DATA SCIENCE ENGINEER

Problem Statement -

We, at Perpetua, are planning a team vacation. Planning holidays is exciting but exhausting. Shortlisting hotels by their location, price-points, amenities, and reviews requires meticulous research. Combing through tons of reviews to identify potential jackpot deals or avoiding pitfalls for the crappy hotel-stays is a herculean task.

Thankfully, we have access to hotel reviews procured from different websites. The dataset contains unstructured text in the form of reviews and a bunch of other metadata. However, we do not have time to read, comprehend, and analyze each and every review without inducing bias into the whole process.

To address this problem, we have managed to assign ratings on a scale of **1-5** based on the review text for a sampled data only.

What Needs to be done -

- As the official Data Guru at Perpetua, we seek your help in reaching a data-driven decision for shortlisting the best places to go on a vacation.
- We would like you to -

Analyze the Data -

- The data is too large to investigate manually.
- Analyze and Derive **TWO** Actionable Insights from the Dataset to help us decide where we should go on vacation.

Build an ML Model -

- We can always scrape more reviews from the web. However, to evaluate the hotels, we need an unbiased evaluator that can objectively rate the hotels.
- Develop and Implement a scalable model for predicting hotel rating using `review_text` and `review_title`.
- Feel free to engineer new features and choose a model wisely.

Build an API service

- We would like to enable wider adoption of the work you have done by exposing the model via an API.
- Design and Implement an API to predict the ratings given the required inputs.
- Use a framework of your choice (Flask, Django, etc)

Send us -

- a runnable Github repo or the codebase in a zipped file.

Brownies Points for -

- We cannot wait to go on a Holiday. Earn brownie points by submitting the assignment within the first 8 hours.
- We live by the mantra of "Extracting every single drop of information" from the dataset. Suggest ways to better utilize the dataset and earn bonus points.

Let the Holidays Begin !!