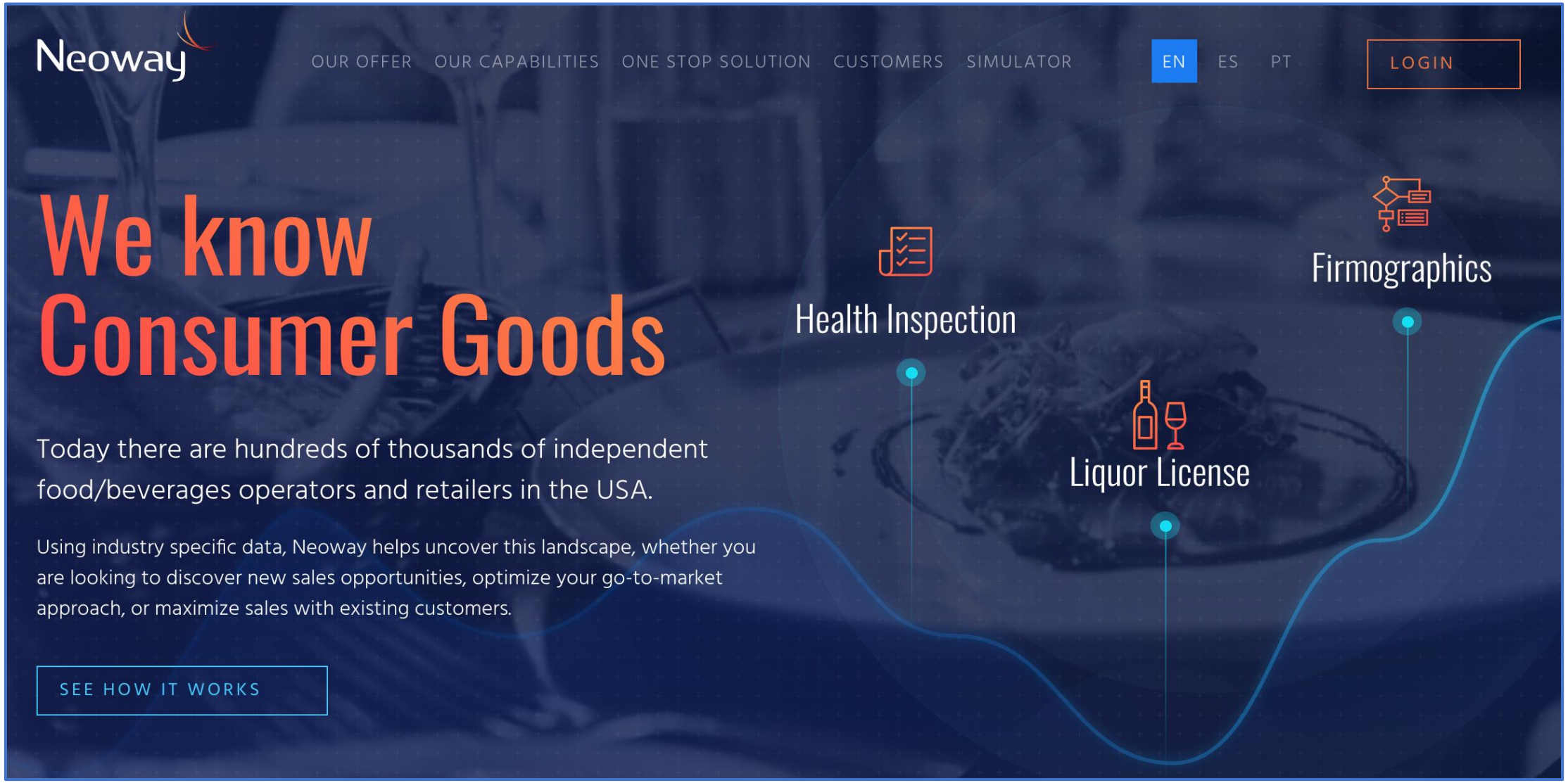
Classifying Food and Beverage Establishments from Website Data

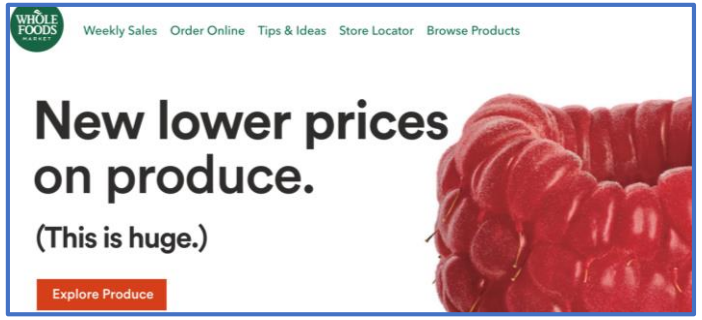
Gaurav Chawla, Jason Kuo, Nanshan Li and Andres Potapczynski

COLUMBIA UNIVERSITY
Data Science Institute

Neoway delivers customer insights based on firmographic data publicly available on the web



They want to uncover certain food/beverages categories from establishment webpages



Webpage

Restaurant

Type of Cuisine

(American, Mexican, Chinese, etc.)

Supermarket / Grocery

Organic / Non-Organic

Bar

Predominant type of alcohol served

(ex: wine, cocktail)

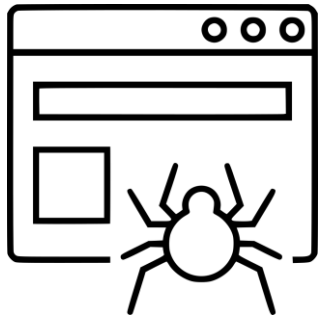
Liquor Store

Predominant type of alcohol sold

(ex: wine, cocktail)

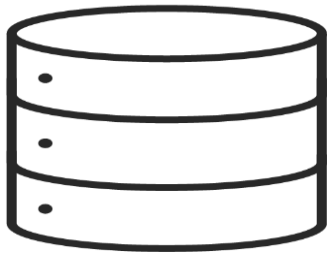
For this, they have gathered close to 1M HTMLs from ~200K establishments

Capturing the HTMLs for different establishments



Gathered the HTML contents via a web crawling bot

Storing the HTMLs



Stored the HTMLs indexed by the initial URL in the cloud

Adding metadata (for a subset)

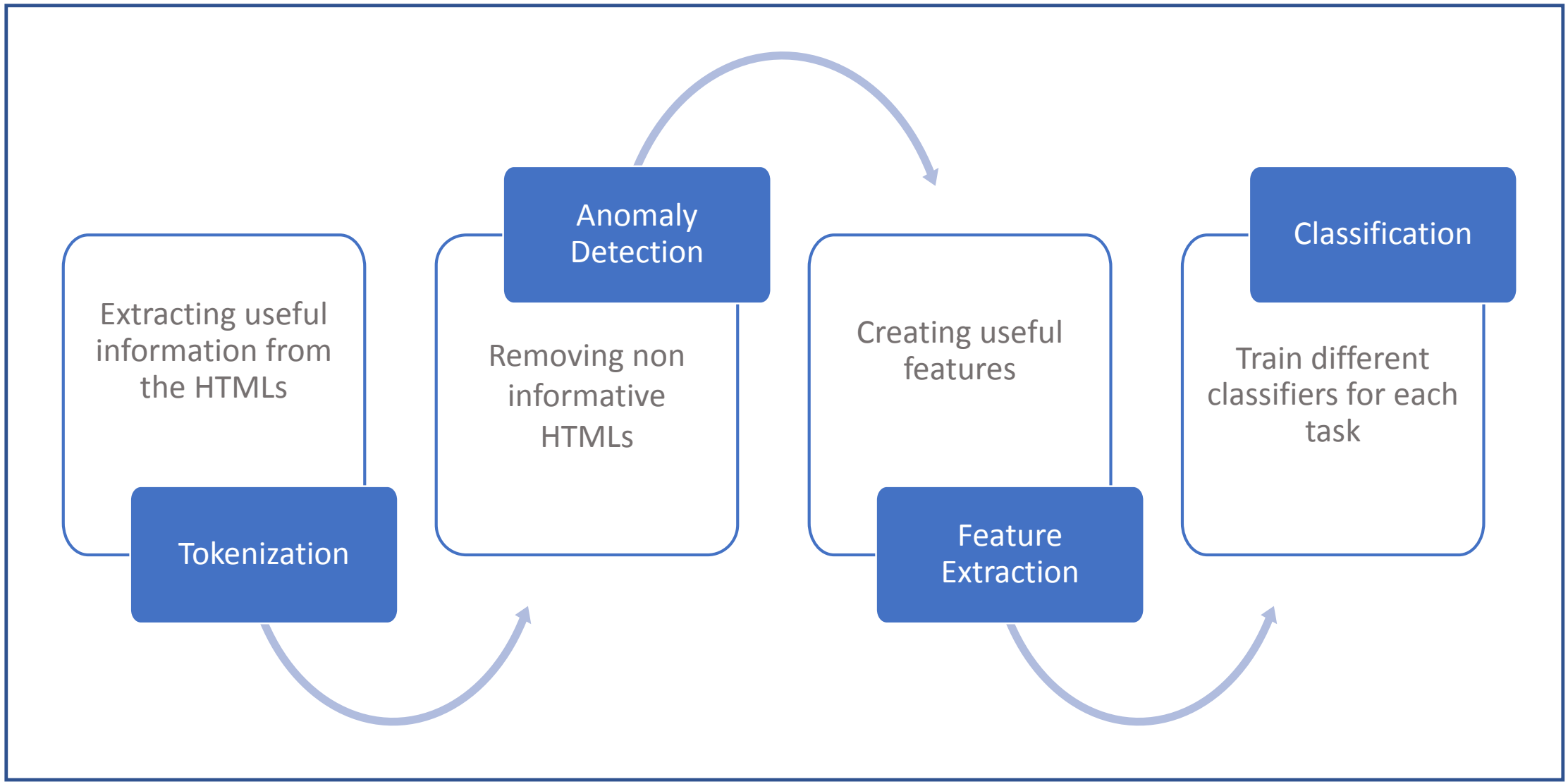
Augmented the database with information from Google and others

