UCL DEPARTMENT OF GEOGRAPHY

YEAR 2020-21

MODULE CODE:	**GEOG0051**
MODULE NAME:	**Mining Social and Geographic Datasets**
COURSE PAPER TITLE:	**Machine Learning Analysis for Cambridge Mobility Data and Calgary Venue Review Data**
WORD COUNT:	**2497**

Your essay, appropriately anonymised, may be used to help future students prepare for assessment. Double click this box to opt out of this ☒

Coursework Part One

Mobility Patterns Analysis in Cambridge

Introduction

Mobility analysis can provide recommendations in the marketing, advertisement, traffic and urban planning fields, by understanding spatio-temporal movement of human beings (Barbosa et al., 2018). Generally, the analysis can be conducted at micro-mobility scale (namely individual-level) or macro-mobility scale (namely population-level). Several previous studies suggested that individual trajectories perform to be regular and predictable, which is possible to be used for constructing generative individual mobility models (Barbosa et al., 2018 and González, Hidalgo & Barabási, 2009). In addition, investigating population-level data can model aggregate mobility of large number of individuals and reproduce the origin-destination flows (Barbosa et al., 2018). In urban planning field, mobility analysis can contribute to the decisions of location selection of urban features (Cheng, Li & Yu, 2007 and Elsamen & Hiyasat, 2017).

Apart from that, at macro scale, several articles have indicated the relationship between urban morphology and street centrality (Strano et al., 2012). For example, to attract customer footfalls, the site for shopping malls should satisfy accessibility and transport connectivity, which can be reflected by street centrality. Recently, Cambridge government would like to construct a new shopping mall and seek advice on placement location. Therefore, this essay will analyze people's movement and investigate the relationship between check-in frequencies and network centrality for Cambridge. Finally, the author will take the role of a consultant to the authorities to provide suggestions for location selection of a new shopping mall in Cambridge.

Data and Methodology

Gowalla is a now-defunct online application found in 2007, providing location-based services where users can "discover, capture and share places and events with friends" by check-ins (Cuddy & Glassman, 2010). By 2010, Gowalla had approximately 340,000 users (ibid.). The dataset applied in the research is a subset of Gowalla users located in Cambridge, UK. In general, the dataset consists of user ID, check-in-date, check-in-time and geolocation. Therefore, it is able to trace the movements of certain individuals on designated date and draw possible routes on the map, according to the check-ins.

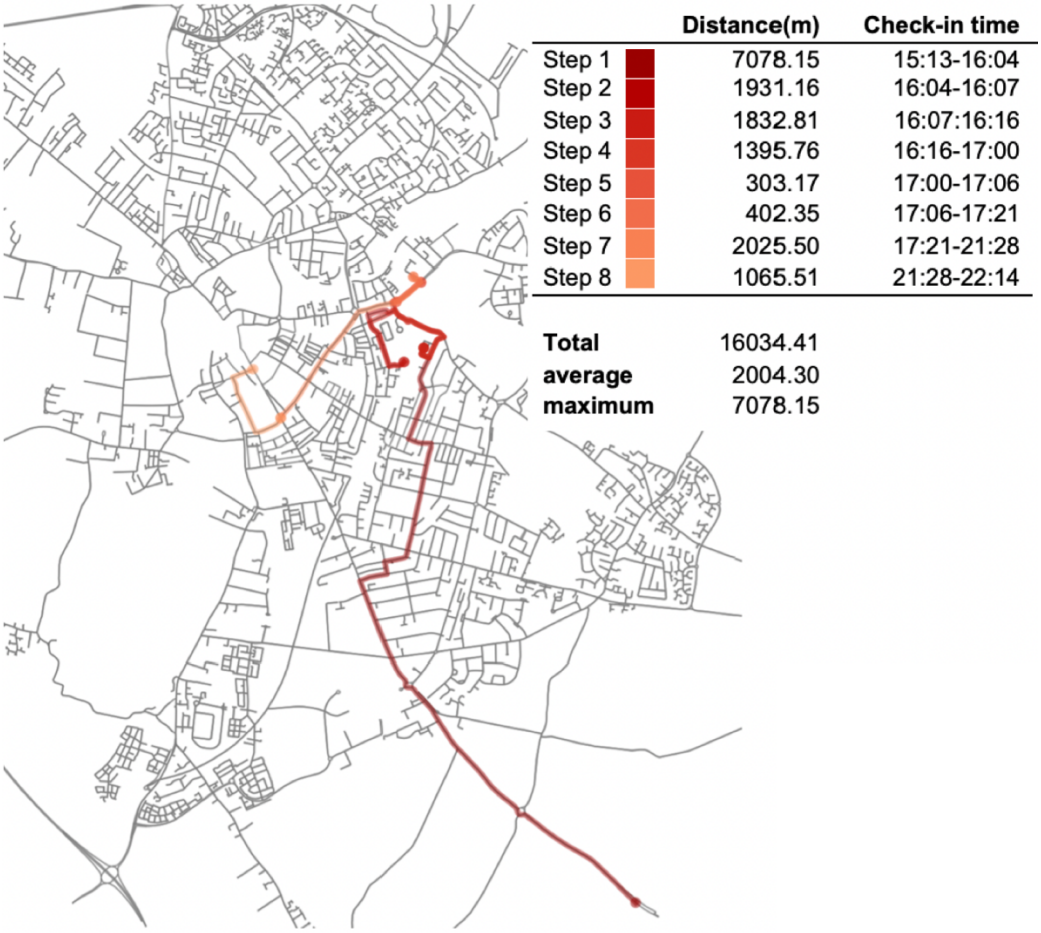
The research on mobility patterns in Cambridge will emphasize both micro scale analysis and macro scale analysis to thoroughly analyze people's movement. At the beginning, detailed individual-level assessment for the selected two users is conducted and their differences are characterized. After that, the relationship between the general pattern of check-in frequencies and closeness centrality measured for the City of Cambridge is described. In addition, the paper also makes supplementary investigations for street centrality in Cambridge, in order to contribute to the urban planning decision-making process. Finally, based on the completed analyses, the recommendations for location selection of the new shopping mall are proposed.

Micro Scale Analysis

According to Barbosa et al. (2018), each individual can be characterized by a high probability to visit several highly frequented locations and an interrelated time-independent travel distance. Besides, despite of the diversity of individual travel history, a person probably follows a similarly reproducible pattern (ibid.). Therefore, it is possible to discover the potential characteristics and implications of particular persons' trajectories and estimate their individual mobility patterns.

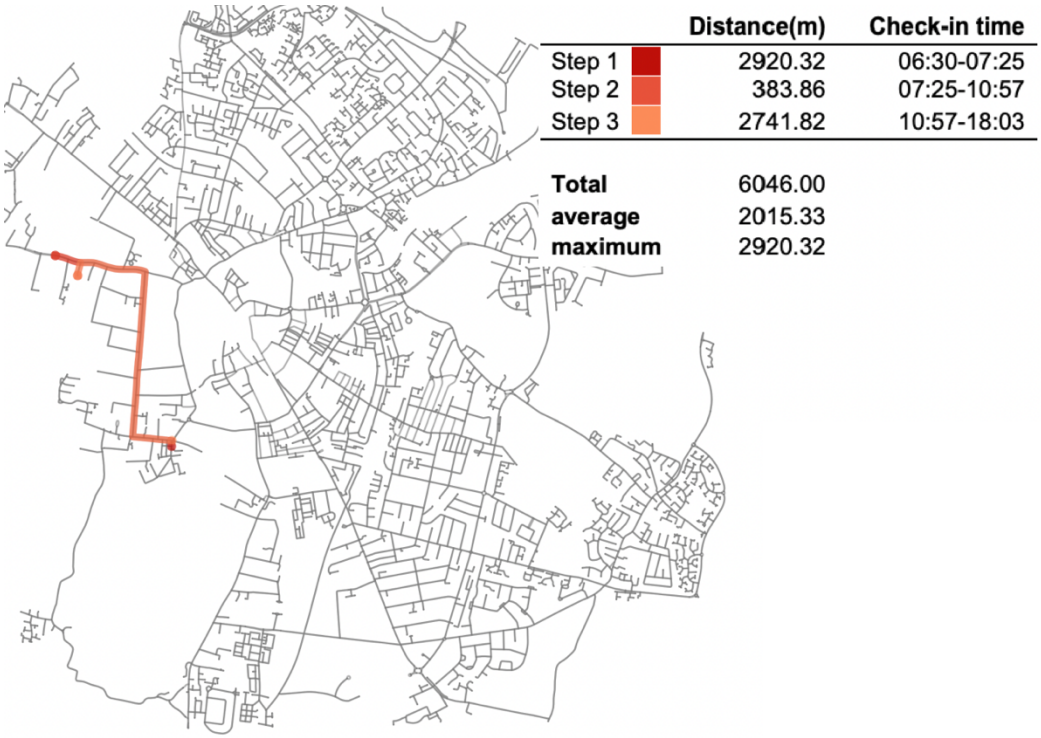
In the research, the individual daily travels of two designated Gowalla users are characterized and compared. As displayed in Figure 1 and Figure 2, the paths of these two users have been visualized using the OSMnx library in Python with detailed routes for each user attached in Appendices. Overall, user one visited more places, and the total travel displacement was more than twice of user 2, accounting for 16034.41m. Comparing their maximum displacement, over similar period of time (around 1 hour), user 1 travels three times as far as user 2, which might be relevant to the road condition, travel methods and their motivations.

