n0vah917 / dsc-capstone Public

0 stars 0 forks

Star Unwatch

Code Issues Pull requests Actions Projects Wiki Security Insig

main

n0vah917 Add files via upload ... now 15

View code

README.md

Classifying a Song's Genre Using Lyrics

Project Overview/Summary

The focus of this project is to identify key attributes that make a restaurant "good" or "bad". This would be of value to restaurants who are looking to improve their business, but have no idea where to start. By using the "voice of the people" through the reviews they leave on restaurant businesses, direct feedback and routes to improve service can be identified.

Restaurant review data sourced from a web scraped Yelp dataset was used in order to develop a NLP (Natural Language Processing) model that differentiates good reviews from bad reviews. Stopwords were removed, and all document tokens in the model were lemmatized in order to consolidate via root words. The TF-IDF (Term Frequency - Inverse Document Frequency) Vectorizer is used in conjunction with the Multinomial Naive Bayes classifier, in order to create a model that successfully classifies a review as "good" or "bad".

Additionally, using the review dataset, a rudimentary proof-of-concept recommendation system was created in order to provide restaurant suggestions for a given category, using using user_IDs, business_IDs, and review ratings for reference. This system is meant to serve as a follow-up to the NLP model, to provide a user facing solution that assists choosing restaurants.

Differentiating review types has massive implications on the ability to identify the competence of a restaurant. The resulting model was able to successfully assign a review type with an accuracy of 84% on training data and 83% on testing data.

Business Problem

In an industry as sensitive as food service, reviews, whether on a customer level or a professional level can make or break a restaurant in a customer's mind. Review services like Yelp and Google Reviews are instrumental for tourists and locals alike when informing decisions on where to eat for a night out. Any combination of factors can lead people to leave a review on a business's page, whether they be positive/supportive, detrimental/harsh, or somewhere in between. Owners with failing restaurant businesses may be at a loss when trying to find avenues of improvement. Significant investment in new hardware,

Without a culinary degree being a given in the restaurant industry, there is an interminable range of quality and service standards the average diner may run into; many times, the choice is a shot in the dark. From personal experience, I have gone to restaurants whose food can be seen as barely edible, but have excellent review aggregates. On the contrary, when going to restaurants that have subpar quality decor/cleanliness, the food I end up eating ends up being phenomenal. Unassuming exteriors like Sal's Pizza in Brooklyn, Peck Peck Chicken in Teaneck NJ, and Bagel Twist in Teaneck, NJ have m. Other times, in the face of mediocre food, some restaurants have outstanding service from their staff, which ends up bumping up the overall perception of satisfaction in my mind.

TV shows like Gordon Ramsay's Kitchen Nightmares offer the average viewer a look behind the curtain, into where underperforming restaurants are going wrong. In a similar vein, looking at review data gives context into the quality of restaurants in a given neighborhood, and provide insight into whether a restaurant is worth supporting or not. However, a cumulative 5-star review via Yelp can only hold so much information, and its reliability is dependent on a plethora of other factors that vary from restaurant to restaurant.

Thankfully, Yelp has an abundance of additional information, some being information in the form of review text. Yelp's review system enables users to write essay-length summaries of their experiences with restaurants. Yelp restaurant data spanning 2010 to 2014 will be aggregated and modeled via Natural Language Processing (NLP) in order to classify an individual review as bad or good. Looking at the importance of features within the predicting model will inform key features within restaurants that should be scrutinized when looking to improve overall satisfaction. To supplement this analysis, a recommendation system will be created using user reviews, in order to provide a user-informed method of restaurant selection.

