Topic Modeling Yelp Reviews By Sentiment

Capstone Project 2 • Miguel Montano

Problem Statement

The Problem

Thanks to resources like Yelp, consumer perspectives of individual businesses are currently plentiful, but finding actionable insight remains the proverbial needle in the haystack
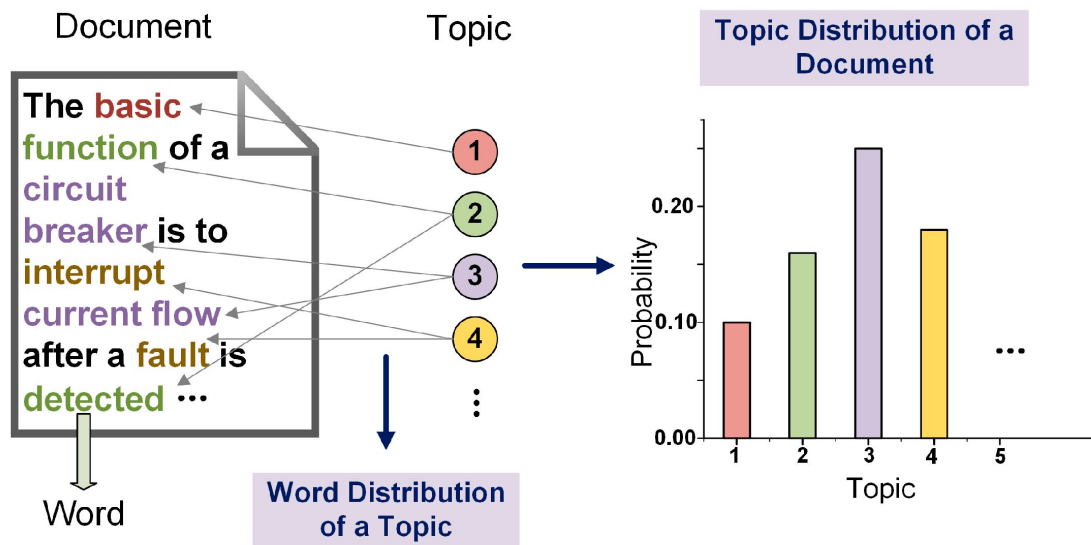
Potential Clients

- Individual Businesses being reviewed on Yelp

- Analysts seeking sentiment specific topics in a sector

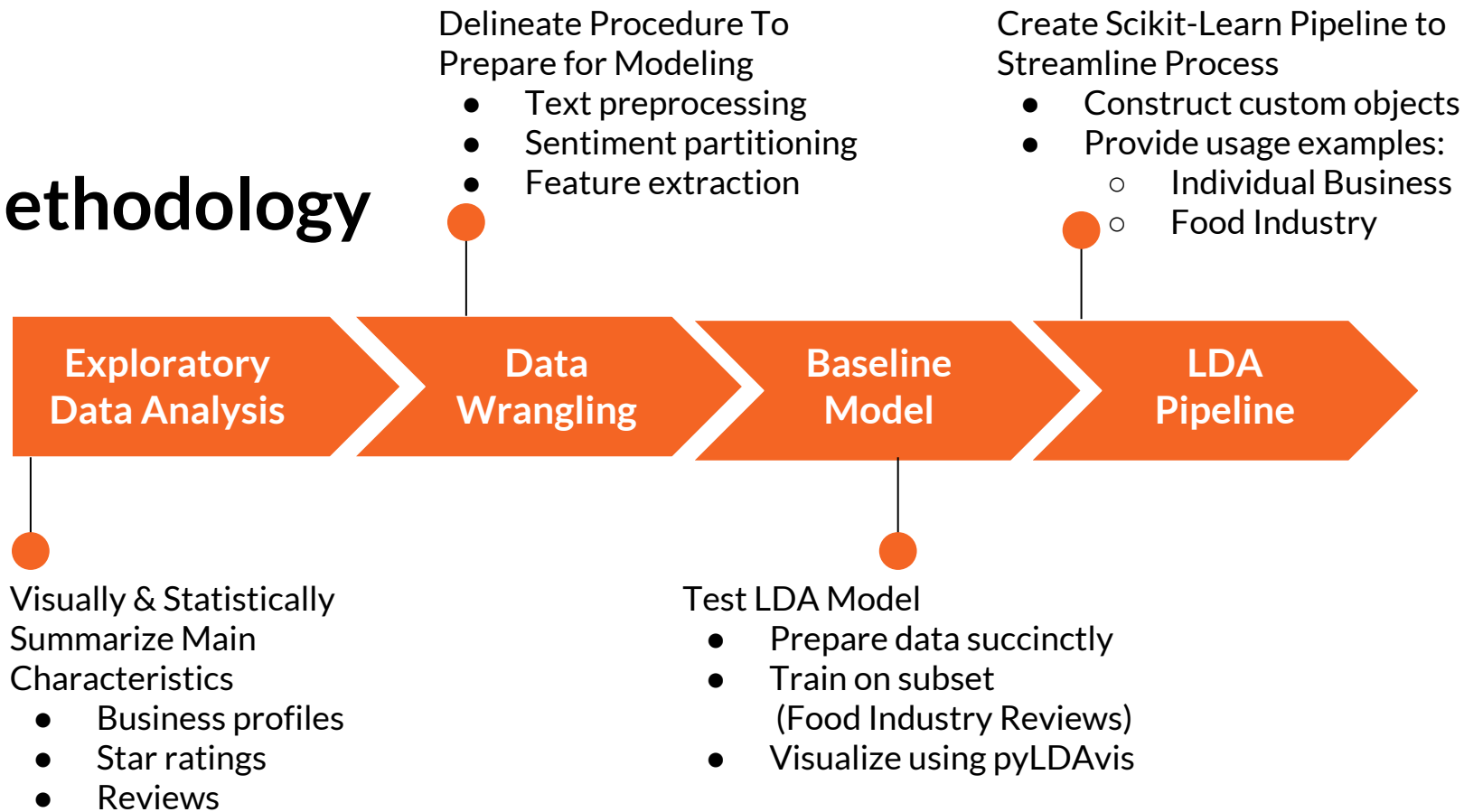- Special interest groups dedicated to local industries

A Solution

I propose an approach that utilizes the existing rating system in Yelp reviews to infer sentiment, attributing low and high ratings to negative and positive sentiment respectively, and creating a Scikit-Learn pipeline that trains a Latent Dirichlet Allocation Topic Model for each set, extracting topics that are being discussed when users write reviews with strong opinions

Background - Latent Dirichlet Allocation (LDA)

Document

The **basic** **function** of a **circuit** **breaker** is to **interrupt** **current flow** after a **fault** is **detected** ...

Word

Topic

1

2

3

4

Word Distribution of a Topic

Topic Distribution of a Document

Probability

0.20

0.10

0.00

1 2 3 4 5

Topic

...

- 'generative probabilistic model'

- Sets of observations can be described by unobserved groups, 'Topics', that explain similarities between parts of the data

- If observations are words collected into documents, each document is a mix of some topics and each word's presence is attributable to a topic

- Works as a way of 'soft clustering' the documents, as each can belong to a combination of topics, which is key in this case as reviewers may write about more than one subject

Methodology

Delineate Procedure To Prepare for Modeling
- Text preprocessing
- Sentiment partitioning
- Feature extraction

Create Scikit-Learn Pipeline to Streamline Process
- Construct custom objects
- Provide usage examples:
  - Individual Business
  - Food Industry

Exploratory Data Analysis → **Data Wrangling** → **Baseline Model** → **LDA Pipeline**

Visually & Statistically Summarize Main Characteristics
- Business profiles
- Star ratings
- Reviews

Test LDA Model
- Prepare data succinctly
- Train on subset (Food Industry Reviews)
- Visualize using pyLDAvis

Exploratory Data Analysis

Datasets were originally in a JSON format,
and each was imported into Pandas DataFrames

Links to Data:
- https://www.kaggle.com/yelp-dataset/yelp-dataset
- https://www.yelp.com/developers/documentation/v3/all_category_list
- https://gist.github.com/pbojinov/a87adf559d2f7e81d86ae67e7bd883c7

- **Business Profiles**
  - **Overview**
  - **Deep Dive into Top States**
    - Location of Businesses
    - Business Count
    - Average Star Rating
    - Distribution of Ratings
    - Average Number of Reviews Per Star Rating
- **Categories**
- **Food Industry Subset**
  - **Business Profiles**
    - Average Star Rating Per Year
  - **Reviews**
    - Average Length of Reviews Per Star Rating
    - Distribution of Review Lengths By Year

