HIGH SCHOOL EDITION
Journal of Student Research
Volume 13 Issue 1 (2024)

Web Crawling Tripadvisor to Develop a Restaurant Recommendation System for Los Angeles

Jimin Han[1] and Sergei Ievlev[#]

[1]QSI International School of Shenzhen, China
[#]Advisor

ABSTRACT

The widespread use of online review websites has revolutionized how consumers choose restaurants, particularly in popular tourist destinations like Los Angeles, where a vast range of dining options is readily available. However, the sheer abundance of similar cuisine offerings can be overwhelming. To address this challenge, this study used Python Selenium to web crawl Tripadvisor for gathering data about Los Angeles restaurants. Relevant information from user reviews was extracted and analyzed utilizing natural language processing techniques to classify restaurants based on cuisine, price, and customer reviews and ratings. This classification allowed for the identification of distinct dining preferences, providing insights into restaurant selection in tourist-heavy areas. With the application of cosine similarity, the analysis further led to the development of a recommendation system specific to consumers' needs and preferences. This study thus offers a new approach to improving restaurant discovery and decision-making in busy urban centers in Los Angeles.

Introduction

Tripadvisor, a platform that helps tourists decide on their hotels, activities, and restaurants, provides numerous reviews written by consumers from all around the world. Having gained popularity among many customer bases, the website has proven itself very useful. In fact, as of September 2023, the website recorded 156.4 million views, becoming one of the most visited travel and tourism websites worldwide (*Most Visited Travel and Tourism Websites Worldwide 2023 | Statista,* 2023). The extensive use of the website not only reflects its popularity but also highlights its reliability as a data source.

Los Angeles, a city in the United States, is the main subject of this research. The city is a popular destination for tourists worldwide, as evidenced by 46.2 million visitors who spent US$21.9 billion in 2022 (Li, 2023). Most importantly, Los Angeles is known for its diverse cultures (*California Dreaming: Cultural Diversity - One of the Golden State's Greatest Assets*, 2021). This mix of culture, in turn, has led to the development of various cuisines in the area. With 25,292 different restaurants (Gase et al., 2019), the city offers an ideal background for the research to be done. Therefore, using the data from Tripadvisor, this study aims to provide a restaurant recommendation system for the tourists and locals in Los Angeles.

Previously, several studies have been conducted on developing restaurant recommendation systems (Munaji & Emanuel, 2021; Mahajan, Joshi, Khedkar, Galani, & Kulkarni, 2021). However, this research diverges by focusing extensively on Los Angeles using a distinct methodology. This research applies web crawling techniques on Tripadvisor and represents its findings using visual aids like Word Clouds. Furthermore, with the detailed and area-specific analysis, this study uses cosine similarity, a popular method for showing resemblance, to build its own system.

Univariate Analysis of the LA Restaurant Data

Data Introduction

In this study, a total of 6,791 sample sets were gathered by using Python Selenium to web crawl Tripadvisor from December 20th to 25th, 2022. This data was then organized into a table, categorizing each by the number of stars, the number of reviews, the cuisine type, and the price range.

Table 1. Sample Data Collected During the Web Crawling Process

Restaurant Name	Stars	Number of Reviews	Cuisine	Price Range
n/naka	5.0	173	Japanese	High
Raffaello Ristorante	4.5	278	Italian	Medium
Providence	4.5	796	Seafood	High
Brent's Deli Northridge	4.5	1556	American, Deli	Medium
Langer's	4.5	836	American	Medium

Analysis of the data revealed that the average star rating for each restaurant varied from 1.0 to 5.0, with 5.0 representing the highest rating. Specifically, the average star rating for restaurants in Los Angeles was 4.2, and the most common star rating was 4.0, represented by 2,309 samples, followed by a 4.5 rating, with 2,234 samples.

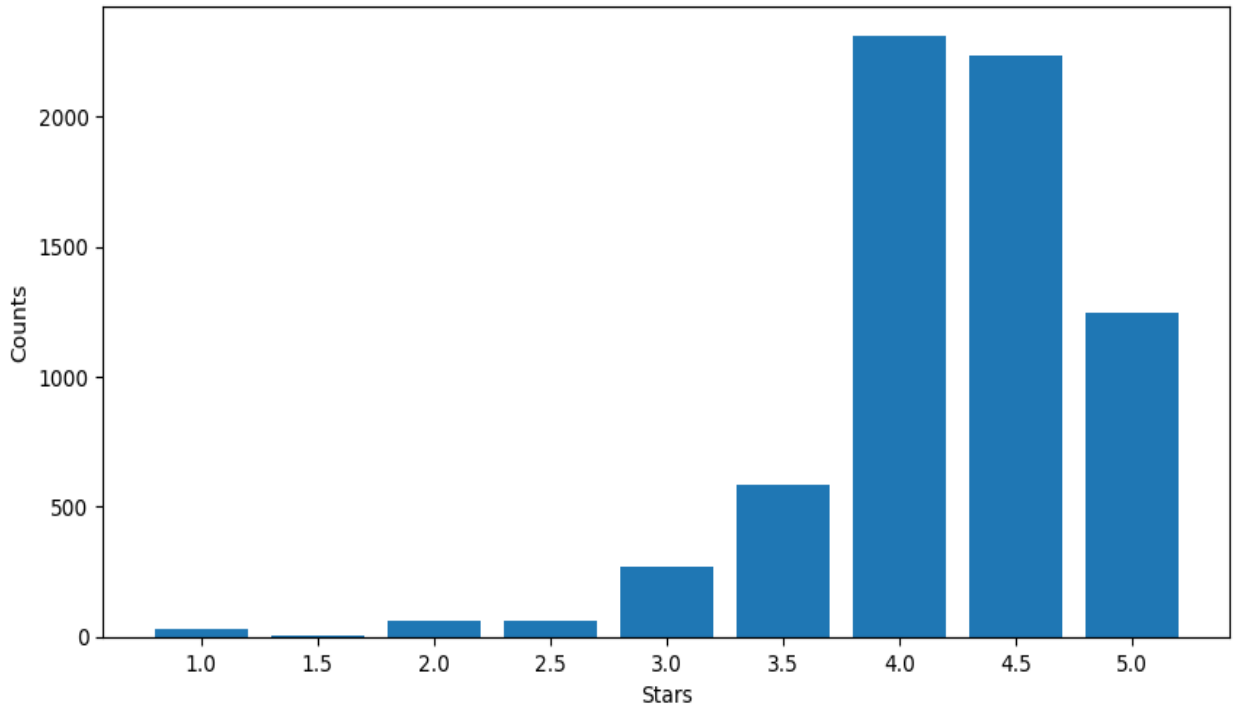

Figure 1. Distribution of Stars in LA Restaurants. As seen in the bar graph, 4.0 and 4.5 restaurants were dominant in the dataset.

