Performing Summarization Using a Hugging Face Pipeline

Skills covered: Automatic Text Summarization, Artificial Intelligence

Muhammad Jamil

Senior Software Engineer

Artificial Intelligence Technical

Using Colab to work with Hugging Face Transformers

- **Task**
  - In this lab, we'll see how we can write Python code in a Colab notebook to access and use a pre-trained transformer model from the Hugging Face Transformers library. We'll pick a pre-trained model that works well for text summarization, we'll fine-tune that model on a dataset that we load, that is, the CNN Daily Mail dataset and we'll generate summaries and compute the ROUGE score for that model on our data.

Tasks Libraries Datasets Languages Licenses
Other

Filter Tasks by name

Multimodal

- Image-Text-to-Text
- Visual Question Answering
- Document Question Answering
- Video-Text-to-Text
- Any-to-Any

Computer Vision

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

Models 4,809

t5-small

Full-text search

Sort: Trending

google-t5/t5-small
Translation • Updated Jun 30, 2023 • 6.35M • 318

google/flan-t5-small
Text2Text Generation • Updated Oct 10, 2023 • 324k • 251

allenai/t5-small-squad2-question-generation
Text2Text Generation • Updated Jan 24, 2023 • 725 • 39

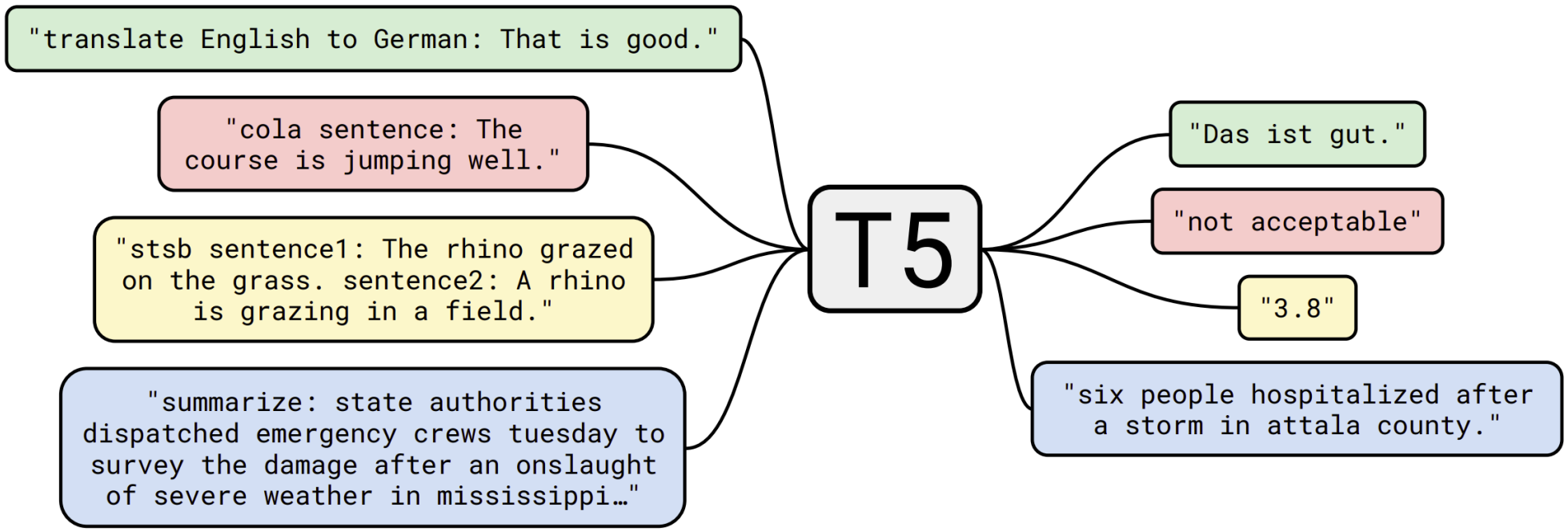
flax-community/arabic-t5-small
Text2Text Generation • Updated Nov 29, 2023 • 78 • 7

google/t5-v1_1-small
Text2Text Generation • Updated Jan 24, 2023 • 22.3k • 23

somosnlp-hackathon-2022/t5-small-spanish-nahuatl
Translation • Updated Jan 24 • 31 • 12

paust/pko-t5-small
Text2Text Generation • Updated Sep 14, 2022 • 786 • 5

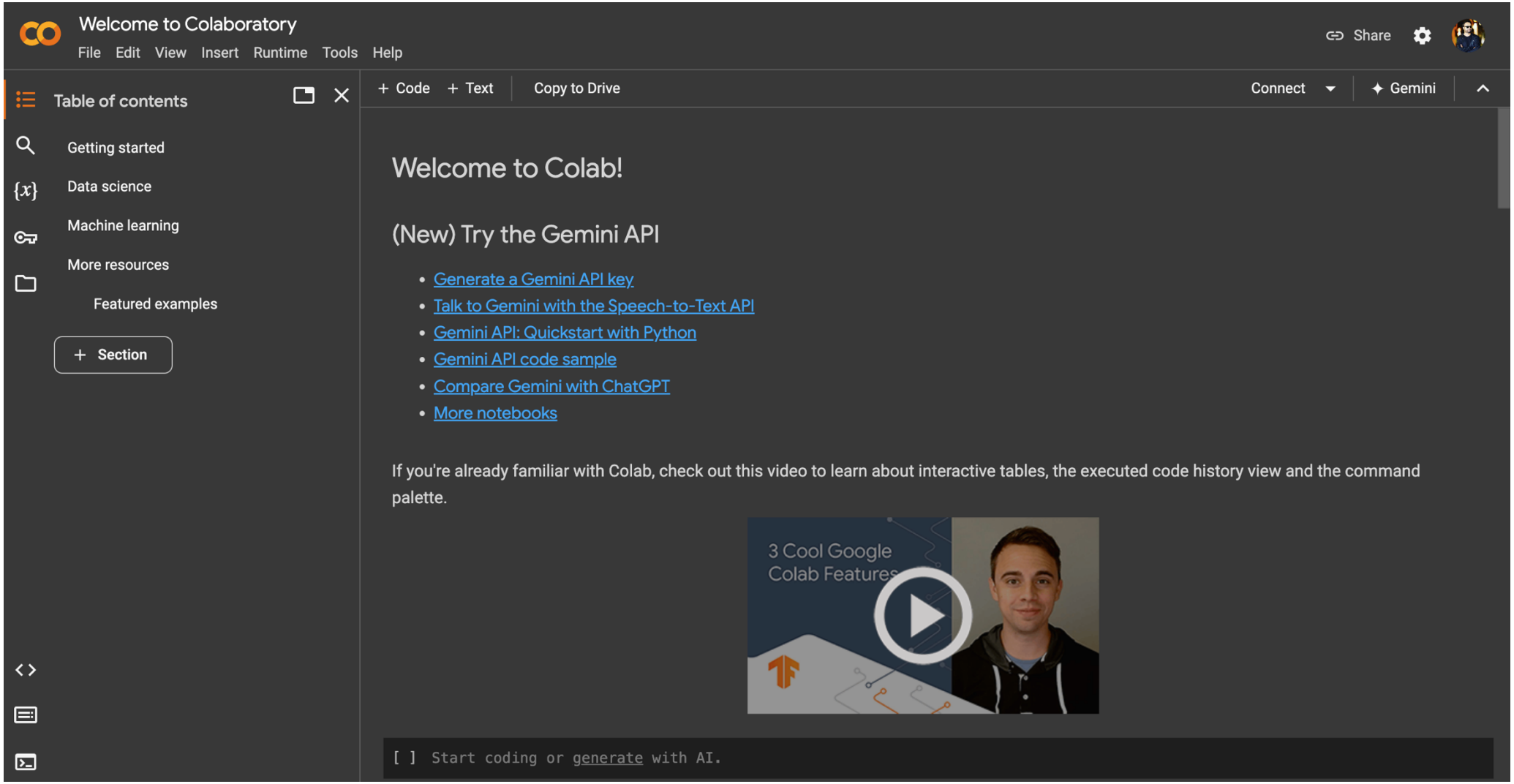
Model Card for T5 Small



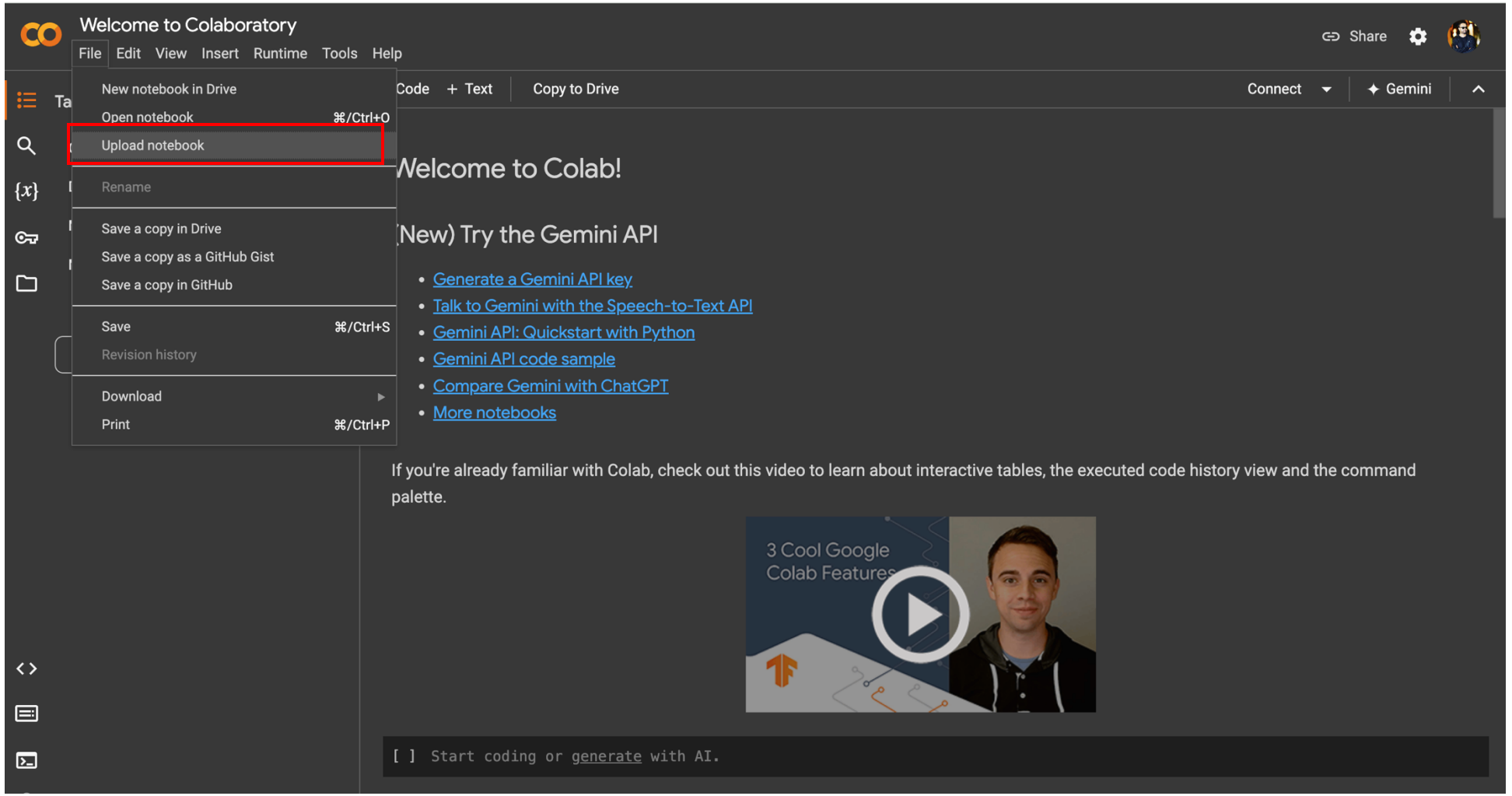
Model Description

- With T5, we propose reframing all NLP tasks into a unified text-to-text-format where the input and output are always text strings, in contrast to BERT-style models that can only output either a class label or a span of the input. Our text-to-text framework allows us to use the same model, loss function, and hyperparameters on any NLP task.

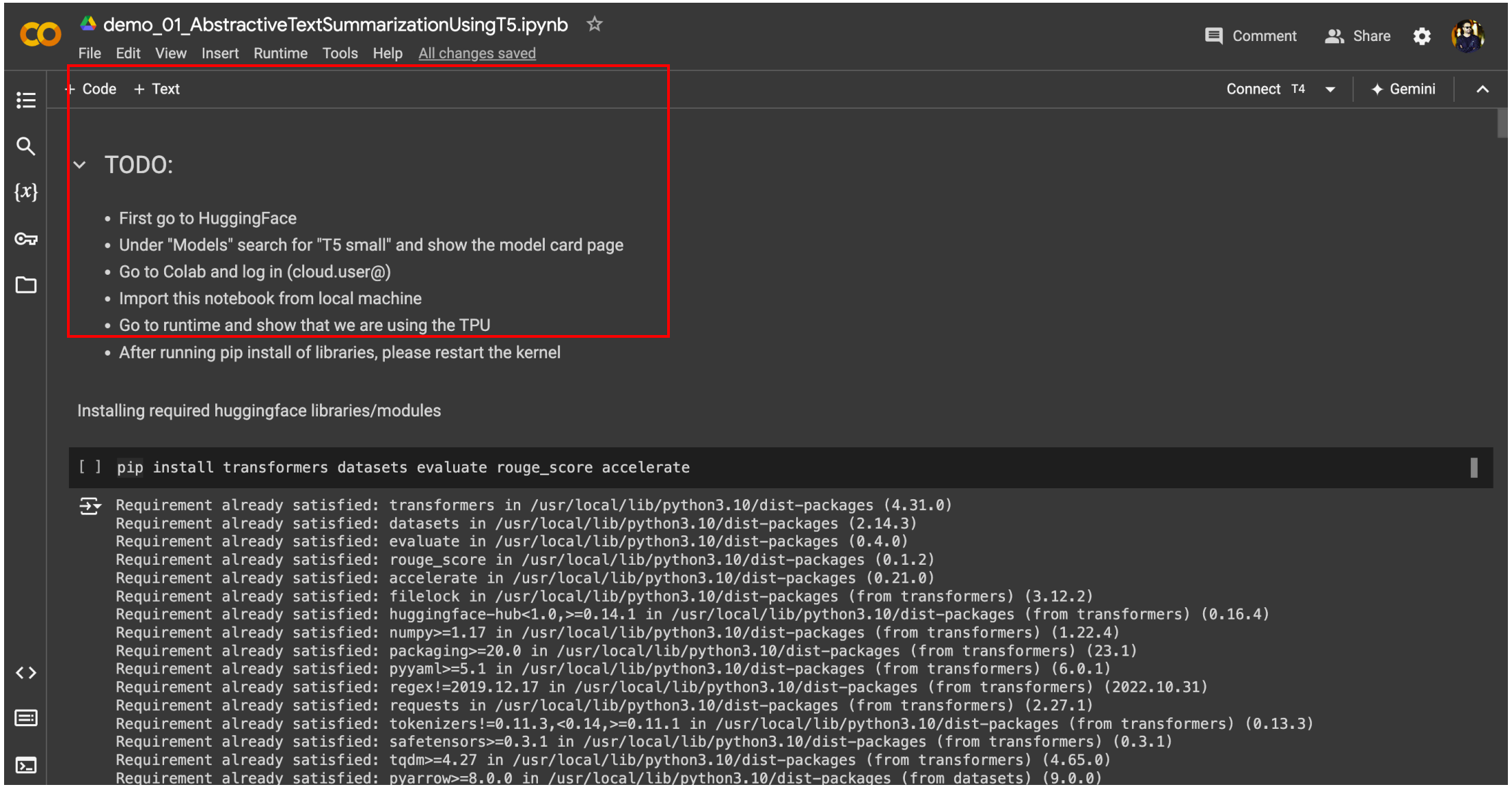
Colab Interface



Opening a notebook



TODO:


demo_01_AbstractiveTextSummarizationUsingT5.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment
Share
+ Code + Text
Connect T4
Gemini
TODO:
First go to HuggingFace
Under "Models" search for "T5 small" and show the model card page
Go to Colab and log in (cloud.user@)
Import this notebook from local machine
Go to runtime and show that we are using the TPU
After running pip install of libraries, please restart the kernel
Installing required huggingface libraries/modules
[ ] pip install transformers datasets evaluate rouge_score accelerate
Requirement already satisfied: transformers in /usr/local/lib/python3.10/dist-packages (4.31.0)
Requirement already satisfied: datasets in /usr/local/lib/python3.10/dist-packages (2.14.3)
Requirement already satisfied: evaluate in /usr/local/lib/python3.10/dist-packages (0.4.0)
Requirement already satisfied: rouge_score in /usr/local/lib/python3.10/dist-packages (0.1.2)
Requirement already satisfied: accelerate in /usr/local/lib/python3.10/dist-packages (0.21.0)
Requirement already satisfied: filelock in /usr/local/lib/python3.10/dist-packages (from transformers) (3.12.2)
Requirement already satisfied: huggingface-hub<1.0,>=0.14.1 in /usr/local/lib/python3.10/dist-packages (from transformers) (0.16.4)
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.10/dist-packages (from transformers) (1.22.4)
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.10/dist-packages (from transformers) (23.1)
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.10/dist-packages (from transformers) (6.0.1)
Requirement already satisfied: regex!=2019.12.17 in /usr/local/lib/python3.10/dist-packages (from transformers) (2022.10.31)
Requirement already satisfied: requests in /usr/local/lib/python3.10/dist-packages (from transformers) (2.27.1)
Requirement already satisfied: tokenizers!=0.11.3,<0.14,>=0.11.1 in /usr/local/lib/python3.10/dist-packages (from transformers) (0.13.3)
Requirement already satisfied: safetensors>=0.3.1 in /usr/local/lib/python3.10/dist-packages (from transformers) (0.3.1)
Requirement already satisfied: tqdm>=4.27 in /usr/local/lib/python3.10/dist-packages (from transformers) (4.65.0)
Requirement already satisfied: pyarrow>=8.0.0 in /usr/local/lib/python3.10/dist-packages (from datasets) (9.0.0)

T5 Small Model Card Page

Hugging Face | Search models, datasets, users... | Models | Datasets | Spaces | Posts | Docs | Pricing

google-t5 / t5-small | like 319

Translation | Transformers | PyTorch | TensorFlow | JAX | Rust | ONNX | Safetensors | c4 | 5 languages | t5 | text2text-generation | summarization | text-generation-inference | Inference Endpoints | 8 papers | License: apache-2.0

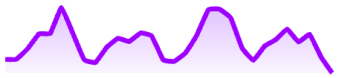
Model card | Files and versions | Community 27 | Train | Deploy | Use this model

Edit model card

Model Card for T5 Small

Table of Contents

Downloads last month
6,282,687



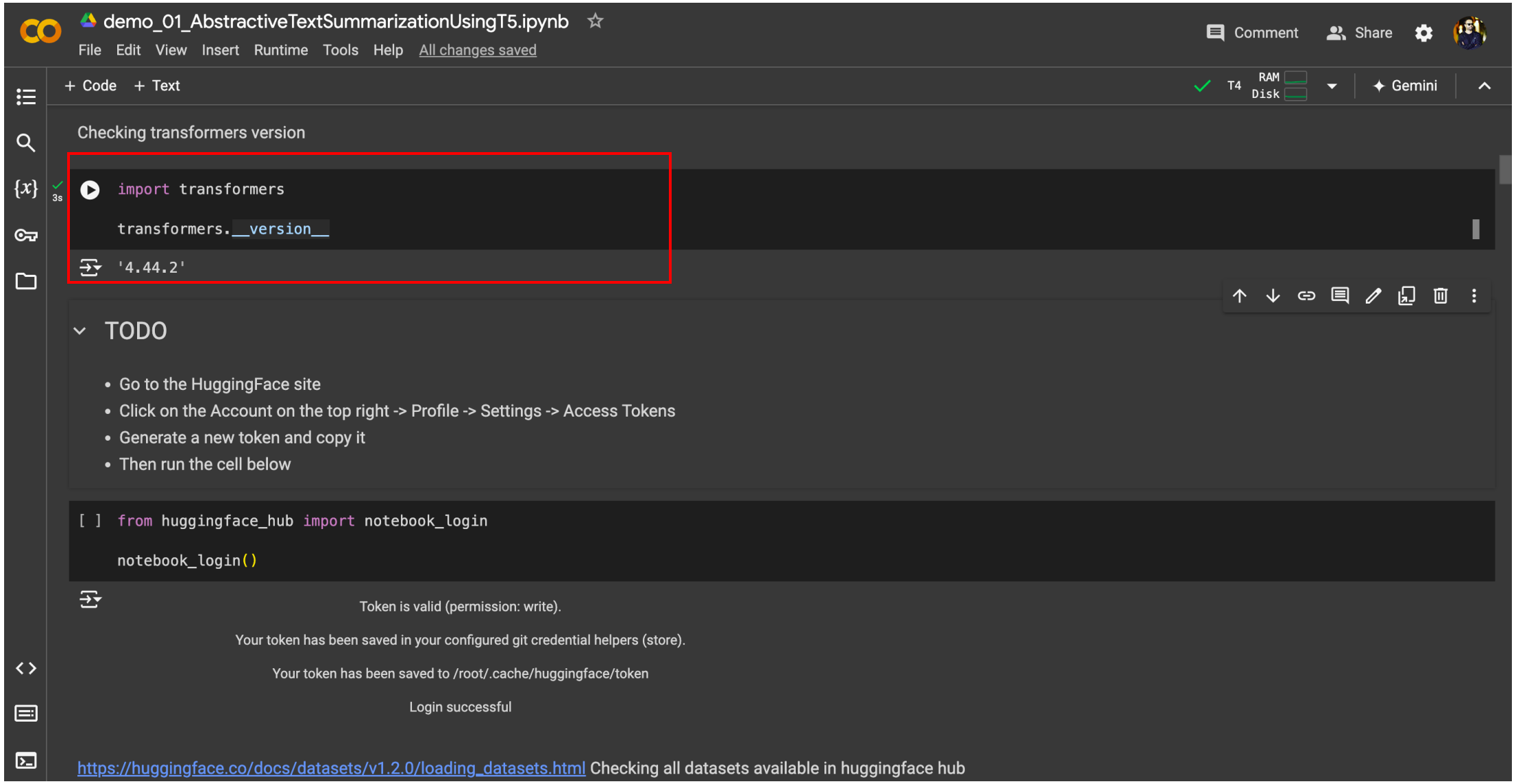
Safetensors | Model size 60.5M params | Tensor type F32

