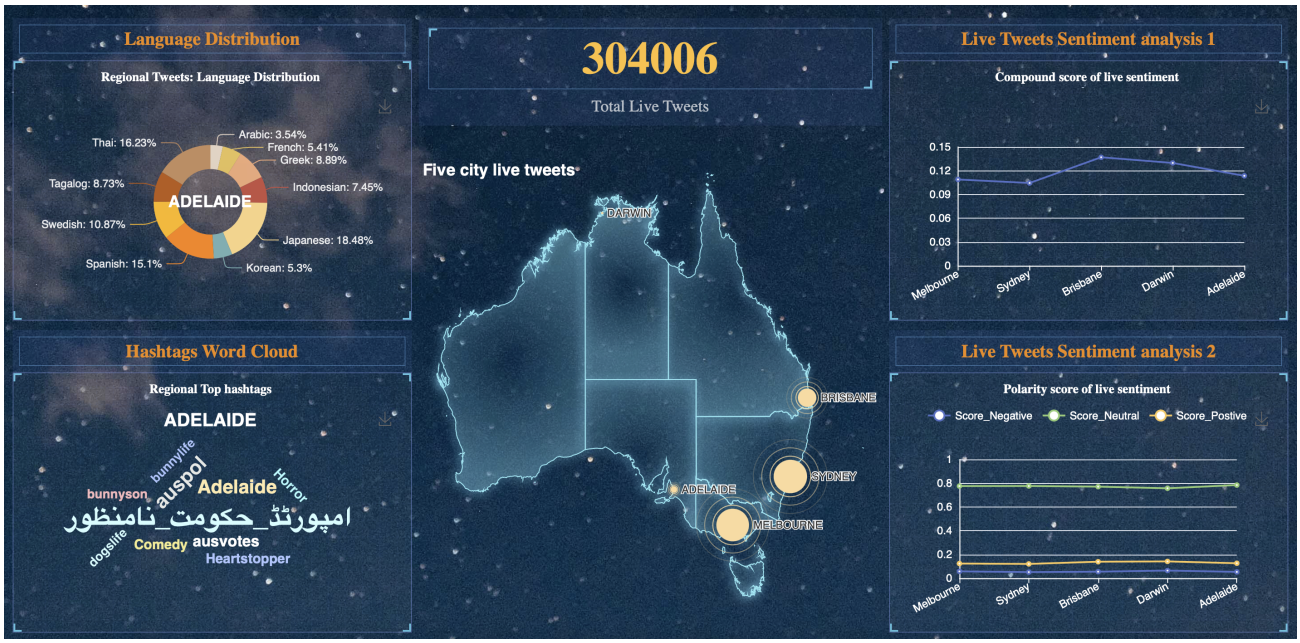

Figure 8: Website about 5 cities

6.2.1 Map chart with scatter

Map chart for geographic data visualisation is used to show the number of real-time tweets that are crawling from Twitter for 5 cities, which helps present the information directly and visually to the users, giving them a more intuitive visual experience. In this web application, the map of Australia which contains seven geographical regions, New South Wales, Victoria, Western Australia, South Australia, the Northern Territory, and Tasmania, is generated using Echarts through GeoJSON data that retrieved from the web source and is utilised as a background. Data related to these seven regions will not be shown as they were not investigated.

As mentioned before, the coordinates of the center of each city will be displayed on the map, and their sizes are related to the number of tweets collected. These scatters are great and very straightforward in the presentation showing the relationship between what team members believe are Australia's livable cities and the total number of tweets in the region, making it easier for users to distinguish trends quickly and clearly in the data, such as the cities with the highest and lowest number of tweets. As shown in Figure 8, scatter points are painted oranges as circles apparently, the larger the number of tweets, the larger the displayed points. Furthermore, the maps, i.e., scatter points, also played a role as interactive tools that allows users to switch and display data from other visual charts for 5 cities by clicking on the points.

Additionally, each city has its corresponding tooltip to show the number of live tweets. When the mouse hovers over the scatter point of each city, the user can see the current number of tweets for each city. The combination of statistics and images makes it easier for users to compare and understand.

6.2.2 WordCloud: Hashtags

WordCloud is used to display the top 10 hashtags by city directly. It is clear to see from Figure 20-24, hashtags that appear more frequently will appear larger in the word cloud, and vice versa. Similar to scatter points on the map, users can hover over each word to find out the frequency.

6.2.3 Pie chart: Language distributions

Pie charts are deployed to illustrate the distribution of languages by city, since they provide more visual and accurate representation of the percentage of each language. The colors contrasted with the background are also used in sections that highlight the percentage of various languages. Moreover, the size of the pie chart is displayed differently on different devices, which may result in an incomplete display of the data. Users can hover over each section to see the name of each language and the percentage of their use in the city.

6.2.4 Line charts

Two types of line charts are employed to show the sentiment analysis on the tweets, specially, real-time tweets from five cities and old tweets from selected Melbourne suburbs.

The basic line chart is utilized to represent the compound score of the sentiment analysis of the tweets. This clearly shows whether people's overall sentiment tends to be positive or negative, which in turn gives insight into the people's attitudes towards everyday life in the area they live in.

For further analysis, the stack line chart is used in conjunction with the basic line charts as portrayed in Figure 26. This also helps the users to easily find more detailed information such as the negative score, positive score, and neutral score in the sentiment analysis. This also provides a direct and comparative visualization on the distribution of score percentage for these three indicators.

