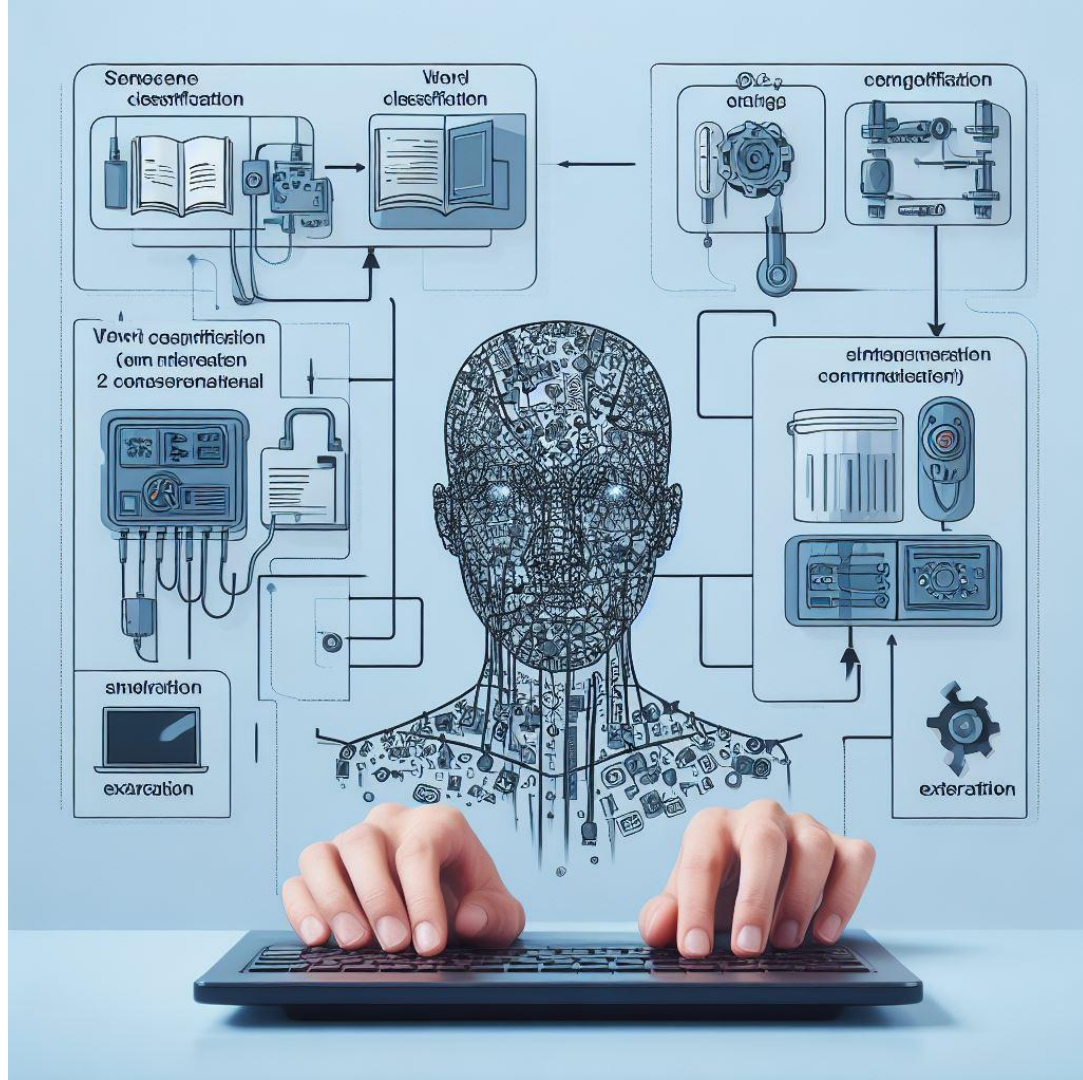


NLP專案的 底層邏輯

臺灣大學
應用深度學習 - 實作三
林彥廷

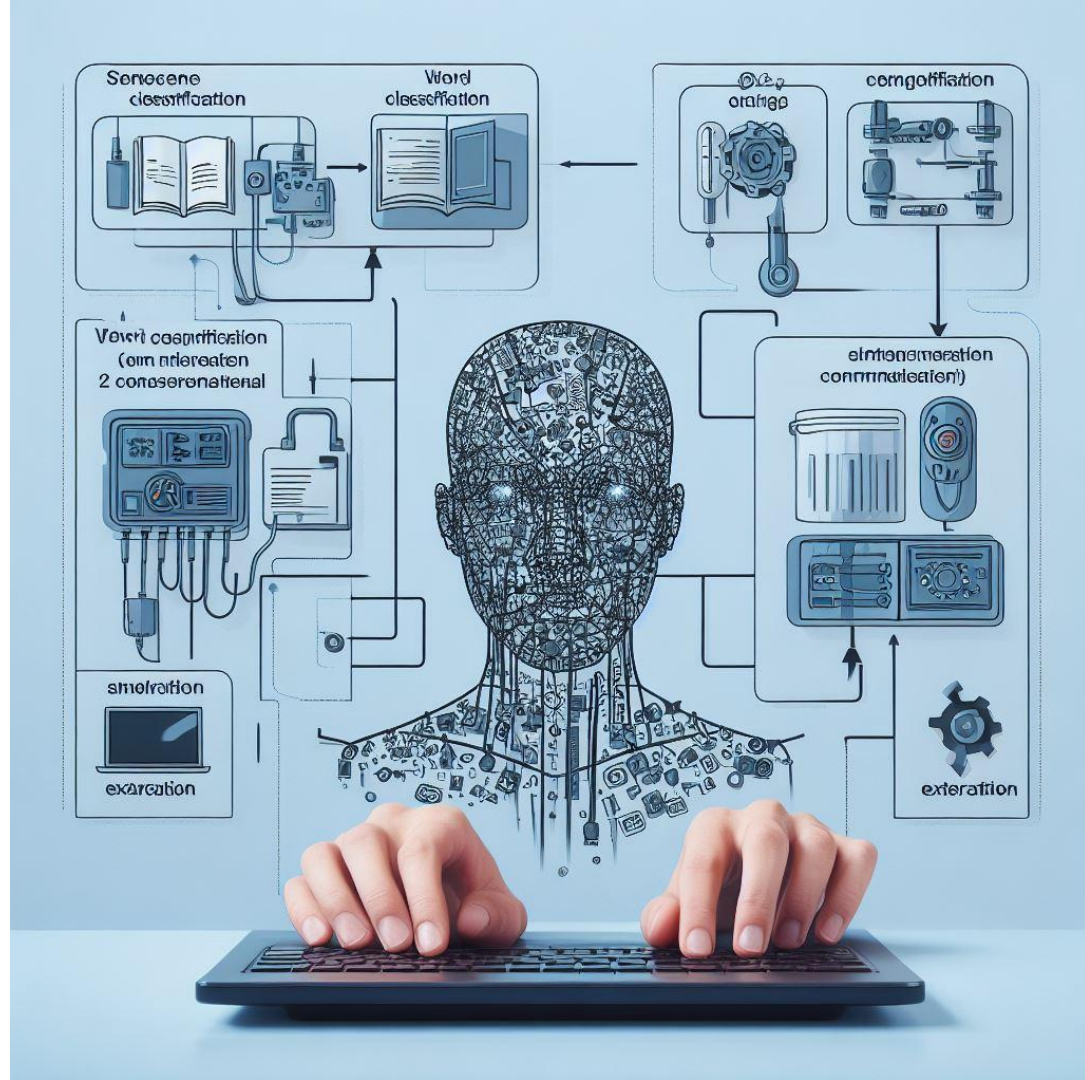


The Underlying Logic of NLP Project

NTU

ADL - Recitation 3

[Yen-Ting Lin](#)



回顧: 專案的一生




模型

回顧: 專案的一生

- 分類**整句**
- 分類句中每個**單詞**
- **生成**文本內容
- 從文本中**提取**答案

你會學到什麼？

- 如何準備四大任務的資料、訓練、預測
- 使用 Huggingface 生態系 

四大任務

- 分類**整句**
- 分類句中每個**單詞**
- **生成**文本內容
- 從文本中**提取**答案

四大任務

- 分類整句
- 分類句中每個單詞
- 生成文本內容
- 從文本中提取答案

專案的一生

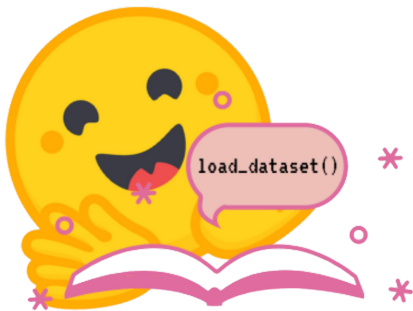
資料



模型

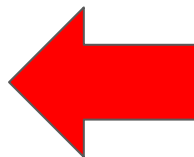


Demo



四大任務

- 分類整句
- 分類句中每個單詞
- 生成文本內容
- 從文本中提取答案



分類單詞

資料

Ousted **WeWork** founder **Adam Neumann** lists his **Manhattan** penthouse for **\$37.5 million**

[organization] [person] [location] [monetary value]

分類單詞

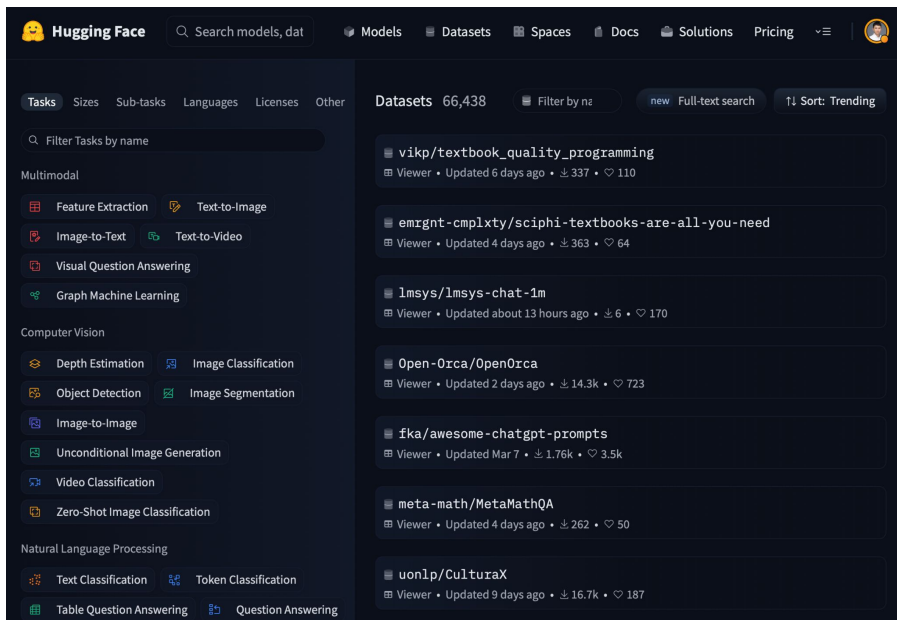
資料

- 如何找資料？

分類單詞

資料

- 如何找資料？ <https://huggingface.co/datasets>



分類單詞

資料

Natural Language Processing



Text Classification



Token Classification



Table Question Answering



Question Answering



Zero-Shot Classification



Translation



Summarization



Conversational



Text Generation



Text2Text Generation



Fill-Mask



Sentence Similarity



Table to Text



Multiple Choice



Text Retrieval

分類單詞

資料

Datasets 506

Filter by na

new Full-text search

↑↓ Sort: Trending

Open-Orca/OpenOrca

Viewer • Updated 2 days ago • 14.3k • 723

conll2003

Viewer • Updated Apr 5 • 49.5k • 67

xtreme

Viewer • Updated Jun 1 • 33.3k • 57

albertvillanova/universal_dependencies

Viewer • Updated Dec 15, 2022 • 2 • 5

分類單詞

資料

Datasets: con112003 like 67

Tasks: Token Classification Sub-tasks: **named-entity-recognition** **part-of-speech** Languages: English Multilinguality: monolingual

