Topic Models for Yelp Reviews

Springboard Data Science Career Track
Capstone Project 2

Introduction

Goals of this Project

- Provide an efficient means of summarizing the content of reviews.
- Specifically, construct models that automatically generate dominant “topics” in reviews.

Motivation for the Project

- Yelp is useful for customers, but could be more useful for business owners.
- A proposal for how to add value for business owners: enable business owners to have greater awareness of customer experiences.

Data Acquisition and Wrangling

The Data

- Approximately 6.5 million reviews of a variety of different businesses
- Acquired directly from Yelp

Data Wrangling

1. Restricted data to reviews of *restaurants*
2. Eliminated rows with missing or duplicate values
3. Restricted data to reviews in the English language

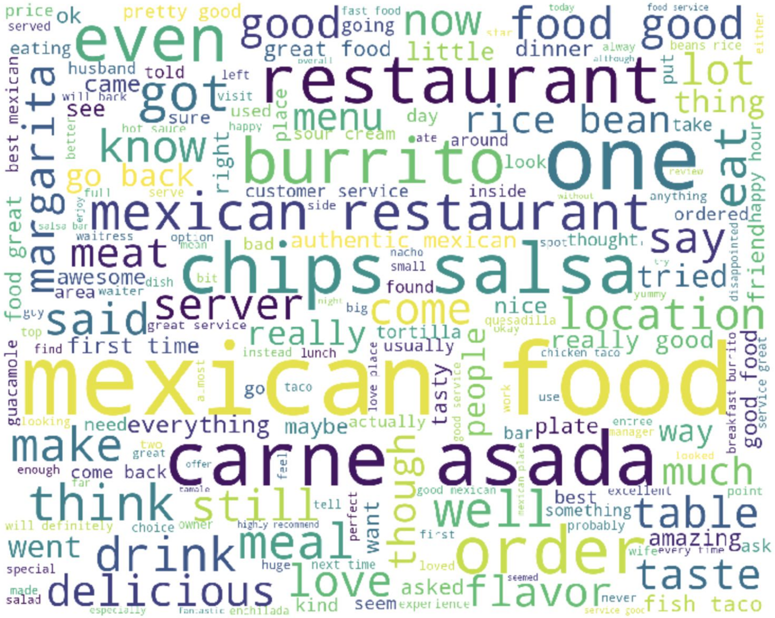
Storytelling and Inferential Statistics

Similarity of Word Clouds for All Reviews and for All Las Vegas Restaurants

Data Exploration

Word Cloud for Mexican Restaurants

All Reviews in the category [Mexican, Restaurants]



Constructing Artificial Topics

- Topics:
  - Positive sentiment: ‘good’, ‘delicious’, ‘yummy’, ‘tasty’, ‘superb’, ‘best’, ‘great’, ‘amazing’, ‘awesome’
  - Negative sentiment: ‘bad’, ‘disgusting’, ‘gross’, ‘nasty’, ‘terrible’, ‘worst’, ‘horrible’
- Results:

Star Rating	Proportions for Entire Dataset	Proportions for ‘Positive Sentiment’ Dataset	Proportions for ‘Negative Sentiment’ Dataset
5.0	39.3%	43.3%	12%
4.0	26.1%	29.4%	15.6%
3.0	13.3%	13.7%	18.4%
2.0	9.4%	7.8%	19.1%
1.0	11.9%	5.9%	34.9%

Statistical Analysis

- Question: Are the differences between mean star rating of Sushi Bars and mean star rating of Mexican restaurants statistically significant?
- T-test of the difference in mean star rating
- Null Hypothesis: The mean star rating is the same for the two categories.
- Alternative Hypothesis: The mean star rating is not the same for the two categories.
- Alpha = 0.05
- Results: | t-score: 13.33 | p-value ≅ 0 |
- Conclusion: The difference in mean star rating is statistically significant.

Baseline Modeling

Text Preprocessing, Vector Representation, Baseline Models

- **Text Preprocessing**
  - Remove punctuation
  - Make all text lowercase
  - Remove stopwords
  - Lemmatize the word tokens
- **Vector Representation**
  - Represented with a bag of words frequency vectorization (Gensim implementation)
- **Baseline Models**
  - Four models using samples of size 1_000, 10_000, 100_000, and 500_000

Extended Modeling

Methods of Evaluation

- Measuring the “Coherence” of the topics
  - Two Metrics Used:
    1. c_v (ranges from -1 to 1)
    2. U_mass (ranges from -14 to 14)
- Using my own judgment of the quality of the topics
  - Used scale from 1 to 5
  - A Primary Question: To what extent can an informative label be assigned to a topic that summarizes the mutual relevance of the heavily weighted words in that topic?

Model Construction

- <u>**Stage 1**</u>
- Constructed 12 Models
- Used samples of size 1_000, 10_000, and 100_000
- For each sample, constructed models with 5, 10, 25, and 50 topics
- Calculated coherence scores (c_v and u_mass) for each model

- <u>**Stage 2**</u>
- Constructed 12 more Models
- Used samples of size 1_000, 10_000, and 100_000
- Removed words appearing in more than 30% of documents or less than 5 times in corpus
- For each sample, constructed models with 2, 5, 7, and 10 topics
- Calculated coherence scores (c_v and u_mass) for each model, and used my own judgment to assess quality of 4 models

Summary of Findings

Model Performance

- Overall, model performance was unexceptional.

Evaluations for Filtered Models of Sample Size 100_000

	u_mass	c_v	my evals
Filtered Sample 100_000, 2 Topics	-1.38848	0.309944	2.5
Filtered Sample 100_000, 5 Topics	-1.55123	0.303365	2.2
Filtered Sample 100_000, 7 Topics	-1.52561	0.315137	2
Filtered Sample 100_000, 10 Topics	-1.56588	0.318677	2.3

- ('u_mass' ranges from -14 to 14, 'c_v' ranges from -1 to 1, and 'my evals' ranges from 1 to 5)

Conclusions and Future Work

Future Directions

- Gaining a deeper understanding of content in the dataset to identify kinds of topics to expect
- Using n-grams (for different values of n) as the features in the vector representation of the text
- Varying other parameters in Gensim LDA model implementation (e.g. ‘alpha’, ‘eta’, ‘gamma_threshold’, etc.)
- Filtering more or less words from the dataset prior to model construction
- Using other coherence metrics available in Gensim’s implementation (e.g. c_uci, c_npmi, etc.)
- Exploring other methods for evaluating topic models (e.g. perplexity)
- Exploring and gaining deeper understanding of pyLDAvis visualizations

Recommendations for Client

Recommendations

- Explore some of the future directions mentioned previously to improve (1) the methods for evaluating model performance, and (2) improving the model performance itself.
- First recommended step: varying some of the other parameters in Gensim's implementation of LDA.