# Yelp Review Analysis

Group Tattarrattat
Shucen Liu, Xibai Wang, Can Liu,
Lily Zhifan Zhou

# Project Goal and Results

Facilitate customers' restaurant searching process

Suggest similar restaurants based on user inputs (category, location, price)

Visualize data into diagrams of food score and service score, with all restaurants represented as dots

Create a table with other information of these similar restaurants


CHICAGO
RESTAURANT WEEK
JANUARY 30 - FEBRUARY 12
Cafe Ba-Ba-Reeba!

# System Structure

1. Crawler
2. Create Database
3. Sentiment Analysis
4. Topic Modeling
5. Recommendation
6. Visualization
7. Website

# Crawler

Name, Category, Address

Price Range, Reviews

900+ Most Popular Restaurants

（mainly located in Downtown and Hyde Park)

200 Reviews Each

1. Oriole

92 reviews

$$$$ · American (New)

Near West Side, West Loop

661 W. Walnut St.
Chicago, IL 60661
(312) 877-5339

I cannot possibly explain enough how truly magical of an experience we had at Oriole. Chef Noah Sandoval and Chef Genie Kwon are truly master artists in their craft. Not to be cliche… read more


2. Girl & the Goat

5780 reviews

$$$ · American (New)

Near West Side, West Loop

809 W Randolph St
Chicago, IL 60607
(312) 492-6262

Delicious brunch!!! I had the shrimp grits and it blew my mind. Cute atmosphere and fast service. read more

**Sarah S.**
Chicago, IL
1 friend
34 reviews

3/9/2017

This place has the best tasting menu we have had in Chicago and we have tried most. The food is unique and delicious. The wine pairing was well thought out and fit every course very well. The service was great. I highly recommend.

Was this review ...?

Useful | Funny | Cool

**Shari B.**
Chicago, IL
92 friends
168 reviews
110 photos
Elite '17

3/8/2017

1 check-in

Oriole is amazing. It's at a very discreet location, to say the least. And they don't have valet parking, fyi. After you enter the restaurant, you are kept in an interesting old elevator shaft before they take you into the dining room. I was offered a spiked hot apple cider while we waited for the rest of our group to arrive. I was pleasantly surprised to see that this place had a more laid-back atmosphere than I expected and they also played some good reggae music as the evening ended. My table passed on wine pairings

# Django Database

2 Models:

- Restaurant
- Rating
  - Maps to Restaurant

# Sentiment Analysis

Use NLTK (Natural Language Toolkit) and its subpackage SentiwordNet

Calculate positive, negative, and objective scores (sum=1) for each word

Take the average of all related words for a certain keyword

```
In [3]: related = list(swn.senti_synsets('like'))

In [4]: related
Out[4]:
[SentiSynset('like.n.01'),
 SentiSynset('like.n.02'),
 SentiSynset('wish.v.02'),
 SentiSynset('like.v.02'),
 SentiSynset('like.v.03'),
 SentiSynset('like.v.04'),
 SentiSynset('like.v.05'),
 SentiSynset('like.a.01'),
 SentiSynset('like.a.02'),
 SentiSynset('alike.a.01'),
 SentiSynset('comparable.s.02')]
```

# Sentiment Analysis

Process and Funny Issues:

Wanted to get the scores for each word, but NLTK suggests a bunch of related words, their different parts of speech, and usage frequency

```
In [7]: category
Out[7]:
{'JJ': 'a',
 'JJR': 'a',
 'JJS': 'a',
 'NN': 'n',
 'NNP': 'n',
 'NNPS': 'n',
 'NNS': 'n',
 'PRP': 'n',
 'RB': 'r',
 'RBR': 'r',
 'RBS': 'r',
 'VBD': 'v',
 'VBG': 'v',
 'VBN': 'v',
 'VBP': 'v',
 'VBZ': 'v'}
```

```
In [9]: tag = nltk.pos_tag(['like'])

In [10]: tag
Out[10]: [('like', 'IN')]
```

# Sentiment Analysis

So, wanted to extract the word and part of speech to see if it matches with the one in use, then take the most frequent one → Use nltk.pos_tag

