Lab 2 Report

1. Team Details
Team Name:

Name	USC ID
Chenxiao Yu	6024079123
Yiqing Hong	4395913002

2. Github link: https://github.com/AiChiMoCha/SP25_DSCI560/tree/main/lab2

3. Datasets
Shortlisted domain: Machine Learning
Topic choice reason: There are many public resources of ML online, so we can easily get the training data we need.

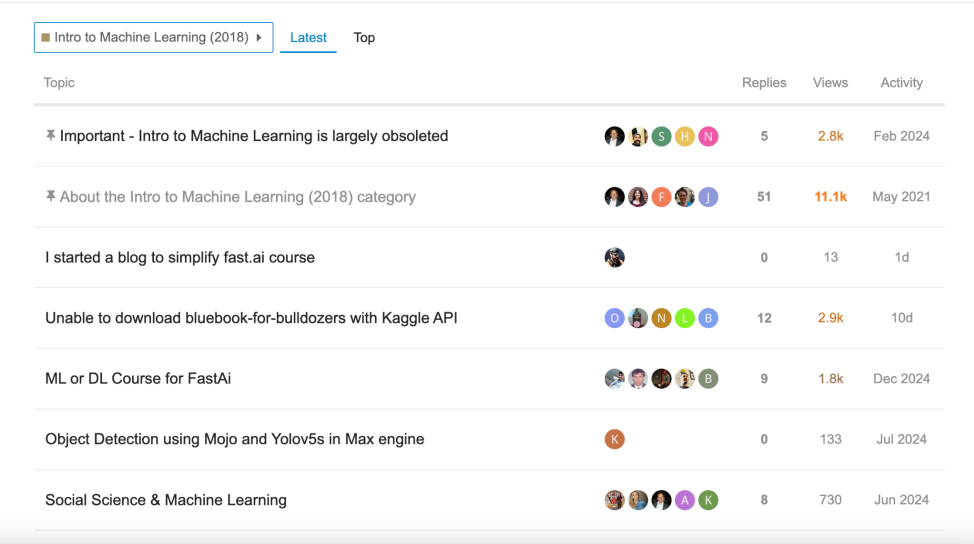
- Stanford CS299 Machine Learning course materials

Syllabus and Course Schedule

Event	Date	Description	Materials and Assignments
Introduction and Pre-requisties review (3 lectures)			
Lecture 1 [YouTube]	6/24	• Introduction and Logistics • Review of Linear Algebra	**Class Notes** • Introduction [pptx] • Linear Algebra (section 1-3) [pdf]
Lecture 2 [YouTube]	6/26	• Review of Matrix Calculus • Review of Probability	**Class Notes** • Linear Algebra (section 4) [pdf] • Probability Theory [pdf] • Probability Theory Slides [pdf]
Lecture 3 [YouTube]	6/28	• Review of Probability and Statistics • Setting of Supervised Learning	**Class Notes** • Supervised Learning [pdf] • Probability Theory [pdf]
Supervised Learning (8 lectures)			
Lecture 4 [YouTube]	7/1	• Linear Regression • [Stochastic] Gradient Descent ([S]GD) • Normal Equations	**Class Notes** • Supervised Learning (section 1-3) [pdf]

