Data Mining
Term Project

Prof. Dr. Şule Gündüz Öğüdücü

Term Project

- The term project is finding patterns in a dataset which size is suitable for data mining applications using Python programming language
- Define a problem which can be solved using the data mining techniques that are covered in the course
  - implement a solution presented in class or a recent data mining conference and evaluate the results
  - analyze and/or improve state of the art

Project Deliverables and Grading

- Project Topic
- Project Proposal (-/10 points Bonus)
- Intermediate Report (30%)
- Demo/Presentation (60%)
- Final Report (10%)

Project Topic

- Each student should submit a project topic (10/04/2022)
  - A brief summary of the project topic and goal
    - the problem you want to solve
- Each student will receive a notification that the project topic is accepted or it should be modified/changed (10/07/2022)

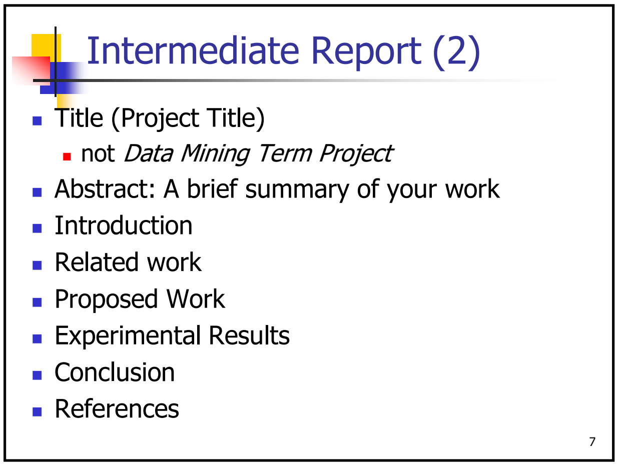
www3.itu.edu.tr/~sgunduz/courses/webmining/

Project Proposal (1)

- Each student should submit a project proposal (10/18/2022)
  - 1-2 page document using the provided IEEE conference template on the course web site
    - when working in Overleaf, the template is available at https://www.overleaf.com/gallery/tagged/ieee-official
- The proposal should include information on the following:
  - The data set: it will be selected by the students themselves
  - The problem you want to solve: why is it important?
  - The method you are planning to apply to solve the problem
  - The evaluation method
  - The time plan: Describe the steps and what you will accomplish at each step, backup plan

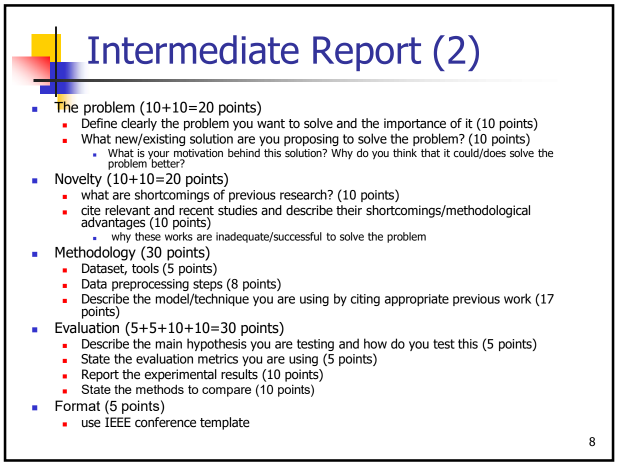
Intermediate Report (1)

- Each student should submit an intermediate report (11/29/2022): 105 points
  - 5-6 page document using the provided IEEE conference template on the course web site
- Define the current stage of your implementation
  - At the time of the intermediate report, I expect that you have **an initial working prototype** of your proposed model, and that you are able to report **early results** for that model.