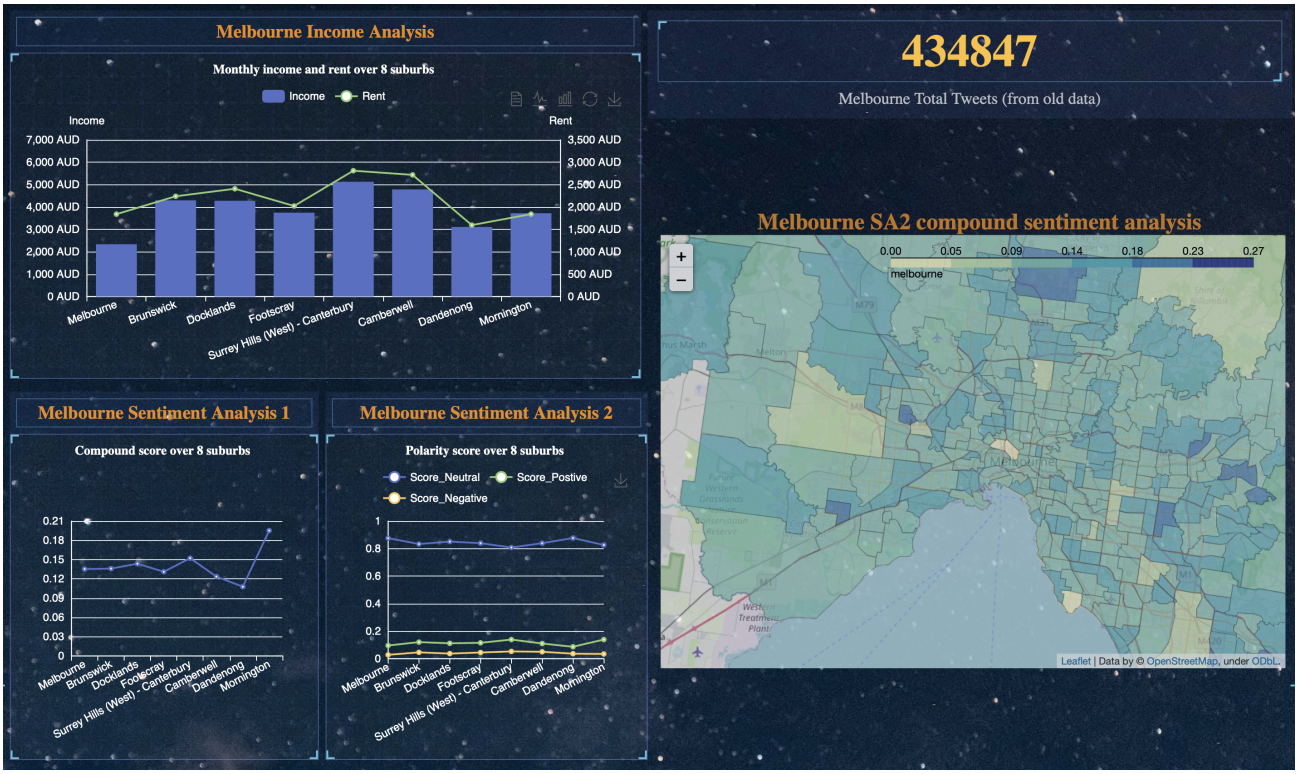

Figure 9: Website about Melbourne suburbs

Intuitively, there are 3 lines with different colors displayed on the chart. The blue line represents the negative score, the green line represents the natural score, the yellow line represents the positive score. Users can simply click on the legend to select a single line to observe, or decide which lines need to be stacked together for comparison. Moreover, users can hover over the data points to observe the scores displayed in the tooltips.

6.2.5 Mixed line and bar chart

The monthly income and monthly rent of people living in the 8 Melbourne suburbs are shown by a mixed line and bar chart. Monthly income is shown by the blue bar graph, while monthly rent is displayed by the green line graph. Users can also interact with these charts. By clicking on the legend, users can select an indicator to visualize the difference, or consider the impact on livability by combining the two factors. In addition, as users move their mouse over the bar or line chart, they can also find the values inside the hover tooltips, which helps them make comparisons.

6.2.6 Folium map

Folium is applied to visualize Twitter data, such as tweet count and average sentiment score in each SA2. An interactive Folimun map as an HTML file is rendered on the web application.

7 Deployment Guide

In this project, the version of Ansible used is 2.12.4.

1. To begin with, please clone the repository to local from:
   `https://github.com/KeYANG813/COMP90024Group33.git`
2. Please make sure you are connecting to the `Unimelb AnyConnect VPN` before running the ansible playbook, even though you are using the campus network. Because we need to use the melbourne student account for openstack authentication and get the proxy. Then we can have the right to deploy the project on MRC.
3. To deploy the whole system locally, please enter the folder Ansible using the command:
   `cd Ansible`
4. Then please copy our MRC private key to your local ssh file by the command:
   `copy test.pem ~/.ssh/test.pem`
5. After finising set up, now ready to run the Ansible playbooks.
   - To create instances on MRC and set up Docker for all instances:
     `./run-nectar.sh`

     After running run-nectar.sh, four new instances(instance1, instance2, instance3 and instance4) will be created on MRC with corresponding four new IP address. Then you have to change the IP address in hosts file. The instance1, instance2 and instance3 are the datanodes for CouchDB cluster. The instance4 are processing node for Web application. And the instance1 should be regarded as the master node for database cluster and instance2, instance 3 should be regarded as the workers of database cluster.

   - To continue, running the following command:
     `./install-docker.sh`
   - To set up the CouchDB container on Docker and CouchDB clusters:
     `./deploy-couchdb-cluster.sh`
   - To deploy the twitter crawler in instance 1, 2 and 3:
     `./run-crawler-cloud.sh`
   - To start backend and frontend, as well as deploy the web application in instance 4:
     `./run-app.sh`

8 Scenarios Analysis

Scenarios overviews
There are five fundamental aspects of great, livable cities: robust and complete neighborhoods, accessibility and sustainable mobility, a diverse and resilient local economy, vibrant public spaces, and affordability.[5] The liveability of a city is generally measured by indicators that provide quality of life, such as food, housing, transport, health care, opportunity and a safe and stable environment. Our purpose is to investigate the aspects of liveability by considering one or more of indicators.

1. The different languages distribution used in tweets in different cities.
2. The Top 10 common hashtags of tweets from different cities.
3. The Live sentiment Analysis for different cities.

4. The comparison between the people's emotional feelings and their income in the different suburbs of Melbourne.

Chosen of cities
We implement social media data analytics scenarios (scenario 1, 2 and 3) by comparing the twitter data among five cities in Australia: Melbourne, Sydney, Brisbane, Darwin and Adelaide.

8.1 Twitter language usage for different cities

This scenario aims to explore the commonly used twitter languages in five cities of Australia: Melbourne, Sydney, Adelaide, Brisbane and Darwin and compare them with official Population Census data from AURIN.