Business Profiles Overview

```
BUSINESSES OVERVIEW

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 188593 entries, 0 to 188592
Data columns (total 16 columns):
address          188593 non-null object
attributes       162807 non-null object
business_id      188593 non-null object
categories       188052 non-null object
city             188593 non-null object
hours            143791 non-null object
is_open          188593 non-null int64
latitude         188587 non-null float64
longitude        188587 non-null float64
name             188593 non-null object
neighborhood     188593 non-null object
postal_code      188593 non-null object
review_count     188593 non-null int64
stars            188593 non-null float64
state            188593 non-null object
state_name       188003 non-null object
dtypes: float64(3), int64(2), object(11)
memory usage: 23.0+ MB
```

```
Descriptive Statistics:
              is_open          latitude         longitude
count   188593.000000   188587.000000    188587.000000
mean         0.830391       38.506793       -97.490873
std          0.375290        5.122684        17.693360
min          0.000000      -71.753941      -180.000000
25%          1.000000       33.630878      -112.279276
50%          1.000000       36.143595      -111.777460
75%          1.000000       43.593106       -79.982958
max          1.000000       85.051129       115.086769

                 stars   review_count
count   188593.000000  188593.000000
mean         3.631550      31.797310
std          1.016783     104.124212
min          1.000000       3.000000
25%          3.000000       4.000000
50%          3.500000       9.000000
75%          4.500000      24.000000
max          5.000000    7968.000000
```
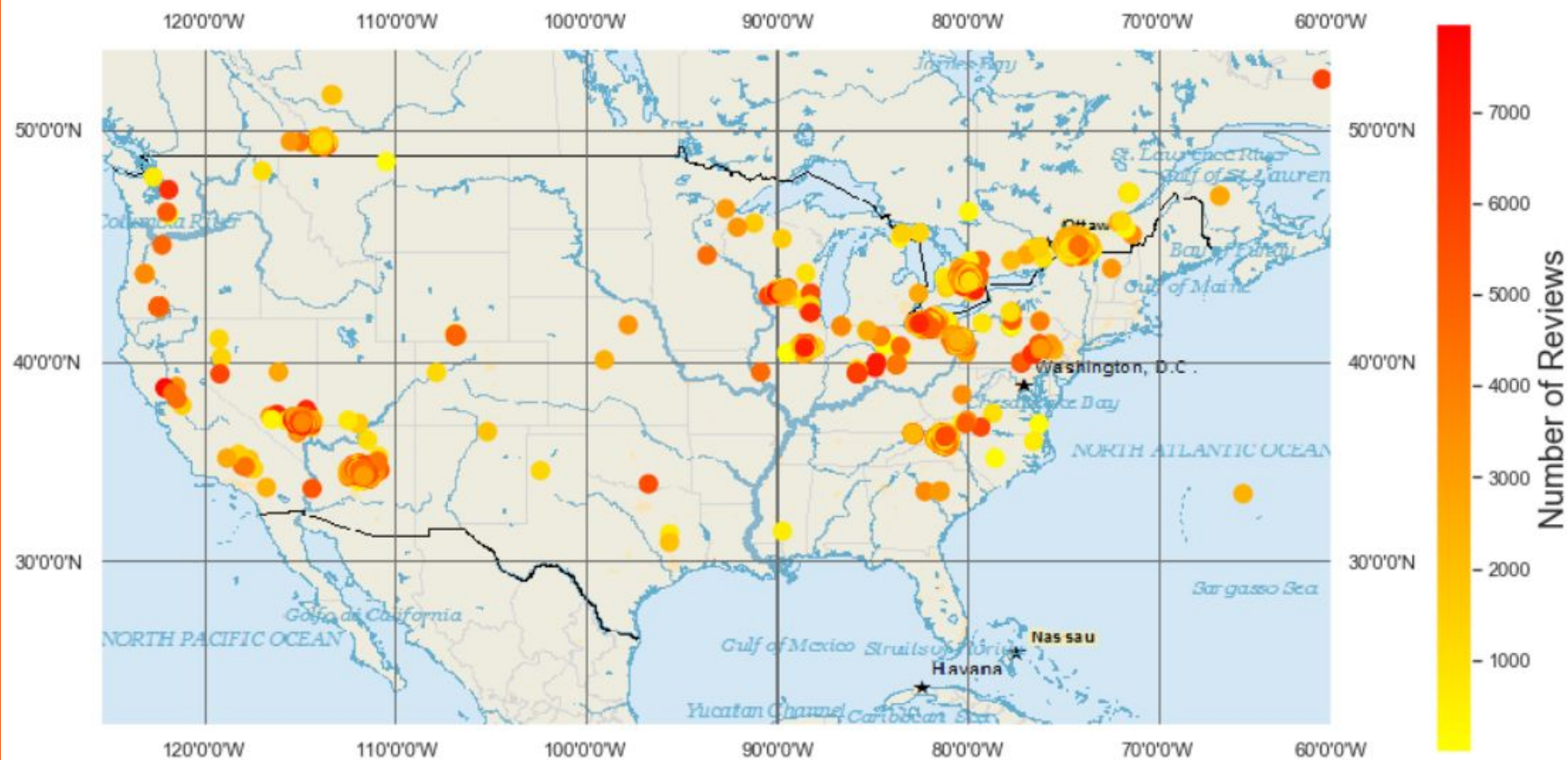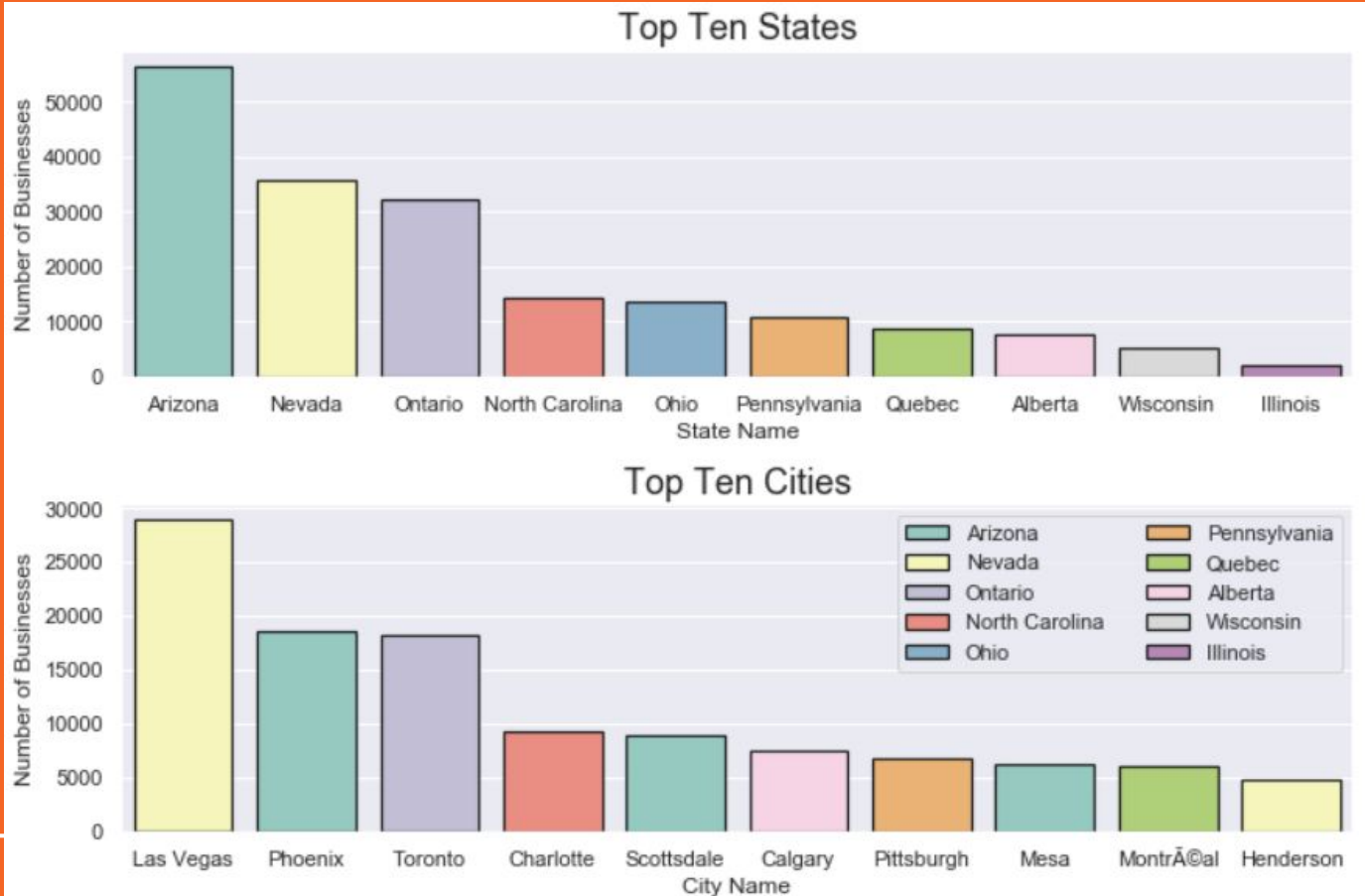
Highlights:
- 5,996,750 reviews are accounted for in Business DataFrame (99.996% of All Reviews)
- 99.14% of All Businesses are in the Top 10 States
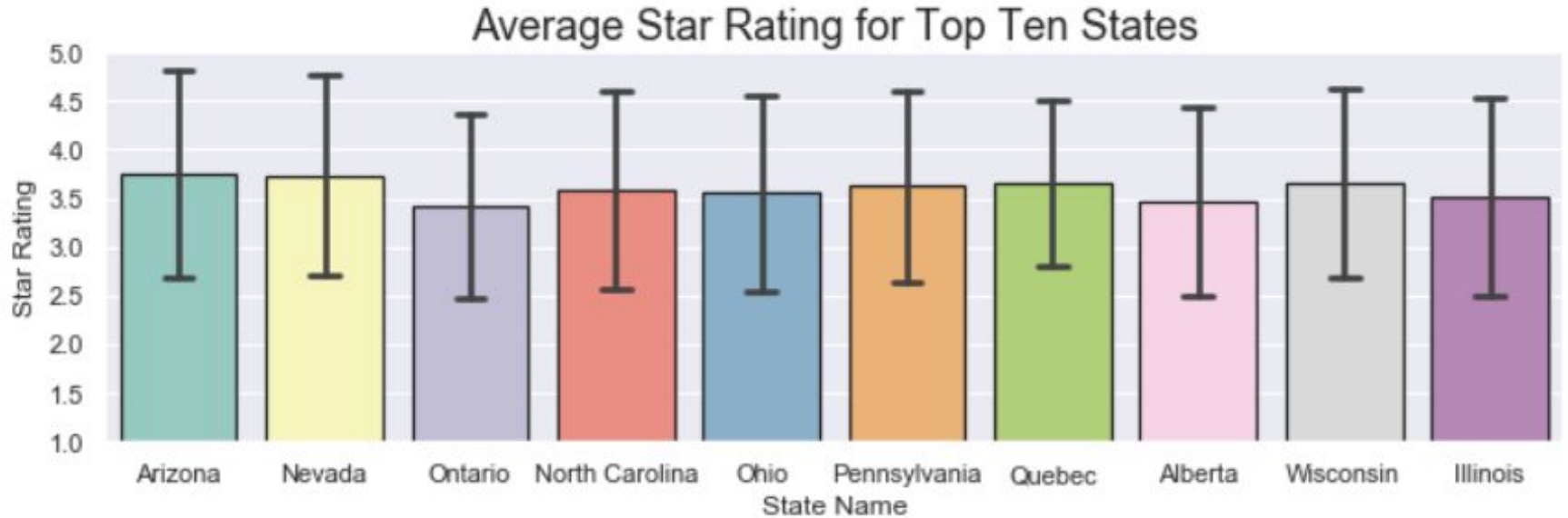
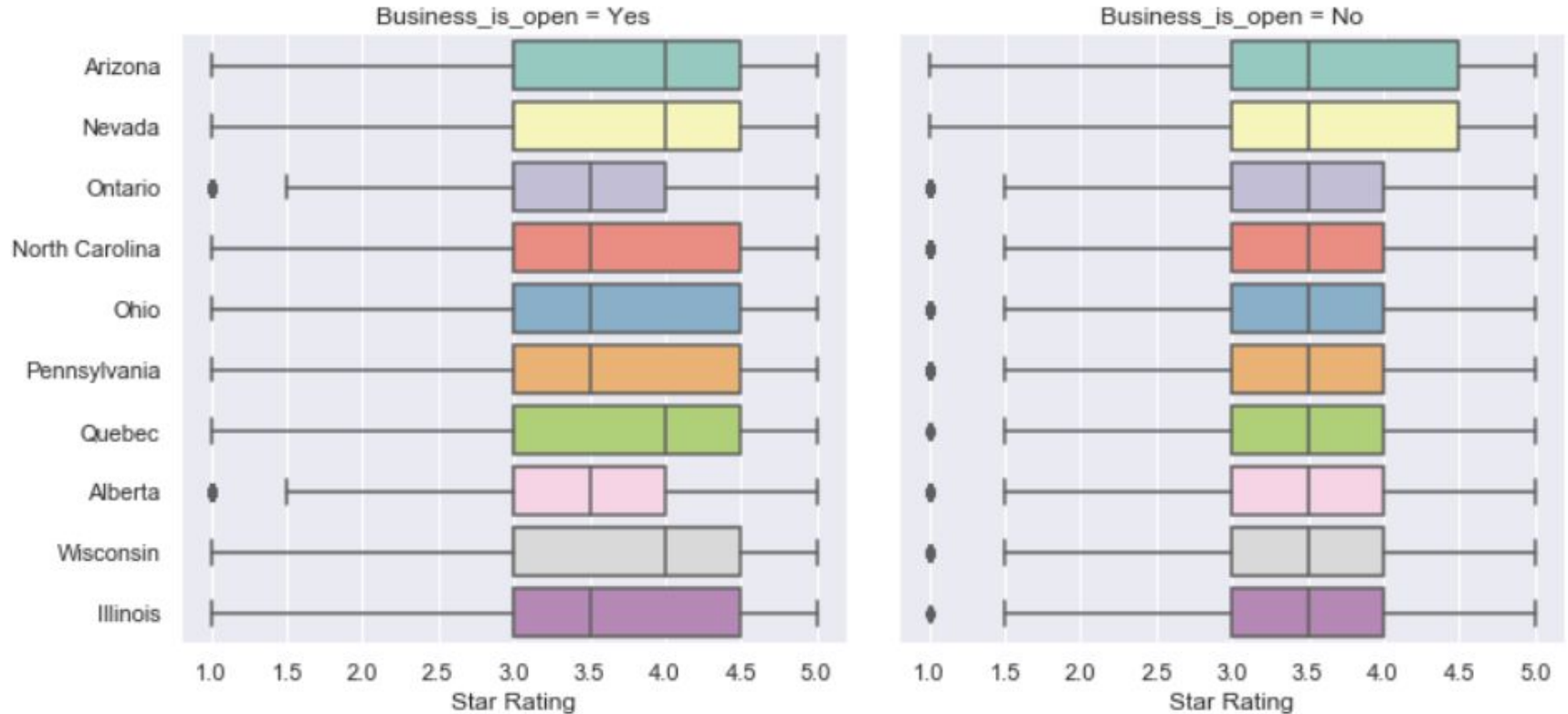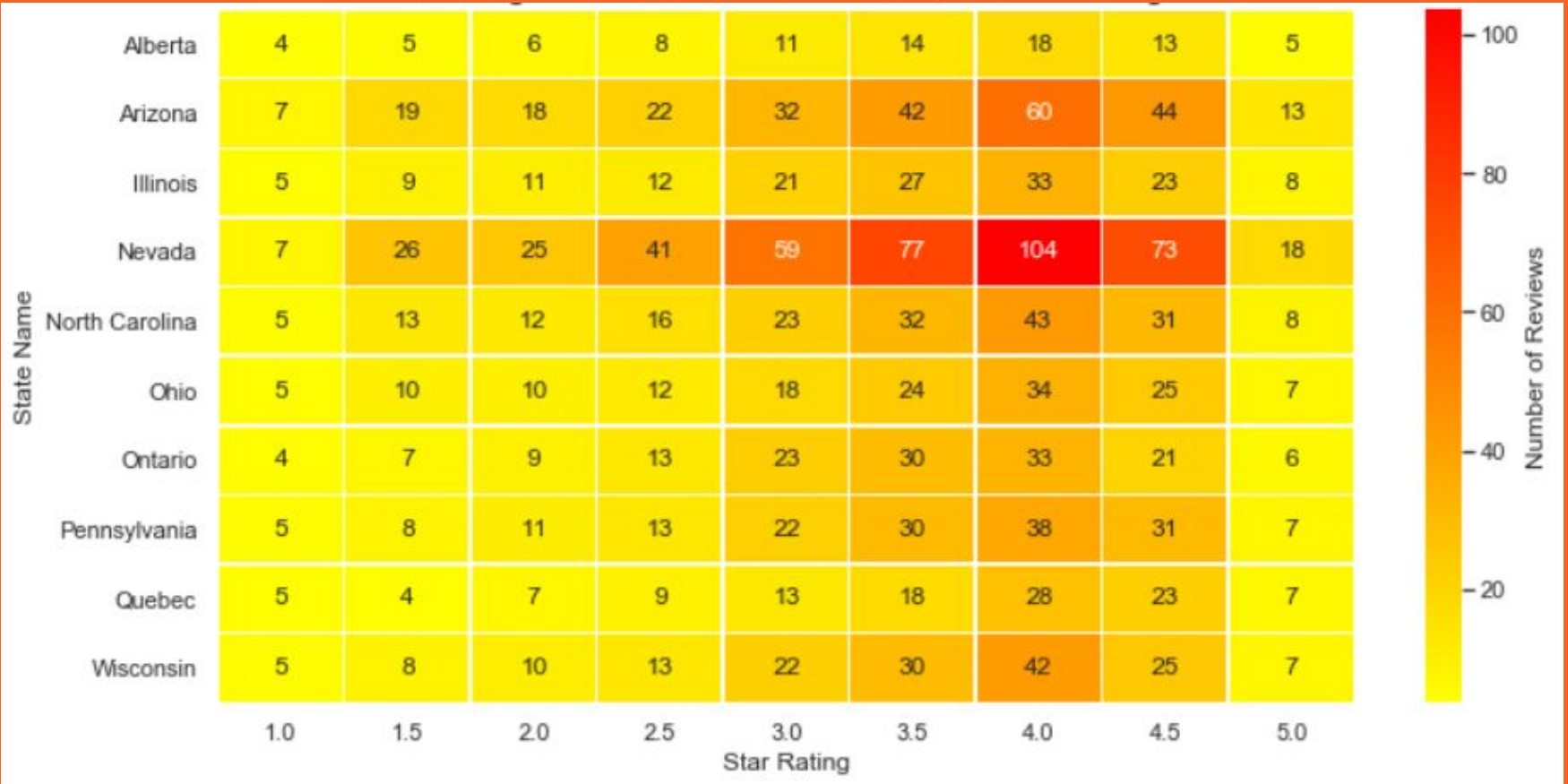# Location of Businesses in the Top States

