

# Natural Language Processing and Large Language Models

Corso di Laurea Magistrale in Ingegneria Informatica



Lesson 12

## HuggingFace

Nicola Capuano and Antonio Greco

DIEM – University of Salerno



# Outline

- Overview
- Setup
- Pipeline
- Model selection
- Common models
- Gradio





# Overview

# Hugging Face

- <https://github.com/huggingface/education-toolkit>



## Education Toolkit

---

👋 Welcome!

We've assembled a toolkit that anyone can use to easily prepare workshops, events, homework or classes. The content is self-contained so that it can be easily incorporated in other material. This content is **free** and uses well-known Open Source technologies ( `transformers` , `gradio` , etc).

Apart from tutorials, we also share other resources to go further into ML or that can assist in designing content.

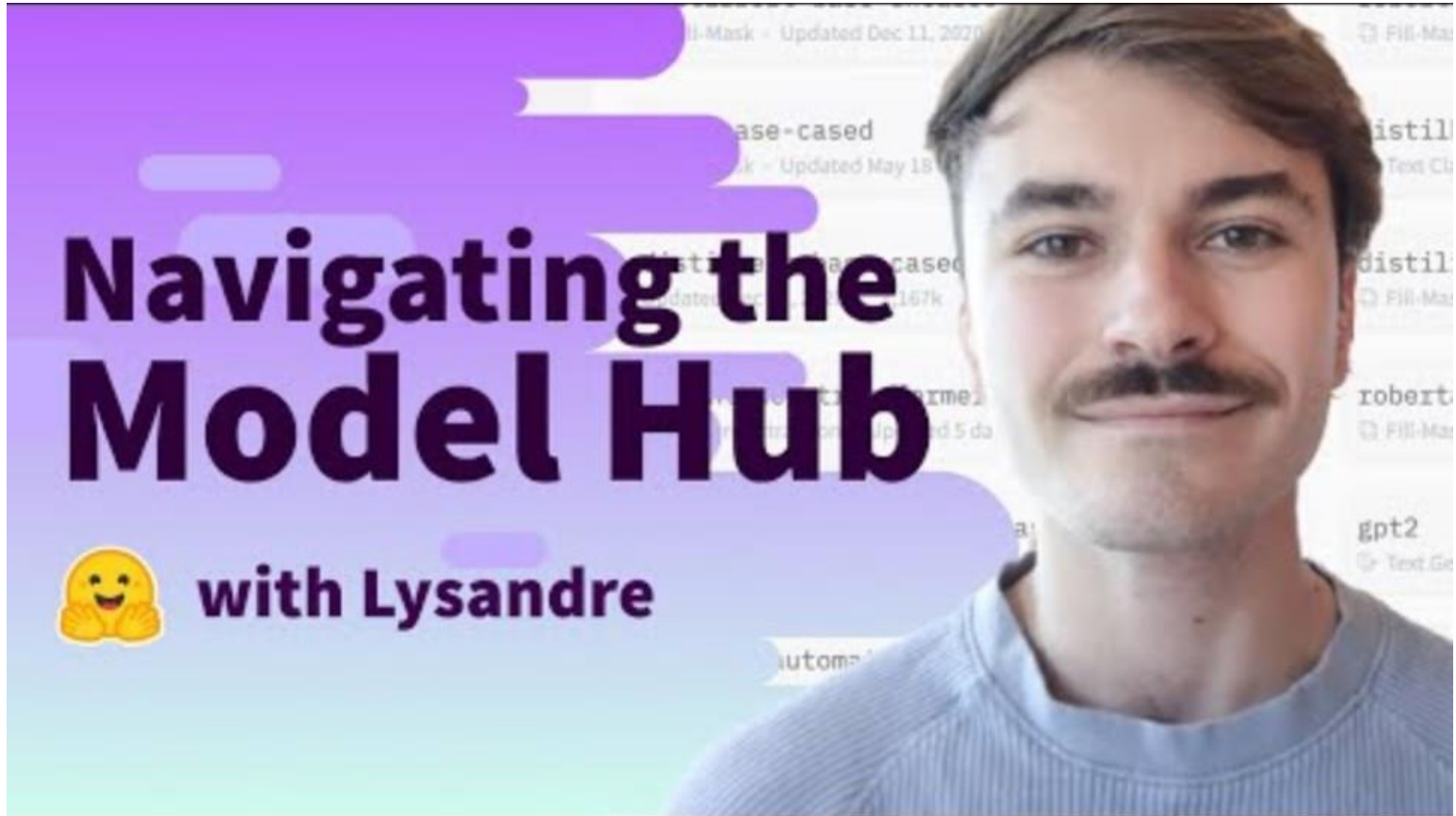
# Hugging Face



**Hugging Face**

- The Hugging Face Hub (<https://huggingface.co/>) hosts:
  - Models
  - Datasets
  - Spaces for demos and code
- Key libraries include:
  - datasets: Download datasets from the hub
  - transformers: Work with pipelines, tokenizers, models, etc.
  - evaluate: Compute evaluation metrics
- These libraries can use PyTorch and TensorFlow

