Member-only story

Simple abstractive text summarization with pretrained T5 — Text-To-Text Transfer Transformer

Text summarization with T5 Text-to-Text transformer

Ramsri Goutham · Follow
Published in Towards Data Science
3 min read · Apr 16, 2020

Listen Share

Image from Pixabay and Stylized by AiArtist Chrome Plugin

T5 is a new transformer model from Google that is trained in an end-to-end manner with text as input and modified text as output. You can read more about it here.

It achieves state-of-the-art results on multiple NLP tasks like summarization, question answering, machine translation etc using a text-to-text transformer trained on a large text corpus.

Today we will see how we can use huggingface's transformers library to summarize any given text. T5 is an abstractive summarization algorithm. It means that it will rewrite sentences when necessary than just picking up sentences directly from the original text.

Install these libraries in your jupyter notebook or conda environment before you begin :

```
!pip install transformers==2.8.0
!pip install torch==1.4.0
```

Text input:

```
The US has "passed the peak" on new coronavirus cases, President
Donald Trump said and predicted that some states would reopen this
month.The US has over 637,000 confirmed Covid-19 cases and over 30,826
deaths, the highest for any country in the world.At the daily White
House coronavirus briefing on Wednesday, Trump said new guidelines t
reopen the country would be announced on Thursday after he speaks to
governors."We'll be the comeback kids, all of us," he said. "We want
to get our country back."The Trump administration has previously fixed
May 1 as a possible date to reopen the world's largest economy, but
the president said some states may be able to return to normalcy
earlier than that.
```

Summary from T5:

```
The us has over 637,000 confirmed Covid-19 cases and over 30,826
deaths. President Donald Trump predicts some states will reopen the
country in april, he said. "we'll be the comeback kids, all of us,"
the president says.
```

Analyzing the output:

If you see the algorithm has intelligently summarized with mentioning **April** although it was never mentioned in the original story**.** It also replaced **he** with **the president.** Cut short and removed the extra information after the comma in the sentence “over 30,826 deaths ….”

The code :

```
import torch
import json
from transformers import T5Tokenizer, T5ForConditionalGeneration, T5Config

model = T5ForConditionalGeneration.from_pretrained('t5-small')
tokenizer = T5Tokenizer.from_pretrained('t5-small')
device = torch.device('cpu')

text ="""
The US has "passed the peak" on new coronavirus cases, President Donald Trump said and predicted t

The US has over 637,000 confirmed Covid-19 cases and over 30,826 deaths, the highest for any

At the daily White House coronavirus briefing on Wednesday, Trump said new guidelines to reopen th

"We'll be the comeback kids, all of us," he said. "We want to get our country back."

The Trump administration has previously fixed May 1 as a possible date to reopen the world's large
"""


preprocess_text = text.strip().replace("\n","")
t5_prepared_Text = "summarize: "+preprocess_text
print ("original text preprocessed: \n", preprocess_text)

tokenized_text = tokenizer.encode(t5_prepared_Text, return_tensors="pt").to(device)


# summmarize
summary_ids = model.generate(tokenized_text,
                                    num_beams=4,
                                    no_repeat_ngram_size=2,
                                    min_length=30,
                                    max_length=100,
                                    early_stopping=True)

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

print ("\n\nSummarized text: \n",output)

# Summarized output from above ::::::::::
# the us has over 637,000 confirmed Covid-19 cases and over 30,826 deaths.
# president Donald Trump predicts some states will reopen the country in april, he said.
# "we'll be the comeback kids, all of us," the president says.
```

