Yewno

Data Science Assignment

Thanks for your interest in Yewno. At Yewno, we don’t believe in arbitrary, onerous “what is the difference between supervised and unsupervised learning?” type interviews. At work every day, you’ll be dealing with a range of challenges, some modeling, some developing, some testing, hopefully all fun. The objective of this assignment is to see how you deal with challenges in a realistic setting, rather than in an artificial one hour interview.

The process goes like this:
1. You: Thoroughly read the exercise below, if you have any questions, email haris@yewno.com.
2. You: Complete the exercise within 3 days of receiving this document.
3. We: Contact you to setup a time to chat about your submission.

We do not expect you to develop the ultimate solution to the problem below. We are however interested in seeing a number of things:
1. How do you approach a problem?
2. How do you manage your work?
3. Are you aware of potential pitfalls your solution might have and could you propose alternative paths?

Introduction

Data and data processing is the foundation of Yewno. With our goal to ingest the world’s knowledge, we are working to consume both public and private data sources in both batch and streaming methods. Both data pipelines are built around sets of algorithms that are ran against the datasets to build the Yewno inference engine.

One of the key roles within Yewno will be developing cutting edge machine learning algorithms to extract useful, decision-making, knowledge from raw data. Scalability of the algorithms is a must, and this role will work closely with the data engineering team to tune the algorithms into performant, production-ready systems.

Task

Develop a system for *plagiarism detection*: you are given a set of textual documents and you are asked to detect which documents contain significant overlapping in content. It is up to you to define what "*significant*" and "*overlapping*" mean. You can also use examples of plagiarized portions of documents.
You are allowed to use whatever programming language/library you feel the most comfortable with. Briefly describe your choices in terms of data preprocessing, feature extraction, algorithms used and text similarity metrics.

Additional questions:
1. How would you assess the performances of your system?

2. How could malicious authors potentially fool your system?
3. Is your system scalable w.r.t. number of documents / users? If not, how would address the scalability (in terms of algorithms, infrastructure, or both)?

When you are finished, send us a link to the code repository -Github or BitBucket are great. Please be sure to **save the outputs of your test** run so we can take a look. Remember, we care for as much about **how you think** about the problem as the code itself! Document the code as needed and be ready to discuss your project.

Above all, have fun and reach out if you have any questions. The task is designed to take approximately 1-2 days to complete.

Datasets

You are free to take whatever dataset you prefer. We suggest using randomly sampled academic papers from arxiv (www.arxiv.org).