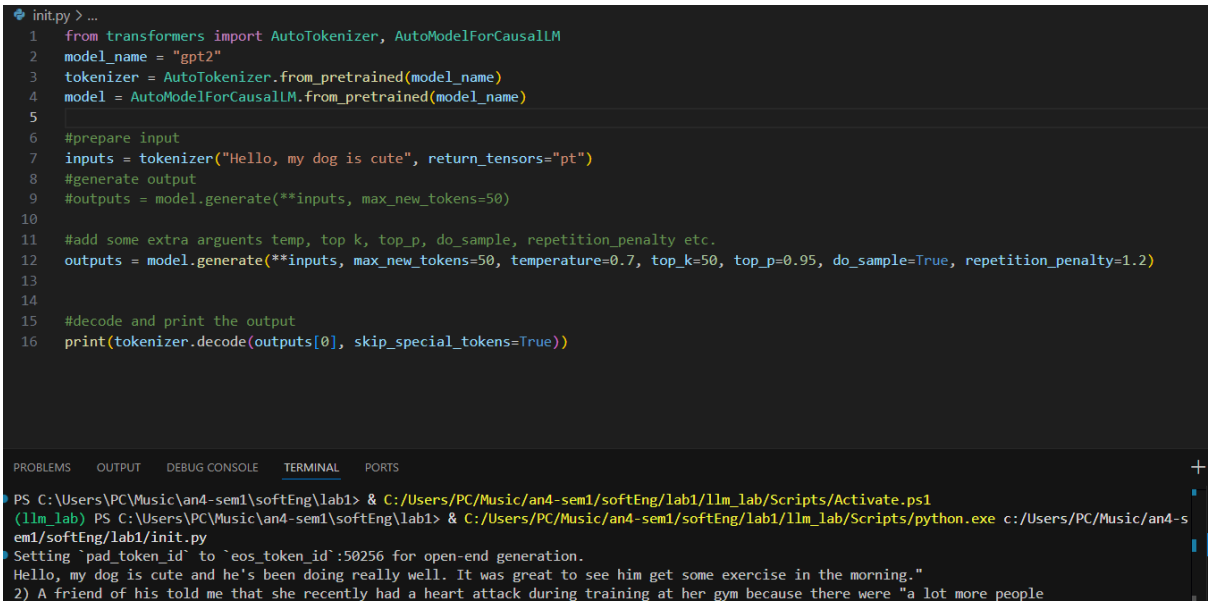
file title: **Solution for HW 1**
student: **FINICHIU Eduard - Adelin**
group: **1241 EB** (FILS)

