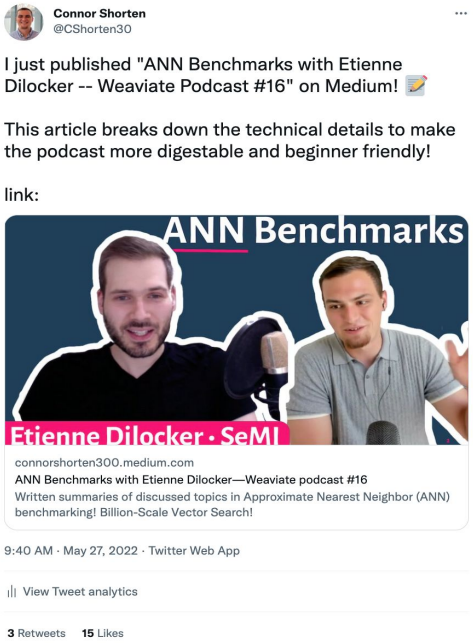
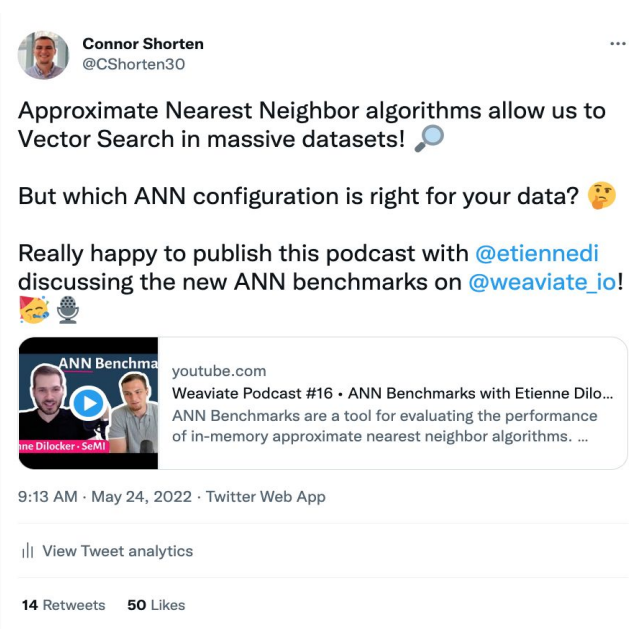
Vector Search for Data Scientists

A Case Study with Weaviate Twitter Analytics

Twitter Analytics





Overview

- Segmentation in Data Science
- Vector Representations of Data
- Vector Search for Semantic Segmentation
- Weaviate Example for Twitter and YouTube Analytics
  - How to load Pandas or CSV data into Weaviate
  - Nearest Neighbor Tweets
  - Cluster Analysis with Classification Probe
- Future Work
  - Fine-tuning, Retrieval or Re-Ranking?

Descriptive Statistics

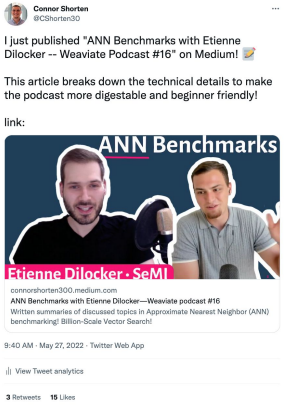
How is my data distributed?

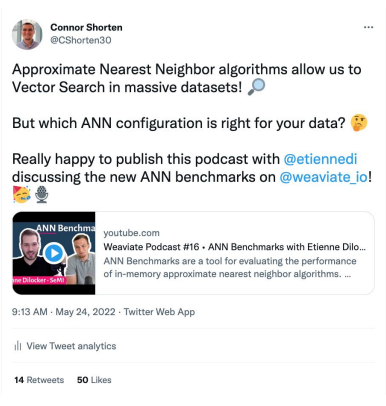
Are there any outliers in my data?

Are my variables correlated with each other?