# Hugging Face – Model Hub



<https://huggingface.co/models>

# Hugging Face - Datasets

- The Hub (<https://hf.co/datasets>) hosts around 3000 datasets that are open-sourced and free to use in multiple domains.
- On top of it, the open-source datasets library allows the easy use of these datasets, including huge ones, using very convenient features such as streaming.
- Similar to model repositories, you have a dataset card that documents the dataset. If you scroll down a bit, you will find things such as the summary, the structure, and more.
- Example: <https://huggingface.co/datasets/nyu-mll/glue>



# Setup



# Setup – Google Colab

- Using a Colab notebook is the simplest possible setup; boot up a notebook in your browser and get straight to coding!
  - `!pip install transformers`
  - `import transformers`
- This installs a very light version of Transformers. In particular, no specific machine learning frameworks (like PyTorch or TensorFlow) are installed. You can also install the development version, which comes with all the required dependencies for pretty much any imaginable use case:
  - `!pip install transformers[sentencepiece]`

# Setup – Virtual environment

Download and install Anaconda:

<https://www.anaconda.com/download>

- `conda create --name nlpllm`
- `conda activate nlpllm`
- `conda install transformers[sentencepiece]`

# Setup – Create a Hugging Face account

- Most of the functionalities relies on you having a Hugging Face account.
- It is recommend creating one

**Welcome**

[Skip to feed →](#)

**Create a new model**  
From the website

**Hub documentation**  
Take a first look at the Hub features

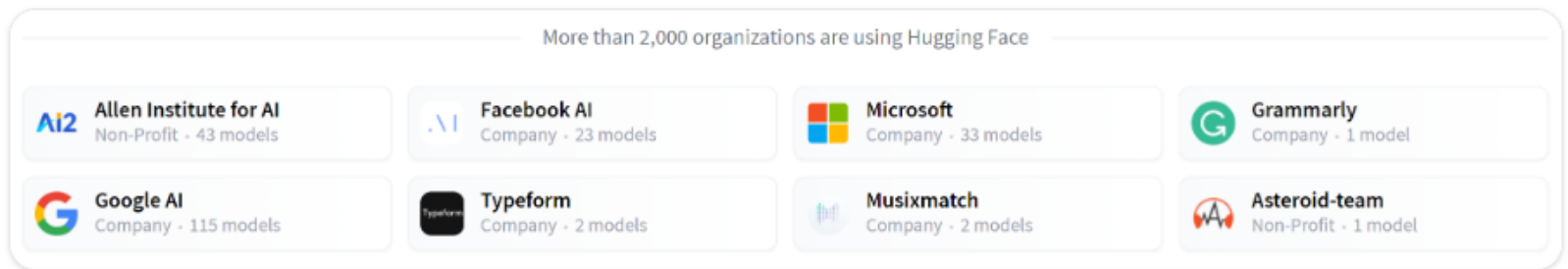
**Programmatic access**  
Use the Hub's Python client library



# Pipeline

# Pipeline

- The Hugging Face Transformers library provides the functionality to create and use the models.
- The Model Hub contains thousands of pretrained models that anyone can download and use.
- The most basic object in the Hugging Face Transformers library is the `pipeline()` function.
- It connects a model with its necessary preprocessing and postprocessing steps, allowing us to directly input any text and get an intelligible answer.



# Pipeline





# Model selection

# Model selection

(CNN)

A magnitude 6.7 earthquake rattled Papua New Guinea early Friday afternoon, according to the U.S. Geological Survey. The quake was centered about 200 miles north-northeast of Port Moresby and had a depth of 28 miles. No tsunami warning was issued...

<Article 1  
summary>

**NLP task behind this app:**

**Summarization**

**Extractive:** Select representative pieces of text.

**Abstractive:** Generate new text.

**Find a model for this task:**

**Hugging Face Hub** → 176,620 models.

Filter by task → 960 models.

Then...? Consider your needs.



# Model selection

The screenshot shows the Hugging Face website's model selection interface. It includes a search bar, navigation tabs for Models, Datasets, Spaces, Docs, and Solutions, and a sidebar for filtering by task, license, language, etc. The main content area displays a list of models, with 'bert-base-uncased' and 'jonatasgrosman/wav2vec2-large' visible. Four callout boxes highlight specific features: filtering by task, sorting by popularity, filtering by model size, and checking git release history.

