SATRIO RIZKI RAMADHAN (11121413) 4KA20

Automated Article Generation

Dalam praktikum ini Anda akan mencoba untuk membuat sebuah artikel dengan library yang telah di training sebelumnya atau yang dikenal sebagai pre-trained model. Apabila Anda akan mengunduh file praktikum dan menjalankan pada mesin local Anda, silahkan pastikan dependency library lainnya telah terinstall. Pada praktikum ini Anda akan menggunakan model GPT2 dari OpenAI.

```
pip install transformers
```

```
Requirement already satisfied: transformers in /usr/local/lib/python3.10/dist-packages (4.44.2)
Requirement already satisfied: filelock in /usr/local/lib/python3.10/dist-packages (from transformers) (3.16.1)
Requirement already satisfied: huggingface-hub<1.0,>=0.23.2 in /usr/local/lib/python3.10/dist-packages (from transformers) (0.24.7)
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) (24.1)
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) (2024.9.11)
Requirement already satisfied: requests in /usr/local/lib/python3.10/dist-packages (from transformers) (2.32.3)
Requirement already satisfied: safetensors>=0.4.1 in /usr/local/lib/python3.10/dist-packages (from transformers) (0.4.5)
Requirement already satisfied: tokenizers<0.20,>=0.19 in /usr/local/lib/python3.10/dist-packages (from transformers) (0.19.1)
Requirement already satisfied: tqdm>=4.27 in /usr/local/lib/python3.10/dist-packages (from transformers) (4.66.5)
Requirement already satisfied: fsspec>=2023.5.0 in /usr/local/lib/python3.10/dist-packages (from huggingface-hub<1.0,>=0.23.2->transformers) (
Requirement already satisfied: typing-extensions>=3.7.4.3 in /usr/local/lib/python3.10/dist-packages (from huggingface-hub<1.0,>=0.23.2->trans
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) (2024.8.30)
```

```
import tensorflow as tf
from transformers import GPT2LMHeadModel, GPT2Tokenizer
```

Set Up Tokenizer dan Model

Pada praktikum ini karena kita menggunakan model yang telah di latih sebelumnya, kita perlu mendefinisikan dua kompenen sebelum dapat membuat artikel secara otomatis yaitu tipe tokenizer dan model yang akan digunakan.

```
tokenizer = GPT2Tokenizer.from_pretrained("gpt2-large")
```

```
/usr/local/lib/python3.10/dist-packages/huggingface_hub/utils/_token.py:89: UserWarning:
The secret `HF_TOKEN` does not exist in your Colab secrets.
To authenticate with the Hugging Face Hub, create a token in your settings tab (https://huggingface.co/settings/tokens), set it as secret in y
You will be able to reuse this secret in all of your notebooks.
Please note that authentication is recommended but still optional to access public models or datasets.
  warnings.warn(
tokenizer_config.json: 100%  26.0/26.0 [00:00<00:00, 691B/s]
vocab.json: 100%  1.04M/1.04M [00:00<00:00, 8.76MB/s]
merges.txt: 100%  456k/456k [00:00<00:00, 10.5MB/s]
tokenizer.json: 100%  1.36M/1.36M [00:00<00:00, 15.1MB/s]
config.json: 100%  666/666 [00:00<00:00, 8.89kB/s]
/usr/local/lib/python3.10/dist-packages/transformers/tokenization_utils_base.py:1601: FutureWarning: `clean_up_tokenization_spaces` was not se
  warnings.warn(
```

```
model = GPT2LMHeadModel.from_pretrained("gpt2-large", pad_token_id=tokenizer.eos_token_id)
```

```
model.safetensors: 100%  3.25G/3.25G [00:33<00:00, 91.7MB/s]
generation_config.json: 100%  124/124 [00:00<00:00, 5.67kB/s]
```

Proses Tokenisasi

Pada praktikum ini karena kita menggunakan model yang telah di latih sebelumnya, kita perlu mendefinisikan dua kompenen sebelum dapat membuat artikel secara otomatis yaitu tipe tokenizer dan model yang akan digunakan. Proses tokenisasi pada dasarnya adalah pemisahan frasa, kalimat, paragraf, atau seluruh dokumen teks menjadi unit yang lebih kecil, seperti kata atau istilah individual. Masing-masing unit yang lebih kecil ini disebut token. Dalam tokenization, unit yang lebih kecil dibuat dengan menempatkan batas kata. Batas kata adalah titik akhir dari sebuah kata dan awal dari kata berikutnya. Token ini dianggap sebagai langkah pertama untuk proses stemming dan lemmatization

```
blog_title = "Indonesia Capital City"

input = tokenizer.encode(blog_title, return_tensors='pt')

input
```

```
tensor([[5497, 1952,  544, 9747, 2254]])
```

