Data Mining Capstone Task 3

February 27, 2020

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

1 Introduction:

In this task, popular dishes in the Yelp Review Academic Challenge Dataset for specific cuisines were extracted using frequent phrase mining algorithms. The analysis and findings are part of Task-3 for Data Mining Capstone course, wherein the Yelp restaurant reviews were explored to identify popular dishes for some specific cuisines. The goal of this task was to come up with a dish recognizer which would identify common dishes present in any cuisine.

In the following, the implementation of the sub tasks for this task are explained.

1.1 Tools and Libraries:

The exploration and analysis of Yelp restaurant dataset was performed with Python 3 using Jupyter notebook and with Java 12. Phrasal mining algorithms such as **SegPhrase**, **AutoPhrase** and **TopMine** were used. Several python libraries such as **Gensim**, **scipy, numpy** and **Scikit-learn** were used to analyze and cleanse the review texts.

1.2 Application:

The first step was to gather all the relevant reviews. All reviews for the categories of **"American (New)", "Chinese", "Indian", "Italian", "Mexican"** and **"Mediterranean"** were extracted. To accomplish the phrase mining efficiently, a sample of 50,000 reviews for each category were retained.

The following tasks, as described in subsequent sections in this report, were performed:

1. Manual Tagging of dishes – *See chapter 2*
2. Mining Additional Dish Names – *See chapter 3*
    a. Mutual Information (MI) calculation
    b. Dish name mining with TopMine
    c. Dish name mining with SegPhrase
    d. Dish name mining with AutoPhrase
    e. Dish name mining with Word2Vec

In the following, the results of the analysis for each of the above are explained in detail.

2 Task 3.1: Manual Tagging

In this sub-task, candidate dish names were manually tagged and labelled for SegPhease. This label list was already provided with the task but was modified to correct some of the labels. For example, the false positive label of "**spring training**" (among others) in "American_(New).label" was changed to **0** while false negative label of "**deep dish pizzas**" was changed to **1**. Similarly, in the label files for

Chinese, Indian, Italian, Mexican and Mediterranean, the False-Positive and False-Negative labels were also corrected to True Negative and True positive accordingly.

As a first set, the provided label files were not used, rather new label files were generated by using a different sample of 5000 reviews for each category. These generated table files were exhaustively analyzed to gather more true positive tables to complement the provided and corrected table files. Some labels such as "**cole slaw**" & "**mash potatoes**" (among others) were discovered in American (New) cuisine and were added to the main label file. This was also done for the other 5 categories.

Some ambiguous labels were also removed increase quality of generated results. Labels similar to "**shrimp and pork**" in Chinese label file or " **their garlic sauce**" in the Mediterranean label file could be confusing in the phrase mining and were either corrected by removing the unwanted words or removed from the file altogether.

These manually tagged labels are used in dish name mining with **SegPhrase** (as discussed in chapter 3.3).

3 Task 3.2: Mining Additional Dish Names

In this subtask several phrase mining algorithms were used to extract dish names from the data set. Each algorithm produced a different quality of resulting dish name list and was analyzed to evaluate the general accuracy of the algorithm. The following approaches for Dish Name recognition were applied:

a. **Mutual Information** calculation – *Chapter 3.1*
b. Dish name mining with **TopMine** – *Chapter 3.2*
c. Dish name mining with **SegPhrase** – *Chapter 3.3*
d. Dish name mining with **AutoPhrase** – *Chapter 3.4*
e. Dish name mining with **Word2Vec** – *Chapter 3.5*

It is worth mentioning here that the data for the dish name extraction sub tasks were already preprocessed and tokens were already converted in their base lemma form and the stop words were removed.

3.1 Task 3.2 (A): Mutual Information calculation

The Mutual information between the frequent words in all of the 6 selected categories (both combined and separately) was calculated. The idea was to compute the co-occurrence strength between the words, to estimate the probability for them being in a multi word phrase. However, this information was not useful since each of the unigram word was associated with another with a MI score, which was not easy to interpret. Moreover, dish names mostly are not uni-grams so the results were not very useful.

3.2 Task 3.2 (B): Using TopMine Algorithm

Next the TopMine algorithm was applied on the review texts of each of the categories. TopMine is an unsupervised quality frequent phrase mining algorithm.

