craigslist

create a posting

my account

search craigslist

# “the best AUD” team

## team members

kai-duan chang

li-ci chuang

su-tien lee

yen-tsz huang

yi-hsuan hsu

meghan harris

## final group project

12/8/2021

## background

As of 2018, Craigslist has surmounted over 60 million users with over 50 billion pageviews per month and over 80 million ads posted on Craigslist monthly (expandedramblings.com). So, what is Craigslist? Craigslist is one of the most popular classified advertisements website used for viewing and posting advertisements. Craigslist's goal is to help people easily find, buy, or sell anything. However, with these incredibly high user-visitation numbers, comes neglected improvement on user-interface areas within less popular pages. As hired consultants for Craigslist, we were asked to improve the ad-viewers' experience by analyzing Craigslist's business model and process of the platform to find improvements and solutions for this proposal with a demo implementation.

After reviewing Craigslist's platform and interface, we discovered an area of improvement that would greatly improve the user-experience. Within Craiglist's site "lost+found", there is no option for posts to be automatically labeled as lost or found. Currently, advertisers/posters manually label their posts as lost or found which is an additional step that takes time. In addition to this, it's not required that advertisers/posters label their post as lost or found, so the current sort by functionality doesn't encompass all the posts. By having the ad be automatically labeled as lost or found, it helps the advertisers to save time and ensures all posts are accurately labeled. As users looking within this section are clearly visiting this site with the clear intention of either viewing the site in terms of "lost" or "found", we identified this as a problem that we could solve.

Not only would this greatly improve the advertising experience, but also would improve the user-experience as well. This subsection improvement is a win-win that benefits all parties once the improvement has been implemented.

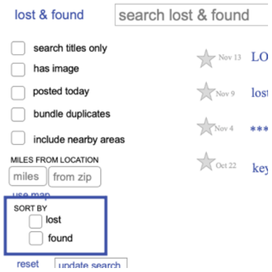
## business analysis

Once we defined Craiglist's problem, we created three project objectives that supported solving the business problem.

1. Provide users with the ability to search for all ads via lost or found
2. Provide users with quick links to specific posts based on the most popular topics
3. Identify images based on keyword

### First Project Objective:

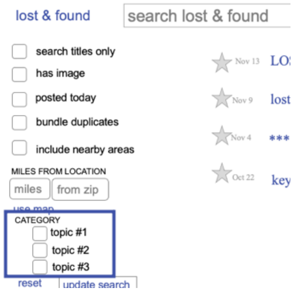
The first objective defined was to provide users with the ability to search for all ads via lost or found. To accomplish this objective, we first had to define how to determine if a post is lost or found. Upon starting this project, we quickly discovered that some post titles are mis-labeled as the opposite of what type of post they are. For example, a post would be labeled as “lost cat”, however the post content clearly described that they found a cat and phrased the title in terms for the user to find. To avoid this issue, we text mined based on the post content instead of the title as the post content typically had more words and more context for the model to catch. In addition to this, not all of the posts used the terms “lost” or “found. To combat this issue, we identified key words associated with “lost” or “found”; for example, “missing”, “stolen”, “left” were words defined for lost and “searching” and “looking for” were words used for “found”.



*Example of goal for objective #1*

## Second Project Objective:

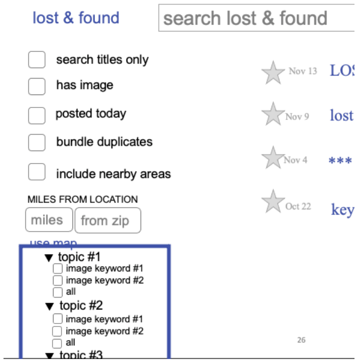
The second project objective was to provide users with quick links to specific posts based on the most popular topics. This project objective was set as a further improvement towards Craiglist’s current problem. Based on user behavior, we know that typically users come to this page to look for something specific. By quick linking the most popular topics within the sidebar, users would be able to immediately click on the posts related to what the most popular posts are. This not only saves time for the users to sift through irrelevant topic posts, but it also improves the user experience. However, with this project objective, we had to determine which topics within the lost+found are the most popular. Luckily, this can be done easily within the data analysis section.



