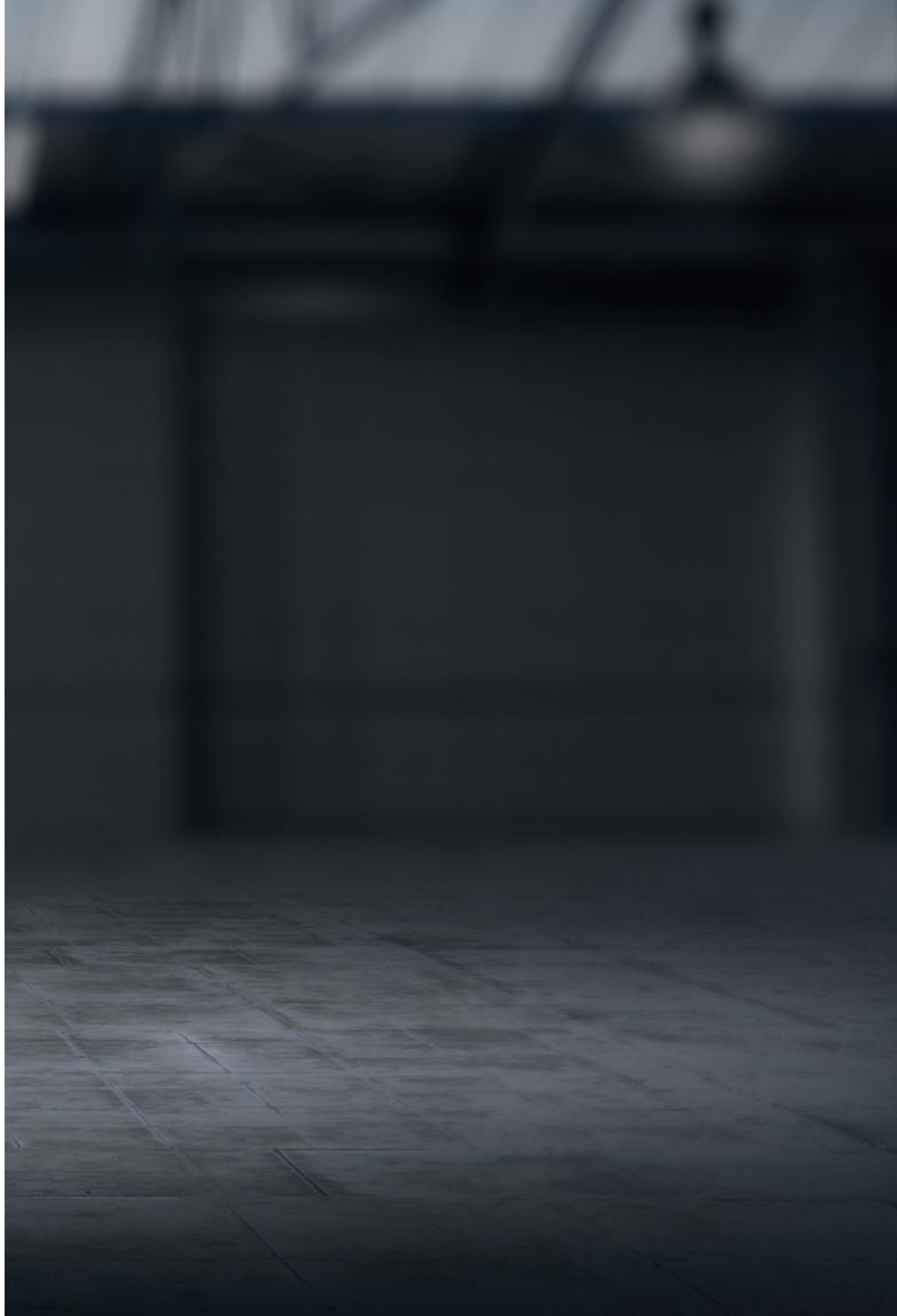
Interpreting spatial heterogeneous of Housing rent in Dallas using open textual data

GISC_6325_RS_Fundamentals
Yalin Yang



Content

- Introduction
- Research Questions
- Study Area
- Methodology
- Results and Conclusion
- Reference

Introduction

- Gentrification is one of the inevitable consequences that come up with rapid urbanization in the past decades (Baker, 2019). It refers to wealthier groups to move into a place with and brings irreversible effects of original residents usually with lower income, resulting an increase in rents and displacement of the poor (Lees, Slater, & Wyly, 2010). An estimated 2.7 million renters in the U.S. faced eviction in 2015 (American Information Research Services). More than 20 million renters pay more than 30% of their total income on rents, according to data from the U.S. Census. Unaffordable rent is one of the most severe social problems in the United States nowadays (The National Low-Income Housing Coalition). Gentrification would take place in neighborhoods once an extensive rent gap has existed (Smith, 1979). Rental market understanding is essential to recognize rent gaps and opportunities to mitigate potential impacts of gentrification. Nevertheless, traditional data sources for housing rent lag timely updates to reflect the current rental market, are limited to fixed reporting areal units and suppress the spatial heterogeneity of rent within the areal unit. Instead of quantitative analysis of reported rent data, this study collects textual data from a housing rent information platform (Craigslist) and applies natural language processing to summarize narratives that relate rent and housing properties using the City of Dallas as an example. Such narrative analysis can provide the rich context and new insights into the rental market.

