Springboard

craigslist

SPRINGBOARD CAPSTONE PROJECT TWO
WHAT'S FOR FREE ON CRAIGSLIST?

Author: Josh Mayer

Date: November 15, 2017

Springboard

craigslist

OUTLINE

- Introduction & Problem Statement
- The Dataset
- Analysis & Findings
- Statistical Inference
- Machine Learning
- Conclusions

☆ La Z Boy love seat (Parker)

☆ FREE SCRAP METAL (Hudson)

☆ Free Wine Cabinet - holds 260+ bottles of wine (Stapleton)

Springboard

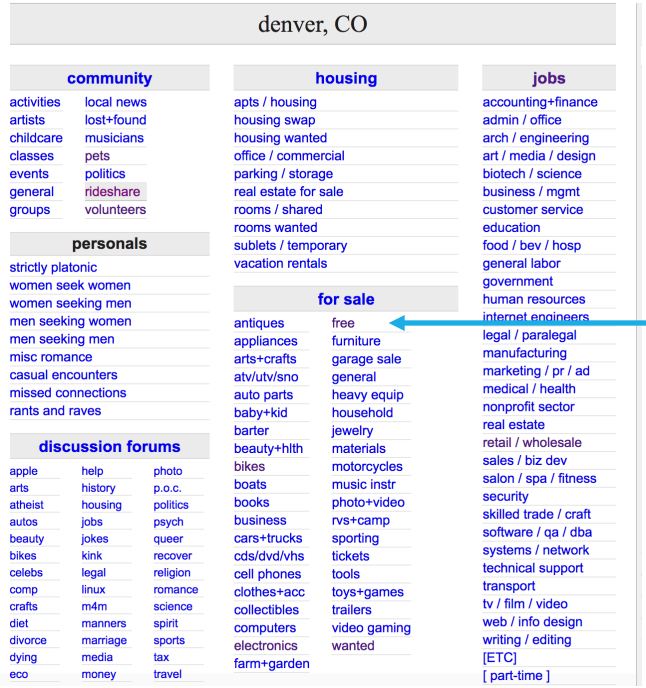
craigslist

INTRODUCTION & PROBLEM STATEMENT

Craigslist was founded in 1995 and has become one of the most powerful sources for classified advertisements found on the web today.

THE PROBLEM

Oftentimes products are misclassified or the category itself is too broad to be useful.



denver, CO

community

activities, artists, childcare, classes, events, general, groups, local news, lost+found, musicians, pets, politics, rideshare, volunteers

personals

strictly platonic, women seek women, women seeking men, men seeking women, men seeking men, misc romance, casual encounters, missed connections, rants and raves

discussion forums

apple, arts, atheist, autos, beauty, bikes, celebs, comp, crafts, diet, divorce, dying, eco, help, history, housing, jobs, jokes, kink, legal, linux, m4m, manners, marriage, media, money, photo, p.o.c., politics, psych, queer, recover, religion, romance, science, spirit, sports, tax, travel

housing

apts / housing, housing swap, housing wanted, office / commercial, parking / storage, real estate for sale, rooms / shared, rooms wanted, sublets / temporary, vacation rentals

for sale

antiques, appliances, arts+crafts, atv/utv/sno, auto parts, baby+kid, barter, beauty+hlth, bikes, boats, books, business, cars+trucks, cds/dvd/vhs, cell phones, clothes+acc, collectibles, computers, electronics, farm+garden, free, furniture, garage sale, general, heavy equip, household, jewelry, materials, motorcycles, music instr, photo+video, rvs+camp, sporting, tickets, tools, toys+games, trailers, video gaming, wanted

jobs

accounting+finance, admin / office, arch / engineering, art / media / design, biotech / science, business / mgmt, customer service, education, food / bev / hosp, general labor, government, human resources, internet engineers, legal / paralegal, manufacturing, marketing / pr / ad, medical / health, nonprofit sector, real estate, retail / wholesale, sales / biz dev, salon / spa / fitness, security, skilled trade / craft, software / qa / dba, systems / network, technical support, transport, tv / film / video, web / info design, writing / editing, [ETC], [part-time]

Source: https://denver.craigslist.org/

PRIMARY OBJECTIVE

The goal is to explore the "For Sale/Free" category, leverage unsupervised learning to cluster and categorize products listed.