*Example of goal for objective #2*

## Third Project Objective:

The third project objective is to include a third component that utilizes image recognition to provide keyword sub-topics within the most popular topics. With this project objective, the goal was to find the keywords that have the most images and can be further distinguished from other posts within the same topic. For this, we would find posts with images within a popular topic to further provide a subsection within the topic. The purpose of this objective is to improve the user-experience as we know the user-motivation on this page is to look for something either lost or found urgently. Within the most popular topics, it further improves user-experience to provide quick links to related keywords that image recognition detected the image to be.



*Example of goal for objective #3*

### Web Scraping

Before starting the model construction, we scrapped around 1,000 Lost and Found postings from Craigslist, including 490 postings in the West Lafayette area and 595 postings in Los Angeles, to ensure that enough data acquired for the model training. We chose to include Los Angeles, CA in addition to the West Lafayette, IN dataset because we wanted to ensure we had enough data to produce an accurate model for Craigslist. We chose LA as we feel it is representative of a typical city and would have a lot of data for us to include.

We found the lost and found postings in the West Lafayette area by using the zip-code 47906 and including the posting-distance to be within 250 miles; for LA, we used the zip-code of 90036 and included the posting-distance to be within 120 miles. To extract the data, we first created a for loop to retrieve the titles of the top 5 pages among the postings in the areas respectively and saved them as lists. Secondly, we retrieved the contents of the postings for each title in the above-mentioned list, saved them as another list, converted and saved them into csv files respectively.

### Classification Models

- **Data Processing**

Once we had the list of both WL and LA posts, we labeled them into “lost” or “found” category manually (put “lost” = 1, put “found” = 0), so that each of the 995 posts would have defined labels to do the classification training and prediction. After that, we did data partitioned, split the data into train set and test set with 70% and 30% respectively, and separated the label and post content.

- **TFIDF Vectorizing**

To turn the content text into vector to fit the model, we needed to vectorize them. First, we tokenized and normalized them with the Lemmatize function, where we used a pre-defined database to lookup lemma and removed punctuations or filtered out all special characters. Then we did TD-IDF conversion to weight our terms. Since we would like to drop less important words to reduce the dimension, we set the minimum document frequency equal to 5, which removed all the words with frequency less than 5. In addition to this, we also wanted to keep order information, so we included bag of 2-grams in our argument. Upon completing these steps, the data was prepared to be tested in different models.

- **Model**

Our objective is to train a model to automatically label the post into “lost” or “found” category, so it will be a classification problem. We tried Naïve Bayes, Decision Tree, Random Forest and SVM, four models to do the classification.

### Analysis of Models

#### Model Accuracy

We predicted the classification with our test set, and the table below listed accuracy rate from our four models:

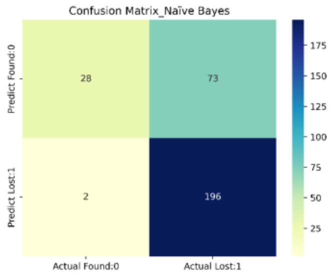
| | |
|---|---|
| Naive Bayes model Accuracy | 74.92% |
| Decision Tree Model Accuracy | 70.57% |
| Random Forest Model Accuracy | 66.22% |
| SVM model Accuracy | 77.93% |

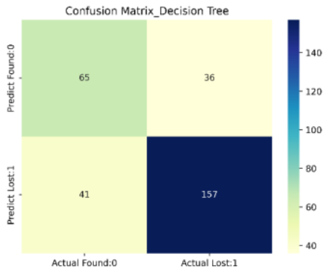
We selected the SVM model as our classification model because it is advanced and has the highest accuracy rate. It’s worth mentioning that the Naive Bayes model is a classic and simple model that may not be suitable for advanced analysis for thousands of posts in the future. And the poor performance of Decision Tree may be due to its overfitting issue because it might just memorize the outcome and be very fitted into training data.

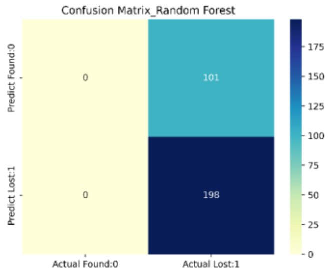
#### Confusion Matrix

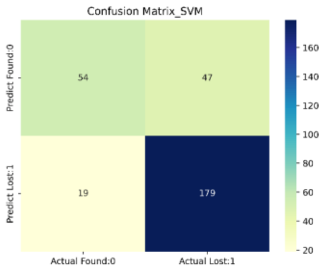
Besides the model accuracy, we also wanted to see the detailed distribution of our prediction given if the result is “found” or “lost” to accurately assess if our model misclassified anything. For example, if the post was predicted to be “found” when it is actually “found”. Therefore, we used confusion matrix to check this.  Picture 1 to Picture 4 are the charts for our four models.

