Invitation to Participate in Data Science/Engineering Assessment

Dear Candidate,

We are delighted to extend an invitation for you to participate in a pivotal assessment designed to showcase your technical prowess and innovative thinking as a Data Scientist or Data Engineer. This assessment is a key phase in our selection process and offers you a prime opportunity to demonstrate your capabilities in handling real-world challenges.

Assessment Overview

Your task is to develop a pipeline, utilizing either a rule-based or NLP-based approach, that processes text data extracted via OCR from laboratory test results or medical records. The goal is to transform this unstructured, potentially faulty OCR output into a structured Python list of dictionaries, adhering to the specified format:

```
[{"parameter": "iron", "value": 5.3, "unit": "mmol/mL"}, …]
```

Key Considerations

- **Data Unstructuredness and OCR Faults:** The input data will be highly unstructured with potential inaccuracies due to OCR errors.
- **Uniformity in Units and Parameter Names:** Ensure that units and parameter names match the specifications in the attached file X1.
- **Parameter Uniqueness:** Each parameter name must be unique within the dataset, discarding duplicates and retaining only the first instance.
- **Date Sensitivity:** Focus solely on the latest test results, disregarding earlier data to ensure relevance and accuracy.
- **Data Integrity:** Aim for a data integrity level exceeding 90% to ensure reliability.

In our quest to not only attract but also reward exceptional talent, we are pleased to announce that the candidate presenting the most effective and innovative solution will not only be offered a contract with Entropy Technologies but will also receive compensation for the time invested in the assessment task. We recognize the effort and expertise required to excel in this challenge, and we are committed to ensuring that high-quality submissions are duly acknowledged and rewarded. This is a testament to our appreciation of your professional dedication and a reflection of our company's values in fostering a culture of excellence and innovation.

Note

Please note that fine-tuning external models, such as openAI API-based models, is not appropriate for this task. The goal is to enable the development of an internal processing pipeline in which all data processing and handling takes place on internal servers, independent of third-party services.

An open conversation and questions regarding the assessment task during the duration of the task are encouraged and are a chance not only to get on the right track but also to informally present your approach and tactics.

Sample data for the OCR task can be found at:

https://drive.google.com/drive/folders/18lbbitkIqu9f3rLoDJGI_xcU-xKgUWS6?usp=share_link

Submission Details

You have 7 days from the receipt of this email to complete and submit your assessment.

Please email us a link to your repository containing the completed task to info@entropy-technologies.com. Include your first name and the position you're applying for in the subject line of your email.

In your submission, share your progress, highlighting both the challenges you overcame and those that remain. This insight will be invaluable for the subsequent live interview stage. A well-maintained readme file can do the job.

Support and Questions

We understand the complexities of this task and encourage open communication. Should you have any questions or require clarification, feel free to reach out to us at

info@entropy-technologies.com. Your queries will be addressed promptly, ensuring you have all the information needed to excel in this assessment.

Closing Note

This assessment is not just an essential step in the selection process; it's an opportunity for you to engage with real-world data challenges and demonstrate your innovative solutions. We look forward to reviewing your submission and potentially welcoming you to our team.

Thank you for considering this opportunity with Entropy Technologies Digital Pty Ltd. We wish you the best of luck and are excited about the possibility of working together to tackle the fascinating challenges at the intersection of data science and healthcare.