Inference API | Cold

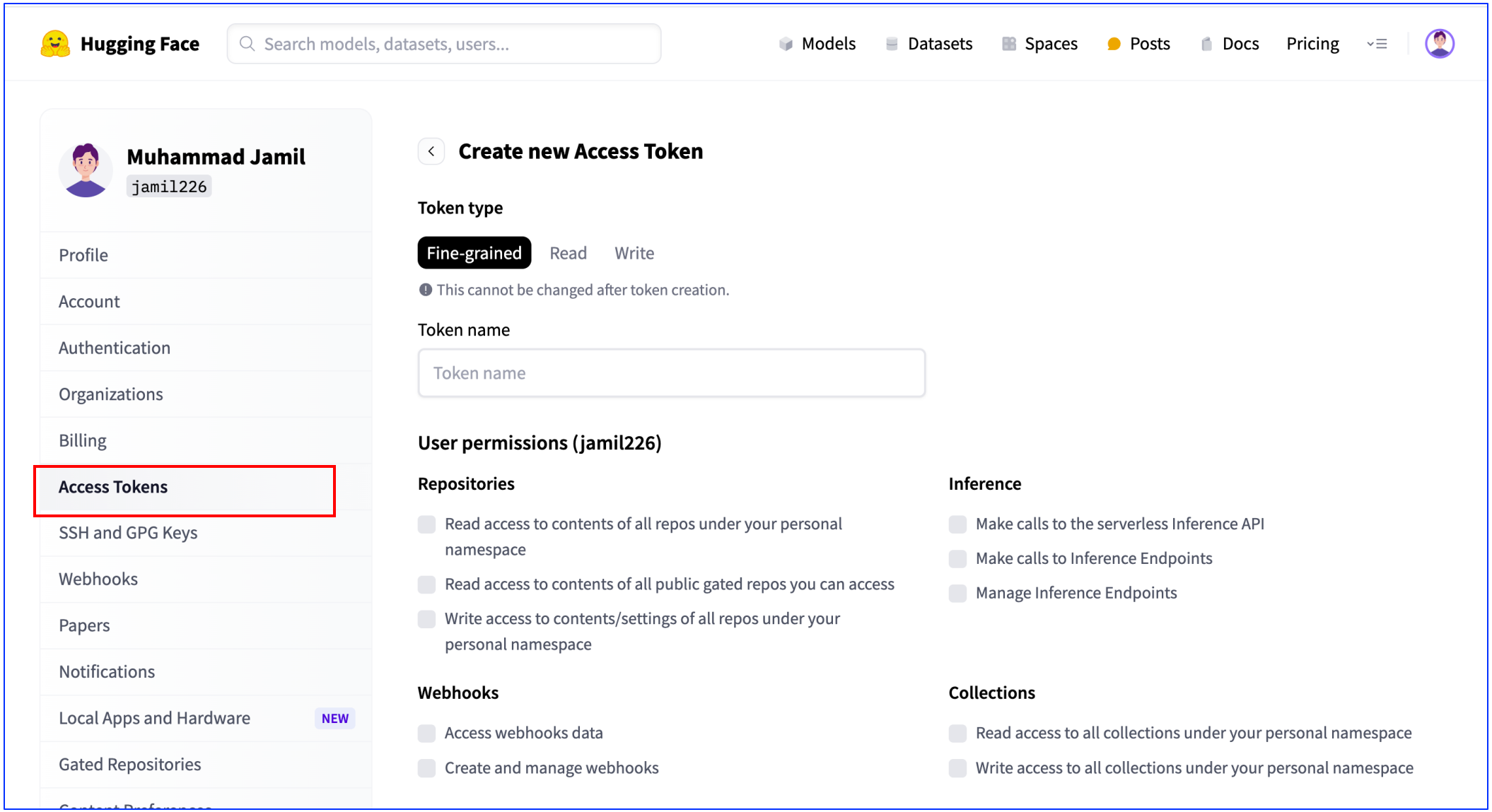
Translation | Examples

My name is Sarah and I live in London

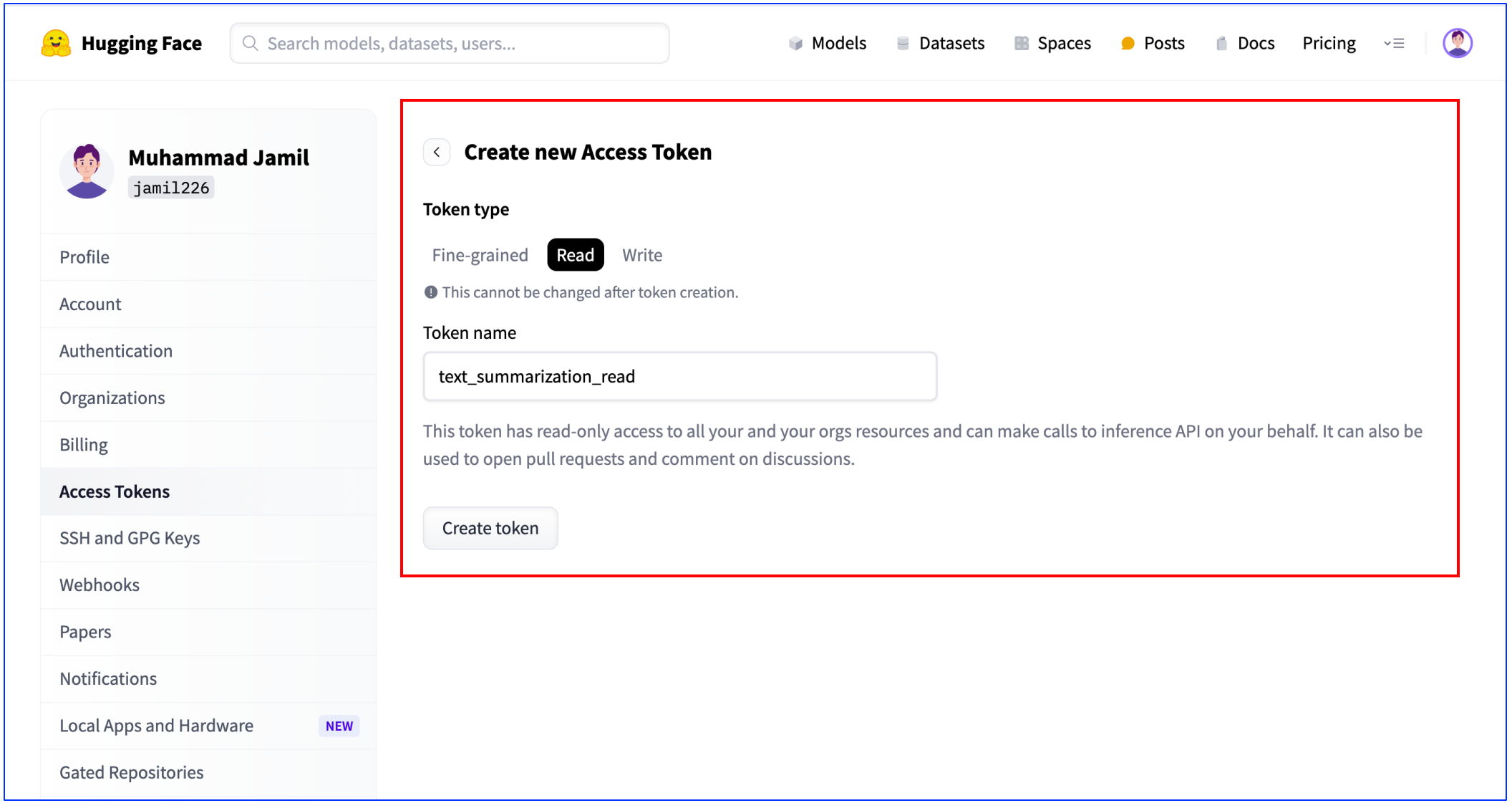
Checking Transformer Version



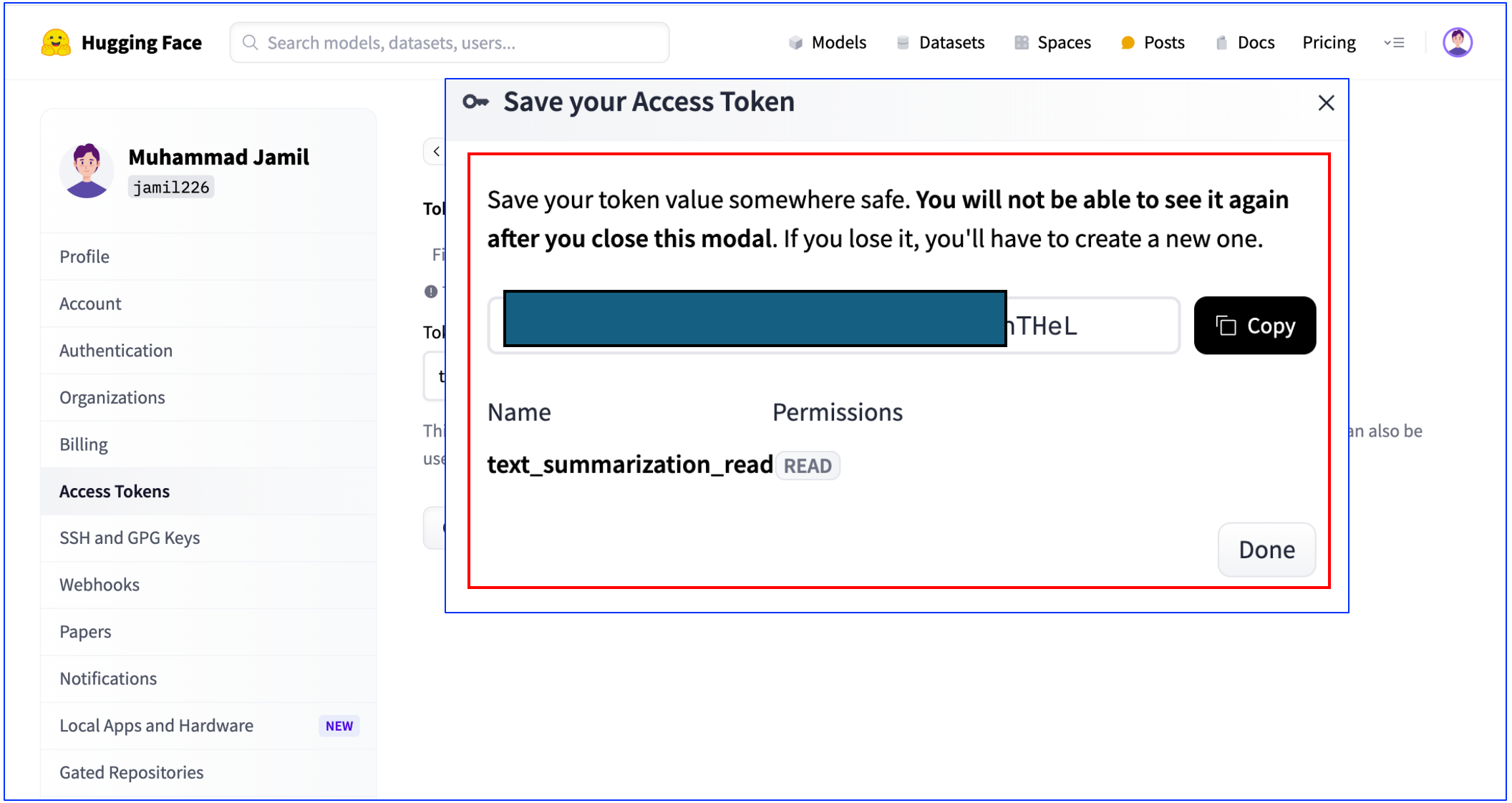
Hugging Face Access Tokens



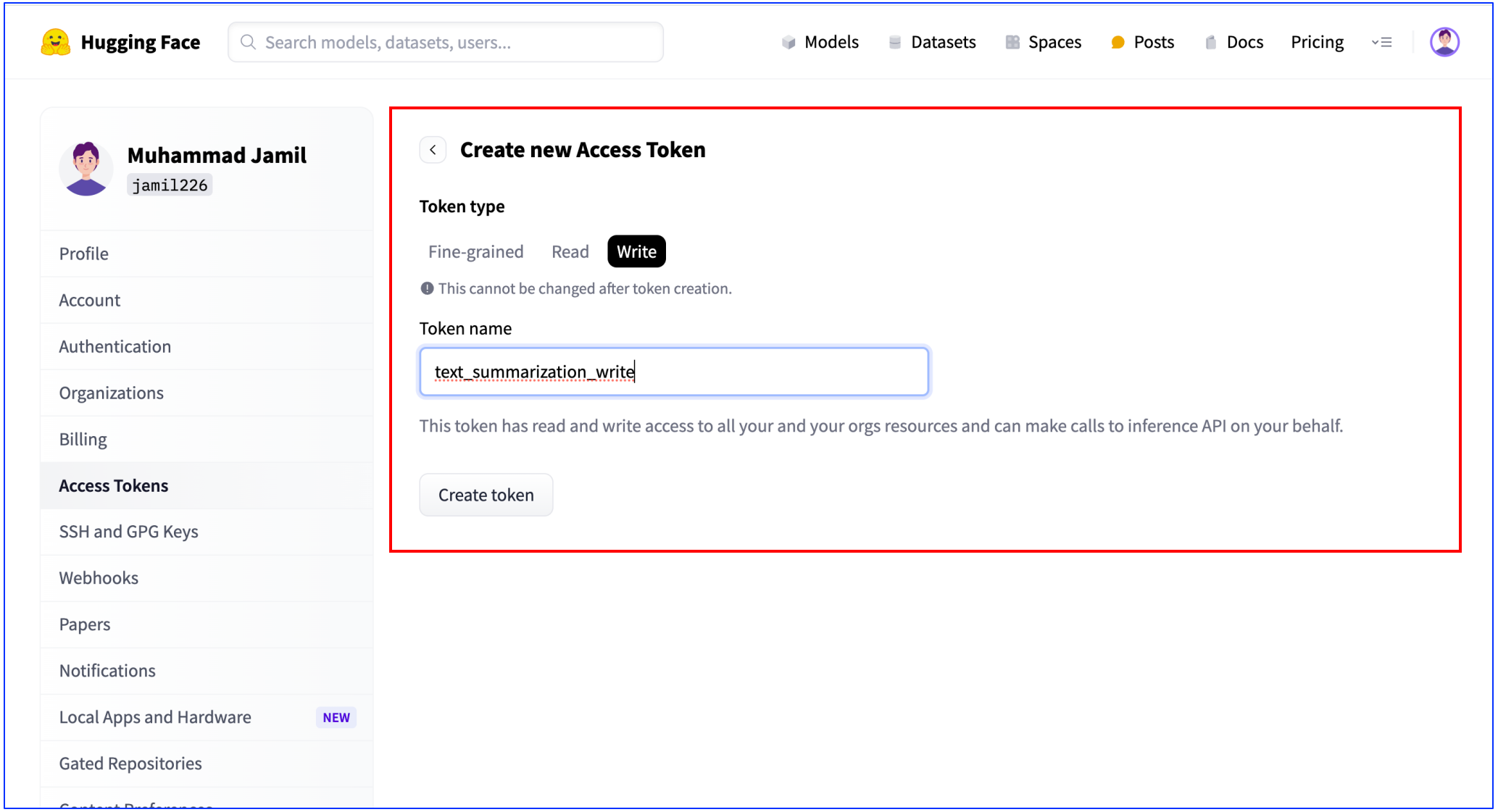
Creating a new Access Token with Read Permission



