


Vector Search for Data Scientists

A Case Study with Weaviate Twitter Analytics


Twitter Analytics

 **Connor Shorten**
@CShorten30

I just published "ANN Benchmarks with Etienne Diloocker -- Weaviate Podcast #16" on Medium! 📝

This article breaks down the technical details to make the podcast more digestible and beginner friendly!

link:



Etienne Diloocker · SeMI

connorshorten300.medium.com


ANN Benchmarks with Etienne Diloocker—Weaviate podcast #16

Written summaries of discussed topics in Approximate Nearest Neighbor (ANN) benchmarking! Billion-Scale Vector Search!

9:40 AM · May 27, 2022 · Twitter Web App

||| View Tweet analytics

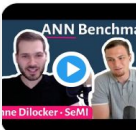
3 Retweets 15 Likes

 **Connor Shorten**
@CShorten30

Approximate Nearest Neighbor algorithms allow us to Vector Search in massive datasets! 🔍

But which ANN configuration is right for your data? 🤔

Really happy to publish this podcast with @etiennedi discussing the new ANN benchmarks on @weaviate_io! 🎧

 youtube.com
Weaviate Podcast #16 · ANN Benchmarks with Etienne Dilo...
ANN Benchmarks are a tool for evaluating the performance of in-memory approximate nearest neighbor algorithms. ...

9:13 AM · May 24, 2022 · Twitter Web App

||| View Tweet analytics

14 Retweets 50 Likes

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

 Connor Shorten
@CShorten30

I just published "ANN Benchmarks with Etienne Diloquer -- Weaviate Podcast #16" on Medium! 🎧

This article breaks down the technical details to make the podcast more digestible and beginner friendly!

link:




ANN Benchmarks
Etienne Diloquer - SeMI

connorshorten300.medium.com
ANN Benchmarks with Etienne Diloquer -- Weaviate podcast #16
Written summaries of discussed topics in Approximate Nearest Neighbor (ANN) benchmarking! Billion-Scale Vector Search!

9:40 AM · May 27, 2022 · Twitter Web App

|| View Tweet analytics


3 Retweets 15 Likes

 Connor Shorten
@CShorten30

Approximate Nearest Neighbor algorithms allow us to Vector Search in massive datasets! 🗨️

But which ANN configuration is right for your data? 🤔

Really happy to publish this podcast with @etiennedi discussing the new ANN benchmarks on @weaviate_io!



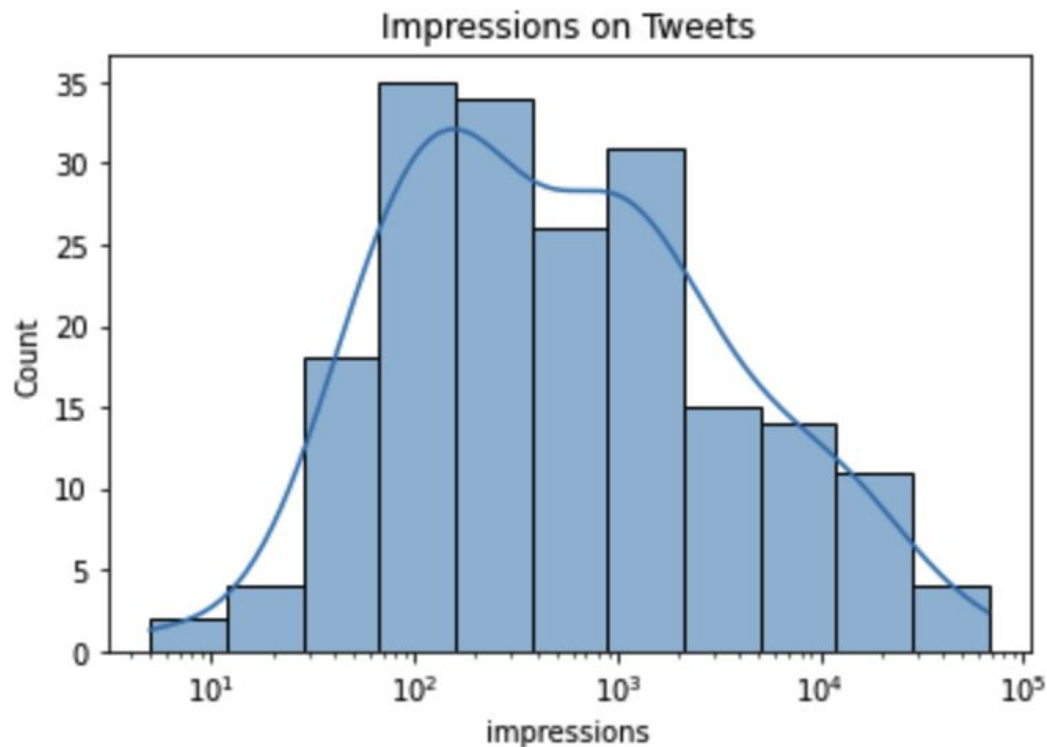
youtube.com
Weaviate Podcast #16 - ANN Benchmarks with Etienne Diloquer
ANN Benchmarks are a tool for evaluating the performance of in-memory approximate nearest neighbor algorithms. ...