!OBS: THE REPORT (EX3) IS WRITTEN AFTER THE SOLUTIONS FOR EX 1 AND EX 2

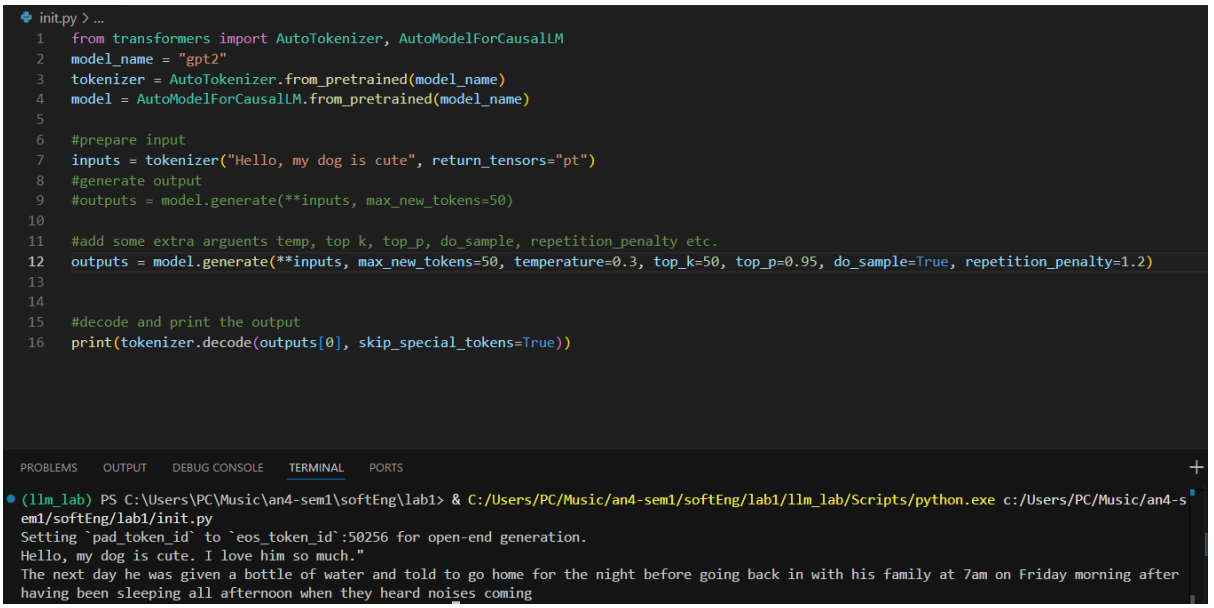
EXERCISE 1

1. **Try dif. temp. values and see which one gives best results!**

TEMP = 0.7



TEMP = 0.3



TEMP = 0.95

```python
from transformers import AutoTokenizer, AutoModelForCausalLM
model_name = "gpt2"
tokenizer = AutoTokenizer.from_pretrained(model_name)
model = AutoModelForCausalLM.from_pretrained(model_name)

#prepare input
inputs = tokenizer("Hello, my dog is cute", return_tensors="pt")
#generate output
#outputs = model.generate(**inputs, max_new_tokens=50)

#add some extra arguents temp, top k, top_p, do_sample, repetition_penalty etc.
outputs = model.generate(**inputs, max_new_tokens=50, temperature=0.95, top_k=50, top_p=0.95, do_sample=True, repetition_penalty=1.2)


#decode and print the output
print(tokenizer.decode(outputs[0], skip_special_tokens=True))
```

```
(llm_lab) PS C:\Users\PC\Music\an4-sem1\softEng\lab1> & C:/Users/PC/Music/an4-sem1/softEng/lab1/llm_lab/Scripts/python.exe c:/Users/PC/Music/an4-sem1/softEng/lab1/init.py
Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.
Hello, my dog is cute. And I'm sure there's something to this story we're going through tonight," he said during the phone call from his local veterinarian in North Carolina (I've only seen one person so far).
"It could be that she got bitten by
```

Conclusion: In my case, higher temperature gave more “normal” results, even though higher temperature usually means higher “creativity.