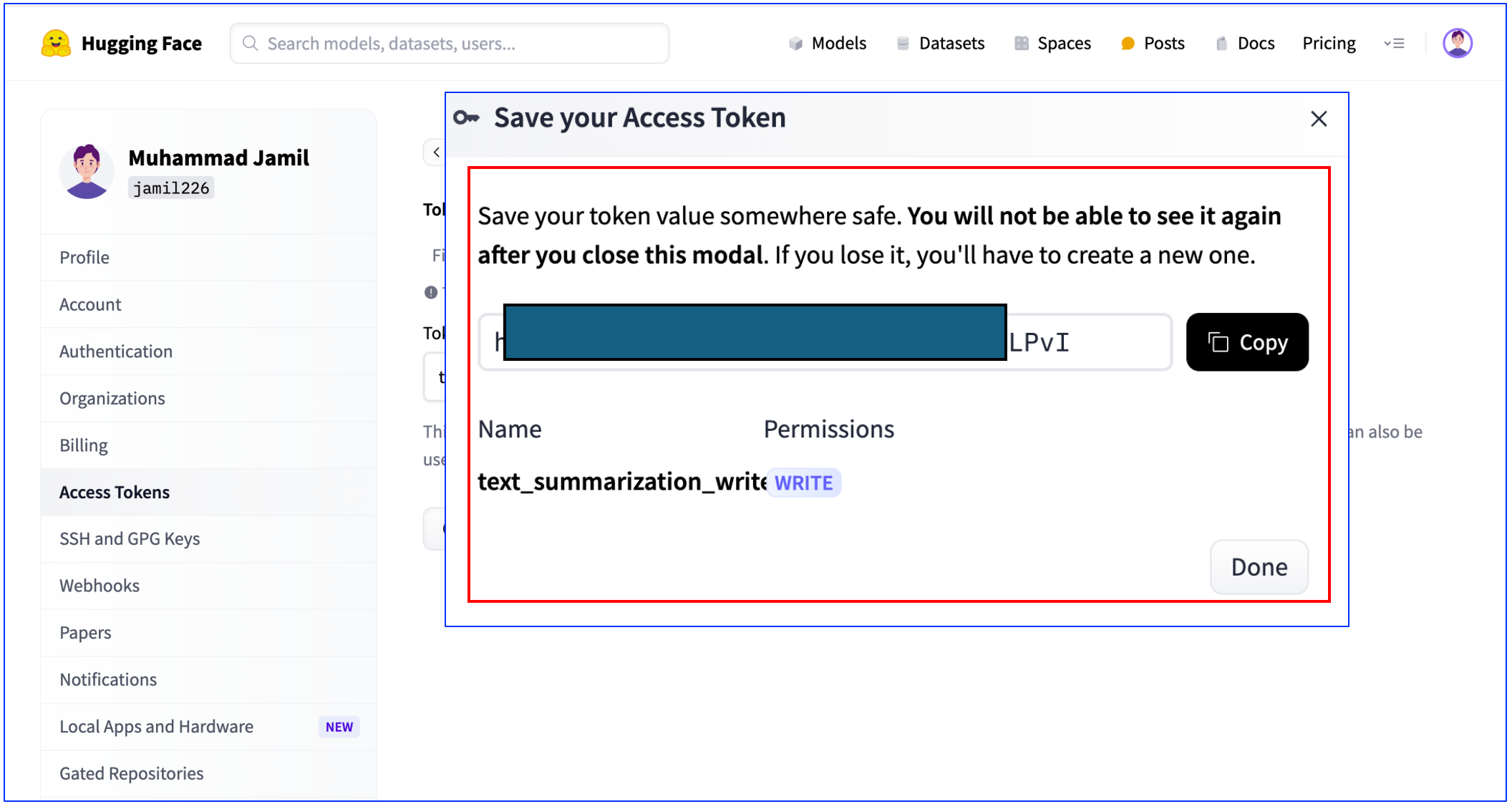
Token Credentials



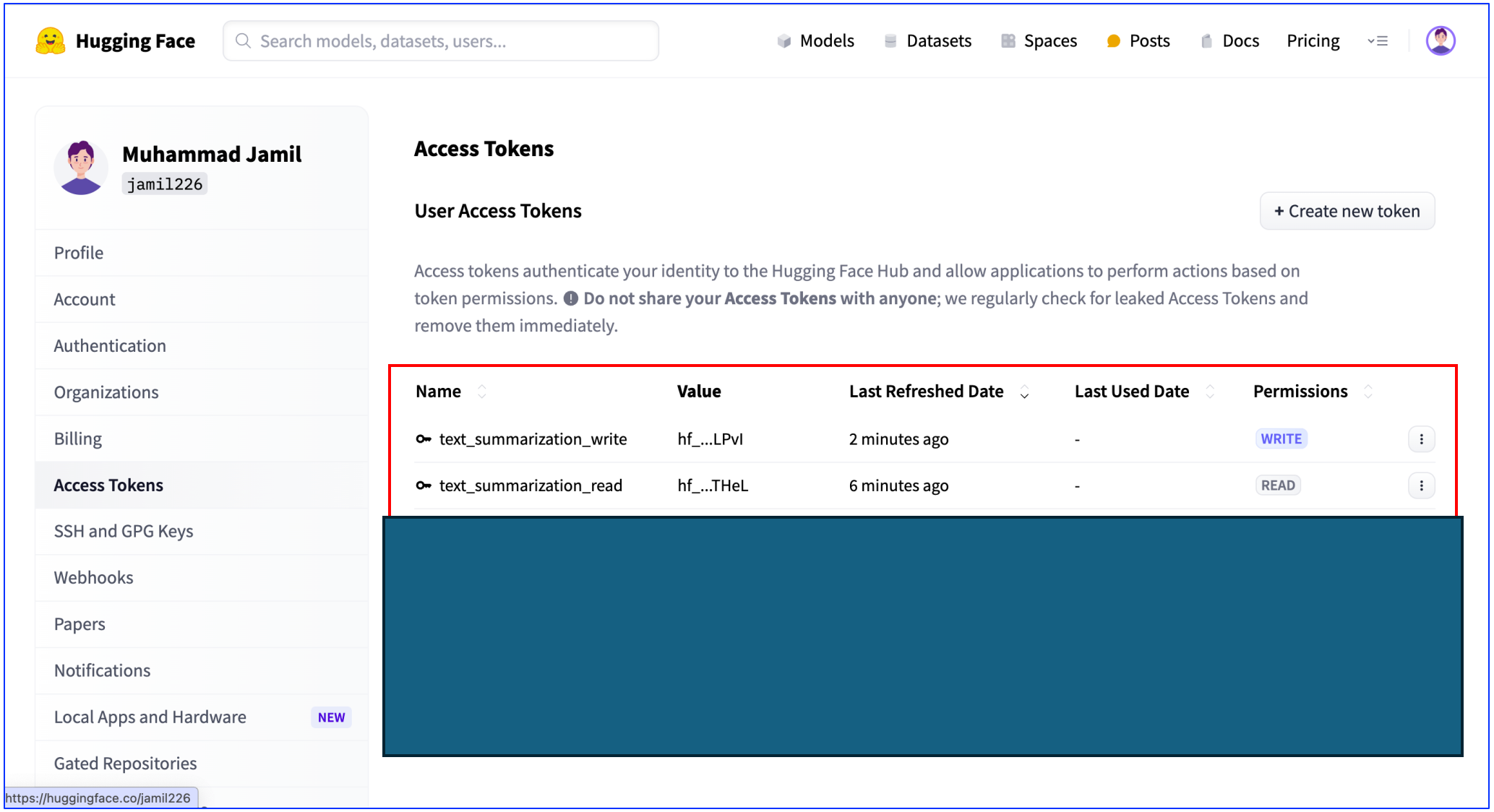
Creating a new Access Token with Write Permission



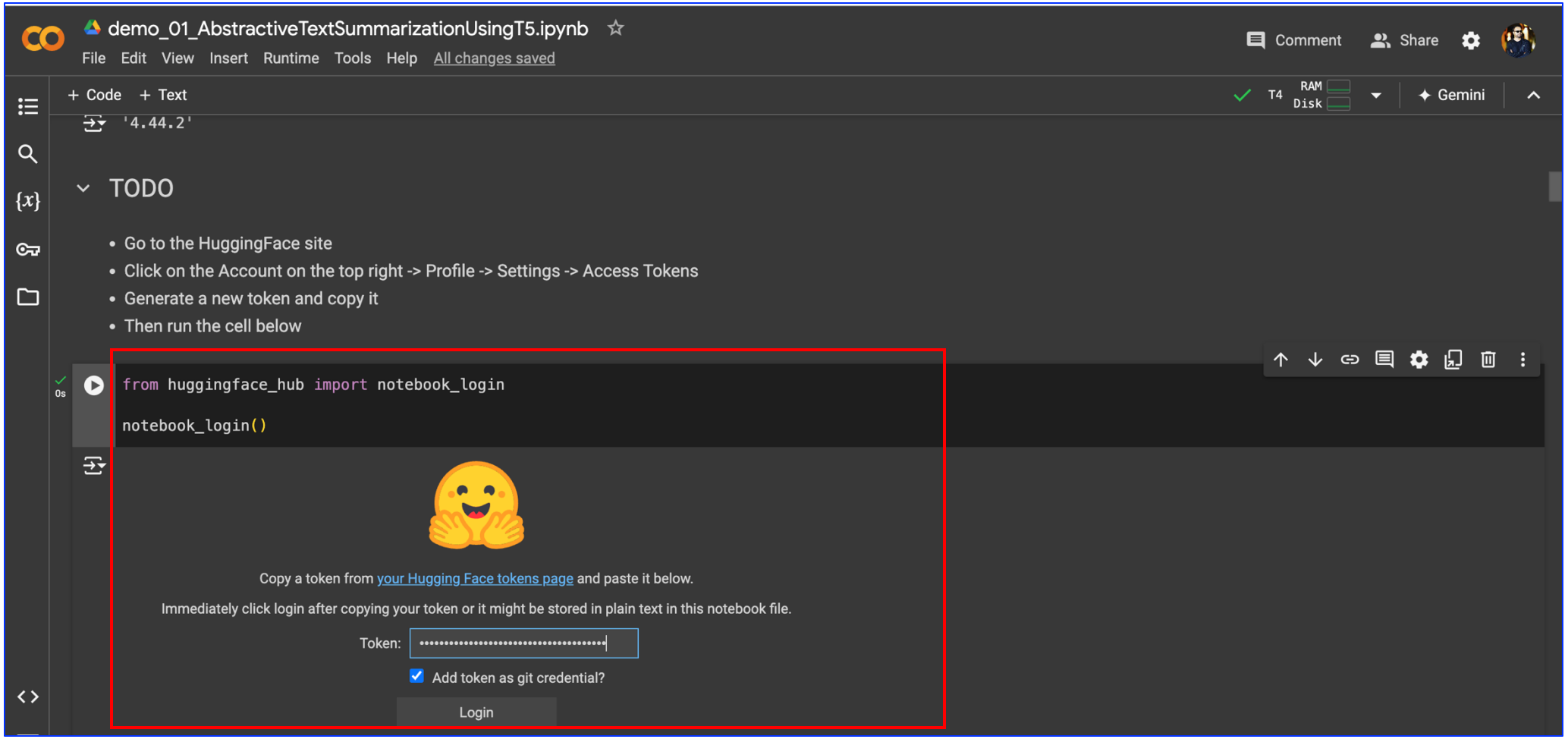
Token Credentials



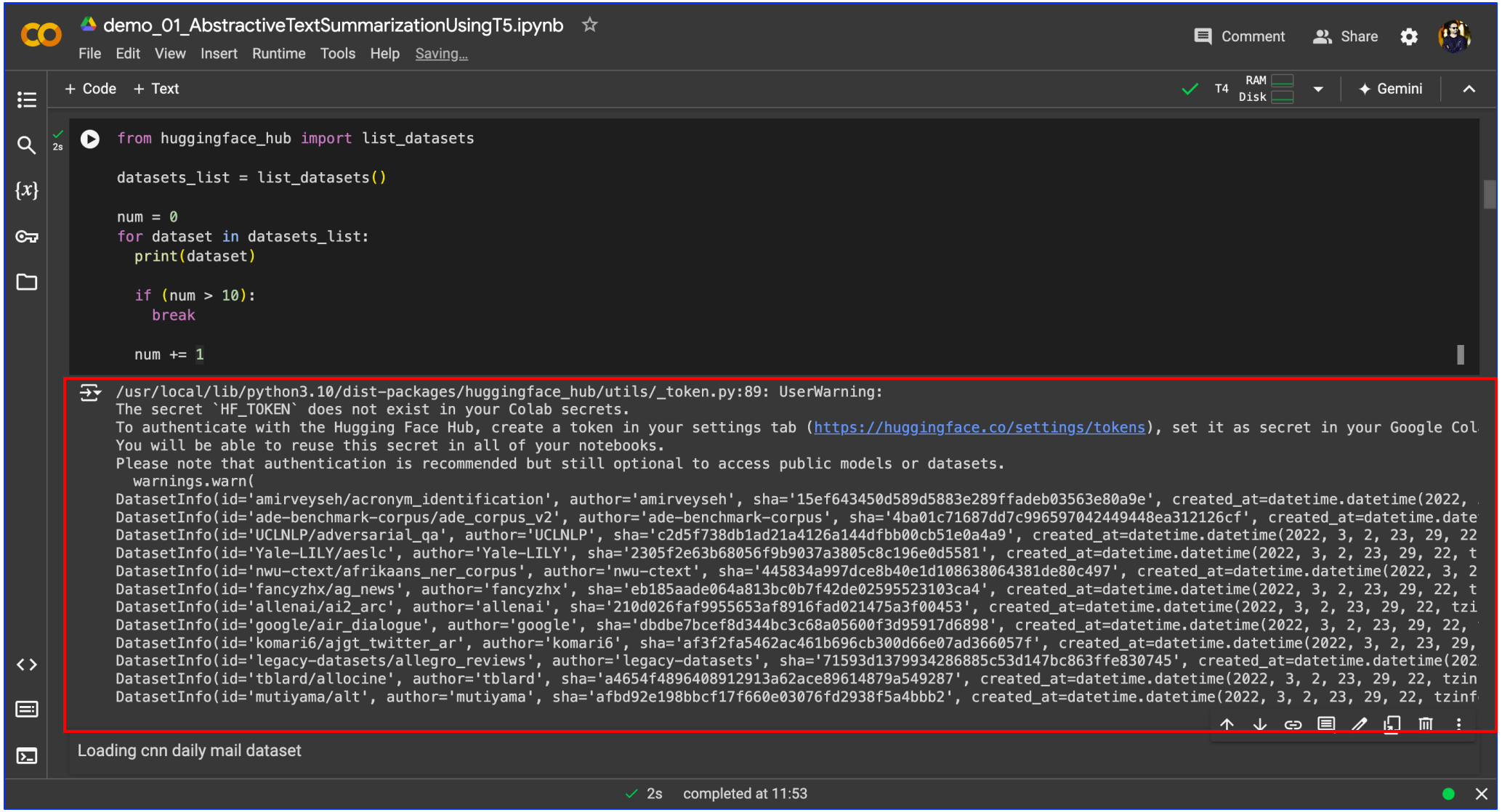
Token List


Hugging Face
Search models, datasets, users...
Models
Datasets
Spaces
Posts
Docs
Pricing
Muhammad Jamil
jamil226
Profile
Account
Authentication
Organizations
Billing
Access Tokens
SSH and GPG Keys
Webhooks
Papers
Notifications
Local Apps and Hardware
NEW
Gated Repositories
https://huggingface.co/jamil226
Access Tokens
User Access Tokens
+ Create new token
Access tokens authenticate your identity to the Hugging Face Hub and allow applications to perform actions based on token permissions. Do not share your Access Tokens with anyone; we regularly check for leaked Access Tokens and remove them immediately.
Name
Value
Last Refreshed Date
Last Used Date
Permissions
text_summarization_write
hf_...LPvI
2 minutes ago
-
WRITE
text_summarization_read
hf_...THeL
6 minutes ago
-
READ

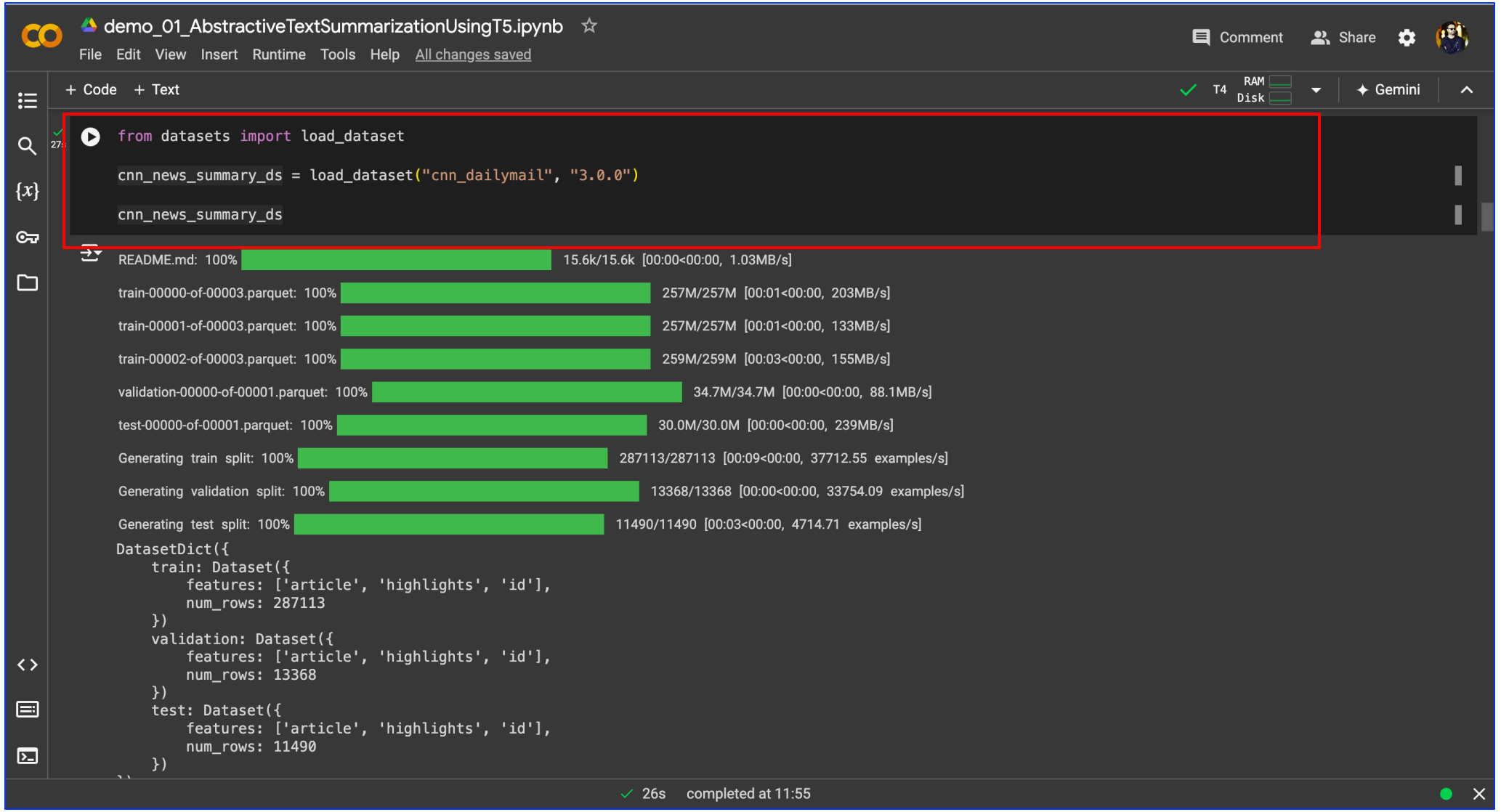
Notebook Login Passing Access Token having Read Permission



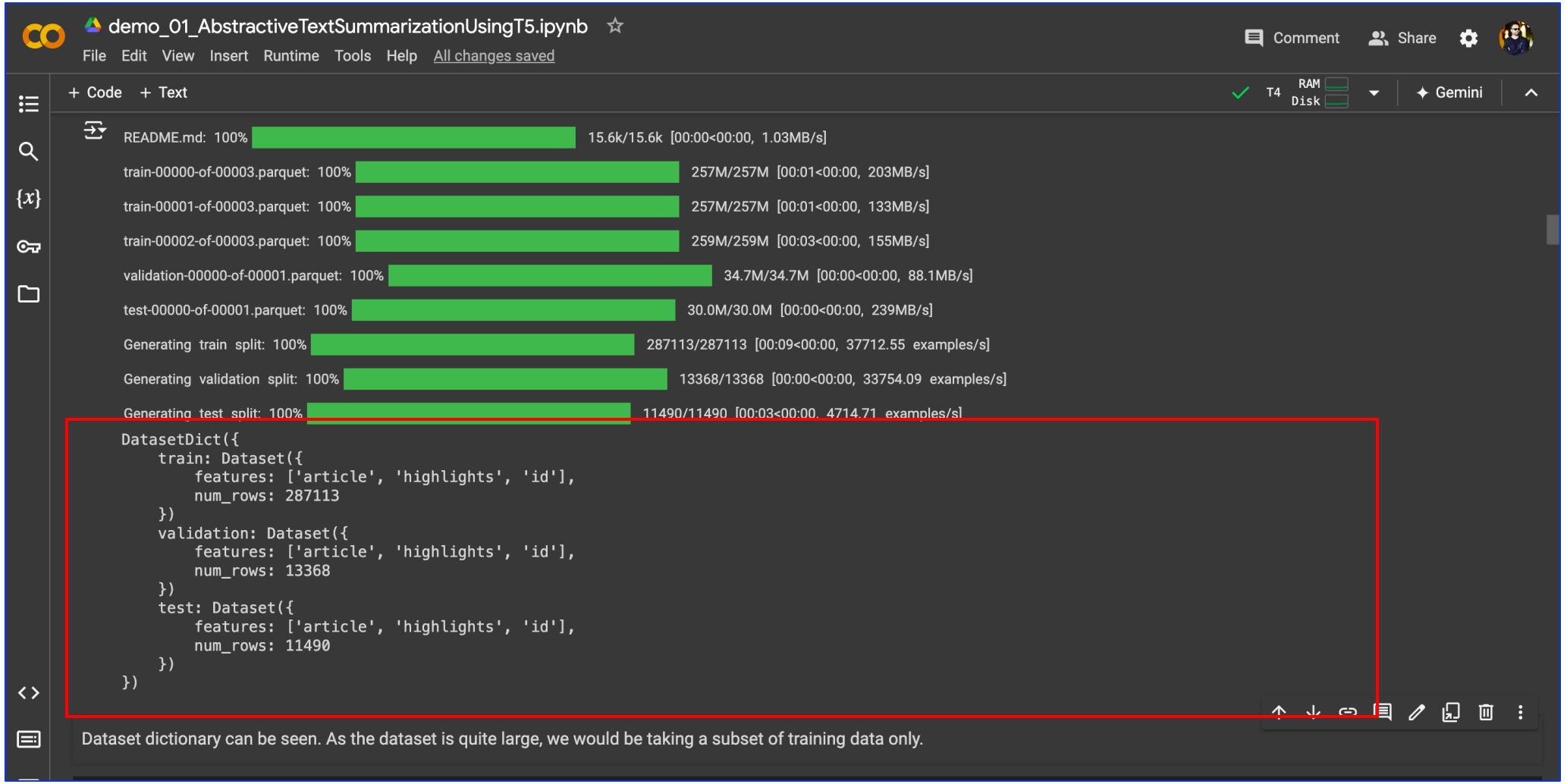
Hugging Face Dataset List (10 Only)