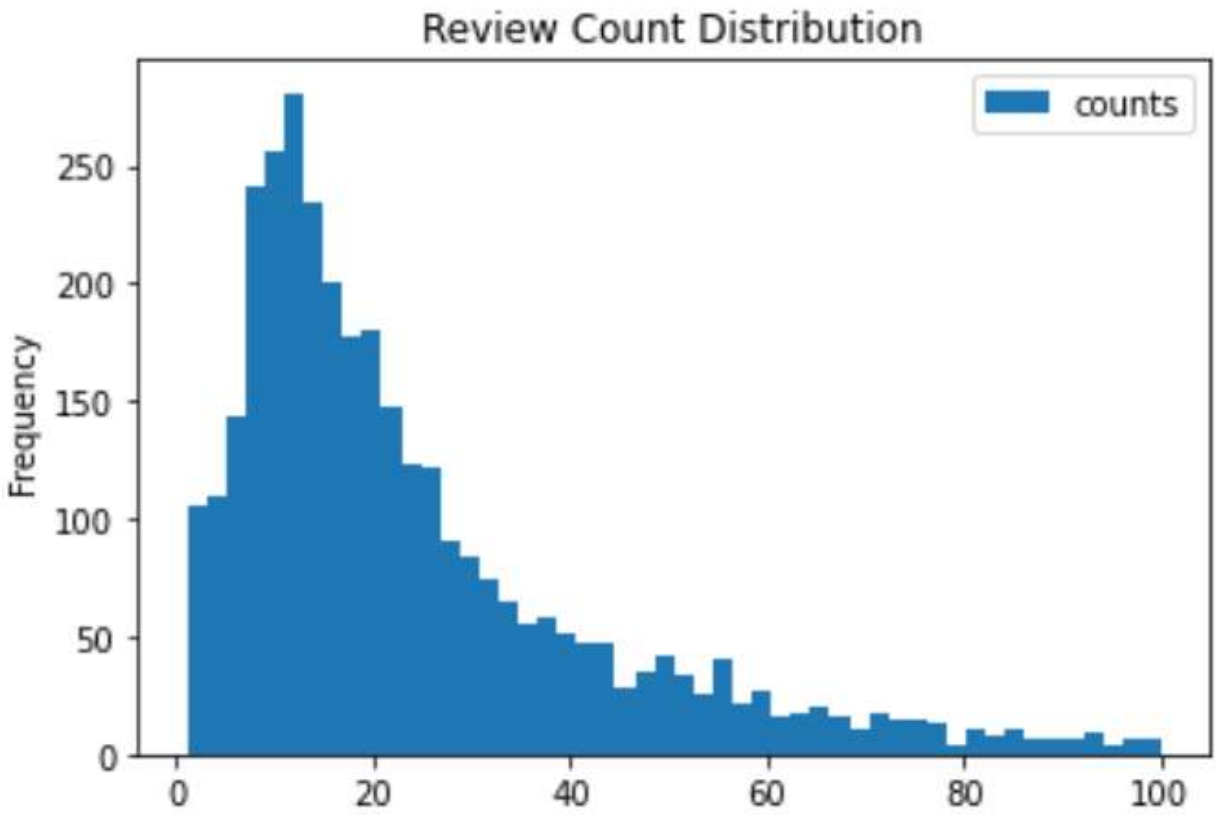
Data Understanding

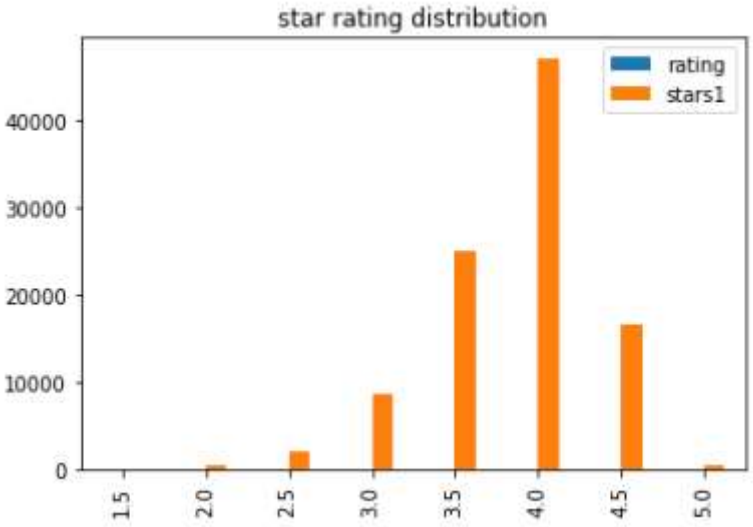
Aggregate data from the review site, Yelp, was found pre-scraped from a Kaggle dataset. All packages for interpretation/modeling are imported in this step. Each row in the initial dataset represents an individual review, with its user, business, review text, review rating, and supplementary elements that help further identify the business. Upon observation, the review text is represented in the 'text' column, as a long string separated by spaces and punctuation.

