ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DIPARTIMENTO DI
SCIENZE STATISTICHE "PAOLO FORTUNATI"

10 - Hugging Face and Transformers Library

Giovanni Della Lunga
giovanni.dellalunga@unibo.it

Halloween Conference in Quantitative Finance

Bologna - October 25-26-27, 2023

Introduction

What is Hugging Face?

- Hugging Face is a community specializing in Natural Language Processing (NLP) and artificial intelligence (AI).
- Founded in 2016, the company has made significant contributions to the field of NLP by democratizing access to state-of-the-art machine learning models and tools.

What is Hugging Face?

- Hugging Face has a strong community focus.
- They provide a platform where researchers and developers can share their trained models, thereby fostering collaboration and accelerating progress in the field.

Hugging Face Library

Hugging Face Transformers as a Library

- Hugging Face‘s Transformers library is an open-source library for NLP and machine learning.
- It provides a wide variety of pre-trained models and architectures like BERT, GPT-X, T5, and many others.
- The library is designed to be highly modular and easy to use, allowing for the quick development of both research and production projects.
- It supports multiple languages and tasks like text classification, question-answering, text generation, translation, and more.

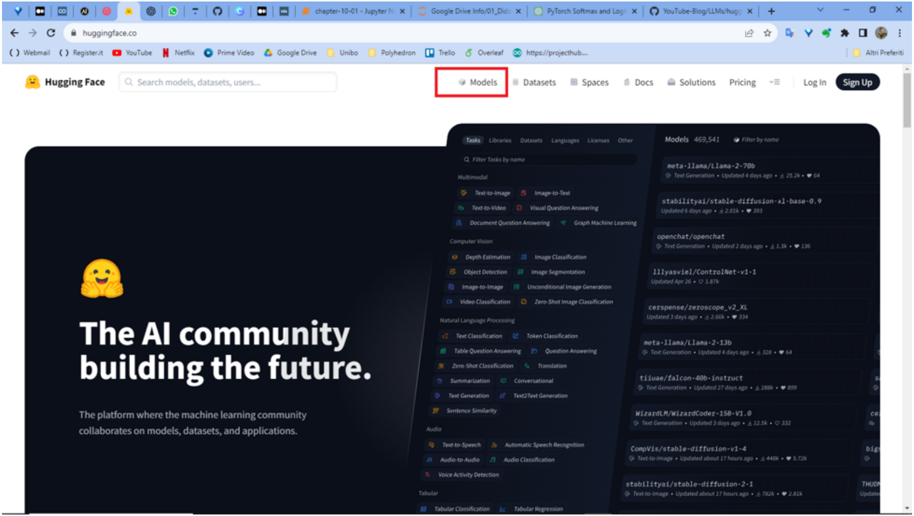
Transformers

- Transformers is a Python library that makes downloading and training state-of-the-art ML models easy.
- Although it was initially made for developing language models, its functionality has expanded to include models for computer vision, audio processing, and beyond.
- Two big strengths of this library are: 1) it easily integrates with Hugging Face's Models, Datasets, and Spaces repositories, and 2) the library supports other popular ML frameworks such as PyTorch and TensorFlow.
- This results in a simple and flexible all-in-one platform for downloading, training, and deploying machine learning models and apps.

Models

- There is a massive repository of pre-trained models available on Hugging Face (more than 250000 at the time of writing this).
- Almost all these models can be easily used via Transformers.

Models



The AI community building the future.

The platform where the machine learning community collaborates on models, datasets, and applications.

