Task 3: Dish Recognition

In this task phrase mining techniques are applied to the Yelp dataset to discover dish names for a particular cuisine. In this case dish name phrases will be mined from the Yelp Chinese cuisine reviews.

Preparing the Data

This task uses two source files from the Data Mining Capstone course. The first is a list of candidate phrases for the Chines cuisine dish names, *Chinese.label.* The auto-labeling process of SegPhrase framework automatically generated the labels (dish names) in this file. The second file is the Chinese review corpus, *Chinese.txt*.

Pre-processing of the data included cleaning the review corpus by removing all blank lines in Chinese.txt so that each line contained a single review. This was done in R.

Next, the candidate dish labels were visually reviewed in a text editor. The file contains a list of labels and a corresponding rating of either 1 if it was computed by SegPhrase to be a positive candidate phrase or a 0 if it was computed to be a negative candidate phrase. Any candidate phrase that was visually assessed as a false positive was deleted from the file. Any candidate phrase that was assessed to be a false negative was changed to a positive phrase by changing the 0 value to 1. In addition to the resulting label list several additional dish names were manually added the files as positive candidate phrases. See Appendix A for the contents of the modified labels file.

Mining Dish Names

Phrase Mining Using SegPhrase

SegPhrase is a state of the art phrase mining framework that integrates two steps - *phrase quality assessment* and *phrasal segmentation*. Phrase quality assessment uses a set of labels to compute frequent phrases of reasonable quality. The phrasal segmentation step uses the quality phrases to guide the segmentation of the corpus to rectify the phrase quality estimation. Rectification examines the context of word sequences in text to determine if it a quality phrase, which is an advancement over using term frequency alone to determine the quality of a phrase.

The first step in mining Chines dish names was running the SegPhrase algorithm on the Chinese review corpus. The documentation for SegPhrase was quite sparse, but after analyzing the code it appears that the train.sh script incorporates both the phrase quality assessment and phrasal segmentation steps. This experiment was conducted under that assumption.

For this activity, the SegPhrase *train.sh* script was executed using modified list of labels (candidate phrases) and review corpus, described above, as inputs. The following are the parameters that were changed from default values in *train.sh*:

```
RAW_TEXT='Chinese.txt'
AUTO_LABEL=0
DATA_LABEL='data/Chinese.label'
```

The output file from the algorithm, *ranking_1.csv*, provided the list of ranked quality phrases.

Phrase Mining Using ToPMine

ToPMine is framework for phrase mining and topic modeling. Phrase mining is based on term frequency. ToPMine was used to mine additional dish names from the review corpus. The following are the parameters used in the TopMine *run.sh* script.

```
inputFile='../rawFiles/Chinese.txt'
minsup=10
maxPattern=8
topicModel=2
numTopics=5
gibbsSamplingIterations=500
thresh=4
optimizationBurnIn=100
alpha=2
optimizationInterval=50
```

The resulting output file, *topPhrases.txt,* contained the candidate dish names mined.

Compiling the Final Dish Name List

To compile the final Chinese dish name list the output files from SegPhrase and TopMine (*ranking_1.csv* and *topPhrases.txt* respectively) along with the label file were read into R. Both the SegPhrase and TopMine phrases were assessed to determine at which point in each list the phrases no longer represented strong dish name candidates. For the SegPhrase list the occurred after the first 3,000 phrases. For TopMine this occurred after 3,700 phrases. So the final dish list was compiled by combining the positive labels from the label file, the first 3,000 SegPhrase terms and the first 3,700 TopMine phrases. A more scientific approach would be to use the phrase quality metric or frequency as a threshold to select the top phrases.

Conclusions

SegPhrase and TopMine, both state of the art phase mining techniques, produce quality phrases. While both result sets contain a mixture of dish name and non-dish name phrases, SegPhrase's top ranked results contain more dish names (see table below – non-dish names are highlighted in red font). This may be due to SegPhrase's use of user-provided labels (dish names) into the algorithm as a feature to derive quality phrases. The dish names that both algorithms produced are useful in that they do represent actual Chinese dish names and prove that phrase mining can produce quality results.

