Final Report: Best Business Types

Summary

This project aimed to identify the best types of businesses. This was done by addressing the research question: What are the top types of businesses in New York with consistently high ratings, given that they are clustered into groups that are within walking distance of one another? New York metadata from the Google Local Data (2021) dataset (Yan et al., 2022) is the dataset used to answer this question. The data was analyzed with Hierarchical Clustering with Euclidean Distance, Divide and Conquer, GroupBy, and MapReduce algorithms. Results from the analysis were to indicate which categories of businesses are generally popular within local areas.

Introduction

Background

Features from the Google Local Data (2021) dataset that were particularly important are the longitude and latitude, as they were used for the Divide and Conquer process to create location-based subsets for the clustering (using GroupBy). This is because Hierarchical Clustering does not handle large datasets very well due to its O(n2) complexity (GeeksforGeeks, 2023). Longitude and latitude were also used to calculate the Euclidean Distance between businesses for the Hierarchical Clustering. The Euclidean Distance is necessary to identify which businesses should be grouped (i.e., are close together) for the clustering.

Additionally, the average rating and the categories features of the dataset were used to identify the highest-rated business in each cluster and its types. This was done with the MapReduce algorithm. MapReduce is apt at handling large datasets via parallel processing (Map-reduce, n.d.).

Motivation

Places of business are located everywhere in the world, with New York being one city that boasts an enormous number. Each business has many customers and therefore has many reviews. Additionally, due to New York being such a popular destination, the customers would be from many different backgrounds. Each business also falls under a limited number of categories. This combination of elements results in an extremely large dataset that can be used to discover, via customer ratings, which type of businesses are most successful. Note that this is based on the assumption that better businesses would have higher average ratings. I am personally interested in finding out which businesses perform better because I am about to finish my studies. So, knowing which category of businesses are likely to be successful would help me in evaluating where I would like to find employment. Others in a similar position to myself may also find that information useful in their searches.

Research Question

The research question is: What are the top types of businesses in New York with consistently high ratings, given that they are clustered into groups that are within walking distance of one another? It is relevant to the Google Local Data (2021) dataset since a large portion of that dataset is reviews from random individuals who have visited the locations they are reviewing. Those reviews (ratings) are appropriate indications of the quality of the businesses.

The purpose of grouping locations by longitude and latitude, and applying Hierarchical Clustering was to only compare businesses with their direct competitors. There was the assumption that direct competitors would be those businesses within walking distance of each other. This helped ensure that the top types identified were because those categories consistently have high ratings, no matter their competition. Once the highest-rated types in each cluster had been identified with MapReduce, they were collected.

Experimental Design and Methods

The sequence of steps to address the research question is as follows.

Firstly, the dataset was divided into smaller sections based on location using GroupBy (for Divide and Conquer). This separated the data into subsets within grids delineated by longitude and latitude coordinates.

Next, the Hierarchical Clustering with Euclidean Distances was carried out on the subsets, with a location/cluster joining another cluster if the minimum distance between their centroids was less than walking distance (0.4 km from Yang & Diez-Roux (2012)). Calculating the distance between two locations from longitude and latitude values was done using the Haversine function.

Finally, for each cluster, the top 1 business with the highest average rating was identified using MapReduce. The ratings for these businesses from each cluster were compared to find the unique values. The categories for these were taken as the answers to the research question.

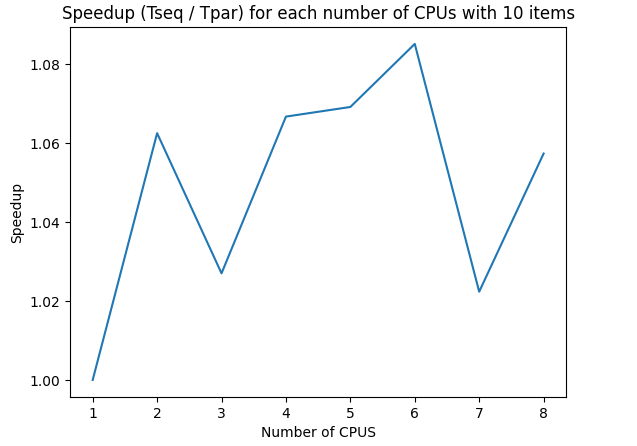
The code included:

* **Urllib.request:** to import data from the URL supplied by Yan et al. (2022)
* **ijson library**: to read downloaded data into a safe JSON format
* **convert\_to\_jsonl()**: to convert the downloaded data to .jsonl format
* **process\_line()**: to create a DataFrame to contain the data
* **separate\_into\_smaller\_groups()**: used to implement the Divide and Conquer.
* **calculate\_distance()**: to apply the haversine function to get the physical distance between two locations.
* **merge\_iteration() and apply\_clustering()**: were used for the Hierarchical Clustering. They used some helper functions for .foldby() and .map().
* **get\_max\_category\_in\_cluster()**: used to find the top categories in each section/cluster
* **Dask library**: to handle large sets of data

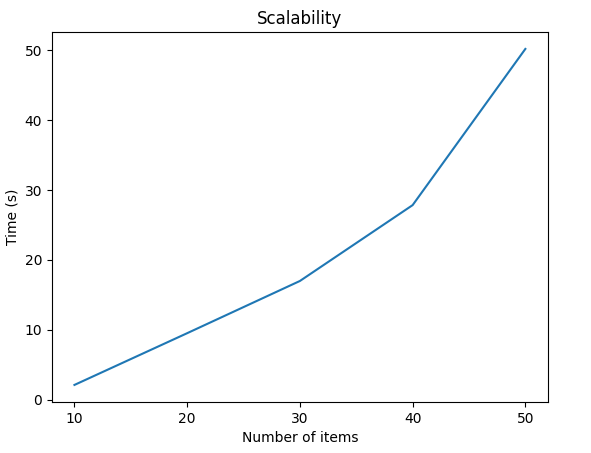
Results

This analysis aimed to identify top business types by rating in New York. The largest analysis attempted was on 100 items on 6 processors. It was on 6 processors because 6 has the best speedup according to Figure 1. The analysis took more than 30 minutes, likely because Hierarchical Clustering is an O(n2) algorithm and the implementation is not sufficiently parallel. Figure 2 and Figure 3 also indicate that the process is not scalable since they depict an almost exponential curve. So the analysis was not run on the full dataset (~33,000,000 items) as sufficient resources were not present and the chosen algorithms were not ideal. The analysis of smaller subsets produced results such as River, Peninsula, Home inspector, Painter, Repair Service, Stores, Restaurants, Concert hall, and Event Venue. Based on this data the best businesses are those that are very service-oriented or tourist destinations.

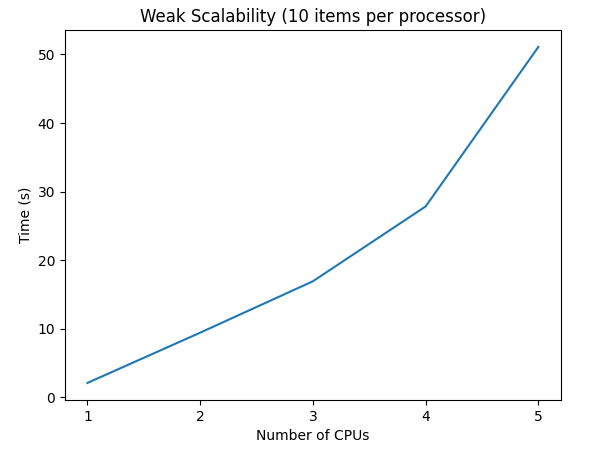
**Figure 1:** Speedup (Tseq / Tpar) for each number of CPUs with 10 items



**Figure 2**: Scalability showing the time taken (s) for the processes to complete with 1 CPU and an increasing number of items



**Figure 3**: Weak Scalability showing the time taken (s) for process with 10 items per CPU



Conclusion

The research question was not answered. This is because the entire dataset was not analyzed. The cause behind not analyzing the entire data sets is that the process was not sufficiently parallel and would consume a lot of resources (time, money, and processors). Subsets analyzed were not completely randomly chosen and as such are not likely representations of the entire dataset. Therefore the results from analyzing the subsets cannot be used to answer the research question satisfactorily.

If one assumes that the results are correct, the implication would be that the best businesses (and places to seek employment) are those that are very service-oriented or tourist destinations. It is far more likely that the results imply more individuals submit ratings to Google Reviews about places they visited for enjoyment. Additionally, it is reasonable that individuals would submit higher ratings for businesses that focus on customers. So it is likely that ratings on Google Reviews are not a good metric to evaluate the success of all types of businesses.

In the future, to answer the research question, the ideal first step would be identifying a more suitable dataset for evaluating business. Perhaps one that contains statistics such as employee numbers, revenue, environmental impact, etc, along with their categories and locations. Secondly, the hierarchical clustering will have to be replaced with another algorithm with lower complexity to ensure the analysis is more scalable. This replacement algorithm would need to group businesses in some way that ensures they are only being compared with direct competitors.

