1 Introduction

Background

The digital space is crucial to bringing in business for restaurants, and many platforms exist for broadcasting food opinions that affect this bottom line. These platforms are so diverse and numerous that it may seem impossible to disentangle whether a single mention by a large news organization, such as the New York Times, or positive early reviews by customers on Yelp, are responsible for kicking off a restaurant's success.

Profit margins are already notoriously slim and closures a regular occurrence in the competitive and stressful restaurant industry. Making headway towards determining how different media platforms affect restaurant popularity and reviews could allow businesses to better adapt to the changing ways with which people consume, share, and decide where to spend their dollars in the food industry.

Project

In this project, I compare the influence of two major sources of restaurant reviews -- 1) the New York Times (NYT) dining reviews and 2) Yelp.com -- on New York City (NYC) restaurants from 2014 to 2018. In addition to tracking how reviews from each platform individually affect the restaurant, I also determine how reviews from one platform (NYT) affect the other (Yelp) by performing time-series analyses. For instance, will a restaurant awarded a perfect set of stars and "NYT Critics Pick" subsequently see an increase in positive Yelp reviews?

This work is enriched with additional NLP analyses to determine whether certain dishes or other aspects of a restaurant (ex. service), when mentioned in NYT, affect the text of subsequent Yelp reviews. In the process, an additional dimension of information is added which recognizes that changes to restaurants may not necessarily be reflected in ratings, but also in what customers tend to order or attend to.

2 Data Acquisition & Cleaning

Code: https://github.com/diana-xie/yelp-nyt-NLP/blob/master/data_wrangling.ipynb

This project focuses on restaurants in NYC with a NYT review dating between June 2014 to June 2018, as well as an active Yelp business page with Yelp reviews. Yelp data was obtained through a combination of the Yelp API and our own Yelp web scraper, and NYT data was also obtained through another web scraper that I built to scrape their dining section.

Although Yelp was founded in 2005, there were several turning points in its timeline that may confound its impact on local businesses. In its early years, the number of users and reviews experienced rapid growth. For instance, in 2006 Yelp only had ~1 million unique monthly users and 100,000 reviews. In 2008, when the Yelp mobile app was released, this statistic was up to ~15.8 million unique monthly users (and from 12 to 24 cities, compared to 2007). Finally, in 2012 Yelp's stock began publicly trading - this is the year I originally intended to start with. However, our Yelp scraper ran into issues and I were only able to obtain reviews starting from June 2014.

2.1 Extracting NYT dataset

Basic restaurant info (Scraper #1)

To first form the database of restaurants, I used NYT's dining section. From here, I imported XML (all the way down to 2014) and used BeautifulSoup to extract basic information on each restaurant, such as their name and NYT rating.

Then, I manually appended NYT ratings (i.e. "stars" and/or "Critic's Pick). Given the complicated XML layout for displaying its two-part rating results, I opted to manually enter this information into our database, rather than automating its collection. This decision was made to ensure data collection accuracy, as well based on the fact that I only had 334 reviews to enter - thus, the process was relatively quick.

NYT review text (Scraper #2)

Then, I coded a web scraper to extract the actual text of each NYT review. Since some NYT dining reviews did not have the same format as the first iteration of our scraper, I expanded our scraper to accommodate two different formats of NYT dining review webpages.

2.2 Extracting Yelp dataset

Yelp API

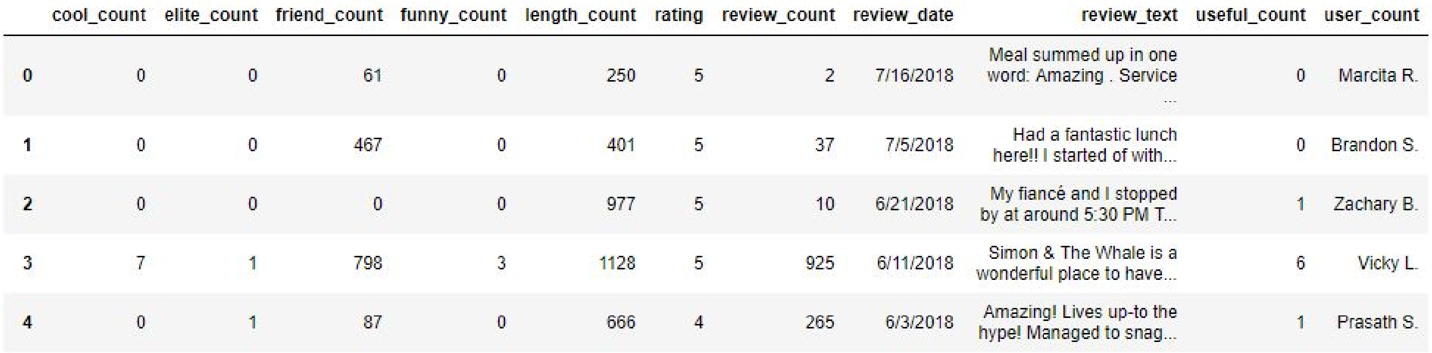
I initially intended to use Yelp's well-known academic dataset. However, it only contains a comprehensive set of reviews for a handful of cities, and NYC is not significantly represented. Thus, I started with Yelp's API, which I accessed using the Python library YelpAPI. Accessing Yelp's API requires an API key, which was obtained by registering our project as an app here.

Using our generated list of NYT restaurants, I conducted a search query to Yelp's API to match with the restaurant's Yelp information. This information was cross-checked with the NYT name to ensure that the API's top-result Yelp restaurant was indeed the restaurant reviewed by the NYT.

The most valuable information I were able to obtain was the restaurant's Yelp link and whether it was closed. However, Yelp's API does not actually provide review information (except a few snippets), such as review text and rating. As a result, I coded our own web scraper to extract individual review text/information per restaurant (next section).

Yelp web scraping (Scraper #3)

Since review information was not provided by Yelp's API, I coded our own web scraper to scrape review text, rating, and other associated information (including user metrics) (**Fig 2**). Our web scraping algorithm was adapted from Parkhar Thapak's code here, which was originally designed to scrape peripheral review information such as "Funny"/"Useful" votes, number of reviews the user had written, whether the user was "Yelp Elite" status, etc. I added in extraction for review text, date review was written, and rating assigned by Yelp user.

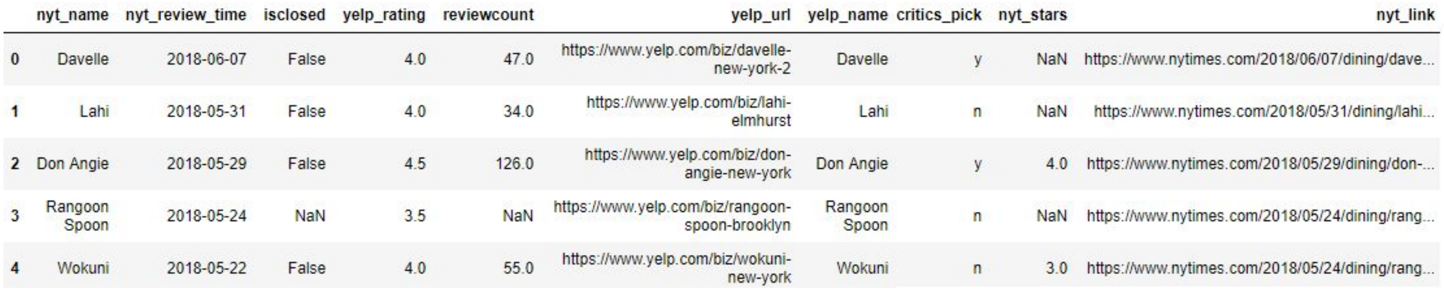


	cool_count	elite_count	friend_count	funny_count	length_count	rating	review_count	review_date	review_text	useful_count	user_count
0	0	0	61	0	250	5	2	7/16/2018	Meal summed up in one word: Amazing . Service ...	0	Marcita R.
1	0	0	467	0	401	5	37	7/5/2018	Had a fantastic lunch here!! I started of with...	0	Brandon S.
2	0	0	0	0	977	5	10	6/21/2018	My fiancé and I stopped by at around 5:30 PM T...	1	Zachary B.
3	7	1	798	3	1128	5	925	6/11/2018	Simon & The Whale is a wonderful place to have...	6	Vicky L.
4	0	1	87	0	666	4	265	6/3/2018	Amazing! Lives up-to the hype! Managed to snag...	1	Prasath S.

