Food trend

East Cost(NYC) vs West Coast(LA)

Chris Lee

Agenda

1. Data collecting - Collect posts using web API (1,000 posts from each subreddit)

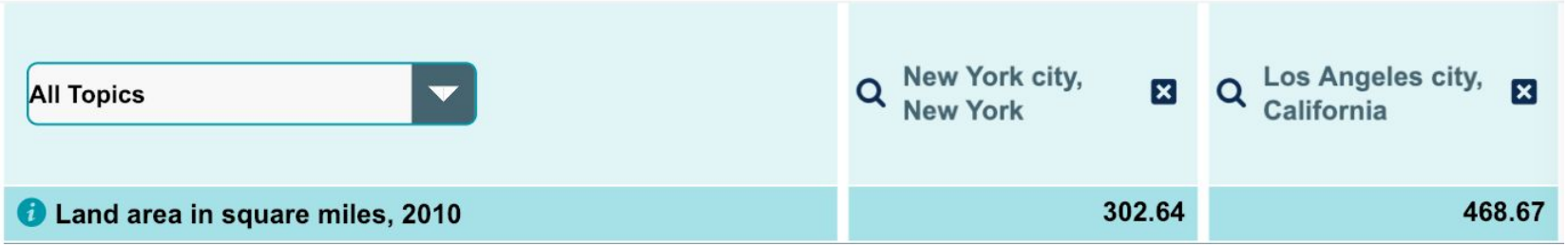
2. Data cleaning - Clean the data and convert text into words by using vectorizer

3. Modeling & Evaluating - Use classification model and evaluate
   (RandomforestClassification Model +)

4. Conclusion & Recommendation - Interpret the result and make a recommendation

Population and Land area

All Topics	New York city, New York	Los Angeles city, California
Population estimates, July 1, 2017, (V2017)	8,622,698	3,999,759



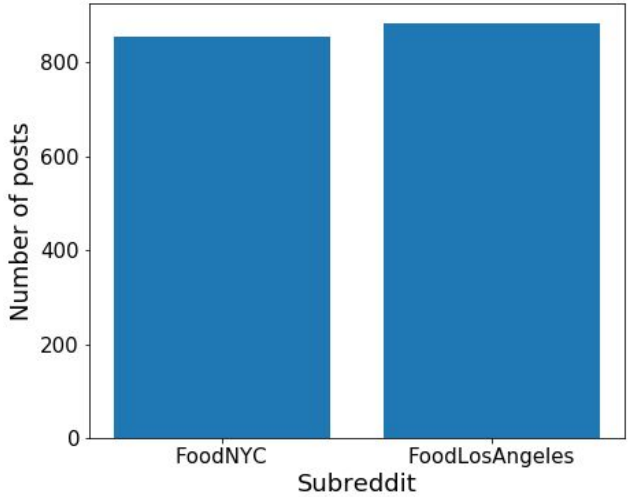
All Topics	New York city, New York	Los Angeles city, California
Land area in square miles, 2010	302.64	468.67

QuickFacts, *Census,* Last updated : 12/19/18, Retrieved from https://www.census.gov/quickfacts/fact/table/US/PST045218

1. Data Collecting - Reddit API

Two Sub-reddits

- FoodNYC - 856 posts
- FoodLosAngeles - 882 posts



- **URL** + **.json** = html text document
- 25 post per request - Iterate 40 times to get 1,000 posts

2. Data Cleaning (Natural Language Processing)

- Tokenizing - the process of breaking a stream of text up into words, phrases, symbols, or other meaningful elements called tokens

  ex) ‘How are you?’ -> ‘How’, ‘are’, ‘you’, ‘?’