Table 2. Summary Table for Stars

Statistics	Stars
Count	6791.0
Mean	4.2
Min	1.0
25%	4.0
50%	4.5
75%	4.5
Max	5.0

Figure 1 and Table 2 show a left-skewed distribution of the data, implying that most restaurants received a star of 4.0 or higher.

On the other hand, the number of reviews varied more significantly compared to the number of stars, with the highest frequency being 2,513 reviews for "Hard Rock Cafe", an American cuisine restaurant, and the lowest being just one review. The average number of reviews was a mere 35.8, suggesting that most restaurants had only a few reviews.

Table 3. Top Five Restaurants with the Most Reviews

Restaurant Name	Stars	Number of Reviews	Cuisine	Price Range
Hard Rock Cafe	4.0	2513	American, Bar	Medium
The Ivy	4.0	2146	American	High
Saddle Ranch Chop House	4.5	1787	American, Steakhouse	Medium
The Griddle Cafe	4.5	1588	American, Cafe	Medium
Brent's Deli Northridge	4.5	1556	American, Deli	Medium

Table 4. Summary Table for the Number of Reviews

Statistics	Number of Reviews
Count	6791.0
Mean	35.8
Min	1.0
25%	2.0
50%	7.0
75%	25.0
Max	2513.0

Table 4 shows a right-skewed distribution of the data, contrasting to the relationship illustrated in the previous table for stars. Additionally, there was a distinguishable trend in the cuisine types among the sampled restaurants: many had multiple cuisine options listed. However, for the purposes of this research, only the first listed cuisine type, considered the primary cuisine, was used to represent each restaurant. This approach revealed that the most common cuisine type was "Not Defined", with 1,808 samples, followed by "American" cuisine, totaling 1,064.

The price range was also a crucial determinant of the restaurants' ratings. Restaurants labeled as "Cheap" had price ranges of less than US$10, those with a "Medium" price range fell between US$10 and US$30, and "High" price restaurants were those with price ranges greater than US$30. The most common price range among these three was "Medium", with a total of 2,632 samples. However, the most frequent price range overall was "Not Defined", responsible for 3,058 samples.

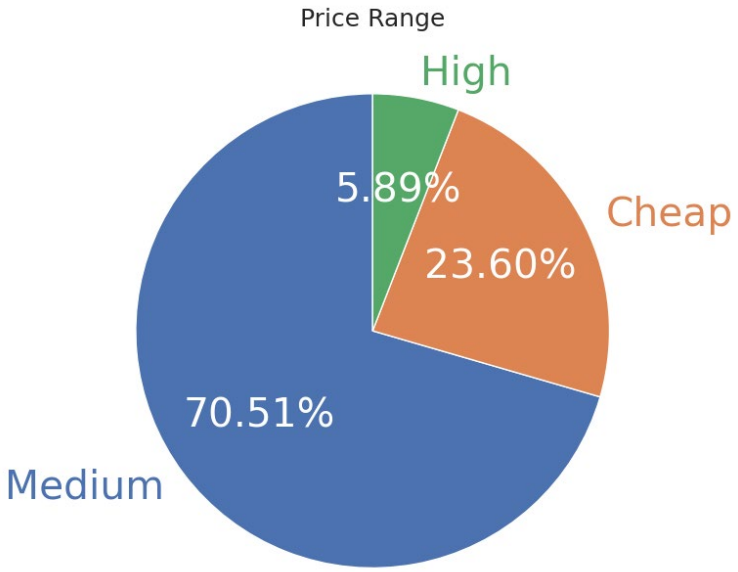

Figure 2. Pie Chart Representing the Proportion of Each Price Range. The majority of the collected samples were in the "Medium" price range.

Multivariate Analysis

Correlation Between the Factors

The analysis of various factors affecting the ratings of Los Angeles restaurants led to the inquiries regarding the correlation between these elements. One area of interest was the relationship between the type of cuisine and its star ratings. Theoretically, the number of stars should have a directly proportional relationship with the number of samples. However, the study presents different findings. For instance, while "Chinese" cuisine ranked seventh among cuisine types with stars greater than or equal to 4.0, it was third on the list for cuisine types with stars less than or equal to 2.5. This suggests that Chinese cuisine restaurants were not rated as highly relative to restaurants serving other types of cuisine, and thus the directly proportional relationship is not universal.

Table 5. Type of Cuisine and Count of Appearance of Stars Greater Than or Equal to 4.0

Type of Cuisine	Count of Stars Greater Than or Equal to 4.0
Not defined	1457
American	947
Mexican	532
Italian	399
Japanese	355
Asian	289