EXERCISE 2

2. **Run the code with 3 diff. models of your choice (keep them small)!**

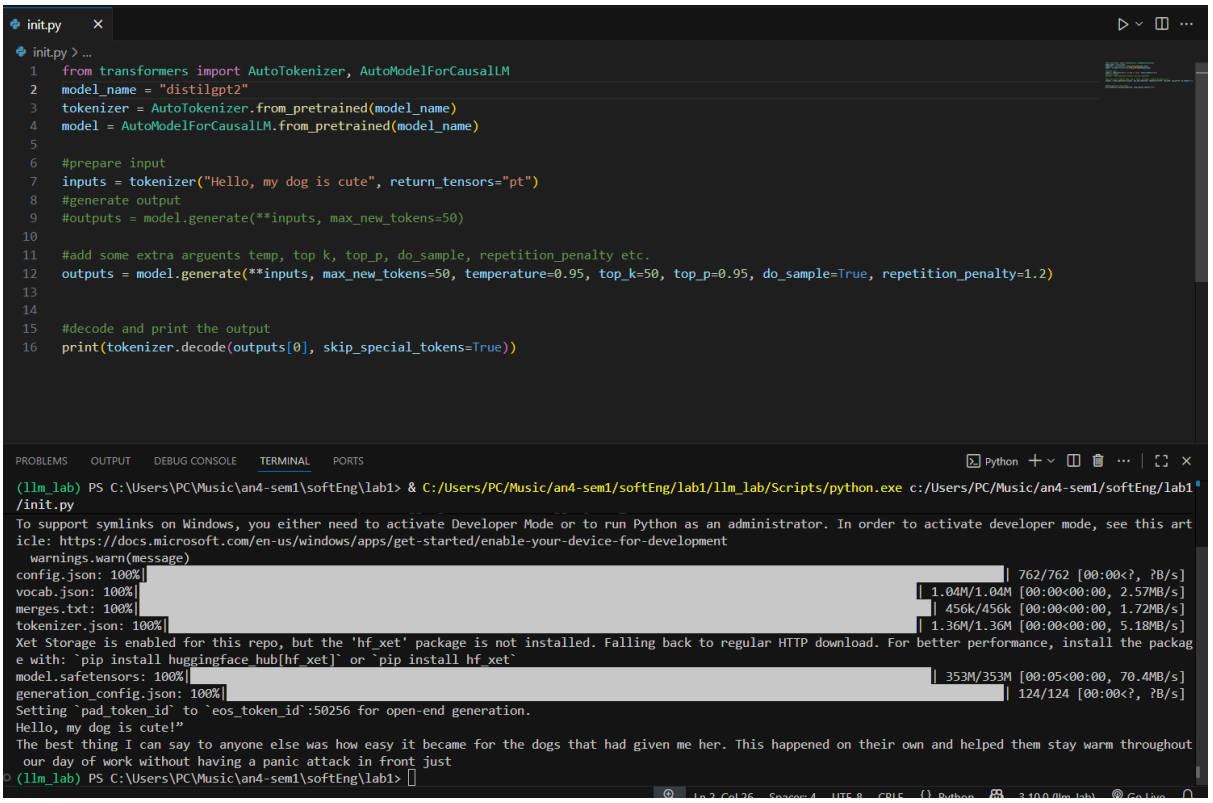
I will run 3 versions:

1. "gpt2" - **done in class**

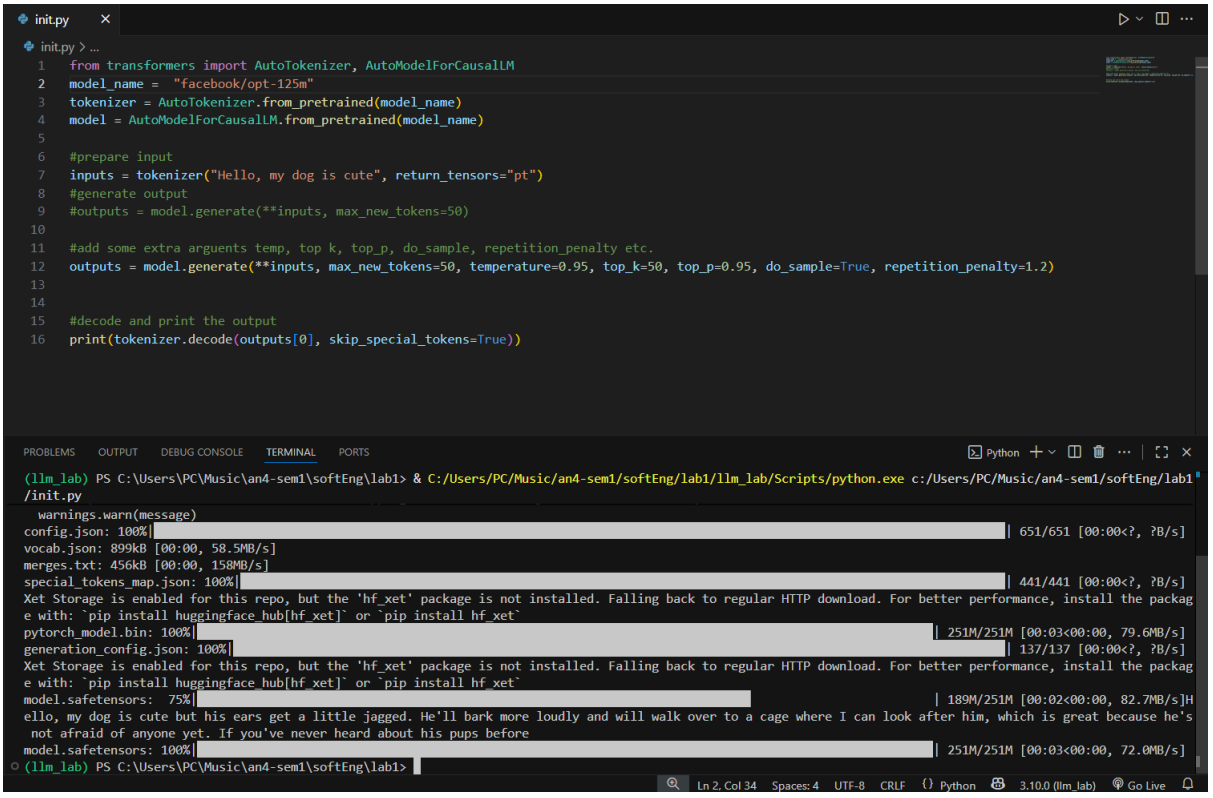

```python
from transformers import AutoTokenizer, AutoModelForCausalLM
model_name = "gpt2"
tokenizer = AutoTokenizer.from_pretrained(model_name)
model = AutoModelForCausalLM.from_pretrained(model_name)

#prepare input
inputs = tokenizer("Hello, my dog is cute", return_tensors="pt")
#generate output
#outputs = model.generate(**inputs, max_new_tokens=50)

#add some extra arguents temp, top k, top_p, do_sample, repetition_penalty etc.
outputs = model.generate(**inputs, max_new_tokens=50, temperature=0.95, top_k=50, top_p=0.95, do_sample=True, repetition_penalty=1.2)


#decode and print the output
print(tokenizer.decode(outputs[0], skip_special_tokens=True))
```

```
(llm_lab) PS C:\Users\PC\Music\an4-sem1\softEng\lab1> & C:/Users/PC/Music/an4-sem1/softEng/lab1/llm_lab/Scripts/python.exe c:/Users/PC/Music/an4-sem1/softEng/lab1/init.py
Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.
Hello, my dog is cute. And I'm sure there's something to this story we're going through tonight," he said during the phone call from his local veterinarian in North Carolina (I've only seen one person so far).
"It could be that she got bitten by
```