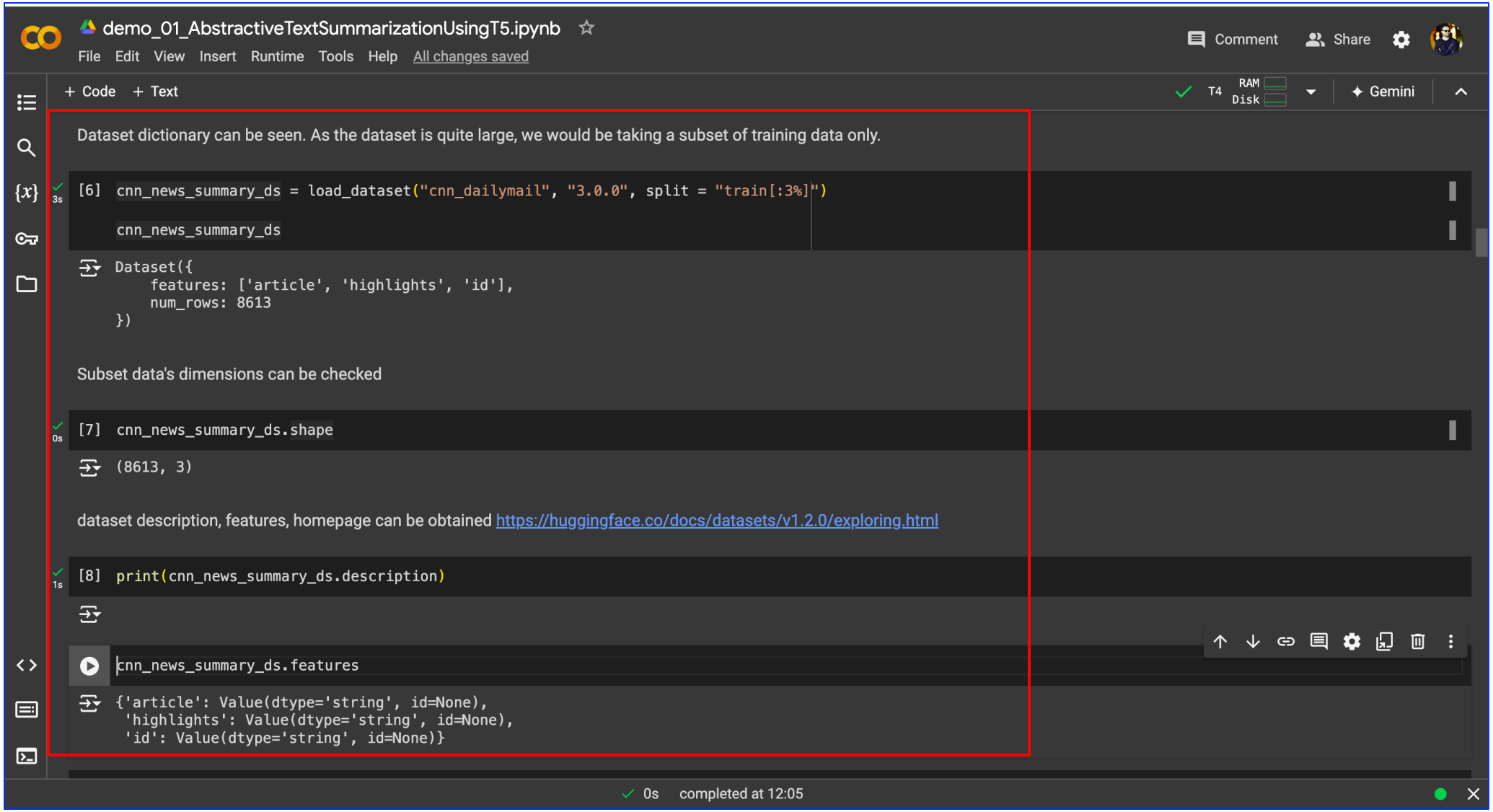
Importing CNN Daily Mail Dataset (Available on HF)



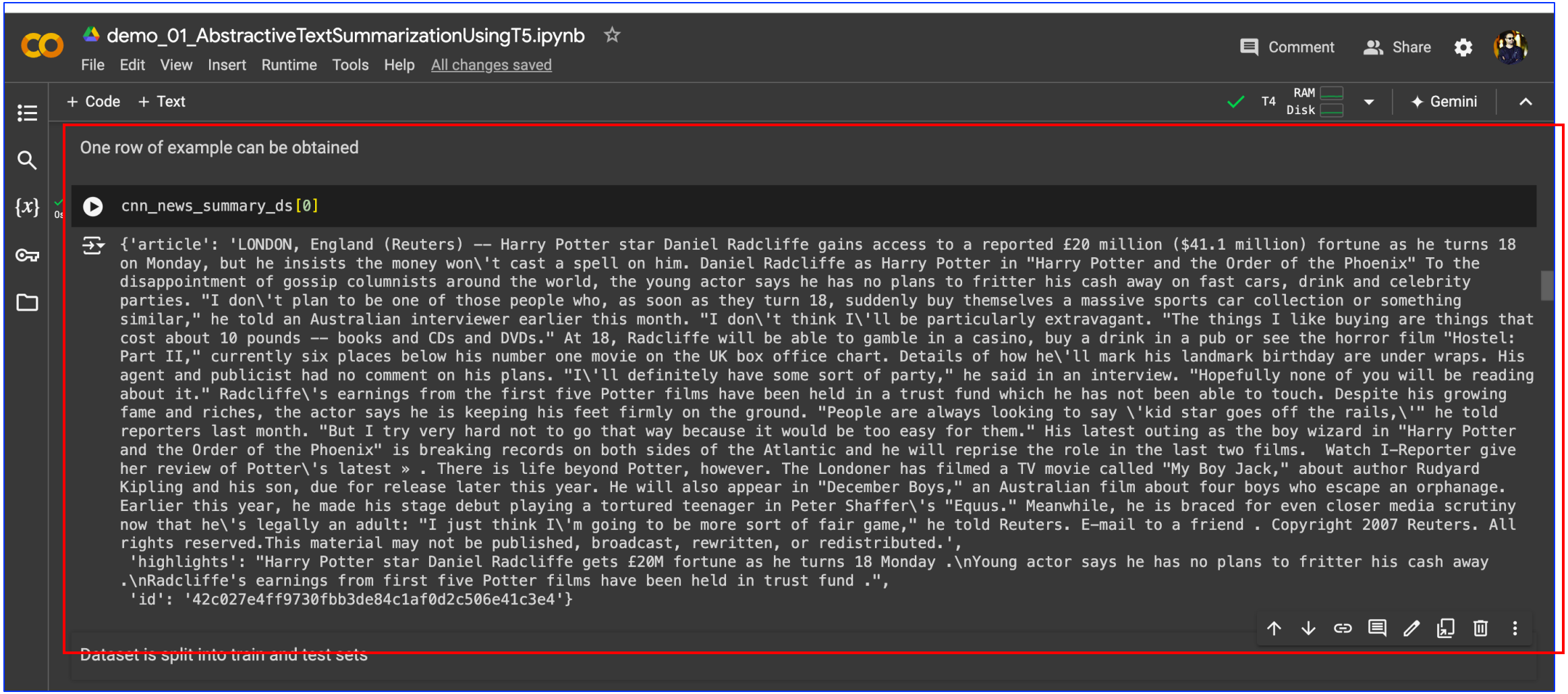
Daily Mail Dataset Information



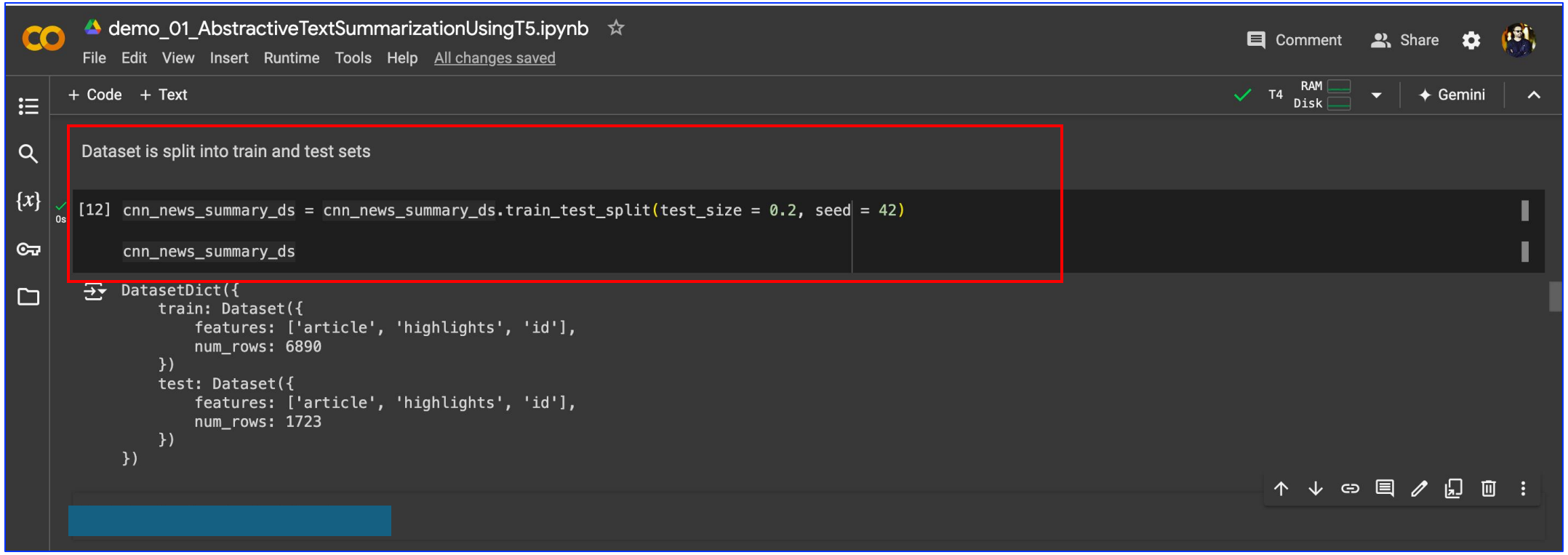
Working with T5 Transformer Model's Subset



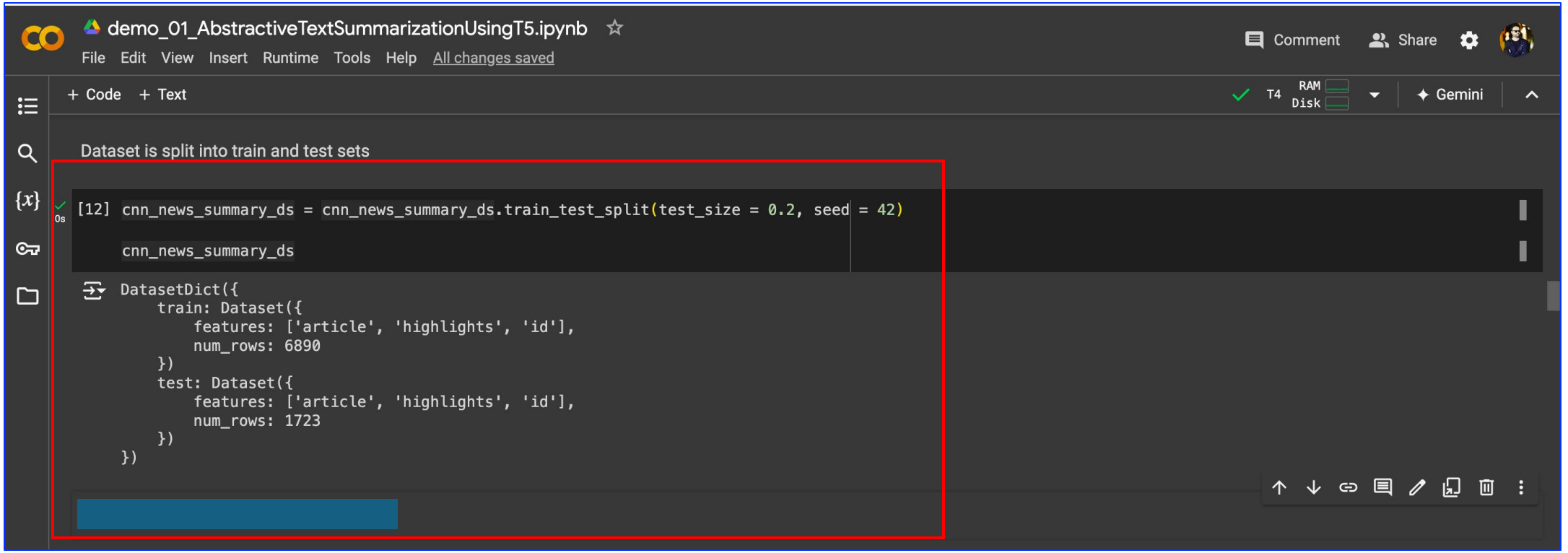
Obtaining First Row of the Dataset



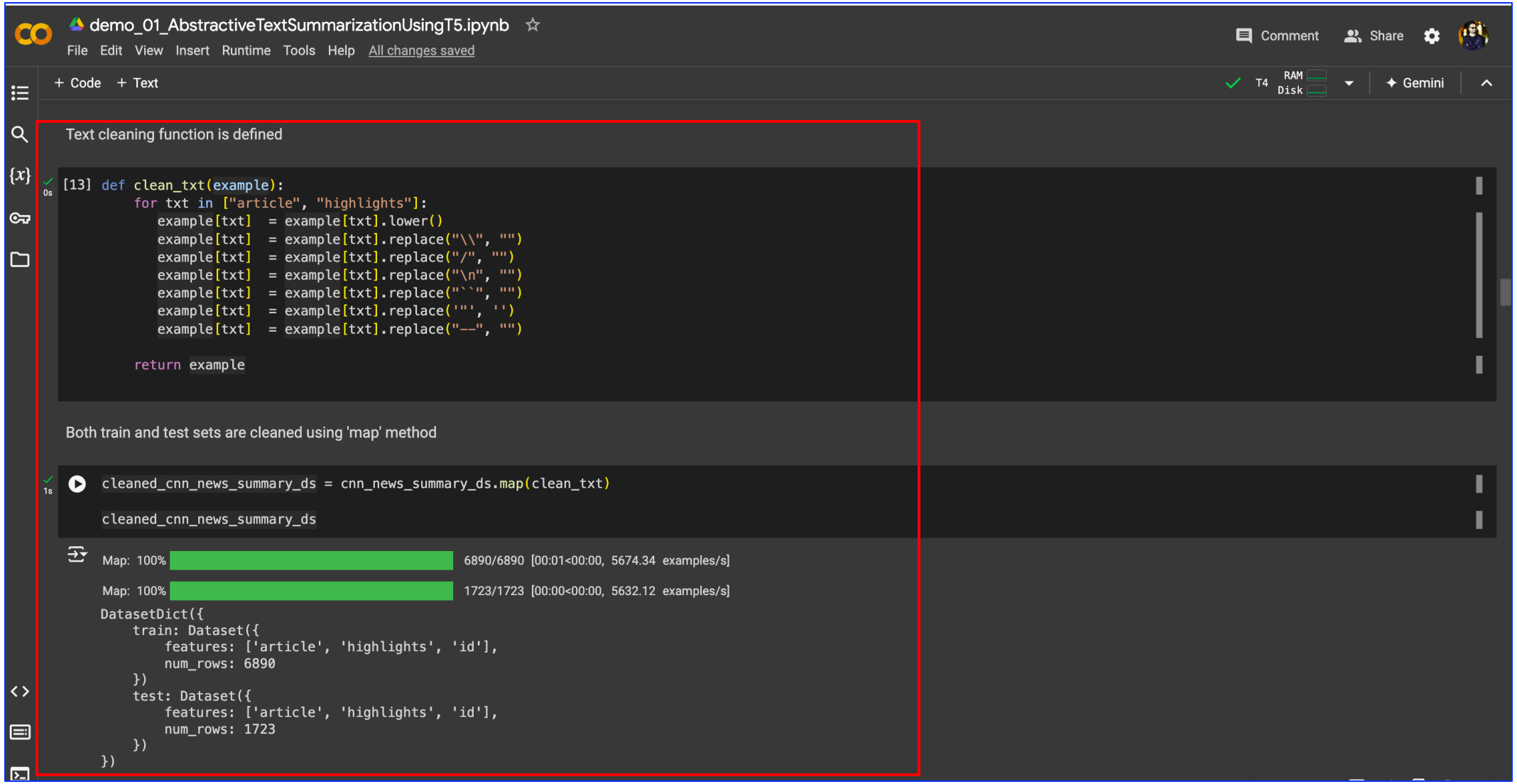
Splitting Dataset for Training and Testing