Running Ad-hoc methods

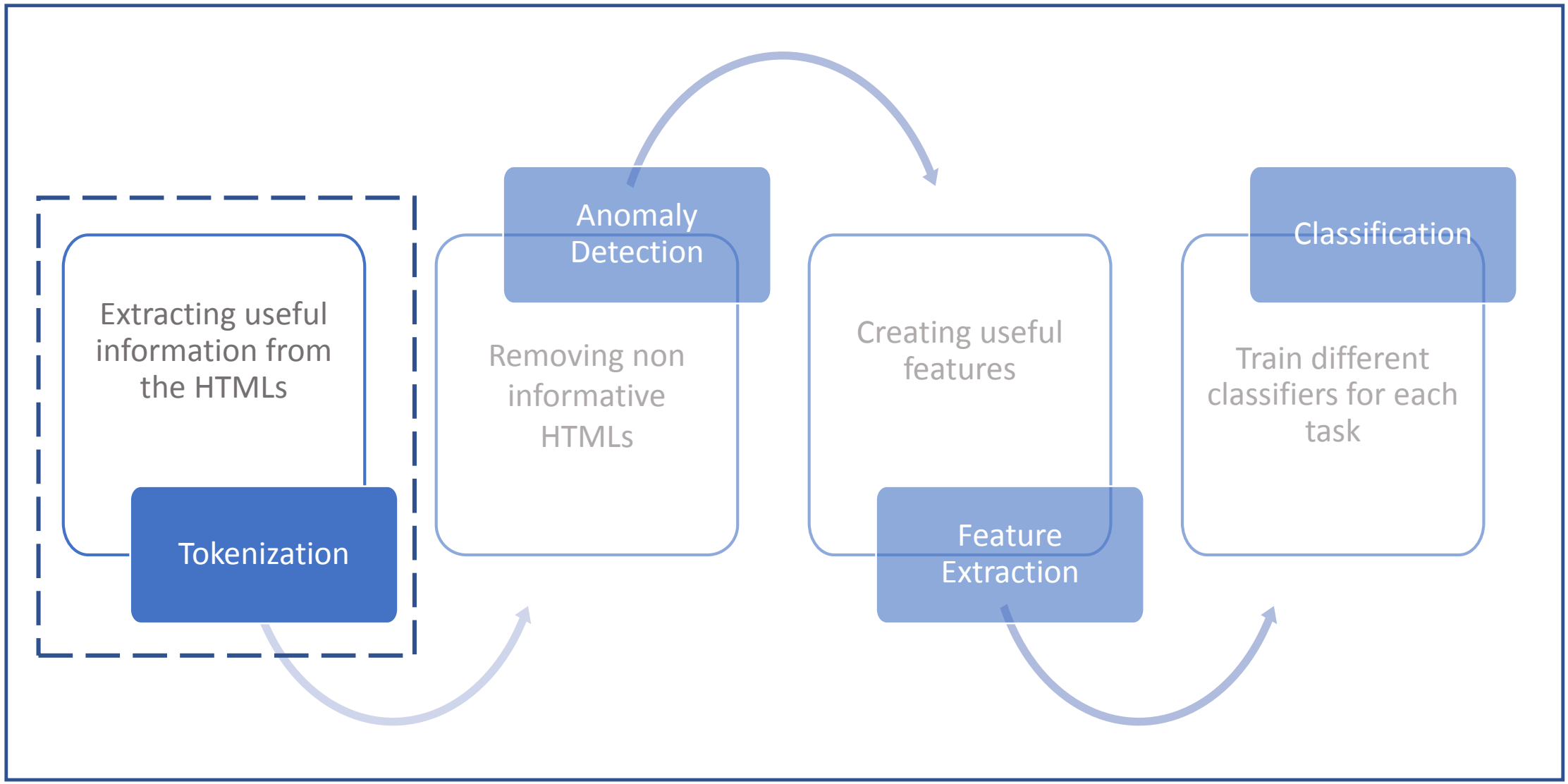
.[RegEx]*

Labelling using keywords via RegEx

We designed a pipeline to help them improve the previous process



We designed a pipeline to help them improve the previous process



Store the tokenized data in a single DB to accelerate the model prototyping and shareability

Original HTML

b'b\'<!DOCTYPE HTML>\\n<html lang="en">\\n<head>\\n<meta charset="UTF-8">\\n<title>Windy City Grille</title>\\n\\n<meta name="norton-safeweb-site-verification" content="gyww51iscbqhtbjem97962ns68l-po3i2sx7fld-g-vazvflhxysybdz0px8lzrz10oinhmrm2zlipwmiw5ybtlcw3te5nim-cbq5pag7cmmt0at4wv56llujdcudrji" />\\n<meta name="keywords" content="Restaurant, Wyoming MI, Wyoming, Michigan, 49519, Byron Center Ave SW, Karadchy, Chicago Style, Chicago, Gyro, Sandwiches, Grille, feta cheese, Vienna, food challenge, wall of fame, wall of shame, eating contest, man vs. food, catering, cater, caterer in wyoming michigan, caterer in wyoming mi">\\n<meta name="description" content="The Windy City Grille is a family owned and operated establishment priding ourselves in delivering authentic food of the highest quality at a great price.">\\n<meta name="OWNER" content="Will Karadchy">\\n<meta name="ALIAS" content="Windy City Grille">\\n<meta name="AUTHOR" content="bluevortex.net">\\n<meta HTTP-EQUIV="Pragma" CONTENT="cache">\\n'

Tokenized HTML

Windy, city, grille, home, gift, certificate, coupon, restaurant, overview, letter, history, menu, main, dish, side, challenge, photo, meal, social, medium, facebook, twitter, foursquare, contact, application, windy, city, grille, privacy, policy, gyro, menu, strip, cooked, lamb, beef, tomato, onion, cucumber, sauce, pita, fold, chicken, gyro, menu, grilled, chicken, tomato, onion, cucumber

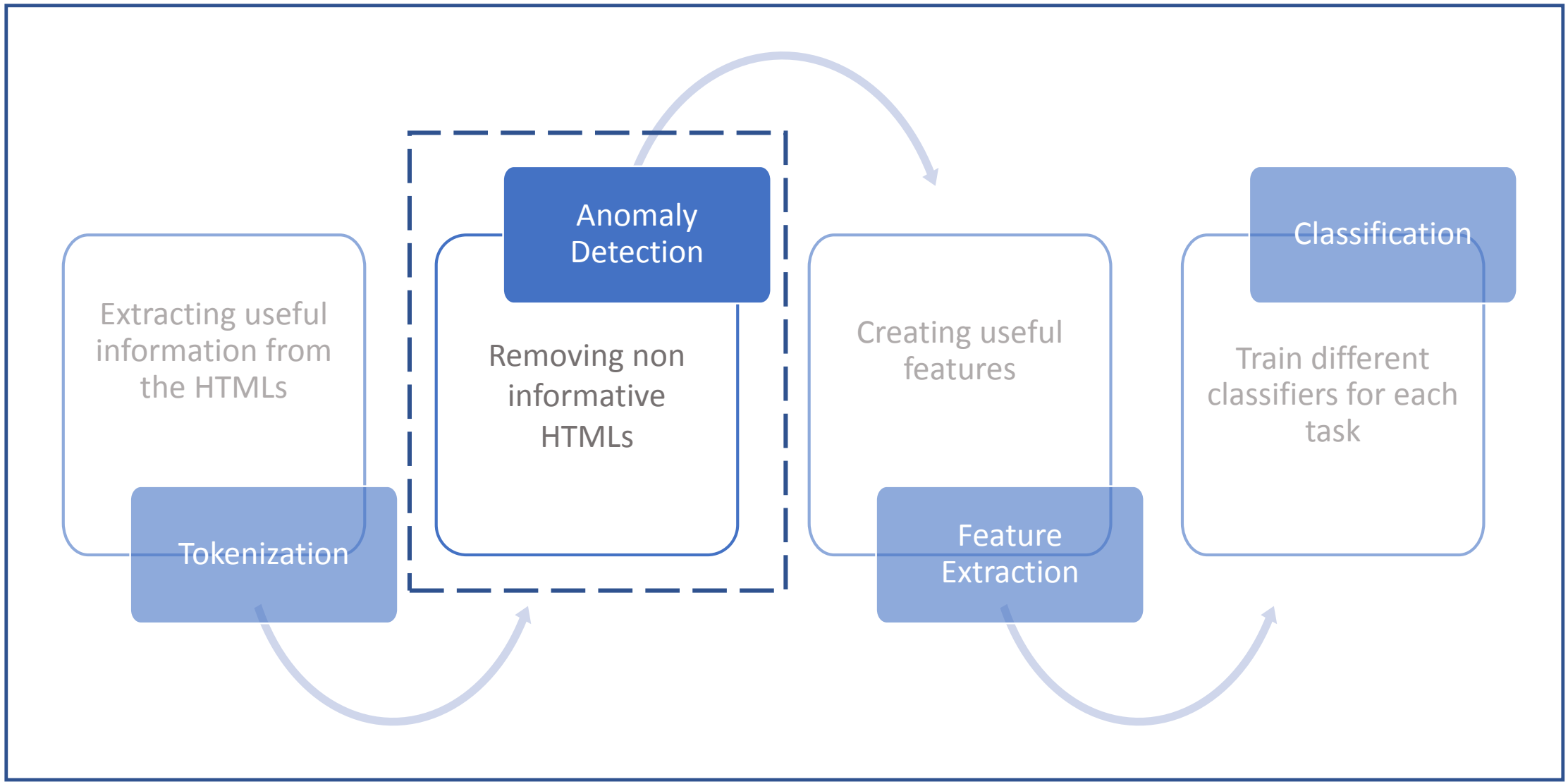
Approach

- Extracted HTML body via XML parser
- Employed NLTK package for lemmatizing, handling punctuation and capitalization, etc
- POS tagging removed sentence filler words
- Created a parallelizable implementation for speed

