Summarize results of this task, list qualities of generated models (e.g., in terms of accuracy), and rank their quality in relation to each other.

For our group project we are looking in depth at the Yelp dataset from a variety of perspectives.

**Category 1: Recommendations**

**Collaborative Filtering**

Firstly, we are looking at what is the best way to give a recommendation to a user of where to eat?

Since these are classification techniques, we will be evaluating performance via a confusion matrix. As with a confusion matrix we are able to retrieve to retrieve metrics like accuracy, false positives, true positives, and etc. As we tune, we will determine if more metrics will be necessary to help us ensure we have the most optimal model.

We are experimenting with a couple of different modeling techniques. Firstly, we are trying collaborative filtering approach.

With collaborative filtering we are using the idea of making automatic predictions about the interests of a user by collecting taste or information about many users. We are doing this features based on the category based on

city',

'hours', 'name', 'review\_count', 'stars\_x', 'state', 'AgesAllowed',

'Alcohol', 'Ambience', 'BYOB', 'BYOBCorkage', 'BestNights',

'BikeParking', 'BusinessAcceptsBitcoin', 'BusinessAcceptsCreditCards',

'BusinessParking', 'ByAppointmentOnly', 'Caters', 'CoatCheck',

'Corkage', 'DietaryRestrictions', 'DogsAllowed', 'DriveThru',

'GoodForDancing', 'GoodForKids', 'GoodForMeal', 'HappyHour', 'HasTV',

'Music', 'NoiseLevel', 'Open24Hours', 'OutdoorSeating',

'RestaurantsAttire', 'RestaurantsCounterService', 'RestaurantsDelivery',

'RestaurantsGoodForGroups', 'RestaurantsPriceRange2',

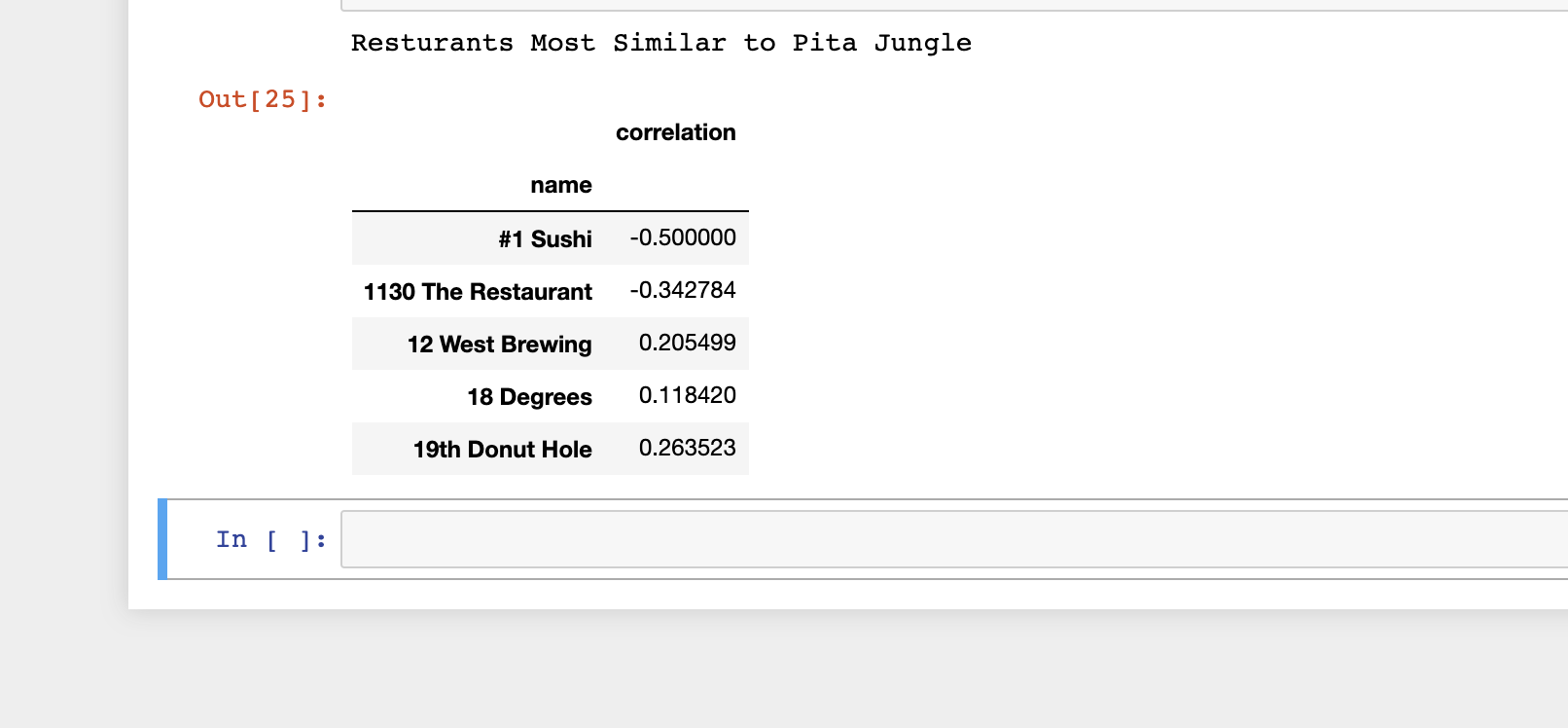
'RestaurantsReservations', 'RestaurantsTableService',

