Using social media data for urban analysis in Singapore

Koh, Wee Boon

2017

Koh, W. B. (2017). Using social media data for urban analysis in Singapore. Master's thesis, Nanyang Technological University, Singapore.

<http://hdl.handle.net/10356/73050>

<https://doi.org/10.32657/10356/73050>

NANYANG TECHNOLOGICAL UNIVERSITY

Using Social Media Data for Urban Analysis in Singapore

Koh Wee Boon

A thesis submitted to
Nanyang Technological University
in fulfillment of the requirement for the degree of
Master of Engineering

School of Computer Science and Engineering

2017

Abstract

Effective land use planning is essential for countries where land is scarce. In Singapore, land use planning is managed by the Urban Redevelopment Authority (URA). A revised land use plan is released by the URA every four to five years with the goal to create self-sufficient neighbourhoods and reduce commute time between places. Identifying identical land zones can assist city planners to draft out similar development objectives, reducing the time required to produce the land use plan. For business owners, identifying similar land zones can allow them to find the next potential area of expansion that fits their first success case. This work provides an analysis and a framework to the problem of similar urban region query. In particular, the analysis and the proposed query framework investigate the use of urban and social media data to find regions that are topically consistent and exhibits similar demographics. To achieve this, an Urban Region Similarity Analysis System (URSAS) is developed. This work shows that social media data can help to improve the quality of the query result as opposed to only considering urban information. Additionally, this work shows that the proposed query framework can recognize the type of function for a given zone based on urban and social media data.

Acknowledgements

I would like to thank and acknowledge my supervisor, Associate Professor Cong Gao, for his valuable feedback and guidance during my course of Master study. Furthermore, his thoughtful criticism has made my research more meaningful and an enjoyable process.

Additionally, I would also like to thank the Final Year Project students for their contributions, especially in the data preparation stages, in this research.

Lastly, I would also like to thank my family and school for their constant support and encouragement.

Contents

Abstract	i
Acknowledgements	ii
List of Figures	vi
List of Tables	viii
Abbreviations	x
1 Introduction	1
1.1 Motivation	1
1.2 Background	3
1.2.1 Urban Land Use Planning	3
1.2.2 Location-based Social Networks (LBSN)	6
1.2.3 Region Similarity	7
1.3 Urban Land Boundaries Definition	9
1.4 Objective	10
1.5 Organization of This Thesis	11
2 Related Work	13
2.1 Urban Data and Analysis	13
2.1.1 Urban Data	13
2.1.2 Urban Analysis	15
2.1.3 Crowdsourcing and Resident Participation for Urban Analysis	17
2.2 Urban Planning	18
2.2.1 Goals and Challenges	18
2.2.2 Urban Land Use	19
2.3 Clustering	20
2.3.1 Techniques	21
2.3.2 Text Clustering	21
2.4 Clustering Multiple Features	23
2.4.1 Text Similarity	24
2.4.2 Spatial Distance	25
2.4.3 Probability Distribution Similarity	26
2.5 Summary	27

3	Problems and Techniques	28
3.1	Problems	28
3.1.1	Urban and Social Media Data Analysis	28
3.1.2	Similar Urban Region Query	29
3.1.3	Functional Zones Analysis and Recognition	29
3.2	Techniques	30
3.2.1	Observing Correlation Between Urban Data and Tweet	30
3.2.2	Finding Tweet Clusters	31
3.2.3	Proposed Region Similarity Approach	32
3.2.4	Validating Region Similarity as Perceived by Users	34
3.2.5	Functional Zones Analysis and Recognition	35
3.3	Aspects of a Region	36
3.3.1	Master Plan (MP) Boundaries	36
3.3.2	Population Distribution	37
3.3.3	Point of Interests (POI) Distribution	38
3.3.4	Tweet	39
3.4	Summary	39
4	Urban Region Similarity Analysis System (URSAS)	40
4.1	Implementation	40
4.1.1	Input and Storage	44
4.1.2	Preprocessor	45
4.1.3	Similarity Analyser	46
4.1.4	Clusterer	46
4.1.5	Visualizer	47
4.1.6	Defining Land Use	50
4.2	Preliminary Experiments	51
4.2.1	Clustering Tweet	51
4.2.2	Single Similarity Measures	52
4.2.3	Double Similarity Measures	53
4.3	Findings and Observations	54
4.4	Final Prototype Features	54
4.4.1	Region Analysis	55
4.4.2	Region Similarity	58
4.4.3	Multiple Input Sources	61
4.4.4	Functional Region	62
4.5	Summary	63
5	Urban and Social Media Data Analysis	64
5.1	Different Views of Data	64
5.1.1	Tweet	64
5.1.2	Population Data	65
5.1.3	POI Data	65
5.2	Urban Data Analysis	66
5.2.1	POIs Dependency on Residents	66
5.2.2	Residents Dependency on POIs	68
5.2.3	Volume Dependency of Tweet, POI and Resident	69

5.3	Social Media Data Analysis	70
5.3.1	Dependency of Tweet on Residents	70
5.3.2	Dependency of Tweet Term on Residents	71
5.3.3	Dependency of Tweet on POIs	72
5.4	Summary	75
6	Similar Regions Query	76
6.1	Objective	76
6.2	Experiment Setup	77
6.2.1	Dataset	77
6.2.2	Preliminaries	77
6.3	Single Aspect and Measures	78
6.3.1	Results	78
6.3.2	Findings and Observations	79
6.4	Double Aspects and Measures	80
6.4.1	Results	80
6.4.2	Findings and Observations	81
6.5	Triple Aspects and Measures	82
6.5.1	Results	83
6.5.2	Findings and Observations	84
6.6	Running Time	85
6.7	Summary	86
7	Functional Zones Analysis and Recognition	87
7.1	Objective	87
7.2	Experiment Setup	88
7.3	Functional Zone Analysis	88
7.3.1	Residential Zones	88
7.3.2	Business Zones	89
7.3.3	Commercial Zones	90
7.3.4	Findings and Observations	90
7.4	Functional Zone Recognition	91
7.4.1	Results	92
7.4.2	Findings and Observations	92
7.5	Summary	93
8	Conclusions and Future Work	94
8.1	Conclusions	94
8.2	Future Work	96
8.2.1	Semi-supervised Query Framework	96
8.2.2	Functional Zones Prediction	96
8.2.3	Automatically Generating the Boundaries	97
A	Preliminary Experiment Results	111
B	Queries with Single Aspect	116
C	Queries with Double Aspects	118

List of Figures

1.1	URA Master Plan 2014	4
1.2	URA Master Plan Boundaries 2014	5
1.3	Proposed Query Framework	11
3.1	URA Master Plan 2014: Central Region	37
4.1	URSAS Modules	41
4.2	URSAS Architecture	41
4.3	Web Visualization Tool: Overview	44
4.4	Storing Records as Documents in JSON format	45
4.5	Web Visualization Tool: Density Visualization	48
4.6	Web Visualization Tool: Tweet View Options	49
4.7	Web Visualization Tool: Residents and POI Distributions	50
4.8	Web Visualization Tool: Defining Land Use	51
4.9	Region Analysis: Tweet Analysis Overview	56
4.10	Region Analysis: POI Analysis Overview	57
4.11	Region Analysis: Population Analysis Overview	58
4.12	Region Similarity: User Evaluation	59
4.13	Region Similarity: Evaluation Result	60
4.14	Functional Region Analysis	63
5.1	Correlation of Healthcare Facilities and Age Group Residents in Regions	67
5.2	Correlation of POIs and Age Group Residents in Subzones	68
5.3	Correlation of Hourly Distribution of Tweet and Age Group Resident in Subzones	71
5.4	Correlation of the Hourly Distribution of the Term "orchard" and Age Group Residents in Subzones	71
5.5	Correlation of the Hourly Distribution of the Term "shopping" and Age Group Residents in Subzones	72
5.6	Correlation of the Hourly Distribution of Tweet and POI in Subzones	73
5.7	Correlation of the Hourly Distribution of Term "orchard" and POIs in Subzones	74
5.8	Correlation of the Hourly Distribution of Term "shopping" and POIs in Subzones	74
7.1	Correlation of the Hourly Distribution of Tweet and POI in Residential Functional Zone	89
7.2	Correlation of the Hourly Distribution of Tweet and POI in Business Functional Zone	90

7.3	Correlation of the Hourly Distribution of Tweet and POI in Commercial Functional Zone	91
A.1	Tweet Clustering Result	111
A.2	Single Similarity Measure: Tweet	112
A.3	Single Similarity Measure: Female Residents	113
A.4	Single Similarity Measure: Female Residents Distributions	113
A.5	Single Similarity Measure: POI	114
A.6	Single Similarity Measure: POI Distributions	114
A.7	Double Similarity Measure: Tweet	115

List of Tables

1.1	Sample URA Feedbacks	4
1.2	Sample Tweets	7
3.1	An Example Female Population Distribution on Each Subzone by Age . .	38
3.2	An Example of POIs	38
4.1	DBScan Tweet Experiment Setup	52
4.2	Single Similarity Measure: Tweet Experiment Setup	52
4.3	Single Similarity Measure: Female Residents Experiment Setup	53
4.4	Single Similarity Measure: POI Experiment Setup	53
4.5	Double Similarity Measure: Tweet Experiment Setup	54
5.1	Highest Correlation Value	68
5.2	Proportionality of Tweet, POI and Resident in Subzones	69
6.1	Planning Area: Query Aspects Performance	78
6.2	Subzone: Query Aspects Performance	79
6.3	Planning Area: Measures Performance	79
6.4	Subzone: Measures Performance	79
6.5	Planning Area: Aspects Performance	80
6.6	Subzone: Aspects Performance	80
6.7	Planning Area: Tweet Representation Performance	81
6.8	Subzone: Tweet Representation Performance	81
6.9	Planning Area: Measure Performance	81
6.10	Subzone: Measure Performance	82
6.11	Planning Area: Overall	83
6.12	Subzone: Overall	83
6.13	Planning Area: POI, Population and Tweet with $w_i = 0.1$	84
6.14	Subzone: POI, Population and Tweet with $w_i = 0.1$	84
6.15	Average Running Time of Different Similarity Measures	85
7.1	Accuracy of URSAS Considering Only a Single Aspect	92
7.2	Accuracy of URSAS with POI and Tweet	92
B.1	Planning Area: Query Performance	116
B.2	Subzone: Query Performance	117
C.1	Planning Area: Combining Tweet with POI, Tweet with Population and POI with Population	118

C.2 Subzone: Combining Tweet with POI, Tweet with Population and Population with POI	119
--	-----

Abbreviations

URSAS	Urban Region Similarity Analysis System
URA	Urban Redevelopment Authority
MP	Master Plan
LBSN	Location-Based Social Networks
DGP	Development Guide Plan
NLP	Natural Language Processing
LDA	Latent Dirichlet Allocation
POI	Point Of Interest
DTM	Document Term Matrix
TDM	Term Document Matrix
TF	Term Frequency
IDF	Inverse Document Frequency
TF-IDF	Term Frequency Inverse Document Frequency
JSC	Jaccard Similarity Coefficient
SMC	Simple Matching Coefficient
KL	Kullback-Leibler
JS	Jensen-Shannon
KML	Keyhole Markup Languages
JSON	JavaScript Object Notation
GeoJSON	Geographical JavaScript Object Notation
nDCG	Normalized Discounted Cumulative Gain
CSV	Comma Separated Values
URL	Uniform Resource Locator

Chapter 1

Introduction

1.1 Motivation

Land use planning is particularly important when space is limited. The Urban Redevelopment Authority (URA)[1–4] is a land use planning authority in Singapore. Every four to five years, URA review and revise their detailed land use plan, which is known as the Master Plan (MP).

In recent years, there have been an increasing number of studies on urban land use. Taking into account limited budget, it is unfeasible to focus all the development efforts into a single neighbourhood. On the other hand, an evenly distributed development effort may have a negative impact on neighbourhoods that required more attention. Studies on urban land use, like those in [5–7], have pointed out that over developing a region can lead to expensive housing. Another challenge is the development of infrastructure for residents to have access to public transport easily. Identifying similar urban regions can assist city planners in tackling these challenges. For example, if there are a similar number of residents in given two regions, the amount and placement of public transport facilities should also be identical. At the same time, housing value should remain affordable in those residential areas that are considerably further away from the central commercial and business districts.

A recent study[8] conducted in Nantong, China, presented the problems of the fragmented use of residential areas due to urbanisation and limited land. Naturally, cities with limited land tend to construct residential zones with a mixture of different functions.

This phenomenon can be observed when a residential zone is used for multiple functions like for commercial, business and education purposes. Due to the mixture of functional zones, it is difficult for city planners to find similar regions. Rather than focusing on the type of zones defined in the MP, it is more desirable to consider the Point of Interests (POI) to determine the function and find similar regions.

Studies like [9] and [10] have focused their efforts on finding regions with similar urban functions. In particular, these studies have used human mobility and POIs to determine the function of a region. In other words, regions with similar types of human mobility patterns and POIs are likely to have the same function. Although these studies have shown viable ways to detect functional zones, they do not consider the distribution of the residents in those areas.

Other studies like [11, 12] have used population census to compare urban growth and detect similar land use. It can be agreed that the types of land use are partially driven by the residents in the area. However, these studies do not consider the socioeconomic activities of the visitors in that region. For example, a visitor can travel to a region to shop. It is clear that visitors travel to a region to utilize the function in that area. The topics in social media posts like Tweets can be used to discover the socioeconomic activities of the visitors in a region.

The studies in [13, 14] have shown the possibilities of using Tweets to identify different land use functions in a region. Using the temporal, spatial and the text information in Tweets, these studies have mapped out the different functions of a region at a different time of a day. However, the boundaries of the land use identified by their proposed approaches are dependent on the Tweets. Such cases are undesirable as the true boundaries of the functions are restricted by the availability of Tweets at the edges of these regions. It is better to use clearly established boundaries, such as those defined in the MPs, to find similar regions.

Even though urban analysis is a relatively well-studied field, the number of systems to assists city planners to find similar regions is still small. The work in [15] has proposed a system that is able to visualize and analyse the different types of urban information to assists city planners in the development of smart cities. Although the system is able to make use of a wide range of urban and social media data, it is not beneficial to other types of users like business owners. From a commercial perspective, it is essential for

business owners to know the demographics and function of a region. For example, a restaurant owner might want to open his shops in regions with demographics that best fit his target market. An urban analysis system should allow business owners to query for similar regions and use it as an indicator for the potential areas of expansion. However, it can be agreed that such a system cannot have a rigid way of retrieving similar urban regions. Like a web search engine, there are many factors that affect the results that are perceived as relevant by a user. For example, two regions may be similar in terms of the population demographics but dissimilar in terms of the facilities. Hence, it is also necessary for such a system to allow users to customize their queries, in terms of the aspects to consider, when looking for similar regions.

1.2 Background

This work analyses the effectiveness of urban and social media data to query for similar regions. To do so, it is important to identify the challenges faced and the strategic objectives in the land use planning process. This section will give a brief introduction to urban land use planning, location-based social networks (LBSN) and region similarity.

1.2.1 Urban Land Use Planning

Ideally, as land is limited, redevelopment of existing estates can mitigate the impact of congestion due to the fast population growth by migration. As such, for obvious reasons, the MP does not necessarily only contain new development plans. Instead, depending on the rapid changing factors, such as median population age, it may also contain redevelopment plans for existing areas to meet the needs of the residents. Consequently, this results to the rapid redevelopment of existing estates in Singapore. Ultimately, these plans are meant to create self-sufficient neighbourhoods for residents to work, relax and shop in close proximity to their homes. When the commuting needs of the general population are reduced, it becomes less congested for those with real commute needs in downtown areas.

Public consultations to review the MP are conducted through exhibitions, focus groups discussions and public forums. Alternatively, the public can feedback to URA through online channels like [16]. Between 27th April 2012 and 26th May 2012, URA collected

339,811 feedbacks. A small sample of the feedbacks is shown in Table 1.1. Evidently, it can be difficult and costly to conduct these public consultations often. Furthermore, the feedbacks gathered may not be sufficient to justify any significant changes. Although the number of feedbacks is large, it is difficult for the authorities to filter them. Additionally, most of these comments are related to the environment and have little significance for land use. Manually labelling and consolidating these feedback were time-consuming and may cause the initial intended meaning to be lost.

Type	Case Summary	Postal Code
Cockroaches	A lot of cockroaches at the void deck, near the lift lobby.	460529
Hygiene	Bird faeces on the chairs and tables	228210
Littering	High rise littering at beach road.	190001

TABLE 1.1: Sample URA Feedbacks

Careful and strategic land use planning is crucial in improving the livelihood of the citizens while allowing population growth. The MP 2014 is accessible to the public at [1] and is shown in Figure 1.1. Different land use, such as residential, business and commercial estates, are depicted by different colours in the MP. Additionally, the plan divides Singapore into regions, namely, North, North-east, East, West and Central. Each region is further divided into smaller planning areas such as Bishan, Ang Mo Kio and Toa Payoh. Each planning area is further broken down into subzones. These are shown in Figure 1.2.

FIGURE 1.1: URA Master Plan 2014

Planning area plays an important role for the URA during zoning of specific land use. Each planning area has a development guide plan (DGP) which contains the objective and development intensity. All these information is released through the URA annual reports[17]. This allows the public to know what kind of development and where it will take place in the following years. Real estate developers will have a clearer understanding of where they should develop next. Home owners will be able to know what to expect in their neighbourhood in the coming years. Business owners can also make use of the opportunity to gain the first mover advantage. Additionally, the planning areas are used by different authorities for administration purposes[2–4]. For example, planning area is used by the police department for jurisdiction boundaries purposes. Crimes that occur in a planning area are handled by the police centre in that vicinity.

FIGURE 1.2: URA Master Plan Boundaries 2014

The MP is made up of the many DGPs drawn out for each planning area. It is a long and tedious process as each DGP takes approximately two years to complete. In general, this process consists of six basic steps. Firstly, data is collected from each planning area to understand the needs and land availability for development. Secondly, these data are analysed to find out the strength and weakness of each planning area. This will help formalize the strategic objectives for each DGP. Thirdly, a rough outline of the plan is drawn. In this step, public consultation sessions will be held to explain and gather

feedbacks from developers and members of the public. This is to ensure that the plan addresses their needs and requirements appropriately. Next, the feedbacks from public consultations are incorporated into the outline plan. This forms the draft of the MP. Following this, the fifth step is to organize an exhibition, lasting about two weeks, for the public to lodge any additional objections to the plan. Lastly, after evaluating and attending any objections from the public, the final MP is drawn. This MP, together with the respective DGPs, will replace existing ones.

Evidently, it is unlikely to have a comprehensive process that addresses the needs and concerns of every member of the public. A better and less manual approach may be to gather and aggregate social media data. These data can be used as part of the public consultation to evaluate the outline plan. In addition, these social media data often contains location information that can be used to determine where it was uploaded. For this reason, many social media platforms, like Twitter, are often termed as location-based social networks. It is hoped that such a data gathering approach can reduce the tediousness and time required to draw out the respective DGPs.

1.2.2 Location-based Social Networks (LBSN)

A large amount of textual and graphical contents are generated by individuals on social media platforms on a daily basis. Two notable examples are Twitter[18] and Instagram[19]. On Twitter itself, it is estimated that there are about 500 million Tweets posted daily[20]. There are many ways in which textual data can be interpreted to give useful information. For example, Natural Language Processing (NLP) [21, 22] techniques or statistical approach like Latent Dirichlet Allocation (LDA)[23–25] can be used to find the semantics and topics for a piece of a text document. However, users on these social media platforms often use multi-lingual, unstructured and short texts. For this reason, NLP techniques that are designed to handle structured content, like news articles, often fail to capture useful information. One possible way to mitigate this problem is to retrieve monolingual text only and performing the necessary filtering steps. Unsurprisingly, this limitation has restricted the text mining applications on such platforms to getting trending topics and popular word mentions. In addition, this type of applications has limited use in terms of land use planning. This is largely due to the fact that most text contents on these social media platforms are meaningless by itself without proper filtering process.

Social media platforms, like Twitter, are termed as LBSN[26, 27] due to the fact that users can tag a location information to their posts. With the availability of GPS-enabled mobile devices, most content generated are tagged with geographical coordinates as shown in Table 1.2. The addition of this geographical information gives more meaning to the content. For example, consider the text, "The train has broken down". From the text, it is obvious that there is a disruption of transportation somewhere. With the geographical coordinates, it can be used to determine where the disruption has occurred. Using this information, commuters can choose alternative transportation to reach their destinations while the relevant authorities can take necessary measures to respond to the disruptions.

Latitude	Longitude	Text
1.301	103.859	I'm at Arab Street in Singapore
1.297	103.801	Just posted a photo @ Management Development Institute of Singapore University Campus
1.311	103.752	ShiftingGoodbye Pandan Loop will miss this place?? @ 200 Pandan loop Pantech 21

TABLE 1.2: Sample Tweets

In recent years, there has been much attention given to mining social media data such as using Twitter and Yelp[28] to track public health issues[29]. Another example is tracking the viewership of YouTube[30] using Twitter streams[31]. Incorporating additional features like spatial and temporal information has sparked new applications like event detection[32–34] and topics tracking[35]. In another work, temporal information is incorporated for real-time applications like city traffic monitoring[36].

1.2.3 Region Similarity

When planning for a region, it is important to identify the attributes of that area[2–4, 17]. Intuitively, there should be more housing facilities and wider roads for regions with higher population density. Otherwise, there will be congestion and in turn, longer commute time. However, it is counter productive if all regions were developed in the same manner as such an approach requires significant financial backings. Excessive development reduces the available funding that could be used for areas with far more urgent needs. To a city planner, it is important to balance the development budget as well as meeting the needs of each region.

