SCE331 DM Project Description

Project Grading: Accounts for 30% of your total grade
Final evaluation composed of **a full report** (22%), peer reviewed **presentation** (6%), participation of **peer reviewing** (2%) .
*** If you do not submit a project presentation or a project report, grade bellow C will be given even if you did everything else.**
** Presentations will be peer graded and a rubric will be provided.

Project Presentation Submission:

- **Video recording** one per team of the project presentation by **Dec. 10th (Sunday) 23:55**
- Late submission: every late date will result in reduction of 10% until Dec. 14일(화) **23:55** after which you will get **0%**
- How long?: **6 min** (if longer, presentation will be review only up to 6min)

Project Peer Evaluation Submission:

- Peer Review 7~10 presentations and submit grade in Google Forms
- Peer Review Deadline: **Dec. 22th(Fri.) 23:55**

Project Report Submission:

- Submit a report by **Dec. 10th (Sunday) 23:55** per team
- Late submission: every late date will result in reduction of 10% until Dec. 14일(화) **23:55** after which you will get **0%**
- Format: Convert to **Pdf** and submit through Ajou Blackboard with title as (teammate1_teammate2_teammate3_short-title.pdf)

Project Requirements: With the selected data, you are required to do the following:

1. Description of the Application (what problem is being solved)
2. Description of the Data used (format of data and information it contains, how it was obtained, preprocessing steps if needed)
3. Description of the Data Mining process used
   A. Must use two data mining algorithms learned in class and solves data mining problem we learned in class
4. Must contain the results that show you have succeeded in what you set out to do.

Submission list

1. Ajou Git (https://git.ajou.ac.kr/) or Github link of your project by until Dec. 14일(화) **23:55**
2. Presentation video (under 6 min) **Dec. 10th (Sunday) 23:55**
3. Report paper **Dec. 10th (Sunday) 23:55**
   i. Formal report that includes introduction (problem), method (solution approach), results, and conclusion (summary).
   ii. 8~10 page, 10 points; 1.5 spaced; Times New Roman

Data Samples:
Data Location & Data Descriptions:

- fourNewsGroups.tar.gz - Four newsgroups for document classification: See http://alias-i.com/lingpipe/demos/tutorial/cluster/read-me.html Natural Language Clustering section. "The four newsgroups data is a subset of 178 newsgroup posts balanced among the groups alt.atheism, misc.forsale, soc.religion.christian, and talk.religion.misc, a particularly challenging subset."
- muc34.tar.gz -"Message Understanding Conferences (MUC)" data for more description: https://www-nlpir.nist.gov/related_projects/muc/muc_data/muc_data_index.html
- twitter-sentiment-analysis2.zip - twitter sentiment challenge from Kaggle - from https://www.kaggle.com/c/twitter-sentiment-analysis2/data extracted on 2019-11-07
- sample.zip - "Applied Text Analysis with Python" sample data (you can also get 7.2 GB raw data through text github. https://github.com/foxbook/atap)

Other Corpus resource you may want to look at :

- https://www.informatik.tu-darmstadt.de/ukp/research_6/data/index.en.jsp
- https://github.com/dstl/baleen/wiki/Available-Corpora
- https://nlp.stanford.edu/links/statnlp.html#Corpora
- Google ngram: http://storage.googleapis.com/books/ngrams/books/datasetsv2.html

Project Report Structure

1. **Title**
   - The title must be clear, appropriate for the topic, and not too long (keep it less than 45 characters)
2. **Abstract**
   - Abstract is a self-contained piece of writing that can be understood independently from the essay or project.
   - It should be one to two paragraph; no figures & reference
   - Content of the abstract should contain: 1. Problem/Motivation/Objective 2. A statement of the problem and objectives 3. Methods or Approach you (will) use; 4. Summary results 6. Conclusions and comments
3. **Introduction**
   - Background information about the project's goal. Importance, related work, etc.
4. **Methods**
   - Describe the data (include the data description provided in UCI repository in your own words)
   - Basic properties of the data (mean, variance, normality, etc. and basic data visualizations)
   - Restate the objective of the project in terms of the data analysis.
   - In this section provide a clear, explicit and thorough description of the statistical analysis methods you use for the result presentation

   〈**Methods Contents**〉
   - *Architecture and Environment*:
     - Project environment (OS, software, hardware, languages, organizations, etc)

       Ex〉 Colab default setting with Python 3
   - *Analysis Methods Used and Evaluation Criterion*
5. **Results (Final report only):**
   - Describe the product of your project.
   - Evaluate your project according to the "evaluation criterion" you have specified.
6. **Conclusion:**
   - Make overall summary
7. **References**
   - Use IEEE citation style

     (IEEE Citation Reference: http://www.ieee.org/documents/ieeecitationref.pdf)
   - Using (free) citation manager program (+ plug in for word) is an easier way.
     - Suggestions: Mendeley Desktop.

   〈Example:〉

   in paragraph [1]

   [1] Y. Huang and S. Du, "Weighted support vector machine for classification with uneven training class sizes," in *Machine Learning and Cybernetics*, 2005, no. August, pp. 18-21.