Jr Data Scientist - Evaluation -1

Deadline - Within 7 days from the date of assignment.

Thank you for applying, and welcome to the first round of evaluation.

Index

-  Please carefully read the instructions and document the notebook/code well - as we will be running that in order to evaluate the assignment.
- Expectation from the evaluation: This is stage 1 evaluation - wherein the goal is just to see how well you fare in running and improving simple models. How you fare in working with APIs - with given instructions. And please read every line closely.
-  As with anything, please read this entire document before attempting the assessment.

Why read the next section

Before you proceed with the evaluation, which may take hours of your time, it's very important that you read about what we do and what we are looking for - so that it aligns with your goals.

A Brief Preface

1. **What we are looking for?**

An ideal candidate should be extremely good with python programming. It's also important to deploy the data apps - either directly or via pipelines.

2. **About this role** 

As a Jr Data Scientist, you would play an important role in improving our products. Our goal is to understand how search works. In this process, we collect a lot of data, and find patterns. For example - at the time of this writing, why does Google Chrome does not appear in top 3 when you search for ***browser*** on playstore?

What does a user think before deciding which app they'd like to download? What are the factors that attribute to that decision and so on. Some questions in the evaluation would require to figure this out.

We hire people with broad sets of technical skills. You'd find that you are handling many of the projects by your own. It's tough with a very steep learning curve.

This role requires data gathering, extensive writing  - for example - requirement specification. This also includes providing analysis at scale and ofcourse - come up with forecasts, or other optimization methods to improve our products.

3. **About Evaluations**

There are about two evaluations - this is the first set and if you are selected for the next round, then you'd be required to complete another evaluation. These evaluations would help unearth some interesting questions.

We, at Nextlabs, wish you a good luck.

All of us makes mistakes - and so do we. In case you have any feedback around it - then please share it with us. We'd like to know how we can make the evaluation process better.

Part 1 (Three Questions)

1. Write a regex to extract all the numbers with orange color background from the below text in italics.

{"orders":[{"id":1},{"id":2},{"id":3},{"id":4},{"id":5},{"id":6},{"id":7},{"id":8},{"id":9},{"id":10},{"id":11},{"id":648},{"id":649},{"id":650},{"id":651},{"id":652},{"id":653}],"errors":[{"code":3,"message":"[PHP Warning #2] count(): Parameter must be an array or an object that implements Countable (153)"}]}

2. Here's the list of reviews of Chrome apps - scraped from Playstore. DataSet Link

Problem statement - There are times when a user writes ***Good***, ***Nice App or any other positive text***, in the review and gives 1-star rating. Your goal is to identify the reviews where the semantics of review text does not match rating.

Your goal is to identify such ratings where review text is good, but rating is negative- so that the support team can point this to users.

Deploy it using - Flask/Streamlit etc and share the live link.

Important
In this data app - the user will upload a csv and you would be required to display the reviews where the content doesn't match ratings. This csv will be in the same format as the DataSet Link

Bonus Points - If you deploy the app with Authentication.

3. Ranking Data - Understanding the co-relation between keyword rankings with description or any other attribute. Here's the dataset.

Suggested questions:

1. Is there any co-relation between short description, long description and ranking? Does the placement of keyword (for example - using a keyword in the first 10 words - have any co-relation with the ranking)?
2. Does APP ID (Also known as package name) play any role in ranking?
3. Any other pattern or good questions that you can think of and answer?

Part 2 (Two Questions)

Check if the sentence is Grammatically correct: Please use any pre-trained model or use text from open datasets. Once done, please evaluate the English Grammar in the ***text*** column of the below dataset.
DataSet Link

Optional - if you can indicate the grammatical accuracy of sentences in percentage or on number scale (1-10), that would be an added plus - but is not essential.

More Bonus points (You can write answers to these in ReadMe)

1. Write about any difficult problem that you solved. (According to us difficult - is something which 90% of people would have only 10% probability in getting a similarly good solution).