T5_transformers_summarization.py hosted with ❤ by GitHub view raw

The same code inline as some users have reported that the Github Gist wasn't loading :

```
import torch
import json
from transformers import T5Tokenizer, T5ForConditionalGeneration,
T5Config

model = T5ForConditionalGeneration.from_pretrained('t5-small')
tokenizer = T5Tokenizer.from_pretrained('t5-small')
device = torch.device('cpu')

text ="""
The US has "passed the peak" on new coronavirus cases, President
Donald Trump said and predicted that some states would reopen this
month.

The US has over 637,000 confirmed Covid-19 cases and over 30,826
deaths, the highest for any country in the world.

At the daily White House coronavirus briefing on Wednesday, Trump said
new guidelines to reopen the country would be announced on Thursday
after he speaks to governors.

"We'll be the comeback kids, all of us," he said. "We want to get our
country back."

The Trump administration has previously fixed May 1 as a possible date
to reopen the world's largest economy, but the president said some
states may be able to return to normalcy earlier than that.
"""


preprocess_text = text.strip().replace("\n","")
t5_prepared_Text = "summarize: "+preprocess_text
print ("original text preprocessed: \n", preprocess_text)

tokenized_text = tokenizer.encode(t5_prepared_Text,
return_tensors="pt").to(device)


# summmarize
summary_ids = model.generate(tokenized_text,
                                    num_beams=4,
                                    no_repeat_ngram_size=2,
                                    min_length=30,
                                    max_length=100,
```

```
                                    early_stopping=True)

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

print ("\n\nSummarized text: \n",output)

# Summarized output from above ::::::::::
# the us has over 637,000 confirmed Covid-19 cases and over 30,826
deaths.
# president Donald Trump predicts some states will reopen the country
in april, he said.
# "we'll be the comeback kids, all of us," the president says.
```

The summarized output as mentioned above is -

```
The us has over 637,000 confirmed Covid-19 cases and over 30,826
deaths. President Donald Trump predicts some states will reopen the
country in april, he said. "we'll be the comeback kids, all of us,"
the president says.
```

The key point to note in the code above is that we prepend our text with **"summarize: "**

corresponding string eg: **"translate English to German: "** for translation task.

Happy NLP exploration and if you loved the content, feel free to find me on Twitter.

If you want to learn modern NLP using transformers, check out my course Question generation using NLP

Machine Learning | Data Science | Artificial Intelligence | NLP | Naturallanguageprocessing