SegPhrase	TopMine
stir fry	Chinese food
fried rice	fried rice
brown sauce	Chinese restaurant
fortune cookie	egg rolls
hong kong	orange chicken
sea bass	lunch specials
food court	food is good
white rice	Panda Express
hot pot	hot and sour soup

chow mein	Mongolian beef
bok choy	pretty good
san francisco	dim sum
beef stew	crab puffs
char siu	Chinese place
bitter melon	love this place
panda express	egg drop soup
chicken wings	good food
spare ribs	chow mein
brown rice	spring rolls
dim sum	Kung Pao Chicken
bamboo shoots	great food
xo sauce	lo mein
soy sauce	wonton soup
steamed rice	soy sauce
food poisoning	noodle soup

I found the challenge of this task to be in post-processing of the phrase lists to extract just the dish names. These algorithms are not domain specific – their intent is produce quality phrases from a corpus. Additional processing is required to extract a more specific list of dish names from the broad list of phrases. Possible areas for further exploration could include investigating the topic produced by ToPMine and applying clustering or topic modeling as a post-processing step SegPhrase to further mine dish names from the general result set.

Task 3: Dish Recognition

Appendix A: Chines Dish Name Labels

sesame seeds 1	chow mein 1	hot and sour soup 1
sweet and sour sauce 1	hoisin sauce 1	lo mien 1
fried rice 1	peking duck 1	wonton soup 1
hot sauce 1	bean curd 1	plum sauce 1
winter melon 1	pulled pork 1	pancakes 1
hash browns 1	kung pao chicken 1	mongolian chicken 1
fried chicken 1	bone marrow 1	sweet and sour pork 1
spring roll 1	sesame seed 1	Kung Pao chicken 1glutinous rice 1
soy sauce 1	chicken wings 1	Crispy fried chicken 1
dim sum 1	xo sauce 1	Mapo tofu 1
beef tongue 1	green pepper 1	Buddha jumps over the wall 1
frog legs 1	miso soup 1	Cantonese seafood soup 1
tomato sauce 1	fortune cookie 1	ginger duck 1
bitter melon 1	rice cake 1	Bang Tofu 1
vanilla ice 1	spare ribs 1	Bright Pearl Abalone 1
prime rib 1	shark fin soup 1	Caterpillar Fungus Duck 1
stir fry 1	duck sauce 1	Crab and Fish Stomachs 1
foie gras 1	chinese noodles 1	Crab-apple Flower Cake 1
chinese sausage 1	rice noodle 1	Dried Pot Tofu 1
rice noodles 1	tom yum 1	Five Colours Fish Cake 1
mashed potatoes 1	hot pot 1	Flower Mushroom Frog 1
chili sauce 1	dungeness crab 1	Fried Pumpkin Dumplings 1
green beans 1	green onion 1	Fried Tofu Curd Balls 1
steamed rice 1	star anise 1	Fuli Roast Chicken 1
stir fried 1	pork belly 1	Hay Wrapped Fragrant Ribs 1
fried fish 1	fish ball 1	Jade Rabbit Sea Cucumber 1
hainanese chicken rice 1	fried egg 1	Lotus Seed Pod Fish 1
duck soup 1	shrimp paste 1	Phoenix Tail Shrimp 1
sticky rice 1	sweet potato 1	Potato Croquet 1
salad bar 1	jasmine rice 1	Silver Fish Fried Egg 1
shaved ice 1	mu shu pork 1	Soy Braised Mandarin Fish 1
soy milk 1	refried beans 1	
chicken soup 1	oyster sauce 1	
noodle soup 1	brown sauce 1	
sea bass 1	green tea 1	
fried dough 1	white rice 1	
ice cream 1	coconut milk 1	
iced tea 1	jasmine tea 1	
brown rice 1	fish sauce 1	
bok choy 1	chop suey 1	
california roll 1	lotus root 1	
shark fin 1	pork ribs 1	
peanut sauce 1	scrambled eggs 1	
egg roll 1	general tso's chicken 1	
general chicken 1	sushi roll 1	
bubble tea 1	wonton strips 1	