One interesting thing is that all the models seem to do a greater job in identifying “lost” posts than “found” posts. An explanation for this is that most people are worried when they lose their stuff/pets; due to this, they would use stronger words to show their emotion which makes model easier to capture the categories. Also, the sample size matters as well. For example, the number of “lost” posts are larger than the “found” posts in the dataset. Because of this, it makes sense it does a better job classifying the “lost” posts.









- **Topic Model**

In addition to the classification model that classifies the posts, we want to capture the top topics within lost and found. To do this, we had to remove words like “found”, "missing", "lost" “contact", "please", "info", "show" and "home" as these are words related to our posting topics that we already know and don’t need to categorize. In 3 topic models, we listed top 5 keyword and found that dog and cat appear across our models, meaning that they are the most possible focus in posts. Also, key and wallet are also popular in postings. Below are the listed key topics with 5 keywords.

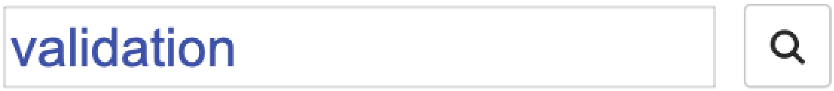
**Topic 0:**
cat old dog female name
**Topic 1:**
near cat back dog one
**Topic 2:**
key dog wallet around street

- **Image Recognition**

Once we found the top topics from the posts, we utilized image recognition to provide even more value to users. Because we saw the words “dog” and “cat” consistently across our top topics, we realized this is due to people not categorizing on specific breeds or detail. Therefore, if we can recognize the breed, color, or other characteristics, we can create the tags automatically. This enables website users to investigate all the related posts easily and timely when looking for dogs or cats. For example, we used image recognition on “dog” and “cat” from images to further identify the image. The image recognition did just so, the top 3 ranking keywords for both images are listed below.

('n02113624', 'toy_poodle', 0.5213419), ('n02113712', 'miniature_poodle', 0.43107972), ('n02113799', 'standard_poodle', 0.012640942)

('n02124075', 'Egyptian_cat', 0.24735513), ('n04033995', 'quilt', 0.09394641), ('n02123597', 'mouse', 0.07160977)

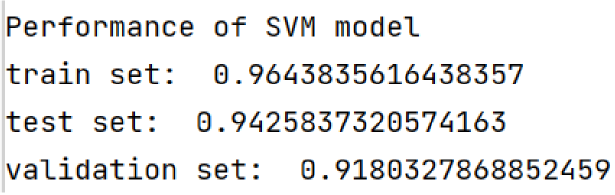


validation

To ensure our model was performing correctly, we separated our data into separate train, test, and validation sets. Separating the data into an additional validation set allows us to objectively evaluate the performance of the models we have run earlier to find and optimize the best model for our business problem. To first allow us to validate our model, we manually went through our dataset with over 1,000 datapoints and coded our predicted column “y” with 1 if it should be categorized as lost or 0 if it should be categorized as found. If the post content could not be identified as being labeled lost or found, we removed this from the dataset; however, there weren’t many that needed to be removed.

Once this was done, our data set was ready to be split into train, test, and validation sets. To do this, we utilized the train_test_split function within from sklearn.model_selection. We first randomly split our data into 30% test and 70% train sets. Then, to get our validation set, we split the train set into 75% train and 25% validation. This allows the model to have a new set of data to have the model unbiasedly estimate the accuracy of the models.