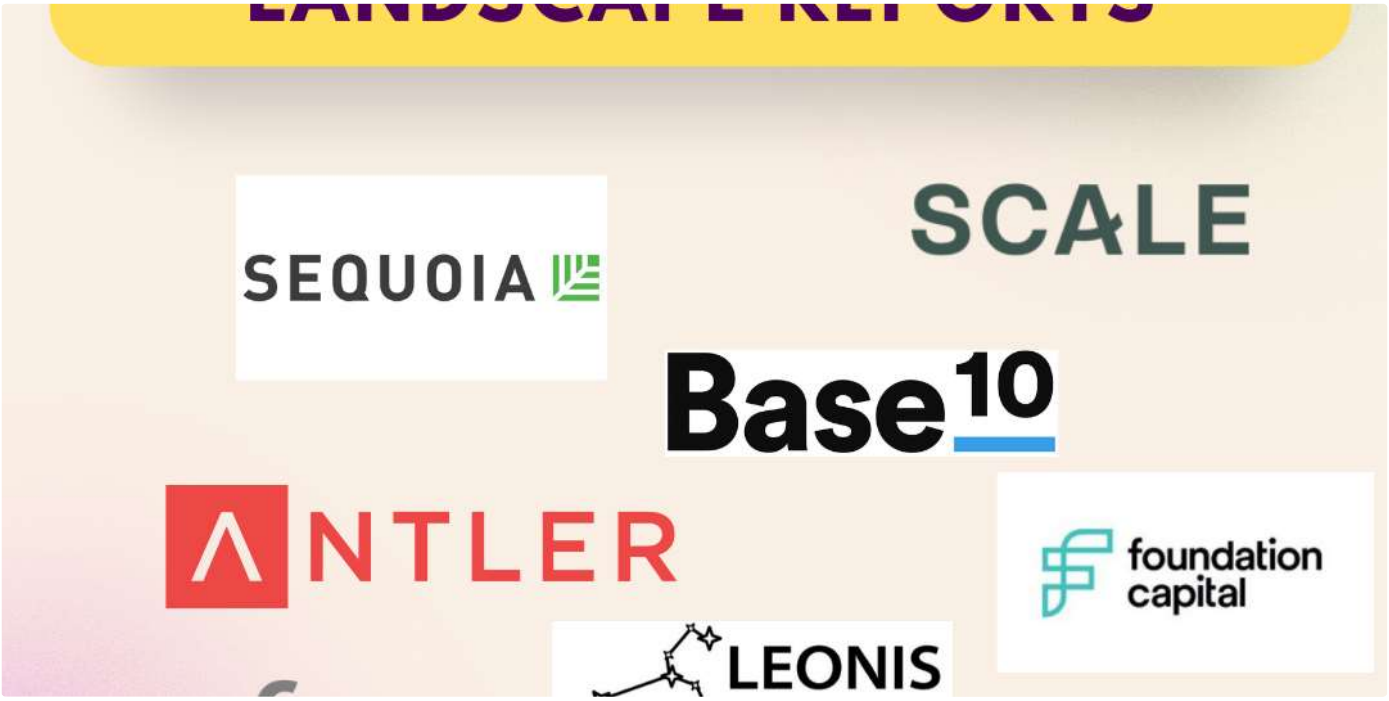
Follow

Written by Ramsri Goutham

1.4K Followers · Writer for Towards Data Science

Teaches NLP courses at https://www.learnnlp.academy/ ◆ Building AI SaaS Apps: https://questgen.ai/ and https://supermeme.ai/