The user_id and business_id seem to be randomly generated strings, but are unique to each individual user/restaurant respectively. In addition to these, there are supplementary attributes that may be relevant for future analyses, such as votes for any particular review, and date, but are not relevant for the NLP model and recommendation system.

The field most critical to differentiating good vs. bad reviews is the 'stars1' field, which represents a review score from a scale of 1 to 5 for each review, in increments of 1/2 star.

There are a massive amount of reviews in the dataset to parse through, with around 50k individual users that have written reviews, and ~3.5k unique businesses. That being said, there are more than enough records for the NLP model.





Data Preparation

The first step in prepping the dataset to be summarized is to break each review (currently stored as a text string in the overall table) into individual tokens, separating by commas, then re-populating as a list in each row in the dataframe. Below, the regular expression pattern is used within the tokenizer to split words, looking for non-letter characters as boundaries while preserving apostrophes within words where relevant.

Out of curiosity, the distribution of a review length by each star rating is depicted in the histogram below. Interestingly, reviews with 1.5 stars have the highest review length on average. This implies that horrendous restaurant experiences cause the average user to write out all they have to say about their experience.

Similarly, average review length by price range is depicted below. As the dollar signs increase, the average review length increases as well. This may be due to a 1 dollar sign restaurant being viewed as "lower stakes" as a relatively lower amount of money was spent, but with regard to a 4 dollar sign restaurant, it is a significantly higher investment. Customers would want to be more informed about a costly restaurant before risking trying it.

All words are converted to lowercase in the following step. It would not be desirable to count 2 words as distinct if their only differing factor is case (e.g. "I'm" vs "i'm".)

In the following step, English stopwords were imported and a function was created to remove all stopwords within each review text list. When doing a form of classification, it is important to remove stopwords as there are bound to be sets of words that are common amongst any body of text (e.g. "and","the"). Keeping these words makes it harder for the eventual model to differentiate between two sets of text and make the proper classification.

The 'stars1' field is used to map good and bad reviews, at least on a rudimentary level. As a rule of thumb, the model will first consider all reviews with a star count of less than 3.5 as a "bad" review, and greater than 3.5 as a "good review. This designation is not permanent and will be changed in the event of needing further differentiation.

Data Modeling

The text data has now been cleaned and is ready to be modeled. In order to initialize modeling, the full dataset is split into training and testing groups so that the models can be evaluated in the presence of "new data". Review text is represented by X, and the assigned rating (good/bad) defined by the preceding step is represented by y.

Model 1: TF-IDF Vectorizer + Multinomial Naive Bayes

In order to further prep the data, a term-frequency inverse document frequency vectorizer is initialized. Vectorizers are necessary for this process, as they provide a method of converting text data, currently in the form of lists, into a form that a model can understand. A TF-IDF vectorizer, also known as Term Frequency-Inverse Document Frequency, will be used. This vectorizer is ideal for use in the case of multiple review documents, and due to the nature of the calculation, it implies rare words contain more information than common words. Term Frequency (TF) represents the presence of a word in a document as a fraction, while Inverse Document Frequency (IDF) represents the prevalence of the word across all documents. By multiplying the two, a value for the importance each word has in the entire document base is obtained.

The max_df is specified as 1000, which implies that words are only considered as identifiers for the model if they appear in less than 1000 distinct documents. The min_df is specified as 30, which implies that words are only considered as identifiers for the model if they appear in more than 30 distinct documents. Words that appear too frequently/not frequently enough are not desirable, as their inclusion may decrease resulting performance.

