Weaviate

Vector Search for Data Scientists

A Case Study with **Twitter Analytics**

Powered by  SeMI Technologies

Common questions in Data Science

#1 - How is my data distributed?

#2 - Are there outliers in my data?

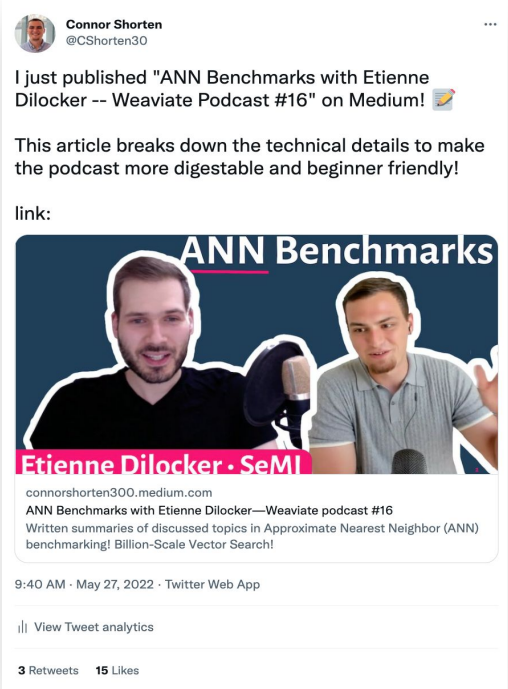
#3 - Are my variables correlated with each other?

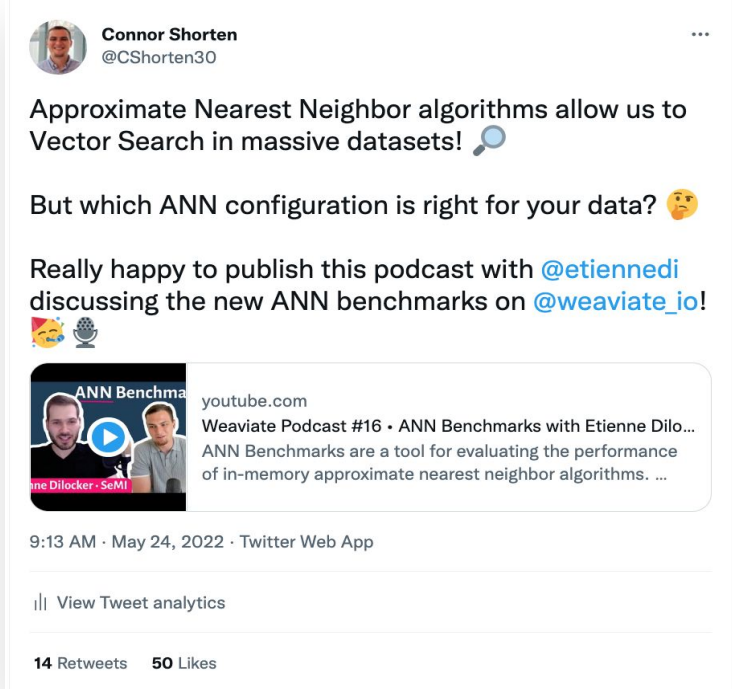
Vector Search

#1 - Can we capture the semantics in vector representations?

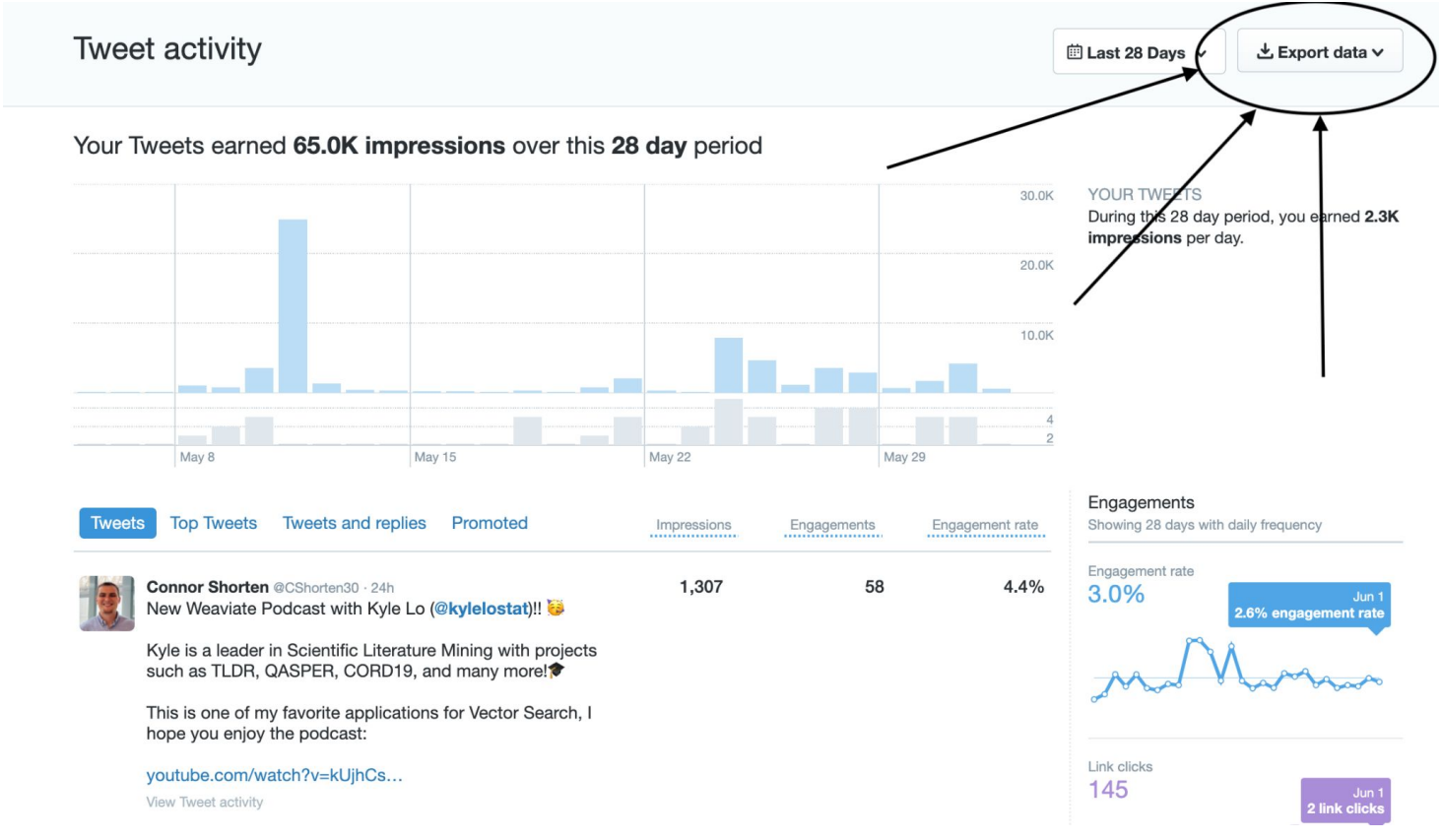
#2 - What can we learn about our data from semantic clusters?

Social Media Clicks and Twitter Analytics





Twitter Analytics



Twitter Analytics CSV Data

Tweet Text	Time	Impressions	Engagements	Engagement Rate	Retweets	Replies	Likes	User Profile Clicks	Url Clicks
I just published "ANN Benchmarks with Etienne Dilcoker -- Weaviate Podcast #16 on Medium..	May 27th, 1:34pm	1905	50	2.6%	3	1	15	2	18
Approximate Nearest Neighbor algorithms allow us to Vector Search in massive datasets! ...	May 24th, 1:13pm	7182	252	3.5%	14	1	50	27	36

Feature Engineering:
Contains Emoji?
Character Count?
Word Count?
Contains "Weaviate"?

