databricks

14 DAYS

AI CHALLENGE

DAY 14

Topic:

AI-Powered Analytics: Genie & Mosaic AI

Challenge:

1. Use Genie to query data with natural language
2. Explore Mosaic AI features
3. Build simple NLP task
4. Create AI-powered insights

Search data, notebooks, recents, and more...   CTRL + P

Catalog

Type to search...

For you   All

⌄ My organization
  > ⬡ workspace
  > ⬡ system
  > ⬡ ecommerce
⌄ Delta Shares Received
  > ⬡ samples

File Edit View Run Help   Python ⌄   Tabs: ON ⌄   ⭐   Last edit was 6 hours ago   ▶ Run all   ○ Serverless ⌄   Schedule   Share

▶   ✓ 01:28 PM (37s)   1

```
%pip install torch transformers
```

Collecting torch
  Downloading torch-2.10.0-cp312-cp312-manylinux_2_28_aarch64.whl.metadata (31 kB)
Collecting transformers
  Downloading transformers-4.57.6-py3-none-any.whl.metadata (43 kB)
Requirement already satisfied: filelock in /usr/local/lib/python3.12/dist-packages (from torch) (3.18.0)
Requirement already satisfied: typing-extensions>=4.10.0 in /databricks/python3/lib/python3.12/site-packages (from torch) (4.12.2)
Requirement already satisfied: setuptools in /usr/local/lib/python3.12/dist-packages (from torch) (74.0.0)
Collecting sympy>=1.13.3 (from torch)
  Downloading sympy-1.14.0-py3-none-any.whl.metadata (12 kB)
Collecting networkx>=2.5.1 (from torch)
  Downloading networkx-3.6.1-py3-none-any.whl.metadata (6.8 kB)
Requirement already satisfied: jinja2 in /databricks/python3/lib/python3.12/site-packages (from torch) (3.1.5)
Requirement already satisfied: fsspec>=0.8.5 in /databricks/python3/lib/python3.12/site-packages (from torch) (2023.5.0)
Collecting huggingface-hub<1.0,>=0.34.0 (from transformers)
  Downloading huggingface_hub-0.36.0-py3-none-any.whl.metadata (14 kB)
Requirement already satisfied: numpy>=1.17 in /databricks/python3/lib/python3.12/site-packages (from transformers) (2.1.3)
Requirement already satisfied: packaging>=20.0 in /databricks/python3/lib/python3.12/site-packages (from transformers) (24.1)
Requirement already satisfied: pyyaml>=5.1 in /databricks/python3/lib/python3.12/site-packages (from transformers) (6.0.2)
Collecting regex!=2019.12.17 (from transformers)
  Downloading regex-2026.1.15-cp312-cp312-manylinux2014_aarch64.manylinux_2_17_aarch64.manylinux_2_28_aarch64.whl.metadata (40 kB)
Requirement already satisfied: requests in /databricks/python3/lib/python3.12/site-packages (from transformers) (2.32.3)

💡1

▶   ✓ 01:33 PM (7s)   2

```
from transformers import pipeline
import mlflow
```

/local_disk0/.ephemeral_nfs/envs/pythonEnv-e1fcb4b6-a796-4390-9c98-fbac2194d501/lib/python3.12/site-packages/torch/_vmap_internals.py:9: FutureWarning: `isinstan

🏠 Workspace ⌄   🔲 Day14_DataBricks ✕   ➕   ⬚ ⌃

📋   **Catalog**   ↻ ➕ ✕   File Edit View Run Help   Python ⌄ Tabs: ON ⌄ ☆ Last edit was 6 hours ago   🔲 ▶ Run all   ● Serverless ⌄   Schedule   Share   💬

📁   Type to search...   ⇅   ▶   ✓ 3 minutes ago (12s)   2   ⠿
                                                                                                                              ⠿
⬚   from transformers import pipeline   🕐
         For you   All   import mlflow   {x}
      ⌄ My organization
         › 🔲 workspace   /local_disk0/.ephemeral_nfs/envs/pythonEnv-75ca6bb9-5ef6-4a05-aaed-ad536e325529/lib/python3.12/site-packages/torch/_vmap_internals.py:9: FutureWarning: `isinstan   ⚡
         › 🔲 system   ce(treespec, LeafSpec)` is deprecated, use `isinstance(treespec, TreeSpec) and treespec.is_leaf()` instead.   ℹ️
         › 🔲 ecommerce   from torch.utils._pytree import _broadcast_to_and_flatten, tree_flatten, tree_unflatten
      ⌄ Delta Shares Received
         › 🔲 samples   💡1