“IN” represents prepositions

```
In [7]: category
Out[7]:
{'JJ': 'a',
 'JJR': 'a',
 'JJS': 'a',
 'NN': 'n',
 'NNP': 'n',
 'NNPS': 'n',
 'NNS': 'n',
 'PRP': 'n',
 'RB': 'r',
 'RBR': 'r',
 'RBS': 'r',
 'VBD': 'v',
 'VBG': 'v',
 'VBN': 'v',
 'VBP': 'v',
 'VBZ': 'v'}
```

```
In [11]: tag = nltk.pos_tag(['creamy'])

In [12]: tag
Out[12]: [('creamy', 'NN')]
```

# Sentiment Analysis

Another thing is some words might get categorized into a part of speech it doesn't have

"Creamy:" adjective → noun

# Sentiment Analysis

Apparently, this way is not going to work. So tried to use nltk.simplify, which directly translates the NLTK POS to regular POS

A lot of the POS still doesn’t match

# Sentiment Analysis

```
In [34]: related = list(swn.senti_synsets('evil'))

In [35]: related[0]
Out[35]: SentiSynset('evil.n.01')

In [36]: related[0].pos_score()
Out[36]: 0.75

In [37]: related = list(swn.senti_synsets('sweet'))

In [38]: related[0].pos_score()
Out[38]: 0.0

In [39]: related[0].neg_score()
Out[39]: 0.0

In [40]: related[0].obj_score()
Out[40]: 1.0
```

Some NLTK words have really weird scoring results, such as “evil” “sweet”, etc.

# Sentiment Analysis

```
In [3]: related = list(swn.senti_synsets('like'))

In [4]: related
Out[4]:
[SentiSynset('like.n.01'),
 SentiSynset('like.n.02'),
 SentiSynset('wish.v.02'),
 SentiSynset('like.v.02'),
 SentiSynset('like.v.03'),
 SentiSynset('like.v.04'),
 SentiSynset('like.v.05'),
 SentiSynset('like.a.01'),
 SentiSynset('like.a.02'),
 SentiSynset('alike.a.01'),
 SentiSynset('comparable.s.02')]
```

Eventually, decided to take the average of all related words. Recap -- “like”

Gives a much more reliable and making-sense result

```
In [26]: pos_score/len(related)
Out[26]: 0.2840909090909091

In [27]: neg_score/len(related)
Out[27]: 0.0227272727272728

In [28]: obj_score/len(related)
Out[28]: 0.6931818181818182
```

# Topic Modeling

1. Train Topic Models
2. Represent appearance of words with vectors
3. Compute inner product (“distance” between models and given sentences)
4. Determine topics
5. Test

# Train Topic Model

- Lists containing topical vocabularies
- Training data from Yelp Reviews
- Tokenize sentences + stem words
- Words with high frequencies + do not overlap

```
food = ['food', 'taste', 'dish', 'savory', 'sweet', 'salty', 'eat', 'flavor']
service = ['service', 'friendly', 'quick', 'attitude', 'staff', 'efficient']
ambience = ['clean', 'location', 'space', 'classy', 'room', 'look']
price = ['price', 'cheap', 'expensive', 'quite', 'inexpensive', 'affordable', 'bill','overpriced']
```

# Construct vector

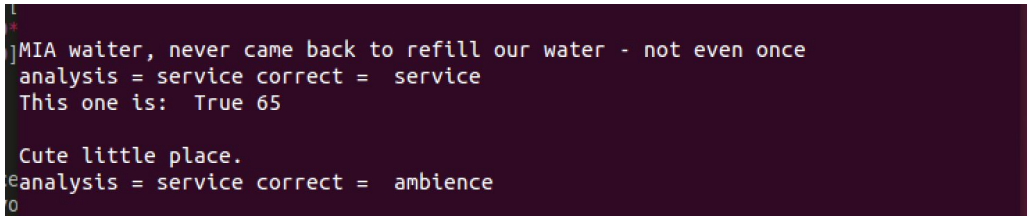
1. “1” if present, “0” otherwise
2. Vectors for both trained models and test data

# Compute distance

1. Inner Product
2. Sort the results
3. Normalize

# Determine Topics

1. If normalized inner product >= threshold
2. Testing: 60%
3. Larger training data?
4. More “food”; less “price”?
5. Better Threshold?



