In [1]:
```
try:
  from google.colab import drive
  IN_COLAB=True
except:
  IN_COLAB=False

if IN_COLAB:
  print("We're running Colab")
```

```
We're running Colab
```

In [2]:
```
import tensorflow as tf

print("Running TensorFlow version ",tf.__version__)

# Parse tensorflow version
import re

version_match = re.match("([0-9]+)\.([0-9]+)", tf.__version__)
tf_major, tf_minor = int(version_match.group(1)) , int(version_match.group(2))
print("Version {v:d}, minor {m:d}".format(v=tf_major, m=tf_minor) )
```

```
Running TensorFlow version  2.15.0
Version 2, minor 15
```

In [3]:

```
gpu_devices = tf.config.experimental.list_physical_devices('GPU')
if gpu_devices:
    print('Using GPU')
    tf.config.experimental.set_memory_growth(gpu_devices[0], True)
else:
    print('Using CPU')
```

```
Using CPU
```

In [4]:
```
if IN_COLAB:
  !pip install datasets evaluate transformers[sentencepiece]
  !pip install gradio
```

```
Requirement already satisfied: datasets in /usr/local/lib/python3.10/dist-pack
ages (2.18.0)
Requirement already satisfied: evaluate in /usr/local/lib/python3.10/dist-pack
ages (0.4.1)
Requirement already satisfied: transformers[sentencepiece] in /usr/local/lib/p
ython3.10/dist-packages (4.38.2)
Requirement already satisfied: filelock in /usr/local/lib/python3.10/dist-pack
ages (from datasets) (3.13.4)
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.10/dist-p
ackages (from datasets) (1.25.2)
Requirement already satisfied: pyarrow>=12.0.0 in /usr/local/lib/python3.10/di
st-packages (from datasets) (14.0.2)
Requirement already satisfied: pyarrow-hotfix in /usr/local/lib/python3.10/dis
t-packages (from datasets) (0.6)
Requirement already satisfied: dill<0.3.9,>=0.3.0 in /usr/local/lib/python3.1
0/dist-packages (from datasets) (0.3.8)
Requirement already satisfied: pandas in /usr/local/lib/python3.10/dist-packag
es (from datasets) (2.0.3)
Requirement already satisfied: requests>=2.19.0 in /usr/local/lib/python3.10/d
ist-packages (from datasets) (2.31.0)
Requirement already satisfied: tqdm>=4.62.1 in /usr/local/lib/python3.10/dist-
packages (from datasets) (4.66.2)
Requirement already satisfied: xxhash in /usr/local/lib/python3.10/dist-packag
es (from datasets) (3.4.1)
Requirement already satisfied: multiprocess in /usr/local/lib/python3.10/dist-
packages (from datasets) (0.70.16)
Requirement already satisfied: fsspec[http]<=2024.2.0,>=2023.1.0 in /usr/loca
l/lib/python3.10/dist-packages (from datasets) (2023.6.0)
Requirement already satisfied: aiohttp in /usr/local/lib/python3.10/dist-packa
ges (from datasets) (3.9.5)
Requirement already satisfied: huggingface-hub>=0.19.4 in /usr/local/lib/pytho
n3.10/dist-packages (from datasets) (0.20.3)
Requirement already satisfied: packaging in /usr/local/lib/python3.10/dist-pac
kages (from datasets) (24.0)
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.10/dist-p
```

```
ackages (from datasets) (6.0.1)
Requirement already satisfied: responses<0.19 in /usr/local/lib/python3.10/dis
t-packages (from evaluate) (0.18.0)
Requirement already satisfied: regex!=2019.12.17 in /usr/local/lib/python3.10/
dist-packages (from transformers[sentencepiece]) (2023.12.25)
Requirement already satisfied: tokenizers<0.19,>=0.14 in /usr/local/lib/python
3.10/dist-packages (from transformers[sentencepiece]) (0.15.2)
Requirement already satisfied: safetensors>=0.4.1 in /usr/local/lib/python3.1
0/dist-packages (from transformers[sentencepiece]) (0.4.3)