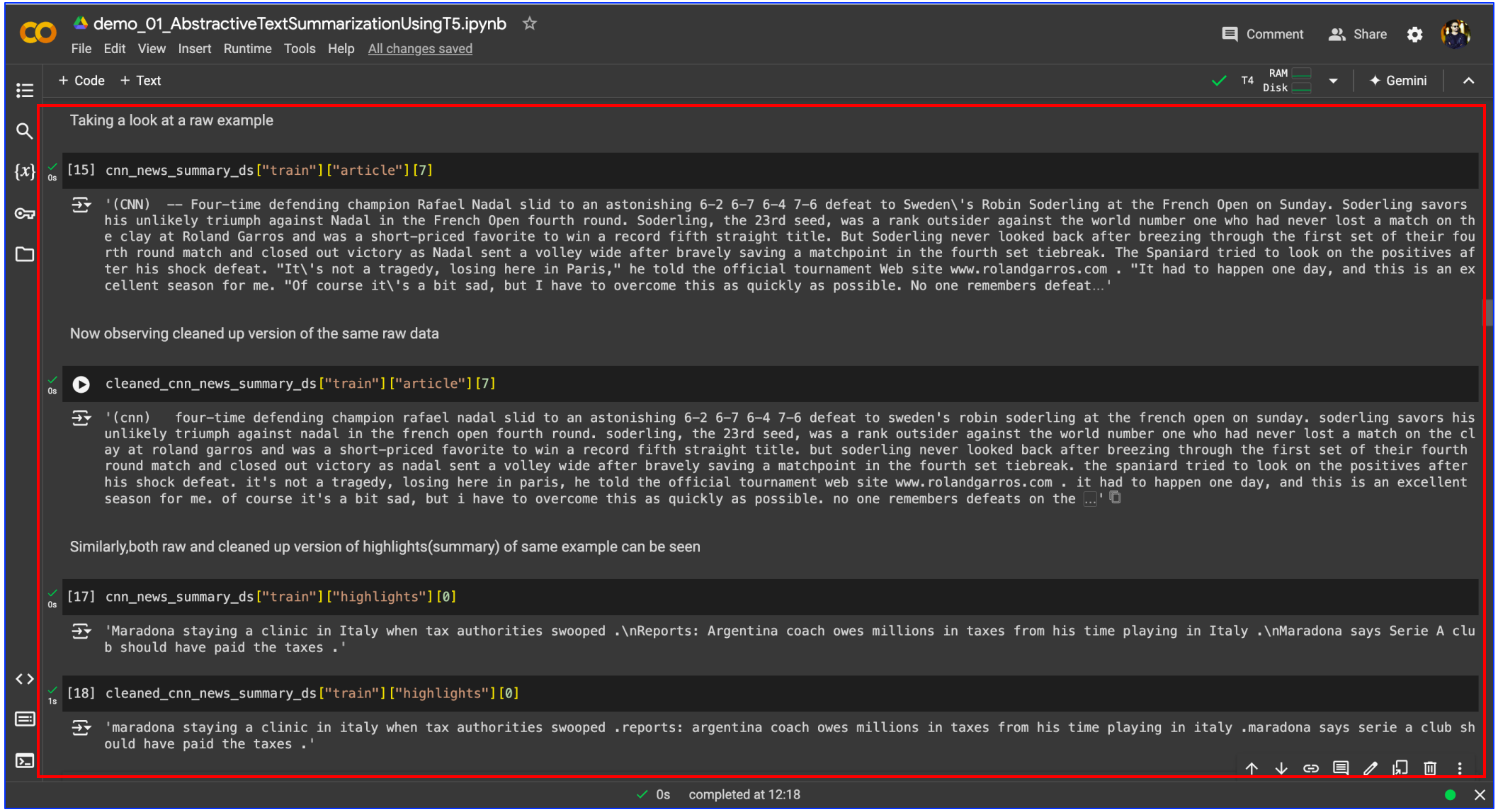
Taring and Testing Dataset Information



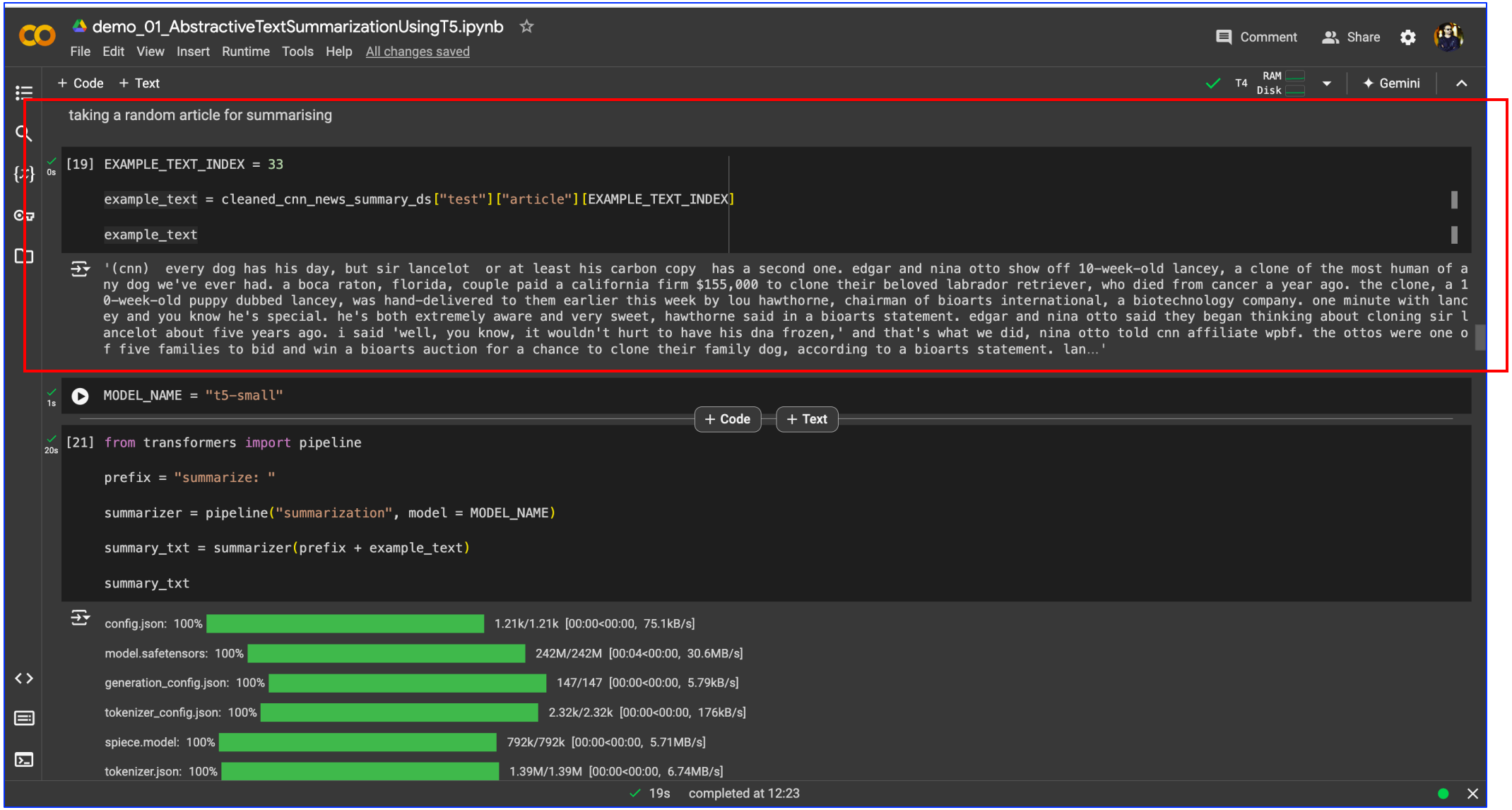
Dataset (subset) Preprocessing



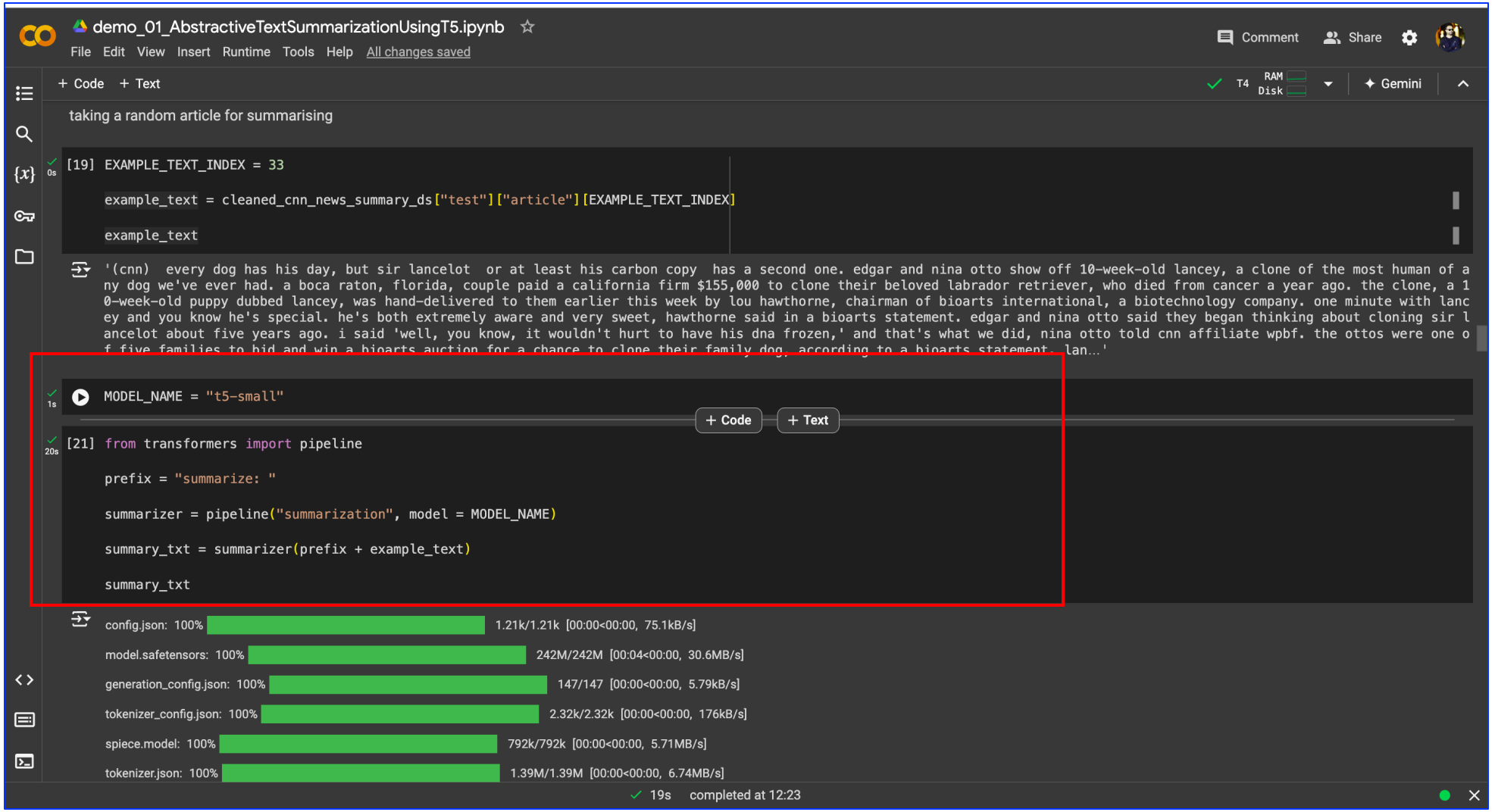
Dataset (subset) Preprocessing – Before and After



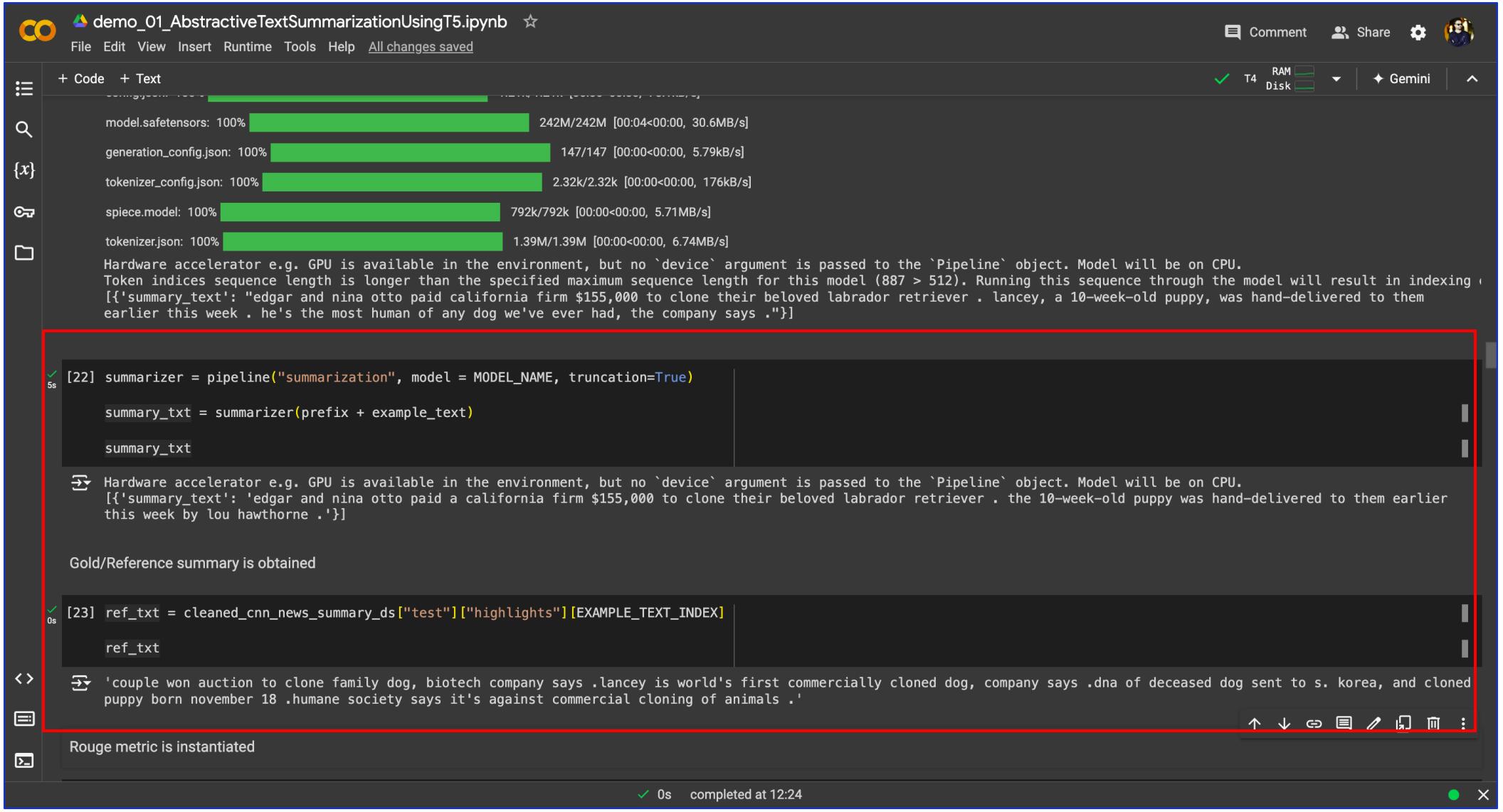
Picking a random article for summary



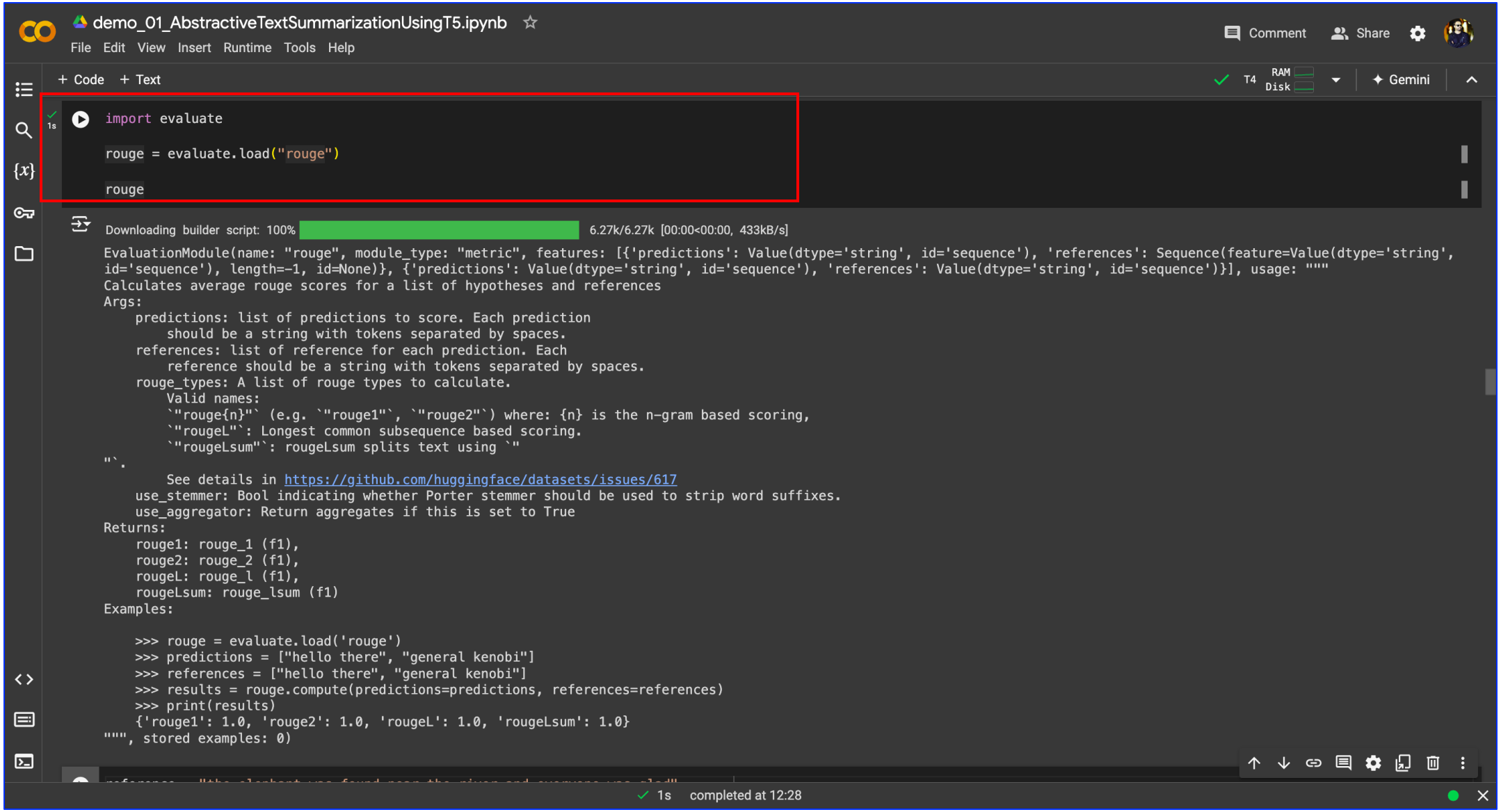
Model Selection



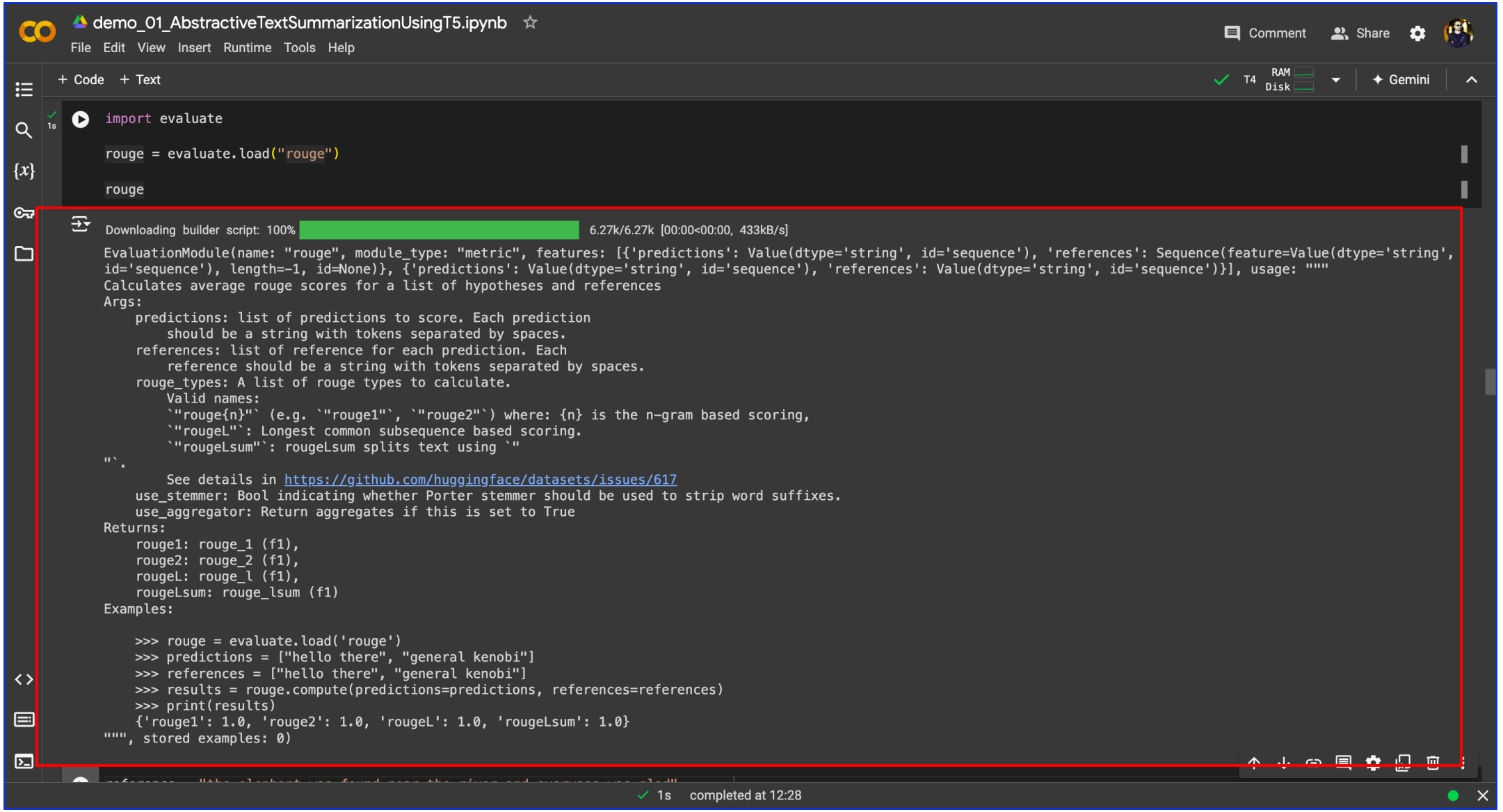
Summarizing the article with the help of the Model



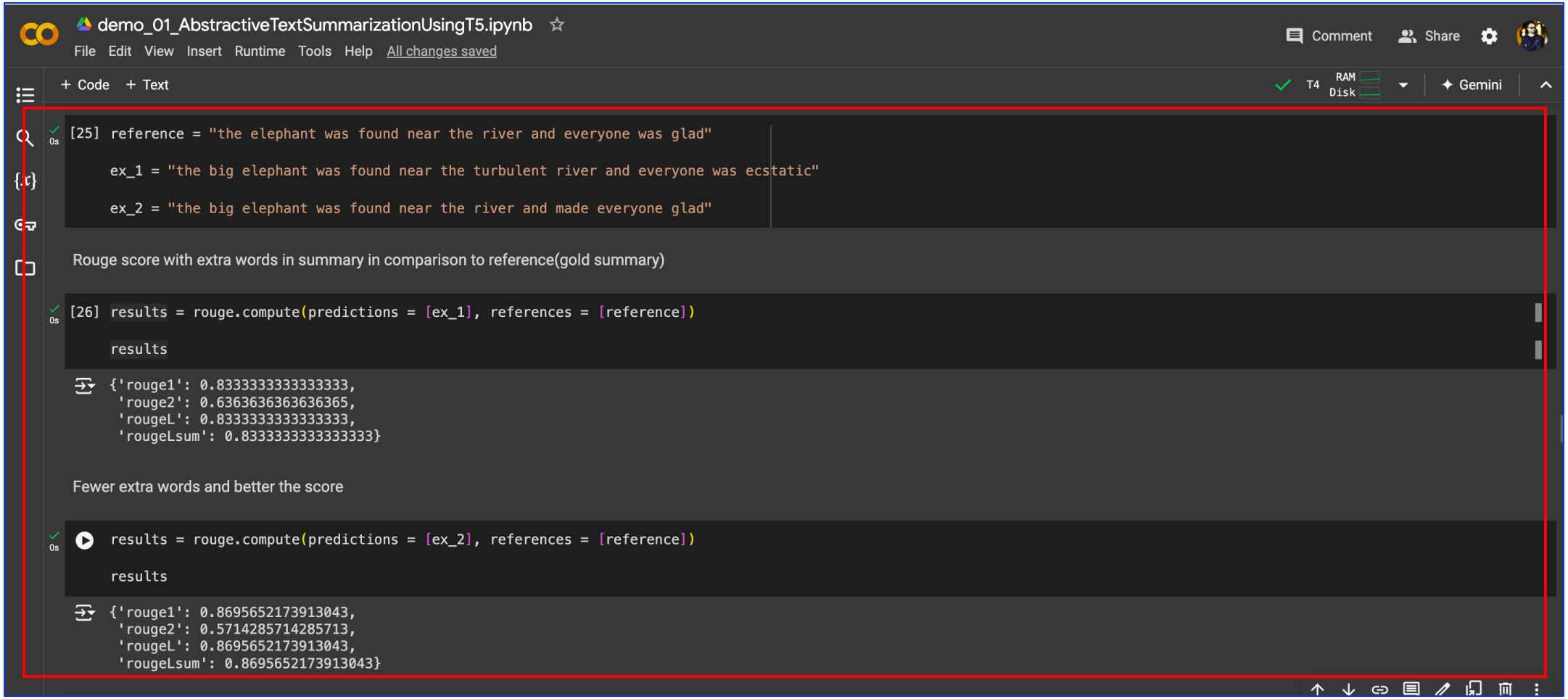
Rouge Score


demo_01_AbstractiveTextSummarizationUsingT5.ipynb
File Edit View Insert Runtime Tools Help
Comment
Share
+ Code + Text
T4
RAM
Disk
Gemini
import evaluate
rouge = evaluate.load("rouge")
rouge
Downloading builder script: 100%
6.27k/6.27k [00:00<00:00, 433kB/s]
EvaluationModule(name: "rouge", module_type: "metric", features: [{'predictions': Value(dtype='string', id='sequence'), 'references': Sequence(feature=Value(dtype='string', id='sequence'), length=-1, id=None)}, {'predictions': Value(dtype='string', id='sequence'), 'references': Value(dtype='string', id='sequence')}], usage: """
Calculates average rouge scores for a list of hypotheses and references
Args:
predictions: list of predictions to score. Each prediction
should be a string with tokens separated by spaces.
references: list of reference for each prediction. Each
reference should be a string with tokens separated by spaces.
rouge_types: A list of rouge types to calculate.
Valid names:
`"rouge{n}"` (e.g. `"rouge1"`, `"rouge2"`) where: {n} is the n-gram based scoring,
`"rougeL"`: Longest common subsequence based scoring.
`"rougeLsum"`: rougeLsum splits text using `"
"`.
See details in https://github.com/huggingface/datasets/issues/617
use_stemmer: Bool indicating whether Porter stemmer should be used to strip word suffixes.
use_aggregator: Return aggregates if this is set to True
Returns:
rouge1: rouge_1 (f1),
rouge2: rouge_2 (f1),
rougeL: rouge_l (f1),
rougeLsum: rouge_lsum (f1)
Examples:
>>> rouge = evaluate.load('rouge')
>>> predictions = ["hello there", "general kenobi"]
>>> references = ["hello there", "general kenobi"]
>>> results = rouge.compute(predictions=predictions, references=references)
>>> print(results)
{'rouge1': 1.0, 'rouge2': 1.0, 'rougeL': 1.0, 'rougeLsum': 1.0}
""", stored examples: 0)
1s completed at 12:28