It is worth noting that Australia is a famous multicultural country. A city can gain many benefits from multiculturalism. Firstly, multiculturalism can bring a variety of different foods [6]. Secondly, If there are more people who speak different kinds of language, it might be easier and more friendly for international students or immigrants to join the society and promote understanding across cultures in cities. Therefore, language distribution apart from English of a certain city could be seen as an indicator of multiculturalism. If there are more languages used in social media in one city, this city could be more livable compared to others. The combinations of population in each city may be quite different.

8.1.1 Result visualization

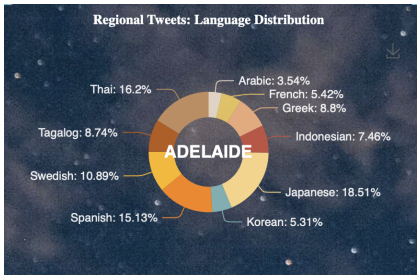

Figure 10: Adelaide Language Pie Chart

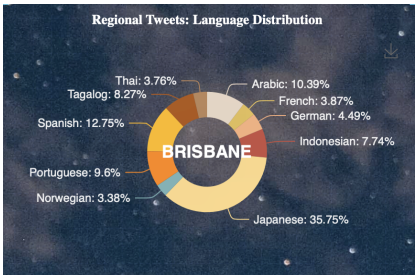

Figure 11: Brisbane Language Pie Chart

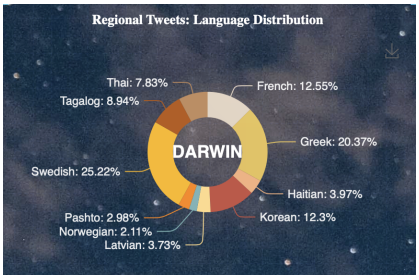

Figure 12: Darwin Language Pie Chart

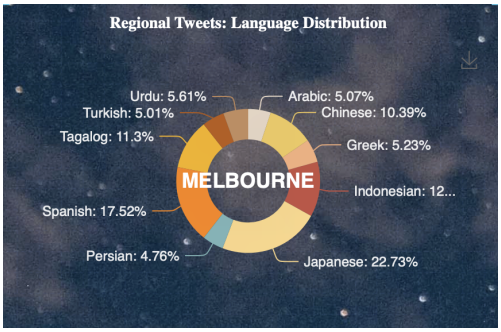

Figure 13: Melbourne Language Pie Chart

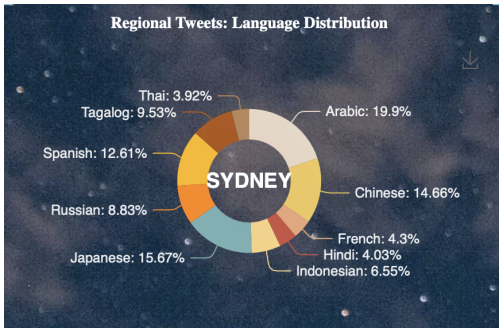

Figure 14: Sydney Language Pie Chart

We are interested in the top 10 commonly used languages apart from English because it will account for majority proportion of language usages for all five cities and make the analyzing results biased.

From the figures, the live tweet data shows that Japanese plays the dominant role in the distribution of frequently used language in Melbourne, Sydney, Brisbane and Adelaide. The Spanish illustrate the similar trend as Japanese in the cities except Darwin. People in all cities are willing to use Tagalog for tweeting, which almost ranks fifth among the top 10 common languages.

Chinese is only frequently used for tweets in Melbourne and Sydney. And the total number of tweets in these two cities are clearly more than the rest of the cities. In Adelaide, Thai has the greatest percentage in language distribution, in contrast to the least proportion of it in Sydney. Swedish, Greek and Korean are the top 3 languages in Darwin. Indonesia ranks fourth and fifth for Melbourne and Brisbane respectively. In Sydney, there are more people using Arabic for posting tweets than there are in Melbourne, Brisbane and Adelaide.

8.1.2 AURIN Census data

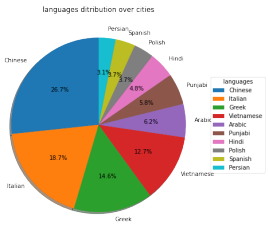

Figure 15: Greater Adelaide pie chart

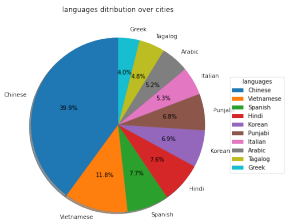

Figure 16: Greater Brisbane pie chart

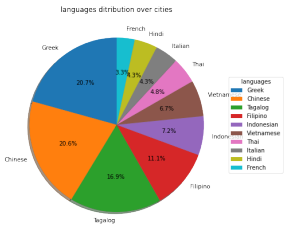

Figure 17: Greater Darwin pie chart

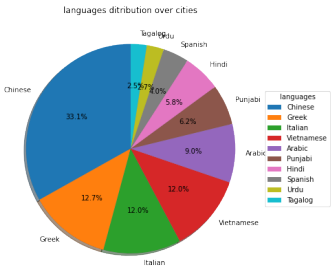

Figure 18: Greater Melbourne pie chart

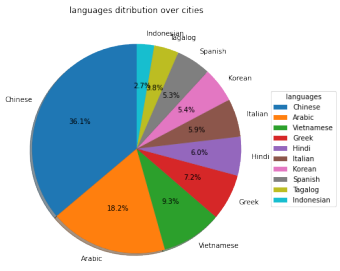

Figure 19: Greater Sydney pie chart

The data about GCCSA-P13 Language Spoken at Home by Sex-Census of population and housing in 2016 are extracted from AURIN platform and we make preprocessing, in order to get the top 10 languages that Australian prefer to use in family except English. We observe that the four cities(Melbourne, Sydney, Brisbane, Adelaide) show similar distributions for common language usage,

which are Chinese, Arabic, Greek, Vietnamese, Italian and Spanish. The most popular languages in Darwin additionally include Tagalog, Filipino, Indonesian. This might be caused by the geographical location of Darwin. It is a coastal city situated along the western shoreline of the Northern Territory where A large number of indigenous dialects remain spoken in this area. It obviously illustrates that Chinese takes up the main proportion across the five cities which can be regarded as the most popular Asian language, followed by Vietnamese. More people in Greater Brisbane prefer to use Spanish(7.7%) which is nearly twice that of Greater Melbourne(4.0%). Greek is the least frequently used language in Greater Brisbane but has the highest ranking in common language of other 4 cities. Also, there are a great number of people who would like to speak Italian in Great Melbourne and Great Adelaide.