9:13 AM · May 24, 2022 · Twitter Web App

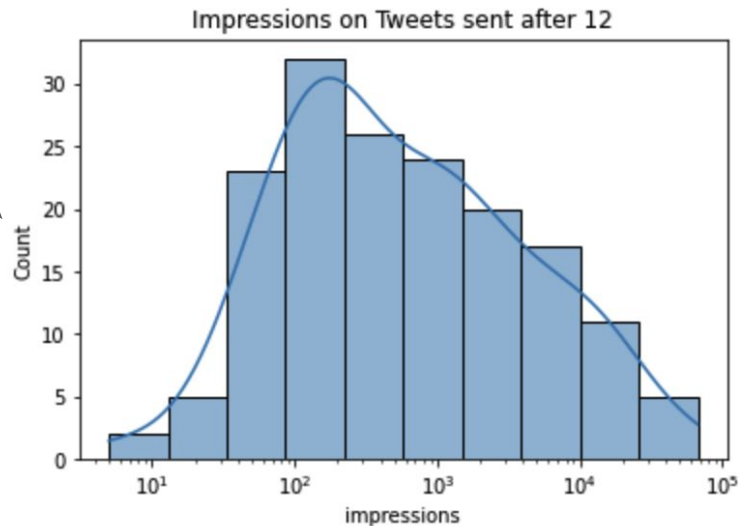
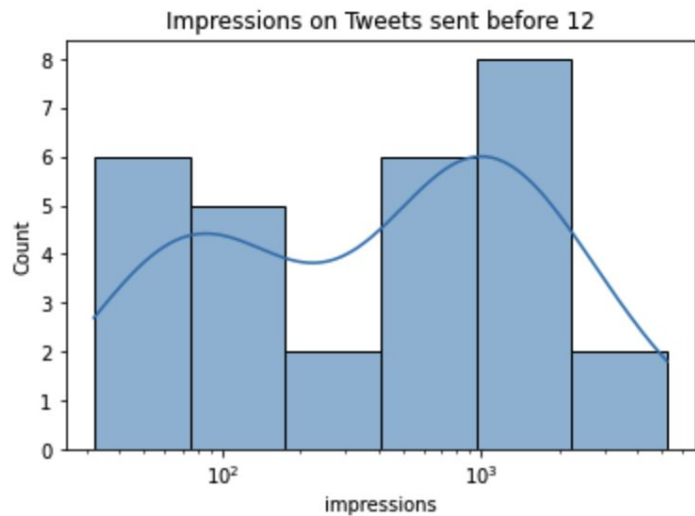
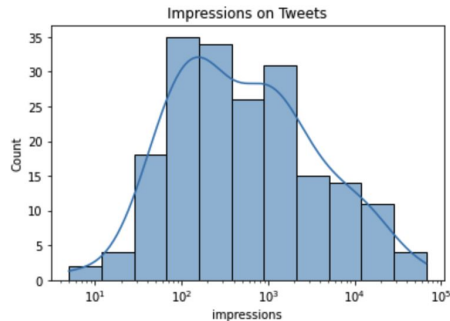
|| View Tweet analytics

