towards
data science

510K Followers · About

Cutting edge semantic search and sentence similarity

Semantic search is a hard problem worth solving in NLP.

Daulet Nurmanbetov May 4 · 9 min read ★

Photo by Anthony Martino on Unsplash

We commonly spend a lot of time looking for a specific piece of information in a large document. And we commonly find if using CTRL + F. The proverbial Google-fu, the art of effectively searching for information on google is a valuable skill in a 21st-century workplace. All of humanity's knowledge is available to us, it is a matter of asking the right question, and knowing how to skim through results to find the relevant answer.

Our brains perform a semantic search, where we review the results and find sentences that are similar to our search query. This is especially true in finance and legal professions as documents get lengthy and we have to resort to searching many keywords to find the right sentence or a passage. To this day, the cumulative human effort spent on discovery is staggering.

Machine learning has been trying to solve this problem of semantic search since the dawn of NLP. A whole area of study -Semantic Search has emerged. And recently thanks to advancements in Deep Learning computers are able to accurately surface relevant information to us with minimal human involvement.

Sentence embedding methods

Natural Language Processing (NLP) field has a term for this, when a word is mentioned we call it a "surface form" take for example the word "*president*" by itself this means the head of the country. But depending on context and time it could mean Trump or Obama.

Advancements in NLP allow us to effectively map these surface forms and capture the context in those words into something called "embeddings" Embeddings are commonly a vector of numbers which have certain peculiar characteristics. Two words with similar meaning would have similar vectors allowing us to compute vector similarities.

Extending this idea, in the vector space, we should be able to compute the similarity between any two sentences. And this is what sentence embedding models achieve. These models convert any given sentences into a vector to be able to quickly compute the similarity or dissimilarity of any pair of sentences.

State of the art Semantic Search — Finding most similar sentences

The idea is not new, The paper that started it all — word2vec proposed representing individual words with vectors back in 2013. However, we came a long way since then with BERT and other Transformer-based models which allow us to capture the context of those words much more effectively.

Here how we stack up on recent embedding models compared to word2vec or GloVe of the past.

STS is a competition of sentence meaning similarity for NLP. Higher is better. Table from SBERT paper

These modified and fine-tuned BERT NLP models are quite good at identifying similar sentences, much better than older predecessors. Let's take a look at what this means in a practical sense.

I have several article headlines from April 2020, And I wish to find the most similar sentences to a set of search terms.

Here are my search terms —

```
1. The economy is more resilient and improving.
2. The economy is in a lot of trouble.
3. Trump is hurting his own reelection chances.
```

And the article headlines I have are the following —

```
Coronavirus:
White House organizing program to slash development time for
coronavirus vaccine by as much as eight months (Bloomberg)
Trump says he is pushing FDA to approve emergency-use authorization
for Gilead's remdesivir (WSJ)
AstraZeneca to make an experimental coronavirus vaccine developed by
Oxford University (Bloomberg)
Trump contradicts US intel, says Covid-19 started in Wuhan lab. (The
Hill)
Reopening:
Inconsistent patchwork of state, local and business decision-making
on reopening raising concerns about a second wave of the coronavirus
(Politico)
White House risks backlash with coronavirus optimism if cases flare
up again (The Hill)
Florida plans to start reopening on Monday with restaurants and
retail in most areas allowed to resume business in most areas
(Bloomberg)
California Governor Newsom plans to order closure of all state
beaches and parks starting Friday due to concerns about overcrowding
(CNN)
Japan preparing to extend coronavirus state of emergency, which is
scheduled to end 6-May, by about another month (Reuters)
Policy/Stimulus:
Economists from a broad range of ideological backgrounds encouraging
Congress to keep spending to combat the coronavirus fallout and don't
believe now is time to worry about deficit (Politico)
Global economy:
China's official PMIs mixed with beat from services and miss from
manufacturing (Bloomberg)
China's Beige Book shows employment situation in Chinese factories
worsened in April from end of March, suggesting economy on less solid
ground than government data (Bloomberg)
```

```
Japan's March factory output fell at the fastest pace in five months,
while retail sales also dropped (Reuters)
Eurozone economy contracts by 3.8% in Q1, the fastest decline on
record (FT)
US-China:
Trump says China wants to him to lose his bid for re-election and
notes he is looking at different options in terms of consequences for
Beijing over the virus (Reuters)
Senior White House official confident China will meet obligations
under trad deal despite fallout from coronavirus pandemic (WSJ)
Oil:
Trump administration may announce plans as soon as today to offer
loans to oil companies, possibly in exchange for a financial stake
(Bloomberg)
Munchin says Trump administration could allow oil companies to store
another several hundred million barrels (NY Times)
Norway, Europe's biggest oil producer, joins international efforts to
cut supply for first time in almost two decades (Bloomberg)
IEA says coronavirus could drive 6% decline in global energy demand
in 2020 (FT)
Corporate:
Microsoft reports strong results as shift to more activities online
drives growth in areas from cloud-computing to video gams (WSJ)
Facebook revenue beats expectations and while ad revenue fell sharply
in March there have been recent signs of stability (Bloomberg)
Tesla posts third straight quarterly profit while Musk rants on call
about need for lockdowns to be lifted (Bloomberg)
eBay helped by online shopping surge though classifieds business hurt
by closure of car dealerships and lower traffic (WSJ)
Royal Dutch Shell cuts dividend for first time since World War II and
also suspends next tranche of buyback program (Reuters)
Chesapeake Energy preparing bankruptcy filing and has held
discussions with lenders about a ~$1B loan (Reuters)
Amazon accused by Trump administration of tolerating counterfeit
sales, but company says hit politically motivated (WSJ)
```