8.1.3 Comparison between live twitter data and AURIN data

We can find that the most frequently used language on live tweets dataset is Japanese, whereas AURIN data shows that Chinese is the most commonly used language for these cities. Moreover, less people spoke Spanish in Sydney, Melbourne and Adelaide from AURIN data, however, more tweets are written in Spanish in these cities. The only similarity they demonstrate is that Arabic is the most popular language in Sydney, no matter for speaking within family or tweeting.

Generally, we conclude that the results of the AURIN data and Twitter data show the great differences.Since people in the family would communicate by the languages commonly used in certain regions(except English). When they post the contents on social media, they would use a more formal language or the language can be noticed and understood quickly by the target groups. In fact, English is still the most popular tweeting language that people in five cities prefer to use. Also, the differences may be caused by the different social media preferences for people. For example, Chinese and Vietnamese may not usually choose Twitter as the first choice for getting breaking news and entertainment to sports and politics, so the frequencies of tweeting, liking or retweeting from them are less as expected. The other possible reason is that Chinese and Vietnamese may use English for tweeting, in order to join in wider discussions.

8.2 Commonly used hashtags from tweets

This scenario explores top 10 hashtags from live tweets in five cities of Australia: Melbourne, Sydney, Adelaide, Brisbane and Darwin. What are the most commonly used hashtags to find out what people care about.

The motivation of this scenario is to find the different topics that people from five cities are interested in. The hashtags can reflect people's online and offline activities, their awareness toward politics and which big event that people care about much (like sport races and holiday celebrities). Furthermore, if there is any horrible thing happening such as illness (e.g. COVID-19), hill fires or floods exist in a city, it could be quickly shown on the hashtag, therefore we could understand the current life status. Therefore, the less of such negative words appear, the more livable or stable a city could be.

8.2.1 Result visualization

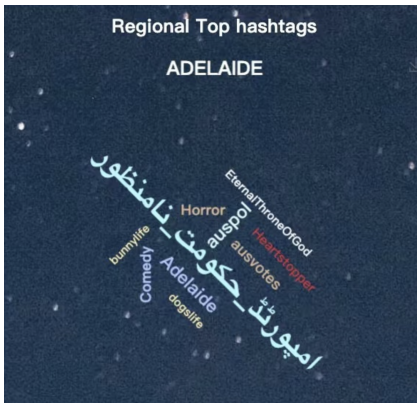

Figure 20: Greater Adelaide Hashtag

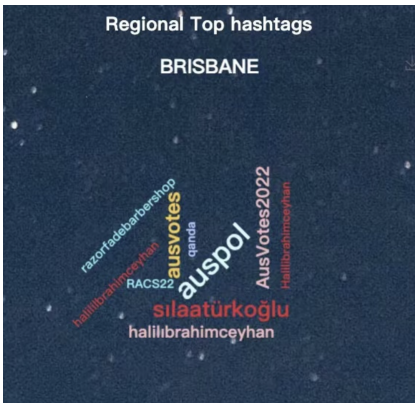

Figure 21: Greater Brisbane Hashtag

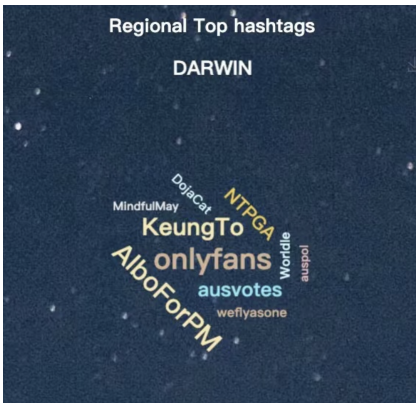

Figure 22: Greater Darwin Hashtag

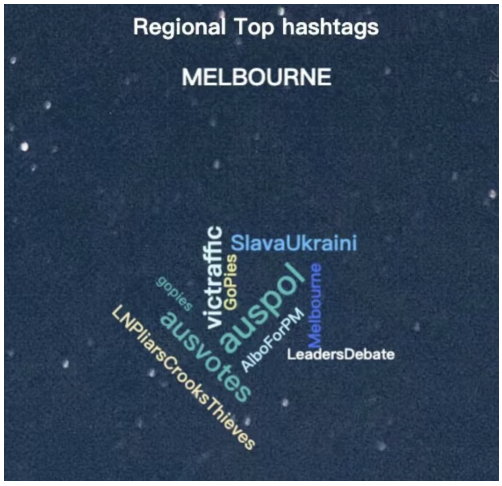

Figure 23: Greater Melbourne Hashtag

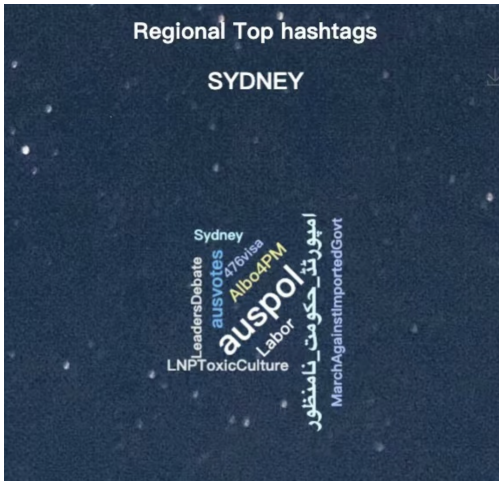

Figure 24: Greater Sydney Hashtag

We observe that people in Adelaide, Brisbane, Sydney and Melbourne often join the free and open discussion of Aussie politics. The greater emphasis is placed on updated political policies by them. There are also more election-related conversations on Twitter in Brisbane and Melbourne, which indicates that people tend to publish their own views about certain important issues. The traffic in Victoria is the second most popular topic that people talk about in Melbourne. Most of the people in Brisbane and Adelaide tweet with EternalThroneOfGod which shows that they would like to talk about their beliefs and religions on social media. In Adelaide, people are more interested in watching the horrible channels and comedy stories, while Brisbane people like the TV series. People in Adelaide also pay more attention to the LGBTQ+ group by hashtag "Heartstopper". The Australian Football League competition triggers more discussions in Melbourne and Adelaide. Darwin people talk much about the labor party and voting, they're also willing to express their emotions for entertainment news.

