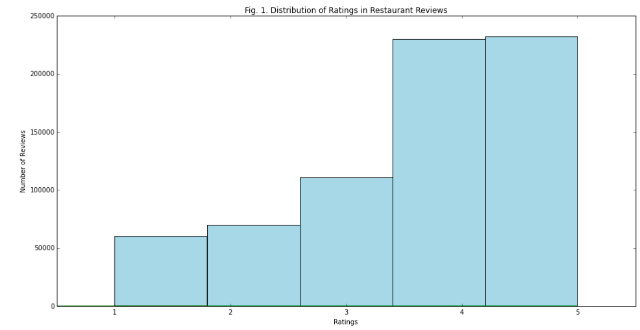
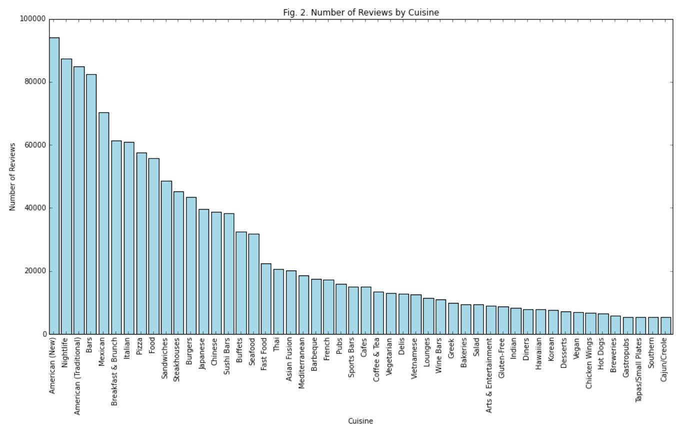
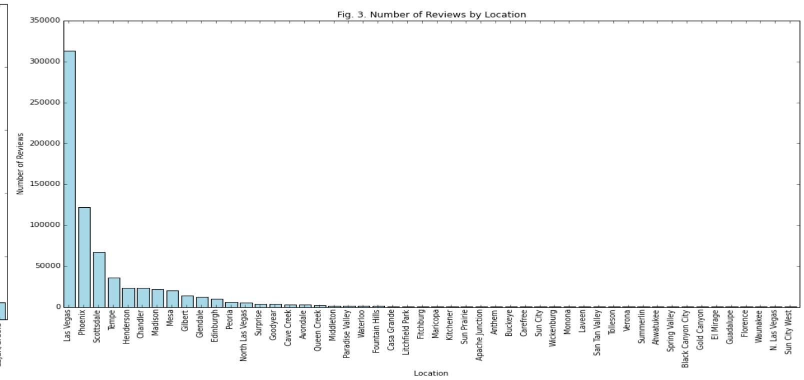
Restaurant Reviews
(Exploratory Analysis of Yelp dataset)

1. Data Cleansing and general exploratory analysis

The dataset contains 42,153 businesses, of which 14,303 are restaurants (identified by presence of 'Restaurants' category). After removing all businesses without 'Restaurants' category, the dataset has 1,125,458 reviews in total with 706,646 reviews for the restaurants.
Initial review of the categories for restaurants showed 240 categories with some of these categories assigned very few businesses, most with irrelevant labels like "Dry Cleaning & Laundry" and "Sporting Goods". Filtering out any categories with less than 10 businesses, leaves 123 categories and 14,057 restaurants with 702,816 reviews. Fig 1 shows the distribution of ratings ('stars' attribute) among reviews.





