Bloomberg Industry Group 2022 Datathon Challenge

What is an embedding?

An embedding is a dense mathematical representation of some input data. For this challenge, the embedding is the first half of Bloomberg INDG's prototype model to assign news articles to various publishing channels.

The output is a 512 element list of 16 floating point numbers from -1 to 1. The list is a compressed representation of the input text fed to the model. For those with experience in deep learning, this is the output of a tanh pooling layer.

The Data

There are two embedding datasets for this challenge

- `cnn_samples.csv` - A small sample of CNN articles from the `cnn_dailymail` dataset along with generated embeddings. The data is made available under the Apache-2.0 License.
- `federal_samples.csv` - A sample of press releases from various US government agencies along with embeddings.

The Challenge

Part One (50 points possible)

There are five mystery articles with embedding shown in `challenge.csv`. For each article, make your best educated guess as to what the article is about. It can be as broad or as precise as you'd like. (Feel free to offer up a phrase, sentence, or even a paragraph to describe your guess. Dissertations not welcome.)

The closest team for each article gets 10 points, the 2nd closest team gets 5 points, the 3rd closest gets 2 points.

Part Two (25 points + 5 style points possible)

Build a ~~binary~~ classifier using the embeddings!

Can you classify product reviews? Can you detect spam?? We don't know!

Classifiers will be judged on methodology (10), technical implementation (10), and novelty (5). **Usefulness of model not a factor.**

Bonus (0 points)

Guess the 6th mystery embedding. It's in `mystery.json` .

Hint: It's not a news article.

Where do I start?

You can do lots of things with embeddings but mostly, they are inputs for other models. Theoretically, similar input data will generate similar embeddings. Though you should probably check for yourself if this is true.

There are numerous ways to measure similarity. The most common methods are euclidian distance and cosine similarity but don't let yourself be constrained to these. If you have a good hypothesis for a novel similarity measure, here's your chance to test it out!

(There are some examples in `similarity-example.py`).

Once you decide on a distance metric, you can use methods like k-Nearest Neighbors to compare data points to groups of neighbors. You can even use graphs to represent relationships between data points. Here's your chance to experiment and go wild!

I need more data points

You'll probably need more data. You can call our API to generate more embeddings. All you need to do is give it an API key and some text.

Each team will be provided an API key which will be throttled to X number of calls per minute. Larger teams may need to self-throttle so everyone on the team can use the API.

Here's how to get an embedding:

Via Code

See `api-example.py` .

Via Postman

You can call the API from a browser using Postman.

1. In the address bar, use the URL: https://datathon.bindgapi.com/channel
2. Change method from GET to POST.
3. Under **headers** make sure you have the following keys:
   - X-API-KEY = "Your API Key Here"
   - Content-Type = "application/json"
4. Under **body**, select raw, JSON, and input the body in the following JSON format:

```
{
    "input": "text goes here"
}
```

Prizes

- **First place**: Bose Sport True Wireless Bluetooth Earbud
- **Second place**: RK ROYAL KLUDGE RK100 Wireless Mechanical Keyboard

FAQs

How do I win?

Get the most points as a team

That's a lot of numbers, do I need to use all of them?

Nope. Regularization is an important technique in statistics and data science.

What's the mystery prize?

It's a mystery.

Previous
« Baker Hughes Challenge: Object Detection in Advanced Manufacturing

Next
CBRE Challenge: Get in Line »