Fig 2. Sample of first five Yelp review extractions for a restaurant. Text is stored in column 'review_text'. Other useful metrics include information about the Yelp user who posted the review (ex. 'review_count', 'friend_count', 'elite_count').

2.3 Summary

In this section, I coded a series of web scrapers, combined with Yelp's API, to produce our database of restaurants and all relevant data for our forthcoming analyses (**Fig 3**). In the process, I took many steps to ensure that data was complete, missing data was filled, and the restaurants were actually located in NYC.



	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...

Fig 3. Sample of restaurant database. Each row represents a restaurant's data. Review text (both NYT & Yelp) are stored in separate databases.

3 Data Exploration

Code: https://github.com/diana-xie/yelp-nyt-NLP/blob/master/data_wrangling.ipynb

To better determine the effect that NYT and Yelp reviews might have on our restaurants, the bulk of our data exploration centered on determining what each rating system meant.

3.1 Discussion of Yelp & NYT: two different review systems

1. Yelp review system: Yelp's rating system operates from 1-5, lowest to highest. A restaurant's average rating is taken simply to be the average of all Yelp users' individual review ratings (i.e. no extra weight put on whether a Yelper is "Elite", etc.). This is the official documentation for what each rating means:

- 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!"

However, the user is free to interpret Yelp's rating system as they'd like and therefore is subjected to less control for bias/variance (both between users and within), compared to a news platform such as NYT that employs full-time critics and relies on the perception of consistency. Several interesting phenomena occur as a result, such as "Yelp 4.0 being the average".

2. NYT review system: NYT's review system for restaurants is actually two-fold and consists of two sub-ratings:

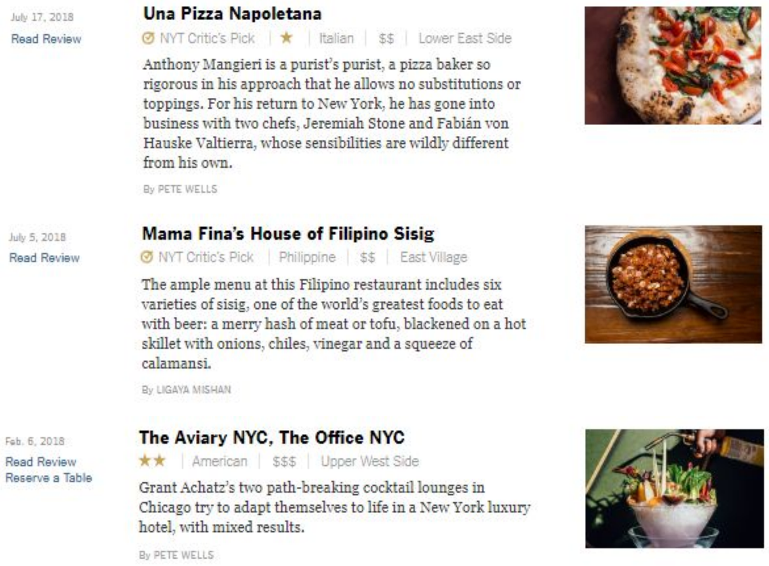
A. Stars: A 1-4 rating system, where only the chief restaurant critic (Pete Wells) can assign star ratings.

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

The star system may be somewhat counterintuitive, as any star (even 1) indicates approval from the chief critic and is consisted of merit. As a result, there are 3 additional ratings that precede stars in this system (in order from least to most approval): "Poor", "Fair", or "Satisfactory".

B. "NYT Critic's Pick" - a seal of approval by any other NYT critic. This rating is binary - either a restaurant receives "NYT Critic's Pick" or does not receive it, and no numeric scale is given for comparison between each instance of "NYT Critic's Pick".

Some restaurants may only have stars, while others may have both a star and a “NYT Critic’s Pick” (Fig 2). Often, a restaurant will simply have “NYT Critic’s Pick” but no stars, as the restaurant was not evaluated by NYT’s chief restaurant critic (**Fig 4**).



July 17, 2018
Read Review

Una Pizza Napoletana

NYT Critic's Pick | ★ | Italian | $$ | Lower East Side

Anthony Mangieri is a purist’s purist, a pizza baker so rigorous in his approach that he allows no substitutions or toppings. For his return to New York, he has gone into business with two chefs, Jeremiah Stone and Fabián von Hauske Valtierra, whose sensibilities are wildly different from his own.

By PETE WELLS

July 5, 2018
Read Review

Mama Fina’s House of Filipino Sisig

NYT Critic's Pick | Philippine | $$ | East Village

The ample menu at this Filipino restaurant includes six varieties of sisig, one of the world’s greatest foods to eat with beer: a merry hash of meat or tofu, blackened on a hot skillet with onions, chiles, vinegar and a squeeze of calamansi.

By LIGAYA MISHAN

Feb. 6, 2018
Read Review
Reserve a Table

The Aviary NYC, The Office NYC

★★ | American | $$$ | Upper West Side

Grant Achatz’s two path-breaking cocktail lounges in Chicago try to adapt themselves to life in a New York luxury hotel, with mixed results.

By PETE WELLS

Fig 4. NYT’s dual sub-rating system. Different combinations of whether stars were awarded and whether a “NYT Critic’s Pick” label was awarded.

<u>Conclusion:</u> Because the Yelp and NYT rating systems are so different from each other, we'll need to calibrate what is considered a "good" or "bad" rating between Yelp and NYT for purposes of determining their impact on a restaurant (and how NYT affects Yelp ratings).

3.2 Yelp & NYT rating distributions

To get a better idea of what constitutes relatively "good"/"bad" ratings within and between NYT and Yelp, I first take a look at their rating distributions. Similarly, I also examine the "NYT Critic's Pick" label to determine how it is assigned (**Fig 6**).

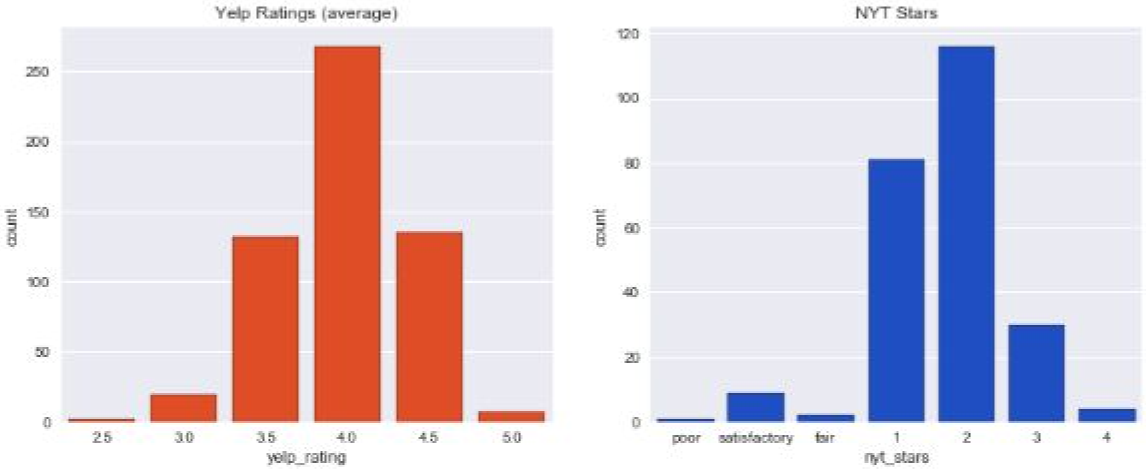


Fig 5. Yelp (average per restaurant) and NYT star rating distributions. Yelp’s average ratings are normally distributed. NYT’s less so, due to the lack of bad reviews under their star system.

