Classifying and Predicting Houston and Austin Subreddits

By Eric Swanson

Project 3. 9/13/2021

Location Map ~150 miles (2.5 hour drive)

Photo by website

Quick Fact Comparisons

	Austin	Houston
Population	961,855	2,304,580
Land size	305 sq mi	637.4 sq mi
economy	86 billion GDP	478 billion GDP
Park area	29,000 acres	56,405 acres
Population growth rate	21% last 10 years	9.8% last 10 years
Reddit sub Users	254,518	258,138

Source: Wikipedia Austin, Houston.

In 2013, Austin was the most active city on Reddit, having the largest number of views per capita.

Problem Statement

Building a classification model to determine which city a reddit post should go in can have many uses. City officials, election candidates can look at what people are enjoying, or complaining about and are policies working. An example might be if you approved spending on upgrading your park system are people using it and enjoying it.

In this project we will look at posts from both subreddits Austin and Houston and using Logistic Regression and Random Forest classification models to determine which city a post text should be classified in. Success will be determined on the accuracy of the model.

Preprocessing EDA

Pulled12,100 post from Austin and was able to keep 6,052
Pulled 15,100 post from Houston and was able to keep 6,044

Dropped rows that had embedded http web links

Removed [\n, &x200b, and apostrophes]

Final Count: 5,441 houston and 5,565 Austin posts

Added the city names to the list of stop words.

