Introduction to Machine Learning
Module 4-c

What is your Class Project?

Team Structure & Roles

Overview of the Project Notebook

Options for Each Team

Expectations for Week 5

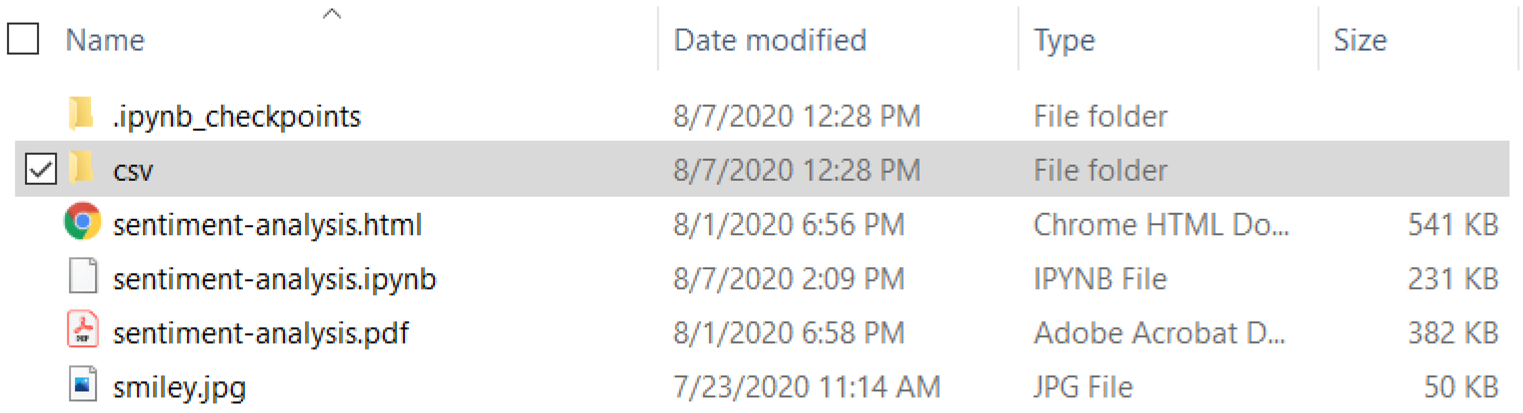
Detailed Overview of Project Notebook

CLASS PROJECT

Download the Project Folder

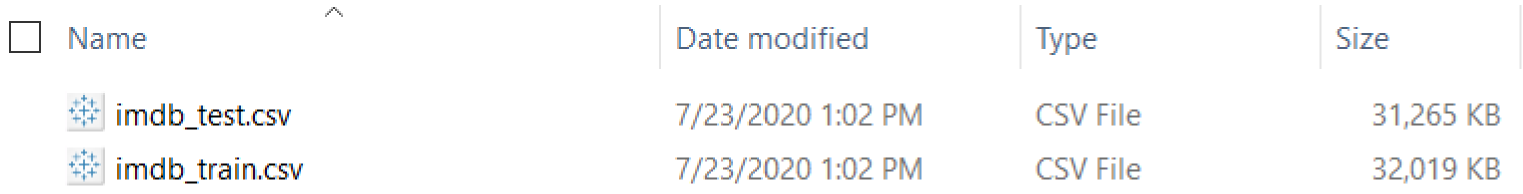
- Grab the project zip file – from Pragmatiq or Google drive link that will be sent to you;
- Unzip the contents carefully to a folder you use to download other notebook contents;

Main directory



□ Name	Date modified	Type	Size
.ipynb_checkpoints	8/7/2020 12:28 PM	File folder	
☑ csv	8/7/2020 12:28 PM	File folder	
sentiment-analysis.html	8/1/2020 6:56 PM	Chrome HTML Do...	541 KB
sentiment-analysis.ipynb	8/7/2020 2:09 PM	IPYNB File	231 KB
sentiment-analysis.pdf	8/1/2020 6:58 PM	Adobe Acrobat D...	382 KB
smiley.jpg	7/23/2020 11:14 AM	JPG File	50 KB

csv sub-directory



□ Name	Date modified	Type	Size
imdb_test.csv	7/23/2020 1:02 PM	CSV File	31,265 KB
imdb_train.csv	7/23/2020 1:02 PM	CSV File	32,019 KB

Class Project Description

- You are given a working Jupyter notebook that is capable of CLASSIFYING movie reviews into one of two categories:
  - 0 ➔ negative review
  - 1 ➔ positive review
  - Note – there is no such thing as a neutral review!
- You will have several tasks to extend this functionality significantly in the next 2 weeks!!

