CRAIGSLIST

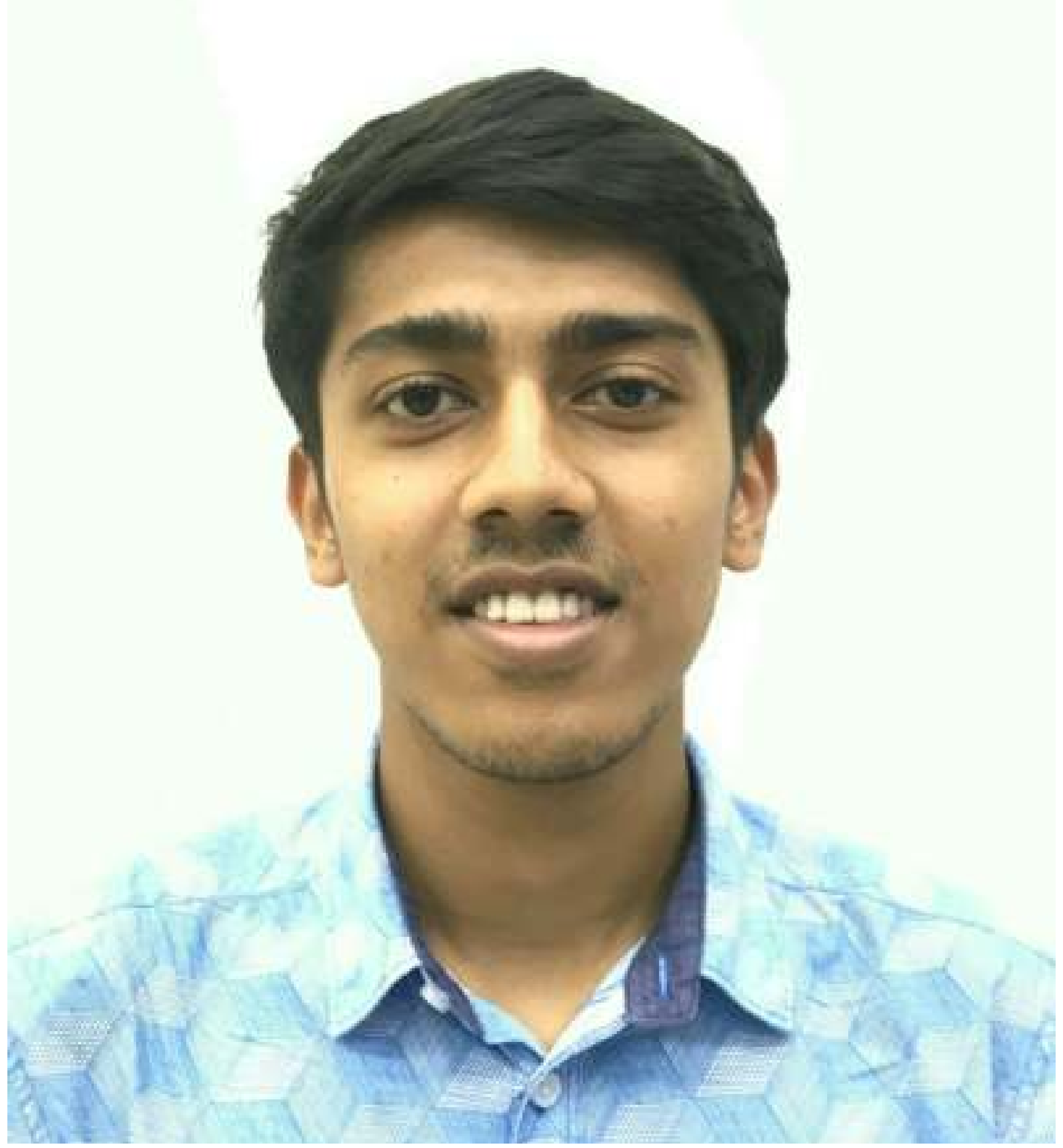
BEAUTY AND HEALTH

Implementing Tag-Based Filtering for enhanced shopping experience

Group members :
Pratik Borkar I Zeeshan Husein Gilani I Monika Madugula I Ashvin Raj I Hanbing Yang I Samridhi Vats

About Us

Pratik Borkar



Zeeshan Husein Gilani

Monika Madugula

About Us

Ashvin Raj

Hanbing Yang

Samridhi Vats

CRAIGSLIST

AN INTRODUCTION

A privately held American company for classified advertisements with sections dedicated to jobs, housing, sales, services, community service, and more.

One of the largest user-generated advertisement websites, operating in 570 cities across 70 countries.

Ebay

Craigslist

Comparative Analysis

Craigslist v/s Ebay

Compared to eBay, the Craigslist website is less appealing and harder to navigate due to its dense and text-heavy layout. This layout can make individual listings harder to notice, especially when compared to eBay's image-focused and categorized listings.

