potloc
Consumer research

Case study.
Junior Data Scientist - Generative AI

🎯 Objective:

Help our Project Support team to automate one of their time-consuming manual tasks. They spend a lot of time classifying open-ended answers (cannot be “yes” or “no”, require the respondent to elaborate) among candidate topics. They come to ask you if you could help them to automatically find out the candidate topics.

Business context:

Text classification relies on two manual steps:

1. Determine candidate topics
2. Classify open-ended answers among these candidate topics

The goal of the project is to automate the first step only.
The current process to determine candidate topics is to read the first 25% of the open-ended answers, then manually choose multiple topics that the Project Support member has found relevant and frequent.

We would like to automate this process so that we can use 100% of the open-ended answers for this first step and save a significant amount of time, in addition to standardizing the process.

The dataset of the test is an extract from Yahoo answers and is composed of the question title, the question content and the best answer. For simplicity of the test, you can predict one topic per row.

Steps:

1. Find a relevant model for Topic Modelling that suits the given dataset;
2. Train the model to predict human-readable topic;
3. Build a simple REST API that we can run locally to predict topics. You can send us the trained model (probably zipped) by email to data-science@potloc.com with instructions to make a successful prediction;
4. Briefly identify next steps to improve the model and the inference.

potloc
Consumer research

Time estimation and evaluation:
~2-3 hours, not more than 3 hours.

The evaluation is not based on the performance of the model but rather on:

- Ability to research and find available solutions - *we won't reinvent the wheel!* 🛞
- Project files organization and git management
- Ability to explain your decisions, not the decisions themselves
- Efficiently save the fine-tuned model to be able to serve it through an API - *let's bring concrete value to our Product* 🏆
- Clarity of code and process - *helps for team collaboration* 🎉
- Identify next steps - *feel free to be creative, sky's the limit* 🚀

Delivery format:
When you feel ready, you can create a public repository on Github and share it to data-science@potloc.com. So that, we will be able to clone your repo and to test your application locally!

Feel free to share any relevant files (notebooks included - thanks to the feature preview Rich Jupyter Notebook Diffs) or ask any questions during the test to data-science@potloc.com.

Enjoy! 💪