Data Mining Project - Week 3 - Dish Recognition

Data Mining Specialization - Coursera / University of Illinois at Urbana-Champaign

- Author: Michael Onishi
- Date: November, 2019

Introduction

The goal of this task is to mine the data set to discover the common/popular dishes of a particular cuisine. Typically when you go to try a new cuisine, you don't know beforehand the types of dishes that are available for that cuisine. For this task, we would like to identify the dishes that are available for a cuisine by building a dish recognizer.

For this task, I chose the Italian cuisine to explore. I used python with Pandas and NumPy for processing and filtering the text files. For new dish names discovery, I used AutoPhrase (Automated Phrase Mining from Massive Text Corpora).

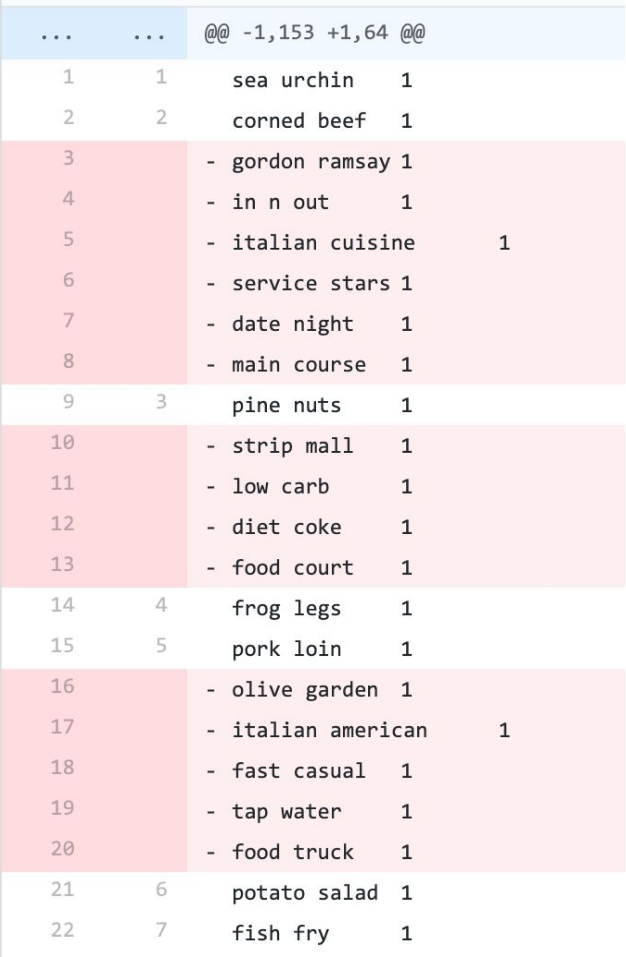
The full notebook with some steps taken here is available at:
https://github.com/michaelonishi/coursera-data-mining-specialization/blob/master/c6-data-mining-project/task3/Data_Mining_Project_Week_3_Dish_Recognizer.ipynb

Task 3.1: Manual Tagging

You are given a list of candidate dish names, which are all frequent (at least 10 times in corresponding corpus), automatically generated by the auto-labeling process of SegPhrase[2]. Some of the dish names are verified by an outside knowledge base such that they are all good phrases, and some of them might be good dish names. However, some of the labels might be wrong. Therefore, your task here is to refine the label list for one cuisine. You could modify/add some phrases.

The original list had a lot of false positives and false negatives. It was a difficult task because I am not used to english named dishes, so I had to search the internet for many phrases to confirm if they represent a dish or not. I also modified some phrases to be a single word, because I think that in simple restaurants some dishes names could be a single word like meatball or salmon.

In total, I removed 140 lines and modified 33 records. Below there is a sample of some cleaning:



```
...   ...   @@ -1,153 +1,64 @@
 1     1      sea urchin    1
 2     2      corned beef   1
 3          - gordon ramsay 1
 4          - in n out      1
 5          - italian cuisine       1
 6          - service stars 1
 7          - date night    1
 8          - main course   1
 9     3      pine nuts     1
10          - strip mall    1
11          - low carb      1
12          - diet coke     1
13          - food court    1
14     4      frog legs     1
15     5      pork loin     1
16          - olive garden  1
17          - italian american      1
18          - fast casual   1
19          - tap water     1
20          - food truck    1
21     6      potato salad  1
22     7      fish fry      1
```

Task 3.2: Mining Additional Dish Names

Once you have a list of dish names, it is likely that many dish names are still missing. In this step, you would expand the list of dishes by using other pattern mining techniques and/or word association methods.

Here I applied AutoPhrase because it has some advantages over SegPhrase like less human effort (only needs positive quality phrases for training), supports multiple languages with automatic detection, POS-guided phrasal segmentation model.

For training, I extracted only the positive quality phrases from the manual reviewed 'Italian.labels' file. That list replaced the 'data/EN/wiki_quality.txt' file from AutoPhrase. Then I formatted all the reviews from Italian restaurants to the expected file format for training (reviews separated by a line break and a line with one single dot). The training file looks like this:

```
The best Italian food in town, hands down.  They've got homemade (...)
.
I always crave Vin Santo, even though I haven't been there for years, I still (...)
.
Vin Santo rules!    This is a great casual restaurant that is owned/operated (....)
.
(...)
```

I applied the following basic text preprocessing from the reviews:

- Removed accents
- Removed http links
- Removed line breaks and tabs
- Removed leading and trailing spaces

Then I ran AutoPhrase with the script 'auto_phrase.sh' with the following changes:

```
MODEL=${MODEL:- "models/yelp"}
# RAW_TRAIN is the input of AutoPhrase, where each line is a single document.
RAW_TRAIN=${RAW_TRAIN:- data/yelp-italian-reviews.txt}
```

In two minutes I had a model trained with a list of quality phrases and a score. I then filtered only quality phrases with scores greater than 0.95 and removed the already known dish names. That gave me 404 potencial new dish names. Here I list the first 20 lines:

1. shrimp
2. pizza
3. steak
4. bruschetta
5. gnocchi
6. peanut butter
7. hip hop
8. ricotta
9. pasta
10. au jus
11. veal
12. risotto

13. crab cakes
14. price
15. chicken
16. chef
17. french toast
18. angel hair
19. baked ziti
20. Pizzeria

We can see that some basic dishes were listed with a little noise. But going further, we could easily see some interesting dishes like:

- chilean sea bass
- applewood smoked bacon
- eggplant parmesan
- fettuccine alfredo
- osso bucco
- minestrone soup
- fingerling potatoes
- butternut squash ravioli
- steamed mussels
- pistachio gelato

Conclusion

This third task was very challenging, because it had very little guidance. The manual was very difficult for me, because I am not used to dish names in English. But it is clear that this starting point with a small but consistent initial quality phrases was worth the effort. The AutoPhrase showed very good results, giving a good list of possible quality phrases that could easily expand the dish names from a cuisine.

References

Jingbo Shang, Jialu Liu, Meng Jiang, Xiang Ren, Clare R Voss, Jiawei Han, "Automated Phrase Mining from Massive Text Corpora", accepted by IEEE Transactions on Knowledge and Data Engineering, Feb. 2018.

Jialu Liu*, Jingbo Shang*, Chi Wang, Xiang Ren and Jiawei Han, "Mining Quality Phrases from Massive Text Corpora”, Proc. of 2015 ACM SIGMOD Int. Conf. on Management of Data (SIGMOD'15), Melbourne, Australia, May 2015. (* equally contributed, slides)