# Identifying Experts in the Yelp Dataset

**Brandon Kates**
Cornell University
Biometry and Statistics
bjk224@cornell.edu

**Brian Cheang**
Cornell University
Computer Science
bmc78@cornell.edu

**Abstract**

High ratings and awards can drive large tourist crowds into local favorite restaurants, often causing restaurants to change (increase prices, new booking rules, impersonal service) to better accommodate the new customer base. As a result, the most popular and highest-rated restaurants may no longer be true local favorites but instead thrive on their popularity with tourists. By identifying local users and local experts to provide ratings reflective of local opinions, Yelp may become more popular with tourists looking to "travel like a local" and enjoy a more authentic experience.

## 1 Introduction

The objective of the project is to build a model to classifying users as experts in locality and category. The trouble with the dataset is that there are no labels for user locality or experts among users. In this paper, we propose both local authority and topical authority models to identify these local experts. The "Local Authority Model" is a Gaussian Mixture Model that identifies clusters in each user's review locations to predict the user's most probable location. The model looks at time spans, and review ratios to improve on the model that was described in the Jindal Thesis. The "Topical Authority Model" employs k-means clustering on a set of users who have written reviews for a specific category to determine if the user is an expert in that category. In the Jindal thesis, she employs supervised learning techniques to learn which users are experts by using 'elite' users as the ground truth. The model described in this paper does not rely on these potentially flawed tags to make predictions.

## 2 Local Authority Model

The objective of the Local Authority Model is to identify Yelp reviews that are written by users that are locals in the area. Because there are no labels in the dataset that indicate whether a review is written by a local or non-local, we need to design our own "Local Authority Model" to predict this information. Our proposed model identifies reviews by locals using information of distance, timespan, and review ratio.

### 2.1 Local Authority Model 1 (LAM1): Distance

The initial step to building the local authority model is to locate a user given their review locations. User locations are predicted by fitting a Gaussian Mixture Model onto a given user's included review locations. As the number of distinct metropolitan areas (12 total areas included in Yelp data set) visited by each user is unknown, the local authority model should start with 12 mixture components and iterate down. However, for the sake of time, our Gaussian Mixture Model starts with 4 clusters and iterates down (3 clusters, 2 clusters...) until the set of clusters are distinct. The center of the most probable mixture component is the predicted location of a user. With all user locations predicted, the distance (in kilometers) between reviewer and business of every review can be calculated using the haversine formula. Our LAM1 will identify reviews by locals by finding all reviews in the dataset with distance value less than a predetermined constant DISTANCE_THRESHOLD.

### 2.2 Local Authority Model 2 (LAM2): Distance + Time Span

If a tourist visits city A for a week and writes several Yelp reviews during his/her stay, the tourist's reviews may be mistakenly labeled as reviews by a local. A potential way to improve LAM1 is to filter out

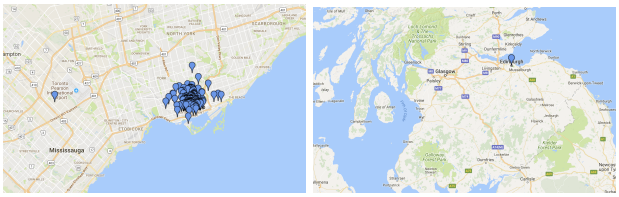

Fig. 1. Review locations of a sample user A


Fig. 2. Identified cluster centers for user A

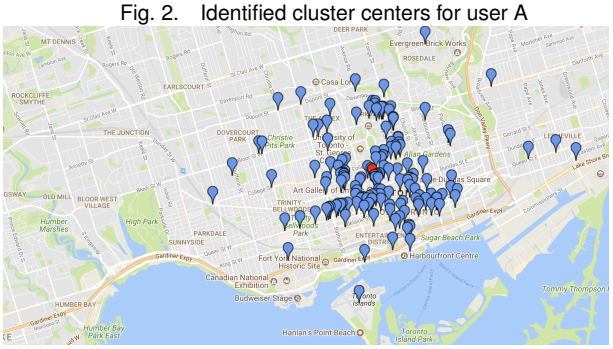

Fig. 3. Predicted location of user A in red

'reviews by locals' that are written over a short time span. This is found by taking the difference between the most recent and least recent date of local reviews written by a user. LAM2 then filters out all reviews that have a 'time_span' value less than a predetermined constant TIMESPAN_THRESHOLD.