A region can be considered similar to another if both display identical population distribution. For example, there should be sufficient neighbourhood academic institutions for regions with a high number of residents who are still at schooling age. Naturally, such regions should receive more attention. By comparing population distribution, city planners can better outline the requirement for each region. For example, suppose that an under developed region has similar population distribution with another developed area. City planners can factor in the lacking facilities, that is otherwise available in a developed region, as part of the requirements for the planning objectives of the under developed area.

Other than the population distribution, two regions can also be considered similar if the residents talk about related topics. For example, residents could feedback about traffic congestion problems. In this case, these regions require infrastructure development or redirection of traffic. The first option for city planners is to simply plan for a widening of roads. Alternatively, city planners can reallocate some of the facilities to a less populated neighbouring region to divert some of the traffic. However, this contradicts the goal of having self-sufficient neighbourhoods. The likely approach will be to either redevelop some of the estates to make way for wider roads or construction of public transport facilities. However, a significant number of feedbacks is required to justify such plans. This may be difficult to achieve if only part of the residents participates in those sessions. The data from the URA feedback channels, as seen in Table 1.1, also contributes to the final MP. City planners can make use of these feedbacks to revise their outline plan for similar regions. These feedbacks display similar text content as those found in Tweets, as seen in Table 1.2. Both contain a short text and location information. In this case, Tweets can be used as a replacement or a complementary data to the feedbacks obtained from URA online channels.

As mentioned earlier that it is essential not to over develop a region and neglect other areas that have a greater need for development. This can be achieved if city planners can identify regions that contain similar facilities. In general, these facilities are individually known as a POI. Two regions can be considered similar if they have similar POI distribution. POI distribution of a region refers to the number of different types of facilities, such as schools and shopping malls, in the area. For example, regions having the similar POI distribution can either means it is over or under developed. Development of leisure facilities in a developed region may be redundant compared to one that has plenty of

housing but little recreational options. In this case, city planners can make use of such information to outline similar objective plans for the DGPs of these regions.

1.3 Urban Land Boundaries Definition

As there are many different terms used by URA in Singapore to split the land use, it is important to list down the differences and the meaning of these terms used. This work will follow the terms used by URA to explain certain concepts. In particular, there are four types of boundaries that are used by URA as follows.

- **Region Boundary** - this boundary type split Singapore into 5 segments namely, North, North-east, Central, East and West regions. It is often used to refer to a generic area of Singapore. The region boundary example is shown in Figure 1.2(a). Each region contains multiple planning areas.
- **Planning Area Boundary** - this boundary type split Singapore into 55 smaller administrative area. It is often used for generic strategic planning purposes. The planning area boundary example is shown in Figure 1.2(b). Each planning area can contain multiple subzones.
- **Subzones Boundary** - this boundary type split Singapore into 323 smaller administrative area. It is often used for more specialized planning purposes. An example is planning a subzone as a residential district. However, it is less specialized than functional zones since a residential district can contain educational facilities. The subzone boundary example is shown in Figure 1.2(c). Each subzone contains multiple functional zones.
- **Functional Zones Boundary** - this boundary type split each subzone further into smaller and more specific function for each piece of land. It is often used to determine the size and function of a specific piece of land. The functional zones boundary example is shown in Figure 1.1, where each function is defined by different colours.

For subsequent chapters, the word "boundary" and "boundaries" will be used when referring to the different types of urban land boundaries data.

1.4 Objective

The main objective of this research is to investigate the effectiveness of using urban and social media data to query for similar land regions. This work developed an Urban Region Similarity Analysis System (URSAS) to handle the similar urban region query problem. The URSAS is based on the similar urban region query framework as shown in Figure 1.3. However, the objective of this work is not to come up with a new way of analyzing social media and urban data. Instead, this work aims to find out the effectiveness of the various existing measures to find similar regions. Effectiveness in this work refers to whether existing techniques can query for similar regions as perceived by users and to what extent. For example, this work aims to find out whether existing techniques are sufficient for similar region queries or there is a need for further studies in this area. In addition, this work aims to find out whether social media data can be used together with urban statistics to retrieve relevant similar regions. Lastly, this work aims to find out whether social media and urban data can identify the functions of each piece of land. Functional recognition of urban regions is a field that should receive more attention as it is applicable to both business owners and city planners. In addition to URSAS, the main contributions and analysis of this work in the field of urban data mining are summarized as follows.

- *Urban and Social Media Data Analysis* - To seek a better understanding of the relation between population dynamics and the facilities in the area. It is found that the population dynamics is one of the key factors for the type of facilities developed in that region. A sub-analysis is also performed to explore the clusters of Tweets posted in Singapore. It is found that the clusters of Tweets posted are not useful in identifying similar regions without considering other aspects like population and POIs.
- *Similar Urban Region Query* - To compare and analyse urban region similarity based on urban data, social media streams and user perceptions. The proposed query framework uses a weighted aspect approach to find similar urban regions. To measure the relevance of the queries, a number of users from different disciplines are asked to evaluate the relevance of the results. It is found that combining aspects that have a positive correlation with one another tends to give more relevant results.

In addition, this work compares the different types of similarity measures. It is found that certain similarity measures are more suitable for one type of aspect. For this reason, a different measure should be used for each aspect to find similar regions.

- *Functional Zones Analysis and Recognition* - To analyse the relationship between the Tweets posted and POIs in a functional zone. In addition, this work will investigate how well the query framework can recognize the function of a functional zone. To test the query framework, each functional zone is compared with every other. The function of the zone is obtained from the most similar neighbour. It is found that the query framework can recognize the function. Additionally, it is found that using more than one aspects in the queries will improve the recognition accuracy.

FIGURE 1.3: Proposed Query Framework

1.5 Organization of This Thesis

This thesis is organized as follows. Firstly, the related works are discussed in Chapter 2. Next, Chapter 3 will list the various problems in this work and techniques used to solve them. The implementation and features of URSAS are described in the following Chapter 4. A preliminary experiment performed using the URSAS is discussed in the same chapter. Next, Chapter 5 presents the analysis of urban and social media data. Using

the analysis, Chapter 6 shows the results of the experiment that uses the proposed query framework to find similar urban regions. Chapter 7 will present the analysis performed on functional zones to find the correlation between Tweets and POIs. Additionally, the same chapter will present how well the proposed query framework can recognize the function of each functional zone. Lastly, Chapter 8 will summarize the findings and recommendations for future works for this study.

Chapter 2

Related Work

This chapter describes the related works of this study. Firstly, the challenges, limitations and existing works on urban data analysis are described. This is followed by a discussion on the objectives and importance of city planning.

2.1 Urban Data and Analysis

Generally speaking, urban data refers to data that originates from a city[37–39]. Some common examples are population demographics[40], vehicle trips[41] and visitor data[42]. Depending on the availability and accessibility of the data, some works have proposed crowdsourcing[43–45] as a substitute or replacement for urban data. Crowdsourcing platforms are particularly useful for data-poor cities.

Urban analysis is the study of the different aspects of a city using related data[37, 39, 41, 46–49]. Whether it is crowdsourced or publicly available data, it can be used in a wide range of applications such as for city planning[40, 50], trips and traffic tracking[36, 42, 43], urban activities monitoring[38, 45] and development intensity detection[44].

2.1.1 Urban Data

The straightforward way to obtain urban data is through government portals like [51]. One example of publicly available urban data is the population census. Such data is used for analysis of population dynamics[40]. However, one limitation is that the data is

often historical. Historical data is less beneficial in real-time and urban mobility analysis. The restriction here is the frequency in which the data is updated, which can result in outdated analysis at best.

Another approach to obtain urban data is through controlled environments like using Bluetooth sensors deployed around studied areas as performed by [42, 46]. Other examples include using Wi-Fi probe to observe visitor counts in [40], issuing phones with specialized logging software in [47, 48] and tracking of a number of mobile phone users in [49]. However, setting up such a data collection infrastructure can be costly. In turn, this restricts the sample size of the study and limits their findings. Furthermore, these types of data collection methods may be useless in certain scenarios. One such example is if the users do not enable Bluetooth connectivity on their mobile phones for the case in [42, 46]. As the resulting analysis may be heavily dependent on the sample and methods of the data collection, the findings in these studies are likely unable to generalize well. A preferable approach is to use data that are more easily available such as from vehicle GPS or census.

Another example of public data is from vehicle sensors. Most urban vehicles, like taxis, comes with GPS. Vehicles with GPS can be treated as a location sensor[41, 43]. In this case, vehicle trips can be tracked and analysed. However, the limitation is that the intention of the commuter is unknown. For example, the intention of a commuter travelling from a residential estate to a downtown area can be for work or leisure. It is difficult to determine the intentions of commuters from their trips data alone. Furthermore, it is also a challenge to visualize the complex and huge amount of data. One such example is [41], which mentioned that there are roughly about 500,000 trips in New York alone. Such analysis is complex and may overwhelm a user. The reason is that there are temporal, spatial and neighbourhood analysis information to consider.

Other than obtaining urban data from government portals, an alternative way is through a third party. One example is the data from mobile devices[37–39] from telecommunication companies. These are not public data as they are often restricted and only available to collaborators. Furthermore, not everyone uses the same service providers. This means that, for a research to be comprehensive, it needs to collect data from all telecommunication companies in a country, which may not be feasible. The resulting analysis is limited as it may only be true for a specific user base like in [46]. Additionally, the restriction of

such data is that it is often aggregated to protect the identity of individuals. As such, it may be difficult to determine the specific purpose and activity of a user at a location at any point in time.

This work proposed the use of Tweet which has location, time and text information. This three information can be used to determine the intentions of a user. In this work, the temporal information is taken into account by considering the hourly distribution of Tweets. It is found that when the temporal information is considered, the relevancy of the results from the proposed query framework increases. However, this work does not consider the spatial information in Tweets. The reason is that the boundaries data have clearly separated the Tweets from each region. Although the number of Tweets gathered is restricted for any given point in time, it is still a significant amount. To generalize the proposed query framework, this work combines dynamic data, like Tweets, with other forms of static information like census and POIs to mitigate outdated analysis.

Evidently, it may be difficult for data-poor countries to set up the necessary data collection infrastructure. Collecting Tweets is less costly to set up and can be used as a substitute for traditional urban data. For countries with access to urban data like the census, Tweets can be used as a complementary information provide a more comprehensive analysis.

2.1.2 Urban Analysis

Trips data can be used to analyse the frequented areas, travel patterns and city traffics. A study[41] in New York incorporates temporal information of taxi trips data to analyse urban travel patterns. Such information can also be used to determine the time and location where congestion is most expected. Treating the city traffic events as a sequence labelling problem, a study[36] in San Fransico Bay uses the spatial and textual information in Tweets to conduct their research. In Zurich, a research[43] took a different approach by using data collected from location and acceleration sensors embedded in mobile devices. Their purpose is to determine the travel behaviour and urban activities of these users. However, the findings and observations from these studies may be limited to the city from which it is conducted. The limitation of these data is that it may not be available in less developed countries.

A work in Paris[42] has proposed using Bluetooth tracking sensors to determine the number of visitors and duration of stay at the Louvre Museum. To achieve that, a number of Bluetooth sensors were deployed within the Museum, at each exhibit, to determine the visit and duration of stay at each node. A similar study[40] has been performed in New York City Lower Manhattan neighbourhood, where Wi-Fi probe is deployed at designated areas to determine the population dynamics. Such analysis is limited as it requires a specific wireless connectivity to be activated on the mobile devices. The sensors will not be able to track a user visit if the wireless connectivity is disabled. Additionally, it is a challenge to deploy a large number of these sensors in urban settings. The reason is due to the need for maintenance and the high cost of deployment for such infrastructure.

In Murcia, Spain, a study[45] uses Foursquare data to observe similarities between user preferences and the facilities within the city. The goal is to monitor the different types of urban activities. On another study[38] in Milan, the intensity and evolution of urban activities were tracked using the data from mobile devices. The limitations of these studies are that the definition of urban activities is often unclear. For example, a user could have visited a cafe as he is a cafe hopper or is hungry and wants to look for the nearest eatery.

Using a different approach, a study[44] in Shanghai proposed combining remote sensing imagery and Sina Weibo data to detect development intensity. However, although the findings are intriguing, a straightforward way to detect development intensity is to find areas with denser social media activities. In addition, it is a challenge to conduct such studies in cities with low development intensity but high social media activities.

The discussion in [50] attempts to bridge urban and information theories together, enabling city planners and policy makers to make informed decisions. However, the biggest challenge is finding the appropriate tools and data to perform the analysis that is meaningful to them.

Although this work is conducted in Singapore, the developed URSAS is adaptable. For less developed countries, the boundaries data can be generated manually using the URSAS. Additionally, the developed URSAS in this work can enable city planners to quickly visualize different aspects of a region.

The works mentioned so far show that it is a challenge to provide a semantics for the different urban activities taking place in each region. In this work, the text content in Tweets is used to find out what users are talking about in each region. However, social media activities itself is hardly an indicator of development intensity. The problem is that not every resident will participate on Twitter. For this reason, this work proposes the addition of census as it does not require user participation. In addition, development intensity in this work is determined using POIs like educational and healthcare facilities. The reason is that one of the major consideration factors for city planners is the ratio of the population and these public facilities.

2.1.3 Crowdsourcing and Resident Participation for Urban Analysis

Urban data are available in many forms as discussed in earlier sections. In recent years, crowdsourcing from social media has become a popular method due to the significantly larger amount of data generated as compared to traditional approaches. This is shown in the taxi trips data from [41, 43], Sina Weibo data from [44], Foursquare data from [45] and Tweets from [36]. Data from other studies like [38, 40, 42] are obtained through a controlled environment and may not necessarily be sufficient to justify their analysis in urban settings.

Urban planning affects the residents living in the area. As pointed out by [52], the quality of life of residents can be improved by involving them in urban planning. In Morocco, a study [53] has presented the significance of collaborations between city authorities and the residents. It is important to iron out any conflict of interests among city authorities and the residents. One example is the redevelopment of existing estates that may lead to the reallocation of the residents. It is a long and tedious process to allow resident participation in city planning. This is a challenge when a city grows larger. Urban plans will become more complex, leading to longer planning process. Some possible ways of resident participation include surveys [52] and focus groups like in [53]. However, both cases are difficult to conduct and may not be comprehensive.

It is more desirable to use social media data as a complementary or substitute to traditional urban data. Traditional ways of gathering urban data through the participation of the residents are time-consuming. It is a challenging task to achieve urban growth

while keeping the residents content. As such, this work uses urban and social media data gathered with minimal participation from residents while addressing urban needs.

2.2 Urban Planning

Ultimately, one of the purposes of urban data and analysis is to have a better understanding of urban needs. Urban needs is a major consideration factor for city planners and policy makers. This section discusses the goals and challenges faced in urban planning.

2.2.1 Goals and Challenges

In general, the goals of urban planning[5–7, 13, 52–56] can be categorized into three major types as follows.

- Urban infrastructure management
- Improve livelihood of the residents
- Efficient and effective land use

Urban infrastructure management refers to the planning of transportation nodes and roads for commuters. The study in [7] proposed the use of Twitter and Facebook uploads to detect origins and destinations of public transport commuters. This results in a network showing the hubs of transportation location. From the network, it is possible for policy makers to determine the potential areas for expansion and existing bottlenecks. The challenge here is that not all commuters will upload information about their origin and destination. For example, a commuter might only start updating his Facebook status the moment he reaches his second stop.

In recent years, researchers in urban planning have diverted their attention to urban green spaces[5, 6]. In other words, it is the development of parks and relaxation areas for the residents. It is shown in [6] that urban green spaces promote healthier lifestyle and lessen the stress level of residents. One reason is that residents are more inclined to exercise and relax outdoors if the parks are within their neighbourhood. However, excessive green spaces can have negative impacts on cities where land is scarce. Land in Singapore

is expensive and is often more worthwhile to develop high-density residential estates to accommodate the population. Furthermore, it is pointed out in [5] that developing urban green spaces in neighbourhoods will increase nearby estate value, leading to higher cost of housing.

Efficient and effective land use does not necessarily mean developing high-density residential, industrial and commercial buildings. There is also a consideration for administration and law enforcement jurisdiction as mentioned in [55]. Another factor of consideration is the size of land use. For example, in [5], the study compares the United States and China in an attempt to find the appropriate amount of urban green spaces to develop. The goal is to avoid raising the property values too high while improving the lives of the residents.

The biggest challenge for city planners is to determine the balance between urban requirements for growth and making public housing affordable for new house owners. This work allows city planners to visualize the development intensity of each region. For example, with the help of functional zones and POIs, city planners can determine the number of parks in each subzone. In this case, if there are already sufficient urban green spaces, city planners should divert their attention to other subzones.

2.2.2 Urban Land Use

The main aim of urban planning is to determine the purpose of a given piece of land. There are two ways of looking at this. Firstly, the study in [13] based on the temporal and textual information on Tweet to plot out the different land uses at different time of the day. The intuition behind is that most modern urban buildings consist of a mix of residential, office spaces and commercial facilities. However, the limitation of the study is that the resulting land use type is dependent on the dominant facilities. For example, during office hours, there may be fewer Tweets from workers compared to shoppers in the same building. Consequently, the result contradicts the initial intuition and has limited application.

In a similar approach, a study[54] in Beijing uses a combination of POIs and Sina Weibo data to determine the changes in land use patterns. The study has shown that a more

accurate mapping of the land use type can be obtained when considering both the socioeconomic activities and the facilities in an area. However, the limitation is that it does not take into account urban data, like population demographics, to complement their approach. It is well known that social media data are tentatively sparse and the resulting analysis may not be able to generalize well.

A different study conducted in Ghana[55], West Africa, looks at the peri-urban land use patterns. Peri-urban lands refer to the intersecting areas between urban and rural areas. Although it is not strictly urban land use, it has identified population growth as a contributing factor. However, their hypothesis is based purely on immeasurable factors like culture and history, which limits their work in terms of relevancy to other cities.

In the field of similar region queries, the work in [57] used a quad-tree filter approach to find similar regions using POIs. Their work has shown an efficient divide and conquer approach to split any given region into smaller pieces. Each of these smaller pieces is treated as an aspect of the region to identify similar areas. However, the application is limited in that it is heavily dependent on the individual geospatial information of the data. If a given data about a region does not have the specific location information, the quad-tree filter approach is simply unnecessary. For example, population dynamics may be given as per region basis rather than a geographical point on the map.

This work does not consider the land use patterns based on Tweet or a quad-tree filter to find the POIs. Instead, this work considers the POIs, volume of Tweet and the population to determine the function of each functional zone. It is shown that the proposed query framework can recognize these functions. A quad-tree filter will require precise location information for each of the aspects considered in this work. However, it is challenging to have the precise number of residents for a given piece of land. Hence, a quad-tree filter may not be suitable for urban analysis.

2.3 Clustering

Ideally, Tweets in the same region should form a cluster due to the fact that they are close to one another and topically similar.

2.3.1 Techniques

Clustering or cluster analysis is a method of grouping similar and separating dissimilar points[22, 58]. This approach is useful when there are observable similarities between data points. Clustering can be done in fuzzy[59, 60], where each data point has a weight assigned to it. The sum of all the weights in each group is 1. Additionally, it be modified to have soft cluster membership[22]. Soft cluster membership means that a point may have a probability vector, where each value represents how likely it belongs to one of the clusters.

KMeans[58, 61, 62] is a clustering technique that given a number of clusters, K , it finds K number of groups in a given data set. However, KMeans is not useful in scenarios where there are no prior or insufficient knowledge to the number of natural groups. Furthermore, the choice of the initial centroids will affect the quality of the clusters[63]. A better approach is to use a density-based clustering approach like DBScan[58, 64]. In DBScan, it is not required to define the number of groups during initialization. Instead, DBScan requires two parameters, the minimum neighbours to form a cluster and the ϵ parameter that defines the distance for a point to be considered as a neighbour of another. Although prior knowledge of the natural groups in the data set is not needed, DBScan requires the ϵ and the minimum points parameters to form a cluster. One drawback is that it is not able to separate cluster of different density. In the interest of this work, it is not able to separate distinct clusters that are close to one another. The reason is that the land regions are densely populated within and around the edges of the boundaries. This makes the Tweets generated in one region relatively close to another. As a result, the clusters in any given two land regions may be merged into one.

It is straightforward to cluster data points with numerical values. However, a Tweet contains both numerical and textual content. For numerical values such as latitude and longitude, a measure like the Euclidean distance can be used. A proper distance metric and representation is needed for textual content.

2.3.2 Text Clustering

One way to represent a document is to first convert it into a bag-of-words model[65] like a document term matrix (DTM) or term document matrix (TDM). A bag-of-words[66]

model views a document as a bunch of terms but does not consider the positional information of the terms. A weighting factor, such as the Term Frequency-Inverse Document Frequency (TF-IDF)[67], can be used to indicate the importance of a term. There are many other variants for calculating term frequency (TF) and inverse document frequency (IDF) weights[66] depending on the application. However, one drawback of the bag-of-words model is the curse of dimensionality problem [68, 69], which is can be observed in a sparse TDM or DTM that are filled mostly with zeros. This has two implications. Firstly, it requires a relatively large amount of memory to store the matrix. Secondly, it becomes difficult to cluster documents when they are similar and dissimilar in so many ways due to sparsity.