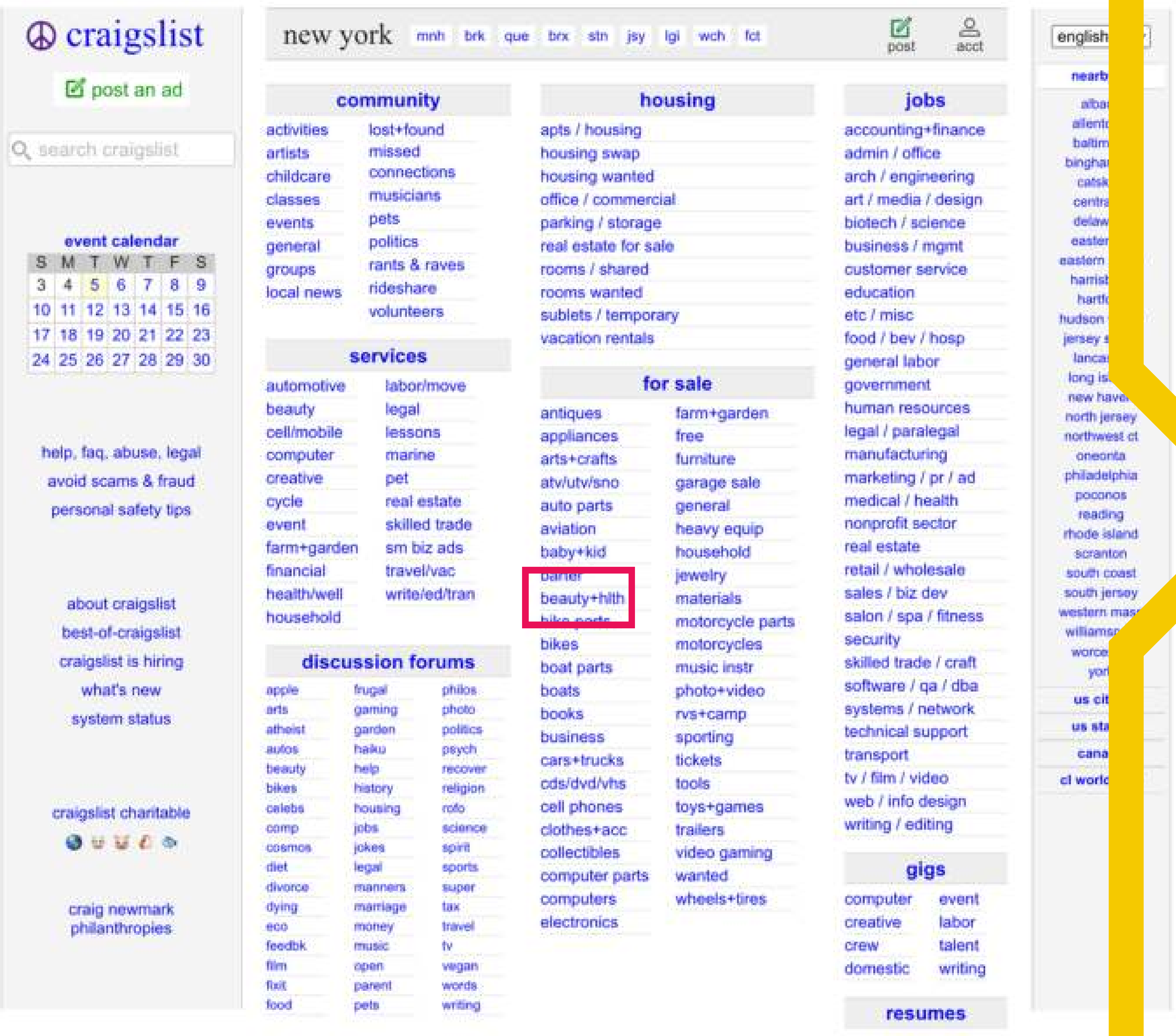
3

Key Issues Identified

- Poor user interface
- Inefficient search capabilities
- Lack of visibility for listing

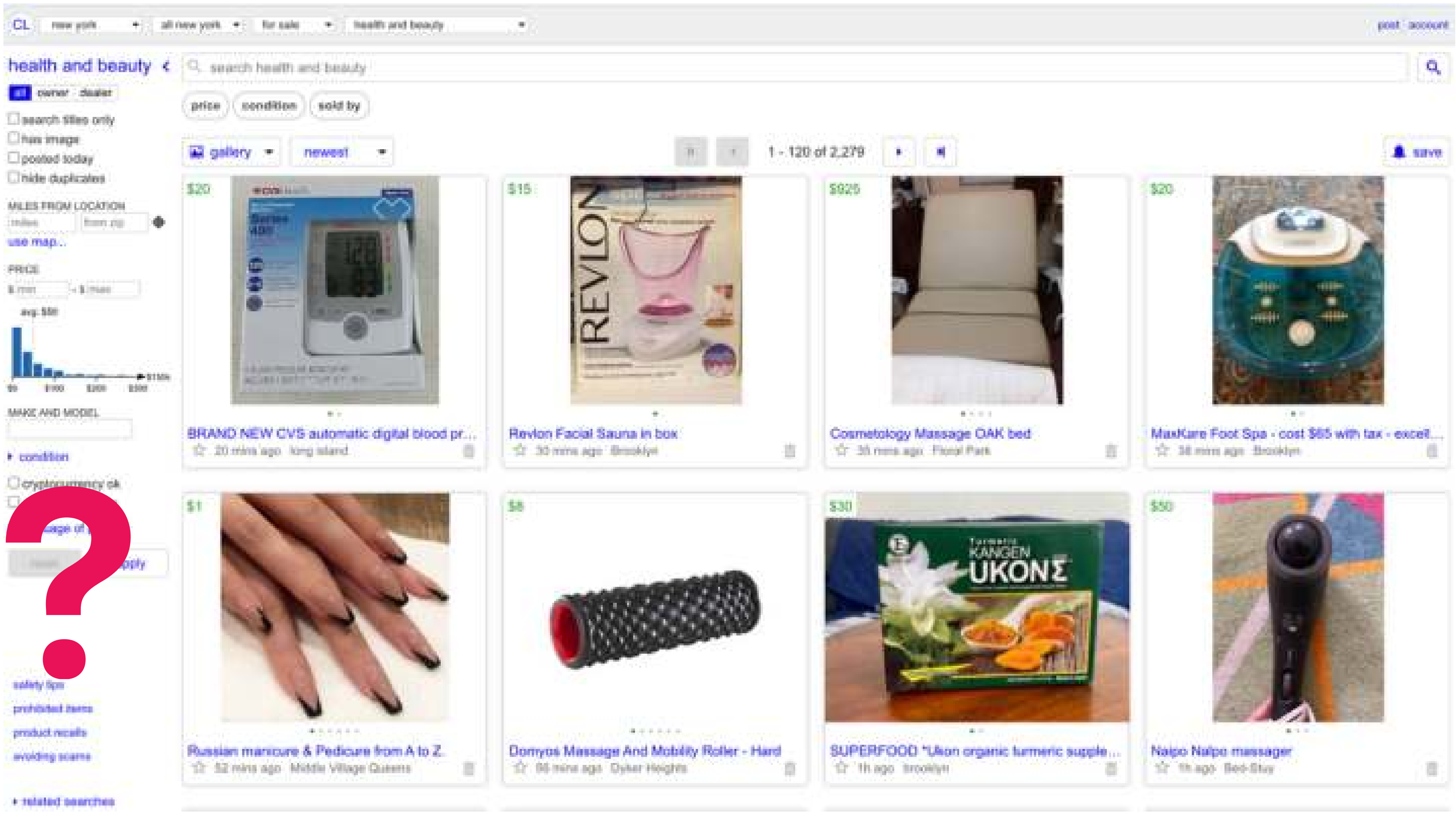
Our Project Objective

To analyze and refine the beauty and health listings on Craigslist for the New York market, utilizing the region's rich data to enhance categorization and improve the user search experience.



Why Beauty and Health?