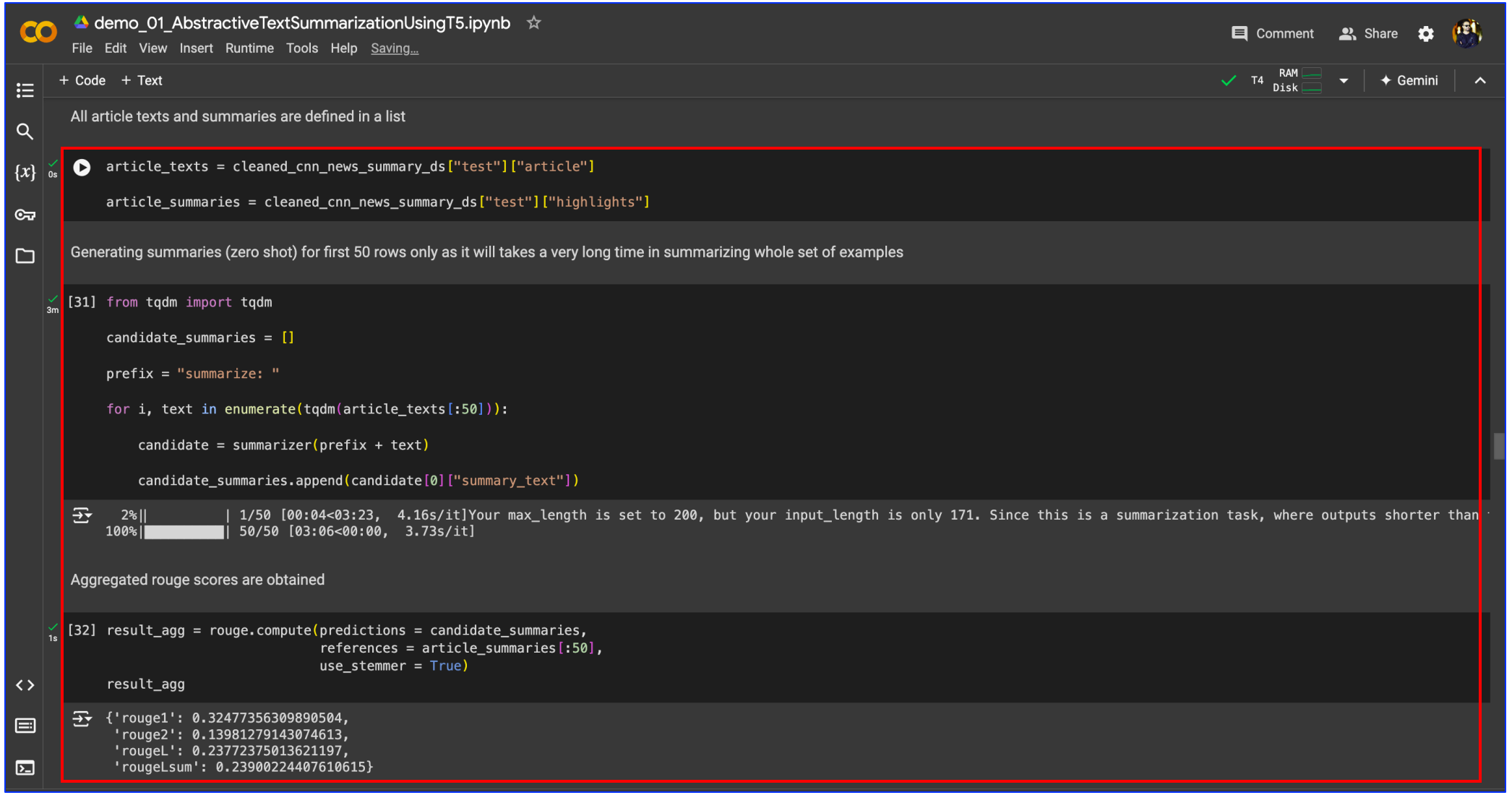
Variants of ROUGE



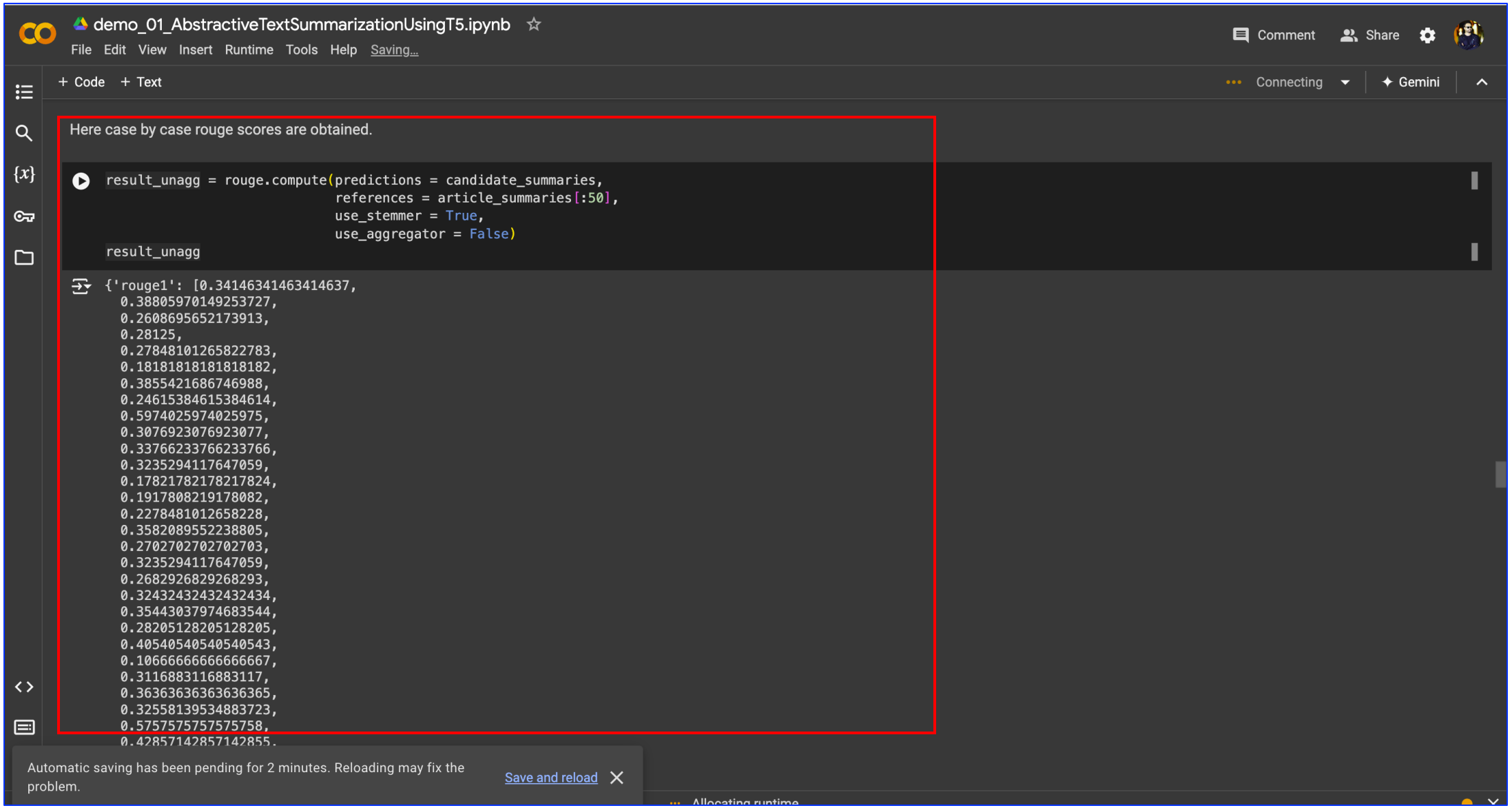
ROUGE score calculation with Reference Sentence and example sentence (summarized)



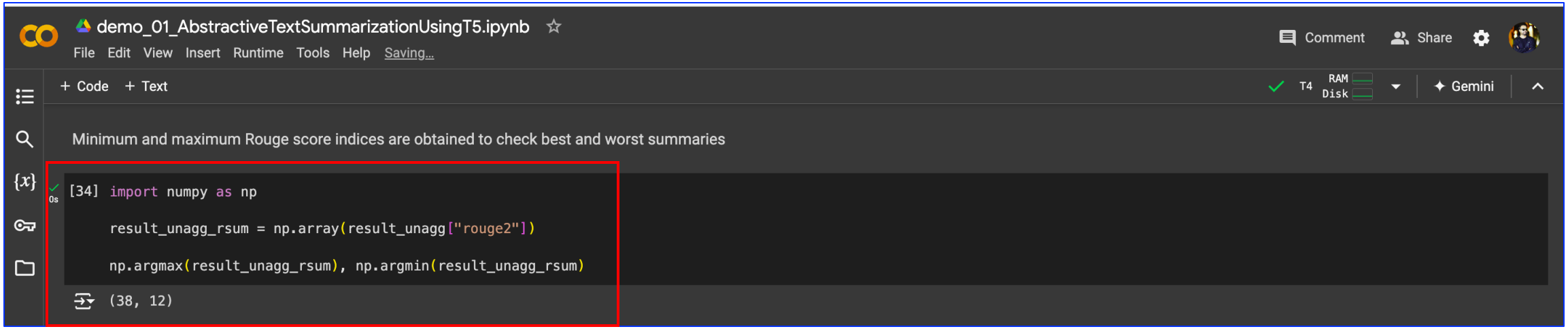
Generating first 50 rows summaries



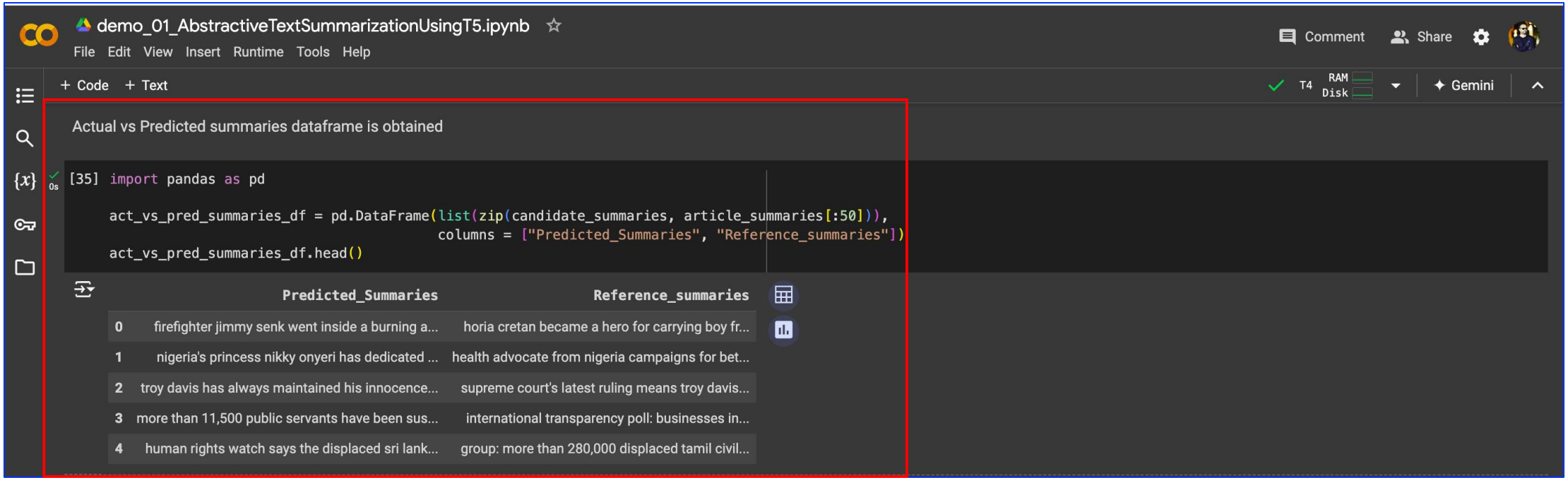
Rouge Scores of first 50 summaries



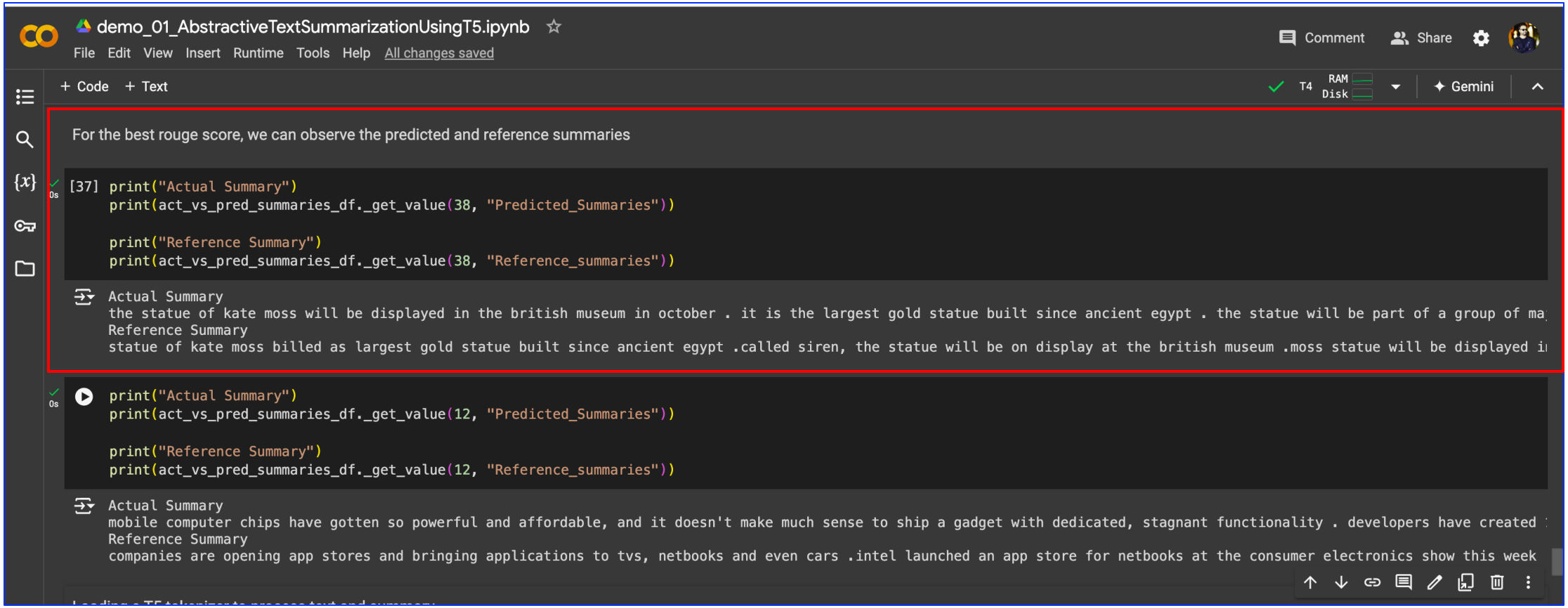
Minimum and maximum Rouge score indices are obtained to check the best and worst summaries



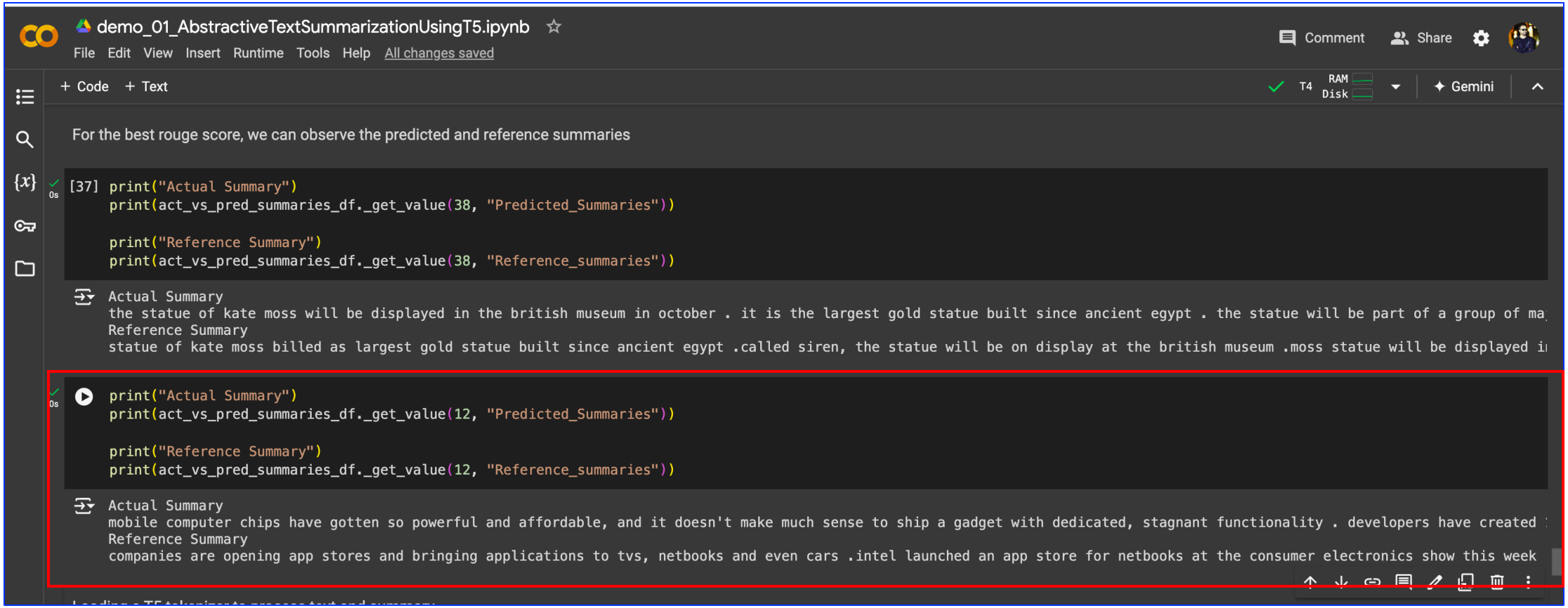
Actual vs Predicted summaries dataframe is obtained



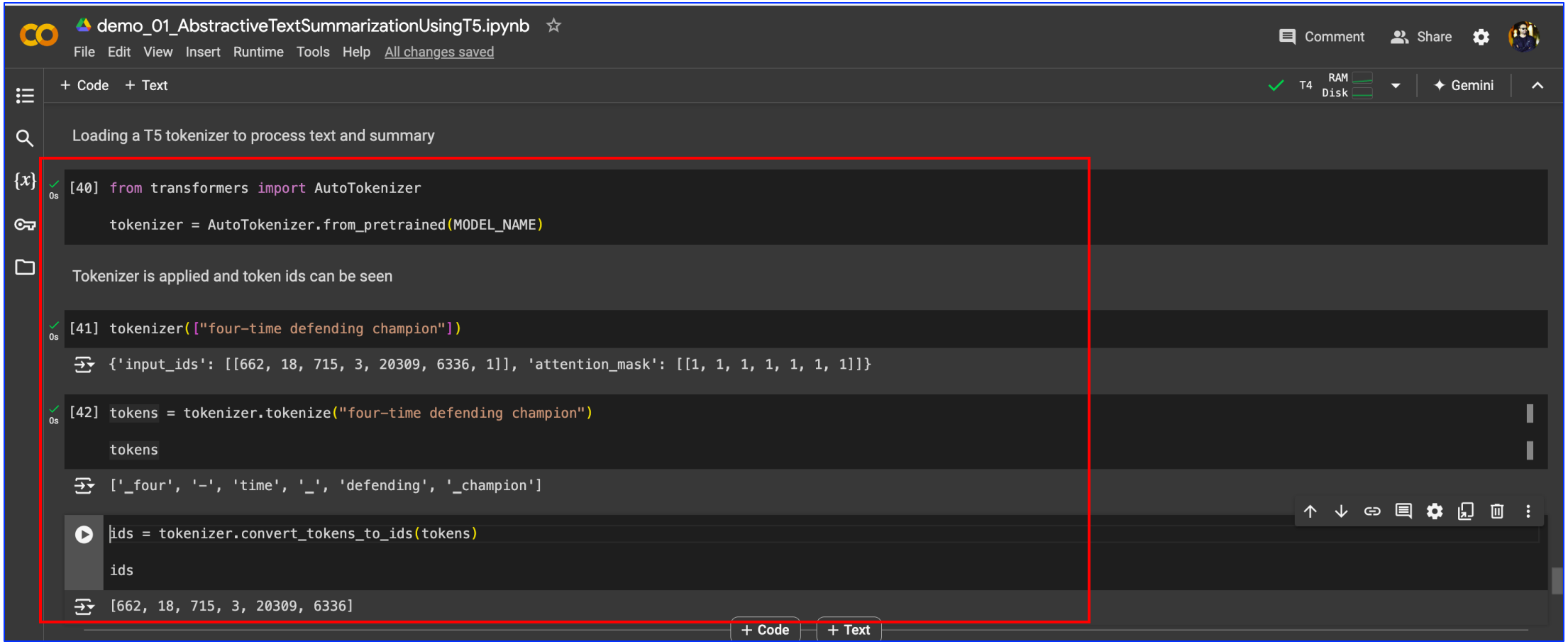
For the best rouge score, we can observe the predicted and reference summaries