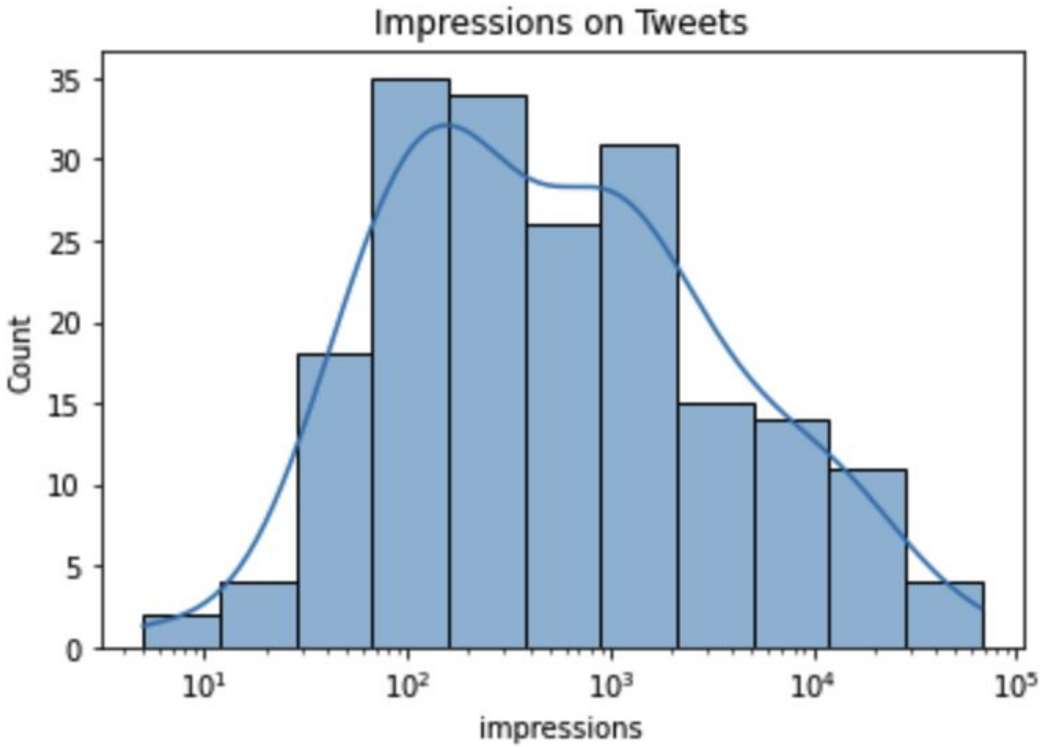
Twitter Analytics: *This presentation will utilize Twitter Analytics data as an example of Data Science and Vector Search*

Tweet Text	Time	Impressions	Engagements	Engagement Rate	Retweets	Replies	Likes	User Profile Clicks	Url Clicks
I just published "ANN…	May 27th, 1:34pm	1905	50	2.6%	3	1	15	2	18
Approximate Nearest Neighbor…	May 24th, 1:13pm	7182	252	3.5%	14	1	50	27	36

