Machine Learning Intern Test

Quantcast, Jan 2024

Congratulations on being selected to participate in our machine learning challenge. The challenge is intended to take 1-4 hours of your time and should be completed and returned no later than 10 days after you receive the email.

Feel free to use any libraries of your choice, or write your own code.

The challenge

You are given a dataset that contains ~20,000 URLs, and for each URL some extracted page contents and a corresponding high dimensional embedding vector in a parquet file.

Considering the vectors in isolation, use unsupervised machine learning methods to try and make sense of the data. The URL and text are intended as supplementary information to help interpret what your algorithm is doing once you have results. They are not intended as inputs to computation.

What does the numerical data alone tell you? How well did your chosen algorithm(s) perform? When considering the supplementary data as well, what do you conclude? Given more time, what would you do next?

Please return a zip file by email (<10 MB) with

- your code (or a link to the code)
- any graphs or tables.
- your findings on the dataset

There is no correct answer. Have some fun and see how far you can get in a few hours.

You can download the data file "dataset.parquet" in parquet binary format here: https://drive.google.com/file/d/1eY3_G3EtoZ9kgVl6DA_mgwwoWOS2lZ_s/view?usp=sharing

The data can be read in python using the `pandas` library, try `pandas.read_parquet()`.