

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-requisites review (3 lectures)			
Lecture 1 [YouTube]	6/24	<ul style="list-style-type: none">• Introduction and Logistics• Review of Linear Algebra	Class Notes <ul style="list-style-type: none">• Introduction [pptx]• Linear Algebra (section 1-3) [pdf]
Lecture 2 [YouTube]	6/26	<ul style="list-style-type: none">• Review of Matrix Calculus• Review of Probability	Class Notes <ul style="list-style-type: none">• Linear Algebra (section 4) [pdf]• Probability Theory [pdf]• Probability Theory Slides [pdf]
Lecture 3 [YouTube]	6/28	<ul style="list-style-type: none">• Review of Probability and Statistics• Setting of Supervised Learning	Class Notes <ul style="list-style-type: none">• Supervised Learning [pdf]• Probability Theory [pdf]
Supervised Learning (8 lectures)			
Lecture 4 [YouTube]	7/1	<ul style="list-style-type: none">• Linear Regression• [Stochastic] Gradient Descent ([S]GD)• Normal Equations	Class Notes <ul style="list-style-type: none">• 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


Intro to Machine Learning (2018) ▾		Latest	Top			
Topic		Replies	Views	Activity		
⚡ Important - Intro to Machine Learning is largely obsoleted		5	2.8k	Feb 2024		
⚡ About the Intro to Machine Learning (2018) category		51	11.1k	May 2021		
I started a blog to simplify fast.ai course		0	13	1d		
Unable to download bluebook-for-bulldozers with Kaggle API		12	2.9k	10d		
ML or DL Course for FastAi		9	1.8k	Dec 2024		
Object Detection using Mojo and YOLOv5s in Max engine		0	133	Jul 2024		
Social Science & Machine Learning		8	730	Jun 2024		

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



MACHINE LEARNING MASTERY

Making developers awesome at machine learning

CLICK TO TAKE THE FREE CRASH-COURSE

GET STARTED BLOG TOPICS EBOOKS FAQ ABOUT CONTACT

Need Help Getting Started with Applied Machine Learning?

These are the Step-by-Step Guides that You've Been Looking For!

What do you want help with?

Foundations	Beginner	Intermediate	Advanced
<ul style="list-style-type: none"> How Do I Get Started? Step-by-Step Process Probability Statistical Methods Linear Algebra Optimization Calculus 	<ul style="list-style-type: none"> Python Skills Understand ML Algorithms ML + Weka (no code) ML + Python (scikit-learn) ML + R (caret) Time Series Forecasting Data Preparation Data Science 	<ul style="list-style-type: none"> Code ML Algorithms XGBoost Algorithm Imbalanced Classification Deep Learning (Keras) Deep Learning (PyTorch) ML in OpenCV Better Deep Learning Ensemble Learning 	<ul style="list-style-type: none"> Long Short-Term Memory Natural Language (Text) Computer Vision CNN/LSTM + Time Series GANs Attention and Transformers

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/>

The screenshot shows the D2L.ai website interface. On the left is a sidebar with a table of contents. The main content area is titled '3.1. Linear Regression' and contains text explaining regression problems and a running example of predicting house prices. Below the text is a code block with the following code:

```

PYTORCH  MXNET  JAX  TENSORFLOW

%matplotlib inline
import math
import time
import numpy as np
import torch
from d2l import torch as d2l

```

On the right is a 'Table Of Contents' sidebar listing the chapters and sections of the book.

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.

