DATS-6501 CAPSTONE PROJECT

yelp DATA CHALLENGE

Wei Hao Fei Wang

Data Description

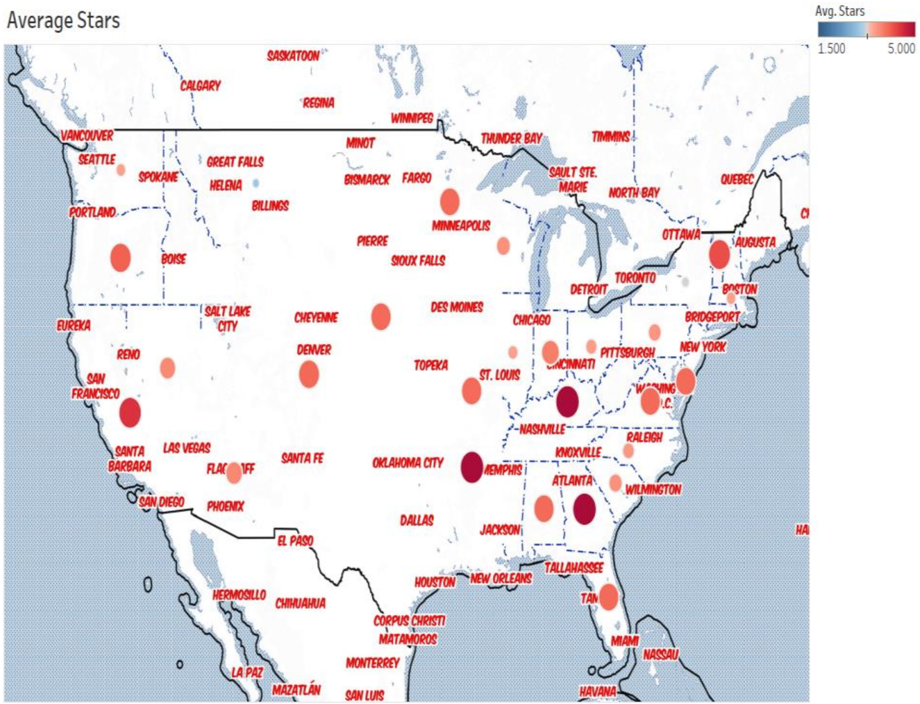
The Yelp dataset containing its businesses, reviews and user data in Json file format. This dataset involves **5,996,996** reviews, **188,593** businesses, **1,185,348** tips by **1,518,169** users, and business attributes like hours, parking, availability, and ambience in **10** metropolitan areas of US.

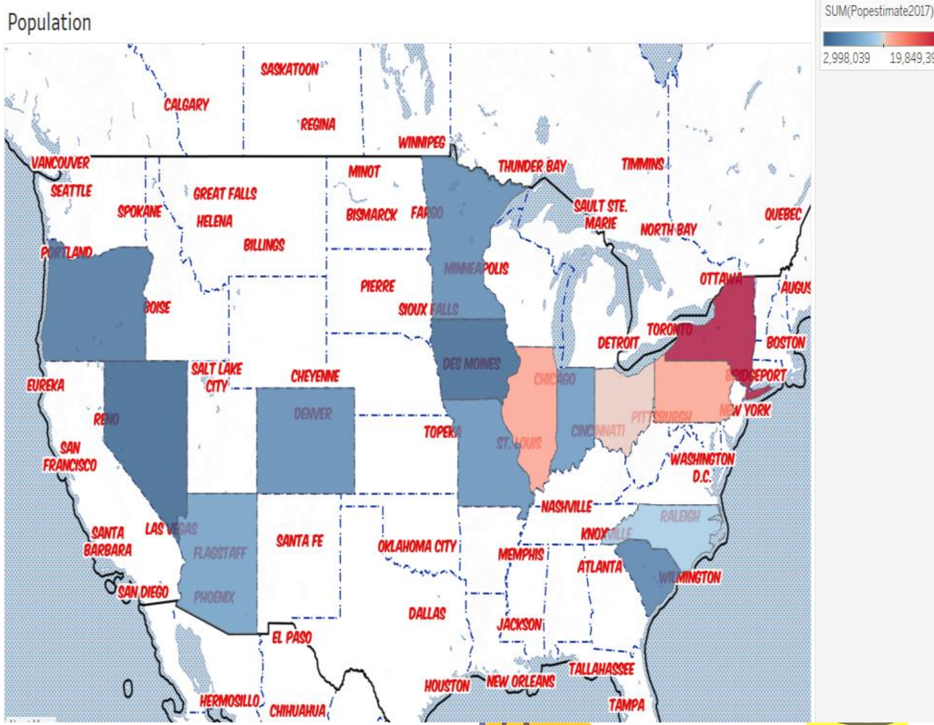
Business file		Review file	
Attribute	**Meaning**	**Attribute**	**Meaning**
business_id	Business ID (index for datasets merging)	busindess_id	Business ID(index for datasets merging)
neighborhood	The neighborhood of the restaurant	date	Date of this review
is_open	Weather the restaurant is open	stars	Yelper rating levels
categories	Classify by cuisine, abstention, etc.	text	Review text
review_count	Number of reviews	cool	"cool" counts this review received
name	Name of the restaurant	funny	"funny" counts this review received
stars	Business/Restaurant rating levels	useful	"useful" counts this review received
postal_code	Restaurant's location		