# Recommendation

**Restaurant:**

**Location:**

Specify branch location. (Optional)

**Rank Neighborhood:**

e.g. 1 (meaning the most important)

**Rank Category:**

e.g. 3 (meaning the least important)

**Rank Price:**

e.g. 2 (order must not repeat)

Submit!

Filter by similar traits:

- Neighborhood
- Cuisine Category
- Price Range

1. [Neighborhood, Category, Price]:

| Name | Neighborhood | Cuisine | Price | Food Score | Ambience Score | Service Score | Price Score |
|---|---|---|---|---|---|---|---|
| Alinea | Lincoln Park | American | $$$$ | 74.0 | 68.25 | 63.05 | 60.98 |
| Alinea | Lincoln Park | American | $$$$ | 74.0 | 68.25 | 63.05 | 60.98 |
| Del Seoul | Lincoln Park | Korean | $ | 59.12 | 71.64 | 57.86 | 70.89 |
| Cupbop + Ramen | Lincoln Park | Korean | $ | 58.25 | 71.65 | 63.67 | 59.77 |
| Chicago Halal | Lincoln Park | Mediterranean | $ | 78.83 | 54.86 | 52.1 | 46.5 |
| Boka | Lincoln Park | American | $$$ | 53.33 | 57.61 | 65.2 | 65.31 |
| Captain's Catch | Lincoln Park | Cajun/Creole | $$ | 59.22 | 76.88 | 77.45 | 70.46 |
| U Rice by Lans | Lincoln Park | Asian | $$ | 65.81 | 70.15 | 70.4 | 63.52 |

2. [Category, Price, Neighborhood]

| Name | Neighborhood | Cuisine | Price | Food Score | Ambience Score | Service Score | Price Score |
|---|---|---|---|---|---|---|---|
| Alinea | Lincoln Park | American | $$$$ | 74.0 | 68.25 | 63.05 | 60.98 |
| Oriole | Near West Side, West Loop | American | $$$$ | 65.56 | 63.39 | 62.42 | 60.85 |
| Girl & the Goat | Near West Side, West Loop | American | $$$ | 65.15 | 66.99 | 58.17 | 64.72 |
| Eden | Near West Side | American | $$$ | 56.15 | 57.72 | 62.54 | 65.71 |
| Alinea | Lincoln Park | American | $$$$ | 74.0 | 68.25 | 63.05 | 60.98 |
| Au Cheval | Near West Side, West Loop | American | $$ | 62.72 | 64.51 | 58.63 | 63.46 |
| Giant | Logan Square | American | $$ | 59.75 | 56.04 | 62.17 | 53.33 |
| Knife & Tine | Lincoln Park, DePaul | American | $$ | 50.29 | 67.67 | 61.1 | 65.96 |