To sum up, since hashtags from Adelaide and Melbourne provided valuable information, we only compare these two cities in that case. We observe that people in Adelaide are more likely to choose

leisure activities in their spare time, which indicates a more free and relaxing life pattern for them. They also place the greater importance on the pursuit of spiritual satisfaction, this possibly due to the abundance in their daily life and stable living with steady incomes, public welfare and government subsidies. People in Adelaide feel free to show their belief and sex orientation also reveal the sufficient religious pluralism and inclusion of this city. With regard to the people in Melbourne,the Victoria traffic specially raises more attention as more people complain about the traffic crashes and delay on the Victoria roads which can be a big issue in people's daily life. Therefore, Adelaide is considered to be more comfortable and vibrant than Melbourne for people to live.

8.3 Live sentiment Analysis for different cities

8.3.1 Steps of sentiment analysis

Briefly, the tweets data obtained and stored inside the CouchDB is stored in ascending order of tweeting time. Therefore, we acquire the data of each database directly in reverse order by Mango Query, aiming to get the tweets of the last 3 days, and thus perform sentiment analysis on the tweets of these 3 days. The rate of data crawling is different for each city, hence we estimate the approximate number of tweets we can get in 3 days for each city. More detail of sentiment analysis method is in [Section 8.4.3].

8.3.2 Sentiment Analysis of tweets in different cities

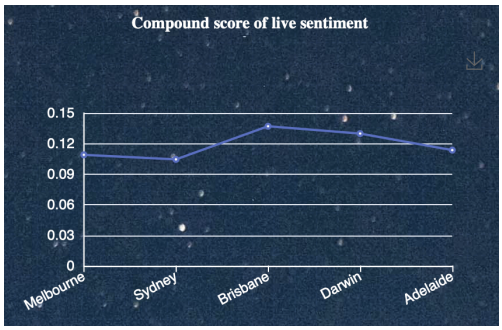

Figure 25: Sentiment Analysis compound

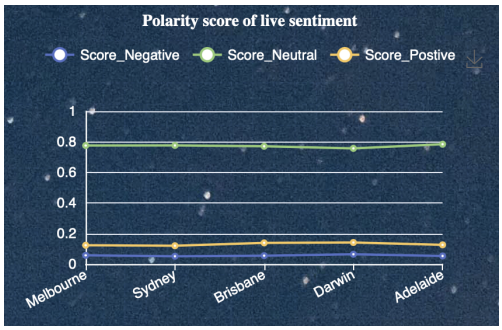

Figure 26: Sentiment Analysis polarity

We can observe that Melbourne obviously has the highest compound score than other four cities. The lowest compound score occurs in Darwin. Darwin also has the lowest positive score, neutral and negative scores. So people in Darwin may be more likely to pour out their emotions on Twitter. In general, there are no significant differences between four types of sentiment scores in the five cities.

8.4 Investigate livability within Melbourne

8.4.1 Dataset analysis

In order to investigate the most liveable areas of Melbourne, we tried to analyze the old Twitter data with 2,500,000 rows of data for various aspects. 434,847 tweets with accurate latitude and longitude are filtered and obtained. SA2 as the statistical level is chosen as we want to discuss the liveability for detailed and small regions. The polygon shapefile of SA2 is also applied to determine whether the coordinates of tweets are within areas of SA2 or not.[7]The preprocessed tweets are stored in a

dictionary.

At the same time we also used other datasets:

- Australian Statistics Geography Standard (ASGS): Volume 1 -Main Structure and Greater Capital City Statistics, Jily, 2016
- ABS -Personal Income -Total Income Distribution (SA2) 2016-2017
- ABS -Jobs in Australia -Employee Jobs and Income By Industry (SA2) 2016-17
- ABS -Data by Region -Family & Community (SA2) 2011-2018

8.4.2 Chosen of suburbs

We aim to investigate the most livable suburbs in Melbourne based on the Statistical Areas Level 2 (SA2), 2016 SA2 boundaries. We have chosen 8 representative areas to analyze, which are Melbourne, Brunswick, Docklands, Footscray, Surrey Hills (West) - Canterbury, Camberwell, Dandenong,Mornington.

We did some investigation on different suburbs in Melbourne. We basically choose suburbs based on distance to "Melbourne's Central Business District", and have the best or worst reputation among all suburbs.

Suburbs (SA2)	**Distance from central**	**Special reason**
Melbourne	Inner suburb (<10 km)	CBD and the most well-known suburb
Brunswick[8]	Inner suburb (<10 km)	"Brunswick is a laid-back multicultural suburb popular with a young, alternative crowd, and known as a destination for live music, upbeat pubs and beer gardens."
Dockland[9]	Inner suburb (<10 km)	"Dockland is well known as the home to Marvel Stadium and many modern buildings, public spaces and artworks."
Footscray[10]	Inner suburb (<10 km)	"The area is generally pretty safe despite what some people say about it, there are some colourful characters around but they are mostly harmless."
Surrey Hills[11]	suburb (10-20km)	"Surrey Hills is a fantastic suburb for those who can afford the price tag of this prestigious location."
Camberwell[12]	suburb (10-20km)	"Camberwell is a top-ranked suburb with its period houses, impressive schools and lively shopping strip."
Dandenong[13]	remote suburb (>20 km)	"Safe, friendly, leafy and right next to the station and parklands."
Mornington[14]	remote suburb (>20 km)	"Mornington is known for its "village" atmosphere and its beaches."

As a preview of this topic, we plot the folium map for count of Tweets(Figure 28), sentiment analysis(Figure 27), Median income(Figure 29) as well as the rent (Figure 30).