Top Locations by Business Count

Top Ten States

Top Ten Cities

Average Star Rating By State

Average Star Rating for Top Ten States

Distribution of Star Ratings for Top Ten States

Average Number of Reviews Per Star Rating For Top States

State Name	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
Alberta	4	5	6	8	11	14	18	13	5
Arizona	7	19	18	22	32	42	60	44	13
Illinois	5	9	11	12	21	27	33	23	8
Nevada	7	26	25	41	59	77	104	73	18
North Carolina	5	13	12	16	23	32	43	31	8
Ohio	5	10	10	12	18	24	34	25	7
Ontario	4	7	9	13	23	30	33	21	6
Pennsylvania	5	8	11	13	22	30	38	31	7
Quebec	5	4	7	9	13	18	28	23	7
Wisconsin	5	8	10	13	22	30	42	25	7

Star Rating

Categories Overview & Top Ten Food Tags

```
CATEGORIES OVERVIEW

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1539 entries, 0 to 1538
Data columns (total 5 columns):
alias                1539 non-null object
country_blacklist     349 non-null object
country_whitelist     518 non-null object
parents              1539 non-null object
title                1539 non-null object
dtypes: object(5)
memory usage: 60.2+ KB
```

```
Top 10 Food Tags:
1. Restaurants
2. Food
3. Nightlife
4. Bars
5. Coffee & Tea
6. Sandwiches
7. Fast Food
8. American (Traditional)
9. Pizza
10. Burgers
```

Highlights:

- There are 118 unique parent tags, and categories can possess multiple parents
- Of the Top 10 Food tags, numbers one and two have a higher frequency in the overall Yelp dataset than the next eight combined, which might be because they are typically used in combination with other more specific tags

Food Industry Business Profiles Overview

```
FOOD INDUSTRY OVERVIEW

<class 'pandas.core.frame.DataFrame'>
Int64Index: 73536 entries, 0 to 188590
Data columns (total 17 columns):
address          73536 non-null object
attributes       71567 non-null object
business_id      73536 non-null object
categories       73536 non-null object
city             73536 non-null object
hours            55299 non-null object
is_open          73536 non-null int64
name             73536 non-null object
neighborhood     73536 non-null object
postal_code      73536 non-null object
review_count     73536 non-null int64
stars            73536 non-null float64
state            73536 non-null object
state_name       73144 non-null object
latitude         73535 non-null float64
longitude        73536 non-null float64
color            73536 non-null object
dtypes: float64(3), int64(2), object(12)
memory usage: 10.1+ MB
```

```
Descriptive Statistics:
            is_open     review_count          stars
count  73536.000000     73536.000000   73536.000000
mean       0.738047        55.236796       3.494764
std        0.439700       145.715852       0.823946
min        0.000000         3.000000       1.000000
25%        0.000000         6.000000       3.000000
50%        1.000000        16.000000       3.500000
75%        1.000000        49.000000       4.000000
max        1.000000      7968.000000       5.000000

           latitude        longitude
count  73535.000000     73536.000000
mean      40.020325       -92.323266
std        5.360234        18.120518
min      -71.753941      -123.587426
25%       35.233611      -112.073403
50%       41.152726       -81.439480
75%       43.695149       -79.431609
max       59.438181       115.086769
```
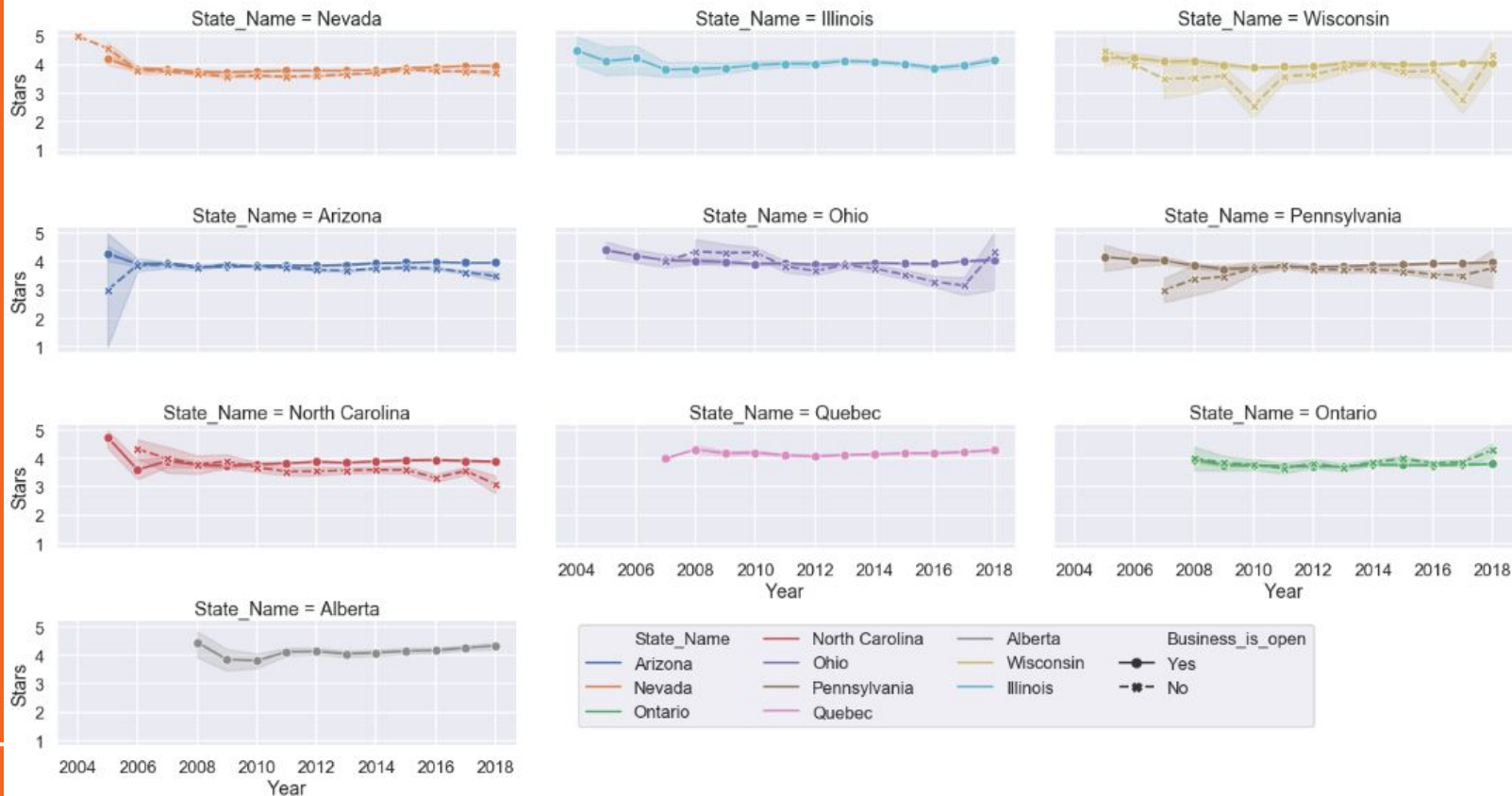
Highlights:
- 38.9% of all businesses profiles belong to the Food industry
- Food establishments have 20 more reviews per business than the population average (55.2 vs. 31.8)
- Closing rate is 9.2% higher for food businesses

Average Food Industry Star Rating Per Year for Top States

Average Star Rating Per Year

Food Industry Reviews

```
FOOD REVIEWS OVERVIEW

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1905325 entries, 0 to 1905324
Data columns (total 9 columns):
business_id      object
cool             int64
date             datetime64[ns]
funny            int64
review_id        object
stars            int64
text             object
useful           int64
user_id          object
dtypes: datetime64[ns](1), int64(4), object(4)
memory usage: 130.8+ MB
```

Highlights:
- 31.7% of all reviews in Yelp academic dataset are for businesses in the food industry
- A majority of reviews do not garner any additional 'cool', 'funny', or 'useful' tags
- If they do have such a tag, it is likely that they are reviews giving favorable ratings of 4-5 stars (70% of all reviews and 68% of reviews with additional tags)

stars	cool mean	cool std	cool min	cool max	cool count	funny mean	funny std	funny min	funny max	funny count	useful mean	useful std	useful min	useful max	useful count
1	0.307	2.144	0	505	162220	0.727	3.710	0	637	162220	1.610	8.598	0	1118	162220
2	0.428	1.872	0	172	157849	0.636	2.178	0	274	157849	1.398	5.250	0	1234	157849
3	0.634	2.603	0	245	241866	0.579	2.849	0	435	241866	1.201	3.741	0	805	241866
4	0.838	3.130	0	229	512665	0.556	3.094	0	566	512665	1.202	3.509	0	245	512665
5	0.608	2.322	-1	208	830725	0.371	3.240	0	991	830725	0.910	2.707	-1	215	830725

Average Length of Food Industry Reviews Per Star Rating

Average Length of Reviews Per Star Rating

State Name	1	2	3	4	5
Alberta	49	53	58	52	47
Arizona	49	53	49	43	33
Illinois	57	57	49	43	37
Nevada	49	54	53	48	37
North Carolina	50	54	52	45	36
Ohio	52	58	55	49	41
Ontario	56	59	58	53	44
Pennsylvania	50	58	56	51	42
Quebec	53	52	48	45	42
Wisconsin	55	57	55	48	41

