Data Mining Capstone Task 7

Final Project Report

May 13, 2020

Asad Bin Imtiaz CS598
Data Mining Capstone
MCS-DS UIUC
Email: aimtiaz2@illinois.edu

1 Introduction:

In this task, an application system is developed to mine the dish references from the reviews from Yelp Reviews from Yelp Academic Challenge Dataset or any user provided review text using state of the art Machine learning & Natural Language Processing models. The analysis and findings are part of Task-7 for Data Mining Capstone course, wherein the Yelp restaurant reviews were explored to extract dish names mentioned in the review for a restaurant and predict the cuisine type.

To develop this application system, important features about dish names and their context in reviews texts are learned to develop classification logic to efficiently and effectively mine the dish references in these texts. The goal of this task was to come up with appropriate data science specific approaches to identify important features which represent dish names and compare their predictions.

The idea was to build up on Task-4/5, where Seg-Phrase and Top-Mine frequent pattern mining tools were used to min dish names. In this Task, instead of using these tools one of the following unsupervised techniques were used to mine the dishes and their results compared:

1. Dish mining with **Named Entity Recognizer**
2. Dish Mining with **Conditional Random Fields**
3. Dish Mining with **Word2Vec** model.

The resulting applications is deployed at: http://18.157.64.217/

The code repository of this project can be found at: https://github.com/AsadBinImtiaz/CS598_DMC

In the following, the implementation of these tasks and their sub tasks are explained.

1.1 Tools and Libraries:

The exploration and analysis of Yelp restaurant dataset was performed with **Python 3**. Several python libraries such as **gensim**, **sklearn**, **scipy, numpy** and **Scikit-learn** were used to analyze and cleanse the review texts.

1.2 Dataset:

For this task, the dataset with around 74,000 reviews was selected. This was done to get enough sample data to train the models and for the application system. The selection criteria was:

- All restaurant reviews for non-cuisine categories (example Event Planning, Grocery, Lounges etc.) were omitted.

- Only those reviews were selected which had a useful vote count of 2 or more. This was done to have some credibility of training data, since the reviews having useful votes can be considered implicitly verified by other users to have meaningful content.
- Reviews for restaurants having 5 or less reviews in total were not selected.
- Only reviews having minimum of 50 words were selected
- At max 10 reviews with score 4 and 5 were selected. This was done to have a balanced review set for positive and negative reviews for later sentiment analysis of extracted dishes. However, due to limited time, dish sentiment analysis was later descoped from the project.

The selected reviews were pickle dumped in data directory of the project as: **review_data.pickle.**

2 Approach

The extraction of Dish-Names is not an easy task, and there was no existing work found for directly mining dish names from text data. The dish mentions are hard to separate from other nouns. It is not easy to separate ***Harry Potter*** and ***Pot Pie***, or ***Sweet Person*** and ***Sweet Potato***, or ***Turkish Grill*** (restaurant name) vs ***Turkey Grill*** (dish name). On top, there can be spelling variations and mistakes (like ***Kebab*** and ***Kabab***), or abbreviations (like ***veggie roll*** vs ***vegetable roll***), or non-capitalizations in dish names (such as red ***chicken*** vs ***Red Chicken***) which could make NLP library wrongly POS tag words (e.g red as adjectives instead of noun in Red Chicken). Moreover, many dish names have conjunction words such as **With** or **And** in them, such as in '***Mac And Cheese***', or in a dish name like '***Thai Beef Curry With Rice***'. Therefore it was important to ensemble several approaches to effectiveluy min dish names.

For this task, following approaches were explored to extract dish names:

1. Using Named-Entity-Recognition (**NER**) with POS tagging
2. Using Conditional-Random-Fields (**CRF**) to classify words are dishes
3. Using **Word2Vec** model to find similarity of phrases with words like 'Dish', 'Food', 'Cuisine' etc.

These approaches are discussed in detail in subsequent chapters.

