README

Derek Nelson
Brian Oh

Original data in "data" folder

user.json
yelp_academic_dataset_business.json
yelp_academic_dataset_checkin.json
yelp_academic_dataset_review.json
yelp_academic_dataset_tip.json

Modification to the data

We used "getUserData.py" to generate of Transition matrix from user data.
-> generated "Iteration_userID", "TransitionMatrixUnmodified"
We used "user_matrix_modification.py" to add farmers and useful vote.
-> generated "Iteration_userIDUnmodified", "TransitionMatrix"
We used "reviewModification.py" to add fake reviews.
-> modified "yelp_academic_dataset_review.json"

User weight for review rank generation

All needed files are in elephant in "/user/u0343930/YelpProject" folder.

We used "getUserWeightUnmodified.py" to generate user weight without spam farm architecture for review rank. To run the program enter the following command line:
nohup spark-submit --driver-memory 8g --executor-memory 8g getUserWeightUnmodified.py &
-> "uwUnModified" folder contains rdd object of user weight

We used "getUserWeightModified.py" to generate user weight with spam farm architecture for review rank. To run the program enter the following command line:
nohup spark-submit --driver-memory 8g --executor-memory 8g getUserWeightModified*.py &*
-> "uwModified" folder contains rdd object of user weight

*nohup allows user to disconnect from cluster, and it finishes all calculation in background.

Data Calculation

Modifying Review Data
This script needs to be run twice once with MainFarmer = 99 and other farmers = 1
The second time needs to be MainFarmer and all other farmers = 0
NOTE: You will need to copy the original data first and also specify the file name as well.

$ python reviewModification.py

If you want to have user weight of one you will need to create data set of (userID, 1)
To do this you can run the extractData with the commented code in main uncommented.

UserData Needs to be processed with PageRank Method before this
Location can be set at the top of this script, or specified here.
#userData = sc.textFile(user_data) \
.map(lambda x: clean_user(x)) \
.saveAsTextFile("/home/derekn/CS6965/yelp_dataset_challenge_academic_dataset/user_data")
#sys.exit(1)

Then save as a text file so you can add farmer information.
Run the script userModify.py to add the farmer informaion to the text file.
To modify the user JSON data use userModifyJSON.py
NOTE: You will need to copy the original data first and also specify the file name as well.

$ python userModify.py
$ python userModifyJSON.py

To run the calculations we first need the user weight made.
Then we can run the following scripts.
NOTE: These scripts output a vector (key, value) for each combination of the data.

$ python extractData.py
$ python simple.py
$ python other.py

They also output statistical data to a file. The files are:
output_extractData.out
output_simple.out
output_other.out

extractData uses our ranking algorithm with time depreciation and advantage rate
simple is our ranking algorithm without time depreciation and advantage rate
other is averaging and useful weight

If you want to just run one or two there is a variable "iteration" that you can modify to select the code you want. THis iteration variable allows you to select the desired initial setting. You can always assert the "doneFlag" to be true to stop iterations in the while loop.

When all these scripts finish then you can run this script.
$ python search.py

This will also output to a file output_search.out which summarizes the results.
NOTE: It displays a result for every possible test.

Again if you want to modify the search just the same as the other three scripts you can do so the same way as above.