Below, a Multinomial Naive Bayes model is instantiated for our model to be fit on, and the plotted confusion matrix helps visualize how well the model did. The MultinomialNB model is essential for text classification, as it involves using conditional probability across the different categorizations, and is generally good for classifying on distinct features.

The corresponding confusion matrix, seen below, shows how the model classifies the test/train data using the Multinomial Naive Bayes classifier through counts of true/false positives/negatives.

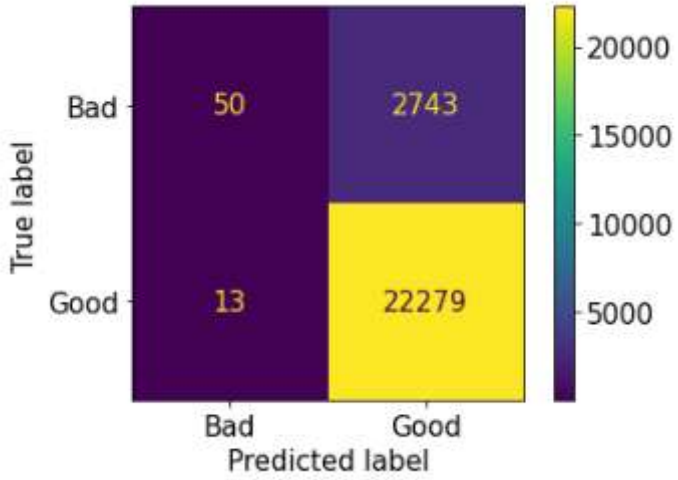
Precision: Precision measures how precise the predictions are. Precision is calculated by taking the number of true Good classifications, and dividing by the number of Good song classifications predicted by the model.

Recall: Recall measures what % of the reviews were actually identified as Good correctly. Recall is calculated by dividing the number of total true Good classifications identified by the model, by the number of actual Good reviews.

Accuracy: Accuracy evaluates all predictions within the model. Accuracy is calculated by dividing the total correct predictions from the model (both Good/Bad) by the total records in the original dataset.

F1 Score: F1 Score takes both precision and recall into account to get a cumulative score. The formula is calculated as follows: 2(precision recall)/(precision + recall). This metric is a good measure of overall model performance, as a high F1 score implies both precision and recall are both high as well.

As seen below, the initial model performs very poorly on the both test data and training data, despite an accuracy overall of 89%. The confusion matrix shows that our model heavily skews towards predicting Good reviews as correctly. This is due to the overwhelming about of commonalities of word frequency between Good and Bad reviews, which can be fixed in subsequent steps.



Model 2: TF-IDF Vectorizer + Multinomial Naive Bayes, Changing Conditions For Bad/Good Reviews

For the second iteration of the model, since the Multinomial Naive Bayes classifier in conjunction with the TF-IDF vectorizer seemed to be an ideal fit for text data, all model parameters will be kept the same in order to compare similar setups.

This time, in an effort to better differentiate Good/Bad reviews while improving overall accuracy score, the overall review set will be limited by filtering out "mediocre reviews" identified as the 3-4.5 star rating range. While this limits the overall dataset, better classification results are anticipated through this step. In the initial model results, it seemed as if most reviews under the initial conditions were identified as "Good" which makes the overall dataset heavily imbalanced. This step mitigates the discrepancy to a great degree as well.

As seen through the heavily improved recall scores within both the test data and training data, we can see evident model improvement with the use of the new condition set. While changing the conditions seemed to increase the overall model performance, additional steps can be implemented to make more clear designations and mitigate false readings.

Model 3: TF-IDF Vectorizer + Lemmatizer

As seen in the prior word count distribution across the Good/Bad split, it was observed that some words like “get” and “got” were considered as distinct words although semantically, they have the same meaning. In an effort to improve prior model performance, lemmatization will be performed in order to consolidate based on their root words. Below, the part of speech (POS) is tagged to each token word, and the root words of each are found. Additionally, more common words amongst the designations are included in order to decrease similarity amongst the genre sets. Word frequencies are revisualized, each designation looking more unique. For the most part, the word frequency still looks relatively similar, but order of frequency is distinctly different between Good/Bad designations.