Star Rating

Number of Tokens Per Review

Distribution of Review Lengths By Year for the Food Industry

Distribution of Review Lengths for the Food Industry
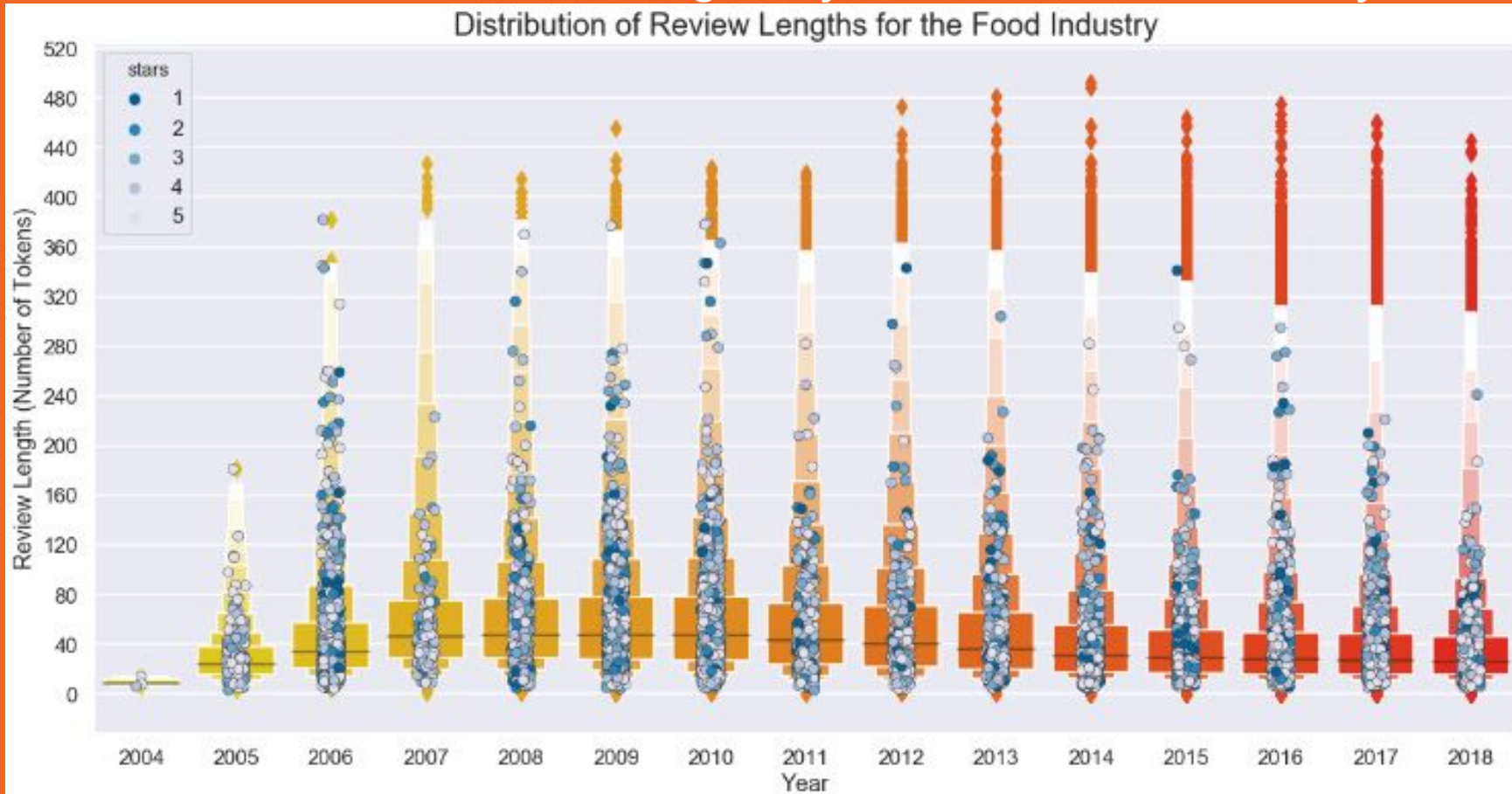
Data Wrangling

- With Gensim, NLTK, and Pandas libraries

- Using Food Industry reviews subset

- Define steps to be implemented in future pipeline

- **Text Preprocessing**
  - **Tokenization**
  - **Normalization**
  - **Stopword Removal**
  - **Lemmatization**
- **Sentiment Partitioning**
  - **'negative' or 'positive'**
  - **Based on star rating**
- **Feature Extraction**
  - **Map words to unique integer ids**
    - Collect word frequency
  - **Create vector space corpora**
    - Each review becomes a 'bag-of-words'
    - Two sets of document-term matrices, one for 'negative', and one for 'positive' reviews
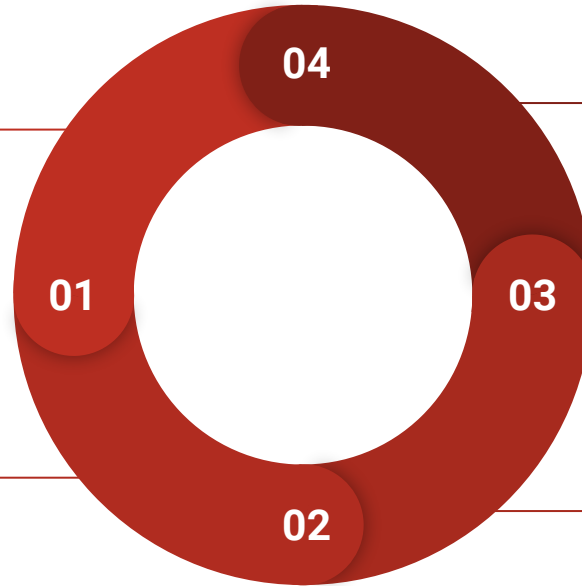
Text Preprocessing

Tokenization

Remove punctuation/special characters and transform passages from one long string into lists of word strings

Normalization

Convert all text to lower-case, expand contractions, remove numerals and accent marks

04

01

03

02

Lemmatization

Eliminate affixes from a word by capturing the canonical forms based on a word's lemma, or chosen representative, in this case assigning a Verb category tag to the tokenized parts of a sentence

Stopwords Removal

Remove words below three characters, and those which contribute little to overall meaning

Sentiment Partitioning

Negative Reviews

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 319968 entries, 13 to 1904179
Data columns (total 14 columns):
business_id          319968 non-null object
cool                 319968 non-null int64
date                 319968 non-null datetime64[ns]
funny                319968 non-null int64
review_id            319968 non-null object
stars                319968 non-null int64
text                 319968 non-null object
useful               319968 non-null int64
user_id              319968 non-null object
year                 319968 non-null int64
State_Name           319968 non-null object
Business_is_open     319968 non-null object
tokens               319968 non-null object
review_length        319968 non-null int64
dtypes: datetime64[ns](1), int64(6), object(7)
memory usage: 36.6+ MB
```

Positive Reviews

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1342441 entries, 0 to 1904177
Data columns (total 14 columns):
business_id          1342441 non-null object
cool                 1342441 non-null int64
date                 1342441 non-null datetime64[ns]
funny                1342441 non-null int64
review_id            1342441 non-null object
stars                1342441 non-null int64
text                 1342441 non-null object
useful               1342441 non-null int64
user_id              1342441 non-null object
year                 1342441 non-null int64
State_Name           1342441 non-null object
Business_is_open     1342441 non-null object
tokens               1342441 non-null object
review_length        1342441 non-null int64
dtypes: datetime64[ns](1), int64(6), object(7)
memory usage: 153.6+ MB
```

Post Text Processing

Negative Food Industry Reviews

```
1904080    Decor is nice but slow service, pastries were not good at all, tiramisu just ok but too thick. C...
1904091    Came in and order at 3.16 pm waited for 30 mins. Still haven't gotten our order. Hopefully next ...
1904147              Good atmosphere and location, but the taste of the coffee and deserts are horrendous.
1904178    We traveled 30 minutes to this spot since we been here once before and wanted to come back to tr...
1904179    OVERHYPED and OVERRATED.\n\nYes it's aesthetically pleasing to the eyes. Nice greenhouse in the ...
Name: text, dtype: object

1904080    [decor, nice, slow, service, pastries, good, tiramisu, litchi, rise, pastrie, flavor, come, choo...
1904091              [come, order, wait, mins, haven, get, order, hopefully, time, come, bavk, better, service]
1904147                         [good, atmosphere, location, taste, coffee, desert, horrendous]
1904178    [travel, minutes, spot, want, come, different, things, yelp, google, show, close, arrive, exactl...
1904179    [overhyped, overrate, aesthetically, please, eye, nice, greenhouse, middle, beautiful, victorian...
Name: tokens, dtype: object
```

Positive Food Industry Reviews

```
1904173    Friends night out and I chose Gabi Coffee and Bakery and everyone loved this place.  5 stars for...
1904174    This is a hidden gem in Las Vegas! The aesthetic and history surely creates an experience to the...
1904175    This is my favorite hang in Las Vegas!!! Because of the indoor atrium and korean decor it's like...
1904176    Little confusing to find cause of just having a big brown door and no sign. But when you walk in...
1904177    Their desserts are super cute and you can tell that the staff puts hard work into them, consider...
Name: text, dtype: object

1904173    [friends, night, choose, gabi, coffee, bakery, love, place, star, vibe, decor, ambiance, wish, h...
1904174    [hide, vegas, aesthetic, history, surely, create, experience, personal, favorite, latte, perfect...
1904175                         [favorite, hang, vegas, indoor, atrium, korean, decor, like, vegas, love]
1904176    [little, confuse, cause, have, brown, door, sign, walk, inside, like, different, little, world, ...
1904177    [desserts, super, cute, tell, staff, put, hard, work, consider, small, detail, cake, enjoy, expe...
Name: tokens, dtype: object
```

Feature Extraction

Gensim Dictionary

A mapping between words and their integer ids, created from the normalized tokens in each corpus (set of documents) by sweeping across them, assigning a unique integer id to each word, and then collecting word counts and other relevant statistics

Document-Term Matrix

Vector Space Corpus

Using a trained Dictionary, tokenized terms in reviews are converted to their integer ids, and paired with their Frequency or TF-IDF weighting, resulting in a sparse vector. This representation is known as a 'Bag-of-Words', as it disregards grammar and even word order but maintains multiplicity

LDA Model

Baseline Model

- **Using Gensim LDA model**
  - Trained on Food Industry reviews

- **One instance for each sentiment**
  - Ten topics per model

- **Unigrams or (bag-of-words) vectors**
  - Ignores context or phrasing
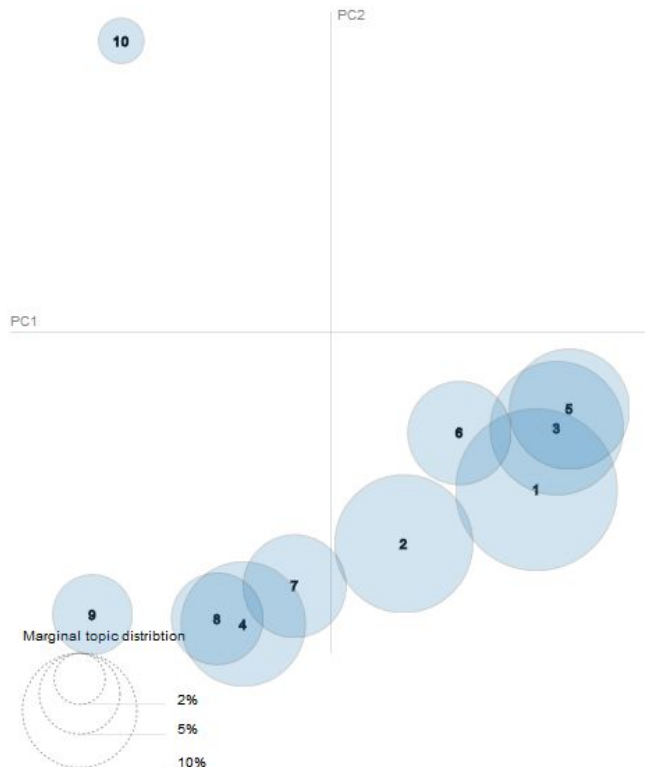
Baseline - Negative Review Topics

```
Topic: 0
Words: 0.045*"food" + 0.033*"place" + 0.024*"good" + 0.023*"service" + 0.018*"price"
Topic: 1
Words: 0.040*"order" + 0.039*"food" + 0.036*"come" + 0.031*"wait" + 0.024*"service"
Topic: 2
Words: 0.024*"order" + 0.021*"fry" + 0.018*"burger" + 0.017*"pizza" + 0.016*"salad"
Topic: 3
Words: 0.019*"hair" + 0.012*"movie" + 0.010*"italian" + 0.010*"asada" + 0.010*"indian"
Topic: 4
Words: 0.028*"soup" + 0.025*"dish" + 0.022*"chicken" + 0.021*"pork" + 0.018*"order"
Topic: 5
Words: 0.031*"table" + 0.017*"restaurant" + 0.016*"seat" + 0.012*"host" + 0.011*"party"
Topic: 6
Words: 0.020*"like" + 0.018*"food" + 0.016*"chip" + 0.016*"order" + 0.016*"tacos"
Topic: 7
Words: 0.024*"place" + 0.020*"like" + 0.014*"drink" + 0.011*"people" + 0.010*"look"
Topic: 8
Words: 0.026*"food" + 0.023*"sushi" + 0.020*"buffet" + 0.017*"roll" + 0.016*"like"
Topic: 9
Words: 0.024*"say" + 0.018*"tell" + 0.016*"order" + 0.014*"time" + 0.013*"ask"
```