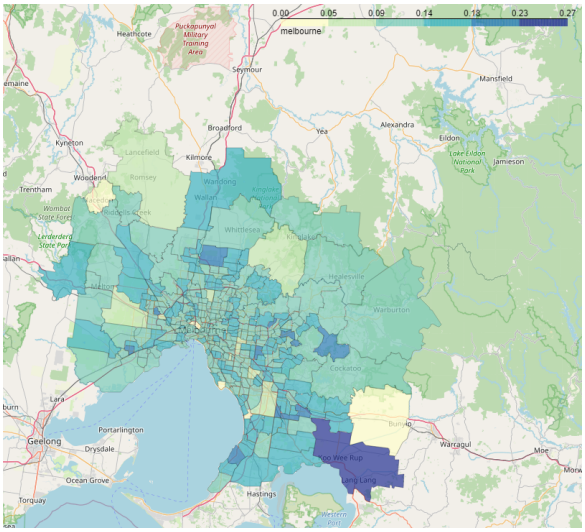

Figure 27: Sentiment score for SA2

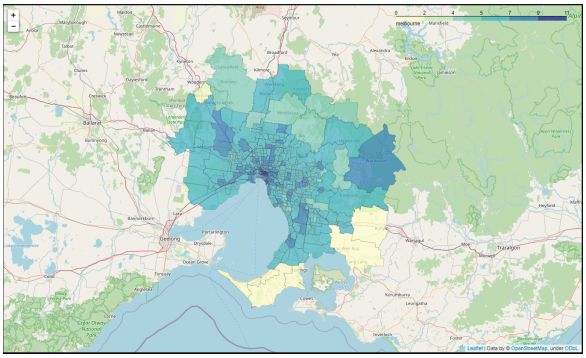

Figure 28: Old tweet count in SA2

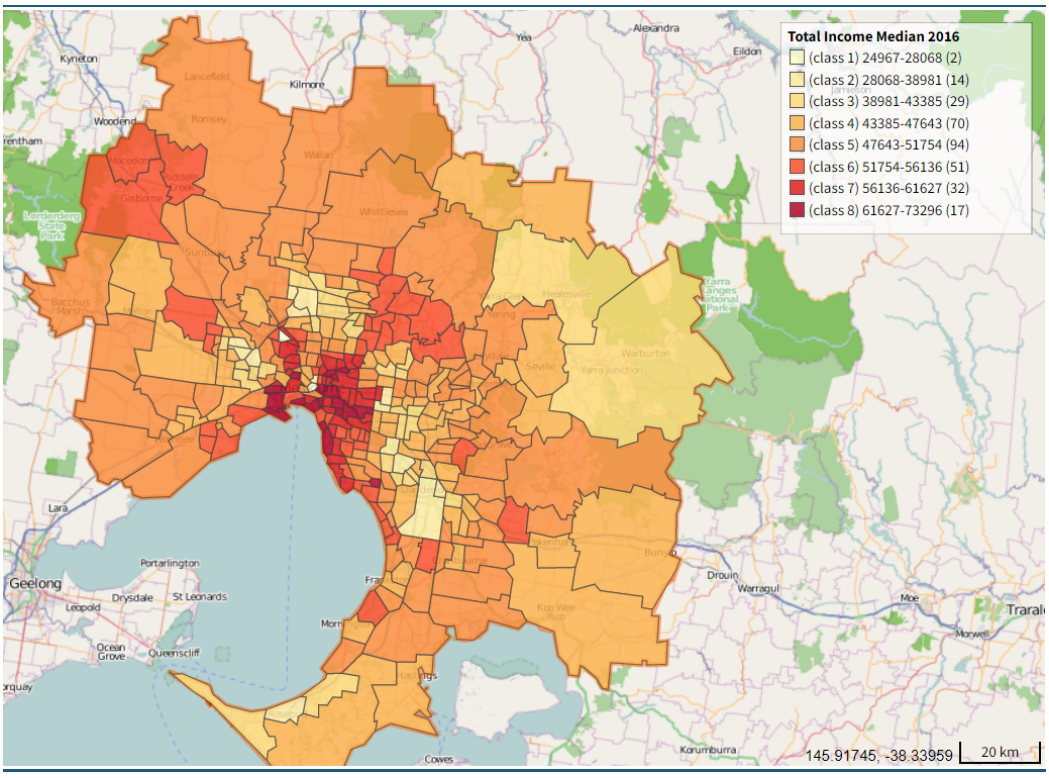

Figure 29: Median income in SA2

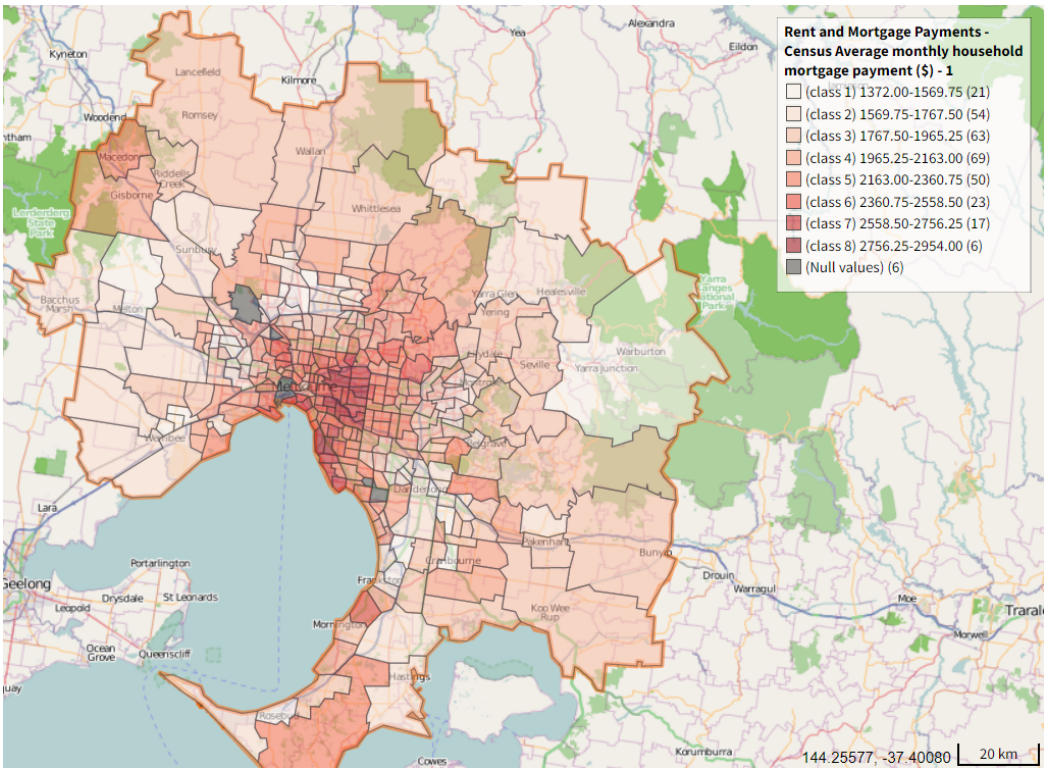

Figure 30: Rent and Mortgage Payment in SA2

8.4.3 Steps of sentiment analysis

We first use the shapefiles of each SA2 in Melbourne to get the polygon boundary data to determine whether the point of the old tweet is in the polygon, and assign tweets to each SA2. Then preprocess the text in the tweet to remove links and emojis.

We use the sentiment analysis in the NLTK package to obtain the sentiment scores of each tweet (the larger the score, the higher the proportion of the positive emotions), and calculate the average score of each area. However, there are very few tweets in some remote areas, causing inaccurate scores. Therefore, we ignored these data and changed them to 0.

8.4.4 Scenario 4 motivation

This scenario intends to investigate the most liveable suburb among 8 selected suburbs in Melbourne. We discuss the aspect of life pressure based on the statistical data of income levels and monthly rents from AURIN, as well as the sentiment analysis from the old tweets.

The motivation of investigating this scenario is to see whether people feel more stressed or experience negative feelings because of the low income level or high rents in a particular suburb. Sometimes, people would like to vent their emotions on Twitter if people encounter problems in their daily lives. Therefore, we apply sentiment analysis on tweets in each suburb, which helps us further analyze their stress in daily life. There are lots of possible reasons for high stress. For example, if a person has a low income but lives in a place where the rent is relatively high, it can be financially stressful for that person.

8.4.5 Analysis of income and rent data from AURIN

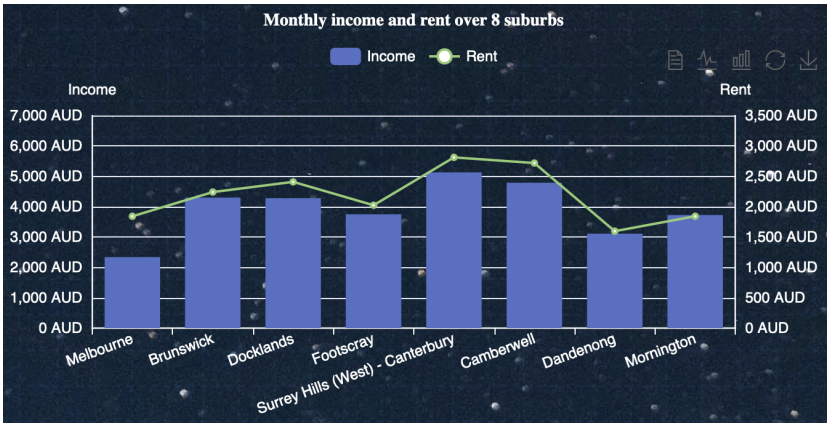

Figure 31: Monthly Income and Rent

