Task 3: Dish Recognition

# Task 3.1 Manual Tagging

## Dataset Selected

Chinese cuisine

## Where

You can find the manually tagged dish names at:

<https://github.com/pauldeng/MOOC/blob/master/Data%20Mining%20Capstone/Task%203/Task%203.1/Chinese.label>

## How

1. Remove a false positive non-dish name phrase
2. Change a false negative dish name phrase to a positive label
3. Add dish names extracted from Wikipedia

# Task 3.2 Mining Additional Dish Names

## Summary

## Mining Details

### Manually Tagged

#### Dataset

Chinese.label: manually tagged Chinese cuisine from Task 3.1

#### Result

By submitting the file to the auto grader, I got 5/10.

You should experiment with TopMine and Segphrase tools.

### Mutual Information

#### Dataset

Categories/Chinese.txt

#### How

1. Execute Java –cp textmining.jar functions.MutualInformation categories ChineseCuisine\_MutualInformation.tsv to generate
2. Format the TSV file to CSV
3. Sort the file based on the words co-occurrence
4. Manually remove the header wrong dish names until the first dish name occurred (“dim sum”)
5. Save the top 10000 lines

#### Result

By submitting the file to the auto grader, I got 5/10.

You should experiment with TopMine and Segphrase tools.

### SegPhrase

#### Dataset

Categories/Chinese.txt: Chinese cuisine reviews

Chinese.label: manually tagged Chinese cuisine from Task 3.1

#### How

1. Download and install SegPhrase
2. Place “Chinese.txt” and “Chinese.label” into SegPhrase/data folder
3. Modify the train.sh file as follow:
   1. RAW\_TEXT=’data/Chinese.txt’
   2. DATA\_LABEL=’data/Chinese.label’
4. Execute the train.sh
5. Format the “salient.csv” file
6. Save the top 10000 lines

#### Result

By submitting the file to the auto grader, I got 9/10. Nice job. You beat our baseline with (TopMine) but you can experiment more. You got lower precision than our baseline (Segphrase).

### ToPMine

Categories/Chinese.txt: Chinese cuisine reviews

Chinese.label: manually tagged Chinese cuisine from Task 3.1

#### How

1. Download and install ToPMine
2. Place “Chinese.txt” into topicalPhrases/rawFiles folder
3. Modify the run.sh file as follow:
   1. inputFile=../rawFiles/Chinese.txt’
4. Execute the run.sh
5. Format the “topiclPhrases/output/outputFiles/topPhrases.txt” file
6. Save the top 10000 lines

#### Result

By submitting the file to the auto grader, I got 5/10. You should experiment with TopMine and Segphrase tools.

### Combine

#### Dataset

topPhrases.txt: from ToPMine

salient.csv: from SegPhrase

Chinese.label: manually tagged Chinese cuisine from Task 3.1

#### How

1. Combine them into 1 list
2. Removed duplications
3. Manually removed a common word (“good”) that is not common in Chinese dish names
4. Top 10000 lines

#### Result

By submitting the file to the auto grader, I got 10/10. You should experiment with TopMine and Segphrase tools.