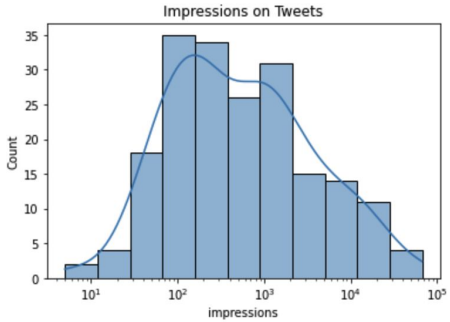


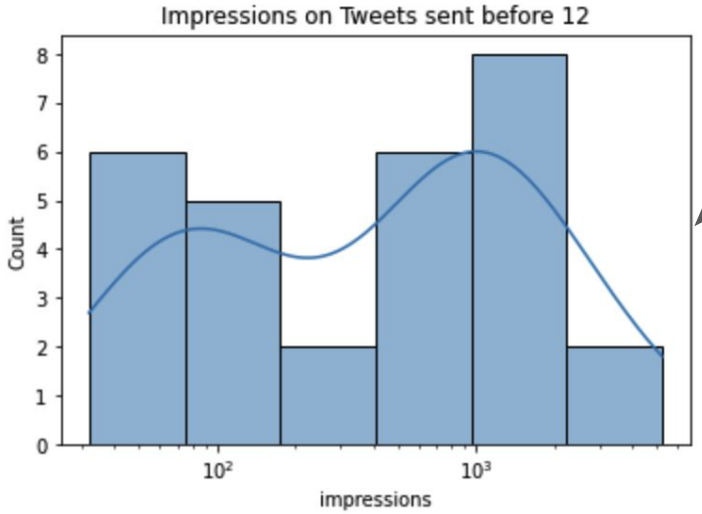


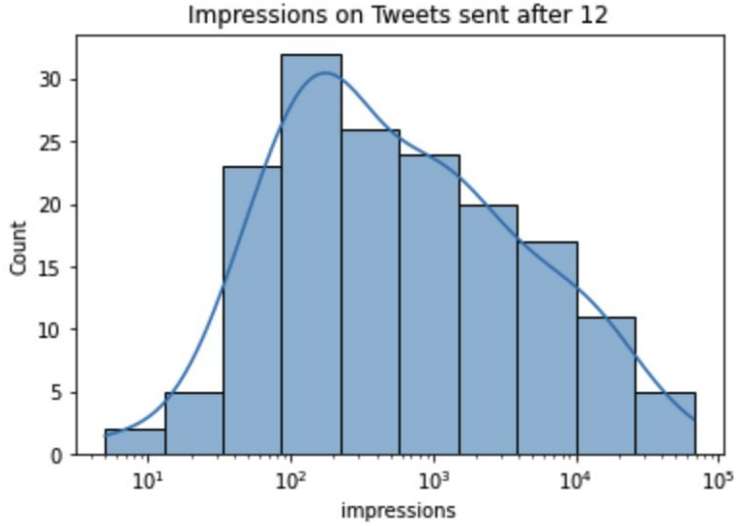
Visualizing Distributions of Values



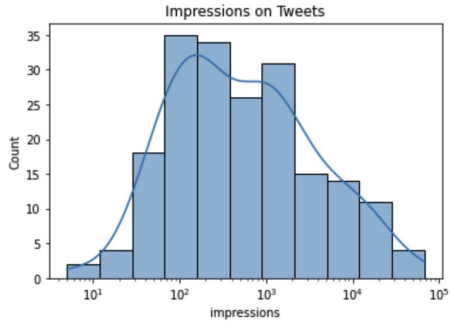