Research Questions

- Based on the background information provided, this study is aiming to solve the following research questions:
- 1. Could we find a data source that provides housing rent data timely?
- 2. Could we find a directed explanation for spatial heterogeneous in the rental market? what kind of features would contribute to higher rents?

Study Area

- In this study, we select Dallas as the study area. Dallas is the ninth most-populous city in the U.S. and third in
  Texas after Houston and San Antonio. Located in North Texas, the city of Dallas is the main core of the largest metropolitan area in the Southern United States and the largest inland metropolitan area in the U.S. that lacks any navigable link to the sea.
- In this study, it has been divided into 14 councils

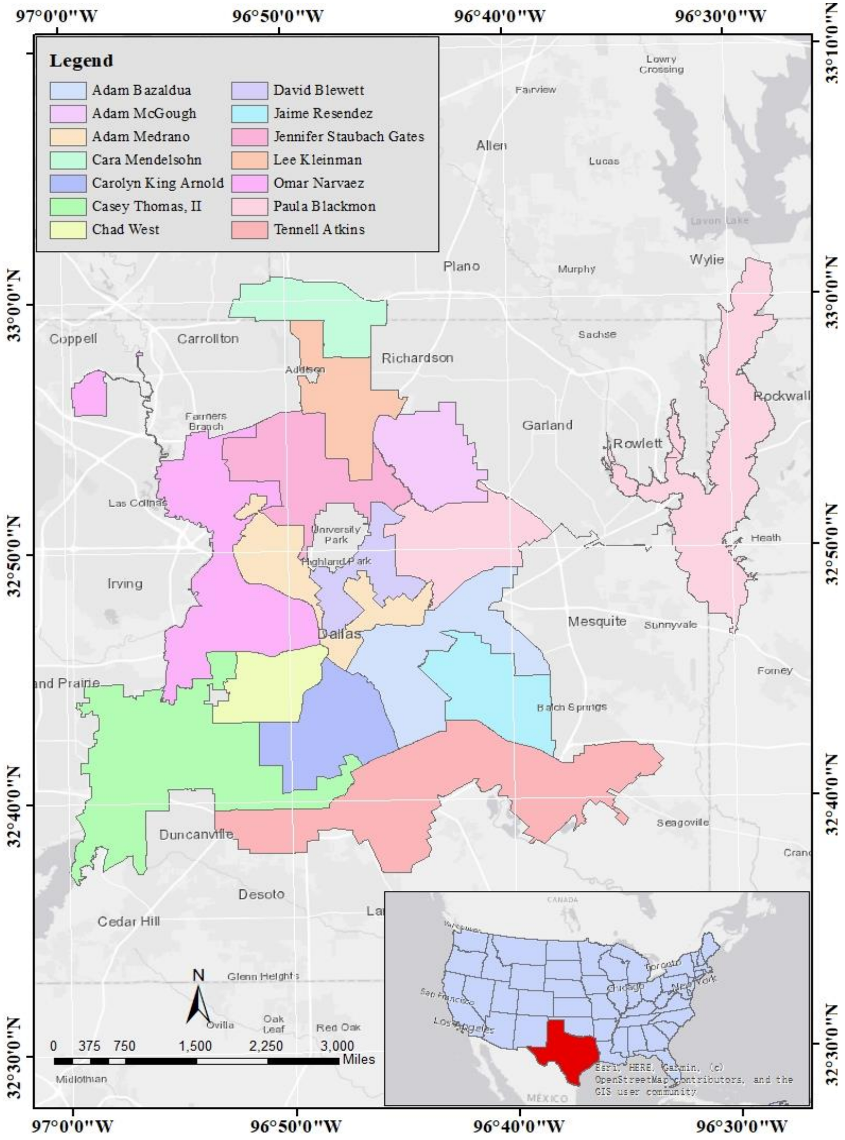
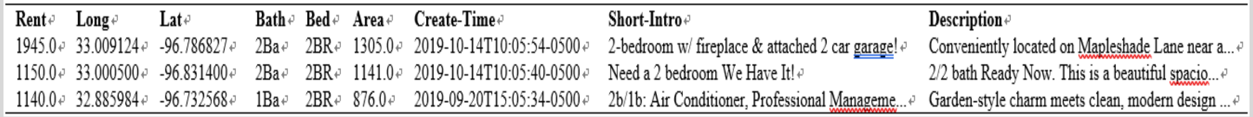


