# Task 3: Dish Recognition

Xiaoming Ji (xj9)

# Overview

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 comparison, two approaches: [SegPhrase](https://github.com/shangjingbo1226/SegPhrase) and [AutoPhrase](https://github.com/shangjingbo1226/AutoPhrase) are used to extract dishes from **Chinese** cuisine. SegPhrase and AutoPhrase are algorithms developed by UIUC and achieved SOTA performance on mining quality phrases. SegPhrase needs a small amount of manually tagged samples. While AutoPhrase reduced human effort by leveraging existing general knowledge bases.

We can see both approaches achieved good results but AutoPhrase performs better in term of finding new phrases with less effort.

# Data Preparation

## Cuisine corpus

Extract all (38715) Chinese restaurants related reviews from the whole dataset using script *processYelpRestaurants.py*.

Tagging

To support SegPhrase, we produced 2 label files by modifying the provided *Chinese.label* file*,*

1. *Chinese.updated.label*:
   1. Removed 128 false positive non-dish name phrases. E.g., los angeles, soft serve.
   2. Changed 6 false negative dish name phrases to a positive label. E.g., general tso's chicken, wonton strips.
2. *Chinese.update.plus.label*:
   1. Based on *Chinese.updated.label*, we added 26 new positive dish name phrases. These phrases are mined from AutoPhrase. Note: adding negative samples don’t seem to improve the results.

Dish Name Knowledge Base

AutoPhrase leverages the existing high-quality phrases, as available from Wikipedia or other trusted sources. Since we are mining dish names, we find 132 Chinese dish names from [List of Chinese Dishes](https://en.wikipedia.org/wiki/List_of_Chinese_dishes) and produced 2 knowledge base files,

1. *Chinese\_dish\_quality.txt*: only contains these wiki Chinese dish names.
2. *wiki\_chinese\_dish\_quality.txt*: we append these wiki Chinese dish names to wiki\_quality.txt (found in <AutoPhrase installation>/data/EN)

# Dish Name Mining

SegPhrase

SegPhrase was run with following parameters,

* SUPPORT\_THRESHOLD=10
* DISCARD\_RATIO=0.00
* MAX\_ITERATION=5
* ALPHA=0.85
* WORDNET\_NOUN=0

See <https://github.com/shangjingbo1226/SegPhrase> for explanations on these parameters.

3 label files are used to mine new Chinese dish names.

1. Chinese.label: the provided/original label file. Shown as **Seg.Chinese**
2. *Chinese.updated.label*: Shown as **Seg.ChineseUpdated**
3. *Chinese.update.plus.label*:Shown as **Seg.ChineseUpdated+**

AutoPhrase

AutoPhrase was run with the following parameters,

* ENABLE\_POS\_TAGGING=1, AutoPhrase will utilize the POS tagging in the phrase mining.
* MIN\_SUP=10, a hard threshold of raw frequency for frequent phrase mining.

3 dish name knowledge base files are used to mine new Chinese dish names.

1. *wiki\_quality.txt*: the wiki quality phases found in AutoPhrase installation. Shown as **Auto.Wiki**.
2. *Chinese\_dish\_quality.txt*: Shown as**Auto.ChineseDish**.
3. *wiki\_chinese\_dish\_quality.txt*: Shown as **Auto.WikiChineseDish**.

Results

The top 40 new Chinese dish names mined by each approach are listed in following figure. New means we don’t count the dish names in the label/knowledge base files. Wrong dish names are marked in red. Some phrases are marked in blue and they are food but not Chinese/Asian dish names.

# 

# Conclusion

By checking the results, we can see both SegPhrase and AutoPhrase can find good quality of Chinese dish names. To conclude,

* Dish names are most commonly mentioned phrases in restaurant reviews. Even without specific label/knowledge base and just use phrases from Wiki, we see large portion of phrases mined by **Seg.Chinese** (Chinese.label are generated from wiki phrases) and **Auto.Wiki** are quality Chinese dish names.
* AutoPhrase is very efficient on finding new dish names. No manual tagging is needed. Although the knowledge base of 132 Chinese dish names is small, the results are still very impressive.
* **Auto.ChineseDish** (only use Chinese dish names as the quality phrases) is the most effective approach to find new Chinese dish names. Only 2 errors found compared with 10+ errors as of produced by other approaches. For further improvement, we can increase the size of the knowledge base.
* **Seg.ChineseUpdated+** performs a slightly better than **Seg.ChineseUpdated** in terms of error rates (10 errors vs. 12 errors). It seems adding more positive tagged samples won’t improve performance significantly. In addition, our testing shows adding negative tagged samples won’t improve performance at all.