Workflow

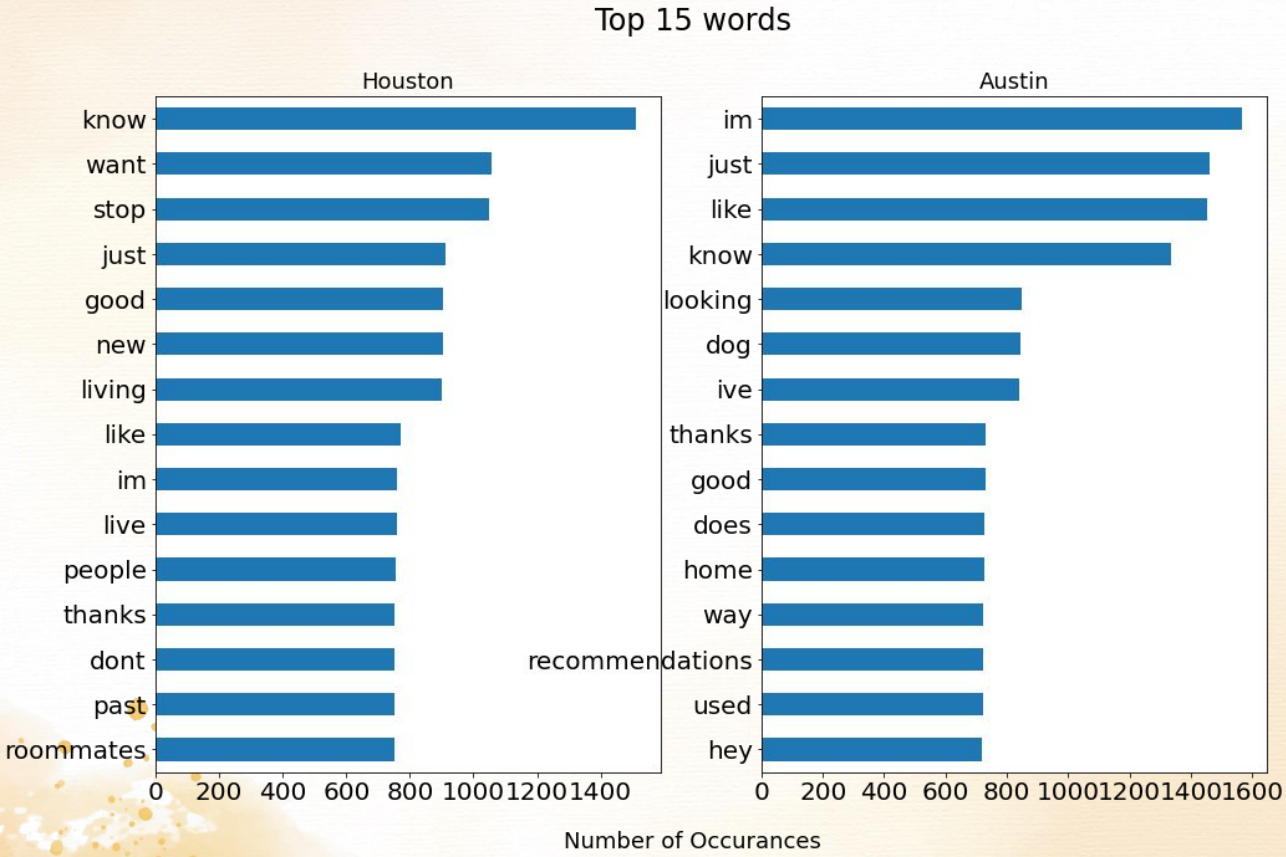
Collect Data and EDA

Tokenize and lemmetize:
Countvectorize
TFIDF

Models:
Random Forest
Logistic Regression

Evaluation:
Accuracy
Feature Importance
Confusion Matrix

Top Words from Countvectorizer



Summary of 20 Most Used Words

Unique Houston	Unique Austin	In Both
Want	Looking	Know
Stop	Dog	Just
New	Ive	Good
Living	Does	Like
Live	Home	Im
People	Way	people
Dont	Recommendations	Thanks
Past	Used	does
Roommates	Hey	
House	Park	
Area	Place	
things	state	

All models had 99% Accuracy For Both Train and Test

Note: because it is a balanced dataset baseline is 50%
Used a Pipeline and Gridsearch limiting Random Forest with both CountVectorizer and TFIDF

Parameters:

CountVectorizer
Min_df: 1,2
Ngram_range (1,1), (1,2)

Random Forest
N_estimators: 100, 150, 200
Max_deth: [none, 1, 2, 3, 4, 5]
Max_features: [sqrt, .5]

CountVectorizer
Min_df: 1
Ngram_range (1,2)

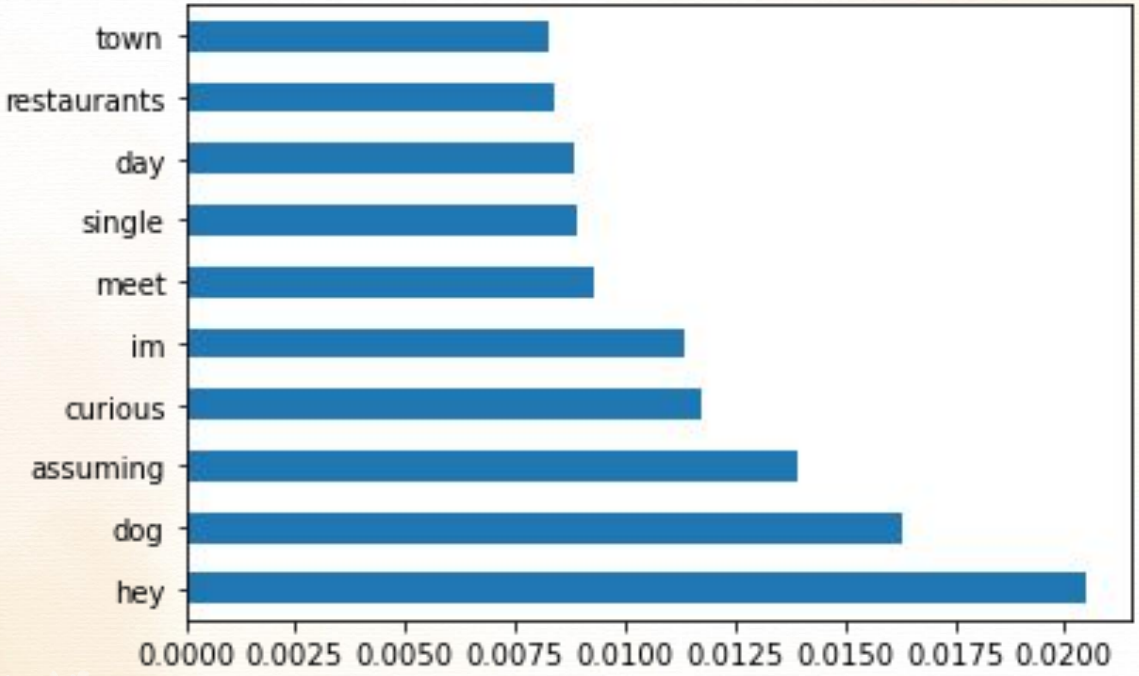
Random Forest
N_estimators: 100
Max_deth: [none]
Max_features: [0.5]