meta-llama/Llama-2-70b
Text Generation • Updated 4 days ago • 25.2k • 64

stabilityai/stable-diffusion-xl-base-0.9
Updated 6 days ago • 2.01k • 393

openchat/openchat
Text Generation • Updated 2 days ago • 1.3k • 136

lllyasviel/ControlNet-v1-1
Updated Apr 26 • 1.87k

cerspense/zeroscope_v2_XL
Updated 3 days ago • 2.66k • 334

meta-llama/Llama-2-13b
Text Generation • Updated 4 days ago • 328 • 64

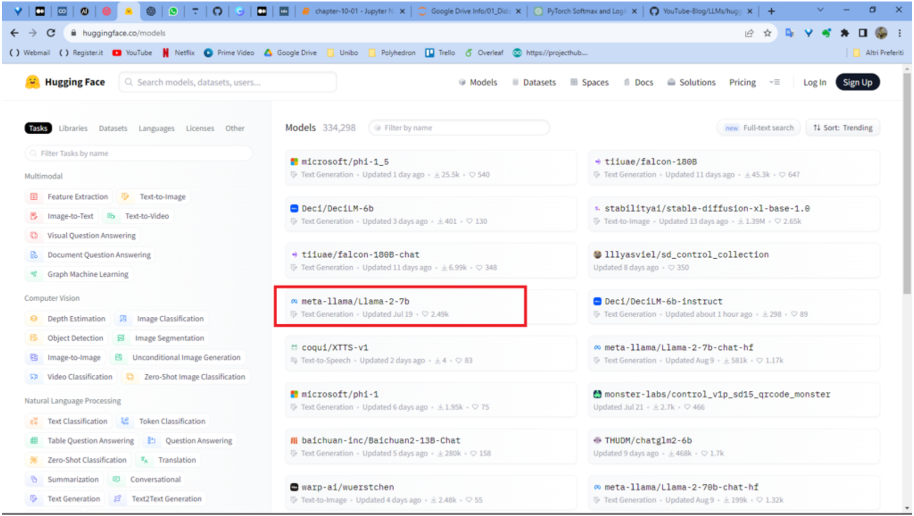
tiiuae/falcon-40b-instruct
Text Generation • Updated 27 days ago • 288k • 899

WizardLM/WizardCoder-15B-V1.0
Text Generation • Updated 3 days ago • 12.5k • 332

CompVis/stable-diffusion-v1-4
Text-to-Image • Updated about 17 hours ago • 448k • 5.72k

stabilityai/stable-diffusion-2-1
Text-to-Image • Updated about 17 hours ago • 782k • 2.81k

Models



Tasks Libraries Datasets Languages Licenses Other

Filter Tasks by name

Multimodal

- Feature Extraction
- Text-to-Image
- Image-to-Text
- Text-to-Video
- Visual Question Answering
- Document Question Answering
- Graph Machine Learning

Computer Vision

- Depth Estimation
- Image Classification
- Object Detection
- Image Segmentation
- Image-to-Image
- Unconditional Image Generation
- Video Classification
- Zero-Shot Image Classification

Natural Language Processing

- Text Classification
- Token Classification
- Table Question Answering
- Question Answering
- Zero-Shot Classification
- Translation
- Summarization
- Conversational
- Text Generation
- Text2Text Generation

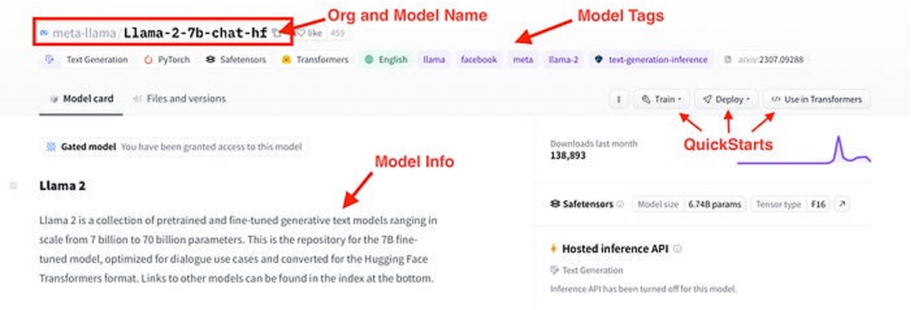
Models 334,298 Filter by name new Full-text search Sort: Trending

- microsoft/phi-1_5 · Text Generation • Updated 1 day ago • 25.5k • 540
- tiiuae/falcon-180B · Text Generation • Updated 11 days ago • 45.3k • 647
- Deci/DeciLM-6b · Text Generation • Updated 3 days ago • 401 • 130
- stabilityai/stable-diffusion-xl-base-1.0 · Text-to-Image • Updated 13 days ago • 1.39M • 2.65k
- tiiuae/falcon-180B-chat · Text Generation • Updated 11 days ago • 6.99k • 348
- lllyasviel/sd_control_collection · Updated 8 days ago • 350
- meta-llama/Llama-2-7b · Text Generation • Updated Jul 19 • 2.49k
- Deci/DeciLM-6b-instruct · Text Generation • Updated about 1 hour ago • 298 • 89
- coqui/XTTS-v1 · Text-to-Speech • Updated 2 days ago • 4 • 83
- meta-llama/Llama-2-7b-chat-hf · Text Generation • Updated Aug 9 • 581k • 1.17k
- microsoft/phi-1 · Text Generation • Updated 6 days ago • 1.95k • 75
- monster-labs/control_v1p_sd15_qrcode_monster · Updated Jul 21 • 2.7k • 466
- baichuan-inc/Baichuan2-13B-Chat · Text Generation • Updated 5 days ago • 280k • 158
- THUDM/chatglm2-6b · Updated 9 days ago • 468k • 1.7k
- warp-ai/wuerstchen · Text-to-Image • Updated 4 days ago • 2.48k • 55
- meta-llama/Llama-2-70b-chat-hf · Text Generation • Updated Aug 9 • 199k • 1.32k

Models

- To see what navigating the repository looks like, let's consider an example.
- Say we want a model that can do text generation, but we want it to be available via the Transformers library so we can use it in one line of code (as we did above).
- We can easily view all models that fit these criteria using the "Tasks" and "Libraries" filters.
- A model that meets these criteria is the newly released Llama 2. More specifically, Llama-2-7b-chat-hf, which is a model in the Llama 2 family with about 7 billion parameters, optimized for chat, and in the Hugging Face Transformers format.
- We can get more information about this model via its model card, which is shown in the next slide.