2.1 Data Preprocessing and cleansing

As always, the first step when processing text data is to clean it and remove unnecessary elements such as stop-words from the data. The text data was taken and following cleansing was performed:

- URLs from the data were removed
- Extra whitespaces and newlines were removed
- Word abbreviations were expanded and punctuations removed
- Stop-words were removed

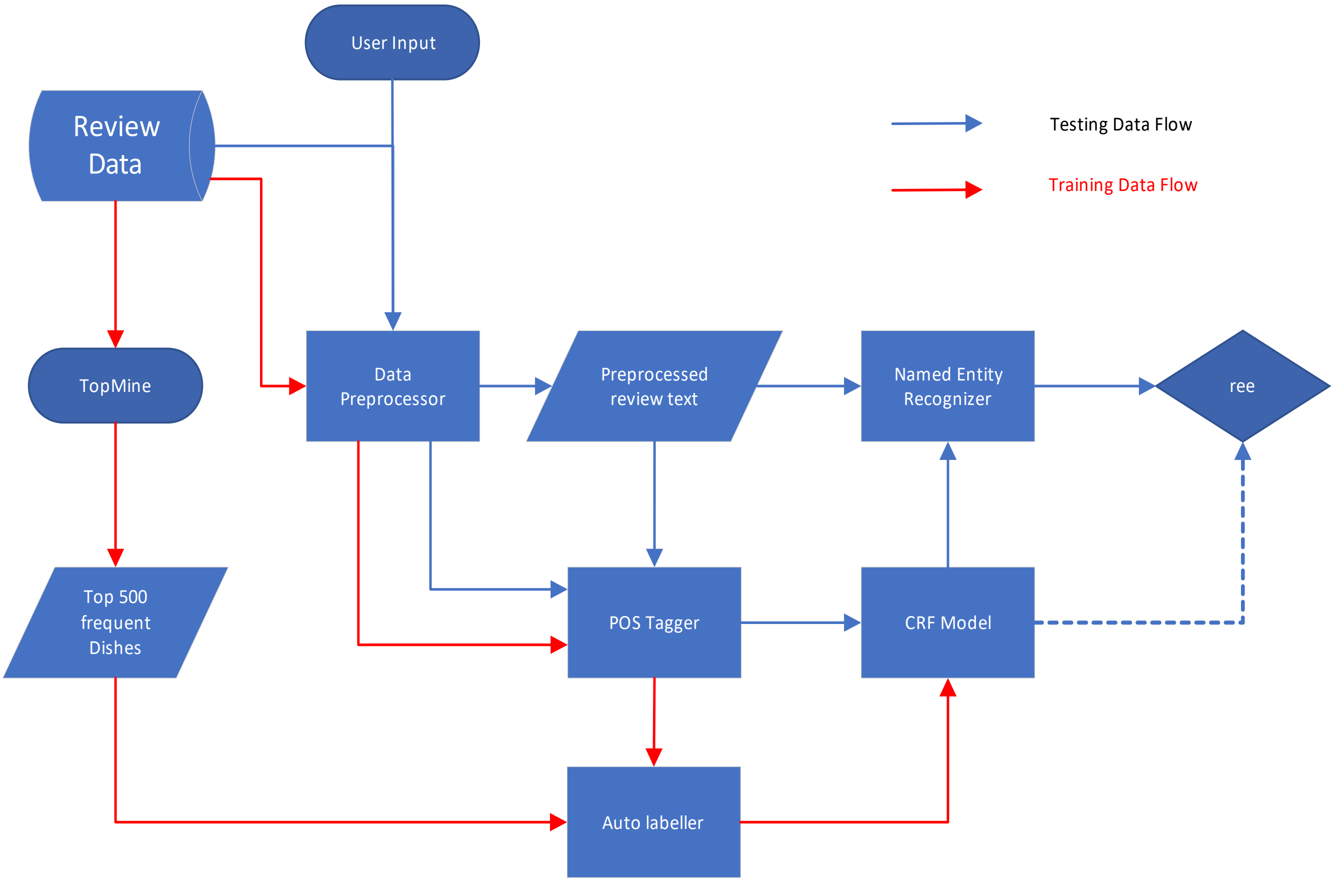
Additionally, from text data, bi-grams and tri-grams were extracted. This included:

- All tokens were converted into base lemma form.
- Merging frequently occurring words into bigrams and trigrams
- Merging Noun-Phrases by creating bi-grams by merging adjectives and nouns, or negations (such as not, barely etc.) and adjectives or adverbs. This gave phrases such as not-good, great-place etc. Please note that not is a stop-word, but is retained in phrases e.g. in not-bad.
- Only words representing Proper-Nouns, Nouns, Verbs, Adjectives, Adverbs, Determiners, interjections, Conjunctions and ad-positions were kept.
- All stop-words were removed.