Recommendations / Next Steps

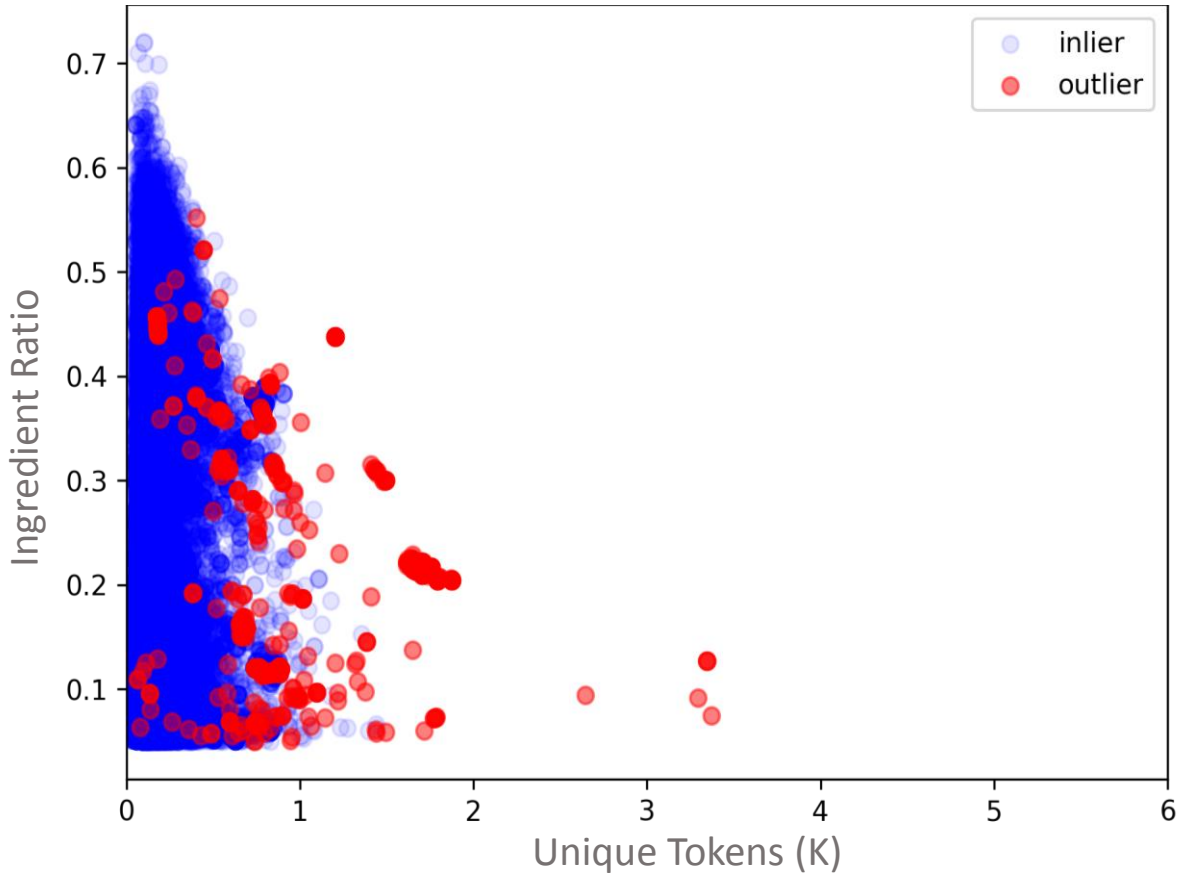
- Generate a consolidated DB that contains all the tokens for each establishment
  - Use a NoSQL DB such as SQLite Dict or Mongo DB

We designed a pipeline to help them improve the previous process



Separate and only use the informative HTMLs to reduce computational overhead

Isolation Forests[1] gives an anomaly score to each observation



Approach

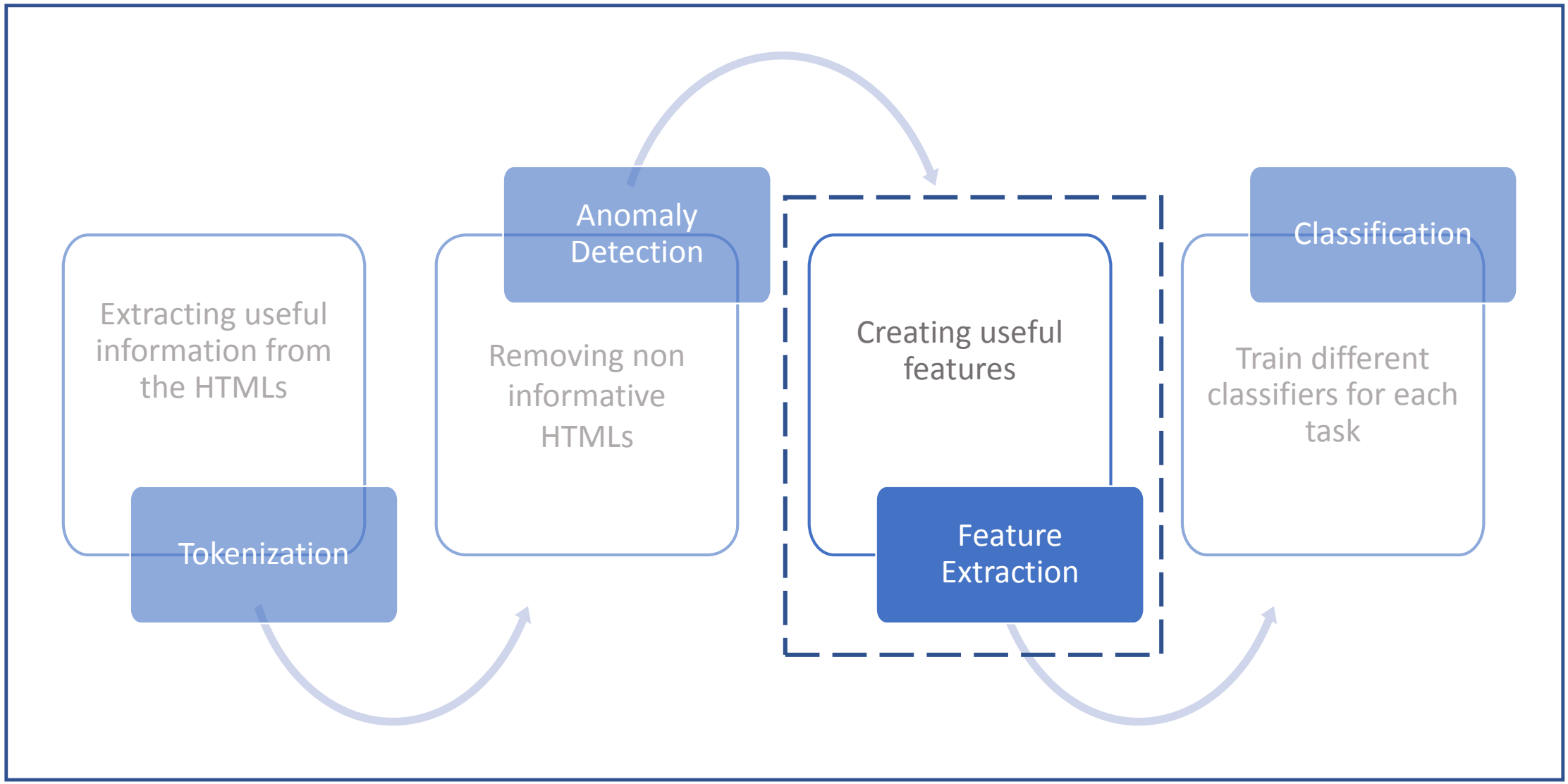
- Construct a series of meta-features (length, # of unique tokens, etc.)
- Define weak thresholds on each feature
- Use Isolation Forests to determine the outliers

Recommendations / Next Steps

- Find the relationship of meta-feature via classification
- Understand the procedure's impact on accuracy

1 Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008, December). Isolation forest. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on* (pp. 413–422). IEEE.