Model Results 99% Accuracy For Both Train and Test

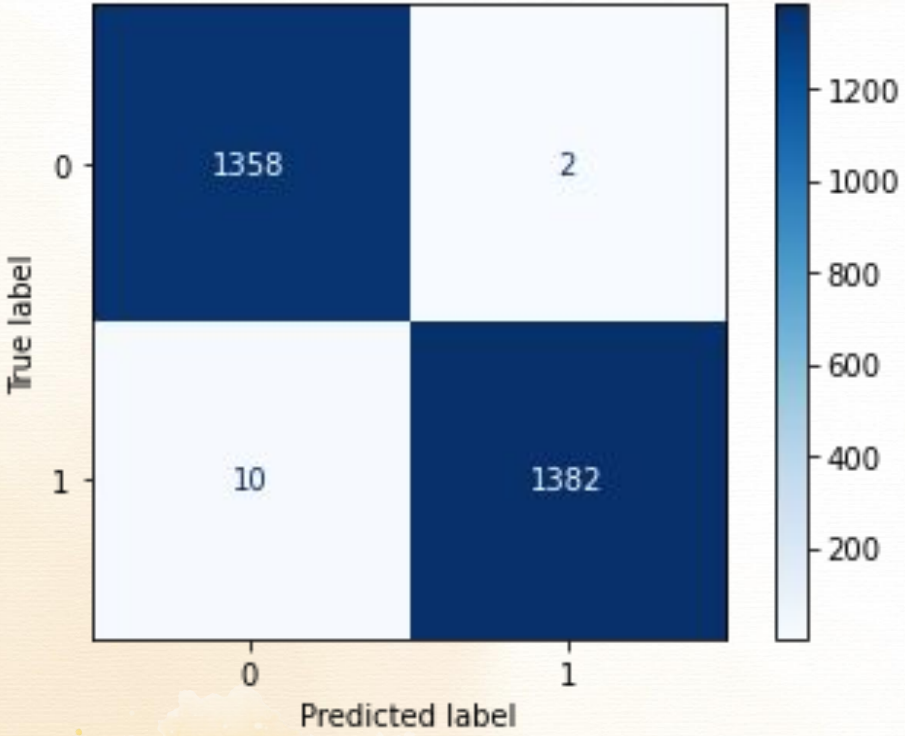
Feature Importance

Logistic Regression	RF with CV	RF TFIDF
assuming 6.989675	hey 0.069636	im 0.013394
little 6.430970	dog 0.060130	like 0.012934
hey 6.044143	assuming 0.050627	hey 0.012555
south 5.397371	question 0.030884	dog 0.011381
yall 4.898522	park 0.024818	good 0.011201
curious 4.812134	passed 0.024511	open 0.009202
open 4.050332	thanks advance 0.022411	need 0.008520
question 3.737335	bars 0.020156	place 0.008224
games 3.306832	tonight 0.018384	assuming 0.007673
floor 3.193121	restaurants 0.018120	right 0.007664

Random Forest Feature Importance



Look at Confusion Matrix



12 not classified correctly:

2 (false Positive) guessed Austin and they were Houston

10 (false negative)guessed Houston that were Austin

Investigate the missed posts

Guessed Austin and was Houston:

Have been trying my luck at freshwater fishing the past few weeks and have come up empty. I've been toe Dwight D park, Buffalo Run, and Tom Bass. Are there any other places to catch largemouth bass?

Investigate the missed posts

Guessed Houston and was Austin:

I’m aware of the spots around Festival Beach but I live in south ~~Austin~~ and would love to avoid crossing the bridge if possible. Thanks for ideas!

I noticed yesterday that the city put some speed bumps on the road into Roy G (Grove St). They start just after the bus turnaround, and they are **no joke**. Quite jarring at 20mph, slow down to 5mph.I guess Im as guilty as anyone else for speeding on that road, but it seems to me that the bumps are not a great addition. Have there been collisions there causing problems, or is it just concerned citizens boofing us?

Just moved out here and want to start exploring. I used to traditional camp back in the NC mountains but obviously it’s so hot here it isn’t the same. Are there any cool spots I could go and park the car and hike during the day then camp in the car?

Sentiment Analysis

Bonus: Used nltk sentiment analyzer to look at all the self text posts -

Houston compound score total: 1,100
Austin compound score total: 1,253

Remember the compound adds up the total count of negative and positive words on a scale form -1 negative to 1 positive. Above is 153 difference.

Conclusion

Overall:
With high accuracy and low false positives and false negatives I think I have a good model for predicting what someone might post about in one city or another.

An Example of use:
Reddit user posts might not be a true representation of the overall population, however an example of spending and sentiment - Houston has spent 200 million on upgrading one of it's largest parks (Memorial park), and allocated an additional 155 million toward upgrading parks around the city yet **"park"** doesn't show up in the top 50 words used for the city.