After application of above pre-processing steps, processed review text for each restaurant was retrieved and was used in subsequent steps. The processed text was divided into training and testing splits based on the restaurants for which labels were already provided for training.

2.2 Modelling for Dish Miner

In this section, the steps to construct dish name extraction models for review texts are described. The general strategy was to mine a few dishes with frequent phrase mining, label the review texts from this dish list which have mentions of these dishes and construct the classifiers based on this training. The entire application flow is depicted in the following diagram:



2.2.1 Labelling Data for training

If dish name extraction is framed as a machine learning problem, we would need a manually labeled review data to train the classification model. This was not feasible to do in limited time of two weeks. Therefore, the top 500 dish names extracted in Task-4/5 were used as seed to label the existing data with dish references. Once the models were trained with this semi-supervised labeling of data, POS tag and NER based rules were learned to mine new dish names and classify them as new dishes.

Luckily the frequent dish names were already present from Task 6 which is used to seed as initial data for auto labelling for CRF modelling.

2.2.2 Text pre-pressing

For NER to work correctly, the data had to be pre-processed and non-useful elements were removed. This included:

1. Removal of URLs from data
2. Removing customized stop words
3. Removing extra spaces
4. Removing New lines
5. Ensuring each line ends with a period nor NER to split text into sentences.
6. Removing symbols from data
7. Converting abbreviations to long words (such as don't to do not and I'll to I will etc.).

Apart from above, for certain preprocessing tasks, lowercase operations were also performed. However, for current NER, Proper Nouns mattered, therefore no word-cases were kept as was.

2.2.3 Parts of Speech (POS) tagging

The processed review text was spitted into sentences, and each token of the sentence was tagged with the POS respective pos tag using python Spacy library. For POS tagging, the large English corpus was used. It was observed that almost all the dish names ended with a Noun, so POS tagging in this way helped in classification of a sequence of words into Food vs Non-Food categories.

2.2.4 Named Entity Recognition (NER)

One way to find dish references was to use named entity recognizer. Python Spacy library was used for NER tagging using the larger English corpus. In almost all cases, a dish reference was a Proper Noun or a Noun Phrase ending with a noun. To enhance the possibility of finding a dish, all nouns which were preceded by an adjective were converted to have the first character capital. Moreover, conjunctions and interjections between two noun phrases or a noun phrase and a noun were also titled. This helped NER recognize these phrases as entities. From the recognized entities, only following entity types were kept:

- Entities representing Persons (dishes like Asada, Gulab Jamun, Junior Stake etc. were qualified)
- Entities representing Products (dishes like Kebab, Samosa etc. were qualified)
- Entities representing Organizations (dishes like Apple, Orange etc. were qualified)
- Entities representing Work of Art (dishes like Thai Special Curry were qualified)

This was done since it was observed that all of dishes recognized were either a Proper person or organization name, or one which is classified as product or work of art. All entities representing geo-Political entities or locations were removed, giving much higher chance of getting a dish name. However, many of the recognized entities were still not dish names, rather person names and adjustive or noun phrases. To filter these out, a black list consisting of around 5000 common names and daily noun phrases was used to filter these entities out. Lastly, the name of the restaurant and the city/location references (such as **Corner of 10th and 15th St**) were removed. TF-IDF on the resulting entity list gave almost always a dish reference.

On web application, the results of NER are combined with CRF model, and are displayed in respective section and are also included in dish name prediction on original review text.

2.2.5 Conditional Random Field (CRF)

To further classify any dish unseen by NER or in seeded 500 dishes from top mine, a CRF model was trained. The text was preprocessed and pos tagged, and text and labelled at training with sample topmine dishes having references in NER sentences. Thereafter, text featurization of POS tagged tokens was performed which included features such as (list not exhaustive):

- Token itself
- Lowercase of token
- Is token a digit
- Is token titled
- Is token uppercased
- Preceding token
- Succeeding token
- POS tag
- First and last few characters
- POS tags of preceding/succeeding tokens

The CRF model was generated on around 2 million auto-labelled sentences, with at least one dish reference. The resulting model gave an F1 score of above 0.90 at training with cross validation of factor 5.

On web application, the results of CRF are displayed in respective section and are also included in dish name prediction on original review text.