Models



Org and Model Name

Model Tags

QuickStarts

Model Info

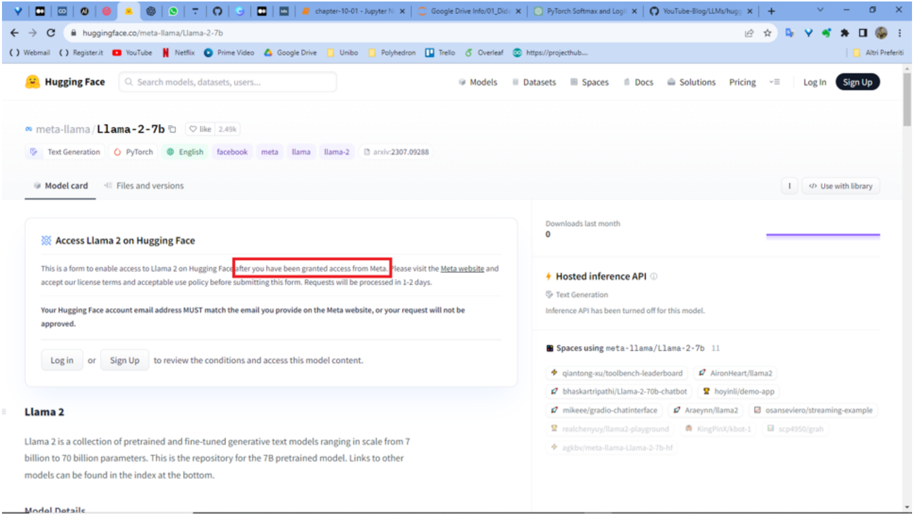
Gated model You have been granted access to this model

Llama 2

Llama 2 is a collection of pretrained and fine-tuned generative text models ranging in scale from 7 billion to 70 billion parameters. This is the repository for the 7B fine-tuned model, optimized for dialogue use cases and converted for the Hugging Face Transformers format. Links to other models can be found in the index at the bottom.

Inference API has been turned off for this model.

Models



Access Llama 2 on Hugging Face

This is a form to enable access to Llama 2 on Hugging Face after you have been granted access from Meta. Please visit the Meta website and accept our license terms and acceptable use policy before submitting this form. Requests will be processed in 1-2 days.

Your Hugging Face account email address MUST match the email you provide on the Meta website, or your request will not be approved.

Log in or Sign Up to review the conditions and access this model content.

Llama 2

Llama 2 is a collection of pretrained and fine-tuned generative text models ranging in scale from 7 billion to 70 billion parameters. This is the repository for the 7B pretrained model. Links to other models can be found in the index at the bottom.

Model Details

Downloads last month
0

Hosted inference API

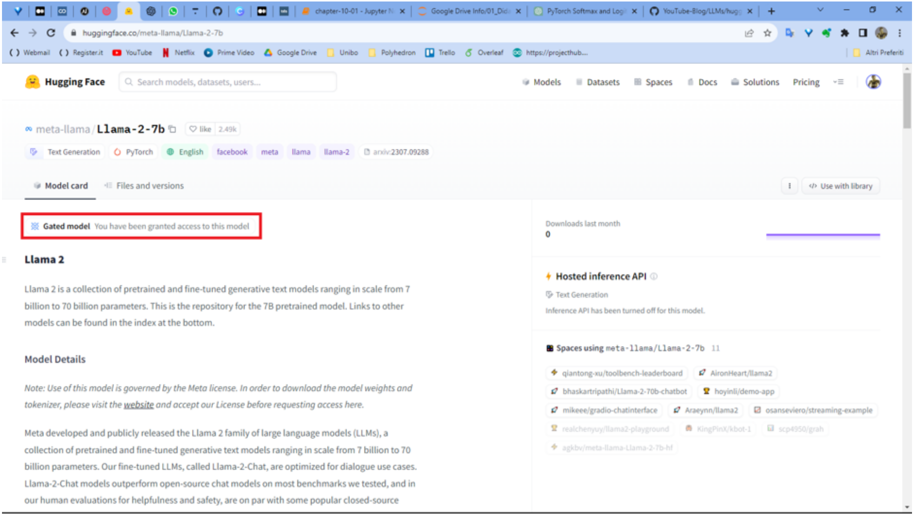
Text Generation

Inference API has been turned off for this model.