Sources: houston parks, culturemap story

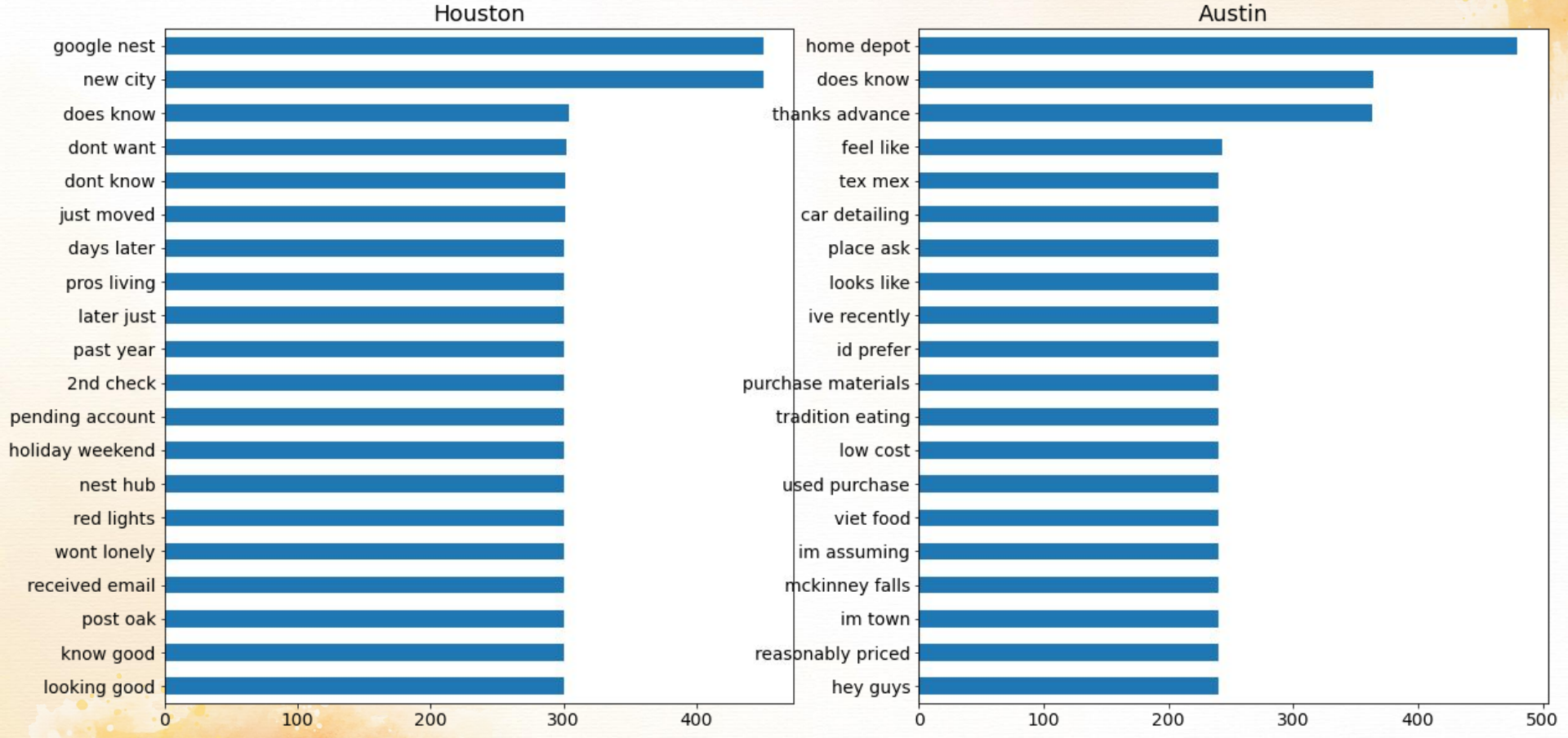
Additional Work

- Another use can be for an individual to see how similar or different their current city is to a potential new location to live.
  - Vibe, or interests
- Look into if any information available on Reddit users base that might indicate any pattern or trend.

