Duplicates

You can find the data in the file Sub-Companies.csv.

We want you to compare the entries and mark the duplicates. Duplicates are entries that have the same company name and address. But other data is of course relevant too.

It’s possible that the names of the companies are not exactly identical but are still the same company. An idea is to use a function to search for phonetic similarities, but you could use something else of course. Sometimes company names are identical and the address is not matching because they are company branches – then they are not duplicates.

My idea would be to do a scoring system to show what percentage of the data is the same, for example. If less than 50% of data match, I wouldn’t mark it. If more than 80% of data match, it’s a duplicate. For everything in between or if I am not completely sure if the two entries are duplicates, I would let the user decide with a simple user interface, if it should be marked or not. But if you have another idea, you are welcome to do it and present it.

In the column CustomerNo you see our customer numbers. The ones that start with Abo are our actual customers and the ones that start with GS are the one we got from yellow pages. If you find a company, where the customer number starts with Abo, with a duplicate in yellow pages, you should mark it somehow special. The reason is that we send out advertisement emails to companies from yellow pages, but we don’t want to do that if the company is already our customer.

We would like you to send us a short precise description of your thoughts, solution and the way how you got to it. Please send it together with the script.