As seen in the following classification report, overall performance does not seem to have improved seen from the slight uptick in false positives/negatives. This may just be chalked up to noise, however.

Model 4: Considering Bigrams/Trigrams

When reading through the individual reviews, it appears that certain word sequences such as "very good" or "not good" were seen multiple times. The issue that arises when vectorizing using each word individually is that the context of where a given word appears can be vastly different. To mitigate this, the sequence of words will be acknowledged to inevitably improve performance, through the use of bigrams and trigrams within the vectorizer. The ngram_range parameter is used to identify all sequences of 1-3 words that appear within the dataset.

As seen from the output report, performance slightly improved at mitigating false positives, which is a good sign since the majority of the data is comprised of good reviews anyways. It seems Recall and Precision for Bad reviews increased as well. Overall accuracy went up from 82 to 83 in testing data and 83 to 84 in training data. This shows that the consideration of bigrams/trigrams made a positive improvement on model performance.

The model is in an overall better state/performance level when compared to the original output after filtering on relevant reviews, lemmatizing, and implementing bigrams/trigrams. Below, a list of the most impactful coefficients paired with their respective word sets can be seen in the dataframe below. The absolute value of the coefficients were taken in order to observe the highest magnitude of effect.

Classifier Results Evaluation

When considering a matter as sensitive as the restaurant industry, the distribution of words chosen across the breadth of different reviews have a large impact on how a restaurant is perceived and recognized by cultures and communities. The final model created, using a Multinomial Naive Bayes classifier with a TF-IDF vectorizer, was able to differentiate good/bad reviews with an accuracy of 83% for test data and 84% for training data.

A classification report and confusion matrix of the final classifier model iteration can be seen below across the test/train datasets.

The model performance is relatively strong, when considering the overall imbalance between good and bad reviews. This imbalance is to be expected, however, as seen through the previous plotted distributions, more people tend to leave relatively positive reviews for restaurants. In an ideal scenario, there would be an equal distrbution of good/bad restaurant reviews, but people may not want to leave bad reviews often, potentially out of fear of ruining someone's livelihood.

Several of the identified words/phrases below seemed to highlight features and experiences regarding the restaurants in the dataset, both good and bad. The themes around the experience-oriented phrases revolve around money spent or bad experiences from attending staff (e.g. speaking to managers, hospitality, sub-par service, saving money, taking forever, horrible service). It seems as if these issues, when apparent from the customer end, may even be reflected through the experience with the food. Inefficiencies with any restaurant or business are hard to ignore, and they translate down to the inevitable service a customer gets.

	feature	abs_value
2335	hooter	11.095267
566	bull	11.095267
1773	frozen hot chocolate	11.095267
1994	golden corral	11.095267
5137	venetian	11.095267
4181	serendipity	11.095267
1068	dance floor	11.095267
1772	frozen hot	11.095267
3933	resort fee	11.095267
1199	dim sum	11.008127
2654	linq	10.964657
2959	mimi	10.926317
956	corral	10.893200
4484	sport bar	10.859702
4606	sugar factory	10.818226
4452	speak manager	10.784332
5235	waste time money	10.658306
4022	rudely	10.657740
4252	service terrible	10.643969
379	beer cold	10.624604
4226	service lack	10.619672

Recommendation System

To supplement the NLP analysis, a follow up component to the analysis is the creation of a recommendation system based on other users' reviews. While knowing key points of improvement proves to be a good solution for restaurants, a recommendation system is a client facing solution that also involves user input. These recommendations, supplemented by the tendencies found in the NLP analysis, provides a holistic, well-informed method of gauging restaurant value/worth.

