

Recommending similar listings on craigslist

MGMT 59000-005 Final Presentation

Unstructured Data Analysts

Akarsh Sinha, Girish Sharma, Keerthana Nemili,
Mandeep Singh Rahi, Manmeet Walia

TABLE OF CONTENTS



01

Business Problem



02

Methodology



03

Pre-processing
and Exploratory Data
Analysis



04

Models & Comparison



05

Recommendations



06

Appendix



01 Business Problem

Problem Statement

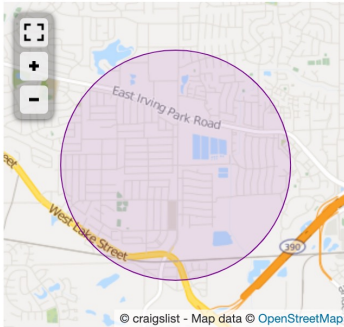



Craigslist does not provide recommendations of similar products & services

- Lack of recommendations would lead to customer drop-offs if they are unable to find the desired listings
- Drop-offs would result in lower sales and ultimately low ad-sales.

[reply](#) [favorite](#) [hide](#) [flag](#) [share](#) [Posted 25 days ago](#) [print](#)

iPhone 12 in Black 64 GB - Brand New sealed - \$580



condition: **new**

make / manufacturer: **Apple**

mobile OS: **apple iOS**

model name / number: **iphone 12**

iPhone 12 in Black 64 GB - Brand New sealed - \$580

factory unlocked for all mobile carriers.
I only sell locally and accept cash.

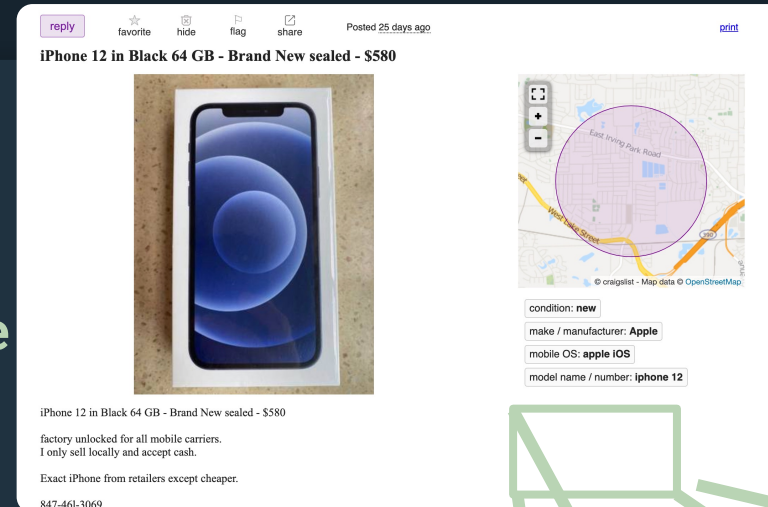
Exact iPhone from retailers except cheaper.

847-461-3069

Objective

Building a recommendation system to show 5 similar listings on the product page

- **Low resistance in the buying process**, as visitors would be redirected to recommended product page
- Will result in **increase in ad revenue**.
- Firms like Amazon attribute **~35%** of their revenues to cross selling.



- ★ Dec 6 Unlocked Apple iPhone 14 Pro Max with Box and accessories \$450 (day) 151.3mi
- ★ Dec 6 Brand new 14 Pro max for sale 256GB fully unlocked \$500 (spi) 153mi
- ★ Dec 6 iPhone 13 Pro Max 256gb (unlocked) with charger \$400 (spi) 153mi
- ★ Dec 6 Unlocked Apple iPhone 14 Pro Max with Box and accessories \$450 (evv) 166.5mi
- ★ Dec 6 Best ever iPhone 13 pro max unlocked 512GB \$300 (dil) 116.7mi



02 Methodology

Methodology



Web Scraping



~1200 Cell phone listings were scraped from Craigslist.