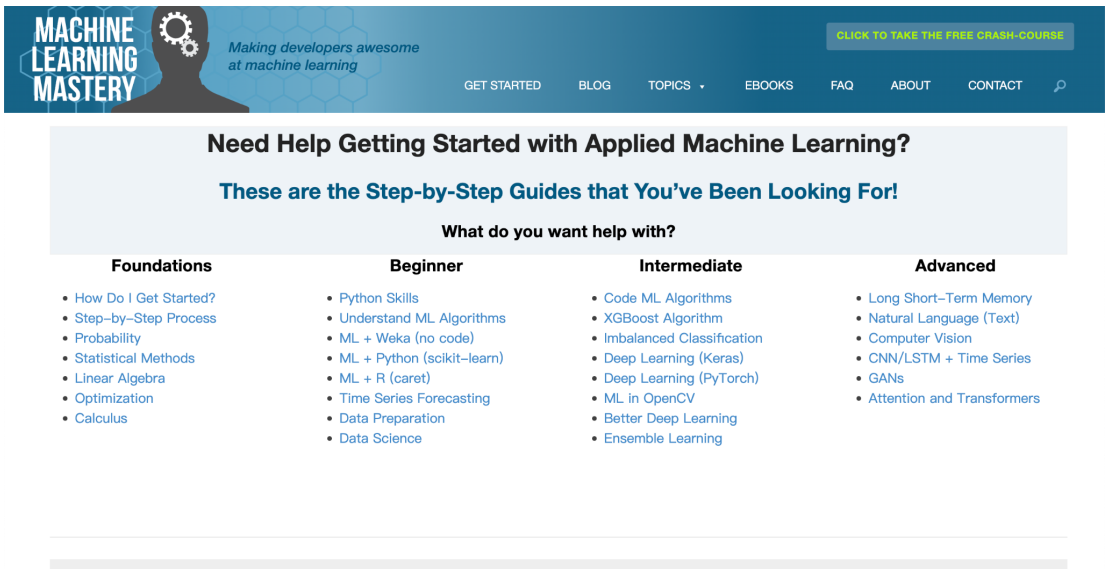
Link: https://cs229.stanford.edu/syllabus-summer2019.html
Description: This website includes course structure, video links and class notes. CS229 provides a broad introduction to statistical machine learning (at an intermediate / advanced level) and covers supervised learning (generative/discriminative learning, parametric/non-parametric learning, neural networks, support vector machines); unsupervised learning (clustering, dimensionality reduction, kernel methods); learning theory (bias/variance tradeoffs, practical); and reinforcement learning among other topics.
Reason: The course content covers machine learning topics from basic to advanced levels. This comprehensive coverage provides the depth and breadth to train a chatbot capable of answering most questions in machine learning related fields.

- fast.ai machine learning forum



Link: https://forums.fast.ai/c/ml1/13
Description: This is a forum where many machine learning learners ask ML related questions and make comments. There are also online course materials. People post questions about the course and someone answers the questions.
Reason: Forum questions come from learners of different levels, covering a variety of question types from beginners to intermediate and advanced users. The training data can help the robot answer questions of different depths. By understanding the learners' questioning methods and language habits, the robot can understand the questions more naturally and provide answers that are close to the user's language style. In addition, there are often code examples and picture explanations in the discussion, which helps the robot provide more intuitive answers when answering questions.