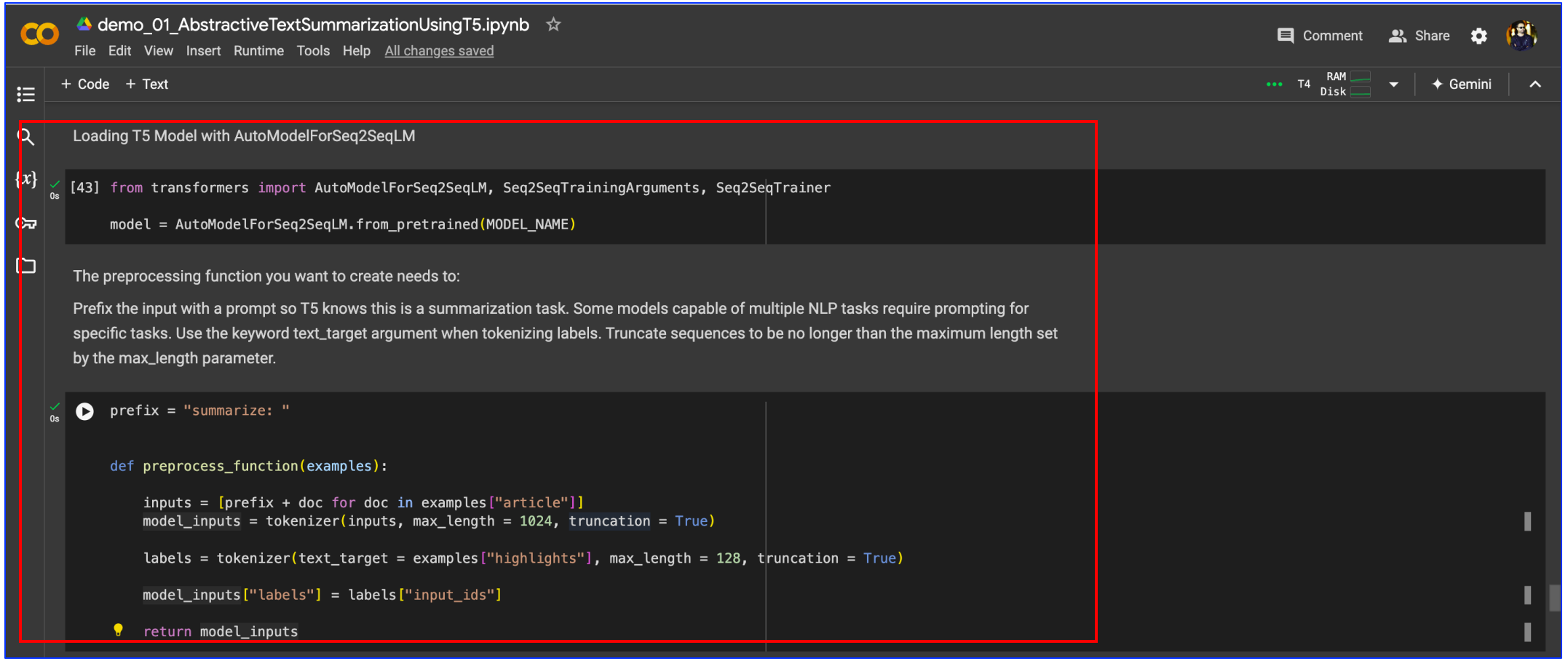
For the best rouge score, we can observe the predicted and reference summaries



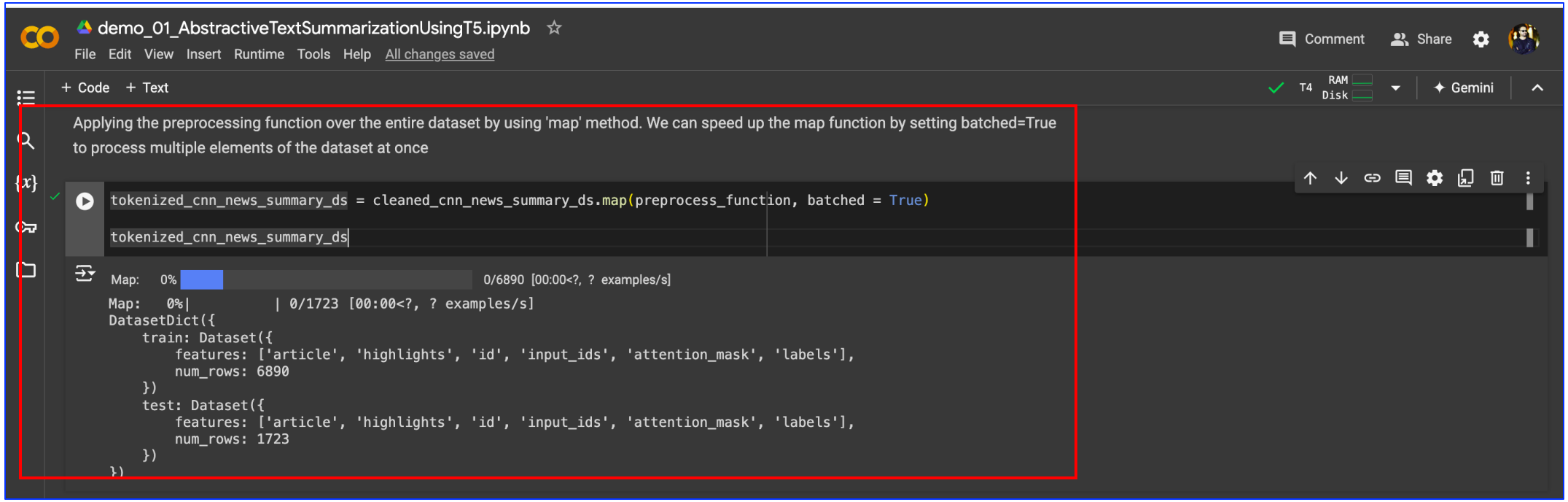
Loading a T5 tokenizer to process text and summary



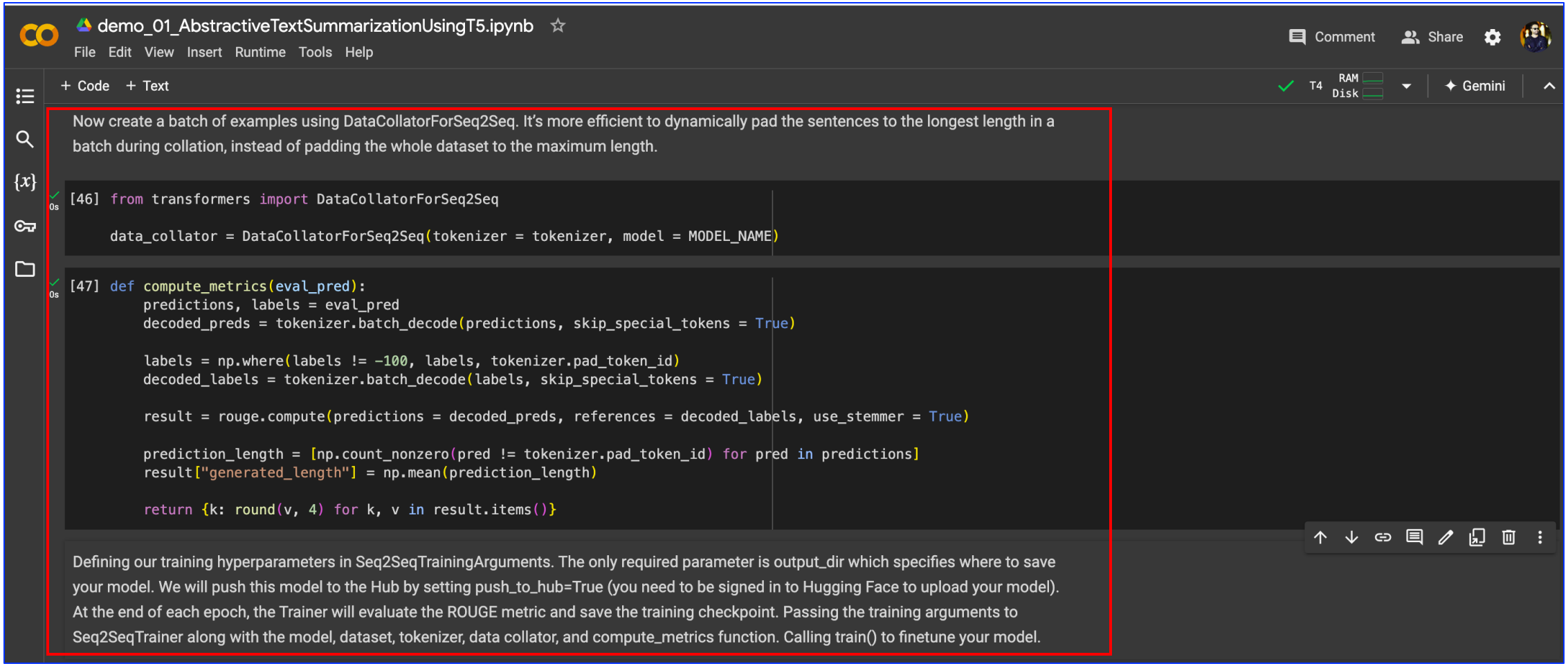
Loading T5 Model with AutoModelForSeq2SeqLM



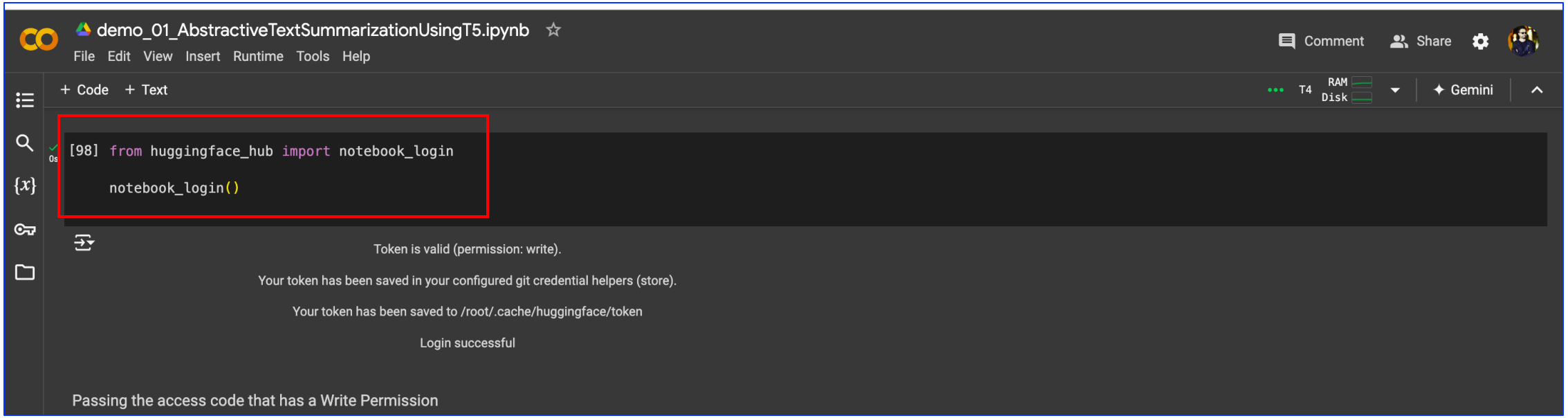
Applying the preprocessing function over the entire dataset by using 'map' method.



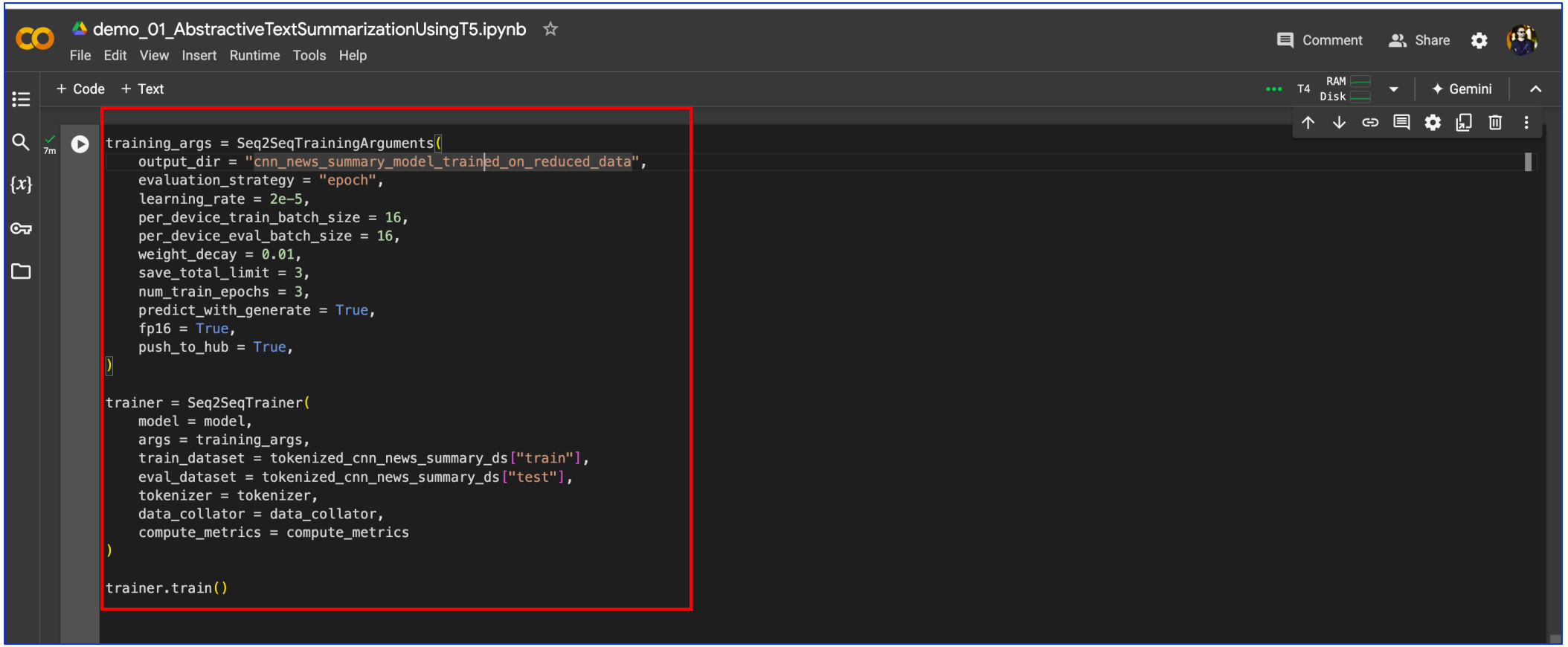
Now create a batch of examples using DataCollatorForSeq2Seq.



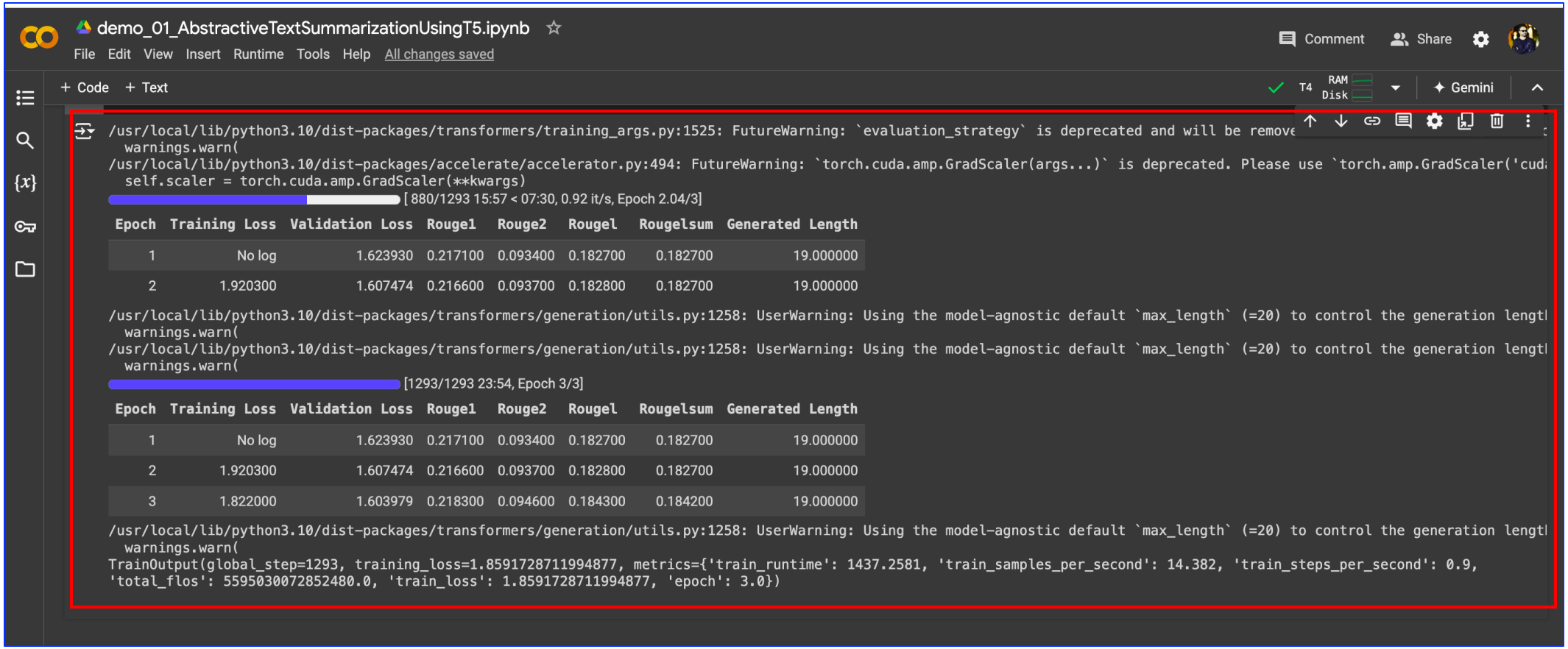
Passing the access code that has a Write Permission



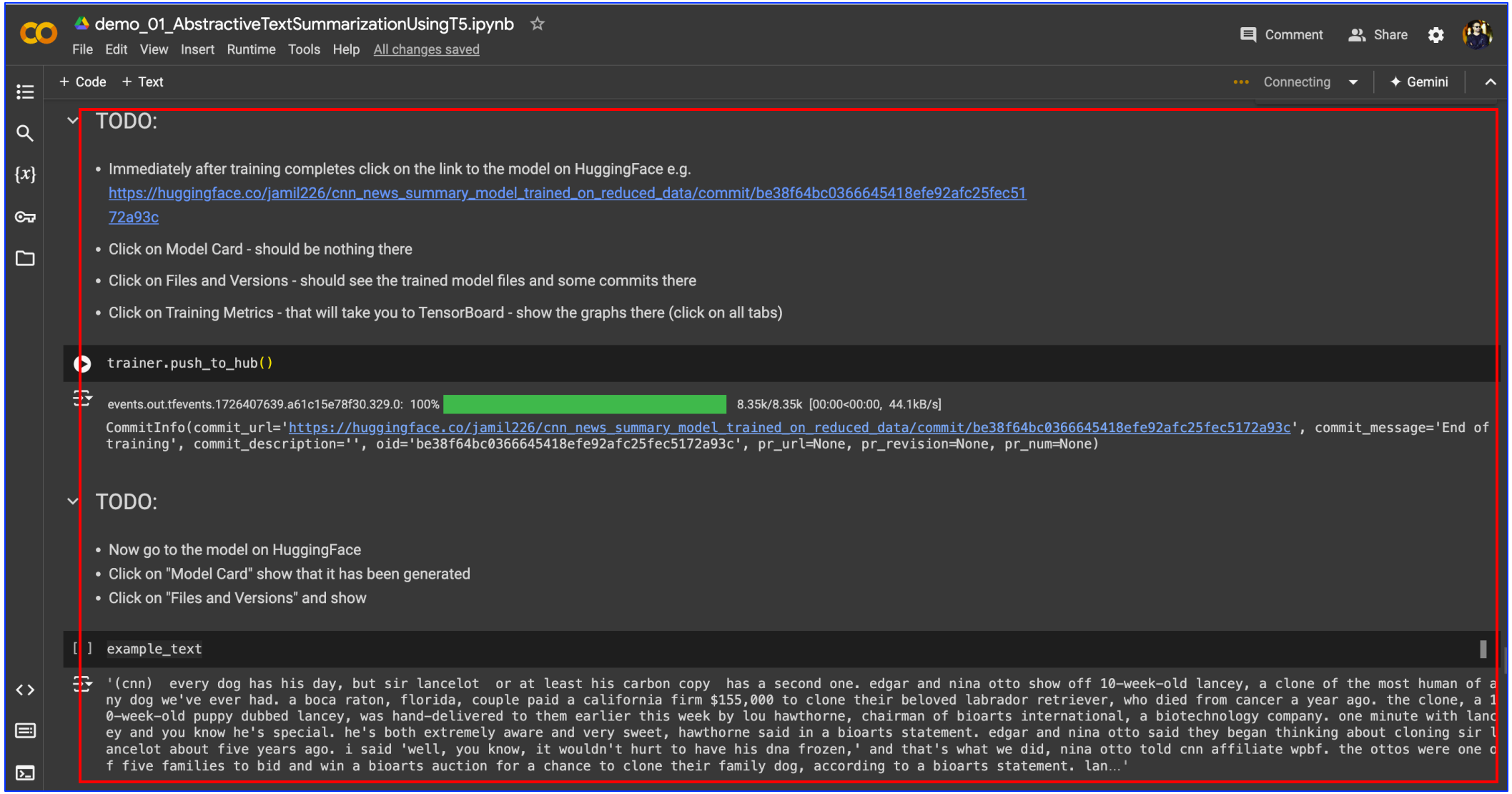
Fine-Tuning a Seq2Seq Model for Text Summarization