- Listing titles
- Location (lat,long)
- Price

Exploratory Data Analysis



- Frequency of cell-phone models
- Clustering the kind of listings on Craigslist

Pre-Processing



- Tokenized Listings
- Lemmatization
- Stop Words removal

Word Frequency Normalization



- TF-IDF

Modelling



- KNN
- Locality Sensitivity Hashing (LSH)

Evaluation



- Cosine distances
- Relative hit accuracy



03

Pre-Processing & Exploratory Data Analysis

Scraping

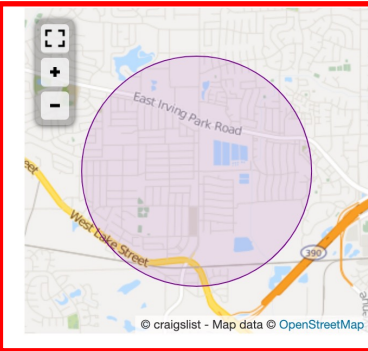



Scraped 1.2k cell-phones listings

- **Location:** Lafayette ± 250 miles
- **Keyword:** Cell phones
- **Data Scraped:**
 - ID
 - Title
 - Description
 - Link
 - Location (Latitude, Longitude)
 - Price
- **Libraries Used:**
 - Selenium
 - BeautifulSoup

reply favorite hide flag share Posted 25 days ago print

iPhone 12 in Black 64 GB - Brand New sealed **\$580**



condition: **new**

make / manufacturer: **Apple**

mobile OS: **apple iOS**

model name / number: **iphone 12**

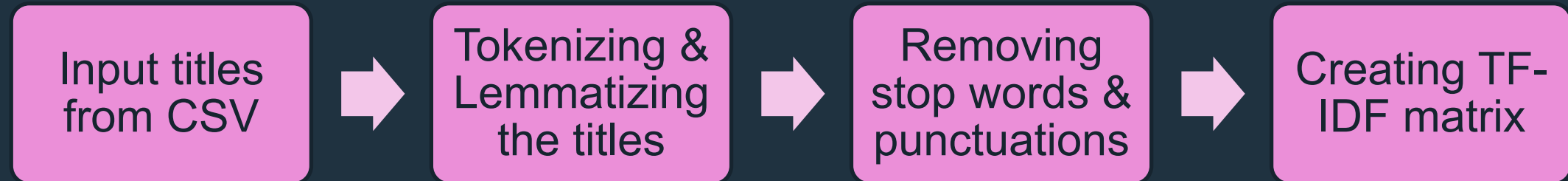
iPhone 12 in Black 64 GB - Brand New sealed - \$580

factory unlocked for all mobile carriers.
I only sell locally and accept cash.

Exact iPhone from retailers except cheaper.

