Restaurant Reviews and Social Media

MIDS W205-2016 Spring
Group 4
Marlea Gwinn, Sue Yang, Siddharth Singhal

Problem Overview

"What restaurant should I go to and what should I order?"

- Restaurant reviews are

  blobs and streams of text => tough to parse for a human reader

- Often browsable by:
  - Recency or
  - Rating or
  - Certain Demographic of author or
  - Higher ranked critics

but **true insights in food reviews remain** **hidden** **in the text material** posted!

- Studies show that **online ratings are one of the most trusted sources** of consumer confidence in e-commerce **decisions**
- But research consistently suggests that they are **systematically biased and easily manipulated**.

No easy-way !

To parse through reviews and find out what the popular food items are.

Better exploration of reviews!

We bring-out the true-insights shared within reviews, often hidden in streams of data by

- Text-analysis
- Multiple data sources
- Independent and unbiased view-point

We add more facets to the stream of restaurant reviews, to enable better, unbiased and truer review-browsing experiences!

Dataset

Yelp Challenge Dataset

- **2.2M** reviews and **591K** tips by **552K** users for **77K** businesses
- **566K** business attributes, e.g., hours, parking availability, ambience.

Twitter Streaming Data

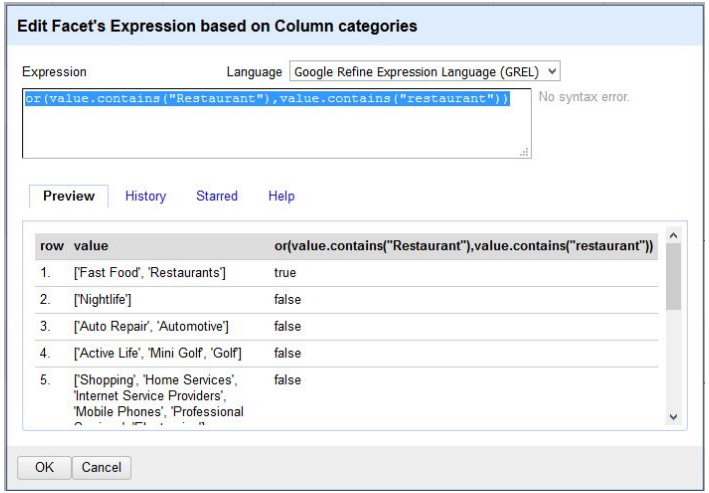
- Twitter streaming API allows us to pull social media data for given restaurants

Data Cleaning

Business (77K items)

- **OpenRefine -- remove unwanted data**

  E.g. yelp_academic_dataset_business.csv contains business with multiple categories such as barber shop, auto shop, home service... etc. We only keep the restaurant category.



Edit Facet's Expression based on Column categories

Expression

Language: Google Refine Expression Language (GREL)

```
or(value.contains("Restaurant"),value.contains("restaurant"))
```

No syntax error.

row	value	or(value.contains("Restaurant"),value.contains("restaurant"))
1.	['Fast Food', 'Restaurants']	true
2.	['Nightlife']	false
3.	['Auto Repair', 'Automotive']	false
4.	['Active Life', 'Mini Golf', 'Golf]	false
5.	['Shopping', 'Home Services', 'Internet Service Providers', 'Mobile Phones', 'Professional	false

