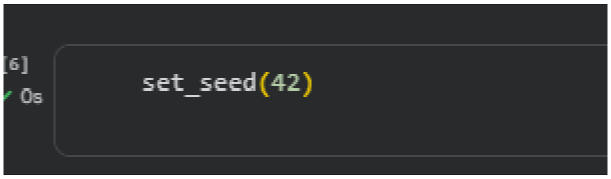
UE23CS342BA4 - Generative AI and its Applications

Handson - unit 1 Date: 22/01/2026

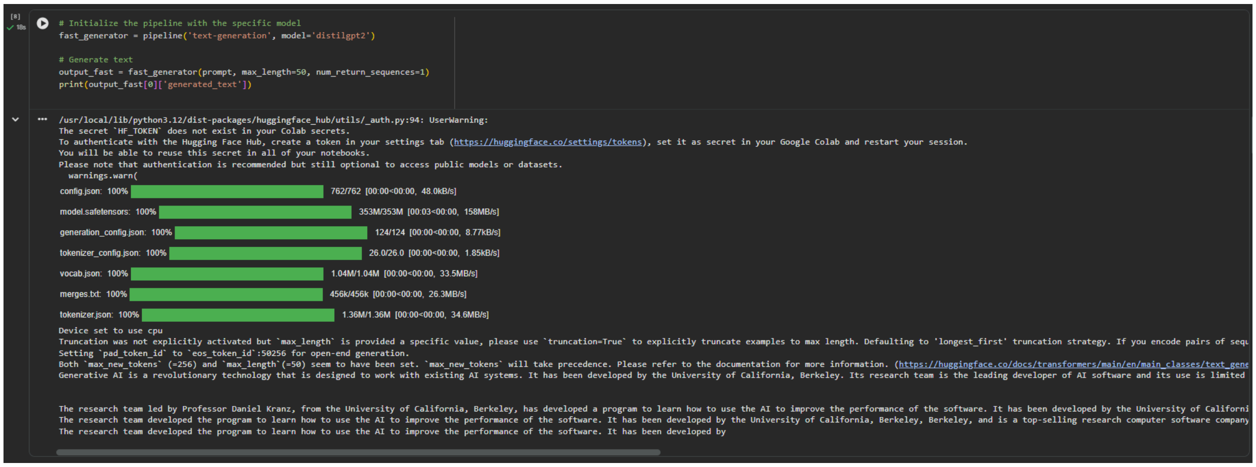
Name: Bhuvi Prashanth	SRN: PES2UG23CS130	Section: B