Springboard

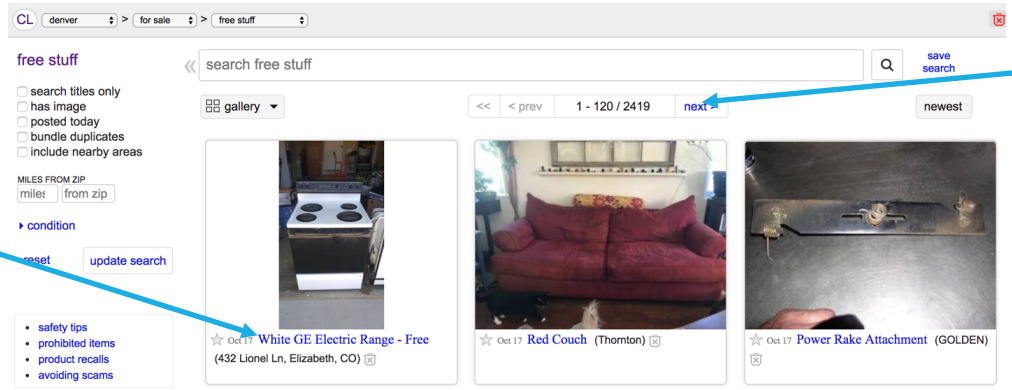
craigslist

DATA ACQUISITION & WRANGLING

Data was scraped from the Denver, Boulder, Colorado Springs, and Fort Collins (Greater Denver Metro) Craigslist sites via the Python package "Scrapy".

1) Initiate a "Scrapy" spider/crawler to acquire data based on HTML tag.

2) Grab the title for each listing on the page.



3) Parse the pages until all listings have been scraped.

FINAL DATASET

3,927 Listings

DATA WRANGLING

- Duplicate Records – 133 records (3% of the acquired records) were dropped since they contained duplicate title AND timestamp. New timestamps are created if listings are "reposted".
- Non-letter Removal – remove numbers, punctuation, emojis, etc. in order to have only letters.
- Lowercase Conversion – all letters were converted to be lowercase for comparison.

Springboard

craigslist

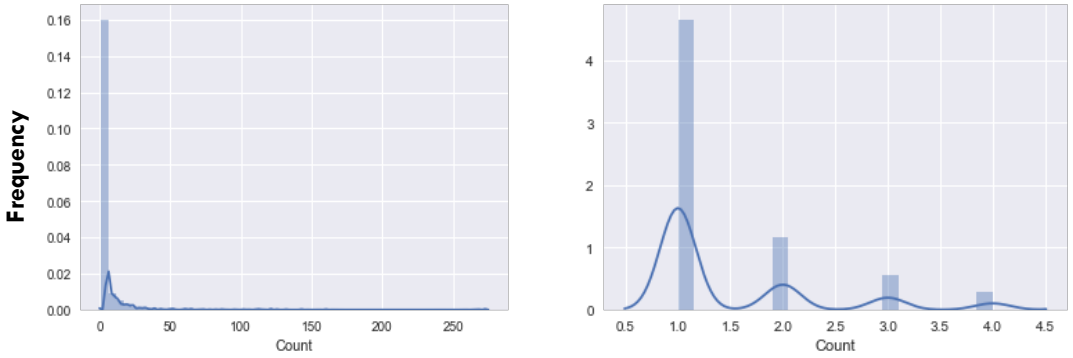
EXPLORATORY DATA ANALYSIS

The dataset acquired was highly skewed towards infrequently found words.

DATASET CHALLENGES

- 2,585 unique words were found in 3,927 unique listings.
  - The average word is found 4.51 times with a min count of 1 and max count of 274 counts.
- 57% of the words are found only one time throughout the many listings.
- 82% of the words are found fewer than 5 times.

FREQUENCY VS UNIQUE WORD COUNT



Charts above: frequency of all words (left) along side the frequency of words found fewer than 5 times (right).

Springboard

craigslist

FURTHER ANALYSIS

Analyzing the most (and least) frequent words begins to show some of the potential categories.

We can already start to see some of the potential categories based on word frequency alone (e.g. “wood” and ”firewood” are amongst the top ten words found).

MOST FREQUENT VS INFREQUENT WORDS

COUNT	TOP 10 WORDS
274	wood
267	couch
160	tv
143	boxes
132	chair
125	dirt
121	pallets
121	curb
115	moving
68	firewood

COUNT	BOTTOM 10 WORDS
1	ability
1	newspaper
1	newer
1	neutral
1	network
1	net
1	nepro
1	needing
1	necklaces
1	mtb

Charts above: the most frequent word found in the listings was “wood” at 274 (left chart). There are many words with a count of w (right chart).