### 2.3 Local Authority Model 3 (LAM3): Distance + Review Ratio

Only reviews from 12 metropolitan areas were included in the dataset. As a result, predicting user locations using their provided review locations only may not be very accurate. One possible way to improve LAM1 is to filter out 'review by locals' that are written by users with a very small ratio of their reviews in their predicted location. Review ratios of users can be found by dividing the number of reviews written by a user in their predicted area by the total review written by the user. With this information, LAM3 filters out 'review by locals' that are written by users with 'review_ratio' less than a predetermined RATIO_THRESHOLD.

### 2.4 Data

Because of time constraints, our Local Authority Model was run on a subset of users in the Yelp dataset. Although there are over a million users included in the dataset, the majority of the users only have a few reviews included in the dataset (Only 94,392 users in the dataset have more than 8 reviews included). Evaluation of our Local Authority Models is done using a dataset of 90,000 users with 2,368,680 reviews in total.

### 2.5 Evaluation

Assuming locals in different areas have their own unique tastes, some businesses will differ in ratings by locals and ratings by tourists. With this assumption, our Local Authority Model is significant if it can identify a group of local users such that the average ratings of businesses by these local users is more accurate than the Yelp ratings of businesses for predicting local ratings. The baseline is the predictive ability of Yelp business ratings. We quantify this by producing a mean-squared-error of using the Yelp business ratings to predict actual user review ratings. Then we filter all reviews to find "Local Reviews" and again produce mean-squared-errors of using the average ratings of restaurants to predict actual user review ratings in the subset of local reviews. We assume that if our model is able to identify local users into such that the mean-squared-errors of "Local Rating Prediction" is lower than the baseline Yelp rating prediction, then our model is more indicative of local opinions than the current Yelp star ratings. The mean-squared-error of using Yelp ratings to estimate all user review ratings in our dataset is MSE = 1.3002 so our goal is to create Local Authority Model with a lower MSE for predicting local ratings.

#### 2.5.1 Local Authority Model 1 Evaluation

By optimizing the MSE of LAM1, we find that at DISTANCE_THRESHOLD = 0.7295875, the model identifies 157,159 total reviews written by 46,764 different users and has the minimal MSE of 1.19238. This suggests that as DISTANCE_THRESHOLD increases beyond 0.7295875 kilometers, the radius of locality is too large and increasing numbers of non-local reviews are mislabeled leading to average ratings that are less representative of local tastes. On the other hand, as DISTANCE_THRESHOLD decreases below 0.7295875 kilometers, the MSE increases. This suggests there are too few locals and average ratings become determined by personal tastes instead of the local taste. The results show that at DIS-

TANCE_THRESHOLD = 0.7295875, LAM1 (MSE = 1.19238) has the minimal MSE for our dataset.

### 2.5.2 Local Authority Model 2 Evaluation

After plotting the relationship between TIMESPAN_THRESHOLD and MSE of LAM2, we find that there is a general trend of the MSE decreasing as TIMESPAN_THRESHOLD increases. This seems logical since locals who have been in the city for the longest are probably most understanding of the local taste. However, MSE fluctuates and does not decrease steadily as TIMESPAN_THRESHOLD increases. However, the minimal MSE (0.985588) can be found at TIMESPAN_THRESHOLD = 3609.375 (32,317 total reviews by 9,344 different users). Using DISTANCE_THRESHOLD = 0.7295875 and TIMESPAN_THRESHOLD = 3609.375, LAM2 has the minimal MSE of 0.985588 for our dataset.

### 2.5.3 Local Authority Model 3 Evaluation

After plotting the RATIO_THRESHOLD and MSE of LAM3, we find that the optimum RATIO_THRESHOLD is at 0. MSE of LAM3 slowly increases as RATIO_THRESHOLD increases from 0. This result is surprising since it suggests that filtering out reviews written by users with low ratios of reviews in the city does not help identify the local taste of the city but instead makes the prediction worse. The results show that MSE only increases as RATIO_THRESHOLD increases from 0 so there is no RATIO_THRESHOLD that can minimize the baseline MSE of our dataset.

### 2.5.4 Evaluation Results

After optimizing our Local Authority Model on distance, time span, and review ratio, we found that the MSE of our Local Authority Model is minimized at DISTANCE_THRESHOLD = 0.7295875 km, TIMESPAN_THRESHOLD = 3609.375 days, and RATIO_THRESHOLD = 0. These results suggest that 9,344 users that live within 0.73 kilometers of a restaurant and have written reviews in the area for at least 3,609 days have the most similar tastes.