Size Categories: 10K<n<100K Language Creators: found Annotations Creators: crowdsourced Source Datasets: extended|other-reuters-corpus

License: other

Dataset card Files Community 10

Dataset Viewer Auto-converted to Parquet API Go to dataset viewer

Split

train (14k rows)

Search this dataset

id	tokens	pos_tags	chunk_tags	nex_tags
string · lengths	sequence	sequence	sequence	sequence
1→2				
				0.1%
0	["EU", "rejects", "German", "call", "to", "boycott", "British", "lamb", ".: "]	[22, 42, 16, 21, 35, 37, 16, 21, 7]	[11, 21, 11, 12, 21, 22, 11, 12, 0]	[3, 0, 7, 0, 0, 0, 7, 0, 0]

Downloads last month **49,482**

Use in dataset library

Edit dataset card

Train in AutoTrain

Papers with Code

Evaluate models

HF Leaderboard

Homepage:

分類單詞

資料

Homepage:
aclweb.org

Size of downloaded dataset files:
983 kB

Size of the auto-converted Parquet files:
1.82 MB

Number of rows:
20,744

分類單詞

資料

```
{  
  "chunk_tags": [11, 12, 12, 21, 13, 11, 11, 21, 13, 11, 12, 13, 11, 21, 22, 11, 12, 17, 11, 21, 17, 11, 12, 12, 21, 22, 22, 13, 11, 0],  
  "id": "0",  
  "ner_tags": [0, 3, 4, 0, 0, 0, 0, 0, 0, 7, 0, 0, 0, 0, 0, 7, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],  
  "pos_tags": [12, 22, 22, 38, 15, 22, 28, 38, 15, 16, 21, 35, 24, 35, 37, 16, 21, 15, 24, 41, 15, 16, 21, 21, 20, 37, 40, 35, 21, 7],  
  "tokens": ["The", "European", "Commission", "said", "on", "Thursday", "it", "disagreed", "with", "German", "advice", "to", "consumers",  
    "to", "shun", "British", "lamb", "until", "scientists", "determine", "whether", "mad", "cow", "disease", "can", "be", "transmitted", "to",  
    "sheep", "."]  
}
```

分類單詞

資料

NER Tag Legend with Definitions and Mapping



0 - O - Other



1 - B-PER - Beginning of Person



2 - I-PER - Inside of Person



3 - B-ORG - Beginning of Organization



4 - I-ORG - Inside of Organization



5 - B-LOC - Beginning of Location



6 - I-LOC - Inside of Location



7 - B-MISC - Beginning of Miscellaneous



8 - I-MISC - Inside of Miscellaneous

分類單詞

資料

Commission
The European said on Thursday it disagreed with German advice to consumers shun British lamb

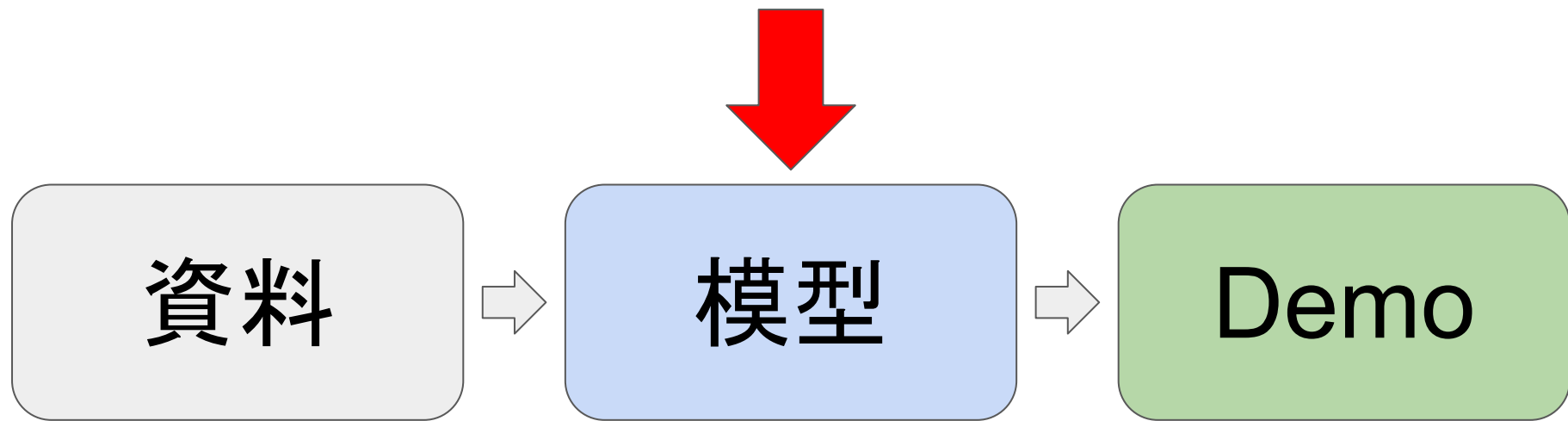
0 3 4 0 0 0 0 0 0 7 0 0 0 0 0 7 0

3 - B-ORG - Beginning of Organization

4 - I-ORG - Inside of Organization

7 - B-MISC - Beginning of Miscellaneous

專案的一生



怎麼訓練？

模型

The screenshot shows the GitHub interface for the HuggingFace `transformers` repository. At the top, the repository name `huggingface / transformers` is displayed. Below the navigation bar, the repository is identified as `transformers` (Public). The main content area shows a list of recent commits, with the most recent one by `Rocketknight1` adding an `add_generation_prompt` argument to `apply_chat_te...`. The commit message is truncated. The commit hash is `8b46c5b`, and it was made 1 minute ago. The repository has 14,115 commits. On the right side, the 'About' section describes the repository as 'Transformers: State-of-the-art Machine Learning for Pytorch, TensorFlow, and JAX.' and provides a link to `huggingface.co/transformers`. Below the link, there are tags for various topics: `python`, `nlp`, `machine-learning`, `natural-language-processing`, `deep-learning`, `tensorflow`, `pytorch`, `transformer`, `speech-recognition`, `seq2seq`, `flax`, and `pretrained-models`.