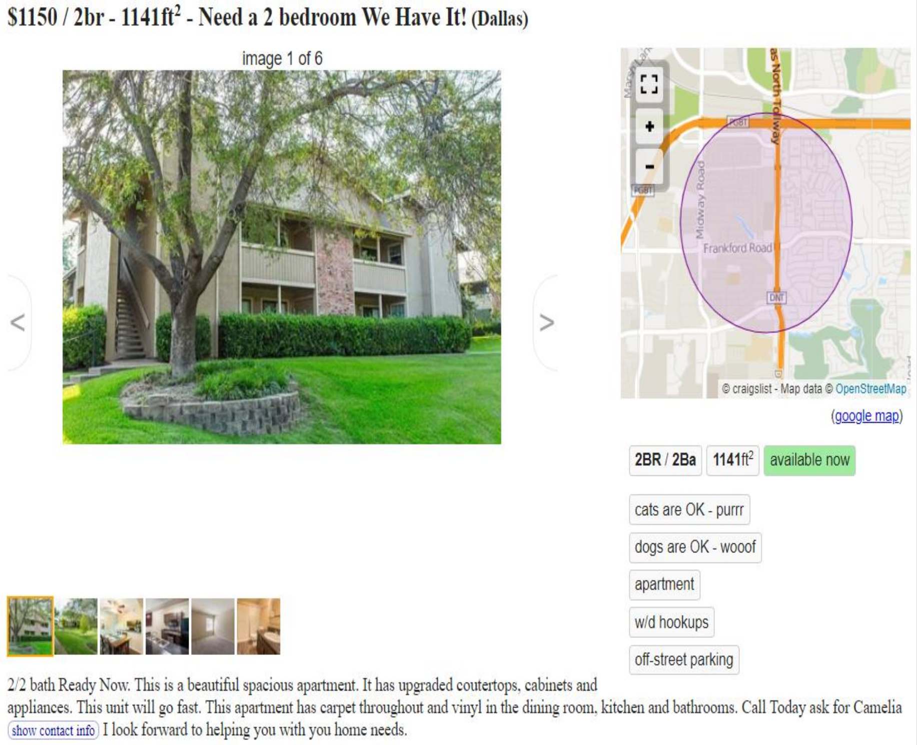
Fig. 1 Study Area

Methodology

- **Data Collecting**
- For solving those three problems, this study collects the data from Craigslist using python. More than 40,000 rental Housing advertisements were extracted from Craigslist. Each record contains features including the rent price per month, the number of bedrooms and bathrooms it has, available area, and several descriptive advertise slogan. The original website as shown in figure 2. And its record used in this study as shown below:



Rent	Long	Lat	Bath	Bed	Area	Create-Time	Short-Intro	Description
1945.0	33.009124	-96.786827	2Ba	2BR	1305.0	2019-10-14T10:05:54-0500	2-bedroom w/ fireplace & attached 2 car garage!	Conveniently located on Mapleshade Lane near a...
1150.0	33.000500	-96.831400	2Ba	2BR	1141.0	2019-10-14T10:05:40-0500	Need a 2 bedroom We Have It!	2/2 bath Ready Now. This is a beautiful spacio...
1140.0	32.885984	-96.732568	1Ba	2BR	876.0	2019-09-20T15:05:34-0500	2b/1b: Air Conditioner, Professional Manageme...	Garden-style charm meets clean, modern design ...



$1150 / 2br - 1141ft² - Need a 2 bedroom We Have It! (Dallas)

image 1 of 6

2BR / 2Ba 1141ft² available now

cats are OK - purrr

dogs are OK - wooof

apartment

w/d hookups

off-street parking

2/2 bath Ready Now. This is a beautiful spacious apartment. It has upgraded coutertops, cabinets and appliances. This unit will go fast. This apartment has carpet throughout and vinyl in the dining room, kitchen and bathrooms. Call Today ask for Camelia show contact info I look forward to helping you with you home needs.

Fig. 2 Housing rent posted on Craigslist

Methodology

- **Data Preprocessing**
- **Lemmatization**
- Because we collect data from websites, there has a large redundancy. Therefore, for the first step, we should pay effort for textual data cleaning. After splitting by whitespace and removing punctuation, Lemmatization is performed in this study. In computational linguistics, lemmatization is the algorithmic process of determining the lemma of a word based on its intended meaning. Unlike stemming, lemmatization depends on correctly identifying the intended part of speech and meaning of a word in a sentence, as well as within the larger context surrounding that sentence, such as neighboring sentences or even an entire document. As a result, developing efficient lemmatization algorithms is an open area of research.