## 3 Topical Authority Model

Each business in the Yelp dataset contains a list of categories that the business belongs to. For example, an Italian restaurant might contain the labels, 'Italian', 'Restaurant', and 'Pizza'. There are more than

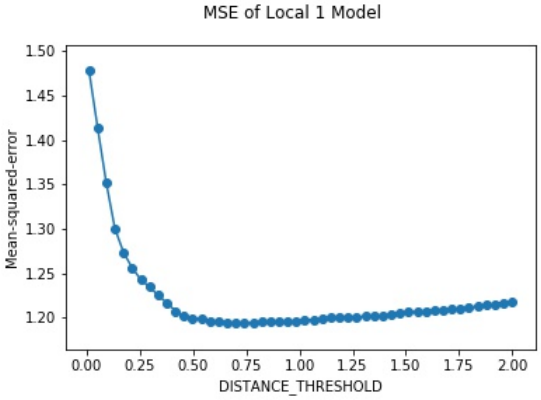


Fig. 4. Mean-squared-error of LAM1

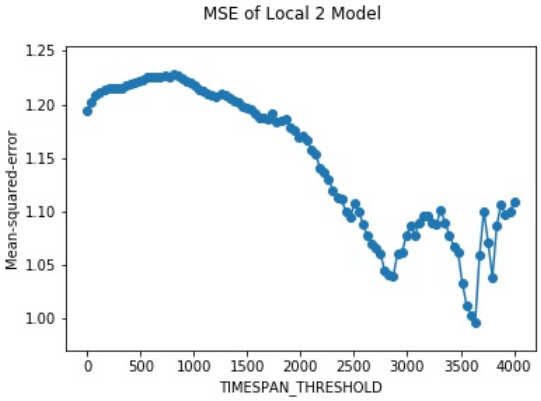


Fig. 5. Mean-squared-error of LAM2

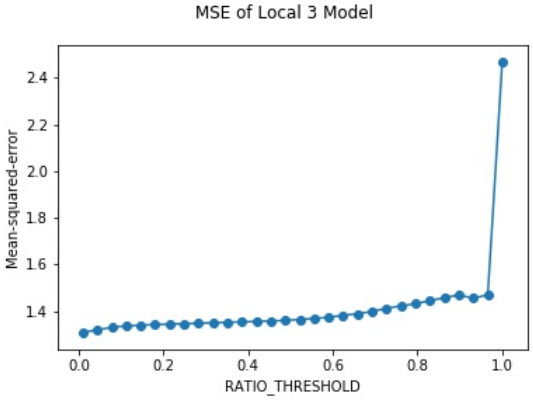


Fig. 6. Mean-squared-error of LAM3

1200 categories in the dataset, but we will focus on only some of the categories with the greatest number of businesses. The model classifies users who have written reviews about a business in the category as experts vs. non-experts. In order to determine if a user knows enough information about a category to be considered an expert, we define a feature set that is used to classify users.

The topical authority model uses the k-means clustering algorithm with 2 clusters on our chosen feature sets in order to classify users as experts or non-experts

in the specific category.

### 3.1 Feature Selection

- **Yelp Age:** This is the amount of time since the user created their Yelp account. We quantify this as the number of months since creation. A brand new user on Yelp probably does not have the expertise of users who have been using the app for much longer.
- **Total Reviews:** The total number of reviews that a user has written. While the quality of the reviews is unknown, this gives insight into a user's activity level.
- **Category Reviews:** The number of reviews the user has written about this category. Gives insight into how many times the user has interacted with this category of restaurant. We are also interested in the number of reviews that the user has written in a category relative to the total number of reviews that they have written. If most of a user's reviews are about a specific category, that might make them more of an expert.
- **Average Rating:** Average star rating given out by this user to businesses.
- **Standard Deviation of Ratings:** When coupled with the average, this can give some insight into the consistency of the user's ratings.
- **Funny, Useful, and Cool Votes:** These are the number of useful, funny, and cool votes that a user receives for a review in the specific category. A user who has been given a bunch of useful votes is probably more knowledgeable than one who has a bunch of cool and funny votes.
- **Unique Businesses:** This is the number of unique businesses that the user has reviewed. This feature shows us the spread of a user's knowledge in a specific category. If a user only reviews the same restaurant over and over, then that might only make them an expert on that specific restaurant, and not the category as a whole.
- **Number of Friends:** This is the number of friends that a user has on Yelp. This metric can give us an idea of how well-connected a user is on Yelp.

