AI Developer Homework Assignment

Problem

We have a lot of titles and abstracts of scientific papers, which we want to categorise based on their subject automatically. Sometimes the paper can have only one category, however, some of the papers cover multiple domains, which means they will have multiple subjects. Try to explore the data and build a model predicting the subject(s) of the papers based on their title and abstract. The second (bonus) objective is to prepare the models for inference by creating a web service with the models.

Data

You are provided with a file containing the data about the papers and their titles/abstracts - **cc_data.parquet**, a file for making predictions - **cc_test.parquet**. The names of the columns should give a hint about what they represent.

Tasks

1. Create a single Jupyter notebook, which shows all of your work. Submit this notebook with the output cells present.
2. Do an exploratory data analysis of the provided dataset.
3. Build a data preprocessing pipeline, which can be used for your machine learning model.
4. Build a machine learning model that predicts the subject(s) of the papers.
5. Focus on creating a fully working training pipeline. The quality of the predictions should be your second priority.
6. Use the model to predict the subjects for the items in the **cc_test.parquet** file. Export these predictions to **predictions.parquet** file.

Bonus Task

- Prepare the model/models for inference by creating a web service with the model/models. You can use any tool that you see fit for this task such as Django, Flask, TensorFlow Serving, etc.

A short documentation of design decisions and assumptions must be provided in the code itself or the README.

Requirements

- You should have Pickle (or similar) files for each of your trained models. There should be a clear and easy way to load these files to your Jupyter notebook and use them for predictions.
- Use Python 3 in this project.
- Make sure to export the list of Python modules to a **requirements.txt** file. Export modules that are only needed for your code to run.
- Submit your solution as a private GitHub repository and invite karolis.mariunas@cybercare.cc as a collaborator.

Recommendations

- Walk us through your thought process in the comments of your code.
- Your solution should use Python software development best practices.
- There should be an easy way to run your solution. README is a perfect place to provide instructions for how to run your solution.

Evaluation Criteria

The solution is evaluated based on these criteria (in order):

1. How many tasks were completed.
2. The demonstration of your machine learning skills.
3. The demonstration of your data analysis skills.
4. Your knowledge of software engineering best practices.