Key Takeaways: **“Vector Search for Data Scientists”**

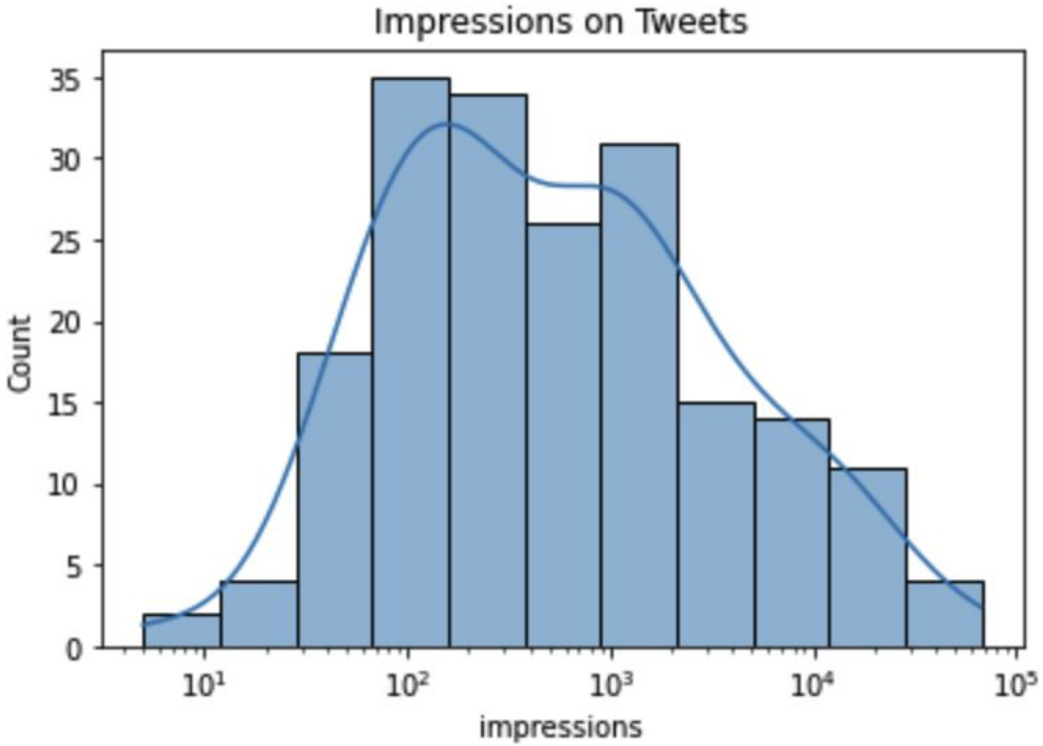
1. Segmentation in Data Science
2. Vector Representations of Data
3. Vector Segmentation
4. Weaviate for Twitter Analytics
5. Research Questions and Discussion

Slides, Colab Notebook, Video Presentation available on:

github.com/CShorten/Vector-Search-for-Data-Scientists

Key Takeaway #1 - **Segmentation in Data Science**

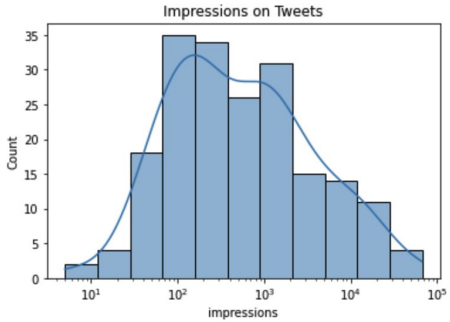
Visualizing Distributions of Values

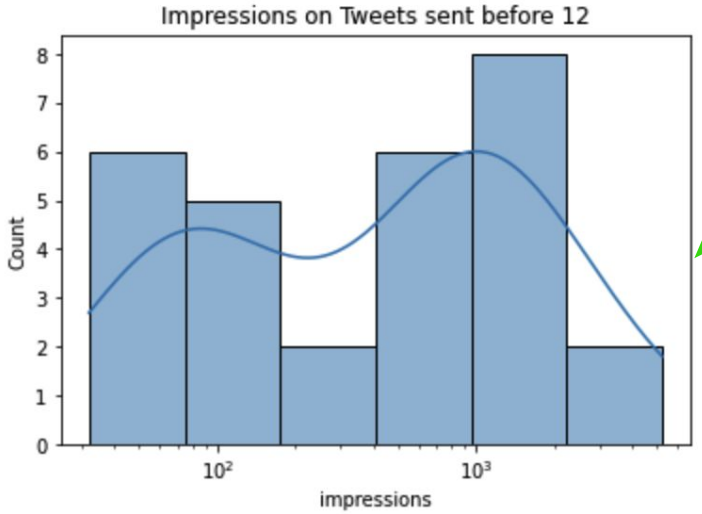


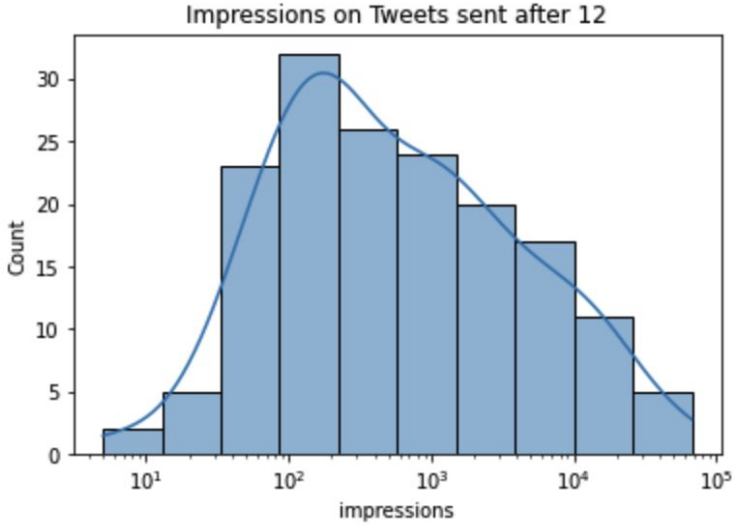
Segmentation in Data Science

- What Time was the Tweet sent?
- Is there a URL Link in the Tweet?
- Symbolic vs. Vector Segmentation

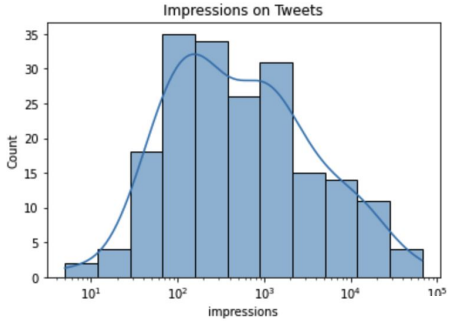
What Time was the Tweet sent?

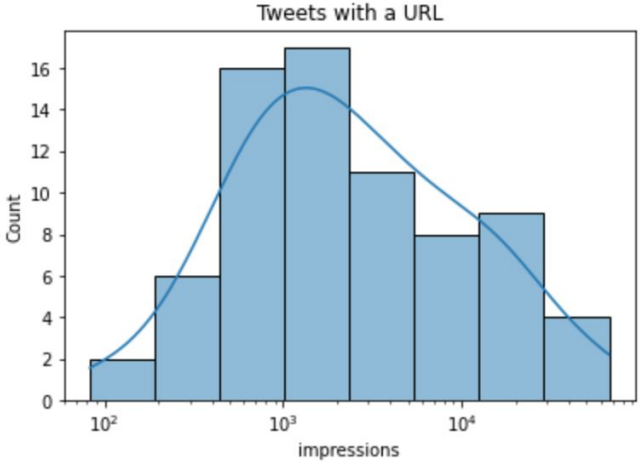


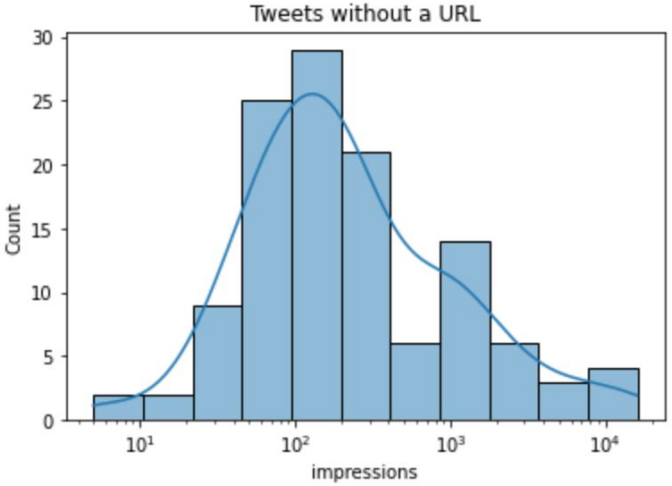




Is there a URL Link in the Tweet?