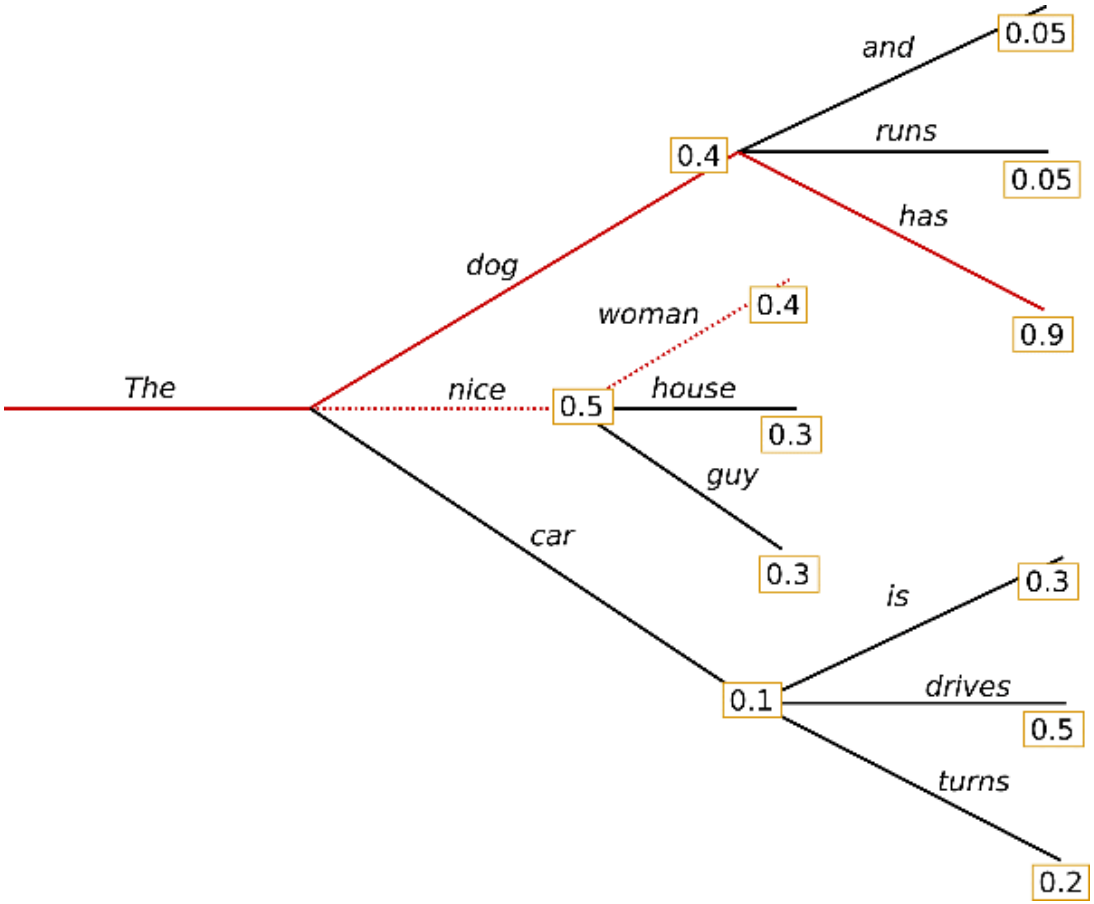
Let's Generate!

Pada tahapan ini Anda sudah siap untuk membuat model pertama Anda

Beam Search

Pencarian dengan metode beam search mengurangi risiko kehilangan urutan kata probabilitas tinggi yang tersembunyi dengan menjaga jumlah hipotesis yang paling mungkin pada setiap langkah waktu dan akhirnya memilih hipotesis yang memiliki probabilitas tertinggi secara keseluruhan. Dengan pendekatan ini, algoritma beam search dapat mengatasi kekurangan yang dimiliki oleh algoritma seperti greedy search

Berikut ilustrasi dari algoritma beam search




```
output = model.generate(input, max_length=500, num_beams=5, no_repeat_ngram_size=2, early_stopping=True)

print(tokenizer.decode(output[0], skip_special_tokens=True))
```

```
Indonesia Capital City

KUALA LUMPUR (Reuters) - Indonesian President Joko "Jokowi" Widodo said on Friday he would not be pressured into changing the country's consti

"I will not change the constitution. I am not going to change it," Joo said in an interview with Reuters in his office in the capital, Kota Ki
```