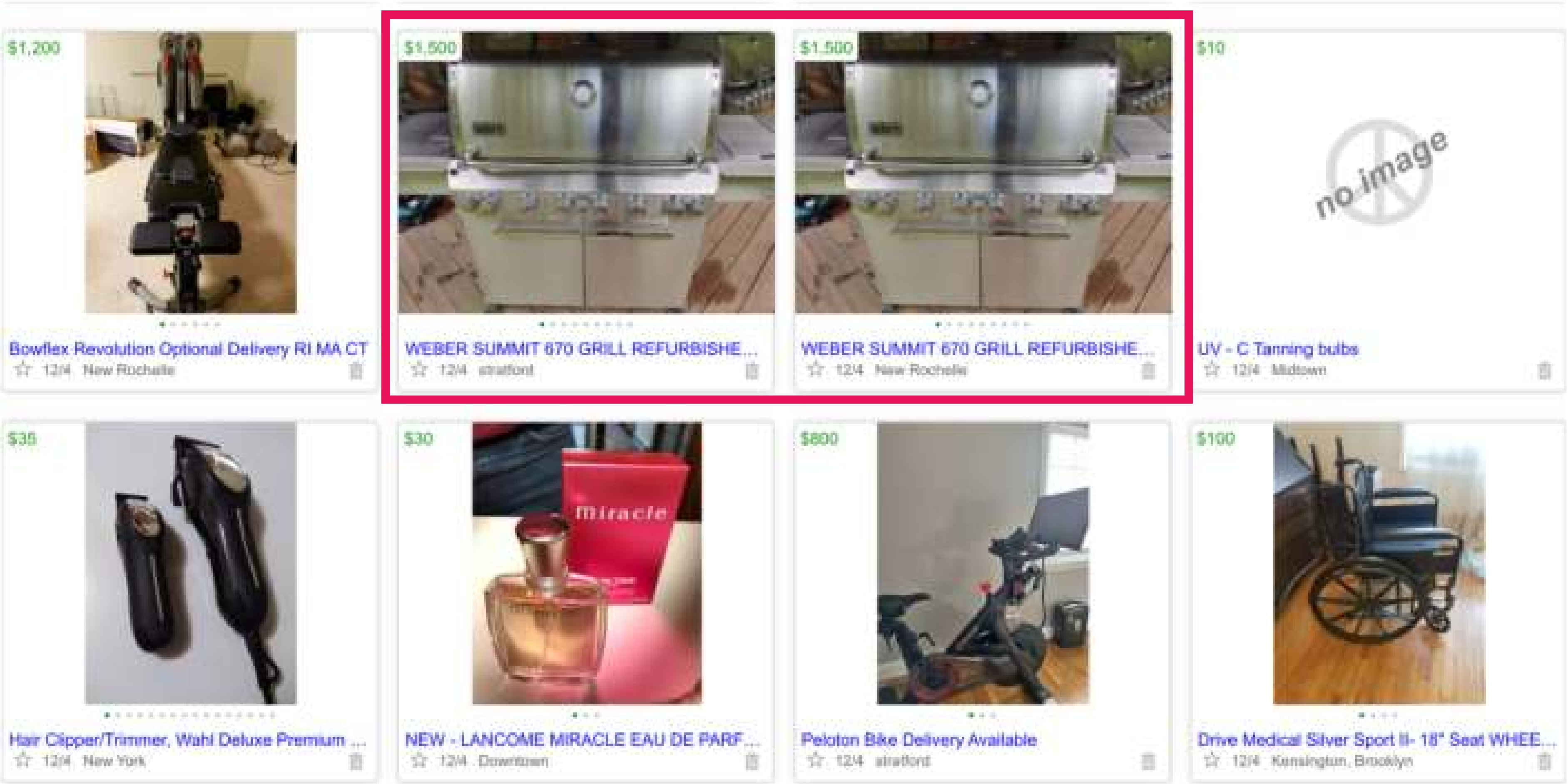
In a category as diverse as beauty and health, where the distinctions between subcategories can be nuanced—from organic skincare products to therapeutic services—filters and tags serve as essential tools for streamlining the search process. However, these are currently missing on Craigslist.



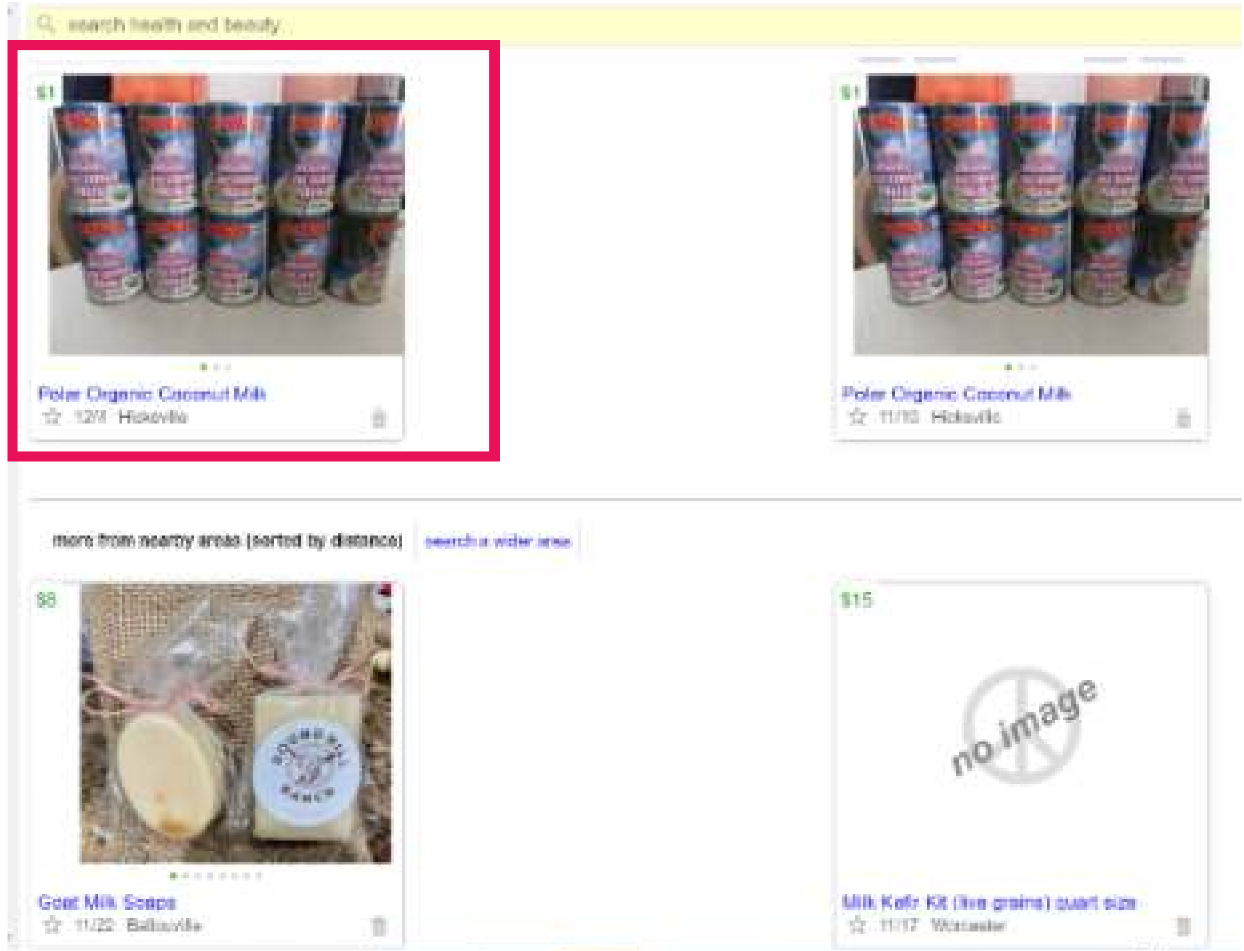
https://newyork.craigslist.org/search/haa#search=1~gallery~0~0

Why Beauty and Health?

The lack of filters and tags on Craigslist leads to a chaotic Beauty and Health section, mixing various items from weighing scale to trimmers and even coconut milk, leading to a poor search experience.