Can we split Impressions based on the Semantics of the content?

Weaviate Podcast

Weaviate Tutorial

AI Weekly Update

How can we segment analytics based on the semantics of...

- Text
- Images
- Code
- Audio
- Video
- Graph-Structure
- Biological Sequences
- ... !

Summary of Takeaway #1 Segmentation in Data Science

We visualize the **Distribution** of our data to get a sense of it.

For example we see that **Impressions** are somewhat **Normally Distributed**.

Is that also true for Tweets sent at **3 AM?**

What about Tweets related to **Deep Learning for Robotics?**

Key Takeaway #2 - **Vector Representations of Data**

Symbols compared to Vectors

Symbols

Category - [0, 1, 0, 0, 0, 0]

Numeric - 52

Boolean - True

Vectors

[0.1, 0.8, 0.34, 0.8, ... 0.2]

Vector Representations of Data

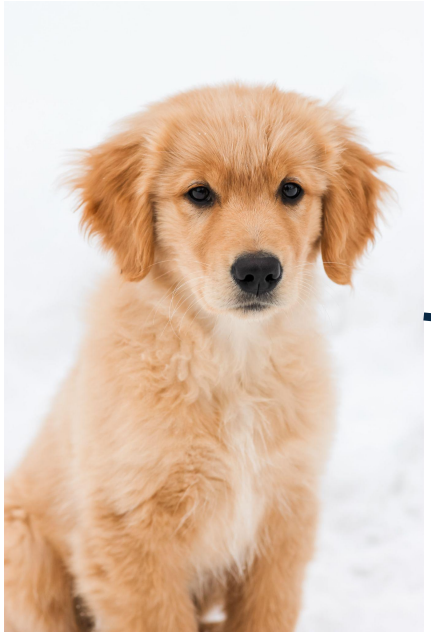
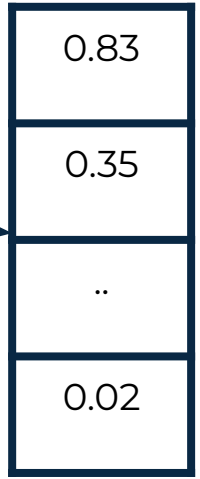
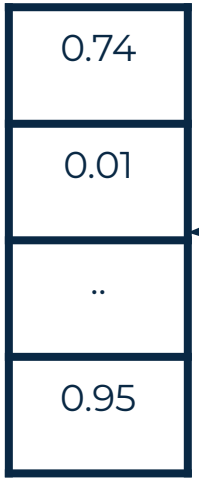


Photo by Shayna Douglas on Unsplash





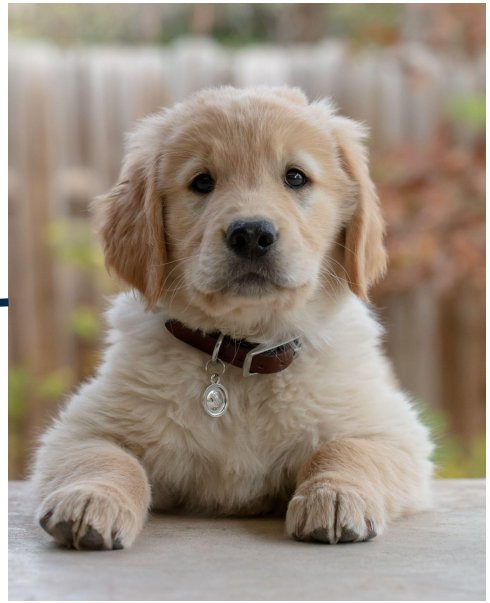


Photo by Bill Stephan on Unsplash

Are these puppies similar?
Let's ask Vector Distance!

$$\text{L2 Distance} = \Sigma \left|\left| a_i - b_i \right|\right|^2$$

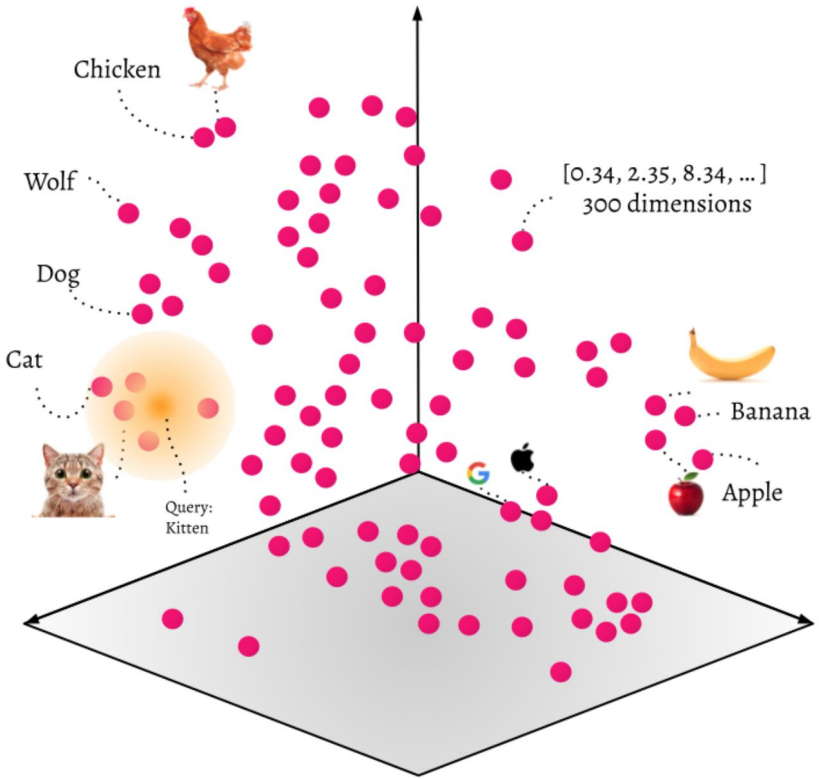
Vector Name	Value 1	Value 2	Value 3
Puppy1	4	8	10
Puppy2	2	9	11
Airplane	1	20	20

L2 Distance (Puppy1, Puppy2) = $(4-2)^2 + (8-9)^2 + (10-11)^2 = 6$

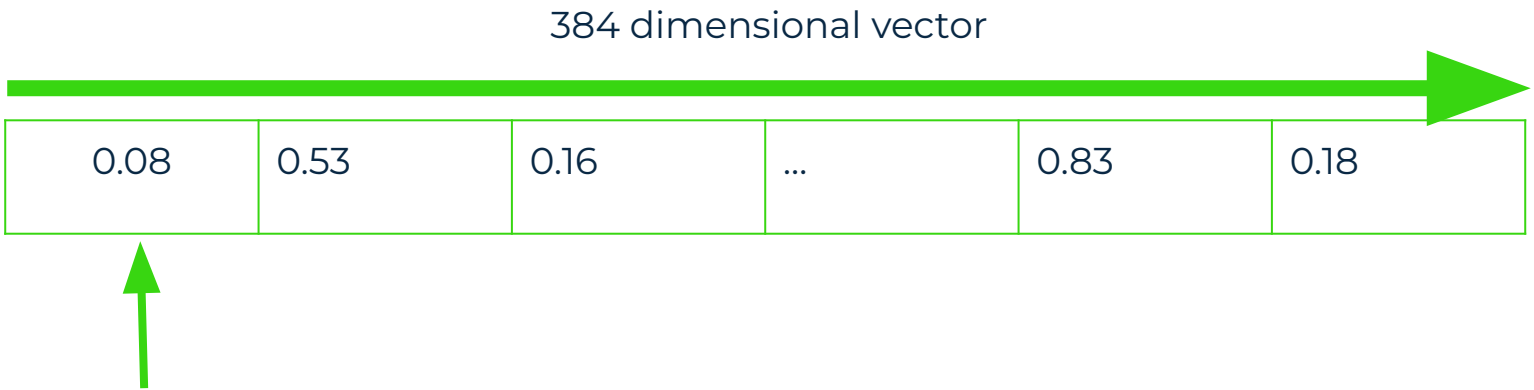
L2 Distance (Puppy1, Airplane) = $(4-1)^2 + (8-20)^2 + (10-20)^2 = 253$

$6 << 253$, Puppy1 is thus much more semantically similar to Puppy2 than Airplane

Capturing Semantics in Vector Representations



How do Vectors represent real-world objects?