For the TopMine settings, support level was set to 10 and max tokens in the phrase was set to 6 and number of iterations set to 100. The results of the TopMine in terms of top 10 phrases for each of the categories is shown in the table below. The score here is the frequency of the phrase in the data sample.

TopMine					
Top 10 Frequent Phrases					
America New	Score	Chinese	Score	Italian	Score
happy hour	2613	*dim sum*	3612	*good pizza*	2242
chicken waffle	2607	*chinese food*	2974	*margherita pizza*	2027
food good	2604	*pho kim long*	1894	*bread basket*	1814
mac cheese	1553	*fry rice*	1812	*spaghetti meatball*	1789
sweet potato fry	1152	*egg roll*	1462	*pasta dish*	1330
fry chicken	977	*spring roll*	1254	*bottle wine*	976
cheesecake factory	885	*sweet potato fry*	1152	*garlic bread*	939
french toast	833	*orange chicken*	1097	*tomato sauce*	846
wait minute	769	*cheesecake factory*	885	*caesar salad*	730
short rib	768	*hot sour soup*	863	*olive oil*	697
Indian	**Score**	**Mexican**	**Score**	**Mediterranean**	**Score**
garlic naan	2615	*fish taco*	2059	*pita jungle*	2542
chicken tikka masala	2607	*chips salsa*	2039	*great food*	2301
spaghetti meatball	1789	*rice bean*	1398	*hookah lounge*	1588
tandoori chicken	1506	*gallo blanco*	1166	*flat bread*	1472
mango lassi	1062	*black bean*	1056	*beef carpaccio*	1423
bottle wine	976	*street taco*	1051	*olive ivy*	1323
chicken curry	972	*highly recommend*	953	*bread basket*	1142
butter chicken	948	*chicken enchilada*	896	*short rib*	927
gulab jamun	914	*corn tortilla*	891	*olive oil*	813
tikka masala	852	*chile relleno*	825	*ice cream*	795

As can be seen, the results are reasonably okay with a few phrases not representing dish names in the topic representing dishes. For example, in the Indian, Chinese and Italian categories, all of the top 10 results are popular dishes of respective cuisines. However, there are some inaccuracies such as "**happy hour**" in American (New) cuisine or "**Highly recommend**" in Mexican cuisine, which are obviously not dish names.

Let's see if the results could be improved with SegPhease.

3.3 Task 3.2 (C): Using SegPhrase

In this sub task, the Dish name mining was done with SegPhrase (Jialu Liu, n.d.).

As first step, the Auto-Label setting was turned on for a sample of 5000 reviews to gather some more quality dish name phrases for enriching the provided label files. This process of enriching label files is discussed in chapter 2.

Once the label files were enriched and false positive label were removed and true positive added, these table files were used used for training with the Auto-Label setting turned off. Following settings were used:

- The Support threshold was set to 10
- The max Iteration number was set to 5
- The max number of tokens in a phrase to 6
- The wordnet noun was set to 0

The algorithm was performed on all 6 category texts iteratively. These text files comprised of 50,000 preprocessed reviews for each of the cuisine type. The Pre-processing included token lemmatization, stop-word removal and retention of specific POS tags; namely Nouns, Verbs, Adverbs, Adjectives and Determiners.

The results in terms of top 10 phrases for each of the category was generated is shown in the table below:

SegPhrase					
Top 10 Frequent Phrases					
America New	**Score**	**Chinese**	**Score**	**Italian**	**Score**
zucchini fry	0.93738169	*fried egg*	0.91576738	*cooked perfection*	0.92405985
sage fried chicken	0.93391903	*thai curry*	0.90796381	*balsamic vinegar*	0.91959047
mac cheese	0.93374460	*dim sum*	0.89776206	*mario batali*	0.91952784
turkey burger	0.93001348	*pho kim long*	0.89667780	*filet mignon*	0.91922584
fry chicken waffle	0.93001348	*orange chicken*	0.89557540	*al dente*	0.91326477
chop salad	0.92984120	*pork belly*	0.89455911	*pizza margherita*	0.91114584
bread pudding	0.92878398	*taste kung pao*	0.89417565	*ice cream*	0.90978081
goat cheese	0.92824001	*hoisin sauce*	0.89326790	*spaghetti meatball*	0.90800980
chicken breast	0.92820474	*chicken soup*	0.89222804	*olive oil*	0.90563586
cheesecake factory	0.92776630	*sushi sushi roll*	0.89102507	*smoked salmon*	0.90354018
Indian	**Score**	**Mexican**	**Score**	**Mediterranean**	**Score**
garlic naan	0.95800768	*shrimp chorizo*	0.90386185	*pita jungle service food*	0.92110210
brown rice	0.95748908	*chips salsa*	0.90386185	*mixed green salad*	0.92073173
mango lassi	0.95685180	*fish taco*	0.90386185	*salad veggie burger*	0.92057974
rice pudding kheer	0.95377389	*chicken enchilada*	0.90386185	*spaghetti meatball*	0.91917863
flat bread	0.95102835	*horchata rice*	0.90206697	*lentil soup*	0.91884344
chicken tikka masala	0.95011965	*corn tortilla*	0.90206697	*hummus chicken appetizer*	0.91884344
tandoori chicken	0.94932060	*appetizer tortilla bean*	0.90206697	*rice gyro meat*	0.91884344
mango pudding carrot	0.94842000	*guacamole flight*	0.90205552	*cook al dente*	0.91880826
chilli kebab platter	0.94803171	*appetizer chips salsa*	0.90205552	*chicken filet mignon*	0.91880826
chicken curry	0.94654702	*shrimp tacos*	0.90126569	*flatbread ravioli*	0.91878728

As can be seen, the results are remarkably good and the top 10 high quality phrases are almost always a dish name in all of the cases. The manually tagged labelled list helped the miner to retain only the high-quality phrases.

3.4 Task 3.2 (D): Using AutoPhrase

In this sub task, the Dish name mining was done with **AutoMine** (Jingbo Shang, n.d.), which is claimed to have more improvements over **SegPhrase**. The Support threshold was set to 10, the max Iteration number to 5 and the max number of tokens in a phrase to 6. The algorithm was performed on all 6 category texts and the results in terms of top 10 phrases for each category is shown in the table below:

AutoPhrase					
Top 10 Frequent Phrases					
America New	**Score**	**Chinese**	**Score**	**Italian**	**Score**
pulled pork	0.90167371	*sugar cane*	0.93337092	*blood orange*	0.91719167
poached egg	0.89993675	*papaya salad*	0.92881902	*mamma mia*	0.91652395
peanut butter	0.89989261	*dim sum*	0.92875140	*le provencal*	0.91257753
cole slaw	0.89825726	*sam woo*	0.92771841	*lagunitas ipa*	0.90978012
eggs benedict	0.89425803	*chen wok*	0.92667058	*meatpacking district*	0.90860496
smoked salmon	0.89412178	*pho kim chicken*	0.92618466	*broccoli rabe*	0.90785202
mgm grand	0.89401837	*shishito pepper*	0.92567369	*pizza margherita*	0.90755681
hash browns	0.89357116	*sesame seed*	0.92537112	*poppy seed*	0.90747147
duck confit	0.89337578	*cajeta flan*	0.92514461	*hip hop*	0.90738095
cotton candy	0.89337577	*bachelorette party*	0.92503198	*eggs benedict*	0.90694854
Indian	**Score**	**Mexican**	**Score**	**Mediterranean**	**Score**
kabli pulao	0.94743472	*mariachi band*	0.90694339	*taco bell*	0.94004745
belly dancing	0.94301010	*volcano cookie*	0.90634061	*puff pastry*	0.93946604
seekh kabab	0.94288975	*passion fruit*	0.90592059	*philly cheesesteak*	0.93915983
daal maharani	0.94280114	*churro tot*	0.90584621	*rice pilaf*	0.93873404
health inspection	0.94253531	*fortune cookie*	0.90530442	*fremont street*	0.93864248
kebab mahal	0.94196539	*chips salsa*	0.90470102	*ali baba*	0.93842364
chicken tikka masala	0.94020639	*planet hollywood*	0.90469890	*avgolemono soup*	0.93814974
sear tofu	0.94006193	*shu mai*	0.90398738	*cheesecake factory*	0.93779585
baingan bartha	0.93984807	*goat cheese*	0.90371827	*hole wall*	0.93760772
grocery store	0.93983429	*hot dogs*	0.90298360	*hanger steak*	0.93747120

Despite the fact that the **AutoPhrase** phrase mining was unsupervised and no manual labels were provided (in contract to **SegPhrase**), the results reasonably good, though some of the phrases were obviously far from being dish names.

