**Capstone Project #2: What’s for “Free” on Craigslist?**

Milestone Report

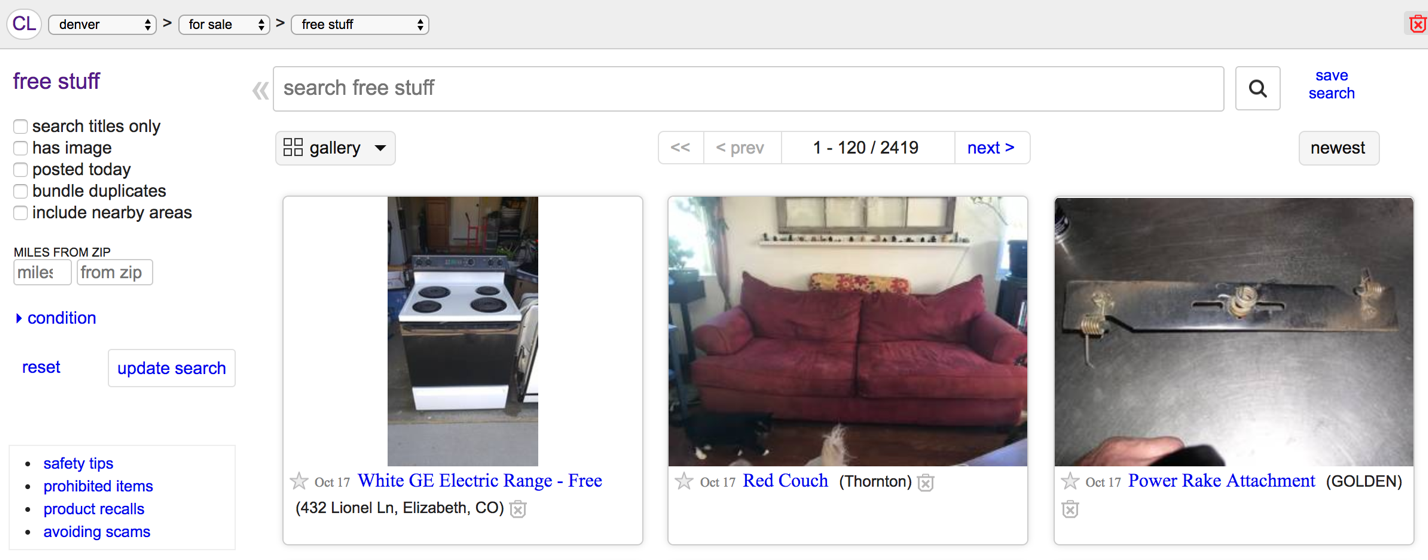
Date: October 19, 2017

**Introduction and Objective**

Craigslist was founded in 1995 and has become one of the most powerful sources for classified advertisements found on the web today. The site is commonly used when searching for new and used products, jobs, and can even be used as a service to meet other people.

One of the most interesting aspects of the Craigslist site is its simplicity in design and user experience. It is fairly simple to set up an advertisement and I have often utilized the site when selling items that I no longer need (typically when I moving to/from apartments). Within minutes a user can list an item for sale.

One of my pain points in the user experience, however, lies in the browse functionality of some of the categories. Oftentimes products are misclassified or the category itself is too broad to be useful. To test this theory this I will explore the “For Sale/Free” category and leverage unsupervised learning techniques to cluster and explore the actual products advertised.



*Photo above: screenshot taken of the Denver Craigslist “free stuff” category that shows a variety of products with no ability to filter.*

**Target Audience and Why**

My target audience will initially be the savvy online shopper. In my opinion, Craigslist users tend to be deal seekers so having more detailed information on what is offered for “Free” may be of particular interest. This would allow users to understand what is commonly offered for “Free” and may sway them to consider Craigslist before purchasing something in those categories. E.g. if “moving boxes” are commonly offered for free why would anyone purchase them?

Craigslist has individual sites that are based on metro area since it uses a “meet-in-person” model. Everything is local meaning that no transactions occur online (e.g. exchange of money for shipment of goods). This translates to a localized market of goods and services where offerings may drastically differ across metros. It may also be of interest to compare the findings of the model for Denver metro to another metro in a different area of the country.

**Dataset Acquisition**

Data was scraped from the Denver Craigslist site via the Python package “Scrapy” which was initially built to scrape Craigslist sites for job postings and is easily modifiable to handle other Craigslist categories.

Creating a Scrapy spider for the Denver “free stuff” page was an easy task. I simply had to modify the appropriate URL and then find the appropriate HTML tags and variables to modify in order to return the appropriate results.