There's a fairly normal distribution to the Yelp average ratings, a good sign for avoiding a skewed dataset that would indicate bias in how Yelp users review restaurants overall. It also seems that I should set our threshold for a “good” Yelp rating to 4.5, rather than 4.0 - in agreement with our earlier discussion of “Yelp 4.0 being the average”.

With NYT, the number of restaurants with awarded stars is expected be to much lower, since stars can only be awarded by NYT's chief restaurant critic and any star awarded by the NYT is considered prestigious.

3.3 NYT stars vs. Yelp stars

We'd like to see if there's a relationship between Yelp and NYT stars. Our hypothesis is that average Yelp ratings will generally align with NYT ratings, with less variability at the extremes ("Poor"/"Fair" or 3-4 NYT stars).

Recall that with NYT star rating system, at least 1 star awarded is already deemed as a positive review by a prestigious food critic. Since I have converted the stars to accommodate string-ratings such as "Poor" (0), "Fair" (1), and "Satisfactory" (2), our actual stars begin with nyt_stars = 3 (which would be 1 star) (**Fig 6**).

- **"Poor" & "Fair"**: There is a very small sample size (1 "Poor", 2 "Fair"). However they both correspond to at least 1 restaurant w/ a poor Yelp rating (3.5 or 3.0).
- **"Satisfactory"**: The range widens, but there are no restaurants above a 4.0 Yelp average.
- **1-2 NYT stars**: With the onset of NYT stars, the range widens to include 4.5 Yelp averages.
- **3 NYT stars**: There are no longer any restaurants w/ Yelp averages < 3.5.
- **4 NYT stars**: No restaurant drops below a 4.5 Yelp average.

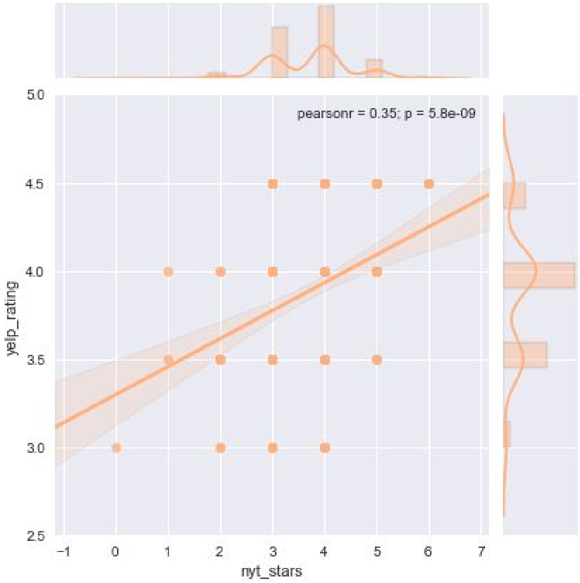
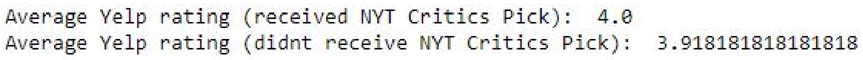


Fig 6. NYT stars vs. Yelp stars. Each data point is Yelp and NYT stars for a restaurant with both. # restaurants w/ “Poor”: 1, “Fair”: 2, “Satisfactory”: 10, 1-star: 86, 2-star: 122, 3-star: 31, 4-star: 4.

<u>**Conclusion**</u>: Our results support our hypothesis that Yelp average ratings generally align with NYT evaluations. It is particularly notable that restaurants with 4 NYT stars never has an average Yelp rating below 4.5 - restaurants at the highest tier of critical acclaim likely have strong Yelp ratings.

3.4 NYT "Critic's Pick" vs. Yelp averages

In the previous section, I examined what NYT's star rating system meant by pegging it to Yelp stars. Now I similarly disentangle the meaning behind NYT's "Critic's Pick" label. To do so, I separate restaurants by whether they received NYT’s “Critic’s Pick” or did not receive. Then, for each group I determined the average Yelp rating (**Fig 7**).



```
Average Yelp rating (received NYT Critics Pick):  4.0
Average Yelp rating (didnt receive NYT Critics Pick):  3.918181818181818
```

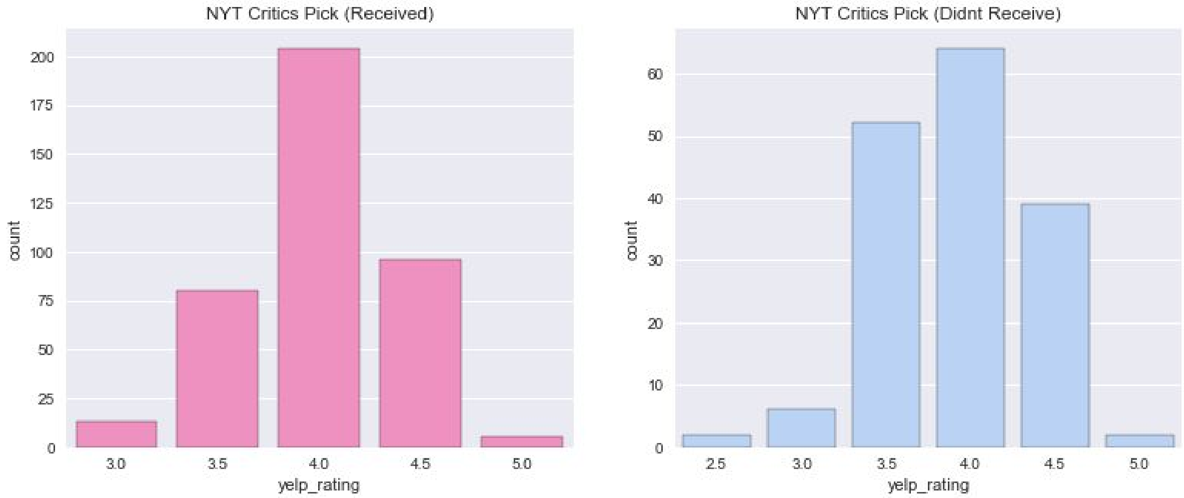


Fig 7. Average Yelp ratings for NYT Critic’s Pick (didn’t vs. did receive) restaurants. The difference (in average Yelp review) between restaurants receiving or not receiving NYT’s “Critic’s Pick” is insignificant.

The difference in average Yelp rating between restaurants receiving and not receiving “Critic’s Pick” (4.0 vs. 3.92) is so small that it’s likely insignificant. Additionally, each group’s Yelp average distribution is normal - clearly there’s variability and unpredictability in whether a NYT “Critic’s Pick” necessarily indicates that the restaurant will have a high Yelp rating.

Conclusion: Since the Yelp average between restaurants receiving and not receiving “Critic’s Pick” is marginal, it probably isn't useful to use "NYT Critic's Pick" in our time-series analysis of whether NYT impacts subsequent Yelp reviews (**Fig 8**).