Our first feature set includes all of the above features except for the number of friends, while the second feature set includes all of the above.

### 3.2 Category Selection

In order to use our feature set, we first selected the categories that we wanted to train the model. Table 1 gives us categories with the most number of businesses. Table 2 shows the categories with the most number of businesses, given that they contain the label 'Restaurants'.

| Category | No. of Businesses |
|---|---|
| Restaurants | 51613 |
| Shopping | 24595 |
| Food | 23014 |
| Beauty & Spas | 15139 |
| Home Services | 13202 |
| Health & Medical | 12033 |
| Nightlife | 11364 |
| Bars | 9868 |
| Automotive | 9476 |
| Local Services | 9343 |

Table 1. Top 10 Categories in the Yelp Dataset

| Category | No. of Businesses |
|---|---|
| Restaurants | 51613 |
| Food | 9599 |
| Nightlife | 6969 |
| Bars | 6690 |
| Sandwiches | 5864 |
| Fast Food | 5792 |
| American (Traditional) | 5737 |
| Pizza | 5652 |
| Italian | 4411 |
| Burgers | 4236 |

Table 2. Top 10 Categories for businesses containing restaurant category

The categories presented here can be quite different from each other. Therefore, it would not make sense to

use one model to train on all users and classify them as experts for multiple different categories. Our approach runs the model on each category one by one.

### 3.3 Unsupervised Learning Model

Our first approach at a learning model involved training a supervised learning model with the Yelp 'Elite' tag as a label for an expert, and our feature set described above. However, that posed a few problems. Namely, Yelp users nominate themselves for the tag, and there is no validation to say that they have more knowledge than anyone else, only that they have a desire to have a virtual tag. Additionally, the elite tags are independent of category, so an elite user who has written no reviews about a category still has a great chance of being classified as an expert in that category.

Therefore, we moved on to an unsupervised learning model involving running k-means cluster with two clusters on the feature set. Our assumption is that the cluster with fewer users are the experts. This assumption should hold for the categories that we chose to run the data on because they all have a huge number of businesses and a huge number of users associating reviews with those businesses. We want to observe meaningful differences between the two clusters, to ensure that the users in the expert cluster are in some way different from those in the non-expert cluster. An example of the algorithm run on the 'pizza' category can be observed in fig 4.

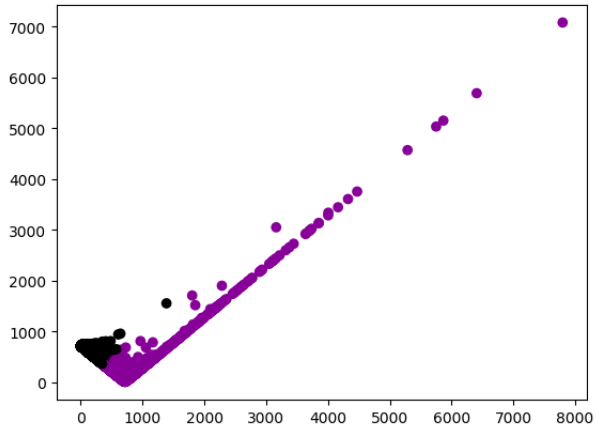


Fig. 7. K-means cluster on feature set for 'pizza' category. Black dots represent users classified as expert, purple are non-experts

### 3.4 Evaluation

We describe the design of our evaluation model for the clustering algorithm. We want to show that we are confident that the algorithm separates users into two groups in a meaningful way. We find both the mean squared error, and the mean absolute error of the average star rating for users in each business. For each business, separate users into expert and non-expert, and then calculate the average star rating for experts and non-experts for each business. We calculate the MAE and MSE business by business with the average star-rating for each group in the business.

In order to see that there is a meaningful selection of expert and non-expert, we should observe that the selection of experts from each business is not just a random sample of all the users who have written reviews about that business. In that case, we should observe very small MSE and MAE. If the values are larger, then that could represent that we have two distinct groups of users. This is just one way to evaluate the model. One issue that comes up is illustrated by fig 5. Over 80% of the businesses have between 0 and 5 experts classified for that business (the figure only shows pizza restaurants but the relationship extends to all categories that we have observed). Additionally, a majority of businesses have no experts at all, and so we do not include those when making our predictions.