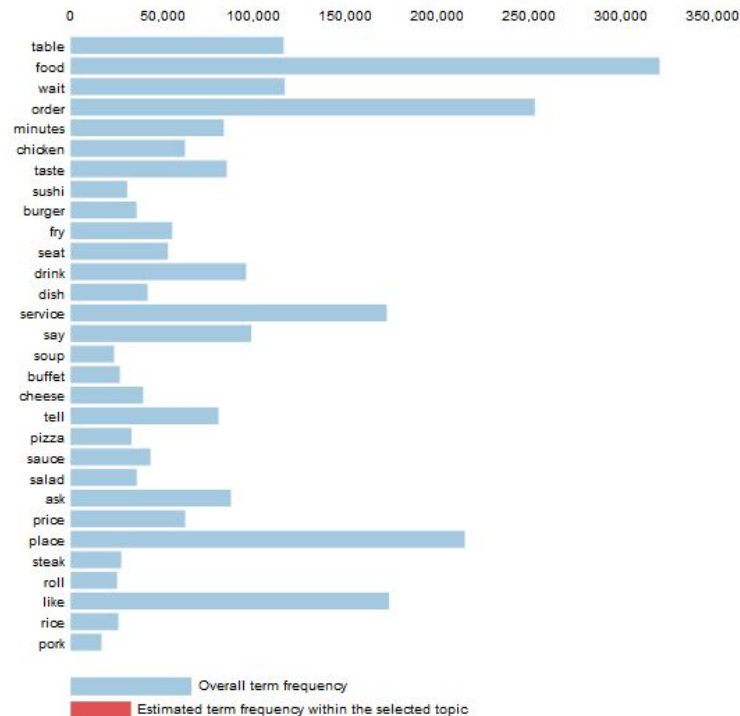# Baseline - Negative Review Topics

Intertopic Distance Map (via multidimensional scaling)

Top-30 Most Salient Terms₁

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t: see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w): see Sievert & Shirley (2014)

Baseline - Positive Review Topics

```
Topic: 0
Words: 0.025*"breakfast" + 0.019*"coffee" + 0.016*"cream" + 0.015*"buffet" + 0.013*"brunch"
Topic: 1
Words: 0.020*"place" + 0.016*"like" + 0.014*"drink" + 0.012*"good" + 0.008*"pretty"
Topic: 2
Words: 0.032*"fry" + 0.027*"good" + 0.025*"chicken" + 0.022*"burger" + 0.019*"order"
Topic: 3
Words: 0.026*"order" + 0.020*"food" + 0.020*"time" + 0.020*"come" + 0.015*"wait"
Topic: 4
Words: 0.045*"place" + 0.029*"food" + 0.028*"love" + 0.027*"best" + 0.020*"time"
Topic: 5
Words: 0.046*"tacos" + 0.027*"chip" + 0.027*"mexican" + 0.021*"salsa" + 0.020*"taco"
Topic: 6
Words: 0.070*"great" + 0.061*"food" + 0.042*"place" + 0.042*"service" + 0.040*"good"
Topic: 7
Words: 0.092*"pizza" + 0.024*"italian" + 0.021*"pasta" + 0.019*"crust" + 0.018*"sauce"
Topic: 8
Words: 0.018*"order" + 0.018*"good" + 0.018*"roll" + 0.016*"rice" + 0.016*"dish"
Topic: 9
Words: 0.013*"steak" + 0.012*"dish" + 0.011*"dinner" + 0.011*"dessert" + 0.009*"restaurant"
```

Baseline - Positive Review Topics

Intertopic Distance Map (via multidimensional scaling)

Top-30 Most Salient Terms[1]

Marginal topic distribtion
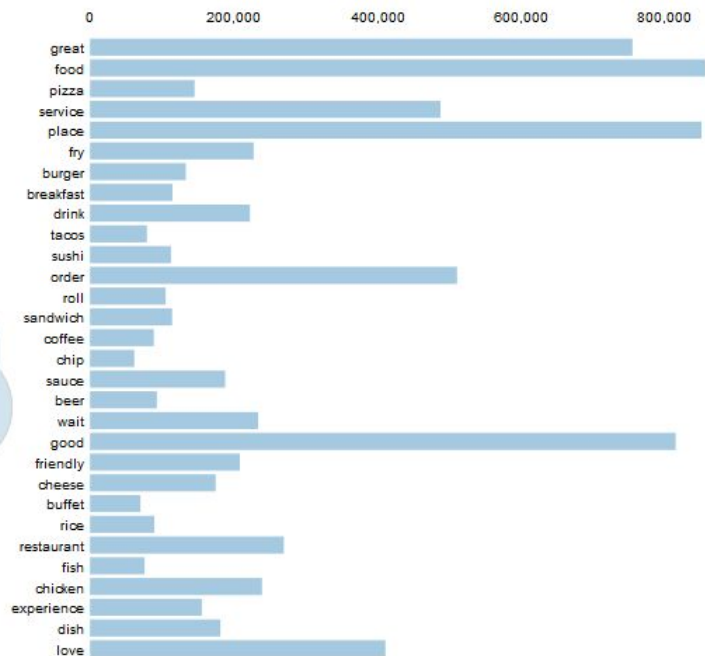
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t: see Chuang et. al (2012
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w): see Sievert & Shirley (2014)

LDA Pipeline

❖ Sequentially applies the transformations necessary for generating a topic model

❖ Custom objects were built by wrapping Gensim modules and Pandas functions, while inheriting Scikit-Learn BaseEstimator and TransformerMixin classes

- **Components**
  - **Transformers**
    - Yelp Review Selector
    - Preprocess
    - Vectorize
  - **Estimator**
    - LDA Estimator

- **Implementation**
  - **LDA Metrics**
    - Combination of graphical and statistical evaluations
  - **Individual Business Example**
  - **Food Industry Example**

Pipeline Components

Vectorize

Create a vector space of features, where each feature is a sparse vector extracted from each document, in this case preprocessed Yelp reviews, consisting of terms and term weights

LDA Estimator

Trains an instance of Gensim's Latent Dirichlet Allocation Model, with the option to create a pyLDAvis figure from the result

Preprocess

Preprocess text data, in this case the Yelp reviews themselves, per specifications; with the option to expand final vector space with n-grams

Yelp Review Selector

Selects Yelp reviews based on specified criteria, with the option of returning all review data in a pandas DataFrame, or simply the review text filtered by sentiment as a pandas Series

LDA Metrics

Statistical Measures

- Approximate how well the probability model predicts topic based on given samples:
  - Variational Bound
  - Log Perplexity

- Attempts to represent numerically how interpretable topics are to humans:
  - Coherence Model

Graphical Representations

- Topic Terms and Weights

- Most Representative Review Per Dominant Topic

- Distribution of Topic Contributions

- Dominant Topic Frequency

- Dominant Topic Distribution Among Documents

- pyLDAvis Visualization

LDA Pipeline - Individual Business Example

Construction

- A random business with at least 500 reviews was selected, 'Postino Arcadia' from Phoenix, Arizona

- Two separate pipelines were constructed, one for negative (3 or less star) reviews, and one for positive (4-5 star) reviews

- Using term frequency weighting, possessing trigrams, and removing the two most frequent tokens, as well as any tokens occurring in less than two reviews

- Ten passes were run over the training corpora, and four topics were drawn

Scores

```
Individual Business - Negative Reviews (Trigrams)

Variational Bound: -29457
Log Perplexity: -6.383
Coherence Score: 0.544
Coherence Per Topic
Topic 3: 0.587
Topic 2: 0.562
Topic 0: 0.536
Topic 1: 0.492

Individual Business - Positive Reviews (Trigrams)

Variational Bound: -41914
Log Perplexity: -9.082
Coherence Score: 0.536
Coherence Per Topic
Topic 2: 0.594
Topic 0: 0.543
Topic 1: 0.540
Topic 3: 0.466
```
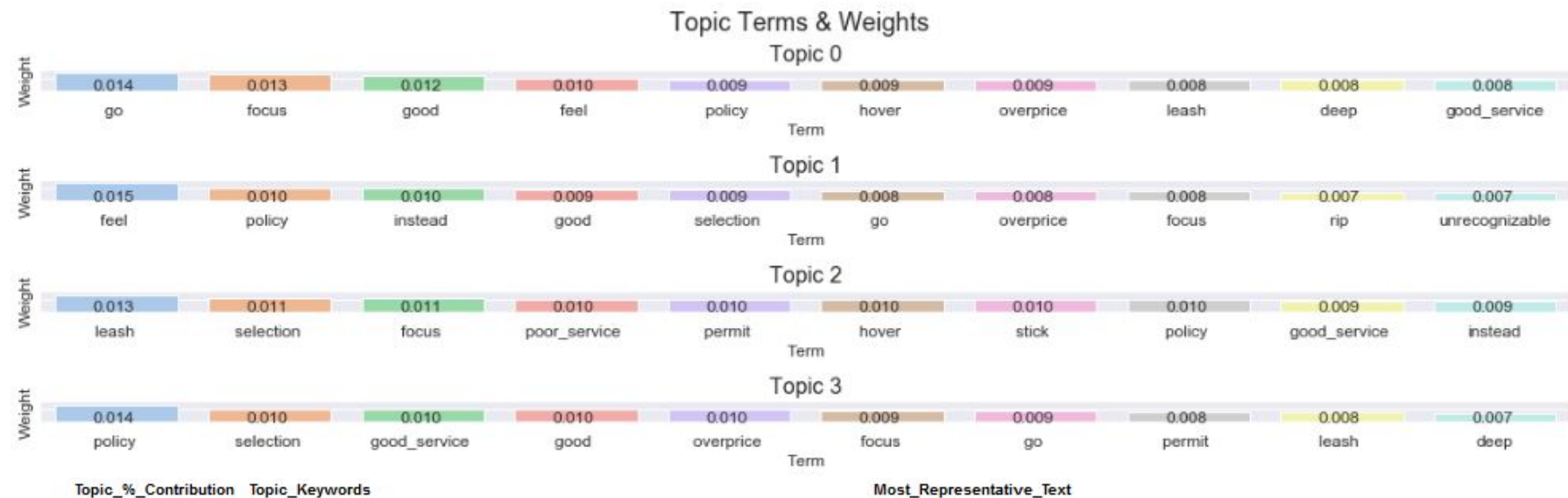
Individual Business Example - Negative Review Topics

Topic Terms & Weights

Topic 0

Weight									
0.014	0.013	0.012	0.010	0.009	0.009	0.009	0.008	0.008	0.008
go	focus	good	feel	policy	hover	overprice	leash	deep	good_service

Term

Topic 1

Weight									
0.015	0.010	0.010	0.009	0.009	0.008	0.008	0.008	0.007	0.007
feel	policy	instead	good	selection	go	overprice	focus	rip	unrecognizable

Term

Topic 2

Weight									
0.013	0.011	0.011	0.010	0.010	0.010	0.010	0.010	0.009	0.009
leash	selection	focus	poor_service	permit	hover	stick	policy	good_service	instead

Term

Topic 3

Weight									
0.014	0.010	0.010	0.010	0.010	0.009	0.009	0.008	0.008	0.007
policy	selection	good_service	good	overprice	focus	go	permit	leash	deep

Term

Topic	Topic_%_Contribution	Topic_Keywords	Most_Representative_Text
0.0	0.9990	go, focus, good, feel, policy, hover, overprice, leash, deep, good_service	This was my favorite restaurant. But not anymore. Today is my Birthday & thanks to them my day i...
1.0	0.9984	feel, policy, instead, good, selection, go, overprice, focus, rip, unrecognizable	We decided to visit this place while in the area based on all of the strong reviews from our fel...
2.0	0.9982	leash, selection, focus, poor_service, permit, hover, stick, policy, good_service, instead	Postino is normally one of our favorite local spots to grab a drink and some dinner. Last night ...
3.0	0.9991	policy, selection, good_service, good, overprice, focus, go, permit, leash, deep	I won't even venture to guess why so many people like this place. But a few notes: no, it is not...