3.5 - Summary

In this section, I had two aims: 1) do basic data exploration, and 2) determine the relationship between NYT and Yelp’s different rating systems. For the first aim, I explored Yelp and NYT rating distributions and confirmed their differences. For the second aim, I found that NYT’s “Critic’s Pick” did not have a meaningful relationship with Yelp ratings, but that NYT’s star rating system might be a more useful metric when exploring the impact of NYT reviews on subsequent Yelp reviews.

Un-pictured analyses:

- # of Yelp reviews written for restaurant - fairly normal distribution; good representation of restaurants of all review levels
- Yelp rating vs. # of Yelp reviews - # of Yelp reviews does not bias restaurant’s average Yelp rating
- Average Yelp rating for closed restaurants - surprisingly normal distribution; there are certainly other factors involved in why businesses closed
- # of NYT dining reviews published online, per year - # of NYT restaurant reviews fairly consistent after 2011; re-assuring that there was no one year in which NYT reviews may receive less attention due to “too many other reviews”

4 Determining NYT influence on Yelp reviews

Code: https://github.com/diana-xie/yelp-nyt-NLP/blob/master/data_NYT_influence.ipynb

After doing basic data wrangling and EDA, I begin tracking methods for determining whether the onset of NYT review has affected subsequent Yelp ratings.

4.1 - Week-by-week time-series analysis

Binning by week

I binned each restaurant's Yelp reviews by week and took the average Yelp rating and number of reviews posted to Yelp. Our hypothesis was that in the weeks following a positive NYT review, the weekly average Yelp rating would increase.

Our rationale for binning by week was to take into account the difficulty of changing an average Yelp rating. Because a restaurant’s Yelp average is the average of all its ratings through its history, it would take many ratings over time to detect a significant change in average rating. Restaurants with many prior ratings would be especially biased against, in terms of discovering ratings boosts, as it would take even more new ratings to detect any change to their Yelp averages.

As a result, I averaged all new Yelp reviews occurring each week so that each week would have its own average - such a measure would be much more sensitive to the sudden onset of an event that could at least temporarily affect Yelp ratings, such as a positive (or negative) NYT review.

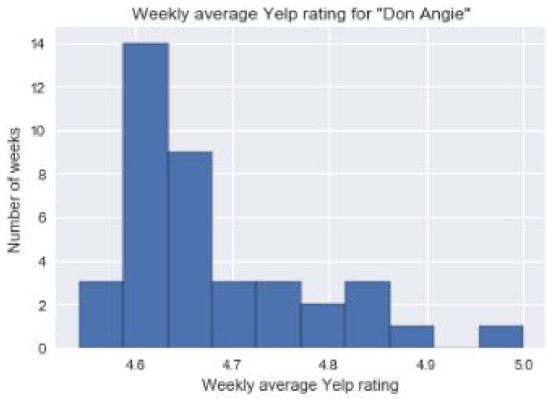


Fig 8. Weekly average Yelp rating. For a single restaurant, “Don Angie”. In most weeks, average Yelp rating was around 4.6.

A case study

Since the analysis was complicated, I examined a case study of a random restaurant with 130 reviews and a strong NYT star rating (2/4 stars) to see our results. The restaurant I picked was “Don Angie”, a restaurant opened in Nov 2017, reviewed by NYT in May 2018, and totalling 130 Yelp reviews in June 2018.

Distribution of weekly average Yelp ratings

A plot of our weekly average Yelp ratings for “Don Angie” shows that the frequencies per each weekly rating that I observe are not normally distributed (**Fig 8**). If the weekly averages were normally distributed, I would be able to determine outliers that would indicate weeks where some event could have caused a sudden increase in positive or negative reviews.

Timeline of weekly average Yelp ratings (timeseries plot)

In the previous section, where I generated a histogram, chronological information was not present. Here, I visualize the timeline of weekly average Yelp ratings. Doing so might allow us to better visualize any salient changes to the Yelp average.

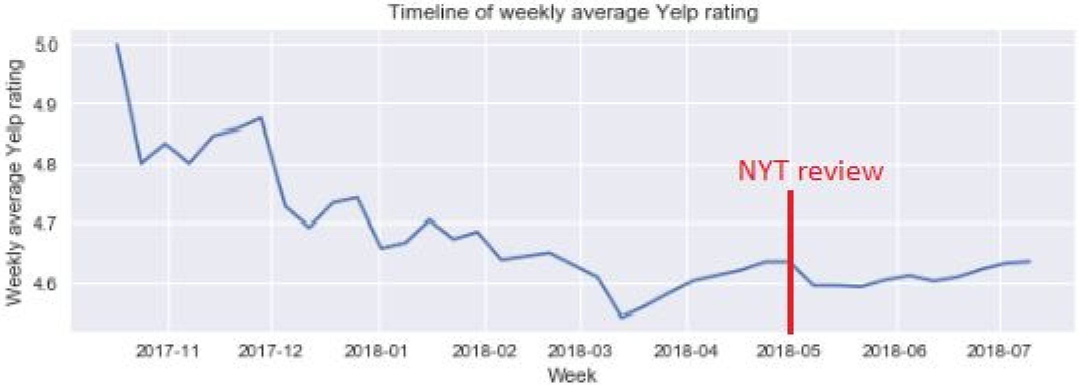


Fig 9. Timeline of weekly average Yelp rating. X-axis: each week. Y-axis: the average Yelp rating for any new reviews posted that week. Single restaurant: “Don Angie”. NYT review: 5/29/2018.

What I are looking for is a noticeable change in the weekly average following the restaurant's NYT review publication (specifically, change in "slope" between points). Here, it is 5-29-2018. There doesn't appear to be much of a change around that time - this would indicate the weekly average Yelp rating didn't change much after its NYT review was published (**Fig 9**).

<u>Conclusion</u>: Although we've only presented one case study here, it's clear that there is a lot of variation in weekly average Yelp rating that is not attributable to NYT review publication (see time-series plot). Therefore, it would be difficult to confidently determine whether a NYT review produced a change in weekly Yelp averages just based on week-by-week change in Yelp reviews. As a result, we'll shelve this form of analysis for now.

4.2 - Examine critical time window before/after NYT review

Since I did not have much luck with a week-by-week time-series analysis, I take a single larger time window before and after a restaurant’s NYT review is published. Within these two windows, I take the average Yelp rating and conduct a bootstrap test to determine whether the change is significant.

The idea behind restricting pre- vs. post-NYT Yelp rating comparisons to time windows is that any effects that might be caused by a NYT review might wash out after some time. For instance, if I don't restrict the time window after a NYT review, there could be years' worth of Yelp reviews after a NYT review is published, where such subsequent Yelp reviews may have nothing to do with the NYT review. Thus, I experimented with different lengths of time windows and eventually settled on 6 months after doing EDA on individual restaurants.

4.3 - Bootstrapping approach

I are interested in determining whether a significant post-NYT change occurred to the average Yelp ratings of restaurants. Since there is no “control” group, I perform a bootstrap analysis on each restaurant. For each restaurant, I apply our bootstrap function to obtain a p-value, allowing us to determine whether our observations support a NYT review affecting the restaurant’s subsequent Yelp ratings.

Bootstrap function

1. **Generate a bootstrap replicate.** First, I use bootstrap sampling (here, 1000 times) from a restaurant's "population" of reviews. Then I take the average of these 1000 randomly selected reviews to obtain a a simulated Yelp average.
2. **Repeat 1000x.** I repeat #1, also 1000 times, to form a null distribution of consisting of 1000 sample means = 1000 Yelp averages. By drawing randomly from the restaurant's 6-month window reviews, I form a distribution in which the reviews are randomly included irrespective of whether they occurred before or after a NYT review. Hence, a "null hypothesis" distribution where the population is under the assumption that there is no difference between the "before"/"after" NYT reviews.
3. **Calculate p-value.** I then calculate p-value of our actual observed post-NYT Yelp average (again, for the restaurant in question) and determine how likely I would observe this value under the null hypothesis.

