Predicting Venue Like Count Using Natural Language Processing of Venue "Tip" Text

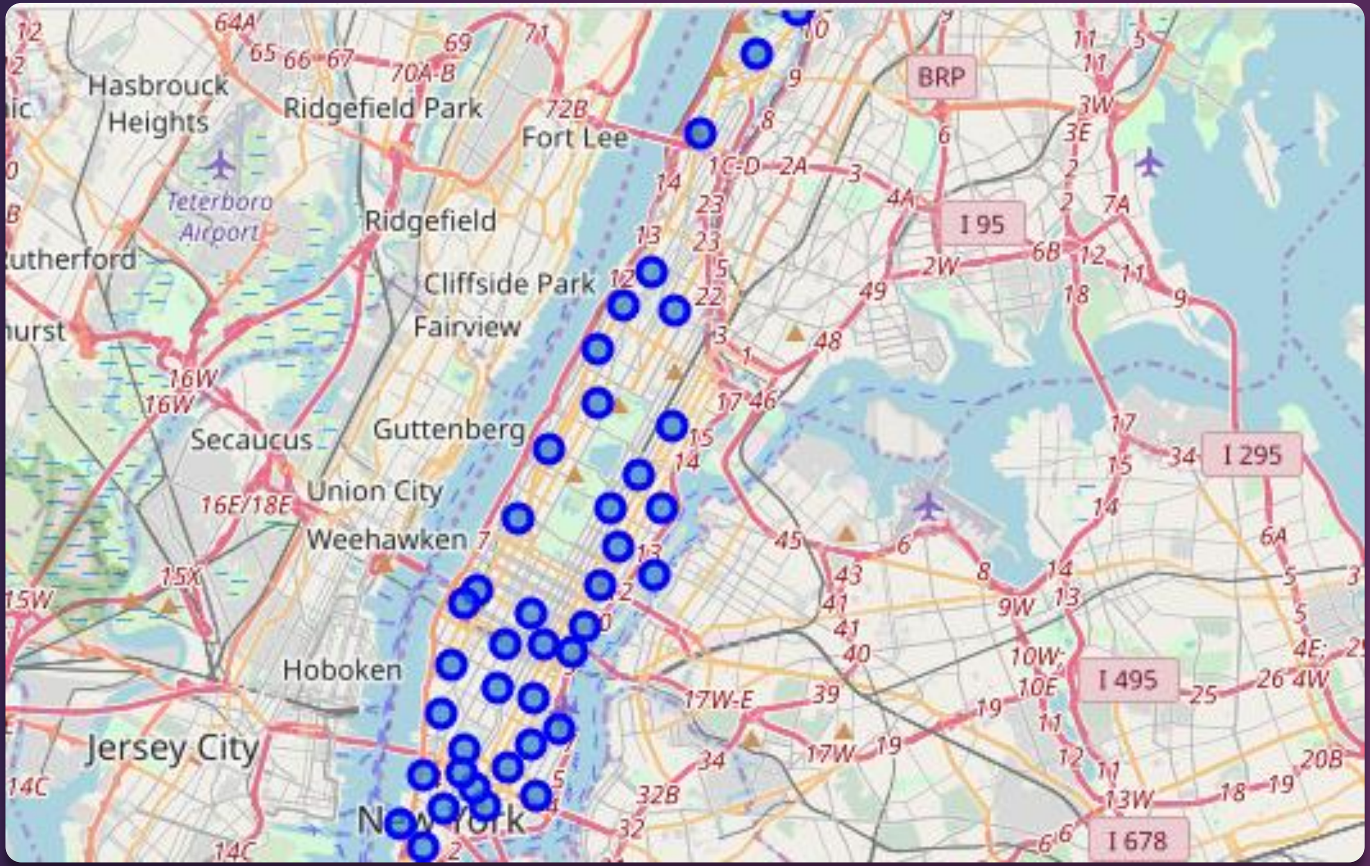
ANDREW BARBER

Introduction

- When it comes to user recommendations, utilizing features of the items in question (rating of movies, comments on products, like count for a venue, etc.)
- In our case, we hoped to utilize the textual information associated with Manhattan venues to help guide users to high-quality establishments in the absence of venue-features that more directly describe popularity