We designed a pipeline to help them improve the previous process



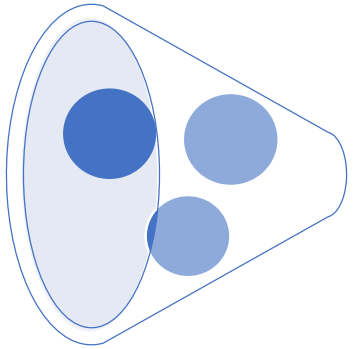
Use Topic Modeling and Doc2Vec to get a dense feature representation

Feature construction process

Words

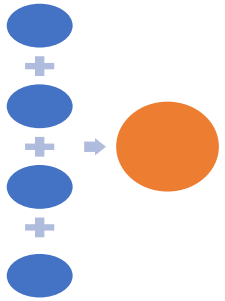
Topics

TF-IDF Filtering



- Reduces the vocabulary from 16K to ~ 2K
- Decreases training complexity downstream

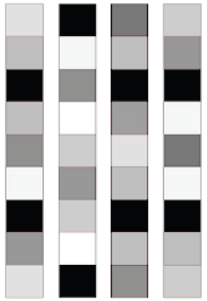
Topic Modeling



- Uncovers the general topics present in the data

Doc2Vec

- Generates a semantic dense representation



75x training time reduction

Understand the topics present in the data and create features that preserve its semantics

LDA[1] and NMF[2] helped uncover topics in the data

LDA Topic Examples with top words

- Topic 44: pizza, cheese, chicken, sauce, mozzarella, tomato, onion, italian, garlic
- Topic 46: margarita, tacos, taco, mexican, menu, location, specials, happy hour
- *Topic 71: twitter, facebook, instagram, google, email, skip, press, online, menu, location*
- Topic 76: chicken, rice, shrimp, sauce, beef, pork, fried, vegetable, onion, spicy

---

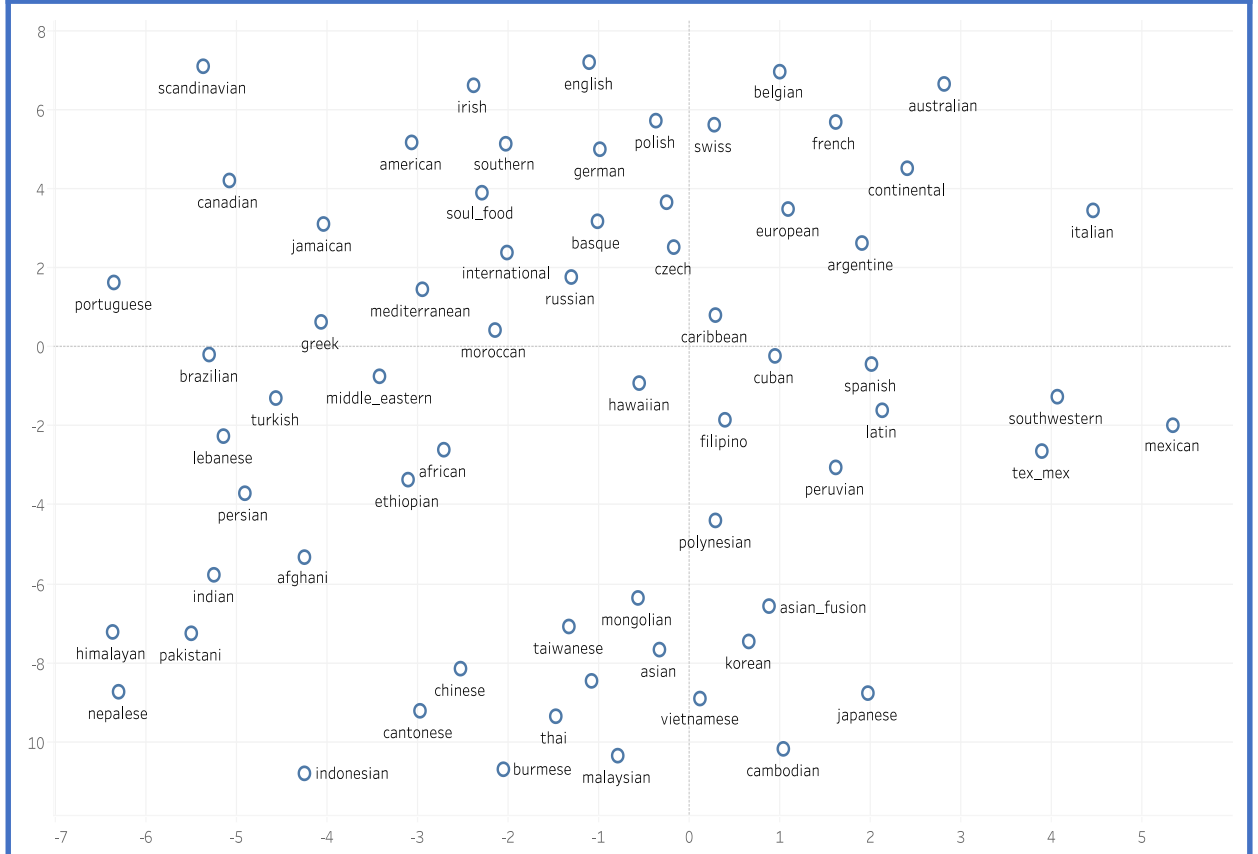
NMF Topic Examples with top words

- Topic 1: cheese, bacon, onion, tomato, lettuce, salad, cheddar, chicken, choice, potato
- Topic 4: pizza, slice, topping, crust, pepperoni, cheese, large, pasta, order, phone
- Topic 6: mexican, authentic, margarita, family, salsa, tacos, good, recipe, nachos
- *Topic 83: good, greate place, time, friendly staff, family, atomoshpere, town, friend*

Recommendations / Next Steps

- Augment establishment categorization based on the topics uncovered

Doc2Vec[3] correctly captured the semantics of the data

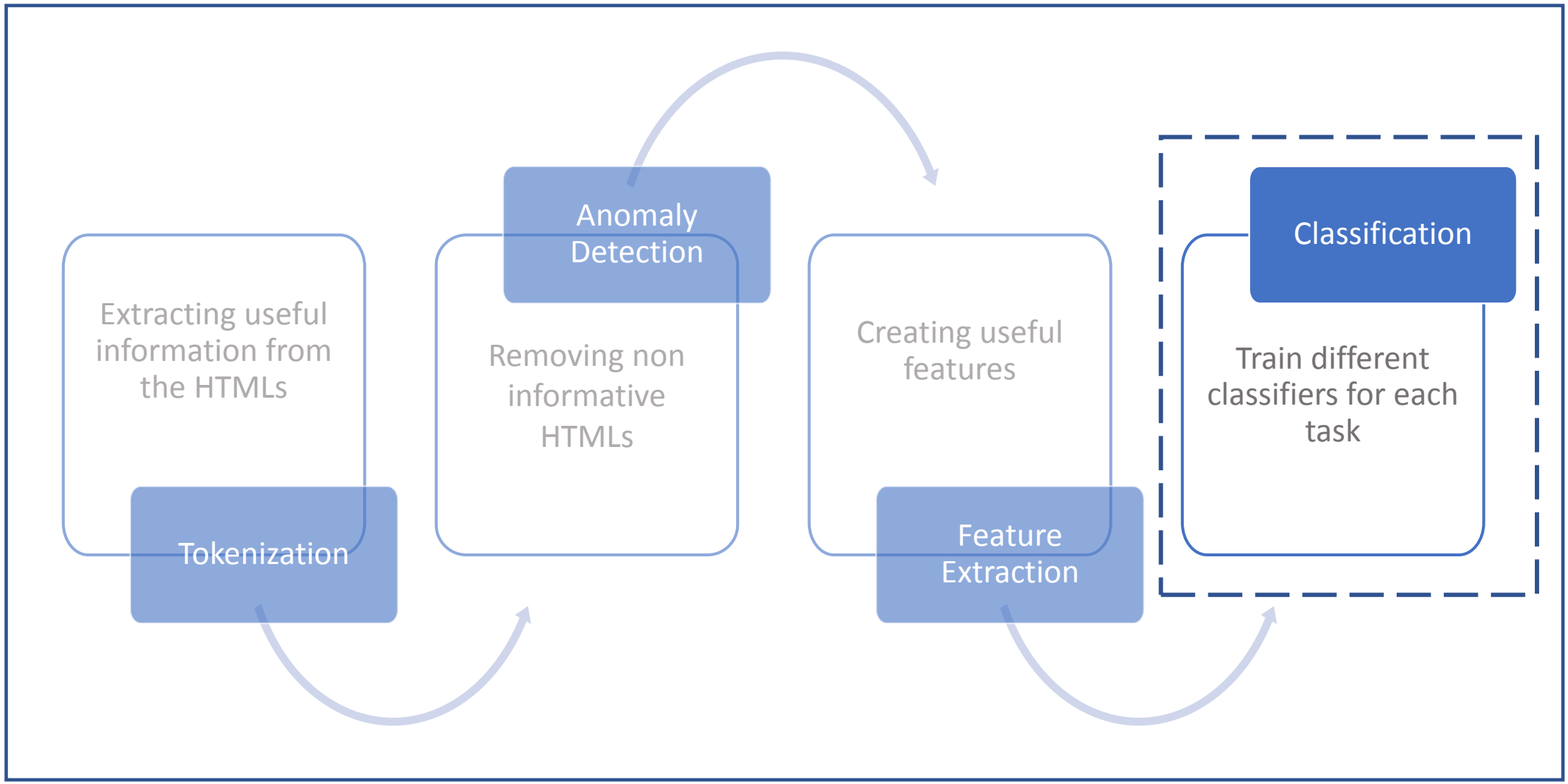


Recommendations / Next Steps

- Use a smaller and dense set of features to save time
- Try state-of-the-art word embedings like ELMo or Bert

1 Blei, et. al. Latent Dirichlet Allocation, 2 Dhillon I. Generalized Nonnegative Matrix Approximations, 3 Mikolov T. and Le Q. Distributed Representations of Sentences and Documents.