Once we had our validation sets, we evaluated all our models based on the train, test, and validation datasets. We found that all the models improved significantly with the addition of a validation set. However, the best model was the SVM model (which was also the best model from our earlier model analysis). It should be noted that with the addition of the validation set that our accuracy improved almost over 20%!



```
Performance of SVM model
train set:  0.9643835616438357
test set:  0.9425837320574163
validation set:  0.9180327868852459
```

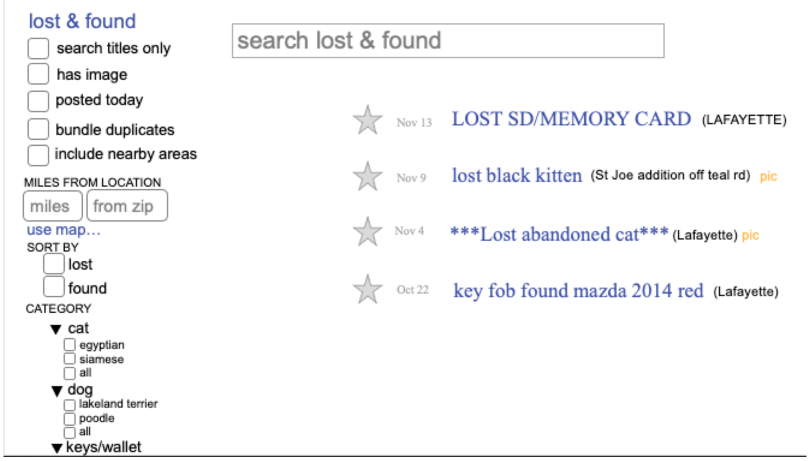
# conclusion

In conclusion, our analysis allows Craigslist to improve their user-experience by allowing ads within the lost + found section to be automatically filtered into lost or found without manual entry. Not only this, but we also took it a step further to allow users the option to “quick link” to the most popular topics within the lost + found section to allow users to find posts much faster and easier. In addition to this, we utilized image recognition on posts within the most popular topics to further label and distinguish posts from another to further improve user-experience. For example, with more detailed information being identified, such as the breeds of the animals, the users could use them as the advanced filters to search for their lost or found items more conveniently and precisely.

We wholeheartedly believe that this project and model improves user-experience and brings value both to the user and to Craigslist. We not only considered what best benefits the user and Craigslist respectively, but also researched and understood the user behavior to best optimize value within the “lost+found” page. By increasing the amount of efficient and accurate posts, this then increasing the credibility, which could in turn increase the number of users. With an increased number of users, Craigslist will be able to have an increase in advertisers, which would increase the amount of revenue for Craigslist.

# appendix

Our final mock-up with recommended improvements:



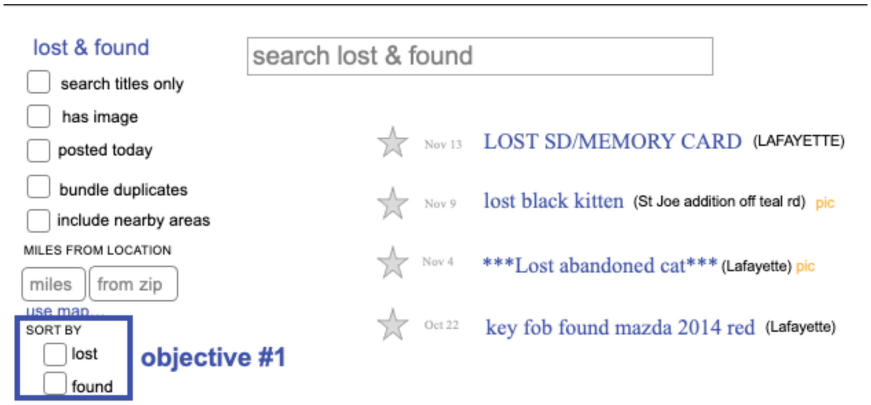
☆ Nov 13 LOST SD/MEMORY CARD (LAFAYETTE)

☆ Nov 9 lost black kitten (St Joe addition off teal rd) pic

☆ Nov 4 ***Lost abandoned cat*** (Lafayette) pic

☆ Oct 22 key fob found mazda 2014 red (Lafayette)