### 3.5 Evaluation Results

Our evaluation results for the topical expert model is shown in appendix A, and appendix B. The first table shows the first feature set run on the top 100 categories in the Yelp Dataset. We observe that a majority of the MSE and MAE for each category are between 0 and 1. Therefore, we do not observe a very large difference in star predictions for experts and non-experts in most of the categories.

## 4 Further Studies

This paper describes a method to predict both a users knowledge of a location and in a specific category of business. The work is novel in that it employs time span analysis and review ratios in order to look at a user's location. Additionally, we employ an unsupervised learning approach to determining a user's expertise in a specific type of business.

To improve the local authority model, we want to explore using other information included in the Yelp dataset to better identify locals. One possible improvement is to incorporate Yelp friendship data into our

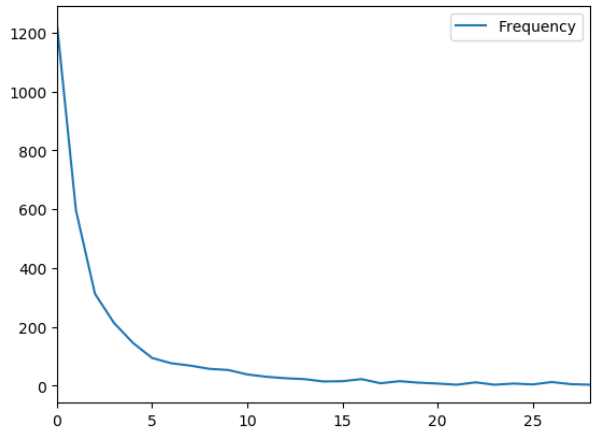


Fig. 8. Frequency of the number of experts in each pizza restaurant.

model to predict user localities using information about their Yelp friends.

To improve the topical authority model, we want to do phrase-level sentiment analysis on the text similar to the second reference and use information from that analysis as features in our model. Additionally, we could use collaborative filtering similar to the third reference to get more information about a users tastes. Essentially, we would like to look at the actual review text to make predictions about a user's expertise on a subject.

Our final improvement is to combine the two models. The goal is for a Yelp user to look at a local restaurants reviews, and see a section where they can view local experts reviews determined by this model. This improves the quality of information that a user receives about a restaurant. In this paper, we were focused on developing the two models independently to maximize their success.

## References

[1] Jindal, Tanvi. Finding Local Experts from Yelp Dataset. University of Illinois at Urbana-Champaign. Published 2015-04-27.

[2] Zang, Yonfeng. Incorporating Phrase-level Sentiment Analysis on Textual Reviews for Personalized Recommendation. State Key Laboratory of Intelligent Technology and Systems Department of Computer Science and Technology Tsinghua University, Beijing, 100084, China. Published February 2015.

[3] Chee Hoon Ha. Yelp Recommendation System Using Advanced Collaborative Filtering. Stanford University.