14 Retweets 50 Likes

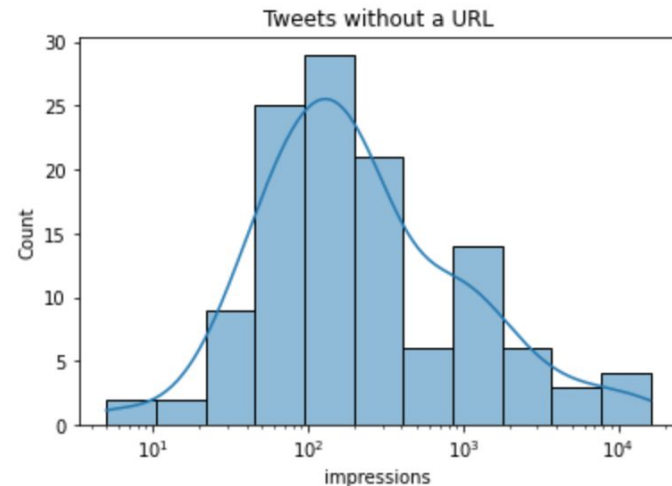
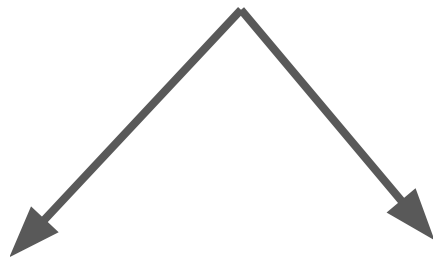
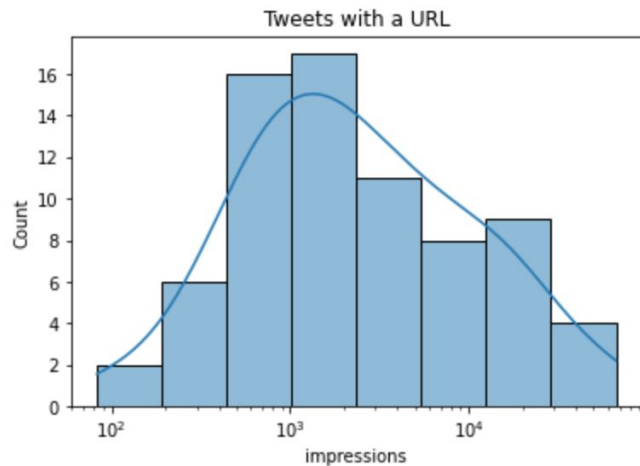
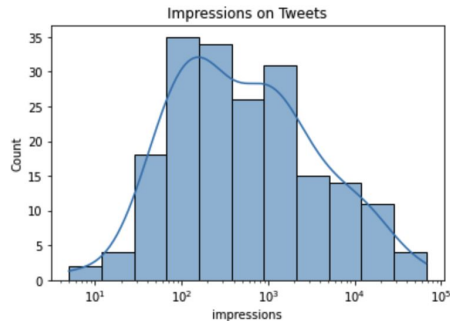
Visualizing Distributions of Values