Now, we designed a pipeline to help them improve the previous process



Overall Results and models used for each classification task

Establishment Type
Classification

- Multi-class
- Naive Bayes
- F1 score: 90%

- Problem
- Model
- Performance

Restaurant

Cuisine Classification

- Multi-label Classification
- Binary Relevance with Logistic Regression (one-hot)
- F1 score: 70%
- Jaccard Similarity: 66%
- MicroAvg - Pre: 86%
- MicroAvg - Re: 60%

Supermarket / Grocery

Organic vs Not organic Classification

- Binary Classification
- NN
  - (64ReLu-128ReLu-64ReLu-2Sigmoid
- Accuracy: 92%
- Precision: 43%
- Recall: 75%

Bar

Type of Alcohol Classification

- Categorical Regression
- Ridge Regression (per category)
- RMSE Cocktail: 0.07
- RMSE Wine: 0.02

Liquor Store

Type of Alcohol Classification

- Categorical Regression
- Ridge Regression (per category)

Overall Results and models used for each classification task

Establishment Type
Classification

- Multi-class
- Naive Bayes
- F1 score: 90%

- Problem
- Model
- Performance

Restaurant

Cuisine Classification

- Multi-label Classification
- Binary Relevance with Logistic Regression (one-hot)
- F1 score: 70%
- Jaccard Similarity: 66%
- MicroAvg - Pre: 86%
- MicroAvg - Re: 60%

Supermarket / Grocery

Organic vs Not organic Classification

- Binary Classification
- NN
  - (64ReLu-128ReLu-64ReLu-2Sigmoid
- Accuracy: 92%
- Precision: 43%
- Recall: 75%

Bar

Type of Alcohol Classification

- Categorical Regression
- Ridge Regression (per category)
- RMSE Cocktail: 0.07
- RMSE Wine: 0.02

Liquor Store

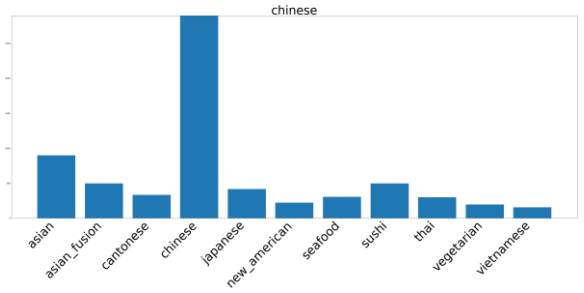
Type of Alcohol Classification

- Categorical Regression
- Ridge Regression (per category)

Cuisine classification carried considerable difficulties

Task: Multi-label Classification

- 99 non-mutually exclusive cuisines
- Implied hierarchy not present
- Multi-label output required
- Highly imbalanced data

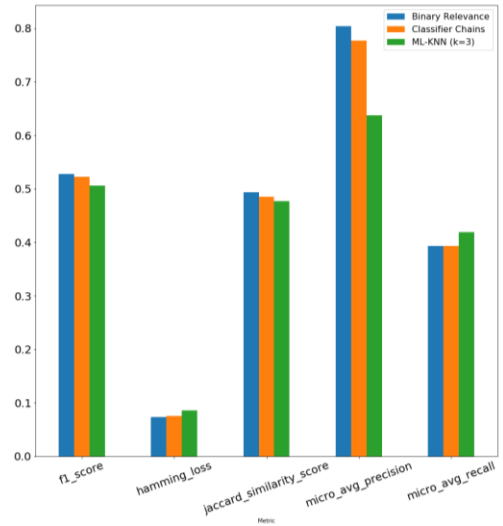


Metrics

- Hamming loss
- Jaccard Similarity Score
- Micro-average Precision & Recall
- F1 Score

Three main approaches:

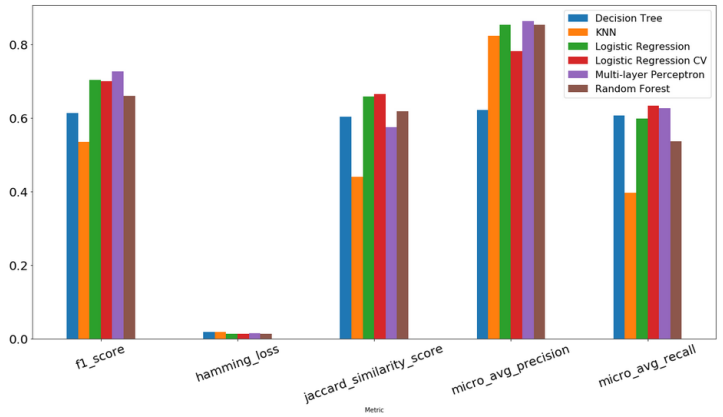
- Binary Relevance
  - Assumes independence between cuisines
- Classification Chains
  - Output is added to input of the next classifier: n! permutations
- Multi-label KNN
  - Lazy KNN method



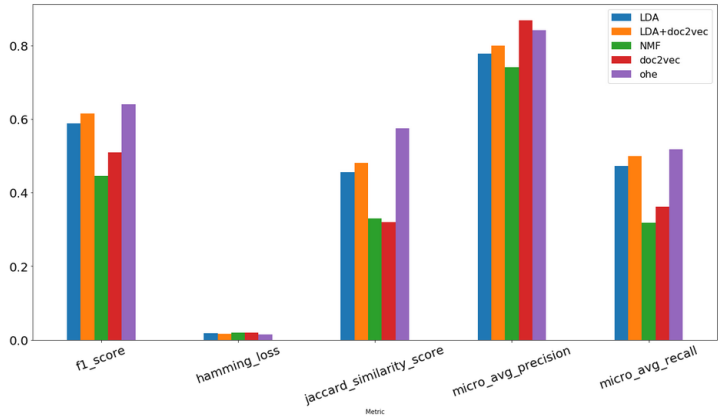
Cuisine classification carried considerable difficulties

Results on 66 Regional Cuisines

- Binary Relevance approach:
  - Decision Tree
  - KNN
  - Logistic Regression
  - Multi-layer Perceptron
  - Random Forest



- Comparing results from different feature representations on Logistic Regression BR:
  - LDA
  - LDA + Doc2Vec
  - NMF
  - Doc2Vec
  - One hot encoding



Recommendations

- Clean up existing cuisine labels
- Develop potential hierarchy and orthogonal label framework:
  - Regional cuisines
  - Food
  - Restaurant format
  - Dietary restrictions
- Framework / spatial understanding of cuisines can be used to improve approach taken in this task

Overall Results and models used for each classification task

Establishment Type
Classification

- Multi-class
- Naive Bayes
- F1 score: 90%

- Problem
- Model
- Performance

Restaurant

Cuisine Classification

- Multi-label Classification
- Binary Relevance with Logistic Regression (one-hot)
- F1 score: 70%
- Jaccard Similarity: 66%
- MicroAvg - Pre: 86%
- MicroAvg - Re: 60%

Supermarket / Grocery

Organic vs Not organic Classification

- Binary Classification
- NN
  - (64ReLu-128ReLu-64ReLu-2Sigmoid
- Accuracy: 92%
- Precision: 43%
- Recall: 75%

Bar

Type of Alcohol Classification

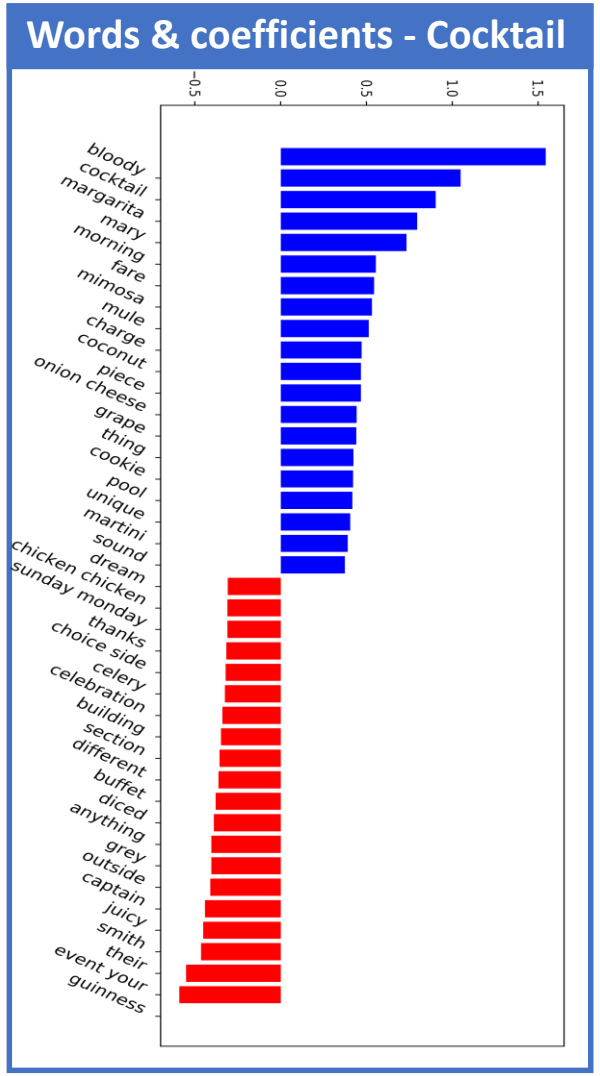
- Categorical Regression
- Ridge Regression (per category)
- RMSE Cocktail: 0.07
- RMSE Wine: 0.02