Variations on bootstrapping data

Although our bootstrap function was the same in all cases, I experimented with three variations on how to obtain bootstrapped data itself:

1. **All data up to 6 months post-NYT**. Our actual sample statistic (that I calculate p-value on) is the restaurant's overall Yelp average, 6 months after the NYT review. That means this average takes into account all reviews in the restaurant's history, leading up to the end of 6 months post-NYT.
2. **Only data beginning with NYT review, up to 6 months post-NYT.** Our actual sample statistic of the restaurant's post-NYT Yelp average is calculated only on Yelp reviews posted ***after*** the NYT review was published.
3. **No time window restriction**. Ignore time window and take as our actual sample statistic the Yelp average of all reviews post-NYT. In other words, no 6-month restriction, but starting from the NYT review date and up to the most recent Yelp review, even if it was as recent as 2018.

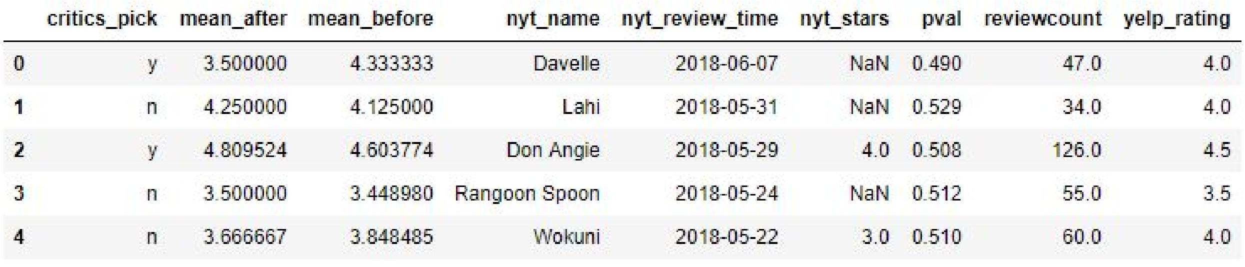
From each version, I gathered what were "significant" restaurants - i.e. restaurants with a statistically significant change (α = 0.05) from pre- to post-NYT Yelp average, according to the bootstrapping algorithm used. However, I recognized that the restaurants classified as "significant" could change depending on what I took as the test statistic - hence, our 3 variations of bootstrapping.

I found that #1 may potentially wash out possible post-NYT changes in Yelp reviews, since the test statistic, in addition to incorporating 6-months of post-NYT reviews, would also incorporate Yelp reviews written *before* the NYT review was published. #2 was much more effective at obtaining "significant" restaurants. However, the p-values were suspiciously low, so I followed up with #3. The same restaurants remained statistically significant, but I reverted back to #2, as the number of Yelp reviews pre- and post-NYT could be dramatically asymmetrical and I should rely on a set pre- and post-NYT time window to at least attempt to control for volume of reviews.

In conclusion, I decided on #2 for the bootstrapping approach going forward.

4.3 - Bootstrapping algorithms

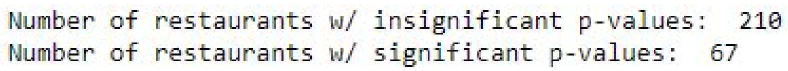
After conducting EDA to steer us towards the appropriate bootstrapping process, I automated the bootstrapping function to all restaurants.



	critics_pick	mean_after	mean_before	nyt_name	nyt_review_time	nyt_stars	pval	reviewcount	yelp_rating
0	y	3.500000	4.333333	Davelle	2018-06-07	NaN	0.490	47.0	4.0
1	n	4.250000	4.125000	Lahi	2018-05-31	NaN	0.529	34.0	4.0
2	y	4.809524	4.603774	Don Angie	2018-05-29	4.0	0.508	126.0	4.5
3	n	3.500000	3.448980	Rangoon Spoon	2018-05-24	NaN	0.512	55.0	3.5
4	n	3.666667	3.848485	Wokuni	2018-05-22	3.0	0.510	60.0	4.0

Fig 10. Sample of our "NYT influence" data so far. Each row is a restaurant from our database.

P-values: significant vs. insignificant



```
Number of restaurants w/ insignificant p-values:  210
Number of restaurants w/ significant p-values:  67
```

About 43% of all our restaurants have insignificant p-values, which means our statistical test/bootstrap process inferred that NYT reviews did not have an effect on these restaurants' ratings.

Inspecting “significant” restaurants

Our group of “significant” restaurants consists of those with a significant change in post-NYT Yelp average. However, we’ve yet to take into account what this group looks like, in terms of what their NYT reviews are.

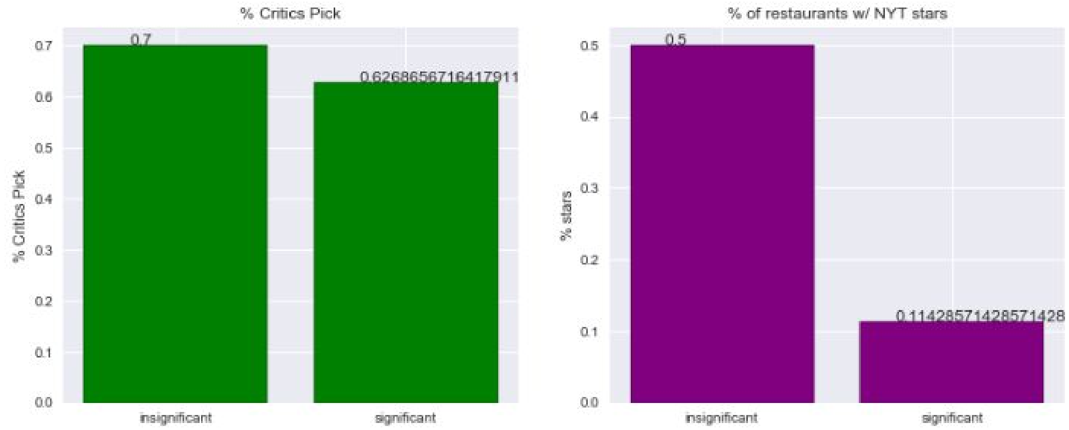


Fig 11. Comparisons between “insignificant” & “significant” restaurants for NYT data. Counterintuitively, “significant” restaurants have lower % “Critic’s Pick” and NYT stars.

Using the “insignificant” restaurants as our control group, I see that our “significant” restaurants actually have a lower rate of both “Critic’s Pick” and NYT stars - a counterintuitive finding (**Fig 11**). This difference may be partly due to the fact that there are much fewer "significant" restaurants (67), compared with “insignificant” (210).

I can also flip the analysis and look at “good” and “bad” NYT review categories, irrespective of our bootstrapping results (**Fig 12**). Here, a “good” NYT review is one giving “Critic’s Pick” and/or stars, and “bad” is a “Poor”, “Fair”, or “Satisfactory” rating. I hypothesize that restaurants with a good NYT review are more likely to receive a post-NYT review increase in their Yelp average.