Individual Business Example - Negative Review Topics

Distribution of Topic Contributions

Dominant Topic Frequency

Dominant Topic Distribution Among Documents

Topic 0.0 — 23.9%
Topic 1.0 — 26.6%
Topic 2.0 — 14.7%
Topic 3.0 — 34.9%

Individual Business Example - Negative Review Topics

Intertopic Distance Map (via multidimensional scaling)

Top-30 Most Salient Terms[1]

Marginal topic distribtion

2%
5%
10%

Overall term frequency
Estimated term frequency within the selected topic

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Individual Business Example - Positive Review Topics

Topic Terms & Weights

Topic 0

Weight									
0.007	0.005	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004
nice	thank	lounge	bruschetta_time	different_varieties	late	edna	service_atmosphere	reviewer	away_service

Term

Topic 1

Weight									
0.006	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.004	0.004
lounge	thank	edna	nice	hard_time	service_atmosphere	food_love	different_varieties	reviewer	bruschetta_time

Term

Topic 2

Weight									
0.006	0.005	0.005	0.004	0.004	0.004	0.004	0.004	0.004	0.004
different_varieties	reviewer	thank	service_atmosphere	edna	food_love	bruschetta_time	lounge	wide	delicious_bruschetta

Term

Topic 3

Weight									
0.006	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.004	0.004
food_love	edna	optimal	lounge	different_varieties	nice	service_atmosphere	reviewer	thank	give

Term

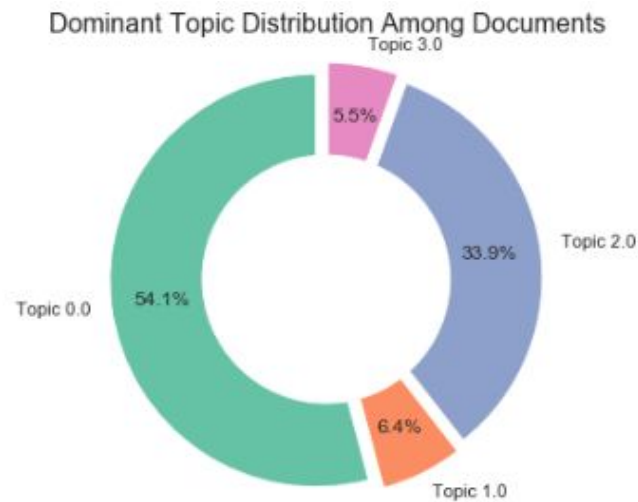| Topic	Topic_%_Contribution	Topic_Keywords	Most_Representative_Text
0.0	0.9959	nice, thank, lounge, bruschetta_time, different_varieties, late, edna, service_atmosphere, revie...	Visiting Postino Wine Cafe makes me feel like I am in another city... an urban, metro, pedestria...
1.0	0.9831	lounge, thank, edna, nice, hard_time, service_atmosphere, food_love, different_varieties, review...	Nothing worth writing about except the bruschetta. Most of the wine tastes like cheap box wine a...
2.0	0.9924	different_varieties, reviewer, thank, service_atmosphere, edna, food_love, bruschetta_time, loun...	I came here to meet a friend who I hadn't seen in ages. Some how the waitress assumed we just ne...
3.0	0.9911	food_love, edna, optimal, lounge, different_varieties, nice, service_atmosphere, reviewer, thank...	Food was ok bread was stale on my sandwich. But service was not good at all! God for bid you ask...

Individual Business Example - Positive Review Topics

Distribution of Topic Contributions

Dominant Topic Frequency

Dominant Topic Distribution Among Documents

Topic 3.0 — 5.5%
Topic 2.0 — 33.9%
Topic 0.0 — 54.1%
Topic 1.0 — 6.4%

Individual Business Example - Positive Review Topics

Intertopic Distance Map (via multidimensional scaling)

Top-30 Most Salient Terms[1]
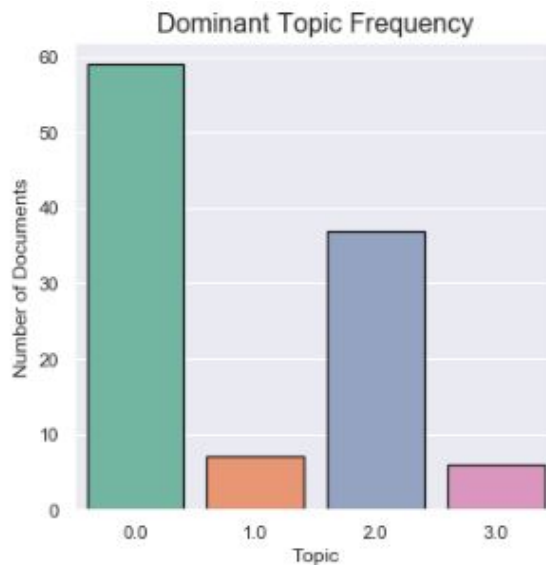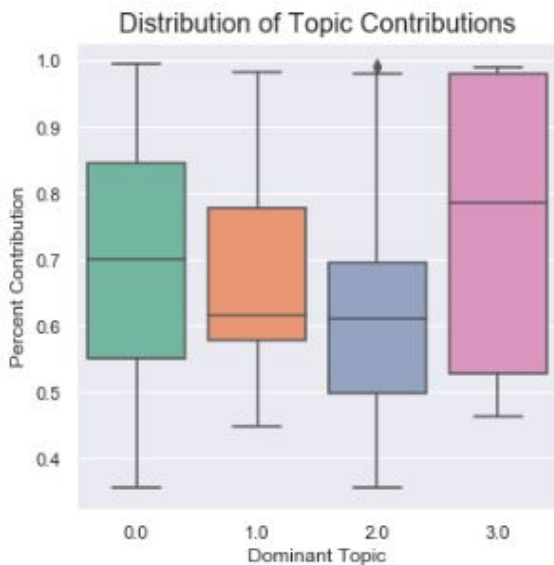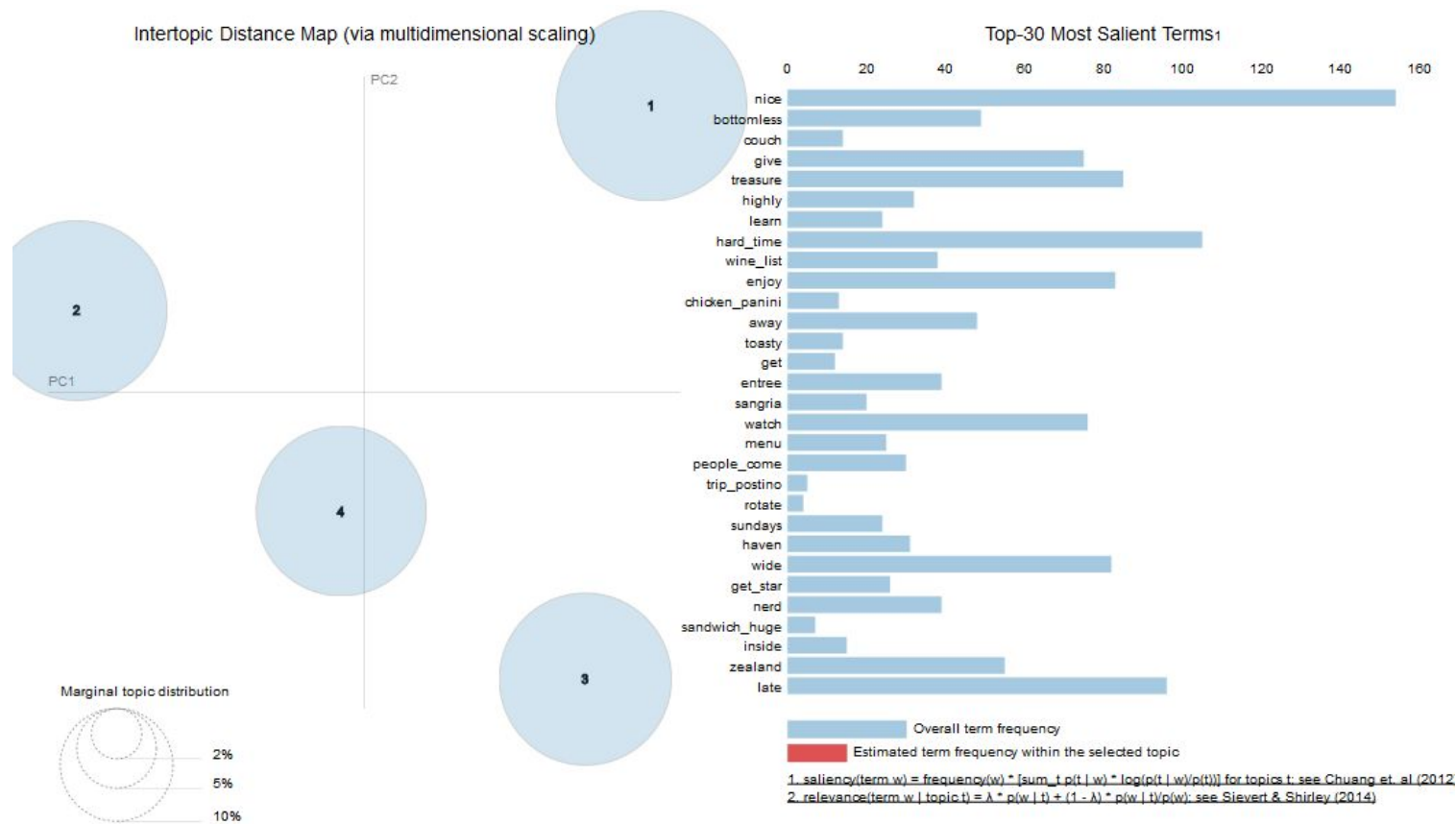
LDA Pipeline - Food Industry Example

Construction

Scores

- For a more global analysis using dataset with a wide variety of constituents, in this case taking another look at the Food Industry reviews used in the baseline model

- Two separate pipelines were constructed, one for negative (1-2 star) reviews, and one for positive (4-5 star) reviews

- Using term frequency weighting, possessing trigrams, and removing the three most frequent tokens, as well as any tokens occurring in less than ten reviews

- Larger sample size allows for splitting into train/test sets, as may be necessary in real-world scenarios, where trained models may need to be tested against new emerging data

```
Food Industry - Negative Reviews (Trigrams)

Variational Bound: -5452816
Log Perplexity: -11.207
Coherence Score: 0.542
Coherence Per Topic
Topic 6: 0.668 Topic 0: 0.544
Topic 3: 0.645 Topic 7: 0.474
Topic 9: 0.578 Topic 1: 0.471
Topic 2: 0.563 Topic 8: 0.470
Topic 4: 0.553 Topic 5: 0.454

Food Industry - Positive Reviews (Trigrams)

Variational Bound: -16619774
Log Perplexity: -11.544
Coherence Score: 0.525
Coherence Per Topic
Topic 5: 0.598 Topic 7: 0.521
Topic 4: 0.582 Topic 9: 0.503
Topic 0: 0.550 Topic 3: 0.476
Topic 1: 0.549 Topic 6: 0.476
Topic 2: 0.544 Topic 8: 0.456
```
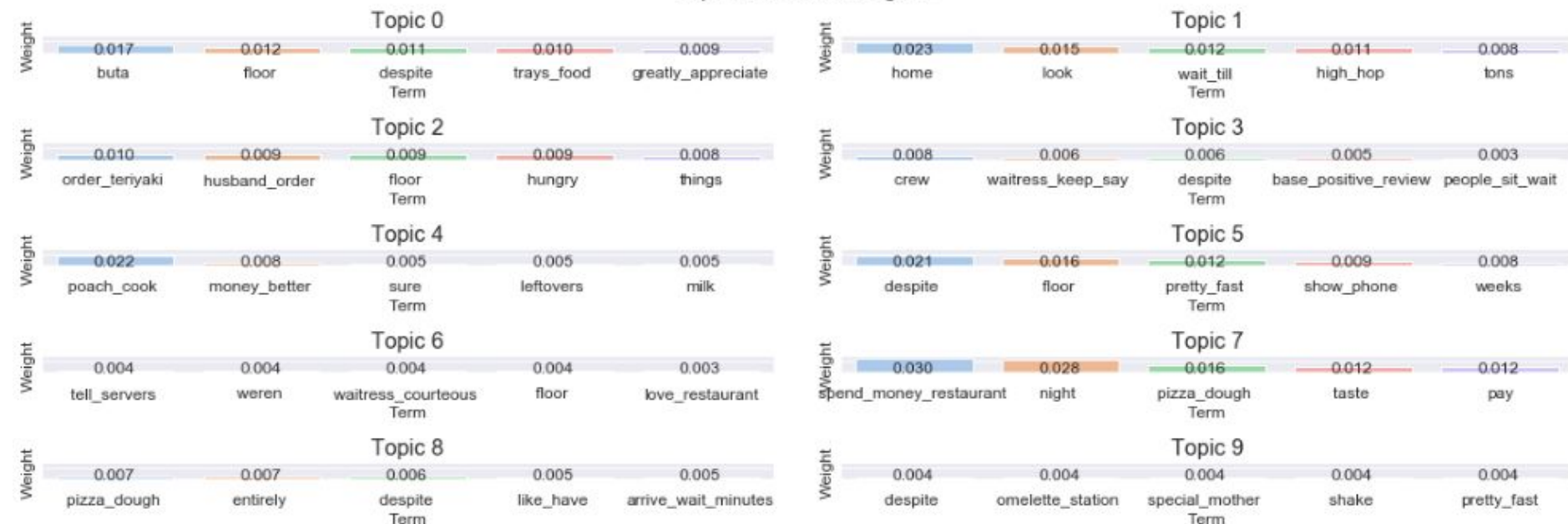
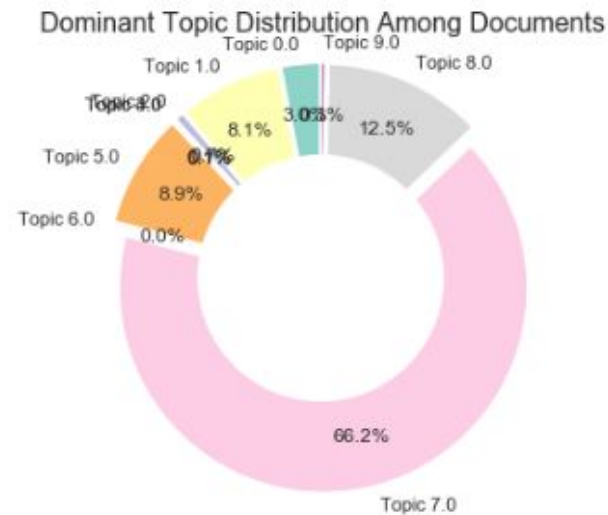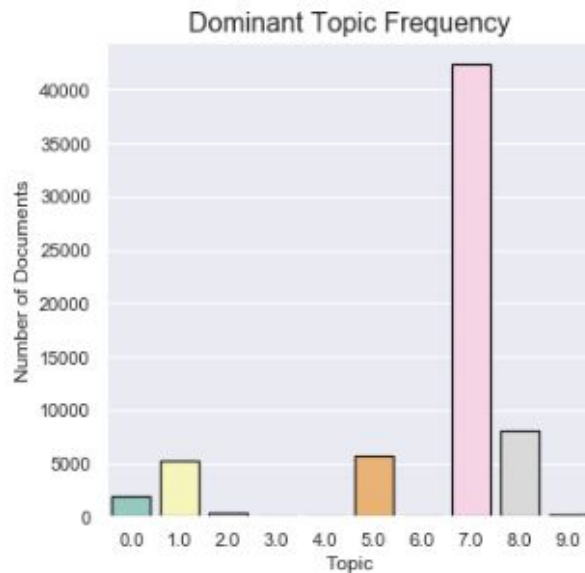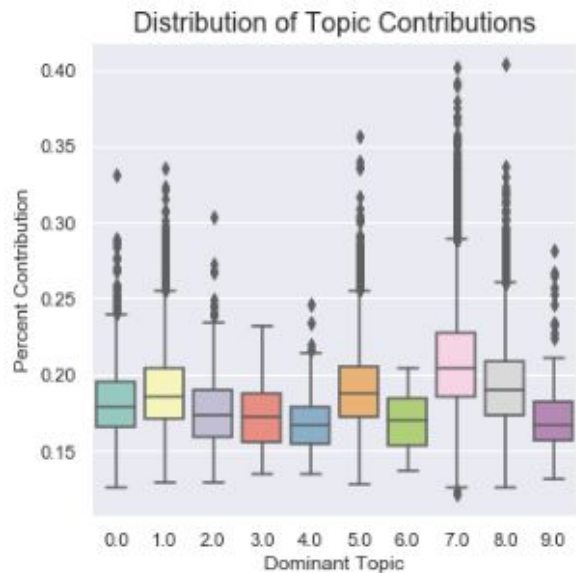# Food Industry Example - Negative Review Topics