After running the similarity of each query to each embedding, here are the top 5 similar sentences for each of my search terms:

```
======================
Query: The economy is more resilient and improving.


Top 5 most similar sentences in corpus:
Microsoft reports strong results as shift to more activities online
drives growth in areas from cloud-computing to video gams (WSJ)
(Score: 0.5362)
Facebook revenue beats expectations and while ad revenue fell sharply
in March there have been recent signs of stability (Bloomberg)
(Score: 0.4632)
Senior White House official confident China will meet obligations
under trad deal despite fallout from coronavirus pandemic (WSJ)
(Score: 0.3558)
Economists from a broad range of ideological backgrounds encouraging
Congress to keep spending to combat the coronavirus fallout and don't
believe now is time to worry about deficit (Politico) (Score: 0.3052)
White House risks backlash with coronavirus optimism if cases flare
up again (The Hill) (Score: 0.2885)
======================
Query: The economy is in a lot of trouble.


Top 5 most similar sentences in corpus:
Inconsistent patchwork of state, local and business decision-making
on reopening raising concerns about a second wave of the coronavirus
(Politico) (Score: 0.4667)
eBay helped by online shopping surge though classifieds business hurt
by closure of car dealerships and lower traffic (WSJ) (Score: 0.4338)
China's Beige Book shows employment situation in Chinese factories
worsened in April from end of March, suggesting economy on less solid
ground than government data (Bloomberg) (Score: 0.4283)
Eurozone economy contracts by 3.8% in Q1, the fastest decline on
record (FT) (Score: 0.4252)
China's official PMIs mixed with beat from services and miss from
manufacturing (Bloomberg) (Score: 0.4052)
======================
Query: Trump is hurting his own reelection chances.


Top 5 most similar sentences in corpus:
Trump contradicts US intel, says Covid-19 started in Wuhan lab. (The
Hill) (Score: 0.7472)
Amazon accused by Trump administration of tolerating counterfeit
sales, but company says hit politically motivated (WSJ) (Score:
0.7408)
```

```
Trump says China wants to him to lose his bid for re-election and
notes he is looking at different options in terms of consequences for
Beijing over the virus (Reuters) (Score: 0.7111)
Inconsistent patchwork of state, local and business decision-making
on reopening raising concerns about a second wave of the coronavirus
(Politico) (Score: 0.6213)
White House risks backlash with coronavirus optimism if cases flare
up again (The Hill) (Score: 0.6181)
```

You can see how uncannily accurate the model is able to pick out the most similar sentences.

The code I used can be found below —

The example above is simple, but illustrates an important point of semantic search. It would take a human couple of minutes to find the most-similar sentences. It gives us the ability to find specific information in a text without human involvement, this means we can search phrases we care about in thousands of documents at computer speed.