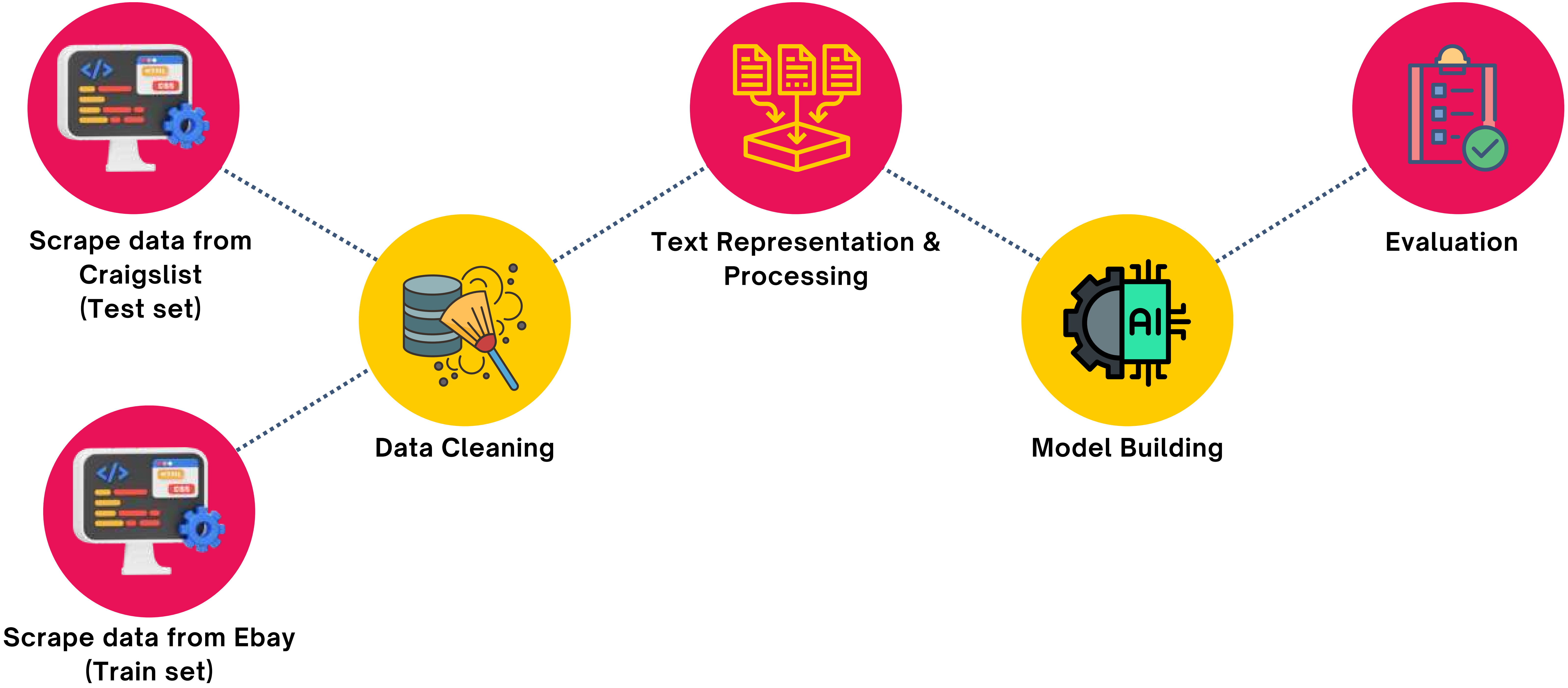


https://newyork.craigslist.org/brx/hab/d/new-york-weber-summit-670-grill/7694277751.html



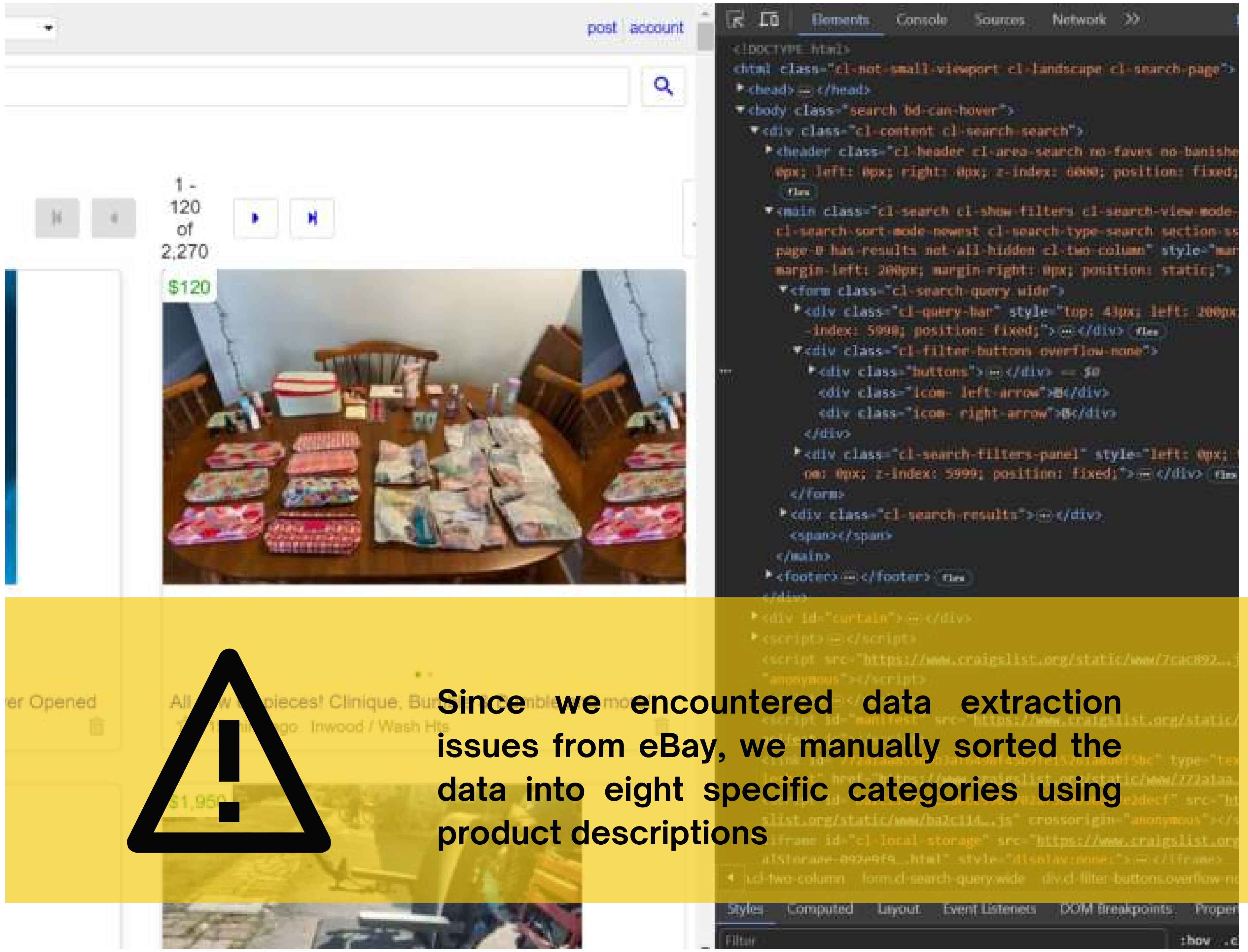
https://newyork.craigslist.org/lgi/hab/d/hicksville-polar-organic-coconut-milk/7686963252.html

Process Flow



Data Collection

- Utilized Selenium and BeautifulSoup to navigate and parse web data
- Retrieved approximately 1680 product URLs across 19 pages for a comprehensive dataset
- Extracted detailed product information, including Product descriptions, Title and IDs
- Compiled and organized all data into CSV format



Since we encountered data extraction issues from eBay, we manually sorted the data into eight specific categories using product descriptions

