

Matching company entities with company profiles

At XING, we have different types of information about companies. In this assignment, we analyze samples from two different tables that provide information about companies:

- `company_entities.tsv`: lists the legal company entities
- `company_profiles.tsv`: lists the company profiles on XING

Both tables have the following structure:

Column	Description
<code>id</code>	ID of the company entity or company profile. Careful: the entity IDs are different from the profile IDs, e.g. ID 4 in <code>company_entities</code> may refer to <i>XING SE</i> while ID 4 in <code>company_profiles</code> may refer to <i>Siemens</i>
<code>company_name</code>	name of the company
<code>website_url</code>	URL that may link to the company website
<code>foundation_year</code>	year when the company was founded
<code>city</code>	name of the city where the company is located
<code>country</code>	code of the country where the company is located

We assume that there is a 1:n relationship between company profiles and company entities, i.e. one company profile may refer to zero or more company entities.

Task: Your task is to analyze the given dataset and...

1. Design a lightweight algorithm that allows for matching company entities and company profiles. Scala or Java preferred, but it is up to you.
2. Use the `ground_truth.tsv` to validate your algorithm. What kind of numbers would you report?
3. Small, basic REST app to wrap your matching algorithm. It should return the matched companies given company profile data.

Remarks

- We do not expect that you find a sophisticated, optimal solution. A lightweight, simple algorithm is perfectly fine. However, we expect that you will be able to justify why the algorithm makes sense with respect to the given data.
- You are responsible for managing your time (e.g. if you want to invest 1 hour that's fine or if you want to invest 4 hours that's also fine)
- In a next video call, we would like to discuss your solution. Please send us your code before the call, to give us some time to go over it beforehand.