Topic Terms & Weights

Topic 0

Weight				
0.017	0.012	0.011	0.010	0.009
buta	floor	despite	trays_food	greatly_appreciate

Term

Topic 1

Weight				
0.023	0.015	0.012	0.011	0.008
home	look	wait_till	high_hop	tons

Term

Topic 2

Weight				
0.010	0.009	0.009	0.009	0.008
order_teriyaki	husband_order	floor	hungry	things

Term

Topic 3

Weight				
0.008	0.006	0.006	0.005	0.003
crew	waitress_keep_say	despite	base_positive_review	people_sit_wait

Term

Topic 4

Weight				
0.022	0.008	0.005	0.005	0.005
poach_cook	money_better	sure	leftovers	milk

Term

Topic 5

Weight				
0.021	0.016	0.012	0.009	0.008
despite	floor	pretty_fast	show_phone	weeks

Term

Topic 6

Weight				
0.004	0.004	0.004	0.004	0.003
tell_servers	weren	waitress_courteous	floor	love_restaurant

Term

Topic 7

Weight				
0.030	0.028	0.016	0.012	0.012
spend_money_restaurant	night	pizza_dough	taste	pay

Term

Topic 8

Weight				
0.007	0.007	0.006	0.005	0.005
pizza_dough	entirely	despite	like_have	arrive_wait_minutes

Term

Topic 9

Weight				
0.004	0.004	0.004	0.004	0.004
despite	omelette_station	special_mother	shake	pretty_fast

Term

Food Industry Example - Negative Review Topics

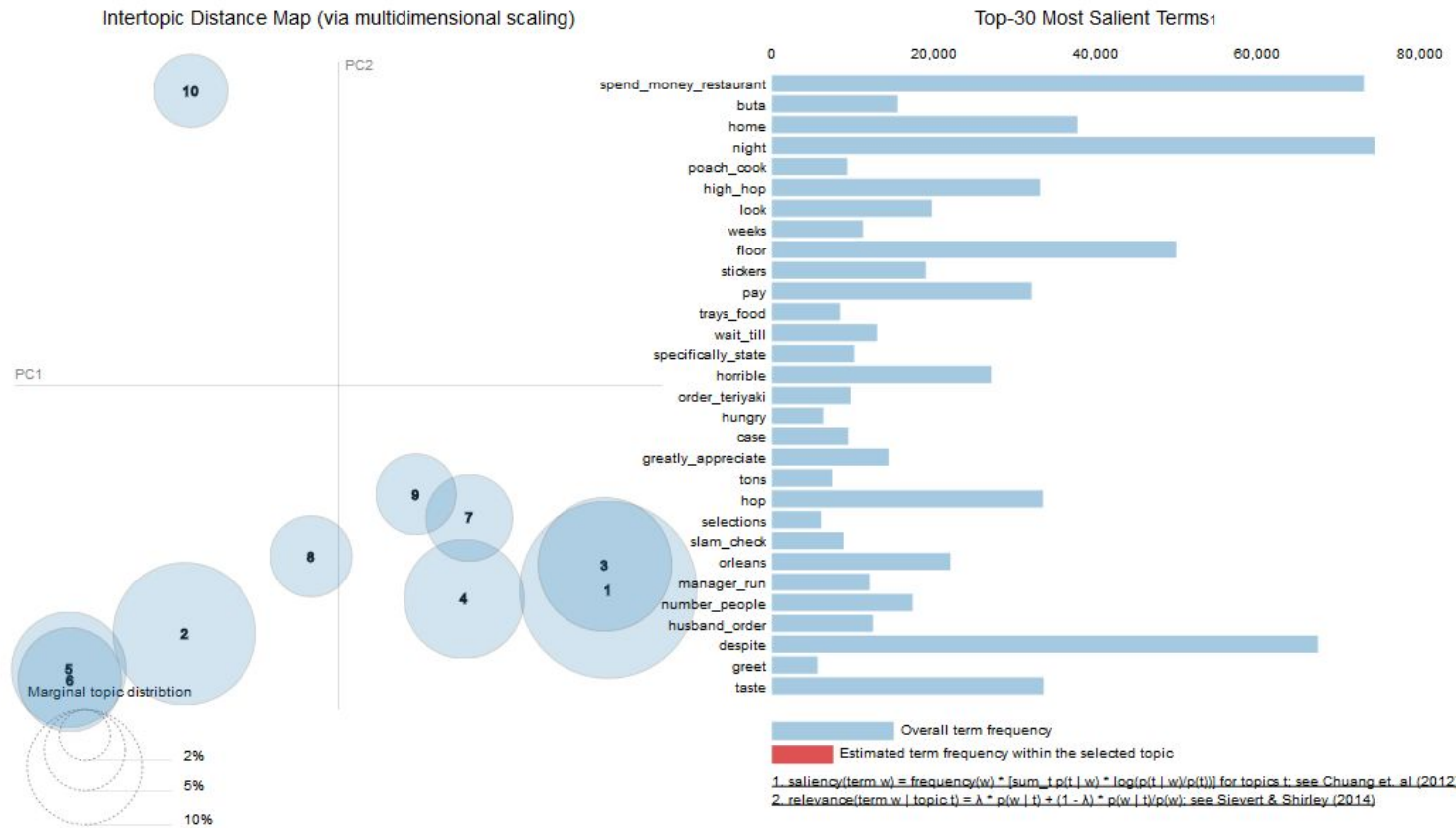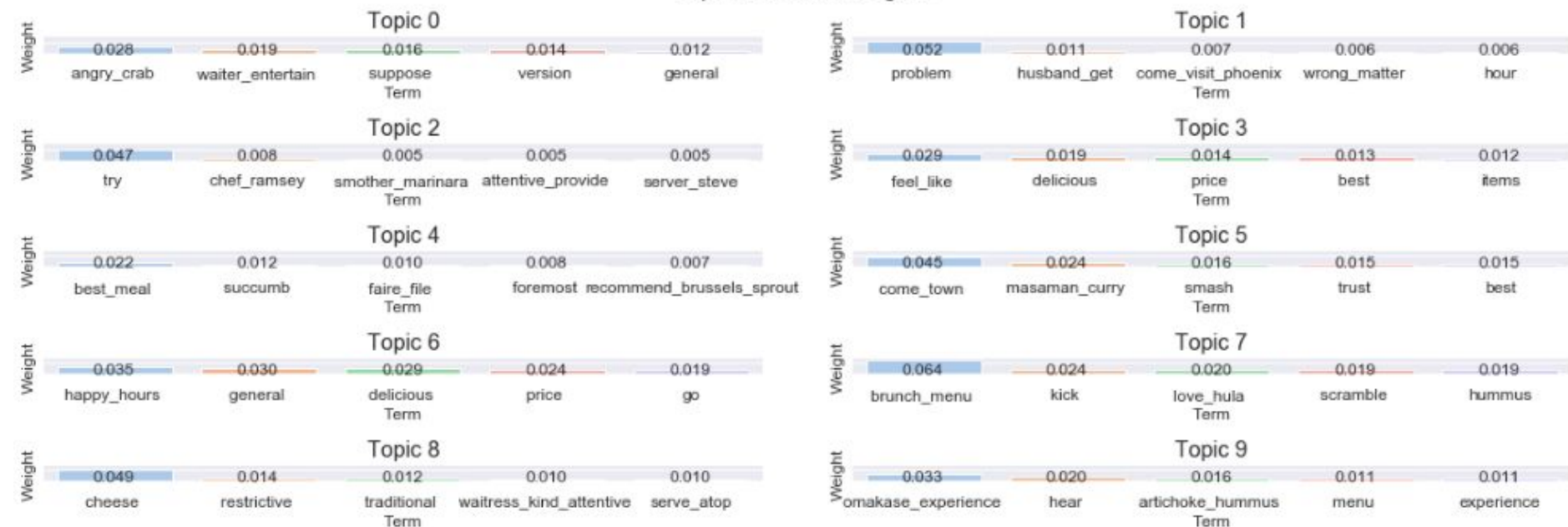| Topic	Topic_%_Contribution	Topic_Keywords	Most_Representative_Text
0.0	0.3308	buta, floor, despite, trays_food, greatly_appreciate, selections, manager_run, greet, decently_f...	Always on the top of the Brunch list, I had to try it, to my dismay, it did not live up to it's ...
1.0	0.3360	home, look, wait_till, high_hop, tons, slam_check, place_try_hard, despite, waitress_come_table,...	Server was a straight bitch , didn't add my blazin rewards after I personally gave her my number...
2.0	0.3039	order_teriyaki, husband_order, floor, hungry, things, night, shift, service_food_good, despite, ...	Originally, I would have rated them five stars because I thought their food was awesome. At the...
3.0	0.2315	crew, waitress_keep_say, despite, base_positive_review, people_sit_wait, potato, price_match, pr...	Ahhh, Grimaldis, I love you so...BUT you really disappointed me today. We were at the mall with...
4.0	0.2456	poach_cook, money_better, sure, leftovers, milk, home, chinese_mexican_food, waitress_come_table...	So this Italian Restaurant was on my to do wish list so I finally gave it a try and I was dissap...
5.0	0.3568	despite, floor, pretty_fast, show_phone, weeks, poorly, specifically_state, couple_time, case, t...	I used to love coming here, the smell of someone else's Gandhi would trigger my own urge to get ...
6.0	0.2046	tell_servers, weren, waitress_courteous, floor, love_restaurant, lack_authenticity, especially, ...	Oyster specials were good and meaty for $2.50 each.\n\nI came here to try the Live Lobster pho t...
7.0	0.4020	spend_money_restaurant, night, pizza_dough, taste, pay, hop, horrible, food_inconsistent, high_h...	There are a lot of high reviews and 5 stars for this place and I am confused as to why. I in no ...
8.0	0.4044	pizza_dough, entirely, despite, like_have, arrive_wait_minutes, serve, stone_cold, pay, waitress...	Avant de commencer, je dois dire que je ne connais rien à la cuisine polonaise et que ce restau...
9.0	0.2819	despite, omelette_station, special_mother, shake, pretty_fast, taste, binge, close_kitchen, yest...	After hearing so many people talk about shake shack we all had to go check it out... After waiti...