Chinese	**257**
Cafe	141
Bar	126
Pizza	125

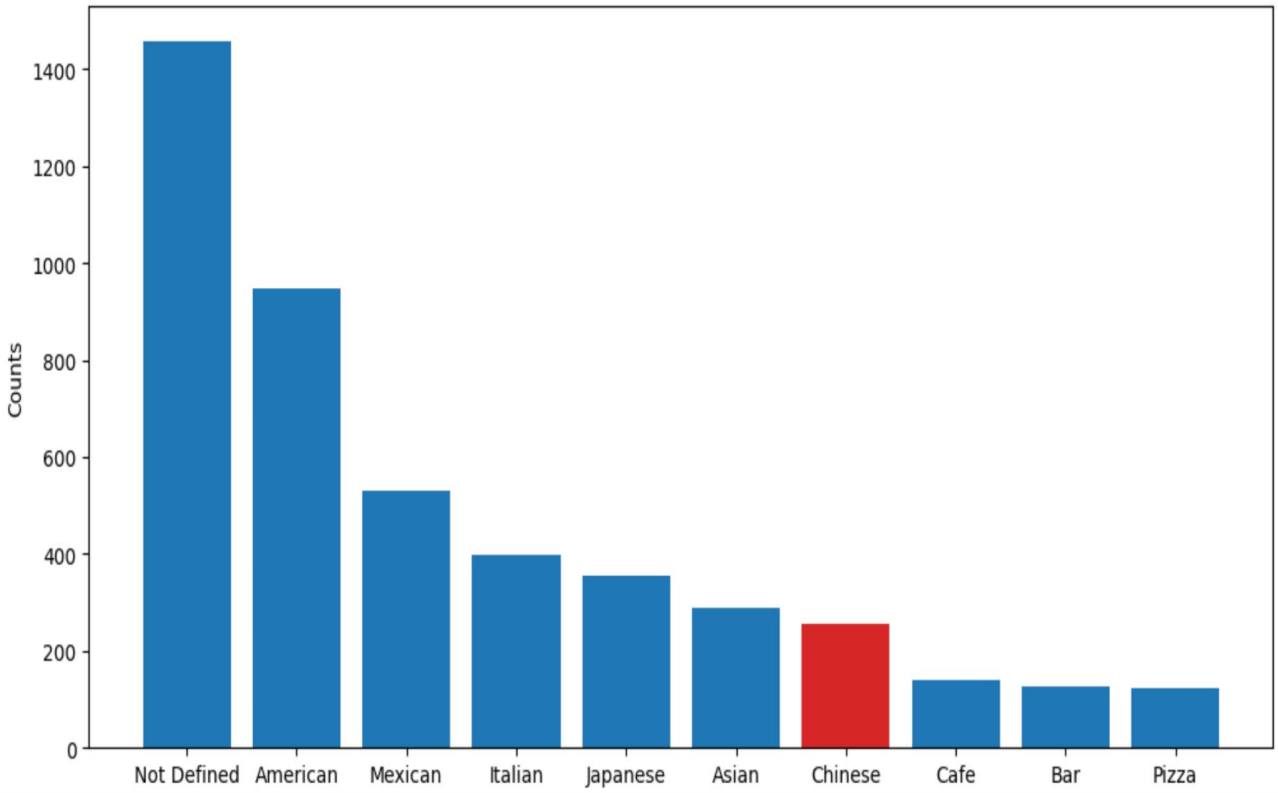

Figure 3. Bar Graph Representing the Relation in Table 5. As demonstrated, Chinese cuisine, marked as red, is ranked seventh.

Table 6. Type of Cuisine and Count of Appearance of Stars Less Than or Equal to 2.5

Type of Cuisine	Count of Stars Less Than or Equal to 2.5
Not defined	84
American	14
Chinese	9
Mexican	7
Italian	6
Japanese	4
Thai	3

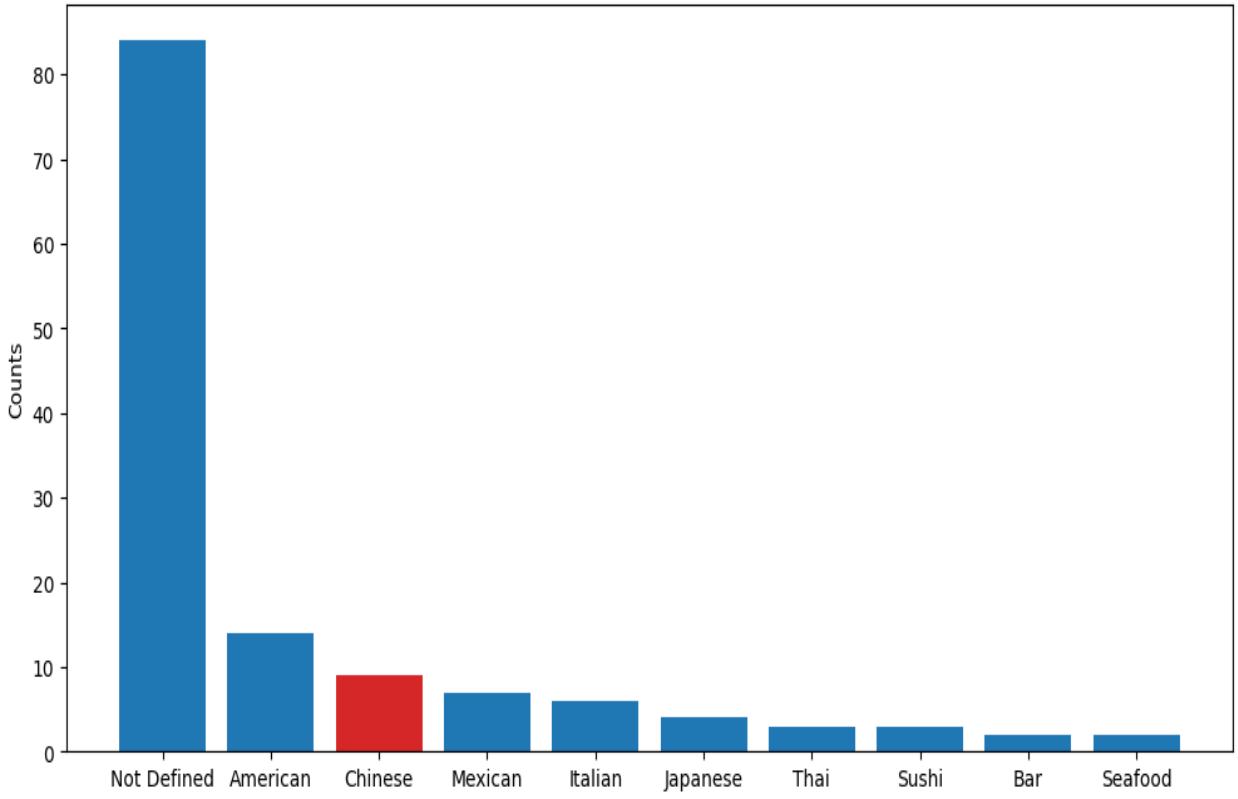

Figure 4. Bar Graph Representing the Relation in Table 6. As demonstrated, Chinese cuisine, marked as red, is ranked third.

