Data Realtor: Real Estate Research Tool

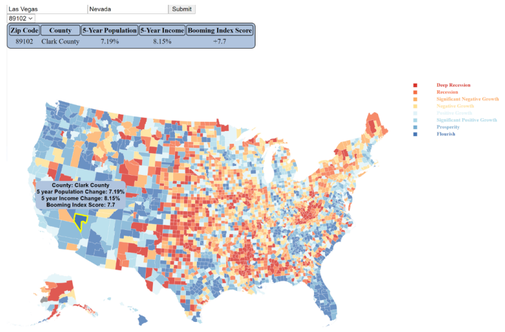
Syefira Shofa (GT) Gerald Richland (GT) Xueyuan Mei (GT) Dohyung Kim (GT) Raghu Kondury (GT)

Summary

Data Realtors is a real estate research tool created to assist property researchers efficiently gather information about a specific location. Rather than focus on information about specific properties, Data Realtor aims to inform users what the environment is like at a given location. This tool will provide a multi-dimensional view of a given location describing local markets, schools, transportation, and insurance. It also provides a trustworthy score for reviews users might find on the internet. Data Realtor combines **machine learning** and **visualization** in order to provide an easy to use and comprehensible user experience.

Local Economy Snapshot – Located in the Search By Zip Code View

Helps homebuyers understand the long-term income and demographic data trends for long term analysis and regional comparison**.** The user is shown the demographic and income change trends and a "Booming Score" which is calculated based on the trends. Demographic and income data was gathered at the county level from 2010 to 2019 from the United States Census Bureau where each year's data can be downloaded as a CSV file.



Review Aggregator

Helps homebuyers understand the living quality of the property through firsthand experiences.

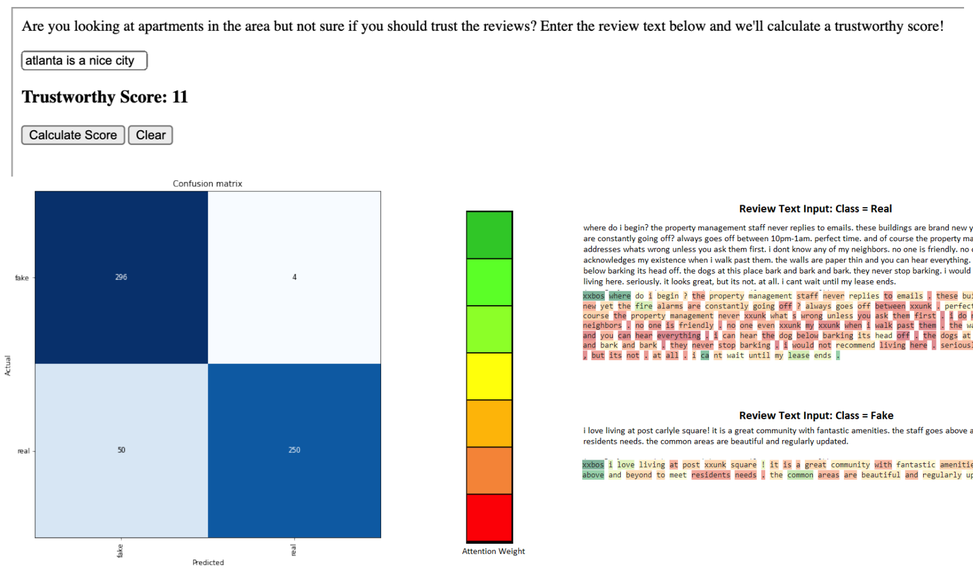
Data: 3000 hand-labeled review text examples scraped from Apartments.com, 50% are real, 50% are fake.

Pre-Process: Split into training, validation, testing, 70/20/10. The texts are then masked, tokenized, and grouped into batches of 64.

Model Used: Sequential bi-directional recurrent neural network architecture, using ASGD Weight-Dropped Long-Short-Term-Memory. Pre-trained on Wikitext 103. Transfer learning was used to train the model for review text classification.

Evaluation: 600 review texts from the test dataset (Class 0 = Fake, Class 1 = Real). Accuracy was 91% and Precision was 85.5%.

Model Analysis: Analyze intrinsic attention of the model for a single review text by looking at the relative change of the output, based on the words in the input text. This is done by calculating the sequential Jacobian matrix of the attention layers, then binding the values to a range of 0 – 1. Values close to 1 are considered sensitive input words that can change the classification output if altered.