847-461-3069

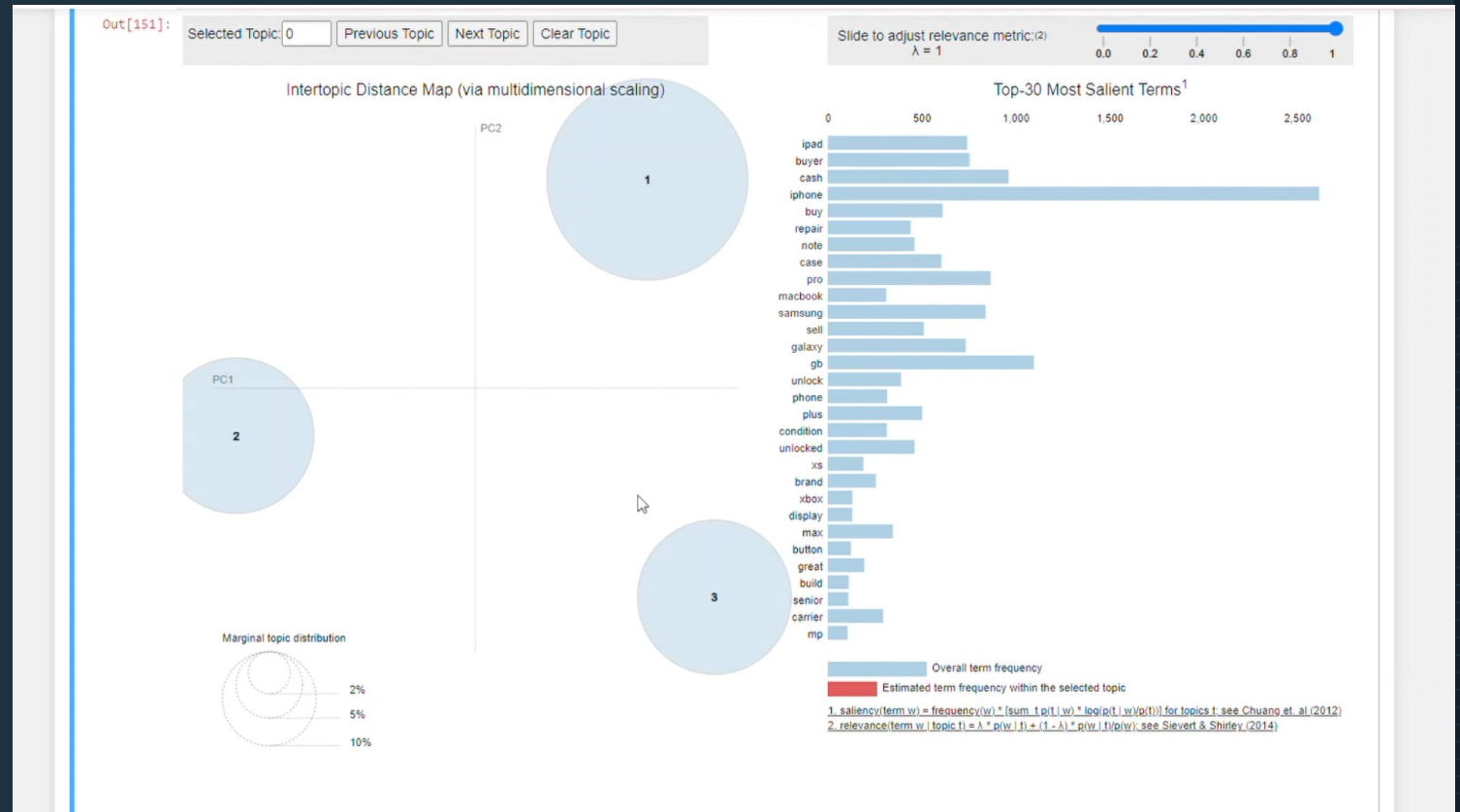
Pre-Processing



EDA

LDA to understand the type of listings

- **Cluster 1:** Contains sales ads of unlocked cell-phones with network carriers.
- **Imp. Keywords:** Unlocked, carrier, verizon, sell.
- **Cluster 3:** Contains buyer ads who buy phones and pay cash. Some of them also provide screen repairs and replacements.
- **Imp. Keywords:** buyer, repair, replacement.