A number of methods have been proposed to reduce dimension. Some examples are feature selection[70], feature extraction[71] and matrix decomposition[66]. In all three methods, some significant terms may also be removed. These methods may be more useful in improving a classifier but less so for unseen data. In practice, it is possible to simply store all the non-zero entries in the TDM as most similarity measures ignore zero values.

Another approach to reduce dimension is to use NLP filters like removing punctuation, white spaces, numbers, websites, stopwords and converting to lower case letters[21, 66]. Another two common NLP filters are stemming, such as Porter Stemmer[72], and lemmatization[66]. However, the impact of dimension reduction by NLP filters is limited. Short texts found in Tweets often contain abbreviations or short forms that have the same meaning[73, 74]. To tackle this issue, a lexical database like WordNet[75] can be used at the expense of increased computation time. However, through various experiments in this work, it is found that the contribution of short forms and abbreviations to size of the data are negligible. As such, this work does not filter abbreviations or short forms.

Topic modelling is a generative probabilistic approach in which the objective is to attempt to assign a probability to every document that it is related to one or more of the given topics[22–24]. Topic modelling is analogous to soft text clustering[22, 58, 67]. Two basic topic modelling approaches are Probabilistic Latent Semantic Indexing (PLSI) and Latent Dirichlet Allocation (LDA)[22–24, 76]. The difference is that LDA models the term-topic and topic-document probability with a Dirichlet distribution prior[77] and

can be observed as a Bayesian version of PLSI[22, 23, 76]. On larger data set, PLSI model parameters grow linearly with the size of the collection. As such, LDA is more commonly used. However, one major drawback of LDA is that early termination may lead to less than optimal solution while a large number of iterations can result to long processing time and over fitting. To tackle this issue, an online variation of the LDA[25] can be used.

2.4 Clustering Multiple Features

When considering multiple features, the clustering approach is slightly different. For example, a Tweet can contain textual and location information. These are considered as two separate features as the dimensional space and bounds of both are different. The location is represented in latitude and longitude values while the text can be represented using a bag-of-words model. When considering the approaches to use, it is important to identify the intended purpose. There are a few approaches to incorporate additional features for clustering. A naive approach can be to normalize all feature values into a single dimensional space. In such cases, a single feature clustering approach can be used.

A less naive but intuitive approach is to use a weighted feature method such as multi-view clustering[78, 79]. In multi-view clustering, each feature is compared independently. For example, when comparing two Tweets t_1 and t_2 , weights can be assigned to the spatial and textual calculation that indicates the importance of each feature. Formally, if the weight is α , the distance of two Tweets can be calculated as shown in Equation 2.1.

$$Dist(t_1, t_2) = (\alpha)(SpatialDist(t_1, t_2)) + (1 - \alpha)(TextDist(t_1, t_2)) \quad (2.1)$$

It is worth noting here that similarity can also be computed using $1 - Dist(t_1, t_2)$. One benefit of using a weighted approach is that it clearly defines which feature is considered to be more important. However, it is a challenge to choose an appropriate α value.

Another approach is to have individual cluster membership for each feature. One such example is subspace clustering[80–82]. The idea behind is to view the data in a different manner. For example, a Tweet can have two different cluster membership label. In

this case, two Tweets can be in the same location-based cluster but different text-based grouping. Evidently, this is useful for finding different groupings for different features but not for a single set of clusters. A possible workaround to have a single set of clusters is to group the data points based on their cluster membership for each feature such as in the case of meta-clustering[83]. In meta-clustering, the final cluster membership is determined by the prior clustering result. However, one drawback of meta-clustering is that data points that belong together in one feature space may not necessarily end up in the same final cluster. Choosing the appropriate approach can result in more meaningful clusters. However, it is also essential to define how two data points are similar.

2.4.1 Text Similarity

There are few ways to measure the similarity between two texts. The naive way is to count the number of terms that occur in both texts. Two such approaches are the Jaccard similarity coefficient (JSC) and simple matching coefficient (SMC) [22, 58, 67]. However, it is worth noting that such approaches will be unable to detect text that are similar in terms of their word distributions.

JSC between two texts is the fraction of the number of intersecting terms over the union of the terms in both texts. For example, consider two sets $\mathbf{A}$ and $\mathbf{B}$. $\mathbf{A}$ represents the terms in the first piece of text. $\mathbf{B}$ represents the terms in the second piece of text. The intersecting terms between the two texts are denoted by $\mathbf{A} \cap \mathbf{B}$. With regards to the bag-of-words model, JSC will only consider the terms that occur in both texts. It is worth noting that the Jaccard distance, $JD(\mathbf{A}, \mathbf{B})$, is simply the complementary of JSC. SMC is similar to JSC but with one key difference. When computing SMC, the terms that are in the collection but do not occur in both texts are also considered. Similarly, the complementary of SMC is simple matching distance (SMD) $SMD(A, B) = 1 - SMC(A, B)$. Note that in the case of SMC, the computation time of the sum of the number of terms that do not occur in both $\mathbf{A}$ and $\mathbf{B}$ increases if the total number of terms is large. As such, SMC may not be suitable for large collection.

Another common method for measuring text similarity is Euclidean distance[22, 58, 66, 67, 84]. Note that unlike other similarity measures, Euclidean distance measures the dissimilarity between two texts. When measuring using Euclidean distance, it is assumed that the vectors lie on a plane of n-dimension.

A less naive method for measuring text similarity is by using the Cosine similarity[66, 84] approach. The idea behind is the Euclidean dot product formula for two non-zero vectors. Cosine similarity of two vectors measures the angle between two vectors. The angle for two of the same vectors is 0° . For two perpendicular vectors, the angle is 90° . The angle for two directly opposite vectors is 180° . With this in mind, observe that $\cos(0^\circ) = 1$, $\cos(90^\circ) = 0$ and $\cos(180^\circ) = -1$. In other words, if two vectors are similar, their values should be close to 1, closer to 0 if they are dissimilar and -1 if exactly opposite. When comparing two texts that are represented by TF-IDF, the values are non-negative. A TF-IDF vector is also known as term vectors. It is also worth noting that the angle between two term vectors cannot be bigger than 90° . The Cosine distance between two documents is simply the complement of Cosine similarity. Note that this is not necessarily true for other types of vectors.

2.4.2 Spatial Distance

Spatial distance for two points is the measure of the length between them in an n -dimensional space. Depending on the context, the spatial distance can also refer to the measurable path between two geographical points represented by latitude and longitude values. There are few ways to measure the distance between two geographical points. For example, one way to measure two geographical points is to consider them as if they lie on a two-dimensional flat space. In this case, the path between the two points can be calculated using the Euclidean or Manhattan distance. It is worth noting that if the points were treated as if they lie on a spherical plane, the length can be calculated by using the Haversine formula[85]. However, in this work, a metric is needed rather than the actual distance. As such, the Haversine formula is omitted from the discussion.

The Manhattan distance or L_1 norm[86] is a way of measuring the path from two points. Unlike Euclidean distance, it does not guarantee to find the shortest direct distance. One main obvious difference is the way it is calculated. The Euclidean distance[22, 58, 67], which is also known as L_2 norm, measures the shortest direct distance between two points. It can be observed that the Euclidean distance of two geographical points, represented in latitude and longitude, satisfy the Pythagorean formula.

2.4.3 Probability Distribution Similarity

There are a few ways to consider two similar distributions. For example, two distributions can be considered similar if they are highly correlated. This means if one increases in value, the other also increases. Alternatively, two distributions can be considered similar if one does not diverge too far from the other.

The L_1 norm[86, 87] and the L_2 norm[22, 58, 67, 87] can also be used to measure distribution similarity. For L_1 and L_2 , two distributions are similar if the value is closer to 0 and non-zero if dissimilar. Another commonly used measure for distribution similarity is the Cosine similarity[66, 84, 87]. However, in this case, if the value is 1 means that the two distributions are same and 0 if entirely different.

Two other common methods for measuring distribution similarity are Kullback-Leibler (KL) divergence[87–89] and Jensen-Shannon (JS) divergence[87, 90–93]. One problem with KL divergence is that it is asymmetrical. This makes it difficult to have a standard way to differentiating two distributions. However, it useful in determining dissimilar distribution as it will result to non-zero values in those cases. If two distributions behave similarly or the same, their KL divergence will result to 0 and 1 if dissimilar. In other words, if a value is closer to 1 implies dissimilar and 0 implies similar distributions. Another problem with KL divergence is that it is only defined if $P(B_i) = 0$ implies $P(A_i) = 0$. This means that if two distributions have a different number of probabilities, it is undefined. Consider if **A** has 3 probabilities and **B** has 2 probabilities. In this case, $P(B_3)$ is undefined. One way to fix this is to assign $P(B_3)$ as 0.

JS divergence is based on KL divergence but it is symmetrical. Additionally, the JS divergence returns finite value. One of the useful properties of JS divergence is that it is bounded between 0 and 1 if base 2 logarithm is used. In KL divergence, if $P(B_i)$ is zero, $\frac{P(A_i)}{P(B_i)}$ is undefined. However, in JS divergence, the average of the two probability distributions is used. In this case, $P(M_i)$ is non-zero if either $P(A_i)$ or $P(B_i)$ is non-zero. This means if **A** has more probabilities than **B** and is non-zero, JS divergence will give a finite value.

2.5 Summary

In summary, urban data, which refers to data originating from an urban area, are used by city planners to make informed decisions. However, with the advancement in technologies, it is possible to use social media data to assist in urban analysis. This work has shown the possibility of combining Tweets, using different methods of measuring similarity, with urban data to query for similar regions. The goals of urban planning are to improve the quality of life of the residents and make effective use of land. Although it is desirable to involve residents in the decision making, it is often a time-consuming process. It is also a challenge for less developed and data-poor countries as it may be difficult for them to conduct a comprehensive data collection that requires the participation of the residents. Gathering data from social media platforms is less costly. This work has shown that it is possible to conduct urban analysis, like querying for similar regions, with minimal involvement from the residents. Furthermore, it is evident that a piece of land can serve a different purpose at different time of a day. Traditional data gathering methods may not be able to clearly identify the different functions of land use effectively. It is shown in this work that the combination of Tweets and urban data can recognize the land use type of a functional zone.

Chapter 3

Problems and Techniques

This chapter will discuss the problems that this work aims to solve and the techniques used. To begin, the main problems are discussed. This is followed by the techniques and dataset used to tackle them.

3.1 Problems

This section presents the problems that this research aims to solve. These problems are the main objective of this research, which is mentioned in [Section 1.4](#).

3.1.1 Urban and Social Media Data Analysis

The first objective of this work is to find out if there is any correlation between data. Specifically, it is to find out whether there is any correlation between the population, Tweets and POIs in an area. A sub-analysis is to investigate the clusters of Tweets in Singapore. The aim is to find out the effectiveness of Tweets for urban analysis in data-poor countries. It can be agreed, and proven by many existing works, that Tweets can show interesting human behaviours and patterns. However, the main limitation of any of these existing works is the lack of applications to urban settings. Urban plans are quite predictable whereas Tweets are unpredictable. This work hopes to use urban data to complement Tweets to narrow this gap between academic interest and commercial applications. It can be shown in this work that the best result can be achieved by

considering both urban and social media data in querying for similar regions. A possible extension of this work is to include a wider range of aspects rather than just Tweets, POIs and population census.

The result of this sub-analysis is covered in Section 4.2. Findings on the correlation between data will be presented Chapter 5. The techniques that are going to be used for this analysis are further described in Section 3.2.

3.1.2 Similar Urban Region Query

The second objective of this work is to find ways to identify similar urban regions. It is in the interest of this work to investigate how to query for similar urban region based on urban and social media data. The main problem in this area is the lack of attention with regards to querying similar urban regions. It may be true that it is rather trivial to simply have query system that output regions that has the identical facilities. However, existing works in urban planning has shown that each region can serve multiple functions. This can be achieved by having in-depth study on the different types of functions that is provided by each region. However, such an attempt may not scalable. This work emphasize that similar regions can be found by considering multiple aspects. It is also shown in this work that such a feat could be achieved using less costly and scalable approach by leveraging on urban and social media data. A possible extension of this part of the work is have a system that allows users to query for certain types of regions and the output is areas that fit the criteria.

Part of this analysis is shown in Section 4.2. The performance of the similar urban region queries is presented in Chapter 6. It is worth noting that the performance of the queries is evaluated based on how relevant the results are to different types of users. The techniques for finding the correlation between data and evaluating the performance of the queries are further elaborated in Section 3.2.

3.1.3 Functional Zones Analysis and Recognition

The third objective is to investigate if there is any correlation between the aspects of a functional zone and the possibility of using these aspects to recognise the function of each functional zone. To achieve this, the functional zones are generated manually

from URSAS. This part of the work hopes to recognize the function of each region, as defined in the MP, by using just Tweets and the facilities. The main challenge in this area is that existing works focuses on different types of data and techniques that are difficult for commoners to relate. Using Tweets and POIs, it is intuitive for any users to relate to the functions recognised by the system. This work has shown that using Tweets and POIs together can have significant increase in the accuracy of the function recognition of regions. The only limitation in this work is the lack of in-depth tuning of the system. A possible extension is to consider the infrastructure development to determine the function. For example, it can be obvious that downtown area should have well developed infrastructure in place.

The manually generated functional zones are further explained in Chapter 4. Findings for the analysis and recognition of functional zones are presented in Chapter 7. The techniques to be used for this problem will be discussed in the following Section 3.2.

3.2 Techniques

This section describes the various techniques that are going to be used to achieve the objectives of this work.

3.2.1 Observing Correlation Between Urban Data and Tweet

Intuitively, the population distribution is likely to have an impact on the existing and future POIs in each region. Similarly, Tweets might have a positive correlation with the POIs in each region. To explore this intuition, the Pearson correlation coefficient[94] is applied to the distribution of each aspect in each region. The formula for the Pearson correlation coefficient is shown in Equation 3.1.

$$\rho_{\mathbf{A},\mathbf{B}} = \frac{\text{cov}(\mathbf{A},\mathbf{B})}{\sigma_{\mathbf{A}}\sigma_{\mathbf{B}}} \quad (3.1)$$

- $\mathbf{A}$ and $\mathbf{B}$ are the distributions
- $\text{cov}(\mathbf{A},\mathbf{B})$ is the covariance of the distributions $\mathbf{A}$ and $\mathbf{B}$

- $\sigma_{\mathbf{A}}$ is the standard deviation of distribution $\mathbf{A}$
- $\sigma_{\mathbf{B}}$ is the standard deviation of distribution $\mathbf{B}$

It can be shown in this work that aspects that are highly correlated with one another tends to retrieve more relevant results to the users. As such, this insight can be used as a guideline to whether the aspects are suitable to be combined.

3.2.2 Finding Tweet Clusters

To analyse whether Tweets can be used in urban planning, a clustering approach was tested. The purpose is to determine if clustering can be used to identify different regions. Intuitively, Tweets that are in close proximity should be within the same clusters. However, if only the spatial distance is considered, the resulting clusters will have Tweets that close to one another but may be topically inconsistent. A better approach will be to consider both the text and location information in Tweet. In this work, the Euclidean distance is used for the location information while the Cosine similarity is used for the text content. The algorithm used for clustering is DBScan. The formula to obtain the similarity value between Tweets is shown in Equation 3.2.

$$Sim(T_i, T_j) = \alpha(L_2(l_i, l_j)) + (1 - \alpha)(CosSim(t_i, t_j)) \quad (3.2)$$

- α is the user-defined weight parameter
- N is the total number of Tweets, $i, j \in N$
- T_i and T_j are Tweets
- l_i and l_j are the coordinates of T_i and T_j respectively
- t_i and t_j are the text of T_i and T_j respectively

However, it can be a challenge to simply use clustering approach to find similar regions. Existing clustering approach like KMeans requires prior knowledge on the natural number of Tweets grouping, which can be difficult to determine. Intuitively, it is possible to assign

the value of K to be the number of boundaries data available. For example, if there are 323 subzone boundaries, the value of K can be set as 323. However, the results will tend to have few dominant large clusters that overlaps multiple boundaries. The rest of the clusters tends to be small and do not result to any meaningful groupings. It can be observed that KMeans requires in-depth parameters tuning for it to clearly separate the Tweets into the defined boundaries found in the MP.

The drawback of clustering in similar region analysis is further emphasized by density-based approach like DBScan. Using DBScan, the challenge is to determine the appropriate ϵ value. Although it is intuitive that a small ϵ leads to specialized clusters, the boundaries defined in the MP do not have fixed shapes or sizes. A possible workaround could be to have a dynamic ϵ value for Tweets that changes according to some condition. However, the possible conditions required for such an attempt are endless as the reasons for the boundaries to be defined in certain manner by the MP are latent. With the current limitations of clustering techniques, a faster and straightforward approach will be to simply label each Tweet based on the boundary that it is uploaded in. However, it is still possible to use the measures used in clustering to find similar regions.

3.2.3 Proposed Region Similarity Approach

This work proposed using the various similarity measures to identify similar regions. Specifically, one way is to consider each aspect of a region independently. This can be achieved by simply applying a suitable similarity measure for the aspect considered. For example, the Cosine similarity is used for the text content in Tweet while the JS divergence is used for population and POI distributions. An alternate way is to consider multiple aspects of a region. In the proposed query framework, users can assign weights to different aspects of a region. The idea behind is that a user may want to place emphasis on certain aspects during a query. Analogous to a web search engine, a user will want to have relevant results. The weight-based similarity measure formula for considering two aspects of a region is shown in Equation 3.3. To consider more than two aspects, the user will be required to set additional weights $w_1, w_2, \dots, w_k$ where $k = (\text{number of aspects} - 1)$ and $\sum_{i=1}^k w_i = 1$.

$$Sim(R_i, R_j) = (w_1)(Sim1(a_{i1}, a_{j1})) + (1 - w_1)(Sim2(a_{i2}, a_{j2})) \quad (3.3)$$

- w_1 is a user-defined weight parameter for the first similarity measure of the first aspect
- R_i and R_j are regions
- a_{i1} is the 1st aspect and a_{i2} is the 2nd aspect of R_i
- a_{j1} is the 1st aspect and a_{j2} is the 2nd aspect of R_j
- Sim1 refers to the measure used to obtain the similarity for the 1st aspect of R_i and R_j
- Sim2 refers to the measure used to obtain the similarity for the 2nd aspect of R_i and R_j

This work also considers the topic distribution in each region. The intuition behind is that two regions are identical if their topic distributions are similar. The topic distributions are generated individually for each region and the number of topics is assumed to be 10. However, since each Tweet will only appear once, the terms for each topic in each region are likely to be different. When analysing two regions, it does not make sense to compare two distributions that refer to two different set of topics. For this reason, a novel approach is proposed in this work to compare the topic distributions. Given two regions, R_1 and R_2 , the following steps are iterated to obtain the similarity between two topic distributions.

1. For a topic in R_1 , use Cosine similarity to find the most similar topic from R_2 based on their topic terms
2. Once a suitable topic is selected from R_2 , it is removed from future selection
3. Repeat 1. and 2. for all other topics in R_1
4. For any topic in R_1 without a corresponding topic in R_2 , remove it from the comparison

5. Apply a suitable similarity measure on the topic distribution of R_1 that has a corresponding topic in R_2
6. If none of the topics in R_1 has a corresponding topic in R_2 , the similarity value between R_1 and R_2 is assumed to be 0

3.2.4 Validating Region Similarity as Perceived by Users

Similar regions can be defined in multiple ways. However, it is subjected to individuals in their perception of similar regions. For example, a region can be similar to one another based on their functions or the type of visitors. As such, users of URSAS are asked to give a label between 1 to 5 for the top 5 similar regions of each area. For each region, there are 6 sets of labels that represent residential, business, commercial, students, working adults and unemployed relevancy. The meaning for each relevancy is further elaborated in Section 4.4.2. A value of 1 represents least similar and value of 5 represents most similar. Regions that are not labelled will be assumed to have a default value of 0.

To validate the proposed approach against the perceived similar regions by users, two measures are used. The measures [66, 84] are normalized discounted cumulative gain (nDCG) and precision. nDCG measures the ranking quality of the results returned by the similar urban queries. The process of getting the nDCG value is straightforward. Firstly, a user submits a query to URSAS. Secondly, upon receiving the results of the query, the user is asked to give a value between 1 to 5 for the top 5 similar regions as perceived by him. Next, once the user is satisfied with his labels, URSAS evaluates the queries against the ranking list. These steps are repeated for different types of users and to obtain the average nDCG and precision value. The formula for nDCG and precision are shown in Equations 3.4 and 3.5.

$$\begin{aligned}
 nDCG_k &= \frac{DCG_k}{IDCG_k} \\
 &= DCG_k \div IDCG_k \\
 &= \sum_{i=1}^k \frac{rel_i}{\log_2(i+1)} \div \sum_{i=1}^x \frac{2^{rel_i} - 1}{\log_2(i+1)}
 \end{aligned} \tag{3.4}$$

- $nDCG_k$ represents the nDCG value, ordered by their similarity level, at k^{th} retrieved region
- rel_i is the i^{th} relevance label of the retrieved region as provided by the user
- x is the number of non-zero labels for the retrieved region up to k

$$precision = \frac{|X \cap Y|}{|Y|} \quad (3.5)$$

- X represents the number of similar regions perceived by user
- Y represents the number of regions retrieved by URSAS

The recall is obtained by dividing the number of regions that users perceived as similar by the total number of true similar regions. Since there is insufficient ground truth to justify the total number true similar regions, recall is not used to evaluate the queries.

3.2.5 Functional Zones Analysis and Recognition

Due to the absence of the individual type of land use boundaries from the relevant authorities, it needs to be manually generated. In this work, the functional zones within the central region boundary are manually generated using URSAS. The correlation between Tweets and POIs in each functional zone is obtained using the Pearson correlation coefficient.