Web-scraping Code Snapshot

```python
def product_details(url_main):
    title = soup.find('span', {'id': 'titletextonly'})
    if title!= None:
        title_temp=(title.get_text(strip=True)).replace(',','')
        product_title.append(title_temp)
    else:
        product_title.append(0)
    prod_price=soup.find('span',{'class':'price'})
    if prod_price!= None:
        price.append(int(prod_price.get_text(strip=True).split('$')[1].replace(',','')))
    else:
        price.append(0)
    Post_info=soup.find('div',{'class':"postinginfos"})
    if Post_info!=None:
        Post_id=Post_info.find('p',{'class': "postinginfo"})
        if Post_id != None:
            post_id.append(int(Post_id.get_text(strip=True).split('post id: ')[1]))
        else:
            post_id.append(0)
        datetime_str = Post_info.find('time')['datetime']
        if datetime_str!= None:
            parsed_datetime = datetime.strptime(datetime_str, "%Y-%m-%dT%H:%M:%S%z")
            post_date.append(parsed_datetime)
        else:
            post_date.append(0)
    else:
        post_id.append(0)
        post_date.append(0)
    post_body=soup.find('section',{'id':"postingbody"})

    if post_body!= None:
        desc_temp=(post_body.get_text(strip=True).split('Post')[1]).replace(',','')
        description.append(desc_temp)
    else:
        description.append(0)
```