Segmenting Analytics with Symbolic Attributes

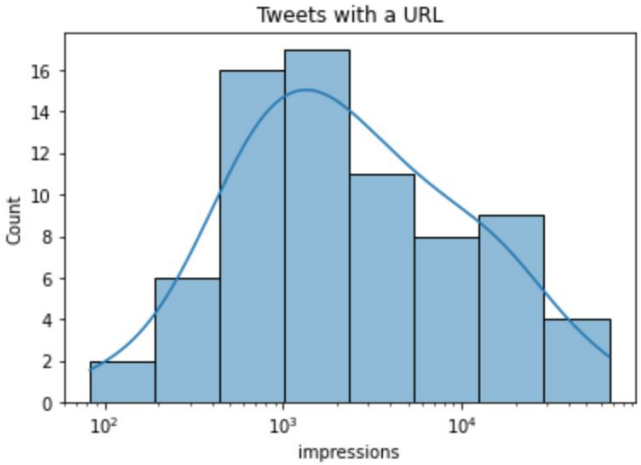


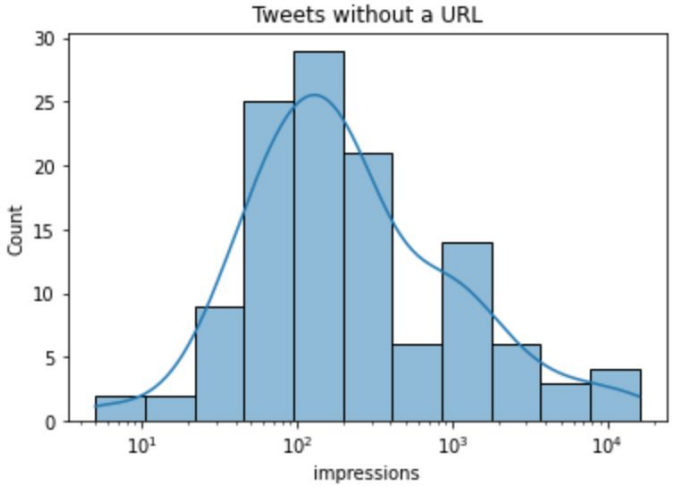




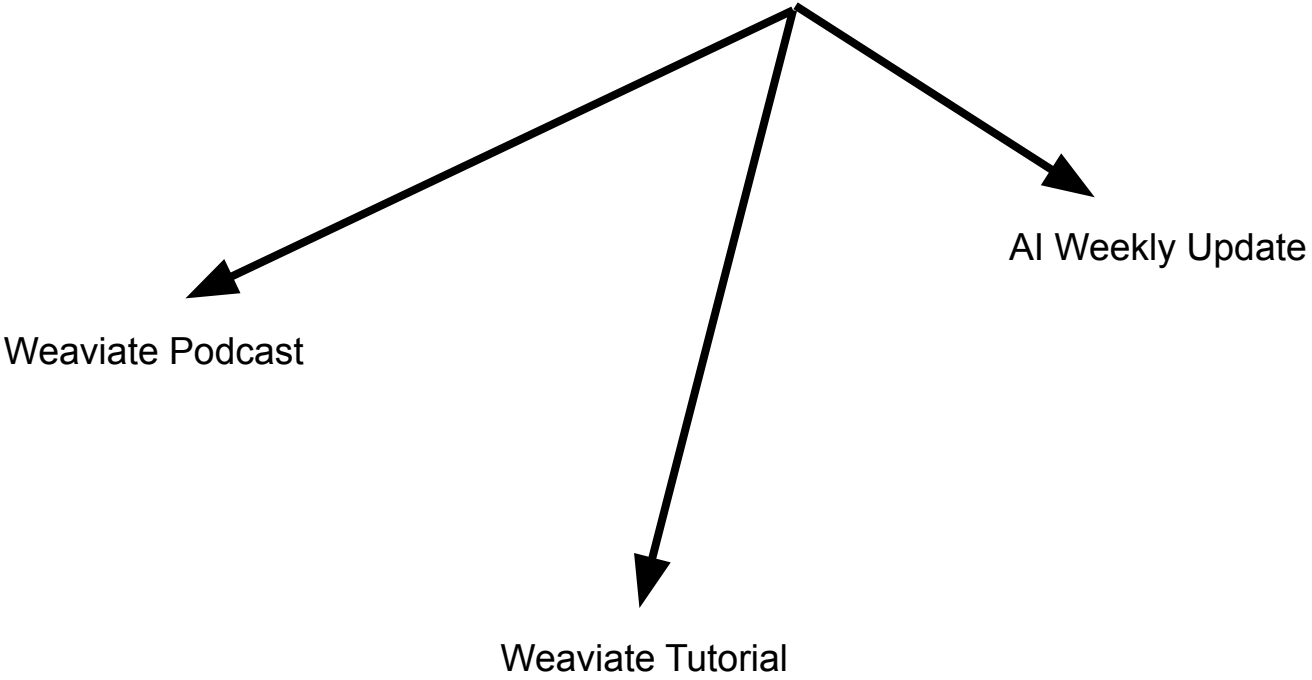
Segmenting Analytics with Symbolic Attributes