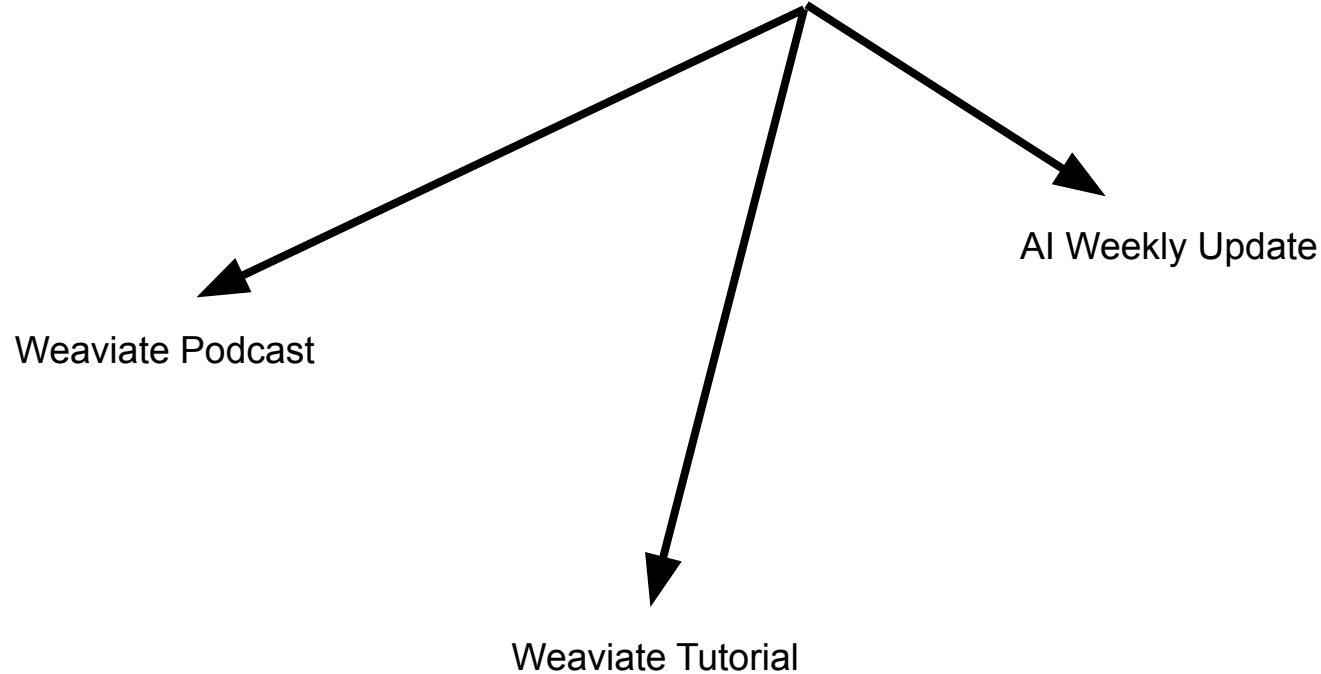
Segmenting Analytics with Symbolic Attributes