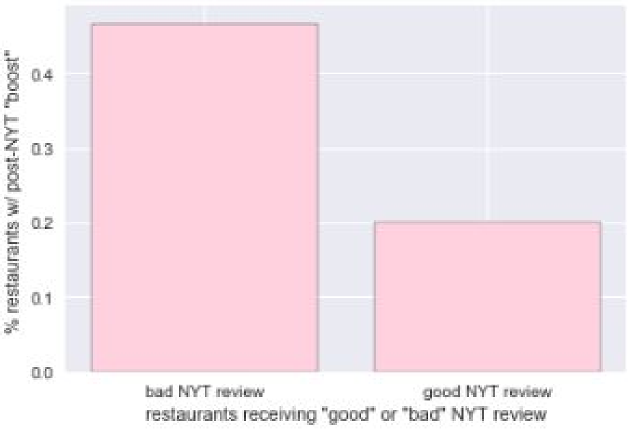


Fig 12. % of restaurants receiving “boost” in post-NYT Yelp ratings (within-category).

However, our results are once again counterintuitive. 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. The sample sizes are also asymmetrical in this case - 55 good-NYT reviewed restaurants, 15 bad - which might confound the analysis.

4.4 - 1-year window without bootstrapping

Since our bootstrapping method for identifying “significant” and “insignificant” restaurants yielded lackluster results, I turn to a final alternative method without bootstrapping. Here, I use a simple cutoff and 1-year pre- and post-NYT Yelp average. I compare the degree to which the Yelp average has changed from the 1-year pre-NYT window to post-NYT 1-year window, where the Yelp average is simply the average of all Yelp reviews posted during their respective time windows.

Our cutoff is a 0.5-fold increase or decrease to a restaurant's average Yelp rating. Such a change could have a significant impact on how the restaurant is perceived on Yelp (ex. 4.0 is perceived as "average", while 4.5 is perceived as "good" in Yelp).

With this new approach, there are now 245 "significant" restaurants and 69 "insignificant" ones. Now, ~78% of our restaurants experienced "significant" Yelp ratings change, post-NYT (in comparison to our bootstrapping method, with 31%).

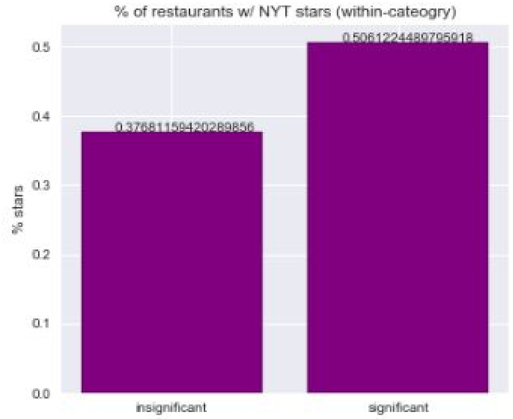


Fig 13. "Significant" vs. "insignificant" % of NYT stars. Significant now has greater % of stars.

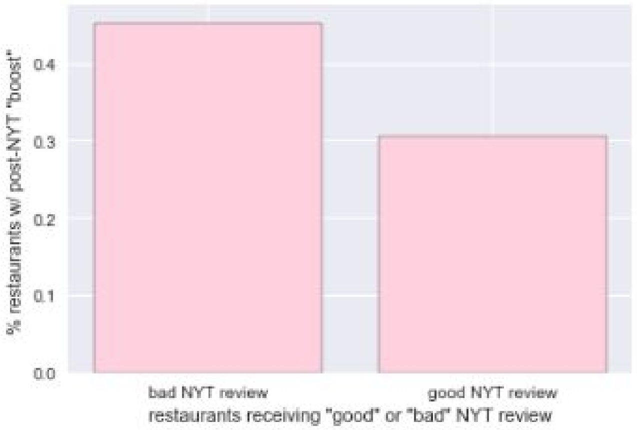


Fig 14. % of restaurants receiving "boost" in post-NYT Yelp ratings (within-category).

Yelp average and % "Critic's Pick" between "significant" and "insignificant" restaurants is still marginal (see notebook for data). However, I now see that "significant" restaurants are more likely to have NYT stars (**Fig 13**). This would align with our expectation that receiving a NYT star (or a bad rating under the star system) would affect subsequent Yelp reviews.

However, as before, having a good NYT review doesn't necessarily mean the restaurant will experience an improvement in Yelp ratings. Here, I again have the counterintuitive finding that among restaurants receiving a good NYT review, they were less likely to experience a post-NYT increase in Yelp averages.

One explanation is that positive NYT reviews may bring in more customers, but this also makes the restaurant more susceptible to higher standards, more reviews at the extreme, etc. I explore this in the next section, when I look at whether NYT influences likelihood of post-NYT increase in Yelp reviews.

4.5 - Impact of NYT reviews on restaurant popularity

Up until now, I have only examined Yelp ratings. Now I examine the possible impact of NYT reviews on restaurant popularity, where popularity is measured by the number of reviews written for a restaurant on Yelp.

The results are much more promising, with those restaurants receiving "good" NYT reviews experiencing more Yelp reviews following their NYT review (i.e. popularity) (**Fig 15**).

The difference between between restaurants receiving "good" & "bad" NYT reviews isn't as large as expected, however. Also, I should again keep in mind that our sample sizes for each category are small.

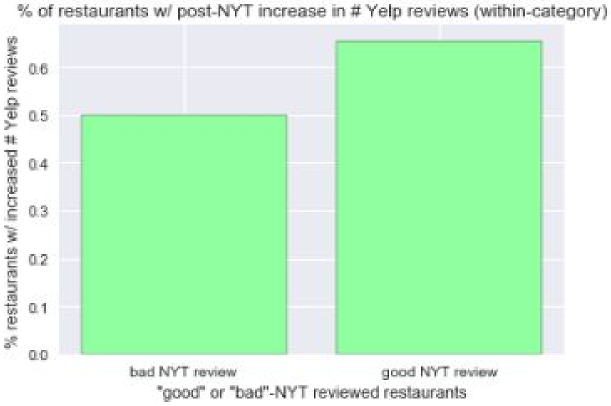


Fig 15. % restaurants w/ increased num. Yelp reviews.

4.6 - Discussion

In this section, I conducted a series of EDA/data visualization to determine whether NYT may have influenced restaurants' Yelp reviews and popularity.

1. **Bootstrapping.** First, I made a bootstrapping algorithm and explored different bootstrapping approaches. I settled on one approach and used it to categorize restaurants by whether their post-NYT Yelp average was significantly impacted or not ("significant" restaurants). Having a positive NYT review ("Critic's Pick" or NYT stars) didn't change whether a restaurant's post-NYT Yelp average would be significantly impacted.

2. **1-year time window.** Since this method didn't work out, I used a more straightforward method involving 1-year pre- and post-NYT time windows. Restaurants receiving NYT stars were more likely to experience a post-NYT increase in Yelp average - however, our sample size was still small.

3. **Restaurant "popularity".** So I finally turned our attention to another metric: number of Yelp reviews. I found that among the "significant" restaurants, those receiving a positive NYT review were more likely to experience an increase in Yelp reviews afterwards, possibly reflecting a boost in restaurant popularity, post-NYT. By extension, one interpretation is that a restaurant could experience more customers, but these customers may be more likely to be discerning upon reading the NYT review - hence, the odd increases/decreases in post-Yelp ratings that would not otherwise form a coherent narrative about how a NYT review can positively impact Yelp reviews.

In the next section, I refine our analysis of NYT influence on Yelp reviews by examining whether NYT reviews affect the language of Yelp reviews. For instance, if a NYT review highlights specific dishes at a restaurant, will post-NYT Yelp reviews be more likely to mention those as well (i.e. dishes more likely to be ordered)?

5 Determining NYT influence on Yelp reviews

Code: https://github.com/diana-xie/yelp-nyt-NLP/blob/master/data_yelp_NLP.ipynb

In the last section, I found mixed results with regards to how NYT may impact Yelp ratings. Now, I turn our attention to how NYT reviews may affect Yelp review *text*. In particular, one hypothesis is that 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.