- **Microsoft Excel -- remove embedded carriage return**

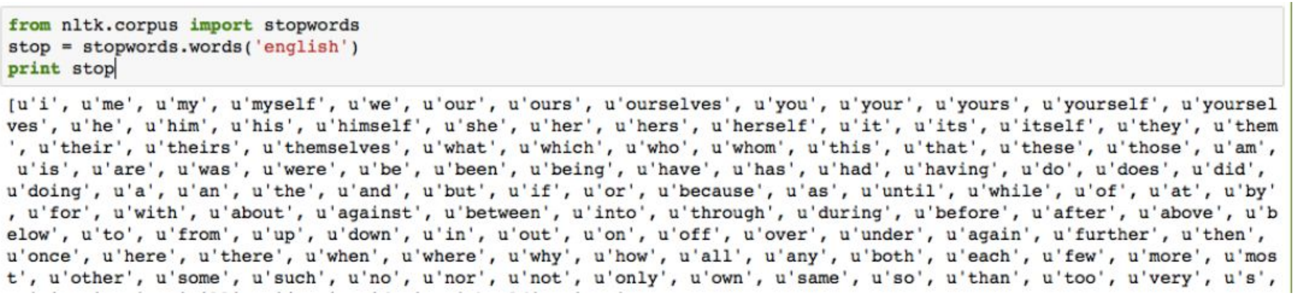
  We found that when using OpenCSVSerde it cannot handle embedded CR within cells.

Data Cleaning (Continued)

Review(2.2M items)

- **Beautifulsoup library -- remove document and tag**
- **nltk -- remove stop words**

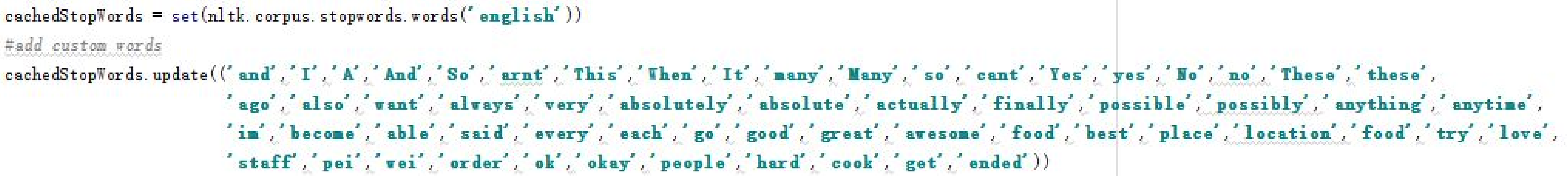
nltk is a nlp python package. Stop words are words which do not contain important significance to be used in Search Queries



```
from nltk.corpus import stopwords
stop = stopwords.words('english')
print stop
```

```
[u'i', u'me', u'my', u'myself', u'we', u'our', u'ours', u'ourselves', u'you', u'your', u'yours', u'yourself', u'yoursel
ves', u'he', u'him', u'his', u'himself', u'she', u'her', u'hers', u'herself', u'it', u'its', u'itself', u'they', u'them
', u'their', u'theirs', u'themselves', u'what', u'which', u'who', u'whom', u'this', u'that', u'these', u'those', u'am',
 u'is', u'are', u'was', u'were', u'be', u'been', u'being', u'have', u'has', u'had', u'having', u'do', u'does', u'did',
u'doing', u'a', u'an', u'the', u'and', u'but', u'if', u'or', u'because', u'as', u'until', u'while', u'of', u'at', u'by'
, u'for', u'with', u'about', u'against', u'between', u'into', u'through', u'during', u'before', u'after', u'above', u'b
elow', u'to', u'from', u'up', u'down', u'in', u'out', u'on', u'off', u'over', u'under', u'again', u'further', u'then',
u'once', u'here', u'there', u'when', u'where', u'why', u'how', u'all', u'any', u'both', u'each', u'few', u'more', u'mos
t', u'other', u'some', u'such', u'no', u'nor', u'not', u'only', u'own', u'same', u'so', u'than', u'too', u'very', u's',
```

We updated stop words list based on result we got



```
cachedStopWords = set(nltk.corpus.stopwords.words('english'))
#add custom words
cachedStopWords.update(('and','I','A','And','So','arnt','This','When','It','many','Many','so','cant','Yes','yes','No','no','These','these',
                        'ago','also','want','always','very','absolutely','absolute','actually','finally','possible','possibly','anything','anytime',
                        'im','become','able','said','every','each','go','good','great','awesome','food','best','place','location','food','try','love',
                        'staff','pei','wei','order','ok','okay','people','hard','cook','get','ended'))
```