The steps to test the recognition capability of the URSAS is simple. Firstly, a functional zone is queried. Next, the function of the most similar neighbour is set as the new function for the queried functional zone. These steps are repeated for all functional zones. To evaluate the recognition capability, the accuracy measure [66, 84] is used. Accuracy is obtained by dividing the number of functional zones, that match the function suggested by URSAS, by the total functional zones.

3.3 Aspects of a Region

In total, there are three aspects of a region considered in this research. To do so, the boundaries of the each region of the MP are required. This is further described in Subsection 3.3.1. The first aspect is the population dynamics of each subzone and planning area by age and gender in 2014. This is described in Subsection 3.3.2. The second aspect is the various types of POIs such as supermarkets and hawker centres locations which are obtained from [51]. However, the locations of the shopping malls were not available on the government public data portal. To overcome this problem, the list of shopping malls were crawled from Wikipedia[95] and reverse geocoded using Google Maps API[96]. This is described in Subsection 3.3.3. The third aspect of the regions considered are the Tweets in Singapore, which are obtained from [97] and further described in Subsection 3.3.4.

3.3.1 Master Plan (MP) Boundaries

The MP boundaries divide Singapore into regions, planning areas and subzones. These boundaries are used to define the regions for similarity comparison. In total, there are 5 regions, 55 planning areas and 323 subzones. These are initially given in KML but are converted and stored in commonly used GeoJSON RFC 7946 standard[98] format for indexing purposes. The boundaries are represented in a series of longitude and latitude coordinates. These coordinates form the geographical polygons that represent the boundaries. For efficient storage purposes, each boundary record can represent more than one polygons. For example, the offshore North-east islands are part of the North-east region. Rather than plotting a large boundary that includes them, it is represented by multiple polygons to depict the separation of the islands by sea. The illustrations for these boundaries are shown in earlier in Subsection 1.2.1.

It is worth noting that the specific land use plan for each plot of land is not available at the point this research is conducted. Both URA and OneMap[99] do not have the necessary services to support the feature. For example, refer to the central region shown in Figure 3.1. It shows the residential, commercial and business land use in the central region. The developed URSAS in this work allows the drawing of polygons on the map. Hence, it is possible to reproduce the functional zones by manually drawing the individual

polygons. This work replicates the functional zones found in the central region boundary as a novel contribution. The purpose is to test whether URSAS is able to recognize the function of each functional zone. Section 4.4.4 will elaborate more on the functional zones.

FIGURE 3.1: URA Master Plan 2014: Central Region

3.3.2 Population Distribution

The population distribution provides demographic information of the residents living in each of the area defined by the MP boundaries data. This data will be compared with the POI distribution to find any potential correlation. However, the population distribution data is only available for planning areas and subzones boundaries. To overcome this problem, the population distribution for each region is simply the sum of all the resident distributions in the individual planning areas that are within each of the region boundaries. Information on the age and gender of the residents are provided. A sample of the data for subzones is shown in Table 3.1. The residents, based on their age in each subzone, forms the population distribution.

Name	0-4	5-9	10-14	15-19	20-24	25-29	30-34	35-39	40-44
Maritime Square	100	90	60	50	50	50	140	150	140
People's Park	10	0	0	0	10	20	20	30	20
Bukit Merah	20	30	30	30	30	40	40	50	50
Chinatown	540	360	190	180	200	290	710	750	460
Name	45-49	50-54	55-59	60-64	65-69	70-74	75-79	80-84	≥ 85
Maritime Square	120	80	80	60	30	20	20	0	10
People's Park	10	10	10	20	20	10	10	10	0
Bukit Merah	40	50	50	50	40	30	20	10	10
Chinatown	350	330	350	370	300	220	200	150	200

TABLE 3.1: An Example Female Population Distribution on Each Subzone by Age

3.3.3 Point of Interests (POI) Distribution

The POI distribution contains the name, nature, address and category of the public facilities in Singapore. In total, there are 15,864 POIs. Addresses of these POIs are given in postal codes. These postal codes are fed into a reverse geocoding service, provided by Google, to obtain their respective longitude and latitude values. However, this data does not contain the information on the intersecting boundary of the POI in each region. This is solved by simply finding the intersecting boundaries for each POI. A sample of the POIs is shown in Table 3.2. The POI distribution in each region is the number of facilities of each category given by the government data portal.

Latitude	Longitude	Name	Category
1.280	103.814	Cherie Hearts	Child Care Centres
1.299	103.855	Amore Fitness - Bugis Junction	Gym
1.316	103.858	60's Hostel	Hotels
1.325	103.929	Bedok Mall	Shopping Malls
1.374	103.885	NTUC Fairprice Co-operative Ltd	Supermarkets

TABLE 3.2: An Example of POIs

3.3.4 Tweet

The Tweets were obtained between 24th June 2014 and 21st September 2015 in Singapore. There are more than 5 million tweets collected. Each Tweet contains a short text and a set of latitude and longitude coordinates that shows where it was uploaded and the time that it was posted. Similar with POIs, each Tweet is queried for the intersecting polygon to define the area in which it is located. However, most Tweet contains short forms and text. To reduce the impact of the size of the corpus during preprocessing, many NLP filters were used to remove URLs and white spaces. This data is used to find the topics and terms that were dominant in each of the regions specified in the boundaries data. A sample of the Tweets can be found in the discussions in Section [1.2.2](#).

3.4 Summary

In summary, this chapter has described the various aspects of a region, the problems to be solved and the proposed techniques. The following Chapter [4](#) will present some of the preliminary experiments performed using the proposed techniques.

Chapter 4

Urban Region Similarity Analysis System (URSAS)

A web-based system was developed in this work, using rapid prototyping, to visualize the aspects and similar areas for each region. This chapter first describes the implementation and the preliminary experiments conducted using the early prototype of URSAS. Next, this chapter will present the final version of URSAS that is used for this work.

4.1 Implementation

There are five major modules in the developed URSAS as shown in Figure [4.1](#). The architecture of URSAS is shown in Figure [4.2](#).

FIGURE 4.1: URSAS Modules

FIGURE 4.2: URSAS Architecture

- (a) MongoDB[100] is a No-SQL database that represents records in JSON
- (b) RoboMongo[101] is MongoDB database management tool
- (c) Spring Mongo Bridge is an interface that allows data transfer between the embedded Tomcat Server[102] in Spring framework[103] and MongoDB

- (d) R[104] is a language and environment for performing statistical data analysis. R server[105] is the server that allows the Tomcat server to perform task in R environment
- (e) the Preprocessor component is a series of R scripts that perform the necessary preprocessing steps before storing the inputs into MongoDB
- (f) the Plotter is another set of R scripts that plots out different charts and figures
- (g) the Input is provided in comma separated values (CSV)
- (h) the Tomcat server is embedded in the Spring framework package that hosts the different services for use on the URSAS
- (i) the Data Definition component is a number of Java[106] classes that define the database schema
- (j) the Repositories component is a number of Spring framework Java classes that define the MongoDB operations
- (k) Maven[107] is a dependencies framework that can be used to define the external libraries used in the server
- (l) the Similarity Analyser component is a number of Java classes that define the similarity measures used in the experiment
- (m) the Clusterer component contains a number of Java classes that define the clustering techniques used in the experiment
- (n) the RestfulAPI Controllers is a number of Java class that define services provided by the server
- (o) Bootstrap[108] is a web framework for designing responsive web pages with accessible charting libraries
- (p) the Chart Libraries component contains a number of external Javascript libraries that allows visualization of data on charts
- (q) the Web Pages consists of HTML and Javascript files that are used to display a web page to the user
- (r) the Mapbox[109] is a cloud-based geographical map library

- (s) the Web Browser is a software that the user uses to interact with the developed tool
- (t) the User is someone who uses the developed tool

The web interface of URSAS is developed in such a way that users can choose the type of visualization to display. Screenshots of URSAS are shown in Figure 4.3. The web interface is designed based on the following objectives.

- **Objective 1** - clear visualization of the aspects in each region defined by the boundaries data
- **Objective 2** - selection of different aspects to be used in the experiment
- **Objective 3** - comparison of the regions based on specific aspect
- **Objective 4** - customization of the measures used during comparison
- **Objective 5** - customization of some of the parameters for the algorithms used

(a) Overview 1

(b) Overview 2

FIGURE 4.3: Web Visualization Tool: Overview

4.1.1 Input and Storage

The input for the developed tool is in CSV format except for the boundaries data. CSV is a format where values are separated by commas. This is useful since the URSAS is made up of components of different languages and platforms. Before storing the input, it goes through the preprocessor for the necessary text filtering.

In MongoDB, records are stored as documents in JSON format. This is useful during data transfer between MongoDB, the Tomcat server and the front-end web pages. Additionally, the boundaries data are stored in GeoJSON format for indexing purposes. Storing data in JSON and the geographical coordinates in GeoJSON format enable geospatial queries. A sample of how the database stores the records is shown in Figure 4.4.

```
{
  "_id" : ObjectId("58fecf0677c85562942c0b1a"),
  "_class" : "com.google.map.polygon.tool.data.ValidTweet",
  "tweetId" : "2982",
  "userId" : "238486507",
  "username" : "Lionel Polaris Song",
  "latitude" : "1.34296636",
  "longitude" : "103.79708469",
  "location" : {
    "type" : "Point",
    "coordinates" : [
      103.79708469,
      1.34296636
    ]
  },
  "datePosted" : "2014-06-24 19:18:45",
  "text" : "I'm at Riders Cafe (Singapore) http://t.co/K04y9zRsLu",
  "cleanedText" : "riders cafe lrb rrb",
  "category" : "default_category",
  "regionName" : "default_region_name",
  "subRegionName" : "default_sub_region_name",
  "visited" : false
}
```

FIGURE 4.4: Storing Records as Documents in JSON format

4.1.2 Preprocessor

The four main functions of the Preprocessor are handled by R scripts and MongoDB geospatial queries. Firstly, it is well known that the text in Tweet is often unstructured, multilingual and contains special symbols and characters. These observations cause a few problems. Text like "I'm" and "I am" are treated differently in a bag-of-words model but both have the same meaning. This results in a larger word matrix and in turn increase computation time. The second problem is that certain types of similarity measure may be unable to accurately identify similar Tweets, resulting in sparse clusters. Additionally, it is common for Tweets to contain a URL link that points to the actual post. This URL link is unique and has no meaning when measuring similarity for two Tweets. The text filtering function is performed using a number of R scripts. In particular, the following text filters are done sequentially in the Preprocessor.

- remove all URL links
- convert all letters to lowercase
- remove numbers
- remove all punctuation, symbols and special characters
- strip whitespace
- remove stopwords
- lemmatize all words

Secondly, the Preprocessor specifies the land regions for the Tweets, POIs and residents aspect. This is handled by the geospatial queries in MongoDB. It is achieved by finding the intersecting geographical polygons boundaries for each data point. Consequently, the respective land region for each aspect is updated in their corresponding records in the database. The density of Tweets, residents and POIs can be observed in this step.

It is possible to generate term and document frequencies beforehand for each region, which is the third function of the Preprocessor. The term and document frequencies are essential for generating the TF-IDF matrix and is performed at the Java Tomcat server. These frequencies are stored in the database for each land region and are later used in the Analyser module. Note that, if required, it is possible to modify this function to handle streaming data.

Lastly, the job of the Preprocessor is to obtain the topics and top words for each land region. The top words can simply be observed from the generated term frequencies. For the topics, the corresponding text is sent to a Java class that runs the LDA. The output is the most probable topic label for each Tweet and the most likely terms for each topic. As a result, each land region in the database is updated with the topics and the corresponding terms.

4.1.3 Similarity Analyser

The Analyser contains the various similarity measures. At this point, it is assumed that the Preprocessor has done the necessary text filtering and eliminate any noisy data. The job of the analyser is to provide a value of similarity between each data. For example, if Cosine similarity is used to compare two Tweets, it should return a value indicating the level of similarity. The similarity value and the measure used, for any land regions against all others, are stored in the database. This prevents repeated calculation for regions that have been compared before. Customization of the measure, such as whether to consider the text and spatial information in Tweet, is also done at this module.

4.1.4 Clusterer

Analogous to the Similarity Analyser, the Clusterer contains various clustering methods. The scope of the Clusterer is simply to group similar data points together based on

the similarity value indicated by the Similarity Analyser. Parameters of the clustering algorithm, such as the number of clusters, are defined in this module. The resulting cluster memberships for the different algorithms are stored in the database.

4.1.5 Visualizer

The objectives of the web interface mentioned earlier in Section 4.1 are achieved in the last module, the Visualizer. This module is responsible for allowing users to customize and view the results of the queries. Customization of the looks and feel of the front-end web pages are done in this module. Another point to note is that this module does not perform any calculation. It is only responsible for displaying the results.

The first objective, which is to clearly visualize the attributes of each land region, is achieved by showing the number of Tweets, residents and POI on the top right-hand side of the map. The third objective is achieved by using different colours. Darker colours imply higher density as evident in Figure 4.5 that shows the density of Tweets, residents and POI.

(a) Tweet

(b) Population

(c) POI

FIGURE 4.5: Web Visualization Tool: Density Visualization

For the selection of data, URSAS allows users to choose either Tweet or POI and population distributions view. For each view, users can choose to view it in either region, planning area and subzone boundaries. The Tweets are available in the first view, as shown in Figure 4.3(a), where users can observe the top words in the word cloud and

the individual posts. On the right tab, there are options to observe top words and topics distribution, choice of similarity measures and the aspects to compare. Lastly, users can choose two kinds of similarity with an alpha weight that is used to define the significance of one measure. These options are shown in Figure 4.6. Most similar regions are in darker shades of green. With these options, objective three and four are achieved. Objective five is achieved because it allows users to customize the weights for similarity measures.

(a) Top Topics and Words

(b) Single Similarity Measure

(c) Weighted Double Similarity Measures

FIGURE 4.6: Web Visualization Tool: Tweet View Options

From the visualization in Figure 4.6(b), it is observed that many regions are similar to one another based on the Cosine similarity of their TF-IDF values. It is worth noting here

that, interestingly, the similarity level reduces when combined with the JS divergence of their term frequency.

The second view allows comparison of the population and POI distributions between land regions. It is separated from the Tweet view since it does not make sense to perform topic modelling and word frequency analysis. Instead, a more meaningful approach will be to show the distributions in bar charts. This is shown in Figure 4.7. It fulfils objective one since it allows viewing of the different distributions in each area. Objective two, three, four and five are achieved as users can choose the data, similarity measures and parameters.

FIGURE 4.7: Web Visualization Tool: Residents and POI Distributions

4.1.6 Defining Land Use

Other than the basic components to allow urban and social media analysis, a sub module was also developed to allow users to draw the boundaries on the map. The reason for this sub module is to enable manual generation of any boundaries that are not available from the relevant authorities. Apart from being able to define the colour, it is also able to define the boundary shape and size. This sub module is illustrated in Figure 4.8. The green coloured polygon is a previously drawn boundary that represents parks. The red coloured polygon is at the drawing stage, in which the opaque and translucent white coloured dots represent the movable and extensible corner edges respectively.

FIGURE 4.8: Web Visualization Tool: Defining Land Use

4.2 Preliminary Experiments

A series of preliminary experiments were conducted to evaluate the effectiveness of the clustering approaches and similarity measures to identify similar regions. This section shows the result of the analysis performed. Firstly, the result of clustering Tweets based on their spatial and text similarity is shown in Section 4.2.1. Next, Section 4.2.2 illustrates the use of the individual aspects of a region to identify similar regions. Lastly, Section 4.2.3 shows the results when multiple aspects are used to identify similar regions.

4.2.1 Clustering Tweet

This part of the experiment is to determine if a simple clustering approach can clearly separate Tweets from different regions. The experiment setup is summarized in Table 4.1.

The results are shown in Appendix A where each dot represents a Tweet and their cluster membership is represented in different colours on the map. The results show that it is difficult to determine the appropriate α parameter. Small α value, which indicates less importance on the location, leads to clusters with points scattered around. High α value does not necessarily improve the clusters since it is evident that some points are still scattered. Small ϵ value, which indicates lesser neighbours that a point requires to

Options	Value
Algorithm	DBScan
Eps	0.01, 0.03, 0.05, 0.08
α	0.3, 0.5, 0.8
minPts	3
Spatial Distance	Euclidean
Text Similarity	Cosine Similarity
Text Representation	TF-IDF

TABLE 4.1: DBScan Tweet Experiment Setup

become a core point, results to more topically consistent groups. It can be observed that, assuming α is constant, varying eps does not have much significance on the cluster membership.

4.2.2 Single Similarity Measures

Evidently, it is difficult to have a clustering approach that clearly separates individual land areas from one another. This part of the experiment is conducted with the prior knowledge of the boundaries of the land areas. The main focus of this part of the experiment is to determine the suitability of each aspect of a region as a means to find similar regions.

This part of the experiment uses Tweets, POIs and the population aspects to analyse the different measures. The Tweet aspect is used with varying similarity measures and text representation. Since the Tweets are clearly separated into individual land areas during preprocessing, spatial information is ignored. In this case, the TF-IDF is computed based on the assumption that any two regions, during the comparison, are the whole collection of documents. The experiment setup for Tweets is summarized in Table 4.2.

Options	Value
Boundaries	Subzones
Subzone	Moulmein
Similarity Measures	Cosine Similarity, Euclidean Distance, JS Divergence
Text Representation	TF-IDF, Raw Term Frequency

TABLE 4.2: Single Similarity Measure: Tweet Experiment Setup

The second series of this part of the experiment uses the distributions of the female residents. Options used for this data is shown in Table 4.3.

Options	Value
Boundaries	Subzones
Subzone	Tampines West
Similarity Measures	Cosine Similarity, Euclidean Distance, JS Divergence

TABLE 4.3: Single Similarity Measure: Female Residents Experiment Setup

The third data used in this part of the experiment is the POI distribution. Options set up for this experiment is shown in Table 4.4.

Options	Value
Boundaries	Subzones
Subzone	Tampines West
Similarity Measures	Cosine Similarity, Euclidean Distance, JS Divergence

TABLE 4.4: Single Similarity Measure: POI Experiment Setup

The results are shown in Appendix A. Each subzone boundary is shown by white dotted lines. Darker shades of green indicate significant similarity. The red coloured polygon indicates the selected area for comparison. For raw term frequency, it can be showed that most regions are significantly similar to one another when using Euclidean distance and JS divergence. However, regions are significantly similar with one another when using Cosine similarity on the TF-IDF text representation. For the population distribution, Euclidean distance identified all regions as significantly similar to the selected area. Cosine similarity identified lesser similar regions while JS divergence identified the least amount. It is observed that the female resident distributions in Tampines West and East are correlated. For the POI distribution, the Euclidean distance identified all regions as significantly similar to the selected area. Cosine similarity identifies lesser similar regions while JS divergence identifies the least amount. It is observed that the POI distributions in Tampines East and West are not correlated.

4.2.3 Double Similarity Measures

As evident from the results so far, each land area are similar to one another in a different way depending on the data used. Furthermore, it is difficult to pick the appropriate similarity measures. This part of the experiment is to explore the possibility of using two aspects and similarity measures. The experiment setup is shown in Table 4.5. Note

that for JS divergence of Cosine similarity of topic distribution, it represents the topic distribution comparison technique mentioned in Section 3.2.3.

Options	Value
Boundaries	Subzones
Subzone	Moulmein
Similarity Measures	Cosine Similarity, JS Divergence, JS Divergence of Cosine Similarity of Topic Distribution
Text Representation	TF-IDF, Term Frequency
Topic Modeling Algorithm	LDA

TABLE 4.5: Double Similarity Measure: Tweet Experiment Setup

The results are shown in Appendix A. It can be shown that varying similarity measures and α value impact the number of similar subzones significantly.

4.3 Findings and Observations

The results show that it is difficult to use clustering approaches to separate Tweets into the corresponding land area boundaries. An observable reason is that each area is relatively near to one another, making it difficult to separate the clusters.

Additionally, it is shown that the Euclidean distance is less suitable for measuring text similarity. The reason for such is the curse of dimensionality problem that exists in the bag-of-words model. A better approach will be to use the Cosine similarity.

Naturally, each region can be similar to one another on one aspect but differs in another. For example, Tampines West and East are similar based on the distributions of female residents but differs when their POI distributions are used.

Although similar regions can be identified, their true similarity is not defined. To solve this issue, the next step will be to obtain a list of perceived similar regions by users.

4.4 Final Prototype Features

Previous sections have presented the various ways to visualize different aspects of a region. The following subsections will showcase the final prototype features including

the ability to let users evaluate the queries. In the final prototype, the visualization is split to two major views, analysis and similarity. The idea behind is that users tend to first perform analysis on the different aspects of a region. After which, users may find some interesting patterns and decide to query for more similar regions that exhibit identical behaviour.

4.4.1 Region Analysis

The region analysis part of URSAS is meant to provide users with all the different useful views and aggregated aspects of each region. Specifically, it is meant to showcase the different indicators and measures that users may use to observe interesting patterns. Figure 4.9 shows the Tweet analysis overview. Each section of the screen is labelled A to F. Each label is further described as follows.

