Slides and Contents from

1. Lijun Zhang's Slides based on Charu C. Aggarwal's "Data Mining: The Textbook."

Ajou University, Dept. of Software and Computer Engineering

Lecturer: 이슬 Sael Lee

[IT집중교육1] Data Mining

Data Preparation

This lectures Slides are originally from: Lijun Zhang's Slides based on Charu C. Aggarwal's "Data Mining: The Textbook." Chapter 2

Updates on Grading Policy

- Project (35%):
  - There will be a middle evaluation composed of a **proposal** (5%) and final evaluation composed of **a full report** (20%) and **presentation (10%).**
- NOTE: According to instructor's decision, actual scores can be scaled to the evaluation measurements.

Lecture Schedule

Class#	date	소주제	실습주제
1	9/2(금)	DM Intro and Prelim	문제 풀이 (선대 등)
2	9/5(월)	Python - DS & Functions	Python 실습 1
3	9/7(수)	Python - OOP & Libraries (NumPy, Pandas, Matplotlib, etc)	Python 실습 2
4	9/14(수)	Python - Web and Text Processing (NLTK & GloVe)	NLTK - Tokenizing
5	9/16(금)	DM Intro - Data Preparation Overview & Project Details	FLASK
6	9/19(월)	Python - Functional Programming & MapReduce & Spark	Python 실습 4
7	9/21(수)	DM Algo - Association Rules	Python 실습 5
8	9/23(금)	DM Algo - Locally Sensitive Hashing	
9	9/26(월)	DM Algo - Clustering	
10	9/28(수)	DM Algo - Dimension Reduction	
11	9/30(금)	DM App - Recommendation System	
12	10/5(수)	DM App - Stream Mining	HW4-3, HW4-4
13	10/7(금)	DM App - Time Series Analysis	
14	10/10(월) 보강	DM App - Link Analysis	
15	10/12(수)	DM App - Graph Mining	Study for Exam
16	10/14(금)	DM 최신 동향 1	팀 프로젝트
17	10/17(월)	Exam	
18	10/19(수)	DM 최신 동향 2	
19	10/21(금)	DM 최신 동향 3	
20	10/24(월)	DM 최신 동향 4	
21	10/26(수)	프로젝트 발표 평가	

Project Requirements

Grading Rubric – Proposal (5%)

1) Done on time and 2) Contains all required content	1) Done on time and 2) Missing 1 required contents	1) Not done on time and 2) Missing 1 of the required contents	1) Not done on time and 2) Missing more than 1 of the required contents	Not done
5	4	3	2	0

- Due date for the Proposal is Oct 5th (Wednesday)

Grading Rubric – Project Report Paper (20)

1) Finished on time and 2) Contains all required content (Content is almost flawless and fully suitable for class content.) 3) Application or Algorithm shows novelty	1) Finished on time and 2Contains all required content (Content has a few errors, but is still suitable.) 3) Lacks novelty but shows solid work.	1) Finished on time and 2) Missing more than 2 of the required contents (Content has errors, but is still suitable.) 3) Lacks novelty and work can be improved.	1) Not Finished on time or 2) Missing more than 2 of the required contents (Content has many errors or is unsuitable for class) 3) Lacks novelty and work need to be improved	Not submitted
20	16	12	8	F

Grading Rubric - Project presentation (10)

Project presentation (10):

- Active participation of the peer grading (2)
- Peer graded based on quality of the presentation (8)

INDICATOR	EXCEEDS	MEETS	DEVELOPING	NOT DONE
	(8)	(6)	(4)	0
Clarity of expression and enthusiasm for topic	Speaker is animated and enthusiastic about the topic. Speaker enhances the project by using clear speech, excellent eye contact, and appropriate tone, volume, pace, transitions and vocabulary throughout presentation. Good posture, and natural hand gestures.	Speaker is mildly enthusiastic about the topic, and at times, uninspiring. Speaker uses clear speech, maintains good eye contact, and proper spacing, tone, volume, pace, transitions and vocabulary during the presentation. Good posture, and hand motions.	Speaker is emotionless about the topic. Speaker uses unclear speech, little eye contact, inappropriate volumeor language, and/or choppy spacing and monotone in more than half of the presentation. Posture, or gestures distract the audience.	not done
Familiarity with information	Speaker is very confident with the material and maintains excellent focus throughout the presentation. Speaker is well-prepared, indicating multiple rehearsals.	Speaker is somewhat confident with the material, maintains good focus during the presentation.Speaker is somewhat prepared, indicating more rehearsals necessary.	Speaker is not confident with the material, loses focus, and reads directly from thePower Point throughout presentation. It is clear that rehearsal is lacking and inadequate	not done
Design of PowerPoint slides	Graphics and text are colorful or contrasted, creative, and uniquely and consistantly sized, framed, and/or spaced with appropriate size and quantity. Presentation interest of the viewer.	Graphics and textare somewhat colorful, creative, and uniquely sized, framed, and/or spaced. Some text is crowded and/or too small. Some graphics show little depth and detail. Presentation is interesting to the viewer.	Slides are somewhat dull, uninteresting and not unique. Text is too crowded and/or illegible. Presentation is somewhat monotonous and repetitive to the viewer.	not done
Inclusion of introduction, conclusion, and work citation	Introduction creatively establishes importance of topic and the name of the presenter, and contains pictures to reveal content of presentation. Conclusion briefly summarizes the importance of the presentationand has a related graphic. Work citations for text and graphics are properly formatted and represented	Introduction establishes topic and presenter. Conclusion summarizesthe presentation. Work citations for text and graphics are included, but may not be complete	Introduction briefly establishes topic and presenter. Conclusion weakly sums up the topic. Work citations are minimal.	not done