Does this represent how much of a “brand” this is?

We aren’t sure! But there are research fields such as “Multimodal Neurons” from OpenAI, and the general field of **Disentangled Representation Learning** that are making great strides in understanding this.

Can we compress these vectors?

Sometimes!

Ideas like Binary Passage Retrieval (shown above) - fp32 to Binary values

Ideas like Product Quantization - 384-d vector mapped to 32-d

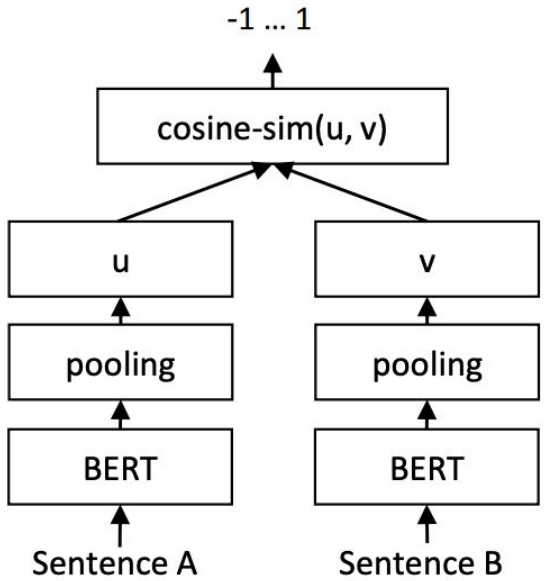
Semantic Similarity with Vector Representations

Sentence-BERT:
Sentence Embeddings
using Siamese
BERT-Networks

Authored by

Nils Reimers and Iryna
Gurevych

Published 2019



Positive and Negative Pair Sampling

Query Point

The **Miami Heat** are an American professional basketball team based in Miami. The Heat compete in the National Basketball Association (NBA) as a member of the league's Eastern Conference Southeast Division. The club plays its home games at FTX Arena, and has won three NBA championships.

The franchise began play in the 1988–89 season as an expansion team. After a period of mediocrity, the Heat gained relevance in the mid-1990s when Pat Riley became team president and head coach. Riley constructed the trades of Alonzo Mourning and Tim Hardaway, which propelled the team into playoff contention. Mourning and Hardaway led the Heat to four consecutive division titles prior to their departures in 2001 and 2002, respectively. The team also experienced success after drafting Dwyane Wade in 2003.

Positive (Semantically Similar)

Led by Wade and, following a trade for former NBA Most Valuable Player (MVP) Shaquille O'Neal, the Heat won their first NBA title in 2006, after Riley named himself head coach for a second stint. After the departure of O'Neal two years later, the team struggled for the remainder of the 2000s. Riley remained team president, but was replaced as head coach by Erik Spoelstra. In 2010, the Heat signed former league MVP LeBron James and NBA All-Star Chris Bosh, creating the "Big Three" along with Wade. During their four years together, Spoelstra, James, Wade, and Bosh led the Heat to the NBA Finals in every season, culminating in back-to-back championships in 2012 and 2013. All three departed by 2016, and the team entered a period of rebuilding. After acquiring All-Star Jimmy Butler in 2019, the Heat returned to the NBA Finals in 2020. The Heat acquired six-time NBA All-Star Kyle Lowry in 2021.

The Heat hold the record for the NBA's third-longest winning streak, 27 straight games, set during the 2012–13 season. Six Hall of Famers have played for Miami, and James won two consecutive NBA MVP Awards while playing for the team.

Negative (Semantically Different)

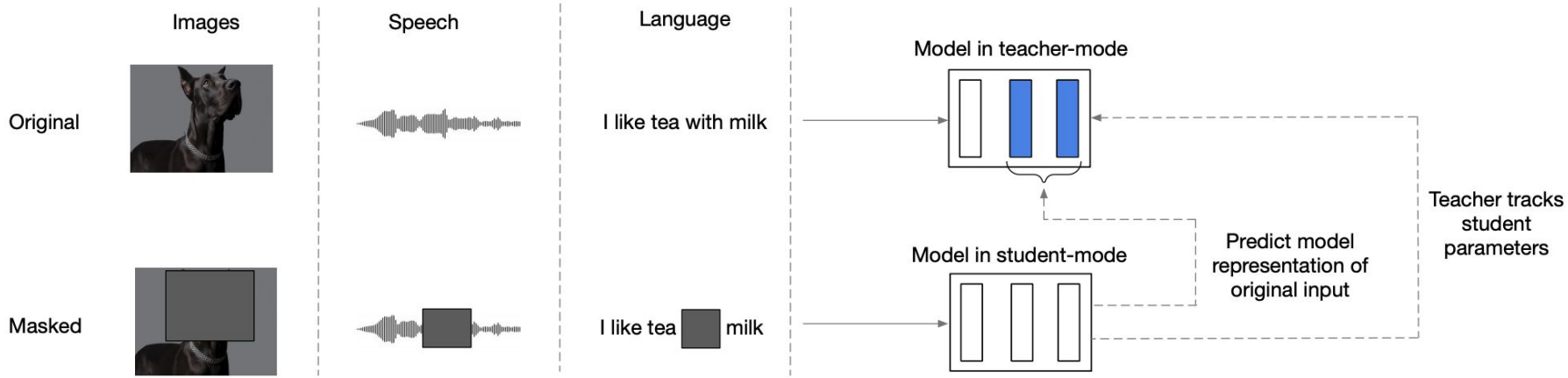
Deep learning (also known as **deep structured learning**) is part of a broader family of machine learning methods based on artificial neural networks with representation learning. Learning can be supervised, semi-supervised or unsupervised.[2]

Deep-learning architectures such as deep neural networks, deep belief networks, deep reinforcement learning, recurrent neural networks and convolutional neural networks have been applied to fields including computer vision, speech recognition, natural language processing, machine translation, bioinformatics, drug design, medical image analysis, climate science, material inspection and board game programs, where they have produced results comparable to and in some cases surpassing human expert performance.[3][4][5]

Artificial neural networks (ANNs) were inspired by information processing and distributed communication nodes in biological systems. ANNs have various differences from biological brains. Specifically, artificial neural networks tend to be static and symbolic, while the biological brain of most living organisms is dynamic (plastic) and analogue.[6][7]

Another strategy - Data2Vec, Baevski et al. 2022

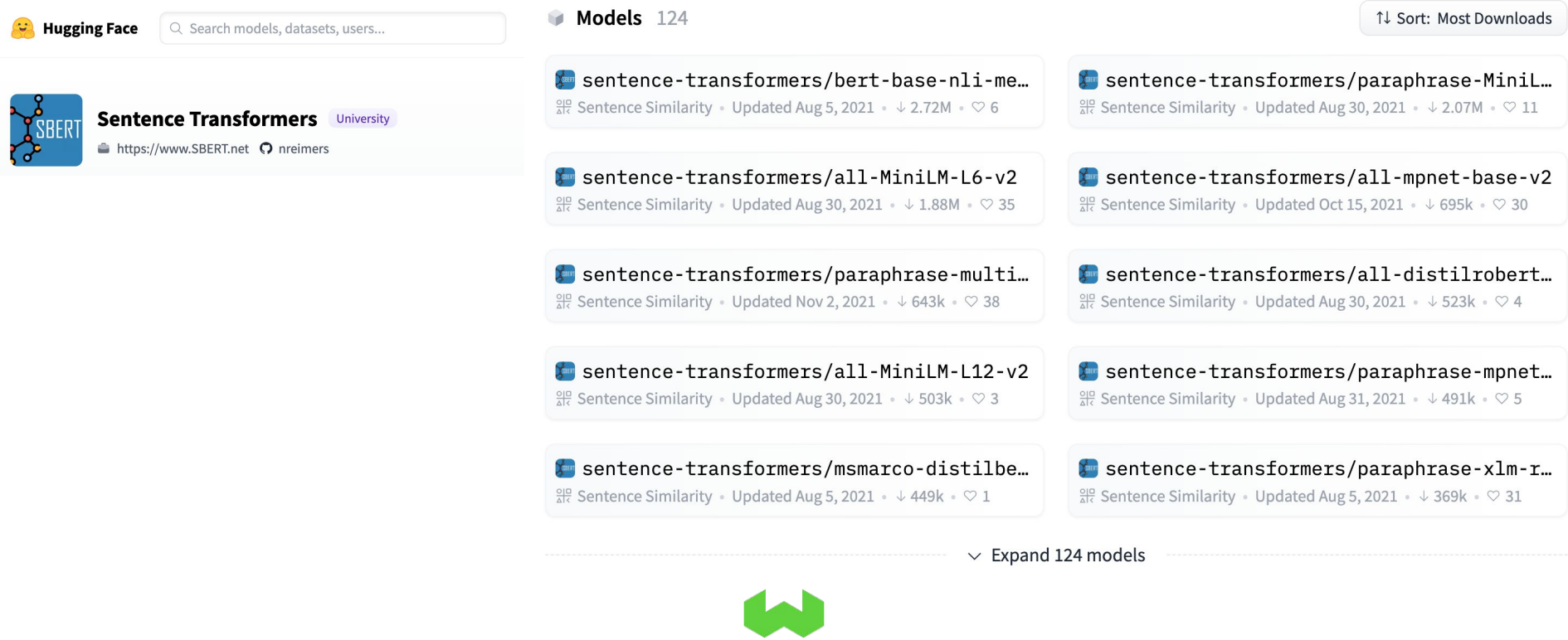
data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language



