Homework 3 (due 9/23 at 9AM)

In this assignment, you will write two (plus extra cerdit one) MapReduce programs and make it works on yelp reviews dataset similar to this week's Lab practice.

<The data scientist's challenge>

You are a new hire in the data science team at Yope, an interent company that provide business yellow page information and user reviews. After first week's training, you have enjoyed this relaxing atmosphere and all-you-can-drink beer fountain at work. Now your boss comes to you with your first several missions.

[1] The most wired review

The Mission:
Your boss asks you to search the company database for the online review that is most unique (Maybe they are random generated content and need to be removed). You decide to use mapreduce to find the review with the most words that are not used in any other review.

Instruction:
In this part, you will follow the code instruction in file <unique_word_template.py>. Finish each TODO section and make it run to output the review id in which contains the most unique words. (Those words only appear in one review, not appear in others)

Sample Output:

```
"WXshgoreBsq124bnadiu3nq" 120
```

[2] Duplicate user detection

The Mission:
Your company has a problem with users trying to cheat the system by setting up multiple accounts to make reviews. Your boss asks you to find a way to detect when two accounts actually belong to the same underlying person. Thus, you theorize that two accounts with reviews that always rate similar business will belong to the same person (Some company got paid to write reviews for business). You want to write a program to flag those accounts

Instruction:
In this part, you will follow the code instruction in file <user_similarity_template.py>. Finish each TODO section and make it run to output pairs of user ids <user_id_1, user_id_2> and its Jaccard score if the Jaccard similarity >= 0.5. Please reference Jaccard similarity definition on wikipedia (http://en.wikipedia.org/wiki/Jaccard_index) or coursera (https://class.coursera.org/nlp/lecture/184)

Sample Output:

```
["xx2j7XRWLFN7QavS5jcbFw", "yZEz7ZHDt9MU7ehH9ecCQA"] 0.5
["xyuqtuV71y2F3uD9z97jPw", "zvfGidbeZn9A_kkRWK6JLw"] 0.6
```

(Extra-credit Mission) [3] Duplicate user detection - A more accurate model

The Mission:
After presenting to your boss about the result of mission [2], both of you discovered the fact that not every user pair with high business similarity can be identified as same user (Maybe they just live in same area with similar habit to go the same restaurant). However, you also noticed the fact that a user tend to use similar word set for reviewing. So you come up another approach to identify duplicate account: If two users use similar words in their reviews (Jaccard similarity > 0.5), the two user id are suspected owned by one person. You want to write a program to output those

accounts.

Instruction:

In this part, there is no code template. Your mission is to create a program that calculate the word set similarity by Jaccard function, and output those user pairs whose Jaccard similarity >= 0.5. For example: User1 use word set (W1, W2, W3) in all reviews, User2 use word set (W1, W2, W4) in all reviews. The Jaccard Similarity = len ([W1,W2])/ len ([W1,W2,W3,W4]) = 2/4 = 0.5. Your program should output <user_id_1, user_id_2> and its Jaccard Score.

Sample Output:

```
["yLYR5tt0_mRvDwVmkZ7Frw", "zXJ0IhDQzpJXQ8mPdCqbsg"] 0.6
["yM0Bv8lqQrx9Yo0_1rtqxg", "ypqCBI5rMe_ecGE1iZpQaQ"] 0.9
```

<How to submit your homework>

Please extensively test your assignment. When it is complete, put it in a script named hw3_part1.<lastname>.py and hw3_part2.<lastname>.py (and hw3_part3.<lastname>.py if you want to earn extra credit). Upload these files using the file upload tool available athttps://www.ischool.berkeley.edu/uploader/?s=i206 Login with your I School userid and password and follow the directions there. If you wish to modify a file you have already submitted, you may do so by resubmitting it using the same tool. As long as it has the same filename, it will overwrite the existing file. We will mark the last file(s) submitted before the due date/time.

<How to run your program>

PS1. Please download the code templates and dataset here:https://www.dropbox.com/s/vn70pgv2sjcl9uh/I206HW3.zip?dl=0

PS2. Please make sure you have installed mrjob library on your python environment. Detail instruction can be found here: https://github.com/Yelp/mrjob

1) Go to mrjob github webpage: https://github.com/Yelp/mrjob
2) At right panel, click "download zip" button
3) Unzip the file, run command line console, go to the folder you unzipped.
4) type "python setup.py install" and finish the installation. (if authorization issues happen, try "sudo python setup.py install" instead)

PS3. To compile your program and make it run on your local environment, please use following command: python <your python file> yelp_academic_dataset_review.json