```
# Étape 1: Installer les bibliothèques nécessaires
# Exécutez cette cellule pour installer les dépendances
!pip install transformers datasets torch pandas scikit-learn

Requirement already satisfied: transformers in
/usr/local/lib/python3.10/dist-packages (4.47.1)
Collecting datasets
  Downloading datasets-3.2.0-py3-none-any.whl.metadata (20 kB)
Requirement already satisfied: torch in
/usr/local/lib/python3.10/dist-packages (2.5.1+cu121)
Requirement already satisfied: pandas in
/usr/local/lib/python3.10/dist-packages (2.2.2)
Requirement already satisfied: scikit-learn in
/usr/local/lib/python3.10/dist-packages (1.6.0)
Requirement already satisfied: filelock in
/usr/local/lib/python3.10/dist-packages (from transformers) (3.16.1)
Requirement already satisfied: huggingface-hub<1.0,>=0.24.0 in
/usr/local/lib/python3.10/dist-packages (from transformers) (0.27.0)
Requirement already satisfied: numpy>=1.17 in
/usr/local/lib/python3.10/dist-packages (from transformers) (1.26.4)
Requirement already satisfied: packaging>=20.0 in
/usr/local/lib/python3.10/dist-packages (from transformers) (24.2)
Requirement already satisfied: pyyaml>=5.1 in
/usr/local/lib/python3.10/dist-packages (from transformers) (6.0.2)
Requirement already satisfied: regex!=2019.12.17 in
/usr/local/lib/python3.10/dist-packages (from transformers)
(2024.11.6)
Requirement already satisfied: requests in
/usr/local/lib/python3.10/dist-packages (from transformers) (2.32.3)
Requirement already satisfied: tokenizers<0.22,>=0.21 in
/usr/local/lib/python3.10/dist-packages (from transformers) (0.21.0)
Requirement already satisfied: safetensors>=0.4.1 in
/usr/local/lib/python3.10/dist-packages (from transformers) (0.4.5)
Requirement already satisfied: tqdm>=4.27 in
/usr/local/lib/python3.10/dist-packages (from transformers) (4.67.1)
Requirement already satisfied: pyarrow>=15.0.0 in
/usr/local/lib/python3.10/dist-packages (from datasets) (17.0.0)
Collecting dill<0.3.9,>=0.3.0 (from datasets)
  Downloading dill-0.3.8-py3-none-any.whl.metadata (10 kB)
Collecting xxhash (from datasets)
  Downloading xxhash-3.5.0-cp310-cp310-
manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (12 kB)
Collecting multiprocess<0.70.17 (from datasets)
  Downloading multiprocess-0.70.16-py310-none-any.whl.metadata (7.2
kB)
Collecting fsspec<=2024.9.0,>=2023.1.0 (from
fsspec[http]<=2024.9.0,>=2023.1.0->datasets)
  Downloading fsspec-2024.9.0-py3-none-any.whl.metadata (11 kB)
Requirement already satisfied: aiohttp in
/usr/local/lib/python3.10/dist-packages (from datasets) (3.11.10)
```

```
Requirement already satisfied: typing-extensions>=4.8.0 in
/usr/local/lib/python3.10/dist-packages (from torch) (4.12.2)
Requirement already satisfied: networkx in
/usr/local/lib/python3.10/dist-packages (from torch) (3.4.2)
Requirement already satisfied: jinja2 in
/usr/local/lib/python3.10/dist-packages (from torch) (3.1.4)
Requirement already satisfied: sympy==1.13.1 in
/usr/local/lib/python3.10/dist-packages (from torch) (1.13.1)
Requirement already satisfied: mpmath<1.4,>=1.1.0 in
/usr/local/lib/python3.10/dist-packages (from sympy==1.13.1->torch)
(1.3.0)
Requirement already satisfied: python-dateutil>=2.8.2 in
/usr/local/lib/python3.10/dist-packages (from pandas) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in
/usr/local/lib/python3.10/dist-packages (from pandas) (2024.2)
Requirement already satisfied: tzdata>=2022.7 in
/usr/local/lib/python3.10/dist-packages (from pandas) (2024.2)
Requirement already satisfied: scipy>=1.6.0 in
/usr/local/lib/python3.10/dist-packages (from scikit-learn) (1.13.1)
Requirement already satisfied: joblib>=1.2.0 in
/usr/local/lib/python3.10/dist-packages (from scikit-learn) (1.4.2)
Requirement already satisfied: threadpoolctl>=3.1.0 in
/usr/local/lib/python3.10/dist-packages (from scikit-learn) (3.5.0)