EDA



42%

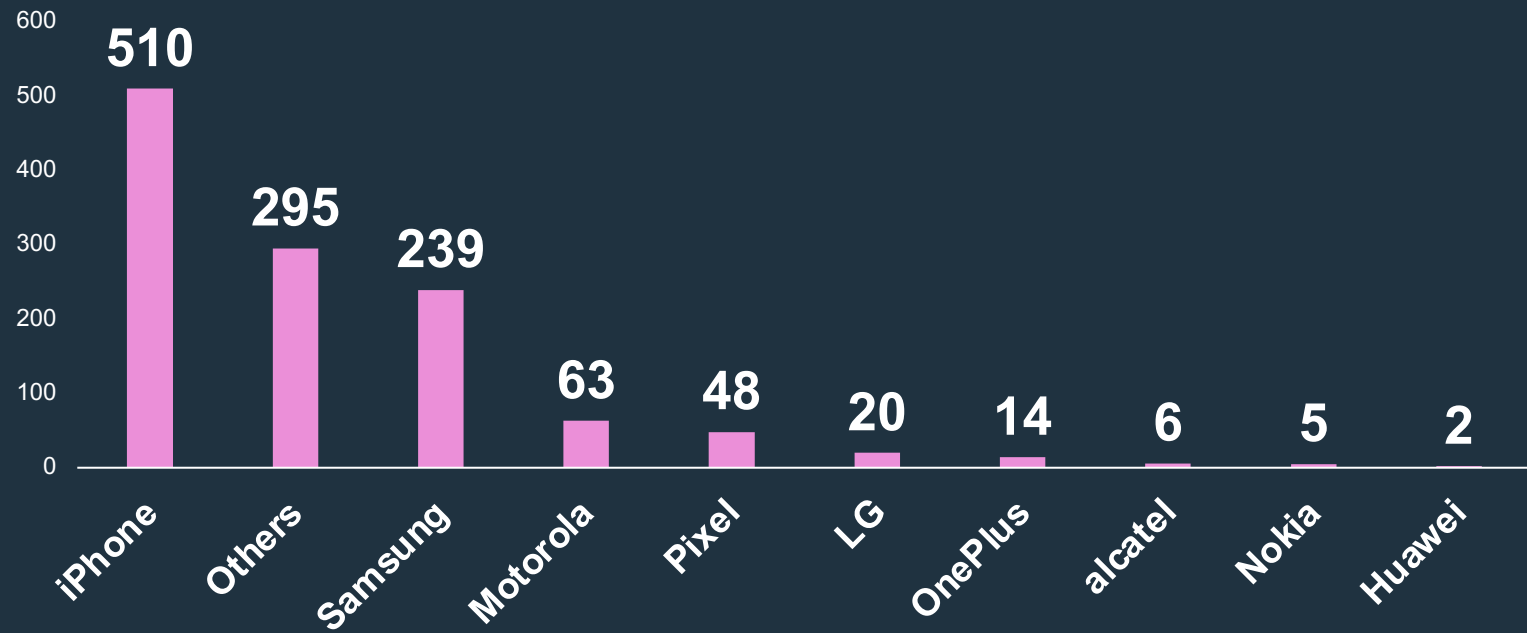
SAMSUNG

20%

Others

38%

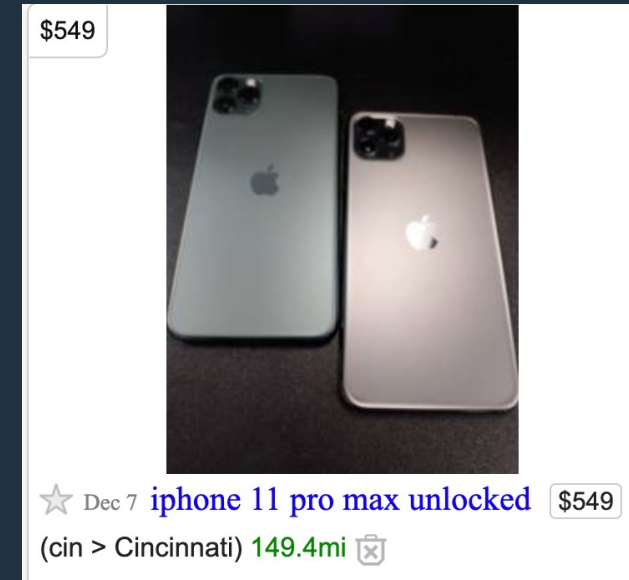
Frequency of Cell-Phone Models



Word Frequency Normalization

TF-IDF word weightage matrix to be inputted into subsequent models.

- TF IDF with 1,2,3 and 4 grams, with a minimum document frequency of 5.
- Took till 4 grams to exactly catch the cell phone's model
- Lesser grams included because full cell phone title not always available in listings



