Using Scikit-LLM for text similarity

Danial Khilji · Follow

2 min read · Jun 6

AI generated image

Since the launch of ChatGPT, its recognized as a huge leap of advancement in the field of AI not just because its powerful but also due to its huge reach to people, and as a result there has been a huge surge in new GPT integrated products in the market.

Scikit-LLM is another new product that provides seamless integration between scikit-learn and GPT models to give better GPT models access to data scientists. This package will help everyone speed up their projects and explore new more advanced GPT models.

In this article, I am going to show you how easy it is to use GPT model to calculate text similarity. I will use:

- Text to calculate GPT embeddings
- and then calculate their cosine similarity

Lets dig into coding without wasting more time!

Install the package:

```
!pip isntall scikit-llm
```

Load the important libraries:

```
from skllm.preprocessing import GPTVectorizer
from skllm.config import SKLLMConfig
import os

SKLLMConfig.set_openai_key(os.getenv("API_KEY"))
SKLLMConfig.set_openai_org(os.getenv("ORG_NAME"))
```

I am going to use GPTVectorizer to calculate text embeddings. SKLLMConfig is used to call OpenAI API credentials. If you save your credentials in a .env file (called environment variables), then you need to os.getenv to call the variables as I had done above. Otherwise you can also simply write your API credentials like:

```
SKLLMConfig.set_openai_key("123abc")
SKLLMConfig.set_openai_org("123abc")
```

Create a vector object:

```
model = GPTVectorizer()
```

Use .fit method to calculate vectors:

```
vectors = model.fit_transform(["how old are you?", "what is your age?"])
```

Reshape the vectors:

```
vector_1 = np.array(vectors[0]).reshape(1, -1)
vector_2 = np.array(vectors[1]).reshape(1, -1)
```

Use the cosine_similarity function to calculate the similarity:

```
cosine_similarity(vector_1, vector_2)
```

```
0.94797838
```

The above code is just a simple instructions on how to calculate similarity but ofcourse you will need to evaluate the performance based on your current best model results to give you a better idea. I believe this package capabilities will be improved further and it will unleash huge potential in the field of NLP.

Share your experience and feedback. Please like and subscribe!

Artificial Intelligence | NLP | Data Science | Python | Scikit Learn

Characteristics of Word Embeddings
✦ · 11 min read · Sep 4, 2021

The Word2vec Hyperparameters
✦ · 6 min read · Sep 3, 2021

The Word2ve
✦ · 15 min rea

View list

Written by Danial Khilji

Follow

43 Followers

Data Scientist @choreograph (WPP company)

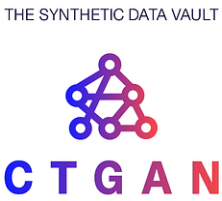
More from Danial Khilji

Danial Khilji

Cosine similarity using GPT models

Photo by Alec Favale on Unsplash

3 min read · Sep 19



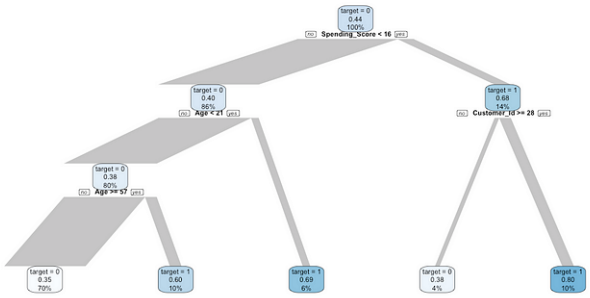
Danial Khilji

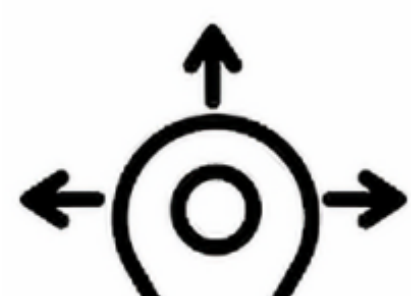
Generating tabular data using CTGAN

There's tons of reasons in the field of Data Science when you need to have new sample...

2 min read · Jul 14, 2022

22 26





Danial Khilji

Exploring auto EDA packages in R (part 1)

✨ · 5 min read · May 20, 2022

9 1

Danial Khilji

Icons in Tableau

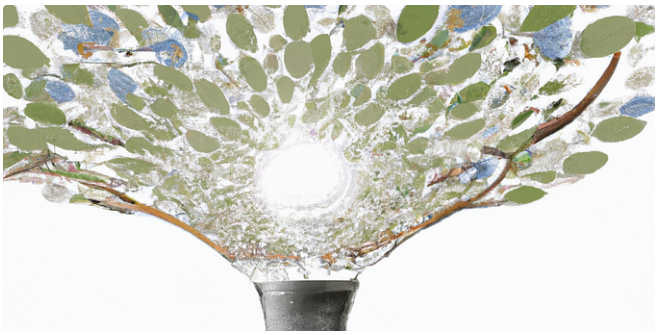
Ever wondered how people make such cool visualizations in Tableau? and using icons?