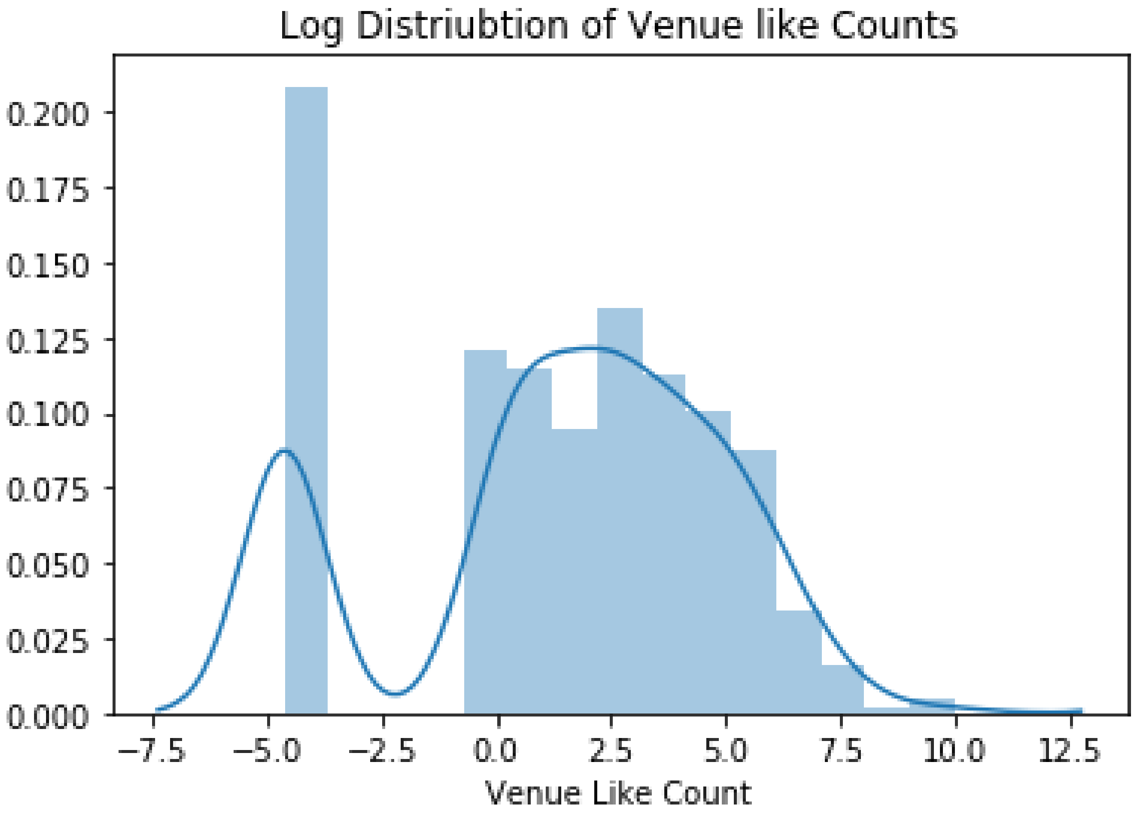
Problem Definition and Data

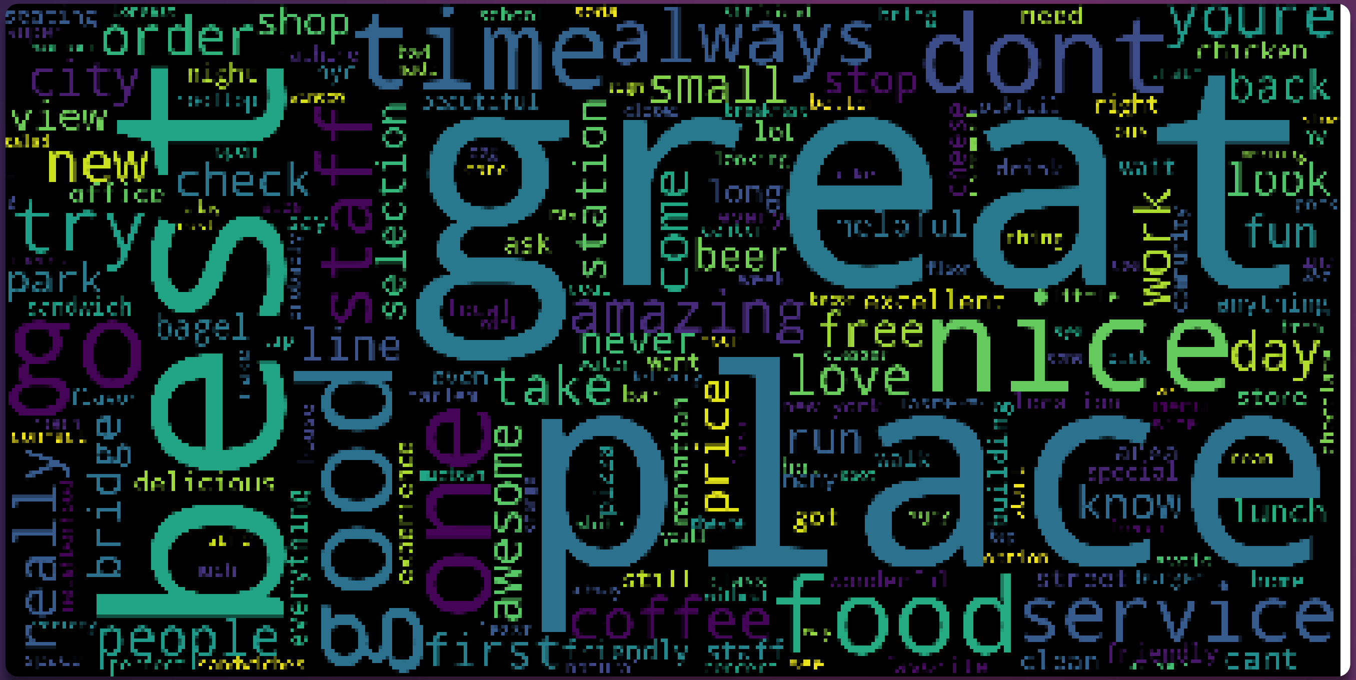
- The purpose of this project is to use natural language processing of venue comments ("tips") to predict corresponding venue like counts
- Data was taken from venues scattered around the 40 major neighborhoods of Manhattan (https://cocl.us/new_york_dataset)
- Data for the closest 12 venues was then extracted from https://developer.foursquare.com/, including venue like counts and "Tips"
  - Because like counts followed log-normal distribution, using log(like counts) would result in a more useful model



Manhattan Neighborhoods

Manhattan Venues


Log Distriubtion of Venue like Counts
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0
12.5
Venue Like Count