Food Industry Example - Negative Review Topics

Distribution of Topic Contributions | Dominant Topic Frequency | Dominant Topic Distribution Among Documents

Food Industry Example - Negative Review Topics

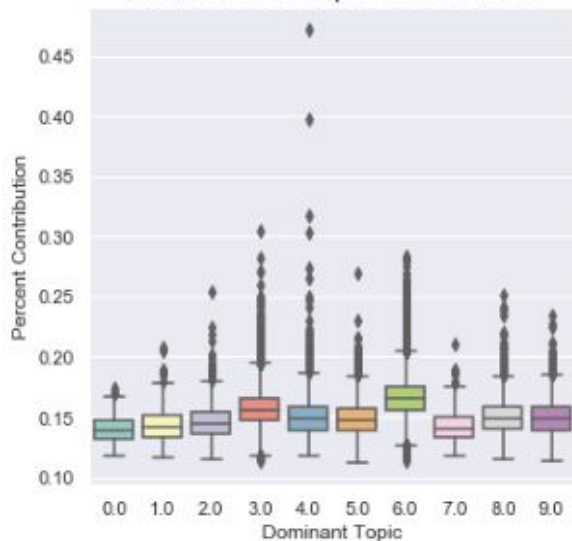# Food Industry Example - Positive Review Topics

Topic Terms & Weights

Topic 0

Weight				
0.028	0.019	0.016	0.014	0.012
angry_crab	waiter_entertain	suppose	version	general
		Term		

Topic 1

Weight				
0.052	0.011	0.007	0.006	0.006
problem	husband_get	come_visit_phoenix	wrong_matter	hour
		Term		

Topic 2

Weight				
0.047	0.008	0.005	0.005	0.005
try	chef_ramsey	smother_marinara	attentive_provide	server_steve
		Term		

Topic 3

Weight				
0.029	0.019	0.014	0.013	0.012
feel_like	delicious	price	best	items
		Term		

Topic 4

Weight				
0.022	0.012	0.010	0.008	0.007
best_meal	succumb	faire_file	foremost	recommend_brussels_sprout
		Term		

Topic 5

Weight				
0.045	0.024	0.016	0.015	0.015
come_town	masaman_curry	smash	trust	best
		Term		

Topic 6

Weight				
0.035	0.030	0.029	0.024	0.019
happy_hours	general	delicious	price	go
		Term		

Topic 7

Weight				
0.064	0.024	0.020	0.019	0.019
brunch_menu	kick	love_hula	scramble	hummus
		Term		

Topic 8

Weight				
0.049	0.014	0.012	0.010	0.010
cheese	restrictive	traditional	waitress_kind_attentive	serve_atop
		Term		

Topic 9

Weight				
0.033	0.020	0.016	0.011	0.011
omakase_experience	hear	artichoke_hummus	menu	experience
		Term		

Food Industry Example - Positive Review Topics

| Topic	Topic_%_Contribution	Topic_Keywords	Most_Representative_Text
0.0	0.1744	angry_crab, waiter_entertain, suppose, version, general, chef_ramsey, surely, original, price, c...	Its a bit difficult to say if the food here is any good or not. I'm writing this review while go...
1.0	0.2078	problem, husband_get, come_visit_phoenix, wrong_matter, hour, menu, quality_quantity, nutella_ba...	I came in on thursday to order a 50th anniversary cake. The lady helping me was friendly and he...
2.0	0.2538	try, chef_ramsey, smother_marinara, attentive_provide, server_steve, north_phoenix, boyfriend_ta...	The atmosphere is really elegant at this place especially at the top floor. Our server did a rea...
3.0	0.3052	feel_like, delicious, price, best, items, run, meat_combo_platter, cheese, second_night, toast	Famous Daves... Famous Daves....\nYes, I know a few.\n\nDavid Lynch \nDavid Cassidy \nDavid Bore...
4.0	0.4726	best_meal, succumb, faire_file, foremost, recommend_brussels_sprout, dress, artichoke_hummus, fe...	Buffet-Restaurants haben fast alle Hotels in Las Vegas. Im teuren Vegas eine Möglichkeit, gut u...
5.0	0.2690	come_town, masaman_curry, smash, trust, best, rito, flavorful, ahead_time, delicious, hand	I am no expert on soul food, but the fried chicken is finger lickin' good. Please don't sue me, ...
6.0	0.2841	happy_hours, general, delicious, price, go, plastic_bag, say_hour_wait, food_general, feel_like,...	This is definitely the best bar in the area. It has a ton of British charm, European and domesti...
7.0	0.2109	brunch_menu, kick, love_hula, scramble, hummus, kinda_slow, remind_mexican, potato_casserole, fe...	I totally agree with the other Yelpers! It is the best authentic Chinese food in Vegas. Prices...
8.0	0.2511	cheese, restrictive, traditional, waitress_kind_attentive, serve_atop, taste, crisp_fresh, best,...	Sunday at Noon and we were able to get a table immediately even though it was a packed house. F...
9.0	0.2350	omakase_experience, hear, artichoke_hummus, menu, experience, terrace_cafe, pineapple_fry, mexic...	Hints: Bone-In Rib Steaks\nBone-In Rib Steaks\nBone-In Rib Steaks\nBone-In Rib Steaks\nBone-In R...

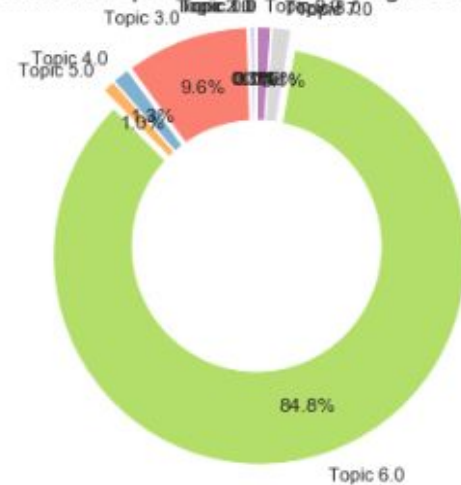# Food Industry Example - Positive Review Topics

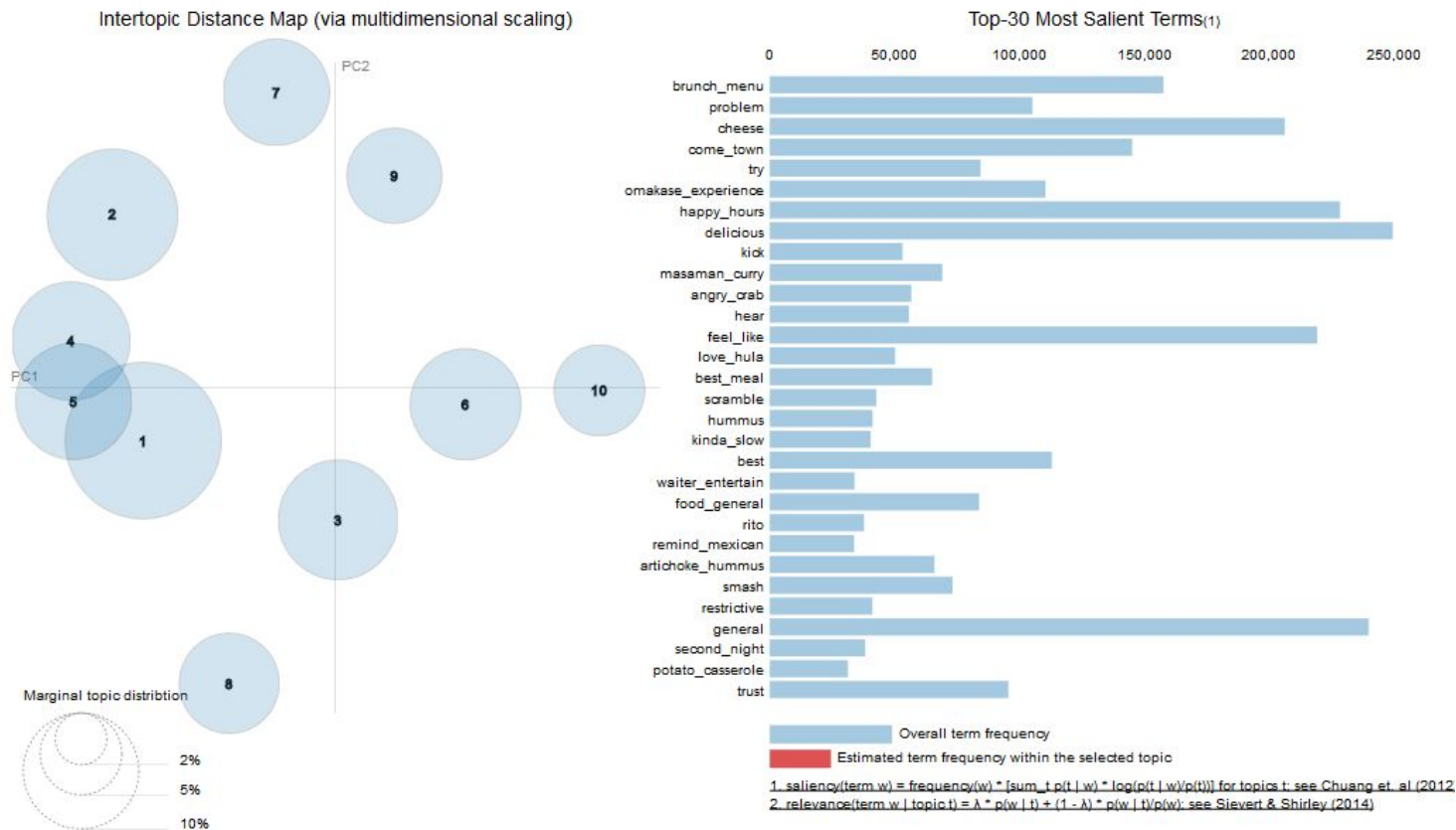Distribution of Topic Contributions

Dominant Topic Frequency
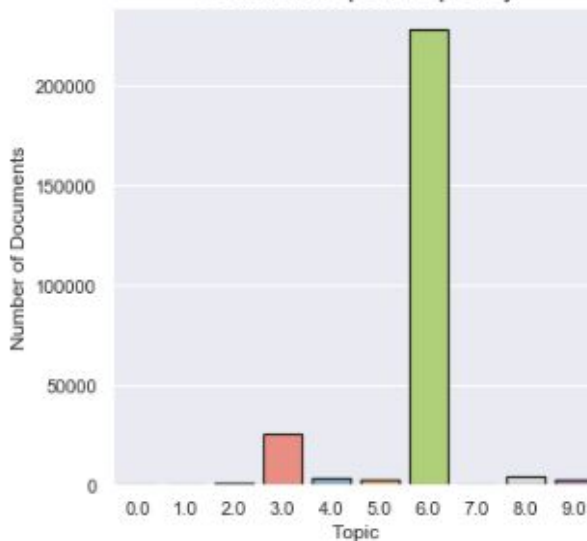
Dominant Topic Distribution Among Documents

Food Industry Example - Positive Review Topics

Client Recommendations

Based on the possible implementations of the product constructed and the results that may obtained from it, the following are some insightful uses that I would promote to any future clients:

- **To determine the direction of favorability for a particular product or initiative**
  - Success of a loss-leader in creating profitable upsell scenarios
  - Response to rebranding or new marketing strategy
  - Reception of an aesthetic change such as a remodel or relocation

- **To assess global versus local analysis**
  - Alongside a market penetration analysis to boost insight
  - Verifying Consistency in customer service across localities for a chain of businesses
  - Distinguishing between circumstantial and pervasive issues for a particular industry

Conclusion

In constructing this text analysis pipeline I encountered several problems consistent with common parables in reference materials, a notable one being that the quality of the topic model output rests tremendously on the quality of the preprocessing steps taken prior to training. The initial extraction of the text data from the overall Yelp academic dataset, and the subsequent transformation into a serviceable vector that can be used to train an LDA model, were both some of the most computationally expensive and consequential aspects of the overall procedure. Minor changes in these steps, such as filtering the vocabulary or adding n-grams to the corpus, often led to drastically differing results, which themselves proved uniquely challenging to interpret. Moving forward I would add more visualizations such as word clouds to assist in readability, as the success of each topic model created hinges on a combination of the user's domain knowledge and the ease in which it can be interpreted