Methodology

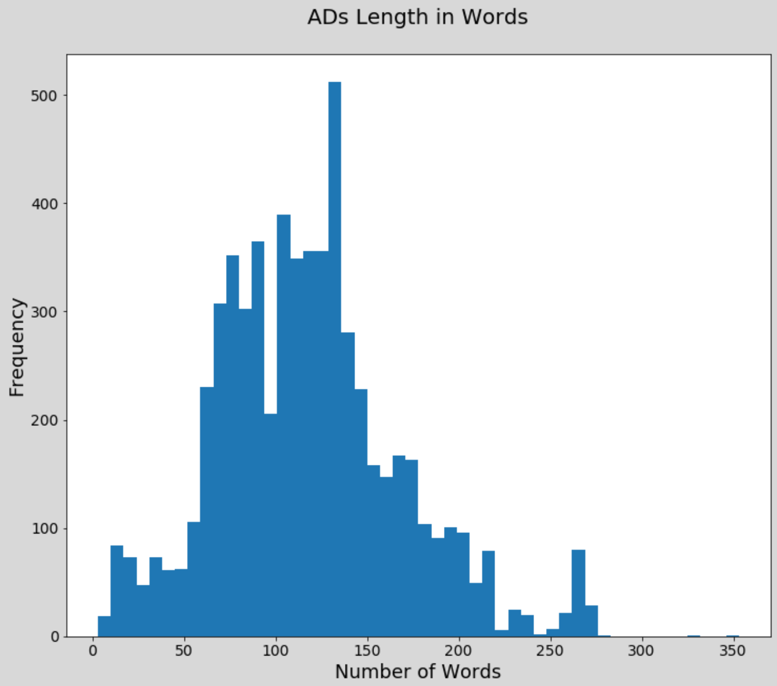


Fig. 3 Frequency of records across the number of words it contains

- **Data Preprocessing**
- **Stop words**
- In computing, stop words are words which are filtered out before processing of natural language data. In this study, the stop words provided by the Natural Language Toolkit (NLTK) are utilized as the filtering standard, which is also the most common package for textual data cleaning.
- After removing duplicates and filtering out records without a specified location, the number of distinct words across the frequency of records are generated as figure 3. The range of the number of unique words each advertisement contains spans from 3 to 350. For getting rid of outliers, in this study, only those records contain the number of words large than 10 and less than 300 are utilized.

Methodology

- **Data Preprocessing**
- **Filter out errors, spams**
- Due to the constraint from the study area and the specificity of our data source, we further filter datasets by its location and rent from quantile 2% to 98%.
- After all steps for data preprocessing, a total of 6087 records are analyzed in this study. The spatial distribution of all records as shown in figure 4.

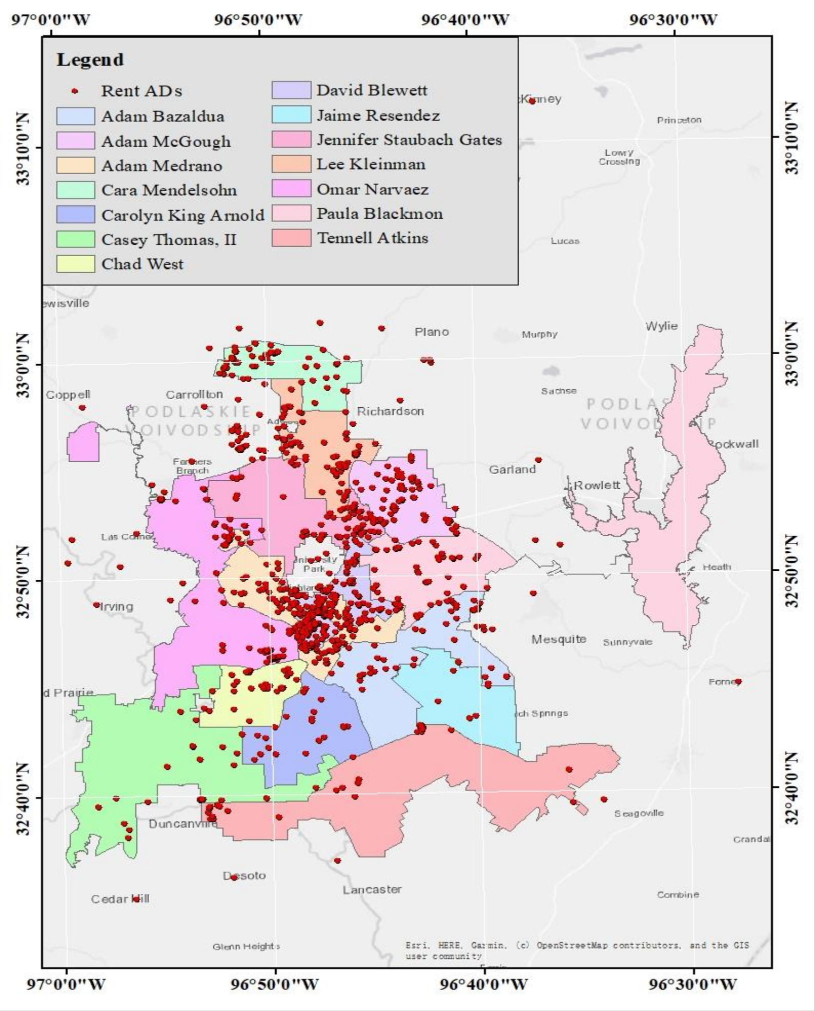
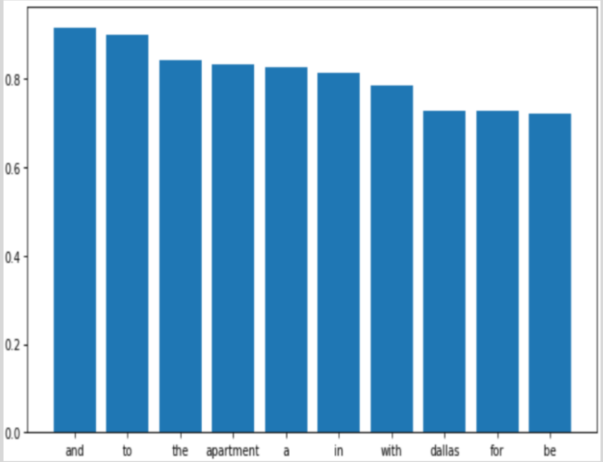


