**What’s for “Free” on Craigslist?**

**(Practice with Unsupervised Learning Techniques)**

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 prove this I will explore the “For Sale/Free” category and leverage unsupervised learning techniques to cluster and explore the actual products advertised.

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.

The Data

Data will need to be scraped from the Denver Craigslist site. I plan to leverage the Python library “Scrapy” which was initially built to scrape Craigslist sites for job postings. The primary focus will be on the title of the ad listed in the Denver “For Sale/Free” category. It is unknown how many unique records exist but initial research shows that there are at least 2,500 unique listings in the Denver “For Sale/Free” category.

Approach

1. Leverage the Python library “Scrapy” to scrape data from the Denver Craigslist site.
2. Modify parameters in the “Scrapy” library to be able to scrape data from the “For Sale/Free” category specifically. Returned results will focus on the titles of the posted advertisements.
3. Initial data wrangling will require minimally removal of “Stop Words” (e.g. “The”, “Free”) in order to properly analyze the data.
4. Data analysis will involve analysis of frequent words in order to generally understand their sentiment and relation to potential categories and clusters.
5. Potential to leverage the “Word2Vec” library in order to better optimize clusters.
6. Implement unsupervised learning techniques: bag of words, k-means.
7. Repeat steps for a different metro area and compare results.

Deliverables

Associated code, paper, and presentation. In addition, I plan to publish a blog post as my final presentation.