Do we need to train our own models?

No! There are many pre-trained models that work very well for a broad range of data!

Great place to get started: Sentence Transformers



Summary of Takeaway #2 Vector Representations of Data

Data such as Images, Text, Code, … can be represented as **Vectors** with Deep Learning models.

These models are trained to **maximize semantic similarity** with **massive** collections of data.

We often do not need to train the models ourselves for **particular data domains** to reach reasonable performance.

Key Takeaway #3 - **Vector Segmentation**

We can segment analytics based on the semantics of...

- Text
- Images
- Code
- Audio
- Video
- Graph-Structure
- Biological Sequences
- ... !

Can we split Impressions based on the Semantics of the content?

Weaviate Podcast

Weaviate Tutorial

AI Weekly Update

More Examples

House Hunting

Symbols: # of bedrooms, # of bathrooms, square feet, city

→ With Vectors we can encode:

- Visual style
- Neighborhood structure
- *More flexible interface to define features with text*

e-Commerce Products

Symbols: “Shoes”, “T-Shirt”, “Pants” or colors

→ With Vectors we can encode visual styles

Movies

Symbols can differentiate between genres like “Children”, “Action”, or “Sci-Fi”

→ With Vectors we can encode:

- Themes
- Characters
- Storylines

Scientific Papers

Symbols: “Biology”, “Machine Learning”

→ With Vectors we can encode

- Nuance of the ideas
- Writing style

Music

Symbols can differentiate between genres like “Hip Hop”, “Dance”

→ With Vectors we can encode:

- Tone
- Lyrics
- Instruments

“That’s the magic of deep learning: turning meaning into vectors, then into geometric spaces, and then incrementally learning complex geometric transformations that map one space to another. All you need are spaces of sufficiently high dimensionality in order to capture the full scope of the relationships found in the original data.”

- Francois Chollet, Deep Learning with Python, 2nd edition

Summary of Takeaway #3 Vector Segmentation

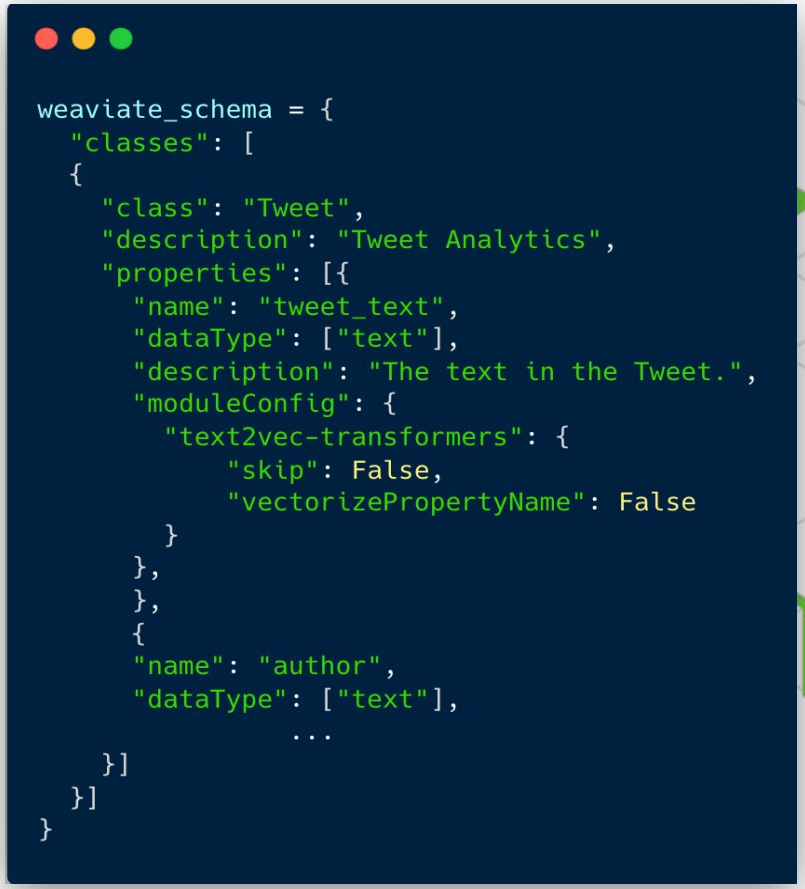
Vector representations, also known as embeddings, enable an **Interface** to split analytics based on the **Semantics** of the content.

This content could be **Text, Images, Code, Audio, Videos, ...**

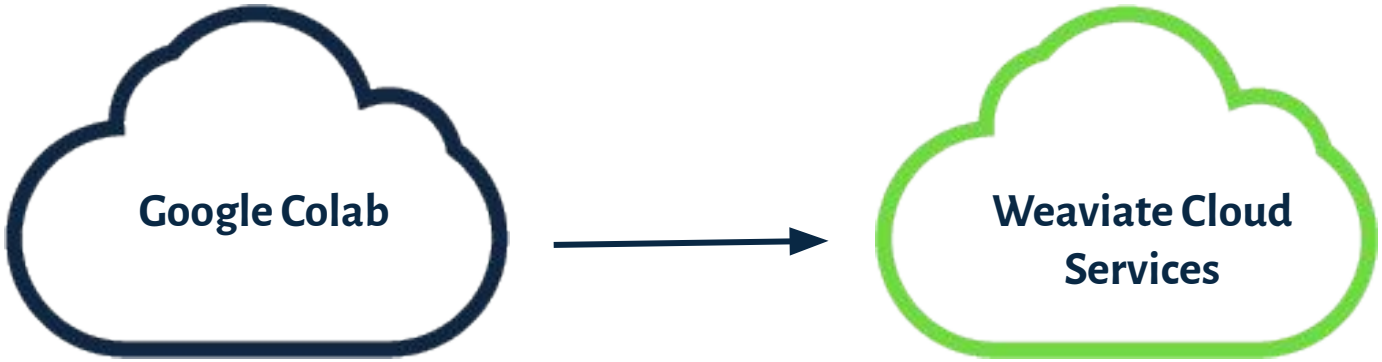
Key Takeaway #4 - **Weaviate for Twitter Analytics**

Twitter Analytics

Tweet Text	Time	Impressions	Engagements	Engagement Rate	Retweets	Replies	Likes	User Profile Clicks	Url Clicks
I just published “ANN Benchmarks with Etienne Dilcoker -- Weaviate Podcast #16 on Medium..	May 27th, 1:34pm	1905	50	2.6%	3	1	15	2	18
Approximate Nearest Neighbor algorithms allow us to Vector Search in massive datasets! ...	May 24th, 1:13pm	7182	252	3.5%	14	1	50	27	36


weaviate_schema = {
"classes": [
{
"class": "Tweet",
"description": "Tweet Analytics",
"properties": [{
"name": "tweet_text",
"dataType": ["text"],
"description": "The text in the Tweet.",
"moduleConfig": {
"text2vec-transformers": {
"skip": False,
"vectorizePropertyName": False
}
},
},
{
"name": "author",
"dataType": ["text"],
...
}]
}]
}

