LAB2: TRANSFORMERS AND BERT

LECTURER : JIRAWAN CHAROENSUK

jirawan.charo@ku.th

AGENDA

- Transformers
  - https://huggingface.co/
  - Pipeline
- BERT
  - Fine-Tune
- BERT multilingual base model
- BERT-th
- WangchanBERTa: Pre-trained Thai Language Model

TRANSFORMERS

https://huggingface.co/

The AI community building the future.

Build, train and deploy state of the art models powered by the reference open source in machine learning.

Star 87,466

More than 5,000 organizations are using Hugging Face

Allen Institute for AI
Non-Profit • 154 models

Meta AI
Company • 667 models

Graphcore
Company • 36 models

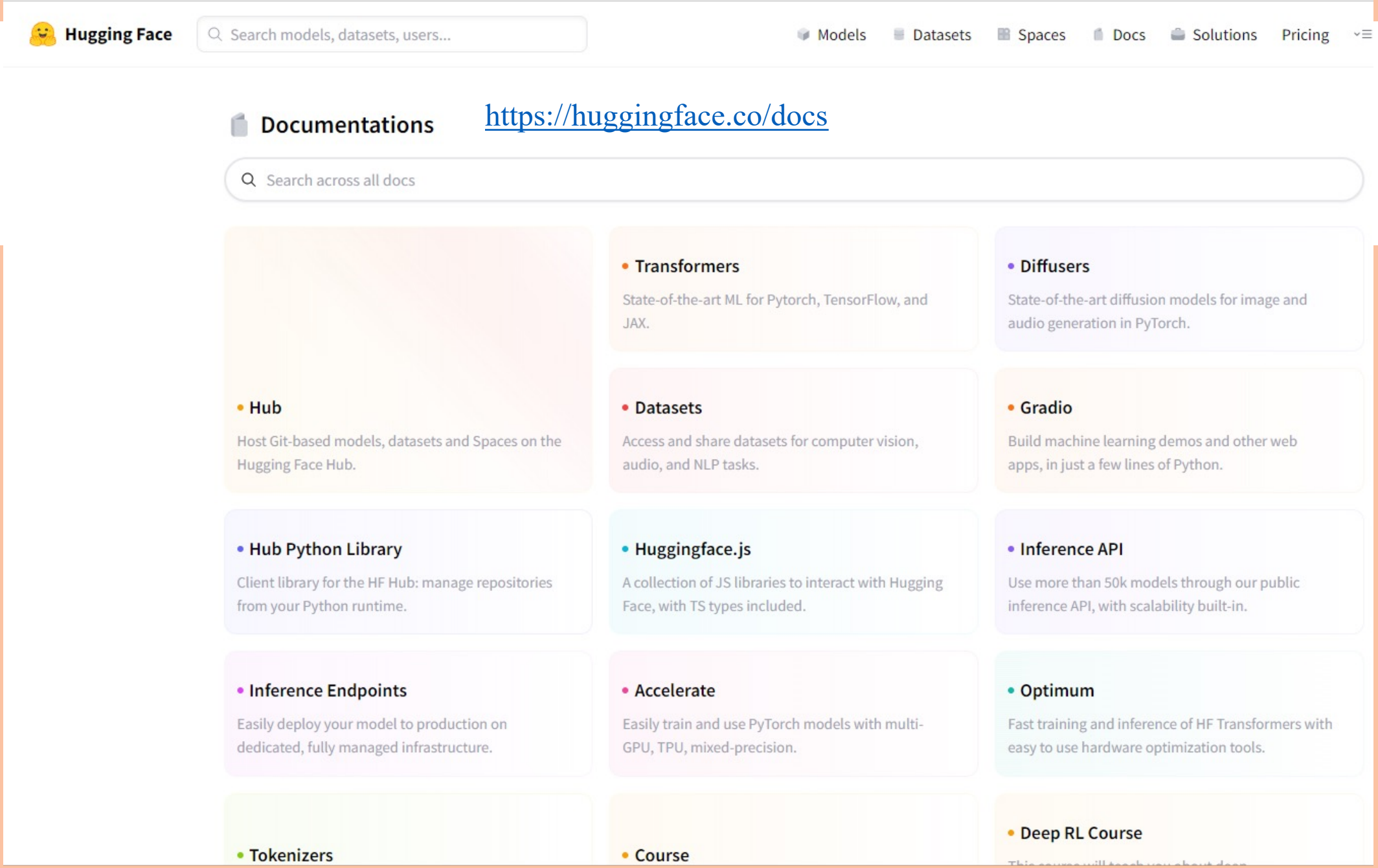
Google AI
Company • 580 models

