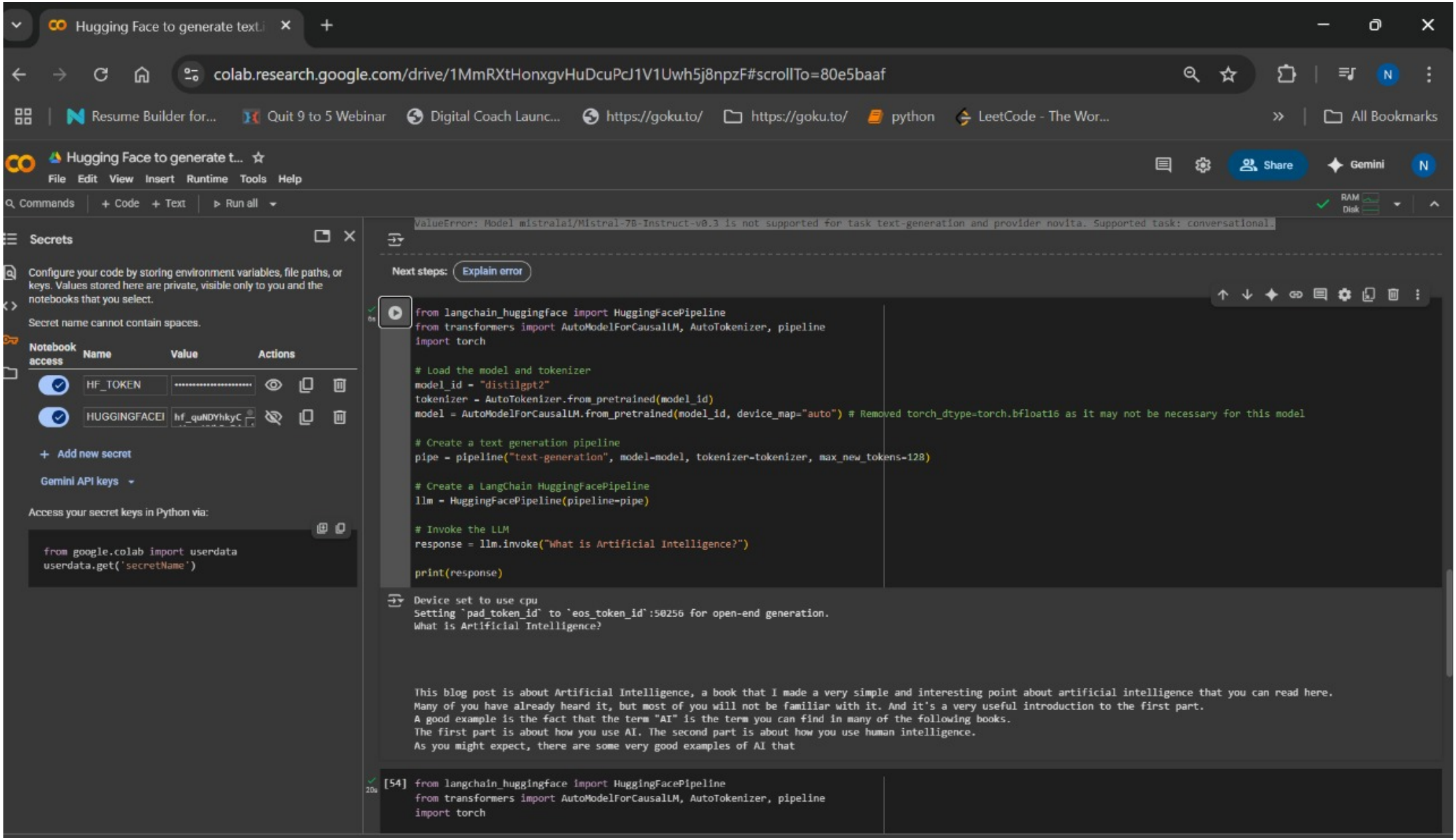
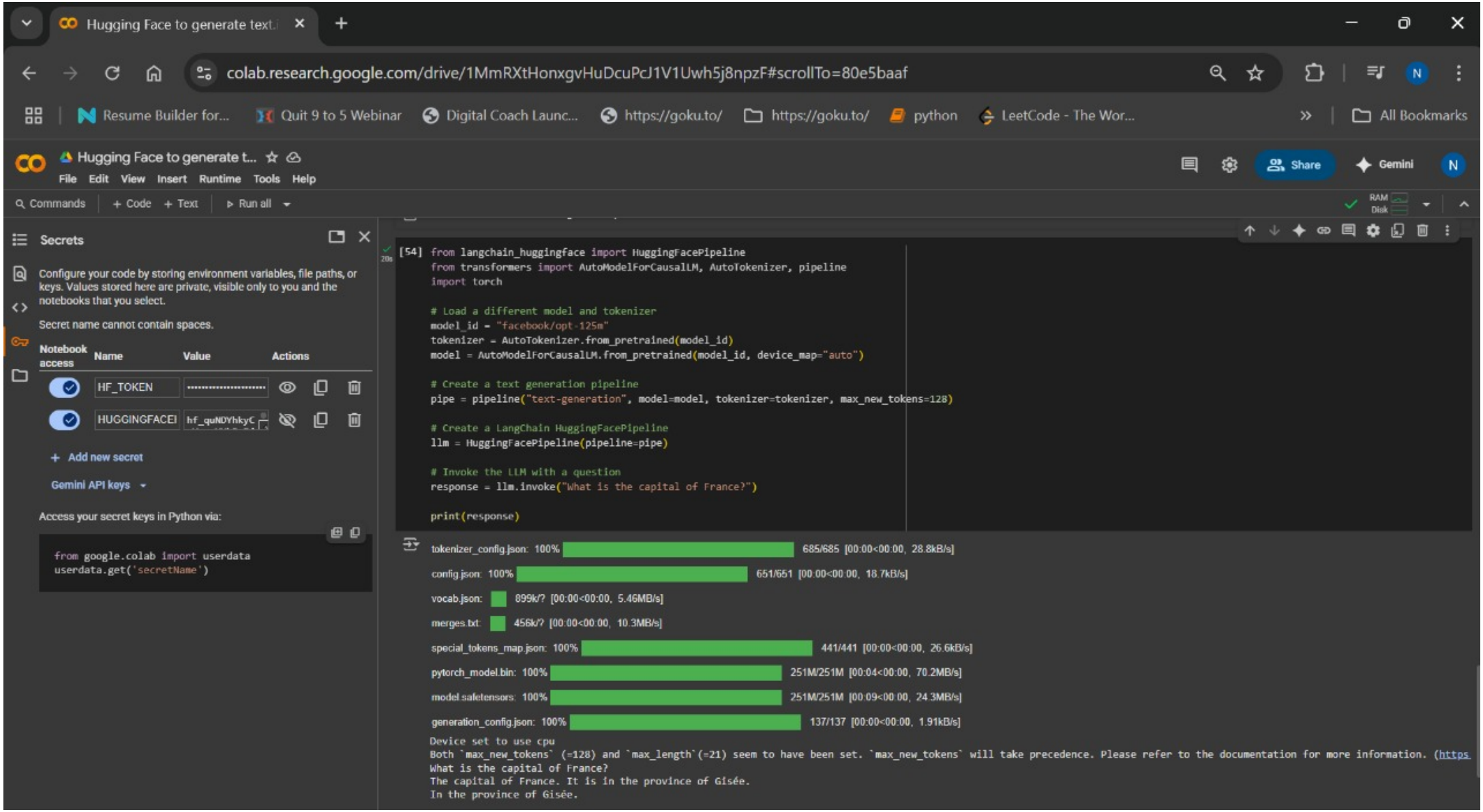

Hugging Face to generate text
colab.research.google.com/drive/1MmRXtHonxgvHuDcuPcJ1V1Uwh5j8npzF#scrollTo=80e5baaf
Resume Builder for...
Quit 9 to 5 Webinar
Digital Coach Launc...
https://goku.to/
https://goku.to/
python
LeetCode - The Wor...
All Bookmarks
Hugging Face to generate t...
File Edit View Insert Runtime Tools Help
Share
Gemini
Commands
+ Code
+ Text
Run all
RAM
Disk
Secrets
Configure your code by storing environment variables, file paths, or keys. Values stored here are private, visible only to you and the notebooks that you select.
Secret name cannot contain spaces.
