

The impact of New York Times on Yelp reviews

Diana Xie

Data Science Career Track
Second Capstone Project

Special thanks to my Springboard mentor:



Tommy Blanchard
Data Science Lead,
Fresenius Medical Care

The problem

- The digital space - crucial these days to restaurants' business, which is already notoriously low-profit margin
- Many platforms exist for broadcasting food opinions that affect this bottom line, both traditional (NYT) and crowd-sourced (Yelp)
- Does a single mention from NYT, or positive early reviews from Yelp, kick off a restaurant's success?
- Does NYT (i.e. an influential food critic) influence the more crowd-sourced Yelp reviews?

The
New York
Times



The project

- Determine the influence of NYT reviews on Yelp reviews for a large dataset of NYC restaurants
- Determine the separate influence of NYT and Yelp reviews on restaurant popularity

Use: time-series analysis, NLP

The data

WHAT

- **NYT reviews:** rating, review text, review date, date of review, etc.
- **Yelp reviews:** average rating, review text, date of review, etc.

HOW

- **Web scraping:** Built two custom web scrapers (NYT and Yelp)

RESULT

- A **restaurant database** of all restaurants in NYC with a NYT review (June 2014 to June 2018)
- A **Yelp review database** of complete Yelp info and review text for each of these restaurants, up to July 2018

	nyt_name	nyt_review_time	isclosed	yelp_rating	reviewcount	yelp_url	yelp_name	critics_pick	nyt_stars	nyt_link
0	Davelle	2018-06-07	False	4.0	47.0	https://www.yelp.com/biz/davelle-new-york-2	Davelle	y	NaN	https://www.nytimes.com/2018/06/07/dining/dave...
1	Lahi	2018-05-31	False	4.0	34.0	https://www.yelp.com/biz/lahi-elmhurst	Lahi	n	NaN	https://www.nytimes.com/2018/05/31/dining/lahi...
2	Don Angie	2018-05-29	False	4.5	126.0	https://www.yelp.com/biz/don-angie-new-york	Don Angie	y	4.0	https://www.nytimes.com/2018/05/29/dining/don...
3	Rangoon Spoon	2018-05-24	NaN	3.5	NaN	https://www.yelp.com/biz/rangoon-spoon-brooklyn	Rangoon Spoon	n	NaN	https://www.nytimes.com/2018/05/24/dining/rang...
4	Wokuni	2018-05-22	False	4.0	55.0	https://www.yelp.com/biz/wokuni-new-york	Wokuni	n	3.0	https://www.nytimes.com/2018/05/24/dining/rang...

Sample of restaurant database.

	nyt_name	nyt_review_time	isclosed	yelp_rating	reviewcount	yelp_url	yelp_name	critics_pick	nyt_stars	nyt_link
0	Davelle	2018-06-07	False	4.0	47.0	https://www.yelp.com/biz/davelle-new-york-2	Davelle	y	NaN	https://www.nytimes.com/2018/06/07/dining/dave...
1	Lahi	2018-05-31	False	4.0	34.0	https://www.yelp.com/biz/lahi-elmhurst	Lahi	n	NaN	https://www.nytimes.com/2018/05/31/dining/lahi...
2	Don Angie	2018-05-29	False	4.5	126.0	https://www.yelp.com/biz/don-angie-new-york	Don Angie	y	4.0	https://www.nytimes.com/2018/05/29/dining/don...
3	Rangoon Spoon	2018-05-24	NaN	3.5	NaN	https://www.yelp.com/biz/rangoon-spoon-brooklyn	Rangoon Spoon	n	NaN	https://www.nytimes.com/2018/05/24/dining/rang...
4	Wokuni	2018-05-22	False	4.0	55.0	https://www.yelp.com/biz/wokuni-new-york	Wokuni	n	3.0	https://www.nytimes.com/2018/05/24/dining/rang...

Sample of Yelp review database.

Yelp & NYT: two different review systems

Yelp review system: 1-5 scale, 1 = worst, 5 = best.



- 1 = "Eek! Methinks not."
- 2 = "Meh. I've experienced better."
- 3 = "A-OK."
- 4 = "Yay! I'm a fan."
- 5 = "Woohoo! As good as it gets!"