- Lemmatizing / Stemming - forms of shortening words that attempt to return their *lemma*, or the base/dictionary form of a word

  ex) running -> run, ran -> run, cats -> cat

3. Modeling and Evaluation

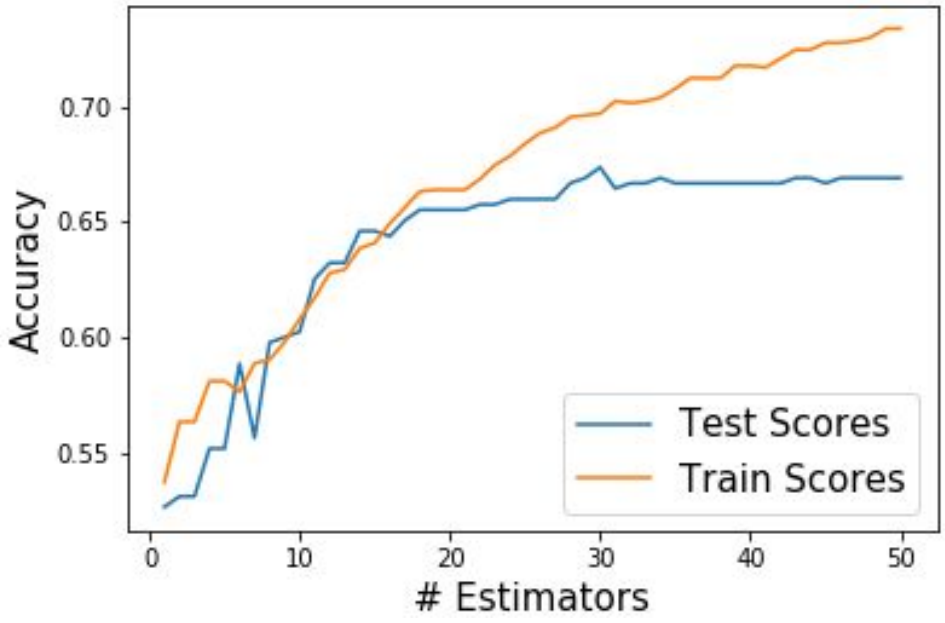
Top 5 model by Accuracy score

(with default parameters setting)

1. Logistic Regression
2. ExtraTree
3. Multinomial Naive Beyes
4. AdaBoost
5. GradientBoosting

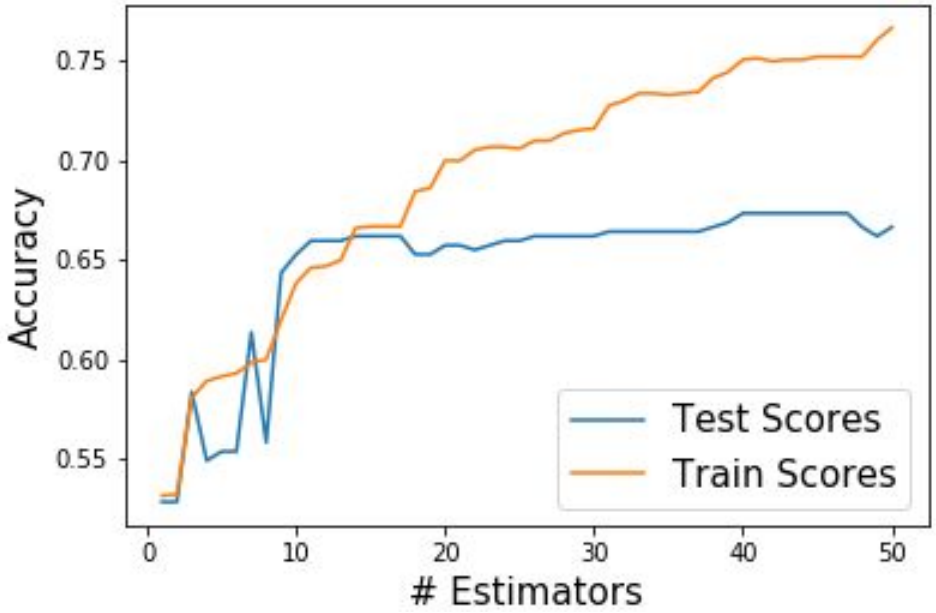
	Accuracy
LogisticRegression	0.735632
ExtraTree	0.712644
MultinomialNaiveBeyes	0.708046
AdaBoost	0.673563
GradientBoosting	0.673563
BaggingTree	0.645977
RandomForest	0.622988
DecisionTree	0.611494
KNeighbor	0.593103
SupportVectorMachine	0.498851

