Assignment No 1

CS-4080 Mining Massive Datasets

Problem and Dataset Introduction

This dataset consists of 'circles' (or 'friends lists') from Facebook. The dataset includes node features (profiles), circles, and ego networks. In addition to the files, the paper titled “Learning to Discover Social Circles in Ego Networks” is attached. This paper uses the same dataset to discover the circles of the given node (you are not required to do this). This paper gives you a good understanding of the problem that you are going to solve. Especially the concepts of the “circle”, “overlap” and so on. Please refer to Figure 1 and Figure 2 of the paper. Figure 2 explains the feature set used in the file ‘x.featnames’. Following quotation from the paper helps you understand the problem and dataset.

> *Our personal social networks are big and cluttered, and currently there is no good way to organize them. Social networking sites allow users to manually categorize their friends into social circles (e.g. ‘circles’ on Google+, and ‘lists’ on Facebook and Twitter), however they are laborious to construct and must be updated when-ever a user’s network grows.*

Facebook data has been anonymized by replacing the Facebook-internal ids for each user with a new value. Also, while feature vectors from this dataset have been provided, the interpretation of those features has been obscured. For instance, where the original dataset may have contained a feature "political=Democratic Party", the new data would simply contain "political=anonymized feature 1". Thus, using the anonymized data it is possible to determine whether two users have the same political affiliations, but not what their individual political affiliations represent.

File	Description
facebook.tar.gz	Facebook data (10 networks, anonymized)
readme-Ego.txt	Description of files

Your Task

You are required to write and “map.py” and “reduce.py” files to compute the following:

1. The most popular person among the 10 Facebook users. (user with biggest circle)
2. The user with most similarity with the user identified in part 1 (user with the biggest circle overlap with the given user)
3. [BONUS] Identify the most common feature(s) on which user in the part 1 is identified. (Common features of the biggest circle).

Submission

You are required to submit a zipped archive with the <*assignment1_rollno.zip*>. The zip archive must contain the following:

- map.py
- reduce.py
- reduce output
- screenshot of the Hadoop cluster output
- one page word document to comment on your findings.