Requirement already satisfied: sentencepiece!=0.1.92,>=0.1.91 in /usr/local/li
b/python3.10/dist-packages (from transformers[sentencepiece]) (0.1.99)
Requirement already satisfied: protobuf in /usr/local/lib/python3.10/dist-pack
ages (from transformers[sentencepiece]) (3.20.3)
Requirement already satisfied: aiosignal>=1.1.2 in /usr/local/lib/python3.10/d
ist-packages (from aiohttp->datasets) (1.3.1)
Requirement already satisfied: attrs>=17.3.0 in /usr/local/lib/python3.10/dist
-packages (from aiohttp->datasets) (23.2.0)
Requirement already satisfied: frozenlist>=1.1.1 in /usr/local/lib/python3.10/
dist-packages (from aiohttp->datasets) (1.4.1)
Requirement already satisfied: multidict<7.0,>=4.5 in /usr/local/lib/python3.1
0/dist-packages (from aiohttp->datasets) (6.0.5)
Requirement already satisfied: yarl<2.0,>=1.0 in /usr/local/lib/python3.10/dis
t-packages (from aiohttp->datasets) (1.9.4)
Requirement already satisfied: async-timeout<5.0,>=4.0 in /usr/local/lib/pytho
n3.10/dist-packages (from aiohttp->datasets) (4.0.3)
Requirement already satisfied: typing-extensions>=3.7.4.3 in /usr/local/lib/py
thon3.10/dist-packages (from huggingface-hub>=0.19.4->datasets) (4.11.0)
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/pyth
on3.10/dist-packages (from requests>=2.19.0->datasets) (3.3.2)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.10/dist-
packages (from requests>=2.19.0->datasets) (3.7)
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.1
0/dist-packages (from requests>=2.19.0->datasets) (2.0.7)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.1
0/dist-packages (from requests>=2.19.0->datasets) (2024.2.2)
Requirement already satisfied: python-dateutil>=2.8.2 in /usr/local/lib/python
```

```
3.10/dist-packages (from pandas->datasets) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.10/dist-
packages (from pandas->datasets) (2023.4)
Requirement already satisfied: tzdata>=2022.1 in /usr/local/lib/python3.10/dis
t-packages (from pandas->datasets) (2024.1)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.10/dist-pack
ages (from python-dateutil>=2.8.2->pandas->datasets) (1.16.0)
Requirement already satisfied: gradio in /usr/local/lib/python3.10/dist-packag
es (4.27.0)
Requirement already satisfied: aiofiles<24.0,>=22.0 in /usr/local/lib/python3.
10/dist-packages (from gradio) (23.2.1)
Requirement already satisfied: altair<6.0,>=4.2.0 in /usr/local/lib/python3.1
0/dist-packages (from gradio) (4.2.2)
Requirement already satisfied: fastapi in /usr/local/lib/python3.10/dist-packa
ges (from gradio) (0.110.1)
Requirement already satisfied: ffmpy in /usr/local/lib/python3.10/dist-package
s (from gradio) (0.3.2)
Requirement already satisfied: gradio-client==0.15.1 in /usr/local/lib/python
3.10/dist-packages (from gradio) (0.15.1)
Requirement already satisfied: httpx>=0.24.1 in /usr/local/lib/python3.10/dist
-packages (from gradio) (0.27.0)
Requirement already satisfied: huggingface-hub>=0.19.3 in /usr/local/lib/pytho
n3.10/dist-packages (from gradio) (0.20.3)
Requirement already satisfied: importlib-resources<7.0,>=1.3 in /usr/local/li
b/python3.10/dist-packages (from gradio) (6.4.0)
Requirement already satisfied: jinja2<4.0 in /usr/local/lib/python3.10/dist-pa
ckages (from gradio) (3.1.3)
Requirement already satisfied: markupsafe~=2.0 in /usr/local/lib/python3.10/di
st-packages (from gradio) (2.1.5)
Requirement already satisfied: matplotlib~=3.0 in /usr/local/lib/python3.10/di
st-packages (from gradio) (3.7.1)