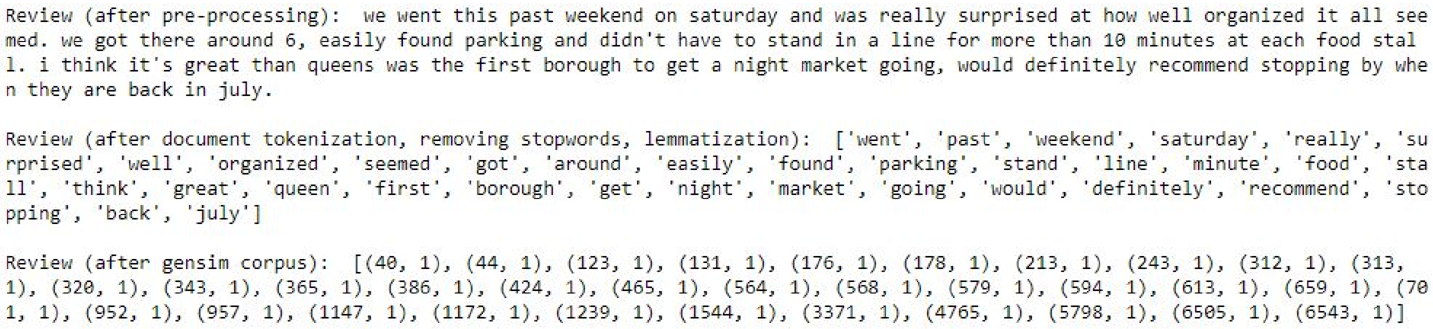
5.1 Yelp reviews - NLP corpus

First, I generate a pre-processed, tokenized list of documents to create a corpus of Yelp reviews, where each document is a review. Our pre-processing steps include:

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

I also experimented with a library called TextBlob, which corrects misspellings that were responsible for duplicate words (i.e. confounded word frequency). However, this added ~20 hours to execute, so I left it in our code but did not run it.

After tokenizing, I used gensim to create a corpus, where each token was mapped to a unique numeric ID and word count (i.e. bag of words, BoW) in order to set up a structure for inputting to NLP algorithms.



```
Review (after pre-processing):  we went this past weekend on saturday and was really surprised at how well organized it all see
med. 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 stal
l. i think it's great than queens was the first borough to get a night market going, would definitely recommend stopping by whe
n they are back in july.

Review (after document tokenization, removing stopwords, lemmatization):  ['went', 'past', 'weekend', 'saturday', 'really', 'su
rprised', 'well', 'organized', 'seemed', 'got', 'around', 'easily', 'found', 'parking', 'stand', 'line', 'minute', 'food', 'sta
ll', 'think', 'great', 'queen', 'first', 'borough', 'get', 'night', 'market', 'going', 'would', 'definitely', 'recommend', 'sto
pping', '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), (70
1, 1), (952, 1), (957, 1), (1147, 1), (1172, 1), (1239, 1), (1544, 1), (3371, 1), (4765, 1), (5798, 1), (6505, 1), (6543, 1)]
```

Fig 16. Visualizing a sample review under our different processing steps leading up to gensim corpus.

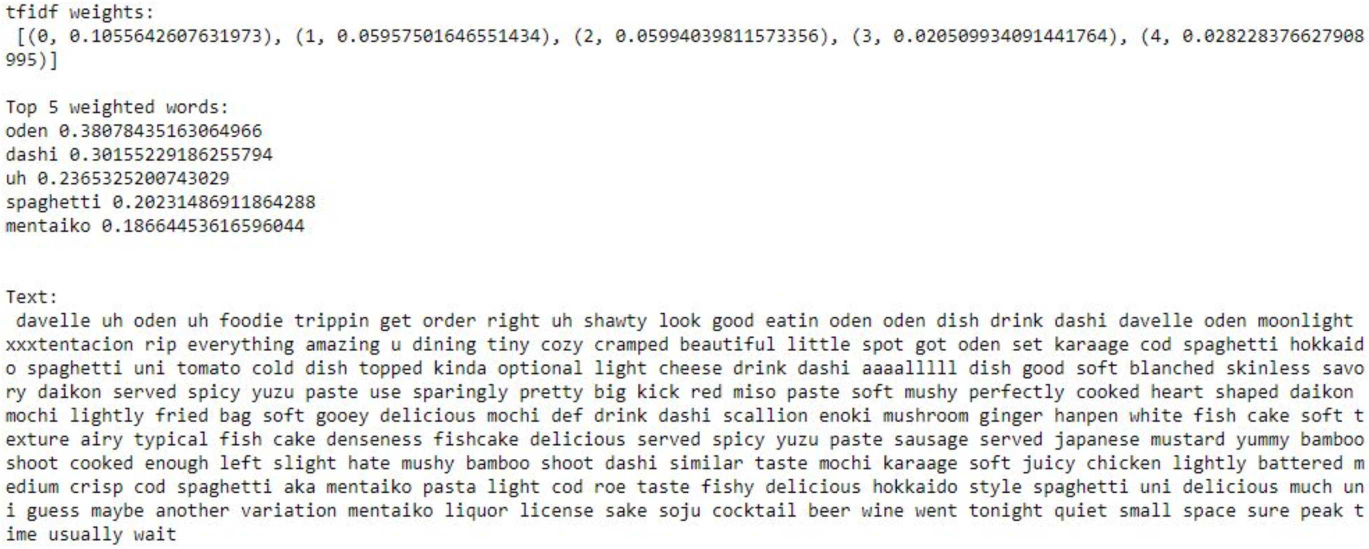
5.2 tf-idf EDA

Next, I experimented with gensim's tf-idf to identify important words in each document. It is with such a method that I may be able to identify cuisine-specific words to address our hypothesis that NYT influences Yelp review mentions of specific dishes.

Weighting “important” words in tf-idf is accomplished by down-weighting shared words (between documents) beyond simply stopwords, ensuring that common words don't show up as key words. Conversely, document-specific words are weighted highly.

Experimenting with tf-idf on a single Yelp review

I generate tf-idf weights for a single document (Yelp review) to see how tf-idf performs. The tf-idf model is generated on the entire corpus of documents (i.e. reviews).



```
tfidf weights:
 [(0, 0.1055642607631973), (1, 0.05957501646551434), (2, 0.05994039811573356), (3, 0.020509934091441764), (4, 0.028228376627908
995)]

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
xxxtentacion 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 aaaallll 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
```

Fig 17. Visualizing a tf-idf results. For a single Yelp review, “Davelle” restaurant.

It appears that if I take a document to be a single review, tf-idf may pick keywords that are specific to the reviewed restaurant's cuisine.

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

To determine whether a NYT review affected Yelp review text, I take as our metric the frequency of intersecting terms between the NYT review and Yelp reviews, pre- vs. post-NYT review publication.

Data preparation

I applied the earlier 1-year window method to separate pre- and post-NYT Yelp reviews. Then I performed tf-idf on all Yelp reviews. Next, I assembled a second corpus: NYT review text, where each review was a document in the corpus. I performed tf-idf on these reviews to get word weights for NYT reviews as well.

To ensure that stopwords and other irrelevant words were not included in the set of word intersections, I only considered 1) the top 20 tf-idf weighted words among each restaurant's set of pre- and post-NYT Yelp reviews, and 2) the top 100 tf-idf weighted NYT review words.

Determine intersection of NYT review words w/ Yelp reviews, pre- vs. post-NYT.

Our calculation for post-NYT increase in intersecting terms was: (difference)/(# of pre-NYT intersecting terms), where difference = (# of post-NYT words) - (# of pre-NYT words). So a 1-fold increase = doubling of intersecting terms.

