|  |  |  |  |
| --- | --- | --- | --- |
| **Module** | **ITC 4541 – Web Science (Level 6)** | | |
| **Term** | **Spring Semester 2019** | | |
| **Assessment** | **Project** | **Weight** | **60%** |
| **Relevant**  **Learning Outcomes** | *LO1. Adapt or combine network measures to construct social or generalised information network models*  *LO2. Design methods to mine the structural and content information in social or generalised information network models*  *LO3. Formulate techniques that are based on structural or content information to build recommender systems or systems that extract higher level modalities* | | |
| **Duration** | *(N/A)* | | |
| **Deliverables** | *Project Report and Code* | | |
| **Method of Submission** | * *The project report will be submitted to TurnitIn* * *The project code will be submitted to the course assignment point* * *A hard copy of the report will be given to the instructor* | | |
| **Deadline:** | *Late assignments will be penalized by 10% of their original grade for each day the assignment is late after the deadline. The maximum penalty for late submissions fewer than 7 days will be a grade of 40. Assignments submitted 7 or more days after the deadline will receive a mark of zero.* | | |



## General Instructions:

* You may not use code or solutions from any source (i.e. originating from colleagues, internet etc.). You may consult various sources to obtain the state of the art or interesting ideas.
* The report that you will compose must be self-contained: description of the problem to be solved, algorithms used, observations/results, conclusions.
* For every question you need to reach some conclusions or make some observations.
* Presentation matters: try to use diagrammed, tables, references to tables and diagrams good formatting

Your task is to analyse a trending subject. It can be about politics, the economy, life style etc. The subject will be agreed between you and the instructor, and it will be specified with some keywords. Then you will collect data from Twitter, while the number of Tweets must be in the order of 100.000. Fewer tweets may hinder your analysis.

**Q1 Influence-35%**

The first part of your analysis is to build a network that is made of user to user contacts.

1. (15%) Build a user to user network and save it in file: ‘network-lastname.data’. Then investigate whether the user to user network follows a power-law distribution.
2. (10%) The second part of your analysis is to rank the users based on their centrality on the user to user network. Use the following centrality measures: degree centrality, page rank, and betweenness. Save the results in descending order of centrality in csv files: ‘degree.data’, ‘pageRank.data’ and ‘betweeness.data’ and compare. Each list should contain the user id, and the user screen name.
3. (10%) Present the top-10 influential users for each category and characterise them. The characterisation should be related to the content of their posts. A suggestion would be to use tags for each user e.g. user1, username:[tag1, tag2, …]. The list should not be too long or too short.

**Q2 Communities of users-30%**

You are to discovery communities of users, i.e. users that tend to cluster often together.

1. (10%) Discover communities on the user to user network.
2. (10%) Characterised each community based on small number of tags, and present the top-10 communities
3. (10%) Does the size communities of users follow a power law distribution?

**Q3 Recommender system-35%**

Design and build a collaborative filtering system to suggest to a user other user to connect to. The phases of your design are the following:

1. (10%) User to user pairwise topic similarities. The results are to be stored in a csv-file: “pair-similarities.data”
2. (10%) K-most similar users to each user. The results are to be stored in file: “k-neighbours.data”
3. (15%) form and evaluate the recommendation