More from Ramsri Goutham and Towards Data Science



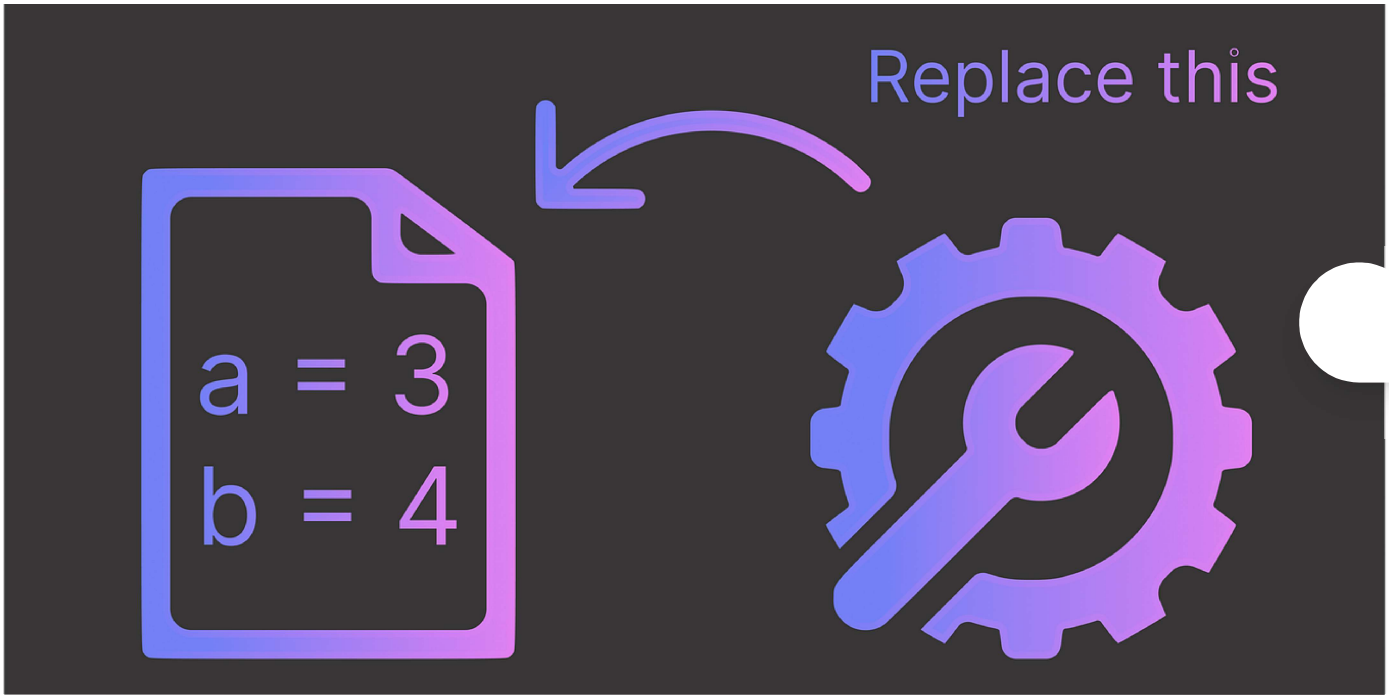
Ramsri Goutham

The landscape of generative AI landscape reports

9 VCs who published generative AI reports

· 2 min read · Jan 15

165 1



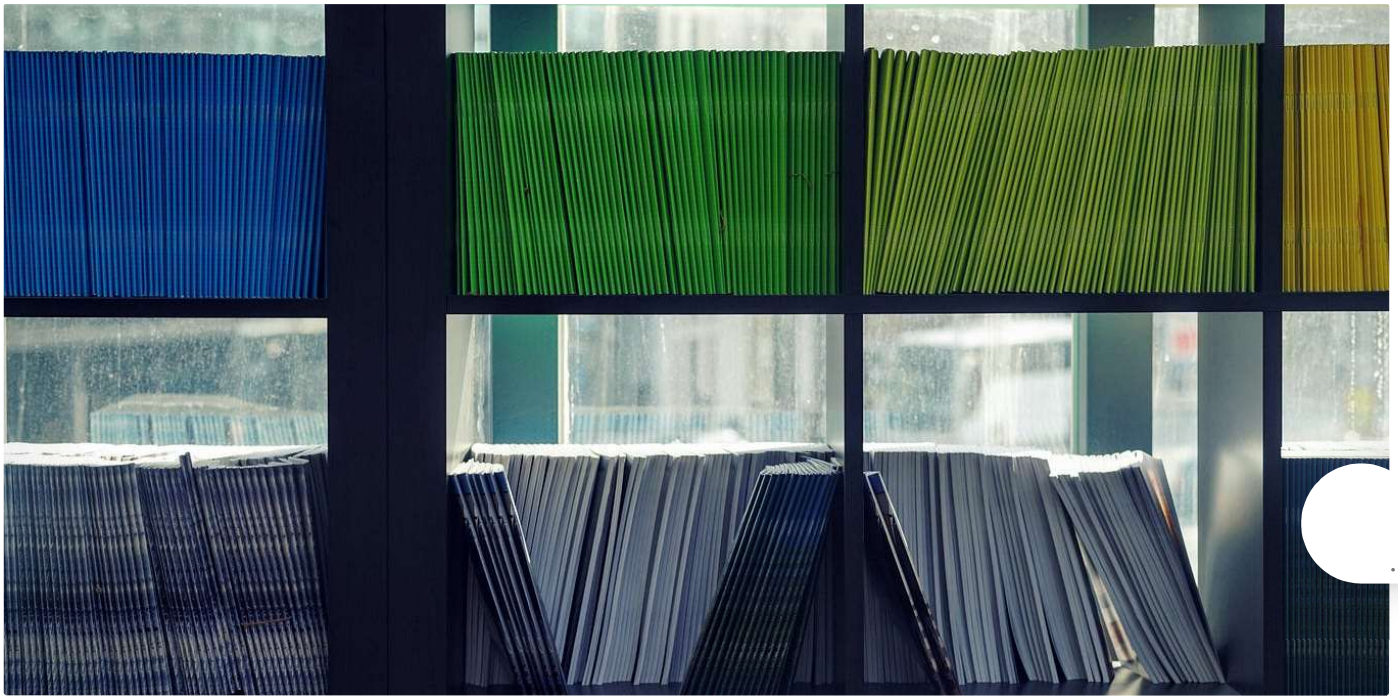
Khuyen Tran in Towards Data Science

Stop Hard Coding in a Data Science Project—Use Config Files Instead

And How to Efficiently Interact with Config Files in Python

· 6 min read · May 26

1.8K 20



Jacob Marks, Ph.D. in Towards Data Science

How I Turned My Company's Docs into a Searchable Database with OpenAI

And how you can do the same with your docs

15 min read · Apr 25

Ramsri Goutham

5 tools to build full-stack apps in Pure Python

No web development experience or front-end knowledge is necessary!

