REVEAL

Data Scientist Intern Tech Case

Expected time: 2 ~ 3 hrs

Context

Reveal develops a solution that aims at finding companies present in multiple CRMs, to build a bunch of different features for our customers. Here are the main definitions of the keywords we use:

- **B2B**: Business to Business, e.g. companies whose product or service targets other companies (as opposed to B2C or C2C, where C stands for individual customer)
- **CRM**: Customer Relationship Management, a tool used by companies to manage the lifecycle of their targets & customer
- **Prospect**: a company that is in a CRM but not yet a customer
- **Customer**: a company that has already purchased / subscribed to a product or service
- **Match**: a pair of CRM records representing a company that are representing similar entities. It could be an exact match (representing exactly the same company), or a close match (entities of the same group for instance)

Overview

In the two following attached files you will find 2 datasets of companies and some of their attributes. They represent the data available in the CRMs of 2 of our users.

dataset_A.csv dataset_B.csv

Your goal is to write an algorithm that will use the data from the "companies" table and compute matches based on the company attributes

1. only compute matches between companies from different sources
2. you can use any attribute, but in general name only is not good enough
3. you can output the result in the form of a CSV file with the columns of your choice
    - make sure we can identify in the match exactly what records are involved
    - you can include extra data, such as a score, or reasons that explain the match etc...

Evaluation

You can take the time you need to prepare that exercise. Once ready, you can send us your solution, code and csv with found matches. You can either send us a zip folder or a link to a github repository.

The following criteria will be used to evaluate the code you write:

- Precision: percentage of identified matches that are correct matches
- Recall: percentage of actual matches that are listed in your results
- Performance: how fast is your implementation?
- Code modularity and readability

Do not hesitate to join any documentation to explain your choices, what you would have improved if you had more time, …

Once reviewed, we will set up a short interview to discuss your solution and give you some feedback.