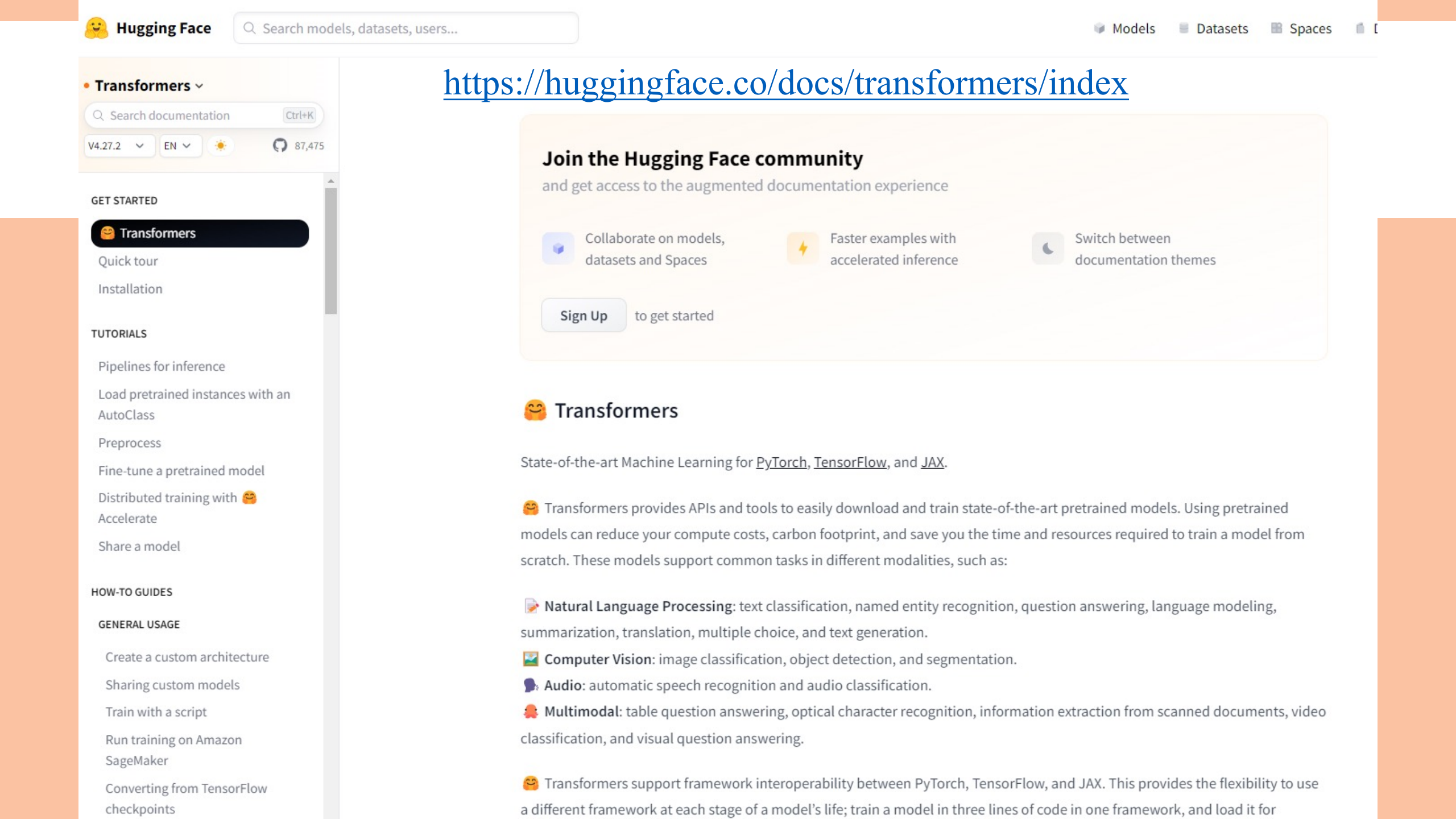
Intel
Company • 90 models

SpeechBrain
Non-Profit • 69 models

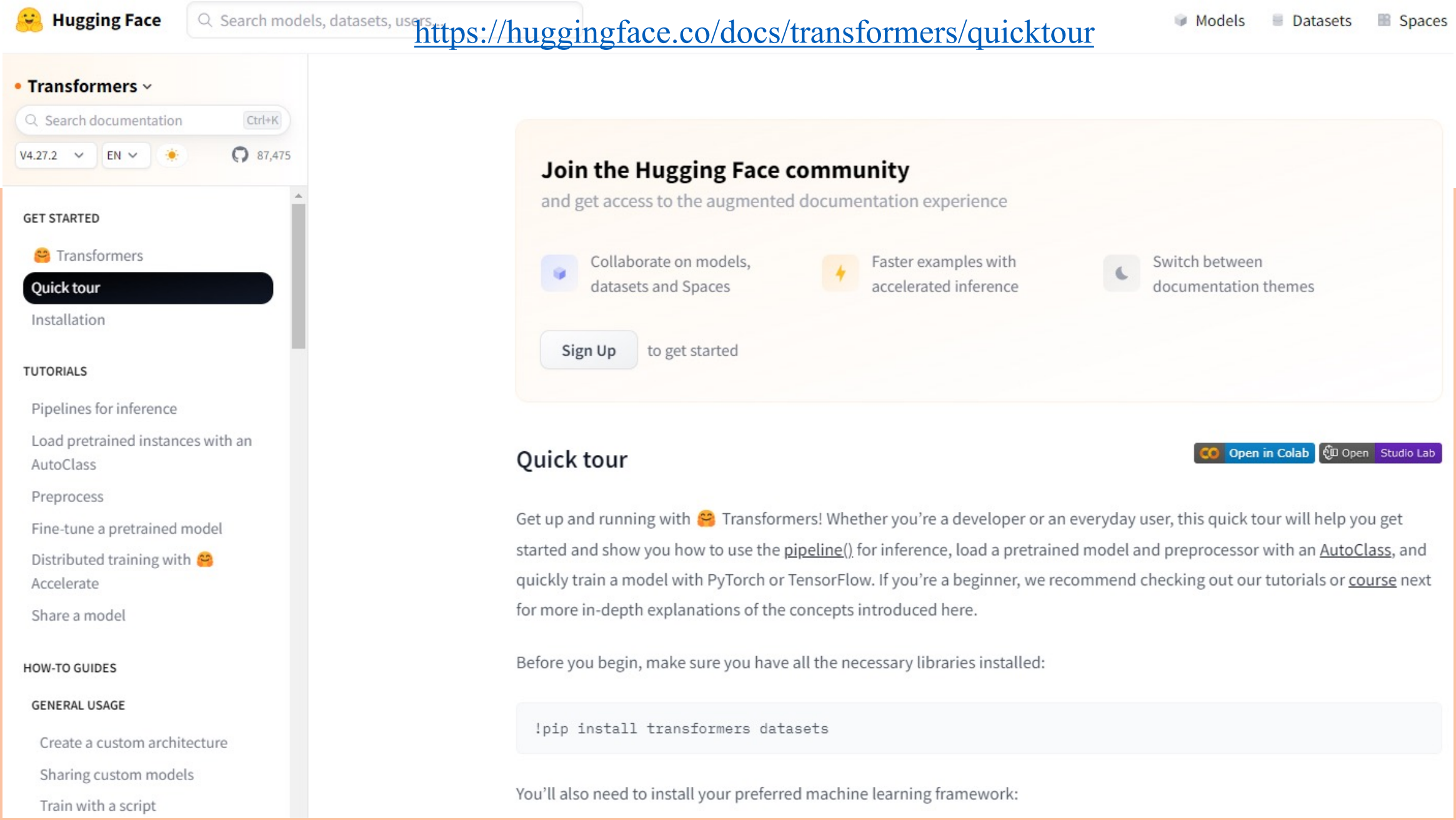
Microsoft
Company • 243 models

Grammarly
Company • 1 model


Hugging Face
Search models, datasets, users...
Models
Datasets
Spaces
Docs
Solutions
Pricing
Documentations
https://huggingface.co/docs
Search across all docs
Transformers
State-of-the-art ML for Pytorch, TensorFlow, and JAX.
Diffusers
State-of-the-art diffusion models for image and audio generation in PyTorch.
Hub
Host Git-based models, datasets and Spaces on the Hugging Face Hub.
Datasets
Access and share datasets for computer vision, audio, and NLP tasks.
Gradio
Build machine learning demos and other web apps, in just a few lines of Python.
Hub Python Library
Client library for the HF Hub: manage repositories from your Python runtime.
Huggingface.js
A collection of JS libraries to interact with Hugging Face, with TS types included.
Inference API
Use more than 50k models through our public inference API, with scalability built-in.
Inference Endpoints
Easily deploy your model to production on dedicated, fully managed infrastructure.
Accelerate
Easily train and use PyTorch models with multi-GPU, TPU, mixed-precision.
Optimum
Fast training and inference of HF Transformers with easy to use hardware optimization tools.
Tokenizers
Course
Deep RL Course