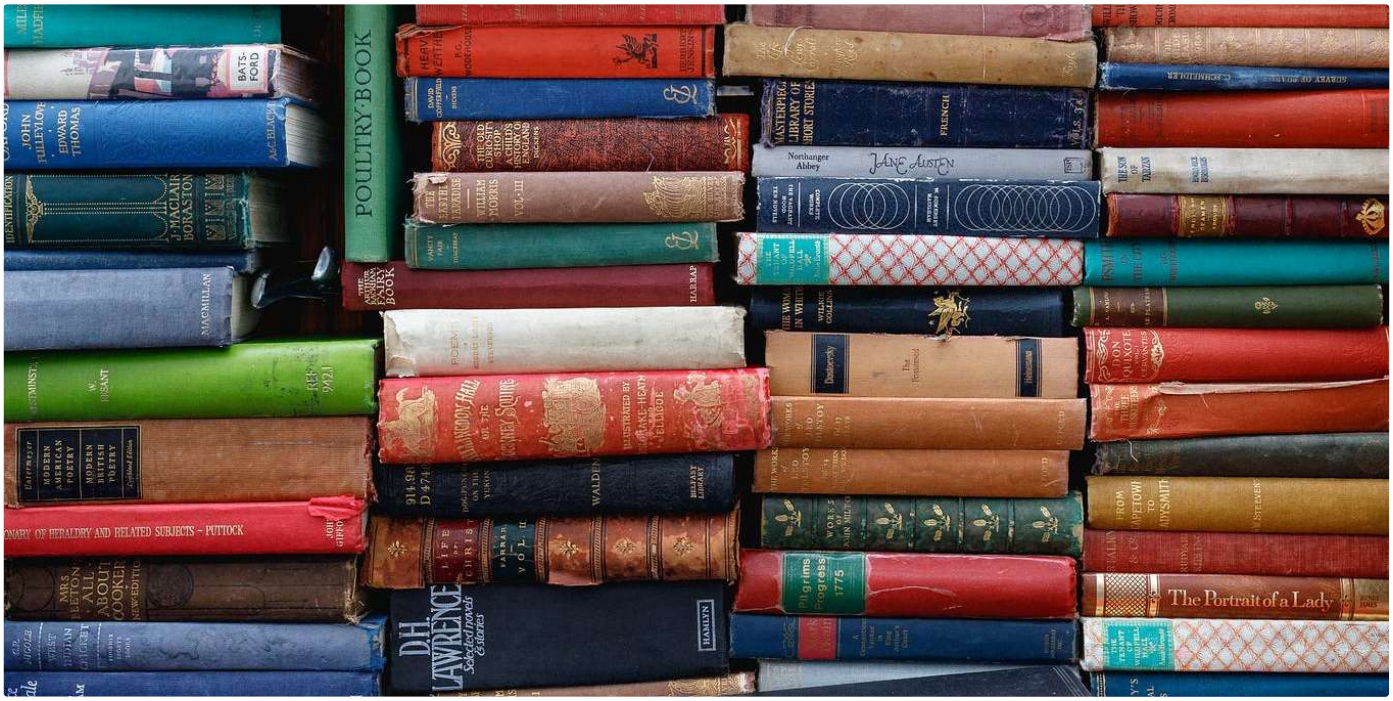
✦ · 3 min read · Jan 23

26 1

See all from Ramsri Goutham

See all from Towards Data Science

Recommended from Medium



Jay Peterman in Towards Data Science

Make a Text Summarizer with GPT-3

Quick tutorial using Python, OpenAI's GPT-3, and Streamlit

✦ · 11 min read · Jan 24

167 1

Arun Jagota in Towards Data Science

Thread Summarization Using NLP

Extractive summarization using POS tagging, NER, and sentiment analysis

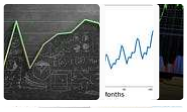