To be specific, further considering the check-in locations and potential events, different travel patterns can be explained. For user 1, coming from A2 (London to Dover) road, the user did not seem to be familiar with the city roads. After went through many repeated routes in the afternoon, the person returned to the residence after eating at around 6pm. For user 2, the displacement seems simple, since the activity scope was concentrated near Grange Road, which is a north–south street with several colleges belonging to the University of Cambridge situated on. Therefore, user 2 might be a student, taking classes through the daytime and returning to the dormitory afterwards.



	Distance(m)	Check-in time
Step 1	7078.15	15:13-16:04
Step 2	1931.16	16:04-16:07
Step 3	1832.81	16:07:16:16
Step 4	1395.76	16:16-17:00
Step 5	303.17	17:00-17:06
Step 6	402.35	17:06-17:21
Step 7	2025.50	17:21-21:28
Step 8	1065.51	21:28-22:14
Total	16034.41	
average	2004.30	
maximum	7078.15	

Figure 1: Travel routes of user 1 (Source: Author's own)



	Distance(m)	Check-in time
Step 1	2920.32	06:30-07:25
Step 2	383.86	07:25-10:57
Step 3	2741.82	10:57-18:03
Total	6046.00	
average	2015.33	
maximum	2920.32	

Figure 2: Travel routes of user 2 (Source: Author's own)

Macro Scale Analysis

For macro scale analysis, the aggregate data from multiple Gowalla users in City of Cambridge are visualized. As indicated in micro scale analysis, people with similar profile would probably have similar reproducible travel pattern, such as commuters between cities or students traveled within fixed areas. In that case, the most frequented visited venues might possibly constitute a certain pattern, reflecting the populated places. Therefore, analysis on macro scale human mobility can identify the inherent similarity and investigate the corresponding impact on mobility-driven phenomena, such as urban planning.

From the research on individual mobility to aggregate mobility, the number of nodes as origin or destination of short journeys being visited are calculated and processed. As Figure 3 represents, the certain attributes reflecting the visit frequency of locations have been integrated into the street network nodes and plotted. Comparing to the Local Plan (Cambridge City Council, 2018), the points with high possibility to be origins or destinations happen to be concentrated in the city center, confirming the absorption capacity of urban areas and explaining the presented pattern in Figure 3 to a certain extent.

Figure 3: Possibility as origin or destination points (Source: Author's own)

Closeness centrality measures the average farness (inverse distance) between certain node and all other nodes, considering the length of the shortest path across nodes (Hansen et al., 2019). Although there are differences between Figure 3 and Figure 4, it is apparent that there is a relationship between frequented-visited locations and high closeness centrality areas. This might because higher closeness centrality always means higher connectivity to other places and more central positions in the network (ibid.), so that more attractive places are located there, attracting more Gowalla users to visit.

Figure 4: Closeness centrality in Cambridge (Source: Author's own)

Betweenness centrality measures the number of times a node being passed through along the shortest path between any couple of nodes (Hansen et al., 2019). According to Strano et al. (2012)., higher betweenness centrality can reflect the spatial accessibility and reveal the "backbone" of highly central routes, which further indicates the location of primary streets. Therefore, considering the population flow and street connectivity, the new shopping mall is suggested to be located around the intersection of Trumpington St Road and Lensfield Road (see Figure 5).

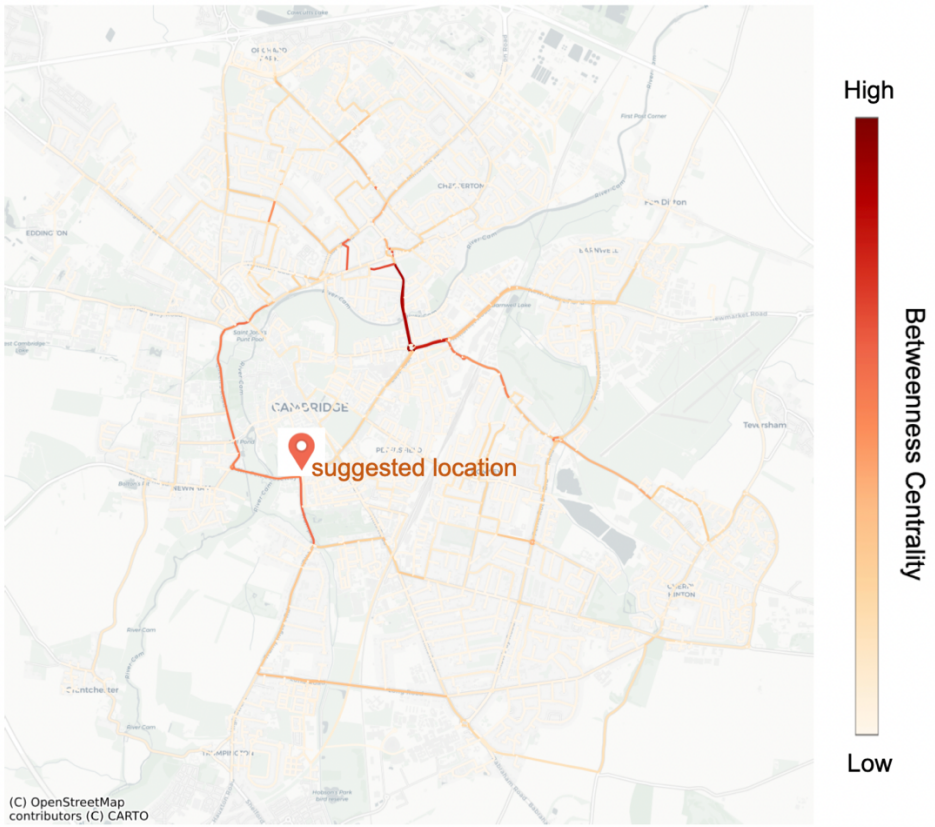


Figure 5: Betweenness centrality in Cambridge and suggested location for the new shopping mall (Source: Author's own)

Conclusion

In conclusion, this research has conducted micro scale analysis and macro scale analysis, thoroughly discussing the mobility patterns in City of Cambridge, based on Gowalla dataset. From the micro scale analysis, different movements for independent users were discussed and potential similar patterns for people with similar profile were indicated. From individual mobility to aggregate mobility, the relationship between frequented-visited locations and high closeness centrality areas was identified. Finally, integrating with the consideration of betweenness centrality, the location of the new shopping mall was suggested. Considering the limitations, there might be privacy issues existing, although some sensitive information had been removed. Besides, the representativeness of Gowalla dataset should be further justified. This paper can be seen as reference for the urban feature location selection research, although field research, land use policy and project budget need to be further concerned.

Reference List

Barbosa, H. et al. (2018). 'Human mobility: Models and applications', *Physics reports*, 734 (734), pp. 1-74.