After scraping a single page of data I needed to modify the script so that I could scrape as many pages of results that exist. In the Denver Craigslist site this ended at 20 pages and 2,400 results. Each result was returned as a separate CSV file and wrangled via Pandas.

‘Glob’ was used to consolidate the 20 CSV files into a single Pandas data frame.

Dataset #1 (Denver “free stuff”) features only two columns:

1. ‘titles’ – features the text description for each listing title in the “free stuff” category.
2. ‘timestamp’ – features the time and date that the listing was posted to the site.

<class 'pandas.core.frame.DataFrame'>

Int64Index: 2400 entries, 0 to 119

Data columns (total 2 columns):

titles 2400 non-null object

timestamp 2400 non-null object

dtypes: object(2)

memory usage: 56.2+ KB

**Data Wrangling**

Steps required to cleanse and modify the data into a necessary format for analysis.

1. Reset the index of the final data frame since multiple CSV files were loaded with records having the same index numbers.
2. Drop duplicate records which removed 5 records. I removed only duplicates that matched title and timestamp. I also considered removing duplicates where ONLY the title matched but found that these were primarily different listings (e.g. microwave).
3. Conversion of columns to proper types. ‘titles’ to string and ‘timestamp’ to an an actual timestamp object.
4. Convert all words to lowercase for future analysis.
5. Remove all characters and numbers.
6. Remove all stop words. Leverage NLTK ‘stopwords’ as the basis for the list. Add the word ‘free’ for this analysis since almost all of the listing titles are for free items. Parse each word in the titles list and if in the stop words list then remove, otherwise keep in the final list of titles to analyze.

**Exploratory Data Analysis**

Most of the initial data analysis was regarding the shape of the words and their associated counts. More advanced analysis will require unsupervised machine learning models to be created.

Understanding the words found in the postings:

count 1871.000000

mean 3.799572

std 9.784902

min 1.000000

25% 1.000000

50% 1.000000

75% 3.000000

max 176.000000

Name: Count, dtype: float64

Although we have 2,395 unique listings, the listings only contain 1,871 unique words. The average word is found 3.79 times with a min of 1 count and max of 176 counts!

|  |  |
| --- | --- |
| Count | Top 10 Words |
| 176 | wood |
| 163 | couch |
| 114 | tv |
| 97 | dirt |
| 84 | chair |
| 81 | boxes |
| 80 | pallets |
| 71 | scrap |
| 69 | desk |
| 68 | firewood |

|  |  |
| --- | --- |
| Count | Bottom 10 Words |
| 1 | ability |
| 1 | national |
| 1 | name |
| 1 | munchkin |
| 1 | multiple |
| 1 | multi |
| 1 | mufflers |
| 1 | mtn |
| 1 | mths |
| 1 | mtb |

*Charts above: ‘wood’ was found 176 times while ‘ability’ was only found 1 time in 2,395 unique listings.*

Interesting findings already! It appears that the top item listed is wood which I assume to be related to firewood (which is the 10th most frequently found word). The words ‘couch’ and ‘tv’ are ranked 2nd and 3rd respectively and make me wonder if I should consider looking on Craigslist for these items before I purchase them next!

It also appears that there are many infrequent words with a count of 1. Analyzing the distribution of the word frequency shows that there are actually many words with a count less than 25.





The distribution of the words found less than 5 times highlights that the data is skewed by infrequently found words:

count 1575.000000

mean 1.475556

std 0.835524

min 1.000000

25% 1.000000

50% 1.000000

75% 2.000000

max 4.000000

Name: Count, dtype: float64

Wow, 1,575 of the 1,871 words (~85%) are found less than 5 times in the 2,395 unique listings with the average word only having 1.47 counts per 2,395 listings.

Clearly there will need to be a minimal word count for model creation, but I am also worried about having enough data to train a robust model. I may need to consider pulling the ‘free stuff’ listings for multiple cities to have enough data.

**Key Findings and Next Steps**

1. Dataset acquisition was fairly easy with the “Scrapy” package. Data wrangling, however, was a challenge and new experience for me given that the data is text-based. It was interesting to research the concept of “stop words” as well as determine the appropriate list of words to use (which also required customization for this application).
2. One of the key challenges in this data set is regarding the concept of duplicate records since users are allowed to “repost” listings. When a user “reposts” an existing listing, the listing is assigned a new timestamp and therefore is a unique record (in my data set). This does NOT catch the scenario where users actually create two separate, but identical listings. In my research I looked at the actual listings that had identical titles (e.g. microwave) and they were indeed different products.
3. The exploratory data analysis section highlighted that most of the words tend to be infrequent (i.e. found fewer than 5 times). Given that I am only looking at 2,400 unique listings I may have to pull data for other cities in order to get to a level of frequency that allows for a more robust model.