✦ · 13 min read · Jan 8

44 1

Lists

Natural Language Processing

363 stories · 17 saves



Predictive Modeling w/ Python

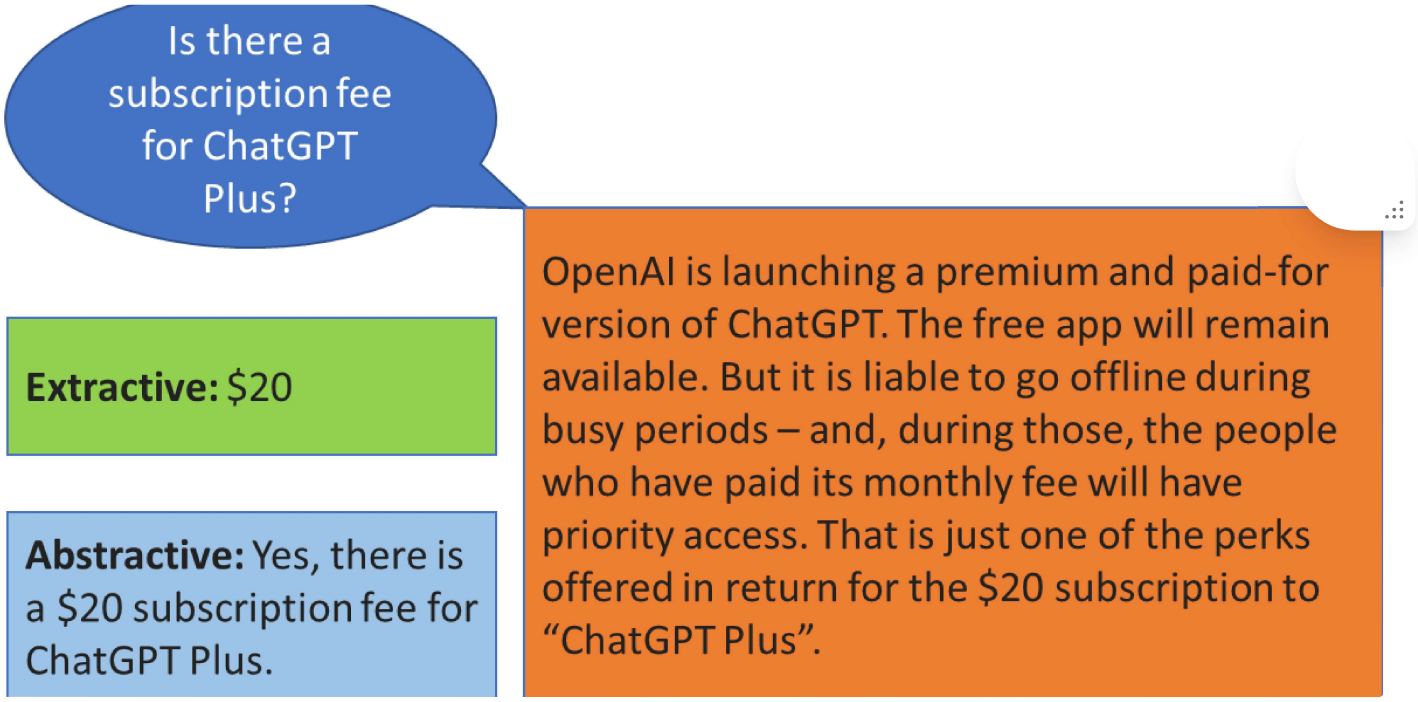
18 stories · 39 saves

Practical Guides to Machine Learning

10 stories · 57 saves

ChatGPT prompts

17 stories · 27 saves



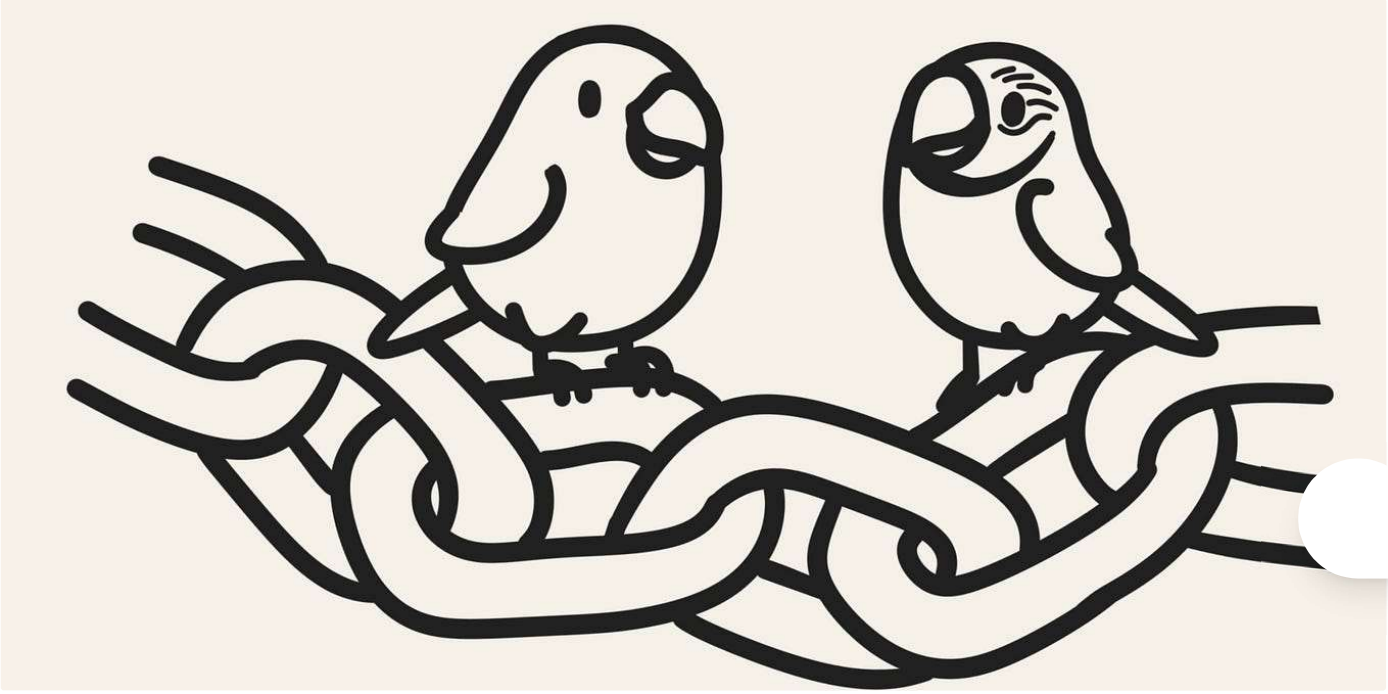
Skanda Vivek in Towards Data Science

Extractive vs Generative Q&A—Which is better for your business?

The arrival of ChatGPT hints at a new era of search engines, this tutorial dives into the 2 basic types of AI based question answering

· 6 min read · Feb 6

57



Ruben Winastwan in Towards Data Science

Semantic Textual Similarity with BERT

How to use BERT to calculate the semantic similarity between two texts

· 11 min read · Feb 15

Wei-Meng Lee in Level Up Coding

Training Your Own LLM using privateGPT

Learn how to train your own language model without exposing your private data to the provider

· 8 min read · May 19

1.1K 9