- Up to individual user interpretation
- "4.0 is the average" Yelp phenomenon

NYT review system: a two-fold system:

- **Stars**: 1-4 scale. Only chief restaurant critic can reward stars.



- Poor
- Fair
- Satisfactory
- 1 = Good
- 2 = Very Good
- 3 = Excellent
- 4 = Extraordinary



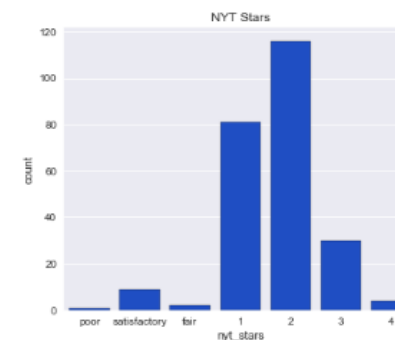
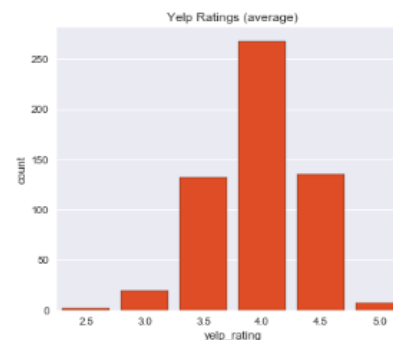
- **"Critic's Pick"**: a seal of approval by any other NYT critic
Yes/no

- Good: Any star (even 1) indicates approval from the chief critic and is consisted of merit
- Bad: "Poor", "Fair", or "Satisfactory"
- Either/both: Some restaurants may only have stars, while others may have both a star and a "NYT Critic's Pick"

Investigating review systems

Because the Yelp and NYT rating systems are different, we explored what was considered a "good" or "bad" rating between them

- Yelp ratings generally align with NYT
- Restaurants at the highest tier of critical acclaim always have strong Yelp ratings - 4-NYT star restaurants never < Yelp 4.5



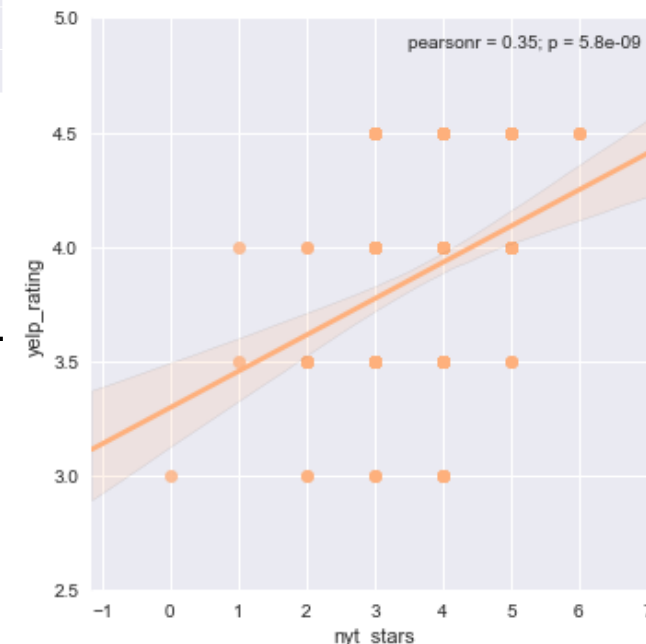
Yelp (average per restaurant) and NYT star rating distributions

Average Yelp rating (received NYT Critics Pick): 4.0
Average Yelp rating (didn't receive NYT Critics Pick): 3.918181818181818



Average Yelp ratings for NYT Critic's Pick (didn't vs. did receive) restaurants.

NYT stars vs. Yelp stars.