Parsing and Modeling

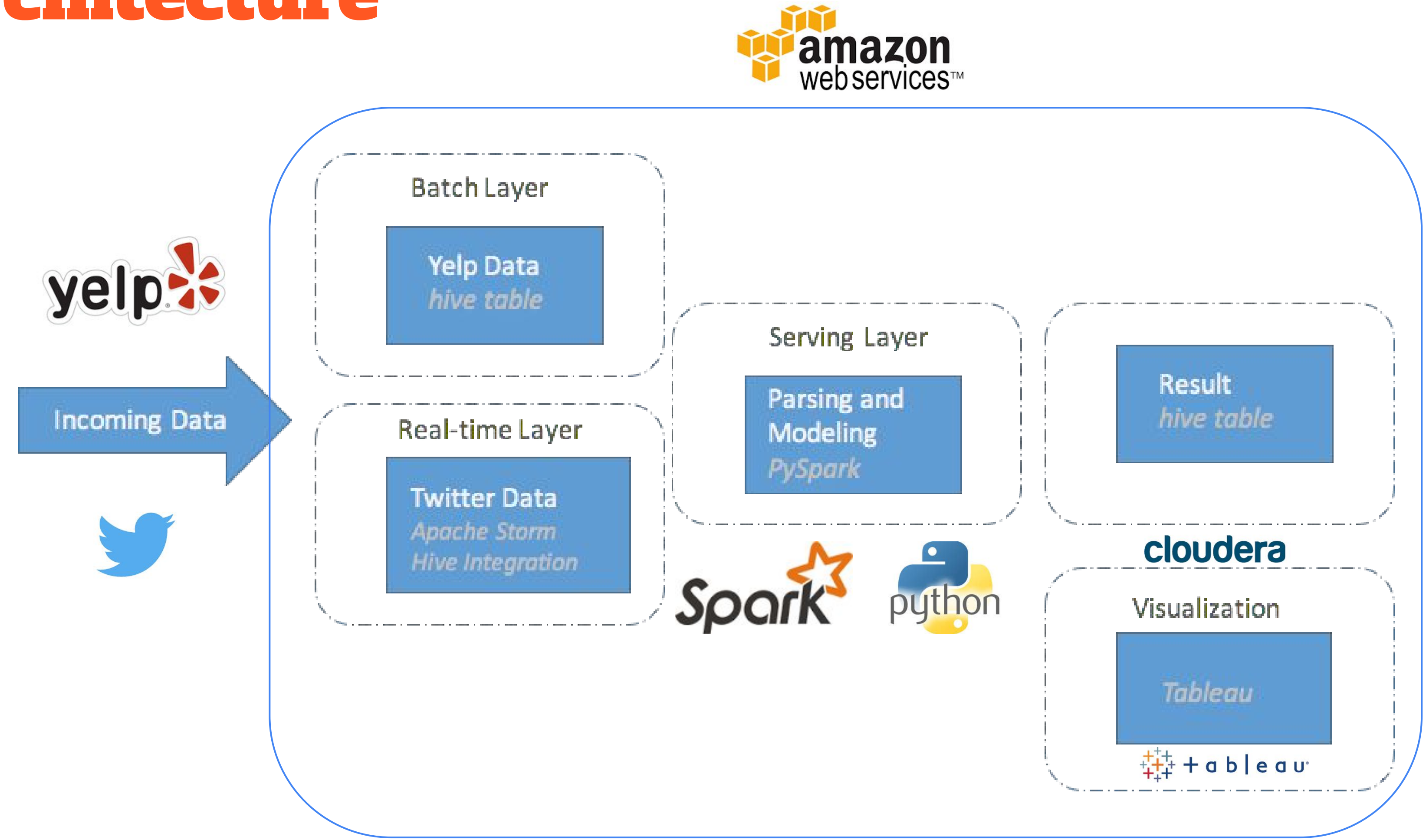
Text(review, twitter) analysis

nltk regexpTokenizer -- Tokenize segments a document and in our case the segments are words. The benefit of using this library is that it allows we define tokenization with regular expression.