According to Figure 31, it demonstrates that Camberwell and Surrey Hills (West) have relatively higher monthly income levels with also higher monthly rent payments. This may imply that although a high monthly income would make them feel fulfilled, the relatively expensive monthly rent of living in these two suburbs may make them stressed. Hence, it may not be cost-effective to live in these two suburbs. Brunswick, Dockland, Footscary and Mornington are middle monthly income suburbs with moderate monthly rent payments. This may indicate a relatively stable and steady standard of living in these suburbs. However, Docklands has a relatively higher monthly rent, people who live here might sometimes feel a bit stressed if they need to pay the rent on time, therefore Docklands might not be a stable suburb to live in. People in Melbourne and Dandenong have lower monthly income levels, however, people in Melbourne need to afford relatively higher rent payments each month. Therefore, people living in Melbourne might have lower disposable income since the cost of living here is relatively higher, which could highly impact their quality of life. Therefore, Dandenong might be a liveable suburb in Melbourne (city) rather than Melbourne (CBD).

In addition, we discuss other interesting comparisons related to liveability based on the monthly income level and rent payment. For those who expect to live in a wonderful coastal area, we chose Docklands, which is close to the CBD, and Mornington, which is far from the CBD, for comparison based on the monthly salary and monthly rent. It is apparent that the monthly income in both Docklands and Mornington is similar. However, people living in Mornington have a lower monthly rent than in Docklands. This may indicate that with similar monthly incomes, they may have more enough money to enjoy life, hence they may have a higher and better quality of life and will not feel much pressure rather than live in Docklands. As a result, Mornington may be a great suburb to live in for sightseeing with low rents.

