Company Classifier

Roscan Rares-Alexandru

March 17, 2025

Abstract

This project focuses on unsupervised multi-label classification of companies based on their textual descriptions. The goal is to assign relevant business categories to each company using a combination of semantic embeddings, vector search, clustering, and zero-shot learning.

1 Introduction

1.1 The task

As part of a technical assessment for an internship application, I was given the task of developing a multi-label clustering approach for companies based on their textual descriptions. The objective was to categorize companies into a predefined set of 220 business categories, leveraging unsupervised learning techniques and embedding-based methods.

1.2 Understanding the assignment

First of all we need to see the data provided by Veridion. The company list and the insurance taxonomy list represent the given data. The dataset consisted of a list of nearly 10,000 companies, each represented by multiple textual attributes. Each company entry included a description, a set of business tags, and categorical information such as sector, category, and niche. These attributes provided a textual representation of the company's activities and industry classification. In addition, the insurance taxonomy contained 220 predefined labels, representing different business categories. The goal was to associate each company with the most relevant labels from this taxonomy based on its textual description and business-related metadata.

Our objective was to identify the most relevant labels for each company based on its textual attributes. It is important to note that the dataset did not include any predefined annotations linking companies to specific labels. As a result, we faced an unsupervised learning problem, requiring a methodology capable of extracting meaningful relationships between companies and business categories without relying on labeled training data.

1.3 Taking a closer look at the dataset

"Our initial step was to understand how companies could be grouped under specific labels. To achieve this, I first conducted a thorough review of all the predefined labels in the insurance taxonomy. By carefully analyzing the label set, I aimed to gain insights into their meaning, scope, and potential overlap, which would later guide our approach to assigning companies to the most relevant categories.

At the same time, I carefully examined the list of companies to identify key patterns in the dataset. From the beginning, I intuitively recognized the following aspects:

- The description was written in natural language and played the most significant role in classifying companies, as it provided a comprehensive textual representation of their activities.
- Business tags acted as keywords that could help categorize companies based on specific terms relevant to their industry.
- The sector was a high-level categorization term with a limited set of values (e.g., Services, Manufacturing), providing a broad classification of the company's industry.
- The category was another defining attribute that could directly map to multiple specific labels from the taxonomy.
- The niche represented the specialized field in which the company operated, offering additional granularity in classification.

2 Thought Process

2.1 Thinking outside the box

> "At Veridion, we run similar algorithms on billions of records. While your solution doesn't need to scale to that level, it would be impressive if it does. For now, however, what matters most is your approach to solving the problem—if your solution is exceptional for the given dataset, we trust that you can scale it effectively using the right tools."

Scalability is not optional—any well-designed solution naturally extends to larger datasets with the right tools and infrastructure.

I see this task as a test of creativity. If one were to focus solely on building a 'robust' solution tailored to the given dataset, achieving the level of scalability required to handle, as stated, 'billions of records' would likely be unrealistic.

Thinking outside the box means putting yourself in the company's position and understanding its actual needs. This becomes evident from the label set, which, despite being relatively small, is highly specialized in certain areas.

For example, we have two distinct labels for swimming pools: 'swimming pool installation services' and 'swimming pool maintenance services'. However, when looking at the automotive industry, there are no equivalent labels for 'car manufacturing', 'car repair', or 'car sales'. This inconsistency suggests that the taxonomy is tailored to specific industries while lacking granularity in others.

When scalability was mentioned, it did not only refer to handling a significantly larger list of companies matching the given labels, but also to the expansion of the label set itself.

We can now imagine that if there are already two distinct labels related to swimming pools, the total number of labels could grow substantially, becoming increasingly specialized. This suggests that scalability is not just about processing more data, but also about maintaining adaptability as the taxonomy evolves.

As both the number of companies and labels increase, the classification task grows combinatorially more complex. Without an efficient optimization strategy, this could lead to an exponential increase in computational cost, making the system impractical at scale

2.2 Scalability, Complexity, and the Automated Labeling Pipeline

Building a scalable supervised learning pipeline presents significant challenges, particularly when manual labeling is not an option. In this context, complexity arises not only from the need to process a large number of companies but also from the expanding taxonomy of labels.