Proposed Research Questions

- Identify top-rated geographic area for different business type.
- Examine the most important attributes affect people's choice among identical restaurants.
- Predict STAR ratings that consumer would score based on business attributes and their reviews.
- Identify the common words that consumer tend to write in their reviews for certain businesses.

Top-Rated Geographic Areas



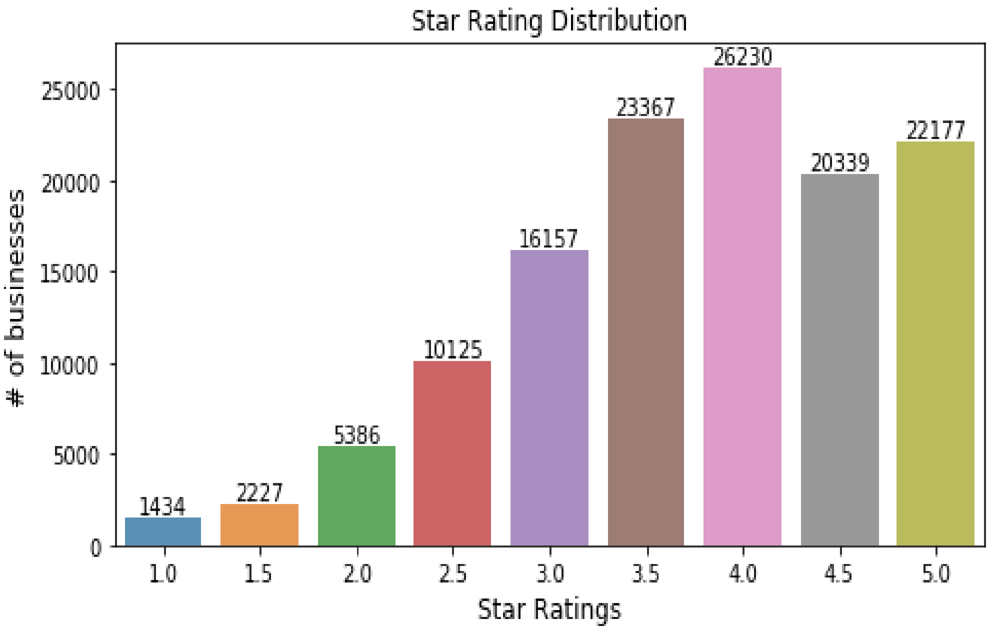


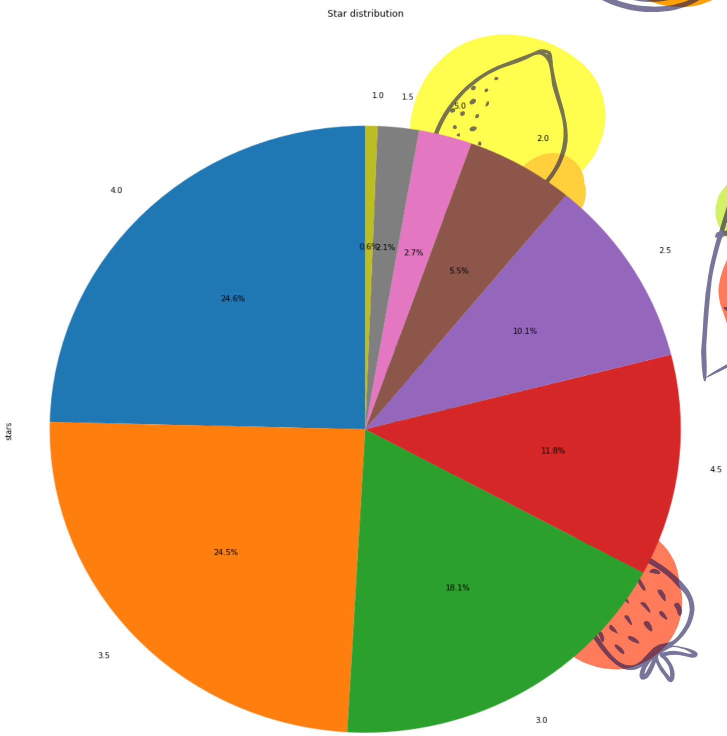
Top-Rated Geographic Areas' Business

categories <chr>	n <int>
Restaurants, Pizza	1092
Pizza, Restaurants	1060
Coffee & Tea, Food	1036
Nail Salons, Beauty & Spas	1015
Beauty & Spas, Nail Salons	981
Food, Coffee & Tea	966
Restaurants, Mexican	932
Mexican, Restaurants	908
Beauty & Spas, Hair Salons	893
Restaurants, Chinese	889

state <chr>	categories <chr>	n <int>
AZ	Restaurants, Mexican	394
ON	Restaurants, Chinese	305
NV	Mexican, Restaurants	181
OH	Pizza, Restaurants	156
PA	Restaurants, Pizza	151
QC	Restaurants, Pizza	102
AB	Restaurants, Pizza	86
NC	Restaurants, Chinese	82
WI	Restaurants, Mexican	37
IL	Restaurants, Mexican	21

Stars Distribution





Data Pre-processing

- Selected the business who have "restaurant" type.
- Flatten all the attributes.
- Delete the attributes which have a lot of Null values
- One hot encoding.

Classification Models

- Label the good restaurants data and bad restaurants data.
  - Good(star 5.0/ star 5.0 and 4.5)
  - Bad(star 1.0,1.5 and 2.0 /star 1.0,1.5,2.0 and 2.5)
- Models
- Split training and testing dataset
- 10 cross validation