huggingface / transformers

Code Issues 656 Pull requests 222 Actions Projects 25 Security Insights

transformers Public

Watch 1.1k Fork 22.5k Starred 113k

main 228 branches 137 tags

Go to file Add file Code

Commits

	Rocketknight1 Add add_generation_prompt argument to apply_chat_te...	8b46c5b 1 minute ago	14,115 commits
	.circleci Docstring check (#26052)		1 hour ago
	.github [AMD] Add initial version for run_tests_multi_gpu (#26346)		yesterday
	docker Integrate AMD GPU in CI/CD environment (#26007)		2 weeks ago
	docs Add add_generation_prompt argument to apply_chat_template (#26...		1 minute ago
	examples Bump pillow from 9.3.0 to 10.0.1 in /examples/research_projects/deci...		4 hours ago

About

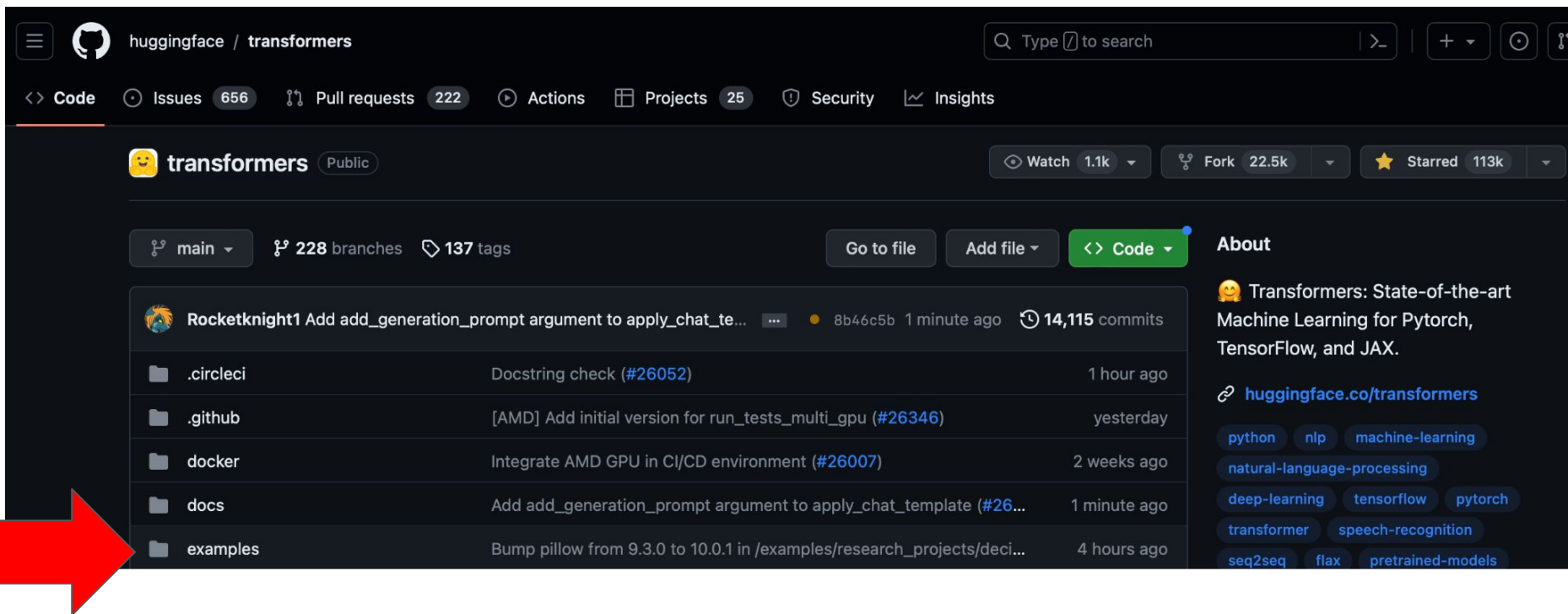
Transformers: State-of-the-art Machine Learning for Pytorch, TensorFlow, and JAX.

huggingface.co/transformers

python nlp machine-learning natural-language-processing deep-learning tensorflow pytorch transformer speech-recognition seq2seq flax pretrained-models

怎麼訓練？

模型



The screenshot shows the GitHub interface for the HuggingFace transformers repository. At the top, the repository name 'huggingface / transformers' is displayed. Below the navigation bar, the repository is identified as 'transformers' (Public). The main content area shows a list of files and directories. A red arrow points to the 'examples' directory. The right sidebar contains the 'About' section, which describes the repository as 'Transformers: State-of-the-art Machine Learning for Pytorch, TensorFlow, and JAX.' and provides a link to 'huggingface.co/transformers'. Below the link are several tags: python, nlp, machine-learning, natural-language-processing, deep-learning, tensorflow, pytorch, transformer, speech-recognition, seq2seq, flax, and pretrained-models.

huggingface / transformers

Code Issues 656 Pull requests 222 Actions Projects 25 Security Insights

transformers Public

Watch 1.1k Fork 22.5k Starred 113k

main 228 branches 137 tags

Go to file Add file Code

File	Commit Message	Commit Hash	Time
.circleci	Docstring check (#26052)	8b46c5b	1 minute ago
.github	[AMD] Add initial version for run_tests_multi_gpu (#26346)		yesterday
docker	Integrate AMD GPU in CI/CD environment (#26007)		2 weeks ago
docs	Add add_generation_prompt argument to apply_chat_template (#26...		1 minute ago
examples	Bump pillow from 9.3.0 to 10.0.1 in /examples/research_projects/deci...		4 hours ago