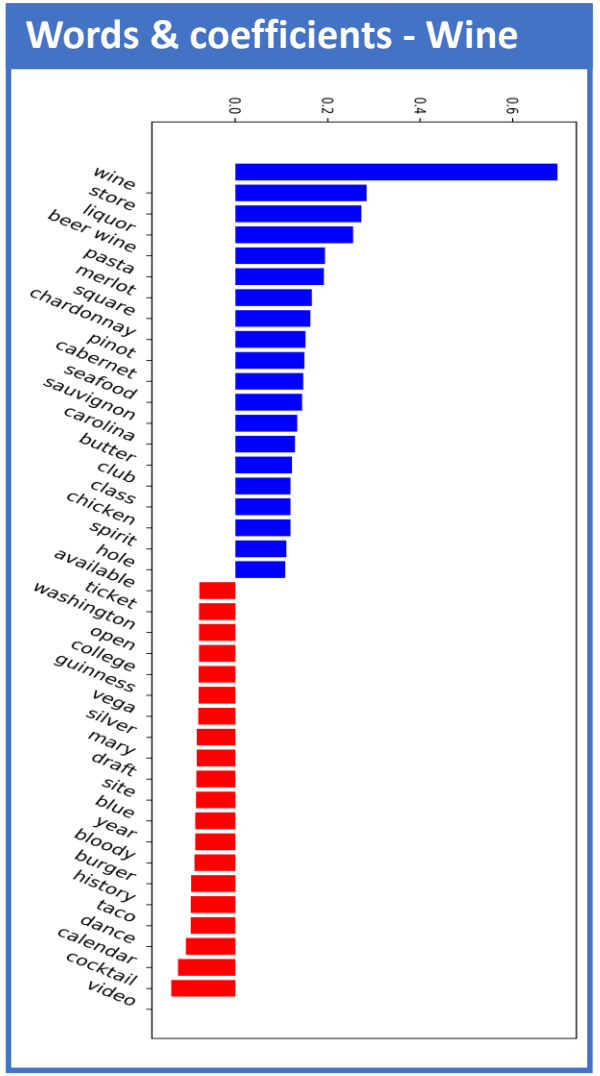
Liquor Store

Type of Alcohol Classification

- Categorical Regression
- Ridge Regression (per category)

The Ridge Regression coefficients correctly capture the relevant words for each class





Approach

- Run a regression model for each type of drink
- Test different regularization schemes like Lasso, Ridge and Elastic Net

Recommendations

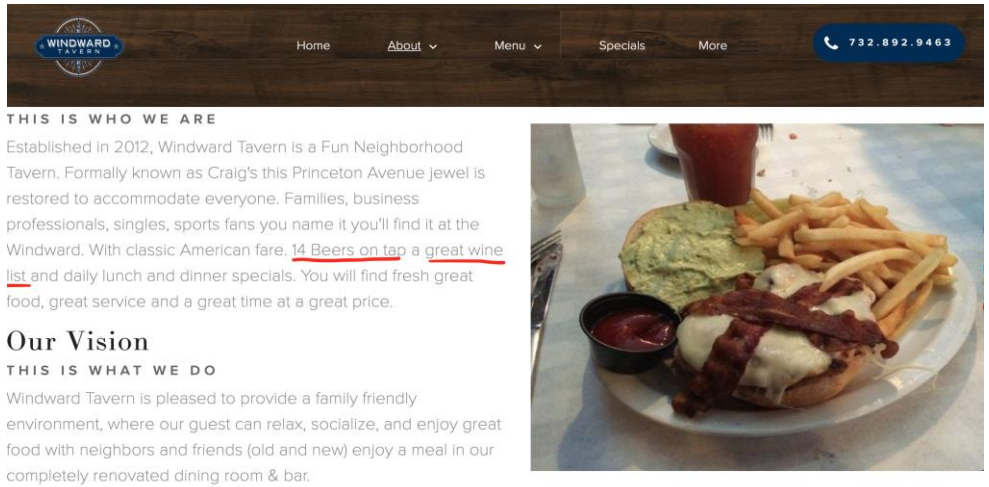
- Use Gridsearch CV to determine the amount of regularization
- Try non-linear models that output a class score (like NN with sigmoid)

Our models handle cases where the labels are incorrect

Cecilia's – Cocktail Bar

Given	Prediction
▪ (Cocktail, Wine) ▪ (0.3 , 0.1)	▪ (Cocktail, Wine) ▪ (0.5 , 0.1)

Winward Tavern



Given	Prediction
▪ (Cocktail, Wine) ▪ (0.6 , 0.2)	▪ (Cocktail, Wine) ▪ (0.2 , 0.2)

Our pipeline has addressed all the previous limitations but still has elements to improve

Ad-hoc limitations

- Heavily dependent on domain knowledge
- Not adaptable to new labels
- RegEx language hard to debug and error-prone

Pipeline results

- **Derives domain knowledge** based on labeled data
- **Scalable** to all types of problems
- **Robust** to mistakes and noise

Next Steps

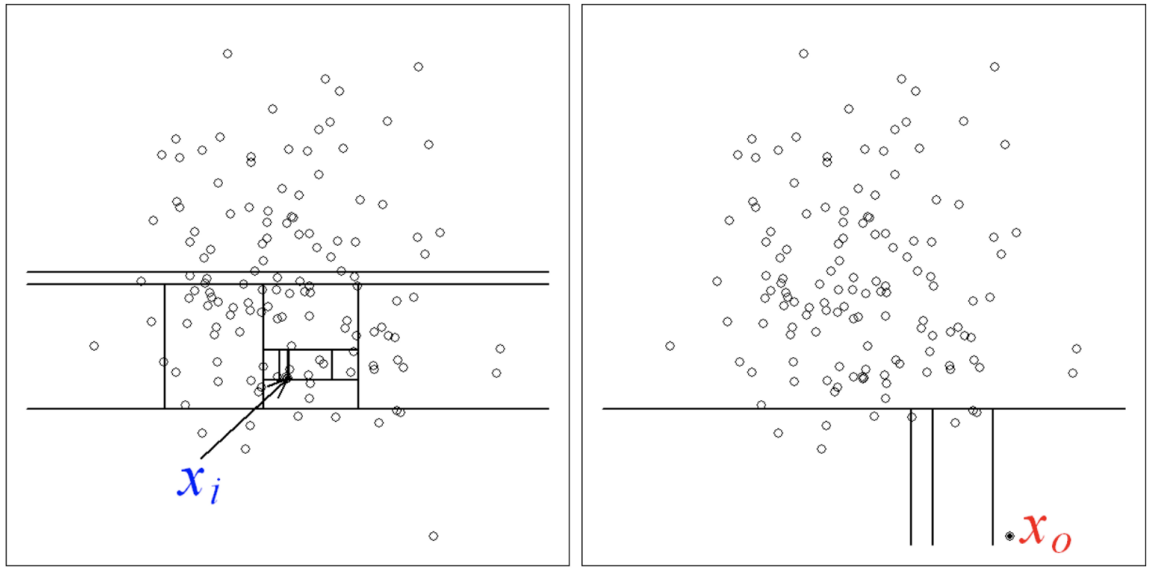
- Run ensemble algorithms like **CatBoost**
- Add **weightings** to deal with the minority class
- Create a **small test set** to validate performance while avoiding noisy labels