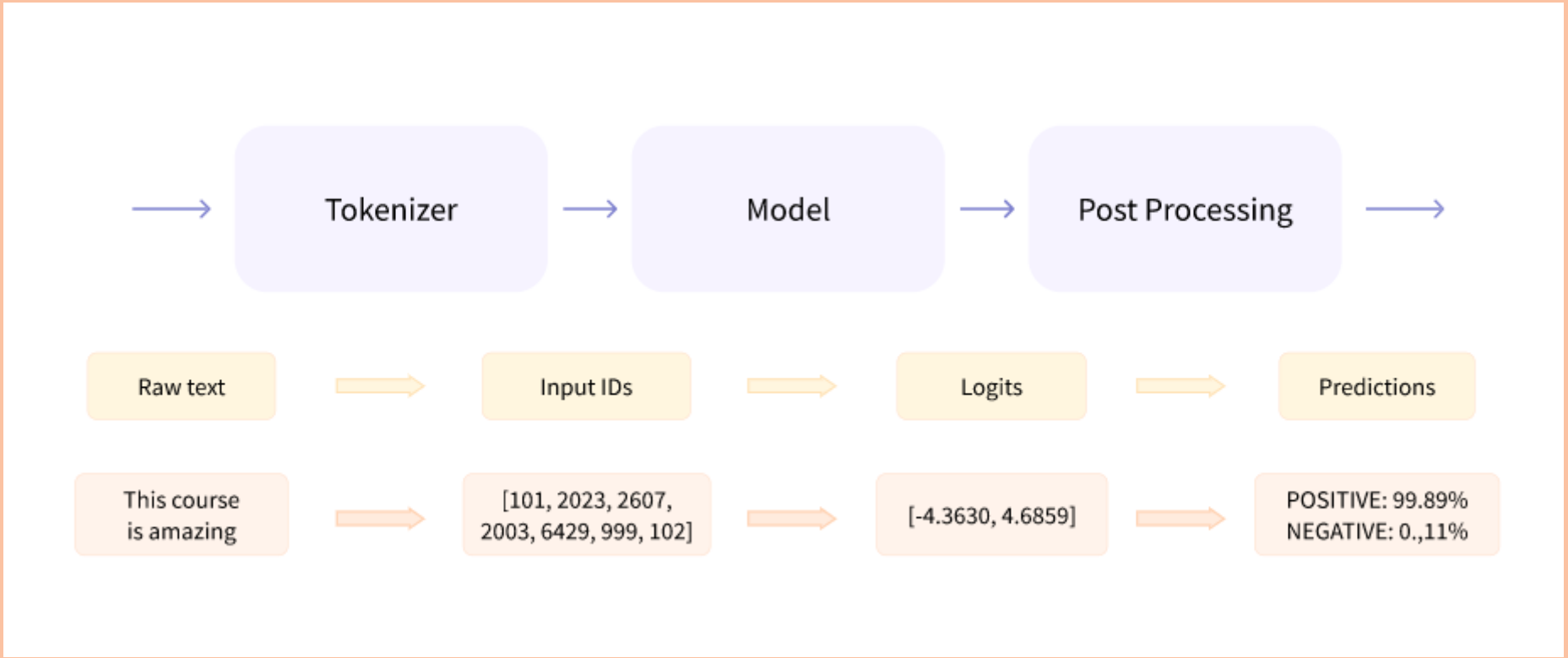

Hugging Face
Search models, datasets, users...
Models
Datasets
Spaces
Transformers
Search documentation
Ctrl+K
V4.27.2
EN
87,475
GET STARTED
Transformers
Quick tour
Installation
TUTORIALS
Pipelines for inference
Load pretrained instances with an AutoClass
Preprocess
Fine-tune a pretrained model
Distributed training with Accelerate
Share a model
HOW-TO GUIDES
GENERAL USAGE
Create a custom architecture
Sharing custom models
Train with a script
Run training on Amazon SageMaker
Converting from TensorFlow checkpoints
https://huggingface.co/docs/transformers/index
Join the Hugging Face community
and get access to the augmented documentation experience
Collaborate on models, datasets and Spaces
Faster examples with accelerated inference
Switch between documentation themes
Sign Up
to get started
Transformers
State-of-the-art Machine Learning for PyTorch, TensorFlow, and JAX.
Transformers provides APIs and tools to easily download and train state-of-the-art pretrained models. Using pretrained models can reduce your compute costs, carbon footprint, and save you the time and resources required to train a model from scratch. These models support common tasks in different modalities, such as:
Natural Language Processing: text classification, named entity recognition, question answering, language modeling, summarization, translation, multiple choice, and text generation.
Computer Vision: image classification, object detection, and segmentation.
Audio: automatic speech recognition and audio classification.
Multimodal: table question answering, optical character recognition, information extraction from scanned documents, video classification, and visual question answering.
Transformers support framework interoperability between PyTorch, TensorFlow, and JAX. This provides the flexibility to use a different framework at each stage of a model's life; train a model in three lines of code in one framework, and load it for

https://huggingface.co/docs/transformers/quicktour



PIPELINE

- The pipeline() is the easiest and fastest way to use a pretrained model for inference