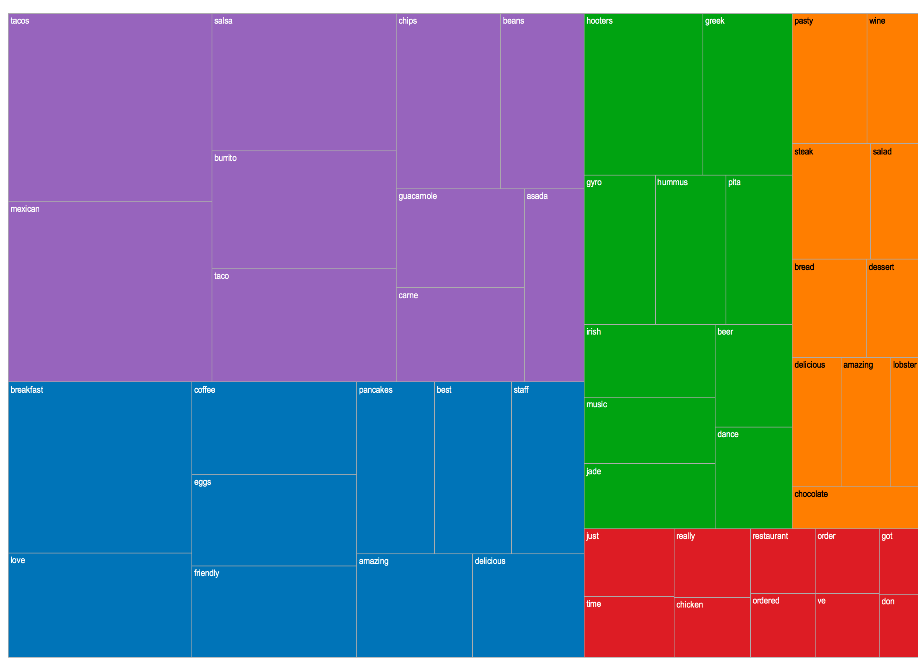


We see a very sharp drop in number of reviews after 15-17 most popular categories, which is to be expected – majority of restaurants stick to few most popular cuisines. Fig. 2 shows the distribution of reviews by category for top 50 categories.

Geographically, most of the reviews in the dataset are in Nevada and Arizona, with overwhelming majority in Las Vegas, NV, as shown in Fig. 3.

2. Review topic extraction (Task 1.1)

Fig. 4. Top 10 words for review topics



I performed topic extraction on the full set of reviews, using LDA. Words appearing in more than 30% of the reviews, words in Pyhon scikitlearn 'english' stopword list, and words appearing less than 5 times in all reviews were ignored.
LDA model was built for 5 topics, doing 10 passes and 1,000 iterations. Fig. 4 shows top 10 words for each of the topics. Words belonging to the same topic are grouped by color, and size of each word's container is proportional to its weight within the topic.
We can see several very clear topics here: Mexican cuisine (purple) - tacos, mexican, salsa, burrito, taco, beans, guacamole, chips, carne, asada; Breakfast (blue) - breakfast, pancakes, coffee, eggs; American cuisine (orange) - wine, steak, salad, pasty, bread, dessert; Mediterranean/Bars (green) - greek, gyro, hummus, pita, hooters, irish, beer, music, dance, and Other (red).
For my first runs of LDA I used cutoffs of words appearing in more than 50% of the reviews and words appearing less than 2 times in all reviews. I used 10 topics, on random subsets of the reviews, which gave much worse results. I also initially ran LDA with fewer iterations – also much worse results, as the number of iterations was insufficient for LDA to converge. However, even increasing number of iterations to 3,000 with 10 topics did not significantly improve results. I got current improvement

by reducing the number of topics to 5 and changing cutoffs to 30% and 5 occurrences respectively. Some results from alternative runs are shown in Appendix A.

3. Positive/Negative Reviews Topic comparison (Task 1.2)

Next, I extracted from the full set of reviews only those for Mexican cuisine (70,406), and split them into two sets: positive reviews (ratings 4 and 5 – 44,157 reviews) and negative reviews (ratings 1 and 2 – 15047 reviews). I dropped reviews rated 3 (11,202 reviews) as neutral to get better separation between positive and negative ratings. On these two sets, I again used LDA for topic extraction with 3 topics, word cutoff at 50% and minimum 5 occurrences, and 1,000 iterations.
The results are in Fig. 5 and 6. In the positive reviews, it's pretty clear that one topic (blue) is about food quality, the other