- **A** - this is the navigation menu that allows users to choose between different boundaries data and aspects. Additionally, the menu allows users to switch between analysis and similarity views. Options highlighted in white represent the current choices.
- **B** - this is the same map that was previously shown in Figure 4.5 with the colours representing the density of the selected aspect.
- **C** - this line graph changes according to the aspect selected. In Tweet analysis overview, it shows the hourly distribution of Tweets for the selected zones.
- **D** - this radar chart shows the distribution of Tweets in the different time period for the selected zones. The time periods are morning, afternoon, evening and night.
- **E** - these two boxes represent the selected zones. The bottom drop-down menu allows users to switch between showing term or Tweet hourly distribution.
- **F** - this sidebar changes according to the aspect selected. In Tweet analysis, this part shows the word cloud, topic distribution and the Tweets in the selected zone.

FIGURE 4.9: Region Analysis: Tweet Analysis Overview

The POI analysis overview, as shown in Figure 4.10, is similar to the Tweet counterpart with the exception of section G, H and I.

- **G** - instead of a line graph, this section shows a bar chart that represents the POI category distribution of the selected zones
- **H** - this radar graph shows each aspect and their proportionality to other zones on the map for the selected zones
- **I** - the sidebar shows the breakdown of POIs for different categories for the selected zone

FIGURE 4.10: Region Analysis: POI Analysis Overview

The population analysis overview page is as shown in Figure 4.11. In this page, the charts and sidebar reflect the information for the resident aspect. These charts and sidebar are labelled J, K and L. Two of the key indicators of effective planning in each zone are the ratios of educational facilities to teens and the seniors to healthcare facilities in the selected zone. Having a high ratio indicates an insufficient number of the facilities and that city planner should allocate more development efforts to that zone. The teens to senior and adult to senior ratios can indicate whether there will be sufficient working adults to support the retired household members. Low ratio can indicate heavier financial burden on each working household member.

- **J** - this line graph shows the age distribution of the selected zones. It is also able to display only the male or female residents if users select the option from the drop-down menu.
- **K** - this radar graph shows the teens, adult and senior breakdown of the population
- **L** - the sidebar shows five important information of the population in the selected zone. The first information is the total number of female and male teens, adults and seniors. The second information is the ratio of educational facilities and teens in the selected zone. This is followed by the ratio of healthcare facilities and seniors. The last two information is the ratio of teens, adults and senior in the selected zone.

FIGURE 4.11: Region Analysis: Population Analysis Overview

4.4.2 Region Similarity

In order to evaluate different similarity measures, URSAS allows users to annotate the results with a score of 1 to 5 with 1 being less relevant and 5 being the most relevant result to what the user is looking for. Note that users need not give a score to all results. They only need to annotate those that they think is relevant. Results that are not annotated by users have a default value of 0. A value of 0 indicates zero relevance. This feature is shown in Figure 4.12. The screen is similar to what is shown in previous Subsection 4.4.1 with the exception of sections M and N. It has also been mentioned in previous Subsection 3.2.3 that similar region can be queried based on multiple aspects with different weights.

- **M** - this shows the results of the query for the requested similarity measure on the selected zone. It also allows users to select the score for different relevancy type.
- **N** - this shows the three inputs that users can customize in their queries. It allows users to select different aspects, measures and the weights assigned to it for the selected zone.

Since individual users may perceive similar regions differently, each score is tied to a relevancy type as shown in section M of Figure 4.12. The intuition behind is based on

FIGURE 4.12: Region Similarity: User Evaluation

the functional zones in each region. This concept of functional zones will be further described in Subsection 4.4.4. The following list describes what each relevancy score represents.

- **Residential Relevancy** - this represents how relevant the result is to the selected zone if both the results and the selected zone are compared by the residential functions like housing options
- **Commercial Relevancy** - this is based on the commercial functions like shopping malls
- **Business Relevancy** - this is based on the business functions like office spaces.
- **Students Relevancy** - this is based on how relevant the results are to students. One example is if both the selected and resulting zones are universities, the score given should be 5. Alternatively, it can refer to regions that are frequented by students. In this case, the score can be 5 for users who want to look for such regions.
- **Working Adults Relevancy** - this is based on working adults. One example is lunch time hot-spots or after work leisure areas.
- **Unemployed Relevancy** - this is based on unemployed or retired members of the public. One example is elders home.

(a) Evaluation Result Page 1

Precision P						
Name	Residential	Commercial	Business	Student	Working Adult	Unemployed
KAKI BUKIT	0.05	0.00	0.00	0.05	0.05	0.05

nDCG Q						
Name	Residential	Commercial	Business	Student	Working Adult	Unemployed
KAKI BUKIT	0.15	0.00	0.00	0.15	0.15	0.15

User Ranking List S			Top 20 Results R		
No	Name	Relevance Score	No	Name	Similarity
1	UPPER PAYA LEBAR	5	1	YUNNAN	97.05%
2	TAMPINES EAST	5	2	WOODLANDS EAST	96.97%
3	KHATIB	5	3	MIDVIEW	96.84%
4	HOUGANG EAST	4	4	WOODGROVE	96.57%
5	YUHUA WEST	3	5	WOODLANDS SOUTH	96.37%
6	KAMPONG UBI	3	6	PASIR RIS DRIVE	95.81%

(b) Evaluation Result Page 2

FIGURE 4.13: Region Similarity: Evaluation Result

Once the users are satisfied with their scores, they can view their evaluation results as shown in Figure 4.13. This part has five sections labelled in O, P, Q, R and S. Since users evaluate the results themselves, it is not possible to calculate recall as there is insufficient information on the total number of true relevant zones.

- **O** - this shows the precision and nDCG values of the input and measures used for the query. It also shows a vertical line chart to show the average precision and

nDCG values. The average for both values is obtained by summing the respective value for the different relevancy types and divide by the number of relevancy types.

- **P** - this shows the precision values for the query based on different relevancy type
- **Q** - this shows the nDCG values for the query based on different relevancy type
- **R** - this shows the top 20 results from the query
- **S** - this shows the user ranking list that is obtained from the user evaluation screen.

It is worth noting that the number of results can be changed to improve the evaluation values. However, it is set at 20 to provide a consistent evaluation result.

4.4.3 Multiple Input Sources

Users can choose to query for similar urban regions based on Tweets, POIs and the population data. For example, a user can choose to query for zones based on the topic distribution of Tweets using Cosine similarity, the age distribution of the residents using JS divergence and the category distribution of the POIs using Euclidean distance. For each of the aspects, there are various ways of representing the data depending on the needs of the user. The measures that users can select to compare each of the aspects are Cosine similarity, KL divergence, JS divergence, Euclidean and Manhattan distance.

For Tweets, users can choose 1 of the 10 possible ways to represent the data. Previous Sections 2.3.2 has elaborated on representing Tweet by term frequency, document frequency, inverse document frequency, TF-IDF and topic distribution. On top of these approaches, URSAS also provides other forms representations. The following list describes the additional representations.

- **Tweet Hourly Distribution** - this compares the hourly distribution of Tweets. It is the distribution that shows, for each hour in a day, the number of the Tweets posted for that hour. If no Tweet is posted in that hour, it is set to 0.
- **Term Hourly Distribution** - this compares the hourly distribution of the top 40 terms with other zones. It is the distribution that shows, for each hour in a day, the number of times a particular term is mentioned in that hour. If there is no mention, it is set to 0.

- **Tweet Burst Period** - this compares the Tweet burst period with other zones. It is the distribution that shows, for morning, afternoon, evening and night, the number of Tweets posted in that period. It is set to 0 if no Tweet is posted.
- **Term Burst Period** - this compares the term burst period with other zones. It is the distribution that shows, for morning, afternoon, evening and night, the number of times a particular term is mentioned in that period. It is set to 0 if there is no mention.
- **Similar Topic Distribution** - this compares the topic distribution with other zones. However, this implementation is based on the most similar terms as elaborated in Section 3.2.3.

For POIs, it is straightforward. Users can choose between category or subcategory distribution. For population data, users can choose 1 of the 6 possible ways. Previous Section 3.3.2 has described the overall, male and female population distribution by age. On top of these representations, URSAS provides 3 additional ways as follows.

- **Age Spread** - this compares the age spread with other zones. It is the distribution that shows, for teen, adult and senior, the number of residents that belong to that group in that zone. It is set to zero if there is no resident.
- **Male Age Spread** - this compares the male age spread with other zones. It is similar to age spread but only for male residents.
- **Female Age Spread** - this compares the female age spread with other zones. It is similar to age spread but only for female residents.

4.4.4 Functional Region

The functional zone represents the specific land use plan for each plot of land. However, functional zones are not available digitally as mentioned earlier in Section 3.3.1. As such, functional zones are manually replicated using URSAS. Figure 4.14 shows the analysis screen for functional zones. This screen has two sections labelled T and U.

- **T** - this shows two drop-down menus that allow users to view different types of functional zones

- U - this map shows the functional zones

The features for the functional zones analysis and similarity view are the same as the region, planning area and subzone counterparts. Users can view each functional zone individually. Similarly, users can perform similar functional zone queries.

FIGURE 4.14: Functional Region Analysis

4.5 Summary

In summary, this chapter has described the implementation URSAS as well as some preliminary experiments performed. It is clear that the resulting clusters are unable to detect the separation of area boundaries in the provided data. Additionally, this chapter has presented the final version of the URSAS that is used for the subsequent chapters. The following Chapter 5 will present the analysis performed on urban and social media data.

Chapter 5

Urban and Social Media Data Analysis

This chapter describes the in-depth analysis performed to explore how each aspect of a region is related with one another. Specifically, this chapter will look at how the three aspects, Tweet, population and POI, are related with one another. The purpose of this data analysis is to explore how to improve the results of finding similar regions. For the following analysis, the correlation values are calculated for each possible combination of the data. However, only those that contain valuable information for finding similar regions are presented in this chapter.

5.1 Different Views of Data

Before performing analysis on each of aspect of a region, it is worth noting that they can be viewed and aggregated differently. The following subsections describe how each of the aspects is viewed and aggregated in this work.

5.1.1 Tweet

Each Tweet contains three key information that is useful for analysis. Specifically, the location, time posted and text. This three information is helpful in gathering the location of certain topics as well as the period it is most relevant.

A first example is when a user might want to know the time period in which there are most Tweets about shopping malls. A straightforward approach is to gather all the Tweets that contain the terms and topics about shopping malls. Subsequently, each Tweet is assigned to the hour of the day based on the time posted.

Alternatively, some users might be interested in the time period that has the most Tweets regardless of topic. In this case, the approach is to calculate the sum of the number of Tweets posted at each hour of a day.

Thirdly, another group of users might want to know the most frequent terms that are mentioned in Tweets. This is achieved by returning the highest count in the term frequency matrix.

Lastly, some users might be interested to know where the Tweets and certain terms are most often mentioned. In this case, the output can be the area that contains a group of Tweets that have the terms that users are interested in.

5.1.2 Population Data

The population data are available for each subzone and planning area. It contains the number, age and gender of the residents in each area. For this work, the male and female residents are combined to obtain the overall resident distribution by age. This is to cater to cases where users might just want to know the total population in certain areas. The age group of the residents forms the population distribution in each region.

5.1.3 POI Data

The POI data contains two key attributes that are beneficial for this analysis. These are the location and the category of the POIs. Using the location, each of the POI can be assigned to a region. The different categories of the various POIs in each region forms the category distribution.

5.2 Urban Data Analysis

The goal of having a self-sufficient neighbourhood is to allow residents to have access to much-needed facilities within their neighbourhood. Intuitively, the number and type of facilities in a neighbourhood should be dependent on the type of residents. For example, a neighbourhood with a high number of young residents should have sufficient educational facilities. On the other hand, elder residents are likely to be in need of healthcare facilities.

However, it is worth noting that not all POIs are related to the type of residents. For instance, government and military facilities are often developed in a region due to some strategic objectives. Since these strategic objectives are mostly immeasurable, it is difficult to find the relationship between these facilities and the population.

The aim of the following subsections is to explore and discover the underlying relation between urban data. Following points outlines the analysis to be performed in the subsequent subsections.

- **Analysis 1** - discover the types of POI that are related to certain type of residents in a region
- **Analysis 2** - discover if there is any relation between certain age group of the population and the POIs in the region
- **Analysis 3** - discover the relative proportionality of POIs, Tweets and residents

5.2.1 POIs Dependency on Residents

This subsection will describe the first analysis that tries to find the POIs that are closely related to certain type of residents. To do so, the Pearson correlation coefficient is used. The formula for the Pearson correlation coefficient can be found in Subsection 3.2.1. This analysis is performed on the three boundaries data that divided Singapore into regions, planning areas and subzones. Two distributions, **A** and **B**, are generated. Distribution **A** = $a_{1i}, a_{2i}, a_{3i}, \dots, a_{ni}$ where a_{1i} is the number of a category, i , of POIs in the first region and n is the total number of regions. **B** = $b_{1j}, b_{2j}, b_{3j}, \dots, b_{nj}$ where b_{1j} is the number

of residents for a specific age group, j , in a region and n is the total number of regions. $\rho_{A,B}$ is the correlation between **A** and **B**.

There are two important findings in this analysis. Firstly, when the region boundary data were used, the correlation value between POIs and the residents were consistently high as evident from Figure 5.1. However, when subzone boundary data is used, the correlation values were more distinguishable as evident from Figure 5.2. The explanation for such behaviour is that there were only 5 observations from the region boundary but there are 323 observations from the subzone boundary data. Having more observations is more desirable for more reliable correlation values. For this reason, only the subzone boundary data are used for subsequent analysis.

FIGURE 5.1: Correlation of Healthcare Facilities and Age Group Residents in Regions

The second finding of this analysis is that healthcare and educational facilities were correlated with a specific age group of residents. Figure 5.2 shows the correlation values between healthcare facilities and the residents increases as the age group increases. The highest correlation value is observed at age group 65 - 69. This means that there are a higher number of healthcare facilities in areas where there are more elder residents. To support the fact that not all POIs are highly correlated with the residents, Figure 5.2 shows that the number of eateries in subzones does not have an increasing correlation value. Although the value is positive, it has relatively low correlation with the population.

On the other hand, elderlies were less correlated with educational facilities. Figure 5.2 shows that the correlation values increases as age group increases. However, the correlation values start to decreases from age group 45 - 49. The highest correlation was observed at age group 40 - 44. This means that for areas with a higher number of younger residents, there are a higher number of educational facilities.

FIGURE 5.2: Correlation of POIs and Age Group Residents in Subzones

5.2.2 Residents Dependency on POIs

This subsection will describe the analysis on whether there are specific age group of residents that affects a specific type of POI in a subzone. It is known from previous Subsection 5.2.1 that subzones with younger residents tend to have more educational facilities. Table 5.1 shows the type of residents and POI that has the correlation value of more than 0.91. The result shows that the number of female residents aged between 40 - 44 has the highest positive correlation with the educational facilities in subzones. However, this does not indicate female residents are the main factor for the number of educational facilities in a subzone. Instead, these values can only be used as an indication of the ratio of residents and number of educational facilities. This analysis shows that there is no significant evidence to support that a specific POI is dependent on a specific type of residents.

Age Group	POI Type	Correlation
40 - 44Female	Education Facilities	0.918201918
45 - 49Male	Education Facilities	0.917675809
10 - 14Male	Education Facilities	0.91420247
10 - 14Female	Education Facilities	0.913536415
40 - 44Male	Education Facilities	0.912847152
45 - 49Female	Education Facilities	0.912066306
25 - 29Female	Education Facilities	0.911765873

TABLE 5.1: Highest Correlation Value

5.2.3 Volume Dependency of Tweet, POI and Resident

This subsection will look at whether the volume of each aspect is proportional to one another in each subzone. The hypothesis here is that the volume of Tweet should be highest where most of the residents are. The aim is to investigate whether the fraction of Tweets is proportional to the fraction of the residents in each subzone. For example, suppose a subzone has $\frac{1}{4}$ of Singapore population, it should also have $\frac{1}{4}$ of the Tweets generated in Singapore. Table 5.2 shows 6 of the subzones and their respective proportionality values. Each value in the table represents the fraction of the individual aspects over their total number. For example, the Tweet column shows the fraction of the number of Tweets appearing in that subzone over the total number of Tweets gathered in Singapore.

Subzone	Tweet	Resident	POI
TAMPINES EAST	0.033404299	0.035327897	0.028241821
WOODLANDS EAST	0.018543894	0.024498691	0.014688268
BEDOK NORTH	0.018213579	0.022781614	0.019416252
YUNNAN	0.018186384	0.018552181	0.013616592
TAMPINES WEST	0.017486648	0.019788993	0.014057871
YISHUN WEST	0.016996138	0.01509737	0.00844733
SENTOSA	0.012683581	0.00036149	0.0027107

TABLE 5.2: Proportionality of Tweet, POI and Resident in Subzones

The first finding of this analysis is that the number of Tweets generated in each subzone is not proportional to the number of residents in that area. A possible explanation for this behaviour is the number of visitors that upload their Tweets in that subzone. Evidently, from the subzone SENTOSA, there are a higher number of Tweets generated from that area as compared to the number of residents. The reason is that SENTOSA is a leisure destination with a large number of visitors. If there are visitors data available for each subzone, it could show a positive correlation with the number of Tweets in that area.

The second finding of this analysis is that the number of residents has a slight impact on the POIs in a subzone. This is evident from the values in Table 5.2 that as the number of residents decreases, the amount of POIs decreases as well. This supports the first analysis in Subsection 5.2.1 that shows educational and healthcare facilities having a high and positive correlation with the residents.

5.3 Social Media Data Analysis

This section will look at how Tweet relates to the population and POI in a subzone. It is known from the previous analysis in Subsection 5.2.3 that the volume of Tweet is not directly proportional to the number of residents. Although most of the volume of Tweet could be contributed by the visitors in a subzone, the residents play a part as well. This section aims to explore the relation between Tweet and urban data. Following points outlines the analysis to be performed in the subsequent subsections.

- **Analysis 1** - discover the age group that is correlated with the time in which the Tweets are posted
- **Analysis 2** - discover the age group that is correlated with a specific term in Tweet
- **Analysis 3** - discover any relation between Tweet and POI

5.3.1 Dependency of Tweet on Residents

This subsection will discuss the analysis on discovering the time period in which is correlated with a specific age group of residents. Three age groups, 15 - 19, 30 - 34 and above 85, were compared with the hourly distribution of Tweet. This is based on the hypothesis that teens are more likely to post Tweets during wee hours than adults. To perform this analysis, the distribution $\mathbf{C}$ is generated where $\mathbf{C} = c_1^k, c_2^k, c_3^k, \dots, c_n^k$ and c_1^k represents the number of Tweets that were posted for a specific hour in a day, k , in a subzone and n is the total number of subzones. Distribution $\mathbf{C}$ were compare with distribution $\mathbf{B} = b_{1j}, b_{2j}, b_{3j}, \dots, b_{nj}$ where b_{1j} is the number of residents for a specific age group, j , in a subzone and n is the total number of subzones. Figure 5.3 shows the correlation value between the hours in which the Tweets are posted for the 3 age groups.

It is evident from the graph that all three age groups have a high correlation with Tweet. Residents aged between 15 - 19 have higher correlation than those aged 30 - 34 and above 85. However, there does not seem to have any patterns in the correlation for the different age groups. All three age groups display similar behaviour.

FIGURE 5.3: Correlation of Hourly Distribution of Tweet and Age Group Resident in Subzones

5.3.2 Dependency of Tweet Term on Residents

This subsection describes the analysis of finding the correlation values between a Tweet term and an age group of residents. This is based on the idea that specific age group of users will use specific terms in Tweet. Two terms, "orchard" and "shopping", were compared with three different age groups, 15 - 19, 30 - 34 and above 85. To achieve this analysis, a new distribution, $\mathbf{D}$, is generated where $\mathbf{D} = d_{1t}^k, d_{2t}^k, d_{3t}^k, \dots, d_{nt}^k$ and d_{1t}^k represents the number of times a term, t , was posted for a specific hour in a day, k , in a subzone and n is the total number of subzones. Distribution $\mathbf{D}$ are compared with $\mathbf{B} = b_{1j}, b_{2j}, b_{3j}, \dots, b_{nj}$ where b_{1j} is the number of residents for a specific age group, j , in a subzone and n is the total number of subzones.

FIGURE 5.4: Correlation of the Hourly Distribution of the Term "orchard" and Age Group Residents in Subzones

It is evident from Figures 5.4 and 5.5 that the terms "orchard" and "shopping" has a low correlation with the 3 age groups. Most of these correlation values were negative.

FIGURE 5.5: Correlation of the Hourly Distribution of the Term "shopping" and Age Group Residents in Subzones

This analysis also found that there is no relation between a certain term in Tweet and a specific age group.

5.3.3 Dependency of Tweet on POIs

This subsection describes the analysis of finding the correlation between the volume of the Tweet posted and the POIs in a subzone. Secondly, specific terms like "orchard" and "shopping" are also compared with the POIs. For this analysis, three distributions, **A**, **C** and **D** were used. Distribution **A** = $a_{1i}, a_{2i}, a_{3i}, \dots, a_{ni}$ is generated where a_{1i} is the number of facilities from a category, i , of POI in the first subzone and n is the total number of subzones. Distribution **C** = $c_1^k, c_2^k, c_3^k, \dots, c_n^k$ is generated where c_1^k represents the number of Tweets that was posted for a specific hour in a day, k , in a subzone and n is the total number of subzones. Distribution **D** = $d_{1t}^k, d_{2t}^k, d_{3t}^k, \dots, d_{nt}^k$ is generated where d_{1t}^k represents the number of times a term, t , was posted for a specific hour in a day, k , in a subzone and n is the total number of subzones.

Many of the Tweets gathered are often check-ins which specify the current location of the users. Based on this observation, this analysis aims to find out if the number of Tweets generated is dependent on the types of POI in the area. Additionally, certain terms in Tweet are likely to be related to the POIs in the area. This analysis shows the correlation between four types of POIs and Tweets. These POIs are shopping malls, eateries, educational and healthcare facilities.