**Appendix A: Evaluation of Feature Set 1 on Top 100 Categories in Yelp Dataset**

| | Categories | MAE | MSE | No. Business | No. Businesses w/ Experts |
|---|---|---|---|---|---|
| 0 | [Restaurants] | 0.598281 | 0.684942 | 51613 | 39058 |
| 1 | [Shopping] | 0.785871 | 1.106890 | 24595 | 12018 |
| 2 | [Food] | 0.621124 | 0.727511 | 23014 | 16759 |
| 3 | [Beauty & Spas] | 0.796088 | 1.145789 | 15139 | 3798 |
| 4 | [Home Services] | 0.981016 | 1.754910 | 13202 | 1939 |
| 5 | [Health & Medical] | 0.959586 | 1.626217 | 12033 | 2595 |
| 6 | [Nightlife] | 0.529242 | 0.553824 | 11364 | 8664 |
| 7 | [Bars] | 0.517152 | 0.530951 | 9868 | 7610 |
| 8 | [Automotive] | 0.941832 | 1.563556 | 9476 | 2619 |
| 9 | [Local Services] | 0.856467 | 1.361653 | 9343 | 1965 |
| 10 | [Event Planning & Services] | 0.605917 | 0.706447 | 8038 | 4503 |
| 11 | [Active Life] | 0.628629 | 0.764121 | 7427 | 3089 |
| 12 | [Fashion] | 0.764410 | 1.033371 | 6299 | 3542 |
| 13 | [Sandwiches] | 0.626335 | 0.721647 | 5864 | 3917 |
| 14 | [Fast Food] | 0.785715 | 1.037209 | 5792 | 3375 |
| 15 | [American (Traditional)] | 0.552005 | 0.577276 | 5737 | 4429 |
| 16 | [Pizza] | 0.663581 | 0.792939 | 5652 | 3179 |
| 17 | [Coffee & Tea] | 0.616216 | 0.685857 | 5565 | 3860 |
| 18 | [Hair Salons] | 0.773942 | 1.089876 | 5395 | 1217 |
| 19 | [Hotels & Travel] | 0.720396 | 0.987315 | 5188 | 3010 |
| 20 | [Arts & Entertainment] | 0.551997 | 0.611507 | 5054 | 3451 |
| 21 | [Home & Garden] | 0.788624 | 1.088657 | 4584 | 1766 |
| 22 | [Auto Repair] | 0.968630 | 1.650734 | 4480 | 1291 |
| 23 | [Italian] | 0.586844 | 0.637572 | 4411 | 2828 |
| 24 | [Burgers] | 0.668776 | 0.822419 | 4236 | 2920 |
| 25 | [Doctors] | 1.148306 | 2.130415 | 4124 | 1013 |
| 26 | [Breakfast & Brunch] | 0.520799 | 0.537607 | 4103 | 3180 |
| 27 | [Mexican] | 0.612205 | 0.681817 | 3913 | 2751 |
| 28 | [Nail Salons] | 0.878989 | 1.321873 | 3884 | 1345 |
| 29 | [Professional Services] | 0.883832 | 1.557725 | 3865 | 532 |
| 30 | [American (New)] | 0.476821 | 0.460479 | 3802 | 3093 |
| 31 | [Chinese] | 0.584070 | 0.631825 | 3775 | 2476 |
| 32 | [Real Estate] | 1.082200 | 2.023524 | 3729 | 500 |
| 33 | [Specialty Food] | 0.557530 | 0.600614 | 3620 | 2186 |
| 34 | [Fitness & Instruction] | 0.680475 | 0.898487 | 3615 | 1028 |
| 35 | [Pets] | 0.800193 | 1.175038 | 3153 | 819 |
| 36 | [Grocery] | 0.634961 | 0.694215 | 3044 | 1886 |
| 37 | [Bakeries] | 0.575669 | 0.632179 | 3014 | 1948 |
| 38 | [Cafes] | 0.577827 | 0.638674 | 2812 | 1749 |
| 39 | [Hair Removal] | 0.757687 | 1.118003 | 2704 | 840 |
| 40 | [Dentists] | 0.819130 | 1.358825 | 2683 | 444 |
| 41 | [Hotels] | 0.634251 | 0.723375 | 2548 | 1703 |
| 42 | [Desserts] | 0.515023 | 0.512716 | 2419 | 1718 |
| 43 | [Skin Care] | 0.758855 | 1.081546 | 2398 | 601 |
| 44 | [Women's Clothing] | 0.760329 | 0.992372 | 2355 | 1270 |
| 45 | [Education] | 0.654132 | 0.815976 | 2209 | 572 |
| 46 | [Japanese] | 0.521920 | 0.495257 | 2186 | 1658 |
| 47 | [Ice Cream & Frozen Yogurt] | 0.605774 | 0.676867 | 2159 | 1378 |
| 48 | [Pet Services] | 0.853963 | 1.287989 | 2113 | 455 |
| 49 | [Day Spas] | 0.743679 | 1.002800 | 2084 | 634 |