2. "distilgpt2" - found it onlinem lighter than gpt 2

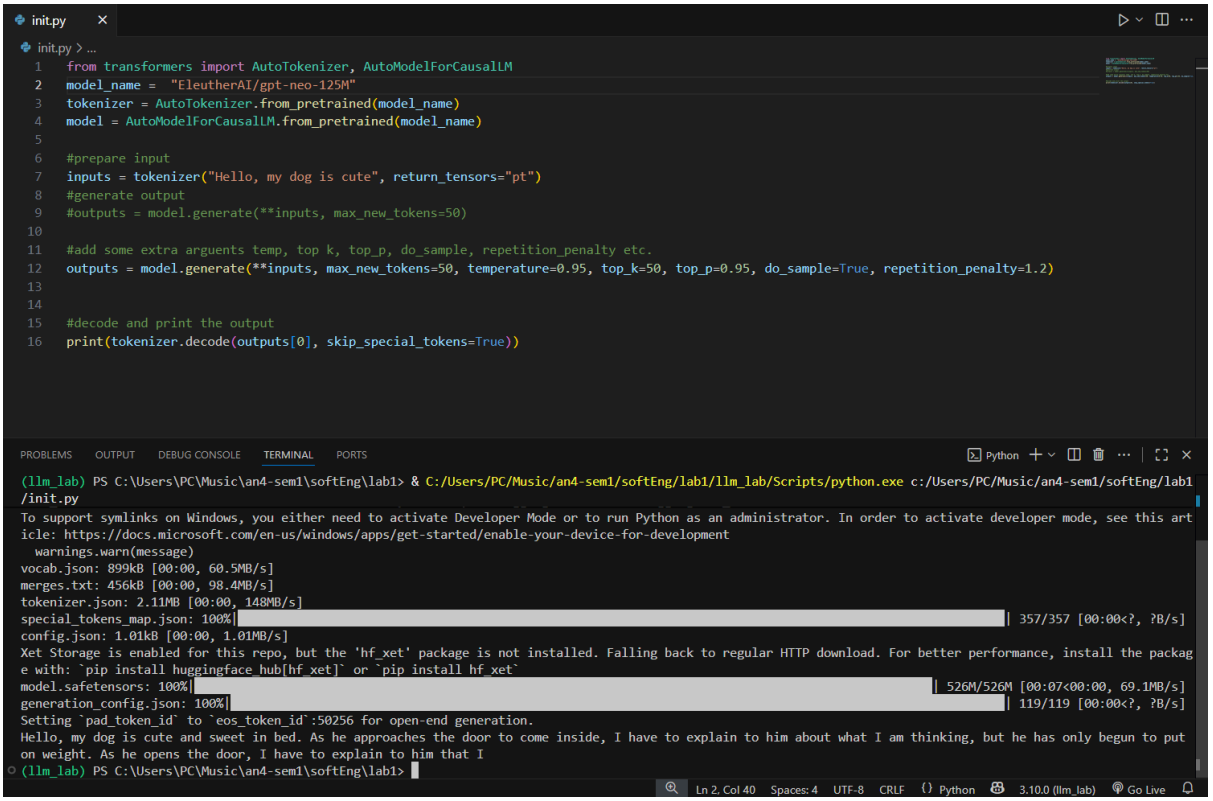


3. "facebook/opt-125m" - for sure the best results yet



4. "EleutherAI/gpt-neo-125M" - results seem to be ok

PHOTO OF RESULTS ON NEXT PAGE



EXERCISE 3

EX3. Write a brief report.

For this small experiment, I tested four causal language models using the same input text: “Hello, my dog is cute”.

The models were GPT-2, DistilGPT-2, OPT-125M, and GPT-Neo-125M. I used the same generation settings for all, except I also tried changing the temperature to see how it affects the results.The results are the following:

GPT-2: This model gave the most natural and fluent results. It continued the sentence in a way that felt like normal conversation. Sometimes it repeated phrases, but overall it sounded the most human.

DistilGPT-2: This is a smaller, faster version of GPT-2. It runs quickly and uses less memory, but the text it generates is usually shorter and sometimes cuts off mid-sentence. It’s less creative.

OPT-125M (Meta/Facebook)**:** Text from OPT felt a bit more “dry” or factual. It seems trained on different data, so the tone feels a bit off compared to GPT-2.

GPT-Neo-125M (EleutherAI)**:** This one produced longer and more imaginative text. It was sometimes interesting but also went off-topic more easily. It’s a bit unpredictable, but can sound creative when the temperature is higher. I like it, since I like creativity.

Temperature was set at values lower than 1 (can be seen in ex 1). With higher temperature, more creativity is allowed. This can lead to aberrations.

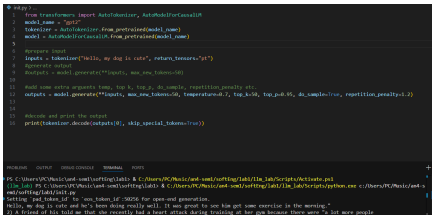
Because the models are light, sometimes (for me) higher temperature means smarter results (since low temperature can give basic responses who sometimes might not sound that smart).

Overall, GPT-2 was the best all-around model, having a good balance between fluency and creativity. DistilGPT-2 is faster and smaller but less expressive. OPT-125M feels

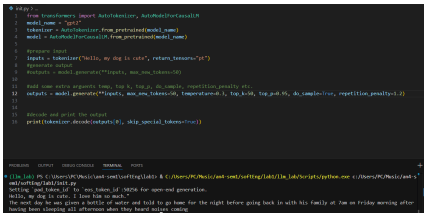
more serious and simple. I think OPT-125M is my favourite. GPT-Neo-125M is creative but not always coherent.

Changing the temperature really affects how the models behave: low values make them safe but boring, and high values make them imaginative but less logical. As I have previously said, I like when models behave in a more creative manner.

TEMP = 0.7



TEMP = 0.3



TEMP = 0.95

distilGPT vs Meta