4 min read · Sep 15, 2022

8

See all from Danial Khilji

Recommended from Medium



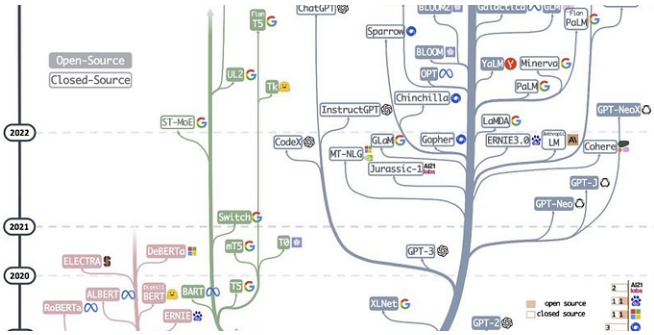
Sirsh Amarteifio

Cluster chatter: HDBSCAN + LLM

HDBSCAN is a density based (hierarchical) clustering algorithm. Clustering algorithms...

5 min read · Jun 13

3



Haifeng Li

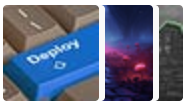
A Tutorial on LLM

Generative artificial intelligence (GenAI), especially ChatGPT, captures everyone's...

15 min read · Sep 14

372

Lists



Predictive Modeling w/ Python

20 stories · 452 saves

Natural Language Processing

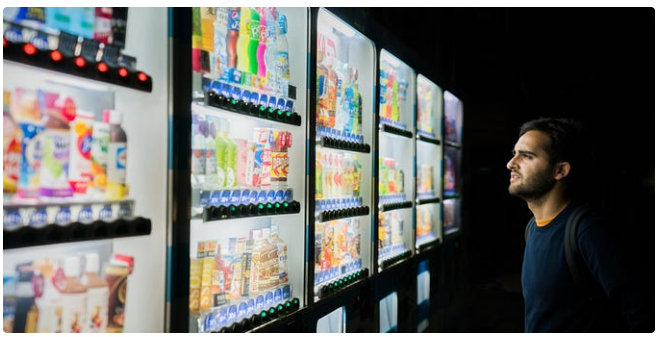
669 stories · 283 saves

ChatGPT

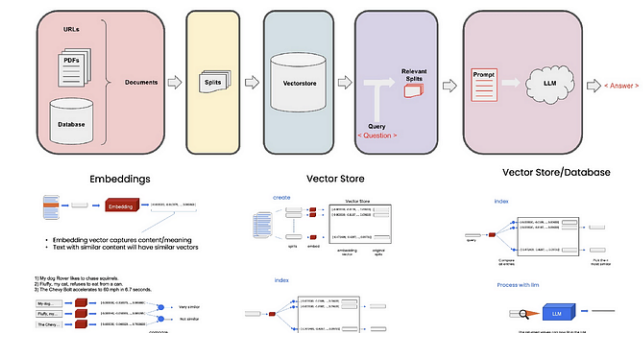
21 stories · 179 saves

Coding & Development

11 stories · 200 saves



Ovbude Ehi



TeeTracker

Recommendation Engines

Practical Techniques: Content-Based Filtering, Collaborative Filtering and...

11 min read · May 8

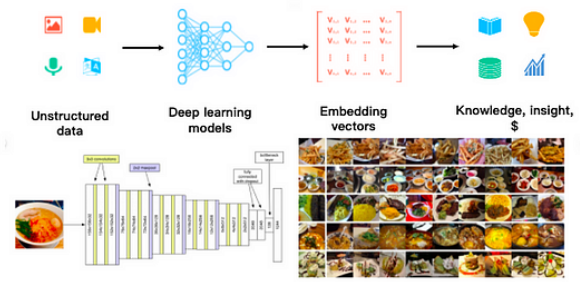
5

Chat with your PDF (Streamlit Demo)

Conversation with specific files

4 min read · Sep 15

56



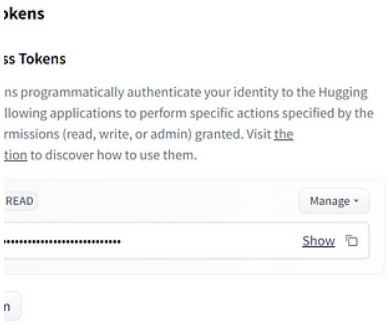
Jayita Bhattacharyya in GoPenAI

Primer on Vector Databases and Retrieval-Augmented Generation...

Vector Databases Generation (RAG) Langchain Pinecone HuggingFace Large...

9 min read · Aug 16

228 1



Ankit

Generating Summaries for Large Documents with Llama2 using...

Introduction

11 min read · Aug 28

103 3

See more recommendations