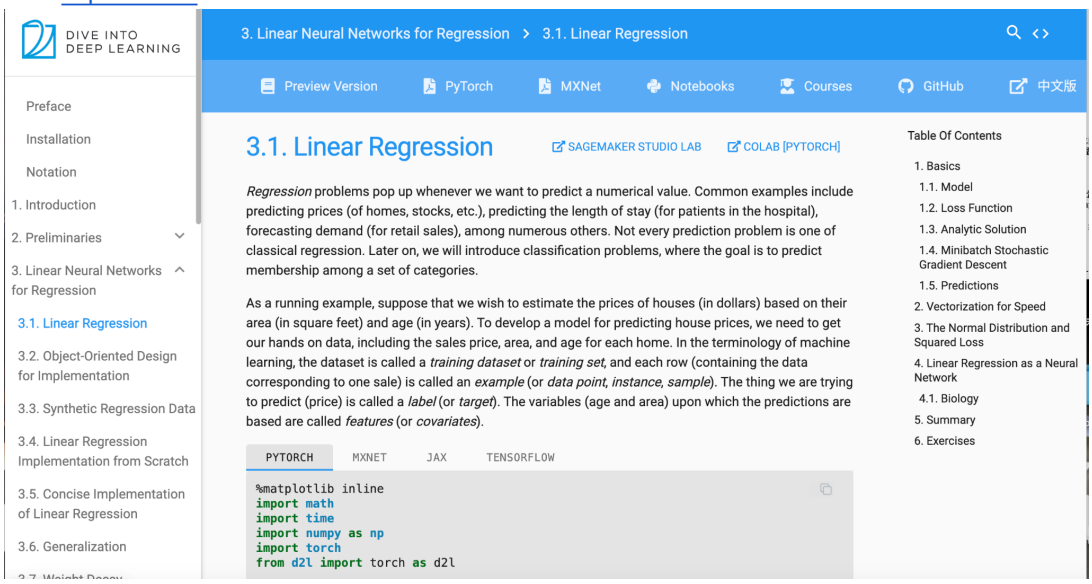
- Machine learning mastery blogs



Link: https://machinelearningmastery.com/start-here/
Description: The blog is structured in a simple, step-by-step manner: from basic theories (such as linear regression) to advanced topics (such as deep learning and reinforcement learning). Each blog usually explains complex concepts in easy-to-understand language, and helps users understand abstract machine learning concepts through concrete examples. The content of the blog is broader than a single ML course, covering many important areas of machine learning.
Reason: The chatbot can provide concise and easy-to-understand answers through this content, which is especially suitable for beginners' questions. The blog focuses on implementing knowledge points through code, and the trained chatbot can quickly generate code examples. Due to the wide range of content, the chatbot can answer machine learning questions from entry to advanced, covering multiple sub-directions

- D2L.ai: Interactive Deep Learning Book with Multi-Framework Code, Math, and Discussions

Link: https://d2l.ai/



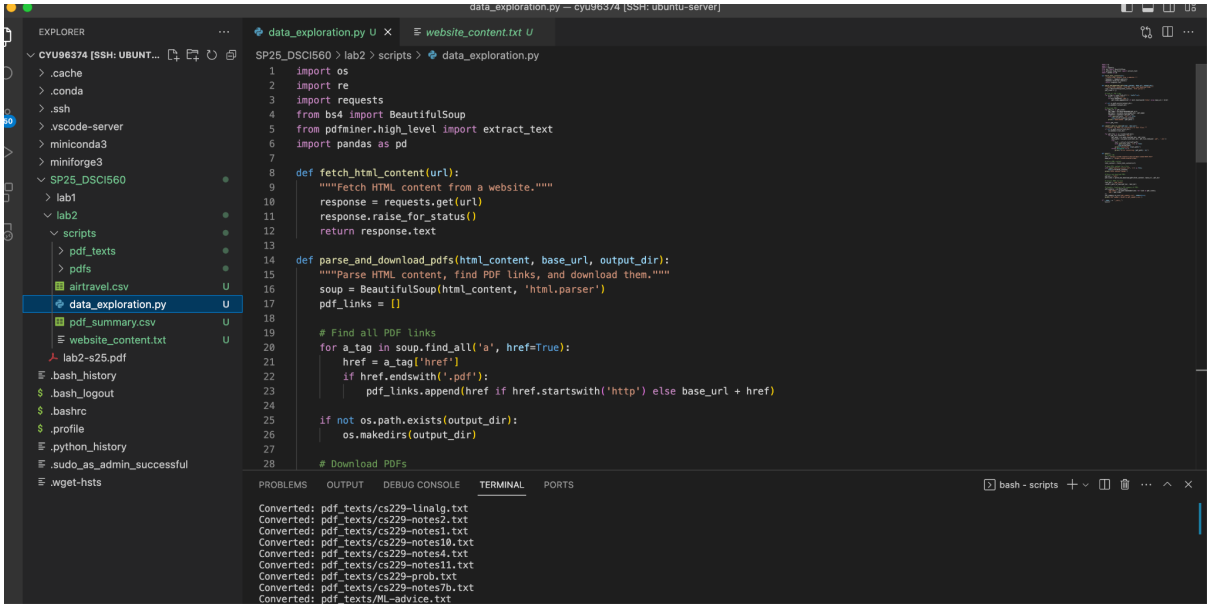
Description:
This open-source book represents our attempt to make deep learning approachable, teaching you the concepts, the context, and the code. The entire book is drafted in Jupyter notebooks, seamlessly integrating exposition figures, math, and interactive examples with self-contained code.
Reason: D2L.ai's courses are presented in the form of Jupyter Notebooks, integrating mathematical formulas, visual charts, and code examples, emphasizing interactivity. These features allow the chatbot to generate answers with formula derivations and code examples, helping users quickly grasp concepts.