FIGURE 5.6: Correlation of the Hourly Distribution of Tweet and POI in Subzones

As evident from Figure 5.6, there were higher correlation values for subzones with healthcare and educational facilities. Eateries and shopping malls tend to have a lesser effect on the number of Tweets.

From Figure 5.7, the term "orchard" has a positive correlation with shopping malls. Between 10:00 am and 23:59 pm, there was a positive correlation with healthcare facilities as well. This can have two possible implications. First, there were many shopping malls in Orchard, a leisure destination in Singapore, as supported by the first observation. Secondly, it may be because there were more mentions of "orchard" after 10:00 am that resulted in the correlation with healthcare facilities.

The term "shopping" has a relatively lower correlation with all four types of POIs in Figure 5.8. However, among the four types of POIs, shopping malls have the highest correlation. Since the term "shopping" is often mentioned with shopping mall check-ins, this result is intuitive.

FIGURE 5.7: Correlation of the Hourly Distribution of Term "orchard" and POIs in Subzones

FIGURE 5.8: Correlation of the Hourly Distribution of Term "shopping" and POIs in Subzones

5.4 Summary

In summary, through the various analysis performed in this chapter, it is clear that the residents have a certain level impact on the POIs in an area. These POIs also affects the Tweets that are posted in that area. However, the residents have little effect on the content and density of the Tweets. Furthermore, there does not seem to be any relation between specific terms mentioned in Tweets and the POIs.

Chapter 6

Similar Regions Query

From previous Chapter 5, it is clear that POI and population data have a high correlation. However, Tweet seems to have low correlation with both data. This chapter describes a series of experiments performed to investigate the effectiveness of URSAS in finding similar urban regions. It is found that POI and population are both essential data to find similar regions. Additionally, it is found that Tweet can improve the results of similar region queries, in terms of precision and nDCG, when combined with POI and population data. This result supports the findings from Chapter 5.

6.1 Objective

The objectives of this chapter are listed below. In short, the purpose of this chapter is to explore the best way to discover similar regions.

- **Objective 1** - to investigate the effectiveness of individual aspects of a region in querying for similar regions
- **Objective 2** - to investigate the effectiveness of the different measures in querying for similar regions
- **Objective 3** - to investigate the effectiveness of combining multiple aspects and measures to query for similar regions

6.2 Experiment Setup

6.2.1 Dataset

To evaluate different aspects and measures, only planning area and subzone boundaries are used. The reason for omitting the region boundary data is due to the fact each of those covers a larger land area. Planning area and subzone boundaries cover a smaller land area are, therefore, more specialized. For functional zones, it is not complete as it only covers the central region. As such, functional zones are not used.

In total, 55 planning areas, 323 subzones, more than 5 million Tweets, 15,864 POIs and 378 sets of population data that are used for this experiment.

For evaluation, 10 different users from different disciplines, such as business owners and city planners, were asked to give ranking scores to the results of the queries.

6.2.2 Preliminaries

To obtain the ranking of the results from the users, the similarity for each type of aspects and measures for each planning area and subzone is computed. The product is that each planning area and subzone has a list of top similar zones according to the measure and aspects used. In total there are 4,950 sets of top similar zones for planning areas and 29,070 sets for subzones computed. This is including the different types of measures and representations. The baseline measure is set as Euclidean distance.

For the similar topic distribution, it computed based on the Cosine similarity between them. One example is that for a given topic in the selected zone, pick the most similar topic to compare on the compared zone based on Cosine similarity.

For the TF-IDF calculation, term frequency and inverse document frequency smooth are used.

Each user is then asked to evaluate the results presented to them. After each user has given the ranking to the results that they feel is relevant, the list is aggregated to obtain the top similar neighbours for each subzone.

Two values are used to evaluate the performance of the queries. The precision and nDCG values are calculated from the average of these results. The recall is not used in this case since the total number of true relevant zones is not available.

It is worth noting here that the objective of this experiment is not to obtain average precision and nDCG values close to 1. Instead, the goal is to observe if there is an increase in average precision and nDCG as more aspects are used. Additionally, it is in the interest of this work to investigate which measure is most suitable to get the good precision and nDCG values.

6.3 Single Aspect and Measures

This section will describe the result of the experiment performed using single aspect of a region. The similarity between any two regions, R_a and R_b where R_{ai} and R_{bi} refer to their i^{th} aspect, is as follows.

$$Sim(R_a, R_b) = Sim(R_{ai}, R_{bi}) \quad (6.1)$$

6.3.1 Results

Tables 6.1 and 6.2 shows the average precision and nDCG of each aspect on querying for similar regions in the planning area and the subzone. The highest values are in bold. It is the average precision and nDCG of all measure and representation types for each aspect. The list of some of the highest precision and nDCG values for each measure and representation type can be found in Appendix B.

Aspect	Avg Precision	Avg nDCG
Tweet	0.166	0.096
POI	0.198	0.102
Population	0.233	0.105

TABLE 6.1: Planning Area: Query Aspects Performance

Tables 6.3 and 6.4 show the performance of each measure. The highest values are in bold.

Aspect	Avg Precision	Avg nDCG
Tweet	0.026	0.041
POI	0.040	0.054
Population	0.144	0.079

TABLE 6.2: Subzone: Query Aspects Performance

Measure	Avg Precision	Avg nDCG
Cosine Similarity	0.283	0.131
Euclidean Distance (baseline)	0.263	0.130
JS Divergence	0.054	0.049
KL Divergence	0.088	0.063
Manhattan Distance	0.271	0.125

TABLE 6.3: Planning Area: Measures Performance

Measure	Avg Precision	Avg nDCG
Cosine Similarity	0.110	0.083
Euclidean Distance (baseline)	0.096	0.074
JS Divergence	0.009	0.017
KL Divergence	0.022	0.023
Manhattan Distance	0.097	0.077

TABLE 6.4: Subzone: Measures Performance

6.3.2 Findings and Observations

There are five key findings and observations from this part of the experiment. Firstly, it is clear that the population data is the most effective at identifying similar regions. This is followed by POI and Tweet.

Secondly, for Tweet, the document frequency, Tweet hourly distribution, inverse document frequency, term burst period distribution, TF-IDF and term frequency representations perform similarly. Topic distribution performs poorly. However, it is evident that by considering similar topics together will improve the results. In addition, it has been shown by [13] that the temporal information can help in finding similar Tweets. As such, the term hourly distribution and similar topic distribution will be tested in the following Section 6.4.

Thirdly, for POI data, regardless of whether it uses a default or category distribution, the performance is the same. As such, only the category distribution will be tested in the following Section 6.4.

Next, the overall and female age distribution in the population data perform identically. In addition, the overall age distribution performs better than other representation types. As such, only the overall age distribution will be tested in the following Section 6.4.

Lastly, Cosine similarity is the overall best measure for finding similar regions. Another measure that does better than the baseline Euclidean distance is the Manhattan distance. Both JS and KL divergence perform undesirably. As such, JS and KL divergence will be excluded from the following Section 6.4.

6.4 Double Aspects and Measures

Previous Section 6.3 discussed the query for similar regions based on single aspect and measure. This section describes the experiment performed based on two aspects of a region. The similarity between any two regions, R_a and R_b where R_{ai} and R_{bi} refer to aspect their i^{th} aspect and R_{aj} and R_{bj} refer to their j^{th} aspect, is as follows. The weight w_i is a user defined parameter for the i^{th} aspect.

$$Sim(R_a, R_b) = (w_i)Sim(R_{ai}, R_{bi}) + (1 - w_i)Sim(R_{aj}, R_{bj}) \quad (6.2)$$

6.4.1 Results

Tables 6.5 and 6.6 show the results for different aspects on planning area and subzone. The list of some of the highest precision and nDCG values can be found in Appendix C.

Aspect	Avg Precision	Avg nDCG
POI,Population	0.318	0.143
Tweet,POI	0.259	0.126
Tweet,Population	0.299	0.135

TABLE 6.5: Planning Area: Aspects Performance

Aspect	Avg Precision	Avg nDCG
POI,Population	0.136	0.100
Tweet,POI	0.055	0.074
Tweet,Population	0.127	0.097

TABLE 6.6: Subzone: Aspects Performance

Although Tweet is not part of the best-performing aspect combination, it is necessary to identify the best way to represent the data. Tables 6.7 and 6.8 show the results for different Tweet representations when combined with the POI or population data.

Representation	Avg Precision	Avg nDCG
Similar Topic Distribution, Age Distribution	0.292	0.133
Similar Topic Distribution, Category Distribution	0.245	0.123
Term Hourly Distribution, Age Distribution	0.306	0.138
Term Hourly Distribution, Category Distribution	0.272	0.129

TABLE 6.7: Planning Area: Tweet Representation Performance

Representation	Avg Precision	Avg nDCG
Similar Topic Distribution, Age Distribution	0.113	0.101
Similar Topic Distribution, Category Distribution	0.053	0.076
Term Hourly Distribution, Age Distribution	0.141	0.093
Term Hourly Distribution, Category Distribution	0.057	0.072

TABLE 6.8: Subzone: Tweet Representation Performance

Since the best-performing combination is the POI and population data, it is essential to know the best measure for the combination. Tables 6.9 and 6.10 show the results for different measures used for the POI and population data combination.

Measure	Avg Precision	Avg nDCG
Cosine Similarity, Cosine Similarity	0.298	0.139
Cosine Similarity, Euclidean Distance	0.324	0.148
Cosine Similarity, Manhattan Distance	0.329	0.146
Euclidean Distance, Cosine Similarity	0.303	0.138
Euclidean Distance, Euclidean Distance	0.329	0.144
Euclidean Distance, Manhattan Distance	0.333	0.143
Manhattan Distance, Manhattan Distance	0.324	0.150
Manhattan Distance, Euclidean Distance	0.320	0.149
Manhattan Distance, Cosine Similarity	0.297	0.134

TABLE 6.9: Planning Area: Measure Performance

6.4.2 Findings and Observations

Four key findings were made in this part of the experiment. Firstly, the results showed that the best aspect combination is the POI and population data. Additionally, it is found that combining two aspects can improve the average precision and nDCG values as compared to just using a single aspect.

Measure	Avg Precision	Avg nDCG
Cosine Similarity, Cosine Similarity	0.112	0.100
Cosine Similarity, Euclidean Distance	0.147	0.108
Cosine Similarity, Manhattan Distance	0.157	0.109
Euclidean Distance, Cosine Similarity	0.114	0.091
Euclidean Distance, Euclidean Distance	0.150	0.099
Euclidean Distance, Manhattan Distance	0.159	0.101
Manhattan Distance, Cosine Similarity	0.106	0.090
Manhattan Distance, Euclidean Distance	0.136	0.098
Manhattan Distance, Manhattan Distance	0.147	0.100

TABLE 6.10: Subzone: Measure Performance

Secondly, it is found that the best representation for Tweet is term hourly distribution. Similar topic distribution follows closely with term hourly distribution but does not perform equally well. As such, similar topic distribution will be excluded from the next Section 6.5.

Thirdly, the best performing measure, based on average precision, is using Euclidean distance on POI category distribution and Manhattan distance on the population age distribution. However, the best measure, based on average nDCG, is using Manhattan distance on both POI and population distance in the planning area. In subzone, the best measure, based on nDCG, is using Cosine similarity on the POI category distribution and Manhattan distance on the population age distribution. Additionally, it is also worth noting that the best measure for Tweet, based on average nDCG and precision for term hourly distribution, is Cosine similarity.

Lastly, the best weight observed from the list of precision and nDCG values in Appendix C for planning areas and subzones is $w_i = 0.1$.

6.5 Triple Aspects and Measures

In previous Section 6.4, the performance of combining two aspects are discussed. This section will discuss the performance of using all three aspects. In this part of the experiment, the similarity between any two regions, R_a and R_b where R_{ai} and R_{bi} refer to aspect their i^{th} aspect, R_{aj} and R_{bj} refer to their j^{th} aspect and R_{ak} and R_{bk} refer to their k^{th} aspect, is as follows. The weights w_i and w_j are user defined parameters for the i^{th} and k^{th} aspect respectively.

$$Sim(R_a, R_b) = (w_i)Sim(R_{ai}, R_{bi}) + (w_j)Sim(R_{aj}, R_{bj}) + (1 - w_i - w_j)Sim(R_{ak}, R_{bk}) \quad (6.3)$$

It is evident that the best w_i value from previous Section 6.4 is 0.1. In this case, the range of w_j must be between 0.1 and 0.8. If set to 0 or 0.9, the result is the same as using only two aspects. The first two aspects are fixed at using the Euclidean distance for POI category distribution and the Manhattan distance for population age distribution as these were the best combination from the previous experiment. Tweets will be the third aspect used and will be fixed at using the Cosine similarity on the term hourly distribution. The Euclidean distance will also be tested on Tweet as the baseline.

6.5.1 Results

In order to investigate whether adding Tweet will yield better results, the values from previous two experiment are included. Tables 6.11 and 6.12 show the performance of combining different aspects of a region. The highest values are in bold.

Aspects	Avg Precision	Avg nDCG
Tweet	0.166	0.096
POIs	0.198	0.102
Population	0.233	0.105
POI,Population	0.318	0.143
Tweet,POI	0.259	0.126
Tweet,Population	0.299	0.135
Tweet,POI,Population	0.320	0.143

TABLE 6.11: Planning Area: Overall

Aspects	Avg Precision	Avg nDCG
Tweet	0.026	0.041
POIs	0.040	0.054
Population	0.144	0.079
POI,Population	0.136	0.100
Tweet,POI	0.055	0.074
Tweet,Population	0.127	0.097
Tweet,POI,Population	0.155	0.098

TABLE 6.12: Subzone: Overall

Tables 6.13 and 6.14 show the precision and nDCG values for varying w_j . The highest values are in bold.

Measure	w_j	Precision	nDCG
Euclidean;Manhattan;Cosine	0.1	0.27	0.13
Euclidean;Manhattan;Cosine	0.2	0.29	0.14
Euclidean;Manhattan;Cosine	0.3	0.31	0.14
Euclidean;Manhattan;Cosine	0.4	0.32	0.14
Euclidean;Manhattan;Cosine	0.5	0.33	0.15
Euclidean;Manhattan;Cosine	0.6	0.34	0.15
Euclidean;Manhattan;Cosine	0.7	0.35	0.15
Euclidean;Manhattan;Cosine	0.8	0.36	0.15
Euclidean;Manhattan;Euclidean	0.1	0.27	0.13
Euclidean;Manhattan;Euclidean	0.2	0.29	0.14
Euclidean;Manhattan;Euclidean	0.3	0.3	0.14
Euclidean;Manhattan;Euclidean	0.4	0.31	0.14
Euclidean;Manhattan;Euclidean	0.5	0.33	0.15
Euclidean;Manhattan;Euclidean	0.6	0.34	0.15
Euclidean;Manhattan;Euclidean	0.7	0.35	0.15
Euclidean;Manhattan;Euclidean	0.8	0.36	0.14

TABLE 6.13: Planning Area: POI, Population and Tweet with $w_i = 0.1$

Measure	w_j	Precision	nDCG
Euclidean;Manhattan;Cosine	0.1	0.06	0.08
Euclidean;Manhattan;Cosine	0.2	0.08	0.08
Euclidean;Manhattan;Cosine	0.3	0.1	0.09
Euclidean;Manhattan;Cosine	0.4	0.12	0.09
Euclidean;Manhattan;Cosine	0.5	0.15	0.1
Euclidean;Manhattan;Cosine	0.6	0.17	0.1
Euclidean;Manhattan;Cosine	0.7	0.2	0.11
Euclidean;Manhattan;Cosine	0.8	0.23	0.11
Euclidean;Manhattan;Euclidean	0.1	0.07	0.07
Euclidean;Manhattan;Euclidean	0.2	0.11	0.09
Euclidean;Manhattan;Euclidean	0.3	0.14	0.1
Euclidean;Manhattan;Euclidean	0.4	0.17	0.1
Euclidean;Manhattan;Euclidean	0.5	0.19	0.11
Euclidean;Manhattan;Euclidean	0.6	0.21	0.11
Euclidean;Manhattan;Euclidean	0.7	0.23	0.11
Euclidean;Manhattan;Euclidean	0.8	0.25	0.11

TABLE 6.14: Subzone: POI, Population and Tweet with $w_i = 0.1$

6.5.2 Findings and Observations

There are three key observations made from this part of the experiment. Firstly, it is evident that adding more aspects of a region can improve the results. For planning area and subzone, the average precision value is highest when considering the POI, population and Tweet data. For the average nDCG value, it is highest when considering POI and population in the subzone.

Secondly, the best measure based on planning area is either using the Euclidean distance or the Cosine similarity on Tweet term hourly distribution with $w_j = 0.8$. In subzone, the best measure is using the Euclidean distance on Tweet term hourly distribution with $w_j = 0.8$. However, in subzone, the Cosine similarity on Tweet does not perform equally well.

Thirdly, it is clear that if querying for similar regions based purely on Tweet will not yield good results. However, combining it with other data seems to improve the results as compared to their single aspect counterparts.

6.6 Running Time

Aspect	Measure	Running Time (ms)
Tweet	Cosine Similarity	5913
Tweet	KL Divergence	6051
Tweet	JS Divergence	7889
Tweet	Euclidean Distance	6003
Tweet	Manhattan Distance	5976
POI	Cosine Similarity	766
POI	KL Divergence	493
POI	JS Divergence	387
POI	Euclidean Distance	383
POI	Manhattan Distance	354
Population	Cosine Similarity	478
Population	KL Divergence	339
Population	JS Divergence	311
Population	Euclidean Distance	292
Population	Manhattan Distance	279

TABLE 6.15: Average Running Time of Different Similarity Measures

Table 6.15 shows the average running time for each similarity measure on different aspect. Lowest running time for each aspect is in bold. Evidently, Cosine similarity has the lowest running time for Tweets while the lowest running time for POI and population aspects is Manhattan distance. If processing time is crucial, the Cosine similarity and Manhattan distance measure should be used.

6.7 Summary

In summary, this chapter has shown that combining multiple aspects will improve the results of similar region queries. Furthermore, the best measures are using the Euclidean distance on POI category distribution with weight $w_i = 0.1$, the Manhattan distance on population age distribution with weight $w_j = 0.8$ and the Cosine similarity on Tweet term hourly distribution with weight 0.1.

Chapter 7

Functional Zones Analysis and Recognition

Previous Chapter 5 presented the idea that POI has a positive correlation with the residents and Tweets within the same area. In Chapter 6, it is shown that combining different aspects can improve the results of similar region queries. Since the population data is not available for functional zones, it is not possible to explore the correlation between POIs and residents in this chapter. However, this chapter will explore if there is any correlation between the volume of Tweet and the POIs within each functional zone. In addition, this chapter will investigate if it is possible to recognise the function of each functional zone based on the POI, Tweet and population data.

7.1 Objective

The objectives of this chapter are listed below. Briefly, the purpose of this analysis is to test the effectiveness of the proposed similar region query framework in recognizing the function for each functional zone.

- **Objective 1** - to investigate whether the aspects of each functional zone is correlated with one another
- **Objective 2** - to investigate whether the proposed similar region query framework can recognize the function of each functional zone

7.2 Experiment Setup

In total, there are 2,555 functional zones with 21 different types of functions, 1,720,860 Tweets and 6,388 POIs in the Central region boundary used for this experiment.

For the analysis part of the experiment, three types of common functional zones were tested. These are residential, business and commercial zones. For each type of functional zone, the correlation between POI and Tweet hourly distribution is computed. The POI types tested are shopping malls, eateries, educational facilities and healthcare facilities.

Next, for the recognition part of the experiment, three types of measures were tested on the term hourly distribution of Tweet and the POI category distribution. The measures used are the Cosine similarity, Euclidean and Manhattan distance as those performed relatively well in Chapter 6. The baseline is Euclidean distance. To evaluate the performance, the accuracy of each recognition is computed. For every recognition, each zone is compared with all others. The function of the top similar neighbour of a functional zone is set as the recognised function. The accuracy is obtained by dividing the number of functional zones that have the same function as the one recognised by URSAS by the total number of functional zones.

7.3 Functional Zone Analysis

Each plot of land can be used for multiple or specialized functions. For example, a plot of land can be reserved for a multi-storey building with housing option and commercial space. On the hand, a plot of land can also be reserved for housing purposes only. The purpose of this analysis is to find out if there is certain POI that is correlated with the volume of Tweet in a specific functional zone type.

7.3.1 Residential Zones

The residential zones are mostly made up of private and public housings. However, to reduce congestion outside these residential zones, facilities like schools, clinics and eateries are often developed within the area to reduce the commute time of the residents.

As such, it is expected that the volume of Tweet in residential zones to be higher where there is a presence of certain POI.

Figure 7.1 shows the correlation for eateries, shopping malls, healthcare and educational facilities on the hourly distribution of Tweets in residential zones.

FIGURE 7.1: Correlation of the Hourly Distribution of Tweet and POI in Residential Functional Zone

7.3.2 Business Zones

The business zones are mostly made up of office spaces. In recent years, most of these business zones are being upgraded to business parks. The difference is that the facilities in these business parks are more comprehensive. For example, it is common to have childcare centres and eateries within or near office buildings.

Figure 7.2 shows the correlation for eateries, shopping malls, healthcare and educational facilities on the hourly distribution of Tweets in business zones.