Springboard

craigslist

MACHINE LEARNING

The primary goal of the machine learning (ML) section is to identify the best algorithm to properly classify the "free stuff" listings on Craigslist.

MACHINE LEARNING APPROACH

1) Text analysis was performed with the term frequency-inverse document frequency (tf-idf) algorithm and the vectorizers were then transformed to the tf-idf matrix.

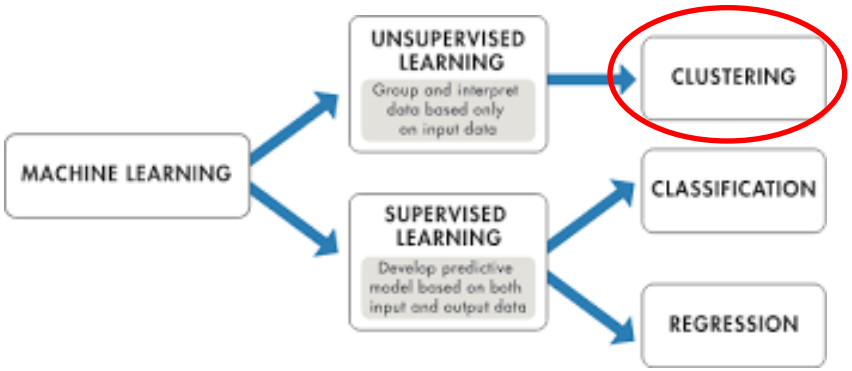
2) Parameter tuning included setting min/max term frequency, "N grams", stop words, and logarithmic data transformations.

3) Perform clustering with the "K-Means" algorithm.

4) Understand model results and identify the best model to name the clusters.

FOCUS ON UNSUPERVISED LEARNING

This is an unsupervised ML exercise and focused on testing and understanding key parameters as relates to the exploratory data analysis findings.



Springboard

craigslist

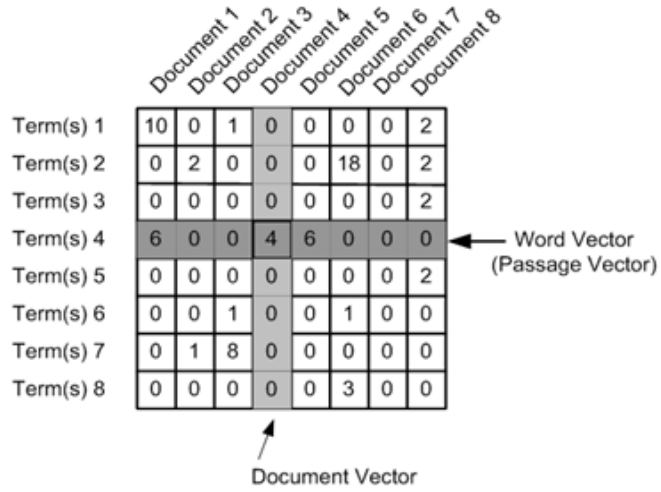
MACHINE LEARNING – TEXT ANALYSIS

Text analysis was performed with the term frequency-inverse document frequency (tf-idf) algorithm and the vectorizers were then transformed to the tf-idf matrix.

STEPS TO CREATE THE TF-IDF MATRIX

1. Count word (term) occurrences by document.
2. Transform into a document-term matrix.
3. Apply the term frequency-inverse document frequency weighting.
4. Perform sensitivity analysis on frequency weighting as well as some of the other key parameters.

VISUAL EXAMPLE TF-IDF MATRIX



Springboard

craigslist

MACHINE LEARNING – PARAMETER TUNING

Sensitivity analysis was performed on key model parameters to determine the impact.

KEY MODEL PARAMATERS

- Min/Max Frequency – this is the frequency within the listings a given term can have to be used in the tf-idf matrix. E.g. If the term is in greater than X% of the listings it probably carries little meaning.

- N grams: accounts for term relativity. As an example, setting this value equal to "2" would mean that I consider unigrams and bigrams, e.g. "leather" and "leather couch" which are both important to me.

- Stop Words: some words are too common (e.g. the, and) to be considered (stop words) for analysis and need to be removed. The only customization made for this model was to include the word "free" in the stop words list as it was found in nearly all listings.

- Logarithmic Transformation: setting this value to "true" provided a log10 transformation of the data, which in this case was sorely needed given the dataset was skewed towards infrequent words.

