MAPC Rental Listings Final Presentation

Yunruo Ni & Gowtham Asokan

Objective

- We needed to label the data manually so that a Machine learning model may be trained to accurately predict duplicate listings in the future.

- We created a labeling matrix for likely duplicates based on the similarity of column values.

Overview of labeling matrix

- We dropped duplicates that were collected repeatedly every week - different posting dates but everything else was the same.
- We created a groupby sort using “ask”, “image_id”, “bedroom”, “location” that created a group ID for each.
- Ran location based clustering to group together rental listings that are in close proximity to each other. Based on the k-means clustering algorithm.
- We grouped the titles that had the first three words as the same to each other.
- Attempt to use tf-idf to create similar title groups within each location cluster proved to be very slow.
- All these metrics should indicate whether a listing is a duplicate or not.

Deduplication

- Deduplicating for the multiple posting dates brought down the dataset from **158,740** to **44,943** entries.

- Each of the entries do not have the exact same values to each other - they are unique.

Group ID based on all information except “title”

- Grouped by based on “ask”, “image_id”, “bedroom”, “location”. (Adding one more column called “gid”) It has total **32,050** different groups.
- Within each group ID we compared the title’s fuzzy similarity to each preceding one.
- We flagged those titles that are not similar to each other (fuzzy ratio score < 60)
- Finally, we got **2674** entries through the whole dataframe flagged as not having similar titles within its group.
- **32,050** group titles are available and within them **2674** were found to be false based on title similarity.

Location based clustering

- We ran a k-means clustering for **70** clusters.

- **10** of the clusters were minimal with 1 rental listing each, the biggest cluster **4070** rental listings.

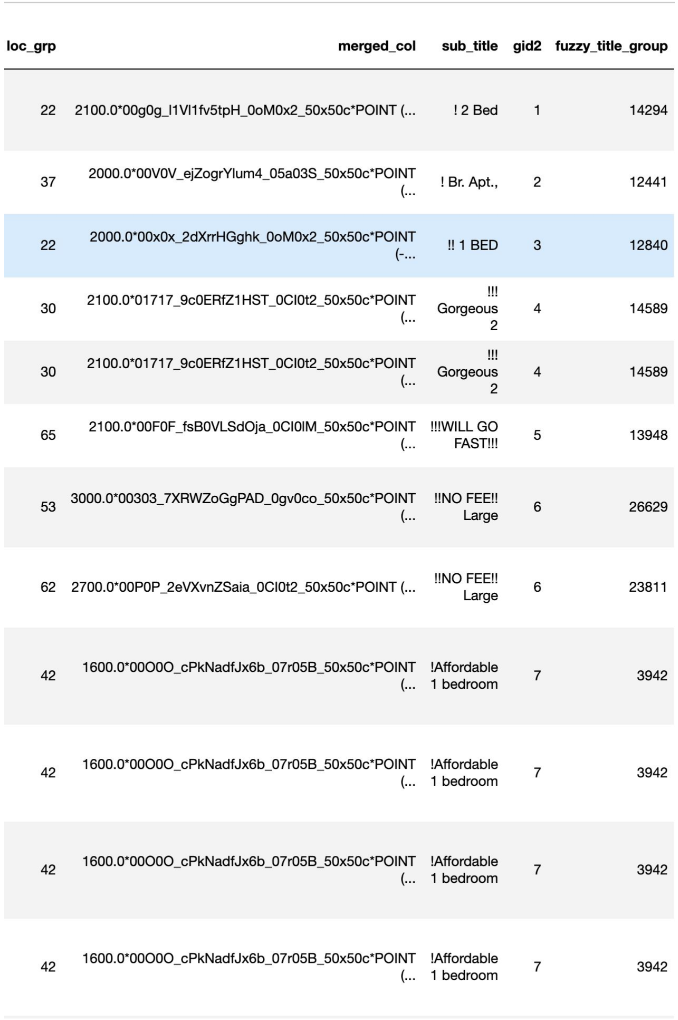
- Median rental listings in a cluster were **132**.

“Fuzzy_Title” Grouping

- We created a column called “sub_title” based on the first three words of the title.

- We then used the sort_values function to order the “sub_title” column.

- This gave us **32,791** title groups based on the “sub_title” column.



loc_grp	merged_col	sub_title	gid2	fuzzy_title_group
22	2100.0*00g0g_l1Vl1fv5tpH_0oM0x2_50x50c*POINT (...	! 2 Bed	1	14294
37	2000.0*00V0V_ejZogrYlum4_05a03S_50x50c*POINT (...	! Br. Apt.,	2	12441
22	2000.0*00x0x_2dXrrHGghk_0oM0x2_50x50c*POINT (-...	!! 1 BED	3	12840
30	2100.0*01717_9c0ERfZ1HST_0Cl0t2_50x50c*POINT (...	!!! Gorgeous 2	4	14589
30	2100.0*01717_9c0ERfZ1HST_0Cl0t2_50x50c*POINT (...	!!! Gorgeous 2	4	14589
65	2100.0*00F0F_fsB0VLSdOja_0Cl0lM_50x50c*POINT (...	!!!WILL GO FAST!!!	5	13948
53	3000.0*00303_7XRWZoGgPAD_0gv0co_50x50c*POINT (...	!!NO FEE!! Large	6	26629
62	2700.0*00P0P_2eVXvnZSaia_0Cl0t2_50x50c*POINT (...	!!NO FEE!! Large	6	23811
42	1600.0*00O0O_cPkNadfJx6b_07r05B_50x50c*POINT (...	!Affordable 1 bedroom	7	3942
42	1600.0*00O0O_cPkNadfJx6b_07r05B_50x50c*POINT (...	!Affordable 1 bedroom	7	3942
42	1600.0*00O0O_cPkNadfJx6b_07r05B_50x50c*POINT (...	!Affordable 1 bedroom	7	3942
42	1600.0*00O0O_cPkNadfJx6b_07r05B_50x50c*POINT (...	!Affordable 1 bedroom	7	3942

Conclusion and Takeaways

- **32,050** group titles from grouping all columns except for “title” and **32,791** from using sort_values on the first three words of the “title” column.
- This indicates that there are roughly 32 thousand rental listing groups based on the various column values.
- Between these two grouping methods **2,836** are not overlapping, hence both methods roughly captured the group titles.
- We hope that these methods can now be used for labeling a rental listing as likely duplicates, etc. Manually labeling within each location cluster and title group should give a high degree of confidence backed by the manual check.