TEXT SPLITTING & EMBEDDING DEMISTYFIED

DOCUMENT EMBEDDING MADE EFFORTLESS

HTTPS://GITHUB.COM/INSIGHTBUILDER

YouTube IN

Search

INSIGHT BUILDER

Kamalraj M M

@insightbuilder 94 subscribers 95 videos

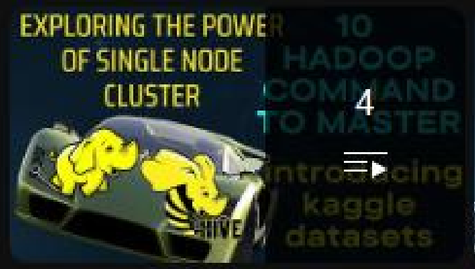
More about this channel >

Subscribe

HOME VIDEOS PLAYLISTS COMMUNITY CHANNELS ABOUT

Created playlists

Sort by



Take Your Big Data Cluster for a Spin with These Essential Data...

Updated today

View full playlist

Build and Own Your Big Data Cluster : Work on ETL, ELT, BI...

Updated today

View full playlist

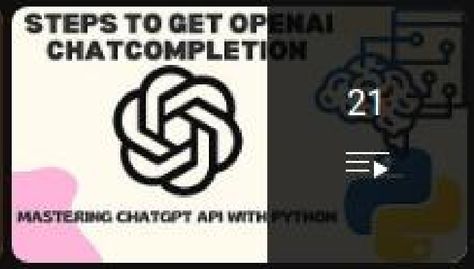
Exploring Top Challenges That Linux Solves: A Comprehensive...

Updated 3 days ago

View full playlist

SQL Mastery : Absolute Basics to Complex Data Transformations

View full playlist



Learn about AI Language Models and Reinforcement Learning

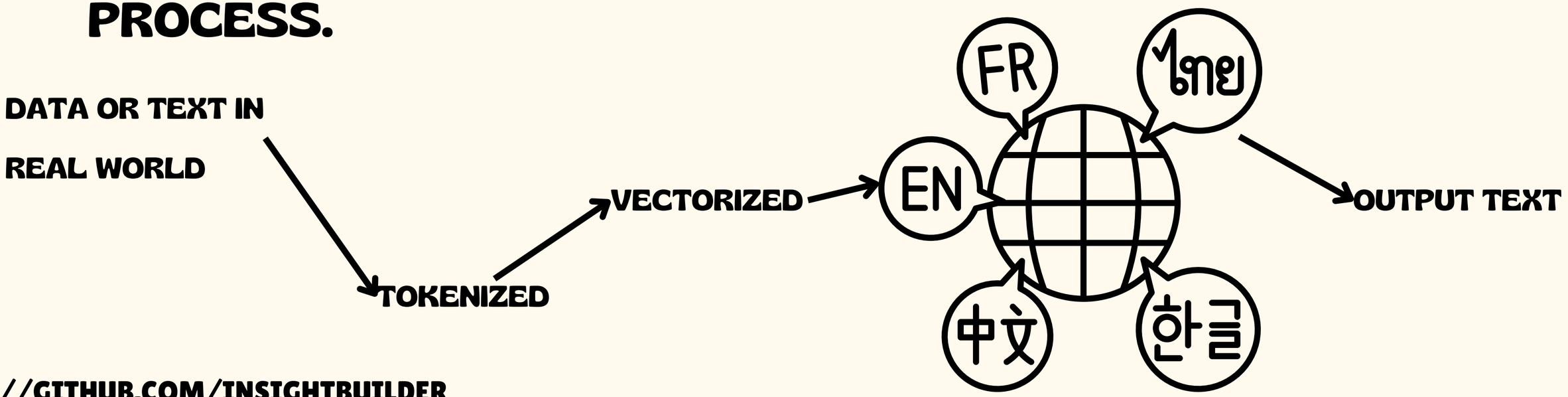
Updated yesterday

View full playlist

HTTPS://GITHUB.COM/INSIGHTBUILDER

WHY USE TEXT SPLITTER

- LLM LIKE ALPACA-LORA OR GPT-4 HAVE A LIMITATION ON THE NUMBER OF TOKEN THEY CAN PROCESS
- DOCUMENTS THAT IS USED IN BUSINESS OR DAILY ACTIVITY HAS MUCH MORE DATA THAN LLM CAN PROCESS.