Requirement already satisfied: aiohappyeyeballs>=2.3.0 in
/usr/local/lib/python3.10/dist-packages (from aiohttp->datasets)
(2.4.4)
Requirement already satisfied: aiosignal>=1.1.2 in
/usr/local/lib/python3.10/dist-packages (from aiohttp->datasets)
(1.3.2)
Requirement already satisfied: async-timeout<6.0,>=4.0 in
/usr/local/lib/python3.10/dist-packages (from aiohttp->datasets)
(4.0.3)
Requirement already satisfied: attrs>=17.3.0 in
/usr/local/lib/python3.10/dist-packages (from aiohttp->datasets)
(24.3.0)
Requirement already satisfied: frozenlist>=1.1.1 in
/usr/local/lib/python3.10/dist-packages (from aiohttp->datasets)
(1.5.0)
Requirement already satisfied: multidict<7.0,>=4.5 in
/usr/local/lib/python3.10/dist-packages (from aiohttp->datasets)
(6.1.0)
Requirement already satisfied: propcache>=0.2.0 in
/usr/local/lib/python3.10/dist-packages (from aiohttp->datasets)
(0.2.1)
Requirement already satisfied: yarl<2.0,>=1.17.0 in
/usr/local/lib/python3.10/dist-packages (from aiohttp->datasets)
(1.18.3)
Requirement already satisfied: six>=1.5 in
/usr/local/lib/python3.10/dist-packages (from python-dateutil>=2.8.2-
```

```
>pandas) (1.17.0)
Requirement already satisfied: charset-normalizer<4,>=2 in
/usr/local/lib/python3.10/dist-packages (from requests->transformers)
(3.4.0)
Requirement already satisfied: idna<4,>=2.5 in
/usr/local/lib/python3.10/dist-packages (from requests->transformers)
(3.10)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/usr/local/lib/python3.10/dist-packages (from requests->transformers)
(2.2.3)
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.10/dist-packages (from requests->transformers)
(2024.12.14)
Requirement already satisfied: MarkupSafe>=2.0 in
/usr/local/lib/python3.10/dist-packages (from jinja2->torch) (3.0.2)
Downloading datasets-3.2.0-py3-none-any.whl (480 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 480.6/480.6 kB 8.3 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 116.3/116.3 kB 367.1 kB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 179.3/179.3 kB 11.7 MB/s eta
0:00:00
ultiprocess-0.70.16-py310-none-any.whl (134 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 134.8/134.8 kB 9.9 MB/s eta
0:00:00
anylinux_2_17_x86_64.manylinux2014_x86_64.whl (194 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 194.1/194.1 kB 11.2 MB/s eta
0:00:00
ultiprocess, datasets
  Attempting uninstall: fsspec
    Found existing installation: fsspec 2024.10.0
    Uninstalling fsspec-2024.10.0:
      Successfully uninstalled fsspec-2024.10.0
ERROR: pip's dependency resolver does not currently take into account
all the packages that are installed. This behaviour is the source of
the following dependency conflicts.
gcsfs 2024.10.0 requires fsspec==2024.10.0, but you have fsspec
2024.9.0 which is incompatible.
Successfully installed datasets-3.2.0 dill-0.3.8 fsspec-2024.9.0
multiprocess-0.70.16 xxhash-3.5.0

import pandas as pd
from sklearn.model_selection import train_test_split
from transformers import AutoTokenizer,
AutoModelForSequenceClassification, Trainer, TrainingArguments
from datasets import Dataset

Exception ignored in: <function _xla_gc_callback at 0x7f54733156c0>
Traceback (most recent call last):
  File
```

```
"/usr/local/lib/python3.10/dist-packages/jax/_src/lib/__init__.py",
line 96, in _xla_gc_callback
    def _xla_gc_callback(*args):
KeyboardInterrupt:
```

```
# Étape 2: Télécharger le dataset depuis GitHub
# Télécharger directement "goodbooks-10k.csv"
!wget https://raw.githubusercontent.com/zygmuntz/goodbooks-
10k/master/books.csv -O goodbooks-10k.csv
```

```
--2024-12-24 19:07:19--
https://raw.githubusercontent.com/zygmuntz/goodbooks-10k/master/books.
csv
Resolving raw.githubusercontent.com (raw.githubusercontent.com)...
185.199.109.133, 185.199.111.133, 185.199.108.133, ...