Word frequency -- Select top 5 words with highest frequency

N-grams -- We used bigram and trigram and measured using Pointwise Mutual Information. The top 5 bigram/trigram collocations are returned.

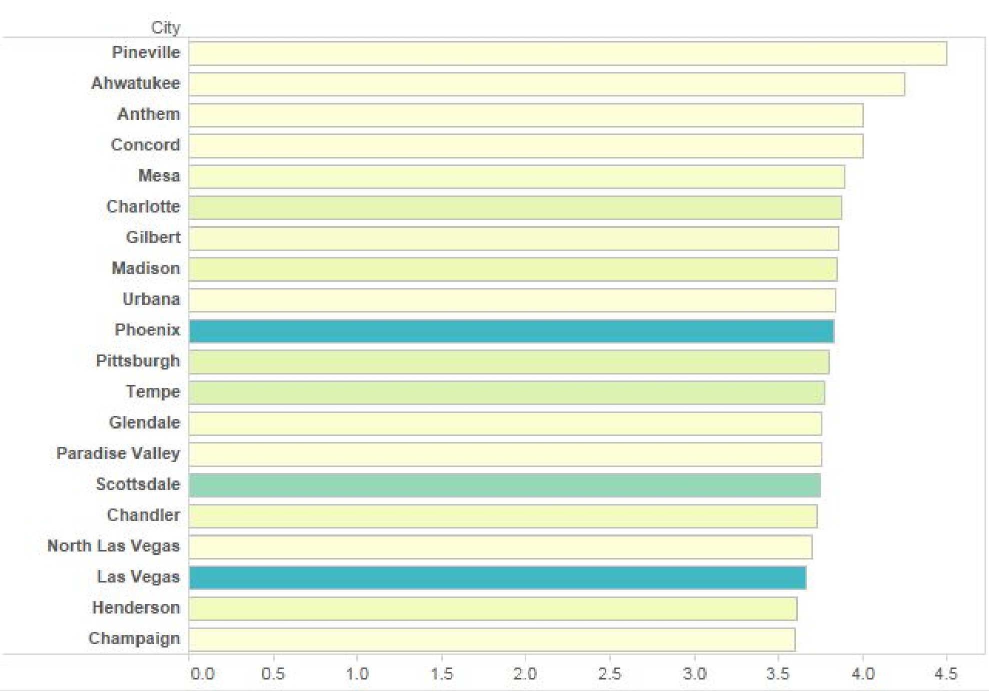
Architecture



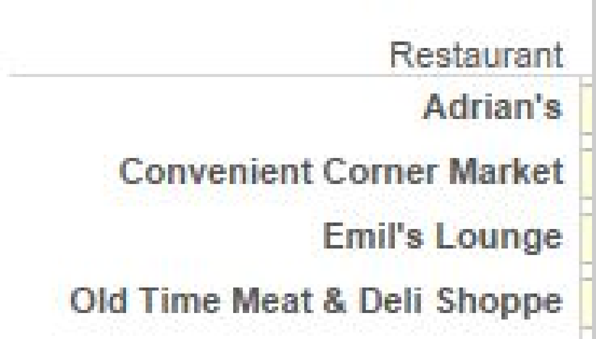
Visuals

Result Visualization

