Exploratory Analysis of Geolocational Data

A PROJECT REPORT

Submitted by

Rahul Nihalani (21MIM10002)
Shwetank Thakur (21MIM10003)
Abhay Prasad (21MIM10022)
Durgesh Kumar Singh (21MIM10067)
Karthik Pandey (21MIM10068)

in partial fulfillment for the award of the degree of

MASTER OF TECHNOLOGY

In

ARTIFICIAL INTELLIGENCE

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING

VIT BHOPAL UNIVERSITY

KOTHRI KALAN, SEHORE
MADHYA PRADESH - 466114

FEBRUARY 2023

VIT BHOPAL UNIVERSITY, KOTHRI KALAN, SEHORE MADHYA PRADESH – 466114

BONAFIDE CERTIFICATE

Certified that this project report titled **"EXPLORATORY ANALYSIS OF GEOLOCATIONAL DATA"** is the bonafide work of "**RAHUL NIHALANI (21MIM10002), SHWETANK THAKUR (21MIM10003), ABHAY PRASAD (21MIM10022), DURGESH KUMAR SINGH (21MIM10067), KARTHIK PANDEY (21MIM10068)"** who carried out the project work under my supervision. Certified further that to the best of my knowledge the work reported at this time does not form part of any other project/research work based on which a degree or award was conferred on an earlier occasion on this or any other candidate.

PROGRAM CHAIR
Dr. AVR Mayuri,
Assistant Professor
School of Computing Science and Engineering
VIT BHOPAL UNIVERSITY

PROJECT GUIDE
Dr. Siddharth Singh Chauhan,
Assistant Professor
School of Computing Science and Engineering
VIT BHOPAL UNIVERSITY

The Project Exhibition I Examination is held on **February 16th,2023.**

ACKNOWLEDGEMENT

First and foremost, I would like to thank the Lord Almighty for His presence and immense blessings throughout the project work.

I wish to express my heartfelt gratitude to **Dr. AVR Mayuri**, Program Chair, Artificial Intelligence, School of Computing Science and Engineering for much of his valuable support encouragement in carrying out this work.

I would like to thank my internal guide **Dr. Siddharth Singh Chouhan**, for continually guiding and actively participating in my project, giving valuable suggestions to complete the project work.

I would like to thank all the technical and teaching staff of the School of Aeronautical Science, who extended directly or indirectly all support.

Last, but not least, I am deeply indebted to my parents who have been the greatest support while I worked day and night for the project to make it a success.

LIST OF ABBREVIATIONS

1. DBSCAN – Density Based Spatial Clustering of Applications with Noise
2. HDBSCAN – Hierarchical Density Based Spatial Clustering of Applications with Noise
3. Eps - epsilon (radius of cluster)
4. EDA – Exploratory Data Analysis
5. ML – Machine Learning
6. GPS – Global Positioning System

LIST OF FIGURES AND GRAPHS

ABSTRACT

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations. Main purpose of this project is to help people who are travelling to an unknown place providing them information regarding the place they are searching for, according to their relevant budget, distance, nearby facilities etc.

In this Project, we have used K-means clustering algorithm on the provided datasets where, Elbow method is used to get the optimal value of K (No. Of clusters). We have also used DBSCAN algorithm on the provided datasets where, Elbow method is used to get the optimal value of eps (Epsilon) with minimal sample size of 10. We have used HDBSCAN algorithm on the provided datasets where, Silhouette score is used to get the optimal value for minimum cluster size.

Table of Contents

CHAPTER-1

PROJECT DESCRIPTION AND OUTLINE

1.1 Introduction

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations. It is basically used to gather information of things, places, and many more.

1.2 Motivation for the work

Motivation of doing this project came in our mind when we heard from our friends and relatives as they faced difficulty in finding desired places for foods, lodging etc. whenever they go to some unknown place on vacation or trip. So, we thought of making such application which will help them in getting the data as quickly as possible and can go to any place without hesitation.

1.3 About Introduction to the project including techniques

In our project, to perform the exploratory data analysis we used algorithm such as K-Means Clustering, DB-SCAN, HDB-SCAN. Using all these algorithms we will make a cluster of reports and then after we will implement it to be used by the user on the front end of our application.

1.4 Problem Statement

This project uses K-Means clustering to find the best accommodations for immigrants by classifying immigrant accommodations based on facility, budget, and location preferences. Obtain, clean, analyze, and perform k-means clustering on geolocation data to recommend accommodations for immigrants in cities.

1.5 Objective of the Project

Our objective was to make a system which is efficient and accurate to provide a proper data of places which user are asking for and show them the most efficient and desired data according to the search of the user and it will help the tourist in enjoying their trip or vacation to the fullest and can visit new places without any worries about food and shelter.

1.6 Organization of the Project

Firstly, all the group members did research on the topic Exploratory Data Analysis and gathered as much information as we can get and then we divided our member in coding group, execution group and data gathering group. Each person was assigned a task and work was done under the group leader which was multilaterally selected. Our group did the work under the direction/supervision of a mentor/guide.

CHAPTER - 2

Related work and Investigation

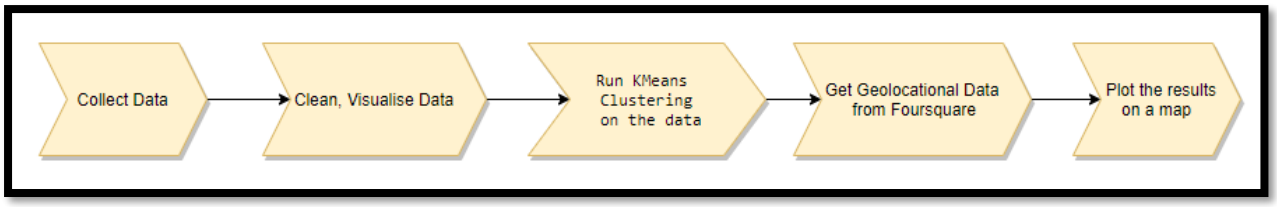
2.1 Introduction:

We apply k-means, DBSCAN and HDBSCAN algorithms to find the best accommodation for tourist in an area (or any specific place) by classifying accommodation for incoming tourist based on their preferences on amenities, budget and proximity to the location.

Implementing the project will take you through the daily life of a data science engineer - from data preparation on real-life datasets, to visualizing the data and running machine learning algorithms, to presenting the results.