**Filter by task, license, language, etc.**

Hugging Face

Models 187,956

bert-base-uncased

jonatasgrosman/wav2vec2-large

**Sort by popularity and updates**

Sort: Most Downloads

Most Downloads

Recently Updated

Most Likes

**Filter by model size (for limits on hardware, cost, or latency)**

Files and versions

pytorch\_model.bin

2.33 GB

LFS

**Check git release history**

github.com/google-research/bert/blob/master/README.md

**BERT**

\*\*\*\*\* New March 11th, 2020: Smaller BERT Models \*\*\*\*\*

This is a release of 24 smaller BERT models (English only, unc

# Model selection

Pick good variants of models for your task.

- Different sizes of the same base model.
- Fine-tuned variants of base models.

Models 5,564  x [new](#) Full-text search

t5-base

🔗 Updated 11 days ago • ↓ 5.76M • ♥ 190

t5-small

🔗 Updated 11 days ago • ↓ 2.17M • ♥ 89

🔗 prithivida/parrot\_paraphraser\_on\_T5

🔗 Updated May 18, 2021 • ↓ 545k • ♥ 97

Also consider:

- Search for examples and datasets, not just models.
- Is the model “good” at everything, or was it fine-tuned for a specific task?
- Which datasets were used for pre-training and/or fine-tuning?

**Ultimately, it's about *your data and users*.**

- Define KPIs.
- Test on your data or users.



# Common models

# Common models

Model or model family	Model size (# params)	License	Created by	Released	Notes
Pythia	19 M – 12 B	Apache 2.0	EleutherAI	2023	series of 8 models for comparisons across sizes
Dolly	12 B	MIT	Databricks	2023	instruction-tuned Pythia model
GPT-3.5	175 B	proprietary	OpenAI	2022	ChatGPT model option; related models GPT-1/2/3/4
OPT	125 M – 175 B	MIT	Meta	2022	based on GPT-3 architecture
BLOOM	560 M – 176 B	RAIL v1.0	many groups	2022	46 languages
GPT-Neo/X	125 M – 20 B	MIT / Apache 2.0	EleutherAI	2021 / 2022	based on GPT-2 architecture
FLAN	80 M – 540 B	Apache 2.0	Google	2021	methods to improve training for existing architectures
BART	139 M – 406 M	Apache 2.0	Meta	2019	derived from BERT, GPT, others
T5	50 M – 11 B	Apache 2.0	Google	2019	4 languages
BERT	109 M – 335 M	Apache 2.0	Google	2018	early breakthrough



©2023 Databricks Inc. — All rights reserved





# Gradio

# Gradio – Demos on the web

	Framework	Language
1. Train a model	 TensorFlow PYTORCH	Python
2. Containerize and deploy the model		Python
3. Store incoming samples		Python
4. Build an interactive front-end		Python



# Gradio – Demos on the web

## News Summarizer

Let Hugging Face models summarize articles for you. Note: Shorter articles generate faster summaries. This summarizer uses bart-large-cnn model by Facebook

URL

Clear

Submit



Screenshot

Flag

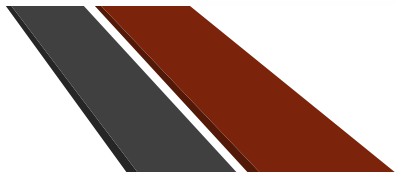
### Examples

<https://www.technologyreview.com/2021/07/22/1029973/deepmind-alphafold-protein-folding-biology-disease-drugs-proteome/>


<https://www.technologyreview.com/2021/07/21/1029860/disability-rights-employment-discrimination-ai-hiring/>


<https://www.technologyreview.com/2021/07/09/1028140/ai-voice-actors-sound-human/>

[view the api](#) 🔑 | built with 




# Gradio – Free hosting on hf.space

 **Hugging Face**

[Models](#) [Datasets](#) [Spaces](#) [Docs](#) [Solutions](#) [Pricing](#) 









---

 **Spaces**  
Discover amazing ML apps made by the community! [Create new Space](#) or [learn more about Spaces](#).





---

Sort: Recently Updated

☆ Spaces of the week 🔥

 Zero Shot Image Classification dt 10 days ago 11	 Poolformer akhaliq 10 days ago 3	 Multilingual TTS Flux9665 12 days ago 5	 XGLM Zero Shot COPA valhalla 22 days ago 6
 Plastic_in_river Kili 21 days ago 4	 Sentence Embeddings Visualization radames 15 days ago 8	 TokenCut akhaliq 12 days ago 6	 Cryptopunks Generator nateraw 10 days ago 4

# All running apps, most recent first

 Manifesto Sa-m 9 minutes ago 2	 Id The Seas snakeeyes021 21 minutes ago 1	 CaffeNet onnx 37 minutes ago	 CoquiTTS (Official) erogol about 1 hour ago 28
--	--	--	--





# Gradio – Build your demo

## Try it yourself

- conda install gradio

<https://bit.ly/34wESgd>

# Natural Language Processing and Large Language Models

Corso di Laurea Magistrale in Ingegneria Informatica



Lesson 12

## HuggingFace

Nicola Capuano and Antonio Greco

DIEM – University of Salerno