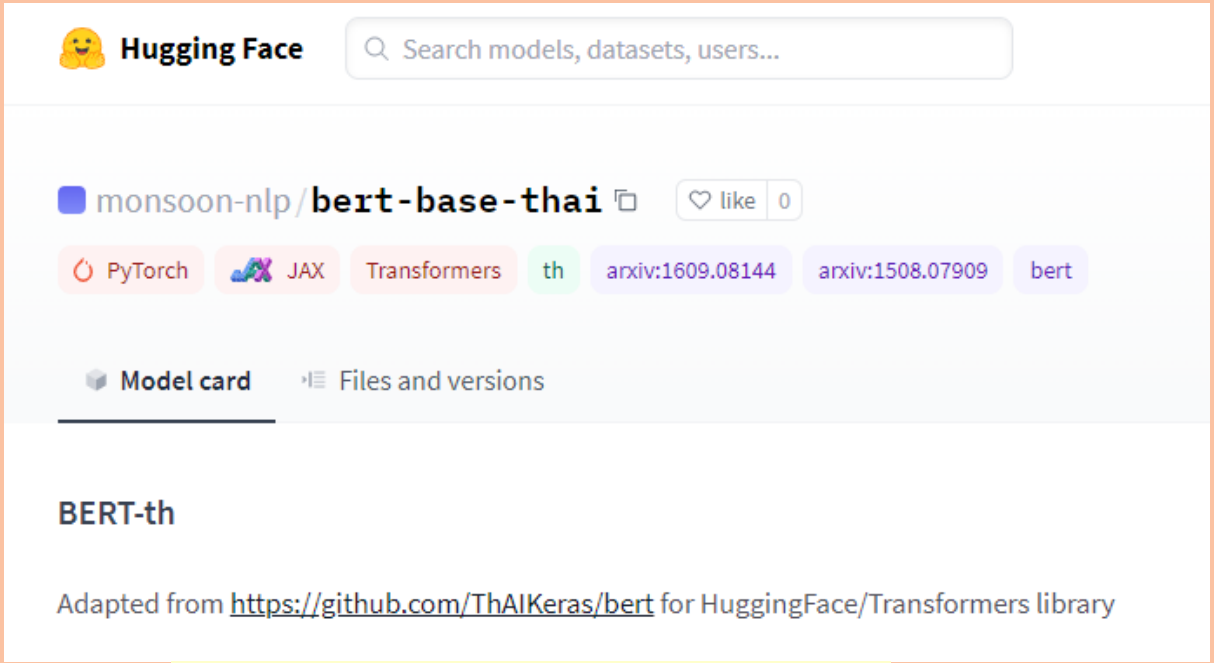
Task	Description	Modality	Pipeline identifier
Text classification	assign a label to a given sequence of text	NLP	pipeline(task="sentiment-analysis")
Text generation	generate text given a prompt	NLP	pipeline(task="text-generation")
Summarization	generate a summary of a sequence of text or document	NLP	pipeline(task="summarization")
Image classification	assign a label to an image	Computer vision	pipeline(task="image-classification")
Image segmentation	assign a label to each individual pixel of an image (supports semantic, panoptic, and instance segmentation)	Computer vision	pipeline(task="image-segmentation")
Object detection	predict the bounding boxes and classes of objects in an image	Computer vision	pipeline(task="object-detection")
Audio classification	assign a label to some audio data	Audio	pipeline(task="audio-classification")
Automatic speech recognition	transcribe speech into text	Audio	pipeline(task="automatic-speech-recognition")
Visual question answering	answer a question about the image, given an image and a question	Multimodal	pipeline(task="vqa")
Document question answering	answer a question about a document, given an image and a question	Multimodal	pipeline(task="document-question-answering")
Image captioning	generate a caption for a given image	Multimodal	pipeline(task="image-to-text")