2. Formally, a vector space V' is a subspace of a vector space V if
    - V' is a vector space
    - every element of V' is also an element of V.

Note that ordered pairs of real numbers (a,b) $a,b \in R$ form a vector space V. **Which of the following is a subspace of V?**

- The set of pairs *(a, a + 1)* for all real *a*
- *The set of pairs (a, b) for all real a ≥ b*
- *The set of pairs (a, 2a) for all real a*
- *The set of pairs (a, b) for all non-negative real a,b*

Important Notes

- Feel free to Google or Stackoverflow (or even go as far as read a book) to understand anything, but please do NOT copy/paste any code/snippet.
- Finish your assignment before the assigned deadline. If you need to extend the deadline, please make sure you drops us an email.
- Document how you deployed the project (in READMe)

Submission Process

Fill this submission form. https://forms.gle/wztJDiVgC9Y1j8wK6. We only consider the submissions through Google Form wef September 14th, 2021.

(Before you fill up the form, please read the below three important pointers)

- Please do record a screencast - running the evaluator through both the problem statement and your solution (While we will not assess the longer videos negatively, we'd prefer the videos to be under 5 minutes - and it should be very concise and to the point). Please do speak about the solution in the screencast.

- We do evaluate ReadMe, documentation. Please ensure that you update that - even the live link should be included in readme - as well as build instructions - just in case we'd like to run it on our local machines. While it's not absolutely essential - we would like you to follow the conventions as much as possible.
- Please read through the submission process and each question carefully. Part 1, Question 2 - asks for deployment, and requires that you share the URL after the deployment is complete.
- If you are shortlisted for the next round of evaluation, we will reach out to you in 7 working days post the date of submission. Unfortunately, we are unable to provide individual feedback.

FAQ:

- Deadline for the test.
    - 7 Days from the date of Assignment.
- Can I use pre-trained model in Chrome reviews problem
    - Yes, you can.
- Can I use a pre-trained model in Grammer problem?
    Yes.
- Can I train the model on partial data from Yelp DataSet?
    - Yes, if for any reason, you are unable to train the model on the full dataset - please feel free to use the partial dataset for the training.
- Can we extend the deadline?
    - Unfortunately, we are unable to - unless there are exceptional circumstances. In case you are unable to do it now, please feel free to re-apply so that we can evaluate on the assessment available then.

Changelog

March 11th 9:34 PM:

- Part -2 question count was mentioned as *two*, corrected it to One.

March 13th 12:24 PM:

- Part -1, question 2 - added a brief note on the expected deployment. Thanks Tej for pointing it out.
- Adding a section for FAQ.

March 17th 8:23 PM:

- Added more FAQs
- March 20th, 11:16 AM. Added a pointer in Submission. Please do record the screencast, and deploy the app (Part 1, Question 2) on a live server.
- Added a minor point to include audio in the screencast - as we've seen very few cases where the screencast didn't have any audio.

April 1 10:03 PM:

- Removed the date of the deadline. We would consider the deadline from 7 days - from the date the assignment was sent to you.

Sep 7th 1:07 PM:

- Specified that we can use pre-trained model for grammar problem.

Sep 14th 12:51 PM:

- Added a Form link. Please fill the same, in order to submit the assignment.

Feb 18 4:21 PM:

- Not mandatory to train the model on the dataset given in the evaluation.

Submitted the assignment, and didn't hear from us? Please add this code to the body of the message - ***QtoJreq07IKYaQa1***

Scenarios in which it can happen:

- Multiple people with the same name.
- All parts in the assignment are not covered.
- No Screencast/deployment URL - in such cases, we presume that it's still work in progress.
- Forgot to add the evaluating team as maintainer (Only in case of Gitlab), or incorrect file permissions in case of Colab.

Please only use this code - if you've already submitted the assignment, but haven't got an acknowledgement.