1 gram	2 grams	3 grams	4 grams
iphone	iphone_11	iphone_11_pro	iphone_11_pro_max



04 Models & Comparison

Feature selection



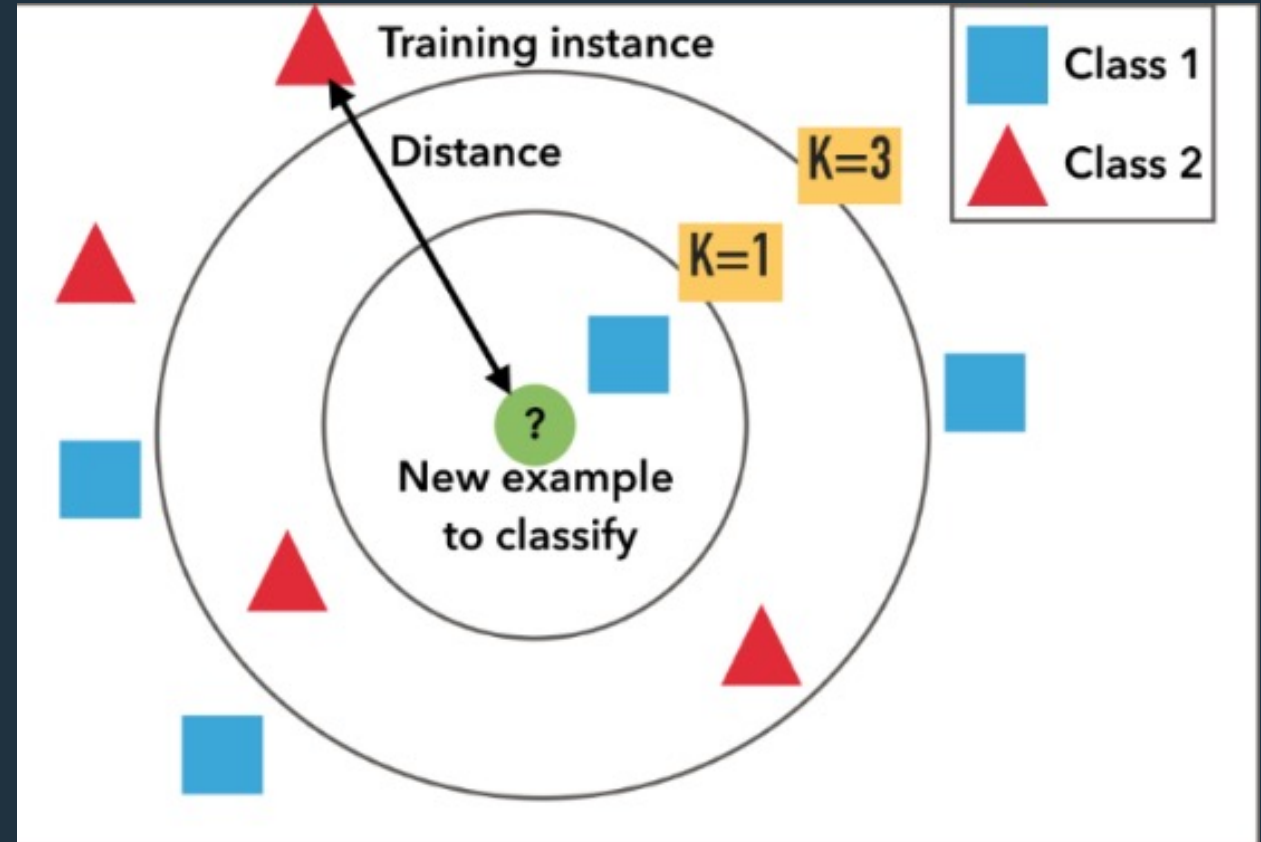
The following features were scraped and added along with TF-IDF vectors to enhance recommendations:

- Longitude and latitude of each ad to localize recommendations
- Price of each ad to show recommendations in similar-price ranges
- These features were normalized for modeling with TF-IDF weights

K-Nearest Neighbors

Based on item-based collaborative filtering

- Non-parametric, lazy learning method.
- Does not assume underlying data distribution but relies on item feature similarity.
- Used Cosine distance for NN-search as Euclidean distance is unhelpful in high dimensions.