**Appendix B: Evaluation of Feature Set 1 on Top 100 Restaurant Categories in Yelp Dataset**

| | Categories | MAE | MSE | No. Business | No. Businesses w/ Experts |
|---|---|---|---|---|---|
| 0 | [Food] | 0.621355 | 0.728023 | 23014 | 16780 |
| 1 | [Nightlife] | 0.529098 | 0.553666 | 11364 | 8664 |
| 2 | [Bars] | 0.517141 | 0.530943 | 9868 | 7610 |
| 3 | [Sandwiches] | 0.626472 | 0.721940 | 5864 | 3917 |
| 4 | [Fast Food] | 0.785715 | 1.037209 | 5792 | 3375 |
| 5 | [American (Traditional)] | 0.552084 | 0.577423 | 5737 | 4428 |
| 6 | [Pizza] | 0.663463 | 0.792701 | 5652 | 3179 |
| 7 | [Italian] | 0.586844 | 0.637572 | 4411 | 2828 |
| 8 | [Burgers] | 0.668778 | 0.822420 | 4236 | 2920 |
| 9 | [Breakfast & Brunch] | 0.520828 | 0.537614 | 4103 | 3180 |
| 10 | [Mexican] | 0.612205 | 0.681817 | 3913 | 2751 |
| 11 | [American (New)] | 0.476821 | 0.460479 | 3802 | 3093 |
| 12 | [Chinese] | 0.584842 | 0.633067 | 3775 | 2478 |
| 13 | [Cafes] | 0.577827 | 0.638674 | 2812 | 1749 |
| 14 | [Coffee & Tea] | 0.616216 | 0.685857 | 5565 | 3860 |
| 15 | [Japanese] | 0.521920 | 0.495257 | 2186 | 1658 |
| 16 | [Chicken Wings] | 0.740284 | 0.934404 | 2019 | 1136 |
| 17 | [Seafood] | 0.506835 | 0.478674 | 1981 | 1487 |
| 18 | [Event Planning & Services] | 0.605917 | 0.706447 | 8038 | 4503 |
| 19 | [Salad] | 0.571266 | 0.584189 | 1850 | 1275 |
| 20 | [Sushi Bars] | 0.517792 | 0.508354 | 1833 | 1332 |
| 21 | [Delis] | 0.612368 | 0.681249 | 1564 | 984 |
| 22 | [Asian Fusion] | 0.520482 | 0.508863 | 1554 | 1115 |
| 23 | [Mediterranean] | 0.580983 | 0.625337 | 1482 | 939 |
| 24 | [Sports Bars] | 0.548548 | 0.542362 | 1781 | 1252 |
| 25 | [Barbeque] | 0.580198 | 0.618040 | 1413 | 982 |
| 26 | [Canadian (New)] | 0.540431 | 0.515630 | 1357 | 749 |
| 27 | [Specialty Food] | 0.557530 | 0.600614 | 3620 | 2186 |
| 28 | [Steakhouses] | 0.492998 | 0.456563 | 1311 | 1013 |
| 29 | [Thai] | 0.561244 | 0.568897 | 1291 | 855 |
| 30 | [Indian] | 0.580127 | 0.596135 | 1289 | 661 |
| 31 | [Pubs] | 0.497740 | 0.483507 | 2024 | 1485 |
| 32 | [Caterers] | 0.585989 | 0.637740 | 1714 | 1005 |
| 33 | [Bakeries] | 0.575669 | 0.632179 | 3014 | 1948 |
| 34 | [Desserts] | 0.515023 | 0.512716 | 2419 | 1718 |
| 35 | [Diners] | 0.559957 | 0.560445 | 1132 | 783 |
| 36 | [Middle Eastern] | 0.637157 | 0.711357 | 958 | 587 |
| 37 | [Greek] | 0.586468 | 0.608289 | 956 | 586 |
| 38 | [French] | 0.498166 | 0.475922 | 952 | 603 |
| 39 | [Vietnamese] | 0.559550 | 0.555898 | 928 | 625 |
| 40 | [Vegetarian] | 0.478209 | 0.447760 | 889 | 662 |
| 41 | [Wine Bars] | 0.454947 | 0.425366 | 978 | 717 |
| 42 | [Beer] | 0.529093 | 0.535560 | 1657 | 1197 |
| 43 | [Wine & Spirits] | 0.529093 | 0.535560 | 1657 | 1197 |
| 44 | [Buffets] | 0.583744 | 0.619237 | 743 | 547 |
| 45 | [Arts & Entertainment] | 0.551997 | 0.611507 | 5054 | 3451 |
| 46 | [Korean] | 0.542197 | 0.545643 | 697 | 540 |
| 47 | [Lounges] | 0.533867 | 0.560422 | 1346 | 968 |
| 48 | [Cocktail Bars] | 0.479695 | 0.462479 | 894 | 667 |
| 49 | [Tex-Mex] | 0.643292 | 0.740124 | 644 | 405 |