Member-only story

Scientific Documents Similarity Search With Deep Learning Using Transformers (SciBERT)

This article is a comprehensive overview of building a semantic similarity search tool for documents with k-NN and Cosine Similarity

Zoumana Keita · Follow

Published in Towards Data Science · 6 min read · Jan 17, 2022

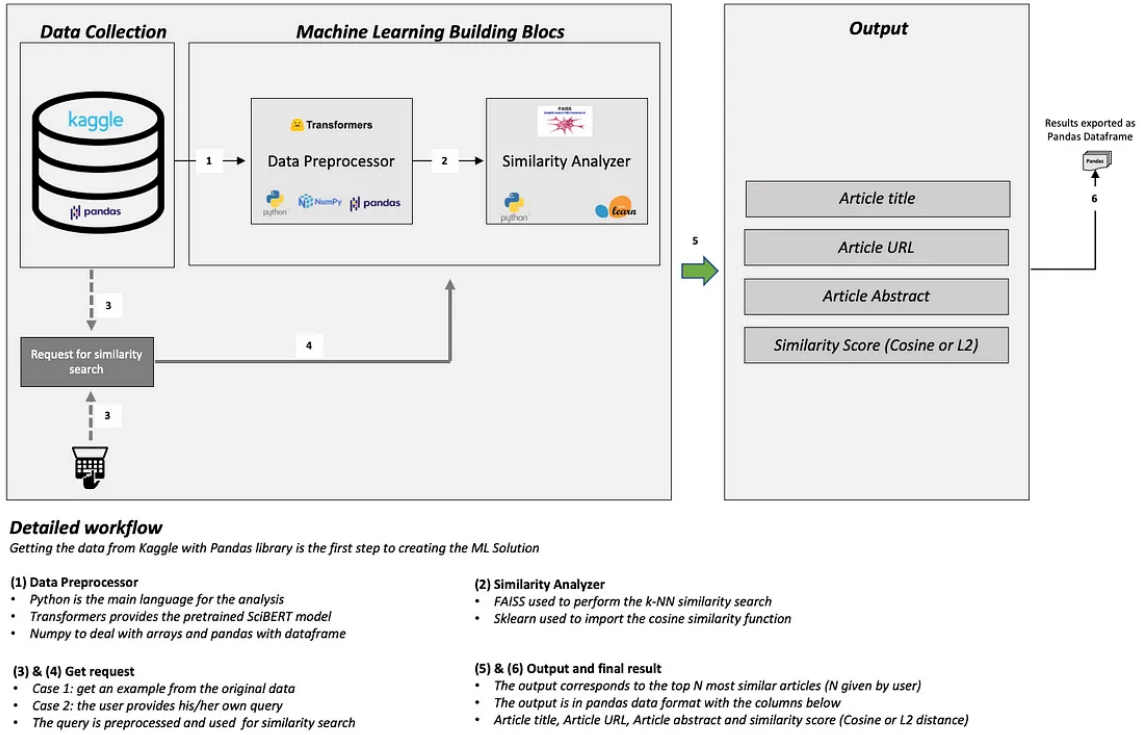
Photo by Maksym Kaharlytskyi on Unsplash

Introduction

When reading an interesting article, you might want to find similar articles from a large corpus of data. Manual processing is obviously not the strategy to go for. So, why not take advantage of the power of Artificial Intelligence to solve such problems? From this article, you will be able to use SciBERT, and two different similarity approaches (Cosine and k-NN) in order to find scientific articles that are most similar in meaning to your specific query.

*Note: **embeddings*** and ***vectors*** will be used interchangeably to mean the same thing throughout the article.

The following workflow gives all the steps of the building of our tool, from data extraction to recommending similar articles.



General Workflow of the Solution Being Solved (Image by Author)

About the Data

- This is the CORD-19 data set, a resource of over 59,000 scholarly articles, including over 48,000 with full text, about COVID-19, SARS-CoV-2, and related coronaviruses.
- This data has been made **freely available** by the White House and a coalition of leading research groups in order to help global research generate insights in support of the ongoing fight against this infectious disease.
- It is downloadable from this page on Kaggle.
- Further details about the dataset can be found on this page.

