

Task 3 Overview

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.

Instructions

Before you begin, make sure you have downloaded the data set and any additional tools you wish to use, as described on the [Data Set and Toolkit Acquisition](#) page.

Some questions to consider when building the dish recognizer are the following:

1. What types of dishes are present in the reviews for a cuisine?
2. Are there any surprising dishes in the list you annotated?
3. What types of dishes were you able to find?

You must complete the following specific tasks.

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]. The list can be found in the [manualAnnotationTask.zip](#) file. 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. Here are some actions you may take:

- Remove a false positive non-dish name phrase (recommended), e.g., hong kong 1 could be removed in Chinese cuisine.
- Change a false positive non-dish name phrase to a negative label, e.g., hong kong 1 could be modified as hong kong 0.
- Remove a false negative dish name phrase, e.g., wonton strips 0 could be removed in Chinese cuisine.
- Change a false negative dish name phrase to a positive label (recommended), e.g., wonton strips 0 could be modified as wonton strips 1.
- Add some new annotated phrases in the same format.

Tip: Notice that the character between a phrase and its label is a tab instead of a space.

Remember that the tools we are using were originally designed for general phrase mining instead of dish name mining. Therefore, it will be much safer if we just remove those ambiguous labels, while aggressively changing them into opposites may lead to some undetermined risks, although it is still worth a try.

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.

For example, ToPMine^[1], as we mentioned in the previous pattern mining course, is an unsupervised frequent pattern-based phrase mining algorithm. It merges consecutive words based on statistical significance (stopwords will be firstly removed and be put back later). The most state of the art framework is SegPhrase^[2]. SegPhrase will need the (refined) labels in the first task. SegPhrase has a classifier to assign a quality score to each phrase candidate based on their statistical features. The classification procedure will be enhanced by phrasal segmentation results. These two parts could mutually enhance each other.

Another approach to possibly extending the dish names is using word association. You have previously learned and implemented methods to judge word associations (paradigmatic & syntagmatic relations), such as Mutual Information. There are also some more state-of-the-art methods such as word2vec[3], which you are welcome to experiment with.

Submission

For Task 3, you need to submit a report.

Report in PDF format. We suggest that it be 1-2 pages long. The report will be peer-graded. Your report will need to include the following elements:

- A brief description of what you did, including the pattern mining/word association techniques used, how you revised the labels we provided and how these modifications improved the results (you can try to judge by simply looking at it or show several case studies), the parameters you used, how you applied the models to the specific cuisine, and any interesting findings discovered during your tries. Your description should be detailed enough to allow others to replicate your work.
- Your opinions about whether the results you generated make sense or are useful in any way.