# Rating Algorithm

Review_sentiment

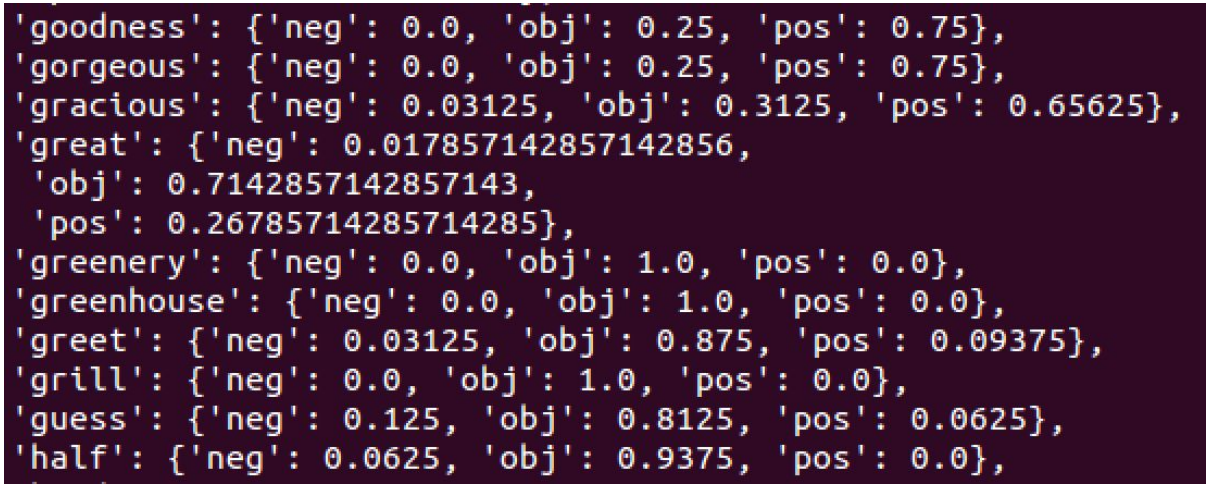
Review_count

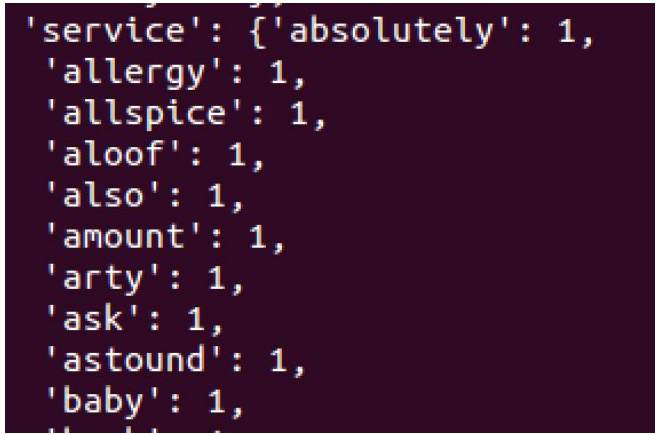
Total Score = Positive + Objective + Negative

%Positive = Positive / Total Score

Score = %Positive / (%Positive + %Negative)

If Total Score = 0, Score = 50.



# Multiple Locations

**Restaurant:**

**Location:**

Specify branch location. (Optional)

1. **MingHin** Cuisine

283 reviews

$$ · Dim Sum, Cantonese, Asian Fusion

The Loop

333 E Benton Pl
Chicago, IL 60601
(312) 228-1333

This restaurant accepts pickup and delivery

Start Order

Ever since the new outlet opened, i have been a regular at **Ming Hin**. The food is authentic and the waiters are always attentive and polite . Unlike some chinese restaurants which… read more