Another notable relationship exists between the type of cuisine and price range. Theoretically, each cuisine should have a similar ratio within each price range. However, this is not always the case in reality. Therefore, to observe the actual relationship between these two attributes, the following graph was drawn.

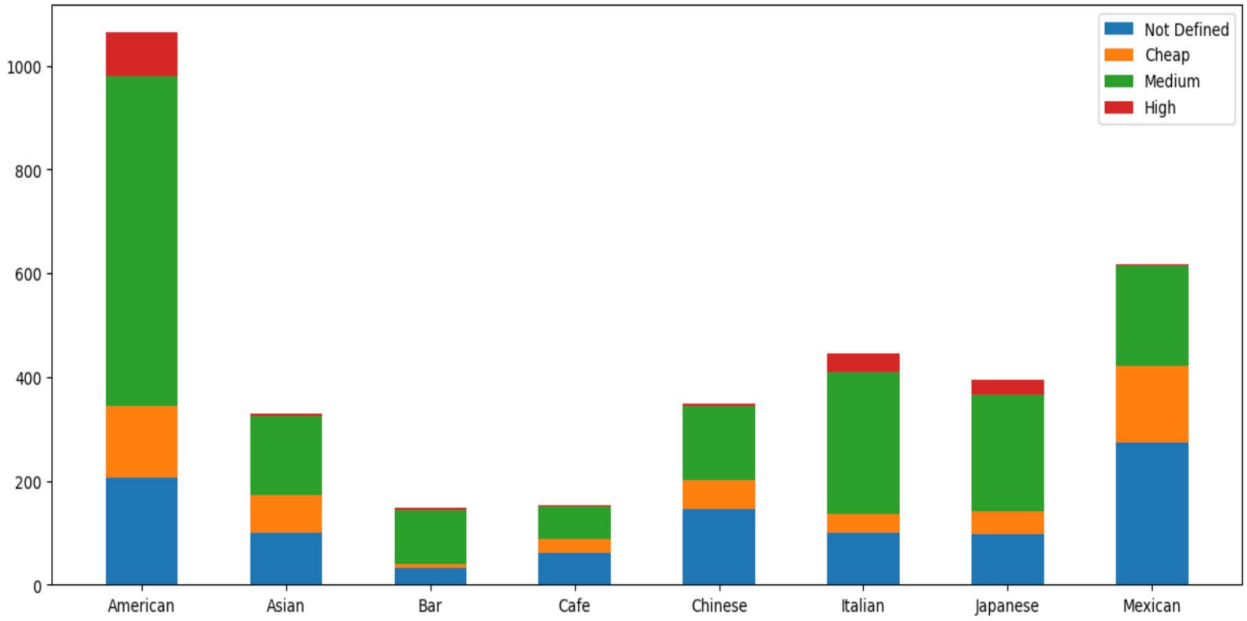

Figure 5. Proportion of Each Price Range in Different Cuisines. As demonstrated, each cuisine shows different proportions of each price range.

To give a more precise comparison, the top three cuisines with the most reviews, American, Mexican, and Italian, were analyzed. For all three cuisines, the "Medium" price range or "Not Defined" were the most common. However, while American and Mexican cuisines noticeably featured the "Cheap" price range more frequently than the "High" price range, Italian cuisine had a slightly more frequent appearance in the "High" price range compared to the "Cheap" price range. Hence, the type of cuisine and price range are not directly related.

Table 7. Counts of Each Price Range for American Cuisine

Price Range	Counts
Medium	637
Not Defined	207
Cheap	136
High	84

Table 8. Counts of Each Price Range for Mexican Cuisine

Price Range	Counts
Not Defined	274
Medium	194
Cheap	147

High	2

Table 9. Counts of Each Price Range for Italian Cuisine

Price Range	Counts
Medium	273
Not Defined	101
High	37
Cheap	35

Lastly, the relationship between the price range and the stars is drawn. One might hypothesize that the number of stars strictly increases with the price, as higher price may imply higher quality. However, this was disproven. Organizing the data by price range for each star rating revealed that the "Medium" price range is the most popular among the three price ranges for every case. To identify a pattern, there was a need to compare the relative abundance of the "High" price range to the "Medium" range at each star level. Doing so derived the following: for a 5.0-star rating, the relative abundance of the "High" price range was 0.097. At 4.5 stars, it was 0.10; at 4.0 stars, it was 0.073; and at 3.5 stars, it was 0.042. From 3.0 stars downward, the count for the "High" price range fell below 5, eventually reaching zero at 1.0 star. Thus, while the general trend suggests that higher prices correlate with higher star ratings, it should be noted that the most expensive restaurants were most prevalent at the 4.5-star level, not 5.0.

Table 10. The Relationship between the Price Range and the Stars

Stars	Not Defined	High	Medium	Cheap
5.0	940	18	185	101
4.5	711	108	1037	378
4.0	806	81	1110	312
3.5	267	10	236	68
3.0	198	3	49	16
2.5	45	0	9	6
2.0	58	0	4	0
1.5	4	0	0	0
1.0	29	0	2	0

To demonstrate the relationship between the attributes including the number of stars, the number of reviews, the cuisine type, and the price range in a clearer manner, the 4x4 correlation matrix was drawn. A cell with the number 0 indicates no linear correlation, the number 1 indicates a perfectly positive linear correlation, and the number -1 indicates a perfectly negative linear correlation. When the matrix was drawn for this study, most of the attributes were, in fact, found to have close to zero correlation with one another: most of the cells had a number value close to zero. Still, the cell showing the correlation between the number of reviews and

price range showed a value of -0.22, whose absolute value is less than the threshold 0.6 but still remarkably high relative to that of numbers in other cells.