Fig. 4 All records used in this study

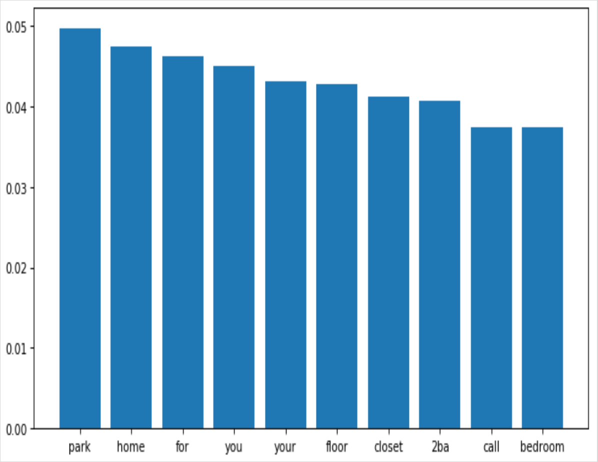
Methodology

- **Deep neural network based on TF IDF score**
- For the first step, the TFIDF array is performed for extracting 10386 distinct words posted on Craigslist. TFIDF Full name is term frequency-inverse document frequency. it is a numerical statistic algorithm that is intended to reflect how important a word is to a document in a collection or corpus.
- For the TF part, it is calculated as tf(t,d) = 1 if t occurs in d and 0 otherwise. It mechanically counts the frequency of word appears in the datasets. But in some situation, high-frequency words may do not related to the price difference. For instance, students A and B said, “I want to buy a cup of tea” and “I want to buy a cup of coffee” respectively, the sale prices are different, but only two words are reasonable for it, “coffee” or “tea”. Hence, the IDF score is developed for marking the contribution to the difference-making of each word. Inverse document frequency is calculated as : $\boldsymbol{idf}(\boldsymbol{t},\boldsymbol{D}) = \log \frac{N}{1+|\{d\in D: t\in d\}|}$ with N represents the number of sentence in the corpus, and $|\{d \in D: t \in d\}|$ represents number of sentence where word t appears.
- Specifically, the top 10 words with significant impact calculated from TF (left) and IDF (right) algorithm in this study are shown as left.

TF Score (Top 10)



IDF Score (Top 10)



Methodology

- **Deep neural network based on TF IDF score**
- According to the results from TF IDF, we establish a classic deep neural network. In this model, each independent variable x represents a distinct word extracted from the TFIDF array. And the dependent variables in this model are housing rents. Three hidden layers with 512 nodes (neurons) respectively are added. The activation function used between two hidden layers is Rectifier and the one between hidden layers and outputs is soft-max. The workflow of this model as shown in figure 5.