Q & A

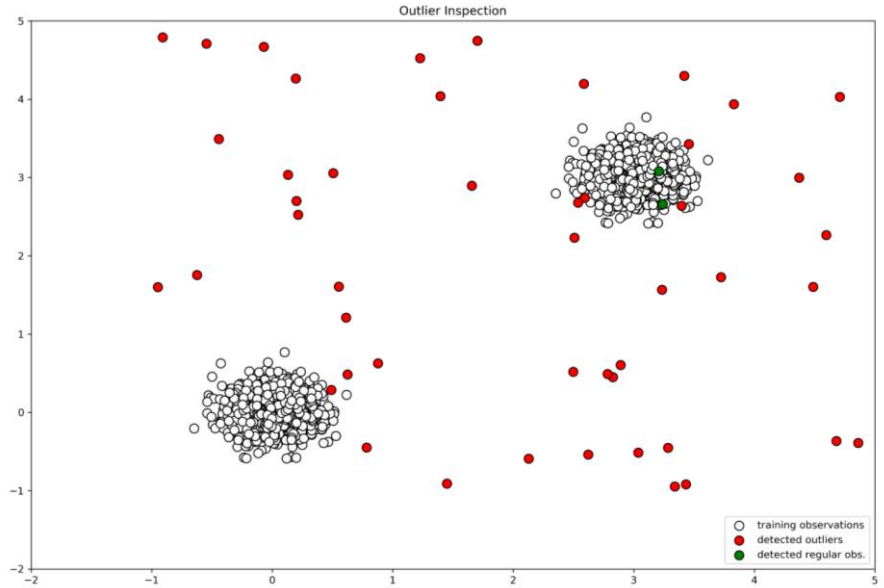
Back-up

Isolation Forest assigns an anomaly score for each observation

At each partition, it selects at random a feature and split

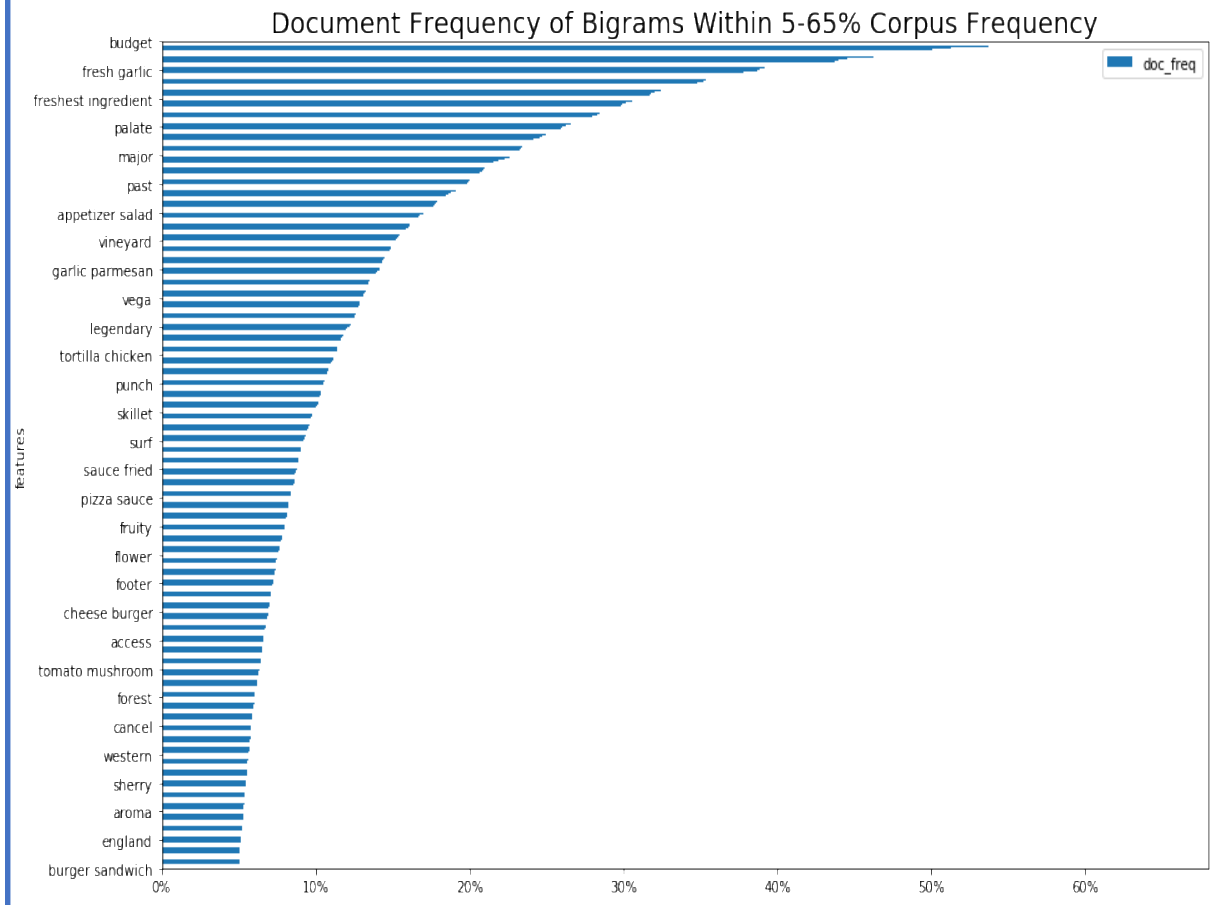


Results of Isolation Forest in two clusters



Use TF-IDF to construct a manageable vocabulary

As expected, the words exhibit a long-tailed distribution



TF-IDF Discussion

- Adjusting the document frequency bounds
  - Setting an upper bound removes uninformative common stop words such as food, order or menu
  - Setting a lower bound removes rare words such as foreign language terms that may overfit model
- Incorporate bigrams in the feature list
  - Captures additional semantics such as pizza sauce or cheese burger that may be lost just looking at single word counts

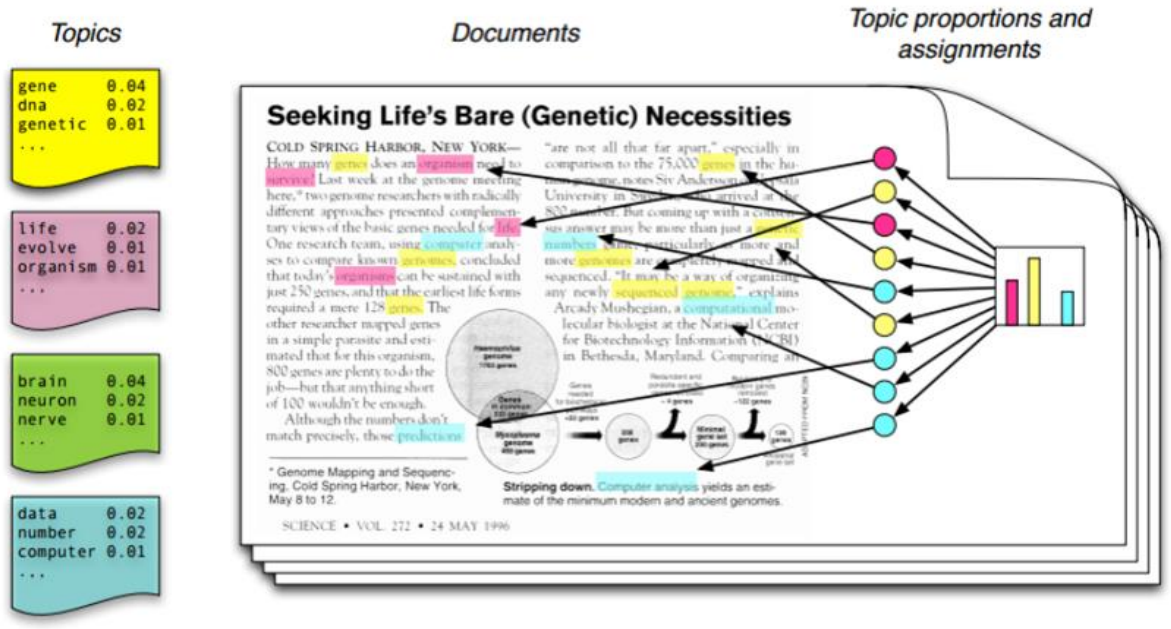
---

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

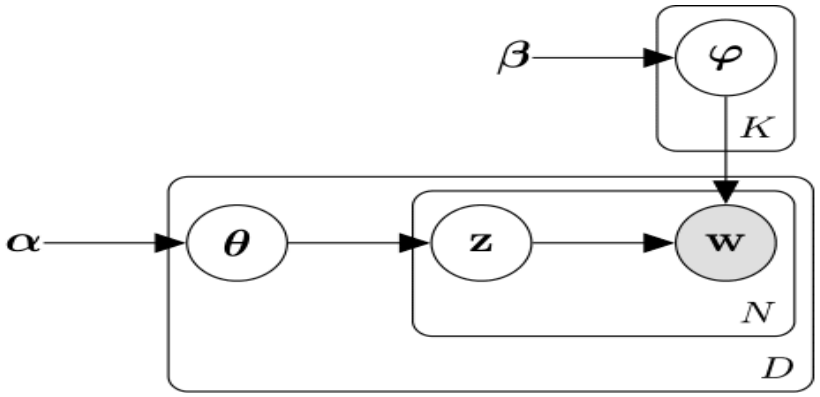
tf_{ij} = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

