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

Phrase	Quanlity Score
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

AUTO_LABEL enabled			**AUTO_LABEL disabled**		
Label	**Quality Score**	**Is dish name**	**Label**	**Quality Score**	**Is dish name**
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.