Before moving further, it is important to consider importing the following libraries that are crucial to the success of this project.

useful_libs.py

The data is downloaded from Kaggle and saved in a folder called ***input.*** Using

the shape function we get (47110, 16), meaning that it contains 47110 articles, and each one has 16 columns.

load_data_scibert.py

16 columns — but which one is the most important?

Below are a few details about the percentage of missing information in the data.

Before removing articles with missing Abstract

paper_id	0.000000
body_text	0.000000
methods	58.439822
results	63.984292
source	10.365103
title	10.439397
doi	12.604543
abstract	12.838039
publish_time	10.365103
authors	11.676926
journal	18.199958
arxiv_id	98.681809
url	10.617703
publish_year	0.000000
is_covid19	0.000000
study_design	0.000000
dtype: float64	

After removing articles with missing Abstract

paper_id	0.000000
body_text	0.000000
methods	53.796698
results	59.466173
source	8.645463
title	8.723394
doi	11.095417
abstract	0.000000
publish_time	8.645463
authors	9.432078
journal	17.473577
arxiv_id	98.487653
url	8.927963
publish_year	0.000000
is_covid19	0.000000
study_design	0.000000
dtype: float64	

Percentage of missing values from the data with a focus on Abstract column (Image by Author)

We will focus our analysis on the **abstract** column for simplicity's sake, also it

is the one with 0% missing data. But you could use other textual columns such as **body_text**; it is up to you. Also, we will use a subset of 2000 observations in order to speed the processing.

Articles' Abstract in the data

Let have a look at some random articles. Here we limit the printing to the first hundred words, because some of them are very long.

show_random_articles.py

```
Article #3252
 --> Title: Tuning antiviral CD8 T-cell response via proline-altered peptide ligand vaccinat
ion
 --> Abstract: AbstractViral escape from CD8+ cytotoxic T lymphocyte responses correlates wi
th disease progression and represents a significant challenge for vaccination. Here, we demo
nstrate that CD8+ T cell recognition of the naturally occurring MHC-I-restricted LCMV-associ
ated immune escape variant Y4F is restored following vaccination with a proline-altered pept
ide ligand (APL). The APL increases MHC/peptide (pMHC) complex stability, rigidifies the pep
tide and facilitates T cell receptor (TCR) recognition through reduced entropy costs. Struct
ural analyses of pMHC complexes before and after TCR binding, combined with biophysical anal
yses, revealed that although the TCR binds similarly to all complexes, the p3P modification
alters the conformations of a ...


Article #36667
 --> Title: A highly conserved WDYPKCDRA epitope in the RNA directed RNA polymerase of human
coronaviruses can be used as epitope-based universal vaccine design
 --> Abstract: BACKGROUND: Coronaviruses are the diverse group of RNA virus. From 1960, six
strains of human coronaviruses have emerged that includes SARS-CoV and the recent infection
by deadly MERS-CoV which is now going to cause another outbreak. Prevention of these viruses
is urgent and a universal vaccine for all strain could be a promising solution in this circu
mstance. In this study we aimed to design an epitope based vaccine against all strain of hum
an coronavirus. RESULTS: Multiple sequence alignment (MSA) approach was employed among spike
(S), membrane (M), enveloped (E) and nucleocapsid (N) protein and replicase polyprotein 1ab
to identify which ...


Article #14038
 --> Title: Development and characterization of a Rift Valley fever virus cell–cell fusion a
ssay using alphavirus replicon vectors
 --> Abstract: Abstract Rift Valley fever virus (RVFV), a member of the Phlebovirus genus in
the Bunyaviridae family, is transmitted by mosquitoes and infects both humans and domestic a
nimals, particularly cattle and sheep. Since primary RVFV strains must be handled in BSL-3+
or BSL-4 facilities, a RVFV cell–cell fusion assay will facilitate the investigation of RVFV
glycoprotein function under BSL-2 conditions. As for other members of the Bunyaviridae famil
y, RVFV glycoproteins are targeted to the Golgi, where the virus buds, and are not efficient
ly delivered to the cell surface. However, overexpression of RVFV glycoproteins using an alp
havirus replicon vector resulted in the ...
```

Three random articles' abstract from the data (Image by Author)

Data Processing & Vectorization

This step aims to vectorize the articles' abstract text so that we can perform the similarity analysis. Since we are dealing with the scientific documents, we will use SciBERT, which is a pre-trained language model for Scientific text data. You can find more information about it on Semantic Scholar.

