Problem assessment

1) Context : Sector codes

As we mentioned, in Belgium (and most countries), companies have to declare a “sector code” when registering. Many companies (around 60%) have declared the “wrong” code. Wrong in this case means either that they have declared a false sector (e.g. “Bakery” while in reality they are a recycling company), or more often that they have declared a code that is correct but too wide for their activity (e.g. “Financial institution” if they are specifically a bank or an insurance company).

This sector code is used for various purposes: tariffs, taxation, insurance, etc. In the EU, these sector codes are called NACE codes. NACE codes are 4 digits codes indicating the sector of a company.

Each country can create more precise codes based on the NACE code. In Belgium, for example, NACEBel codes are 7 digits, with the first four being the EU NACE code, and the last three being country-specific addition. The full list is not necessary, but is available on statbel’s website.

At Inoopa, we have created a model predicting the NACE code of Belgian companies. We can then correct the wrong code of companies and get a precise 7-digit NACEBel for this company. This means that we can tell the exact activity of a company. This has an implication on different aspects, for example:

1) We can ensure that a marketing campaign will reach only relevant companies that will be interested in a product. There is no need to market the new Baking Machine 2000 to a recycling company.

2) We can compute the emissions of a company. For instance, a petroleum company will emit more CO2 on average than an accounting company.

2) Problem Statement

You have financial tabular data of a company as well as their address and a possible URL of their website, and the list of sector codes with their description (that you have found on statbel’s website).

You would like to find out which company has declared a wrong sector code. It is of course possible to look at the website of the company to check if the sector code is correct, but obviously that can be done only on a small sample as there are too many companies in Belgium.

You should find enclosed a small sample of an input example and the list of NACE codes with descriptions to give you an idea.

Describe how you would find the companies with incorrect NACE codes and correct them. List the data would you use - it doesn't have to be specific, even a description of the kind of data you would want available is good.

In particular, we are curious to see for each step:

- Why would you do it this particular way? What alternatives could be tried? Why choose one over the other?
- What could be some problems encountered? How would you solve them? Of course it doesn't need to be exhaustive!
- On what assumptions are you basing your model on and what are the caveats in using them?
- How you would test the performance of your models and what mistakes to avoid.

Lastly, how would you communicate the final results to a colleague with a business background?

You are welcome to write some pseudocode or to lay out which libraries you would use in writing your models, however this is not necessary.

BONUS:
It's possible that the given url for the company is wrongly matched. How would you check if the given url is the correct one for the company's website? How does knowing this affect the models you've described above?