Matching Procedure

Final output:

The goal is to match firm names to the public firm registry in France in order to retrieve the unique firm ID called S*iren*.

Datasets:

1. *rivals_data_fuzzy.csv*:
   contains the firm names that need to be matched to the Siren database. It contains a 2-digit-industry code *apen2*, the firm name *RIVAL_NAME*, the year of the M&A from which the firms are affected, and a firm id *id_rival*. I already took out the firms for which I got an exact match on firm name x industry.
2. *siren_database.csv:*
   contains all registered firms in France. It contains the firm id *Siren* (which is the variable I need from the matching), the firm names *RIVAL_NAME*, the 2-digit industry code *apen2* and a random id_number *id_siren*.

Useful links:

- Public Siren database to check entries manually:
  https://www.sirene.fr/sirene/public/recherche
  Enter the firm name in the field "Raison Sociale" or the firm id number in the field "Unité légale".
- Documentation for the Siren API (in French) in case you want to use the API instead of the datasets provided. You have to create a user account before using the API.
  https://portail-api.insee.fr/catalog/api/2ba0e549-5587-3ef1-9082-99cd865de66f/doc?page=85c5657d-b1a1-4466-8565-7db1a194667b

Procedure

If you would like to use Open AI models for matching, I can provide you with an API key to my account.

Code in Python or R would be really great so that I could make changes or add something later if necessary.

Please try to match the firm names in the rivals_data_fuzzy.csv with the firm names in the *siren_database.csv* file in order to extract the unique firm id *Siren*. Ideally, we want to match on **firm name x industry** because the same firm name can have multiple legal units that are active in different industries. However, we want to identify the actual legal unit that is active in the same industry as the merging parties. Even within the same industry, the same firm name can have multiple legal units.

Obviously, standard fuzzy matching algorithms based on some distance, like for example, the one implemented in Stata do not work because they just match on the distance of the letters within each industry, which gives then wrong matches like DHL = DHI instead of DHL France, NISSAN = ANISS instead of NISSAN FRANCE or NISSAN WEST EUROPE

1. I guess it's the best if you first try to match firms within a given industry, so that we get a firm name x industry match.
2. Afterwards, try to match the unmatched firms just on the firm name. Since the industry code is not a precise code but is based on the main sales share of each firm, a firm can still be active in several industries and be a competitor to the merging parties even if it's not coded as the same industry. Here you will get multiple legal units for one single firm name.