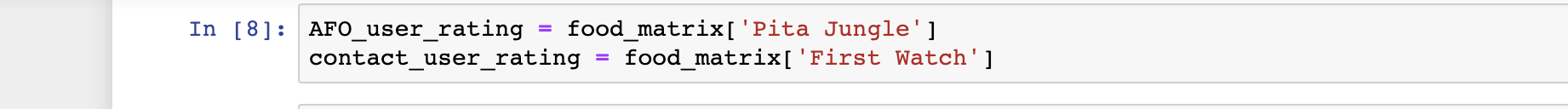
'RestaurantsTakeOut', 'Smoking', 'WiFi\_'free'', 'WiFi\_None',

'WiFi\_u'free'', 'WiFi\_u'no'', 'WiFi\_u'paid'', 'date', 'funny',

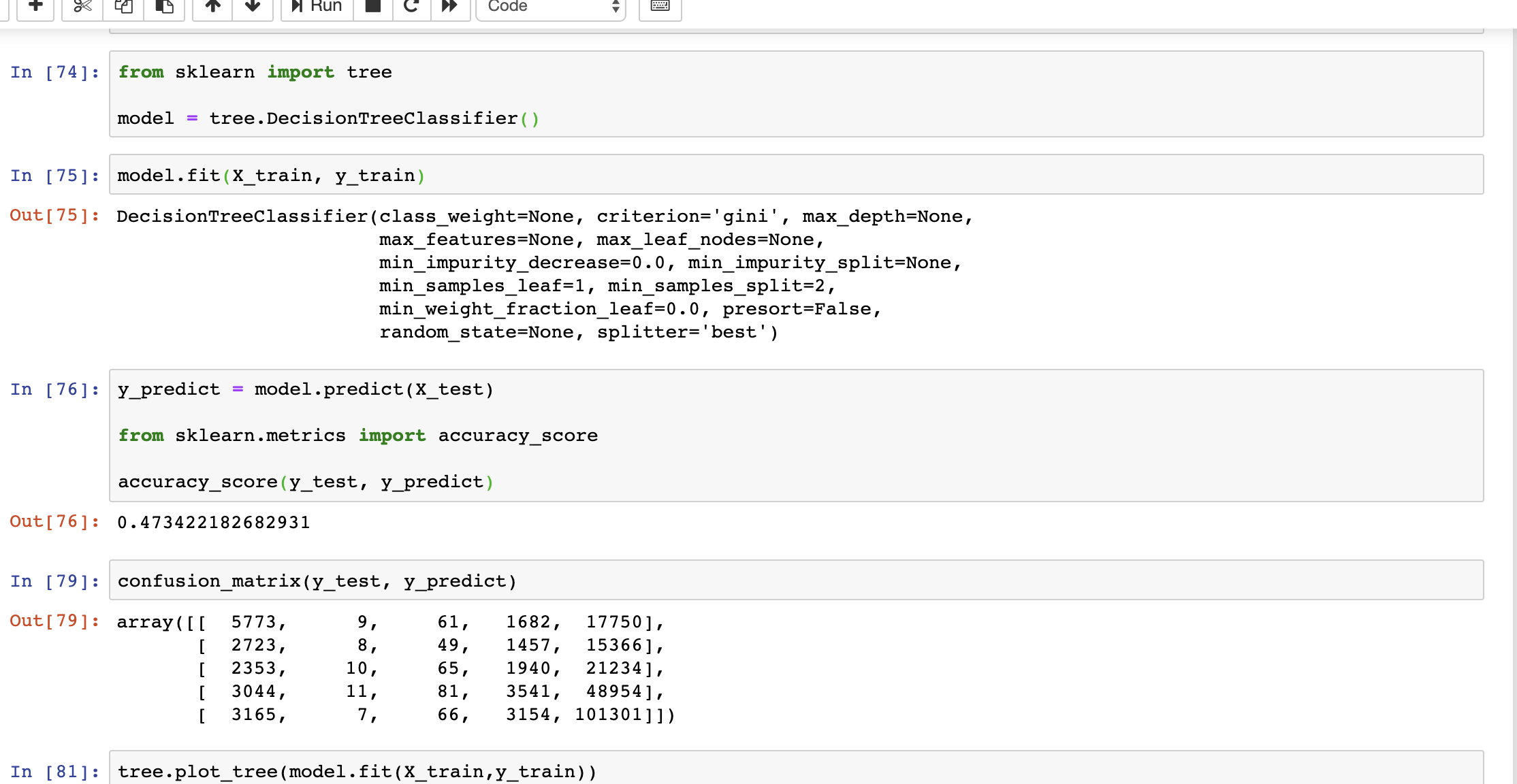
'review\_id', 'stars\_y', 'text', 'useful', 'user\_id', ‘category’

We tried the model with seeing what restaurants are most similar to Pita Jungle. The correlations we got were small.