Top Cities for food



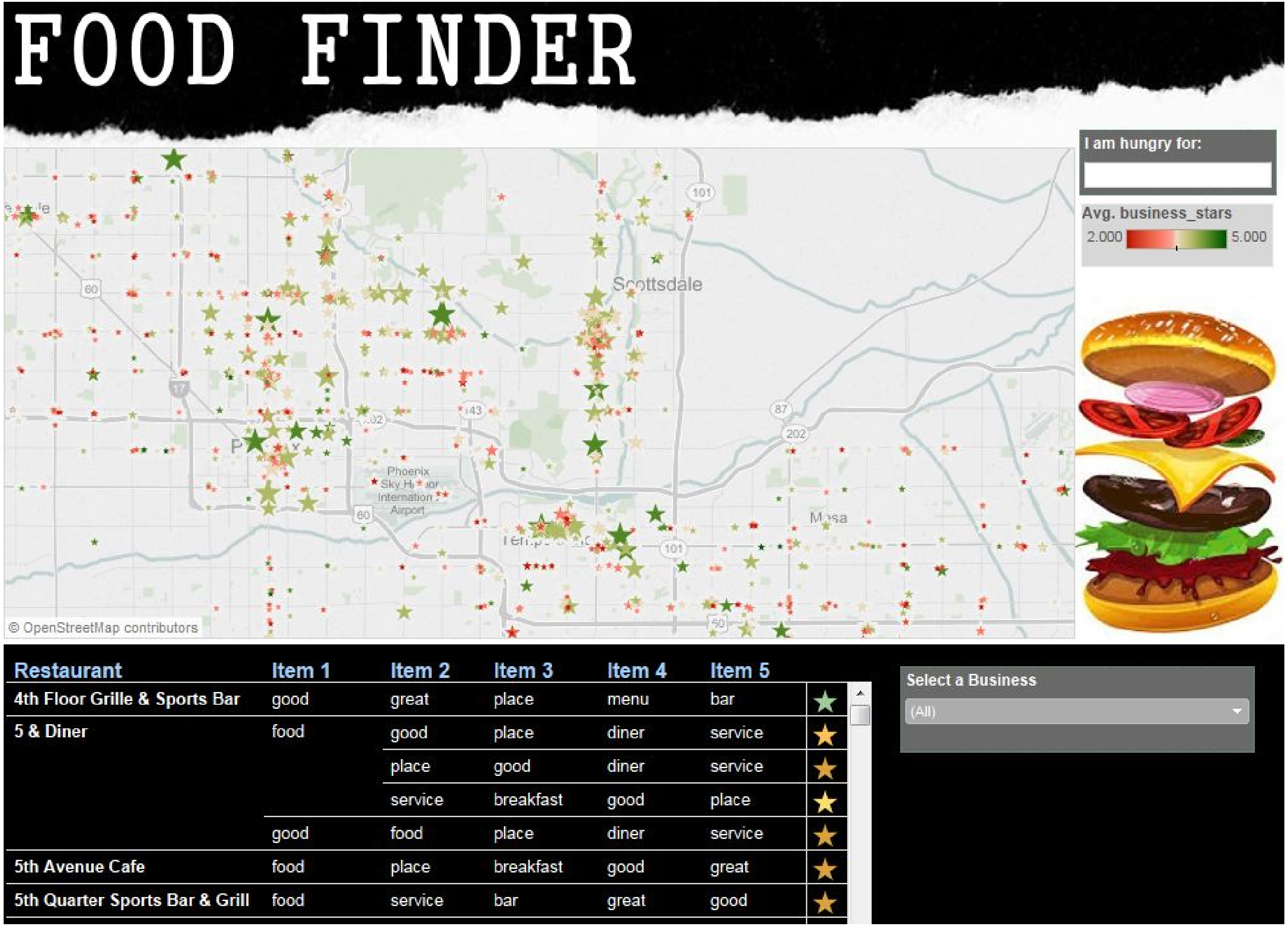
Top Restaurants



Restaurant
Adrian's
Convenient Corner Market
Emil's Lounge
Old Time Meat & Deli Shoppe

5 Star

Over 20 Reviews



FOOD FINDER

Scottsdale
Phoenix Sky Harbor International Airport
Tempe
Mesa
60
17
101
202
143
87
© OpenStreetMap contributors

I am hungry for:

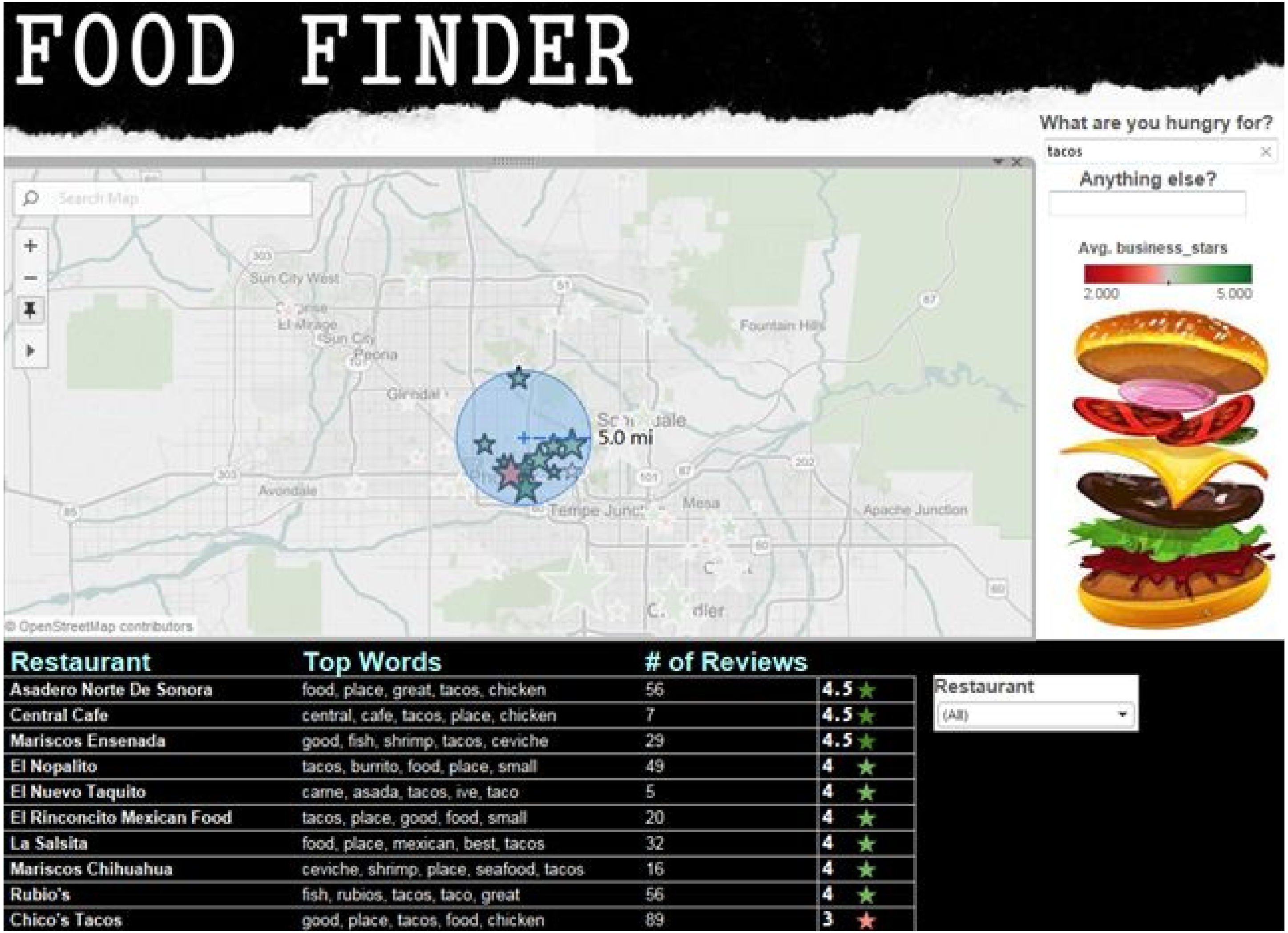
Avg. business_stars

2.000 5.000

Restaurant	Item 1	Item 2	Item 3	Item 4	Item 5	
4th Floor Grille & Sports Bar	good	great	place	menu	bar	★
5 & Diner	food	good	place	diner	service	★
		place	good	diner	service	★
		service	breakfast	good	place	★
	good	food	place	diner	service	★
5th Avenue Cafe	food	place	breakfast	good	great	★
5th Quarter Sports Bar & Grill	food	service	bar	great	good	★

Select a Business

(All)



FOOD FINDER

What are you hungry for?

tacos

Anything else?