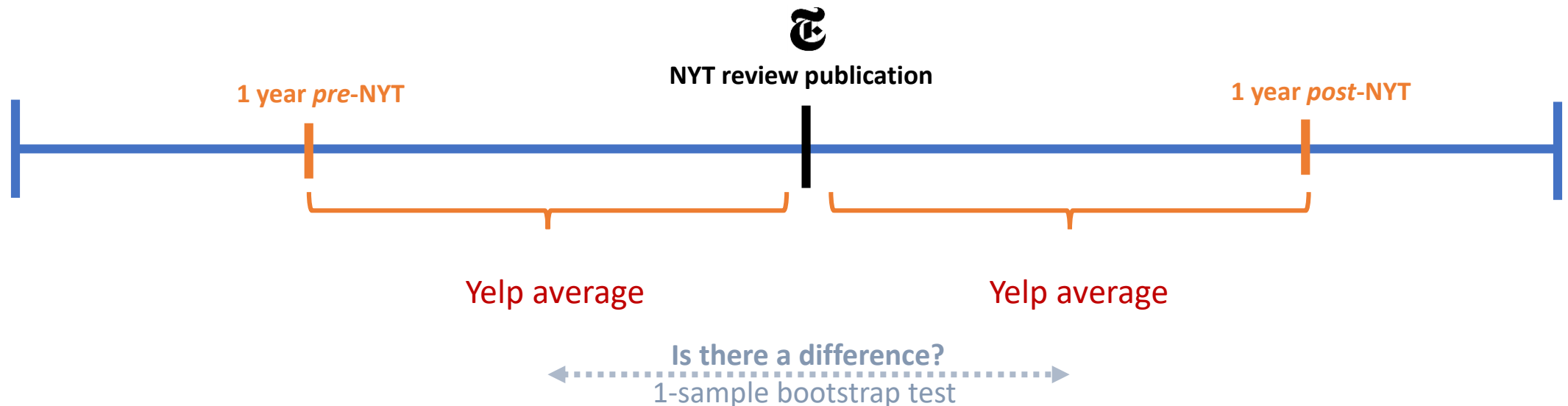
Investigating Week-by-week time-series analysis



- Random case study, “Don Angie” restaurant: 130 reviews, 2/4 NYT stars, opened Nov 2017
- Lots of weekly variation – clearly variations not significantly attributable to NYT review of restaurant
- Weekly analysis insufficient to determine impact of NYT review

Bootstrapping: Pre- vs. post- NYT

- Take a single time window *before* & a single window *after* a restaurant's NYT review is published
- Within these two windows, take the average Yelp rating
- Finally, conduct a bootstrap test to determine whether the change between these windows is significant



Bootstrapping:

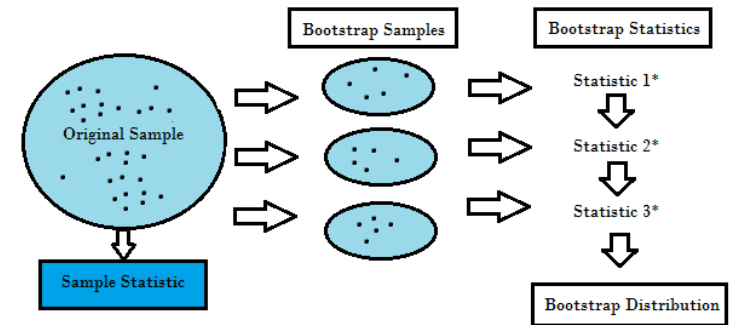
Experimenting w/ different time windows

APPROACH

1. **Generate a bootstrap replicate.** Sample 1000x from a restaurant's "population" of reviews. Take mean = sample mean "Yelp rating".
2. **Repeat #1, 1000x.** Generate 1000 bootstrap replicates = 1000 sample means.
3. **Calculate p-value.** Of actual post-NYT Yelp average, i.e. "How likely do I observe this average, given the null assumption that there is no difference between pre- & post-NYT Yelp reviews?"

