Task 3: Dish Recognition

Brief description

I completed the task using provided tools SegPhrase, TopMine and some subroutine tools implemented in iPhython notebook. I chosen for the task ‘American (New)’ cuisine.

Step to get the goal:

1. I reviewed provided label file simply looking at it line, by lines. I had concerns about some names and used Google to make sure that I didn’t miss any dishes. There was the question what exactly we should understand by ‘dishes’, especially it was critical for sides (brown rice, for instance), ingredients (green beans) and beverages (draft beer, for example). For myself I stopped that beverages are definitely not food, sides I considered as dishes and ingredients I was guided by principle whether it can be order separately or not. I’ve added about 20 names that should be extremely well-know in American cuisine.
2. The next step was to prepare the raw text of reviews for provided tools. I used python and Apache Spark SQL (http://spark.apache.org/sql/) to collect all reviews for business from the category ‘American (New)’. The total number of reviews was **94183**.
3. The first run of SegPhrase with updated labels and mentioned above set of reviews gave me the score 9.0/10.0 and nice message “Nice job. You beat our baseline with (TopMine) but you can experiment more. You got lower precison than our baseline (Segphrase)”. After that I had a long way to get the last one points.
4. I had an idea to extend the first set of labels from an external source. A quite obvious source seemed to me Wikipedia. I found a couple of articles related to American cuisine:
    a. https://en.wikipedia.org/wiki/List_of_hamburgers
    b. https://en.wikipedia.org/wiki/List_of_sandwiches
    c. https://en.wikipedia.org/wiki/List_of_American_breads
    d. https://en.wikipedia.org/wiki/List_of_American_cheeses

    With help of python and PyQuery I managed to parse html and extract there a list of dishes. I tried the extended list of labels for SegPhrase but it doesn’t give any improvement in score.
5. I analyzed the result of SegPhrase and found out that there is a lot of phrases with ‘the’ and some names with ‘a’ and ‘an’. I added additional to trim prepositions from output. Again no change in my score.
6. The next turn was to try to use TopMine. I used the default configuration of TopMine since it seemed to me quite reasonable. TopMine found quite many good phrases that are definitely name of dishes. Furthermore, I noticed that almost all names were in first line with frequency higher 200. It can be explained because name of dishes this is stable set of several words and since we focused only on reviews of restaurants and specially on reviews

for American (New) cuisine, so we will see that names should have high frequency. I scanned manually all these lines and added all new found dishes to labels file and run again SegPrases. It didn't give me desired one point to my score.

7. Then I examined again the result of SegPrase to figure out what exactly is wrong in my output file. Even though there were a lot of new dishes (that I personally didn't know and I had to check them in Google) it contained still a lot of phrases related to restaurants topic but that are not dishes, greatly lot were phrases with word 'night', a day of week, for example Sunday brunch, phrases with the words 'bar', 'beer', 'area', 'menu', 'shop' and 'room'. I decided to add all these words to the stop words list to ignore at all. It doesn't give me any change my score.
8. And then I remembered that I've added removing of preposition from names and I was wondering what if I've disabled it for a while. All of sudden it gave the the last point. That's great to get the target but this change really confused me. Based on that I made the conclusion that the secret test expects to see names with the what is in my opinion is not fully correct.
9. I tried to improve the result using Word2Vec, particularly I relied on scalable implementation of Word2Vec in Apache Spark MLib (https://spark.apache.org/docs/latest/mllib-feature-extraction.html#word2vec). This is quote interesting tool which has quite high precision in my opinion for collected reviews. For example, the synonymous for the word benedict (in the meaning eggs Benedict) were:
    a. scrambled
    b. eggs
    c. benny
    d. poached
10. This result is quite interesting but as for me it is quite hard to use it for expanding of name of dishes since they are set of words, but not one word. Only one idea crossed my mind to find the closest word to 'sandwich' and replace by it all sandwich names to generate a new set of potential names. The most closes synonym was Panini and such way I've added about 170 new dishes (maybe absolutely new and not know for any one else) instead of obviously incorrect in the result file. No surprise it didn't give my any more points in my score.

All intermediate steps can be found in the notebook ipothon in my dropbox:
https://www.dropbox.com/s/ocnhqosp1slq5ev/Capstone%20project%203.html?dl=0

My opinions

As I mentioned above, the score and how I reached it put me in mixed feelings, but in spite of this I find that indeed the obtained result makes sense. I said before that SegPhrase managed to extract from text name of dishes that I didn't know at all and I had to use Google to check them. From other hand I catched myself on thought that I do know many dishes but they didn't cross my mind when Iwas adding it to labales file on the first step. For example, 'hot dog', it is absolutely obvious, but I didn't write. One more interesting thing was that in the original file with labels there were only

one dessert – ‘panna cotta’, but SegPhrase found many other desserts, like ‘red velvet cake’, ‘fig and pecan pie’ and dozen other pies.