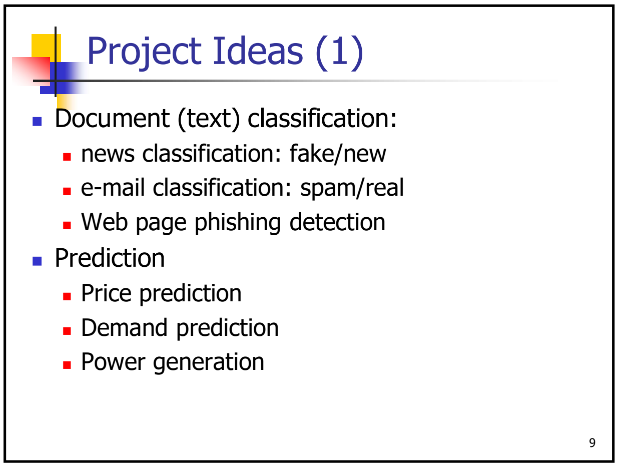
Intermediate Report (2)

- Title (Project Title)
  - not *Data Mining Term Project*
- Abstract: A brief summary of your work
- Introduction
- Related work
- Proposed Work
- Experimental Results
- Conclusion
- References



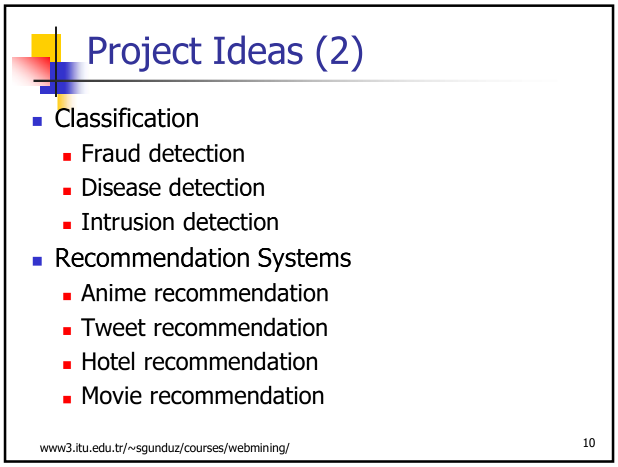
Intermediate Report (2)

- The problem (10+10=20 points)
  - Define clearly the problem you want to solve and the importance of it (10 points)
  - What new/existing solution are you proposing to solve the problem? (10 points)
    - What is your motivation behind this solution? Why do you think that it could/does solve the problem better?
- Novelty (10+10=20 points)
  - what are shortcomings of previous research? (10 points)
  - cite relevant and recent studies and describe their shortcomings/methodological advantages (10 points)
    - why these works are inadequate/successful to solve the problem
- Methodology (30 points)
  - Dataset, tools (5 points)
  - Data preprocessing steps (8 points)
  - Describe the model/technique you are using by citing appropriate previous work (17 points)
- Evaluation (5+5+10+10=30 points)
  - Describe the main hypothesis you are testing and how do you test this (5 points)
  - State the evaluation metrics you are using (5 points)
  - Report the experimental results (10 points)
  - State the methods to compare (10 points)
- Format (5 points)
  - use IEEE conference template



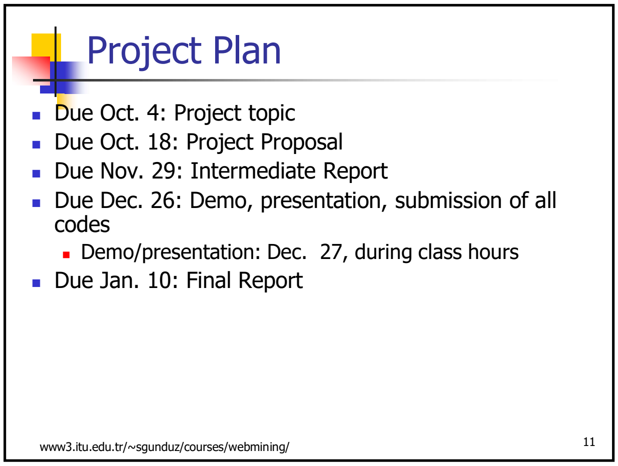
Project Ideas (1)