The main steps involved in this part are:

Load the pre-trained model & tokenizer. When loading the model, we need to set the ***output_hidden_states*** to **True** so that we can extract the embeddings.

load_model_artifacts.py

This function ***convert_single_abstract_to_embedding*** is mostly inspired by the

BERT Word Embeddings Tutorial of Chris McCormick. It aims to create an embedding for a given text data using a pre-trained model.

embedding_from_text.py

Test on a single text data

Here we test the function on the 30th article. You can choose whatever number you want, as long as it exists in the data.

test_embedding.py

Line 6 shows **Embedding shape: (768,).** This means that a single vector is composed of 768 values. We can finally move to the next step by applying the conversion process to all the articles in the data using the ***convert_overall_text_to_embedding*** function. But, before that, we are going to remove some columns from the data using the ***get_min_viable_data*** function, in order to have fewer columns in the result of the final query search.

With the previous helper functions, we can create a new column that will

contain for each article, its corresponding abstract embedding.

	paper_id	title	abstract	url	embeddings
11754	9778c2bdf6be32053f9a450194f767b47c87c891	NaN	Graphical Abstract Highlights d The cryo-EM st...	NaN	[[-0.3626312, 0.34195867, 0.40458974, -0.65087...
20720	98fc84b407a8632ee5ac42431d12c6acf7dfe3d8	Genetic Loci That Influence Cause of Death in ...	A genome scan was conducted to seek evidence f...	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7...	[[-0.49350545, -1.1899799, -0.0245932, -0.8623...
15193	fd0fcdc8d711665338facb36ba94489c3b63a0d7	New Pre-pandemic Influenza Vaccines: An Egg-an...	Highly pathogenic avian H5N1 influenza viruses...	http://europepmc.org/articles/pmc2793094?pdf=r...	[[0.3071721, -0.85410595, 0.43242905, 0.003505...

First 3 rows of the Data With Embeddings column (Image by Author)

So far, so good! Everything is finally set up for similarity search.

Similarity Search

Now, we can perform the similarity analysis between a given ***query*** vector and all the embeddings vectors. The ***query*** processing will be similar to both cosine and k-NN. Below is the function responsible for that.

process_query.py

The cosine similarity between two documents' embedding measures how

similar those documents are, irrespective of the size of those embeddings. It measures the cosine of the angle between the two vectors projected in a multi-dimensional space.

- ***cosine similarity of 1*** means that the two documents are 100% similar
- ***cosine similarity of 0*** means that the two documents have 0% similarity

The following function returns the top N (N is the number of similar articles to return) articles similar to the query text.

cosine_recommendations.py

Now we can call the function to get the top 5 most similar articles.

cosine_results.py

Here is the result of the previous query

title	abstract	url	cos_sim
NaN	Highlights d ENDU-2 nuclease regulates nucleot...	NaN	0.825622
Regioselective synthesis of 6-substituted-2-am...	Abstract A series of 2-amino-5-bromo-4(3H)-pyr...	https://doi.org/10.1016/j.ejmech.2013.06.036	0.773827
Synthesis of 4-aminoquinoline–pyrimidine hybri...	Abstract One of the most viable options to tac...	https://doi.org/10.1016/j.ejmech.2013.05.046	0.761577
Australia was indeed the "lucky country" in th...	Anton Y Peleg, Wendy J Munckhof\nAustralia was...	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7...	0.757679
Antigen delivery systems for veterinary vaccin...	Abstract The recent advances in molecular gene...	https://doi.org/10.1016/j.vaccine.2008.09.044	0.754393

Top 5 most similar Articles to the Query (Image by Author)

For a better view, we can just look at the first 2 most similar articles.