Classification Models

Model	Training Accuracy	Testing Accuracy	Recall- Good	Recall- Bad	Model Parameter
GaussianNB	0.530	0.519	0.06	0.99	
Logistic Regression	0.734	0.705	0.69	0.72	LogisticRegression(C=10.01, class_weight=None,dual=False, fit_intercept=True, intercept_scaling=1, max_iter=100,multi_class='ovr', n_jobs=1, penalty='l2', random_state=None, solver='liblinear', tol=0.0001,verbose=0, warm_start=False)
linear SVC	0.725	0.692	0.66	0.72	LinearSVC(C=0.01, class_weight=None, dual=True, fit_intercept=True,iintercept_scaling=1, loss='squared_hinge', max_iter=1000, multi_class='ovr', penalty='l2', random_state=None,tol=0.0001,verbose=0)
Random Forest	0.823	0.721	0.75	0.70	RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini', max_depth=8,max_features='auto', max_leaf_nodes=None,min_impurity_decrease=0.0, min_impurity_split=None, min_samples_leaf=1, min_samples_split=2,min_weight_fraction_leaf=0.0, n_estimators=500,n_jobs=1, oob_score=False, random_state=42, verbose=0,warm_start=False)
KNN	0.780	0.677	0.59	0.77	KNeighborsClassifier(algorithm='auto', leaf_size=30, metric='minkowski', metric_params=None, n_jobs=1, n_neighbors=5, p=2,weights='uniform')

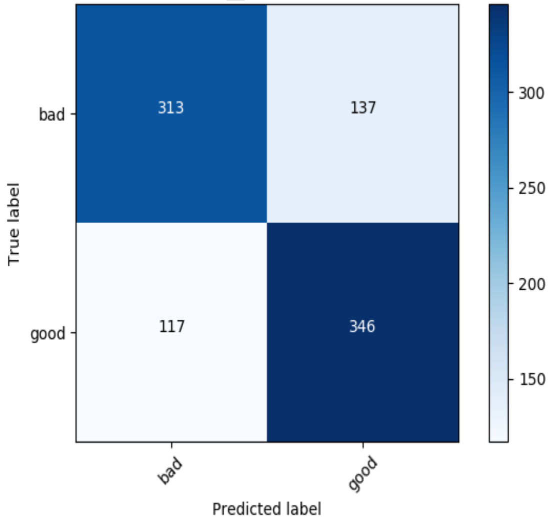
Classification Models - Result

- Less attributes sometimes will improve the model's performance, because there are a lot of Null values in these attributes.
- If we define the bad and good more extremely, we can get higher accuracy.
- The average accuracy of these models is less than 0.8
- In all of the models, we got our best model : Random Forest Model.

