Machine Learning Engineer - Skills Exercise

Thank you for taking the time to do this exercise! It is meant to represent, as accurately as possible, the kind of problem you would be in charge of solving as a Machine Learning engineer at Avalanche. It is a full-time-sized problem, but please spend no more than a few hours on this in the first stage, for your own sake as much as anything else! We will make all the space necessary for discussing what you felt you've left out in the followup conversation.

The Setup

In our surveys, we ask a combination of both open-ended and closed-ended questions. In this exercise, you will be analyzing open text responses from a survey question which asked the following:

- *Why are you cancelling?*

Our team has already analyzed these responses to construct themes that will be useful for our clients to understand in terms of quality (what they are) and quantity (how many of each there are). As is the case with all of our projects, the ultimate goal is to use insights derived from this data to help companies make informed decisions.

The Data

You should have been given a file named *coded_responses.csv*

- Respondent IDs are the unique ids attached to a respondent.
- The text of their response is in the *'response'* column
- *'Theme'* column represents the themes we coded the data for.

Note: There can be responses which have multiple themes, those responses will have multiple rows in this file.

Exercise

As you analyze this data, imagine that you are working on a project in which you will do everything in your power using machine learning to aid analysts in their ability to code the responses, derive and communicate insights. To accomplish this goal, we have divided the exercise into two parts.

Part 1:

Let's assume the themes column is not there and the data we have for now is the question text and all the responses. What machine learning approach can we implement using the response text to make it easier for the team to understand the responses better and come up with the themes? Can we club similar responses together, can we help in finding some of the themes or topics? We would like to see the basic code implementation of your approach in handling this data. It does not need to be a super efficient or best algorithm, we are interested in how you will approach this problem.

Part 2:

Let's assume now the team has come up with a set of themes associated with the responses and we want to display the themes to the users via a dashboard that will aid analysts in their ability to derive and communicate insights.

- *What kind of visualizations could aid the analyst's understanding of the language that distinguishes these themes?*
- *What kind of visualizations will help analysts to communicate their findings with our clients?*

As you consider this problem, please think in particular about how you would incorporate the following. You are not expected to incorporate them, but we hope this will give you a sense of how broadly to consider the problem:

- Simple and more complex approaches
- Human interaction and labeling
- Machine learning
- Advanced NLP techniques

While this data represents one project, such a dashboard would need to serve many different projects with different themes. Please think over how you would approach the task of making such a dashboard usable for all of our analysts regardless of the project they are on.

- *What data or other information will you need access to?*
- *Who will you need to coordinate with?*

- *How will you get feedback?*

We can discuss this along with the work in part 1 during the next interview. Do not worry if you can't conduct all the analyses that you would like to. This is to be expected. We just want to understand how you will approach such a problem, so just be prepared to talk about the work you did.

- You can invoice us a flat fee of $200 honorarium for the time you spend on the work product.
- Please email your invoice to accounting@avalancheinsights.com

The Interview

Please conduct all of your work within a jupyter or colab notebook and share it with us via a link to github or another site for hosting notebooks. The work in this notebook should be well documented to explain your process as well as the thinking behind it. Feel free to use any python packages that you are comfortable with.

This notebook will also serve as a conversation piece for our interview which will loosely follow the following format:

- 15 minutes present the notebook
- 10 minutes of Q&A about the analysis
- 20 minutes of general discussion about data science and Avalanche Insights.

Please don't hesitate to ask any questions, clarifying or otherwise, as you do this exercise and we hope you have fun! 😀