About

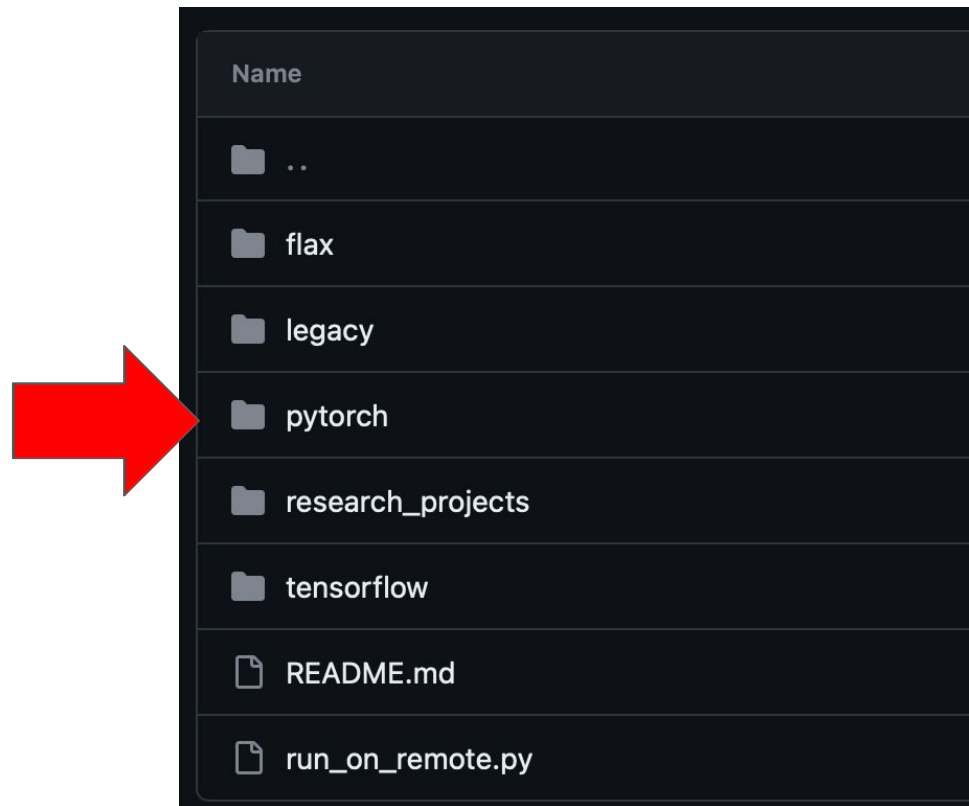
Transformers: State-of-the-art Machine Learning for Pytorch, TensorFlow, and JAX.

huggingface.co/transformers

python nlp machine-learning natural-language-processing deep-learning tensorflow pytorch transformer speech-recognition seq2seq flax pretrained-models