Notebook access
Name
Value
Actions
HF_TOKEN
HUGGINGFACEI
hf_quNDYhkyC
Add new secret
Gemini API keys
Access your secret keys in Python via:
from google.colab import userdata
userdata.get('secretName')
ValueError: Model mistralai/Mistral-7B-Instruct-v0.3 is not supported for task text-generation and provider novita. Supported task: conversational.
Next steps:
Explain error
from langchain_huggingface import HuggingFacePipeline
from transformers import AutoModelForCausalLM, AutoTokenizer, pipeline
import torch
# Load the model and tokenizer
model_id = "distilgpt2"
tokenizer = AutoTokenizer.from_pretrained(model_id)
model = AutoModelForCausalLM.from_pretrained(model_id, device_map="auto") # Removed torch_dtype=torch.bfloat16 as it may not be necessary for this model
# Create a text generation pipeline
pipe = pipeline("text-generation", model=model, tokenizer=tokenizer, max_new_tokens=128)
# Create a LangChain HuggingFacePipeline
llm = HuggingFacePipeline(pipeline=pipe)
# Invoke the LLM
response = llm.invoke("What is Artificial Intelligence?")
print(response)
Device set to use cpu
Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.
What is Artificial Intelligence?
This blog post is about Artificial Intelligence, a book that I made a very simple and interesting point about artificial intelligence that you can read here.
Many of you have already heard it, but most of you will not be familiar with it. And it's a very useful introduction to the first part.
A good example is the fact that the term "AI" is the term you can find in many of the following books.
The first part is about how you use AI. The second part is about how you use human intelligence.
As you might expect, there are some very good examples of AI that
[54] from langchain_huggingface import HuggingFacePipeline
from transformers import AutoModelForCausalLM, AutoTokenizer, pipeline
import torch


Hugging Face to generate text.i
colab.research.google.com/drive/1MmRXtHonxgvHuDcuPcJ1V1Uwh5j8npzF#scrollTo=80e5baaf
Resume Builder for...
Quit 9 to 5 Webinar
Digital Coach Launc...
https://goku.to/
https://goku.to/
python
LeetCode - The Wor...
All Bookmarks
Hugging Face to generate t...
File Edit View Insert Runtime Tools Help
Share
Gemini
Commands
+ Code
+ Text
Run all
RAM
Disk
Secrets
Configure your code by storing environment variables, file paths, or keys. Values stored here are private, visible only to you and the notebooks that you select.
Secret name cannot contain spaces.
Notebook access
Name
Value
Actions
HF_TOKEN
HUGGINGFACEI
hf_quNDYhkyC
Add new secret
Gemini API keys
Access your secret keys in Python via:
from google.colab import userdata
userdata.get('secretName')
[54]
from langchain_huggingface import HuggingFacePipeline
from transformers import AutoModelForCausalLM, AutoTokenizer, pipeline
import torch
# Load a different model and tokenizer
model_id = "facebook/opt-125m"
tokenizer = AutoTokenizer.from_pretrained(model_id)
model = AutoModelForCausalLM.from_pretrained(model_id, device_map="auto")
# Create a text generation pipeline
pipe = pipeline("text-generation", model=model, tokenizer=tokenizer, max_new_tokens=128)
# Create a LangChain HuggingFacePipeline
llm = HuggingFacePipeline(pipeline=pipe)
# Invoke the LLM with a question
response = llm.invoke("What is the capital of France?")
print(response)
tokenizer_config.json: 100% 685/685 [00:00<00:00, 28.8kB/s]
config.json: 100% 651/651 [00:00<00:00, 18.7kB/s]
vocab.json: 899k/? [00:00<00:00, 5.46MB/s]
merges.txt: 456k/? [00:00<00:00, 10.3MB/s]
special_tokens_map.json: 100% 441/441 [00:00<00:00, 26.6kB/s]
pytorch_model.bin: 100% 251M/251M [00:04<00:00, 70.2MB/s]
model.safetensors: 100% 251M/251M [00:09<00:00, 24.3MB/s]
generation_config.json: 100% 137/137 [00:00<00:00, 1.91kB/s]
Device set to use cpu
Both `max_new_tokens` (=128) and `max_length`(=21) seem to have been set. `max_new_tokens` will take precedence. Please refer to the documentation for more information. (https
What is the capital of France?
The capital of France. It is in the province of Gisée.
In the province of Gisée.