The surprise package is imported above, and is used to read in the relevant fields from the review dataset. The surprise package incorporates a data type that is particularly good when it comes to building recommendation systems that are optimized for computational efficiency.

Below, it is seen that there are far more users than there are businesses. That being said, user-user comparison will likely be a more efficient/successful route for this recommender system, which compares similar users to similar users.

Number of users: 20426

Number of businesses: 931

Single Value Decomposition was used to generate a model that outputs user recommendations. Single Value Decomposition (SVD) is used to reduce a matrix into several component matrices, while revealing interesting tendencies about the original matrix. SVD involves the creation of an optimization problem; minimization of the prediction error margin is measured through the root mean squared error (RMSE). A lower RMSE is indicative of improved performance and vice versa.

Grid search is used on the SVD model in order to view the lowest RMSE scores with the tested parameters. n_factors is one of the factors that refers to the number of factors that the matrix will be considered when making predictions. reg_all refers to the regularization term for all parameters.

Below, the RMSE for the ideal model is only .208 for the relevant readers and prodictions. this is relatiely strong, because it measn our model is only off by a fraction of a star for its generated predictions. That being said, it is safe to assume that the system will be successful in assigning predictions to the user

While this method of assignment is invaluable at generating user predictions, it is highly advised for all users to take the next step of parsing through the business pages of the model outputs, in order to get more context before deciding to go to a given restaurant. As seen in the prior analysis the value of reading through review text gives the most informed decision for any restaurant goer.

```
                  business_id           name                       categories
29764  l5FvoKE2-k6JpGVuhH9YVA  Sushi Catcher  Sushi Bars;Japanese;Restaurants
How do you rate this restaurant/business on a scale of 1-5, press n if you have not been :
2
                  business_id                    name  \
92481  SPBZxmt8_nT30rNVnKHYKA  Akaihana Sushi & Grill

                           categories
92481  Sushi Bars;Japanese;Restaurants
How do you rate this restaurant/business on a scale of 1-5, press n if you have not been :
1
                  business_id         name                       categories
35796  6syMU43FKGkcbX2957Ga8A  Sekai Sushi  Sushi Bars;Japanese;Restaurants
How do you rate this restaurant/business on a scale of 1-5, press n if you have not been :
2
                  business_id              name  \
87008  V3ruBXjLGWniPNPQOzRhiw  Makino Restaurant

                           categories
87008  Sushi Bars;Japanese;Restaurants
How do you rate this restaurant/business on a scale of 1-5, press n if you have not been :
3
                  business_id         name              categories
86546  08IFR_ruWR96K3Q6sakI_g  Sushi Creek  Sushi Bars;Restaurants
How do you rate this restaurant/business on a scale of 1-5, press n if you have not been :
4
```

ut[45]:

business_id	categories	name	recrating
6X9iyuM2XdoCGT4q9qv5cA	Sushi Bars;Japanese;Restaurants	JJANGA Japanese Restaurant	4.492273
OwBPjUz2o0J5K3DzcHkBtg	Sushi Bars;Japanese;Restaurants	Soho Japanese Restaurant	4.491927
HpaYCM_NCaul72LLXxC6SA	Tapas/Small Plates;Sushi Bars;Japanese;Restaur...	Yonaka Modern Japanese	4.489452
sNBquLTaV3lbUWkzSUITpw	Sushi Bars;Japanese;Restaurants	Sakana	4.489281

Conclusion

The model was able to successfully assign the majority of restaurant reviews in the dataset, solely through the use of Yelp review text. An 83% accuracy across both test/training data shows that the model performs relatively well, even in the face of new data. The model's important feature outputs, given in the form of coefficients and paired key words/phrases, should be scrutinized in order to understand the most critical elements that should be focused on when it comes to restaurant improvement.

As seen in the ranked keyword importances from the model, several words that were most impactful for the performance of the overall classifier identified a restaurant's features and experiences, both good and bad.

