Machine Learning Engineer

Take-Home Exercise

Background

We receive several million documents a year that need to be classified and distributed to the correct department. To speed up this process, we want to develop a machine learning model to classify documents automatically.

Since we cannot share real data for this task, we use a substitute data set consisting of news headlines.

Goal

1. Build a predictive model to classify headlines into 4 categories (see below)
2. Create a REST API to serve the model
3. Write an example script (Bash or Python) that uses the API from the command line
4. Prepare a presentation of your approach and results.

Data

The news headlines data set can be found at:

https://archive.ics.uci.edu/ml/machine-learning-databases/00359/NewsAggregatorDataset.zip

This dataset consists of the following columns, delimited by \t:

Column	Description
id	Numeric ID
title	News title (to be classified)
url	URL of the original article
publisher	Name of the publisher
category	Category (**b** for business, **t** for science and technology, **e** for entertainment, and **m** for health)
story	Alphanumeric id of the cluster that includes news about the same story
hostname	Hostname
timestamp	Timestamp

Your model does not need to use all attributes. You are free to determine which ones to use and which to discard (using analysis or your own judgment). It's also ok to take a data sample instead of using the entire data set.

Report

Code: You can include a notebook for your exploratory steps, but the final output should also include runnable Python code as a Python file or module.

Presentation: Powerpoint, etc.

Evaluation Criteria

Performance of the final model is not a very important criterion for this exercise; we are more interested in:

- The step-by-step approach and thought process
- The soundness and robustness of the solution and evaluation metrics
- The quality and clarity of the code. Your Python code should be well formatted and (for example) contain docstrings.
- Considerations for possible next steps are also welcome (for example if there are things you would have liked to try if you had had more time)