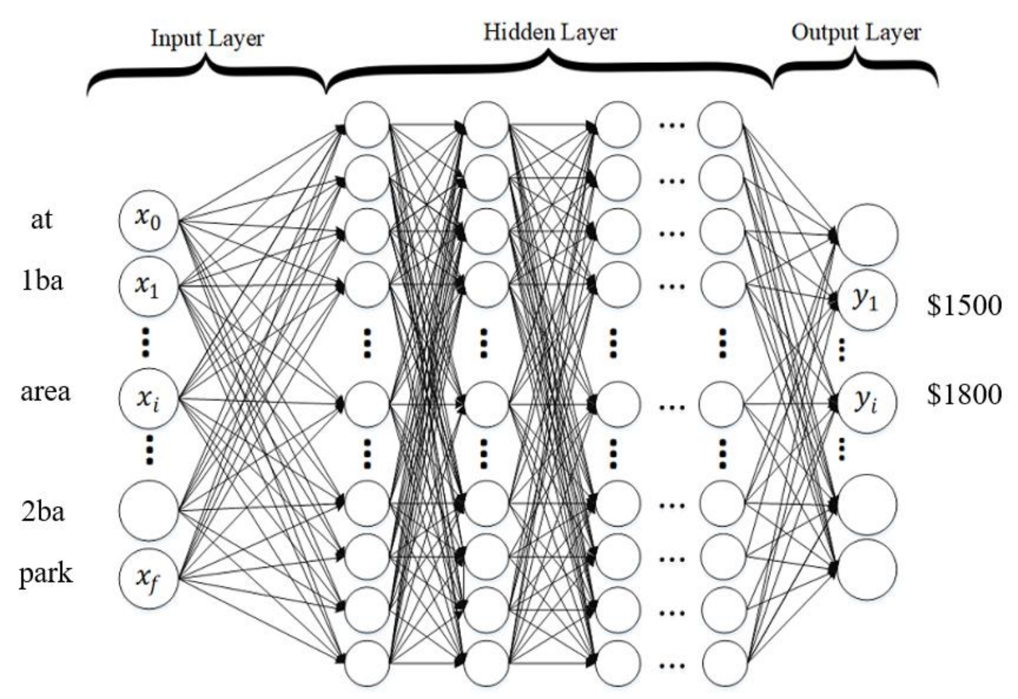
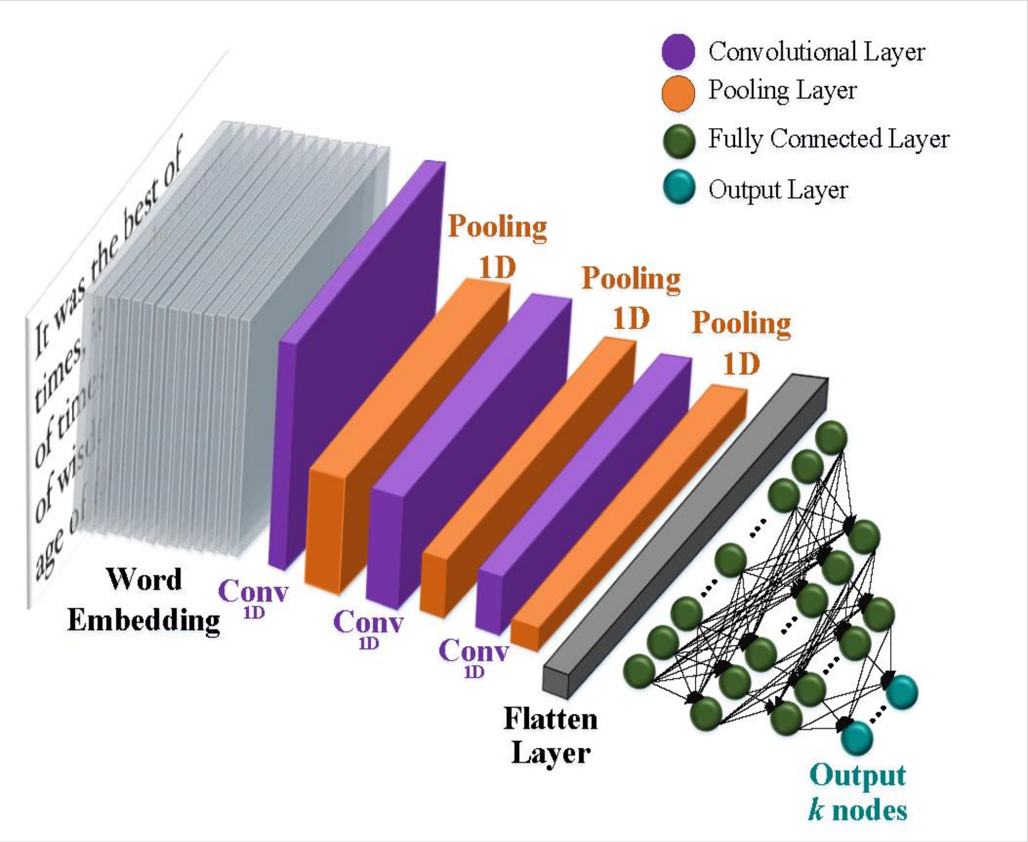


Fig. 5 Workflow for Deep neural network

Methodology

- **Convolutional neural network based on Word Embedding using GloVe**
- The second model relies on convolutional neural network base on word embedding. Word embedding is capable for capturing the context of a word in a document.
- In this study, it is implemented using the frame provided from Stanford university named Global Vectors for Word Representation. This frame converts words to vectors and put a pair of words with similar meaning to the place near to each other. That why we could convert each description to a plot and use convolutional neural network to analysis them.