K-NN sample output

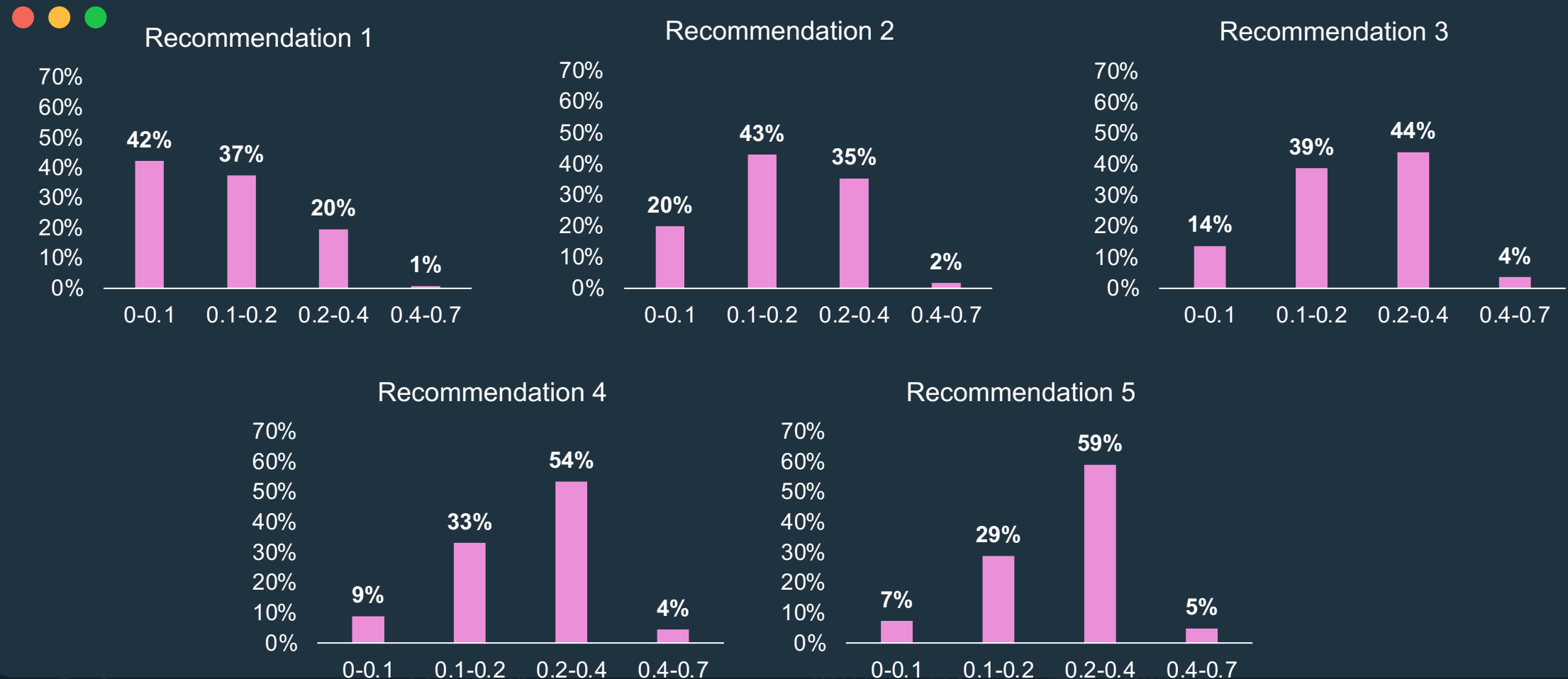


Large no. of ads have close neighbors suggesting accurate recommendations

Original Ad	NEW 10.1 PHABLET - 3G, 2 SIM Slots, 32 GB, Tablet, In Box, Orig \$105 - \$75 (St. Charles)	FACTORY UNLOCKED Pixel 7 Pro 128gb - \$650 (St.Louis)	
Recommendations	NEW 10.1 PHABLET - 2 SIM Slots, 32 GB, Phone Tablet, In Box - \$75 (St Charles)	BRAND NEW SAMSUNG S22+ ***UNLOCKED** - \$650 (KIRKWOOD)	
	NEW 10.1 TABLET - 2 SIM Slots, 32 GB, Phone Tablet, In Box Orig \$105 - \$75 (St Charles)	Google Pixel 7 128 Obsidian Black factory unlocked - \$590	
	Alcatel Joy Tablet - Wifi & Data Capable, 32 GB, W/Box, Sell - \$50 (St charles)	Samsung S22 Ultra 128gb Black Carrier Unlocked - \$550 (Saint Louis)	
	NEW 10.1 TABLET - 2 SIM Slots, 32 GB, Phone Tablet, In Box Orig \$105 - \$65 (St Charles)	Samsung Galaxy S22 ULTRA Verizon - \$650 (Decatur)	
	Alcatel 3T 8 (T-Mobile) tablet - \$80 (Shelby Twp)	Samsung S21 ultra 5G - \$500 (Fairview heights IL)	
Cosine distance bucket	0-0.2	0.2-0.4	
Proportion of ads in cosine distance bucket	55%	42%	0.4-0.7: 3%