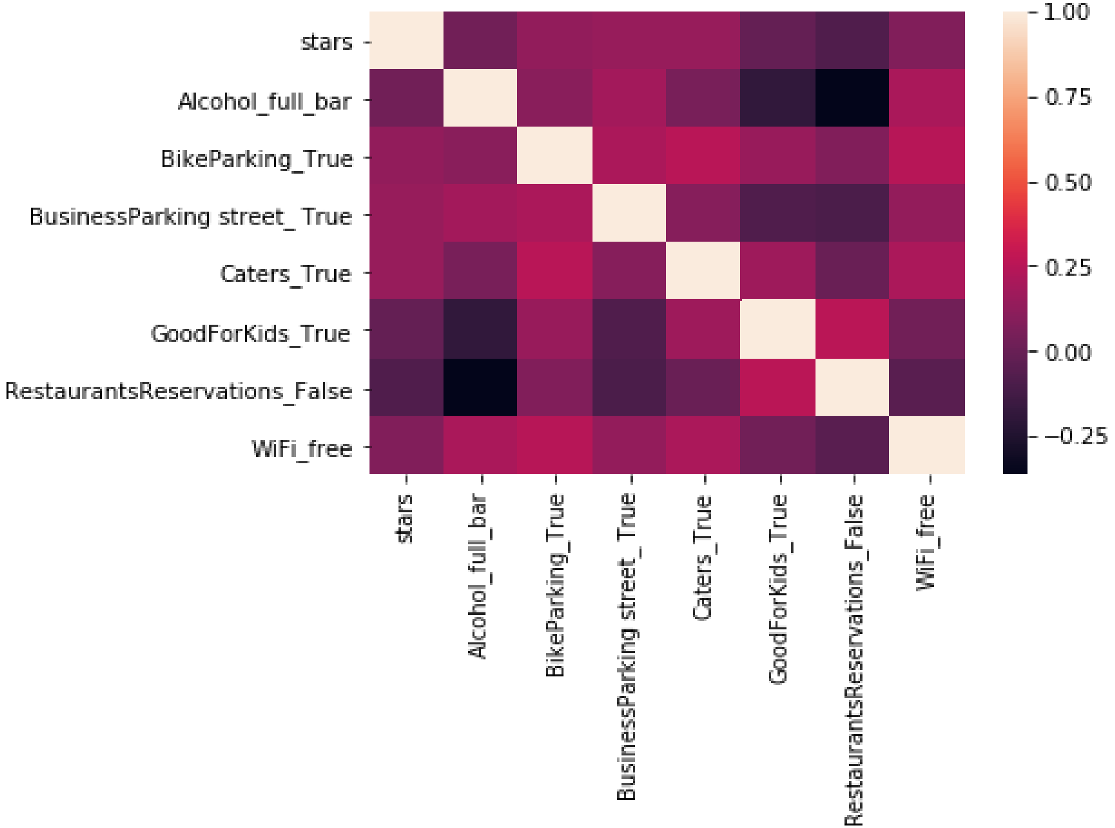
```
Training Accuracy : 0.822838345864661б
Classification Report:
             precision    recall  f1-score   support

        0.0       0.73      0.70      0.71       450
        1.0       0.72      0.75      0.73       463

avg / total       0.72      0.72      0.72       913

Testing Accuracy : 0.7217962760131434
```

0	BusinessParking street_ True
1	RestaurantsReservations_False
2	BusinessAcceptsCreditCards_True
3	RestaurantsAttire_casual
4	RestaurantsGoodForGroups_False
5	RestaurantsReservations_True
6	BikeParking_True
7	NoiseLevel_quiet
8	OutdoorSeating_False
9	OutdoorSeating_True

Confusion matrix__RandomForest Classifier



Correlation



Linear Regression

Feature	Coefficients
BusinessParking street_ True	-4.931739e+12
BusinessParking valet_ True	4.047394e+12
Ambience classy_ True	3.362815e+12
GoodForMeal breakfast_ True	-3.256624e+12
Ambienceromantic_ True	-2.815944e+12
Training MSE: 0.7412492567199715 Test MSE: 0.7572142703112591	