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

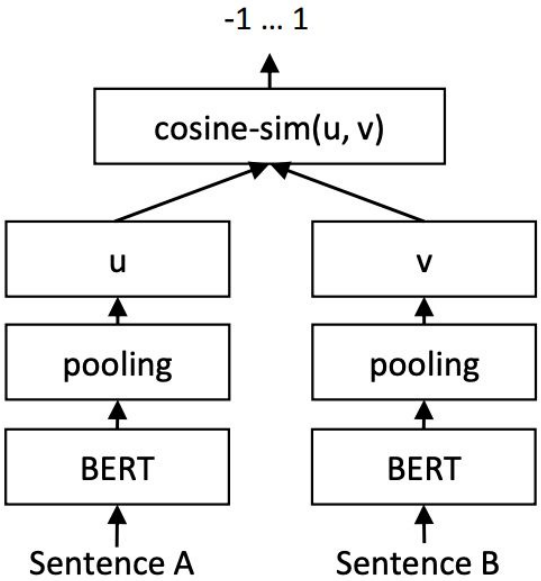


Segmentation from Unstructured Data

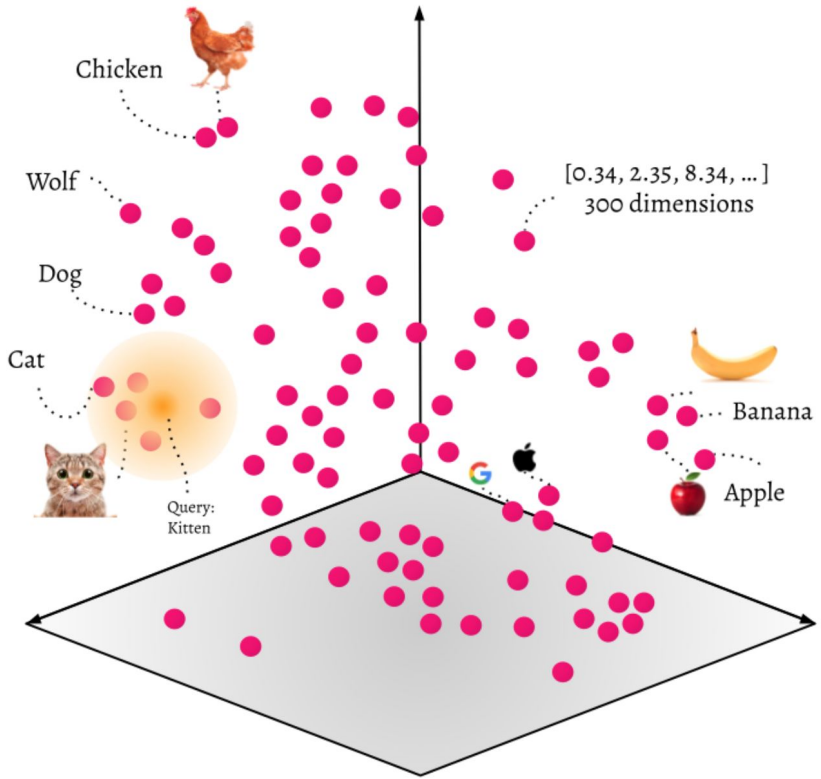
Images, Raw text, Audio, Video, Graph-Structure, Biological Sequences, ...