8.4.6 Sentiment Analysis of tweets in Melbourne

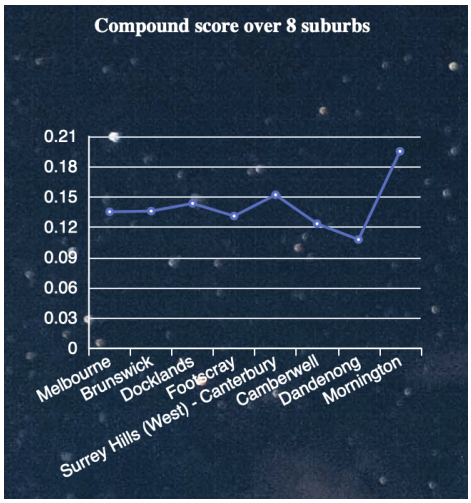

Figure 32: Sentiment Analysis compound

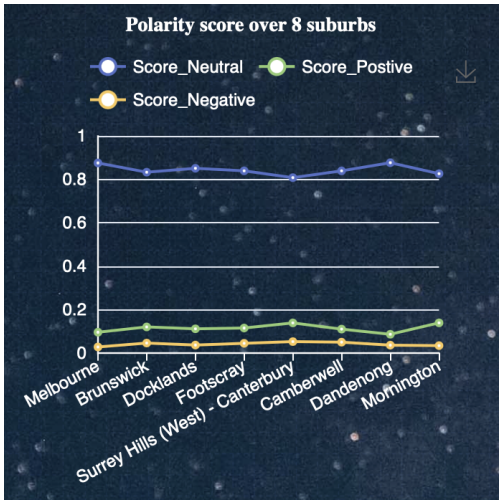

Figure 33: Sentiment Analysis polarity

From the Figure 33, we can observe that the Surrey Hills(West) have the highest positive score and negative scores than the other seven suburbs, which indicates that people are more likely to post a tweet with their specific emotions. People in Dandenong show the lowest positive response in their tweets, which might show that they have negative attitudes toward life as they may complain more about work, living price and housing. Melbourne has the lowest negative score so people may feel less unpleasant and become more satisfied with the current situation.

The neutral scores in all eight suburbs have no obvious difference so people may lack sentiment when they make some object suggestions or comment on important issues in tweets. Compared to the other suburbs, Mornington distinctly has the highest compound score and Dandenong has the lowest compound score.

8.4.7 Comparative Summary

For suburbs such as Camberwell and Surrey Hills (West) which have high income levels, tweets in Surrey Hills (West) obviously have more positive emotions, which might imply that people who live here are happier and more relaxed.

For middle-income suburbs, people living in Docklands, Footscray and Brunswick didn't show relatively more positive or negative emotions, which demonstrates that most of them have balanced lives. There are more tweets showing positive emotion in Mornington. People who live in Mornington would like to spread their happiness, which could have a positive effect on the sense of community.

For low-income suburbs, even though people living in Melbourne and Dandenong didn't show relatively more positive or negative emotions, since people living in these suburbs don't have a relatively high income level, especially people who live in Melbourne need to suffer higher rent payments each month. Therefore these suburbs are not very livable. In conclusion, Camberwell, Surrey Hills (West), Mornington are more livable compared to the rest of suburbs.

9 Conclusion

In this project, we have learnt a lot of useful stuff about cloud computing and had previous experience of using Melbourne Research Cloud to develop our own cloud system. We had heard of Docker and Ansible before, but had no experience using them, so it was also a big challenge for us to learn and use them in our automated deployment. We spent most of our time achieving automation of the overall system. We all agree that it is worth learning from errors since we gained a deep understanding of how to write Ansible Playbooks and how Ansible works with cloud systems. CouchDB is also a powerful database, and MapReduce allows for quick data analysis, such as counting or summing based on specific attributes without spending time on going through all tweets in the dataset (as we did in Assignment 1). Flask is a very friendly and easy-to-used backend development tool, which supports the RESTful API to send and get responses between backend and frontend.

We investigated two related topics about liveability. Firstly, we compared Melbourne with other 4 cities based on multiculturalism, the hot topics people are most concerned about, and the emotional state of live tweets. We concluded that Adelaide may be more liveable than Melbourne since people living in Adelaide are willing to talk about more relaxed topics so people living here may be more comfortable.

Secondly, we only focused on Melbourne and wanted to explore which suburb is the most liveable. We applied the sentiment analysis from tweets and compared the results with the income level and rent payment provided by AURIN to summarize that Camberwell, Surrey Hills (West), Mornington are likely to be better suburbs to live in compared to the rest of suburbs.

References

[1] "Home." Melbourne Research Cloud Documentation. Accessed May 2, 2022. https://docs.cloud.unimelb.edu.au/.

[2] "Creating the app.py File" cloudxlab. Accessed May 11, 2022. https://cloudxlab.com/assessment/displayslide/5947/creating-the-apppy-file

[3] "Using Axios to Consume APIs" Vue.js. Accessed May 11, 2022 https://v2.vuejs.org/v2/cookbook/using-axios-to-consume-apis.html?redirect=true

[4] "Python data, leaflet.js maps" Folium. Accessed May 11, 2022 https://python-visualization.github.io/folium/

[5] "Misssion & Goals-Livable city" Wikipedia. Accessed May 6, 2022 https://www.livablecity.org/missiongoals

[6] "26 Key Pros Cons Of Multiculturalism." Andreas. Accessed May 2, 2022. https://environmental-conscience.com/multiculturalism-pros-cons/.

[7] "Digital boundary files" abs gov. Accessed May 11, 2022 https://www.abs.gov.au/statistics/standards/australian-statistical-geography-standard-asgs-edition-3/jul2021-jun2026/access-and-downloads/digital-boundary-files

[8] "Must-Do Day Trips from Brunswick, Victoria." All In Self Storage. Accessed May 12, 2022. https://allinselfstorage.com.au/must-do-day-trips-brunswick-victoria-storage-brunswick/.

[9] "Docklands - between city life and waterfront view." Mortgage Choice. Accessed May 12, 2022. https://www.mortgagechoice.com.au/stephen.zamykal/our-services/mortgage-broker-docklands/.

[10] "Footscray." Homely. Accessed May 12, 2022. https://www.homely.com.au/footscray-maribyrnong-melbourne-greater-victoria.

[11] "Surrey Hills, Suburb & Moving Guide – TOP 5 Things To Know." Glenferrie Dental. Accessed May 12, 2022. https://glenferriedental.com.au/surrey-hills-suburb-moving-guide-top-5-things-to-know/.

[12] "Camberwell, VIC Suburb & Moving Guide – TOP 5 Things To Know." Glenferrie Dental. Accessed May 12, 2022. https://glenferriedental.com.au/camberwell-vic-suburb-moving-guide-top-5-things-to-know/.

[13] "Dandenong, VIC 3175." Homely. Accessed May 12, 2022. https://www.homely.com.au/dandenong-greater-dandenong-melbourne-greater-victoria.

[14] "Mornington, Victoria." Wikipedia. Accessed May 12, 2022. https://en.wikipedia.org/wiki/Mornington,_Victoria.