LDA uncovers the topics present in the data

Example of LDA in a given corpus

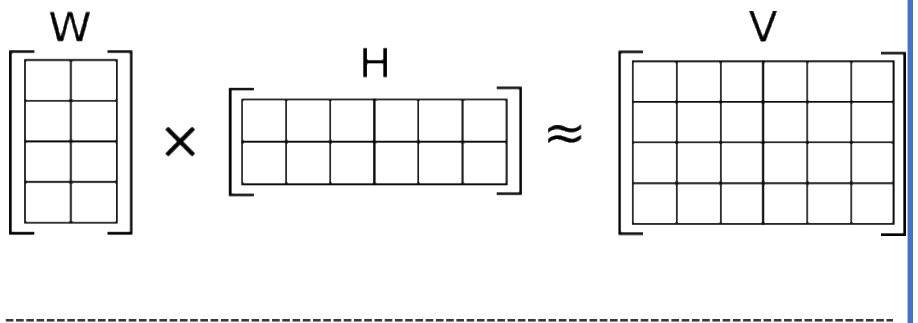


LDA Graphical Model

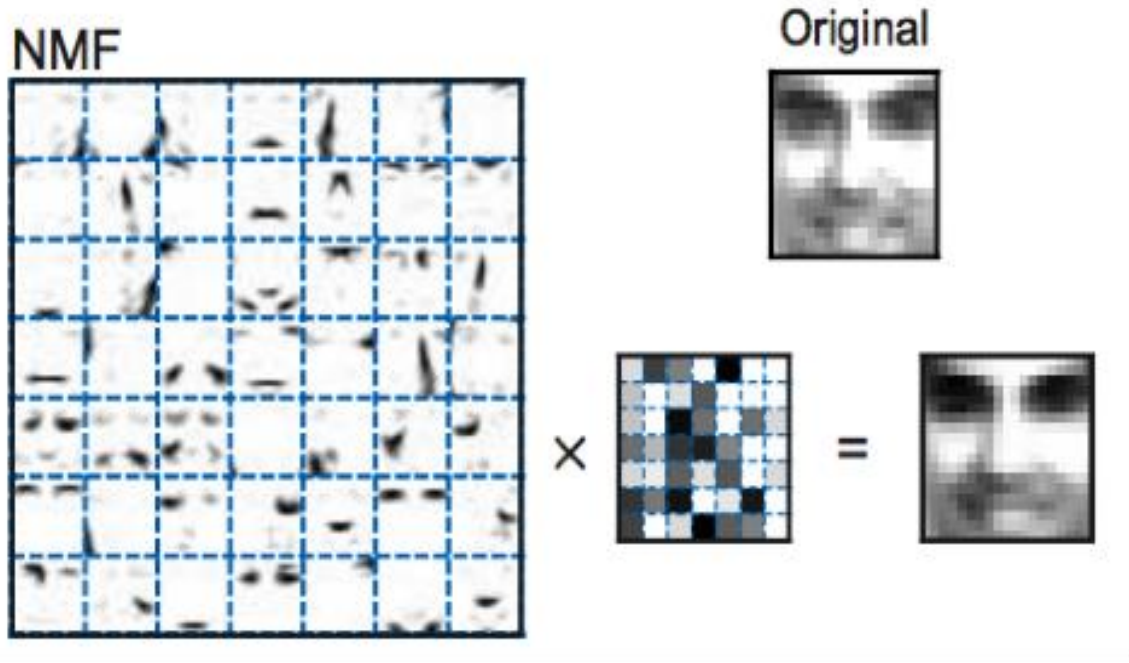


NMF also finds topics in the data

Matrix Factorization

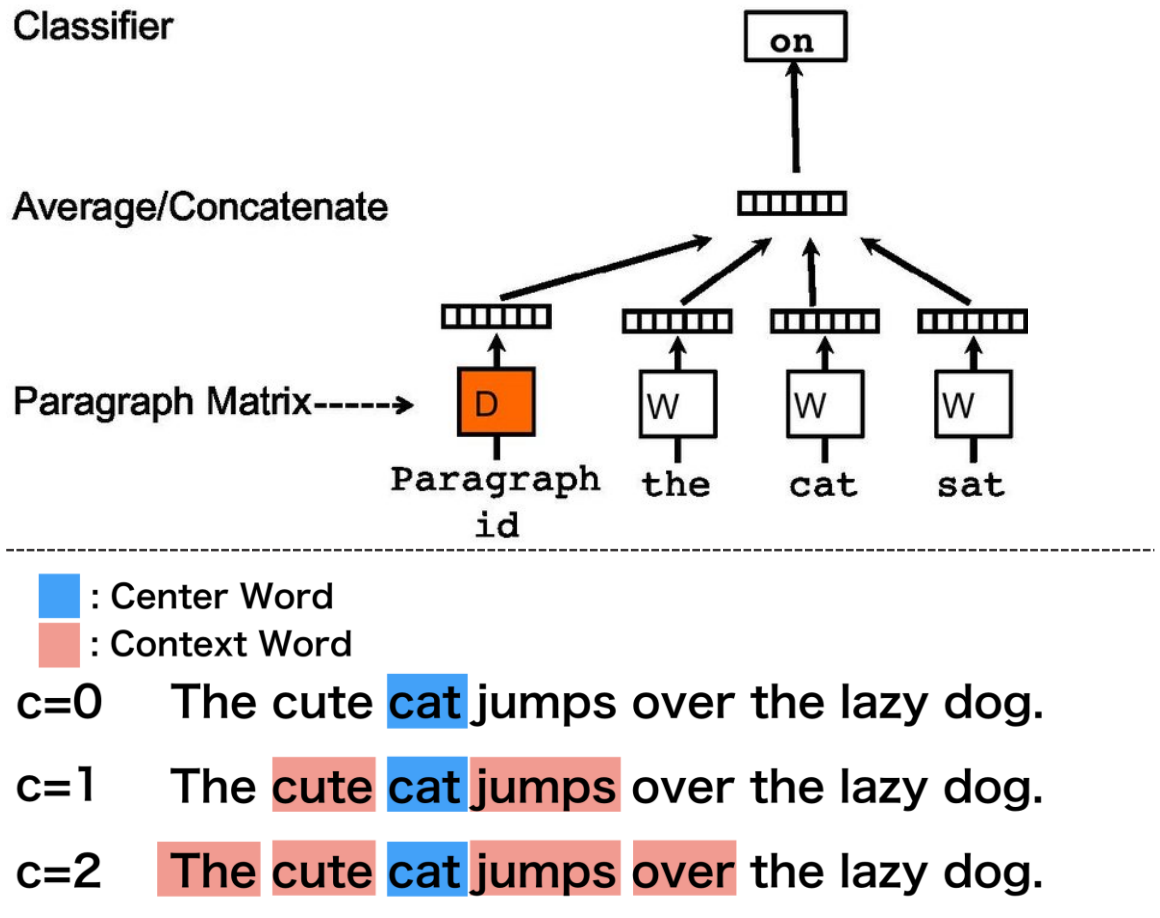


NMF creates "additive" elements

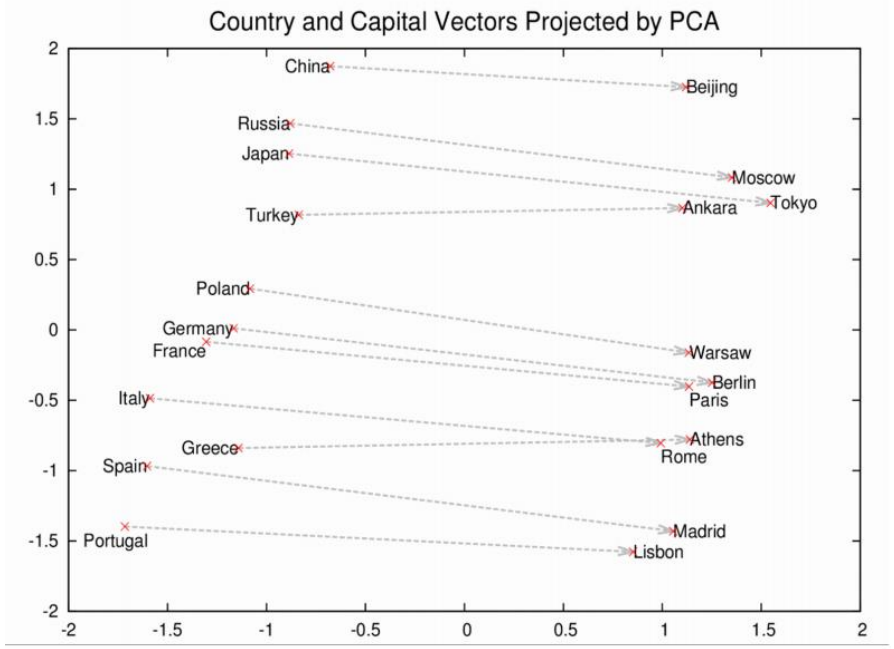


Doc2Vec generates a dense vector representation that preserves the semantics

Doc2Vec Problem



Word2Vec Country – Capital Example



Multi-label classification metrics

Hamming Loss

$$\frac{1}{|N| \cdot |L|} \sum_{i=1}^{|N|} \sum_{j=1}^{|L|} \text{xor}(y_{i,j}, z_{i,j})$$

where $y_{i,j}$ is the target and $z_{i,j}$ is the prediction.

Micro-averaged Precision and Recall

- Microaveraging Precision $Prc^{micro}(D) = \frac{\sum_{c_i \in C} TPs(c_i)}{\sum_{c_i \in C} TPs(c_i) + FPs(c_i)}$
- Microaveraging Recall $Rcl^{micro}(D) = \frac{\sum_{c_i \in C} TPs(c_i)}{\sum_{c_i \in C} TPs(c_i) + FNs(c_i)}$

Jaccard Similarity Score

		A	
		0	1
B	0	M_{00}	M_{10}
	1	M_{01}	M_{11}

$$J = \frac{M_{11}}{M_{01} + M_{10} + M_{11}}.$$

F1 score

$$F1 = 2. \frac{Precision \times Recall}{Precision + Recall}$$

Before, Neoway relied on ad-hoc methods which carried some limitations

Defining a set of rules

- Determined the relevant keywords to look for:
  - Organic
  - …
  - Tortilla
- Construct a set of RegExs to find the keywords
  - /[\w._%+-]+[]/
- Run the search on the database

Inspecting the HTML

```
<html>
 <body>

  <h1>
   Welcome to Whole Foods
  </h1>
   …
  <p>
   Find the best deals on organic
   apples
  </p>

 </body>
</html>
```

Limitations

- Heavily dependent on domain knowledge
- Not adaptable to new labels
- Difficult to handle various forms of the same lemma
- RegEx language hard to debug and error-prone
- Unwieldy supporting large vocabulary