2.2.6 Mining with Word2Vec

As an additional approach, a word2vec model was constructed to mine dishes with similarity to words like 'dish' or 'cuisine'. For this purpose, around 80,000 sentences referencing dishes were used to train the work to vector model. Before training, the text was preprocessed to retain only:

- Nouns
- Proper Nouns
- Adjectives
- Bigrams of two Noun/proper Noun
- Bigrams of adjective followed by a Noun /proper Noun
- Trigrms of 3 nouns//proper Noun
- Trigrams of Adjective following by a bigram of Noun /proper Noun

The resulting Word2Vec model was very rudimentary and often predicted invalid results or missed valid results. Therefore, for final dish name mining in original text, the results of this model are not included, but are still displayed on web interface in separate section.

2.3 Modelling for Cuisine Miner

To train the classifier for cuisine prediction, a Multinomial Naïve Bayes classified was constructed with TFIDF count vectorizer on sentences containing dish references and their original category in reviews. For more accuracy, the target cuisines were mapped to one of the following cuisine types

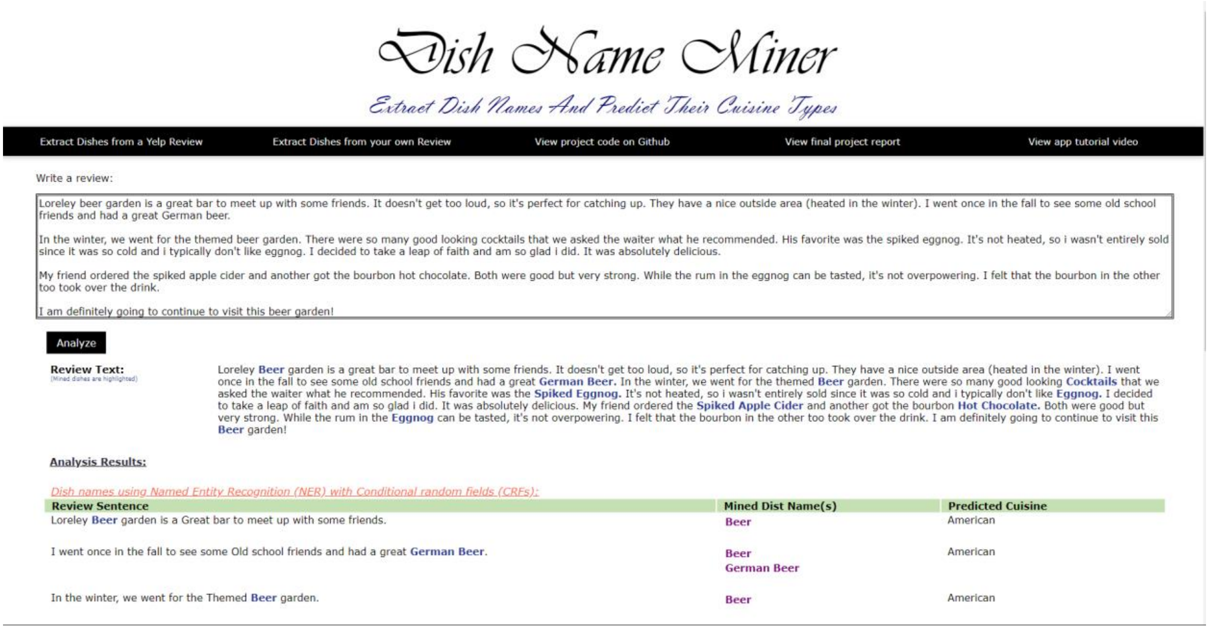
- **American** – All north American dishes
- **Afghan / Indian / Pakistani** – All sub-continental dishes

- **African** – All African dishes
- **Asian fusion** – All Asian dishes except from Middle east and Sub-continent
- **Italian / Pizza** – All dishes from around Italian region
- **French / Western European –** All dishes from around Western European region except UK
- **British** – All dishes from the region around United Kingdom
- **Seafood** – All seafood dishes
- **Fast Food** – All fast food, bakery and barbeque items
- **Bars / Coffee** – All cold and hot drinks
- **Buffets** – All buffet foods
- **Mexican / South American** – All dishes from South American region
- **Mediterranean / Middle Eastern** – All Mediterranean, east European and middle eastern dishes

Mapping dishes into fewer of above categories rather that more that 1000 category combinations in reviews gave an MNB classifier having and accuracy and f1-score of above 0.8 at training.