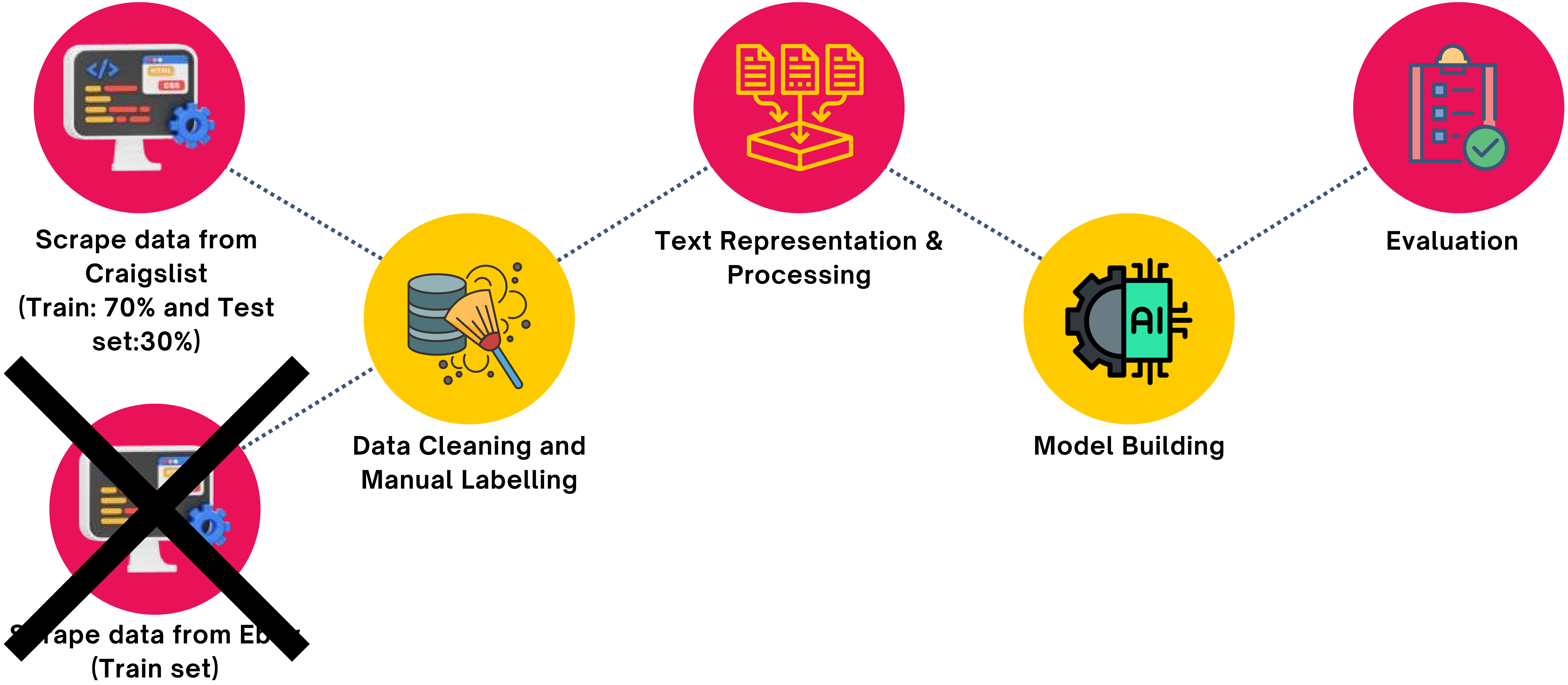
Code for Product details

product_title	Label	Post_ID	Post_date	description	multi_ads	latitude	longitude
PHILIPS Recpironics Breathing Gadget	Health Equipments	7.69E+09	2023-11-13	Like new. Gently used. No problem	1	40.57962	-74.0037
Foam Roller 36" white	Health Equipments	7.69E+09	2023-11-25	The white longer roller left. Great c	0	40.6521	-74.0018
Kolua Wax (large package for hair removal)	Skin Care & Makeup	7.68E+09	2023-11-02	Full. Tried it once. I'd rather go to a	0	40.6521	-74.0018
The Yoga Deck: 50 Poses & Meditations for	Other Health Care	7.69E+09	2023-11-08	The Yoga Deck: 50 Poses & Meditat	0	40.6521	-74.0018
Medical Rolling Portable Folding Adult Mobi	Health Equipments	7.69E+09	2023-11-25	Good conditionComes from a pet f	0	40.62573	-73.9564
Calvin Klein Eyeglasses	Vision Care	7.69E+09	2023-11-15	Used in very good good condition f	0	40.7416	-73.9238
Sit N cycle Exercise Bike	Health Equipments	7.69E+09	2023-11-18	Sit N cycle by Smooth Fitness exerci	1	40.6588	-73.8438
Door Doorway Frame Mount Pull Up Exercis	Health Equipments	7.69E+09	2023-11-25	Iron Gym Proxfit fitness bars for a	0	40.62573	-73.9564
New in Box - L'occitane Verbena EDT - 3.4 o	Fragrances	7.69E+09	2023-11-20	New in box and unused L'occitane	0	40.6424	-73.9758
2 New Adult Girls Womans Blond White Dar	Hair Care	7.69E+09	2023-11-25	$15 each. 2 left - golden and pinkLo	0	40.62575	-73.9564
Curling Iron XTAVA $18	Hair Care	7.69E+09	2023-11-25	Xtava curling wandBox was lost bet	0	40.6816	-73.9798
Root Branch and Blossom	Skin Care & Makeup	7.68E+09	2023-10-29	If purchased individually Body Refr	0	40.7807	-73.7812
Makes Enuresus Alarm	Other Health Care	7.69E+09	2023-11-18	Made in England.	0	40.7807	-73.7812
Conair Double Ceramic 1.5" Flat Iron Electri	Hair Care	7.69E+09	2023-11-25	Used but not abused. Works wellSe	0	40.62573	-73.9564
Covidien Kangaroo Gastrostomy Feeding Tu	Health Equipments	7.69E+09	2023-11-07	REFshow contact infoEFeaturesWit	0	40.7229	-73.8473
Acupuncture	Other Health Care	7.69E+09	2023-11-25	Acupuncture is using a needle to ac	1	40.7443	-73.9781
Cloud Massage Shiatsu Foot Massager	Health Equipments	7.69E+09	2023-11-25	This is a fantastic foot massager. A	0	40.7975	-73.9683
Facial Steamer	Skin Care & Makeup	7.68E+09	2023-11-06	Facial steamer used in spas and it v	1	40.7443	-73.9781
Estee Lauder Skincare Makeup Lot "BRAND	Skin Care & Makeup	7.68E+09	2023-10-29	BRAND NEWSEALED	0	40.6011	-73.9475
Caudalie Resveratrol Lift Anti Wrinkle Firmin	Skin Care & Makeup	7.69E+09	2023-11-08	Brand newNever openedHave a re	0	40.6011	-73.9475
NEW Ray-Ban aviator RB-3625 58mm blue le	Vision Care	7.68E+09	2023-10-29	NEWNEVER USED	0	40.6011	-73.9475
Ray-Ban RB3664CH Chromance Polarized M	Vision Care	7.69E+09	2023-11-21	perfect like new conditionno scratc	0	40.6011	-73.9475
Dior Sauvage Men's Parfum Spray LARGE 20	Fragrances	7.68E+09	2023-10-29	brand newfull bottlewithout box	0	40.6011	-73.9475
Ray-Ban aviator RB-3025 Authentic RARE BL	Vision Care	7.69E+09	2023-11-25	Lenses are in perfect conditionChe	0	40.6011	-73.9475
Ray-Ban aviator RB-3666 Gold 56mm Polari	Vision Care	7.69E+09	2023-11-21	perfect like new conditionno scratc	0	40.6011	-73.9475
Professional Permanent Makeup Machine K	Skin Care & Makeup	7.69E+09	2023-11-14	brand newnever used	0	40.6011	-73.9475
Sonic Electric Toothbrush BRAND NEW SEAL	Health Equipments	7.69E+09	2023-11-10	BRAND NEW SEALED	0	40.6011	-73.9475
Prada PARADOXE Parfum 90ml. Brand New	Fragrances	7.69E+09	2023-11-08	Brand new sealedHave a receipt fr	0	40.6011	-73.9475
Valentino Donna Born in Roma Eau de Parfu	Fragrances	7.69E+09	2023-11-11	Brand new sealedRegular or ENTE	0	40.6011	-73.9475
Digital Upper Arm Blood Pressure Monitor E	Health Equipments	7.69E+09	2023-11-25	Brand newSealed	0	40.6011	-73.9475
BIOSWISS VENTED QUICK DRY BRUSH #9175	Hair Care	7.69E+09	2023-11-23	NEW OLD STOCKPROBABLY PURCH	1	40.6548	-73.6097

products details labelled compl

Scraped Dataset

Revised Process Flow



Labels Identified

01 **Fragrances**

02 **Skin Care & Makeup**

03 **Hair Care**

04 **Health Equipments**

05 **Medications & Supplements**

06 **Other Health Care**

07 **Other Beauty**

08 **Vision Care**

Model Identification - Logistic Regression

```
names_list = []
descriptions_list = []
categories_list = []
column_names = ['Name','Description','Category']
df2 = pd.DataFrame(columns = column_names)
for n,d,c in zip(df['product_title'], df['description'], df['Label']):
    names_list.append(n)
    descriptions_list.append(d)
    categories_list.append(c)
df2['Name'] = names_list
df2['Description'] = descriptions_list
df2['Category'] = categories_list
df2
```

	Name	Description	Category
0	PHILIPS Recpironics Breathing Gadget	Like new. Gently used. No problems at all. Cle...	Health Equipments
1	Foam Roller 36" white	The white longer roller left. Great condition....	Health Equipments
2	Kolua Wax (large package for hair removal)	Full. Tried it once. I'd rather go to a spa bu...	Skin Care & Makeup
3	The Yoga Deck: 50 Poses & Meditations for Body...	The Yoga Deck: 50 Poses & Meditations for Body...	Other Health Care

Name	Description	Health Equipments	Skin Care & Makeup	Other Health Care	Vision Care	Fragrances	Hair Care	Medications & Supplements	Other Beauty	Information
philip recpironics breathe gadget	like new gently use problem clean neat ready use	1	0	0	0	0	0	0	0	philip recpironics breathe gadgetlike new gent...
foam roller white	white long roller leave great condition sunset...	1	0	0	0	0	0	0	0	foam roller whitewhite long roller leave great...
kolua wax large package hair removal	full try id rather go spa work great	0	1	0	0	0	0	0	0	kolua wax large package hair removalfull try i...
yoga deck pose meditation body mind spirit card	yoga deck pose meditation body mind spirit car...	0	0	1	0	0	0	0	0	yoga deck pose meditation body mind spirit car...
medical roll portable fold adult mobility walk...	good conditioncomes pet free smoke free home l...	1	0	0	0	0	0	0	0	medical roll portable fold adult mobility walk...

C	solver	max_iter	penalty	Accuracy	random_state
1	liblinear	10000	l2	0.5565	0
0.05	newton-cg	10000	l2	0.5595	0
1	liblinear	10000	l1	0.3988	0
1	sag	5000	l2	0.5625	0
0.1	lbfgs	5000	l2	0.5595	0
0.05	sag	5000	None	0.5416	0

Model Identification - Logistic Regression

01

Text Preprocessing

Map NLTK POS tags to WordNet tags, clean, tokenize and lemmatize text

02

Word Weighting by TF-IDF

Use TfidfVectorizer to convert text data into TF-IDF features.

03

Model Building

- Set hyperparameters for Logistic Regression
- Use a balanced class weight to handle potential class imbalances.
- Wrap Logistic Regression in OneVsRestClassifier to handle multi-label classification.

04

Model Evaluation

Accuracy on the test data: 0.58

Accuracy: 0.5833333333333334

	precision	recall	f1-score	support
0	0.73	0.84	0.78	135
1	0.80	0.63	0.71	52
2	0.53	0.49	0.51	47
3	0.50	0.17	0.25	6
4	0.77	0.71	0.74	14
5	0.62	0.57	0.59	28
6	0.47	0.33	0.39	27
7	0.56	0.19	0.28	27
micro avg	0.68	0.63	0.65	336
macro avg	0.62	0.49	0.53	336
weighted avg	0.67	0.63	0.63	336
samples avg	0.61	0.63	0.61	336

Model Identification - Pre-trained Model

```python
tokenizer = BertTokenizer.from_pretrained('bert-base-uncased')

def encode_data(tokenizer, texts, labels, max_length=512):
    input_ids = []
    attention_masks = []
    label_list = []

    for text, label in zip(texts, labels):
        encoded_data = tokenizer.encode_plus(
            text,
            add_special_tokens=True,
            max_length=max_length,
            padding='max_length',
            truncation=True,
            return_attention_mask=True,
            return_tensors='pt'
        )
        input_ids.append(encoded_data['input_ids'])
        attention_masks.append(encoded_data['attention_mask'])
        label_list.append(label)

    return torch.cat(input_ids, dim=0), torch.cat(attention_masks, dim=0), torch.tensor(label_list)

X_train_ids, X_train_masks, y_train_tensor = encode_data(tokenizer, X_train.tolist(), y_train_array)
X_test_ids, X_test_masks, y_test_tensor = encode_data(tokenizer, X_test.tolist(), y_test_array)

train_dataset = TensorDataset(X_train_ids, X_train_masks, y_train_tensor)
test_dataset = TensorDataset(X_test_ids, X_test_masks, y_test_tensor)

batch_size = 16
dataset,  batch_size=batch_size)
est_data  taset), bat
n.from_p  els=len(df2
```

Kernel Restarting

The kernel appears to have died. It will restart automatically.