Project Notebook Overview

- Read the HTML file or PDF file in the folder;
- The explanations and references are useful;
- Data extraction / preparation is already done!
- The solution is based on “Bag o Words”;
- It uses unigram / bigram / tf-idf vectorization
- The main classification algorithm is SGD
- The incoming data set is huge!!
- This notebook tries SGD with different options (unigram, bigram, tf-idf)
- Outputs some statistics about performance

Learning Background for Project

- Python code – and use of functions
- Pandas – so you can follow data preparation (although you will not need to change it);
- Some numpy for slicing, etc.
- Sparse matrices (scipy)
- Vectorizing (unigram, bigram, tf-idf) options;
- Matplotlib and Seaborn for visualization

Focus now to accelerate your learning

REVIEW OF PROJECT NOTEBOOK

Composition of Teams, Roles, Responsibilities

PROJECT TEAMS

Class Project

- This is a TEAM Project
- Each of you is part of a team: A through F
  - A ➔ Bajaj, Suresh, Parvathini, Kalatheeswaran, Shaik
  - B ➔ Kamasini, Atla, Shah, Alaparthy, Agarwal
  - C ➔SampathKumar, Trivedi, Dingan, Madikonda, Parsi
  - D ➔ Agarwal, Moulee, Voona, Cunchala, Bolla
  - E ➔ Vasireddy, Dave, K Subramaniam, Kamasani, Rajnikanth
  - F ➔ Sankari, Narayanan, Goud, Cherukupalli, Debnath

Expectations

- Team → try to work together;
- A team of 5 is NOT 5 people working separately;
- Team Work is not easy – but it is important:
  - Communication throughout the next 2 weeks;
  - Different people have different skillsets;
  - Different people have different opinions;
  - Explaining and persuading is a critical life skill!
- This is a 2-week project – set realistic goals;
  - Don't be a hero – you can learn more afterwards;
  - Learn sincerely, work hard, and do your best.

How to Work **Together**

- Use Whatsapp to create a group for **your team**;
- Try to use **free Zoom or Google hangouts** to get together to do things in real-time at least two or three times each week to stay in sync!!
- Developing code as a team can be tricky:
  - – Who is going to work on which problem?
  - – How can you make sure integration is not messy?
  - – Better to plan together first; avoid later frustration.
- This can get messy unless you stay in touch with teammates and plan things out carefully!!!

This is what you will work on

PROJECT TASKS

General Guidance

- **Do not change** anything given to you already!
- That gives you a clear baseline for comparison
- **ADD new functions and code** at the end of the notebook provided.
- That way, you can keep things distinct and easy to demonstrate.

Project Tasks / Deliverables

Task Number	Task Description	Deliverable (what you need to show in 2 weeks)
1	**Add trigram / tf-idf** as an option (in addition to unigram/bigram)	Demonstrate the result for trigram / tf-idf – does it change the result? *
2	**Add StopWords to the algorithm** after researching good stop words and show the result visually.	Demonstrate the result after adding stopwords – does it change the result? *
3	Add a ***new_train_and_show_scores*** function that will try 3 different options for SGDClassifier by adding parameters. Choose the 3 parameters by reading the documentation for SGDClassifier	Show that the function can be called using the different parameters. *
4	Try to **provide integrated visual display** for various options in 1, 2, and 3 above. Idea is to accumulate results in arrays, and render them visually – in an integrated way (e.g., not separate graphs).	Demonstrate specific integrated visualizations to compare the different results side by side using different plots (either Matplotlib or Seaborn). E.g., one graph for confusion matrices for all runs, another for validation score / training score, etc.
5	Create a max 10-slide Project Presentation for your team in PowerPoint	PowerPoint template for Project Report will be posted at the Pragmatiq repository. Don't deviate from it – all slides should be filled in.

* *This deliverable may be adequately addressed if you include it as part of Item 4 (extended visualization)*

Suggested Assignment of Roles

- Everybody has inputs into everything;
- Establish clear accountability and timetable;
- Design and Strategy – everybody participate;
- Stopwords – one person research and provide results
- Python Programming – take different parts, discuss interfaces, implement, and integrate asap;
- Programming Testing and Validation – everybody should test and provide reports;
- PowerPoint presentation – one person with inputs and feedback;

There will be problems ☺

- Be patient – this is supposed to be fun!
- This is not a life-or-death project -- relax…
- Try to discuss problems and resolve them within the team:
  - Communicate
  - Negotiate
  - Compromise and resolve
- If there is a serious issue, reach out privately to one of the course staff, and we can address

Plans for Week 5

- **Week 5 will be done by Appointment** with each team – please be punctual;
- Purpose of the meeting is to **get a progress report from each team,** identify problems, and resolve them right away;
- 25 minutes per team as follows **(all times are EST):**
  - **Team A: 10:00 am ; Team B: 10:30 am;**
  - **Team C: 11:00 am ; Team D: 11:30 am;**
  - **Team E: 12:00 pm ; Team G: 12:30 pm;**
- Please login at the time shown. **BE PUNCTUAL**
- You get your Sunday back after your meeting!
- Be prepared to report on the status of your project (according to the template on the next page)

Project 1-week Status Report Template

- Create a **7-slide PowerPoint Status Report** with the following slide titles:
  - Team (A, B, C, D, E, or F) and Team members
  - Team Roles – who is doing what
  - Tasks completed
  - Tasks remaining / percentage completed / details
  - Technical Issues and Challenges
  - Team Challenges and Problems
  - Solutions for Successful Project Completion