This technology is already being leveraged to find similar sentences between two documents. Or a key piece of information in a quarterly earnings report. With this semantic search, for example, we can easily find daily active users for all social companies like Twitter, Facebook, Snapchat, and others. Even though they define and call the metic differently — Daily Active Users (DAU) or Monthly Active Users (MAU) or Monetizable Active Users (mMAU). Semantic search powered by BERT can find that all these surface forms mean the same thing semantically

— a measure of performance and it’s able to pluck the sentence of interest for us from the reports.

It is not a far fetch idea that hedge funds are leveraging semantic search to parse through and surface metrics within Quarterly reports (10-Q/10-K) and have them available as a quantitative trade signal in an instant once they are published.

The experiment above shows how effective semantic search has gotten in the last year.

Finding similar sentences — Clustering

Another major way one can use these vector embeddings of sentences is for clustering. We can quickly cluster sentences in a single document or multiple documents into similar groups.

Using the above code one can take advantage of a simple k-means from sklearn—

```
from sklearn.cluster import KMeans
import numpy as np


num_clusters = 10
clustering_model = KMeans(n_clusters=num_clusters)
clustering_model.fit(corpus_embeddings)
cluster_assignment = clustering_model.labels_
```

```
for i in range(10):
    print()
    print(f'Cluster {i + 1} contains:')
    clust_sent = np.where(cluster_assignment == i)
    for k in clust_sent[0]:
        print(f'- {corpus[k]}')
```

And again, results are spot-on for a machine. Here is a couple of clusters —

```
Cluster 2 contains:
- AstraZeneca to make an experimental coronavirus vaccine developed
by Oxford University (Bloomberg)
- Trump says he is pushing FDA to approve emergency-use authorization
for Gilead's remdesivir (WSJ)

Cluster 3 contains:
- Chesapeake Energy preparing bankruptcy filing and has held
discussions with lenders about a ~$1B loan (Reuters)
- Trump administration may announce plans as soon as today to offer
loans to oil companies, possibly in exchange for a financial stake
(Bloomberg)
- Munchin says Trump administration could allow oil companies to
store another several hundred million barrels (NY Times)

Cluster 4 contains:
- Trump says China wants to him to lose his bid for re-election and
notes he is looking at different options in terms of consequences for
Beijing over the virus (Reuters)
- Amazon accused by Trump administration of tolerating counterfeit
sales, but company says hit politically motivated (WSJ)
- Trump contradicts US intel, says Covid-19 started in Wuhan lab.
(The Hill)
```

Conclusion

Interestingly, ElasticSeach now has a dense vector field and others in the industry operationalizing the ability to quickly compare two vectors such as Facebook's faiss. This technology is cutting-edge but is quite operational and can be rolled out in a matter of weeks. Cutting edge AI is at the fingertips of anyone knowing what to look for.

If you are interested to learn more feel free to reach out, I am always available for an e-coffee. Stay safe out there.

. . .

Thanks to Nils Reimers' informative post on huggingface discussion the led me to write this.

NLP AI Machine Learning Search Open Source

More from Towards Data Science

Follow

A Medium publication sharing concepts, ideas, and codes.

Read more from Towards Data Science

More From Medium

A Full-Length Machine Learning Course in Python for Free

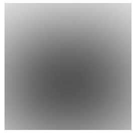
Rashida Nasrin Sucky in Towards Data Science

Noam Chomsky on the Future of Deep Learning

Andrew Kuo in Towards Data Science

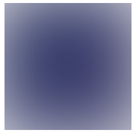
An end-to-end machine learning project with Python Pandas, Keras, Flask, Docker and Heroku

Ryan Lamb in Towards Data Science



Ten Deep Learning Concepts You Should Know for Data Science Interviews

Terence S in Towards Data Science



Classification, regression, and prediction — what's the difference?

Cassie Kozyrkov in Towards Data Science

Kubernetes is deprecating Docker in the upcoming release

Dimitris Poulopoulos in Towards Data Science

How I Switched to Data Science

Rashida Nasrin Sucky in Towards Data Science

Python Alone Won't Get You a Data Science Job

Mohammed Ayar in Towards Data Science