seed

- ***Setting different seed values***

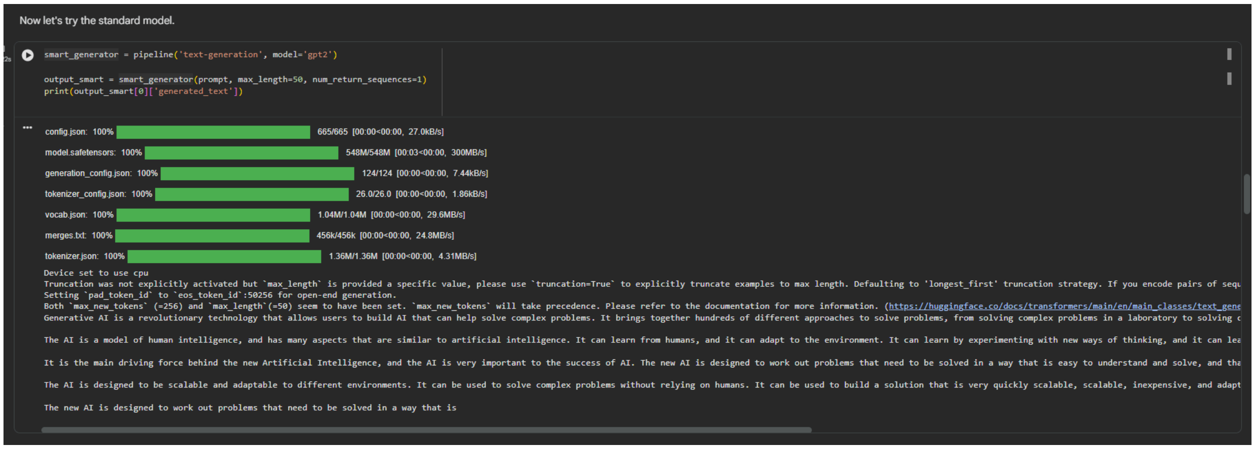


```
set_seed(42)
```



```
# Initialize the pipeline with the specific model
fast_generator = pipeline('text-generation', model='distilgpt2')

# Generate text
output_fast = fast_generator(prompt, max_length=50, num_return_sequences=1)
print(output_fast[0]['generated_text'])
```

```
/usr/local/lib/python3.12/dist-packages/huggingface_hub/utils/_auth.py:94: 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 your Google Colab and restart your session.
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(
config.json: 100% 762/762 [00:00<00:00, 48.0kB/s]
model.safetensors: 100% 353M/353M [00:03<00:00, 158MB/s]
generation_config.json: 100% 124/124 [00:00<00:00, 8.77kB/s]
tokenizer_config.json: 100% 26.0/26.0 [00:00<00:00, 1.85kB/s]
vocab.json: 100% 1.04M/1.04M [00:00<00:00, 33.5MB/s]
merges.txt: 100% 456k/456k [00:00<00:00, 26.3MB/s]
tokenizer.json: 100% 1.36M/1.36M [00:00<00:00, 34.6MB/s]
Device set to use cpu
Truncation was not explicitly activated but `max_length` is provided a specific value, please use `truncation=True` to explicitly truncate examples to max length. Defaulting to 'longest_first' truncation strategy. If you encode pairs of sequ
Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.
Both `max_new_tokens` (=256) and `max_length`(=50) seem to have been set. `max_new_tokens` will take precedence. Please refer to the documentation for more information. (https://huggingface.co/docs/transformers/main/en/main_classes/text_gene
Generative AI is a revolutionary technology that is designed to work with existing AI systems. It has been developed by the University of California, Berkeley. Its research team is the leading developer of AI software and its use is limited


The research team led by Professor Daniel Kranz, from the University of California, Berkeley, has developed a program to learn how to use the AI to improve the performance of the software. It has been developed by the University of Californi
The research team developed the program to learn how to use the AI to improve the performance of the software. It has been developed by the University of California, Berkeley, Berkeley, and is a top-selling research computer software company
The research team developed the program to learn how to use the AI to improve the performance of the software. It has been developed by
```

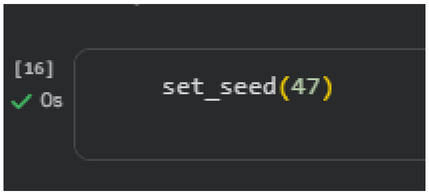


Now let's try the standard model.