Semantic Similarity with Vector Representations

Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks - Nils Reimers and Iryna Gurevych 2019



Segmentation with Vector Similarity



Semantic Similarity with Vector Representations

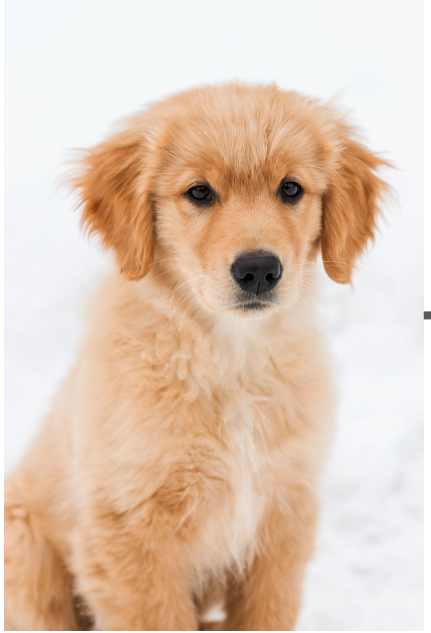
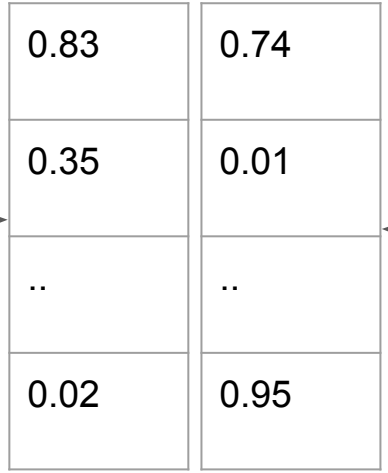


Photo by Shayna Douglas on Unsplash



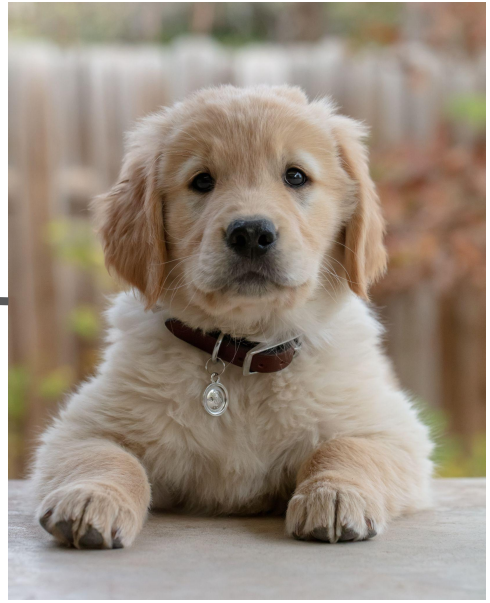
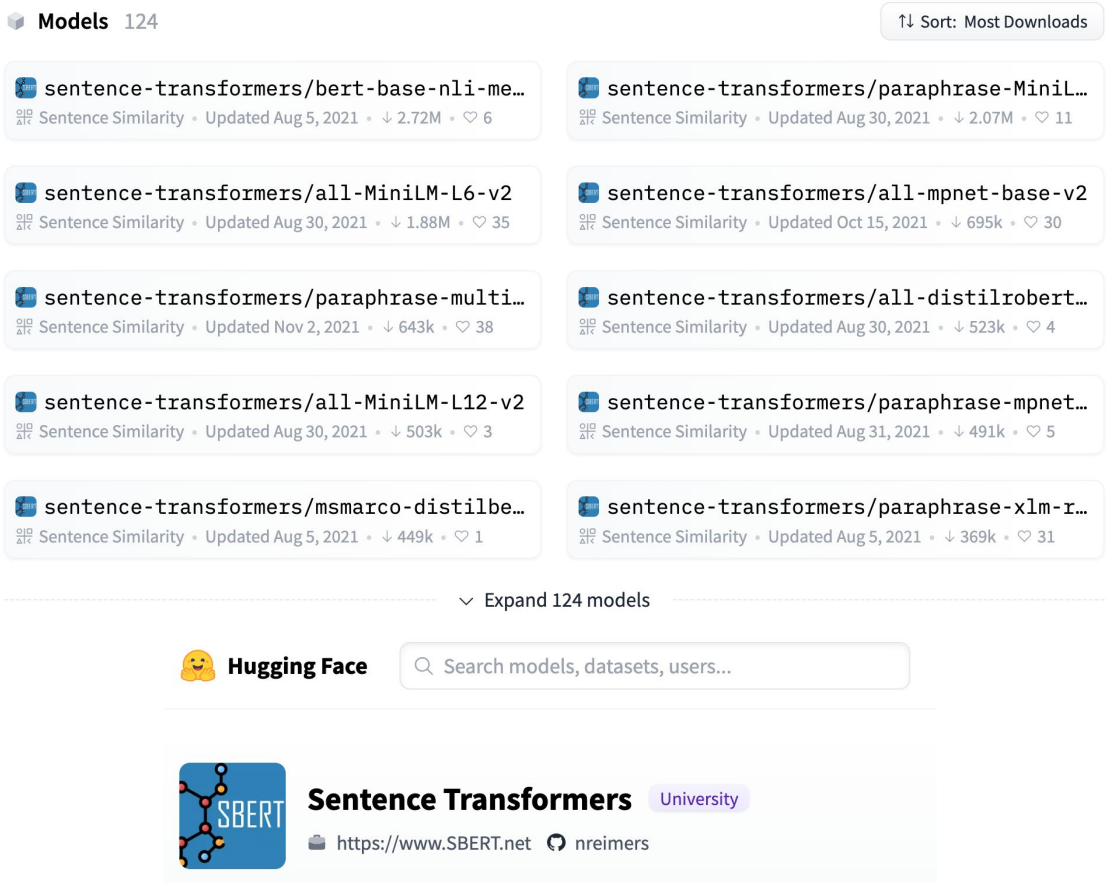


