CS513: Theory & Practice of Data Cleaning

Final Project - Phase I

Kushagra Soni (soni14@illinois.edu)
Pericles Rocha (procha2@illinois.edu)
University of Illinois at Urbana-Champaign

This report details the data set which will be used on our data cleaning project, as well as some of the existing data quality issues with the data and how we expect to address these issues.

Dataset

Our data set (*D*) contains Craigslist listings of apartments for rent in the city of Chicago, Illinois. This data was scrapped manually using a Python program – meaning we didn't get access to the actual, raw data. While this process automates data gathering of web pages at scale, data often lands with quality issues that need to be addressed before it can be useful for analysis.

Target use case (U_1)

The main use case we expect to address with this data is to able to analyze what features influence rent prices the most in the city (U_1) to help users invest in the areas that matter the most for their listings. By performing this analysis, we could build a recommendation system that helps users understand how to maximize the details of their Craigslist listings to increase their odds of getting responses to their listings.

Note: the actual application that performs the recommendations is not a part of this project. We will however describe how we plan to prepare the data for this use case, as well as the two minor use cases.

The data as it was acquired is not of an acceptable quality to allow this use case. As we explain in the dataset description section, some of the listing details are added manually by the users and are contained within the same descriptive field (for example, the listing-title, listing-description, and listing-features columns). Therefore, data cleaning is needed to be able to address the target use case (U_1).

There are also two minor use cases that we envision with this data: one that requires zero data cleaning (U_0), and one that no amount of data cleaning or wrangling will make this dataset suitable for the use case (U_2). Let's look at these cases in detail.

Case (U_0): existing data may be good enough

One use case that we can contemplate here is to show users how long before the actual availability of an apartment a listing gets posted. For example: some listings are made 15 days before the apartment is in fact available for rent. Other listings are done 10, 20, 30, or even 45 days before the apartment will be available to move in. This data is clean enough and offered in two of the dataset's columns: listing-availability and listing-posting-date. By analyzing this data, we can produce averages, or a full bell curve of the distribution of how early apartments are made available before listings, to help users determine the right time to create their listing.

This is important because listings made too early can be ignored and left forgotten in the middle of hundreds of other listings. At the same time, listings for apartments that are made available too soon could cause anxiety from buyers and sellers, preventing a proper negotiation period.

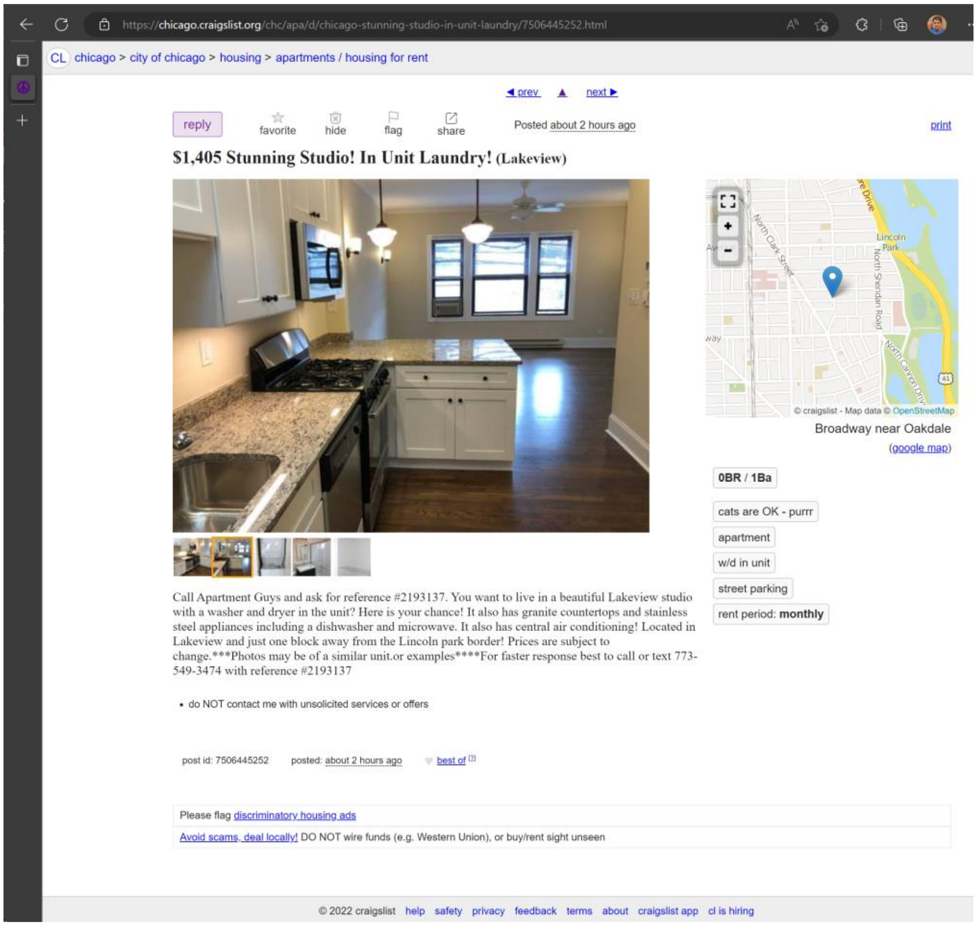
Case (U_2): existing data may never be good enough

Some aesthetical aspects of apartment listings can only be represented by pictures and videos included in listings. Also, the user description provided for listings are subject to the announcer's eyes. A proper recommendation system would require video and photo analysis to "learn" what sorts of visual aids are given for some listings that help make them more interesting for buyers, or how they affect the listing's price.