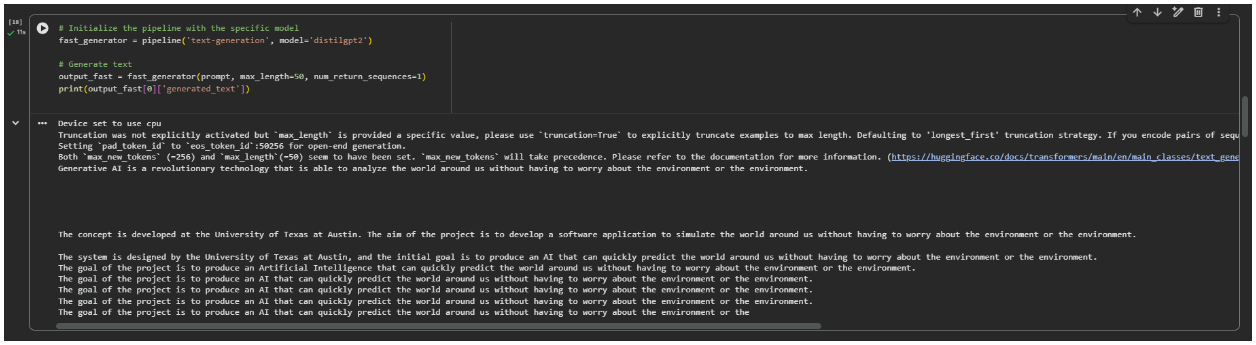
```
smart_generator = pipeline('text-generation', model='gpt2')

output_smart = smart_generator(prompt, max_length=50, num_return_sequences=1)
print(output_smart[0]['generated_text'])
```

```
config.json: 100% 665/665 [00:00<00:00, 27.0kB/s]
model.safetensors: 100% 548M/548M [00:03<00:00, 300MB/s]
generation_config.json: 100% 124/124 [00:00<00:00, 7.44kB/s]
tokenizer_config.json: 100% 26.0/26.0 [00:00<00:00, 1.86kB/s]
vocab.json: 100% 1.04M/1.04M [00:00<00:00, 29.6MB/s]
merges.txt: 100% 456k/456k [00:00<00:00, 24.8MB/s]
tokenizer.json: 100% 1.36M/1.36M [00:00<00:00, 4.31MB/s]
Device set to use cpu
Truncation was not explicitly activated but `max_length` is provided a specific value, please use `truncation=True` to explicitly truncate examples to max length. Defaulting to 'longest_first' truncation strategy. If you encode pairs of sequ
Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.
Both `max_new_tokens` (=256) and `max_length`(=50) seem to have been set. `max_new_tokens` will take precedence. Please refer to the documentation for more information. (https://huggingface.co/docs/transformers/main/en/main_classes/text_gene
Generative AI is a revolutionary technology that allows users to build AI that can help solve complex problems. It brings together hundreds of different approaches to solve problems, from solving complex problems in a laboratory to solving c

The AI is a model of human intelligence, and has many aspects that are similar to artificial intelligence. It can learn from humans, and it can adapt to the environment. It can learn by experimenting with new ways of thinking, and it can lea

It is the main driving force behind the new Artificial Intelligence, and the AI is very important to the success of AI. The new AI is designed to work out problems that need to be solved in a way that is easy to understand and solve, and tha

The AI is designed to be scalable and adaptable to different environments. It can be used to solve complex problems without relying on humans. It can be used to build a solution that is very quickly scalable, scalable, inexpensive, and adapt

The new AI is designed to work out problems that need to be solved in a way that is
```