WHAT IS EMBEDDING

LLM CAN READ NUMBERS IN FORM OF VECTORS. VECTORS THAT ARE READ BY EACH LLM IS DEFFERENT

DATA OR TEXT IN REAL WORLD

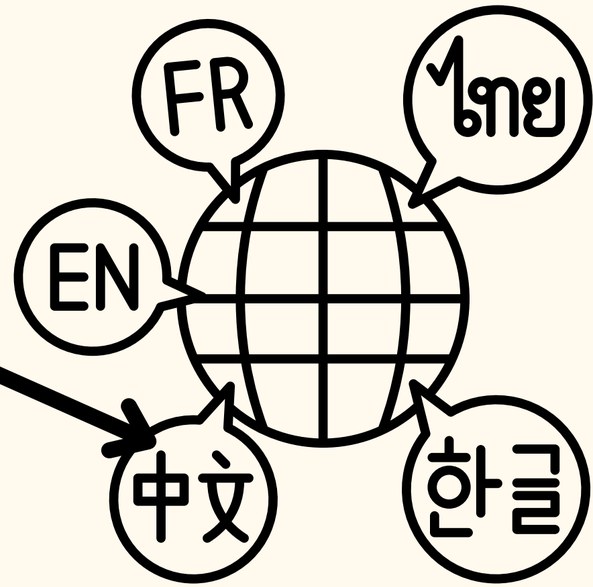
IN DATA TEXT OR REAL WORLD

[8, 5, 2, 1, 87, 97]

[9, 15, 22, 1,0 7, 90]

[119, 59, 68, 10, 2, 51]

[75, 62, 8, 0, 23, 54]



TOKENIZER IS ALSO A MODEL

LANGUAGE WE USE CAN BE STATISTICALLY MODELED. THE WAY SPLITTING THE TEXT IS MODELED BY MANY DIFFERENT ALGORITHMS.

Model	BPE	WordPiece	Unigram
Training	Starts from a small vocabulary and learns rules to merge tokens	Starts from a small vocabulary and learns rules to merge tokens	Starts from a large vocabulary and learns rules to remove tokens
Training step	Merges the tokens corresponding to the most common pair	Merges the tokens corresponding to the pair with the best score based on the frequency of the pair, privileging pairs where each individual token is less frequent	Removes all the tokens in the vocabulary that will minimize the loss computed on the whole corpus
Learns	Merge rules and a vocabulary	Just a vocabulary	A vocabulary with a score for each token
Encoding	Splits a word into characters and applies the merges learned during training	Finds the longest subword starting from the beginning that is in the vocabulary, then does the same for the rest of the word	Finds the most likely split into tokens, using the scores learned during training

THE IMPLEMENTATION OF THESE ALGORITHMS CAN BE FOUND AT HUGGING FACE COURSE.

HTTPS://HUGGINGFACE.CO/COURSE/CHAPTER6/4?FW=PT

THESE ALGORITHMS ARE PART OF THE MODEL PIPELINES

HTTPS://GITHUB.COM/INSIGHTBUILDER

HOW IS THIS LINKED WITH LANGCHAIN

LANGCHAIN LIBRARY HAS WRAPPER AROUND THE ABOVE STEPS

TEXTSPLITTER

- CHARACTER TEXT SPLITTER
- HUGGINGFACE LENGTH FUNCTION
- LATEX TEXT SPLITTER
- MARKDOWN TEXT SPLITTER
- NLTK TEXT SPLITTER
- PYTHON CODE TEXT SPLITTER
- RECURSIVECHARATERTEXTSPLITTER
- SPACY TEXT SPLITTER
- TIKTOKEN (OPENAI) LENGTH FUNCTION
- TIKTOKENTEXT SPLITTER

EMBEDDING ALGO

- AZUREOPENAI
- COHERE
- FAKE EMBEDDINGS
- HUGGING FACE HUB
- INSTRUCTEMBEDDINGS
- JINA
- OPENAI
- SAGEMAKER ENDPOINT EMBEDDINGS
- SELF HOSTED EMBEDDINGS
- TENSORFLOWHUB

VECTOR STORES

- ATLASDB
- CHROMA
- DEEP LAKE
- ELASTICSEARCH
- FAISS
- MILVUS
- OPENSEARCH
- PGVECTOR
- PINECONE
- QDRANT
- REDIS
- WEAVIATE

WHERE TO CONCENTRATE???

WE WILL LOOK AT THE FOLLOWING

TEXTSPLITTER

- CHARACTER TEXT SPLITTER
- HUGGINGFACE LENGTH FUNCTION
- RECURSIVECHARATERTEXTSPLITTER

EMBEDDING ALGO

- HUGGING FACE HUB
- INSTRUCTEMBEDDINGS

VECTOR STORES

- CHROMA

WHY?

1) THEY ARE SIMPLE TO USE

2) AVAILABLE LOCALLY

3) OPEN SOURCED

HTTPS://GITHUB.COM/INSIGHTBUILDER

LETS HEAD TO COLAB

NOW WE ARE TALKING!!! PRACTICE

THANKS FOR WATCHING

HTTPS://GITHUB.COM/INSIGHTBUILDER