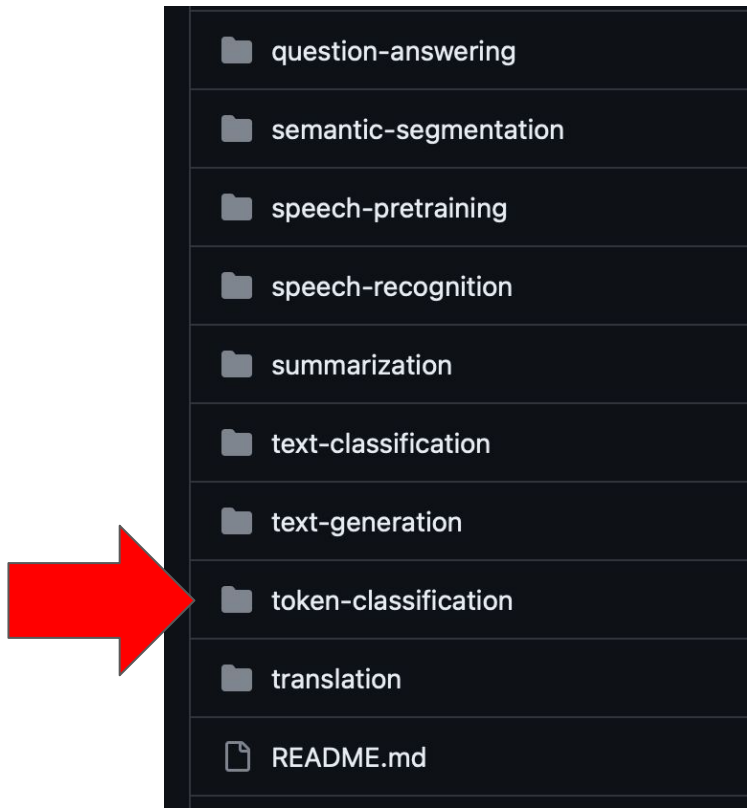
怎麼訓練？

模型

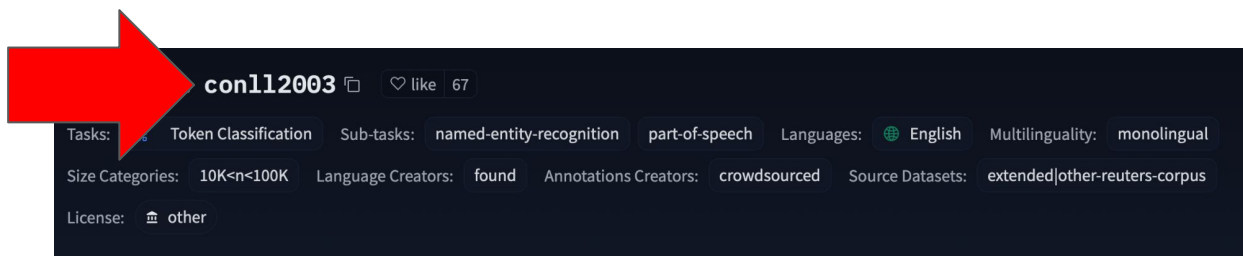


怎麼訓練？

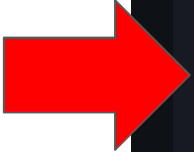
模型



怎麼訓練？

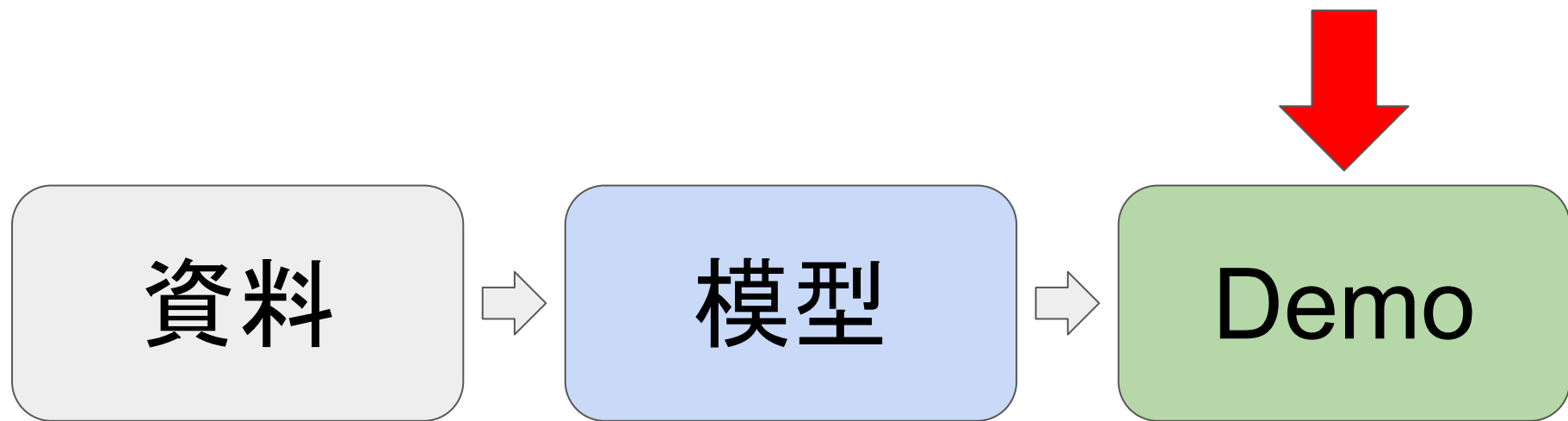


The following example fine-tunes BERT on CoNLL-2003:



```
python run_ner.py \  
  --model_name_or_path bert-base-uncased \  
  --dataset_name conll2003 \  
  --output_dir /tmp/test-ner \  
  --do_train \  
  --do_eval
```

專案的一生



Gradio

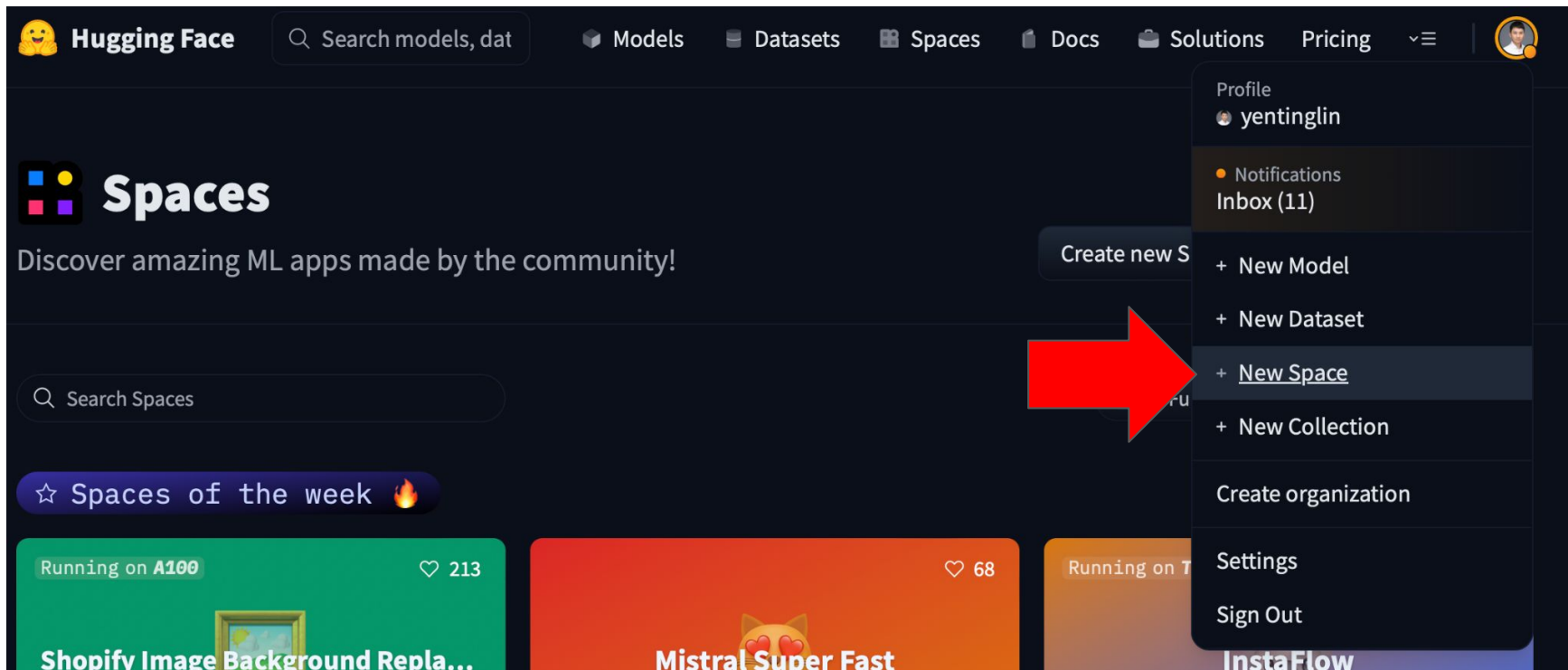
```
import gradio as gr
from transformers import pipeline

classifier = pipeline(task='ner', '你的模型', device=0)

demo = gr.Interface(
    fn=...,
    inputs=gr.Textbox(placeholder="請輸入一段文字..."),
    outputs="label",
    interpretation="default",
    examples=[["Take me to church"]]
)

demo.launch()
```