OK

Your session crashed after using all available RAM. If you are interested in access to high-RAM runtimes, you may want to check out Colab Pro.

Error occurred when installing package 'torch'. Details...

Model Identification - LSTM

01

Text Preprocessing

Convert words to lowercase, remove punctuation and symbols, and eliminate stop words

02

Tokenizing and Padding Documents

Convert labels into a one-hot encoded format, suitable for multi-class classification.

03

Model Building

- An embedding layer,
- A spatial dropout layer (to reduce overfitting)
- An LSTM layer
- A dense output layer with a softmax activation function

04

Model Evaluation

Best Model Accuracy: 0.595

```python
# Load GloVe embeddings
embeddings_index = {}
with open('glove.6B.200d.txt', 'r', encoding='utf8') as f:
    for line in f:
        values = line.split()
        word = values[0]
        coefs = np.asarray(values[1:], dtype='float32')
        embeddings_index[word] = coefs

# Create an embedding matrix
embedding_matrix = np.zeros((len(tokenizer.word_index) + 1, 200))
for word, i in tokenizer.word_index.items():
    embedding_vector = embeddings_index.get(word)
    if embedding_vector is not None:
        embedding_matrix[i] = embedding_vector

## Model Architecture
model = Sequential()
model.add(Embedding(len(tokenizer.word_index) + 1, 200, weights=[embedding_matrix], input_length=250, trainable=False))
model.add(SpatialDropout1D(0.2))
model.add(Bidirectional(LSTM(200, dropout=0.1, recurrent_dropout=0.1)))
model.add(Dense(len(labels), activation='softmax'))

model.compile(loss='categorical_crossentropy', optimizer='adam', metrics=['accuracy'])

# Training the model
epochs = 6
batch_size = 64
history = model.fit(X_train, Y_train, epochs=epochs, batch_size=batch_size,
                    validation_split=0.1, callbacks=[EarlyStopping(monitor='val_loss', patience=3, min_delta=0.0001)])
```

Trials: Model Architecture & Embedding Dimension

GloVe Embedding - Tried 50, 100, 200, 300 dimension

Adding additional LSTM Layer

```
## Model Architecture
model = Sequential()
model.add(Embedding(len(tokenizer.word_index) + 1, 200, weights=[embedding_matrix], input_length=250, trainable=False))
model.add(SpatialDropout1D(0.2))
model.add(Bidirectional(LSTM(100, dropout=0.2, recurrent_dropout=0.2, return_sequences=True)))
model.add(Bidirectional(LSTM(100, dropout=0.2, recurrent_dropout=0.2)))
model.add(Dense(len(labels), activation='softmax'))

model.compile(loss='categorical_crossentropy', optimizer='adam', metrics=['accuracy'])
```

Increasing LSTM Layer from 100 to 128, 200

```
## Model Architecture
model = Sequential()
model.add(Embedding(len(tokenizer.word_index) + 1, 200, weights=[embedding_matrix], input_length=250, trainable=False))
model.add(SpatialDropout1D(0.2))
model.add(Bidirectional(LSTM(200, dropout=0.1, recurrent_dropout=0.1)))
model.add(Dense(len(labels), activation='softmax'))

model.compile(loss='categorical_crossentropy', optimizer='adam', metrics=['accuracy'])
```

Increased epochs from 5 to 6

```
# Training the model
epochs = 6
batch_size = 64
```

Best Model: SVM

01

Text Preprocessing

Convert words to lowercase, tokenize text, remove stopwords, lemmatize words

02

Label Encoding

Uses LabelEncoder to convert categorical labels ('Label' column) into numerical format, which is required for training machine learning models

03

TF-IDF Vectorisation

Convert the preprocessed text data into TF-IDF features

04

Model Training

Tried different kernels and identified RBF as the best one for the SVM model

05

Model Evaluation

Accuracy on the test data: 0.66

```
Number: 0, Label: Fragrances
Number: 1, Label: Hair Care
Number: 2, Label: Health Equipments
Number: 3, Label: Medications & Supplements
Number: 5, Label: Other Health Care
Number: 6, Label: Skin Care & Makeup
Number: 7, Label: Vision Care
```

```python
[26] # Retrieve the best parameters from the grid search
best_params = grid_search.best_params_
print("Best Parameters:", best_params)
```

```
Best Parameters: {'C': 100, 'gamma': 0.01}
```

```python
#Create and train the SVM model with the best parameters
tuned_svm_model = SVC(kernel='rbf', C=best_params['C'], gamma=best_params['gamma'])
tuned_svm_model.fit(X_train, y_train)

# Predict and evaluate the model
y_pred = tuned_svm_model.predict(X_test)
print("Accuracy:", accuracy_score(y_test, y_pred))
print("\nClassification Report:\n", classification_report(y_test, y_pred))
```

```
Accuracy: 0.6666666666666666

Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.67      0.80        18
           1       0.81      0.69      0.74        51
           2       0.72      0.90      0.80       229
           3       0.50      0.38      0.43        37
           4       0.48      0.24      0.32        41
           5       0.43      0.29      0.35        69
           6       0.56      0.70      0.62        50
           7       0.71      0.56      0.63         9

    accuracy                           0.67       504
   macro avg       0.65      0.55      0.59       504
weighted avg       0.65      0.67      0.65       504
```

Kernels Used: SVM

```python
# Create and train the SVM model with a polynomial kernel
svm_poly_model = SVC(kernel='poly')  # Degree can be tuned
svm_poly_model.fit(X_train, y_train)

# Predict and evaluate the model
y_pred_poly = svm_poly_model.predict(X_test)
print("Polynomial Kernel Accuracy:", accuracy_score(y_test, y_pred_poly))
print("\nPolynomial Kernel Classification Report:\n", classification_report(y_test, y_pred_poly))
```

```
Polynomial Kernel Accuracy: 0.48214285714285715

Polynomial Kernel Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.06      0.11        18
           1       1.00      0.04      0.08        51
           2       0.47      0.99      0.63       229
           3       0.33      0.03      0.05        37
           4       1.00      0.02      0.05        41
           5       0.75      0.04      0.08        69
           6       1.00      0.16      0.28        50
           7       1.00      0.11      0.20         9

    accuracy                           0.48       504
   macro avg       0.82      0.18      0.18       504
weighted avg       0.67      0.48      0.35       504
```