Requirement already satisfied: numpy~=1.0 in /usr/local/lib/python3.10/dist-pa
ckages (from gradio) (1.25.2)
Requirement already satisfied: orjson~=3.0 in /usr/local/lib/python3.10/dist-p
ackages (from gradio) (3.10.1)
Requirement already satisfied: packaging in /usr/local/lib/python3.10/dist-pac
```

```
kages (from gradio) (24.0)
Requirement already satisfied: pandas<3.0,>=1.0 in /usr/local/lib/python3.10/d
ist-packages (from gradio) (2.0.3)
Requirement already satisfied: pillow<11.0,>=8.0 in /usr/local/lib/python3.10/
dist-packages (from gradio) (9.4.0)
Requirement already satisfied: pydantic>=2.0 in /usr/local/lib/python3.10/dist
-packages (from gradio) (2.7.0)
Requirement already satisfied: pydub in /usr/local/lib/python3.10/dist-package
s (from gradio) (0.25.1)
Requirement already satisfied: python-multipart>=0.0.9 in /usr/local/lib/pytho
n3.10/dist-packages (from gradio) (0.0.9)
Requirement already satisfied: pyyaml<7.0,>=5.0 in /usr/local/lib/python3.10/d
ist-packages (from gradio) (6.0.1)
Requirement already satisfied: ruff>=0.2.2 in /usr/local/lib/python3.10/dist-p
ackages (from gradio) (0.3.7)
Requirement already satisfied: semantic-version~=2.0 in /usr/local/lib/python
3.10/dist-packages (from gradio) (2.10.0)
Requirement already satisfied: tomlkit==0.12.0 in /usr/local/lib/python3.10/di
st-packages (from gradio) (0.12.0)
Requirement already satisfied: typer<1.0,>=0.12 in /usr/local/lib/python3.10/d
ist-packages (from gradio) (0.12.3)
Requirement already satisfied: typing-extensions~=4.0 in /usr/local/lib/python
3.10/dist-packages (from gradio) (4.11.0)
Requirement already satisfied: urllib3~=2.0 in /usr/local/lib/python3.10/dist-
packages (from gradio) (2.0.7)
Requirement already satisfied: uvicorn>=0.14.0 in /usr/local/lib/python3.10/di
st-packages (from gradio) (0.29.0)
Requirement already satisfied: fsspec in /usr/local/lib/python3.10/dist-packag
es (from gradio-client==0.15.1->gradio) (2023.6.0)
Requirement already satisfied: websockets<12.0,>=10.0 in /usr/local/lib/python
3.10/dist-packages (from gradio-client==0.15.1->gradio) (11.0.3)
Requirement already satisfied: entrypoints in /usr/local/lib/python3.10/dist-p
ackages (from altair<6.0,>=4.2.0->gradio) (0.4)
Requirement already satisfied: jsonschema>=3.0 in /usr/local/lib/python3.10/di
st-packages (from altair<6.0,>=4.2.0->gradio) (4.19.2)
Requirement already satisfied: toolz in /usr/local/lib/python3.10/dist-package
```

```
s (from altair<6.0,>=4.2.0->gradio) (0.12.1)
Requirement already satisfied: anyio in /usr/local/lib/python3.10/dist-package
s (from httpx>=0.24.1->gradio) (3.7.1)
Requirement already satisfied: certifi in /usr/local/lib/python3.10/dist-packa
ges (from httpx>=0.24.1->gradio) (2024.2.2)
Requirement already satisfied: httpcore==1.* in /usr/local/lib/python3.10/dist
-packages (from httpx>=0.24.1->gradio) (1.0.5)
Requirement already satisfied: idna in /usr/local/lib/python3.10/dist-packages
(from httpx>=0.24.1->gradio) (3.7)
Requirement already satisfied: sniffio in /usr/local/lib/python3.10/dist-packa
ges (from httpx>=0.24.1->gradio) (1.3.1)
Requirement already satisfied: h11<0.15,>=0.13 in /usr/local/lib/python3.10/di
st-packages (from httpcore==1.*->httpx>=0.24.1->gradio) (0.14.0)
Requirement already satisfied: filelock in /usr/local/lib/python3.10/dist-pack
ages (from huggingface-hub>=0.19.3->gradio) (3.13.4)
Requirement already satisfied: requests in /usr/local/lib/python3.10/dist-pack
ages (from huggingface-hub>=0.19.3->gradio) (2.31.0)