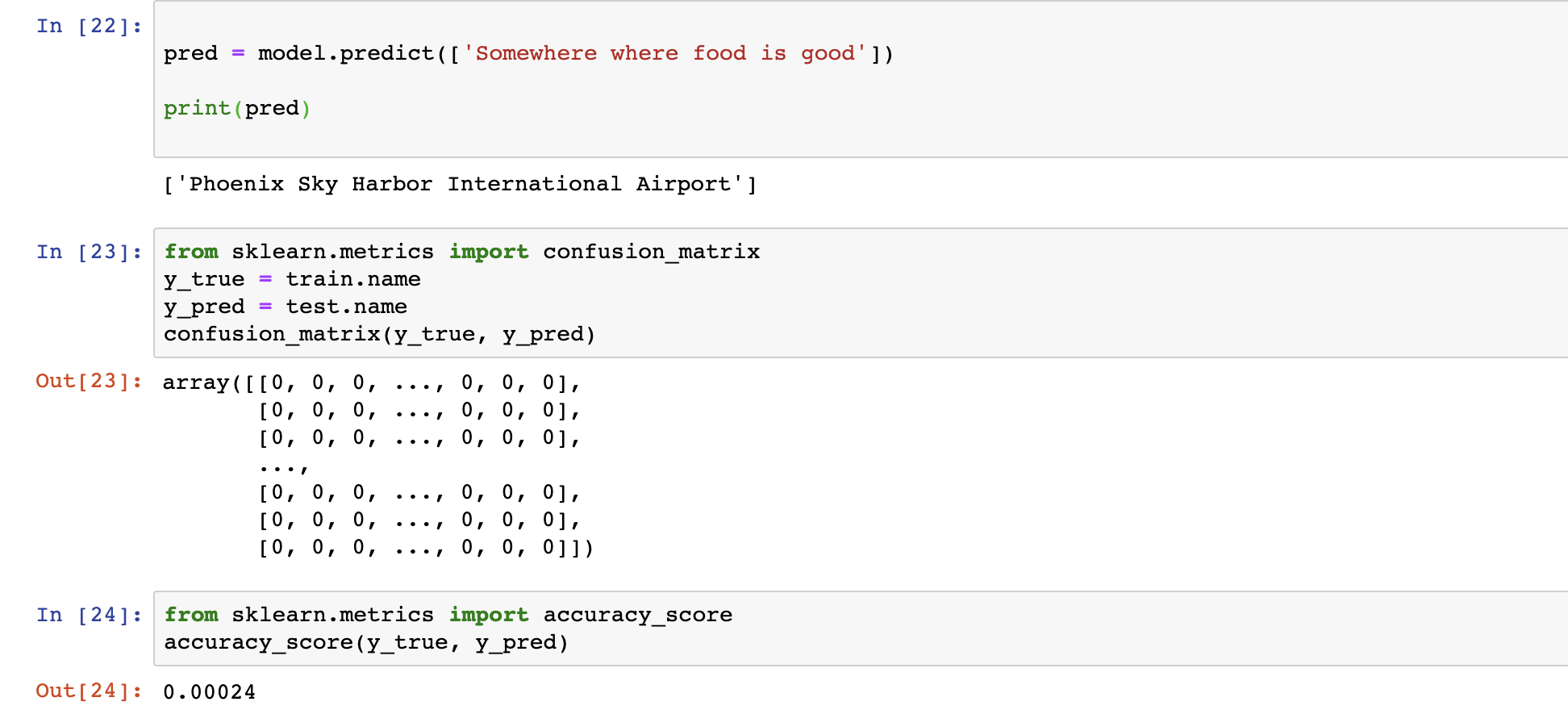
But this approach can be improved with a Memory Based-Collaborative Filtering. Or a model based collaborative filtering which is good with dealing with scalability and sparsity. If we do this we can evaluate with RSME.