In the fast-moving, effort-intense environment that the average person inhabits, it’s a frequent occurrence that one is too tired to fix oneself a home-cooked meal. And of course, even if one gets home-cooked meals every day, it is not unusual to want to go out for a good meal every once in a while, for social/recreational purposes. Either way, it’s a commonly understood idea that regardless of where one lives, the food one eats is an important aspect of the lifestyle one leads.

Food delivery apps aside, managers of restaurant chains and hotels can also leverage this information. This project is a good start for beginners and a refresher for professionals who have dabbled in python / ML before.



2.2 Existing Approaches/Methods:

a) Method 1:

The *k*-**means clustering** method is an unsupervised machine learning technique used to identify clusters of data objects in a dataset. There are many different types of clustering methods, but *k*-means is one of the oldest and most approachable. These traits make implementing *k*-means clustering in Python reasonably straightforward, even for novice programmers and data scientists.

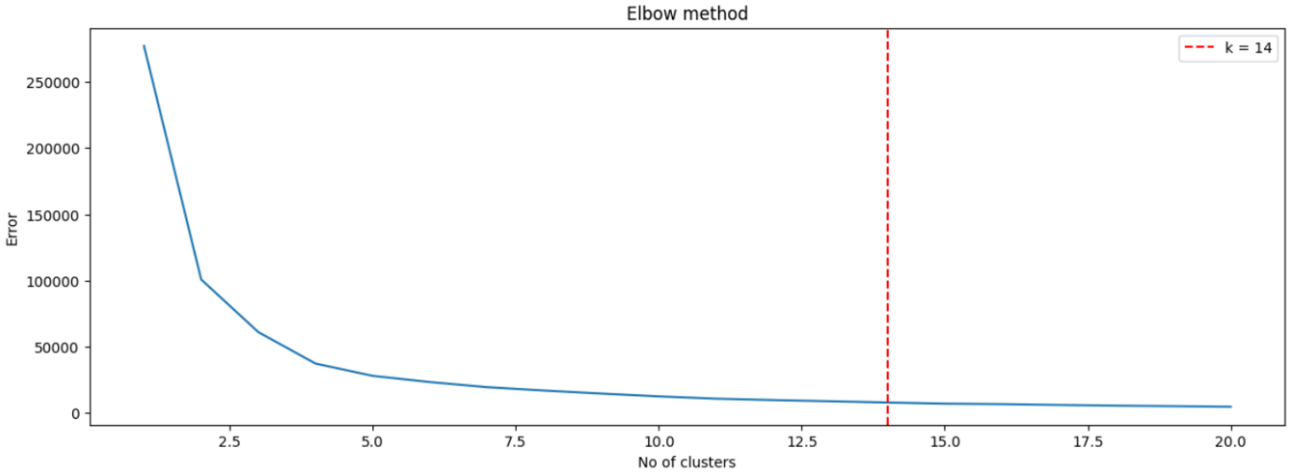


Figure 1.a

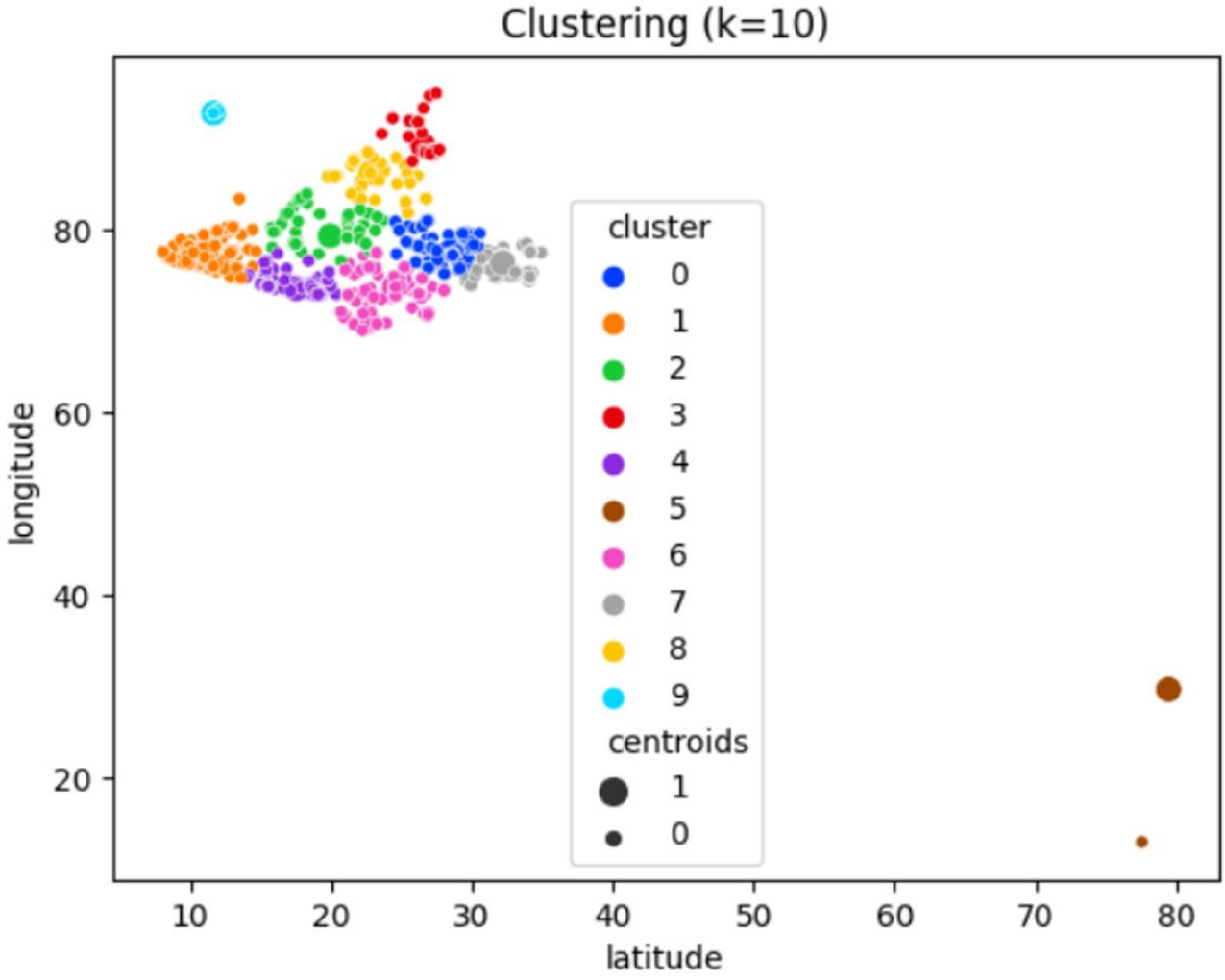


Figure 1.b

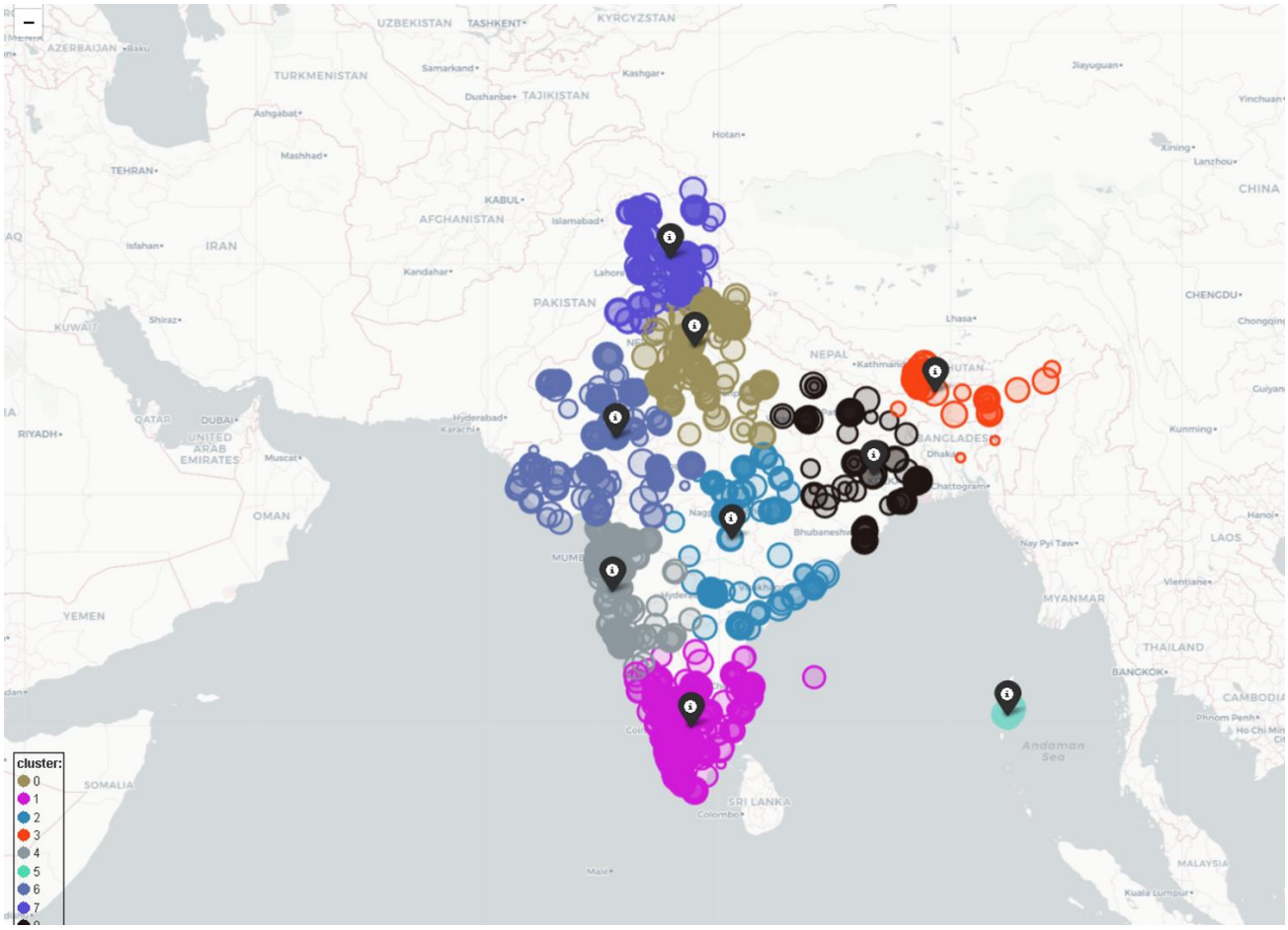


Figure 1.c

b) Method 2:

The **DBSCAN** stands for Density-Based Spatial Clustering of Applications with Noise. It is a popular unsupervised machine-learning algorithm used for clustering, unlike other clustering algorithms such as k-means, DBSCAN does not require the number of clusters to be specified in advance.

Advantages

- Can identify clusters of different shapes and sizes
- Can identify outliers and noisy data Scalable to large datasets

Limitations

- Sensitive to the choice of eps and mints values
- Can struggle with high-dimensional data Not suitable for real-time applications