3.5 Task 3.2 (E): Using Word2Vec

In this sub-task, **Word2Vec** (models.word2vec – Word2vec embeddings, n.d.) in python was used to mine top frequent dish names. Python genism library was used to create a word2vec model for each of the selected category, with 5,000 vector dimensions and 20 iterations. Before creating the model, bigrams were created with a min-count of 4 and threshold of 1.

Once the model was created, the positive distance between the word '**dish**' and bigrams in the model was calculated (only bigrams were returned). The results are shown in the table below. The score here represents the vector distance between the word 'dish' and the bi-gram phrase.

Word2Vec					
Top 10 Frequent Phrases					
America New	**Score**	**Chinese**	**Score**	**Italian**	**Score**
green bean	0.41307586	*stir frie*	0.39408854	*pasta cook*	0.52054375
cast iron	0.39234251	*bell pepper*	0.39146996	*black truffle*	0.51151115
sea bass	0.38351846	*spicy sauce*	0.37574905	*al dente*	0.49955612
cream sauce	0.37390384	*garlic sauce*	0.37300736	*pork belly*	0.46560961
mash potatoe	0.36415756	*clay pot*	0.34477097	*brown butter*	0.45664680
brussel sprout	0.36057016	*mapo tofu*	0.34379196	*creamy polenta*	0.45280433
pan seared	0.35750324	*bean sauce*	0.33908427	*squid ink*	0.44802022
mashed potatoe	0.34987217	*thai style*	0.33260161	*bone marrow*	0.43743646
bread pudding	0.34463316	*singapore noodle*	0.31919032	*lobster gnocchi*	0.43240350
pot pie	0.34302545	*black pepper*	0.31781009	*melted mouth*	0.42854014
Indian	**Score**	**Mexican**	**Score**	**Mediterranean**	**Score**
basmati rice	0.30863297	*goat cheese*	0.49950641	*pork loin*	0.35750413
fish curry	0.30369198	*mole sauce*	0.45765573	*veal cheek*	0.35018638
tender sauce	0.29365849	*chicken breast*	0.43017730	*foie gra*	0.33790028
lamb goat	0.29341763	*spicy sauce*	0.41204765	*sweet corn*	0.32510704
tandoori chicken	0.29241410	*fry egg*	0.41086793	*beef belly*	0.31337315
tikki masala	0.28896201	*cream sauce*	0.40921682	*ricotta ravioli*	0.30675721
lentil soup	0.27102092	*mashed potatoe*	0.39969632	*tender juicy*	0.29658297
chicken makhani	0.26762787	*chile verde*	0.37079218	*wagyu beef*	0.29590678
chili chicken	0.26642936	*dan dan*	0.32840639	*foie gras*	0.29415131
chicken breast	0.26051587	*pork bun*	0.30877316	*poached egg*	0.28740370

It can be seen that Word2Vec model generated quite good results but the quality is not as good as was with SegPhrase. It is worth mentioning here that the data for word2vec consisted of parts of speech (POS) of types Nouns, Verbs, Adverbs & Adjectives only to increase the quality of results.

4 Conclusions

It can be concluded that **SegPhrase** with manual labels generated remarkably good results. The accuracy of **AutoPhrase** was also good considering no manual labels were provided and still reasonable results were achieved.

The **Mutual Information** calculation on uni-grams was barely useful for dish name extraction.

If the data good pre-processed (such as specific POS tags are retained and stop words are removed), **word2vec** also provide acceptable results to mine dish names.

5 References

Jialu Liu, J. S. (n.d.). *Mining Quality Phrases from Massive Text Corpora*. Retrieved from GitHub: https://github.com/shangjingbo1226/SegPhrase

Jingbo Shang, J. L. (n.d.). *AutoPhrase: Automated Phrase Mining from Massive Text Corpora*. Retrieved from GitHub: https://github.com/shangjingbo1226/AutoPhrase

models.word2vec – Word2vec embeddings. (n.d.). Retrieved from Gensim: https://radimrehurek.com/gensim/models/word2vec.html