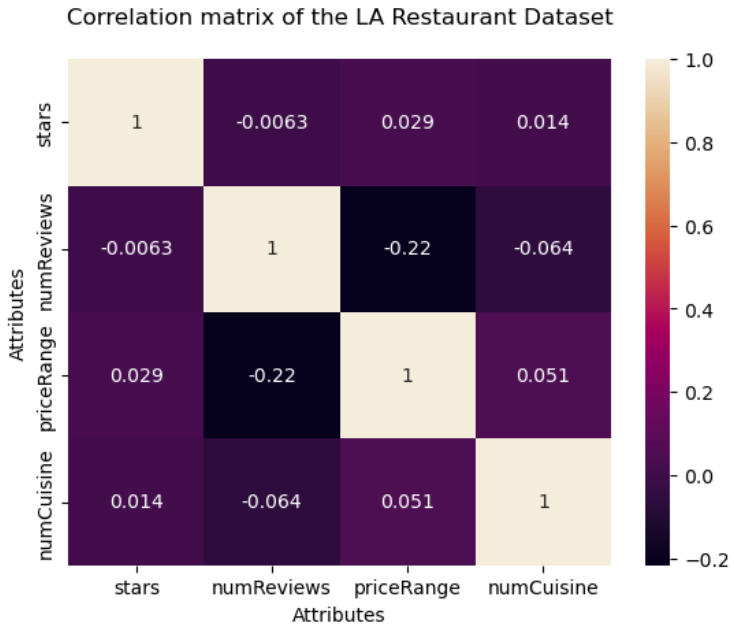

Figure 6. Correlation Matrix of the LA Restaurant Dataset. As illustrated, the matrix implies a deficient correlation between the attributes.

This matrix confirms the finding that the attributes are not directly related to each other. With such a finding, it was concluded that the recommendation system could not be built using the attributes. Therefore, direct text analyses had to be done.

Extracting and Analyzing Texts from the Reviews

Text Data Collection

This section of the analysis aims to demonstrate the differences in texts across various situations. To this end, samples were taken from 136 restaurants, with data previously gathered through the web crawling. In selecting the models, restaurants with an insufficient number of reviews (fewer than 300) were excluded. Additionally, only reviews written in English were chosen for clarity in word analysis. After careful selection of the samples, words were extracted from the main reviews of the restaurants. Subsequently, unnecessary words that do not affect the overall meaning were removed, and the list was organized by counting and ranking the phrases based on their frequency of appearance in the main review.

The analysis began with selecting the restaurant with the most English reviews on the list, "The Ivy", known for American cuisine and an average rating of 4.0 stars from consumers; “The Ivy” had 1,816 English reviews, while “Hard Rock Cafe”, which had the most reviews in total, only had 1,499 English reviews. Following the previously described process, the following list of words was gathered.

Table 11. Top 30 Most Common Texts from "The Ivy"

Texts	Number of Appearance	Texts	Number of Appearance	Texts	Number of Appearance
food	1037	table	298	people	177
ivy	640	nice	243	went	175
service	529	atmosphere	230	one	172
great	520	outside	229	us	166
place	471	beautiful	226	experience	166
good	434	go	223	excellent	165
restaurant	414	staff	205	amazing	153
lunch	339	really	197	always	148
champagne	328	glass	188	complimentary	147
la	301	time	181	delicious	146

From the table, the restaurant is frequently associated with positive words such as "great", "good", "nice", and "beautiful", suggesting a positive social rating. Words like "champagne" and "glass", indicative of American cuisine, also appeared on the list. Moreover, the terms "food" and the restaurant's name, "ivy", were notably frequent. Following this, a Word Cloud was created for visualization using the list.

Figure 7. Word Cloud Made from Table 11. As shown, the positive words and those related to the restaurant can easily be identified.

Factors Affecting Differences in Texts

To demonstrate the distinctions between each list of words, several variances were tested by altering the attributes: the number of stars of a single restaurant, types of cuisine, and restaurants with varying numbers of stars. It should be noted that while testing the variances, other factors remained consistent. Each variance is represented by one restaurant that exemplified the best example.

Text Analysis by Differing the Number of Stars within a Single Restaurant

The long list of texts was grouped by the number of stars, and the process described in Text Data Collection was applied to each group: 5.0, 4.0, 3.0, 2.0, and 1.0 stars. The data was sourced from "The Ivy".

Table 12. Texts vs Number of Appearance from 5.0 Group of “The Ivy”

Texts	Number of Appearance	Texts	Number of Appearance	Texts	Number of Appearance
food	556	table	161	time	104
ivy	344	beautiful	147	go	104

great	344	atmosphere	132	nice	99
service	328	outside	127	best	97
place	250	glass	126	experience	97
champagne	231	staff	124	really	93
restaurant	217	excellent	120	us	90
lunch	191	amazing	115	complimentary	90
la	182	delicious	111	love	88
good	169	always	105	one	88

Figure 8. Word Cloud Made from Table 12

Table 12 shows that the relative appearance of positive words increased in reviews from the 5.0-star group. In fact, new positive words like “best” and “love” emerged. This indicates that reviewers who gave 5.0

stars to "The Ivy" tended to use more positively connoted words. Figure 8, a visual aid to Table 12, illustrates an increase in positive words in the Word Cloud compared to the original reviews.

Table 13. Texts vs Number of Appearance from 4.0 Group of "The Ivy"

Texts	Number of Appearance	Texts	Number of Appearance	Texts	Number of Appearance
food	256	outside	69	staff	48
good	188	atmosphere	68	patio	47
ivy	154	champagne	68	time	45
great	138	la	64	glass	44
place	123	really	59	went	41
service	119	go	54	lovely	41
restaurant	102	people	52	see	41
lunch	98	one	50	little	40
nice	85	beautiful	49	complimentary	40
table	78	us	48	menu	39

Figure 9. Word Cloud Made from Table 13

Table 13 reveals that some positive words, such as “delicious” and “best”, disappeared and were replaced by others in reviews rating "The Ivy" with 4.0 stars. However, many positive words still occurred frequently. This suggests that 4.0-star reviewers were slightly less generous in their word choices than 5.0-star reviewers. Figure 9, a visual aid to Table B-1, shows that many positive words continue to appear in the Word Cloud.