Segmenting Analytics with Symbolic Attributes



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

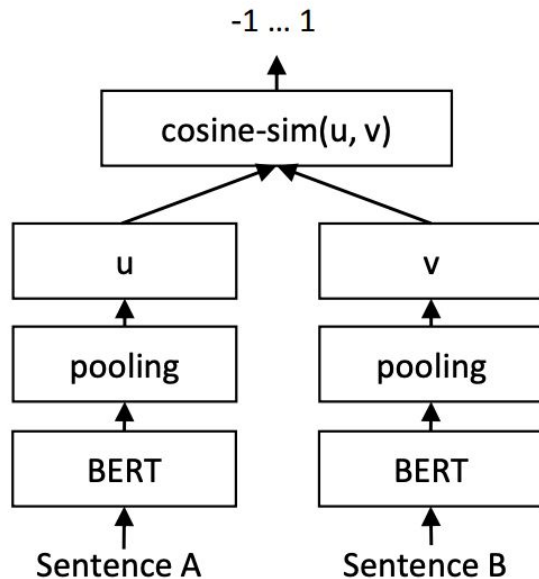


Segmentation from Unstructured Data

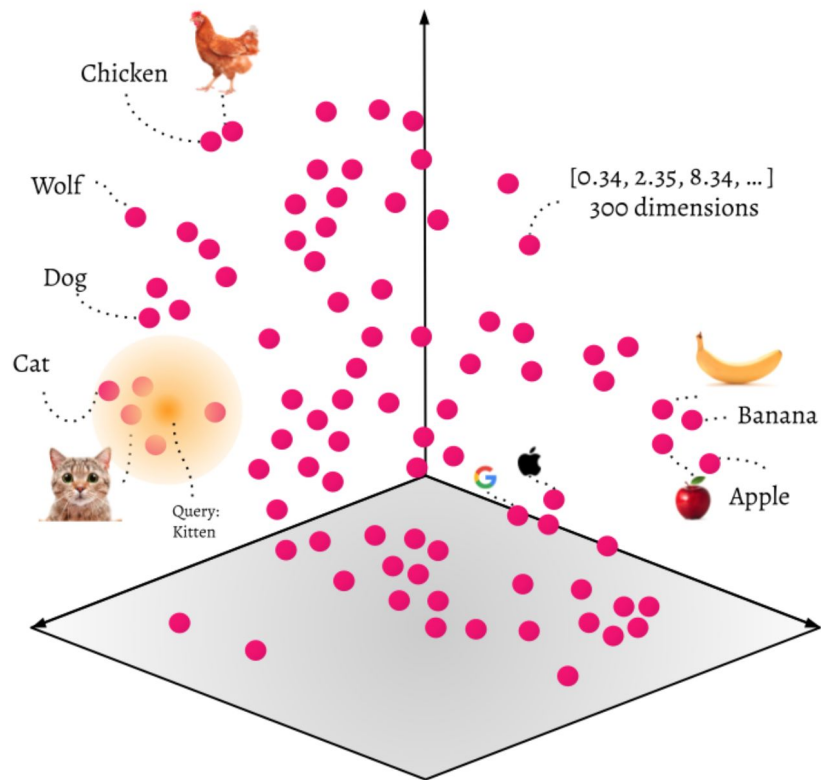
Images, Raw text, Audio, Video,
Graph-Structure, Biological
Sequences, ...

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



0.83	0.74
0.35	0.01
..	..
0.02	0.95

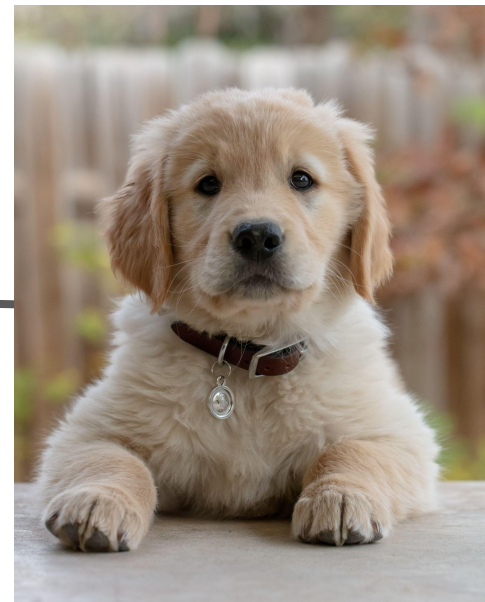


Photo by [Shayna Douglas](#) on [Unsplash](#)