****

**Using a Decision Tree**

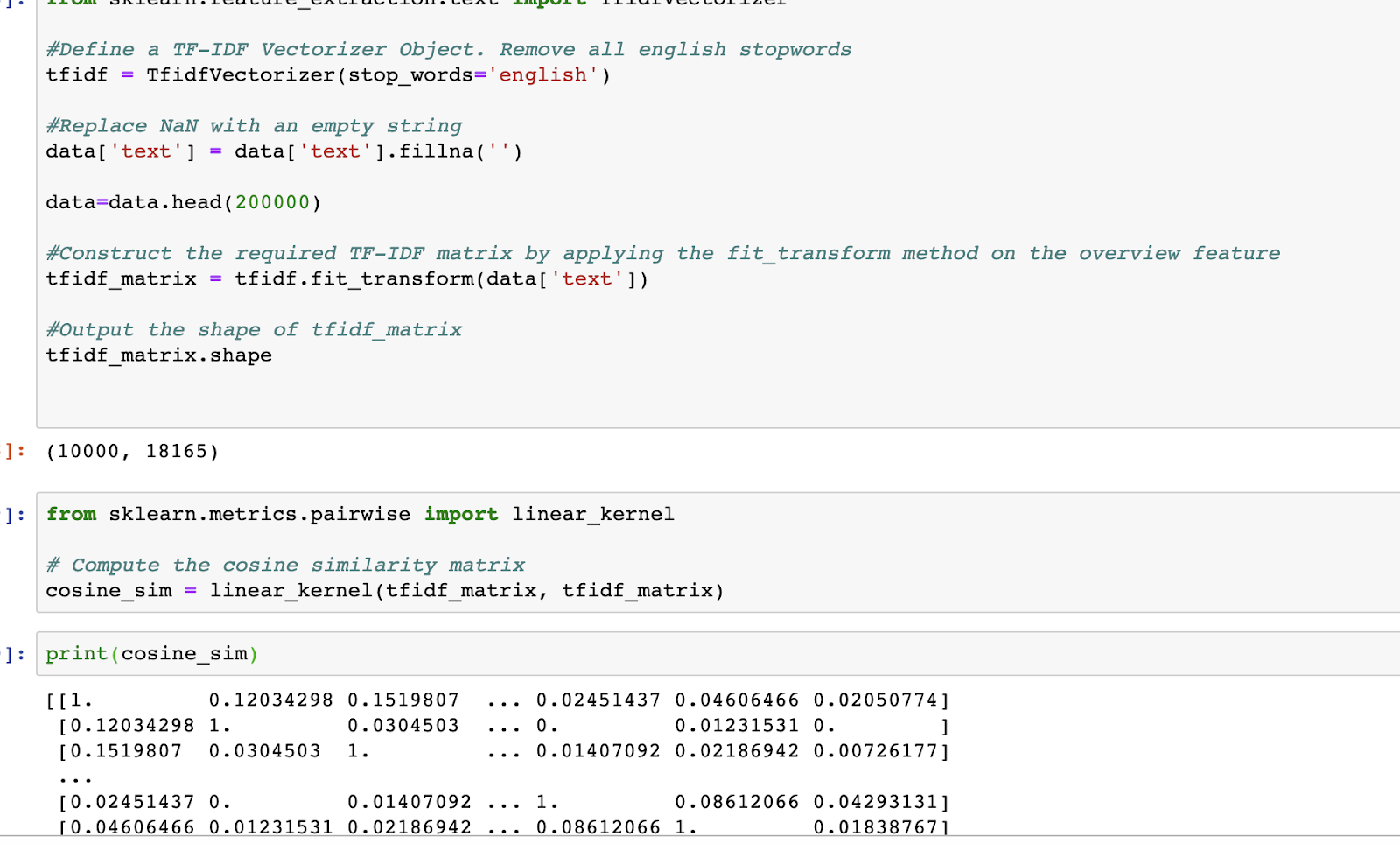
Utilizing the features of business stars and review count and the target feature of review\_stars, we are trying to predict will a user like the restaurant? This approach gives us an accuracy\_score of 47% percent. Which is not so great. This model will still need hyper parameter tuning in order to help give the best choices to a user of where to eat.

**MultiNomial Naive Bayes.**

Using a Multi-nominal naive bayes, based on the cleaned tips, we are trying to classify what words can gear you towards the right recommendation. Unfortunately with this approach. The accuracy was almost close to 0. Right now this approach would need a quite a bit of work. I honestly, feel that maybe we would get more out out of it using reviews. It could be how we preprocessed the data.

Content Based Filtering

Is we are recommending items based on the comparison between the content of the items and a user profile. In this case we are using user\_id, and then all metadata associated to the movie. I do not think this option will work for us as we have omitted metadata about the user. We only are taking into consideration the user\_id. This could explain the low cosine similarity scores we have. To evaluate this technique we were using a RSME, confusion matrix, and recall.



Category 2: Sentiment Analysis:

The metrics we are using is subjectivity and a confusion matrix.

1. For the sentiment analysis we want to predict if reviews/tips are positive or negative? What concerns a customer , what makes a customer neutral, and what makes a customer happy.

To determine if a review is good or bad based we are doing the prediction based on the review rating. (We are trying to predict the ratings using a multinomial bayes)

To determine if a tip is good or bad we are doing the prediction based on the business rating. (We are trying to predict the rating using a multinomial bayes)

To determine the sentiment of a tip or a review we are using the sentiment library from text-blob. We then have the sentiment score and the subjectivity of the score.

To determine what words are positive, negative, neutral in a review we are using the SentimentIntesityAnalyser from nltk and determining polarity scores as thresholds of what we consider what.

Category 3: Photos

For photos based on the labels given in the data set we are trying to determine what labels are most representative of a business. It is still to be determined what metrics we will use for testing and evaluation.