A naïve approach would involve manually annotating data or generating labels dynamically through prompt-based models like ChatGPT. However, these methods do not scale. Instead, our solution relies on a fully automated pipeline designed to maximize efficiency while minimizing computational overhead.

Given the large volume of data, it is crucial to minimize the use of computationally expensive models across extensive portions of the dataset. Instead of applying resource-intensive models indiscriminately, our goal is to leverage computational power strategically, ensuring efficiency without compromising performance.

Personally, I preferred to approach the problem in reverse—starting with the labels rather than the companies. If we were to compare each company against every label individually, this could be done using a simple decision-tree-based approach, such as Random Forest. However, this would likely result in poor accuracy due to the lack of deep semantic understanding. Such a naive approach was immediately ruled out, as delivering an unreliable classification system was not an option.

I was unable to determine the exact origin of the labels or obtain a comprehensive view of the full taxonomy. However, I noticed a pattern: many labels contained recurring keywords, such as services and manufacturing. This observation had significant implications. It meant that I could not rely solely on the full semantic meaning of each label during classification. A key question emerged: if a company was involved in manufacturing, should it automatically be assigned every label containing the word manufacturing? Clearly, such an approach would lead to excessive and inaccurate label assignments, making it necessary to develop a more refined strategy for label selection.

A potential counterargument to this issue is that a more powerful Transformer model, through its embeddings, could theoretically capture the semantic distinctions between labels. By leveraging contextualized word embeddings, such a model should be able to differentiate between a company that manufactures a product and one that provides services related to manufacturing.

While this is true to some extent, it does not fully resolve the problem. Embeddings capture semantic similarity based on context, but they do not inherently enforce a strict logical separation between overlapping terms. If the training data lacks sufficient examples that explicitly distinguish between manufacturing as a production activity versus manufacturing services, even a highly capable Transformer may struggle to disambiguate them perfectly.

Therefore, relying solely on embeddings without additional constraints or post-processing techniques could still lead to misclassifications.

2.3 The NACE codes

As previously mentioned, I understood that the key challenge was handling the large number of labels. But how could this be achieved? Given that this was a multi-label classification task, traditional clustering was not a straightforward option.

The main issue with clustering was defining what constitutes a coherent group. For instance, how could we cluster labels when we have both 'swimming pool cleaning services' and 'legal services' in the same taxonomy? These labels belong to completely different domains, making conventional clustering approaches ineffective.

However, by reframing the problem, an important observation emerged: the ambiguity was caused by the words that accompany the core industry term. Instead of viewing each label holistically, we needed a strategy to disentangle general industry categories from the descriptive modifiers that refine their meaning.

This observation provided a solid starting point. Instead of attempting to cluster the labels based on their full descriptions, we could first group them by industry domain.

By structuring the labels this way, we significantly reduce the search space—there is no need to check whether a company specializing in 'swimming pool cleaning' matches a label related to 'steel manufacturing'. This approach allows us to filter out irrelevant categories early in the process, improving both efficiency and accuracy in label assignment.

But wait, there's more! Why stop here when we can take this idea even further? Perhaps my initial examples didn't fully illustrate the point, but we actually have an entire field that defines the company's sector of activity.

What does this mean? Well, if we've already identified the prestigious company specializing in swimming pool services (yes, I know I keep bringing up swimming pools, but I find it amusing that such a niche industry has specific labels while much broader domains don't), we can leverage its sector field. Since the sector is labeled as 'services', we no longer need to check whether the company builds pools, sells pool equipment, or engages in other unrelated activities.

In other words, this simple step helps us narrow down label assignments efficiently—effectively solving multiple challenges at once. Have we just managed to kill all the birds with one stone?

To illustrate this point, I will refer to the existing CAEN codes in Romania, as we do not have full visibility over the taxonomy used in the label set. CAEN codes are structured precisely by industry domain, with categories such as 'Agriculture, Forestry, and Fishing', 'Construction', and 'Education'.

