Goals:

Finding the most appropriate taxonomies

Validating the taxonomies

Attempt 1:

I’ve started this challenge by sorting based on sector for an easier separation of domain. After that I filtered the companies based on niche and business tags to create an appropriate taxonomy and then tried to validate them using the description.

Problems encountered:

The filtering was not efficient

The processing time was high

A lot of companies didn’t fit in their most appropriate taxonomy

Attempt 2:

I’ve noticed that the only field without NULL values was niche. I’ve counted every word from taxonomy label and sorted them from highest to lowest value. I then started comparing them with the niche through word equality and if the niche had more than half of the words found I added it as a taxonomy. Using that I managed to find a more appropriate filtering than before. After that I started to make another filtering through the business tags to increase the accuracy of the filtering. After that I tried to validate the taxonomy using the company description through word finding.

Problems encountered:

Even longer processing time

A better filter but still not effective

Most companies were filtered in an appropriate taxonomy but the end result was not satisfactory

Attempt 3:

I started a more thorough research and found out about a machine learning method that uses word impact and creates a value corresponding to the impact. I’ve then installed the sklearn library and combined all data into one array. I’ve then used a tf-idf vector to transform them into values, added them into a matrix and separated them into 2 variables: company\_vectors and taxonomy\_vecotrs. After that I computed the similarity between the two variables into another variable called similarities. I’ve defined another 3 variables: top\_n, rate and labels for the most similar values, the rate of acceptance and the final result labels. I then iterated through all similarities and ordered them decreasingly to get the maximum similarity for each company. Then I used the rate to create a percentage based on the most similar taxonomy, then asses the most similar taxonomies for that company. This solution provides a better than average filtering.

Improvements:

In order to improve it a secondary check would be word embedding technique which is more accurate than tf-idf. A very good approach would be doing a first analysis using both of them on the root documents then compare the results of both and create an intersection of the results. For another validation, applying first tf-idf on the root text and then embedding on the tf-idf result then compare it to the intersection and keep the most relevant results using Jaccard Distance.