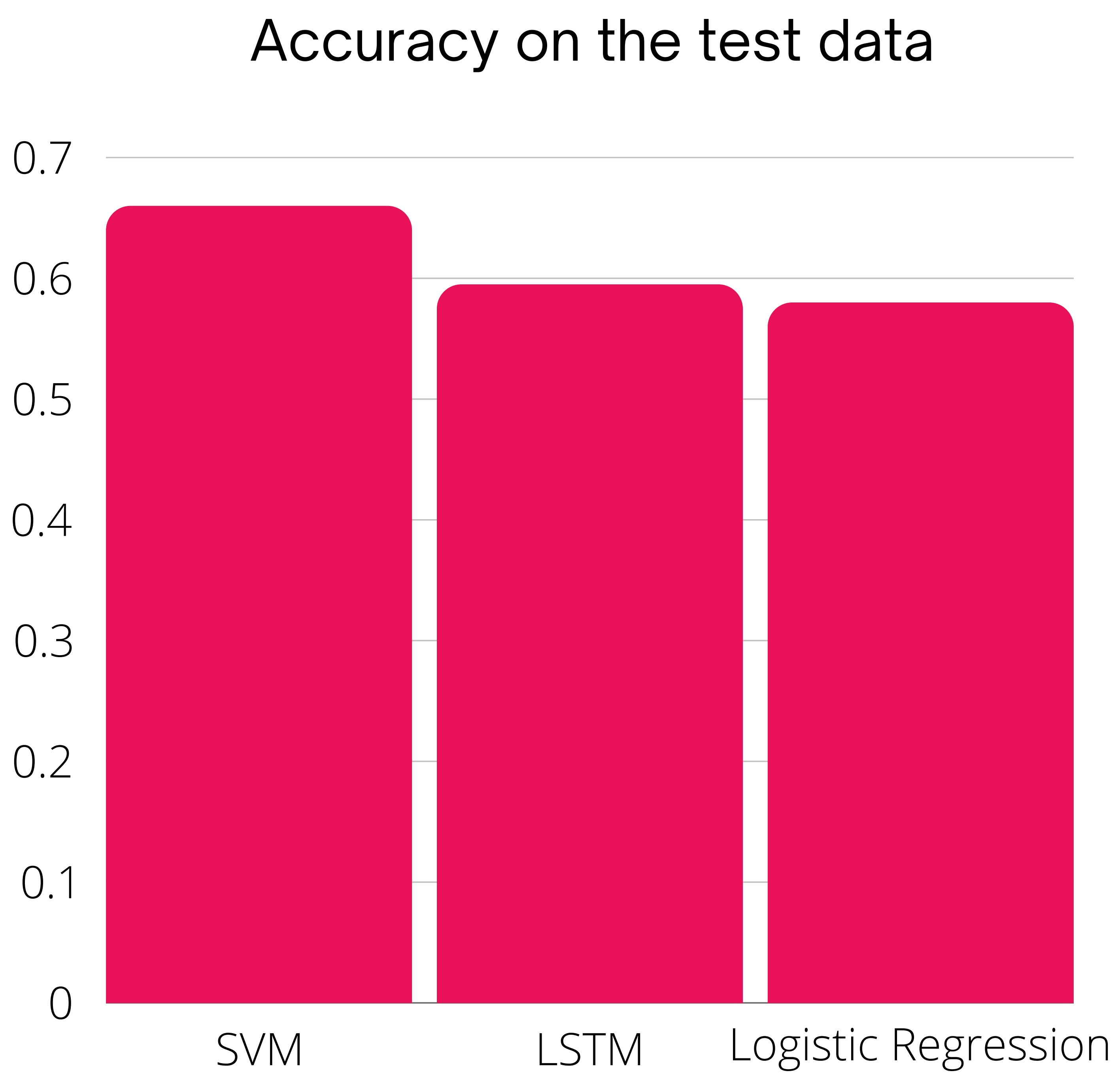
Polynomial kernel

```python
# Create and train the SVM model with a sigmoid kernel
svm_sigmoid_model = SVC(kernel='sigmoid')  # Hyperparameters can be tuned
svm_sigmoid_model.fit(X_train, y_train)

# Predict and evaluate the model
y_pred_sigmoid = svm_sigmoid_model.predict(X_test)
print("Sigmoid Kernel Accuracy:", accuracy_score(y_test, y_pred_sigmoid))
print("\nSigmoid Kernel Classification Report:\n", classification_report(y_test, y_pred_sigmoid))
```

```
Sigmoid Kernel Accuracy: 0.6567460317460317

Sigmoid Kernel Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.67      0.80        18
           1       0.86      0.59      0.70        51
           2       0.63      0.94      0.75       229
           3       0.78      0.38      0.51        37
           4       0.45      0.12      0.19        41
           5       0.61      0.25      0.35        69
           6       0.62      0.68      0.65        50
           7       1.00      0.44      0.62         9

    accuracy                           0.66       504
   macro avg       0.74      0.51      0.57       504
weighted avg       0.67      0.66      0.62       504
```

Sigmoid kernel

Best Model selection



Why is SVM our best model?

- Handling High-Dimensional Data
- Small data size
- Versatile

Demo Implementation: SVM

Evaluating Performance on 5 Random Products from the Test Set

```
Product Index: 1075
Original Text: DRIVE POWER CHAIR ..CIRRUS E C DRIVE POWER CHAIRFOLDS300 LB CAPACITYMINT CONDITIONNEW BATTERIES..COMES WITY CHARGERANTI TIPPING WHEELS
True Label: Health Equipments
Predicted Label: Health Equipments

Product Index: 1526
Original Text: first lady perfume First lady fragrance
True Label: Fragrances
Predicted Label: Fragrances

Product Index: 1377
Original Text: HoMedics BB-2K Bubble Bliss Deluxe Luxury Foot Bubbler with Heat HoMedics BB-2K Bubble Bliss Deluxe Luxury Foot Bubbler with Heathttps
True Label: Health Equipments
Predicted Label: Health Equipments

Product Index: 1632
Original Text: API POND SIMPLY CLEAR Pond Water Clarifier 16-Ounce Bottle (248B) •      Contains one (1) API POND SIMPLY CLEAR Pond Water Clarifier 1
True Label: Other Health Care
Predicted Label: Other Health Care

Product Index: 1204
Original Text: VINTAGE WOOD ROLLER FOOT MASSAGER VINTAGE APRIL BATH AND SHOWER (A B AND S) ROLLER MASSAGER5.5 X 4.5 X 1 3/4"
True Label: Health Equipments
Predicted Label: Health Equipments
```

SUMMARY

- Targeted Beauty and Health section on Craigslist due to disorganization
- Aimed to develop a model to tag products by title and description
- Utilized web crawling to gather and preprocess data for training
- Trained LSTM, SVM, and logistic regression models
- Attained 66% accuracy on test data for the best model

CONCLUSION

- **Navigation Ease**: Our new category filter can help users quickly find products of interest by navigating through a structured hierarchy rather than sifting through an unorganized list.
- **Search Efficiency**: Users can use category filters to narrow down search results, making the shopping experience more efficient.
- **Enhanced product visibility**: The integration of product filters not only enhances overall product visibility but also provides customers with a user-friendly tool for discovering new items. This makes it easier for bargain hunters to find and explore good deals.

Thank You