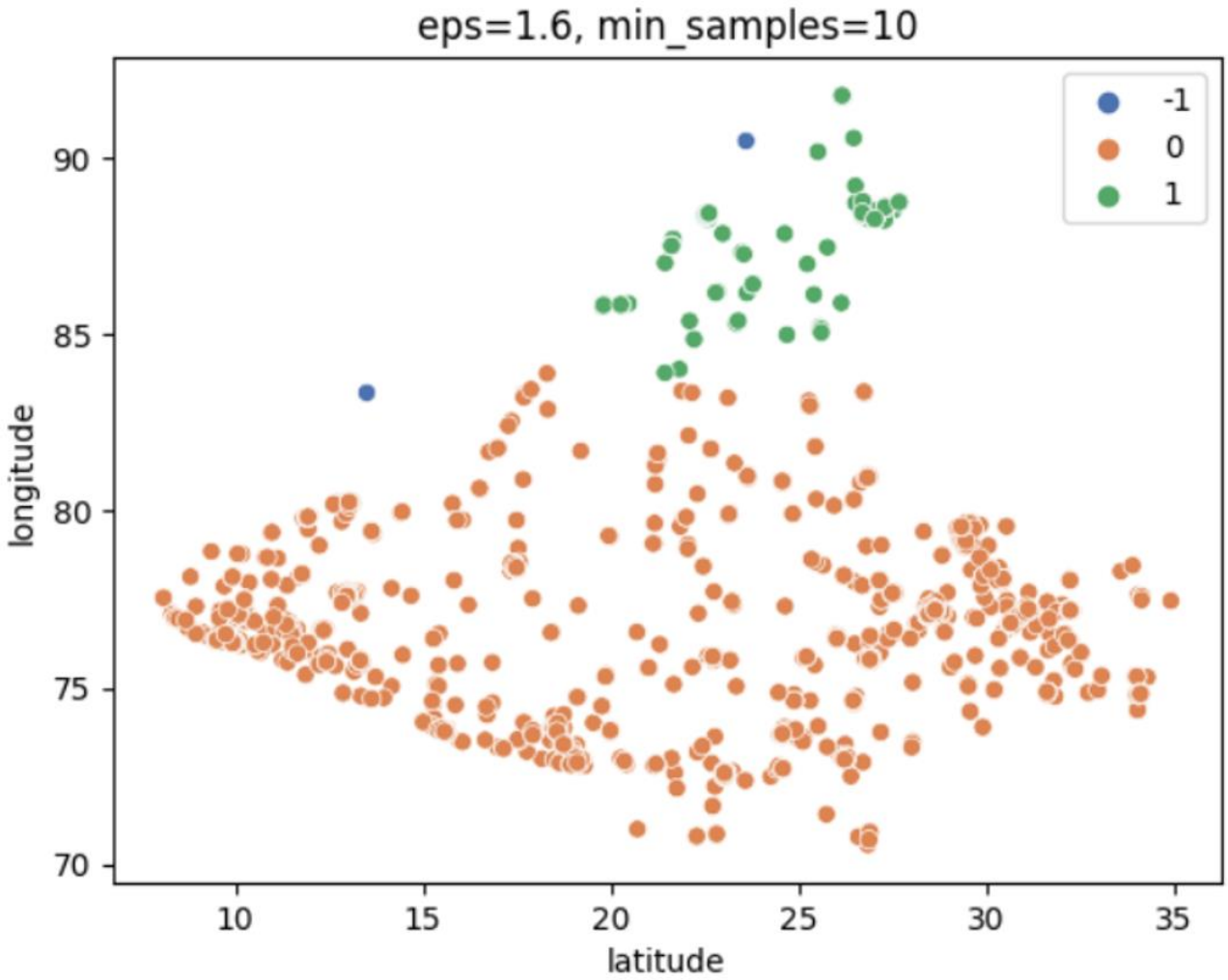


Figure 2.a

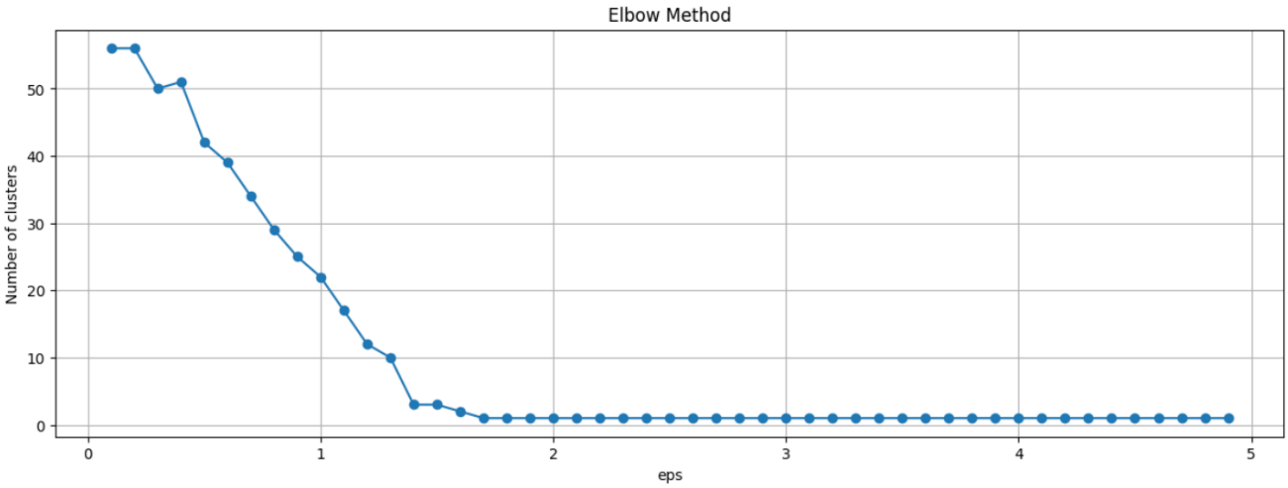


Figure 2.b

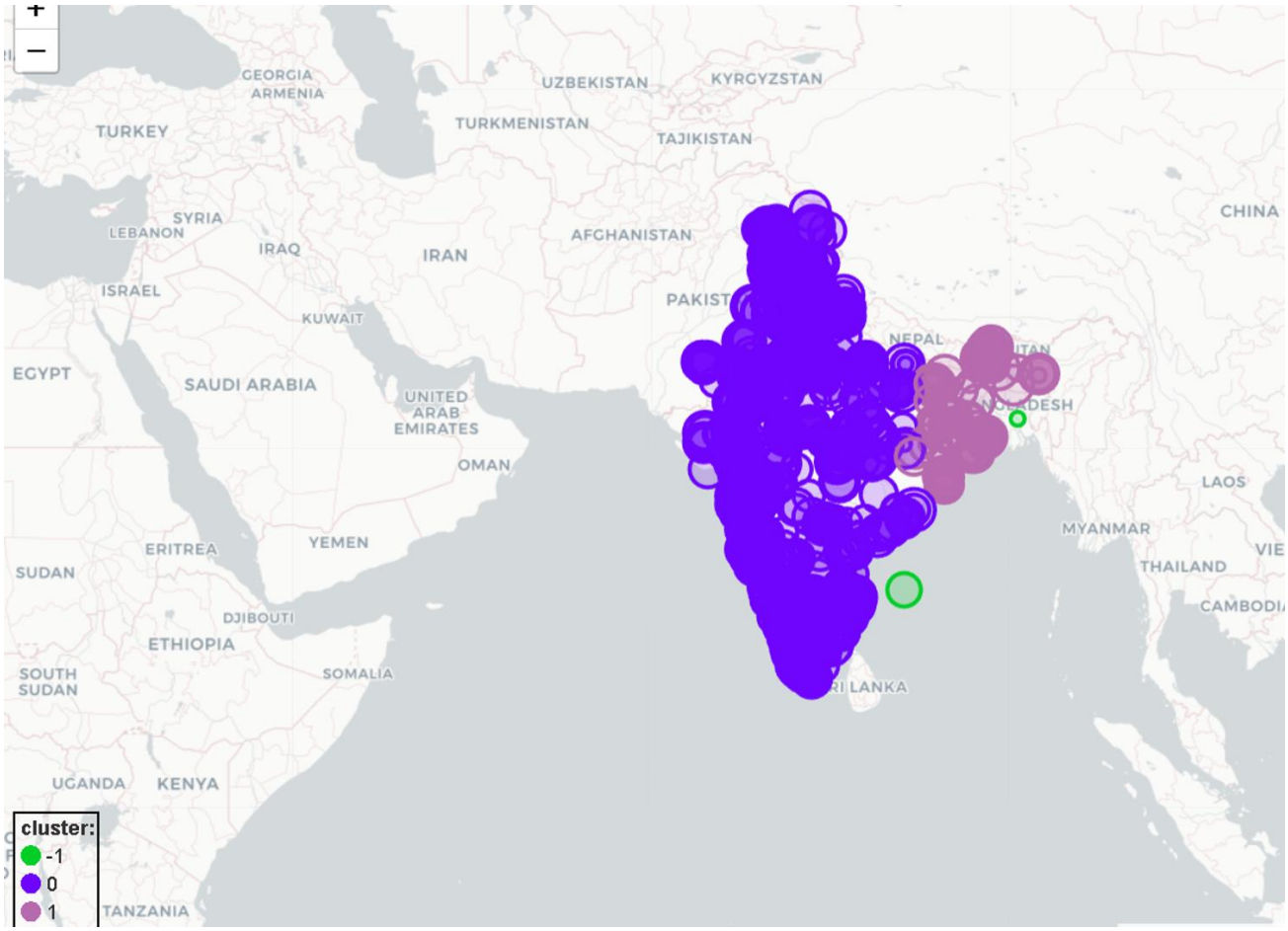


Figure 2.c

c) Method 3:

The **HDBSCAN** is a density-based clustering algorithm that seeks to find clusters of varying densities in a dataset. It is an improved version of the popular DBSCAN algorithm, designed to address its limitations.

Advantages

- HDBSCAN can identify clusters with varying densities, making it useful for datasets with dense and sparse clusters.
- It is computationally efficient, as it only requires a single scan of the data, making it well suited for large datasets.