Fig. 5. Top 10 words for Mexican restaurants positive review topics

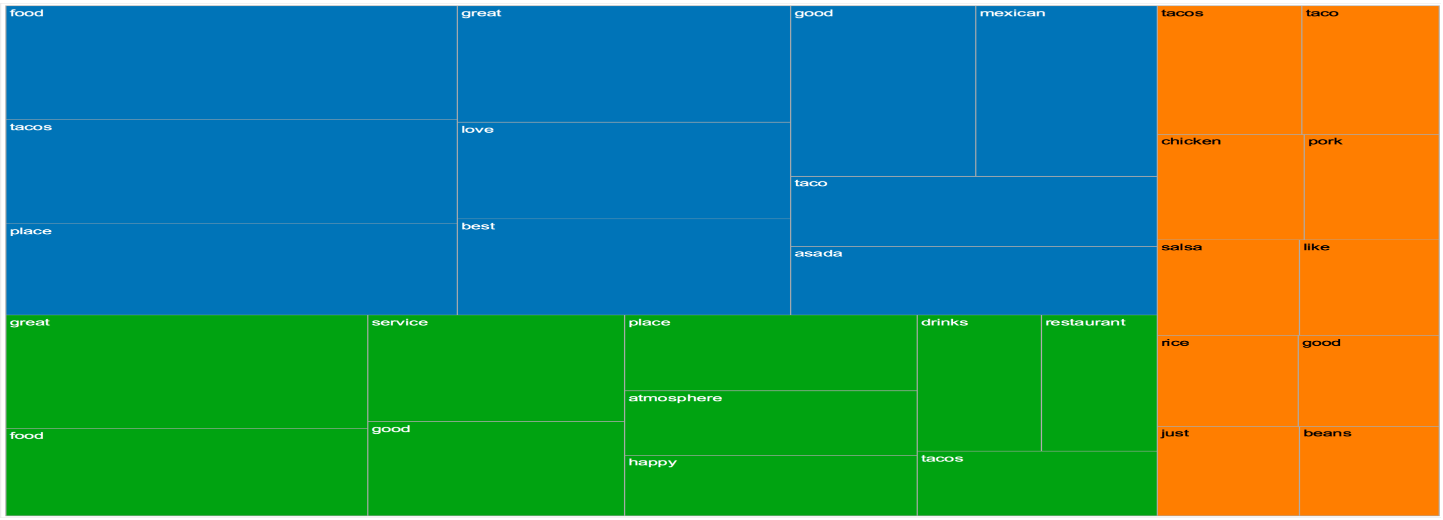
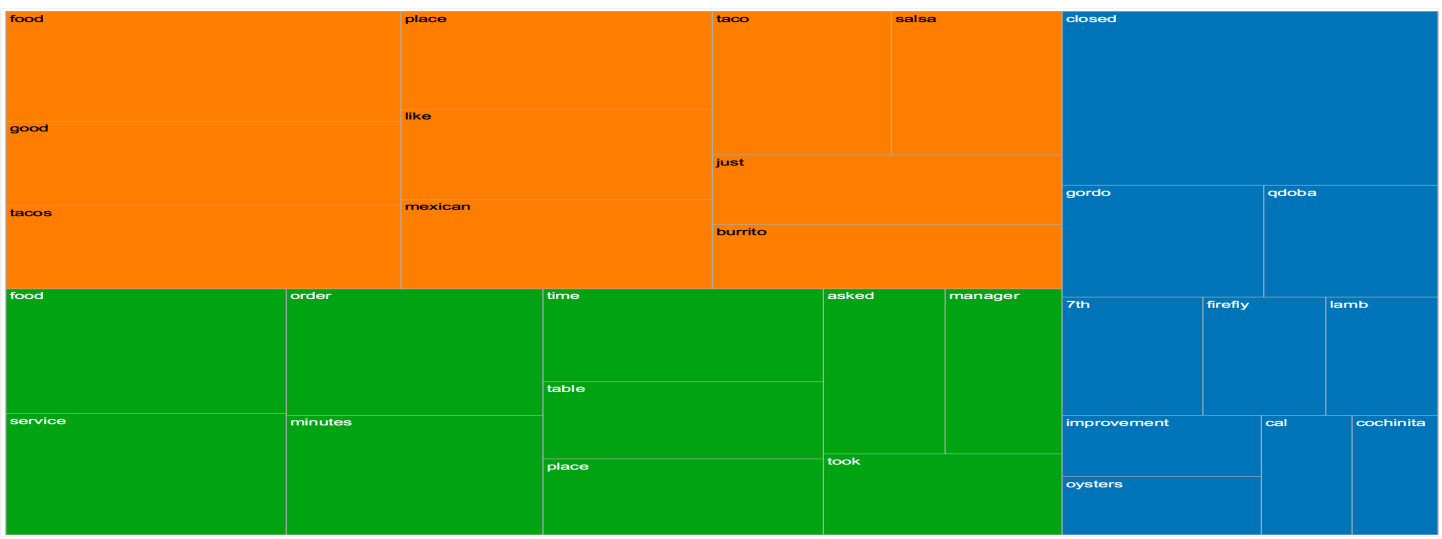