(Speaking of education, this was actually the first category that raised a key question for me: why were there so many educational institutions—universities, colleges, and schools—yet none had a corresponding label? Oa course, later on, the swimming pool example completely baffled me, not going to lie.)

Returning to our classification system, we observe that CAEN codes follow a hierarchical structure. A broad category such as 'Agriculture, Forestry, and Fishing' is further divided into 'Agriculture, Hunting, and Related Services', 'Forestry and Logging', and 'Fishing and Aquaculture'.

Each of these is then subdivided into more specific categories, which in turn are broken down into even more granular subcategories. This multi-level structure ensures a logical and organized classification, making it easier to navigate industries with high levels of specialization.

Essentially, we can represent this structure as a tree, where the root node branches into 27 child nodes—corresponding to the main CAEN categories. Each of these nodes further expands into subcategories, which in turn branch out into more specific classifications, forming a hierarchical structure.

This approach allows us to systematically organize industries, making it easier to group companies based on their sector while reducing unnecessary label comparisons.

A natural follow-up question would be: why do we need a tree structure in the first place?

The answer ties back to an idea I previously dismissed as infeasible in its original form—clustering. Instead of attempting to cluster all labels at once, we can use the CAEN categories as initial clusters. In this setup, a company is first compared only against the top-level clusters based on its description, which carries the most semantic weight.

Once an initial match is found, the company is further assigned to one or more specific CAEN categories using additional metadata fields. This hierarchical approach allows for a progressive refinement of classification, significantly reducing computational complexity.

Moreover, since this is a multi-label problem, a company might fall into two seemingly unrelated CAEN codes. This is not an issue—if two top-level clusters both show high similarity to the company's description, we can compare it within both and assign the most relevant labels accordingly.

With this approach, I believe I have effectively addressed the core complexity challenge and removed the main bottleneck hindering the expected scalability. By structuring the classification process hierarchically, we significantly reduce unnecessary comparisons, making the system both efficient and scalable

2.4 Why trees?

As previously stated, this is the most computationally efficient approach, as it minimizes the number of unnecessary comparisons as much as possible. With this optimization in place, the only remaining challenge is selecting an algorithm—one that does not necessarily need to be highly complex—to determine the direction in which the company should be further evaluated at each step.

The most basic approach would have been straightforward—concatenating all available company fields into a single text input, generating embeddings, and relying on a powerful Transformer model to handle the classification. This brute-force method would push the GPU to its limits, disregarding efficiency, computational cost, or processing time.

If this were a low-stakes hackathon, we might have justified using the most powerful BERT-based model available, optimizing solely for accuracy without concern for scalability or resource constraints. However, in a real-world scenario where efficiency and scalability matter, such an approach is far from viable.

2.5 Alright, you talked a lot but didn't say much—so what's the point?

Using the right algorithm at the right time. In an ideal computational scenario, the first clustering step would use a moderately powerful algorithm, followed by a slightly stronger one in the next step, and so on. As we move deeper into the hierarchy, the differences between child nodes become increasingly subtle.

By the time we reach the leaf level (the actual label), we can afford to use a more capable BERT model, ensuring precise differentiation where it truly matters.

Of course, this is one possible approach. However, we must not forget that we are dealing with an unsupervised model, which essentially makes educated guesses rather than relying on direct supervision.

Additionally, as previously mentioned, sector and category serve as crucial indicators in our step-by-step refinement process. In this case, we need a balanced approach—one that takes into account the description to some extent while also leveraging the tags, sector, category, and niche of the company to guide the classification more effectively.

Unfortunately, this is where things dive deeper into the complexity of writing the actual code. At this point, I can't definitively say what should or shouldn't be done. Simply fine-tuning the weight distribution could take a significant amount of time, and ultimately, the approach depends on the company's specific needs.

Spending dozens or even hundreds of hours trying to find the perfect formula isn't practical. So, I'll admit—this is where the topic starts to go beyond my scope. The lack of critical information (such as the exact number of labels, number of companies, and how frequently the algorithm needs to be applied) leaves me somewhat in the dark.

