Task-3-Report

Task 3.1: Mannual Tagging

In this section, I manually verify and change those tags belongs to Chinese. I adopt two recommended methods. For false positive phase, I removed all those tags such as san francisco bay, debit card, and san gabriel. Besides, I change those false negative dish name phrase to positive label like wonton soup.

Task 3.2: Mining Additional Dish Names

In this section I use word2vec (models.word2vec) in python to mine top frequent dish names. The word2vec model in genism library is used in this section. I set vector to 100, repeat 20 times. Before creating model, I set bigrams with a min-count 5 and threshold 1.

I calculate the positive distance between the word 'dish' and calculate the bigrams between Chinese categories bi-gram phrase and the word. The results are as listed:

Stir fire 0.39408854

Bell pepper 0.39146996

Cream cauce 0.37390384

Garlic sauce 0.37300736

Clay put 0.34477097

Bean sauce 0.34379196

Thai style 0.33260161

Singapore noodle 0.31919032

Black pepper 0.31781009

Melted mouth 0.30877316

Conclusion

Based on my result, the word2vec could generate a relatively good results in this dish name mining tasks.