Data Collection Overview

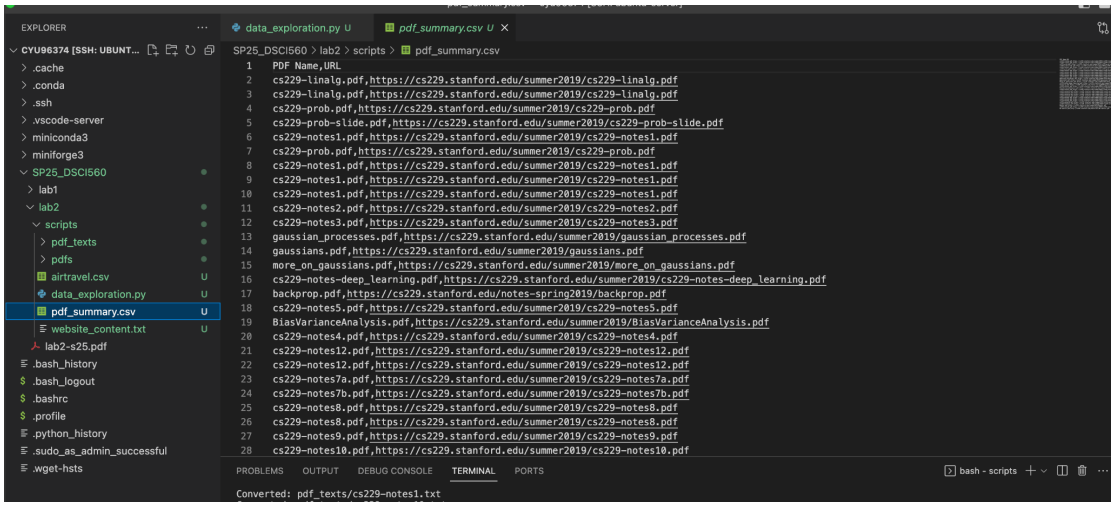
The data collection process involves three types of data:

1. **CSV or Excel:** Structured data stored in tabular format.
2. **ASCII Texts:** Forum postings or website HTML content.
3. **PDF and Word Documents:** Files that require conversion and OCR for text extraction.

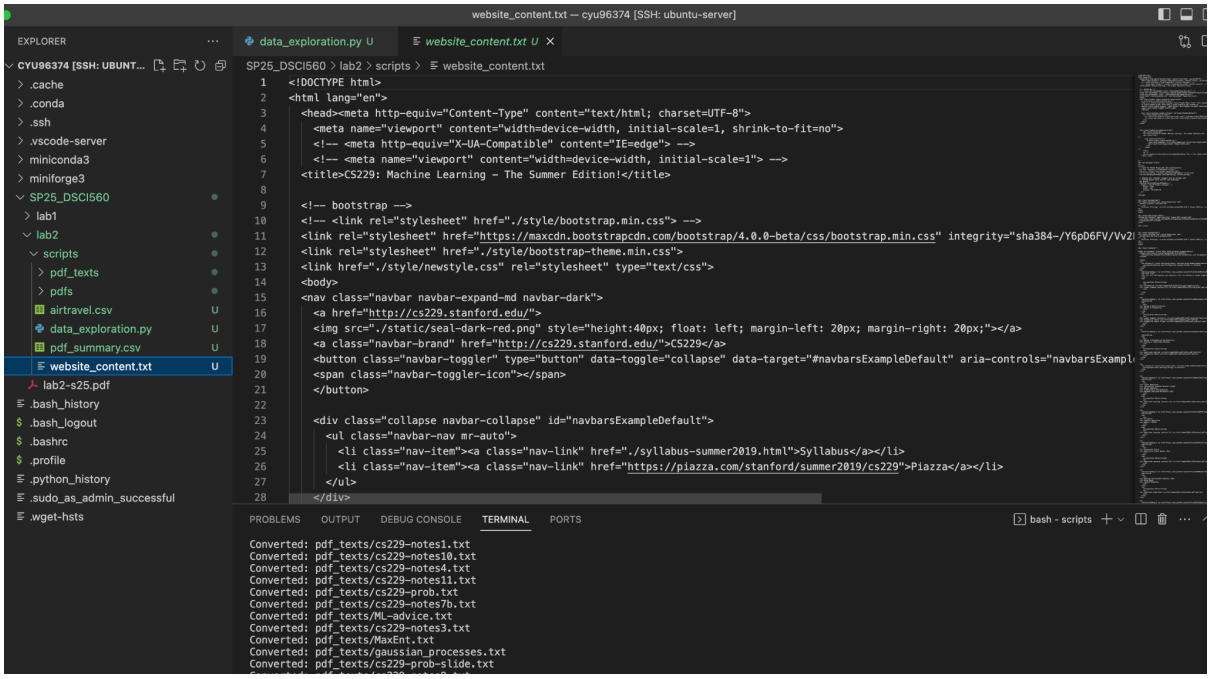
Currently, the code resides in the `SP25_DSCI560/lab2/scripts` folder and performs the following functions:



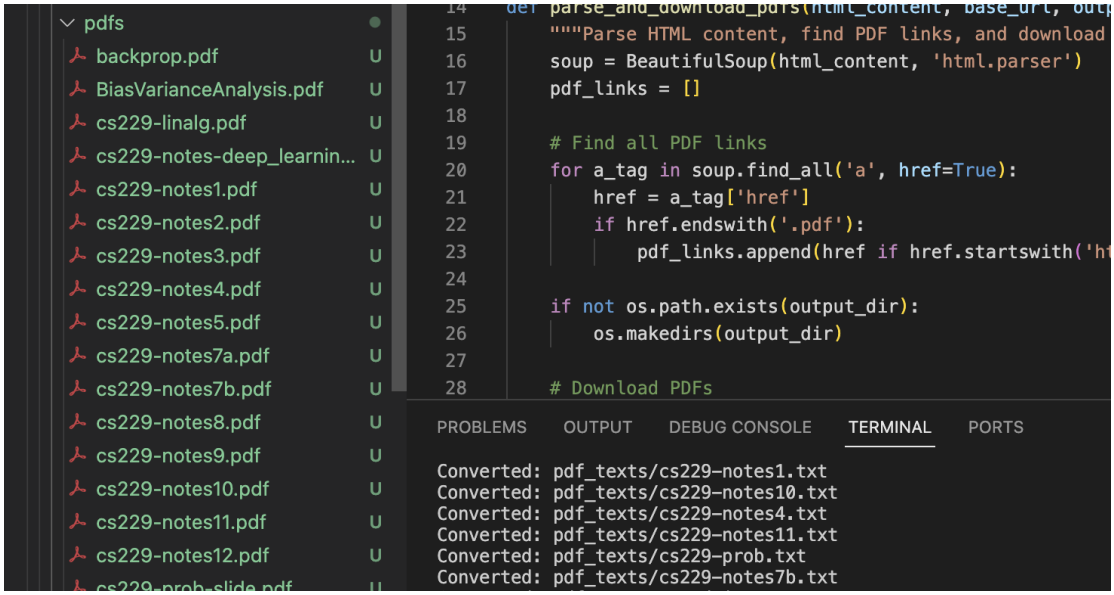
1. **PDF Summary Creation:** Collects all PDF files from the course website and generates a consolidated `pdf_summary.csv` for quick access to their metadata.