Springboard

craigslist

MACHINE LEARNING – K-MEANS CLUSTERING

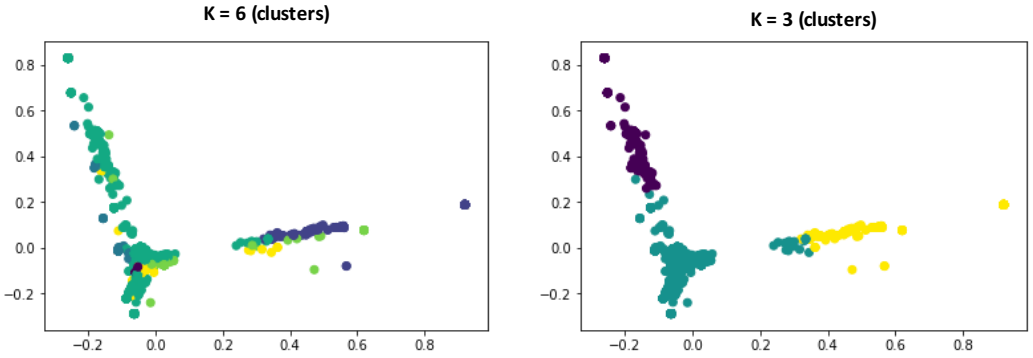
The clustering algorithm of choice for this model was "K-means" which required a pre-determined number of clusters for initialization.

ITERATING OVER POTENTIAL K VALUES

K Value	Sum-of-Squares Score	Silhouette Score
2	-2,273	0.134
3	-2,201	0.145
4	-2,128	0.163
5	-2,081	0.181
6	-1,981	0.202
7	-1,994	0.191
8	-1,893	0.223
9	-1,861	0.233

The silhouette score (SS) ranges from -1 (a poor clustering) to +1 (a very dense clustering) with 0 denoting the situation where clusters overlap. SS scores ranged from 0.134 to 0.233 depending on the K value.

COMPARING VISUAL RESULTS OF K CLUSTERS



Charts above: at k=6 (left) Interesting to see the supposed cluster on the right-hand side of the chart with multiple colors of dots next to each other. At k=3 (right) the clusters visually appear more relevant, however, yellow and green dots are still very close together on the right-hand side of the page. Looking back at the scores for k=3, the SS is 0.145 which means a weak structure.

Springboard

craigslist

MACHINE LEARNING – MODEL RESULTS

Performing sensitivity analysis on key model parameters produced a model with k=3 clusters and the following results.

A few key observations:

- As expected, some of the words are present across clusters. This was to be expected given the low silhouette scores and the visual of the many different colored dots close together.
- Only one bigram is present, meaning that the non-unigram words were not as important as initially thought.
- An attempt at naming the clusters might be:
  - Cluster 1 would be named "Moving Items".
  - Cluster 2 would be named "Scrap Materials".
  - Cluster 3 would be named "Furniture".

NAMING THE CLUSTERS

Cluster 1 "Moving Items"	Cluster 2 "Scrap Materials"	Cluster 3 "Furniture"
Wood	Firewood	Couch
Tv	Wood	Leather
Pallets	Pallets	Leather couch
Chair	Today	Loveseat
Boxes	Come	Recliner
Desk	Scrap	Chair
Dirt	Yard	Reclining
Stuff	Small	Sofa
Table	Mulch	Brown
Moving	Pallet	Bed

Chart above: top 10 terms in each cluster of the model results of k=3.

Springboard

craigslist

RESULTS SUMMARY

KEY FINDINGS

- The data set acquired for the Greater Denver Metro "Free Stuff" analysis was heavily skewed by infrequent terms. Even a log 10 transformation was not able to properly redistribute or normalize the data.

- This does not mean it is "bad data", only that the majority of the postings in the "free stuff" section represent very few terms.

- With a Silhouette Score of 0.14 on the training set and 0.15 on the test set, the K-means clustering model provided weak (at best) structure for the designated clusters.

- In this analysis of the "Greater Denver Metro" Craigslist sites, there was an abundance of items in the Craigslist "free stuff" that was classified into three categories:
  - Category 1 is "Moving Items".
  - Category 2 is "Scrap Materials".
  - Category 3 is "Furniture".

Springboard

craigslist

RECOMMENDATIONS & FUTURE IMPROVEMENTS

Key results show that the Craigslist team could implement a clustering model to help users filter and find the products.

RECOMMENDATIONS & FUTURE IMPROVEMENTS

- Craigslist site administrators should create and enable “tags” or “categories” to assist in filtering the “free stuff” section.
- Perform unsupervised learning on the broader, national data set of all Craigslist site for better understanding of what is offered for free across the country.
- There could be a similar analysis done regarding miscategorized or mislabeled listings where the goal would be to find the outlying data points to understand their value.