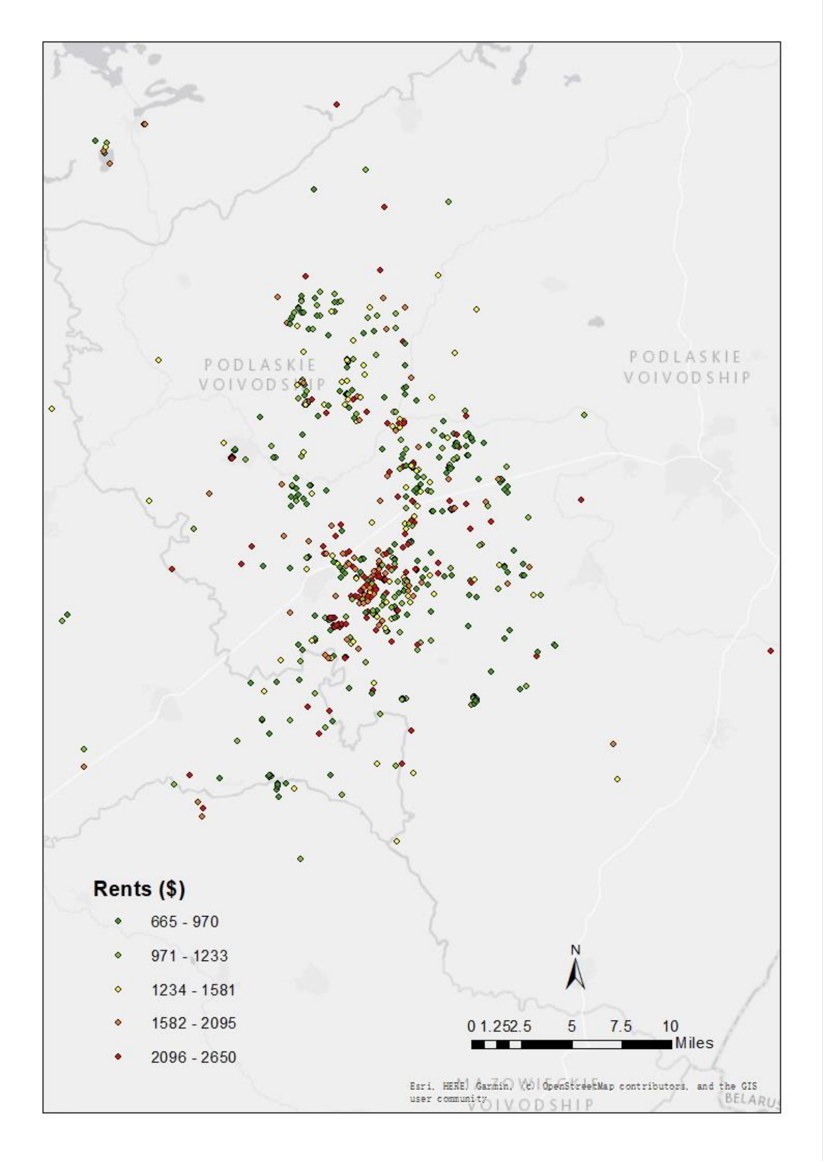
Methodology

- **Validation indicators**
- Three validation indicators are chosen including the mean absolute error (MAE), root-mean-square error (RMSE), and mean absolute percentage error (MAPE).
- In statistics, mean absolute error (MAE) is a measure of difference between two continuous variables. Assume X and Y are variables of paired observations that express the same phenomenon.
- The root-mean-square error is a frequently used measure of the differences between values (sample or population values) predicted by a model or an estimator and the values observed.
- The mean absolute percentage error (MAPE) is a measure of prediction accuracy of a forecasting method in statistics, for example in trend estimation, also used as a loss function for regression problems in machine learning.

Results

- **The result from deep neural network**
- Three validation indicators are performed including the mean absolute error (MAE), root-mean-square error (RMSE), and mean absolute percentage error (MAPE). For this model, those performance metrics are shown here.

