# 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 by AutoPhrase approach.

Dish Name Knowledge Base

AutoPhrase leverages the existing high-quality phrases, as available from Wikipedia. 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. *Dish.quality.txt*: only contains these wiki Chinese dish names.
2. *Wiki.dish.quality.txt*: we append these wiki Chinese dish names to wiki\_quality.txt (found in <AutoPhrase installation>/data/EN)

# Dish Name Mining

SegPhrase

AutoPhrase

# Discussion

<https://en.wikipedia.org/wiki/List_of_Chinese_dishes>