Task Description

We would like you to fine-tune the GPT-2 model on the works of Shakespeare, and host the model for conditional text generation on any cloud computing service of choice by using Huggingface's `Transformers` library and Fast API.

Submission Format:

We require a public git repository to test your implementation. Our test will be the following:

```SHELL
git clone https://your_git_link/your_repo_name
cd your_repo_name
pip install -r requirements.txt
bash start.sh
```

We will then test your Fast API endpoint and get generations by using a line of text as a prompt to the model.

Task Instructions

1. We're going to use the Shakespeare Dataset to finetune a GPT-2 model
You can use the free GPUs offered by Google Colab for this step: Colab

2. Once you've fine-tuned the model, write a Python application using FastAPI to host the model on a Linux/MacOS system.

(You do not need a GPU, you can use a CPU to host the model. We do not care about the speed of the server).

We require a single endpoint that takes a single string input to be feeded to the model. The end point should return a single string as response that is the complete sample sampled by GPT-2. You can use temperature sampling.

3. Please add your comments on any design, procedural, or engineering decisions you have taken to the README file in your repository. If you could not complete any step, please mention the reason and try to comment on possible solutions or alternative implementations.

Disclaimer
The task is only for test purposes, your code will **not** be used by Novus technologies..