BERT

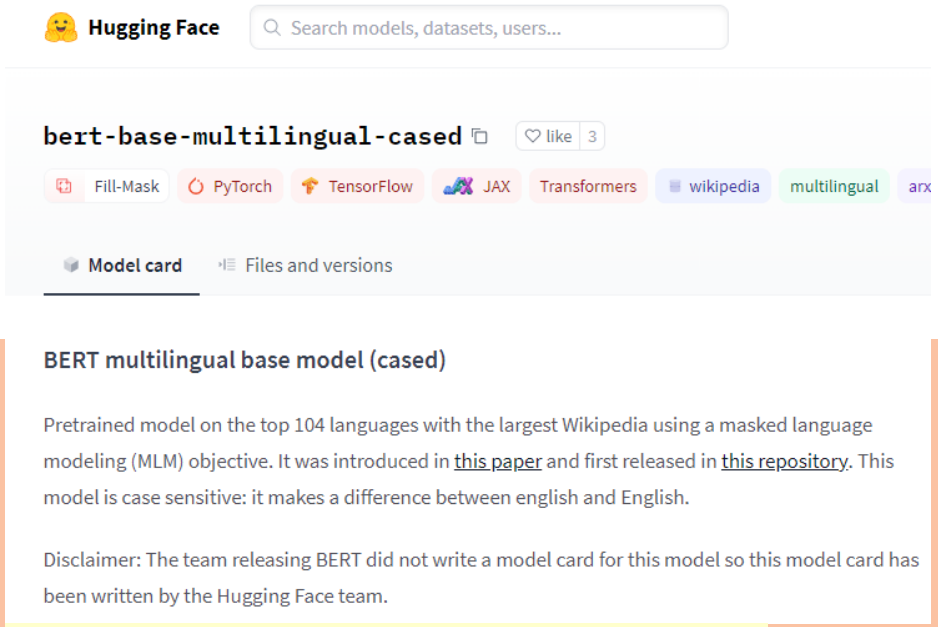
BERT-th, WangchanBerta, BERT multilingual base model

BERT

- BERT multilingual base model
- BERT-th



https://huggingface.co/monsoon-nlp/bert-base-thai



https://huggingface.co/bert-base-multilingual-cased

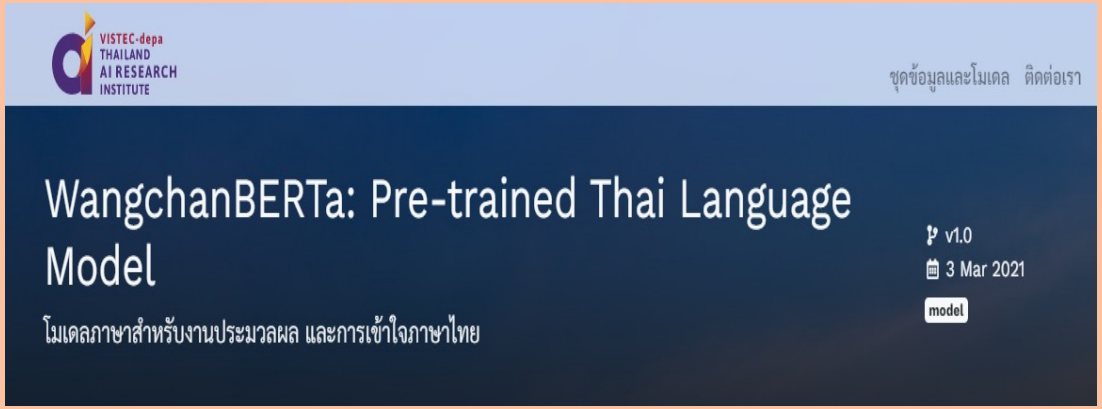
***** New November 23rd, 2018: Un-normalized multilingual model + Thai + Mongolian *****

We uploaded a new multilingual model which does *not* perform any normalization on the input (no lower casing, accent stripping, or Unicode normalization), and additionally inclues Thai and Mongolian.