Cloud Data Upload



There are many other ways to do this as well

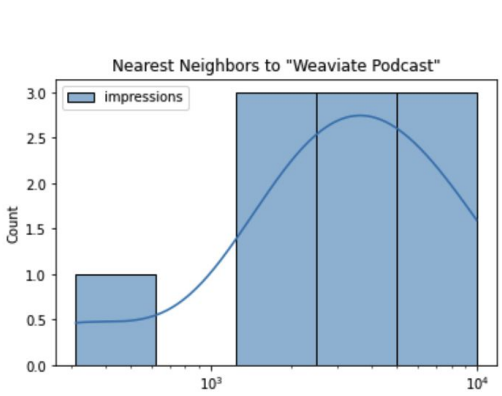
GraphQL Live Demo

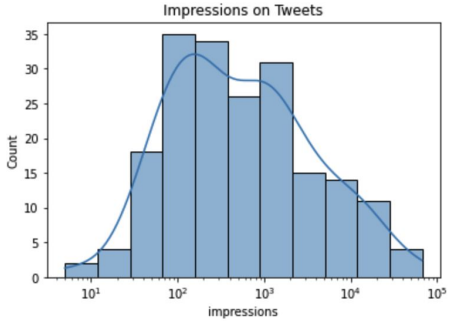
5 Nearest Neighbors to → “Weaviate Coding Tutorial”

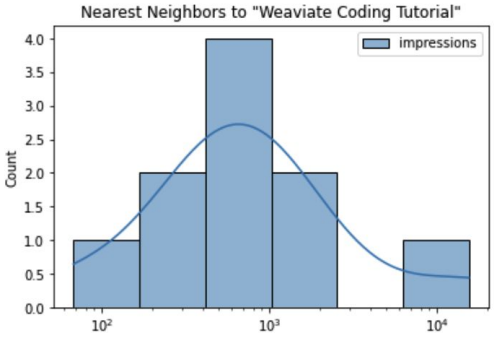
Content	Impressions
“We have 4 Weaviate Podcast Episodes so far […] **how to utilize the Weaviate Database as a Document Store in Haystack pipelines … ”**	311
“We have 2 new coding tutorials on Weaviate YouTube…”	1144
“@weaviate_io Love the integration of this with the GraphQL API!”	378
“Here are some thoughts on combining Weaviate and Haystack! TLDR: Weavaite is a great Vector Search database…”	15563
“Weaviate (@weaviate_io) is also announcing a collaboration with Jina AI (@JinaAI_)! …”	586

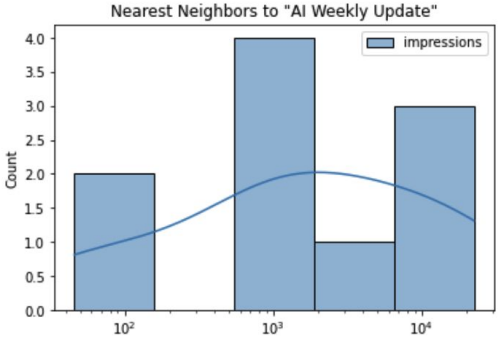
```
def nearText_wrapper(query):
  nearText = {
    "concepts": query
  }
  nearest_neighbor_search = client.query.get("Tweet", ["tweet_text", "impressions"])
  nearest_neighbor_search = nearest_neighbor_search.with_near_text(nearText).do()
  return nearest_neighbor_search["data"]["Get"]["Tweet"]
```

What was the Tweet about?

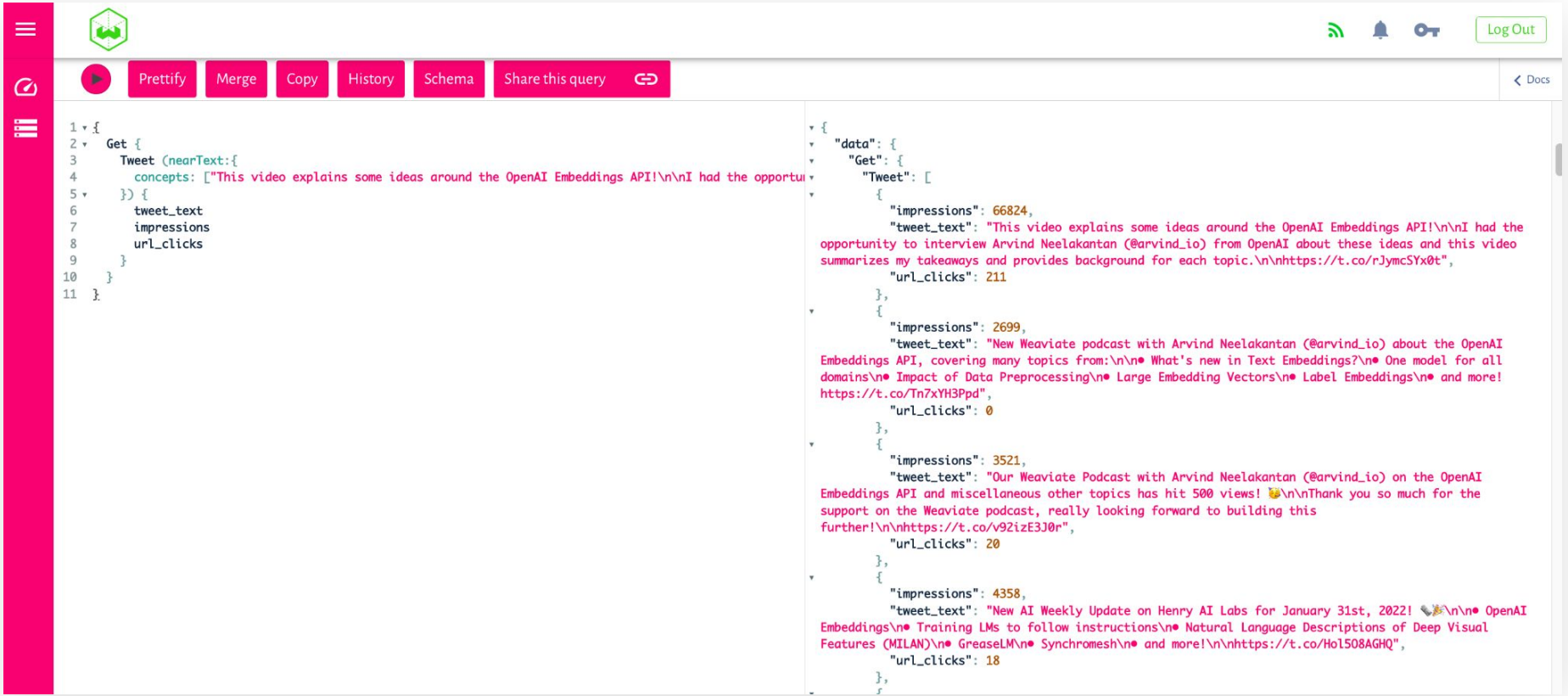




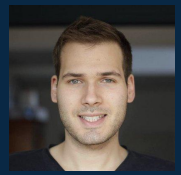




Have I tweeted something like this before?

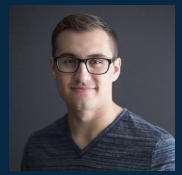


Have any Weaviate Podcast guests tweeted something like this recently?