- Document (text) classification:
  - news classification: fake/new
  - e-mail classification: spam/real
  - Web page phishing detection
- Prediction
  - Price prediction
  - Demand prediction
  - Power generation



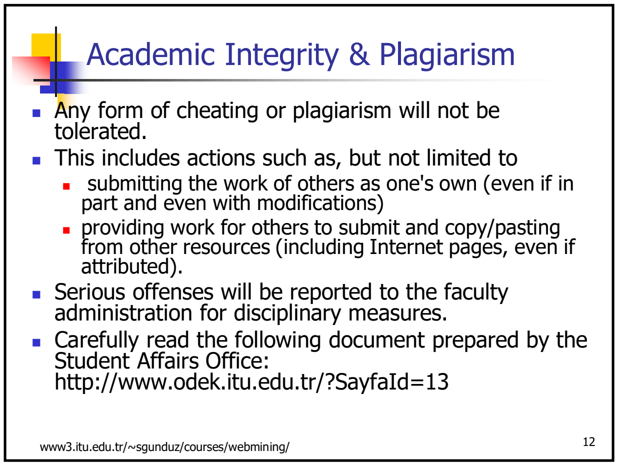
Project Ideas (2)

- Classification
  - Fraud detection
  - Disease detection
  - Intrusion detection
- Recommendation Systems
  - Anime recommendation
  - Tweet recommendation
  - Hotel recommendation
  - Movie recommendation



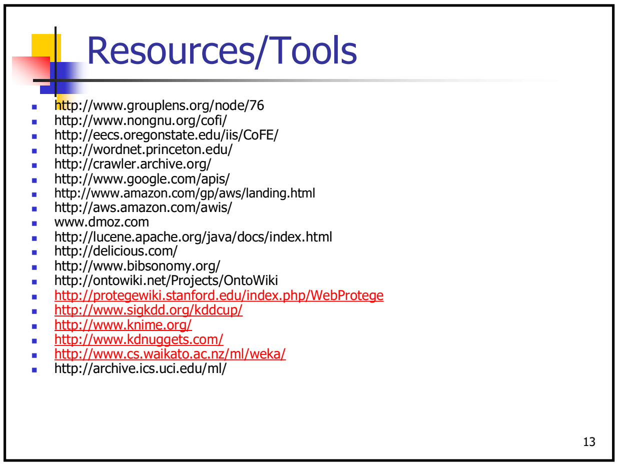
Project Plan

- Due Oct. 4: Project topic
- Due Oct. 18: Project Proposal
- Due Nov. 29: Intermediate Report
- Due Dec. 26: Demo, presentation, submission of all codes
  - Demo/presentation: Dec. 27, during class hours
- Due Jan. 10: Final Report



Academic Integrity & Plagiarism

- Any form of cheating or plagiarism will not be tolerated.
- This includes actions such as, but not limited to
  - submitting the work of others as one's own (even if in part and even with modifications)
  - providing work for others to submit and copy/pasting from other resources (including Internet pages, even if attributed).
- Serious offenses will be reported to the faculty administration for disciplinary measures.
- Carefully read the following document prepared by the Student Affairs Office: http://www.odek.itu.edu.tr/?SayfaId=13



Resources/Tools

- http://www.grouplens.org/node/76
- http://www.nongnu.org/cofi/
- http://eecs.oregonstate.edu/iis/CoFE/
- http://wordnet.princeton.edu/
- http://crawler.archive.org/
- http://www.google.com/apis/
- http://www.amazon.com/gp/aws/landing.html
- http://aws.amazon.com/awis/
- www.dmoz.com
- http://lucene.apache.org/java/docs/index.html
- http://delicious.com/
- http://www.bibsonomy.org/
- http://ontowiki.net/Projects/OntoWiki
- http://protegewiki.stanford.edu/index.php/WebProtege
- http://www.sigkdd.org/kddcup/
- http://www.knime.org/
- http://www.kdnuggets.com/
- http://www.cs.waikato.ac.nz/ml/weka/
- http://archive.ics.uci.edu/ml/