2.6 FAISS: The Nearest Neighbor to Salvation

Okay, maybe I got a bit carried away with the title. The truth is, as a student, I don't fully understand the actual needs of a company. I'm well aware that everything I've said so far could have been completely pointless, my reasoning flawed, and my hopes shattered.

That being said, I still believe that what I presented represents part of the ground truth of the problem. However, it's a broad and ambitious idea—one that I wouldn't have been able to fully implement within the given timeframe. And that's not even mentioning the lack of computational power (which I'll discuss later) or the lack of complete data—after all, I don't even know what the full label set looks like and have only referenced Romanian CAEN codes as a substitute.

Now, back to FAISS—why FAISS? Because it's almost comically easy to scale with it. The way it handles massive datasets with lightning-fast retrieval makes it feel like cheating. When dealing with millions of vectors, FAISS doesn't just work—it makes scalability feel effortless.

Coming back to FAISS, it is often praised in specialized articles for its efficiency and scalability in similarity search across large datasets. For example, a detailed article about FAISS's advantages can be found here and it's not just because the article was written by Meta.

3 The Pipeline

- Preprocessing & Word-Level Clustering – Text fields (descriptions, business tags, etc.) are preprocessed, and a word-level clustering algorithm is applied to detect patterns and group similar terms before generating embeddings.
- Generating Semantic Embeddings – The processed text is transformed into dense vector representations using sentence-transformers/all-mpnet-base-v2, allowing for meaningful comparisons between companies and labels.
- Label Retrieval with FAISS – Label embeddings are indexed using FAISS, enabling fast and scalable nearest-neighbor search, which significantly reduces the number of comparisons needed.
- Zero-Shot Classification for Label Validation – A zero-shot learning model (facebook/bart-large-mnli) is used at the final stage to refine and validate label assignments based on semantic entailment.

3.1 Preprocessing

Depending on how we approach the problem, we have multiple preprocessing options. Given that I used a Transformer model, I chose not to preprocess the text extensively. Since Transformers understand the meaning of a sentence in context, removing words would actually disrupt their ability to extract meaningful representations. Therefore, my preprocessing was limited to removing non-Latin characters and special symbols, ensuring that the text remained clean while preserving its original semantic structure.

Kyoto Vegetable Specialists Uekamo, also known	['Wholesale', 'Dual-task Movement Products', 'Cast Iron Products Manufacturer',

Kyoto Vegetable is a company specializing in selling vegetables, so I highly doubt they have any plans to start casting engine blocks for cars anytime soon. So, what's the next step? Do we keep unnecessary noise in the dataset, or do we clean it wherever possible? The goal is to leave the description intact while validating each business tag separately—since each tag is an element of the business_tags list.

This can be done either by clustering the descriptions or by generating embeddings that include all relevant data except for the specific field being verified. The same approach applies to sector, category, and niche, as these fields can sometimes be inaccurate.

The truth is, I didn't have the chance to fine-tune the model or perfect every single component. The only real issue I encountered during noise cleaning was the time it took—it accounted for nearly a third of the total time required for labeling companies using FAISS + Zero-shot.

Given this, I'm not sure how feasible this step would be at a large scale.