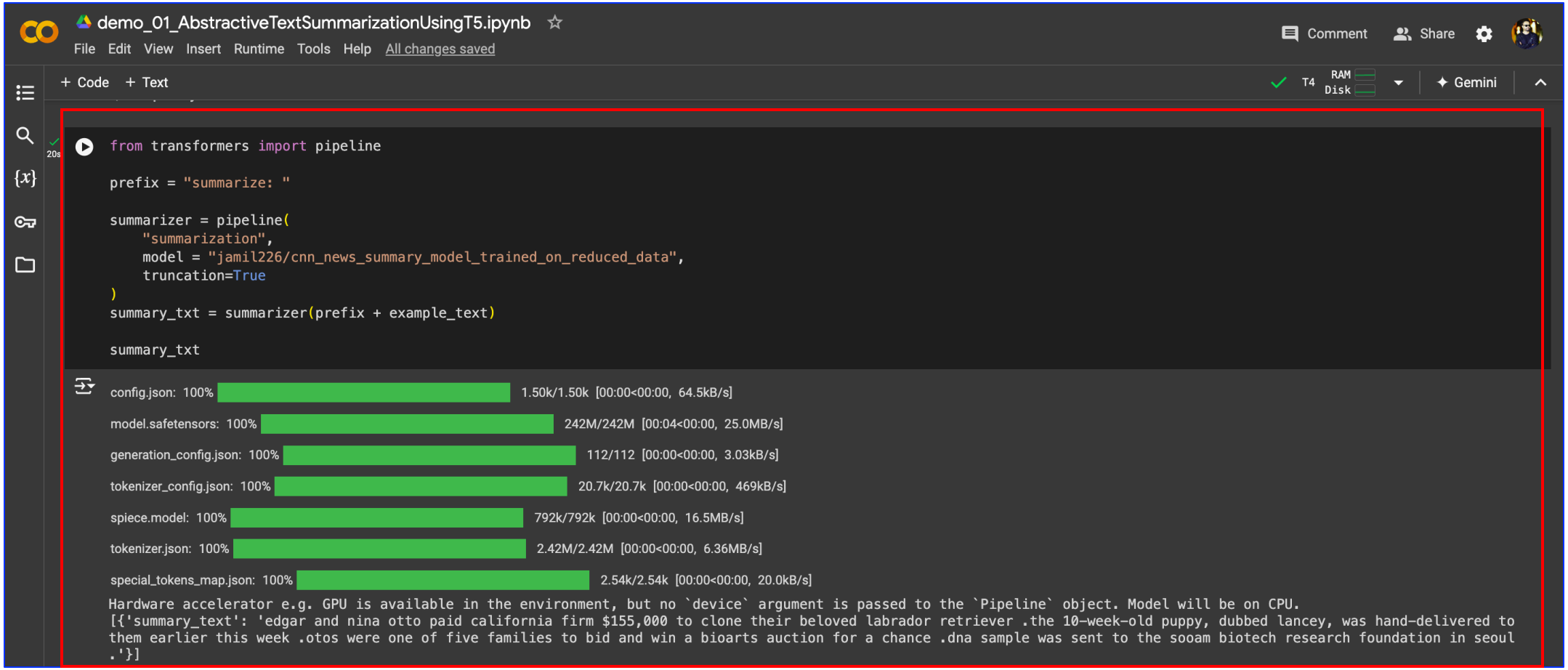
Fine-Tuning a Seq2Seq Model for Text Summarization



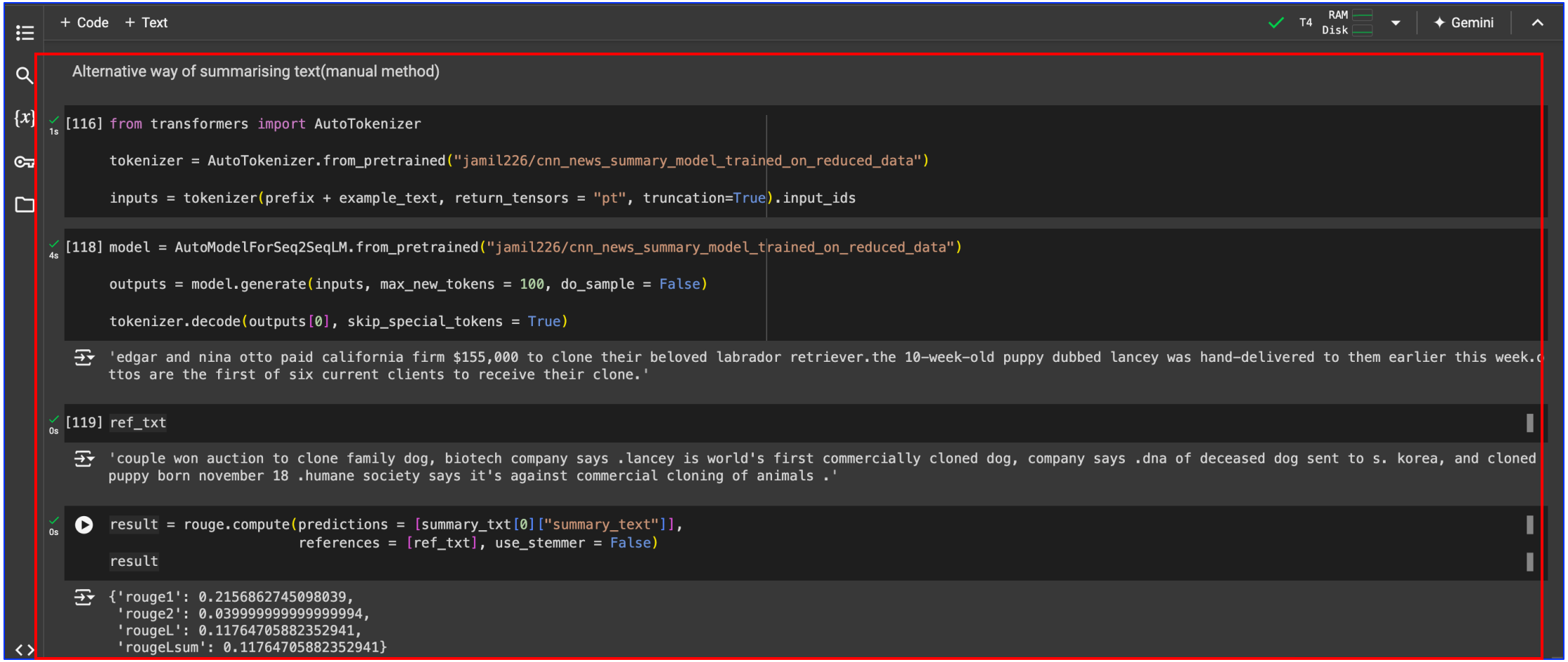
Pushing the fine-tuned model on Hugging Face



Summary of example text using our Fine-Tuned Model



Alternative way of summarising text(manual method)



Summarizing text results (without fine-tuning)

```
[77] summary_txt[0]["summary_text"]
```

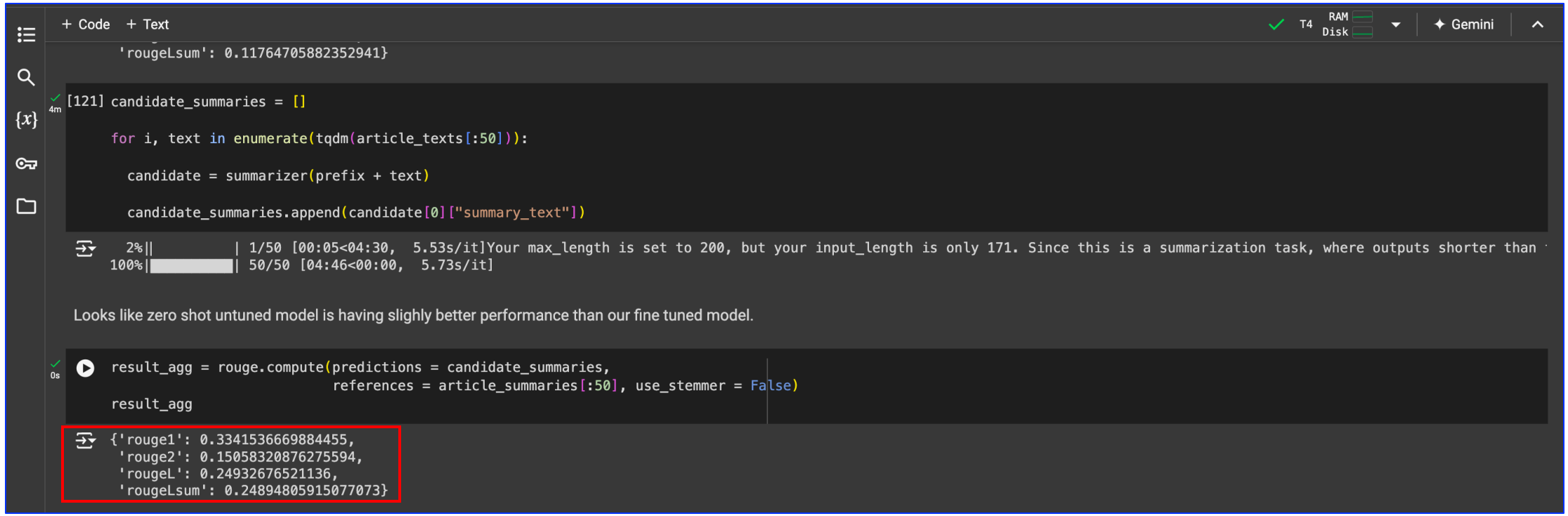
```
'edgar and nina otto paid a california firm $155,000 to clone their beloved labrador retriever . the 10-week-old puppy was hand-delivered to them earlier this week by lou hawthorne .'
```

Rouge score for that text summary is obtained

```
result = rouge.compute(predictions = [summary_txt[0]["summary_text"]],
                       references = [ref_txt], use_stemmer = True)
result
```

```
{'rouge1': 0.13157894736842105,
 'rouge2': 0.027027027027027025,
 'rougeL': 0.07894736842105263,
 'rougeLsum': 0.07894736842105263}
```

Rouge score for 50 summaries



Rouge score of 50 summaries (without fine-tuning)

```
[80] from tqdm import tqdm

     candidate_summaries = []

     prefix = "summarize: "

     for i, text in enumerate(tqdm(article_texts[:50])):

         candidate = summarizer(prefix + text)

         candidate_summaries.append(candidate[0]["summary_text"])
```

```
  2%||          | 1/50 [00:05<04:51,  5.96s/it]Your max_length is set to 200, but your input_length is only 171. Since this is a summarization task, where outputs shorter than
100%|██████████| 50/50 [03:51<00:00,  4.63s/it]
```

Aggregated rouge scores are obtained

```
result_agg = rouge.compute(predictions = candidate_summaries,
                           references = article_summaries[:50],
                           use_stemmer = True)
result_agg
```

```
{'rouge1': 0.32614983426729627,
 'rouge2': 0.1395363723483919,
 'rougeL': 0.2393680577907863,
 'rougeLsum': 0.23889110860526322}
```

Rouge score of 50 summaries (without fine-tuning)

+ Code + Text

T4 RAM Disk Gemini

```
[123] result_unagg = rouge.compute(predictions = candidate_summaries, references = article_summaries[:50],
                                   use_stemmer = True, use_aggregator = False)
```

```
[124] result_unagg_rsum = np.array(result_unagg["rougeLsum"])

      np.argmax(result_unagg_rsum), np.argmin(result_unagg_rsum)
```

```
(44, 23)
```

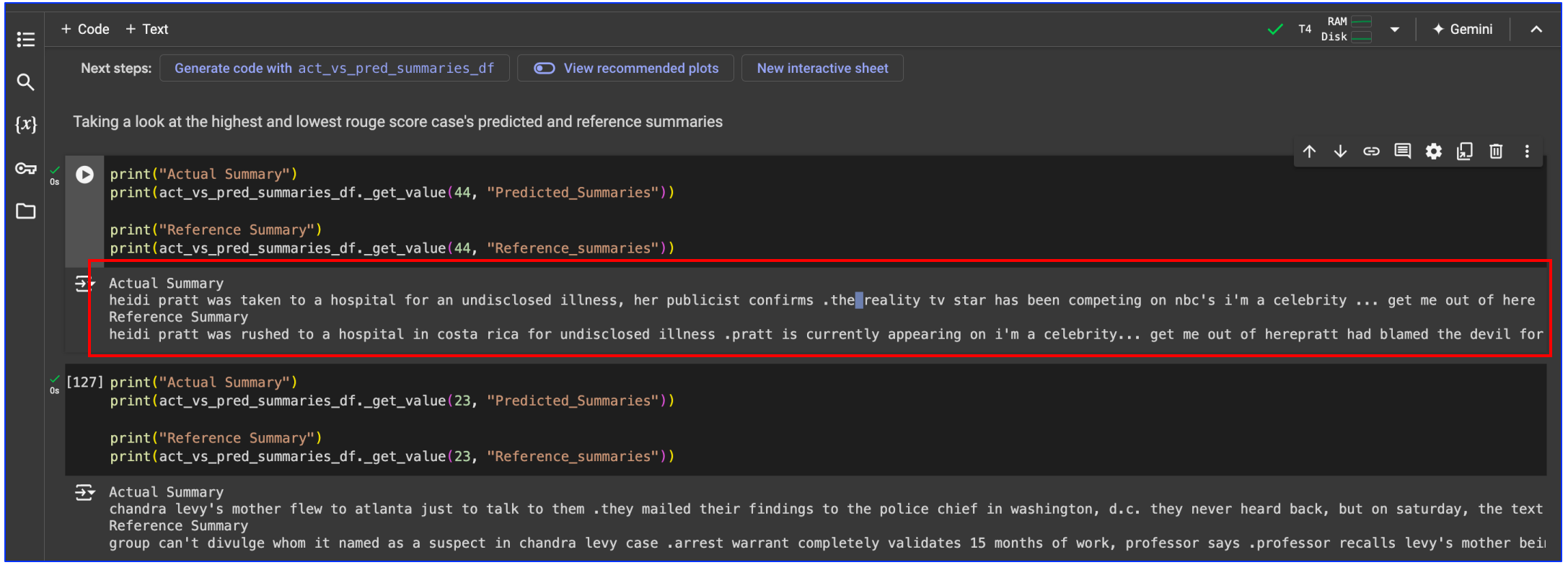
Actual vs Predicted summaries dataframe is obtained

```
act_vs_pred_summaries_df = pd.DataFrame(list(zip(candidate_summaries, article_summaries[:50])),
                                        columns = ["Predicted_Summaries", "Reference_summaries"])
act_vs_pred_summaries_df.head()
```

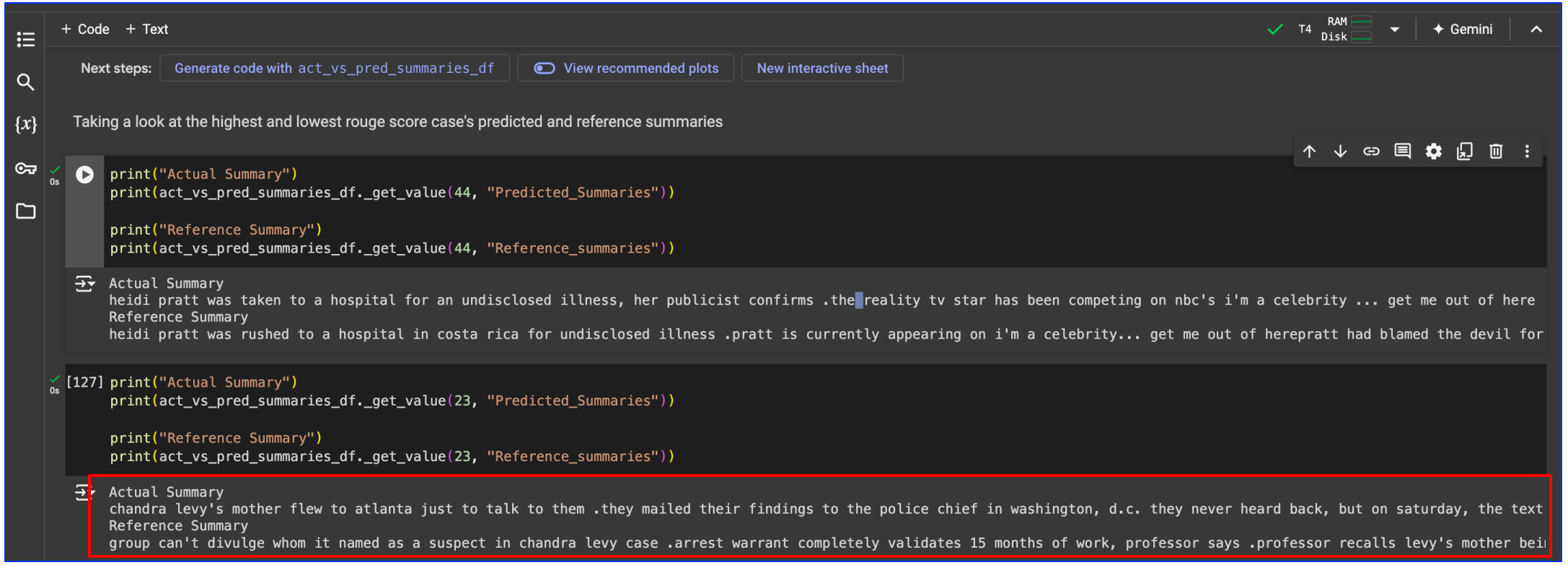
1 to 5 of 5 entries Filter

index	Predicted_Summaries	Reference_summaries
0	bronx firefighter jimmy senk pulls 4-year-old boy to safety .bystander climbed four floors up a fire escape to try to help the boy .the boy has since been released from the hospital .a video has been aired on local, national, and international tv .	horia cretan became a hero for carrying boy from burning building .the other hero is firefighter jimmy senk, who went into the building .senk saw boy disappear back into the smoke, so i masked up and went ini didn't think he was going to make it, senk said .
1	nigeria's princess nikky onyeri speaks out to raise awareness about breast cancer .the impetus behind her drive and persistence is a wrong diagnosis of cancer 15 years ago .	health advocate from nigeria campaigns for better awareness of breast cancer .lobbies nigerian government to do more for women with cancer .was misdiagnosed with cancer 15 years ago; issue still a taboo in africa .
2	troy davis has always maintained his innocence in the 1989 killing of officer mark macphail .the georgia inmate has gained international support for his long-standing claim that he did not murder a savannah police officer nearly two decades ago .he was granted a stay of execution by the supreme court two hours before he was to be put to death last fall .	supreme court's latest ruling means troy davis will continue to sit on death row .davis was convicted in 1991 of murdering a savannah, georgia, police officer .since his conviction, 7 of 9 witnesses against him have recanted their testimony .justices antonin scalia and clarence thomas objected to the court's decision .
3	more than 11,500 public servants suspended or fined for corruption in past two years, mexican government says .ex-anti-drug czar noe ramirez accused of receiving $500,000 per month for passing information to pacific cartel .the problem extends beyond bureaucracy, according to a poll by organization international transparency .	international transparency poll: businesses in mexico more open to paying bribes .organized crime has penetrated state institutions charged with fighting crime .some 40 government agents under investigation for presumed ties to drug cartels .
4	sri lankan government should release more than 280,000 displaced tamil civilians .human rights watch says the displaced were already victims of a protracted and bloody civil war .as many as 70,000 people were killed in the conflict .	group: more than 280,000 displaced tamil civilians living in detention camps .sri lanka says the human rights watch report is overstated .human rights watch says aid workers are prohibited from discussing abuses .group reported health problems created by inconsistent water supply .

Best and Worst Results Comparison



Best and Worst Results Comparison



References

- https://github.com/Jamil226/FA24-AI
- https://huggingface.co/google-t5/t5-small
- https://huggingface.co/jamil226/cnn_news_summary_model_trained_on_reduced_data

Thank You!

jamil@cuisahiwal.edu.pk