Photo by Bill Stephan on Unsplash

Do we need to train our own models?

No! Open-source pre-trained models work very well for a broad range of data!

Semantic Similarity with Vector Representations



Weaviate for Twitter Analytics

What do we want to know about our Tweets?

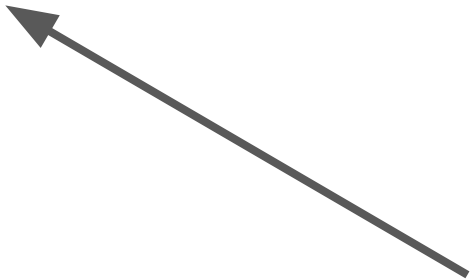
Should I post this?

When might be a better time to post it?

What might be a better phrasing of this tweet?

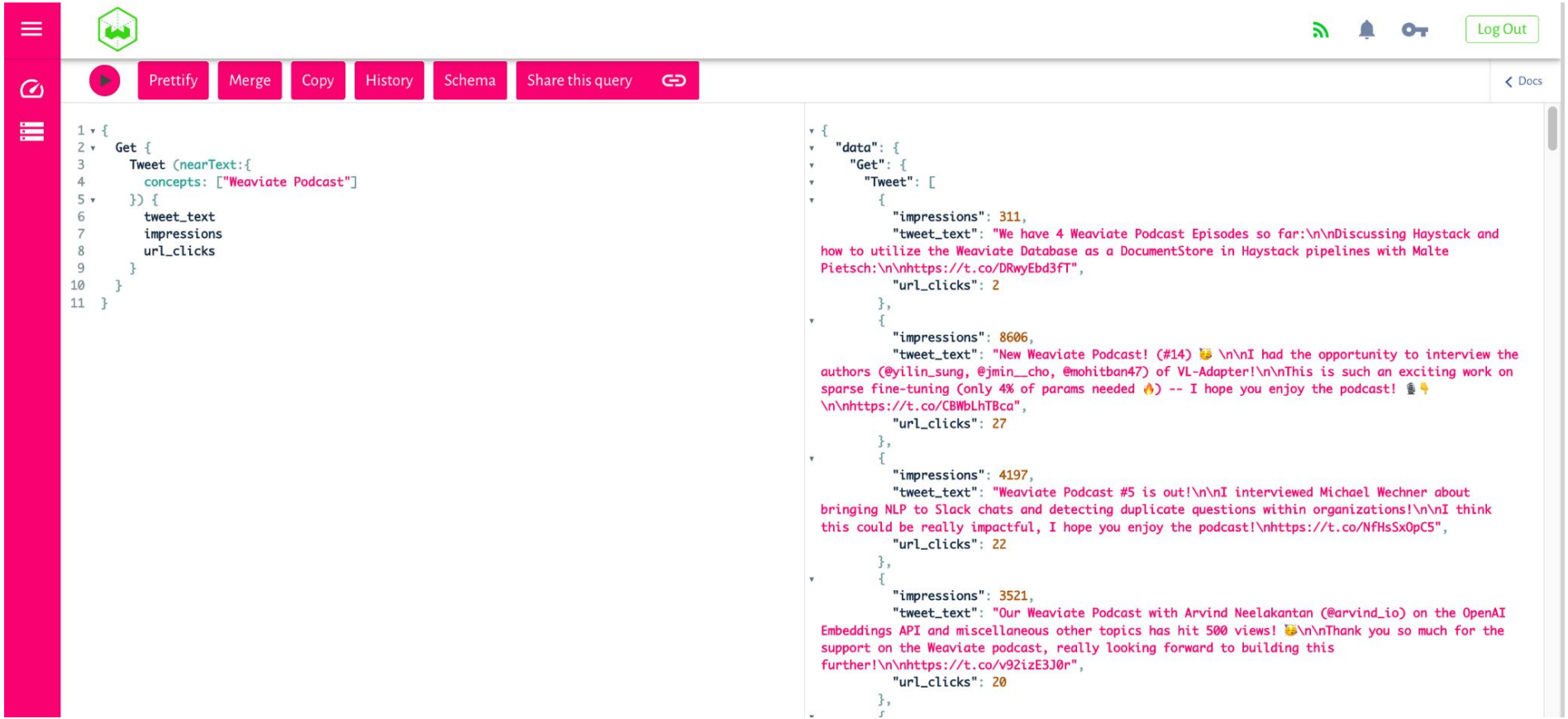
Twitter Analytics CSV

Tweet text	time	impressions	engagements	Engagement rate	retweets	replies	likes	User profile clicks	Url clicks

