17CS651/ 15CS651 - Data Mining and Data Warehousing – Assignment 2020

Project to be completed in groups of 3-4: Explore any data set that is available online to perform association rule mining, classification or clustering using one method that has been taught in class and another method (either one that has been taught in class or any other of your choice) and submit (i) A 4-6 page report (details of the format and content given below), (ii) a link to Github with the team's code neatly documented and a (iii) url to a 2 min youtube video (details of the format and content given below)

Additionally, those who are absent for any test or score <= 15/30 (or <=10/20) are required to submit the solutions (without choice) for that test along with the class project and video.

Timeline for the class project:

Week 6 (Mar 02-08): [Problem statement] Form a team, choose a team name, explore data sets available online, decide on a problem statement (complete section 1 of the report)

Week 7 (Mar 09-15): [Literature review] Go through Google Scholar on what others have done to solve the problem (complete section 2 of the report)

Week 8 (Mar 16-22): [Stocktaking of data] Explore the data set; how many rows? How many columns? What are the features available? What are the other questions we can ask on this data? What code is available for testing on Github? (complete section 3 of the report)

Week 9 (Mar 23-29): [Preparation for T2]

Week 10 (Apr 02-05): [Phase 1 coding] Run any available code on the data/ run library functions on the data; document the results

Week 11 (Apr 06-12): [Stocktaking of results] Can we tune the parameters to achieve better results? What is working? For what cases is the code not working and why?

Week 12 (Apr 13-19) [Phase 2 coding] Implementing changes to the existing code; selecting another algorithm to implement; coding from scratch or using library functions

Week 13 (Apr 20-26) [Phase 2 coding contd] What else can we do better? Parameter turning, testing any other method that is suitable for the data/ problem and documenting the results

Week 14 (Apr 27-May 1) [Documentation] Complete documenting the code, complete the report with sections on experiments and discussions and record and upload the youtube video

Final submission of all components: May 2, 2020

What goes into the 4-6 page report?

- [0.5-1 page] Introduction and background – what is the problem area? Why is it important? What is the specific problem you seek to solve?
- [0.5 -1 page] Previous work – A brief review of only the most relevant predecessor work; what limitations have you identified that you seek to address in your work? What are the assumptions you have made about the data/ problem area or the scope of the problem you seek to solve? What is the code available online (Github, etc.) for the problem?
- [1.5 - 2 pages] Proposed solution – an overview of the various components of your solution (preprocessing + building a model + evaluation) This is to be followed by a detailed explanation of each component and what/ how you have implemented
- [1-1.5 pages] Experimental results and comparison of results + tuning of parameters
- [0.5-1 page] Discussion - detailed explanation of all the insights you have gained into the data (on what cases does the model work well? When does it fail? Is there something that can be done to fix the problems?)
- [0.5 page] Conclusions
- [0.5 page] Contributions – who did what? What did you learn through this project? (Any interesting insights besides those mentioned under the detailed discussion)
- References

What goes into the 2 min Youtube movie?

- [30 sec] Introduce the members of the team
- [30 sec] What problem have you selected and what data set are you using to solve the problem?
- [30 sec] Why is this problem useful?
- [30 sec] What is the approach you have taken/ tools used

[1 min] Anything interesting that you inferred about the data or learnt through the process? Also, the specific role of each member of the team.

Where can we start to look for data/ problems?

- KDNuggets http://www.kdnuggets.com/datasets/index.html
- Kaggle: https://www.kaggle.com/datasets
- UCI Machine Learning Repository https://archive.ics.uci.edu/ml/index.php
- ML Data - http://mldata.org/
- Government data: **https://data.gov.in/**
- Competitions for social good: **https://www.drivendata.org/competitions/**
- Crowd Analytix: **https://www.crowdanalytix.com/community**
- Crowd AI: **https://www.crowdai.org/**