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|
185.199.109.133|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 3286659 (3.1M) [text/plain]
Saving to: ‘goodbooks-10k.csv’

goodbooks-10k.csv   100%[===================>]   3.13M  --.-KB/s    in
0.08s

2024-12-24 19:07:19 (39.4 MB/s) - ‘goodbooks-10k.csv’ saved
[3286659/3286659]
```

```
# Charger les données
books_df = pd.read_csv("goodbooks-10k.csv")
print(books_df.columns)
```

```
Index(['book_id', 'goodreads_book_id', 'best_book_id', 'work_id',
       'books_count', 'isbn', 'isbn13', 'authors',
'original_publication_year',
       'original_title', 'title', 'language_code', 'average_rating',
       'ratings_count', 'work_ratings_count',
'work_text_reviews_count',
       'ratings_1', 'ratings_2', 'ratings_3', 'ratings_4',
'ratings_5',
       'image_url', 'small_image_url'],
      dtype='object')
```

```
# Garder uniquement les colonnes nécessaires
books_df = books_df[["book_id", "title", "authors", "average_rating",
"language_code", "work_text_reviews_count"]]
```

```
#  Étape 3: Prétraitement des données
# Combiner "title" et "work_text_reviews_count" comme entrée, et
"average_rating" comme label
books_df = books_df.dropna()
```

```
books_df["text"] = books_df["title"] + " - " +
books_df["work_text_reviews_count"].astype(str)
books_df["label"] = (books_df["average_rating"] >= 4.0).astype(int)  #
Label 1 pour les livres bien notés

# Séparer les données en train et test
train_texts, test_texts, train_labels, test_labels = train_test_split(
    books_df["text"].tolist(), books_df["label"].tolist(),
test_size=0.2, random_state=42
)

# Étape 4: Préparer le dataset avec Hugging Face
def preprocess_function(examples):
    return tokenizer(examples["text"], truncation=True, padding=True,
max_length=512)

tokenizer = AutoTokenizer.from_pretrained("bert-base-uncased")

train_dataset = Dataset.from_dict({"text": train_texts, "label":
train_labels}).map(preprocess_function, batched=True)
test_dataset = Dataset.from_dict({"text": test_texts, "label":
test_labels}).map(preprocess_function, batched=True)

# Étape 5: Charger un modèle pré-entraîné
model = AutoModelForSequenceClassification.from_pretrained("bert-base-
uncased", num_labels=2)
```


```
{"model_id":"7254dd8cf601465ca26f05ec76f89f59","version_major":2,"vers
ion_minor":0}

{"model_id":"d7301dae30be411ea35a31be1ae33f09","version_major":2,"vers
ion_minor":0}

{"model_id":"ea557e2b8cf74999a14265d1d68a0b5d","version_major":2,"vers
ion_minor":0}
```


```
Some weights of BertForSequenceClassification were not initialized
from the model checkpoint at bert-base-uncased and are newly
initialized: ['classifier.bias', 'classifier.weight']
You should probably TRAIN this model on a down-stream task to be able
to use it for predictions and inference.
```

```
from transformers import TrainingArguments

training_args = TrainingArguments(
    output_dir="./results",
    evaluation_strategy="steps",  # Evaluation plus flexible
    eval_steps=500,  # Evaluer toutes les 500 étapes
    learning_rate=3e-5,  # Légèrement plus grand pour convergence
rapide
    per_device_train_batch_size=16,  # Augmenté pour traiter plus de
```

```
données à chaque étape
    per_device_eval_batch_size=16,  # Aligné avec la taille de lot
d'entraînement
    num_train_epochs=2,  # Réduction à 2 époques
    weight_decay=0.01,  # Légère augmentation pour régularisation
    logging_dir="./logs",
    logging_steps=100,  # Réduction des interruptions fréquentes
    save_steps=1000,  # Sauvegarde moins fréquente
    save_strategy="steps",
    load_best_model_at_end=True,
    fp16=True  # Utiliser la précision mixte pour accélérer sur GPU
)

# Étape 7: Créer un Trainer pour l'entraînement
trainer = Trainer(
    model=model,
    args=training_args,
    train_dataset=train_dataset,
    eval_dataset=test_dataset,
    tokenizer=tokenizer,
)
```

```
<ipython-input-16-8521e9635f01>:2: FutureWarning: `tokenizer` is
deprecated and will be removed in version 5.0.0 for
`Trainer.__init__`. Use `processing_class` instead.
  trainer = Trainer(
```

```
# Étape 8: Entraîner le modèle
trainer.train()
```

```
<IPython.core.display.HTML object>
```

```
Could not locate the best model at
./results/checkpoint-500/pytorch_model.bin, if you are running a
distributed training on multiple nodes, you should activate `--
save_on_each_node`.
```

```
TrainOutput(global_step=892, training_loss=0.5902736689477758,
metrics={'train_runtime': 9043.5159, 'train_samples_per_second':
1.577, 'train_steps_per_second': 0.099, 'total_flos':
387632473419360.0, 'train_loss': 0.5902736689477758, 'epoch': 2.0})
```

```
# Étape 9: Évaluer le modèle
results = trainer.evaluate()
print("Résultats de l'évaluation:", results)
```

```
<IPython.core.display.HTML object>
```

```
Résultats de l'évaluation: {'eval_loss': 0.6361607909202576,
'eval_runtime': 257.6122, 'eval_samples_per_second': 6.925,
'eval_steps_per_second': 0.435, 'epoch': 2.0}
```

```python
# Étape 10: Sauvegarder le modèle
trainer.save_model("./recommender_model")

# Étape 11: Tester le système de recommandation
# Exemple d'entrée utilisateur
user_input = "Fantasy - A magical world full of adventures."
inputs = tokenizer(user_input, return_tensors="pt", truncation=True,
padding=True, max_length=512)
outputs = model(**inputs)
prediction = outputs.logits.argmax(-1).item()
print("Recommandation pour cette description: ", "Recommandé" if
prediction == 1 else "Non recommandé")
```

```
Recommandation pour cette description:  Recommandé
```

```python
user_input = "THE SHARDS"
inputs = tokenizer(user_input, return_tensors="pt", truncation=True,
padding=True, max_length=512)
outputs = model(**inputs)
prediction = outputs.logits.argmax(-1).item()
print("Recommandation pour cette description: ", "Recommandé" if
prediction == 1 else "Non recommandé")
```

```
Recommandation pour cette description:  Non recommandé
```