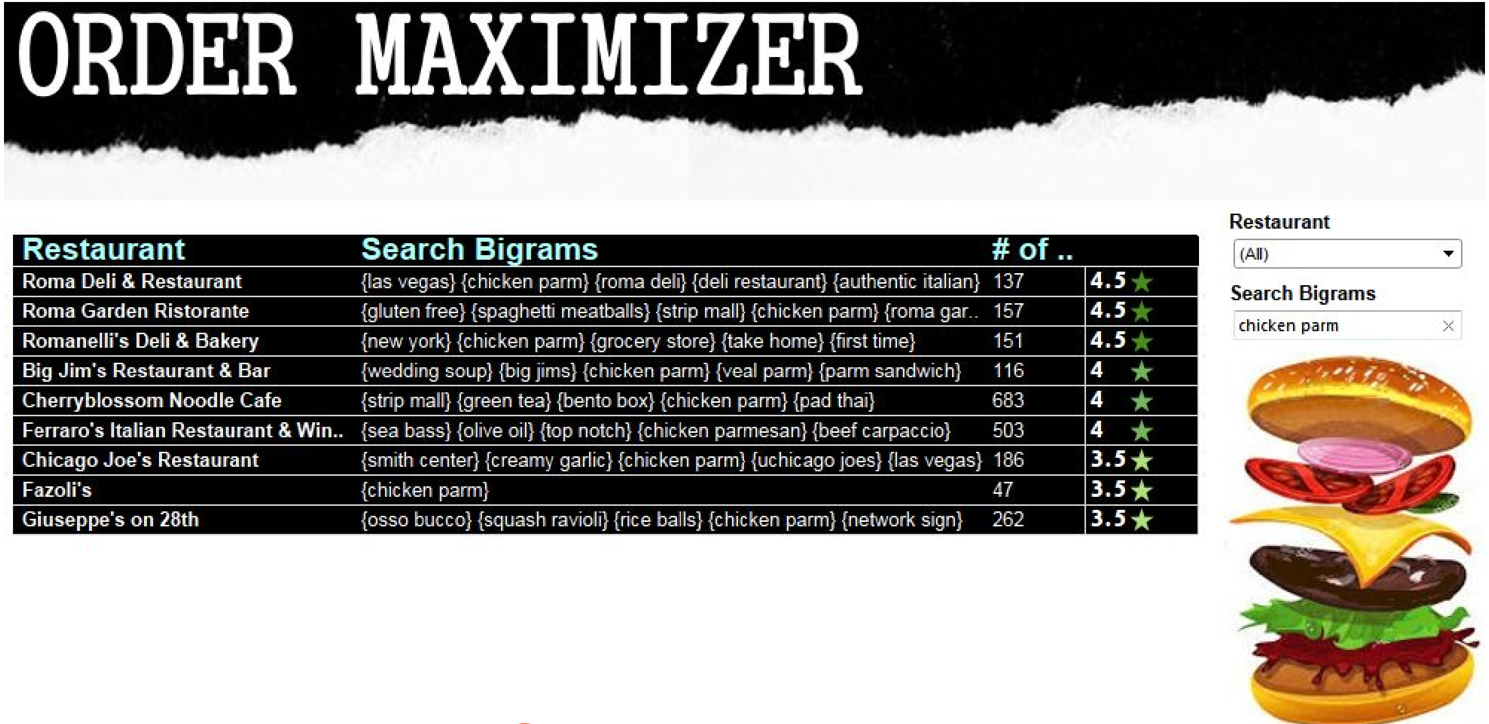
Avg. business_stars

2.000 5.000

Restaurant	Top Words	# of Reviews	
Asadero Norte De Sonora	food, place, great, tacos, chicken	56	4.5 ★
Central Cafe	central, cafe, tacos, place, chicken	7	4.5 ★
Mariscos Ensenada	good, fish, shrimp, tacos, ceviche	29	4.5 ★
El Nopalito	tacos, burrito, food, place, small	49	4 ★
El Nuevo Taquito	carne, asada, tacos, ive, taco	5	4 ★
El Rinconcito Mexican Food	tacos, place, good, food, small	20	4 ★
La Salsita	food, place, mexican, best, tacos	32	4 ★
Mariscos Chihuahua	ceviche, shrimp, place, seafood, tacos	16	4 ★
Rubio's	fish, rubios, tacos, taco, great	56	4 ★
Chico's Tacos	good, place, tacos, food, chicken	89	3 ★