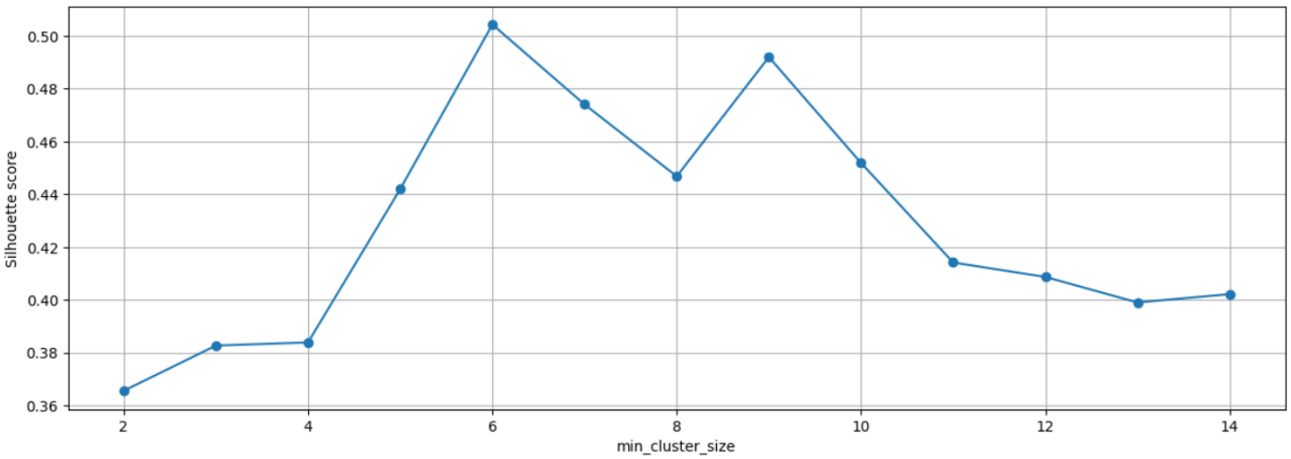


Figure 3.a

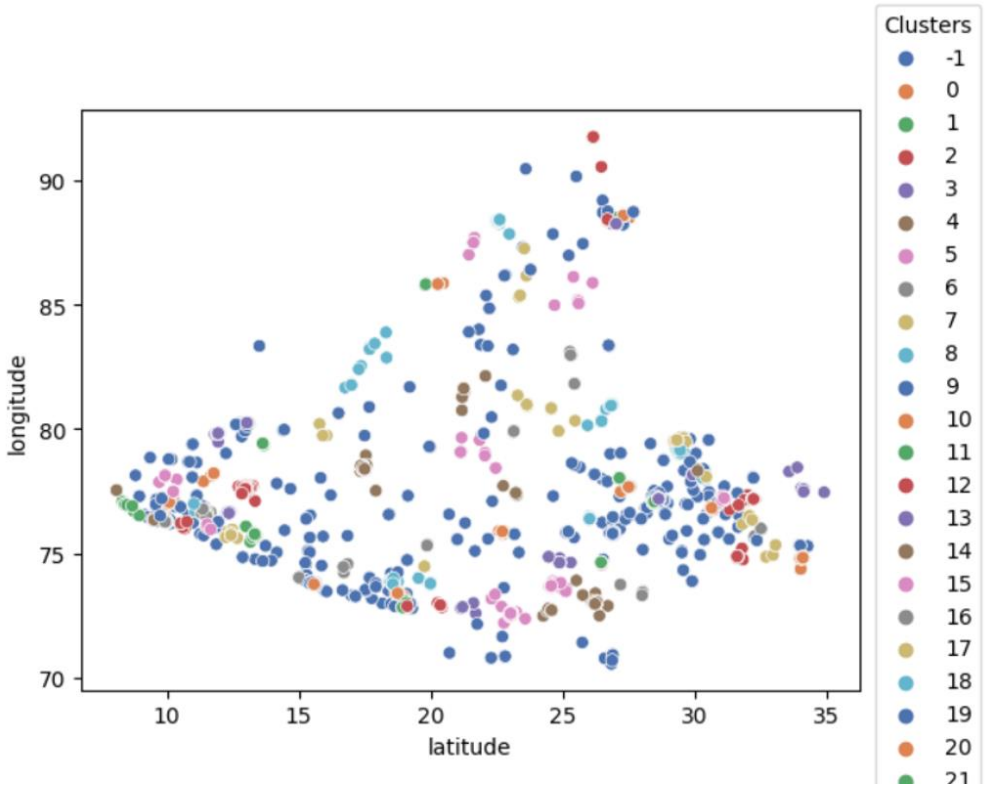


Figure 3.b

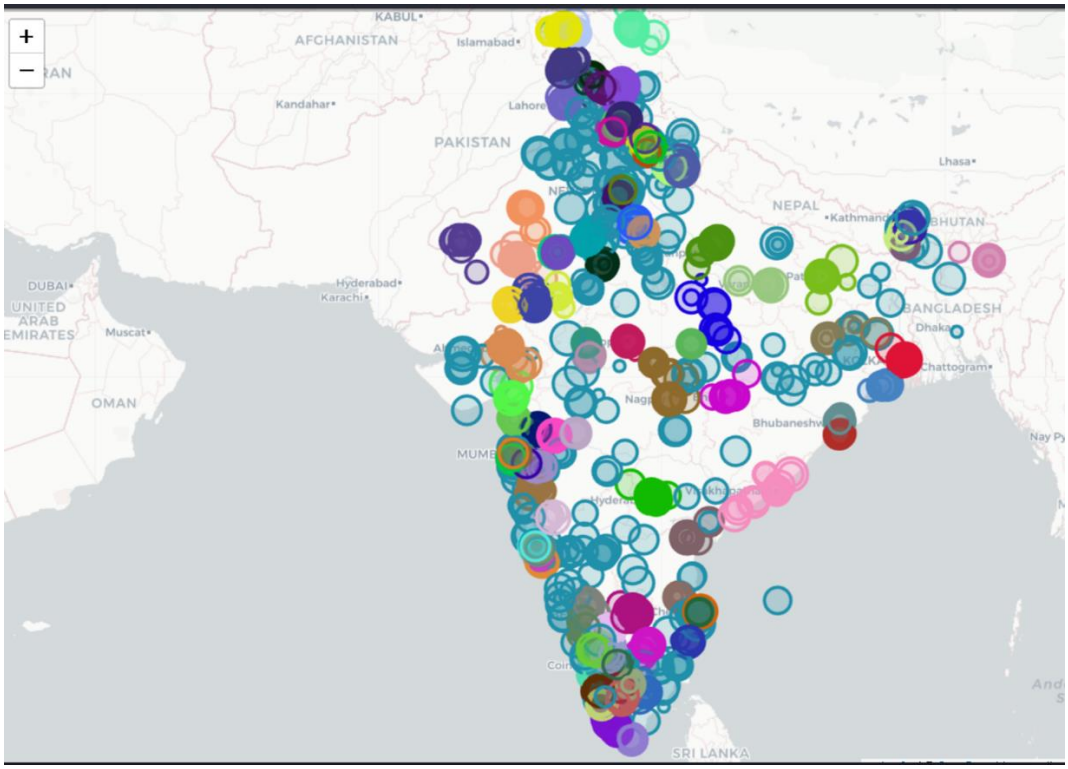


Figure 3.c

2.3 Summary

Clustering is the task of grouping the elements such that observations of same group are more similar to each other than those in another group.

Affinity Propagation is a graph-based algorithm that assigns each observation to its nearest exemplar

Geolocational Analysis is the analysis that processes Satellite images, GPS coordinates and Street addresses and apply to geographic models.

CHAPTER-3

REQUIREMENTS ARTIFACTS

3.1 Introduction

In our project, we have used a machine-learning model which requires a medium-high demand of hardware components. As per our understanding a medium-high end pc is required to run the model seamlessly. Higher the quality of pc, the lesser the time it will take train the model and load it.

3.2 Hardware and Software requirements

Below is the list of minimum and maximum hardware requirements:

1. Minimum requirements:
    a. CPU: i3 11^{th} generation laptop
    b. GPU: Integrated GPU
    c. RAM: 8GB
    d. Storage: HDD or SSD
    e. OS: Windows 10/11
2. Recommended requirements:
    a. CPU: i5 10^{th} generation laptop or above
    b. GPU: Dedicated Graphics Card
    c. RAM: 8GB or above
    d. Storage: SSD
    e. OS: Windows 10/11

3. Hardware used:
    a. CPU: Ryzen 7-5800H laptop
    b. GPU: RTX 3050 laptop or GTX 1650 laptop
    c. RAM: 8/16GB
    d. Storage: M.2 nvme
    e. OS: Windows 10/11

Below is the list of Software requirements:

a. VS code
b. Python and its dependencies

3.3 Specific Project Requirements

3.3.1 Data Requirement

We need a dataset, with latitude and longitude coordinates to train our model. The data set is taken from Kaggle. The data sets are namely “Indian hotels on Goibibo” and “Hotels on Make My Trip”. The hotels listed on Goibibo are 33,344 hotels and the hotels listed on Make My Trip are 1,65,000 hotels. The count of rows and columns of the data set “Indian hotels on Goibibo” are 4000 and 36 respectively before cleaning and 2416 and 5 respectively after cleaning. The count of rows and columns of the data set “Hotels on Make My Trip” are 20036 and 33 respectively before cleaning and 151 and 5 respectively after cleaning.