Cambridge City Council (2018). *Local Pan 2018*. Available from: https://www.cambridge.gov.uk/local-plan-2018 (Accessed: 11 April 2021).

Cheng, E., Li, H. & Yu, L. (2007). 'A GIS approach to shopping mall location selection', *Building and environment*, 42 (2), pp. 884-892.

Cuddy, C. & Glassman, N. (2010). 'Location-Based Services: Foursquare and Gowalla, Should Libraries Play', *Journal of electronic resources in medical libraries*, 7 (4), pp. 336-343.

Elsamen, A. & Hiyasat, R. (2017). 'Beyond the random location of shopping malls: A GIS perspective in Amman, Jordan', *Journal of retailing and consumer services*, 34, pp. 30-37.

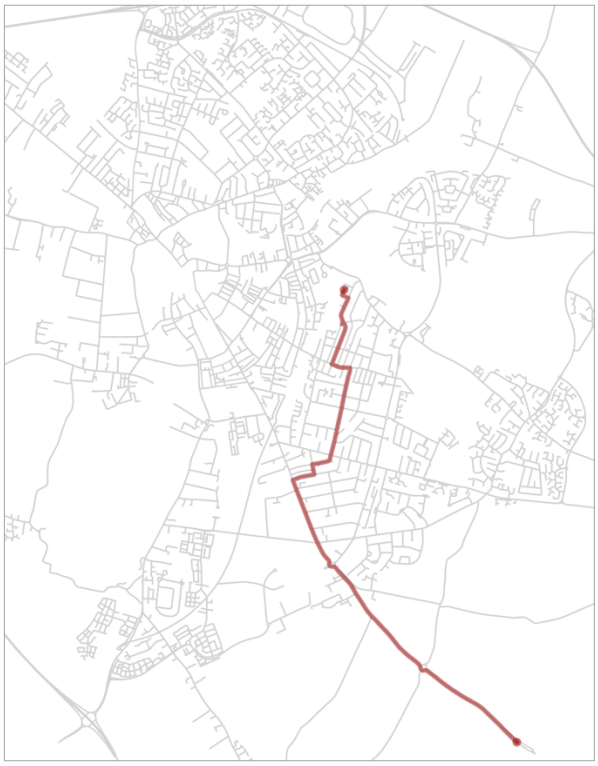
González, M., Hidalgo, C. & Barabási, A. (2009). 'Understanding individual human mobility patterns', *Nature (London)*, 458 (7235), p. 238.

Hansen, D et al. (2019). *Analyzing Social Media Networks with NodeXL, 2nd Edition / Hansen, Derek.* O'REILLY.

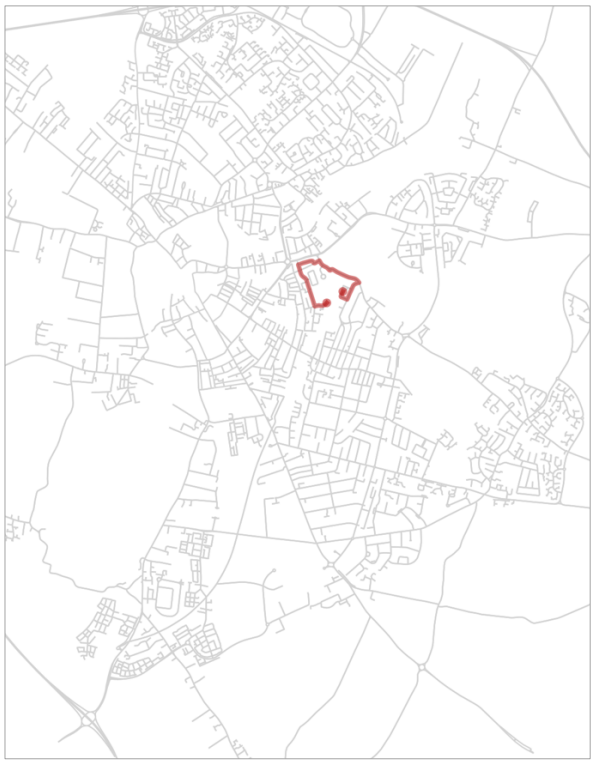
Strano, E. et al. (2012). 'Elementary processes governing the evolution of road networks', *Scientific reports*, 2 (1), p. 296.