Restaurant

(All)



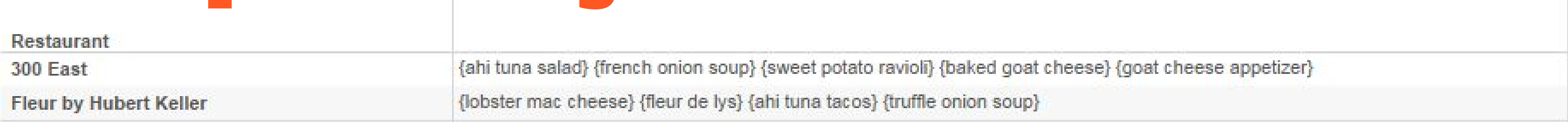
ORDER MAXIMIZER

Restaurant	Search Bigrams	# of ..	
Roma Deli & Restaurant	{las vegas} {chicken parm} {roma deli} {deli restaurant} {authentic italian}	137	4.5 ★
Roma Garden Ristorante	{gluten free} {spaghetti meatballs} {strip mall} {chicken parm} {roma gar..	157	4.5 ★
Romanelli's Deli & Bakery	{new york} {chicken parm} {grocery store} {take home} {first time}	151	4.5 ★
Big Jim's Restaurant & Bar	{wedding soup} {big jims} {chicken parm} {veal parm} {parm sandwich}	116	4 ★
Cherryblossom Noodle Cafe	{strip mall} {green tea} {bento box} {chicken parm} {pad thai}	683	4 ★
Ferraro's Italian Restaurant & Win..	{sea bass} {olive oil} {top notch} {chicken parmesan} {beef carpaccio}	503	4 ★
Chicago Joe's Restaurant	{smith center} {creamy garlic} {chicken parm} {uchicago joes} {las vegas}	186	3.5 ★
Fazoli's	{chicken parm}	47	3.5 ★
Giuseppe's on 28th	{osso bucco} {squash ravioli} {rice balls} {chicken parm} {network sign}	262	3.5 ★

Restaurant: (All)

Search Bigrams: chicken parm

Examples of Trigrams:



Restaurant	
300 East	{ahi tuna salad} {french onion soup} {sweet potato ravioli} {baked goat cheese} {goat cheese appetizer}
Fleur by Hubert Keller	{lobster mac cheese} {fleur de lys} {ahi tuna tacos} {truffle onion soup}

Limitations and Challenges

- Not using Yelp API
    - Rate Limits and return a snippet of reviews; needed all reviews rather than just a handful
- Not using location data to match both Yelp and Twitter
    - Computing challenge: adjacency based on geo-coordinates and accuracy
- Not matching restaurants across data-sources
    - Yelp BusinessIDs are different from Twitter Business IDs, there is no API to get all business IDs on Twitter.
    - Twitter partners with Zagat and OpenTable for restaurant data, but neither of these have open APIs

Future Directions

- Adding
  - More locations
  - Geo-data analysis
  - More data-sources for reviews and restaurants
  - More stopwords (name of restaurant)
- Making sense of
  - More stopwords (name of restaurant)
  - Phrases, sentences and more sophisticated NLP
  - Emojis such as

    🍞 □ 🍖 🍔 🍟 🍕 □ □ □ □ 🍲 🍱 🍘 🍙 🍢 🍣 🍨 🍩 ☕ □ 🍷 🍸 🍹 🍺
  - Japanese kamojis such as

    (˘▽˘)っ　▼█▼□ //　(★‿★)　ℓ ʕɓoçöLåŧë (￣w￣)Ψ　(‥)つ―{}@{}@{}-
- Making most integrations work with live-data
- Making a user-friendly front-end application

Thanks!