FIGURE 7.2: Correlation of the Hourly Distribution of Tweet and POI in Business Functional Zone

7.3.3 Commercial Zones

Commercial zones are mostly made up of shopping malls. Most shopping malls in Singapore contain eateries. Clinics are also sometimes found in commercial zones. These clinics are usually more costly and targeted at residents with higher income.

Figure 7.3 shows the correlation for eateries, shopping malls, healthcare and educational facilities on the hourly distribution of Tweets in commercial zones.

7.3.4 Findings and Observations

Four key findings and observations are made from this analysis. Firstly, the healthcare facilities and educational facilities in residential zones have a high positive correlation with the Tweets within the area. This means that both types of POI have an impact on the volume of Tweet in residential zones. Such behaviour is supported by the fact that residential zones contain healthcare and educational facilities.

Secondly, the correlation between the four types of POI tested and the hourly distribution of Tweets in business zones are not consistent. This means that it is unclear whether

FIGURE 7.3: Correlation of the Hourly Distribution of Tweet and POI in Commercial Functional Zone

these four types of POIs have an impact on the volume of Tweet in business zones. One reason for such behaviour can be due to the fact that these POIs are uncommon in business zones.

Thirdly, the shopping malls are highly correlated with the volume of Tweet in commercial zones. This means that shopping malls may be one of the major factors for Tweets posted in commercial zones. A possible explanation is that most commercial zones contain shopping malls.

Lastly, the correlation between the POIs and the volume Tweet can be used to determine the type of facilities in the zones. It is also clear that there exists a correlation between POIs and the volume of Tweet in functional zones. Hence, it makes sense to make use of both POIs and Tweets to recognize the function of a functional zone.

7.4 Functional Zone Recognition

Intuitively, the function of each zone can be determined by the aspects. For example, a residential zone is expected to have housings whereas a commercial zone should have

shopping malls. Based on this idea, it is possible to recognise a function for a functional zone based on the POIs. Using what has been observed from previous sections, this part of the experiment will test the ability of the URSAS to recognise the function of each functional zone.

7.4.1 Results

Table 7.1 shows the accuracy of URSAS in recognizing the functions of the functional zones based on a single aspect. The best measure for POI is the Euclidean distance and Tweet is the Cosine similarity. These two measures are used to combine POI and Tweet with varying w_i . The accuracy of combining two aspects is as shown in Table 7.2.

Aspects	Measure	Accuracy
POI	Euclidean	0.303
Tweet	Euclidean	0.078
POI	Manhattan	0.299
Tweet	Manhattan	0.078
POI	Cosine	0.225
Tweet	Cosine	0.263

TABLE 7.1: Accuracy of URSAS Considering Only a Single Aspect

Aspects	Measure	Accuracy
POI, Tweet $w_i = 0.1$	Euclidean, Cosine	0.299
POI, Tweet $w_i = 0.2$	Euclidean, Cosine	0.308
POI, Tweet $w_i = 0.3$	Euclidean, Cosine	0.319
POI, Tweet $w_i = 0.4$	Euclidean, Cosine	0.322
POI, Tweet $w_i = 0.5$	Euclidean, Cosine	0.325
POI, Tweet $w_i = 0.6$	Euclidean, Cosine	0.323
POI, Tweet $w_i = 0.7$	Euclidean, Cosine	0.323
POI, Tweet $w_i = 0.8$	Euclidean, Cosine	0.319
POI, Tweet $w_i = 0.9$	Euclidean, Cosine	0.314

TABLE 7.2: Accuracy of URSAS with POI and Tweet

7.4.2 Findings and Observations

From the results, it is evident that combining POI and Tweet will improve the accuracy of the results. In particular, when w_i is set to 0.5, the accuracy is highest.

7.5 Summary

In summary, this chapter has shown that certain POI has a positive correlation with the hourly distribution of Tweets in certain types of functional zones. This work has shown that educational facilities and the hourly distribution of Tweet in residential zones have a high positive correlation. It is also shown that educational facilities are prevalent in residential zones. Furthermore, it can be observed that the functions of the functional zones can be recognised by using the category distribution of the POIs. In the second part of this chapter, this work has shown that the combination POIs and Tweet can identify the functions of the functional zones with higher accuracy than when only one of the aspect is used.

Chapter 8

Conclusions and Future Work

8.1 Conclusions

In recent years, with the advancement of technologies, urban data has become more accessible. In most countries, it is possible for the public to obtain urban data from government portals. For this reason, it is now easier to conduct urban analysis. However, there are still some data-poor countries that do not have the financial capability or manpower to conduct a census. These countries can consider using social media data to perform urban analysis. This work has discussed that social media data are less costly and can be gathered with minimal involvement from the population. This work has also discussed some of the goals and challenges in urban planning. One of the challenges in urban planning is the balance of development efforts. Excessive development can lead to expensive public housings. On the other hand, insufficient development may hinder population growth. Identifying similar regions can allow city planners to find areas with identical development intensity. In turn, this allows them to draft out similar planning objectives.

Some preliminary works were conducted to investigate the suitability of certain concepts in urban analysis. Clustering was tested to investigate if it is possible to segregate regions based on Tweet. It is found to be a challenging task. In countries where the population density is high, it is difficult to separate clusters of Tweets as the regions are closed to one another. Various similarity measures were examined in the preliminary work. The similarity measures examined include the Euclidean distance, Manhattan

distance, Cosine similarity, KL divergence and JS divergence. Each measure has its own advantages and pitfalls. It is found that the Euclidean and Manhattan distance work well on data with a small number of dimension. The Cosine similarity works better on data with a large number of dimension. This work also investigated the various correlation between urban and social media data. It is found that POIs have an impact on the volume of Tweet in a region. In addition, the residents can affect the development of certain types of POI in a region.

The urban region query system developed in this work is used to find similar areas. It is shown in this work that individual aspect of a region can find different types of similar regions. When Tweet is used, the system returns topically consistent regions. Similarly, when the population data is used, the system returns regions with identical demographics. When the POI data is used, the system returns regions with similar facilities. However, if all aspects are used, the system can return results that are more specific to individual users. This work also shows that the system can recognise the function of each functional zone. It is found that the recognition accuracy is highest when two aspects of a functional zone are considered.

There are three main objectives in this work. Firstly, to investigate the effectiveness of different data types in finding similar urban regions. Secondly, to explore the how to find similar urban regions by taking into account multiple aspects. Lastly, to analyse and recognise the function of a functional zone. To achieve these objectives, this work has proposed a number of approaches. These approaches are summarized as follows.

- This work proposed using social media data in urban analysis. It can be costly to conduct public consultation sessions to gather urban data. By using social media data, urban planning authorities can reduce the number of public consultation sessions and time required to draft out urban land use plans.
- This work proposed a similar region query framework. In particular, this work proposed using multiple related aspects to find similar regions. It can be shown that when both urban and social media data are considered, the precision and nDCG values of the queries are improved.
- This work proposed a weighted similarity approach for multiple aspects query. It is presented in this work that certain aspects are better at finding similar regions. In

this case, larger weights can be assigned to aspects that performed better. Smaller weights can be assigned to aspects that do not perform equally well.

- This work has developed an urban region similarity analysis system. An urban region query system can allow city planners and business owners to conduct urban analysis. In particular, city planners can use the system to estimate the development intensity of each region. Business owners can use the system to find the regions that fit their target market.

8.2 Future Work

This work has used different types of boundaries data to conduct urban analysis. However, these boundaries are rigid and may be less useful in some scenarios. For this reason, the developed urban region query system in this work allows users to generate their own boundary data. Users can conduct their own urban analysis using their own boundary data. In addition, there are a number of ways that this work can be extended. The potential extensions of this work are described in the following subsections.

8.2.1 Semi-supervised Query Framework

This work has presented a number of experiments to determine the most appropriate similarity measure for different representation types of urban and social media data. These experiments are conducted sequentially so as to make full use of the findings at each stage. Using the same idea, the proposed query framework can be extended to take into account the relevant similar regions as perceived by users. It is discussed in this work that individual users perceive similar regions differently. By comparing the results of each query with the ranking lists provided by the users, it is possible to automatically adjust the weights for different aspects to fit the requirements of the users. This automation can be done using a semi-supervised deep-learning approach.

8.2.2 Functional Zones Prediction

This work has presented the idea that the proposed query framework can recognise the function of each functional zone based on the POIs and Tweets. A simple nearest

neighbour approach is used in this work to recognise the function. One possible extension of this work is to consider the correlation between certain POI and functions to predict an undefined functional zone. A possible classifier is a decision tree. It has also been shown in this work that certain category of POI is correlated with certain types of functional zones. Furthermore, the Tweets in the same type of functional zones are similar in terms of their text content.

8.2.3 Automatically Generating the Boundaries

Boundaries data varies by countries. It may difficult for some countries to come up with detailed land use plans. As evident from this work, it is possible to identify the function of a functional zone based on Tweet and POI. Using the same idea, it is possible to generate boundaries based on the POIs and Tweets in an undefined region. One possible way to do so is to use Voronoi tessellation[110].

Bibliography

- [1] Singapore Urban Redevelopment Authority, 2016. URL <https://www.ura.gov.sg/uol/master-plan.aspx>.
- [2] K. H Cheong-Chua. *Environment and the city sharing Singapore experience and future challenges*. Institute of Policy Studies Times Academic Press, Singapore, 1995. ISBN 9812100806.
- [3] P Shekhu. *Planning Singapore from plan to implementation*. Singapore Institute of Planners, Singapore, 1998. ISBN 9810405731.
- [4] Sumiko Tan. *Home,work,play*. Urban Redevelopment Authority, Singapore, 1999. ISBN 9810417063.
- [5] Jennifer R. Wolch, Jason Byrne, and Joshua P. Newell. Urban green space, public health, and environmental justice: The challenge of making cities ‘just green enough’. *Landscape and Urban Planning*, 125:234 – 244, 2014. ISSN 0169-2046. doi: <http://dx.doi.org/10.1016/j.landurbplan.2014.01.017>. URL <http://www.sciencedirect.com/science/article/pii/S0169204614000310>.
- [6] Catharine Ward Thompson, Jenny Roe, Peter Aspinall, Richard Mitchell, Angela Clow, and David Miller. More green space is linked to less stress in deprived communities: Evidence from salivary cortisol patterns. *Landscape and Urban Planning*, 105(3):221 – 229, 2012. ISSN 0169-2046. doi: <http://dx.doi.org/10.1016/j.landurbplan.2011.12.015>. URL <http://www.sciencedirect.com/science/article/pii/S0169204611003665>.

- [7] T. Moyo and W. Musakwa. Using Crowdsourced Data (twitter and Facebook) to Delineate the Origin and Destination of Commuters of the Gautrain Public Transit System in South Africa. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, pages 143–150, June 2016. doi: 10.5194/isprs-annals-III-2-143-2016.
- [8] Xue Liu, Helin Liu, Wanzhen Chen, and Zhonghao Zhang. Spatiotemporal fragmentation of urban residential land use: A case study from china. *Sustainability*, 9(7), 2017. ISSN 2071-1050. doi: 10.3390/su9071096. URL <http://www.mdpi.com/2071-1050/9/7/1096>.
- [9] N. J. Yuan, Y. Zheng, X. Xie, Y. Wang, K. Zheng, and H. Xiong. Discovering urban functional zones using latent activity trajectories. *IEEE Transactions on Knowledge and Data Engineering*, 27(3):712–725, March 2015. ISSN 1041-4347. doi: 10.1109/TKDE.2014.2345405.
- [10] Nicholas Jing Yuan, Yu Zheng, and Xing Xie. Discovering regions of different functions in a city using human mobility and pois. In *KDD 2012*. KDD 2012, August 2012. URL <https://www.microsoft.com/en-us/research/publication/discovering-regions-of-different-functions-in-a-city-using-human-mobility-and-pois>
- [11] Annemarie Schneider and Curtis E. Woodcock. Compact, dispersed, fragmented, extensive? a comparison of urban growth in twenty-five global cities using remotely sensed data, pattern metrics and census information. *Urban Studies*, 45(3):659–692, 2008. doi: 10.1177/0042098007087340. URL <http://dx.doi.org/10.1177/0042098007087340>.
- [12] Sebastián Martinuzzi, William A. Gould, and Olga M. Ramos González. Land development, land use, and urban sprawl in puerto rico integrating remote sensing and population census data. *Landscape and Urban Planning*, 79(3):288 – 297, 2007. ISSN 0169-2046. doi: <http://dx.doi.org/10.1016/j.landurbplan.2006.02.014>. URL <http://www.sciencedirect.com/science/article/pii/S0169204606000478>.
- [13] Vanessa Frias-Martinez and Enrique Frias-Martinez. Spectral clustering for sensing urban land use using twitter activity. *Eng. Appl. Artif. Intell.*, 35:237–245, October 2014. ISSN 0952-1976. doi: 10.1016/j.engappai.2014.06.019. URL <http://dx.doi.org/10.1016/j.engappai.2014.06.019>.

- [14] V. Frias-Martinez, V. Soto, H. Hohwald, and E. Frias-Martinez. Characterizing urban landscapes using geolocated tweets. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing*, pages 239–248, Sept 2012. doi: 10.1109/SocialCom-PASSAT.2012.19.
- [15] M. Mazhar Rathore, Awais Ahmad, Anand Paul, and Seungmin Rho. Urban planning and building smart cities based on the internet of things using big data analytics. *Computer Networks*, 101:63 – 80, 2016. ISSN 1389-1286. doi: <http://dx.doi.org/10.1016/j.comnet.2015.12.023>. URL <http://www.sciencedirect.com/science/article/pii/S1389128616000086>. Industrial Technologies and Applications for the Internet of Things.
- [16] Singapore Urban Redevelopment Authority Feedback Channel, 2016. URL https://www.ura.gov.sg/feedbackWeb/contactus_feedback.jsp.
- [17] Singapore Urban Redevelopment Authority, 2014-2016. URL https://www.ura.gov.sg/uol/publications/corporate/ar/ar15_main.
- [18] Twitter, 2016. URL <https://twitter.com>.
- [19] Instagram, 2016. URL <https://www.instagram.com>.
- [20] Internet Live Stats, 2017. URL <http://www.internetlivestats.com/twitter-statistics/>.
- [21] E. Kumar. *Natural Language Processing*. I.K. International Publishing House, 2011. ISBN 9789380578774.
- [22] Cheng Xiang Zhai Charu C. Aggarwal. *Mining Text Data*. Springer-Verlag New York, 2012.
- [23] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435. URL <http://dl.acm.org/citation.cfm?id=944919.944937>.
- [24] Jonathan K. Pritchard, Matthew Stephens, and Peter Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000. ISSN 0016-6731. URL <http://www.genetics.org/content/155/2/945>.

- [25] Matthew D. Hoffman, David M. Blei, and Francis Bach. Online learning for latent dirichlet allocation. In *In NIPS*, 2010.
- [26] Yu Zheng. Tutorial on location-based social networks. In *21st International conference on World Wide Web*, 2012.
- [27] Zhou Xiaofang Yu Zheng. *Computing with Spatial Trajectories*, chapter Location-based social networks: Users, page 308. Springer-Verlag New York, 2011.
- [28] Yelp, 2016. URL <https://www.yelp.com/>.
- [29] Noémie Elhadad, Luis Gravano, Daniel Hsu, Sharon Balter, Vasudha Reddy, HaeNa Waechter, and BORNE OUTBREAKS. Information extraction from social media for public health. In *KDD at Bloomberg: The Data Frameworks Track (KDD 2014)*, 2014.
- [30] YouTube, 2016. URL <https://www.youtube.com/>.
- [31] Adiya Abisheva, Venkata Rama Kiran Garimella, David Garcia, and Ingmar Weber. Who watches (and shares) what on youtube? and when?: Using twitter to understand youtube viewership. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining, WSDM '14*, pages 593–602, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2351-2. doi: 10.1145/2556195.2566588. URL <http://doi.acm.org/10.1145/2556195.2566588>.
- [32] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web, WWW '10*, pages 851–860, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-799-8. doi: 10.1145/1772690.1772777. URL <http://doi.acm.org/10.1145/1772690.1772777>.
- [33] Maximilian Walther and Michael Kaisser. Geo-spatial event detection in the twitter stream. In *Proceedings of the 35th European Conference on Advances in Information Retrieval, ECIR'13*, pages 356–367, Berlin, Heidelberg, 2013. Springer-Verlag. ISBN 978-3-642-36972-8. doi: 10.1007/978-3-642-36973-5_30. URL http://dx.doi.org/10.1007/978-3-642-36973-5_30.

- [34] Yuan Liang, James Caverlee, and Cheng Cao. A Noise-Filtering Approach for Spatio-temporal Event Detection in Social Media. In *Proceedings of the 37th European Conference on Information Retrieval Research: Advances in Information Retrieval*, volume 9022 of *Lecture Notes in Computer Science*, pages 233–244, 2015. doi: 10.1007/978-3-319-16354-3_25.
- [35] Younghoon Kim, Jiawei Han, and Cangzhou Yuan. Toptrac: Topical trajectory pattern mining. *KDD : proceedings. International Conference on Knowledge Discovery and Data Mining*, 2015:587–596, 2015.
- [36] Pramod Anantharam, Payam Barnaghi, Krishnaprasad Thirunarayan, and Amit Sheth. Extracting city traffic events from social streams. *ACM Trans. Intell. Syst. Technol.*, 6(4):43:1–43:27, July 2015. ISSN 2157-6904. doi: 10.1145/2717317. URL <http://doi.acm.org/10.1145/2717317>.
- [37] Gennady Andrienko, Dimitrios Gunopulos, Yannis Ioannidis, Vana Kalogeraki, Ioannis Katakis, Katharina Morik, and Olivier Verscheure. Mining urban data (part c). *Information Systems*, 64:219 – 220, 2017. ISSN 0306-4379. doi: <https://doi.org/10.1016/j.is.2016.09.003>. URL <http://www.sciencedirect.com/science/article/pii/S0306437916304100>.
- [38] Carlo Ratti, Dennis Frenchman, Riccardo Maria Pulselli, and Sarah Williams. Mobile landscapes: Using location data from cell phones for urban analysis. *Environment and Planning B: Planning and Design*, 33(5):727–748, 2006. doi: 10.1068/b32047. URL <http://dx.doi.org/10.1068/b32047>.
- [39] J. Reades, F. Calabrese, A. Sevtsuk, and C. Ratti. Cellular census: Explorations in urban data collection. *IEEE Pervasive Computing*, 6(3):30–38, July 2007. ISSN 1536-1268. doi: 10.1109/MPRV.2007.53.
- [40] Constantine E. Kontokosta and Nicholas Johnson. Urban phenology: Toward a real-time census of the city using wi-fi data. *Computers, Environment and Urban Systems*, 64:144 – 153, 2017. ISSN 0198-9715. doi: <https://doi.org/10.1016/j.compenvurbsys.2017.01.011>. URL <http://www.sciencedirect.com/science/article/pii/S0198971516300928>.
- [41] Nivan Ferreira, Jorge Poco, Huy T. Vo, Juliana Freire, and Cláudio T. Silva. Visual exploration of big spatio-temporal urban data: A study of new york city taxi trips.

- IEEE Transactions on Visualization and Computer Graphics*, 19(12):2149–2158, December 2013. ISSN 1077-2626. doi: 10.1109/TVCG.2013.226. URL <http://dx.doi.org/10.1109/TVCG.2013.226>.
- [42] Y. Yoshimura, A. Krebs, and C. Ratti. Noninvasive bluetooth monitoring of visitors’ length of stay at the louvre. *IEEE Pervasive Computing*, 16(2):26–34, April 2017. ISSN 1536-1268. doi: 10.1109/MPRV.2017.33.
- [43] Dongyoun Shin. Urban sensing by crowdsourcing: Analysing urban trip behaviour in zurich. *International Journal of Urban and Regional Research*, 40(5):1044–1060, 2016. ISSN 1468-2427. doi: 10.1111/1468-2427.12416. URL <http://dx.doi.org/10.1111/1468-2427.12416>.
- [44] Michael Jendryke, Timo Balz, Stephen C McClure, and Mingsheng Liao. Putting people in the picture: Combining big location-based social media data and remote sensing imagery for enhanced contextual urban information in shanghai. *Computers, Environment and Urban Systems*, 62:99 – 112, 2017. ISSN 0198-9715. doi: <https://doi.org/10.1016/j.compenvurbsys.2016.10.004>. URL <http://www.sciencedirect.com/science/article/pii/S019897151630285X>.
- [45] Taras Agryzkov, Pablo Martí, Leandro Tortosa, and José F. Vicent. Measuring urban activities using foursquare data and network analysis: a case study of murcia (spain). *International Journal of Geographical Information Science*, 31(1):100–121, 2017. doi: 10.1080/13658816.2016.1188931. URL <http://dx.doi.org/10.1080/13658816.2016.1188931>.
- [46] Eamonn O’Neill, Vassilis Kostakos, Tim Kindberg, Ava Fatah gen. Schiek, Alan Penn, Danaë Stanton Fraser, and Tim Jones. *Instrumenting the City: Developing Methods for Observing and Understanding the Digital Cityscape*, pages 315–332. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006. ISBN 978-3-540-39635-2. doi: 10.1007/11853565_19. URL http://dx.doi.org/10.1007/11853565_19.
- [47] Nathan Eagle and Alex Sandy Pentland. Eigenbehaviors: identifying structure in routine. *Behavioral Ecology and Sociobiology*, 63(11):1689–1689, 2009. ISSN 1432-0762. doi: 10.1007/s00265-009-0830-6. URL <http://dx.doi.org/10.1007/s00265-009-0830-6>.