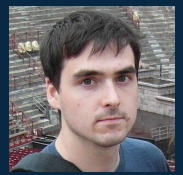




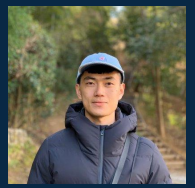




Weaviate
podcast



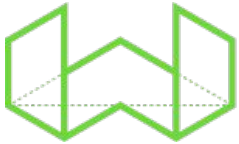




import tweepy
auth = OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token_key, access_token_secret)
api = tweepy.API(auth)
this_persons_tweets = api.user_timeline(username, count=100)

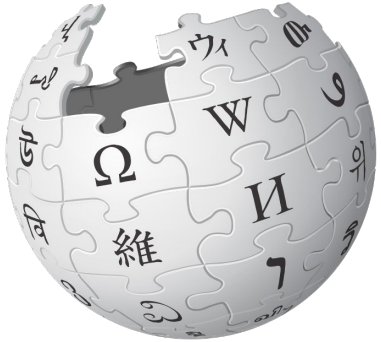
Tweet, Author, Likes



Weaviate

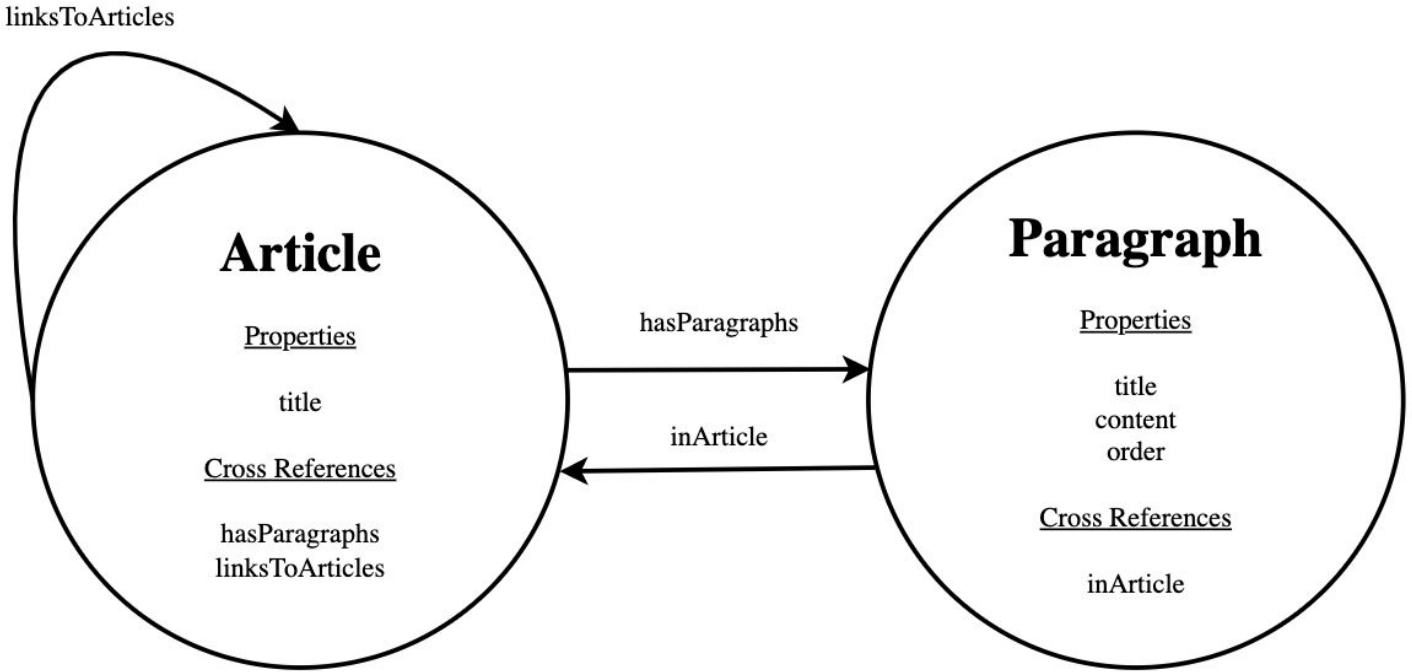
GraphQL Live Demo

GraphQL Wikipedia Demo

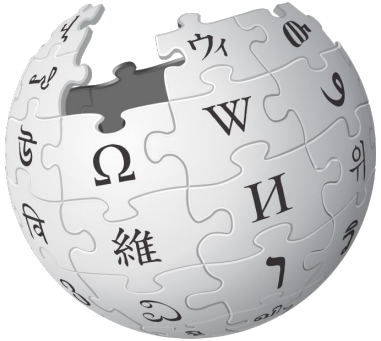


WIKIPEDIA
The Free Encyclopedia

Wikipedia Live Demo - Graph Data Model



GraphQL Wikipedia Demo



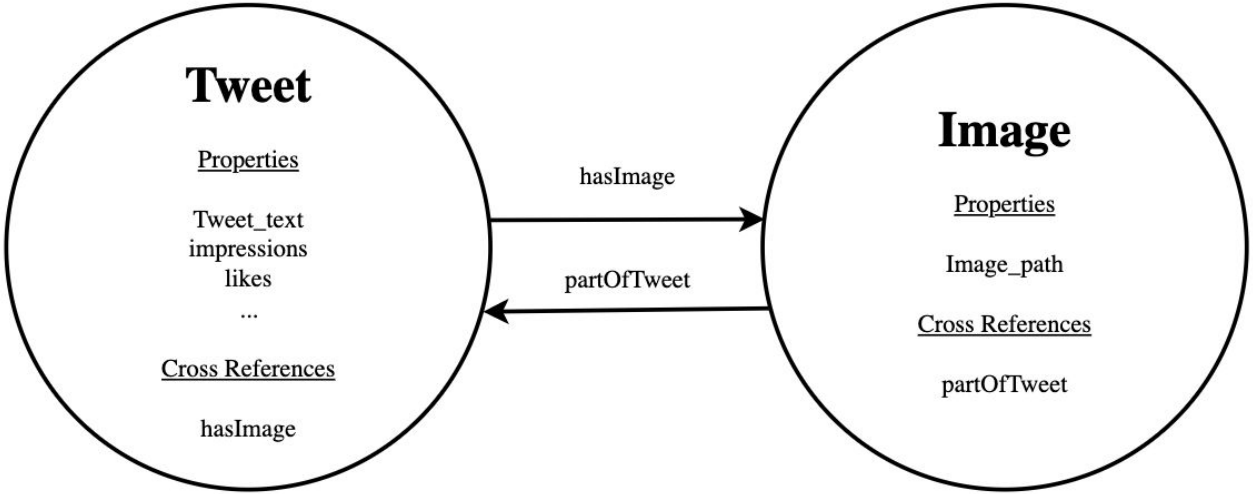
WIKIPEDIA
The Free Encyclopedia

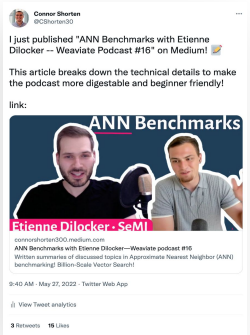
SCAN ME

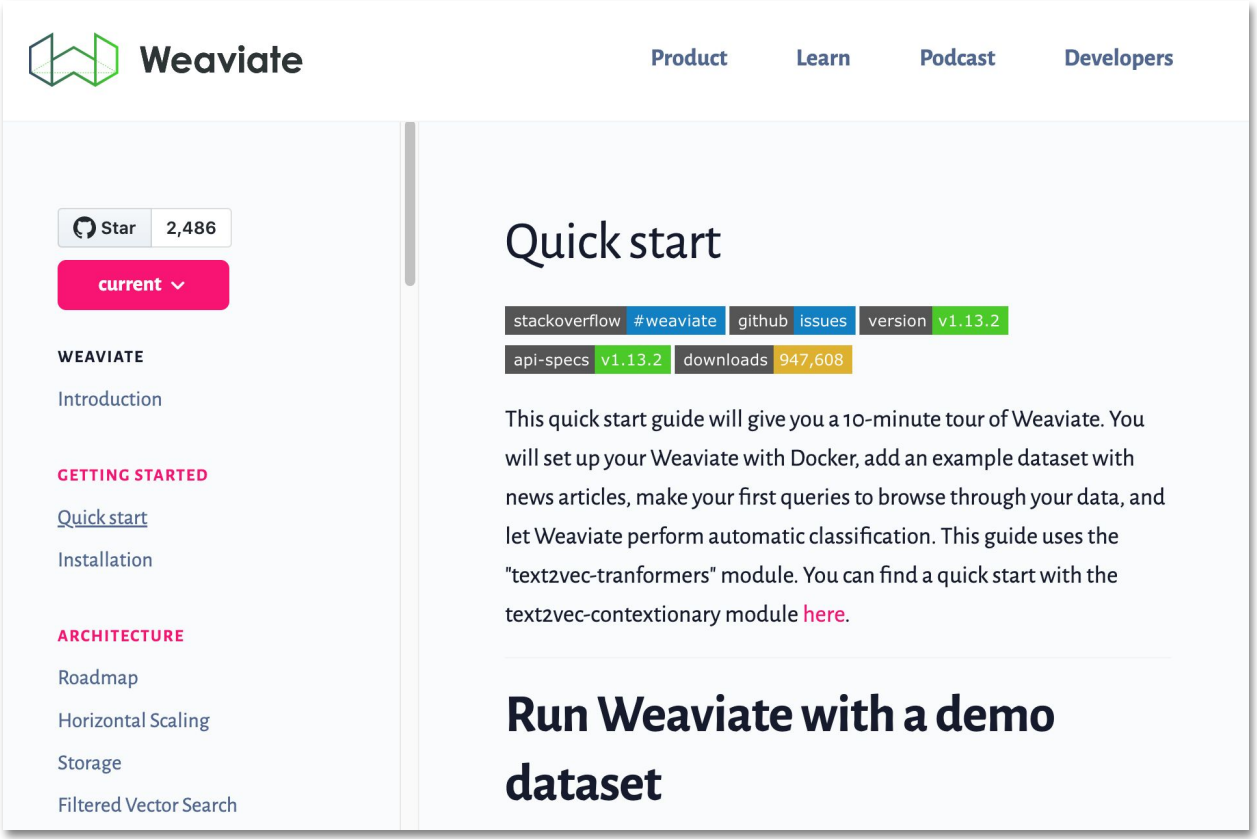
- Weaviate is a Vector Search **Database**, rather than a **Library** such as Facebook's FAISS or ANNOY from Spotify