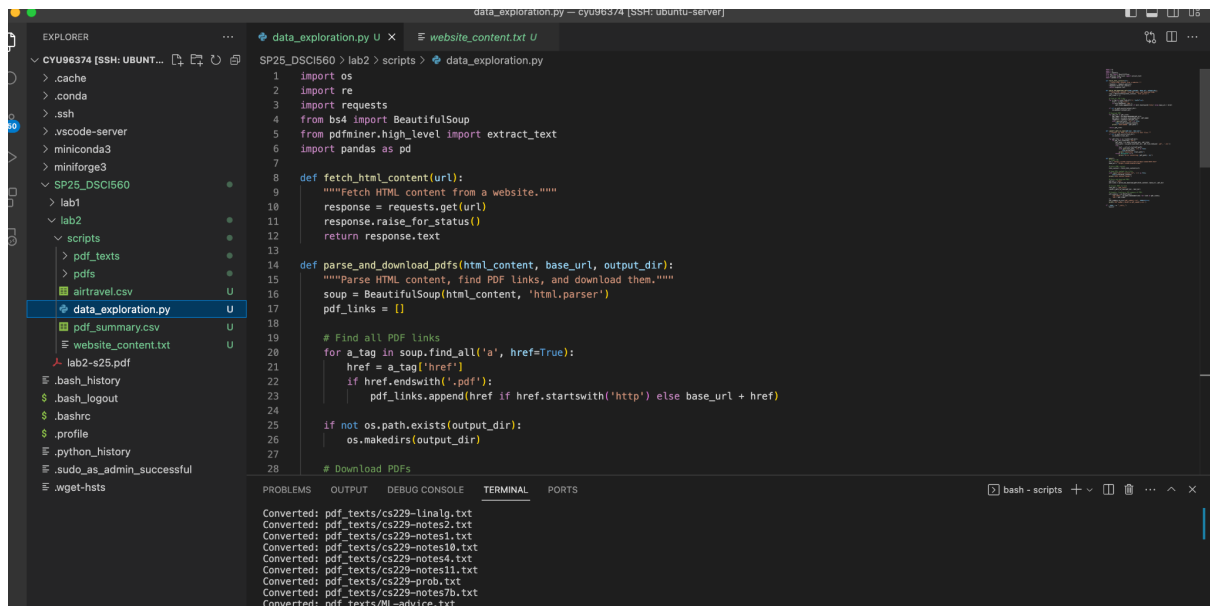
Reason: D2L.ai's courses are presented in the form of Jupyter Notebooks, integrating mathematical formulas, visual charts, and code examples, emphasizing interactivity. These features allow the chatbot to generate answers with formula derivations and code examples, helping users quickly grasp concepts.

Data Collection Overview

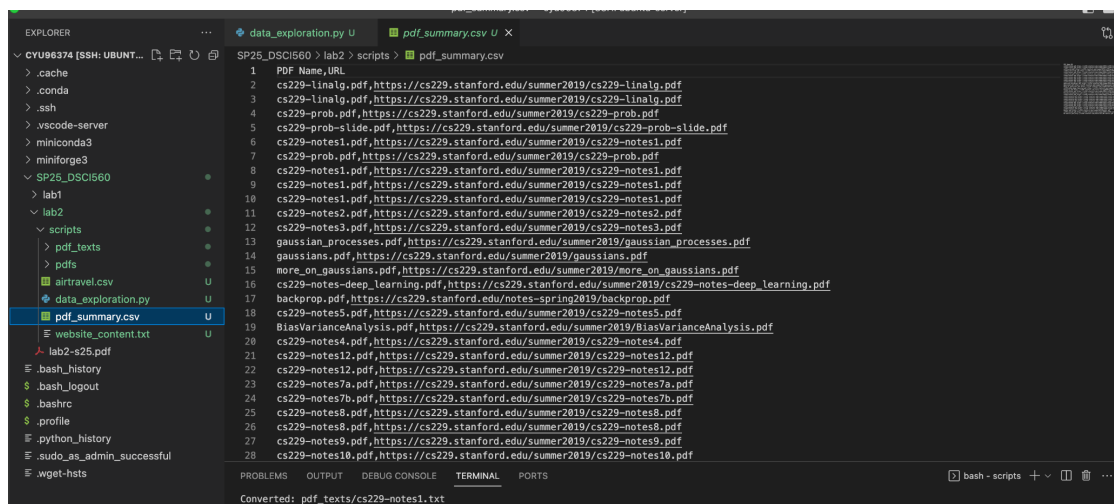
The data collection process involves three types of data:

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`.

```
1 <!DOCTYPE html>
2 <html lang="en">
3 <head><meta http-equiv="Content-Type" content="text/html; charset=UTF-8">
4 <meta name="viewport" content="width=device-width, initial-scale=1, shrink-to-fit=no">
5 <!-- <meta http-equiv="X-UA-Compatible" content="IE=edge"> -->
6 <!-- <meta name="viewport" content="width=device-width, initial-scale=1"> -->
7 <title>CS229: Machine Learning - The Summer Edition!</title>
8
9 <!-- bootstrap -->
10 <!-- <link rel="stylesheet" href="/style/bootstrap.min.css" -->
11 <link rel="stylesheet" href="https://maxcdn.bootstrapcdn.com/bootstrap/4.0.0-beta/css/bootstrap.min.css" integrity="sha384-Y6p66FVv2"
12 <link rel="stylesheet" href="/style/bootstrap-theme.min.css">
13 <link href="/style/newstyle.css" rel="stylesheet" type="text/css">
14 <body>
15 <nav class="navbar navbar-expand-md navbar-dark">
16 <a href="http://cs229.stanford.edu/">
17 
18 <a class="navbar-brand" href="http://cs229.stanford.edu/">CS229</a>
19 <button class="navbar-toggler" type="button" data-toggle="collapse" data-target="#navbarsExampleDefault" aria-controls="navbarsExamp
20 <span class="navbar-toggler-icon"></span>
21 </button>
22
23 <div class="collapse navbar-collapse" id="navbarsExampleDefault">
24 <ul class="navbar-nav mr-auto">
25 <li class="nav-item"><a class="nav-link" href="/syllabus-summer2019.html">Syllabus</a></li>
26 <li class="nav-item"><a class="nav-link" href="https://pliazza.com/stanford/summer2019/cs229">Piazza</a></li>
27 </ul>
28 </div>
```

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

```
14 def parse_and_download_pdfs(html_content, base_url, output_dir):
15     """Parse HTML content, find PDF links, and download them into the pdfs folder.
16     soup = BeautifulSoup(html_content, 'html.parser')
17     pdf_links = []
18
19     # Find all PDF links
20     for a_tag in soup.find_all('a', href=True):
21         href = a_tag['href']
22         if href.endswith('.pdf'):
23             pdf_links.append(href if href.startswith('/') else base_url + href)
24
25     if not os.path.exists(output_dir):
26         os.makedirs(output_dir)
27
28     # Download PDFs
```

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

The screenshot shows a VS Code editor interface. On the left, the 'EXPLORER' sidebar displays a file tree for a project named 'data_exploration.py'. Under the 'scripts' folder, there is a 'pdf_texts' subfolder containing numerous text files, including 'backprop.txt', 'BiasVarianceAnalysis.txt', 'cs229-linalg.txt', 'cs229-notes-deep_learn...', 'cs229-notes1.txt', 'cs229-notes2.txt' (selected), 'cs229-notes3.txt', 'cs229-notes4.txt', 'cs229-notes5.txt', 'cs229-notes7a.txt', 'cs229-notes7b.txt', 'cs229-notes8.txt', 'cs229-notes9.txt', 'cs229-notes10.txt', 'cs229-notes11.txt', 'cs229-notes12.txt', 'cs229-prob-slide.txt', 'cs229-prob.txt', 'gaussian_processes.txt', and 'gaussians.txt'. The main editor window displays the content of 'cs229-notes2.txt', which is a lecture note titled 'CS229 Lecture Notes' by Andrew Ng, covering 'Part IV: Generative Learning algorithms'. The text discusses learning algorithms that model $p(y|x; \theta)$, the conditional distribution of y given x , and mentions logistic regression and the perceptron algorithm. The terminal at the bottom shows a list of converted files: 'Converted: pdf_texts/cs229-notes1.txt', 'Converted: pdf_texts/cs229-notes10.txt', 'Converted: pdf_texts/cs229-notes4.txt', 'Converted: pdf_texts/cs229-notes11.txt', 'Converted: pdf_texts/cs229-prob.txt', 'Converted: pdf_texts/cs229-notes7b.txt', 'Converted: pdf_texts/ML-advice.txt', 'Converted: pdf_texts/cs229-notes3.txt', 'Converted: pdf_texts/MaxEnt.txt', and 'Converted: pdf_texts/gaussian_processes.txt'.

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