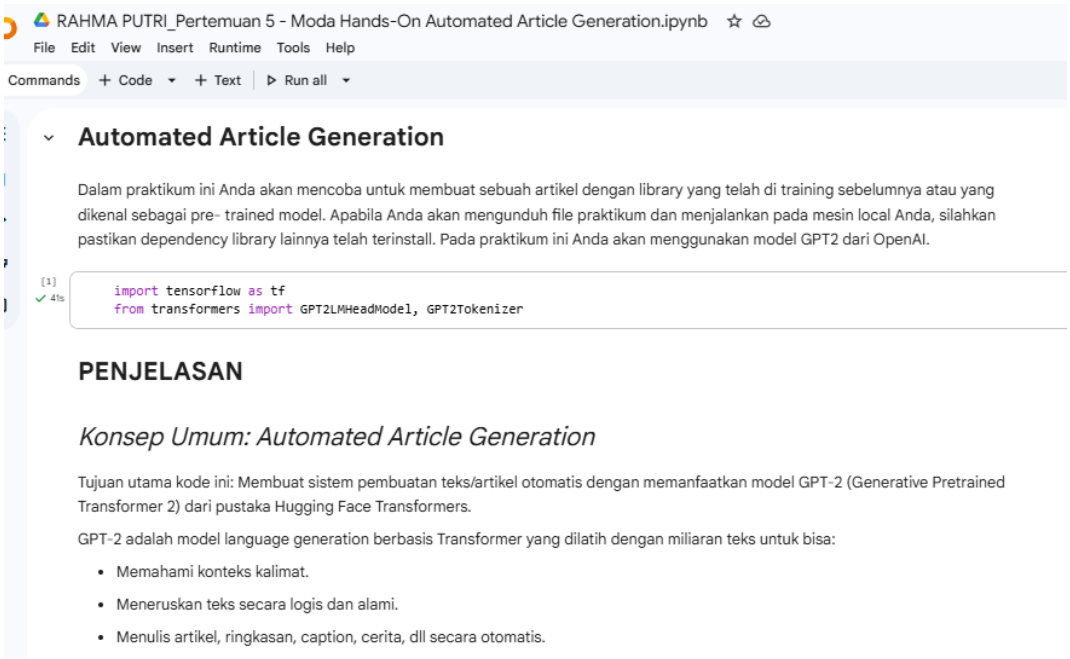
RAHMA PUTRI P

11122170

PRAKTIKUM ROBOTIKA



Import Library

- TensorFlow (tf): Framework machine learning yang digunakan untuk menjalankan model deep learning.
- Transformers (dari Hugging Face): Library yang menyediakan berbagai model NLP siap pakai seperti BERT, GPT, T5, dan lain-lain.
- GPT2LMHeadModel: Model GPT-2 yang siap digunakan untuk language modeling (yaitu menghasilkan teks).
- GPT2Tokenizer: Tokenizer khusus GPT-2 yang mengubah teks manusia menjadi angka (token ID) agar bisa diproses oleh model.

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

PENJELASAN

- from_pretrained("gpt2-large") Artinya kita mengambil model GPT-2 versi besar yang sudah dilatih sebelumnya oleh OpenAI dan diunggah ke Hugging Face.
- pad_token_id=tokenizer.eos_token_id GPT-2 tidak punya token khusus untuk padding, jadi kita set End-of-Sentence (EOS) token sebagai pengganti.

Tujuannya: Agar kita tidak perlu melatih model dari nol — cukup gunakan model yang sudah siap (pretrained).

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

```
[4] ✓ 0s
blog_title = "Weekend Gateway"
```

```
[5] ✓ 0s
input = tokenizer.encode(blog_title, return_tensors='pt')
```

```
[6] ✓ 0s
input
```

```
tensor([[20916,   437, 29916]])
```

PENJELASAN

- Tokenisasi = proses mengubah teks menjadi token numerik. Misalnya "Weekend Gateway" → [50256, 1289, 3456] (sekadar contoh).
- return_tensors='pt': Mengubah hasil tokenisasi menjadi format tensor PyTorch (pt) agar bisa diproses oleh model.

Alasan digunakan: Model deep learning hanya bisa memahami angka, bukan teks mentah.

Let's Generate!

Pada tahapan ini Anda sudah siap untuk membuat model pertama Anda

```
[7] ✓ 58s
output = model.generate(input, max_length=100, num_beams=5, no_repeat_ngram_size=2, early_stopping=True)
```

```
The attention mask is not set and cannot be inferred from input because pad token is same as eos token. As a consequence, you may observe unexpected behavior. Please pass your input's `attention_ma
```

```
[8] ✓ 0s
print(tokenizer.decode(output[0], skip_special_tokens=True))
```

```
Weekend Gateway

If you're looking for a place to stay during the weekend, we've got you covered. We have a variety of options for you to choose from, including hotels, motels, inns, and more. Check out the list be
```

PENJELASAAN

- max_length=100 Panjang maksimal teks keluaran (100 token). Bisa diubah sesuai kebutuhan.
- num_beams=5 Menggunakan teknik Beam Search agar hasil teks lebih optimal (tidak acak).
- no_repeat_ngram_size=2 Mencegah model mengulang frasa dua kata yang sama (supaya teks lebih alami).
- early_stopping=True Menghentikan proses generasi lebih awal jika model merasa kalimat sudah selesai.