Assignment 4

The deadline for the fourth assignment is June 21st, before the end of the day. Turn in the code, with instructions on how to run it. The report should include detailed observations on the results. For late submissions the late policy on the page of the course will be applied. Details for the turn-in, and how to write reports are on the Assignments web page of the course. The will be an oral examination of the Assignment in the week following the assignment submission.

Question 1

In class we proved that for the MAX-COVERAGE problem the Greedy algorithm that iteratively selects the set that covers the most items that have not already been covered, has approximation ratio $\left(1-\frac{1}{e}\right)$. Consider a variation of the MAX-COVERAGE problem, where the collection of sets is partitioned in K categories, and each subset belongs to exactly one category. We want to select K sets, where we have exactly one set from each category, so as to maximize the number of covered elements.

Prove that in this case the approximation ratio of the Greedy algorithm is $\frac{1}{2}$. Construct an input in which the Greedy algorithm achieves this approximation ratio.

Question 2 (Classification)

In this question you will practice with algorithms for classification.

You will use the file "clinton_trump_tweets.txt" that you used for Assignment 3. The goal is to create a classification model that predicts if a user is a follower of Trump or Clinton.

This question has two parts. In the first part you will use the data (together with their classes in the file "clinton_trump_user_classes.txt") both for training and testing. Preprocess the data, and remove all users that have less than 10 tweets (of any form). For the remaining users, use all available information in the file to extract features that you consider useful for classification. Use your imagination for the feature extraction from the full information you have about the users (you can also use the python tools we described in the tutorials). Train four classifiers that we saw in class (Decision Trees, SVM, Logistic Regression, k-NN). Use 5-fold cross validation to evaluate the classifiers. In your report describe the features that you used and report the accuracy of the classifiers on the 5-fold cross validation.

In the second part of the question, you will use the classifier you built in the first part of the question to do the same prediction for a new set of users that are not part of the data that you have seen. To this end, a competition was created in Kaggle for the course (here is the link to the competition). Create an account with your university email. You will be given access to the competition of the course and you will be able to submit a solution to the competition. There is a ranking where you can see your position with respect to the other

solutions. A good position will boost your grade for the question. More importantly, your classifier should achieve a good accuracy score with respect to the other solutions. Use any model you prefer. Add to your report your results in the Kaggle competition.

In the report, you can also discuss about experiments that you did and did not work out, or how you improved upon a solution that did not work well (add also the numerical results).

Question 3 (Network Analysis)

In this question you will practice with algorithms for network analysis.

You will use the file "clinton_trump_tweets.txt" that you used for the previous question. Using this file you will create a graph of who retweets whom. Create the graph by adding an edge (xxx,yyy) if the user with screen name xxx retweeted a message from user with handle @yyy. For example, the line

```
Saint   Saint2205       1537088066      70      133                     Fri Oct 28 20:27:12
EEST 2016   792055126408048640      en      null    27      0       RT @greeneyes0084:
Wikileaks Email: Hillary Campaign Struggles to Reach F**king Dumb Young People
https://t.co/S6QZhY9rBH via @realalexjo
```

in the file will result in the creation of the edge (saint2205,greeneyes0084) in the graph.

Create this graph, and then remove all nodes that are not in the file (that is, nodes that they have not tweeted anything). Also, iteratively prune the nodes with degree less than 10. Take the subgraph consisting of the largest connected component of this graph. This is the graph you will work with.

Our goal is to study this graph. First, compute the fraction of Trump and Clinton followers in the graph. Then, run the PageRank and HITS algorithms and report the first 10 nodes for each algorithm. Study the accounts of these nodes, and try to explain why these nodes may be important in the network.

Second, we will try to separate the followers of Trump and Clinton using algorithms for community detection. You will implement the algorithm of Girvan-Newman, where we iteratively remove the edge with the largest betweenness centrality to find two communities. Run one iteration of the algorithm and compute the time it takes. How much would it require to remove 100, 500, or 1000 edges? Modify the algorithm to remove edges in batches of size K (each time the edges with the highest betweenness), and try for different values of K (20, 50, 100, 250, 500). Report the running time of the algorithm, the confusion matrix, and the success with which the algorithm manages to separate the followers of Trump and Clinton.

Finally, run the PageRank and HITS algorithms for the subgraphs of each partition you created, and report again the top-10 nodes for each subgraph. Discuss the nodes and compare with the previous results.

For this question it will be helpful to use the networkx library and the built-in functions for computing the subgraphs of the connected components, betweenness centrality, PageRank and HITS.

Question 4

The goal of this question is to explore if there is a cultural gap between the followers of Trump and Clinton, and to try to characterize it. We know that the two groups differ in their political beliefs, and there are certain hashtags that are very characteristic in separating the two (e.g., the #maga hashtags for Trump followers, and #imwithher for Clinton followers). Beyond politics, can we use the tweets to understand the cultural profile of the average follower? For example, is it the case that the Trump followers tweet more about Nascar or baseball, while Clinton followers tweet more about theater and city life?

This is an open-ended question in which you are asked to improvise. You can use any of the techniques we saw in class, and you can preprocess and modify the data as you see fit. Propose and define your approach to the problem, implement it, and report your results. Your approach and formulation of the problem will count heavily in your final grade.