Gradio



The screenshot shows the Hugging Face website interface. At the top, the navigation bar includes the Hugging Face logo, a search bar for models and datasets, and links to Models, Datasets, Spaces, Docs, Solutions, Pricing, and a user profile icon. The user profile icon has a dropdown menu open, listing options: Profile (yentinglin), Notifications, Inbox (11), + New Model, + New Dataset, + New Space (highlighted with a red arrow), + New Collection, Create organization, Settings, and Sign Out. Below the navigation bar, the 'Spaces' section is featured with the text 'Discover amazing ML apps made by the community!'. A search bar for Spaces is present. A 'Spaces of the week' section displays three featured Spaces: 'Shopify Image Background Repla...' (Running on A100, 213 likes), 'Mistral Super Fast' (68 likes), and 'InstaFlow' (Running on T4).

Hugging Face Search models, datasets, Spaces, Docs, Solutions, Pricing

Spaces
Discover amazing ML apps made by the community!

Search Spaces

Spaces of the week 🔥

- Running on **A100** 213 likes
Shopify Image Background Repla...
- 68 likes
Mistral Super Fast
- Running on **T4**
InstaFlow

User menu options:
Profile: yentinglin
Notifications
Inbox (11)
+ New Model
+ New Dataset
+ New Space
+ New Collection
Create organization
Settings
Sign Out

Gradio



Create a new Space

Spaces are Git repositories that host application code for Machine Learning demos. You can build Spaces with Python libraries like Streamlit or Gradio, or using Docker images.

Owner

yentinglin

Space name

test

License

License

Select the Space SDK

You can choose between Streamlit, Gradio, and Static for your Space. Or pick Docker to host any other app.



Streamlit



Gradio







Docker






10 templates










Static







Gradio

 Spaces:  yentinglin / **test**  private No application file 

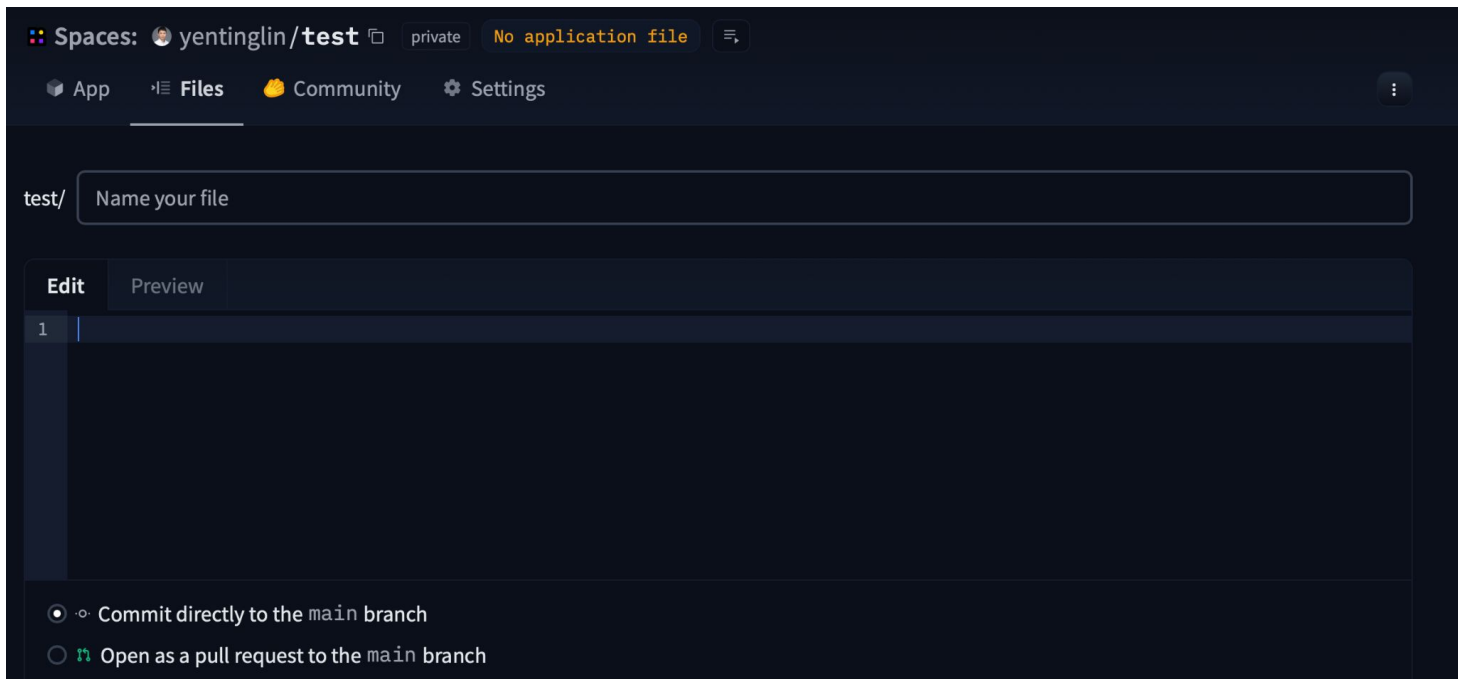
 App  **Files**  Community  Settings 

 main  test  1 contributor  History: 1 commit  + Add file 

 yentinglin initial commit 26062ec less than a minute ago

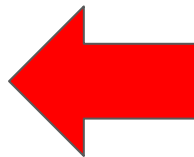
 .gitattributes 	1.52 kB		initial commit	less than a minute ago
 README.md 	225 Bytes		initial commit	less than a minute ago

Gradio





四大任務


- 分類整句
- 分類句中每個單詞
- 生成文本內容
- 從文本中提取答案



文本生成 - 摘要


資料


Datasets: cnn_dailymail


 like

109


Tasks:

 Summarization

Sub-tasks:

news-articles-summarization

Languages:

 English

Multilingual

Size Categories:

100K< n < 1M

Language Creators:

found


Annotations Creators:


no-annotation


Sources:


1

License:


 apache-2.0


 **Dataset card**


 Files


 Community

5

 **Dataset Viewer**

 Auto-converted to Parquet

 API


 Go to dataset viewer

Subset

1.0.0 (312k rows)

Split

train (287k rows)

 Search this dataset

article string	highlights string	id string
LONDON, England (Reuters) -- Harry Potter star...	Harry Potter star Daniel...	42c027e4ff9730fbb3de84c1af0d2c506e41c3e4
Editor's note: In our Behind the Scenes series...	Mentally ill inmates in...	ee8871b15c50d0db17b0179a6d2beab35065f1e9

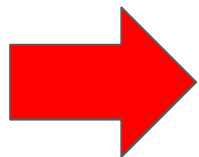
文本生成 - 摘要

資料

article string	highlights string
LONDON, England (Reuters) -- Harry Potter star Daniel Radcliffe gains access to a reported £20 million (\$41.1 million) fortune as he turns 18 on Monday, but he insists the money won't cast a spell on him. Daniel Radcliffe as Harry Potter in "Harry Potter and the Order of the Phoenix" To the disappointment of gossip columnists around the world, the young actor says he has no plans to fritter his cash away on fast cars, drink and celebrity parties. "I don't plan to be one of those people who, as soon as they turn 18, suddenly buy themselves a massive sports car collection or something similar," he told an Australian interviewer earlier this month. "I don't think I'll be particularly extravagant. "The things I like buying are things that cost about 10 pounds	Harry Potter star Daniel Radcliffe gets £20M fortune as he turns 18 Monday . Young actor says he has no plans to fritter his cash away . Radcliffe's earnings from first five Potter films have been held in trust fund .

模型

怎麼訓練？



semantic-segmentation

speech-pretraining

speech-recognition

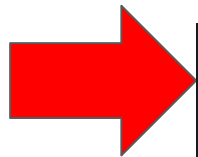
summarization

text-classification

text-generation

模型

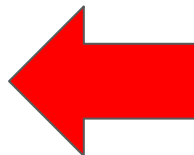
怎麼訓練？



```
python examples/pytorch/summarization/run_summarization.py \  
  --model_name_or_path t5-small \  
  --do_train \  
  --do_eval \  
  --dataset_name cnn_dailymail \  
  --dataset_config "3.0.0" \  
  --source_prefix "summarize: " \  
  --output_dir /tmp/tst-summarization \  
  --per_device_train_batch_size=4 \  
  --per_device_eval_batch_size=4 \  
  --overwrite_output_dir \  
  --predict_with_generate
```



四大任務




- 分類整句
- 分類句中每個單詞
- 生成文本內容
- 從文本中提取答案




抽取式問答



資料




Datasets: squad   like 136

Tasks:  Question Answering Sub-tasks: **extractive-qa** Languages:  English Multilinguality: 


Language Creators: **crowdsourced** **found** Annotations Creators: **crowdsourced** Source Datasets: **exte**

License:  cc-by-4.0




Dataset card  Files  Community 5

Dataset Viewer  Auto-converted to Parquet  API  Go to dataset viewer

Split

train (87.6k rows) 

Search this dataset

title string · lengths	context string · lengths	question string · lengths	answers sequence
 3 59	 151 3.71k	 1 25.7k	
1182 University_of_Notre_Dame	Architecturally, the school has a...	To whom did the Virgin Mary...	{ "text": ["Saint...



抽取式問答

資料

context	question	answers
string · lengths	string · lengths	sequence
		
507↔863 58...	1↔2.57k 100%	
Architecturally, the school has a Catholic character. Atop the Main Building's gold dome is a golden statue of the Virgin Mary. Immediately in front of the Main Building and facing it, is a copper statue of Christ with arms upraised with the legend "Venite Ad Me Omnes". Next to the Main Building is the Basilica of the Sacred Heart. Immediately behind the basilica is the Grotto, a Marian place of prayer and reflection. It is a replica of the grotto at Lourdes, France where the Virgin Mary	To whom did the Virgin Mary allegedly appear in 1858 in Lourdes France?	{ "text": ["Saint Bernadette Soubirous"], "answer_start": [515] }

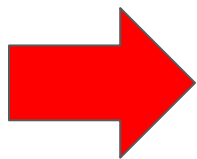
抽取式問答

資料

context	question	answers
string · lengths	string · lengths	sequence
		
507↔863 58...	1↔2.57k 100%	
Architecturally, the school has a Catholic character. Atop the Main Building's gold dome is a golden statue of the Virgin Mary. Immediately in front of the Main Building and facing it, is a copper statue of Christ with arms upraised with the legend "Venite Ad Me Omnes". Next to the Main Building is the Basilica of the Sacred Heart. Immediately behind the basilica is the Grotto, a Marian place of prayer and reflection. It is a replica of the grotto at Lourdes, France where the Virgin Mary	To whom did the Virgin Mary allegedly appear in 1858 in Lourdes France?	{ "text": ["Saint Bernadette Soubirous"], "answer_start": [515] }

怎麼訓練？

模型



language-modeling

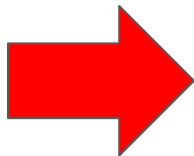
multiple-choice

question-answering

semantic-segmentation

speech-pretraining

怎麼訓練？



```
python run_qa.py \  
  --model_name_or_path bert-base-uncased \  
  --dataset_name squad \  
  --do_train \  
  --do_eval \  
  --per_device_train_batch_size 12 \  
  --learning_rate 3e-5 \  
  --num_train_epochs 2 \  
  --max_seq_length 384 \  
  --doc_stride 128 \  
  --output_dir /tmp/debug_squad/
```

四大任務專案的一生

- 分類整句
- 分類句中每個**單詞**
- **生成**文本內容
- 從文本中**提取**答案

