Data Engineer test assignment

This test is designed to evaluate your skills with machine learning (ML) models and workflows, with a focus on natural language processing (NLP) based on the scenario outlined below.

For this test you are expected to submit:

- The code you write to complete the test (described in Part 1)
- A document containing replies to questions on workflow description and general considerations on the task (described in Part 2)

HYPOTHETICAL SCENARIO

Imagine you are working in a company that provides services to academic institutions. One of the services the team plans to develop is to classify a corpus of academic papers in real-time. The team wants academic institutions to access the service via website, or mobile application. They can upload an archive of files in .zip or .tar and the files could be of different types - .doc, .docx, .pdf, .html, .txt. The team expects that they will start with a small number of users first (e.g., 3-4 users uploading the files daily), but plan to scale up to 30000 users per day. Assume that the average .zip or .tar archive will have around 20,000 papers to classify. You decide that you want to focus on abstracts' classification first. The team still has not decided on the product output, i.e., how and in what format results will be provided to the clients.

PART 1: PROTOTYPE THE MODEL

In this first part, you are asked to prototype a simple ML model. You can use a programming language and platform of your choice, as long as it makes it feasible to share the code with us. Share the code with us, together with the input data and instructions on how to run it (including installation of libraries).

- Setup a ML model that performs the following task: classify the topic of research articles of four categories – “Computer Science”, “Mathematics”, “Physics”, “Statistics” - based on their abstract.
    - Please use the train and test set provided to you as an attachment.
    - Please ignore the other categories that are part of the dataset.
- Include the computation of at least two evaluation metrics on the test dataset, one of which is the area under the ROC curve, and the other one a metric of your choice relevant to this problem. Please use these metrics to showcase the identification of the topic of “Statistics”.

Please note that the evaluation of Part 1 will rely more on the overall workflow rather than on the detailed performance of the model.

PART 2: WORKFLOW DESCRIPTION AND CONSIDERATIONS

In this second part, you are asked to provide the reasoning behind your choices in Part 1, and to think about possible evolutions of this model. Please share with us a document with replies to the following questions:

- Which programming language and platform do you choose for a prototype of the model? Why?
- How did you decide on the architecture of the model?
- Which metric, apart from the ROC curve, did you choose to evaluate performance, and why?
- What is the performance of your model? Please provide the value of ROC and your selected metric here, in particular for the performance when tagging the topic "Statistics".
- What are the main limitations that you have encountered in Part 1?
- Would the workflow you have designed be reliable also as the number of users will grow? Please motivate your reply.
- Given the scenario outlined, what deployment strategies do you suggest? Please detail examples of workflows/platforms/pipelines that you think could be relevant.
- How do you expect that the increase in data will impact the performance of your model? How could you verify that and what re-training strategies you could develop?
- What will be your strategy to handle different file types? Do you anticipate developing a standalone solution to the problem, or use available services?
    - Please argument your point