Query text below

```
Graphical Abstract Highlights d The cryo-EM structure of full-length human NPC1 was determin
ed at 4.4 Å resolution d Structure-guided biochemical analysis of cholesterol transfer from
NPC2 to NPC1 d Low-resolution cryo-EM structure of NPC1 bound to GPcl of Ebola virus was obt
ained d A trimeric GPcl binds to one NPC1 through the crystal structure-revealed interface**
```

Query text data from the original dataframe

Most two similar articles from the previous top_articles dataframe

1st Article's Abstract Cosine Similarity: 0.82

```
'Highlights d ENDU-2 nuclease regulates nucleotide metabolism and germ cell proliferation in
worms d ENDU-2 expression is induced by nucleotide imbalance and other genotoxic stresses d
ENDU-2 inhibits CTP synthase phosphorylation by repressing PKA and HDA-1 in the gut d ENDU-2
function may be conserved in mammalian cells'
```

2nd Article's Abstract Cosine Similarity: 0.77

```
'Abstract A series of 2-amino-5-bromo-4(3H)-pyrimidinone derivatives bearing different subst
ituents at the C-6 position were synthesized using a highly regioselective lithiation–substi
tution protocol, and the effect of structural variation at the C-6 position on their antivir
al activity in cell culture was evaluated. Although some of the derivatives were found to be
active against various virus strains, they were effective only close to their toxicity thres
hold.'
```

Extract of the most two similar articles from the recommended dataframe (Image by Author)

FAISS is a library developed by Facebook AI Research. According to their wikipage:

> Artificial Intelligence Machine Learning Naturallanguageprocessing Data Science Cosine Similarity
>
> fc la er
>
> tains algorithms that search in sets of vectors of any size, up to ones that possibly do not fit in RAM

Below are the steps to build the search algorithm using the previously built embeddings

- create the flat index. The index uses the L2 (Euclidean) distance metrics to measure the similarity between the query vector and all the vectors

eddings).

- add all the vectors to the index
- define the number **k** of similar documents we want
- run the similarity search to get the result

Written by Zoumana Keita

Follow

8.2K Followers · Writer for Towards Data Science

Data Scientist at IFC - The World Bank Group || Videos about AI, Data Science, Programming or Tech 👉 https://www.youtube.com/@zoumdatascience

---

More from Zoumana Keita and Towards Data Science

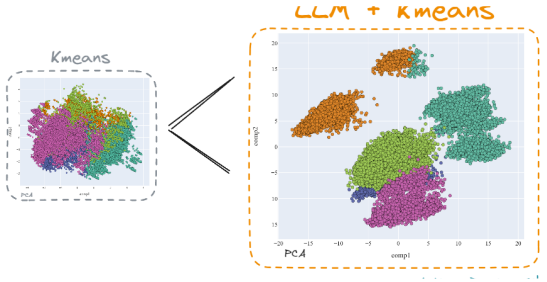
Zoumana Keita in Artificial Corner

4 Hidden Python Treasures That You Probably Didn't Know Existed

Learning these libraries will take your Python skill to the next level — no doubt.

✨ · 7 min read · Aug 3

511 5



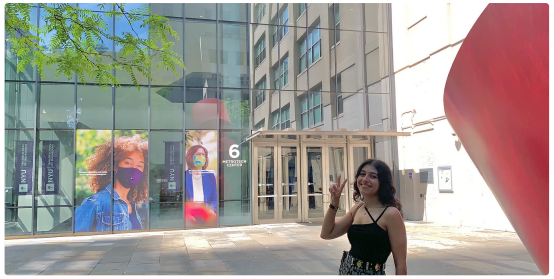
Damian Gil in Towards Data Science

Mastering Customer Segmentation with LLM

Unlock advanced customer segmentation techniques using LLMs, and improve your...