Objective #1 mock-up:


lost & found

search lost & found

□ search titles only

□ has image

□ posted today

□ bundle duplicates

□ include nearby areas

MILES FROM LOCATION

miles | from zip

use map…

SORT BY

□ lost

□ found

objective #1

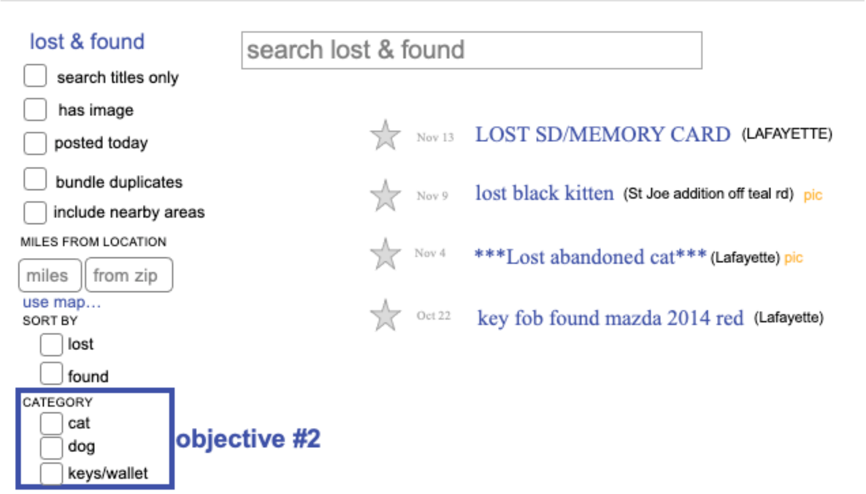
☆ Nov 13 LOST SD/MEMORY CARD (LAFAYETTE)

☆ Nov 9 lost black kitten (St Joe addition off teal rd) pic

☆ Nov 4 ***Lost abandoned cat*** (Lafayette) pic

☆ Oct 22 key fob found mazda 2014 red (Lafayette)

Objective #2 mock-up:


lost & found

search lost & found

□ search titles only

□ has image

□ posted today

□ bundle duplicates

□ include nearby areas

MILES FROM LOCATION

miles | from zip

use map…

SORT BY

□ lost

□ found

CATEGORY

□ cat

□ dog

□ keys/wallet

objective #2

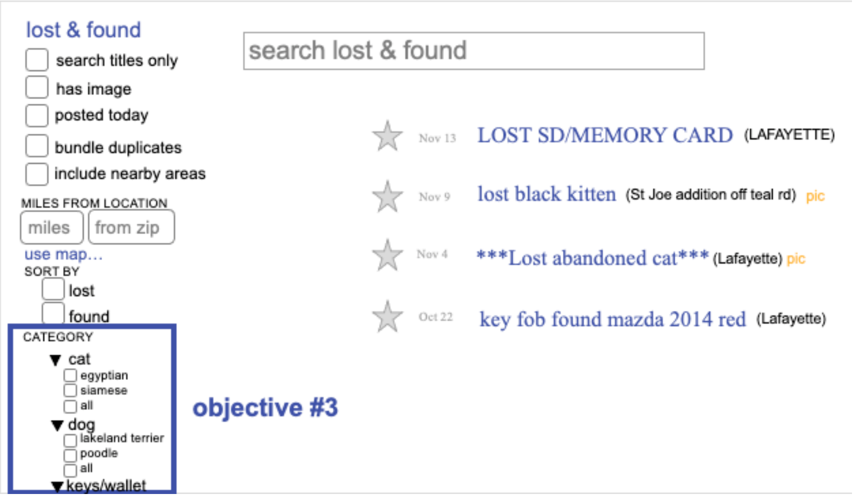
☆ Nov 13 LOST SD/MEMORY CARD (LAFAYETTE)

☆ Nov 9 lost black kitten (St Joe addition off teal rd) pic

☆ Nov 4 ***Lost abandoned cat*** (Lafayette) pic

☆ Oct 22 key fob found mazda 2014 red (Lafayette)

Objective #3 mock-up:



lost & found

□ search titles only

□ has image

□ posted today

□ bundle duplicates

□ include nearby areas

MILES FROM LOCATION

miles | from zip

use map…

SORT BY

□ lost

□ found

CATEGORY

▼ cat
- □ egyptian
- □ siamese
- □ all

▼ dog
- □ lakeland terrier
- □ poodle
- □ all

▼ keys/wallet

objective #3

search lost & found

☆ Nov 13 LOST SD/MEMORY CARD (LAFAYETTE)

☆ Nov 9 lost black kitten (St Joe addition off teal rd) pic

☆ Nov 4 ***Lost abandoned cat*** (Lafayette) pic

☆ Oct 22 key fob found mazda 2014 red (Lafayette)