K-NN accuracy

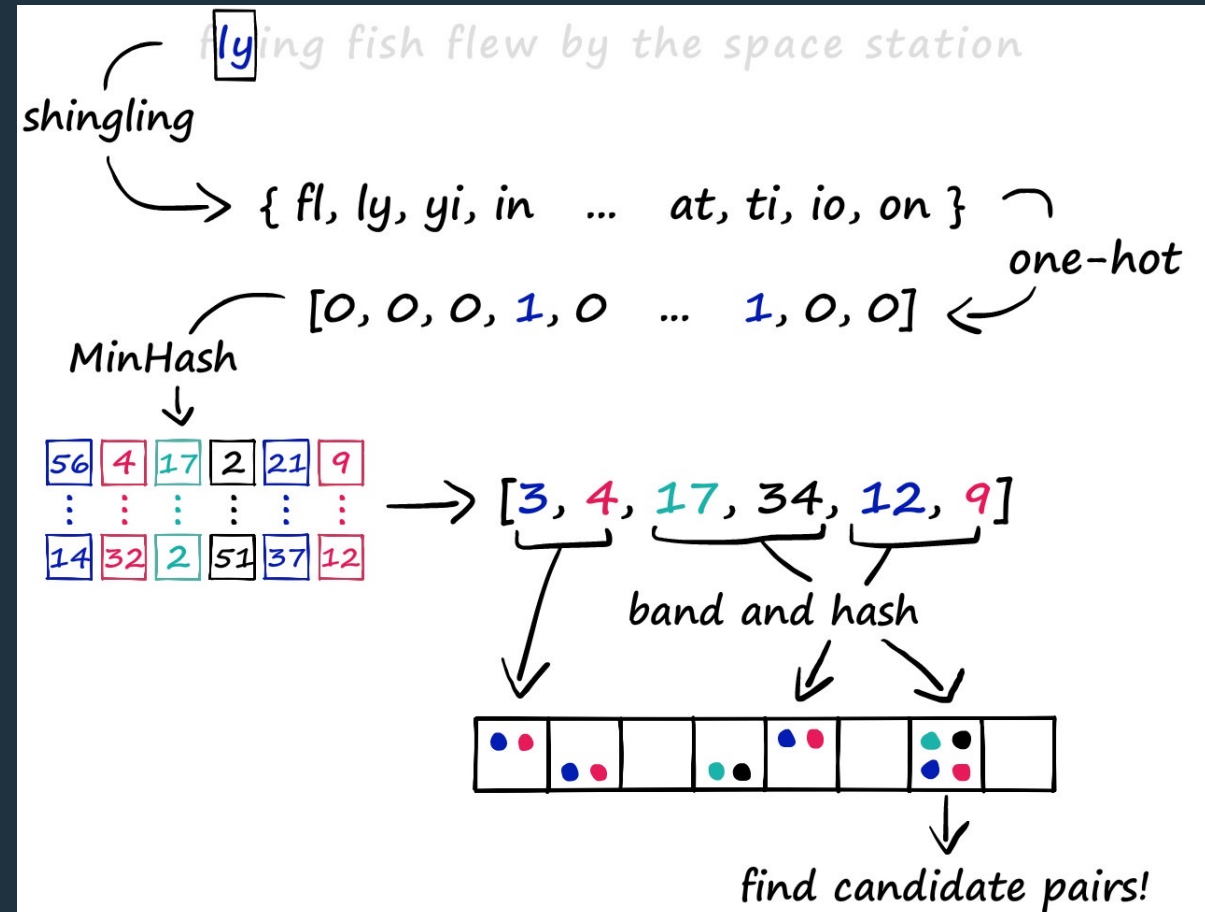
Proportion of ads in each cosine distance bucket



Locality-Sensitive Hashing

"Grouping" the data for faster computation

- Segments and hashes the same sample several times
- When it finds a pair of vectors hashed to the same value *at least once*, it tags them as candidate pairs.
- Based on Jaccard Similarity of sets



LSH performance

Reduces time complexity to find approximate nearest neighbors



Naïve: $60M \times 60M = 3.6 \times 10^{14} s = 11M$ years

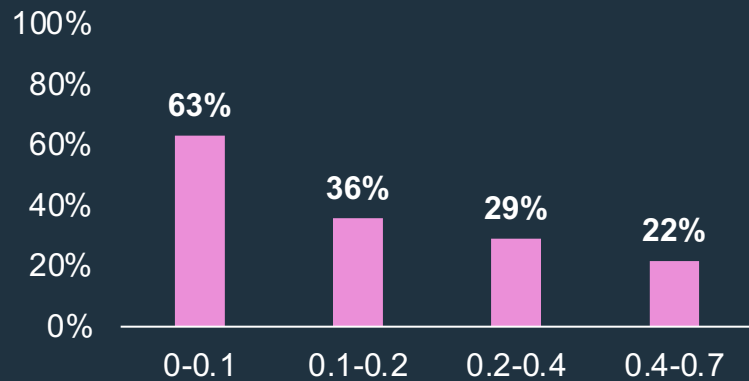
Smart: $60M \times \log(60M) = 46M s = 14$ years