23 min read · Sep 26

3.1K 25



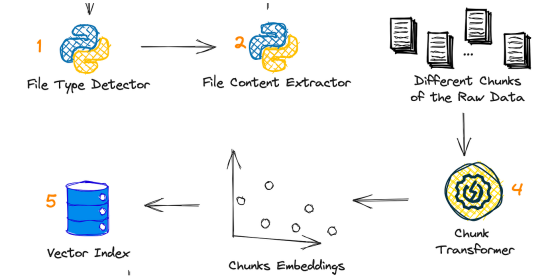
Khouloud El Alami in Towards Data Science

Don’t Start Your Data Science Journey Without These 5 Must-D...

A complete guide to everything I wish I’d done before starting my Data Science...

✦ · 18 min read · Sep 24

2.4K 23



Zoumana Keita in Towards Data Science

How to Chat With Any File from PDFs to Images Using Large...

Complete guide to building an AI assistant that can answer questions about any file

✦ · 9 min read · Aug 5

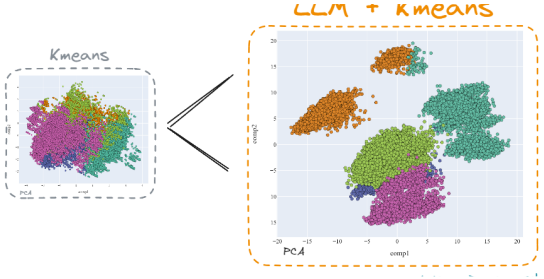
1.5K 13

See all from Zoumana Keita

See all from Towards Data Science

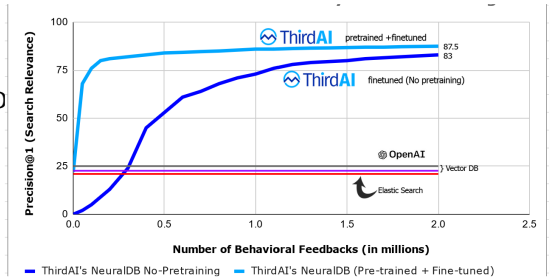
Recommended from Medium

faiss_setup.py



erfo

.



Damian Gil in Towards Data Science

Mastering Customer Segmentation with LLM

Unlock advanced customer segmentation techniques using LLMs, and improve your...

23 min read · Sep 26

3.1K 25

Anshu in ThirdAI Blog

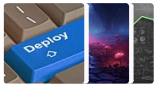
Demystifying LLM-Driven Search: Stop Comparing Embeddings or...

A comprehensive evaluation of a commercial semantic search system reveals that the...

5 min read · 5 days ago

34 3

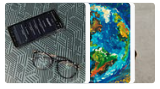
Lists



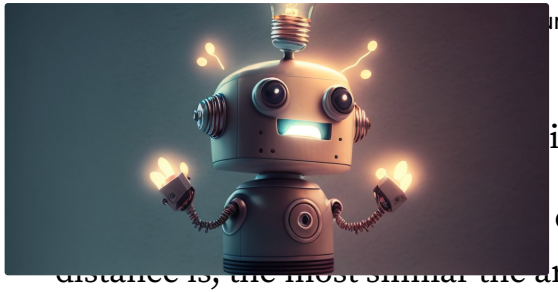
Predictive Modeling w/ Python
20 stories · 490 saves

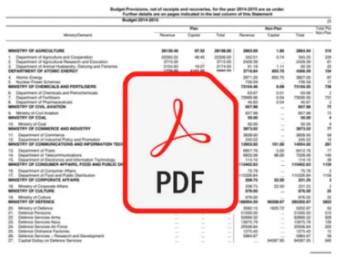
Natural Language Processing
706 stories · 312 saves

Practical Guides to Machine Learning
10 stories · 561 saves



ChatGPT prompts
26 stories · 501 saves





Justin Swansburg

How to Use LLMs to Build Better Clustering Models

A quick guide on a new way to leverage large language models to improve the separation...

9 min read · May 13

-- 2

Ryan Nguyen in Towards AI

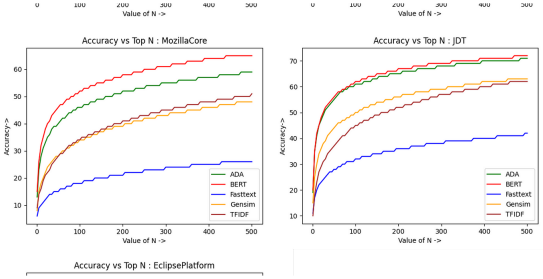
RAG Pipeline Pitfalls: The Untold Challenges of Embedding Table

A typical journey of building the RAG pipeline from Zero to Something and a guide on...