description	business_tags	sector	category	niche	candidate_labels	candidate_sim_s	zero_shot_score	final_labels
Welchcivils is a civil engineering and construction company that specializes in designing and building utility network	['Construction S	Services	Civil Engineering	Other Heavy and	['Pipeline Constr	[0.5577945, 0.49	{'Pipeline Constr	['Residential Electrical Services', 'Commercial Construction Services', 'Elect
Kyoto Vegetable Specialists Uekamo, also known as Iwa-machi, is a company based in Kyoto, Japan that specialize	['Wholesale', 'Du	Manufacturing	Fruit & Vegetabl	Frozen Fruit, Jui	['Food Processin	[0.44543204, 0.4	{'Food Processin	[]
Loidholdhof Integrative Hofgemeinschaft is a company that offers a range of services and products to its customers.	['Living Forms', 'I	Manufacturing	Farms & Agricult	All Other Miscell	['Bakery Product	[0.50124466, 0.4	{'Bakery Product	[]
PATAGONIA Chapa Y Pintura is an auto body shop located in Comodoro Rivadavia, Chubut Province, Argentina. Th	['Automotive Boc	Services	Auto Body Shop	Automotive Body	['Coffee Process	[0.3259633, 0.30	{'Coffee Process	[]
Stanica WODNA PTTK Swornegacie is a cultural establishment located in Swornychgaciach, Poland. It is a popular	['Cultural Activiti	Services	Boat Tours & Cr	Scenic and Sigh	['Fishing and Hu	[0.26750863, 0.2	{'Fishing and Hu	[]
BIQ Benefícios is a Brazilian company that specializes in generating benefits for both individuals and businesses. W	['Healthy Food C	Services	Payment Proces	Financial Transa	['Food Processin	[0.4409913, 0.40	{'Food Processin	['Financial Services']
EnvCompli is a company that provides personalized Industrial General Permit (IGP) Management services for Storm	['Source Control'	Services	Environmental E	Environmental C	['Environmental	[0.5045509, 0.49	{'Environmental	['Environmental Consulting', 'Testing and Inspection Services']
Corrugated Beam Form Ltd is a company based in Walsall, England, specializing in the supply of concrete products.	['Cut Sheets', 'Fc	Manufacturing	Building Material	Engineered Woc	['Pallet Manufact	[0.49314192, 0.4	{'Pallet Manufact	[]
The company specializes in the production and distribution of packaging materials, including cardboard sealing tape	['C Packaging S	Wholesale	Paper & Cardbo	Industrial and Pe	['Paper Producti	[0.59287596, 0.4	{'Paper Producti	['Printing Services']
Indira Caterz And Event Managerzz is a company categorized under restaurant supply stores. It is located in Alappu	['Event Organiza	Wholesale	Restaurant Equi	Other Commerci	['Catering Servic	[0.5208752, 0.45	{'Catering Servic	[]
Halle Heat Exchange Equipment Co., Ltd is a Chinese company that specializes in the production and sales of heat	['Technology Pro	Manufacturing	Industrial Air Sol	Power Boiler and	['Boiler Installatic	[0.4711706, 0.43	{'Boiler Installatic	['Industrial Machinery Installation']
Ballymacarbry Elders Housing Association Ltd. is a registered charity in Ireland that aims to provide quality, age-frier	['Sustainable Ho	Services	Housing Prograr	Other Communit	['Community Enç	[0.39544904, 0.3	{'Community Enç	['Single Family Residential Construction']
An Nhien Scientific Equipment Co., Ltd. (AnNhien) is a company that specializes in the distribution of scientific equip	['Sterilization Prc	Manufacturing	Laboratory Equi	Analytical Labor	['Laboratory Ser	[0.5232743, 0.43	{'Laboratory Ser	[]
Anovei is a company that specializes in air treatment, purifier leasing, and nasal filtration research and development,	['Innovative Solu	Manufacturing	Medical Supply I	Irradiation Appar	['Air Duct Cleani	[0.389279, 0.336	{'Air Duct Cleani	[]

