#### **Data Mining (DS-239)**

#### **Project Topics**

You need to pick a project topic of your choice where you can apply Data Mining algorithms/techniques. Usually we can choose any application-oriented project of your own interests, this can range from computer science to biology.

Once you identify the project, it is important to explore more about the data availability and related work done. Most of projects need multiple data sources to meet required target, for this you either need to acquire the data through APIs or through crawling web pages, otherwise you need to rely on data repositories like Kaggle, NIPS, UCI, Crawdad etc. As data crawling or acquisition may take considerable time and need good pre-processing, time management is key.

#### **Some quick notes on specific types of projects:**

* + 1. **Preprocessed datasets:** Even if you are not crawling or using API for data collection, and using pre-processed datasets, still you need to carry out data cleaning and visualization techniques. This will help you get familiar with the problem and achieve the target.
    2. **Mirror or replication of existing project:** Even if you are replicating existing projects or doing mirror tasks, we expect you to apply multiple models/techniques to achieve target. There should be considerable comparison debate offered at the end of your project to compare different techniques and benefits over each other in term of accuracy metrices we studied.

#### **Project Parts: Proposal, Milestone, Poster, & Final Report**

**Submission**: We’ll be using KATIE for submission of all four parts of the final project. We’ll announce when submissions are open for each part. You should submit on KATIE as a group: that is, for each part, please make one submission for your entire project group and tag your team members.

#### **Evaluation**

We will not be disclosing the breakdown of the 40% that the final project is worth amongst the different parts, but the midterm (presentation, demo and report) and final (Final report, Poster, Demo and Presentation) will combine the majority of the grade. Projects will be evaluated based on:

* + 1. The technical quality of the work. (i.e., Does the technical material make sense? Is data make sense to achieve target? Is data quality ensured? Are required attributes made available for model building? Are the things tried reasonable? Are the proposed algorithms or applications clever and interesting? Is model accuracy calculated effectively? Are different models/techniques compared well enough to materialize the results?)
    2. Significance. Is target significantly achieved and make sense?
    3. The novelty of the work. You need to ensure the originality of your work, project should not be a copied solely from any source. Even you are working on mirror/replication project, you should be in position to justify contribution.

In order to highlight these components, it is important you present a solid discussion regarding the learnings from the development of your method, and summarizing how your work compares to existing approaches.

#### **Deliverables (Deadlines will be announced in class and on KATIE):**

#### **Project Proposals:**

In project proposals, you will introduce your problem statement, data set required and general milestones you want to achieve during full phase of project. So, your project proposal will have following parts in it:

* + - 1. **Project title and team members:**  First you need state your project title with team member introductions.
      2. **Motivation:** What problem you are tackling? What application you can develop out of it? And most importantly what is the target you want to achieve through this?
      3. **Data:** You will discuss the data set? If it is pre-processed data or you need to acquire it through API/Crawling?
      4. **Milestones:** You need to state some of your milestones with dates for your projects, this will help you to stay on routine.

#### **Midterm Report and presentation:**

In Midterm project presentation you need to give 10-minute oral presentation with 5-minute Q&A along with report with following parts in it:

* + - 1. Original problem statement: You need to introduce with your initial problem statement as given in project proposal.
      2. Introduction to data sets including data source, size (complete volume), data attributes, types, statistical summary, data quality (Noise, Outliers, Duplicate and missing values), data cleaning, visualization techniques used (Histograms, Scatter plots, Heat maps, Reg plot etc.). Feature construction, feature extraction and feature removal as per your initial analysis for model development.
      3. Preliminary experiments: You need to describe any preliminary experiments you carried out on your data.
      4. Next steps you are planning to achieve the target?
    1. **Final report, poster and presentation**

You will prepare a poster illustrating your complete task, it will give everyone a good opportunity to see each other work. You need to present your poster for about 3-minute following with your final presentation of about 10 Minutes. There will be 5 minutes for Q & A session for each group.

You need to submit your final poster, code/libraries and report day before your presentation. Final report will have following components in it:

* + - 1. **Project Title and team members:** You need to write project name and team members name
      2. **Data:** You need to attach all the data pre-processing, quality and relevance work from midterm report. Additionally, you will discuss how you choose sample size? How you defined Training and Test sets?
      3. **Model/Algorithms:** This portion will discuss what model you used? why you choose a particular model? Why its suitable? What is efficiency and accuracy of model? Comparison with other techniques/models? Show predictions on test cases? You need to provide R score, MSE, Confusion matrix etc. scores along with possible visualizations to support your findings.
      4. **Significant findings and contributions:** You need to state significant findings and novel contributions to the problem.

At any point if you struggle with your project and stuck, please do visit My office anytime you need to discuss. My doors are always open to advice you as it’s a win-win situation. All the best.