VARIATIONS

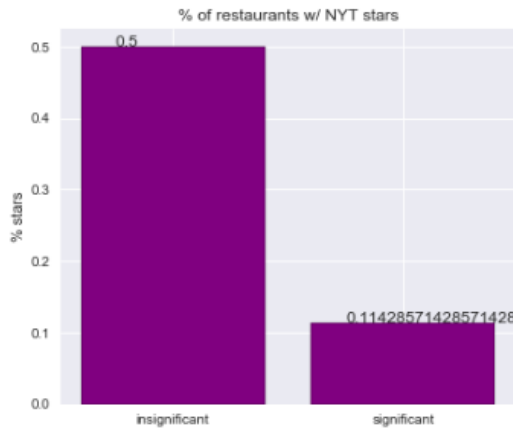
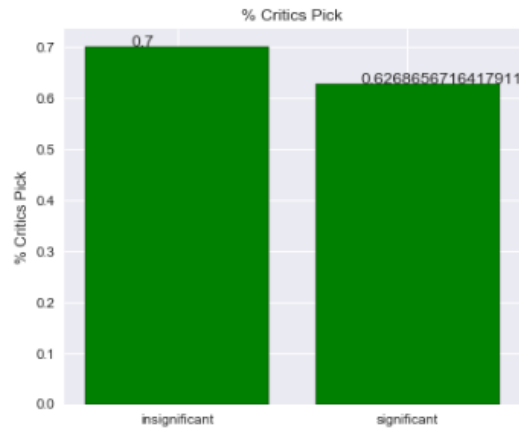
- All data up to 6 months post-NYT.
- Only data beginning with NYT review, up to 6 months post-NYT.
- No time window restriction.



Bootstrapping "how-to"

<http://www.statisticshowto.com/bootstrap-sample/>

Bootstrapping: Results



Comparisons between “insignificant” & “significant” restaurants for NYT data.



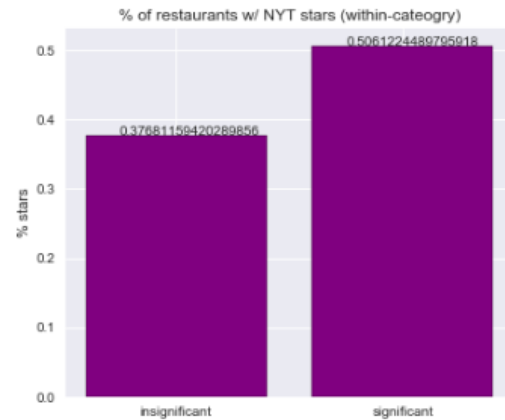
% of restaurants receiving “boost” in post-NYT Yelp ratings

- Counterintuitively, “significant” restaurants have lower % “Critic’s Pick” and NYT stars.
- Also counterintuitively, among restaurants receiving a good NYT review, a lower % of them had a post-NYT increase in Yelp ratings compared with “bad” NYT-reviewed restaurants.

CONCLUSION: Try another approach to comparing pre- vs. post-NYT average Yelp rating

1-year window without bootstrapping

- The bootstrapping method for identifying “significant”/“insignificant” restaurants yielded lackluster results
- Here, simply compare 1-year pre- and post-NYT Yelp average (no bootstrapping)
- Cutoff for “significant” restaurant: 0.5-fold increase or decrease to a restaurant’s average Yelp rating



“Significant” vs. “insignificant” % of NYT stars.



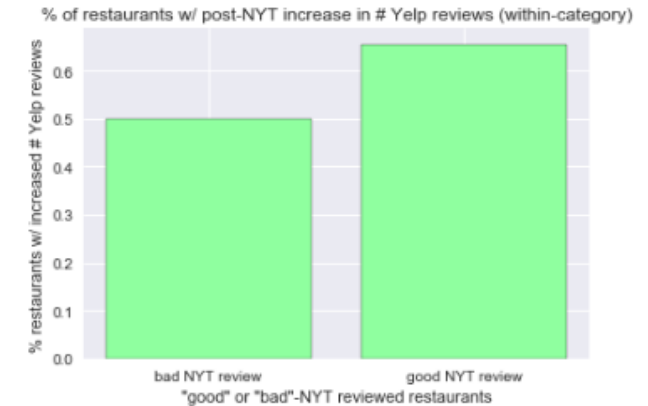
% of restaurants receiving “boost” in post-NYT Yelp ratings

CONCLUSION: Still counterintuitive findings.

We turn to another metric for comparing pre- & post-NYT Yelp reviews.

