Matching company entities with company profiles

Francesco Cafagna

1 Assumptions

- The first two columns of the csv files storing companies or profiles (respectively, their ID and the name) are compulsory. If not provided an exception is thrown.
- For each profile *at least one* match must be returned

2 Findings

I am not sure that the groundTruth is 100% precise:

- In the groundTruth, for profile "*74657, Sparda-Bank München eG, https://www.sparda-m.de, 1930, München, DE*" we just have a match with "*330389003, Sparda-Bank München eingetragene Genossenschaft, www.sparda-m.de, , Rosenheim, DE*". I think there should be a match with each of the following companies:

companyid	name	websiteurl	year	city	ctry
330389003	Sparda-Bank München eingetragene Genossenschaft	www.sparda-m.de		Rosenheim	DE
325811417	Sparda-Bank München eingetragene Genossenschaft	www.sparda-m.de		Traunstein	DE
325830557	Sparda-Bank München eingetragene Genossenschaft	www.sparda-m.de		Freilassing	DE
325827975	Sparda-Bank München eingetragene Genossenschaft	www.sparda-m.de		Dachau	DE
325958929	Sparda-Bank München eingetragene Genossenschaft	www.sparda-m.de		Weilheim i. OB	DE
325836349	Sparda-Bank München eingetragene Genossenschaft	www.sparda-m.de		Garmisch-Partenkirchen	DE
326477986	Sparda-Bank München eingetragene Genossenschaft	www.sparda-m.de		Freising	DE
325804503	Sparda-Bank München eingetragene Genossenschaft	www.sparda-m.de		Mühldorf a. Inn	DE

- Below you can see the matches provided for "*Atos IT Solutions and Services GmbH Austria, http://at.atos.net ,NULL, Wien, AT*". As you can see in the second line, there is a match with *Siemens*. I think this is wrong and the reason why such a match exists in the ground truth is that the address of Atos is Siemensstraße 92 (check it on their webpage):

companyid	name	websiteurl	year	city	ctry
341599498	Atos IT Solutions and Services GmbH	de.atos.net/de-de/	2010	München	DE
300237765	Siemens Aktiengesellschaft Österreich	www.siemens.at	1879	Wien	AT

3 Implementation

With a few queries, I discovered that matching with equality on the url is a very good approach (65% matches found and only 3% of them are false positives). I tried to remove the domain from the URLs and I got (69% matches found and 5% of them are false positives). I sticked with such a choice.

Main idea: Since the url is such a good indicator, I use this as starting point. If for the query profile P, I have one or more companies $E_1, \ldots, E_m$ sharing the same website, I return as matches for P the result of $matchCompaniesOnNameSimilarity(E_i)$, i.e., E_i plus all other entities with a similar name. If no company exists with the same website, I return $matchCompaniesOnNameSimilarity(P)$.

The function $matchCompaniesOnNameSimilarity(P)$ returns all companies with a name similar to the one of P. The similarity is computing using weighted jaccard similarity over grams of size

$GRAM\_SIZE$. For selecting the matches after I have computed the distances, I do not use an absolute threshold since I assumed that for each profile we must return at least one entity: if the absolute threshold is too high, we might return no match for a given profile; if it is too low, we might return many false positives. Therefore, I use a relative threshold meaning that, if s is the highest similarity found for the profile P, I return all those company entities $E_1, ..., E_m$ having $similarity >= s * threshold$.

I exclude grams between spanning over multiple words (i.e., grams including the end of a word and the beginning of the next one): in such a way, the order of appearance of the words in a given company name does not matter.

To avoid that infrequent grams have too much power in finding the matches, I have tried to Log_{10} the gram frequencies. After running a couple of tests with this, I obtained as result the same number of correct matches but a lower fraction of wrong matches. So I decided to keep the log in my implementation.

For efficiency, I use the less-frequent grams to index candidate companies. This means that the companies selected to be compared to the query profile, are only those sharing one or more non frequent grams (on average around 500). Without this step, every company sharing at least one gram is a candidate and, in our scenario, up to 30k companies were compared in the worst case to a given profile, making the runtime extremely slow (1 hour).

4 Experimental Evaluation

I decide to set a GRAMSIZE=4 since it offers a good compromise between runtime performance and accuracy of the result.

	GRAMSIZE=3	GRAMSIZE=4	GRAMSIZE=5
one correct	0.85	0.87	0.87
one wrong	0.17	0.16	0.16
best	0.83	0.85	0.85
no match	0.02	0.009	0.014
time	130 sec	210 sec	350 sec

Table 1: *relative threshold = 0.9 and maxOccurrences = 300*

As you can see in the following table, the threshold does not affect the number of companies for with no match is returned. This is so for two reasons: first, because there are one or two profiles with a name of just 3 letters which does not get mapped to any gram of size 4; second, because there are profiles including only frequent grams, and their corresponding company entities will never be fetched and compared against the query profile, independent on the threshold. I chose a threshold of 0.9 since it maximises the *one wrong* and the best metrics and keeps low the *one wrong* value.

	threshold=0.5	threshold=0.7	threshold=0.9	threshold=1
one correct	0.89	0.88	0.87	0.86
one wrong	0.38	0.21	0.16	0.16
best	0.70	0.82	0.85	0.85
no match	0.009	0.009	0.009	0.009

Table 2: *GRAMSIZE = 4 and maxOccurrences = 300*

Here I show how the algorithm performs when we increase the number of grams to consider to filter out useless candidate. You can see that the first three metrics stay stable, independent if we consider only the grams appearing at most in 100 company names or those appearing up to in 700 company names. I decided to set maxOccurrences to 300 since it keeps at the same time both the *no match* and the *time* low.

Francesco

	maxOccurrences=100	maxOccurrences=300	maxOccurrences=500	maxOccurrences=700
one correct	0.86	0.87	0.87	0.87
one wrong	0.15	0.16	0.17	0.16
best	0.84	0.85	0.85	0.85
no match	0.03	0.009	0.007	0.007
time	171 sec	210 sec	312 sec	362 sec

Table 3: *relative threshold = 0.9 and GRAMSIZE = 4*