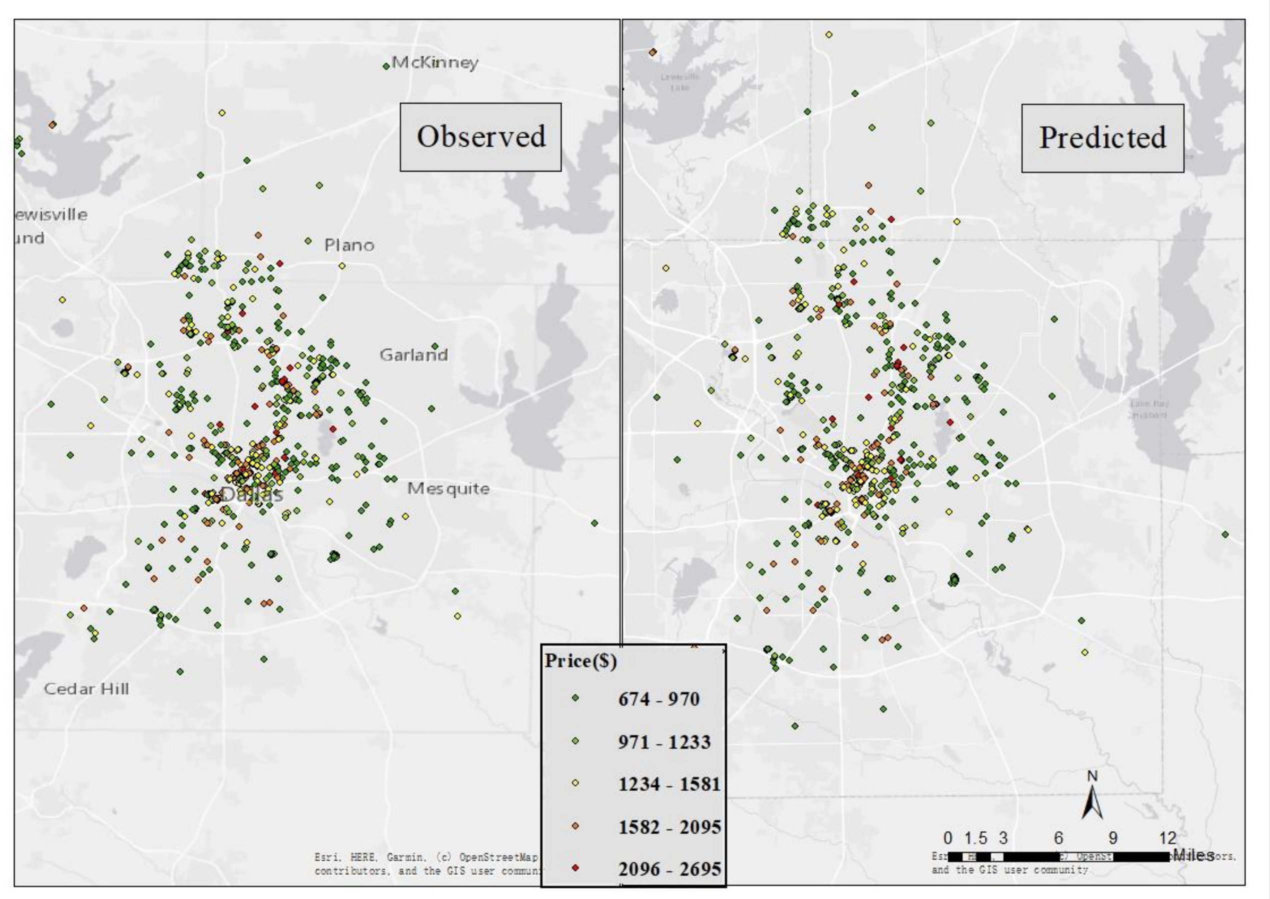
RMSE	306.728
MAE	189.326
MAPE(%)	14.599

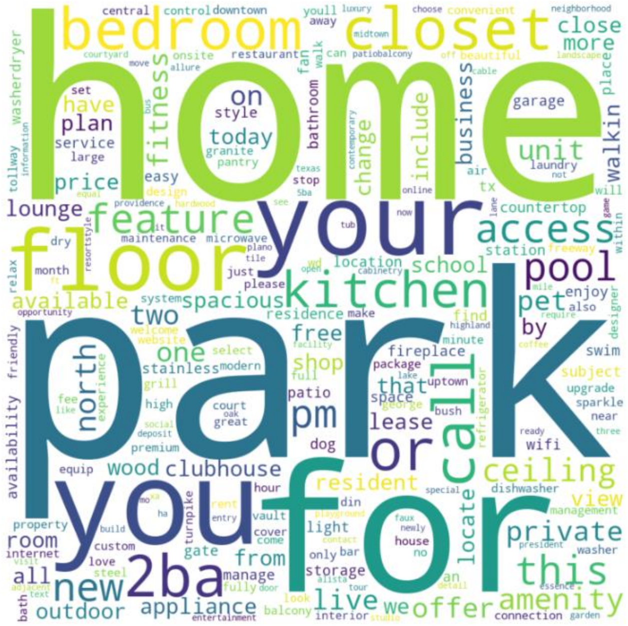


Results

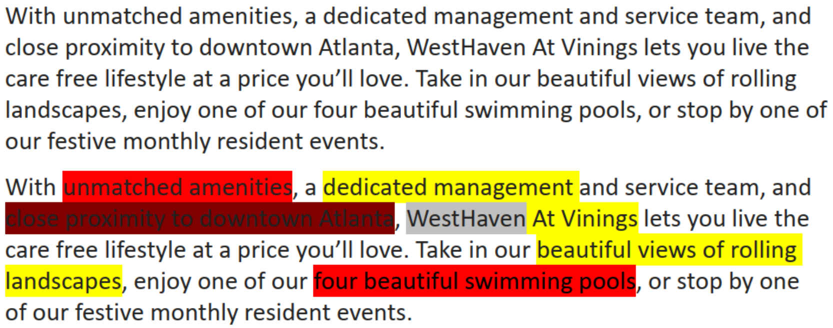
- **The result from deep neural network**
- Three validation indicators are performed including the mean absolute error (MAE), root-mean-square error (RMSE), and mean absolute percentage error (MAPE). For this model, those performance metrics are shown here.

RMSE	272.356
MAE	150.189
MAPE(%)	11.322





Word Cloud



With unmatched amenities, a dedicated management and service team, and close proximity to downtown Atlanta, WestHaven At Vinings lets you live the care free lifestyle at a price you'll love. Take in our beautiful views of rolling landscapes, enjoy one of our four beautiful swimming pools, or stop by one of our festive monthly resident events.

With unmatched amenities, a dedicated management and service team, and close proximity to downtown Atlanta, WestHaven At Vinings lets you live the care free lifestyle at a price you'll love. Take in our beautiful views of rolling landscapes, enjoy one of our four beautiful swimming pools, or stop by one of our festive monthly resident events.

Saliency Map

Conclusions

- In the conclusion part, I want to emphasize that getting a predictive model is not a key for this study. A model with low error rate also proves the accuracy of variable weight it contains. Those weight of variables, in other words, the effect of narrative materials, is what we care about here.
- In this study, we already build two models with acceptable error rate from textual datasets. The interpretation of weight of each word from those two models are essential for next step. In this study, we use both word cloud and saliency map to visualize the result.
- With those two methods, we could visually see what kind of features are correlated to higher housing rent.

Reference

- Nelson, Arthur C., et al. "Office rent premiums with respect to light rail transit stations: Case study of Dallas, Texas, with implications for planning of transit-oriented development." Transportation Research Record 2500.1 (2015): 110-115.
- Zhou, Xiaolu, Weitian Tong, and Dongying Li. "Modeling Housing Rent in the Atlanta Metropolitan Area Using Textual Information and Deep Learning." *ISPRS International Journal of Geo-Information* 8.8 (2019): 349.
- Pennington, Jeffrey, Richard Socher, and Christopher Manning. "Glove: Global vectors for word representation." Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014.
- Johnson, Rie, and Tong Zhang. "Supervised and semi-supervised text categorization using LSTM for region embeddings." arXiv preprint arXiv:1602.02373 (2016).
- Dos Santos, Cicero, and Maira Gatti. "Deep convolutional neural networks for sentiment analysis of short texts." Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. 2014.