Spaces using meta-llama/Llama-2-7b 11

qiantong-xu/toolbench-leaderboard
AironHeart/llama2
bhaskartripathi/Llama-2-70b-chatbot
hoyinli/demo-app
mikeee/gradio-chatinterface
Araeynn/llama2
osanseviero/streaming-example
realchenyuy/llama2-playground
KingPinX/kbot-1
scp4950/grah
agkbv/meta-llama-Llama-2-7b-hf

Models



Gated model You have been granted access to this model

Llama 2

Llama 2 is a collection of pretrained and fine-tuned generative text models ranging in scale from 7 billion to 70 billion parameters. This is the repository for the 7B pretrained model. Links to other models can be found in the index at the bottom.

Model Details

Note: Use of this model is governed by the Meta license. In order to download the model weights and tokenizer, please visit the website and accept our License before requesting access here.

Meta developed and publicly released the Llama 2 family of large language models (LLMs), a collection of pretrained and fine-tuned generative text models ranging in scale from 7 billion to 70 billion parameters. Our fine-tuned LLMs, called Llama-2-Chat, are optimized for dialogue use cases. Llama-2-Chat models outperform open-source chat models on most benchmarks we tested, and in our human evaluations for helpfulness and safety, are on par with some popular closed-source

Downloads last month
0

Hosted inference API

Text Generation

Inference API has been turned off for this model.

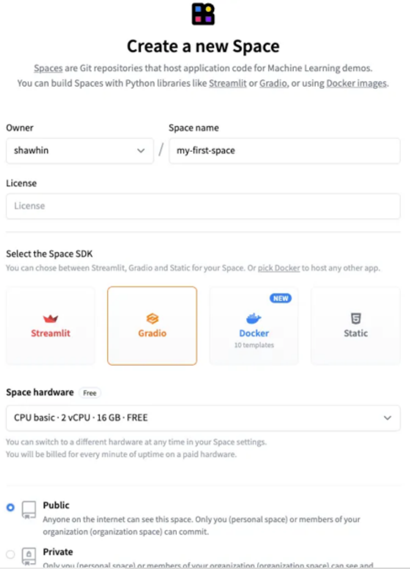
Spaces using `meta-llama/Llama-2-7b` 11

- qiantong-xu/toolbench-leaderboard
- AironHeart/llama2
- bhaskartripathi/Llama-2-70b-chatbot
- hoyinli/demo-app
- mikeee/gradio-chatinterface
- Araeynn/llama2
- osanseviero/streaming-example
- realchenyuy/llama2-playground
- KingPinX/kbot-1
- scp4950/grah
- agkbv/meta-llama-Llama-2-7b-hf

Hugging Face Spaces

Hugging Face Spaces

- To go one step further, we can quickly deploy this UI via Hugging Face Spaces.
- These are Git repositories hosted by Hugging Face and augmented by computational resources.



Create a new Space

Spaces are Git repositories that host application code for Machine Learning demos.
You can build Spaces with Python libraries like Streamlit or Gradio, or using Docker images.

Owner: shawhin / Space name: my-first-space

License: License

Select the Space SDK

You can chose between Streamlit, Gradio and Static for your Space. Or pick Docker to host any other app.

Space hardware Free

CPU basic · 2 vCPU · 16 GB · FREE

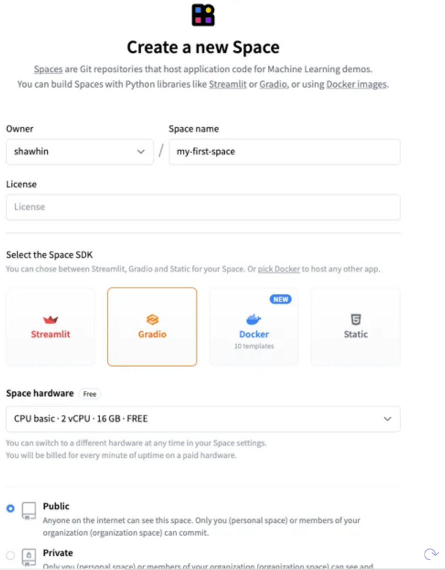
You can switch to a different hardware at any time in your Space settings.
You will be billed for every minute of uptime on a paid hardware.

● **Public**
Anyone on the internet can see this space. Only you (personal space) or members of your organization (organization space) can commit.

○ **Private**
Only you (personal space) or members of your organization (organization space) can see and

Hugging Face Spaces

- To make a new Space, we first go to the Spaces page and click "Create new space".
- Then, configure the Space by giving it the name e.g. "my-first-space" and selecting Gradio as the SDK. Then hit "Create Space".



Create a new Space

Spaces are Git repositories that host application code for Machine Learning demos.
You can build Spaces with Python libraries like Streamlit or Gradio, or using Docker images.