The data set should contain the rows and columns of the original as well as the cleaned dataset after dropping the missing data. The most crucial columns required for our model are latitude and longitude to pinpoint the accurate address of the entered data.

3.3.2 Functions Requirements

The actions which are recognized by the Machine-Learning model should be displayed on the Map. There should be a pre-defined action for which the model will search and give the output accordingly.

3.3.3 Look and feel requirements

There should be an output map on the screen which shows the address of the desired location detected by the model. The screen should not be bulky or filled with unnecessary things, otherwise the user will feel overwhelmed.

3.4 Summary

Overall, the system on which the model is to be trained and run should be a medium-high end pc. More the quality of PC, the better it will perform. Talking about the performance, model should be able to detect the coordinates and display the right address on the map. The interface should be clean and should only contain the required information.

CHAPTER – 4

DESIGN METHODOLOGY AND ITS NOVELTY

4.1 Goal and Methodology

Goal: The goal of the exploratory analysis of geolocational data is to uncover patterns, relationships, and trends in the data that may not be immediately obvious. This process helps us gain a better understanding of the data and identify areas for further investigation.

Methodology:

1. Data collection: Obtain geolocational data, such as GPS coordinates, addresses, or zip codes.
2. Data cleaning: clean data for errors and inconsistencies.
3. Visualization: Create maps and heatmaps to identify patterns.
4. Descriptive statistics: Calculate summary statistics.
5. Cluster analysis: Group similar observations based on geolocational data.
6. Spatial analysis: Analyse relationships between different locations.
7. Further investigation: Identify areas for further analysis based on results.