Table 14. Texts vs Number of Appearance from 3.0 Group of “The Ivy”

Texts	Number of Appearance	Texts	Number of Appearance	Texts	Number of Appearance
food	135	great	30	special	22
ivy	76	la	29	people	21
place	54	really	29	worth	20
restaurant	52	table	28	beautiful	20
good	50	like	26	staff	19

service	50	went	25	little	19
go	44	atmosphere	24	nothing	18
nice	43	see	23	experience	18
average	38	outside	23	ok	17
lunch	31	champagne	22	one	17

Figure 10. Word Cloud Made from Table 14

Table 14 indicates a significant decrease in positive words in reviews from those who gave 3.0 stars. Words like “average” and “ok” became more prevalent, suggesting partial satisfaction with the restaurant. Some positive words like “great” and “good” remained. Figure 10, a visual aid to Table 14, shows a reduction in the size of positive words.

Table 15. Texts vs Number of Appearance from 2.0 Group of “The Ivy”

Texts	Number of Appearance	Texts	Number of Appearance	Texts	Number of Appearance

food	45	experience	12	better	9
ivy	31	lunch	11	say	8
restaurant	25	really	10	years	8
place	23	ordered	10	made	8
service	21	people	10	one	8
la	15	table	10	back	8
go	15	disappointed	9	overpriced	8
good	15	would	9	think	8
get	15	nice	9	special	8
went	14	going	9	beautiful	7

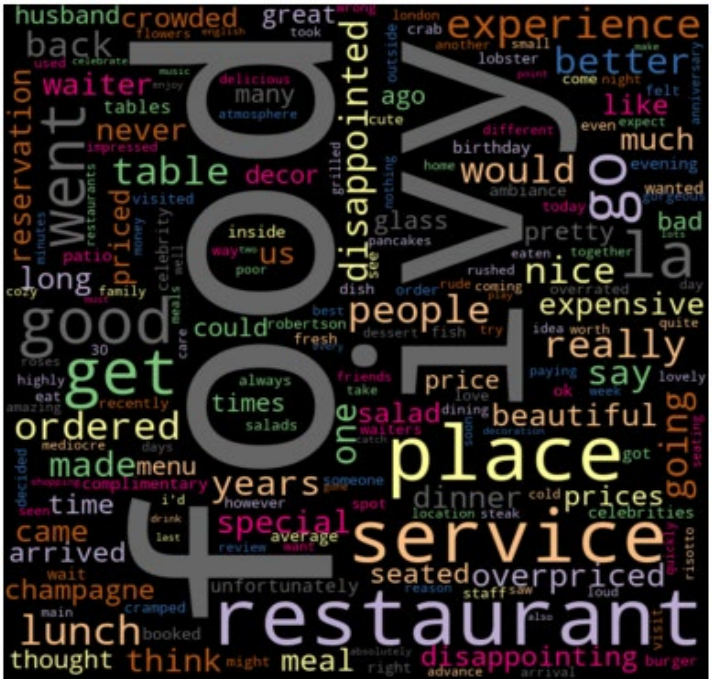

Figure 11. Word Cloud Made from Table 15

Table 15 demonstrates a decrease in positive words and an increase in negative words in reviews from the 2.0-star group. Words expressing dissatisfaction like “disappointed” and “overpriced” appeared, and words like “great” vanished. Figure 11, a visual aid to Table 15, shows smaller positive words and the emergence of negative words in the Word Cloud, although positive words still dominate in size.

Table 16. Texts vs Number of Appearance from 1.0 Group of "The Ivy"

Texts	Number of Appearance	Texts	Number of Appearance	Texts	Number of Appearance
food	45	went	11	rude	8
ivy	35	people	11	experience	8
place	21	staff	10	would	8
table	21	told	9	lunch	8
restaurant	18	waiter	9	us	7
like	14	back	9	even	7
time	13	years	9	nice	7
good	12	overpriced	9	experience	7
service	11	one	9	seated	7
la	11	ordered	8	worst	7

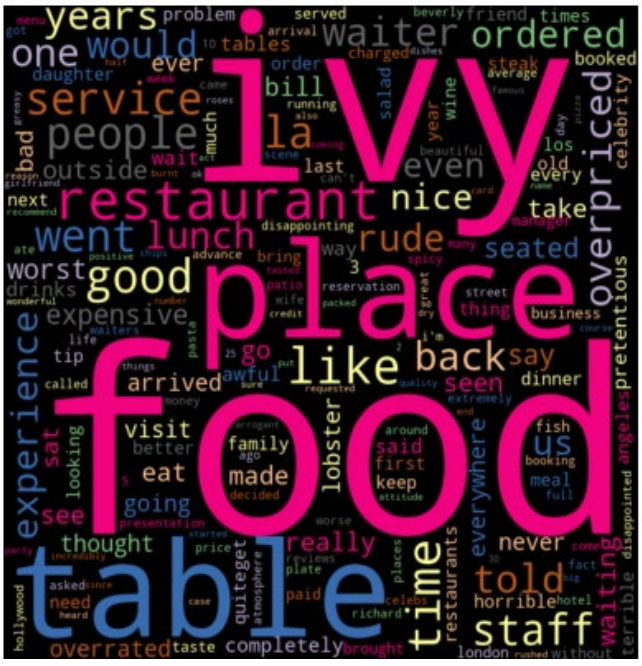

Figure 12. Word Cloud Made from Table 16

Table 16 continues to reveal more frequent appearances of negative words such as “overpriced”, “rude”, and “worst” from the 1.0-star group, contrasting with earlier words like “delicious”, “nice”, and “great”. Figure 12, a visual aid to Table 16, shows positive and negative words of similar size in the Word Cloud. This trend confirms a pattern: as ratings decrease, negative words increase, and positive words decrease.

