# Note

To minimize manual task, I chose Indian cuisine as there are only a few labels to adjust.

# Task 1

As part of task 1, I manually adjusted several tags for the labels provided. According to the instruction, I made changes by only: 1: Remove false positive non-dish labels from list; 2: Change false negative dish names to a positive label.

Changes are included in following list.

|  |  |
| --- | --- |
| Removed non-dish false positive labels | false negative dish labels |
| las vegas | dipping sauce |
| south indian cuisine | tikka masala |
| credit card | coconut chicken |
| south asian | appetizer we |
| san francisco |  |
| mother india |  |
| belly dancing |  |
| strip mall |  |
| south indian |  |
| date night |  |
| bay area |  |
| middle eastern |  |
| belly dancer |  |
| mount everest |  |
| south india |  |
| indian cuisine |  |
| india gate |  |
| that's not |  |
| if there was |  |
| a traditional |  |
| hands down the best |  |
| the reviews on yelp |  |

# Task 2

In this task, I used SegPhrase to generate additional dish names not included in provided labels.

Environment:

|  |
| --- |
| Windows Subsystem for Linux |
| Python 2.7 |
| G++ 4.8 |

Parameters for SegPhrase:

**RAW\_TEXT: 'Indian.txt'** #the Indian cuisine reviews generated from Task1. There are about 30000 paragraphs in total.

**AUTO\_LABEL=1** # I generated labels with both AUTO\_LABEL set to 0 and 1. I will make a comparison in next session.

**WORDNET\_NOUN=0**

**DATA\_LABEL='data/Indian\_update.label'** # the updated labels manually adjusted by me. 140 labels in total.

**KNOWLEDGE\_BASE='data/wiki\_labels\_quality.txt'**

**KNOWLEDGE\_BASE\_LARGE='data/wiki\_labels\_all.txt'**

As there are 5 iterations, A total of 1380 labels generated from SegPhrase, with the maximum phrase length of 6 words.

# Analysis

A snapchat of first 30 phrases generated with AUTO\_LABEL =1 are provided as following:

| **Phrase** | **Quanlity Score** |
| --- | --- |
| fast\_food | 0.99973 |
| main\_course | 0.999165 |
| tandoori\_chicken | 0.998969 |
| south\_indian | 0.998969 |
| indian\_paradise | 0.998969 |
| food\_court | 0.998969 |
| tomato\_soup | 0.99852 |
| tikki\_masala | 0.99852 |
| tika\_masala | 0.99852 |
| strip\_mall | 0.99852 |
| south\_asian | 0.99852 |
| san\_francisco | 0.99852 |
| salad\_bar | 0.99852 |
| rice\_pudding | 0.99852 |
| palak\_paneer | 0.99852 |
| nicely\_decorated | 0.99852 |
| mt\_everest | 0.99852 |
| middle\_eastern | 0.99852 |
| mango\_custard | 0.99852 |
| main\_courses | 0.99852 |
| left\_overs | 0.99852 |
| las\_vegas | 0.99852 |
| la\_carte | 0.99852 |
| indian\_cuisine | 0.99852 |
| iced\_tea | 0.99852 |
| ice\_cream | 0.99852 |
| guru\_palace | 0.99852 |
| gulab\_jamun | 0.99852 |
| gluten\_free | 0.99852 |
| family\_owned | 0.99852 |

Compare the provided label and labels generated from SegPhrase with AUTO\_LABEL enabled, the latter have more labels generated, although there are still many labels not related to dish names.

To make a comparison , I also generated labels from existing labels tagged in Task1. And The top most 30 phrases are as following:

|  |  |
| --- | --- |
| **Phrase** | **Quanlity Score** |
| food\_court | 0.999792746 |
| fast\_food | 0.999047619 |
| chicken\_tikka\_masala | 0.999017682 |
| white\_rice | 0.99893979 |
| tandoori\_chicken | 0.99893979 |
| hot\_sauce | 0.998677546 |
| veggie\_dishes | 0.998470292 |
| chicken\_wings | 0.998164725 |
| gluten\_free | 0.99808931 |
| tikka\_masala\_fries | 0.997957471 |
| tikka\_masala | 0.997957471 |
| rice\_pudding | 0.997957471 |
| poor\_service | 0.997957471 |
| mt\_everest | 0.997957471 |
| masala\_dosa | 0.997957471 |
| goat\_curry | 0.997957471 |
| garlic\_naan | 0.997957471 |
| flat\_bread | 0.997957471 |
| dipping\_sauce | 0.997957471 |
| chicken\_tikka\_marsala | 0.997957471 |
| chicken\_tikka | 0.997957471 |
| chicken\_tiki\_masala | 0.997957471 |
| basmati\_rice | 0.997957471 |
| weekend\_buffet | 0.997934753 |
| ice\_cream | 0.997882057 |
| gulab\_jamun | 0.997882057 |
| tomato\_soup | 0.997845292 |
| chinese\_food | 0.997520019 |
| street\_food | 0.997499825 |
| tomato\_sauce | 0.997332471 |

In the labels generated from existing tagging, more dish names are generated and tend to have a higher quality score. To make a clearer visualization of all labels, see following table:

| **AUTO\_LABEL enabled** | | | **AUTO\_LABEL disabled** | | |
| --- | --- | --- | --- | --- | --- |
| Label | **Quality Score** | **Is dish name** | **Label** | **Quality Score** | **Is dish name** |
| fast\_food | 0.99972973 |  | food\_court | 0.999792746 |  |
| main\_course | 0.999165232 |  | fast\_food | 0.999047619 |  |
| tandoori\_chicken | 0.998968987 | 1 | chicken\_tikka\_masala | 0.999017682 | 1 |
| south\_indian | 0.998968987 |  | white\_rice | 0.99893979 | 1 |
| indian\_paradise | 0.998968987 |  | tandoori\_chicken | 0.99893979 | 1 |
| food\_court | 0.998968987 |  | hot\_sauce | 0.998677546 | 1 |
| tomato\_soup | 0.99852007 | 1 | veggie\_dishes | 0.998470292 | 1 |
| tikki\_masala | 0.99852007 | 1 | chicken\_wings | 0.998164725 | 1 |
| tika\_masala | 0.99852007 | 1 | gluten\_free | 0.99808931 | 1 |
| strip\_mall | 0.99852007 |  | tikka\_masala\_fries | 0.997957471 | 1 |
| south\_asian | 0.99852007 |  | tikka\_masala | 0.997957471 | 1 |
| san\_francisco | 0.99852007 |  | rice\_pudding | 0.997957471 | 1 |
| salad\_bar | 0.99852007 |  | poor\_service | 0.997957471 |  |
| rice\_pudding | 0.99852007 | 1 | mt\_everest | 0.997957471 | 1 |
| palak\_paneer | 0.99852007 | 1 | masala\_dosa | 0.997957471 | 1 |
| nicely\_decorated | 0.99852007 |  | goat\_curry | 0.997957471 | 1 |
| mt\_everest | 0.99852007 |  | garlic\_naan | 0.997957471 | 1 |
| middle\_eastern | 0.99852007 |  | flat\_bread | 0.997957471 | 1 |
| mango\_custard | 0.99852007 | 1 | dipping\_sauce | 0.997957471 | 1 |
| main\_courses | 0.99852007 | 1 | chicken\_tikka\_marsala | 0.997957471 | 1 |
| left\_overs | 0.99852007 |  | chicken\_tikka | 0.997957471 | 1 |
| las\_vegas | 0.99852007 |  | chicken\_tiki\_masala | 0.997957471 | 1 |
| la\_carte | 0.99852007 |  | basmati\_rice | 0.997957471 | 1 |
| indian\_cuisine | 0.99852007 |  | weekend\_buffet | 0.997934753 |  |
| iced\_tea | 0.99852007 | 1 | ice\_cream | 0.997882057 | 1 |
| ice\_cream | 0.99852007 | 1 | gulab\_jamun | 0.997882057 | 1 |
| guru\_palace | 0.99852007 |  | tomato\_soup | 0.997845292 | 1 |
| gulab\_jamun | 0.99852007 | 1 | chinese\_food | 0.997520019 | 1 |
| gluten\_free | 0.99852007 |  | street\_food | 0.997499825 |  |
| family\_owned | 0.99852007 |  | tomato\_sauce | 0.997332471 | 1 |
|  | **Total** | **11** |  | **Total** | **25** |

So SegPhrase generated labels are more valuable with existing refined labels.