Data mining and network analysis ITI8370

Home assignment 3

Text Data Mining

Submission deadline 12.12.21. 23:59

General requirements:

- No plagiarism in any form. Please cite all the sources you used.
- Prepare your solution in such a way that it may be executed on any computer with R-studio.
- Prepare a short write-up with the analysis of achieved results. Maximum 2 pages 12pt. PDF format only, submit by TalTech Moodle, strictly no e-mail submissions.
- Submit your code by means of https://gitlab.cs.ttu.ee, provide the lecturer and teaching assistant (dmgolo@taltech.ee) with the developer access for your project.
- During the home assignment defense you will have to demonstrate your solution and will be asked few questions. Note it is mandatory to attend / participate in the defense of the home works.
- If you are unsure about using some third-party function contact your teacher.
- **All the exercises are mandatory except exercise 3.**
- Assignments are accepted up to one week after the deadline with the penalty of 10% for each day except Saturday and Sunday.
- NB! Students should be able to demonstrate that their implementations are able to perform.
- File naming convention: HA_3_ Name_Surname.pdf
- R codes and functions naming convention. Distance, feature selection and silhouette functions: {student_initials}_nameofthefunction.R. Main codes {student_initials}_ex_{number}_nameofthecode.R. Please avoid using capital letters.
- Please indicate (using bold letters in the beginning of your report) if you are willing to present your work in the class or online.
- The initial plan is to have defense in the hybrid mode.
- The date of the defense 14.12.2021 lecture time.

The following conditions should be satisfied:

The students are expected to demonstrate suitability and goodness of their solutions without guidance on behalf of the lecturer.

Exercise 1. NB! Points a and b are mandatory!!!

For the text data set of your choice:

a. Depict centroids of the clusters
b. Depict decision boundary between different classes.

NB! Do not reproduce the example shown during the practice!

Exercise 2.

Implement your own functions in R to compute:

1. Local clustering coefficient.
2. Degree centrality.
3. Degree prestige.
4. Gregariousness of a node.
5. Closeness centrality and proximity prestige.
6. Betweenness Centrality
7. Common neighbor based measure.
8. Jaccard Measure.

Verification of your code will be done using the files: Dataset1-Media-Example-NODES.csv and Dataset1-Media-Example-EDGES.csv from https://kateto.net/networks-r-igraph You are allowed to build your own networks and graphs to verify/debug and demonstrate your solution! Please use write up to explain the teacher how to use your implementation and what are the limitations if any.

Exercise 3. This exercise does not give any points to the present home assignment! It is proposed to the students willing to increase the grade of the home assignment 1 or home assignment 2. And gives at most 5 points which will be added to the grade of chosen home assignment.

Program in R your own implementation of the reservoir sampling method and demonstrate how it works.

Good luck!