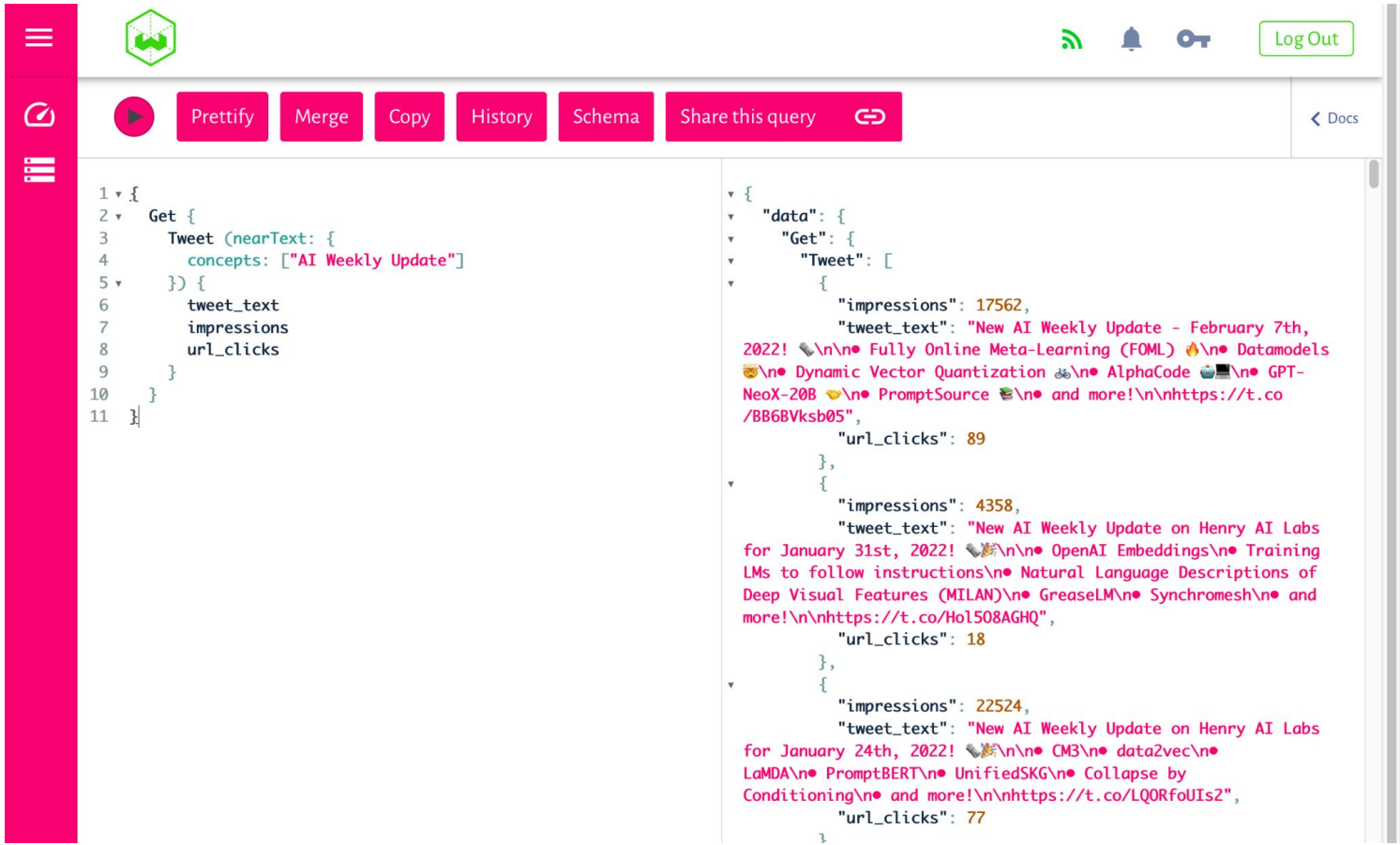


Column to be vectorized with a pre-trained sentence transformer

Classify with Label Embeddings



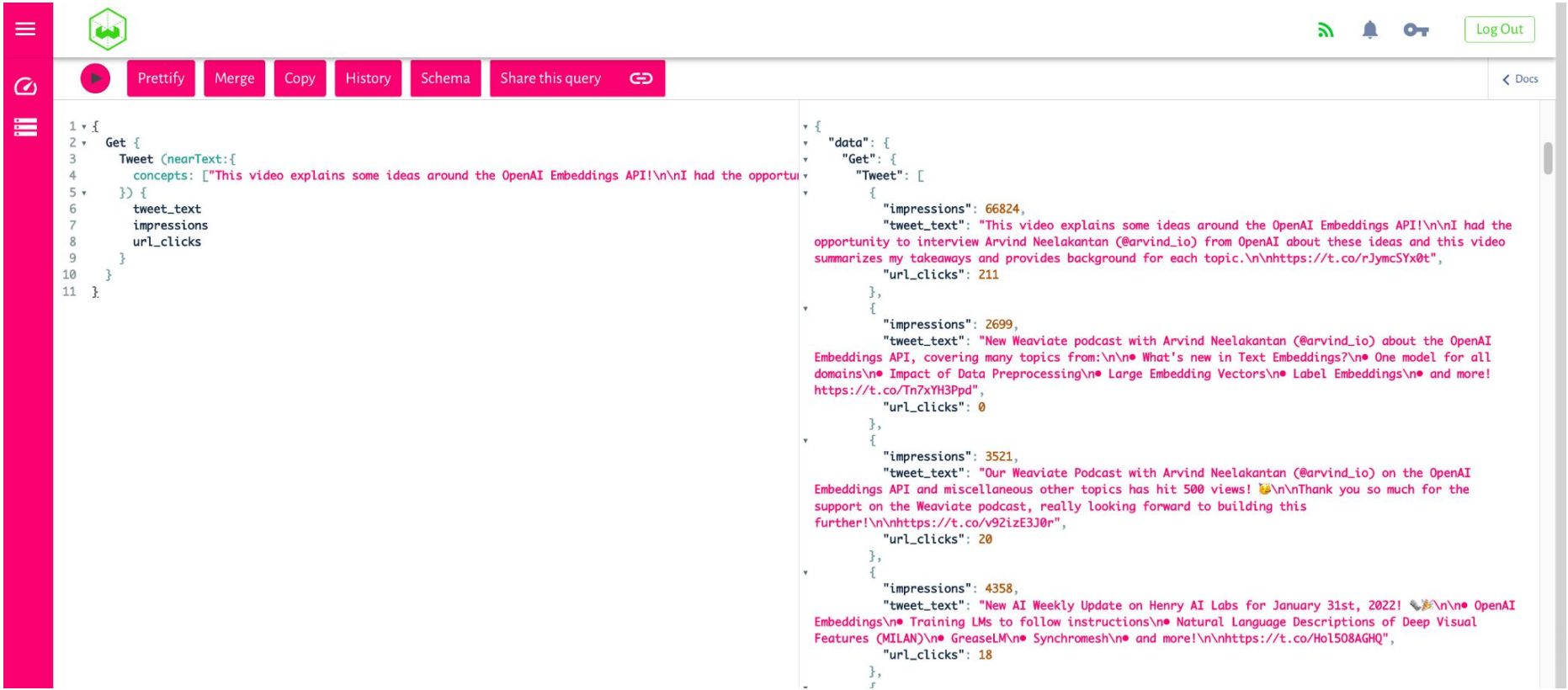
Query in Weaviate



Add histogram for
“Weaviate Podcast” -- “Vector Search”

Have I tweeted something like this before?

Have I tweeted something like this before?



Has anyone tweeted something like this before?

Work in Progress

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?

Technical Details of how this is setup

Pandas DataFrame → Weaviate

A look under the hood of client.from_pandas

Weaviate Schema Setup

```
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 }
            }
        },
        …
```

Batch upload

```
Def add_tweet(batch: Batch, data: dict) -> str:
    Tweet_object = {
        “Tweet_text”: data[“tweet_text”],
        “Hour”: data[“hour”],
        …
    }
    batch.add_data_object(
        Data_object = tweet_object,
        Class_name = “Tweet”,
        Uuid = tweet_id
    )
```

Discussion Topics

- How does Vector Search differ from Classification or Regression models?

- Should I fine-tune my embedding model?

- Large-Scale Vector Search with Approximate Nearest Neighbors (ANN)

Thank you!