Owner: shawhin / Space name: my-first-space

License: License

Select the Space SDK

You can chose between Streamlit, Gradio and Static for your Space. Or pick Docker to host any other app.

Space hardware Free

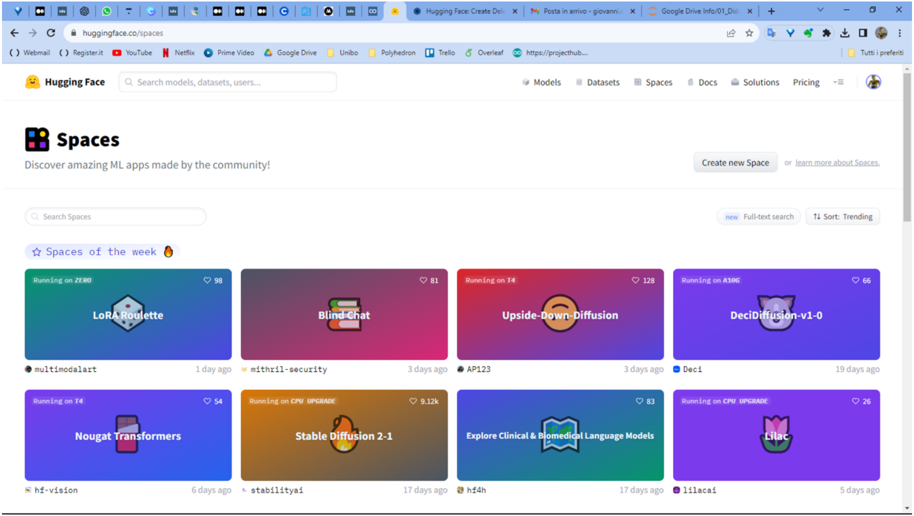
CPU basic · 2 vCPU · 16 GB · FREE

You can switch to a different hardware at any time in your Space settings.
You will be billed for every minute of uptime on a paid hardware.

● **Public**
Anyone on the internet can see this space. Only you (personal space) or members of your organization (organization space) can commit.

○ **Private**
Only you (personal space) or members of your organization (organization space) can see and

Hugging Face Saces



Spaces

Discover amazing ML apps made by the community!

Create new Space or learn more about Spaces.

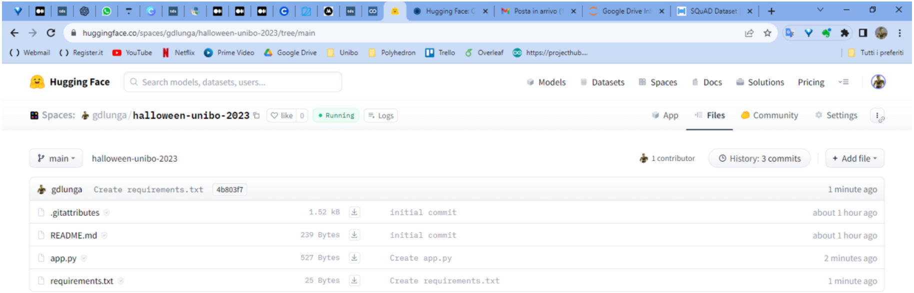
Spaces of the week 🔥

Space	Status	Likes	Author	Updated
LoRA Roulette	Running on ZERO	98	multimodalart	1 day ago
Blind Chat		81	mithril-security	3 days ago
Upside-Down-Diffusion	Running on T4	128	AP123	3 days ago
DeciDiffusion-v1-0	Running on A10G	66	Deci	19 days ago
Nougat Transformers	Running on T4	54	hf-vision	6 days ago
Stable Diffusion 2-1	Running on CPU UPGRADE	9.12k	stabilityai	17 days ago
Explore Clinical & Biomedical Language Models		83	hf4h	17 days ago
Lilac	Running on CPU UPGRADE	26	lilacai	5 days ago

Hugging Face Spaces

- Next, we need to upload **app.py** and **requirements.txt** files to the Space.
- **The app.py file houses the code we used to generate the Gradio UI, and the requirements.txt file specifies the app‘s dependencies**.
- Finally, we push the code to the Space just like we would to GitHub. The end result is a public application hosted on Hugging Face Spaces.