- Weaviate has a **Graph-like Data Model**

Expanding Twitter project with Graph Model






Weaviate
Product
Learn
Podcast
Developers
Star 2,486
current
WEAVIATE
Introduction
GETTING STARTED
Quick start
Installation
ARCHITECTURE
Roadmap
Horizontal Scaling
Storage
Filtered Vector Search
Quick start
stackoverflow #weaviate
github issues
version v1.13.2
api-specs v1.13.2
downloads 947,608
This quick start guide will give you a 10-minute tour of Weaviate. You will set up your Weaviate with Docker, add an example dataset with news articles, make your first queries to browse through your data, and let Weaviate perform automatic classification. This guide uses the "text2vec-tranformers" module. You can find a quick start with the text2vec-contextionary module here.
Run Weaviate with a demo dataset

Summary of Takeaway #4
Weaviate for Twitter Analytics

We can segment **Impressions on Twitter** based on the content of the tweet **without manual labeling!**

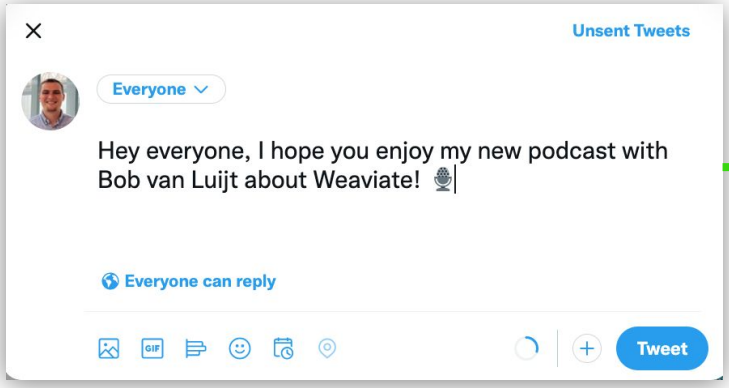
Weaviate is a **Vector Search Database** that can be used to store and search through semantic embeddings of data.

Key Takeaway #5 - **Research Questions and Discussion**

Research Questions and Discussion

- Should I fine-tune my embedding model?
- Large-Scale Vector Search with Approximate Nearest Neighbor (ANN) Algorithms
- How does Vector Search differ from Classification or Regression models?

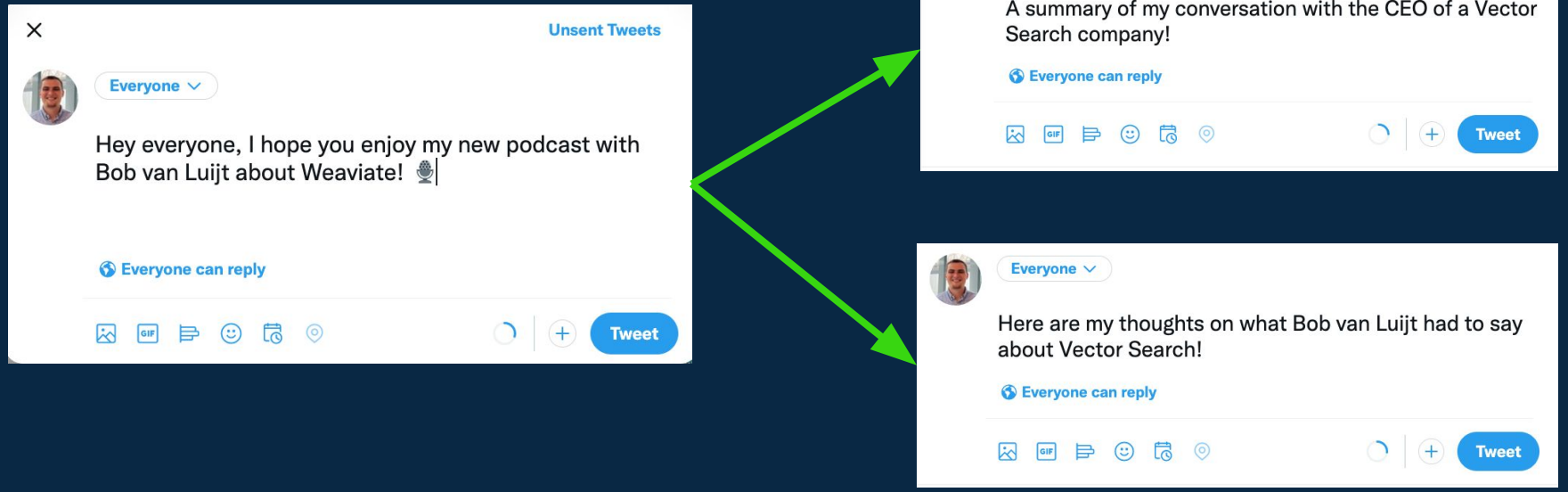
Vector Search versus Regression on Impressions



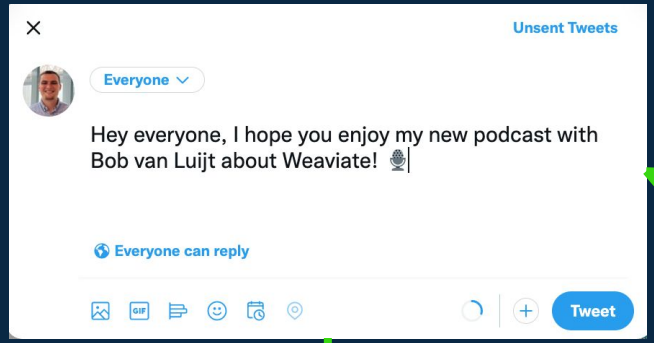
Model Prediction

8,530 **Impressions**

Interpretability of Vector Search



Interpretability of Vector Search and Prediction



Model Prediction

8,530 **Impressions**

What do we want to know about our Tweets?

Should I post this?

When might be a better time to post it?

What might be a better phrasing of this tweet?

Expanding from individuals to teams

- Has anyone on my team tweeted something like this recently?
- Who on our team would be best fit to tell this story?
- What topics should we be tweeting about?

Summary of Takeaway #5 Research Questions and Discussion

How can we **improve** these systems? What looks **promising**?

Key Takeaways: **“Vector Search for Data Scientists”**

1. Segmentation in Data Science
2. Vector Representations of Unstructured Data
3. Vector Segmentation
4. Weaviate Example for Twitter Analytics
5. Research Questions and Discussion

Slides, Colab Notebook, Video Presentation available on:
github.com/CShorten/Vector-Search-for-Data-Scientists

Connect with us!

Weaviate Slack Channel

YouTube: Weaviate · Vector Search Engine

Weaviate Podcast

Twitter @weaviate_io

Thank you for Watching!

Special thanks to **Sebastian Witalec** in advising the development of this presentation and **Svitlana Smolianova** for visual styling.