· 9 min read · Oct 8

-- 1

Note: I decided to break down all the steps on purpose in order to make sure you understand them properly. But you can put everything together into a single function.



Avinash Patil

Embeddings: BERT better than ChatGPT4?

In this study, we compared the effectiveness of semantic textual similarity methods for...

4 min read · Sep 18

-- 2

AL Anany

The ChatGPT Hype Is Over — Now Watch How Google Will Kill...

It never happens instantly. The business game is longer than you know.

· 6 min read · Sep 1

-- 424

faiss_show_recommendations.py

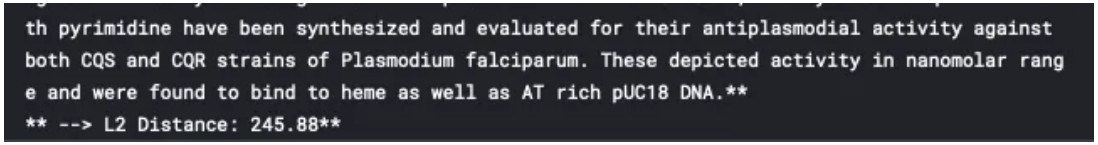
** Article #0 **
** --> Abstract :
Graphical Abstract Highlights d The cryo-EM structure of full-length human NPC1 was determin ed at 4.4 Å resolution d Structure-guided biochemical analysis of cholesterol transfer from NPC2 to NPC1 d Low-resolution cryo-EM structure of NPC1 bound to GPcl of Ebola virus was obt ained d A trimeric GPcl binds to one NPC1 through the crystal structure-revealed interface**
** --> L2 Distance: 0.00**

** Article #1683 **
** --> Abstract :
Highlights d ENDU-2 nuclease regulates nucleotide metabolism and germ cell proliferation in worms d ENDU-2 expression is induced by nucleotide imbalance and other genotoxic stresses d ENDU-2 inhibits CTP synthase phosphorylation by repressing PKA and HDA-1 in the gut d ENDU-2 function may be conserved in mammalian cells**
** --> L2 Distance: 181.52**

** Article #1021 **
** --> Abstract :
Abstract A series of 2-amino-5-bromo-4(3H)-pyrimidinone derivatives bearing different substi tuents at the C-6 position were synthesized using a highly regioselective lithiation–substit ution protocol, and the effect of structural variation at the C-6 position on their antivira l activity in cell culture was evaluated. Although some of the derivatives were found to be active against various virus strains, they were effective only close to their toxicity thres hold.**
** --> L2 Distance: 230.64**

** Article #348 **
** --> Abstract :
Anton Y Peleg, Wendy J Munckhof
Australia was indeed the "lucky country" in the recent worldwide SARS epidemic 229**
** --> L2 Distance: 244.74**

** Article #1712 **
** --> Abstract :
Abstract One of the most viable options to tackle the growing resistance to the antimalarial drugs such as artemisinin is to resort to synthetic drugs. The multi-target strategy involvi ng the use of hybrid drugs has shown promise. In line with this, new hybrids of quinoline wi



5 most similar articles similar to the query (Image by author)

Observation

- The first document has L2 = 0, which means 100% similarity. This is obvious because the query was compared with itself, but we can simply remove it from the analysis.

Conclusion

In this article, we've studied the whole pipeline of implementing a semantic similarity search tool using SciBERT embedding, Cosine, and k-NN similarity approaches. Also, I hope it has been beneficial to you as well!

You can find below additional resources to further your learning. Follow me on YouTube for more interactive sessions!

Additional Ressources

Article Source Code on GitHub

SciBERT: A Pretrained Language Model for Scientific Text

BERT Word Embeddings Tutorial

Facebook AI Research

Bye for now 🏃