Hugging Face Spaces



huggingface.co/spaces/gdlunga/halloween-unibo-2023/tree/main

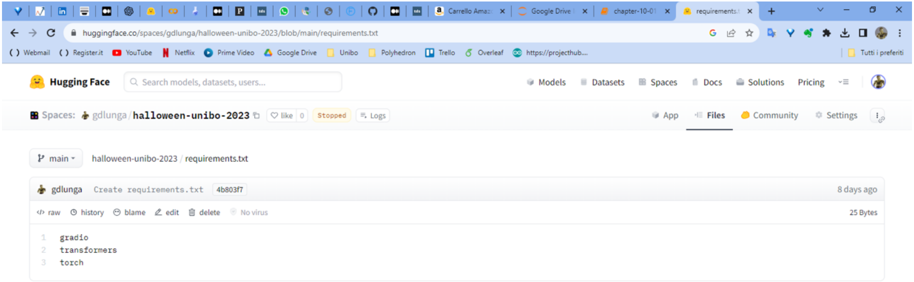
gdlunga / **halloween-unibo-2023**

halloween-unibo-2023

gdlunga `Create requirements.txt` 4b803f7 1 minute ago

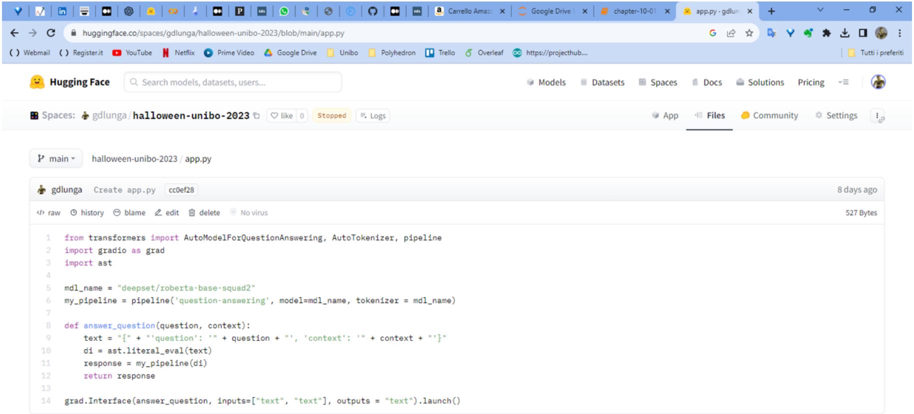
.gitattributes	1.52 kB	initial commit	about 1 hour ago
README.md	239 Bytes	initial commit	about 1 hour ago
app.py	527 Bytes	Create app.py	2 minutes ago
requirements.txt	25 Bytes	Create requirements.txt	1 minute ago

Hugging Face Spaces


huggingface.co/spaces/gdlunga/halloween-unibo-2023/blob/main/requirements.txt
Hugging Face
Search models, datasets, users...
Models
Datasets
Spaces
Docs
Solutions
Pricing
Spaces: gdlunga / halloween-unibo-2023
like 0
Stopped
Logs
App
Files
Community
Settings
main
halloween-unibo-2023 / requirements.txt
gdlunga
Create requirements.txt
4b803f7
8 days ago
raw
history
blame
edit
delete
No virus
25 Bytes

```
gradio
transformers
torch
```

Hugging Face Spaces


huggingface.co/spaces/gdlunga/halloween-unibo-2023/blob/main/app.py
Hugging Face
Search models, datasets, users...
Models
Datasets
Spaces
Docs
Solutions
Pricing
Spaces: gdlunga / halloween-unibo-2023
like 0
Stopped
Logs
App
Files
Community
Settings
main
halloween-unibo-2023 / app.py
gdlunga Create app.py cc0ef28
8 days ago
raw
history
blame
edit
delete
No virus
527 Bytes

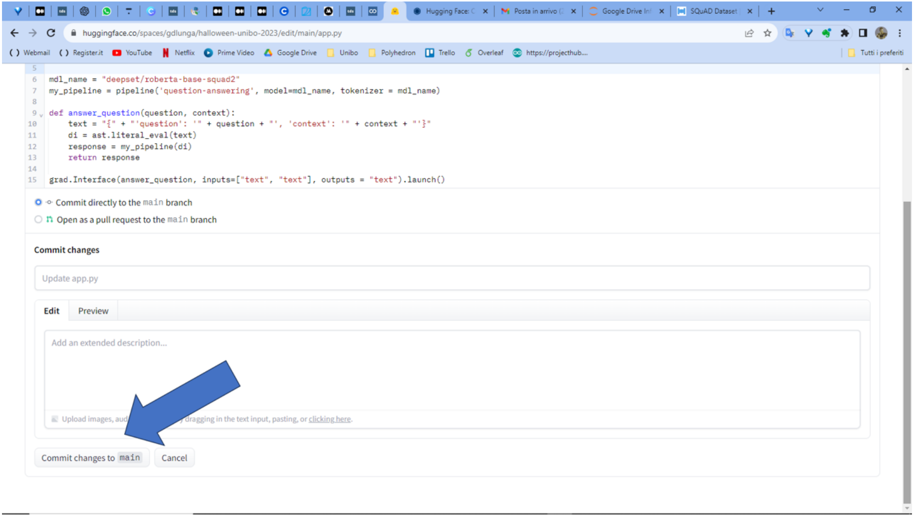
```python
from transformers import AutoModelForQuestionAnswering, AutoTokenizer, pipeline
import gradio as grad
import ast

mdl_name = "deepset/roberta-base-squad2"
my_pipeline = pipeline('question-answering', model=mdl_name, tokenizer = mdl_name)

def answer_question(question, context):
    text = "{" + "'question': '" + question + "', 'context': '" + context + "'}"
    di = ast.literal_eval(text)
    response = my_pipeline(di)
    return response

grad.Interface(answer_question, inputs=["text", "text"], outputs = "text").launch()
```

