Challenge Closer

Thank you for making the time to come to interview with the Data Science team at Closer. We hope you find the process both challenging and rewarding.

The challenge is designed to allow you to show your technical capabilities. We would like you to demonstrate that you can frame and solve real-world problems involving hidden patterns and lots of data with a clear understanding of the statistical, machine learning, software development and big data issues arising.

The challenge comprises an NLP project.

You will receive detailed information of how to get the necessary data for each problem in a separate email.

As deliverables, we expect you to submit the code you used to solve the problems. You can use any open source language in which you feel comfortable with, but we strongly encourage the usage of Python. Besides the submission of the code, we expect that you take us through your solution and recommendations (mentioning both achievements and difficulties) in a small discussion (we discourage PowerPoint slides usage, but you may of course bring written notes). You can expect that, during your presentation, we will ask you questions about your approach. Additionally, we may ask you to reconsider your approach given updated information or alternative scenarios. We will also assess the data structures that you used in both exercises.

You should assume that we are familiar with the cases material.

Some advice:

- The problem is not trivial: if your answer is trivial, you may have misunderstood the problem.
- Make sure your answer demonstrates clear and structured thinking. How you approach the problem is just as important as the answers you get.
- If you have any questions, please email them in advance.

Exercise 2

ACME, the very same company of Exercise 1, created a unit of root-cause analysts that verify and study each situation and identify the root-cause of each complaint. Moreover, similarly to the situation experienced in Exercise 1, its analysts are taking too long to analyse and identify the cause of each situation. Therefore, ACME contacted for you to build a support tool that will help its analysts in making a more efficient job.

For this exercise, you are provided another dataset: **complaints_data.csv**. Its fields are self-explanatory and you shall understand their meaning easily.

In the mentioned dataset, you are provided with the corpus of the client's complaint, and the issue (and sub-issue) that the analysts already classified.

The goal that ACME wants to achieve is to have a tool that, given a non-classified complaint, it provides the following:

- Possible root-cause of the complaint
- Similar complaints

Additionally, ACME also wants a second opinion in the issue (and sub-issue) classification for the already classified complaints.

You should consider that there are two root-causes for each complaint: an apparent one and a real one. Therefore, the real root-cause analysis might be hidden in the corpus of each complaint. If you think necessary, you can use additional datasets to enrich the data.

Hint: tackle different dimensionality reduction techniques to achieve better results.