Requirement already satisfied: tqdm>=4.42.1 in /usr/local/lib/python3.10/dist-
packages (from huggingface-hub>=0.19.3->gradio) (4.66.2)
Requirement already satisfied: contourpy>=1.0.1 in /usr/local/lib/python3.10/d
ist-packages (from matplotlib~=3.0->gradio) (1.2.1)
Requirement already satisfied: cycler>=0.10 in /usr/local/lib/python3.10/dist-
packages (from matplotlib~=3.0->gradio) (0.12.1)
Requirement already satisfied: fonttools>=4.22.0 in /usr/local/lib/python3.10/
dist-packages (from matplotlib~=3.0->gradio) (4.51.0)
Requirement already satisfied: kiwisolver>=1.0.1 in /usr/local/lib/python3.10/
dist-packages (from matplotlib~=3.0->gradio) (1.4.5)
Requirement already satisfied: pyparsing>=2.3.1 in /usr/local/lib/python3.10/d
ist-packages (from matplotlib~=3.0->gradio) (3.1.2)
Requirement already satisfied: python-dateutil>=2.7 in /usr/local/lib/python3.
10/dist-packages (from matplotlib~=3.0->gradio) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.10/dist-
packages (from pandas<3.0,>=1.0->gradio) (2023.4)
Requirement already satisfied: tzdata>=2022.1 in /usr/local/lib/python3.10/dis
t-packages (from pandas<3.0,>=1.0->gradio) (2024.1)
Requirement already satisfied: annotated-types>=0.4.0 in /usr/local/lib/python
```

```
3.10/dist-packages (from pydantic>=2.0->gradio) (0.6.0)
Requirement already satisfied: pydantic-core==2.18.1 in /usr/local/lib/python
3.10/dist-packages (from pydantic>=2.0->gradio) (2.18.1)
Requirement already satisfied: click>=8.0.0 in /usr/local/lib/python3.10/dist-
packages (from typer<1.0,>=0.12->gradio) (8.1.7)
Requirement already satisfied: shellingham>=1.3.0 in /usr/local/lib/python3.1
0/dist-packages (from typer<1.0,>=0.12->gradio) (1.5.4)
Requirement already satisfied: rich>=10.11.0 in /usr/local/lib/python3.10/dist
-packages (from typer<1.0,>=0.12->gradio) (13.7.1)
Requirement already satisfied: starlette<0.38.0,>=0.37.2 in /usr/local/lib/pyt
hon3.10/dist-packages (from fastapi->gradio) (0.37.2)
Requirement already satisfied: attrs>=22.2.0 in /usr/local/lib/python3.10/dist
-packages (from jsonschema>=3.0->altair<6.0,>=4.2.0->gradio) (23.2.0)
Requirement already satisfied: jsonschema-specifications>=2023.03.6 in /usr/lo
cal/lib/python3.10/dist-packages (from jsonschema>=3.0->altair<6.0,>=4.2.0->gr
adio) (2023.12.1)
Requirement already satisfied: referencing>=0.28.4 in /usr/local/lib/python3.1
0/dist-packages (from jsonschema>=3.0->altair<6.0,>=4.2.0->gradio) (0.34.0)
Requirement already satisfied: rpds-py>=0.7.1 in /usr/local/lib/python3.10/dis
t-packages (from jsonschema>=3.0->altair<6.0,>=4.2.0->gradio) (0.18.0)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.10/dist-pack
ages (from python-dateutil>=2.7->matplotlib~=3.0->gradio) (1.16.0)
Requirement already satisfied: markdown-it-py>=2.2.0 in /usr/local/lib/python
3.10/dist-packages (from rich>=10.11.0->typer<1.0,>=0.12->gradio) (3.0.0)
Requirement already satisfied: pygments<3.0.0,>=2.13.0 in /usr/local/lib/pytho
n3.10/dist-packages (from rich>=10.11.0->typer<1.0,>=0.12->gradio) (2.16.1)
Requirement already satisfied: exceptiongroup in /usr/local/lib/python3.10/dis
t-packages (from anyio->httpx>=0.24.1->gradio) (1.2.0)
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/pyth
on3.10/dist-packages (from requests->huggingface-hub>=0.19.3->gradio) (3.3.2)
Requirement already satisfied: mdurl~=0.1 in /usr/local/lib/python3.10/dist-pa
ckages (from markdown-it-py>=2.2.0->rich>=10.11.0->typer<1.0,>=0.12->gradio)
(0.1.2)
```