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

 **Models** 124

Sort: Most Downloads

 **sentence-transformers/bert-base-nli-me...**

  Sentence Similarity • Updated Aug 5, 2021 • ↓ 2.72M • ♥ 6

 **sentence-transformers/paraphrase-MiniL...**

  Sentence Similarity • Updated Aug 30, 2021 • ↓ 2.07M • ♥ 11

 **sentence-transformers/all-MiniLM-L6-v2**

  Sentence Similarity • Updated Aug 30, 2021 • ↓ 1.88M • ♥ 35

 **sentence-transformers/all-mpnet-base-v2**

  Sentence Similarity • Updated Oct 15, 2021 • ↓ 695k • ♥ 30

 **sentence-transformers/paraphrase-multi...**

  Sentence Similarity • Updated Nov 2, 2021 • ↓ 643k • ♥ 38

 **sentence-transformers/all-distilrobert...**

  Sentence Similarity • Updated Aug 30, 2021 • ↓ 523k • ♥ 4

 **sentence-transformers/all-MiniLM-L12-v2**

  Sentence Similarity • Updated Aug 30, 2021 • ↓ 503k • ♥ 3

 **sentence-transformers/paraphrase-mpnet...**

  Sentence Similarity • Updated Aug 31, 2021 • ↓ 491k • ♥ 5

 **sentence-transformers/msmarco-distilbe...**

  Sentence Similarity • Updated Aug 5, 2021 • ↓ 449k • ♥ 1

 **sentence-transformers/paraphrase-xlm-r...**

  Sentence Similarity • Updated Aug 5, 2021 • ↓ 369k • ♥ 31

Expand 124 models



Hugging Face

Search models, datasets, users...



Sentence Transformers

University

 <https://www.SBERT.net>  [nreimers](https://github.com/nreimers)

Weaviate for Twitter Analytics

*What do we want to know about our
Tweets?*

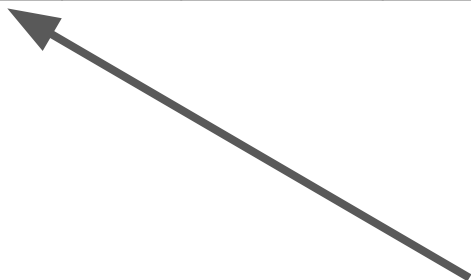
Should I post this?

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

[Log Out](#)[Prettify](#)[Merge](#)[Copy](#)[History](#)[Schema](#)[Share this query](#)[< Docs](#)



```
1 {  
2   Get {  
3     Tweet (nearText:{  
4       concepts: ["Weaviate Podcast"]  
5     }) {  
6       tweet_text  
7       impressions  
8       url_clicks  
9     }  
10  }  
11 }
```

```
{  
  "data": {  
    "Get": {  
      "Tweet": [  
        {  
          "impressions": 311,  
          "tweet_text": "We have 4 Weaviate Podcast Episodes so far:\n\nDiscussing Haystack and  
how to utilize the Weaviate Database as a DocumentStore in Haystack pipelines with Malte  
Pietsch:\n\nhttps://t.co/DRwyEbd3FT",  
          "url_clicks": 2  
        },  
        {  
          "impressions": 8606,  
          "tweet_text": "New Weaviate Podcast! (#14) 🤖 \n\nI had the opportunity to interview the  
authors (@yilin_sung, @jmin_cho, @mohitban47) of VL-Adapter!\n\nThis is such an exciting work on  
sparse fine-tuning (only 4% of params needed 🙌) -- I hope you enjoy the podcast! 🎧🔊  
\n\nhttps://t.co/CBwbLhTBca",  
          "url_clicks": 27  
        },  
        {  
          "impressions": 4197,  
          "tweet_text": "Weaviate Podcast #5 is out!\n\nI interviewed Michael Wechner about  
bringing NLP to Slack chats and detecting duplicate questions within organizations!\n\nI think  
this could be really impactful, I hope you enjoy the podcast!\n\nhttps://t.co/NfHsX0pC5",  
          "url_clicks": 22  
        },  
        {  
          "impressions": 3521,  
          "tweet_text": "Our Weaviate Podcast with Arvind Neelakantan (@arvind_io) on the OpenAI  
Embeddings API and miscellaneous other topics has hit 500 views! 🤖\n\nThank you so much for the  
support on the Weaviate podcast, really looking forward to building this  
further!\n\nhttps://t.co/v92izE3J0r",  
          "url_clicks": 20  
        }  
      ]  
    }  
  }  
}
```