Project Requirements

- Project Type
  - Implementation
  - Uses of Spark is encouraged
- Project Requirements
  - Must contain raw data download (Python `request`) and preprocessing step
  - Must be a Data Mining Application (Recommendation, Advertising, Page rank, Community Detection, Stream Outlier Detection) in a form of web or app
  - Must contain data mining algorithm implementation that we learned in class
  - Should have some novelty in application or in the algorithm
  - Target for **2022 한국소프트웨어종합학술대회Korea Software Congress 2022 (KSC2022)** 2022년 12월 21일(수)~23일(금)

Project Proposal

- Prepare 5 page PPT and present a (2 min max) about
  - What type of data you are going to gather
  - What target application you're going to make
  - What DM algorithms do you think is needed in your application
  - What will be the novelty of your application
  - What have you done so far

Project Report

- 정보확학회 논문지 template으로 5 page 작성.

Must Contain Components of a Project Report

1. Title
2. Abstract
   - (1 or 2 paragraphs)
3. Introduction
   - (0.5 page)
4. Method
   - (1.5 page)
5. Results
   - (1 page)
6. Conclusion
   - (1 paragraph)
7. References

Credit: https://people.ok.ubc.ca/rlawrenc/teaching/writingProposal.html
Compiled by Dr. Ramon Lawrence, Associate Professor of Computer Science at the University of British Columbia - Okanagan Campus and founder of UnityJDBC.

Title

What title will make your title
“Google friendly”?
Not to common but have the key words.

- The title of the project is very significant.
- The title must be
  - Clear,
  - Appropriate for the topic
  - Keep it less than 45 characters
- Take hints from publish works when you do a short survey for the proposal.

Abstract

After searching, abstract will determine whether a person will take his/her time to read your work.

- A self-contained piece of writing that can be understood independently from the essay or project
- One to two paragraph; no figures & reference
- Should contain:
  - Problem/Motivation/Objective
    - A statement of the problem and objectives
  - Methods or Approach you (will) use
    - A summary of methods you (will) employ or your research approach;
  - (Expected) Results
    - The significance of the proposed topic should become clear
  - Conclusions and comments
    - Broader Impact

Objective should be SMART

- All objectives should be **SMART**
- **Specific**
  - Be precise about what you are going to achieve
- **Measurable**
  - Quantify you objectives
- **Achievable**
  - Are you attempting too much?
- **Realistic**
  - Do you have the resource to make the objective happen (human resources, financial, the right context and opportunities)?
- **Timed**
  - State when you will achieve the objective

Introduction

Shows whether the writer know the field they're writing about.
Aligns the reader with the writer by providing adequate info about the topic.

- The Introduction to the project provides a general introduction to the phenomena or issue of interest, and is usually contained in 2 pages.
- The issue or problem under investigation is described,
- Background and/or context for understanding the nature of the issue is provided.
- Provides answers to two main questions:
  - What is the project all about? (Problem Definition and Goal)
  - Why is the project important or worthwhile? (Motivation)
  - What is new compared to existing publications related to the topic? (Short Survey)
- The Introduction will also typically conclude with a brief description of the structure of the remainder of the document.

Method: Project Details

(This section must make sense within the context of the document and be linked with the sections preceding it.)

- In this section provide a clear, explicit and thorough description of how you will complete your project and the timetable for completing each step.

- It is the writer's responsibility to ensure that the proposal is clear about what is being proposed, with whom, where and when. (WHY should already have been explained.)

Method: Project Details - *Data and Environment*

- 2~3 paragraphs + figures

- Describe the data gathering and preprocessing
- Describe the data statistics and distribution (use data visualization tools)

- Describe the project environment
  - Implementation: software, hardware, languages, organizations, etc.

Method: Project Details
Problem and Solution Approach

- This can be long
- What will be the most difficult issues and challenges in the implementation?
- How are you using or extending current tools/systems/algorithms for your problem?
- What makes your project unique (novel)?

Project Details
Evaluation Measures

- Strategy to measure the success
- Explanation of the criteria used to measure the success
- Includes:
  - quantitative indicators (numbers)
  - qualitative indicators (contents)
  - vision of success (what you want to achieve)

Result

- The **Results** Should Justify Your Method (usefulness of your app)
  - Data presented in tables, charts, graphs, and other figures (may be placed among research text or on a separate page)
  - A contextual analysis of this data explaining its meaning in sentence form
  - Report on data collection, recruitment, and/or participants
  - Data that corresponds to the central research question(s)
  - Secondary findings (secondary outcomes, subgroup analyses, etc.)
- Summarize Your **Results.**
- Include Tables and Figures.

Conclusion

- 1 paragraph
- Summarize the project including the problem, motivation, and proposed solution, and re-state important (planned) contributions.

- Emphasize what your project contributes or achieves

? ? ?

Any questions?

? ? ?

Email: sael@ajou.ac.kr