- ***Trying to use a different seed value***

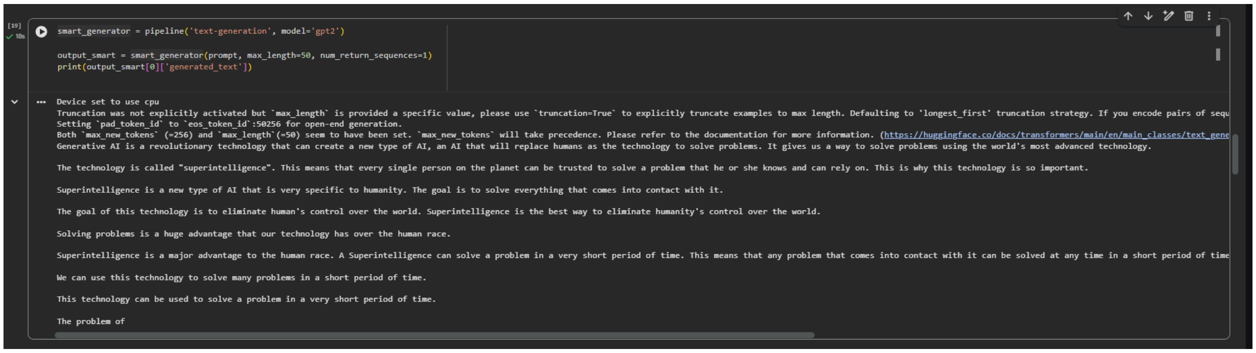


```
set_seed(47)
```



```
# Initialize the pipeline with the specific model
fast_generator = pipeline('text-generation', model='distilgpt2')

# Generate text
output_fast = fast_generator(prompt, max_length=50, num_return_sequences=1)
print(output_fast[0]['generated_text'])
```

```
Device set to use cpu
Truncation was not explicitly activated but `max_length` is provided a specific value, please use `truncation=True` to explicitly truncate examples to max length. Defaulting to 'longest_first' truncation strategy. If you encode pairs of sequ
Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.
Both `max_new_tokens` (=256) and `max_length`(=50) seem to have been set. `max_new_tokens` will take precedence. Please refer to the documentation for more information. (https://huggingface.co/docs/transformers/main/en/main_classes/text_gene
Generative AI is a revolutionary technology that is able to analyze the world around us without having to worry about the environment or the environment.




The concept is developed at the University of Texas at Austin. The aim of the project is to develop a software application to simulate the world around us without having to worry about the environment or the environment.

The system is designed by the University of Texas at Austin, and the initial goal is to produce an AI that can quickly predict the world around us without having to worry about the environment or the environment.
The goal of the project is to produce an Artificial Intelligence that can quickly predict the world around us without having to worry about the environment or the environment.
The goal of the project is to produce an AI that can quickly predict the world around us without having to worry about the environment or the environment.
The goal of the project is to produce an AI that can quickly predict the world around us without having to worry about the environment or the environment.
The goal of the project is to produce an AI that can quickly predict the world around us without having to worry about the environment or the environment.
The goal of the project is to produce an AI that can quickly predict the world around us without having to worry about the environment or the
```



```
smart_generator = pipeline('text-generation', model='gpt2')

output_smart = smart_generator(prompt, max_length=50, num_return_sequences=1)
print(output_smart[0]['generated_text'])
```

```
Device set to use cpu
Truncation was not explicitly activated but `max_length` is provided a specific value, please use `truncation=True` to explicitly truncate examples to max length. Defaulting to 'longest_first' truncation strategy. If you encode pairs of sequ
Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.
Both `max_new_tokens` (=256) and `max_length`(=50) seem to have been set. `max_new_tokens` will take precedence. Please refer to the documentation for more information. (https://huggingface.co/docs/transformers/main/en/main_classes/text_gene
Generative AI is a revolutionary technology that can create a new type of AI, an AI that will replace humans as the technology to solve problems. It gives us a way to solve problems using the world's most advanced technology.

The technology is called "superintelligence". This means that every single person on the planet can be trusted to solve a problem that he or she knows and can rely on. This is why this technology is so important.

Superintelligence is a new type of AI that is very specific to humanity. The goal is to solve everything that comes into contact with it.

The goal of this technology is to eliminate human's control over the world. Superintelligence is the best way to eliminate humanity's control over the world.

Solving problems is a huge advantage that our technology has over the human race.

Superintelligence is a major advantage to the human race. A Superintelligence can solve a problem in a very short period of time. This means that any problem that comes into contact with it can be solved at any time in a short period of time

We can use this technology to solve many problems in a short period of time.

This technology can be used to solve a problem in a very short period of time.

The problem of
```

Difference between distilled and main model outputs here
[Distillation mainly **reduces depth**, not vocabulary or embedding size.]

`distilgpt2` → **faster loading + faster generation**

`gpt2` → **slower but richer generation**

distilgpt2 output

- Shorter ideas
- More repetition
- Less long-range coherence
- Simpler phrasing

gpt2 output

- Better sentence flow
- More global coherence
- Less repetition
- Slightly more “human-like” continuation

1. Main Model (Teacher Model)

The **main model** is the original, full-capacity model trained directly on large datasets.

- **Large size** (many parameters)
- **High accuracy & generalization**
- **High compute & memory requirements**
- Used during **training, research, or cloud inference**

Analogy - A **senior expert** with complete knowledge but slower and expensive to consult every time.

2. Distilled Model (Student Model)

A **distilled model** is a smaller model trained to **imitate the behavior** of the main model using **knowledge distillation**.