Appendices

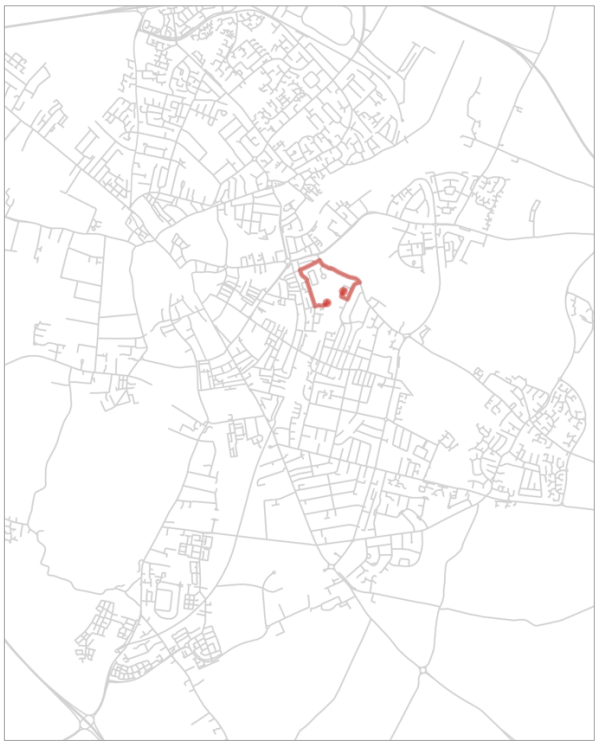
Appendix A: Detailed routes of user 1



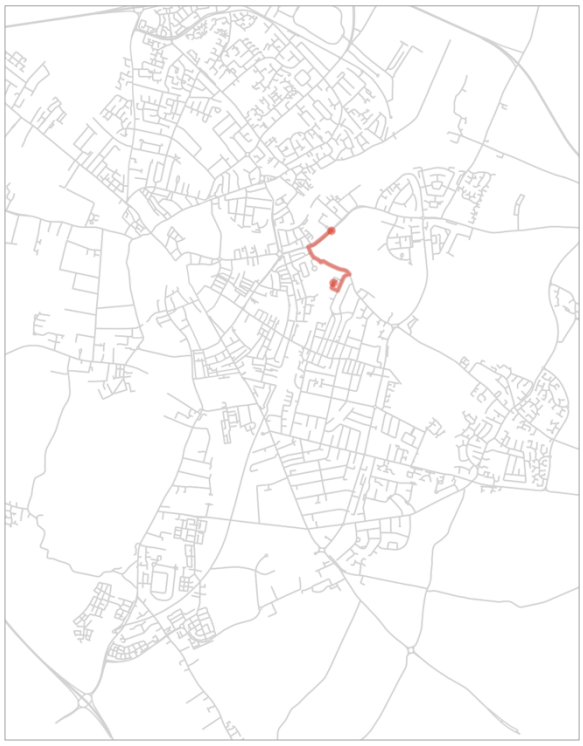
Step 1



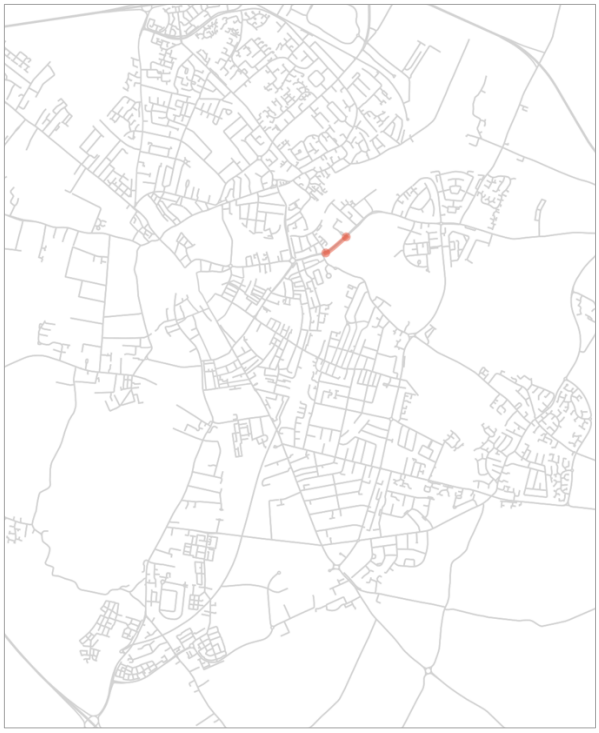
Step 2



Step 3



Step 4



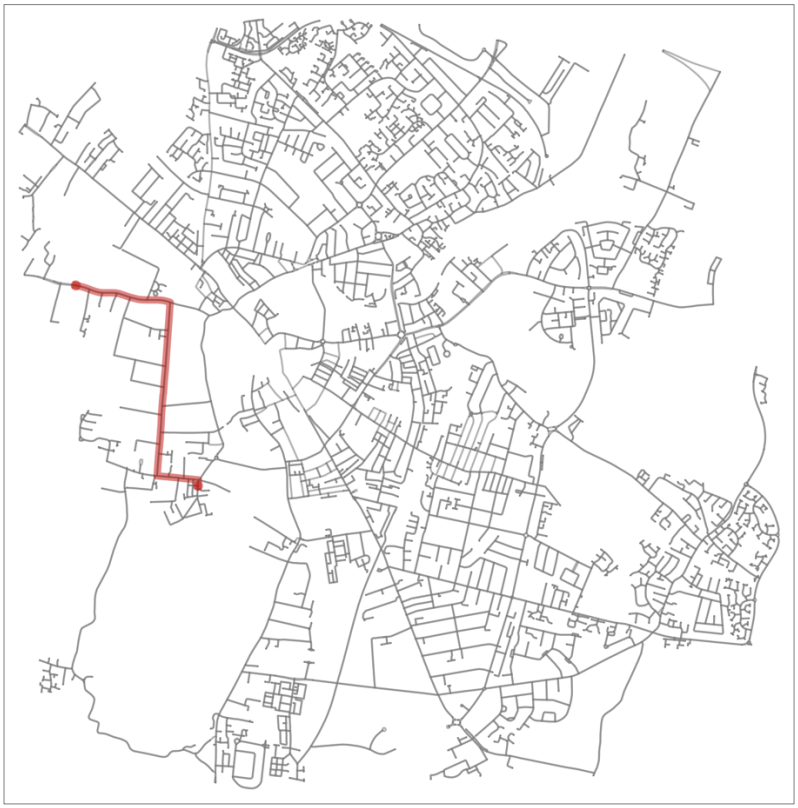
Step 5

Step 6

Step 7

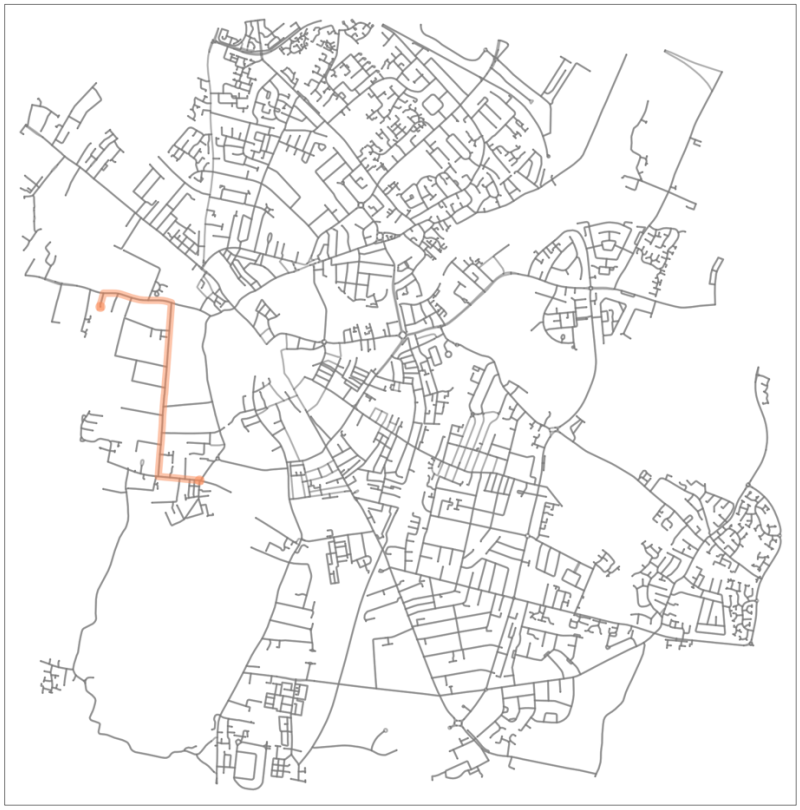
Step 8

Appendix B: Detailed routes of user 2



Step 1

Step 2



Step 3

Coursework Part Two

Machine Learning Analysis with Venue Review Data in Calgary, Canada

Introduction

Data mining technologies have been applied in the market to predict consumers' demands and satisfactory, in order to effectively increase the interest. Recently, online review websites and location-based services became popular, attracting an overwhelming population to write reviews, rate businesses and get advice (Chougule & Mulla, 2015 and Wang et al., 2018). According to Horrigan (2008), consumers are willing to spent from 20% to 99% more on purchasing products with 5-star rather than 4-star. In other words, consumer purchase decisions, product sales and business revenues are all directly or indirectly influenced by the reviews. Through mining the review datasets, review rating prediction models can be conducted, capturing the semantics of texts to predict service qualities of public venues, estimate customer satisfaction and provide suggestions for potential consumers, pioneered by Pang & Lee (2005). Therefore, this essay will combine the textual and non-textual information to improve various machine learning models to discover whether it is possible to accurately predict venue stars in City of Calgary, based on customer reviews.

Research Dataset

As one of the most economically dynamic regions in Canada, namely "the heart of the New West" (Calgary, 2012), City of Calgary is applied as the study area and the adopted social media dataset contains review data of different venues there. In the dataset, text information (e.g., comments) and non-text information (e.g., review count, tags attached such as 'useful', 'funny', 'cool') for 4331 different venues are provided for modelling. Besides, location attributes of these venues are provided for spatial visualization. Figure 1 has displayed the spatial variation of venue stars in the dataset, where more venues have 3 to 5 stars. Overall, peripheral sites, especially the southeast ones, have lower star rating.

