Task 3: Dish Recognition

Help Center

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 will need to submit three different parts.

1. **Annotations file.** Upload the annotations file to the **Task 3.1 part of the Task 3 File Submission** on the Assignments page. Make sure that your submission got accepted by clicking on the feedback button next to the task's name.
2. **Text file.** Upload the text file to the **Task 3.2 part of the Task 3 File Submission** on the Assignments page.The text file should contain at most 10,000 distinct dish names you mined (from any tools or even combinations). We will use some outside knowledge base to calculate the precision in this file and assign scores automatically. The first line in the text file should contain one of the cuisine names we have provided and should be followed by one dish name per line. The valid names for the cuisines are "American," "Chinese," "Indian," "Italian," "Mediterranean," and "Mexican."

   Submit Task 3 Files
3. **Report in PDF format.** We suggest that it is 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.

Submit Task 3 Report

Evaluation

For this task, the file submission parts of the task will be graded automatically, and the report part will be peer-graded. Task 3 is graded out of 30 points total. See the **Task 3 Rubrics** page for a point breakdown.

The evaluation period for the report submissions will begin immediately after the submission deadline. You must evaluate **five** of your peers' report submissions or your own submission score will be penalized by 20%.

Evaluate Task 3 Report

Deadlines

See the Syllabus for detailed information about deadlines for this task.

Getting Help

You can discuss Task 3 with your peers in the Task 3 Discussion forum. We have provided tools in various languages to help you accomplish tasks on the Data Set and Toolkit Acquisition page. You may use these tools or are free to experiment with other ways of accomplishing the task. For information specific to this task, see below.

Useful Resources for Task 3

- **Java:** We have supplied you with a Jar file in task1JavaTools.zip that when given a cuisine file and a list of phrases will compute the Mutual Information for the words in the cuisine file. To learn more about this tool you can find a README accompanied with the tool.
- **ToPMine** (Click on the

Code and Datasets tab on the page to find the ToPMine download.) **SegPhrase**

References

- [1] El-Kishky, Ahmed, et al. "Scalable topical phrase mining from text corpora." *Proceedings of the VLDB Endowment*, 8.3 (2014): 305-316.
- [2] Jialu Liu*, Jingbo Shang*, Chi Wang, Xiang Ren and Jiawei Han, "Mining Quality Phrases from Massive Text Corpora," *Proc. of 2015 ACM SIGMOD Int. Conf. on Management of Data (SIGMOD'15)*, Melbourne, Australia, May 2015. (* equally contributed)
- [3]Goldberg, Yoav, and Omer Levy. "word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method." *arXiv preprint arXiv:1402.3722* (2014).

Created Wed 9 Sep 2015 9:02 AM CDT

Last Modified Mon 14 Sep 2015 10:29 AM CDT