Text Analysis by Differing the Type of Cuisine

The analysis also aimed to show how texts vary in restaurants offering different types of cuisine. Therefore, the previously described process was applied to a restaurant named "Yamashiro Hollywood", known for Japanese cuisine and holding an average rating of 4.0 stars from consumers.

Table 17. Top 30 Most Common Texts from “Yamashiro Hollywood”

Texts	Number of Appearance	Texts	Number of Appearance	Texts	Number of Appearance
food	407	amazing	134	table	76
view	367	yamashiro	128	dinner	71
restaurant	249	sushi	117	best	71
great	243	japanese	107	excellent	70
service	176	beautiful	104	night	68
place	174	go	88	get	66
hollywood	173	nice	85	one	65
views	161	went	81	hill	63
good	159	time	79	location	57
la	153	city	78	really	56

Table 17 reveals a high frequency of words related to Japanese cuisine, such as “sushi” and “Japanese”, in the context of the restaurant "Yamashiro Hollywood". Unlike "The Ivy", there are no mentions of the words indicative of American cuisine such as “champagne” and “glass” in this Japanese restaurant. However, the overall tone of the words in this list remains positive, featuring terms like “great”, “good”, “amazing”, and “beautiful”. Additionally, the term “food” continues to be a dominant element in the list.

Figure 13. Word Cloud Made from Table 17. As shown, the positive words remain dominant.

Text Analysis by Differing the Sample Restaurant with a Different Number of Stars

The analysis was extended to investigate whether variations in the average number of stars given by consumers would impact the texts. To this end, a restaurant named "Spago Beverly Hills", known for American cuisine and with an average consumer rating of 4.5 stars, was selected and analyzed.

Table 18. Top 30 Most Common Texts from "Spago Beverly Hills"

Texts	Number of Appearance	Texts	Number of Appearance	Texts	Number of Appearance
food	666	time	168	amazing	131
service	522	one	162	atmosphere	130
spago	475	table	159	staff	129
great	401	menu	158	always	126
restaurant	352	place	155	go	121
good	274	wolfgang	152	wine	119

dinner	196	experience	147	dining	112
excellent	171	best	142	meal	107
beverly	171	la	139	delicious	107
hills	169	us	131	went	106

Table 18 indicates a prevalence of positive words in the reviews of "Spago Beverly Hills". Compared to "The Ivy", which received an average of 4.0 stars, this restaurant features clearer positive words like “best” in the list. However, aside from a few words, the overall tone remains largely similar. Additionally, words like “wine”, indicative of American cuisine, are present, and “food” continues to be a dominant word in the list.

Figure 14. Word Cloud Made from Table 18. Although it may appear very similar to the original Word Cloud, there is a slight increase in the frequency of positive numbers.

Methods

To develop a viable recommendation system, cosine similarity, which measures cosine of the angle between two non-zero vectors and thus determines how similar the vectors to one another, was utilized. The following represents the equation for cosine similarity for two vectors A and B.

Equation 1: Cosine Similarity

$$cos(\theta) = \frac{A \cdot B}{||A||\,||B||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} {A_i}^2} \times \sqrt{\sum_{i=1}^{n} {B_i}^2}}$$

However, using this algorithm required the word embedding process, which maps words from the reviews to vectors of real numbers. This step began with the removal of unnecessary punctuation, articles such as “a”, “an”, and “the”, and helping verbs such as “is”, “are”, and “am” that do not significantly impact the overall meaning. Additionally, since a single word could have different meanings in various contexts, for instance, “right” meaning both a direction and correctness, trigrams, or sets of three words, were considered each time. For example, according to this rule, the review “The American restaurant is awesome” was filtered to “American restaurant awesome”. Then, the filtered words were grouped into similar categories and expressed as vectors. Finally, cosine similarity was applied to these vectors to determine the similarity between the reviews and ultimately identify restaurants with similar attributes.

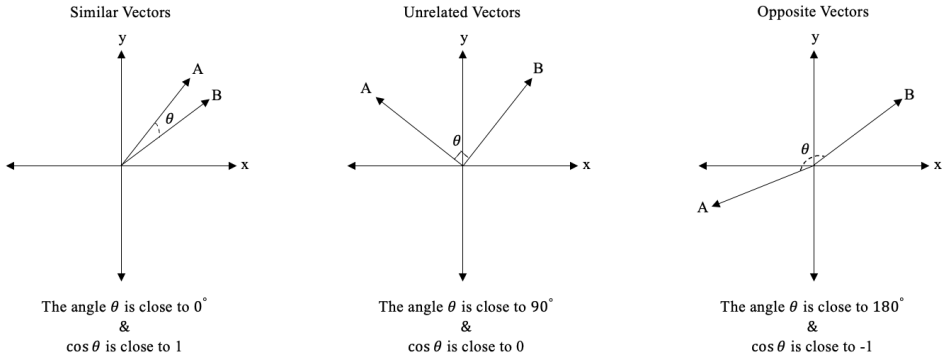

Figure 15. Cosine Similarity and Cosine Distance Functions. As demonstrated, a cosine similarity between two vectors with a degree close to 0 indicates high similarity, a degree close to 90 indicates no relation, and a degree close to 180 indicates an opposite relation.

The following table represents an example of the trigrams that were driven when such an algorithm was applied to the existing data set.

Table 19. Top Five Most Common Trigrams for Four Sample Restaurants

Restaurant	**Top Five Trigrams Based on Frequency of Appearance**				
Umami Burger	awesome burgers truffle	fries delicious burger	fries delicious burgers	twice burgers delicious	burger overcooked fries
Apple Pan	best burgers fries	best burger fries	burgers fries apple	ketchup better burger	steakburger fries apple
IHOP	salad fresh server	best breakfast restaurant	burgers disappointing	butter tasteless burgers	pancakes tasty sandwiches

			crowded		
Denny’s	quickly favorite diner	great restaurant breakfast	salad great diner	fantastic server meals	burgers love fries

Table 19 represents the top five most common trigrams for four different restaurants. As demonstrated, the restaurants "Umami Burger" and "Apple Pan" had a high cosine similarity to one another, as words like “burgers” and “fries” were both associated with positively connoted adjectives such as “delicious” and “best”. Similarly, the restaurants "IHOP" and "Denny’s" shared a high cosine similarity, as the words “breakfast” and “salads” were related to positive words such as “best” and “great”.