4.2 Functional Modules design and analysis

a.) Pandas – It is used for data manipulation, cleaning and preprocessing in exploratory analysis.

b.) NumPy – It is used for numerical computing and efficient array processing in exploratory analysis.

c.) Matplotlib – It is used for data visualization and generating plots in exploratory analysis.

d.) Sklearn – Scikit-learn (sklearn) is used for machine learning algorithms and model building in exploratory analysis.

e.) SciPy – It is used for scientific computing and technical computing in exploratory analysis.

f.) Seaborn – It is used for data visualization and generating statistical graphics in exploratory analysis.

g.) Missingno – It is used for visualizing missing values in a dataset in exploratory analysis.

h.) Folium – It is used for creating interactive maps in exploratory analysis of geolocational data.

i.) Geopy – It is used for geocoding and distance calculation in exploratory analysis of geolocational data.

j.) Requests – Library for sending HTTP requests and handling responses.

4.3 Software and architectural designs

The software design for exploratory analysis of geolocational data using ML involves using data science tools, machine learning libraries, and data visualization libraries in a scalable and modular architecture. The design should enable parallel processing and have a user-friendly interface for easy data manipulation and exploration. The aim is to handle large datasets

efficiently and provide interactive maps and visualizations to uncover insights in geolocational data.

4.4 User Interface Designs

The user interface design for exploratory analysis of geolocational data should allow for easy data collection and cleaning, and provide interactive visualizations for pattern identification. The interface should also facilitate the calculation of descriptive statistics and the ability to perform cluster and spatial analysis. A user-friendly interface that effectively communicates results and allows for further investigation is key for the success of the exploratory analysis.

4.5 Summary

The methodology for exploratory analysis of geolocational data using machine learning involves obtaining a dataset of geolocational information, cleaning and preprocessing the data, and utilizing machine learning techniques such as cluster analysis and spatial analysis to identify patterns and relationships in the data. The novelty in this approach lies in the use of advanced machine learning algorithms to perform in-depth analysis and uncover previously unseen insights. This method can also provide a more automated and efficient way of exploring geolocational data compared to traditional methods.

CHAPTER – 5

TECHNICAL IMPLEMENTAION AND ANALYSIS

5.1 Technical coding and code solutions

In exploratory analysis of geolocation data using ML, data is collected, cleaned and transformed into a format suitable for analysis. Popular Python libraries like pandas, NumPy, and scikit-learn (sklearn) are used for data manipulation and machine learning techniques such as clustering and spatial analysis. Visualization tools like matplotlib and seaborn are used for creating maps and heatmaps.

5.2 Working layout of Forms

Forms play a crucial role in data collection for exploratory analysis of geolocation data in ML. They allow for easy input of geolocational information such as GPS coordinates, addresses, or zip codes. The forms should have clear and concise instructions to ensure accurate data collection. The layout should be user-friendly, with an intuitive and organized structure to facilitate data entry.

5.3 Performance analysis

Performance analysis evaluates the accuracy and efficiency of the model on geolocational data using metrics such as R-squared, mean absolute error, and cross-validation scores. It helps to

identify the strengths and weaknesses of the model, and guide further improvements.

5.4 Summary

Technical implementation involves using various programming languages and libraries to clean, visualize and analyze geolocation data. Performance analysis involves evaluating the model's accuracy and efficiency, and optimizing it using techniques such as feature selection and model tuning. The aim is to obtain meaningful insights and patterns from the data.

Chapter – 6

PROJECT OUTCOME AND APPLICABILITY

6.1 Key implementation outline of the system

EDA stands for exploratory data analysis, which is an approach in data science that involves visualizing and analyzing data to gain insights and make decisions. It helps to understand the nature of data and identify trends, patterns, and anomalies in the dataset.

6.2 Significant project outcome

We managed to plot the address on map on the basis of site review rating. User can filter out on which basis the hotels should be displayed on the map.

6.3 Project applicability on Real-world application

Exploratory analysis of geolocational data can used by travelers mainly. If a person is new to the city and wants to get location of accommodation, restaurant, etc., by using a website or application, a person filters out the locations near him/her.

Chapter – 7

CONCLUSIONS AND RECOMMENDATION

7.1 Outline

- Our system is mainly focused on giving accurate location to the user.
- Dataset used is from Kaggle.
- System should give the output based of options filtered by user.
- Standard data is currently for India, and for accommodations only.
- The system is performing tasks in a convenient manner.

7.2 Limitations/Constraints of the System

- EDA cannot make casual inferences.
- EDA is subjective.
- EDA does not account for missing data.
- EDA does not account for outliers.
- EDA may not be suitable for large datasets.

7.3 Future Enhancements

Further on, we can add more data about restaurants, gyms, parks and many more. For now, we have information regarding only accommodations. On adding more options, user will be more comfortable as everything will be available at one place.

REFERENCES

1. Dipankar Chowdhury, Sigve Hovda, Bjørnar Lund, Analysis of cuttings concentration experimental data using exploratory data analysis, Geoenergy Science and Engineering, Volume 221, 2023, 111254, ISSN 2949-8910, https://doi.org/10.1016/j.petrol.2022.111254.
2. Fatima Zahra Fagroud, Lahbib Ajallouda, El Habib Ben Lahmar, Hicham Toumi, Khadija Achtaich, Sanaa El Filali, IOT Search Engines: Exploratory Data Analysis, Procedia Computer Science, Volume 175, 2020, Pages 572-577, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2020.07.082.
3. O. Aboulola, "A literature review of spatial location analysis for retail site selection," Journal of the Association for Information Systems, 08 2017.
4. Ojokoh, Bolanle, O. Catherine, Olayemi, Babalola, Asegunloluwa, Eyo, and Eyo, "A user-centric housing recommender system," Information Management and Business Review, vol. 10, pp. 17–24, 09 2018.
5. B. Kumar and N. Sharma, "Approaches, issues and challenges in recom- mender systems: A systematic review," Indian Journal of Science and Technology, vol. 9, 12 2016.
6. Jun, H. Jong, Kim, J. Hee, Rhee, D. Young, Chang, and S. Woo, ""seoulhouse2vec": An embedding-based collaborative filtering housing recommender system for analyzing housing preference," Sustainability, vol. 2, no. 7, 2020. https://www.mdpi.com/2071-1050/12/17/6964
7. Sakthivel, M., J. Udaykumar, and V. Saravana Kumar. "Progressive AODV: A Routing Algorithm Intended for Mobile Ad-Hoc Networks." International Journal of Engineering and Advanced Technology (IJEAT), ISSN: 2249-8958, Vol.9 no.2, PP: 70-74, 2019
8. Shishehchi, Saman, Banihashem, Seyed, M. Zin, N. Azan, M. Noah, and S. Azman, "Ontological approach in knowledge-based recommender system to develop the quality of e-learning system," Australian Journal of Basic and Applied Sciences, vol. 6, pp. 115–123, 02 2012.
9. W. Yang, X. Wang, J. Lu, W. Dou and S. Liu, "Interactive steering of hierarchical clustering", IEEE Transactions on Visualization and Computer Graphics, vol. 27, no. 10, pp. 3953-3967, 2020.