Query in Weaviate



Log Out



Prettify


Merge

Copy

History

Schema

Share this query



< Docs






```
1 {
2   Get {
3     Tweet (nearText: {
4       concepts: ["AI Weekly Update"]
5     }) {
6       tweet_text
7       impressions
8       url_clicks
9     }
10  }
11 }
```

```
{
  "data": {
    "Get": {
      "Tweet": [
        {
          "impressions": 17562,
          "tweet_text": "New AI Weekly Update - February 7th, 2022! 🤖\n\n• Fully Online Meta-Learning (FOML) 🔥\n\n• Datamodels 🧠\n\n• Dynamic Vector Quantization 🚲\n\n• AlphaCode 🏆\n\n• GPT-NeoX-20B 🍌\n\n• PromptSource 🧠\n\n• and more!\n\nhttps://t.co/BB6BVksb05",
          "url_clicks": 89
        },
        {
          "impressions": 4358,
          "tweet_text": "New AI Weekly Update on Henry AI Labs for January 31st, 2022! 🤖\n\n• OpenAI Embeddings\n\n• Training LMs to follow instructions\n\n• Natural Language Descriptions of Deep Visual Features (MILAN)\n\n• GreaseLM\n\n• Synchromesh\n\n• and more!\n\nhttps://t.co/Hol508AGHQ",
          "url_clicks": 18
        },
        {
          "impressions": 22524,
          "tweet_text": "New AI Weekly Update on Henry AI Labs for January 24th, 2022! 🤖\n\n• CM3\n\n• data2vec\n\n• LaMDA\n\n• PromptBERT\n\n• UnifiedSKG\n\n• Collapse by Conditioning\n\n• and more!\n\nhttps://t.co/LQORfoUIs2",
          "url_clicks": 77
        }
      ]
    }
  }
}
```

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

Have I tweeted something like this
before?

Have I tweeted something like this before?

Log Out

▶ Prettify Merge Copy History Schema Share this query ↻

< Docs

```
1 {  
2   Get {  
3     Tweet (nearText:{  
4       concepts: ["This video explains some ideas around the OpenAI Embeddings API!\n\nI had the opportu  
5     }) {  
6       tweet_text  
7       impressions  
8       url_clicks  
9     }  
10  }  
11 }
```

```
{  
  "data": {  
    "Get": {  
      "Tweet": [  
        {  
          "impressions": 66824,  
          "tweet_text": "This video explains some ideas around the OpenAI Embeddings API!\n\nI had the  
          opportunity to interview Arvind Neelakantan (@arvind_io) from OpenAI about these ideas and this video  
          summarizes my takeaways and provides background for each topic.\n\nhttps://t.co/rJymcSYx0t",  
          "url_clicks": 211  
        },  
        {  
          "impressions": 2699,  
          "tweet_text": "New Weaviate podcast with Arvind Neelakantan (@arvind_io) about the OpenAI  
          Embeddings API, covering many topics from:\n\nWhat's new in Text Embeddings?\n\nOne model for all  
          domains\n\nImpact of Data Preprocessing\n\nLarge Embedding Vectors\n\nLabel Embeddings\n\nand more!  
          https://t.co/Tn7xYH3Ppd",  
          "url_clicks": 0  
        },  
        {  
          "impressions": 3521,  
          "tweet_text": "Our Weaviate Podcast with Arvind Neelakantan (@arvind_io) on the OpenAI  
          Embeddings API and miscellaneous other topics has hit 500 views! 🥳\n\nThank you so much for the  
          support on the Weaviate podcast, really looking forward to building this  
          further!\n\nhttps://t.co/v92izE3J0r",  
          "url_clicks": 20  
        },  
        {  
          "impressions": 4358,  
          "tweet_text": "New AI Weekly Update on Henry AI Labs for January 31st, 2022! 🥳\n\nOpenAI  
          Embeddings\n\nTraining LMs to follow instructions\n\nNatural Language Descriptions of Deep Visual  
          Features (MILAN)\n\nGreaseLM\n\nSynchromesh\n\nand more!\n\nhttps://t.co/Hol508AGHQ",  
          "url_clicks": 18  
        }  
      ]  
    }  
  }  
}
```

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!