3 Web interface

The classification modules were bundled together using Python Flask in a web application which is running on an AWS EC2 cluster at: http://18.157.64.217/


Dish Name Miner
Extract Dish Names And Predict Their Cuisine Types
Extract Dishes from a Yelp Review
Extract Dishes from your own Review
View project code on Github
View final project report
View app tutorial video
Analyze

Write a review:

Loreley beer garden is a great bar to meet up with some friends. It doesn't get too loud, so it's perfect for catching up. They have a nice outside area (heated in the winter). I went once in the fall to see some old school friends and had a great German beer.

In the winter, we went for the themed beer garden. There were so many good looking cocktails that we asked the waiter what he recommended. His favorite was the spiked eggnog. It's not heated, so i wasn't entirely sold since it was so cold and i typically don't like eggnog. I decided to take a leap of faith and am so glad i did. It was absolutely delicious.

My friend ordered the spiked apple cider and another got the bourbon hot chocolate. Both were good but very strong. While the rum in the eggnog can be tasted, it's not overpowering. I felt that the bourbon in the other too took over the drink.

I am definitely going to continue to visit this beer garden!

Review Text: (Mined dishes are highlighted)

Loreley **Beer** garden is a great bar to meet up with some friends. It doesn't get too loud, so it's perfect for catching up. They have a nice outside area (heated in the winter). I went once in the fall to see some old school friends and had a great **German Beer.** In the winter, we went for the themed **Beer** garden. There were so many good looking **Cocktails** that we asked the waiter what he recommended. His favorite was the **Spiked Eggnog.** It's not heated, so i wasn't entirely sold since it was so cold and i typically don't like **Eggnog.** I decided to take a leap of faith and am so glad i did. It was absolutely delicious. My friend ordered the **Spiked Apple Cider** and another got the bourbon **Hot Chocolate.** Both were good but very strong. While the rum in the **Eggnog** can be tasted, it's not overpowering. I felt that the bourbon in the other too took over the drink. I am definitely going to continue to visit this **Beer** garden!

Analysis Results:

Dish names using Named Entity Recognition (NER) with Conditional random fields (CRFs):

Review Sentence	Mined Dist Name(s)	Predicted Cuisine
Loreley **Beer** garden is a Great bar to meet up with some friends.	**Beer**	American
I went once in the fall to see some Old school friends and had a great **German Beer**.	**Beer** **German Beer**	American
In the winter, we went for the Themed **Beer** garden.	**Beer**	American

The application gives user possibility to mine dishes from the selected review data set or to extract dish names from any review they may write or copy from internet. The prediction is done at the run time and nothing preprocessed is stored. The app code is located in a GitHub repository at: https://github.com/AsadBinImtiaz/CS598_DMC

4 Results

In most cases, the CRF model and CRF+NER combined gave very reasonable results. However, since dish name mining is not an easy task to fully accomplish in 2-weeks' time, sometimes a predicted dish is clearly a mis-prediction, or sometime a prediction is missed at all.

Having said that, the application can help users in quick visualization of mentioned dishes in any review text. Moreover a general cuisine is also predicted for each of the mined dish, allowing user to asses the cuisine of the restaurant and the sentiment of dishes for specific cuisines.

The potential user can be the owners of a restaurant wanting to know the most discussed dishes and their sentiments, or users who are looking at long review texts but wan to know most popular/unpopular dishes in the text.

5 Conclusions

The work can be further extended to find the sentiment of specific dish in a review text and can further be aggregated to most or least popular dishes of a restaurant or chain. The owners can analyze the most discussed dishes, as well as which dishes are highly rated, and what not those which are not so highly rated, and their overall effect on restaurant rating. From this work, one can predict dishes in pain text with reasonably good accuracy. However, each of the tree techniques used have produced their pros and cons.

The NER could mine complicated dish names such as "***Chicken Tikka Masala with Garlic Naan***" but sometimes resulted in inaccurate names such as "Fresh Cutlery"

The CRF model almost always predicted valid dishes, but often splitted long dish names into multiple, such as from *"Chicken Tikka Masala with Garlic Naan"* it may predict "***Chicken Tikka Masala***" and "***Garlic Naan***" as two separate dishes. Sometime is also misses part of the dish name such as from "*Thai Red Curry"* it may predict **"Curry"** only.

Word2Vector seldomly produces complicated dis names and often the results were missed or not correct.