In [5]:
```
import gradio as gr
```

Use an Inference end-point

- Advange:
    - free
    - does not use *local* RAM so can run big models
- paid hosting (https://huggingface.co/pricing#endpoints)
    - don't get charged for a *paused* end-point
- guide (https://huggingface.co/docs/inference-endpoints/index)

In [6]:

```
import requests

from pathlib import Path



def get_API_token(token_file="/content/hf.token"):
    # Check for file containing API token to HuggingFace
    p = Path(token_file).expanduser()
    if not p.exists():
      print(f"Token file {p} not found.")
      return

    with open(token_file, 'r') as fp:
        token = fp.read()

    # Remove trailing newline
    token = token.rstrip()

    return token

API_TOKEN=get_API_token();
```

In [7]:

```
gen_text_key = "generated_text"
input_key = "inputs"
error_key = "error"

models = { "small": "EleutherAI/gpt-neo-1.3B",
           "big":   "EleutherAI/gpt-neox-20b"
}
```

In [8]:

```
import time

headers = {"Authorization": f"Bearer {API_TOKEN}"}

def query(payload, model_string):
  API_URL = f"https://api-inference.huggingface.co/models/{model_string}"

  response = requests.post(API_URL, headers=headers, json=payload)
  return response.json()

def execute_query(q, model_string):
  output = query({  input_key: q }, model_string)

  # Successful output is a list; error output is a dict
  if type(output) is dict:
    out = f"Error: {output[error_key]}"
  else:
    out = output[0][gen_text_key]

  return out



exemplars = [ "this movie was great: positive",
              "one of the best films of the year: positive",
              "just plain awful: negative",
              "I would not see this one again: negative",
              "this movie was great: positive",
              "one of the best films of the year: positive",
              "just plain awful: negative",
              "I would not see this one again: negative",
              "I love this film: positive"
]

sep = " \n "
```

```
exemplar_string = sep.join(exemplars)
few_shot_string =  exemplar_string + sep + "I've heard not so great things about
this one:"

q = few_shot_string

# q = "Can you please let us know more details about your "

# Can run a very large model since execution is on remote end-point
model_string = models["big"]

time_start = time.time()

res = execute_query(q, model_string)

time_end = time.time()

print(f"[{time_end-time_start:3.2f} seconds, using {model_string}]\n {res}")
```

```
[0.22 seconds, using EleutherAI/gpt-neox-20b]
 this movie was great: positive
 one of the best films of the year: positive
 just plain awful: negative
 I would not see this one again: negative
 this movie was great: positive
 one of the best films of the year: positive
 just plain awful: negative
 I would not see this one again: negative
 I love this film: positive
 I've heard not so great things about this one: negative
 it is entertaining: positive
 I would not see this one again: negative
 I love this film: positive
 I've heard not so great things: negative
 it is entertaining: positive
 I've heard not so great things: negative
 enchanced by slow motion visuals: positive
 excellent: positive
 terrific sound design in this one too: positive
 Stallone is a great actor: positive
 I'd turn down a free trip to London to see this movie:
```

Use a pipeline

- runs locally
- so can only use model small enough to fit in RAM

In [9]:
```
model_string = models["small"]
```

In [10]:

```
from transformers import pipeline

num_return = 3
len_return = 30


generator = pipeline('text-generation', model = model_string)
```

```
/usr/local/lib/python3.10/dist-packages/huggingface_hub/utils/_token.py:88: Us
erWarning:
The secret `HF_TOKEN` does not exist in your Colab secrets.
To authenticate with the Hugging Face Hub, create a token in your settings tab
(https://huggingface.co/settings/tokens), set it as secret in your Google Cola
b and restart your session.
You will be able to reuse this secret in all of your notebooks.
Please note that authentication is recommended but still optional to access pu
blic models or datasets.
  warnings.warn(
```