In the case that the same dataset is being reused, it would be better to change the focus of the research question from all businesses to service-oriented businesses such as restaurants and stores. The data will then need to be filtered to exclude businesses that aren’t service-oriented. Also, in that case, it would be best to change the motivation from seeking ideal places of employment to seeking the best places for enjoyment.

Critique of Design and Project

One part of the method that requires improvement is the grouping of businesses by longitude and latitude, and then using Hierarchical Clustering to identify businesses that are direct competitors within those groups. The purpose of grouping businesses before clustering was to reduce the computation required by the clustering. Currently, longitude and latitude are rounded to zero decimal places. This creates relatively large groups for the clustering. An improvement would be creating smaller groups by only rounding longitude and latitude values to one or two decimal places. When increasing the decimal places it will be important to ensure the groups aren’t too small (less than areas with a radius of 0.4km).

Hierarchical Clustering is an algorithm with a high complexity. Using the algorithm meant the analysis process was not very scalable, and could not be used to analyze the entire dataset. Therefore another algorithm with a lower complexity should have been used such as K-Means Clustering which is better for large datasets. Since K-Means Clustering is limited by K (number of clusters required), it would be important to calculate how many clusters of radius 0.4km would reasonably fit in the groups (from the GroupBy). Additionally, the implementation of Hierarchical Clustering was not entirely parallel. .compute() was used multiple times. This operation is used to trigger the execution of delayed computations and therefore forces parts of the process to be linear. This also reduces how scalable an algorithm is. Therefore, occurrences of this would need to be removed. Preferably, .compute() should not be used except at the end to print the output.

Reflection

Tools I found useful for completing the project included the examples and scenarios in lectures. They helped me better understand algorithms (such as MapReduce and Hierarchical Clustering) and were more useful than plain descriptions. The Labs being in Google Collab helped me get familiar with the tools I need for the project ( such as Google Cloud). Being familiar with the Dask library also made it easier to import and perform operations on large amounts of data.

This project has taught me about the complications of handling large datasets. This includes finding and identifying their key properties, and the consequences of missing data. Initially, in the first project proposal, I was planning to use the Google Restaurants dataset from Yan et al. (2023), however, due to incorrect JSON and missing fields in the provided files (as mentioned in my Individual Progress Report) I had to switch datasets. Also, there are many issues with the formatting of data. Often a large amount of time has to be spent on getting data into the correct format and types such as floats, lists, and ints.

Finally, the major learning point for me was that running processes in parallel requires a lot of thought. Often data is not processed in order so some steps cannot be carried out without waiting for processes to finish (or calling costly operations like .compute()). Additionally, it is important to store data in the correct structure (bags, data frames, arrays, etc.) because it greatly affects which operations can be performed. It also affects how optimally the process runs. This was a major point for me because I was unable to implement Hierarchical Clustering efficiently. This was due to not identifying the correct data types initially, and not using parallel processing to its full potential. However, as a consequence, I do believe that this project has given me a good theoretical understanding of the algorithm.

References

GeeksforGeeks. (2023, December 12). *Hierarchical clustering in data mining.* GeeksforGeeks. https://www.geeksforgeeks.org/hierarchical-clustering-in-data-mining/

*Map-reduce*. Map-Reduce - an overview | ScienceDirect Topics. (n.d.). https://www.sciencedirect.com/topics/computer-science/map-reduce

Yan, A., He, Z., Li, J., Zhang, T., & Mcauley, J. (2022). *UCTopic: Unsupervised Contrastive Learning for Phrase Representations and Topic Mining.* Recommender Systems and Personalization Datasets.<https://aclanthology.org/2022.acl-long.426.pdf> .

Yan, A., He, Z., Li, J., Zhang, T., & Mcauley, J. (2022a). *Personalized Showcases: Generating Multi-Modal Explanations for Recommendations.* Recommender Systems and Personalization Datasets.<https://arxiv.org/pdf/2207.00422>

Yan, A., He, Z., Li, J., Zhang, T., & Mcauley, J. (2023). *Personalized Showcases: Generating Multi-Modal Explanations for Recommendations.* Recommender Systems and Personalization Datasets.<https://arxiv.org/pdf/2207.00422>

Yang, Y., & Diez-Roux, A. V. (2012). Walking distance by trip purpose and population subgroups. *American journal of preventive medicine*, 43(1), 11–19. https://doi.org/10.1016/j.amepre.2012.03.015