- [48] Nathan Eagle and Alex (Sandy) Pentland. Reality mining: Sensing complex social systems. *Personal Ubiquitous Comput.*, 10(4):255–268, March 2006. ISSN 1617-4909. doi: 10.1007/s00779-005-0046-3. URL <http://dx.doi.org/10.1007/s00779-005-0046-3>.
- [49] Rein Ahas and Ülar Mark. Location based services—new challenges for planning and public administration? *Futures*, 37(6):547 – 561, 2005. ISSN 0016-3287. doi: <http://dx.doi.org/10.1016/j.futures.2004.10.012>. URL <http://www.sciencedirect.com/science/article/pii/S0016328704001521>.
- [50] Constantine E. Kontokosta. Urban informatics for social good: Definitions, tensions, and challenges. In *Proceedings of the 2Nd International Workshop on Science of Smart City Operations and Platforms Engineering*, SCOPE '17, pages 52–56, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-4989-5. doi: 10.1145/3063386.3064888. URL <http://doi.acm.org/10.1145/3063386.3064888>.
- [51] Singapore Data.gov, 2016. URL <https://data.gov.sg/>.
- [52] Emily Y. Soh and Belinda Yuen. Government-aided participation in planning singapore. *Cities*, 23(1):30 – 43, 2006. ISSN 0264-2751. doi: <http://dx.doi.org/10.1016/j.cities.2005.07.011>. URL <http://www.sciencedirect.com/science/article/pii/S0264275105000855>.
- [53] Heide Studer. *Urban Practices and Planning. A Moroccan Case Study*, pages 333–347. Palgrave Macmillan UK, London, 2016. ISBN 978-1-349-95109-3. doi: 10.1057/978-1-349-95109-3_14. URL http://dx.doi.org/10.1057/978-1-349-95109-3_14.
- [54] Yandong Wang, Teng Wang, Ming-Hsiang Tsou, Hao Li, Wei Jiang, and Fengqin Guo. Mapping dynamic urban land use patterns with crowdsourced geo-tagged social media (sina-weibo) and commercial points of interest collections in beijing, china. *Sustainability*, 8(11), 2016. ISSN 2071-1050. doi: 10.3390/su8111202. URL <http://www.mdpi.com/2071-1050/8/11/1202>.
- [55] Janina Kleemann, Justice Nana Inkoom, Michael Thiel, Sangeetha Shankar, Sven Lautenbach, and Christine Fürst. Peri-urban land use pattern and its relation to land use planning in ghana, west africa. *Landscape and Urban Planning*, pages –, 2017. ISSN 0169-2046. doi: <https://doi.org/10.1016/j.landurbplan>.

- 2017.02.004. URL <http://www.sciencedirect.com/science/article/pii/S0169204617300270>.
- [56] Piyushimita (Vonu) Thakuriah, Nebiyu Y. Tilahun, and Moira Zellner. *Big Data and Urban Informatics: Innovations and Challenges to Urban Planning and Knowledge Discovery*, pages 11–45. Springer International Publishing, Cham, 2017. ISBN 978-3-319-40902-3. doi: 10.1007/978-3-319-40902-3_2. URL http://dx.doi.org/10.1007/978-3-319-40902-3_2.
- [57] Chang Sheng, Yu Zheng, Wynne Hsu, Mong Li Lee, and Xing Xie. Answering top-k similar region queries. In *DASFAA 2010*. DASFAA 2010, April 2010. URL <https://www.microsoft.com/en-us/research/publication/answering-top-k-similar-region-queries/>.
- [58] Micheline Kamber Jiawei Han. *Data Mining: Concepts and Techniques 3rd Edition*. Morgan Kaufmann, 2011. ISBN 9380931913.
- [59] J. C. Dunn. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, 3(3):32–57, 1973. doi: 10.1080/01969727308546046. URL <http://dx.doi.org/10.1080/01969727308546046>.
- [60] James C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Kluwer Academic Publishers, Norwell, MA, USA, 1981. ISBN 0306406713.
- [61] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, pages 281–297, Berkeley, Calif., 1967. University of California Press. URL <http://projecteuclid.org/euclid.bsmsp/1200512992>.
- [62] S. Lloyd. Least squares quantization in pcm. *IEEE Trans. Inf. Theor.*, 28(2): 129–137, September 2006. ISSN 0018-9448. doi: 10.1109/TIT.1982.1056489. URL <http://dx.doi.org/10.1109/TIT.1982.1056489>.
- [63] David Arthur and Sergei Vassilvitskii. K-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '07, pages 1027–1035, Philadelphia, PA, USA, 2007. Society for Industrial and Applied Mathematics. ISBN 978-0-898716-24-5. URL <http://dl.acm.org/citation.cfm?id=1283383.1283494>.

- [64] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, pages 226–231. AAAI Press, 1996. URL <http://dl.acm.org/citation.cfm?id=3001460.3001507>.
- [65] Zellig S. Harris. Distributional structure. *<i>WORD</i>*, 10(2-3):146–162, 1954. doi: 10.1080/00437956.1954.11659520. URL <http://dx.doi.org/10.1080/00437956.1954.11659520>.
- [66] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008. ISBN 0521865719, 9780521865715.
- [67] Anand Rajaraman and Jeffrey David Ullman. *Mining of Massive Datasets*. Cambridge University Press, New York, NY, USA, 2011. ISBN 1107015359, 9781107015357.
- [68] Richard Ernest Bellman. *Dynamic Programming*. Dover Publications, Incorporated, 2003. ISBN 0486428095.
- [69] RICHARD BELLMAN. *Adaptive Control Processes: A Guided Tour*. Princeton University Press, 1961. ISBN 9780691079011. URL <http://www.jstor.org/stable/j.ctt183ph6v>.
- [70] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated, 2014. ISBN 1461471370, 9781461471370.
- [71] Pavel Pudil and Jana Novovicova. Novel methods for subset selection with respect to problem knowledge. *IEEE Intelligent Systems*, 13(2):66–74, March 1998. ISSN 1541-1672. doi: 10.1109/5254.671094. URL <http://dx.doi.org/10.1109/5254.671094>.
- [72] M.F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980. doi: 10.1108/eb046814. URL <http://dx.doi.org/10.1108/eb046814>.

- [73] Beverly Plester, Clare Wood, and Puja Joshi. Exploring the relationship between children's knowledge of text message abbreviations and school literacy outcomes. *British Journal of Developmental Psychology*, 27(1):145–161, 2009. ISSN 2044-835X. doi: 10.1348/026151008X320507. URL <http://dx.doi.org/10.1348/026151008X320507>.
- [74] Kristy Winzker, Frenette Southwood, and Kate Huddleston. Investigating the impact of sms speak on the written work of english first language and english second language high school learners. *Per Linguam*, 25(2), 2011. ISSN 2224-0012. doi: 10.5785/25-2-31. URL <http://perlinguam.journals.ac.za/pub/article/view/31>.
- [75] George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. Introduction to wordnet: An on-line lexical database*. *International Journal of Lexicography*, 3(4):235, 1990. doi: 10.1093/ijl/3.4.235. URL [+http://dx.doi.org/10.1093/ijl/3.4.235](http://dx.doi.org/10.1093/ijl/3.4.235).
- [76] Thomas Hofmann. Learning the similarity of documents: An information-geometric approach to document retrieval and categorization, 2000.
- [77] Samuel Kotz, N. Balakrishnan, and Norman L. Johnson. *Dirichlet and Inverted Dirichlet Distributions*, pages 485–527. John Wiley and Sons, Inc., 2005. ISBN 9780471722069. doi: 10.1002/0471722065.ch49. URL <http://dx.doi.org/10.1002/0471722065.ch49>.
- [78] Steffen Bickel and Tobias Scheffer. Multi-view clustering. In *Proceedings of the Fourth IEEE International Conference on Data Mining, ICDM '04*, pages 19–26, Washington, DC, USA, 2004. IEEE Computer Society. ISBN 0-7695-2142-8. URL <http://dl.acm.org/citation.cfm?id=1032649.1033432>.
- [79] Hua Wang, Feiping Nie, and Heng Huang. Multi-view clustering and feature learning via structured sparsity. In Sanjoy Dasgupta and David Mcallester, editors, *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, volume 28, pages 352–360. JMLR Workshop and Conference Proceedings, May 2013. URL <http://jmlr.org/proceedings/papers/v28/wang13c.pdf>.
- [80] Rakesh Agrawal, Johannes Gehrke, Dimitrios Gunopulos, and Prabhakar Raghavan. Automatic subspace clustering of high dimensional data for data mining

- applications. In *Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data*, SIGMOD '98, pages 94–105, New York, NY, USA, 1998. ACM. ISBN 0-89791-995-5. doi: 10.1145/276304.276314. URL <http://doi.acm.org/10.1145/276304.276314>.
- [81] Lance Parsons, Ehtesham Haque, and Huan Liu. Subspace clustering for high dimensional data: A review. *SIGKDD Explor. Newsl.*, 6(1):90–105, June 2004. ISSN 1931-0145. doi: 10.1145/1007730.1007731. URL <http://doi.acm.org/10.1145/1007730.1007731>.
- [82] J. Hu, Q. Qian, J. Pei, R. Jin, and S. Zhu. Finding multiple stable clusterings. In *2015 IEEE International Conference on Data Mining*, pages 171–180, Nov 2015. doi: 10.1109/ICDM.2015.101.
- [83] Ulrike von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007. ISSN 1573-1375. doi: 10.1007/s11222-007-9033-z. URL <http://dx.doi.org/10.1007/s11222-007-9033-z>.
- [84] Amit Singhal. Modern information retrieval: A brief overview. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 24(4):35–42, 2001. URL <http://singhal.info/ieee2001.pdf>.
- [85] G. Van Brummelen. *Heavenly Mathematics: The Forgotten Art of Spherical Trigonometry*. Princeton University Press, 2013. ISBN 9780691148922. URL <https://books.google.com.sg/books?id=0BCCz8Sx5wkC>.
- [86] Eugene F. Krause. *Taxicab Geometry: An Adventure in Non-Euclidean Geometry*. Dover Publications, 1987. ISBN 0486252027, 9780486252025.
- [87] Lillian Lee. Measures of distributional similarity. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 25–32, Stroudsburg, PA, USA, 1999. Association for Computational Linguistics. ISBN 1-55860-609-3. doi: 10.3115/1034678.1034693. URL <http://dx.doi.org/10.3115/1034678.1034693>.
- [88] S. Kullback and R. A. Leibler. On information and sufficiency. *Ann. Math. Statist.*, 22(1):79–86, 03 1951. doi: 10.1214/aoms/1177729694. URL <http://dx.doi.org/10.1214/aoms/1177729694>.

- [89] L. B. Heilprin. Information theory and statistics. solomon kullback. wiley, new york; chapman and hall, london, 1959. xvii + 395 pp. illus. 12.50. *Science*, 131 (3404):917–918, 1960. ISSN 0036-8075. doi: 10.1126/science.131.3404.917-b. URL <http://science.sciencemag.org/content/131/3404/917.3>.
- [90] Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA, 1999. ISBN 0-262-13360-1.
- [91] D. M. Endres and J. E. Schindelin. A new metric for probability distributions. *IEEE Trans. Inf. Theor.*, 49(7):1858–1860, September 2006. ISSN 0018-9448. doi: 10.1109/TIT.2003.813506. URL <http://dx.doi.org/10.1109/TIT.2003.813506>.
- [92] F. österreicher and I. Vajda. A new class of metric divergences on probability spaces and and its statistical applications. *Ann. Inst. Statist. Math.*, 55:639–653, 2003.
- [93] Ido Dagan, Lillian Lee, and Fernando Pereira. Similarity-based methods for word sense disambiguation. In *Proceedings of the 35th ACL/8th EACL*, pages 56–63, 1997.
- [94] Karl Pearson. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58:240–242, 1895. ISSN 03701662. URL <http://www.jstor.org/stable/115794>.
- [95] Wikipedia, 2017. URL https://en.wikipedia.org/wiki/List_of_shopping_malls_in_Singapore.
- [96] Google, 2016. URL <https://developers.google.com/maps/documentation/javascript/examples/geocoding-reverse>.
- [97] Twitter, 2016. URL <https://dev.twitter.com/rest/public>.
- [98] A. Doyle S. Gillies S. Hagen T. Schaub H. Butler, M. Daly, 2016. URL <https://tools.ietf.org/html/rfc7946>.
- [99] OneMap, 2016. URL <http://www.onemap.sg/api/help/>.
- [100] MongoDB, 2017. URL <https://www.mongodb.com/>.
- [101] Robomongo, 2017. URL <https://robomongo.org/>.

-
- [102] Apache, 2017. URL <http://tomcat.apache.org/>.
 - [103] Pivotal, 2017. URL <https://spring.io/>.
 - [104] The R Foundation, 2017. URL <https://www.r-project.org/about.html>.
 - [105] The R Foundation, 2017. URL <https://www.rforge.net/Rserve/>.
 - [106] Oracle Coporation, 2017. URL <https://www.java.com/en/>.
 - [107] Apache, 2017. URL <https://maven.apache.org/>.
 - [108] Bootstrap, 2017. URL <http://getbootstrap.com/>.
 - [109] Mapbox, 2017. URL <https://www.mapbox.com/>.
 - [110] Franz Aurenhammer. Voronoi diagrams – a survey of a fundamental geometric data structure. *ACM COMPUTING SURVEYS*, 23(3):345–405, 1991.

Appendix A

Preliminary Experiment Results

FIGURE A.1: Tweet Clustering Result

(a) Cosine Similarity Using Term Frequency

(b) Cosine Similarity Using TF-IDF

(c) Euclidean Distance Using Term Frequency

(d) Euclidean Distance Using TF-IDF

(e) JS Divergence Using Term Frequency

(f) JS Divergence Using TF-IDF

FIGURE A.2: Single Similarity Measure: Tweet

FIGURE A.3: Single Similarity Measure: Female Residents

(a) Distribution Comparison

FIGURE A.4: Single Similarity Measure: Female Residents Distributions

FIGURE A.5: Single Similarity Measure: POI

(a) Distribution Comparison

(b) Tampines East POIs

(c) Tampines West POIs

FIGURE A.6: Single Similarity Measure: POI Distributions

(a) Cosine Similarity of TFIDF and JS Divergence of Cosine Similarity of Topic Distribution $\alpha=0.2$

(b) Cosine Similarity of TFIDF and JS Divergence of Cosine Similarity of Topic Distribution $\alpha=0.5$

(c) Cosine Similarity of TFIDF and JS Divergence of Term Frequency $\alpha=0.2$

(d) Cosine Similarity of TFIDF and JS Divergence of Term Frequency $\alpha=0.5$

(e) JS Divergence of Cosine Similarity of Topic Distribution and JS Divergence of Term Frequency $\alpha=0.2$

(f) JS Divergence of Cosine Similarity of Topic Distribution and JS Divergence of Term Frequency $\alpha=0.5$

FIGURE A.7: Double Similarity Measure: Tweet

Appendix B

Queries with Single Aspect

Measure	Representation Type	Precision	nDCG
Cosine Similarity	Age Distribution	0.39	0.16
Cosine Similarity	Female Age Distribution	0.39	0.16
Euclidean Distance	Male Age Distribution	0.35	0.18
Euclidean Distance	Female Age Spread	0.33	0.15
JS Divergence	Male Age Spread	0.04	0.03
JS Divergence	Female Age Distribution	0.01	0.05
KL Divergence	Female Age Spread	0.09	0.06
KL Divergence	Age Distribution	0.06	0.03
Manhattan Distance	Male Age Distribution	0.36	0.15
Manhattan Distance	Female Age Spread	0.35	0.13
Cosine Similarity	POI Category Distribution	0.28	0.13
Euclidean Distance	POI Category Distribution	0.28	0.13
JS Divergence	POI Category Distribution	0.04	0.04
KL Divergence	POI Category Distribution	0.12	0.08
Manhattan Distance	POI Category Distribution	0.27	0.13
Cosine Similarity	Document Frequency	0.25	0.14
Cosine Similarity	Inverse Document Frequency	0.25	0.14
Euclidean Distance	Term Burst Period Distribution	0.26	0.12
Euclidean Distance	TF-IDF	0.26	0.12
JS Divergence	Inverse Document Frequency	0.2	0.11
JS Divergence	Topic Distribution	0.14	0.1
KL Divergence	Document Frequency	0.18	0.12
KL Divergence	Topic Distribution	0.15	0.1
Manhattan Distance	Term Hourly Distribution	0.28	0.12
Manhattan Distance	Term Frequency	0.27	0.13

TABLE B.1: Planning Area: Query Performance

Measure	Representation Type	Precision	nDCG
Cosine Similarity	Male Age Distribution	0.29	0.13
Cosine Similarity	Age Distribution	0.32	0.13
Euclidean Distance	Female Age Distribution	0.27	0.13
Euclidean Distance	Age Distribution	0.29	0.14
JS Divergence	Female Age Distribution	0.02	0.03
JS Divergence	Female Age Spread	0.02	0.03
KL Divergence	Female Age Spread	0.04	0.02
KL Divergence	Male Age Distribution	0.04	0.02
Manhattan Distance	Age Distribution	0.28	0.14
Manhattan Distance	Male Age Distribution	0.26	0.13
Cosine Similarity	POI Category Distribution	0.06	0.08
Euclidean Distance	POI Category Distribution	0.06	0.08
JS Divergence	POI Category Distribution	0.01	0.02
KL Divergence	POI Category Distribution	0.01	0.02
Manhattan Distance	POI Category Distribution	0.06	0.07
Cosine Similarity	Document Frequency	0.06	0.1
Cosine Similarity	Inverse Document Frequency	0.06	0.1
Euclidean Distance	Tweet Hourly Distribution	0.04	0.06
Euclidean Distance	TF-IDF	0.04	0.06
JS Divergence	Inverse Document Frequency	0.03	0.04
JS Divergence	Topic Distribution	0.02	0.03
KL Divergence	Document Frequency	0.03	0.04
KL Divergence	Term Hourly Distribution	0.02	0.04
Manhattan Distance	TF-IDF	0.05	0.07
Manhattan Distance	Term Burst Period Distribution	0.05	0.06

TABLE B.2: Subzone: Query Performance

Appendix C

Queries with Double Aspects

Measure	Representation	w_i	Precision	nDCG
Euclidean;Euclidean	Term Hourly;Category	0.2	0.28	0.14
Euclidean;Euclidean	Term Hourly;Category	0.3	0.28	0.14
Manhattan;Cosine	Term Hourly;Category	0.3	0.28	0.14
Manhattan;Euclidean	Term Hourly;Category	0.2	0.28	0.14
Manhattan;Euclidean	Term Hourly;Category	0.3	0.28	0.14
Cosine;Euclidean	Similar Topic;Age	0.1	0.36	0.15
Cosine;Euclidean	Term Hourly;Age	0.1	0.36	0.15
Euclidean;Euclidean	Term Hourly;Age	0.1	0.36	0.15
Manhattan;Euclidean	Term Hourly;Age	0.1	0.36	0.15
Cosine;Euclidean	Category;Age	0.1	0.36	0.15
Cosine;Manhattan	Category;Age	0.1	0.36	0.15
Euclidean;Euclidean	Category;Age	0.2	0.36	0.15
Euclidean;Euclidean	Category;Age	0.1	0.36	0.15
Euclidean;Manhattan	Category;Age	0.2	0.36	0.15
Euclidean;Manhattan	Category;Age	0.1	0.36	0.15
Manhattan;Euclidean	Category;Age	0.4	0.33	0.16
Manhattan;Euclidean	Category;Age	0.1	0.36	0.15
Manhattan;Euclidean	Category;Age	0.5	0.32	0.16
Manhattan;Manhattan	Category;Age	0.3	0.36	0.15
Manhattan;Manhattan	Category;Age	0.2	0.35	0.16
Manhattan;Manhattan	Category;Age	0.1	0.35	0.16

TABLE C.1: Planning Area: Combining Tweet with POI, Tweet with Population and POI with Population

Measure	Representation	w_i	Precision	nDCG
Cosine;Cosine	Similar Topic;Category	0.1	0.06	0.08
Cosine;Euclidean	Similar Topic;Category	0.1	0.06	0.08
Manhattan;Cosine	Similar Topic;Category	0.1	0.06	0.08
Manhattan;Euclidean	Similar Topic;Category	0.1	0.06	0.08
Cosine;Cosine	Term Hourly;Category	0.1	0.06	0.08
Cosine;Euclidean	Term Hourly;Category	0.1	0.06	0.08
Euclidean;Cosine	Term Hourly;Category	0.1	0.06	0.08
Euclidean;Euclidean	Term Hourly;Category	0.1	0.06	0.08
Cosine;Manhattan	Similar Topic;Age	0.1	0.27	0.11
Euclidean;Euclidean	Term Hourly;Age	0.1	0.26	0.12
Euclidean;Manhattan	Term Hourly;Age	0.1	0.26	0.11
Manhattan;Euclidean	Term Hourly;Age	0.1	0.26	0.12
Manhattan;Manhattan	Term Hourly;Age	0.1	0.26	0.11
Cosine;Cosine	Category;Age	0.1	0.2	0.11
Cosine;Euclidean	Category;Age	0.1	0.25	0.12
Cosine;Manhattan	Category;Age	0.1	0.26	0.12
Euclidean;Cosine	Category;Age	0.1	0.21	0.11
Euclidean;Euclidean	Category;Age	0.1	0.27	0.12
Euclidean;Manhattan	Category;Age	0.1	0.27	0.12
Manhattan;Manhattan	Category;Age	0.1	0.26	0.12

TABLE C.2: Subzone: Combining Tweet with POI, Tweet with Population and Population with POI