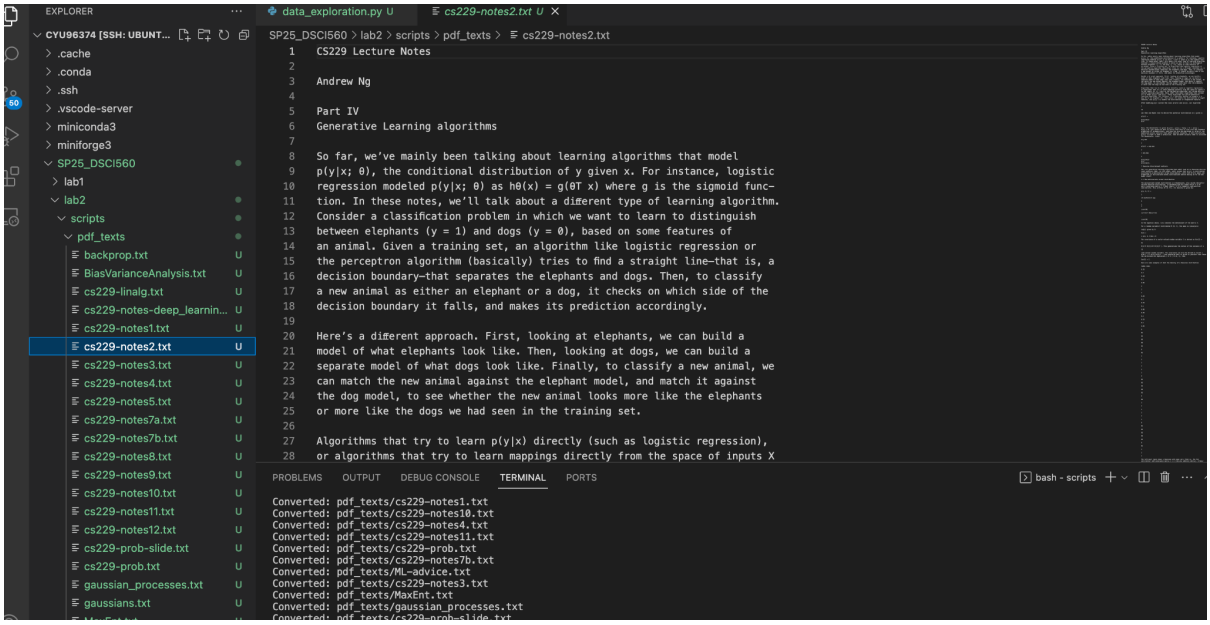
2. **Website Text Extraction:** Extracts textual information from the website and saves it in `website_content.txt`.



3. **PDF Download:** Downloads all course-related PDFs into the `pdfs` folder.



4. **PDF to Text Conversion:** Converts all PDFs in the `pdfs` folder to text files using the `pdfminer` API for subsequent embedding tasks.



5. Current problem and improvement

Current problem of existing chatbot:

- **Currently Unable to Transfer Video contents:** While the Moodle pages contain numerous video resources that are highly beneficial for students' studies, the current system does not support direct handling or transfer of video content. This limitation restricts the inclusion of video materials in the data collection process.
- **Limited Understanding of Context**: Most existing chatbots struggle to understand the nuances of user questions, especially when queries involve multiple layers of meaning or depend on earlier parts of the conversation. They often provide generic or partially correct responses rather than deeply contextualized answers.
- **Inadequate Domain Knowledge**: Many chatbots lack specialized knowledge, particularly for technical fields like machine learning. Their responses often sound superficial or too generalized because they are not trained on high-quality, domain-specific resources.
- **Static and Outdated Responses**: Chatbots often fail to keep up with the latest advancements in fast-evolving fields. Machine learning and deep learning, for instance, introduce new tools, frameworks, and best practices regularly. Many chatbots do not have updated knowledge bases.
- **Lack of Interactivity and Engagement**: Existing chatbots rarely guide users through complex problems interactively. For example, when users want help debugging a piece of code, most bots can only provide static suggestions instead of engaging in iterative problem-solving.
- **Code and Practical Guidance Deficiencies**: While some chatbots provide theoretical answers, they often fall short in offering practical examples or runnable code. For technical fields, this is a crucial limitation.
- **Inability to Explain Concepts Clearly**: Many chatbots struggle to balance technical depth with user-friendliness, often providing explanations that are either too complex or oversimplified.

How our dataset might improve:

- **To address the current limitation of handling video content on Moodle pages,** WhisperX could be utilized as a potential solution.
- **Enhanced Domain Expertise**: incorporating high-quality resources
- **Interactive Problem-Solving**: By training on discussion-based datasets (e.g., Fast.ai Forums), the chatbot learns to mimic a mentor-student dynamic, breaking down problems into smaller steps; offer iterative feedback, allowing users to refine their approach and deepen their understanding

- **Balanced Explanations**: Datasets like Machine Learning Mastery Blogs focus on simplifying complex concepts without losing depth. Training the chatbot with these materials ensures beginners receive clear, digestible explanations, and advanced users get detailed and nuanced responses, complete with code snippets or references.
- **Up-to-Date and Comprehensive Knowledge**: Open-source and community-driven resources like D2L.ai and Fast.ai Forums are regularly updated. Using these ensures the chatbot stays current with recent frameworks, tools, and algorithms.
- **Practical Code Assistance**: Resources like D2L.ai include runnable code, which can be embedded in the chatbot's responses. This makes it capable of offering debugging tips or alternatives when errors arise