2. **MingHin** Cuisine

1455 reviews

$$ · Seafood, Dim Sum, Cantonese

Chinatown

2168 S Archer Ave
Chicago, IL 60616
(312) 808-1999

This restaurant accepts pickup and delivery

Start Order

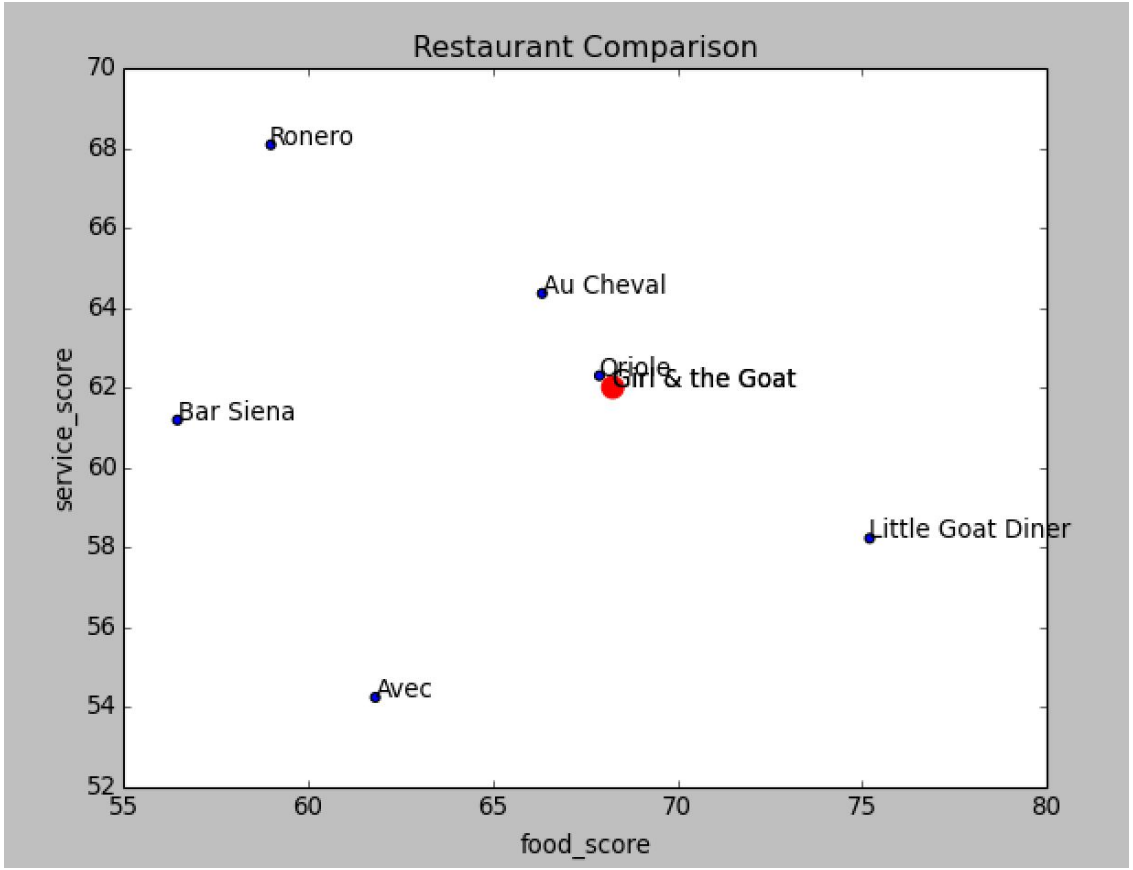
| Name | Neighborhood | Cuisine | Price | Food Score | Ambience Score | Service Score | Price Score |
|---|---|---|---|---|---|---|---|
| MingHin Cuisine | The Loop | Dim | $$ | 62.42 | 70.81 | 58.31 | 65.39 |
| The Dearborn | The Loop | Salad | $$ | 58.66 | 71.39 | 60.64 | 69.63 |
| Spotted Monkey | The Loop | Asian | $$ | 57.38 | 62.37 | 66.84 | 59.41 |
| The Gage | The Loop | Gastropubs | $$$ | 45.18 | 62.94 | 55.28 | 59.42 |
| Brightwok Kitchen | The Loop | Asian | $$ | 53.97 | 58.05 | 68.68 | 63.3 |
| Revival Food Hall | The Loop | Food | $$ | 57.24 | 61.29 | 62.64 | 56.5 |
| The Marq | The Loop | American | $$ | 69.25 | 64.59 | 65.5 | 61.21 |
| The Fat Shallot | The Loop | Food | $$ | 43.59 | 82.04 | 65.3 | 56.05 |

| Name | Neighborhood | Cuisine | Price | Food Score | Ambience Score | Service Score | Price Score |
|---|---|---|---|---|---|---|---|
| MingHin Cuisine | Chinatown | Seafood | $$ | 56.37 | 64.05 | 60.14 | 66.78 |
| Qing Xiang Yuan Dumpling | Chinatown | Chinese | $$ | 68.78 | 64.17 | 73.64 | 67.09 |
| Chi Cafe | Chinatown | Chinese | $ | 71.82 | 60.4 | 57.27 | 62.13 |
| MingHin Cuisine | Chinatown | Seafood | $$ | 56.37 | 64.05 | 60.14 | 66.78 |
| Little Sheep Mongolian Hot Pot | Chinatown | Mongolian | $$ | 62.92 | 56.6 | 77.84 | 63.81 |
| Claypot Noodle Shop | Chinatown | Noodles | $ | 61.3 | 34.8 | 61.91 | 52.82 |
| Legend Tasty House | Chinatown | Chinese | $ | 67.51 | 74.87 | 63.46 | 55.77 |
| Yummy Yummy Noodles | Chinatown | Chinese | $ | 55.74 | 88.01 | 64.77 | 59.69 |

Filtered by [Neighborhood, Category, Price]

# Visualization

Compare food and service



# Q & A