# Preface

1. Code environment
   1. Python 3.7 via Jupyter Notebooks was used to create this solution.

**Word cloud inspiration:** <https://courses.cognitiveclass.ai/courses/course-v1:CognitiveClass+DV0101EN+v1/courseware/89227024130b43f684d95376901b65c8/e7c36d2c4c6840fe8b81b97147ea9c16/>

**LDA Inspiration:** https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/

**TextBlob:** https://textblob.readthedocs.io/en/dev/quickstart.html

# NLP

1. Data exploration
   1. What is the source of this data?

<https://www.yelp.com/dataset>

* 1. Overall data structure: how many rows and columns are in this dataset?

The merged data set has 1,283,301 rows and 23 columns, when we look at just restaurants in Illinois, and drop columns we aren’t going to use, we have 4,808 rows and 18 columns.

* 1. List any data cleaning steps and assumptions you make in your NLP analysis:

The three data frames used were merged on the business ID. We are assuming the mayor only cares about the state of Illinois and restaurants, so that is all we look at in this analysis.

1. Data visualization  
   Create a word cloud to show the most popular keywords or phrases that reviewers use for each cuisine.
   1. For the 10 most popular cuisines, what are the top keywords or phrases used by reviewers?

|  |  |
| --- | --- |
| Cuisine | Top keywords & phrases |
| American | Food, place, great, good, service, time, restaurant, delicious |
| Burger | Food, good, place, order, time, service, burger, great, fries |
| Mexican | Place, food, good, taco, service, great, time, service, ordered |
| Chinese | Food, place, good, menu, restaurant, order, noodle, great, time |
| Barbeque | Place, food, good, great, sauce, side, meat, service, bbq, burnt |
| Korean | Food, place, good, really, restaurant, meat, great, sushi, korean |
| Italian | Pizza, food, place, great, service, good, time, restaurant, table |
| Indian | Food, good, service, place, restaurant, great, naan, chicken |
| Steakhouse | Food, steak, great, good, place, service, ordered, time, drink |
| Seafood | Food, good, place, steak, great, sushi, ordered, service, roll |
| Thai | Food, place, chicken, pad thai, thai, good, restaurant, ordered |

The most popular cuisines were identified. Due to the overlap (ie Mexican, Restaurants and Restaurants, Mexican), restaurants were categorized by seeing if the category contained certain strings (ie “Mexican”). Then, word clouds were created for the top 50 reviewed restaurants for each cuisine in order to reduce processing time. Depending on how “cuisine” is defined, “Burgers” may or may not be included, so there are 12 word clouds below, the overall word cloud, and top 10 or 11 cuisines (depending on if burger is considered a cuisine).

**Most popular cuisines:**

A screenshot of a cell phone

Description automatically generated

**Word Clouds:**

A picture containing newspaper

Description automatically generated

A picture containing newspaper

Description automatically generated

A picture containing newspaper

Description automatically generated

A picture containing newspaper

Description automatically generated

A close up of a newspaper

Description automatically generated

A picture containing newspaper

Description automatically generated

A picture containing newspaper

Description automatically generated

A close up of a newspaper

Description automatically generatedA close up of a newspaper

Description automatically generated

A picture containing newspaper, green

Description automatically generated

A picture containing newspaper

Description automatically generated

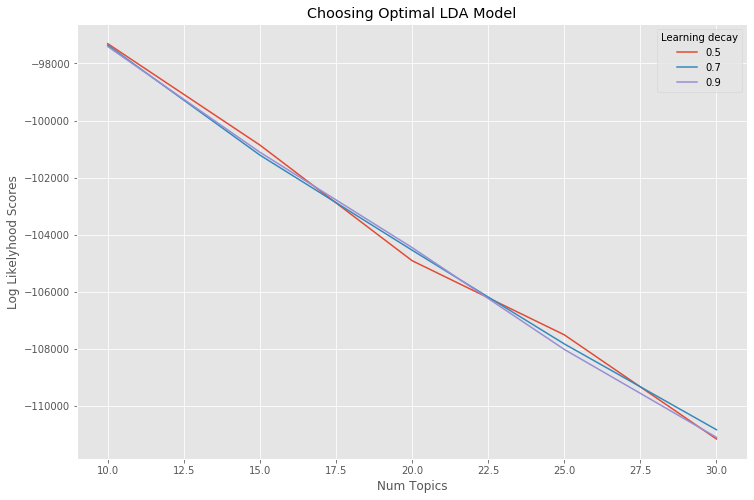
A close up of a newspaper

Description automatically generated

1. Topic modeling   
   Define a set of topics by applying topic modeling algorithms such as LDA on textual reviews. Choose an optimal number of topics in a data-driven fashion such as by using a figure that plots perplexity versus number of topics.
   1. What topic modelling techniques did you apply?  
      LDA (Latent Dirichlet Allocation) topic modeling was used. LDA is used for classifying text to certain topics
   2. How did you select the optimal number of topics? How many topics did you define?

The optimal number of topics (10) was chosen by plotting number of topics vs the log likelihood scores and choosing the number of topics that had the highest log likelihood score. This occurs with a learning decay of .5 and 10 topics, with a log likelihood of -97306. This can be seen in the line chart below.

**Optimal LDA Model:**

****

1. Sentiment & Correlation  
   Calculate sentiment score on each review to answer the question: how strong is the correlation between star rating and sentiment score?
   1. What steps did you take to calculate sentiment score?  
      To calculate sentiment score I used TextBlob’s sentiment to calculate the sentiment score polarity and sentiment score subjectivity for each review. For polarity, positive scores are positive sentiment and negative scores and negative sentiment. For subjectivity, scores closer to one are more subjective and scores closer to zero are more objective.
   2. What steps did you take to calculate the correlation?

To calculate the correlation, I calculated the covariance, Pearson’s R2, and the Spearman coefficient between sentiment score polarity and number of stars, and between sentiment score subjectivity and number of stars, and between sentiment score polarity and number of stars when broken down into negative (stars = 0, 1, or 2) and positive (stars = 3, 4, or 5) reviews. The highest correlation is the R2 between sentiment score polarity and number of stars. The distribution of number of stars can be seen in the bar chart below.

* 1. Is there a correlation between star rating and sentiment score? What’s your reasoning?

There is a moderate correlation between star rating and sentiment score. The highest correlation is the R2 between sentiment score polarity and number of stars, which is .5511. The higher rated reviews seem to have high sentiment polarity scores, this can be seen in the scatter plot below. More improvements can be made regarding text understanding, specifically regarding slang. An example where the current state does not perform well is a review that states “the ozu was mad dank” and leaves 5 stars, but TextBlob gives it sentiment score polarity of -.1186.

**Count of Stars:**

**A picture containing food, drawing, brick

Description automatically generated**

**Stars vs Sentiment Score: A picture containing table, white, group

Description automatically generated**