K-Fold Cross Validation:

```
10-fold RMSEs:
[0.7376264896385099, 0.7534747627960025, 0.7410598543565764, 0.736387835472943, 0.7600223523964008, 0.7342258835840773, 0.7399720340524605, 0.7492238293286, 0.7526578556698972, 0.7547683587490751]
CV RMSE:
0.745991924440831
Std of CV RMSE:
0.012895228941877555
3.4353085182042027
```

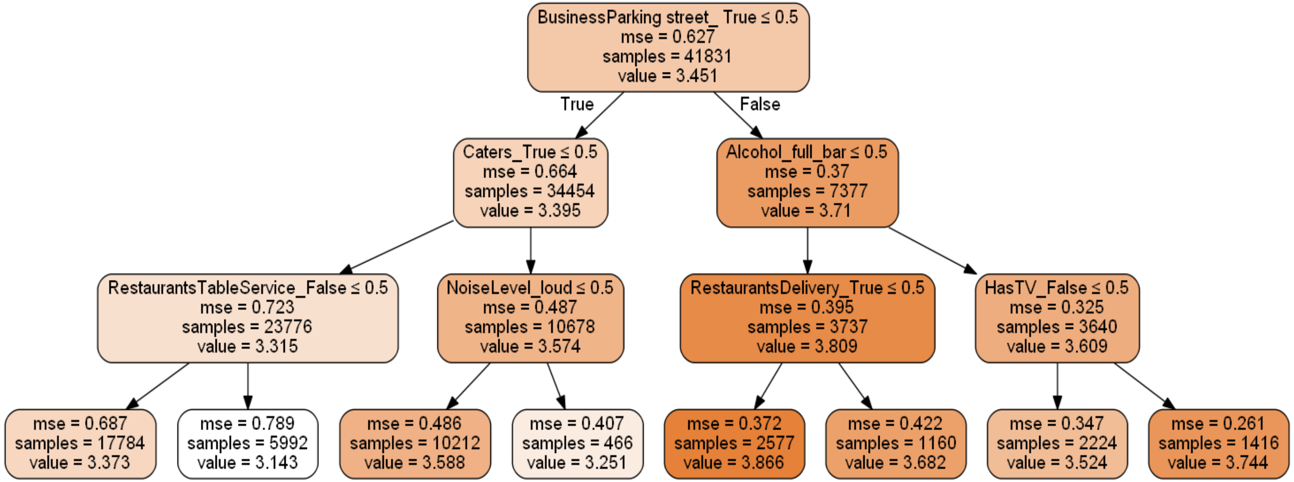
Ridge Regression

Without Polynomial Features:	MSE:0.7458908459319094
With Polynomial Features:	MSE:0.7289958883334875

Support Vector Regression

Training MSE	0.7360020615112289
Test MSE	0.7572142703112591

Decision Tree Regression



Feature	Importance
BusinessParking street_ True	0.395462
Cater_True	0.322624
RestaurantsTableService_False	0.155524
Alchohol_full_bar	0.048029
NoiseLevel_loud	0.033251

Regression Model Selection

Feature	Coefficients
BusinessParking street_ True	-4.931739e+12
BusinessParking valet_ True	4.047394e+12
Ambience classy_ True	3.362815e+12
GoodForMeal breakfast_ True	-3.256624e+12
Ambienceromantic_ True	-2.815944e+12