Fig. 6. Top 10 words for Mexican restaurants negative review topics



(green) is about service quality, and the third (orange) is "other". In the negative reviews, the split is similar: quality of service and wait times (green), quality of food (orange), which has words like "good" and "like", demonstrating the downside of 1-gram tokenization – they came out of "not good" and "don't like" in the review texts. Interestingly the third (blue) topic in negative reviews captures not just "other", but also reviews related to work hours and/or possibly restaurants that went out of business (highest weighted word is "closed").
Similar to Task 1.1, I tried running topic extraction with different parameters. I started with 5 topics, which turned out to be too many, as there was no clear distinction between topics. See Appendix B for those results.

4. Next steps

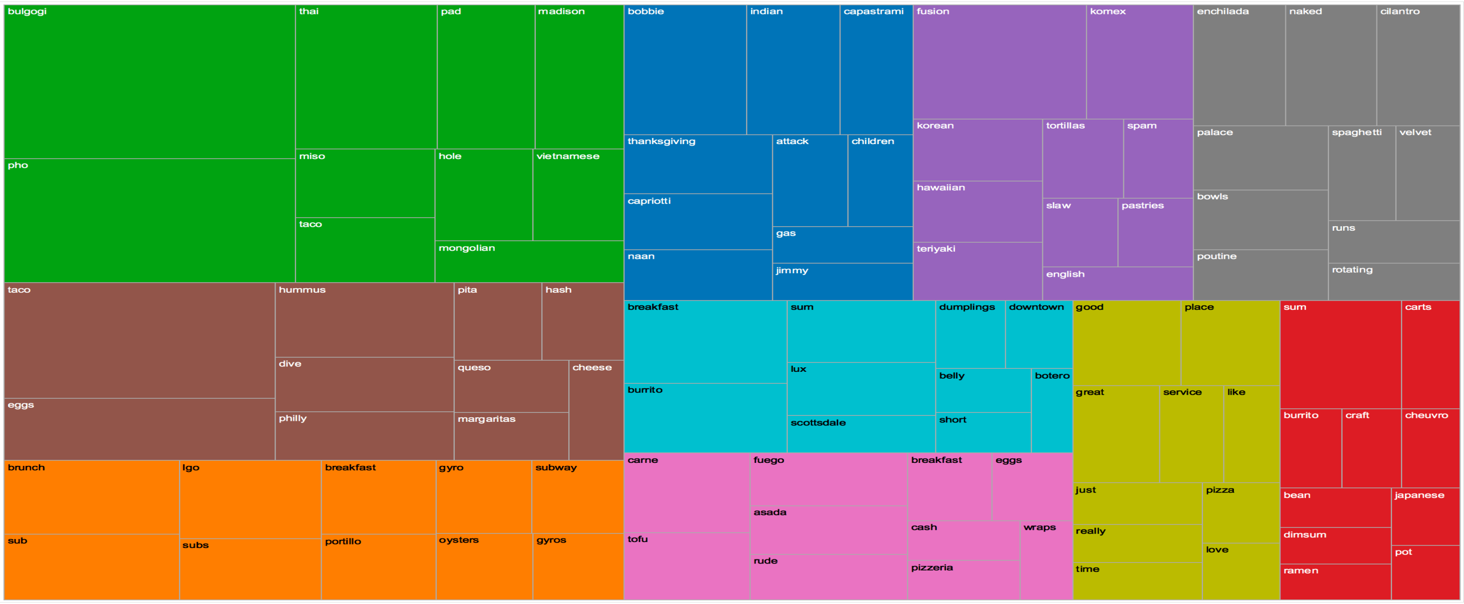
While the resulting topics make sense, they can still be improved. As next steps I plan to use stemming to combine multiple variants of the same word (e.g., in current analysis "taco" and "tacos" are treated as separate words), and switch to 2- and 3-grams.