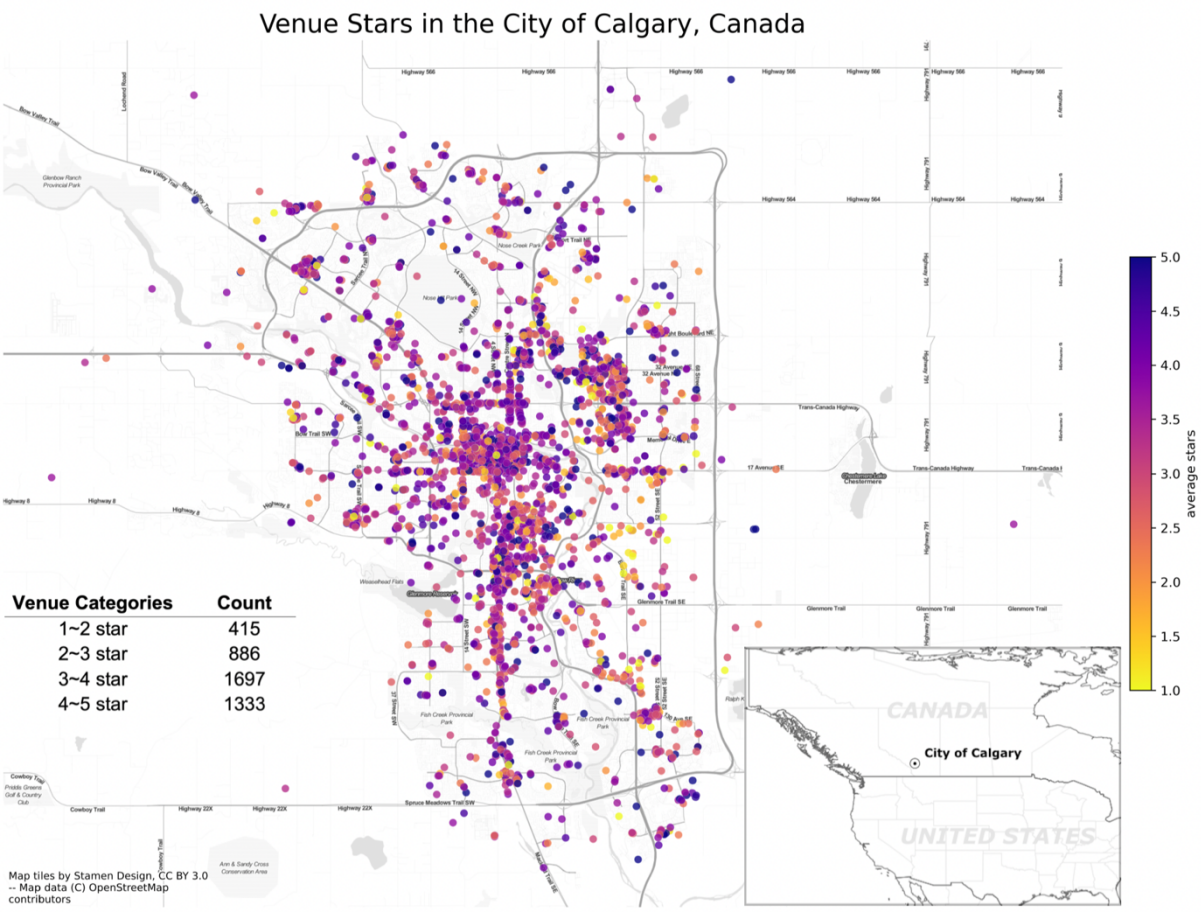


Figure 1: Spatial variation of venue stars in the City of Calgary (Source: Author's own)

Methodology

Considering the methodology, the paper has mutually converted textual and non-textual information and conducted ordinary least squares (OLS) linear regression model and various textual models to predict the venue star ratings and compare their accuracy. For the linear model, review count, tags (e.g., 'useful', 'funny', 'cool') in the dataset have been applied as variables. Apart from that, the length of the reviews and the positive sentiment scores calculated by Valence Aware Dictionary and Sentiment Reasoner (VADER) sentiment analysis over reviews have also been involved. According to Pandey (2018), VADER in Python is a sentiment analysis tool, labeling lexical features (e.g., words) as positive or negative expressions and providing scores based on semantic orientation rules, especially suitable for dealing with social media context. R-squared value of the model is applied to evaluate the model fit and accuracy.

Text analysis models emphasize more on the natural language processing (NLP) techniques with a goal of making computer better understand the textual data. Before building the models, the textual data are pre-processed by removing the uniform resource locators (URLs), mentions, hashtags, ticks, punctuations, numbers and stop words. Besides, because of the potential relationship between the count of different tags and star rating of venues as displayed in Figure 2, in some experimental models, they are converted into text data by multiplying the string-format tags and their logged counts respectively.

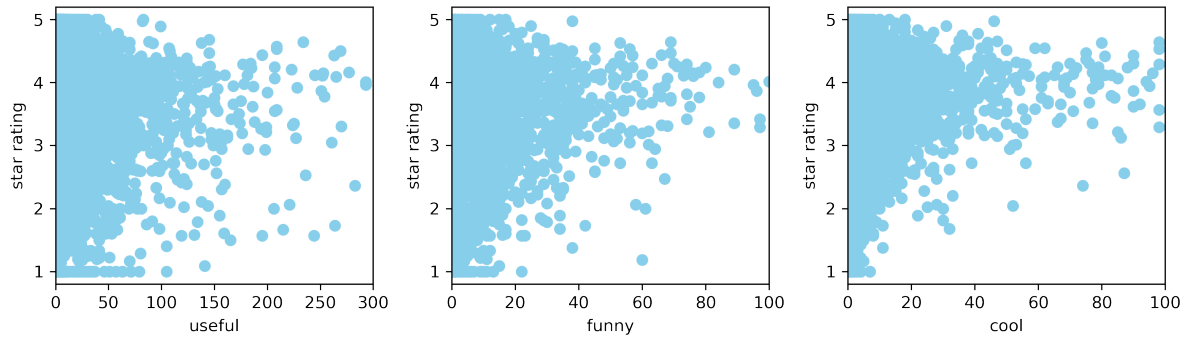


Figure 2: The relationship between number of different tags and venues' star rating (Source: Author's own)

Supervised text classification is applied for the original imbalanced dataset in the textual models. Since NLP packages in Python mainly process individual word (or tokens), there are textual models classify texts into categories based on feature vectors transformed by Count Vectorizer, counting the raw occurrence of different tokens (namely words feature model in Table 1). Besides, another two models notice the different lengths of reviews and adopt Term Frequency, Inverse Document Frequency (TF-IDF) representation to avoid the potential skewness and discrepancies (namely TF-IDF feature model in Table 1). Among them, one of each pair of experimental models would consider the tags attached by customers. In addition, Naive Bayes Classifier is applied to fit the processed tokens and make predictions. To evaluate and compare, F1 score and accuracy score will be calculated. Mentioned by Shung (2018), F1 Score can seek a balance between precision and recall, especially there is an uneven class distribution (detailed formulae attached in Appendix A), while accuracy score can represent the correctly classified samples. Because of the potential deteriorated performance of a standard classifier on an imbalanced dataset (Piri, Delen & Liu, 2018), same models have been conducted on the shrinking balanced dataset by remaining 400 samples for every category.

Results and Discussion

As mentioned in Methodology section, a OLS linear regression model and 8 supervised text classification models are conducted, and Table 1 has summarized the accuracy of each model measured by test dataset (0.2 of the total dataset). As a result, the supervised text classification models have more than 10 percent accuracy than the linear model. Because of the higher accuracy, the subsequent paragraphs mainly focus on the comparison of various text analysis models to predict star rating. Different predictions of test dataset based on different textual models have been combined with their location attributes and mapped in the appendices. By comparison, for the original imbalanced dataset and shrinking balanced dataset, TF-IDF feature model considering the influence of attached tags all get the highest accuracy score (as well as the F1 score), accounting for 0.55 (0.53) and 0.49 (0.45) respectively. Each model has advantages and disadvantages, since there are tradeoffs between the balance and size of the dataset, as well as unknowns about the influence of reviews' length.

Table 1: Accuracy test for applied models (Source: Author's own)

Model	F1 Score	Accuracy	Prediction	Appendix
Linear regression model				
OLS (+sentiment analysis)	-	0.2802 (r^2)	-	-
Textual models with imbalanced dataset				
Words feature model	0.3929	0.4129	Prediction 1	
TF-IDF feature model	0.5249	0.5490	Prediction 2	Appendix B
Words feature model (+tag)	0.3908	0.4106	Prediction 3	
TF-IDF feature model (+tag)	**0.5300**	**0.5536**	Prediction 4	
Textual models with balanced dataset				
Words feature model	0.4270	0.4625	Prediction 5	
TF-IDF feature model	**0.4543**	**0.4938**	Prediction 6	Appendix C
Words feature model (+tag)	0.4280	0.4625	Prediction 7	
TF-IDF feature model (+tag)	**0.4543**	**0.4938**	Prediction 8	

Meanwhile, confusion matrix in Figure 3 and 4 can represent how test data are misclassified by Naive Bayes Classifier into the misclassified categories. For the test data of original imbalanced dataset (with 83, 177, 267, 340 samples respectively), higher star rating can always be correctly classified, especially, 3-4-star venues can be predicted by the TF-IDF feature models accurately, with 70 percent accuracy. Overall, TF-IDF feature models perform much better than words feature models. As shown in Figure 3-a and 3-c, words feature models always underestimate the star rating of the venues, so that several venues have been predicted as the lowest quality ones.