- Training MSE
- Test MSE

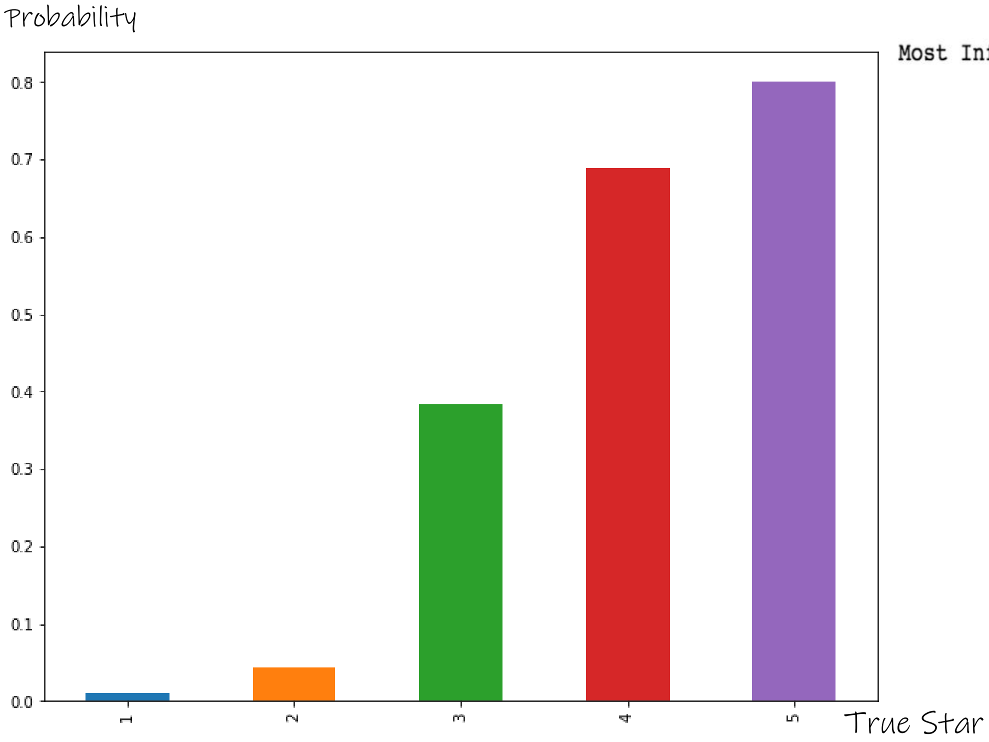
Ridge Regression

- Without Polynomial Features
- With Polynomial Features

Reviews Exploration – Word Cloud


```
good
negative
Displaying 1 of 16376 matches:
e twice locat desert ridg first time good last time veri disappoint order pepp
positive
Displaying 1 of 23305 matches:
look good authent chines food care craft tri p
```

Reviews Exploration – Naïve Bayes




```
Most Informative Features
        appalled = True           neg : pos    =     40.5 : 1.0
        demanded = True           neg : pos    =     26.5 : 1.0
       unhelpful = True           neg : pos    =     24.5 : 1.0
        pathetic = True           neg : pos    =     24.2 : 1.0
        shrugged = True           neg : pos    =     21.9 : 1.0
       cockroach = True           neg : pos    =     21.5 : 1.0
          rudest = True           neg : pos    =     21.3 : 1.0
          argued = True           neg : pos    =     21.2 : 1.0
      deliciously = True          pos : neg    =     21.1 : 1.0
           nerve = True           neg : pos    =     20.7 : 1.0
       defensive = True           neg : pos    =     20.5 : 1.0
          rancid = True           neg : pos    =     20.5 : 1.0
        disgrace = True           neg : pos    =     19.5 : 1.0
        vomiting = True           neg : pos    =     19.5 : 1.0
         arguing = True           neg : pos    =     19.5 : 1.0
       insulting = True           neg : pos    =     18.9 : 1.0
     incompetent = True           neg : pos    =     18.6 : 1.0
            puke = True           neg : pos    =     18.5 : 1.0
       apathetic = True           neg : pos    =     17.5 : 1.0
          hazard = True           neg : pos    =     17.5 : 1.0
```

Reviews Exploration

- Word2Vector

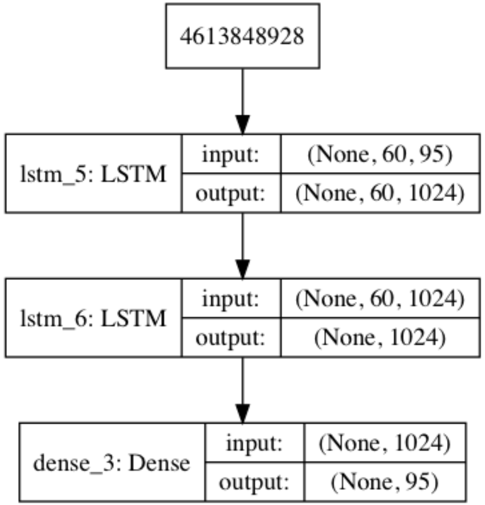
```
word_algebra(add=[u'breakfast', u'lunch'])
```

```
brunch
```

```
get_related_terms(u'service')
```

staff	0.666
food	0.594
customer	0.591
attentive	0.526
atmosphere	0.526
ambience	0.52
prices	0.517
friendly	0.513
prompt	0.501
polite	0.495

- LSTM



Conclusion

What makes a great restaurants ?

- Location
- Parking – Even bike parking
- Service seems more important than the food.
- Generate 5 star reviews by yourself!

Further Work

1. More models or more encoding methods can be tried to improve the accuracy of business.
2. We didn't go too in depth with each model when we trained, maybe there will be more parameters could be tuned in the future.
3. We could establish times series model to examine how previous reviews impact the future review.
4. For the review generator part, word sequence models can be tried.