Hugging Face Spaces



huggingface.co/spaces/gdlunga/halloween-unibo-2023/edit/main/app.py

```
mdl_name = "deepset/roberta-base-squad2"
my_pipeline = pipeline('question-answering', model=mdl_name, tokenizer = mdl_name)

def answer_question(question, context):
    text = "{" + "'question': '" + question + "', 'context': '" + context + "'}"
    di = ast.literal_eval(text)
    response = my_pipeline(di)
    return response

grad.Interface(answer_question, inputs=["text", "text"], outputs = "text").launch()
```

● Commit directly to the main branch

○ Open as a pull request to the main branch

Commit changes

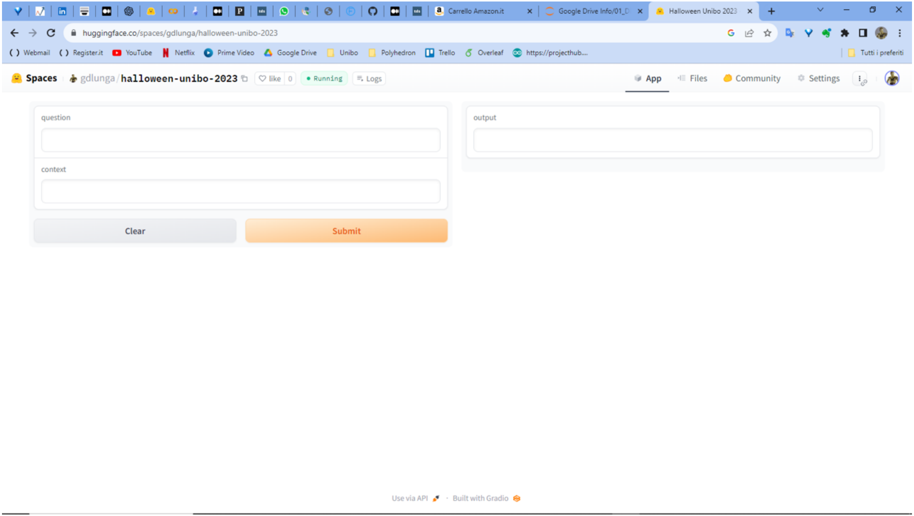
Update app.py

Edit | Preview

Add an extended description...

Upload images, aud dragging in the text input, pasting, or clicking here.

Hugging Face Spaces



huggingface.co/spaces/gdlunga/halloween-unibo-2023

Spaces · gdlunga / halloween-unibo-2023 · like 0 · Running · Logs

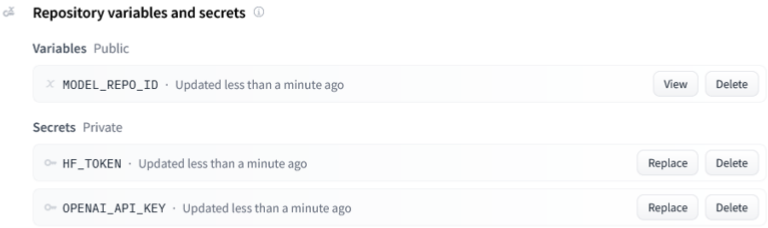
question

context

output

Managing secrets and environment variables

- If your app requires environment variables (for instance, secret keys or tokens), do not hard-code them inside your app!
- Instead, go to the **Settings** page of your Space repository and add a new variable or secret.
- Use variables if you need to store non-sensitive configuration values and secrets for storing access tokens, API keys, or any sensitive value or credentials.



