Feedly Python Coding and ML Challenge

2021 End of Study Internship

You should find a dataset attached in the email containing data from Wikipedia pages about three topics: *Tech*, *Business*, and *Cybersecurity*.

PART 1: Clustering

A. Graph Clustering

One current goal at Feedly is to be able to group articles about the same topic/event. We are not going to tackle this issue with a Machine Learning approach yet, but rather try to group articles about the same topic based on simple rules. In this part (A. Graph Clustering), you are only allowed to use the *Python Standard Library* (*NetworkX* is not part of it).

1. Build a graph whose nodes will be Wikipedia pages of the attached dataset. Two page nodes are considered to be connected if they share at least n tokens.
2. What is the time complexity of building this graph?
3. Find the connected components of the graph.
4. What is the complexity of finding all connected components?
5. Each connected component can be considered as a cluster of pages. How would you choose a title for each cluster?
6. Give one example and comment.

B. End to end Pipeline

You should probably have written different functions and classes until now. At Feedly, we have a clean way of processing data with pipelines: each processing step is a class that handles part of the processing.

1. From your previous work, build a clean and efficient *main* that will take as input a list of Wikipedia pages and will output the clustering results.
2. Run your pipeline with the attached dataset to group articles.
3. Could you evaluate the quality of the results? (The quality does not matter, just the discussion about it).
4. Could you think of an ML algorithm that could achieve comparable or better results?
5. Perhaps you know that your code works just like expected, but we don't! Provide us with a repeatable way to ensure it does! (e.g. unit tests).

PART 2: Classification

1. Build a classification model to infer if an article is about *Tech*, *Business*, or *Cybersecurity*.
2. Which metric would you optimize to maximize customer satisfaction?
3. Evaluate your model with your chosen metric(s).
4. If you were given several models, how would you decide which to deploy into production?

Evaluation

As a real-world problem, there are often many ways to solve and even define the problem. Feel free to make any assumption or initiative that you find relevant for Feedly readers.

What we are evaluating:

There are two equally important aspects that we evaluate:

> First, as we give interns full engineering responsibilities, we value the quality of your code. This challenge aims at evaluating your ability to write clean and expressive code, as well as how you apply your knowledge in algorithms and data structures to real world use cases.
> The second aspect is about ML strategies and creativity: modeling and problem-solving skills. For this aspect, feel free to explain what you would do if you had more time.

What we are not evaluating:

This is not a Kaggle-like challenge where we score model results. This challenge should just be seen as a way to showcase your skills.

Sending your challenge

Please send a zip file with your full name in the filename containing a Jupyter notebook (with both ipynb and pdf versions) and/or Python files.

Send the zip file at stephane@feedly.com by the end of the week you chose. You will receive an acknowledgment email, contact stephane@feedly.com otherwise.