10. J.A. Dos Santos, T.I. Syed, M.C. Naldi, R.J. Campello and J. Sander, "Hierarchical density-based clustering using MapReduce", IEEE Transactions on Big Data, vol. 7, no. 1, pp. 102-114, 2019.
11. Y. Chen, J. Hu, Y. Xiao, X. Li and P. Hui, "Understanding the user behavior of foursquare: A data-driven study on a global scale", IEEE Transactions on Computational Social Systems, vol. 7, no. 4, pp. 1019-1032, 2020.
12. S. Al-Dabooni and D. Wunsch, "Model order reduction based on agglomerative hierarchical clustering", IEEE transactions on neural networks and learning systems, vol. 30, no. 6, pp. 1881-1895, 2018.
13. D. Cheng, Q. Zhu, J. Huang, Q. Wu and L. Yang, "A novel cluster validity index based on local cores", IEEE transactions on neural networks and learning systems, vol. 30, no. 4, pp. 985-999, 2018. A. Psyllidis, J. Yang and A. Bozzon, "Regionalization of social interactions and points-of-interest location prediction with geosocial data", IEEE Access, vol. 6, pp. 34334-34353, 2018.
14. Y.M. Nemani, R. Yadav, M. Patki, O. Padave and M.M. Bhelande, "City Tour Traveller: Based on FourSquare API", City, vol. 5, no. 04, 2018.
15. P. Patel, B. Sivaiah and R. Patel, "Approaches for finding Optimal Number of Clusters using K-Means and Agglomerative Hierarchical Clustering Techniques", 2022 International Conference on Intelligent Controller and Computing for Smart Power (ICICCSP), pp. 1-6, 2022, July.
16. Adams, B. and Janowicz, K. (2012). On the geo-indicativeness of non-georeferenced text. In Proceedings of Sixth International AAAI Conference on Weblogs and social media, ICWSM ’12, Dublin, Ireland.
17. Amitay, E., Har’El, N., Sivan, R., and Soffer, A. (2004). Web-a-where: geotagging web content. In Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR ’04, pages 273–280, Sheffield, United Kingdom. ACM.
18. Backstrom, L., Kleinberg, J., Kumar, R., and Novak, J. (2008). Spatial variation in search engine queries. In Proceeding of the 17th international conference on World Wide Web, WWW ’08, pages 357–366, Beijing, China. ACM.
19. Backstrom, L., Sun, E., and Marlow, C. (2010). Find me if you can: improving geographical prediction with social and spatial proximity. In Proceedings of the 19th international conference on World wide web, WWW ’10, pages 61–70, Raleigh, North Carolina, USA. ACM.

20. Bauer, S., Noulas, A., Seaghdha, D. O., Clark, S., and Mascolo, C. (2012). Talking places: Modelling and analysing linguistic content in foursquare. In Proceedings of the ASE/IEEE International Conference on Social Computing, SocialCom 2012, Amsterdam, Netherlands.
21. Bilhaut, F., Charnois, T., Enjalbert, P., and Mathet, Y. (2003). Geographic reference analysis for geographic document querying. In Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references - Volume 1, HLT-NAACL-GEOREF '03, pages 55–62. Association for Computational Linguistics.
22. Buyukokkten, O., Cho, J., Garcia-Molina, H., Gravano, L., and Shivakumar, N. (1999). Exploiting geographical location information of web pages. In ACM SIGMOD Workshop on The Web and Databases (WebDB'99), pages 91–96.
23. Cheng, Z., Caverlee, J., and Lee, K. (2010). You are where you tweet: a content-based approach to geo-locating twitter users. In Proceedings of the 19th ACM international conference on Information and knowledge management, CIKM '10, pages 759–768, Toronto, ON, Canada. ACM.
24. Crandall, D. J., Backstrom, L., Huttenlocher, D., and Kleinberg, J. (2009). Mapping the world's photos. In Proceedings of the 18th international conference on World wide web, WWW '09, pages 761–770, Madrid, Spain. ACM.
25. Dalvi, N., Kumar, R., and Pang, B. (2012). Object matching in tweets with spatial models. In Proceedings of the fifth ACM international conference on Web search and data mining, WSDM '12, pages 43–52, Seattle, Washington, USA.
26. ACM. Ding, J., Gravano, L., and Shivakumar, N. (2000). Computing geographical scopes of web resources. In Proceedings of the 26th International Conference on Very Large Data Bases, VLDB '00, pages 545–556, Cairo, Egypt.
27. Eisenstein, J., O'Connor, B., Smith, N. A., and Xing, E. P. (2010). A latent variable model for geographic lexical variation. In Empirical Methods in Natural Language Processing, pages 1277–1287, Cambridge, MA, USA. Gelernter, J. and Mushegian, N. (2011). Geo-parsing messages from microtext.
28. Transactions in GIS, 15(6):753–773. Hauff, C. and Houben, G.-J. (2012). Geo-location estimation of flickr images: social web-based enrichment. In

Proceedings of the 34th European conference on Advances in Information Retrieval, ECIR'12, pages 85–96, Barcelona, Spain. Springer-Verlag.

29. Hong, L., Ahmed, A., Gurumurthy, S., Smola, A. J., and Tsioutsiouliklis, K. (2012). Discovering geographical topics in the twitter stream. In Proceedings of the 21st international conference on World Wide Web, WWW '12, pages 769–778, Lyon, France. ACM.
30. Kinsella, S., Murdock, V., and O'Hare, N. (2011). "i'm eating a sandwich in glasgow": modeling locations with tweets. In Proceedings of the 3rd international workshop on Search and mining user-generated contents, SMUC '11, pages 61–68, Glasgow, Scotland, UK. ACM.