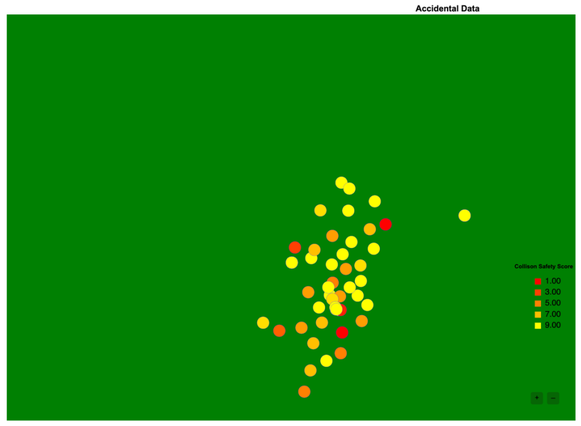


Evaluation

Via a Google survey, we asked for a rank on website satisfaction (1-10) and left a feedback section open. By not specifying to rank x or comment on x, we left it to the user to rank and comment on our website using whatever unbiased, subconscious KPI they had in their head. We learned that Data Realtors was above average (mode rank of 7) and needs to be tuned to focus more on providing user intuition on how to use the website and the information presented to them.

User Interface – Search By Major City

We aimed for the user interface to be as friendly as possible. Thus, when a user selects their respective city, the data is already combined and shown in an intuitive format.



School list

School Name	County Name	Web Site URL	Location Address	Location City	Location ZIP
7 PILLARS CAREER ACADEMY	Clayton County	N/A	4234 Hendrix Dr	Forest Park	30297
A.L. BURRUSS ELEMENTARY SCHOOL	Cobb County	http://alburruss.marietta-city.org	325 Manning Rd SW	Marietta	30064
AARON COHN MIDDLE SCHOOL	Muscogee County	http://tinyurl.com/aaroncohn	7352 Garrett Rd	Midland	31820

Traffic Data

record	hours	city	mean_travel_time_minutes
0	[0, 1, 2, 3, 4, 5, 21, 22, 23]	Atlanta	19.027587039286 16
1	[6, 7, 8, 13, 14, 18]	Atlanta	24.339131230996877
2	[9, 10, 11, 12, 19, 20]	Atlanta	21.540790735926482
3	[15, 16, 17]	Atlanta	28.099624754618603

Car Insurance Local Landscape

Helps homebuyers understand the insurance implications of living at that property. The color legends for this Map ranges from Red (lowest score) to yellow (highest score) where the less the accident count, the higher the weighted score. Car accident data was obtained from Kaggle and zip code data was obtained from Simple Maps.

Local K-12 Education Snapshot

Catered to property researchers with children who are interested in having their children go to a good school. The user can easily look to see which schools are available in a given location, what other nearby school options are and the school website should they decide they want further information. Data source was the integrated Greatschools.org API with scrapping from nces.ed.gov.

Hourly Uber Accessibility Snapshot

Catered to property researchers who plan to use Uber in lieu of self or public transportation and could also serve to inform drivers which hours of the day have the least traffic**.** The user is given a data table showcasing clusters of hours with travel times similar to one another as well as the average travel time for each cluster. Uber data was gathered at the hourly level from Uber Movement using one year's worth of data ranging from quarter 2 of 2019 to quarter 1 of 2020. Each quarter's worth of data was downloaded via csv which contains for each source and district the mean and standard deviation of the travel time.

Clustering Experiment

Experiment: A comparison was made on the KMeans vs MeanShift clustering algorithms. We took the average of the average travel times aggregated by hour and our goal was to cluster the hours with similar travel times so that the end user could see best times to call an Uber and whether or not that hour of the day was similar to another.

MeanShift algorithm: centroid based algorithm, which works by updating candidates for centroids to be the mean of the points within a given region. These candidates are then filtered in a post-processing stage to eliminate near-duplicates to form the final set of centroids. MeanShift automatically sets the amount of clusters.

KMeans algorithm: clusters data by trying to separate samples in n groups of equal variance, minimizing a criterion known as the inertia or within-cluster sum-of-squares. This algorithm requires the number of clusters to be specified. The number of clusters was determined via the elbow method for this experiment.

Result: The output of both clustering models proved to be very similar. As a result, we decided to use the output of the MeanShift model as it would allow us to skip the elbow method step of specifying the number of clusters.