Impact of NYT reviews on restaurant popularity

- Up until now, only examined Yelp ratings.
- Now, measure impact of NYT reviews on restaurant popularity
- Popularity = # of reviews written for a restaurant on Yelp



% restaurants w/ increased num. Yelp reviews.

CONCLUSION: Results are much more promising. Restaurants that received a “good” NYT review experienced *more* Yelp reviews, post-NYT.

Determining NYT influence on Yelp review *text*

Now, turn our attention to review *text*

Hypothesis: a NYT mention of specific dishes may make it more likely that it will be mentioned in subsequent Yelp reviews (i.e. ordered more often)

Approach:

- Generate NLP corpus of reviews
- Perform tf-idf
- Determine NYT-Yelp word intersections
- Compare pre- vs. post-NYT

Review (after pre-processing): we went this past weekend on saturday and was really surprised at how well organized it all seemed. we got there around 6, easily found parking and didn't have to stand in a line for more than 10 minutes at each food stall. i think it's great that queens was the first borough to get a night market going, would definitely recommend stopping by when they are back in july.


Review (after document tokenization, removing stopwords, lemmatization): ['went', 'past', 'weekend', 'saturday', 'really', 'surprised', 'well', 'organized', 'seemed', 'got', 'around', 'easily', 'found', 'parking', 'stand', 'line', 'minute', 'food', 'stall', 'think', 'great', 'queen', 'first', 'borough', 'get', 'night', 'market', 'going', 'would', 'definitely', 'recommend', 'stopping', 'back', 'july']

Review (after gensim corpus): [(40, 1), (44, 1), (123, 1), (131, 1), (176, 1), (178, 1), (213, 1), (243, 1), (312, 1), (313, 1), (320, 1), (343, 1), (365, 1), (386, 1), (424, 1), (465, 1), (564, 1), (568, 1), (579, 1), (594, 1), (613, 1), (659, 1), (701, 1), (952, 1), (957, 1), (1147, 1), (1172, 1), (1239, 1), (1544, 1), (3371, 1), (4765, 1), (5798, 1), (6505, 1), (6543, 1)]

Visualizing a sample review under our different processing steps leading up to gensim corpus.

Yelp reviews - NLP corpus

Generate a pre-processed, tokenized list of documents

- Lowercase
 - Remove non-alphabetic characters/punctuation
 - Remove stop words
 - Lemmatize
 - Misspellings
- 

Also experimented **TextBlob** library

- Corrects misspellings that were responsible for duplicate words (i.e. confounded word frequency)
- However, added ~20 hours to execute, so remains an optional part of code

- Use [gensim](#) to create a corpus
- Create bag of words: each token mapped to a unique numeric ID and word count

tf-idf EDA

Experiment with gensim's **tf-idf** to identify important (hopefully dish-specific) words in each document

- Tf-idf weights “important” words
- Common, shared words between documents are down-weighted – aren’t meaningful key words
- Conversely, document-specific words are weighted highly

```
tfidf weights:  
[(0, 0.1055642607631973), (1, 0.05957501646551434), (2, 0.05994039811573356), (3, 0.020509934091441764), (4, 0.028228376627908995)]
```

```
Top 5 weighted words:  
oden 0.38078435163064966  
dashi 0.30155229186255794  
uh 0.2365325200743029  
spaghetti 0.20231486911864288  
mentaiko 0.18664453616596044
```

```
Text:  
davelle uh oden uh foodie trippin get order right uh shawty look good eatin oden oden dish drink dashi davelle oden moonlight  
xxxxtentacion rip everything amazing u dining tiny cozy cramped beautiful little spot got oden set karaage cod spaghetti hokkaid  
o spaghetti uni tomato cold dish topped kinda optional light cheese drink dashi aaaalllll dish good soft blanched skinless savo  
ry daikon served spicy yuzu paste use sparingly pretty big kick red miso paste soft mushy perfectly cooked heart shaped daikon  
mochi lightly fried bag soft gooey delicious mochi def drink dashi scallion enoki mushroom ginger hanpen white fish cake soft t  
exture airy typical fish cake denseness fishcake delicious served spicy yuzu paste sausage served japanese mustard yummy bamboo  
shoot cooked enough left slight hate mushy bamboo shoot dashi similar taste mochi karaage soft juicy chicken lightly battered m  
edium crisp cod spaghetti aka mentaiko pasta light cod roe taste fishy delicious hokkaido style spaghetti uni delicious much un  
i guess maybe another variation mentaiko liquor license sake soju cocktail beer wine went tonight quiet small space sure peak t  
ime usually wait
```

CONCLUSION: tf-idf is good at identifying dish-specific keywords

Visualizing a tf-idf results. For a single Yelp review, “Davelle” restaurant.