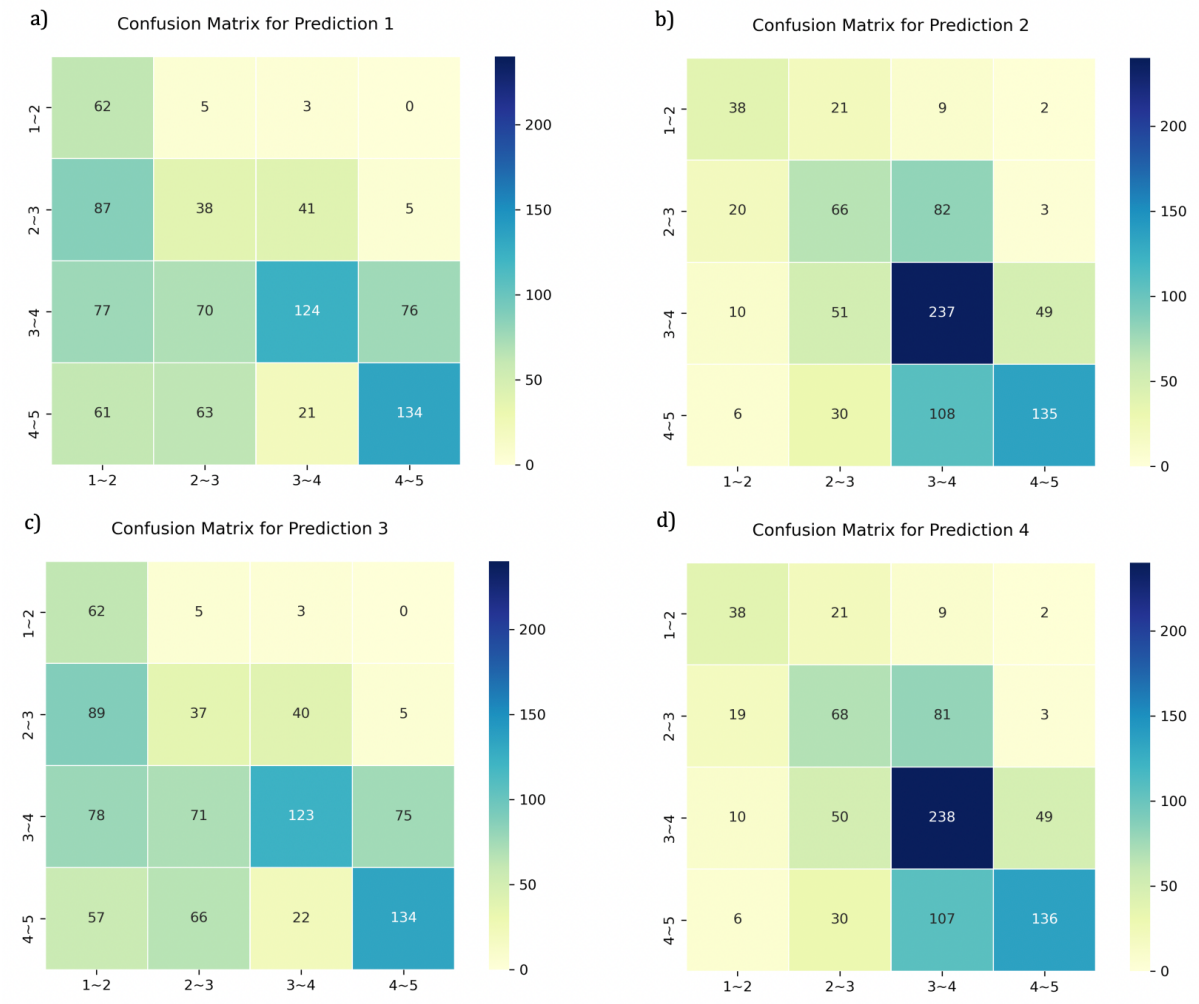


Figure 3: Confusion matrixes for supervised text classification models on imbalanced dataset (Source: Author's own)

For the shrinking balanced dataset, each test dataset has 80 (400*0.2) samples. As displayed in Figure 4, these 4 models have similar performance, where over 70 samples of 1-2-star venues can be predicted correctly. However, over half of 3-4 and 4-5-star venues would be misclassified, while only 12 percent of 2-3-star venues can be identified.

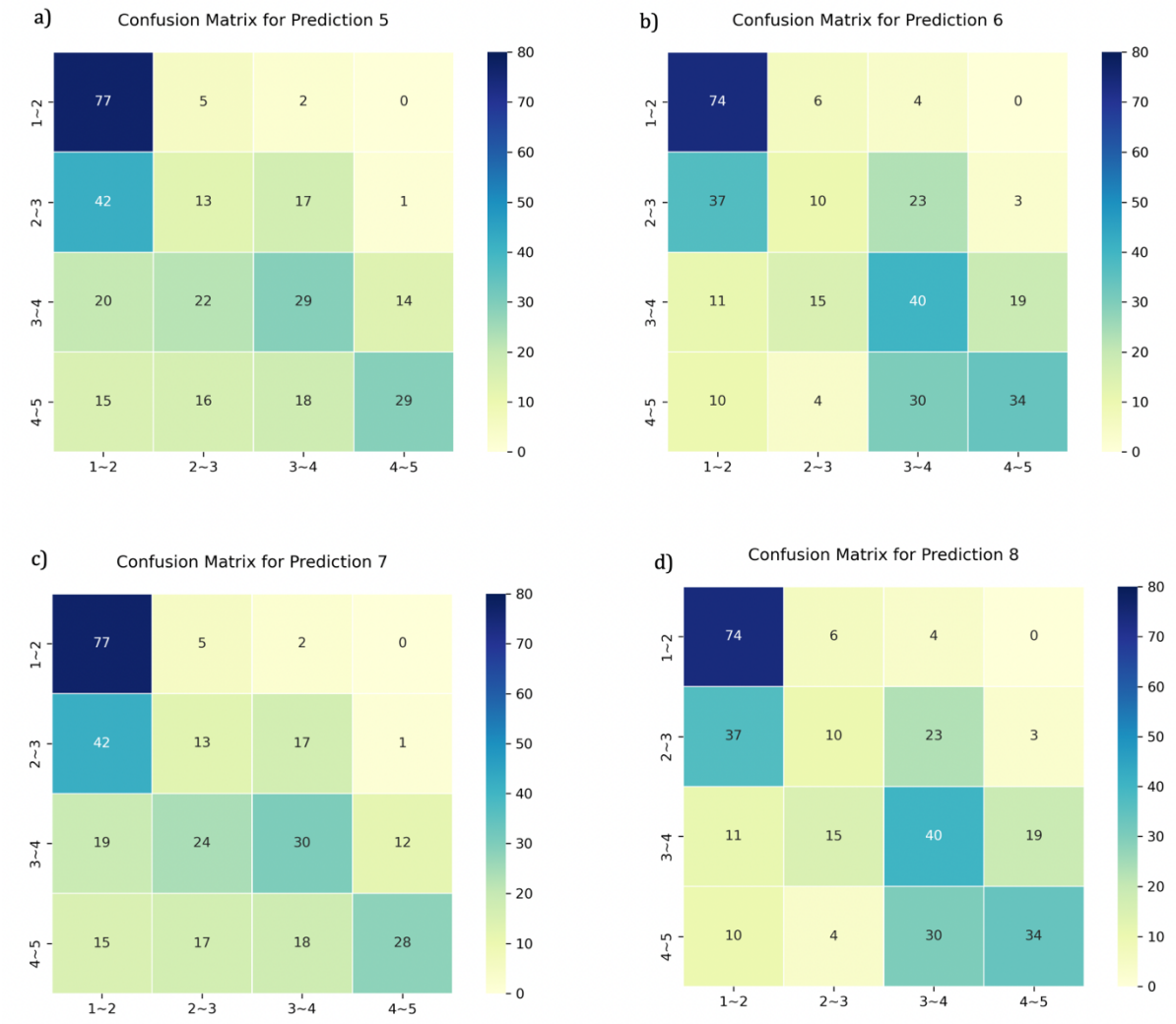


Figure 4: Confusion matrixes for supervised text classification models on the shrinking balanced dataset (Source: Author's own)

In order to further assess the models, the research has visualized their performance (detailed mapping attached in Appendices), with gradient colors from yellow to purple indicating low-star to high-star venues. Figure 5 has represented the actual venue star distribution of test dataset and the best predicted values obtained by TF-IDF feature model with consideration of tags. It is apparent that this specific model can perfectly predict the overall pattern of venue stars based on customer reviews. Similar to the finding by confusion matrixes above, the words feature models for original dataset would misclassify the higher star venues to the lowest-value ones, therefore, much more yellow points can be identified in the mapping of prediction 1 and 3 in Appendix B. Similarly, all the four models for the

processed balanced dataset would also misclassify the higher star venues to the lowest-value ones, although the 1-2-star venues can be perfectly predicted.