No amount of data cleaning on this dataset would help address the scenario on case U_2.

Dataset description

As mentioned, our dataset was acquired manually by using text retrieval and search engine techniques. The apartment listings that contain the data that we scraped can be seen on https://chicago.craigslist.org/search/apa. Image 1 shows how listings are displayed on Craiglist.org.


https://chicago.craigslist.org/chc/apa/d/chicago-stunning-studio-in-unit-laundry/7506445252.html
CL chicago > city of chicago > housing > apartments / housing for rent
◀ prev ▲ next ▶
reply
favorite
hide
flag
share
print
North Clark Street
Lincoln Park
North Sheridan Road
North Cannon Drive
41
© craigslist - Map data © OpenStreetMap
(google map)
best of [?]
© 2022 craigslist help safety privacy feedback terms about craigslist app cl is hiring

Posted about 2 hours ago

$1,405 Stunning Studio! In Unit Laundry! (Lakeview)

Broadway near Oakdale

0BR / 1Ba

cats are OK - purrr

apartment

w/d in unit

street parking

rent period: **monthly**

Call Apartment Guys and ask for reference #2193137. You want to live in a beautiful Lakeview studio with a washer and dryer in the unit? Here is your chance! It also has granite countertops and stainless steel appliances including a dishwasher and microwave. It also has central air conditioning! Located in Lakeview and just one block away from the Lincoln park border! Prices are subject to change.***Photos may be of a similar unit.or examples****For faster response best to call or text 773-549-3474 with reference #2193137

- do NOT contact me with unsolicited services or offers

post id: 7506445252 posted: about 2 hours ago

Please flag discriminatory housing ads

Avoid scams, deal locally! DO NOT wire funds (e.g. Western Union), or buy/rent sight unseen

Image 1: Typical listing on Craiglist.org

Our text retrieval program acquired data from these listings by directly reading the HTML sections of the listing pages, such as the listing page on Image 1. After performing text retrieval from listings and sub-pages (e.g.: listings directory, and then the listing detail pages), all data was saved to a comma-separated file (CSV).

Here are the columns present on our dataset:

Column name	Description
web-scraper-order	An identifier code produced by the search engine application used for text retrieval. Not needed for the analysis.
web-scraper-start-url	Source URL of the listing.
pagination	Source URL of the listing including the page number in the web application's pagination.
listing-title	Title of the listing as provided by the user. May contain price, number of bedrooms, location, and rent conditions (such as pet policy).
listing-description	Description of the listing as provided by the user. It contains listing details such as apartment features, location, and more. Also contains metadata related to availability of a QR code that links to this post.
listing-housing-type	Number of bedrooms and baths in the listing separated by a "/" character.
listing-features	List of apartment features as provided by Craigslist. This is categorical data that can be used to cleanly differentiate listings, but actual features are all contained in this column.
listing-notices	A note that indicates if the person who made the listing is open to receive communication for unsolicited services and offers.
listing-id	An identifier that represents the number of the post in Craigslist.
listing-link	The actual application shows a link to the listing here. However, the data captured shows the listing price and a carousel of pictures available.
listing-link-href	Direct link to the specific posting.
listing-availability	Date when the listing will be available for rent.
listing-posting-date	Date when the listing was posted on Craigslist.
listing-address	Physical address of the property listed.

Also, we're offering additional metadata about this data set:

- **Spatial extent**: this dataset contains apartment listings for the city of Chicago, IL only.
- **Temporal extent**: this dataset contains all listings available on Craigslist until the date this dataset was scrapped (July 7, 2022).
- All the data in the dataset is presented as clear text.

Obvious data quality issues

Our intent for this project is to offer a clean data set that would allow a better understanding of features that help influence listing prices. To perform this analysis, we would need to have the value of the apartment listed as well as any details that help describe this apartment.

While the data needed to support our use-case is available in the dataset, it is not presented cleanly, and is not ready for analysis. For example:

- **Price**: no column presents this data cleanly, separately, in a clean fashion. We need to cleanse one of two columns (listing-title or listing-link) to be able to capture the price on the listing and make it useful for analysis
- **Listing features**: this information is never presented cleanly in the dataset. We need to gather data from the listing-title, listing-description, listing-housing-type, listing-features, and list-address columns to separate and organize listing details. This information is never presented cleanly in any of these columns.
- **Listing posting date**: this information is presented in a format that shows "x days ago" (e.g.: 10 days ago, or 5 days ago). It doesn't include specific dates of the posting. Since we know the date when this data was acquired, we can calculate the actual listing date and offer this information more cleanly to users, allowing them to work with time dimension tables to slice the data.
- **Readability**: the descriptive data provided by users in the listing-title and listing-description columns is not standardized and can contain all-caps or non-caps text. We will attempt to make data in this column more readable by standardizing input formats.

Plan of action

We intend to use text cleaning techniques and natural language processing to clean the dataset. We will load it into a Jupyter notebook and use Python to perform the data cleaning process. This will allow us to show a step-by-step process of how we're cleaning the data, explain the tasks throughout the notebook, and show previews of the data as we clean it.

Work is being split across our team by separating the tasks amongst the members:

- **Data acquisition**: Kushagra Soni
- **Phase I project description report**: Peri Rocha
- **Source control and code management**: Kushagra Soni
- **Data cleaning tasks**: split equally between team members

All work is reviewed by both team members.