description	business_tags	sector	category	niche	candidate_labels	candidate_sim_s	zero_shot_score	final_labels
Welchcivils is a c	['Construction S	Services	Civil Engineering	Other Heavy and	['Pipeline Constr	[0.54348576, 0.5	{'Pipeline Constr	['Pipeline Construction Services', 'Residential Electrical Services', 'Electric Line Constr
Kyoto Vegetable	['Wholesale', 'Du	Manufacturing	Fruit & Vegetabl	Frozen Fruit, Jui	['Agricultural Equ	[0.37806574, 0.3	{'Agricultural Equ	[]
Loidholdhof Inte	['Living Forms', 'I	Manufacturing	Farms & Agricult	All Other Miscell	['Bakery Product	[0.43323463, 0.4	{'Bakery Product	['Bakery Production Services', 'Gardening Services', 'Coffee Processing Services']
PATAGONIA Ch	['Automotive Boc	Services	Auto Body Shop	Automotive Body	['Coffee Process	[0.2948764, 0.25	{'Coffee Process	[]
Stanica WODNA	['Cultural Activiti	Services	Boat Tours & Cr	Scenic and Sigh	['Welding Servic	[0.23207857, 0.2	{'Welding Servic	[]
BIQ Benefícios i	['Healthy Food C	Services	Payment Proces	Financial Transa	['Business Devel	[0.3980954, 0.39	{'Business Deve	['Financial Services', 'Travel Services', 'E-Commerce Services']
EnvCompli is a c	['Source Control'	Services	Environmental E	Environmental C	['Water Treatmei	[0.45728815, 0.4	{'Water Treatmei	['Testing and Inspection Services']
Corrugated Bear	['Cut Sheets', 'Fc	Manufacturing	Building Material	Engineered Woc	['Sheet Metal Se	[0.50389946, 0.4	{'Sheet Metal Se	[]
The company sp	['C Packaging S	Wholesale	Paper & Cardbo	Industrial and Pe	['Paper Producti	[0.5555873, 0.49	{'Paper Producti	['Paper Production Services', 'Sheet Metal Services', 'Stationery Manufacturing', 'Texti
Indira Caterz An	['Event Organiza	Wholesale	Restaurant Equi	Other Commerci	['Catering Servic	[0.49924767, 0.4	{'Catering Servic	['Catering Services', 'Event Planning Services']
Halle Heat Exch	['Technology Pro	Manufacturing	Industrial Air Sol	Power Boiler and	['Roofing Service	[0.4409356, 0.41	{'Roofing Service	[]
Ballymacarbry E	['Sustainable Ho	Services	Housing Prograr	Other Communit	['Residential Plu	[0.28148645, 0.2	{'Residential Plu	['Single Family Residential Construction', 'Volunteer Services']
An Nhien Scienti	['Sterilization Prc	Manufacturing	Laboratory Equi	Analytical Labor	['Laboratory Ser	[0.46862572, 0.4	{'Laboratory Ser	[]
Anovei is a com	['Innovative Solu	Manufacturing	Medical Supply I	Irradiation Appar	['Air Duct Cleani	[0.42436615, 0.3	{'Air Duct Cleani	[]
Db Motor Servic	['Diagnostic Serv	Services	Auto Services	Specialized Autc	['Commercial Ele	[0.38562775, 0.3	{'Commercial Ele	[]
DOO Bački Petr	['Production', 'WI	Manufacturing	Cleaning Equipn	Soap and Other	['Pet Food Manu	[0.35015875, 0.3	{'Pet Food Manu	['Cosmetic Manufacturing']
Reparación Lava	['Home Care', 'Te	Services	Appliance Repai	Appliance Repai	['Confectionery M	[0.422301, 0.361	{'Confectionery M	[]
Ver Peliculas is a	['Metaverse Auct	Services	Alarms, Surveilla	All Other Busine	['Media Producti	[0.3051482, 0.24	{'Media Producti	[]

3.2 The actual labelling

TL;DR: It's inefficient :))) "I am joking around, but honestly, it doesn't seem that great,at least at first glance. If you made it through all of this, I truly appreciate it, but you will not see me fall. To begin with, I want to explain what I did in the labeling stage. I applied the FAISS model, which I may not have explained well until now. As can be seen in the images, the same companies receive the correct labels depending on the weight adjustments, or they receive nothing at all, as there is quite a bit of volatility involved. FAISS transforms a large set of vectors into an optimized structure for fast retrieval. Once indexed, it can find the nearest neighbors of a query vector in just milliseconds, even when dealing with databases containing millions or billions of vectors. To ensure the correctness of the embeddings at this stage, it is necessary to eliminate noise from the tags, sector, category, and niche. I used it as a placeholder for my tree ideea because it's not computational heavy and it is scalable. This requires a bit more fine-tuning; unfortunately, in the meantime, I "ran out of a laptop." Of course, it wasn't strictly because of running the model—it was already quite used, but once it encountered a bit more heat, the battery started to swell.