LSH accuracy

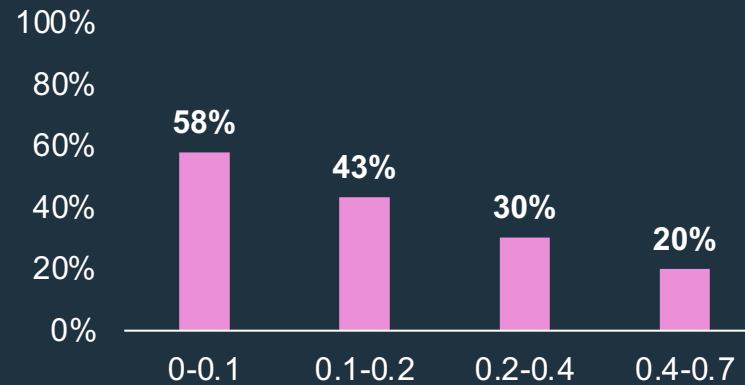
Hit accuracy
relative to k-NN



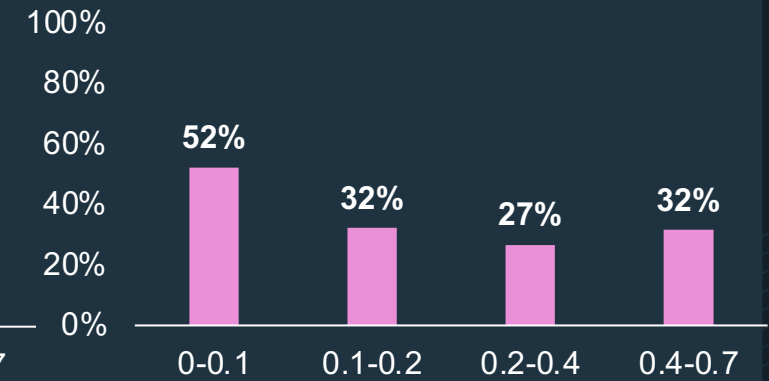
Recommendation 1



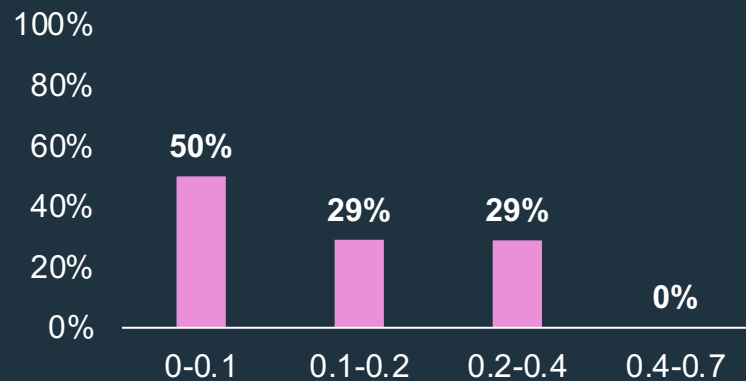
Recommendation 2



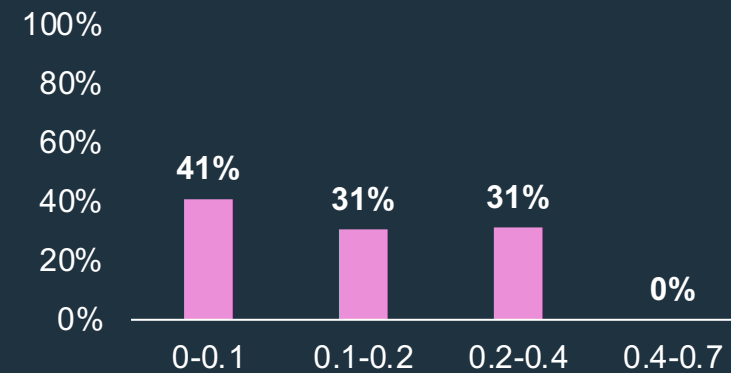
Recommendation 3



Recommendation 4



Recommendation 5





05 Recommendations

Recommendations & Future Work



- Recommendation system would improve customer experience as well as ad-sales. Hence, Craigslist should implement this system.
- Research additional features that can be scraped and would enhance the recommendation system.
- This same model can be customized for revenue generating categories such as job postings, rentals, cars and trucks etc.
- Duplicate listings reduce the quality of the recommendation.
- Model comparison was difficult due to different evaluation metrics for each model.



THANK YOU!

Do you have any questions?