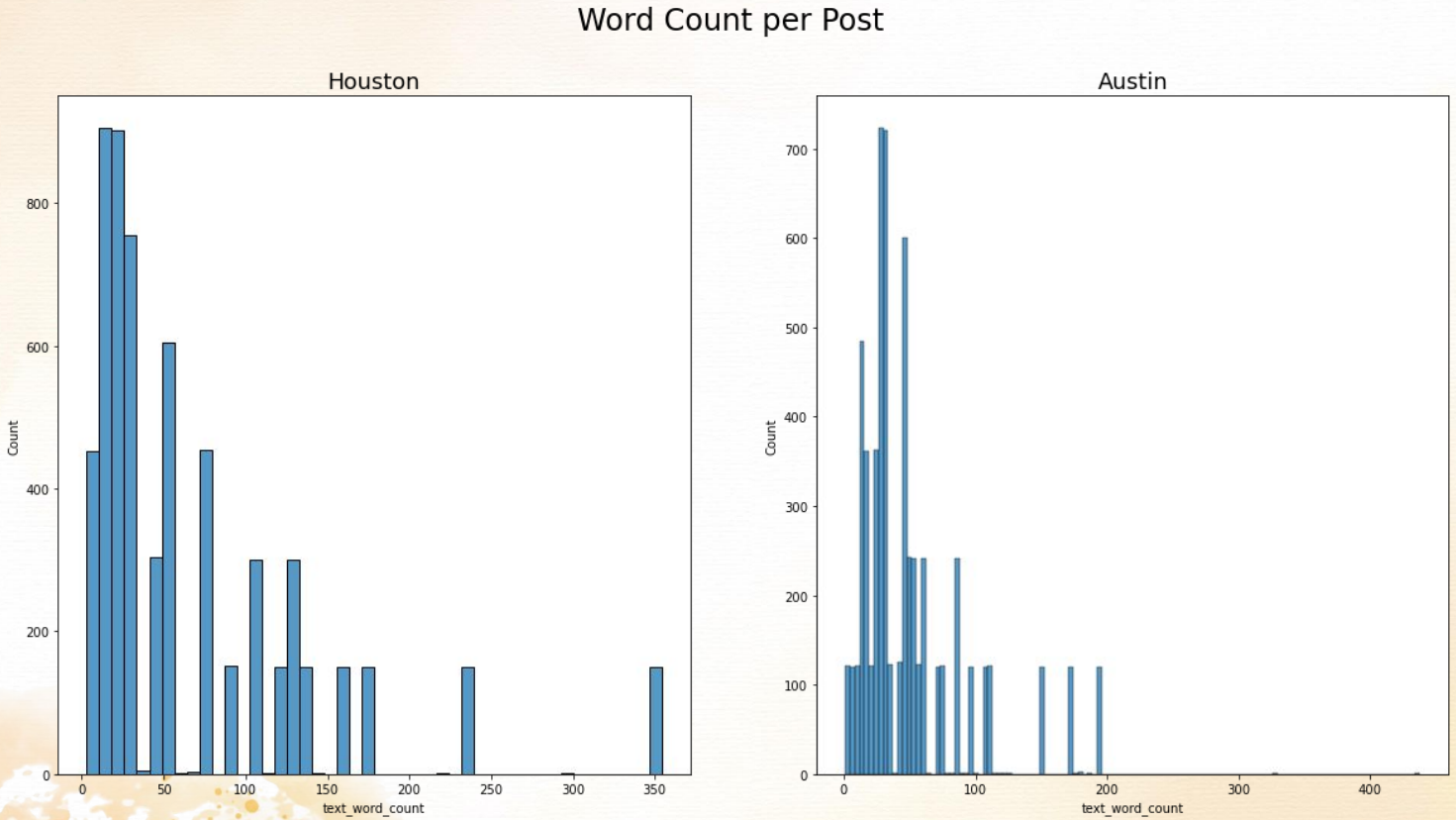
Questions?

Ngrams 2,2

Top 20 words



Word Count per Post



Models:

Random Forest

Logistic Regression

Technical Extra Slides

Parameters run for Random Forest on both CountVectorizer and TFIDF

```
    'tvec__min_df': [1,2],
    'tvec__ngram_range': [(1,1), (1,2)],
    'rf__n_estimators': [100, 150, 200],
    'rf__max_depth': [None, 1, 2, 3, 4, 5],
    'rf__max_features': ['sqrt', .5]
```

```
{'rf__max_depth': None,
 'rf__max_features': 'sqrt',
 'rf__n_estimators': 200,
 'tvec__min_df': 1,
 'tvec__ngram_range': (1, 2)}
```

```
{'cvec__min_df': 1,
 'cvec__ngram_range': (1, 2),
 'rf__max_depth': None,
 'rf__max_features': 0.5,
 'rf__n_estimators': 100}
```