- **Smaller and lighter**
- **Faster inference**
- **Lower memory and power consumption**
- Slightly **lower accuracy**, but often very close
- Ideal for **edge devices, mobile apps, real-time systems**

Analogy - A **well-trained assistant** who learned from the expert and can respond quickly.

Transformer

Encoder - knows how to write
Decoder - knows how to write

GPT-2 (Generative Pre-trained Transformer-2)

- **Decoder-only Transformer model [trained to write only ex. Generating textx]**
- Uses **masked self-attention**
- **Autoregressive**: predicts one token at a time
- Can’t do summarization

What GPT-2 Is Trained For

- Trained **only on next-token prediction**
- Learns **how to write text fluently**
- Generates:
  - Stories
  - Paragraphs
  - Code-like text
  - Conversations

How does gpt2 predict the next word?

1. Softmax layer [**Softmax layer** converts logits into probabilities]
2. Sampling
   - temperature sampling – decides the randomness of the outputs, lower the temperature,better and accurate results
   - top p aggregate chooses smallest set of tokens whose **cumulative probability ≥ p** (%)
   - top k considers only the **top k most probable tokens and** ignores low-probability tokens

3. Context Encoding
   Input text is split into **tokens → embeddings → transformer models**

Tokenization

POS-wise Explanation (Token → Tag → Meaning)

1. Transformers → NNS

- **NNS** = *Plural Noun*
- Refers to **more than one transformer**
- In context: multiple transformer models or architectures
- Role in sentence: **Subject**

2. revolutionized → VBD

- **VBD** = *Verb, Past Tense*
- Indicates an action that **already happened**
- Here: describes the impact transformers had in the past
- Role in sentence: **Main verb**

3. NLP → NNP

- **NNP** = *Proper Noun, Singular*
- Used for **specific named entities**
- “NLP” (Natural Language Processing) is treated as a proper noun
- Role in sentence: **Object of the verb**

4. . → .

- **.** = *Sentence-ending punctuation*
- Marks the end of the sentence
- Role in sentence: **Punctuation**

Named Entity Recognition (NER)

What is NER?

Named Entity Recognition (NER) is an NLP task used to identify and classify named entities in text into predefined categories such as: Person, Organization, Location, Date, Miscellaneous, etc.

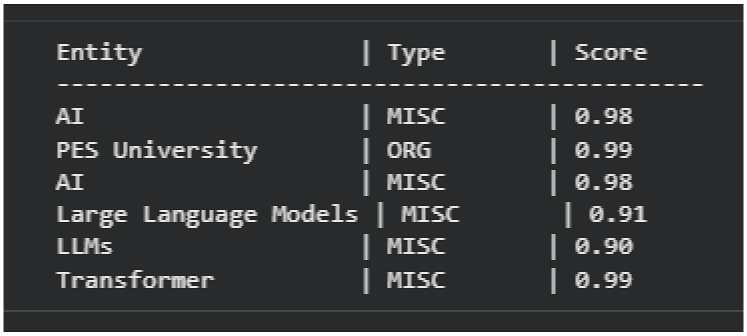
NER helps in extracting structured information from unstructured text.

Applications:

- **Question answering**
- **Resume parsing**
- **Search engines**
- **Document analysis**

BERT for NER

- **Bidirectional context (looks at left & right words)**
- **Strong contextual understanding**
- **Captures meaning of entities based on context**
- **Encoder-based → ideal for classification tasks**



```
Entity               | Type       | Score
-------------------------------------------
AI                   | MISC       | 0.98
PES University       | ORG        | 0.99
AI                   | MISC       | 0.98
Large Language Models | MISC       | 0.91
LLMs                 | MISC       | 0.90
Transformer          | MISC       | 0.99
```

MISC – miscellaneous; ORG – organization

Score	**Meaning**
≥ 0.90	**Very high confidence**
0.75 – 0.89	**Good confidence**
< 0.75	**Less reliable**

BERT vs BART

BERT = understand
BART = understand + generate