Word Cloud of Venue Tip Text Vocabulary

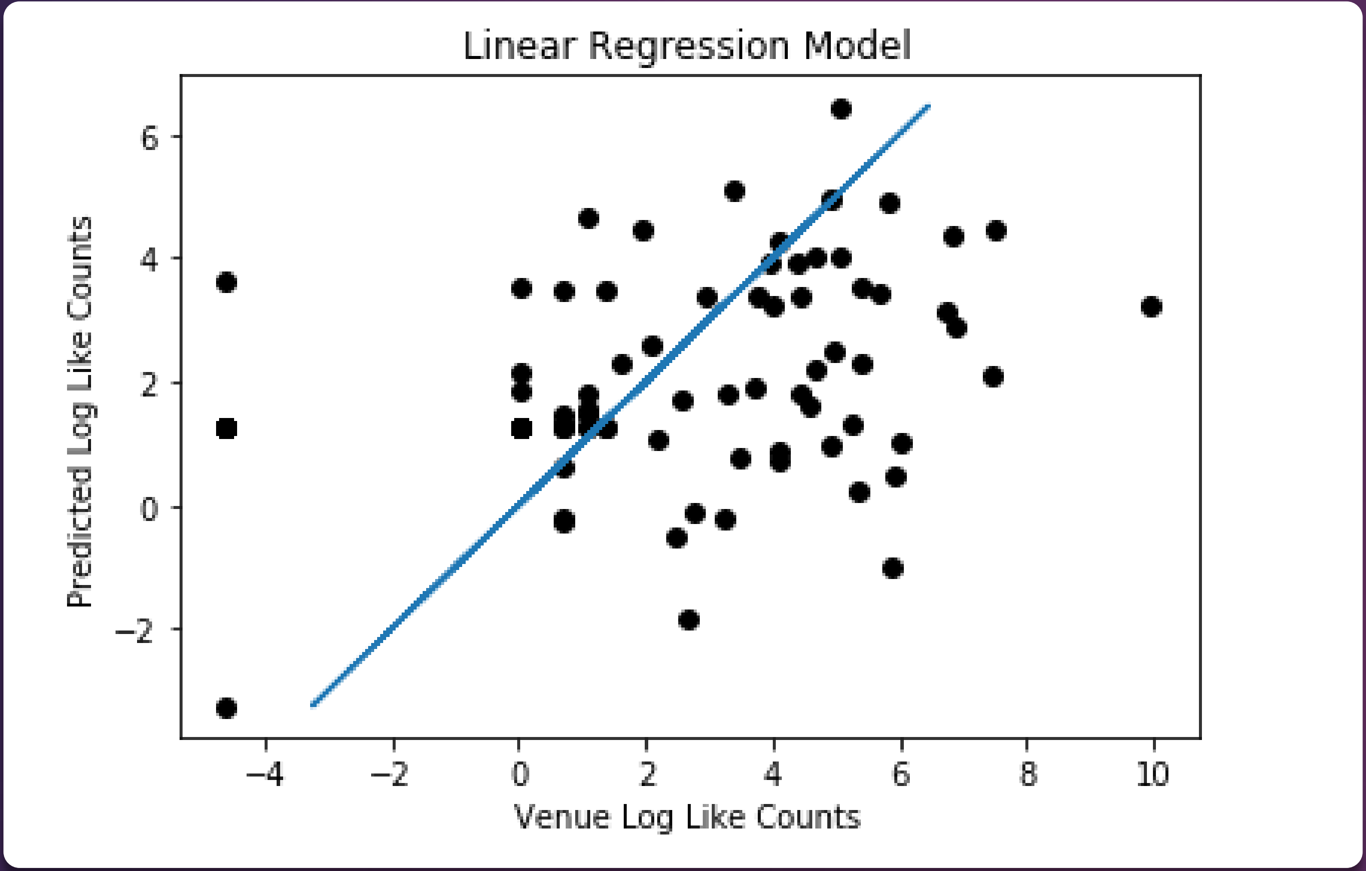
Methodology

- Remove all venues with incomplete like counts (480 venues to 459 venues)
- Remove all punctuations and capitals from tip text
- Tokenize sentences
- Concatenate tokens for each individual venue
- Represent venue tip text using bag-of-words (BOW) model
- Baseline is the average like count of all venues in the test set
- Use BOW tip text representation to implement simple linear regression model
- Evaluate linear regression model against baseline

Evaluation and Results

- Used mean squared error (MSE) between predicted log(like counts) and true log(like counts)
- Results:

Model	MSE
Baseline – Average Like Counts	
Linear Regression (BOW)	



Linear Model Predicted Values Versus True Values

Discussion

- **Limitations:**
  - Dataset is small – only 459 records
  - Dataset is narrow in scope – only looked at venues in Manhattan
- **Future Directions:**
  - Expand the size and scope of the dataset
  - Utilizing different text-representation methods (term-frequency independent-document-frequency (TFIDF), word embeddings, pre-trained word embeddings)
  - Utilizing different regression models (recurrent neural networks such as long-short-term memory, convolutional neural networks)

Conclusion

- Our model successfully predicted venue log like counts with lower MSE than our baseline
- This project presents a promising proof-of-concept when it comes to NLP approaches to user recommendation problems
- There are many future directions for this subject with auspicious outlook