Determine NYT-Yelp word intersection (i.e. shared words)

Goal: To determine whether a NYT review affected Yelp review text

Metric: Frequency intersecting terms between the NYT review and Yelp reviews, pre- vs. post-NYT review publication.

Approach: 1-year time window (from previous analysis)

Tf-idf:

- Top 20 tf-idf weighted pre- & post-NYT Yelp reviews' text
- Top 100 tf-idf weighted NYT review words

Davelle:

```
pre-NYT shared terms: {'oden'}  
post-NYT shared terms: {'daikon', 'uni', 'spaghetti', 'urchin', 'dashi', 'hokkaido', 'mentaiko', 'oden'}
```

Rangoon Spoon:

```
pre-NYT shared terms: {'noodle', 'burmese'}  
post-NYT shared terms: {'spoon', 'tofu', 'rangoon', 'thoke', 'burmese'}
```

2 restaurants' pre- vs. post-NYT intersecting terms

Initial examination

- Calculation for **n-fold change**: (difference)/(# of pre-NYT intersecting terms)
- Non-normal distribution – so we set a cutoff of 0.5-fold change to be a “significant” restaurant change
- Of note: never >1-fold decrease for restaurants’ intersecting NYT terms
- NYT reviews never significantly damaged tendency to order certain dishes

Of restaurants with ≥ 0.5 post-NYT increase in NYT-Yelp intersections:

- ~76% had a "good" NYT review (compared w/ ~70% for “bad”-reviewed)

```
% restaurants w/ 'good' ratings (Critic's Pick and/or NYT stars)
>=1-fold increase, post-NYT:  0.7681159420289855
<1-fold increase, post-NYT:  0.6979591836734694
```

```
Number of restaurants w/ "good" ratings:  69
```

```
Number of "good"-NYT reviewed restaurants:  224
Number of "bad"-NYT reviewed restaurants:  60
```

```
Number of >=0.5-fold increased restaurants:  69
Number of <0.5-fold increased restaurants:  245
```

Permutation test – checking results

Not confident that there's really a difference. So check with a *permutation test*.

- Non-parametric test similar to boot-strapping

Control group: Bad/neutral NYT reviewed restaurants (“Poor”, “Fair”, “Satisfactory”, or no “Critic’s Pick” or stars)

- No change expected in post-NYT NYT-Yelp shared words

Experimental group: “Good” NYT-reviewed restaurants: “Critic’s Pick” and/or NYT stars

Null hypothesis: No difference between “control” & “experimental”; no change in intersecting NYT-Yelp terms

Results:

Control vs. experimental group difference is small: 0.105
(p-value = 0.25) – *can't* reject null hypothesis

CONCLUSION: Insufficient evidence that a positive or neutral/negative NYT review would have an impact on the text of Yelp reviews

Summary

Project aim: Determine the influence of NYT reviews on Yelp reviews, to better understand how 2 major platforms for food reviews/information interact w/ each other and impact restaurants

Conclusion

- There are handful (or more) of cases in which NYT reviews clearly impacted Yelp reviews (in average ratings or popularity)
- However, there are obviously many other factors that can affect restaurants and complicate analyses
- Although project results mixed, generated several materials in the process

Materials produced

- Method of reconciling different restaurant rating systems
- Custom web scrapers to generate 2 valuable databases of NYC restaurant reviews/info
- Method of analyzing how one rating system affect the other (NYT on Yelp) – ratings, restaurant popularity
- Method for leveraging NLP for even further investigations - dish-specific effects