5. Tools Used

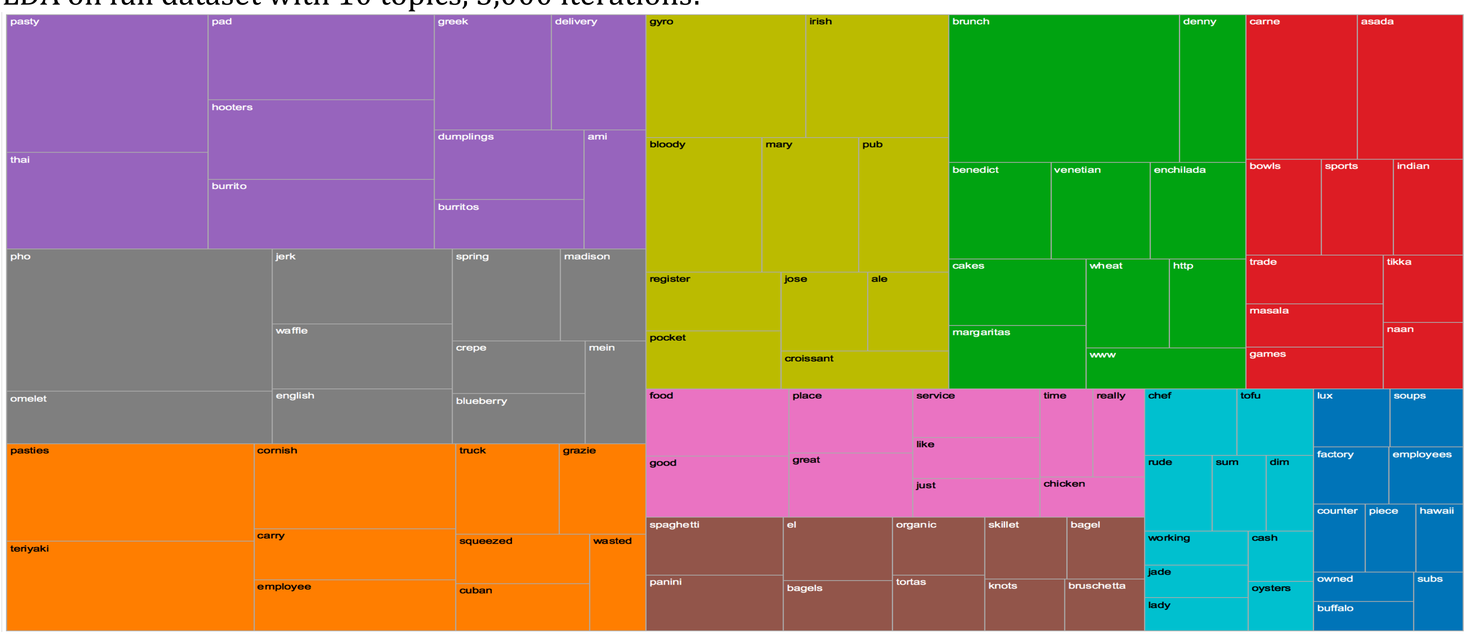
Data processing: Python
Visualization: Python matplotlib library (Fig. 1-3), Tableau (Fig. 4-6)
LDA topics, words and weights were exported to CSV files from Python for visualization in Tableau.

Appendix A. Alternative results for Task 1.1

LDA on a random sample of 100,000 reviews (~14% of the total set of reviews), with 10 topics, 1000 iterations:



LDA on full dataset with 10 topics, 3,000 iterations:



Appendix B. Alternative results for Task 1.2

5-topic LDA with 1000 iterations for Positive/Negative Reviews