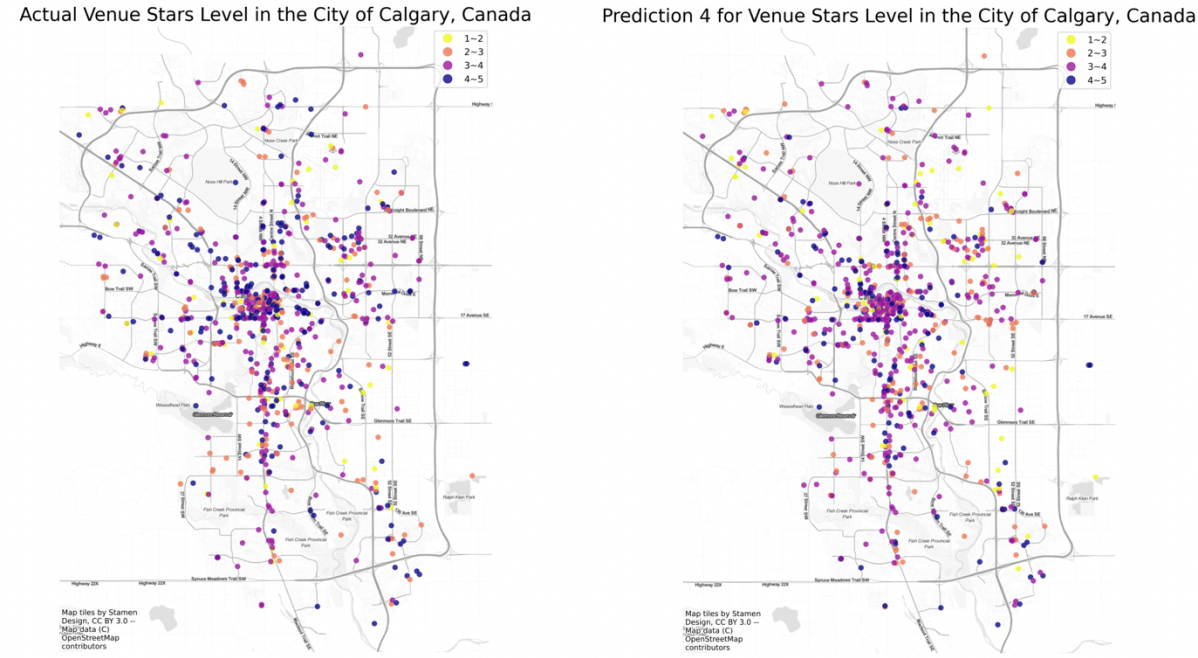


Figure 5: Data visualization of actual venue stars of test dataset and the best prediction (Source: Author's own)

Conclusion

In conclusion, the essay has given an overview of spatial variation of venues' service qualities in the City of Calgary, Canada. Besides, the linear regression model and several textual models have been built and compared, concerning about the tradeoffs between the balance and size of the dataset, as well as the potential influence of reviews' length. It is apparent that the TF-IDF feature model with consideration of tags had better performance and could perfectly predict the overall pattern of venue stars based on customer reviews. Although there are limitations within the models on distinguishing lower star venues, this study can provide inspiration for further review rating prediction research.

Reference List

Calgary (2012). *Calgary, Heart of the New West.* Available from: https://calgaryeconomicdevelopment.com/sites/default/files/pdf/research/reports/whats_new/calgary_advantages_presentation/Calgary_Presentation_August_2012.pdf (Accessed: 9 April 2021).

Chougule, V. & Mulla, A. (2015). 'Predictive Modeling and Sentiment Analysis: Data Mining Approach', *International Research Journal of Engineering and Technology (IRJET),* 2 (8), pp. 163-167.

Horrigan, J. (2008). 'Online shopping', *Pew Internet & American Life Project*. Available from: https://www.pewresearch.org/internet/2008/02/13/online-shopping-2/ (Accessed: 9 April 2021).

Pandey, P. (2018). *Simplifying Sentiment Analysis using VADER in Python (on Social Media Text)*. Available from: https://medium.com/analytics-vidhya/simplifying-social-media-sentiment-analysis-using-vader-in-python-f9e6ec6fc52f (Accessed: 8 April 2021).

Pang, B. & Lee, L. (2005). 'Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales', *Proc. 43rd Meeting of the Association for Computational Linguistics*, pp.115-124.

Piri, S., Delen, D. & Liu, T., (2018). 'A synthetic informative minority over-sampling (SIMO) algorithm leveraging support vector machine to enhance learning from imbalanced datasets', *Decision Support Systems*, 106, pp.15-29.

Shung, K. (2018). *Accuracy, Precision, Recall or F1?* Available from: https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9 (Accessed: 8 April 2021).

Wang et al. (2018). 'Review Rating Prediction on Location-Based Social Networks Using Text, Social Links, and Geolocations', *IEICE transactions on information and systems*, E101.D (9), pp. 2298-2306.

Appendices

Appendix A: Calculation of F1 Score

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$= \frac{True\ Positive}{Total\ Predicted\ Positive}$$

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

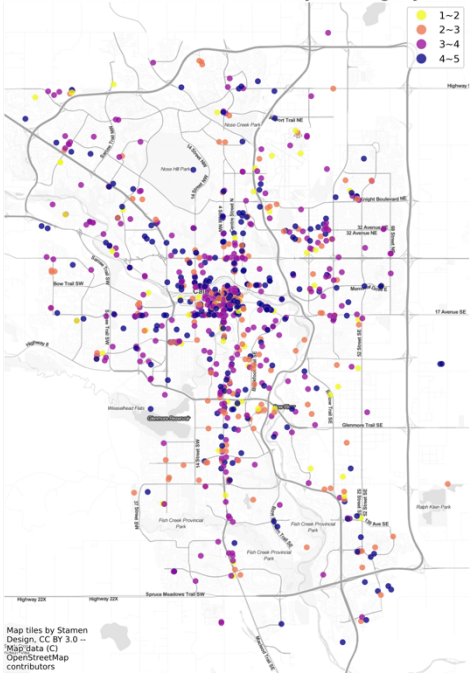
$$= \frac{True\ Positive}{Total\ Actual\ Positive}$$

$$F1 = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

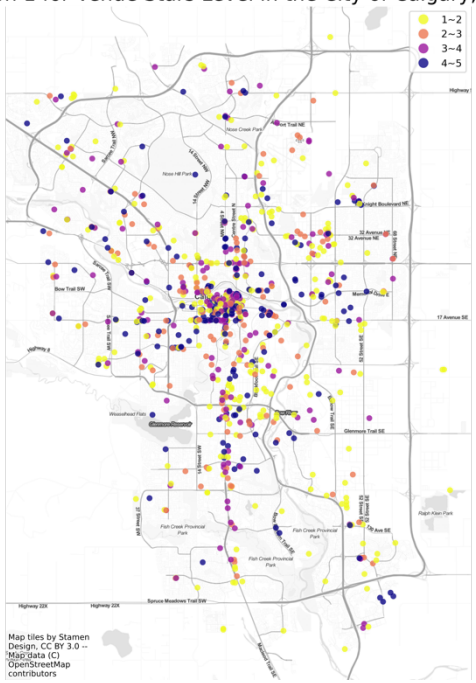
Appendix B:

Text dataset and Predictions
for the Imbalanced Dataset

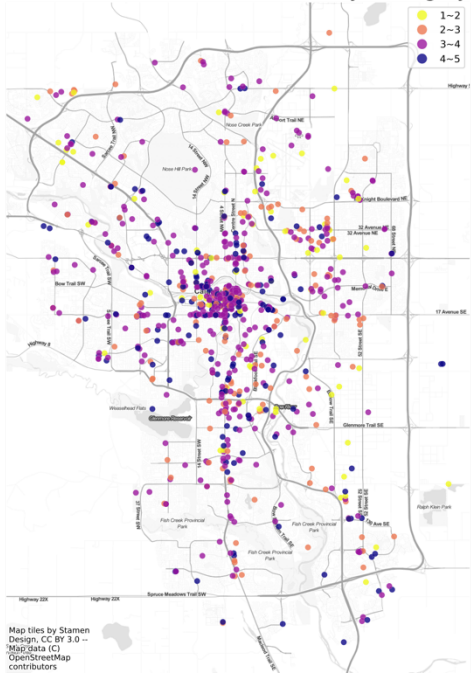
Actual Venue Stars Level in the City of Calgary, Canada



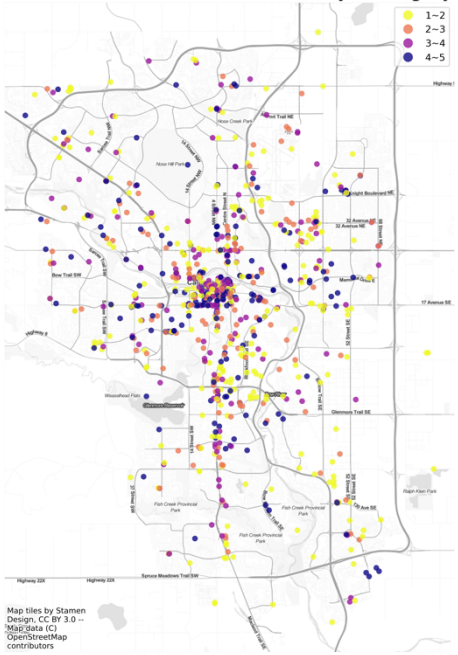
Prediction 1 for Venue Stars Level in the City of Calgary, Canada



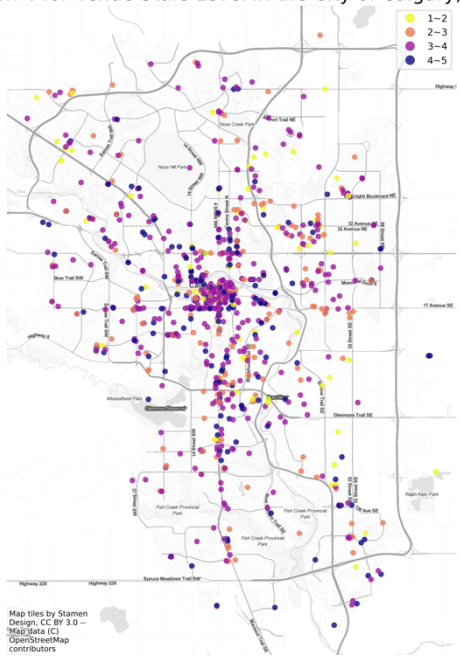
Prediction 2 for Venue Stars Level in the City of Calgary, Canada



Prediction 3 for Venue Stars Level in the City of Calgary, Canada



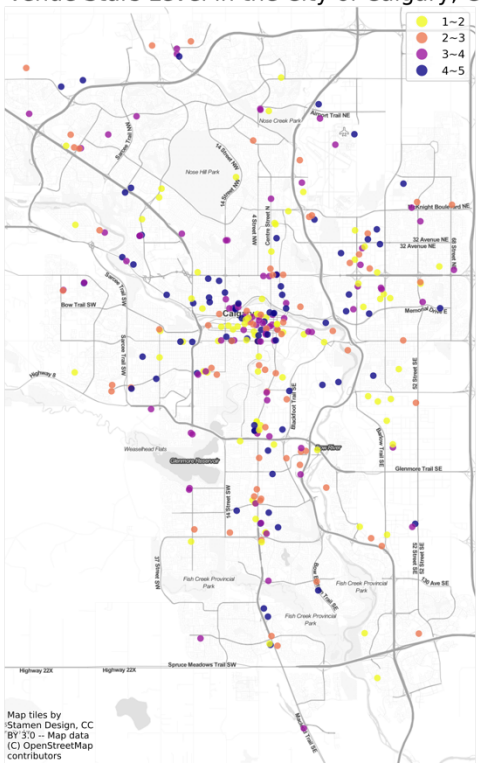
Prediction 4 for Venue Stars Level in the City of Calgary, Canada



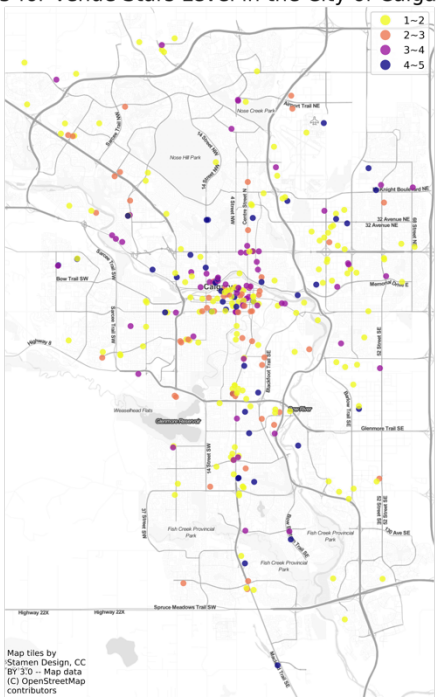
Appendix C:

Text dataset and Predictions
for the Shrinking Balanced Dataset

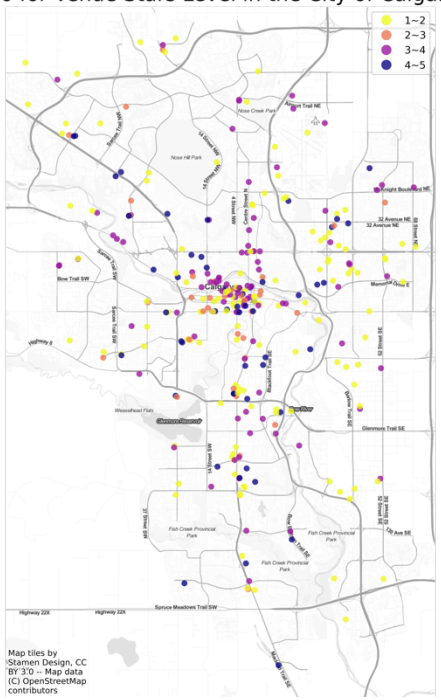
Actual Venue Stars Level in the City of Calgary, Canada



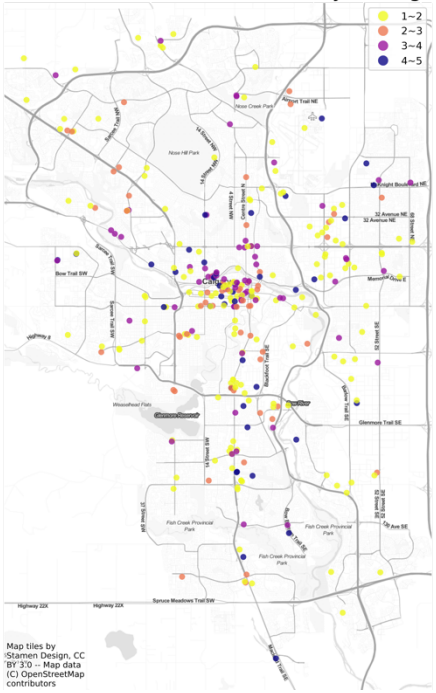
Prediction 5 for Venue Stars Level in the City of Calgary, Canada



Prediction 6 for Venue Stars Level in the City of Calgary, Canada



Prediction 7 for Venue Stars Level in the City of Calgary, Canada



Prediction 8 for Venue Stars Level in the City of Calgary, Canada