As a result, I haven't really had the opportunity to do fine-tuning, and I ran the code on a friend's computer to obtain the output data.

description	business_tags	sector	category	niche	candidate_labels	candidate_sim_s	zero_shot_scores	final_labels
Welchcivils is a c	['Construction S	Services	Civil Engineering	Other Heavy and	['Pipeline Constr	[0.5542776, 0.50	{'Pipeline Constructic	['Residential Electrical Services', 'Commercial Electrical Services', 'Commercial Construction Services', 'Electric Line Construction']
Kyoto Vegetable	['Wholesale', 'Du	Manufacturing	Fruit & Vegetabl	Frozen Fruit, Jui	['Agricultural Equ	[0.3943376, 0.39	{'Agricultural Equipm	[]
Loidholdhof Inte	['Living Forms', 'I	Manufacturing	Farms & Agricult	All Other Miscell	['Bakery Product	[0.47594708, 0.4	{'Bakery Production S	[]
PATAGONIA Ch	['Automotive Boc	Services	Auto Body Shop	Automotive Body	['Coffee Process	[0.34127778, 0.3	{'Coffee Processing S	[]
Stanica WODNA	['Cultural Activiti	Services	Boat Tours & Cr	Scenic and Sigh	['Welding Servic	[0.24252653, 0.2	{'Welding Services': (	['Travel Services']
BIQ Benefícios i	['Healthy Food C	Services	Payment Proces	Financial Transa	['Food Processin	[0.42576373, 0.4	{'Food Processing Se	['Food Processing Services', 'Financial Services']
EnvCompli is a c	['Source Control'	Services	Environmental E	Environmental C	['Environmental	[0.48546916, 0.4	{'Environmental Cons	['Environmental Consulting', 'Commercial Irrigation Systems', 'Testing and Inspection Services', 'Compliance Services']
Corrugated Bear	['Cut Sheets', 'Fc	Manufacturing	Building Material	Engineered Woc	['Sheet Metal Se	[0.46739453, 0.4	{'Sheet Metal Service	['Commercial Construction Services', 'Welding Services']
The company sp	['C Packaging S	Wholesale	Paper & Cardbo	Industrial and Pe	['Paper Producti	[0.5937568, 0.48	{'Paper Production S	['Printing Services']
Indira Caterz An	['Event Organiza	Wholesale	Restaurant Equi	Other Commerci	['Catering Servic	[0.51983654, 0.4	{'Catering Services':	['Catering Services']
Halle Heat Exch	['Technology Pro	Manufacturing	Industrial Air Sol	Power Boiler and	['Boiler Installatic	[0.47372207, 0.4	{'Boiler Installation S	['Industrial Machinery Installation', 'Commercial Communication Equipment Installation']
Ballymacarbry E	['Sustainable Ho	Services	Housing Prograr	Other Communit	['Community Enç	[0.3356004, 0.31	{'Community Engage	['Community Engagement Services', 'Single Family Residential Construction']
An Nhien Scienti	['Sterilization Prc	Manufacturing	Laboratory Equi	Analytical Labor	['Laboratory Ser	[0.50725615, 0.4	{'Laboratory Services	['Veterinary Health Centers', 'Veterinary Services']
Anovei is a com	['Innovative Solu	Manufacturing	Medical Supply I	Irradiation Appar	['Air Duct Cleani	[0.40040606, 0.3	{'Air Duct Cleaning S	[]
Db Motor Servic	['Diagnostic Serv	Services	Auto Services	Specialized Autc	['Commercial Ele	[0.3825289, 0.36	{'Commercial Electric	[]
DOO Bački Petr	['Production', 'WI	Manufacturing	Cleaning Equipn	Soap and Other	['Pet Food Manu	[0.38278466, 0.3	{'Pet Food Manufactu	[]
Reparación Lava	['Home Care', 'Te	Services	Appliance Repai	Appliance Repai	['Confectionery M	[0.4414188, 0.38	{'Confectionery Manu	[]
Ver Peliculas is a	['Metaverse Auct	Services	Alarms, Surveilla	All Other Busine	['Media Producti	[0.32750908, 0.2	{'Media Production S	[]
Gina Crow is an	['Virtual Producti	Services	Video and Audio	Motion Picture a	['Media Producti	[0.30884817, 0.2	{'Media Production S	[]

r.i.p