3. Modeling and Evaluation



	Accuracy
LogisticRegression	0.735632
ExtraTree	0.712644
MultinomialNaiveBeyes	0.708046
AdaBoost	0.673563
GradientBoosting	0.673563
BaggingTree	0.645977
RandomForest	0.622988
DecisionTree	0.611494
KNeighbor	0.593103
SupportVectorMachine	0.498851

3. Modeling and Evaluation



	Accuracy
LogisticRegression	0.735632
ExtraTree	0.712644
MultinomialNaiveBeyes	0.708046
AdaBoost	0.673563
GradientBoosting	0.673563
BaggingTree	0.645977
RandomForest	0.622988
DecisionTree	0.611494
KNeighbor	0.593103
SupportVectorMachine	0.498851

3. Modeling and Evaluation

Gridsearch on Logistic regression for the hyperparameter setup

- penalty = specify the norm used in the penalization / default = l2
- C = Inverse of regularization strength / default = 1.0
- tol = tolerance for stopping criteria / default = 0.0001

3. Modeling and Evaluation

Random Forest Classification

n_estimators: [10, 20, 50, 100],

max_depth: [None, 2, 3, 4],

Max_features: [auto, 0.5]

After tuning -> 0.678434

	Accuracy
LogisticRegression	0.735632
ExtraTree	0.712644
MultinomialNaiveBeyes	0.708046
AdaBoost	0.673563
GradientBoosting	0.673563
BaggingTree	0.645977
RandomForest	0.622988
DecisionTree	0.611494
KNeighbor	0.593103
SupportVectorMachine	0.498851

3. Modeling and Evaluation

Logistic Regression

penalty: [l1, l2],

tol: [0.000001, 0.00001, 0.0001, 0.001, 0.01],

C: [100, 10, 1, 0.1, 0.01]

After tuning -> 0.735632 (did not change!)

	Accuracy
LogisticRegression	0.735632
ExtraTree	0.712644
MultinomialNaiveBeyes	0.708046
AdaBoost	0.673563
GradientBoosting	0.673563
BaggingTree	0.645977
RandomForest	0.622988
DecisionTree	0.611494
KNeighbor	0.593103
SupportVectorMachine	0.498851

4. Conclusion

	coef	Ratio
queen_x	1.284854	3.614141
brooklyn_x	1.193415	3.298325
manhattan_y	1.184725	3.269787
manhattan_x	1.098076	2.998391
star_x	1.037865	2.823184
midtown_x	0.964830	2.624341
infatu_x	0.898599	2.456161
flush_x	0.873470	2.395208
ate_x	0.870935	2.389143
brooklyn_y	0.830592	2.294677
momofuku_x	0.826912	2.286249
omakas_x	0.825746	2.283584
cooki	0.818797	2.267771

< Top features

Ex) Omakase has 2.28 time more chance to be called in NYC than LA.

Bottom features >

Ex) Toast has 0.6 time less change to be called in NYC than LA.

	coef	Ratio
toast_x	-0.513436	0.598436
make_x	-0.514424	0.597845
south_x	-0.516736	0.596464
chili_x	-0.520904	0.593983
lo_y	-0.520937	0.593964
thing_x	-0.524165	0.592049
marin	-0.524833	0.591654
oak_x	-0.528460	0.589512
socal_x	-0.530735	0.588173
place get_x	-0.535366	0.585455
santa_y	-0.536147	0.584998
langer	-0.542471	0.581310
venic_x	-0.544352	0.580218

4. Conclusion

- Meaningful features from Top 50 features (more relevant to NYC)
  - Momofuku, omakaze, chinese, cookie, bagel, bar, rib,
  - restaurant week, food fest, italian, michelin, vegan

- Meaningful features from Bottom 50 features (more relevant to LA)
  - Toast, taco, chili, burger, shake shack, potato, donut, tsujita
  - Hollywood, cafe

5. Recommendation

1. For NYC (East coast)
   - Omakaze, bagel, and italian food are popular in NYC
   - Use food fest, restaurant week, michelin star
   - Better to have vegan menu
2. For West coast
   - Burger, taco, chili are popular in LA
3. For both
   - Japanese restaurant is familiar for EC and WC in general.