Results

To demonstrate how this recommendation system works, a sample restaurant named “26 Beach Cafe”, which has 4.5 stars, 299 reviews, American cuisine, and a cheap price range, was chosen. When the restaurant was entered into the system, several restaurants in Los Angeles were recommended based on different attributes. However, although the recommended restaurants’ star range remained between 4.0 and 4.5 and the main cuisine was consistently “American”, the number of reviews and price range significantly varied, ultimately impacting the value of the cosine similarity. As a result, even the most closely related restaurant to “26 Beach Cafe” had a cosine similarity of less than 0.5, indicating that no restaurant closely resembles another to a great extent.

Table 20. Top Five Restaurants with the Highest Cosine Similarity to “26 Beach Cafe”

Restaurant Name	Stars	Number of Reviews	Cuisine	Price Range	Cosine Value
Umami Burger	4.0	463	American, Bar	Medium	0.329006
Green Street Restaurant	4.5	493	American	Medium	0.300098
Russell's	4.5	425	American	Medium	0.292481
Apple Pan	4.0	368	American, Fast Food	Cheap	0.275913
Bea Bea's	4.5	407	American, Cafe	Medium	0.261681

This table represents the top five restaurants with the highest cosine similarity to a randomly selected restaurant “26 Beach Cafe” in descending order. As demonstrated, the restaurant “Umami Burger” showed the highest level of similarity to “26 Beach Cafe”.

Discussion

The results indicate the potential for this recommendation system to be used in real life. Using such an algorithm, consumers could search more conveniently for restaurants in Los Angeles. However, since the similarity

between two different restaurants may also depend on the personal perceptions of consumers, multiple restaurants with high cosine similarity to the consumer's chosen restaurant would have to be recommended each time. Nonetheless, overall, the recommendation system provides new insights into how customers in Los Angeles can choose their best-fit restaurants.

Conclusion

In this paper, a web crawling system was utilized to collect data on Los Angeles restaurants from Tripadvisor, which was then analyzed. By applying the method, the paper employed natural language processing techniques to identify key factors influencing the success of top-selling restaurants. It also introduced a recommendation system beneficial for consumers selecting dining options in Los Angeles.

The findings indicate that the primary factors influencing a restaurant's rating are the number of stars, the number of reviews, the type of cuisine, and the price range. Additionally, the language used in reviews varied significantly based on the ratings given by reviewers, the types of cuisine, and the overall star ratings. In the process, it was concluded that higher ratings correlated with more positive word choices, specific cuisines influenced the presence of related words in reviews, and restaurants with higher ratings generally had more positive tones in their reviews. Also, because the attributes showed a deficient correlation, direct text analysis had to be conducted. The text analysis, in turn, led to the development of a recommendation system using word embedding and cosine similarity to group restaurants with similar characteristics.

These results address consumers' longstanding questions about the qualities shared by top-selling restaurants and factors to consider when choosing a dining venue. The proposed algorithm, incorporating a recommendation system and natural language processing, can recommend the most suitable restaurants based on consumer preferences. Therefore, the main merit of this study is that it provides customers with a recommendation system based not on randomness, but on 6,791 data points, indicating the potential for developing an automated machine learning recommendation system based on data.

Limitations

In the process of collecting and analyzing data, this study made several assumptions. First, a large portion of data in the collected dataset was labeled “Not Defined”. This might have been due to many restaurants not specifying their type of cuisine or price range in their information details. With the prevalence of such extraneous data, only a small proportion of the collected data could have been considered valid. Moreover, for simplification, this study is specific to Los Angeles and English reviews only. Depending on the region, language, and cultural background of the dataset, the analysis could have led to different conclusions. For instance, because this study focuses on Los Angeles and English reviews, the data from American cuisine could have been exaggerated. Lastly, the use of cosine similarity as the base for the recommendation system leads to recommendations of restaurants with similar menus. Such a recommendation algorithm may have resulted in a lack of diverse results.

Acknowledgments

I would like to express much gratitude to my school math and computer science teacher Mr. Sergei Ievlev for reviewing this research paper.

References

Gase, L. N., Green, G., Montes, C., & Kuo, T. (2019). Understanding the Density and Distribution of Restaurants in Los Angeles County to Inform Local Public Health Practice. Preventing Chronic Disease, 16. https://doi.org/10.5888/pcd16.180278

Li, H. (2023, May 4). L.A. tourists are (mostly) back — except some big spenders - Los Angeles Times. Los Angeles Times. https://www.latimes.com/business/story/2023-05-04/la-fi-tourism-mostly-back-except-biggest-spenders

Mahajan, K., Joshi, V., Khedkar, M., Galani, J., & Kulkarni, M. (2021). Restaurant Recommendation System using Machine Learning. International Journal of Advanced Trends in Computer Science and Engineering, 10(3), 1671–1675. https://doi.org/10.30534/ijatcse/2021/261032021

Mayorquin, J. (2021, March 23). California Dreaming: Cultural diversity - one of the Golden State's greatest assets. ABC7 Los Angeles. https://abc7.com/california-dream-solutions-culture/10327371/

Most visited travel and tourism websites worldwide 2023 | Statista. (2023, October 17). Statista. https://www.statista.com/statistics/1215457/most-visited-travel-and-tourism-websites-worldwide/

Munaji, A. A., & Emanuel, A. W. R. (2021). Restaurant Recommendation System Based on User Ratings with Collaborative Filtering. IOP Conference Series: Materials Science and Engineering, 1077(1), 012026. https://doi.org/10.1088/1757-899x/1077/1/012026