The themes around the experience-oriented phrases revolve around money spent or bad experiences from attending staff (e.g. speaking to managers, hospitality, sub-par service, saving money, taking forever, horrible service). It seems as if these issues, when apparent from the customer end, may even be reflected through the experience with the food. Inefficiencies with any restaurant or business are hard to ignore, and they translate down to the inevitable service a customer gets.

For example, in the FX show The Bear, an accurate look into the life of restaurant workers, depicts a prime when communication, organization, and efficiency are not at the forefront of any operating business. In the scene depicted below, chaos in the kitchen soon ensues when tension between the cooks and chef starts to flare. The aftermath of this scene is not shown, but it is implied that several customers orders were left unfulfilled. Similar to the outputs of the model keywords, long wait times and terrible service are indicative of bad performance.

https://www.youtube.com/watch?v=1K3z62yoiOA

Money-oriented words/phrases were identified as well, implying perceived customer value of the food served also plays a big part in the overall satisfaction (e.g. "full price", "save money", "waste time money"). If a customer doesn't believe that the food they received was worth the money they paid they are less likely to support that restaurant in the future.

Recommendations

In order to improve overall restaurant reviews and customer satisfaction, improvement should be looked at through the lens of service improvement, and food quality. In reality, there is a significant overlap between the factors, but both can be improved through different means.

In terms of organization, different systems can be established, such as the Brigade de cuisine invented by Escoffier, a hierarchal system found in restaurants that keeps the overall kitchen working similarly to a well-oiled machine. This system, when implemented, allows for no margin of error while emphasizing precision and productivity. Confusion and lack of clarity are mitigated to a large degree when kitchen staff has defined and stable roles. Newfound organization would improve customer wait times and overall satisfaction.

Outside of the logistical framework, service improvement can be achieved through front-of-house staff training, emphasizing professionalism and enthusiastic service. Vetting waiters and servers to ensure that temperament and values align is critical in ensuring a customer feels welcome and accepted. Comments around rude service would be reduced when all diner-facing staff cultivates a common friendly culture.

Rude service may also stem from worker compensation, as staff members may feel disgruntled when their work does not reflect what they take home at the end of the day. This is especially evident in America, where the majority of what a waiter makes is based on tips. If overall base wages were improved, waiters/busboys would not feel feel the pressure to "fake" friendliness in order to get paid. Their enthusiasm and desire to work at a given establishment would translate into the service they provide.

The "value" oriented comments are somewhat trickier to mitigate, but can be managed through the kitchen's supply chain. Customers may not feel like they get the "bang for their buck" if they pay too much for a smaller portion than they were expecting. Understandably restaurants do have to price their food in certain ways to keep their overall profit margins high and pay their staff fairly. In order to keep prices low, expensive/costly ingredients can be substituted with cheaper alternatives, keeping the focus on quality and identity of a given dish.

Next Steps

Next steps with the overall model would involve using a more nuanced system of ratings to evaluate, rather than just a 5 star review. Seen through Metacritic movie reviews, a score of 1-100 would be more valuable to interpret, as there is more of a range/scale that users could input. A rating of 4 stars may mean something different to many different users, either average, or very good. A rating of 90 implies several aspects hit the mark, but slight points of improvement can be observed. Using a review system on this scale may improve overall model accuracy.

Another avenue to explore would be to utilize more filtering options within the recommendation system, such as price range, in order to create more nuanced recommendations. Being able to hone in on a more granular level would make the recommendation system more usable/dynamic for the average diner.

Additionally, combining both components of the project by displaying top words of significance/frequency for each restaurant recommendation could provide more context in a decision to dine.

Using a logistic regresssion model on the additional features that were present in the initial dataset, such as presence of restaurant features and votes on individual reviews, would provide a different perspective on the most critical features that influence good/bad reviews. Restaurants would be able to add these features (e.g. a salad bar) in an effort to improve customer perception.

Running the NLP analysis under the grain of restaurant categories may revieal if some cuisine types receive worse criticism than others. Restaurants with differing cuisine types may require differing solutions. The output for each of the category types would provide a solution set for each cuisine.

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Languages

- **Jupyter Notebook** 100.0%