The resulting distribution of n-fold increases to NYT-Yelp word intersections is not normal (**Fig 18**). As a result, we'll have to set our own subjective cut-off for what constitutes a significant increase. I will set it 0.5-fold, which means an increase of intersecting terms of at least 1/2 as many as before, post-NYT.

It's noteworthy that for those restaurants in which a decrease in intersecting NYT-Yelp terms occurred, it's never greater than 1. This means NYT reviews did not cause any unusual n-fold decrease in intersecting terms - such an anomaly would might undermine the validity of using NYT reviews to evaluate its role in publicizing specific dishes.

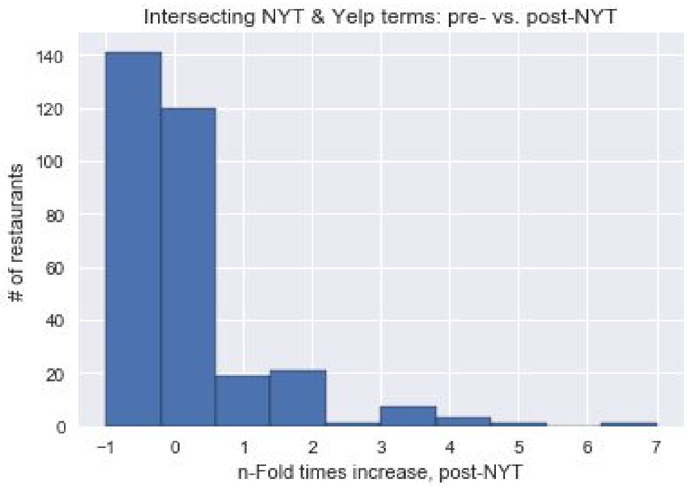
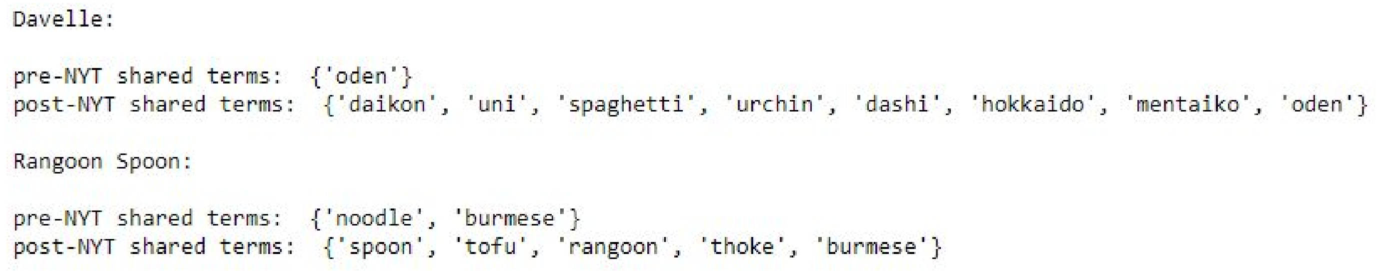


Fig 18. NYT-Yelp word intersections - post-NYT changes. Frequency of post-NYT n-fold increases to NYT-Yelp word intersections.

Here are a few samples of such changes in intersecting terms:



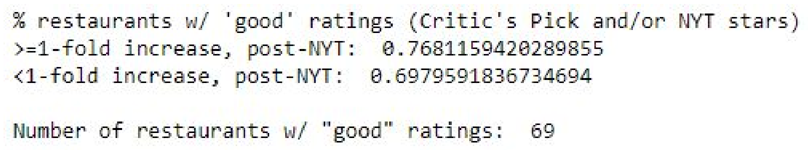
```
Davelle:

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'}
```

"Good" NYT ratings: influence on n-fold increases?

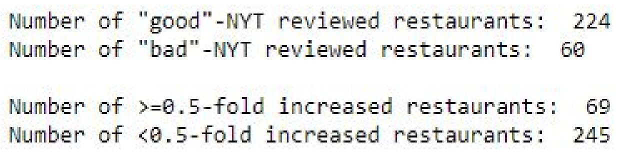
Are "good" NYT ratings (Critic's Pick and/or NYT stars) more likely to experience >=0.5-fold increase in shared NYT & Yelp words?

Among the restaurants w/ at least a 0.5-fold, post-NYT increase in NYT-Yelp shared words, ~76% had a "good" NYT review. Conversely, among <0.5-fold increase restaurants (i.e. insignificant post-NYT change), ~70% had a "good' NYT review.

The difference may seem significant, but I should keep in mind that the sample size per a group is very asymmetrical (numbers below).



```
% 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
```

Fig 19. Results of n-fold NYT-Yelp word intersection analysis.

As a result, I perform a permutation test in the next section to attempt a test for significance.

5.4 Permutation test - NYT-Yelp shared words

Since our distribution of NYT-Yelp word intersections is non-normal, I use a nonparametric test called the permutation test, which is similar to bootstrapping. Our "control" group consists of the bad/neutral NYT reviewed restaurants (“Poor”, “Fair”, “Satisfactory”, or no “Critic’s Pick” or stars), in which no change in post-NYT NYT-Yelp shared words would be expected. The “experimental” group consists of “good” NYT-reviewed restaurants: “Critic’s Pick” and/or NYT stars.

The difference between our control and experimental group is small: 0.105. Furthermore, the p-value is 0.25 - I can’t reject our null hypothesis (null = no difference between “control” & “experimental). Therefore I do not have sufficient evidence that a positive or neutral/negative NYT review would have an impact on the text of Yelp reviews.

5.5 Conclusion

In section 4, I found that restaurants w/ positive NYT reviews were slightly more likely to experience in increase in average Yelp rating afterwards. Since the results were less than expected, I also explored impact on Yelp popularity and found a larger increase.

There were undoubtedly other features at play, and here I explored whether NYT influenced the language of Yelp reviews. More specifically, would mentions of certain dishes in a NYT review increase mentions in Yelp reviews afterwards (i.e. implying that more people ordered the dish)? I used tf-idf to improve our hits of dish-specific terms and compared Yelp reviews, pre- & post-NYT review publication.

When I performed a significance test, I could not find sufficient evidence for a difference between positive and neutral/negative NYT reviews impacting Yelp reviews, at least language-wise. Thus, I cannot say that NYT reviews influence dish-specific mentions in subsequent Yelp reviews.

6 Conclusion

In this project, I aimed to determine the influence of NYT reviews on Yelp reviews. The goal was to better understand, with the increasing importance of the digital space to the restaurant industry, how two major platforms (influential in their own right) could influence the content of the other.

Here, the two platforms I used were NYT and Yelp. I first explored their different rating systems to better understand how they could relate to one another. Next, I performed a series of analyses to determine whether Yelp ratings were affected by NYT reviews. Our results were mixed, so I refined our investigation further by leveraging NLP techniques. Again, our results were mixed - I found a possible effect, but it was not able to be proven to be statistically significant.

The takeaway is that perhaps there are a handful (or more) of cases in which NYT reviews had impacted Yelp reviews, in terms of average ratings or popularity. However, there are obviously many other factors that can affect restaurants and make such a hypothesis as ours difficult to address.

Although the project’s results may have been mixed, in the process I generated several materials. This included a Yelp review web-scraper for scraping text, a functionality that the Yelp API does not provide and a tool that fills gaps in Yelp’s publicly available dataset. Additionally, I coded a NYT review web-scraper for extracting review information and text. I also experimented with, and displayed the results of our bootstrapping algorithm and various other tests that may be of interest to perform in future investigations of how one platform’s content can influence the other.