Big Data Intelligence Project

2024/10

Project and Basics

- This project aims to help you comprehensively understand the content of this course. You are supposed to use what you have learned from this course to finish the project.
- Basically, you are required to form a team, choose a task from the given topics, and finish it. (we suggest **3 members** per team, 2-4 is acceptable)
- At the end of this course, each team is required to give a presentation to introduce your project (the task, your solution, and your learned experience etc.), and others could ask questions, score your projects. The score of your project is highly related to your final score in this course.

How to form a team

- Those who could access the web learning system, could post in the forum, and this could help you find teammates. You could also directly discuss offline/online through WeChat to form a team.

- After you have decided to form a team, please let us know by emailing the TA or directly fill in the online table with your team name and team members, and the task you choose.

- Note: If you really find it hard to cooperate with others (Time Zone problem, Language gap, etc.), please let us know, we will try to help. We encourage teamwork, and we would not agree you to work alone unless there is no other choice.

Topics

- **Online Competitions.**
  - We select some hot and interesting topics from online competition websites (Kaggle, AliTianchi, etc.). These competitions would improve your skills. Moreover, if you really work hard, you would have the chance to win the money award from the sponsor. (See the pdf file)
- **Your own project of applying BDI to your discipline.**
  - Integrating BDI with a specific discipline is also a very important topic. Therefore, students not from the department of information science can choose the combination of big data analytics and a specific domain application as a project. Before you start, please email the TA with a short proposal to roughly introduce your project.
- **Research survey on literature related to big data intelligence**
  - This course could not cover all the frontiers of BDI, you can also choose a frontier area of big data intelligence to conduct literature research survey.

Note: Refer to the project description file for more details.

Timelines

- **Before next class (29th, October): form a team, proposal.**
  - You should fill in the online table with your team information (members, topics).
  - If you choose the personal project, you should also email your project proposal to the TA before next class.
- **Before 10th Dec: finish project.**
  - Please remember to start early on your project.
- **The last three classes (10th Dec. ~ 24th Dec.): Presentation.**
  - Each team is required to give a presentation to introduce your project (task, solution, ranking).
  - We encourage in-depth thinking and more trials about the project, and not only pay attention to the final competition results.
  - You are also required to submit a project report after your presentation within 1 week. Grading will both consider your presentation (and report) and competition results (See the pdf file for more details).

Thanks a lot!

Q&A

Some Competition Examples

1. House Prices - Advanced Regression Techniques

https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques

Ask a home buyer to describe their dream house, and they probably won't begin with the height of the basement ceiling or the proximity to an east-west railroad. But this playground competition's dataset proves that much more influences price negotiations than the number of bedrooms or a white-picket fence. With 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa, this competition challenges you to predict the final price of each home.

2. Google - Unlock Global Communication with Gemma

https://www.kaggle.com/competitions/gemma-language-tuning

With over 7,000 languages and countless cultural differences, AI has the potential to foster global understanding. In a step towards broader linguistic inclusion, we're launching a Kaggle competition focused on adapting Gemma 2, Google's open model family, for 73 eligible languages. These languages were selected to represent a diverse range and to align with the expertise of our judging panel for effective evaluation. Our initial focus on these languages will allow us to establish a robust foundation of techniques and resources that will later enable us to support under-resourced languages.

More detailed task description will be released on the web learning.