                                  # Build a simple NLP task

                                  ▶   ✓ 2 minutes ago (2s)   4

                                      # Load sentiment analysis pipeline (Mosaic AI / HuggingFace compatible)
                                      classifier = pipeline("sentiment-analysis")

                                  No model was supplied, defaulted to distilbert/distilbert-base-uncased-finetuned-sst-2-english and revision 714eb0f (https://huggingface.co/distilbert/distilbert
                                  -base-uncased-finetuned-sst-2-english).
                                  Using a pipeline without specifying a model name and revision in production is not recommended.

                                  config.json: 100% ████████████████ 629/629 [00:00<00:00, 53.5kB/s]

                                  model.safetensors: 100% ████████████████ 268M/268M [00:00<00:00, 297MB/s]

                                  tokenizer_config.json: 100% ████████████████ 48.0/48.0 [00:00<00:00, 9.02kB/s]

                                  vocab.txt: ██ 232k/? [00:00<00:00, 26.5MB/s]

                                  Device set to use cpu

Workspace ∨   Day14_DataBricks ×   +

Catalog

Type to search...

For you   All

∨ My organization
  › 🔒 workspace
  › 🔒 system
  › 🗂 ecommerce
∨ Delta Shares Received
  › 🔒 samples

File Edit View Run Help   Python ∨   Tabs: ON ∨   ☆   Last edit was 6 hours ago

▶ Run all   ● Serverless ∨   Schedule   Share

```python
▶       ✓ 3 minutes ago (1s)                                    5
# Sample product reviews
reviews = [
    "This product is amazing and works perfectly!",
    "Terrible quality, totally disappointed",
    "Good value for money",
    "Worst purchase I have ever made"
]

results = classifier(reviews)
results
```

```
[{'label': 'POSITIVE', 'score': 0.9998794794082642},
 {'label': 'NEGATIVE', 'score': 0.9997342228889465},
 {'label': 'POSITIVE', 'score': 0.9998501539230347},
 {'label': 'NEGATIVE', 'score': 0.9997796416282654}]
```

Create AI-powered insights

```python
▶       ✓ 4 minutes ago (10s)                                   7
sentiment_df = spark.createDataFrame(
    [(reviews[i], results[i]['label'], results[i]['score']) for i in range(len(reviews))],
    ["review_text", "sentiment", "confidence"]
)

sentiment_df.display()
```

› 📊 See performance (1)

▯ Day14_DataBricks ✕   +

Catalog   ↻ + ✕

File Edit View Run Help   Python ⌄   Tabs: ON ⌄   ☆   Last edit was 6 hours ago   ▦   ▶ Run all   ● Serverless ⌄   Schedule   Share

Type to search...   ⇟

For you   All

> ☲ See performance (1)

> ▦ sentiment_df: pyspark.sql.connect.dataframe.DataFrame = [review_text: string, sentiment: string ... 1 more field]

⌄ My organization
  > ⛁ workspace
  > ⛁ system
  > ⛁ ecommerce

Table ⌄   +   🔍 ▽ ⫶ ⊟

	review_text	sentiment	1.2 confidence
1	This product is amazing and works perfectl...	POSITIVE	0.9998794794082642
2	Terrible quality, totally disappointed	NEGATIVE	0.9997342228889465
3	Good value for money	POSITIVE	0.9998501539230347
4	Worst purchase I have ever made	NEGATIVE	0.9997796416282654

⌄ Delta Shares Received
  > ⛁ samples

⤓ ⌄   4 rows | 10.24s runtime   Refreshed 3 minutes ago

Mosaic AI Exploration (MLflow logging)

▶ ⌄   ✓ 08:01 PM (1s)   9   Python 🗑 ✦ ⛶ ⫶

```python
with mlflow.start_run(run_name="sentiment_analysis_model"):
    mlflow.log_param("model", "distilbert-base-uncased-finetuned-sst-2")
    mlflow.log_metric("avg_confidence",
                      sum([r["score"] for r in results]) / len(results))
```

▼ (1) MLflow run

Logged 1 run ⧉ to an experiment ⧉ in MLflow. Learn more ⧉

[Shift+Enter] to run and move to next cell
[Ctrl+Shift+P] to open the command palette
[Esc H] to see all keyboard shortcuts