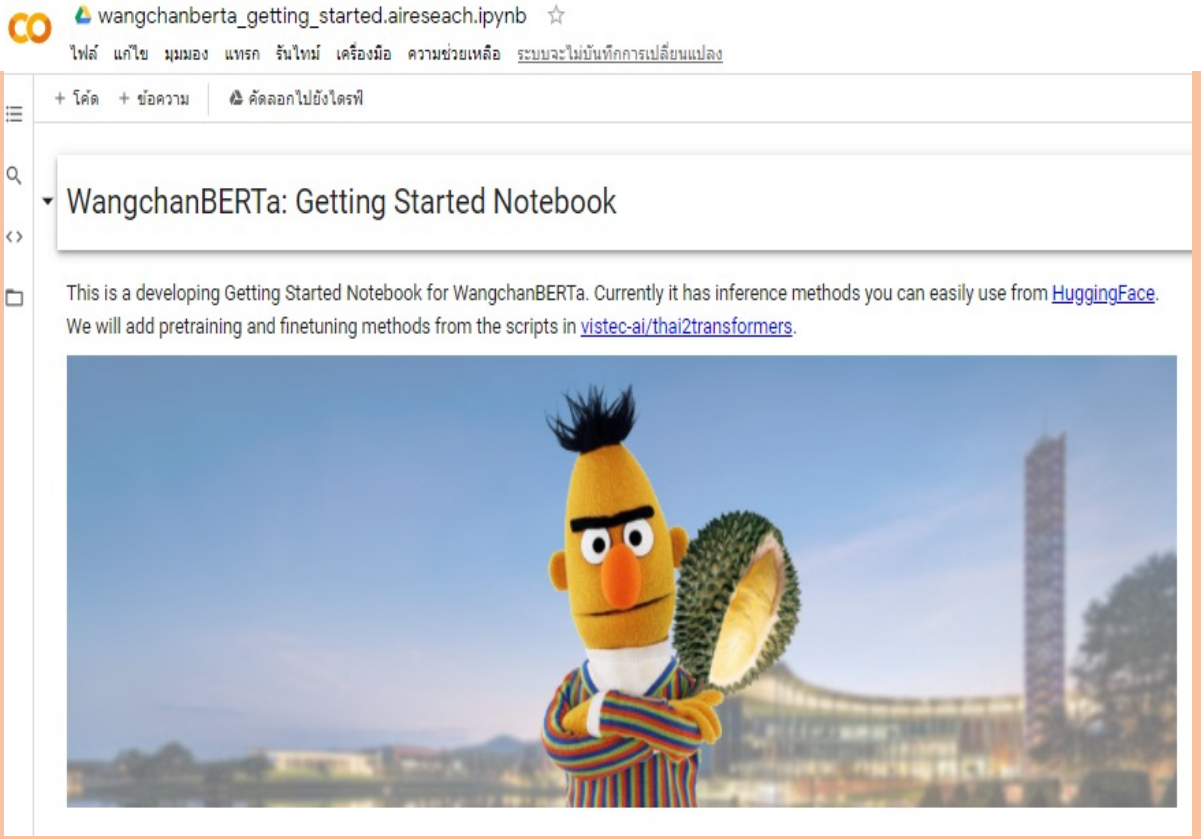
It is recommended to use this version for developing multilingual models, especially on languages with non-Latin alphabets.

This does not require any code changes, and can be downloaded here:

GitHub - google-research/bert: TensorFlow code and pre-trained models for BERT

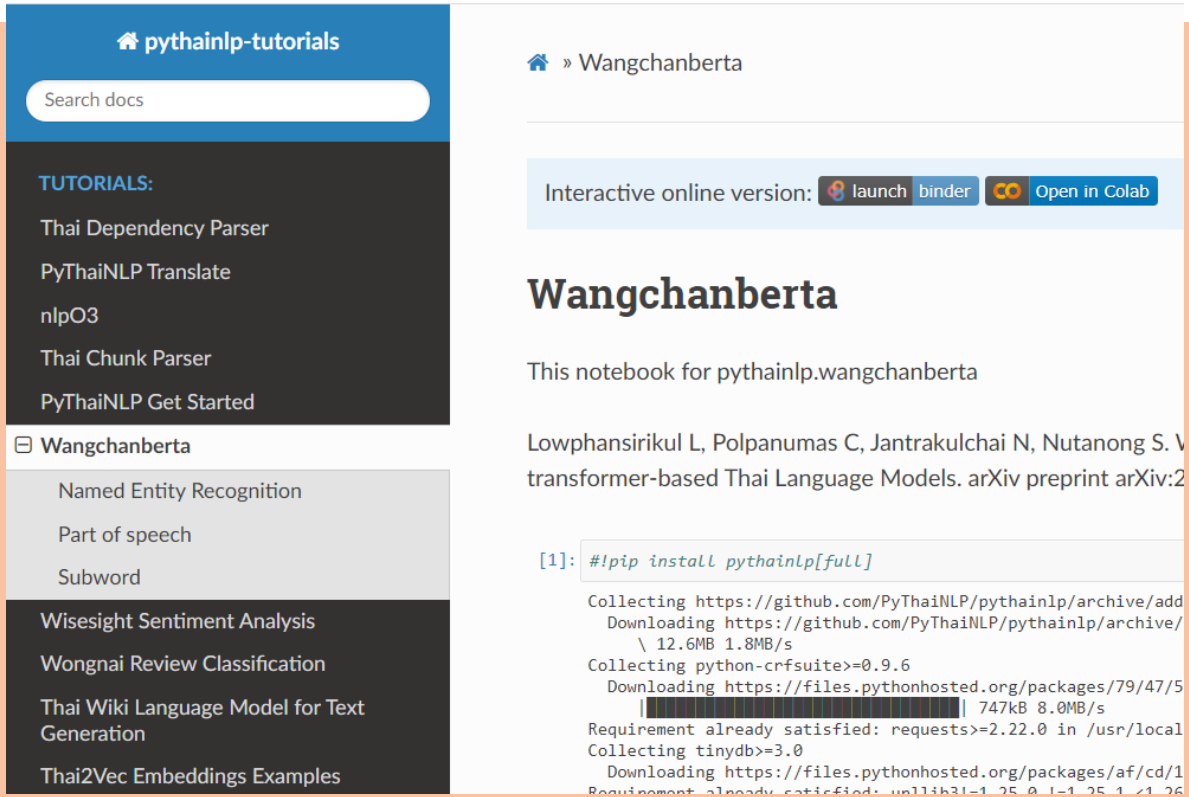


https://airesearch.in.th/releases/wangchanberta-pre-trained-thai-language-model/

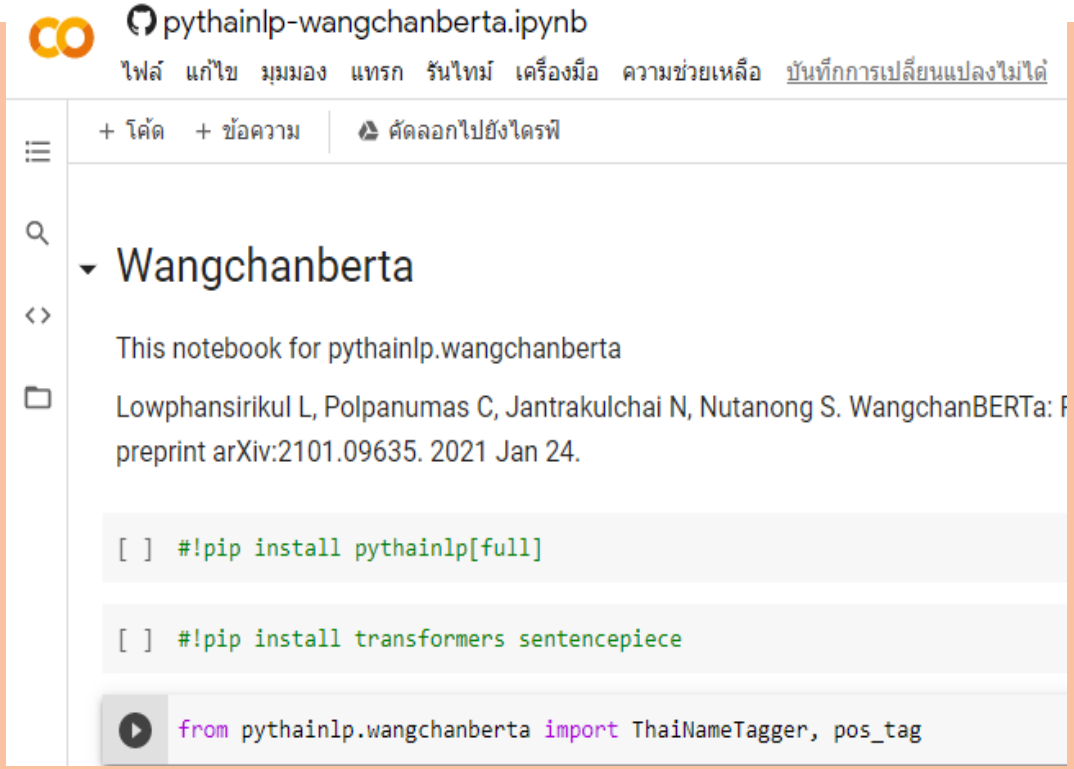


https://colab.research.google.com/drive/1Kbk6sBspZLwcnOE61adAQo30xxqOQ9ko

WANGCHANBERTA



https://pythainlp.github.io/tutorials/notebooks/pythainlp_wangchanberta.html



https://colab.research.google.com/github/PyThaiNLP/tutorials/blob/master/source/notebooks/pythainlp_wangchanberta.ipynb#scrollTo=Wpi77fIZgH6d

QUESTION

https://code-ai.mk/would-you-like-a-full-video-tutorials/