Repository variables and secrets

Variables Public

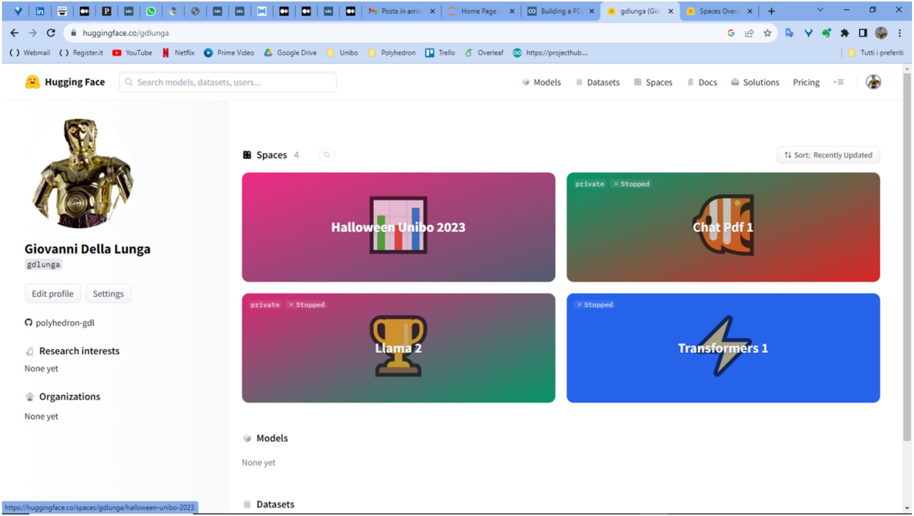
MODEL_REPO_ID · Updated less than a minute ago

Secrets Private

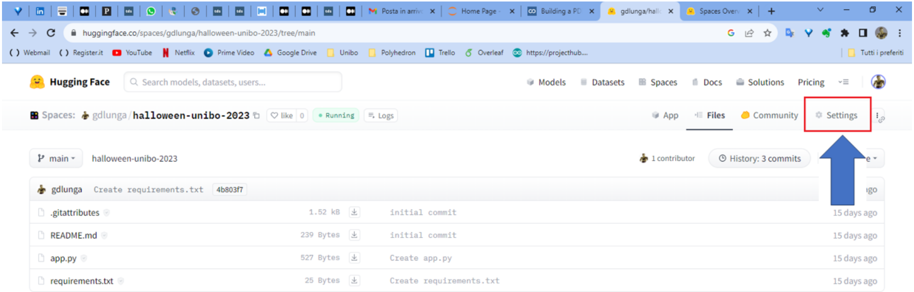
HF_TOKEN · Updated less than a minute ago

OPENAI_API_KEY · Updated less than a minute ago

Managing secrets and environment variables



Managing secrets and environment variables



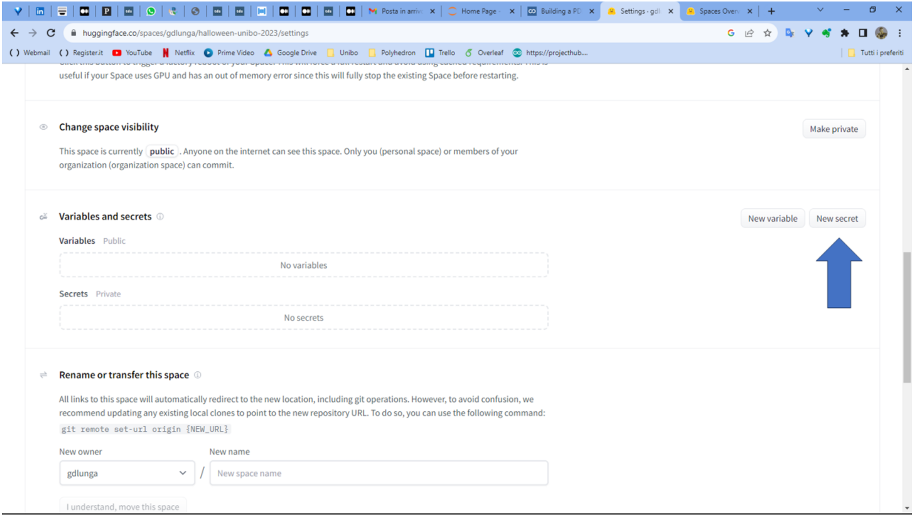
huggingface.co/spaces/gdlunga/halloween-unibo-2023/tree/main

halloween-unibo-2023

gdlunga Create requirements.txt 4b803f7

.gitattributes	1.52 kB	initial commit	15 days ago
README.md	239 Bytes	initial commit	15 days ago
app.py	527 Bytes	Create app.py	15 days ago
requirements.txt	25 Bytes	Create requirements.txt	15 days ago

Managing secrets and environment variables



huggingface.co/spaces/gdlunga/halloween-unibo-2023/settings

useful if your Space uses GPU and has an out of memory error since this will fully stop the existing Space before restarting.

Change space visibility

This space is currently public. Anyone on the internet can see this space. Only you (personal space) or members of your organization (organization space) can commit.

Variables and secrets

Variables Public

No variables

Secrets Private

No secrets

Rename or transfer this space

All links to this space will automatically redirect to the new location, including git operations. However, to avoid confusion, we recommend updating any existing local clones to point to the new repository URL. To do so, you can use the following command:

```
git remote set-url origin {NEW_URL}
```

New owner: gdlunga / New name: New space name