University of Padova

Department of Mathematics "Tullio Levi-Civita"

Master Thesis in Data Science

Health Analysis of Italian Companies through Text Mining Exploration of Supplementary Notes

Supervisor
Prof. Alessandro Sperduti
University of Padova

Company Supervisor
Niccolò Stamboglis
InfoCamere S.C.p.A.

Master Candidate
Jacopo Magliani

Student ID
2040912

Academic Year

2023-2024

"I must not fear. Fear is the mind-killer. Fear is the little-death that brings total obliteration. I will face my fear. I will permit it to pass over me and through me. And when it has gone past I will turn the inner eye to see its path. Where the fear has gone there will be nothing. Only I will remain."
— Frank Herbert, Dune

Abstract

This thesis aims to study the valorization of the data in the supplementary notes filed by companies together with the financial statements. This data, currently unstructured, requires preprocessing for an effective use and to satisfy business purposes such as the provision of digital services to the Chambers of Commerce, Public Administration and other users. This thesis focuses on the exploration of the supplementary notes and the application of natural language processing tools for the valorization of the filed information, with applications to cooperative societies.

Contents

Listing of figures

Listing of acronyms

JSON	JavaScript Object Notation
XBRL	eXtensible Business Reporting Language
XML	eXtensible Markup Language
HTML	HyperText Markup Language
NLP	Natural Language Processing
POS	Part Of Speech tagging
NER	Named Entity Recognition
LLM	Large Language Models
RAG	Retrieval Augmented Generation
BERT	Bidirectional Encoder Representations from Transformers
MLM	Masked Language Model
NSP	Next Sentence Prediction
DeBERTa	Decoding-enhanced BERT with disentangled Attention mechanism
GDES	Gradient Disentangled Embedding Sharing
RTD	Replace Token Detection
REA	Repertorio Economico Amministrativo

1
Introduction

The primary objective of this thesis is to provide a detailed overview of the methodologies applied and the results achieved during the candidate's internship at InfoCamere S.C.p.A. of Padova.

Throughout this experience the main focus was to study the financial situation of Italian companies through an in-depth analysis of the supplementary notes associated with financial data using Text Mining techniques.

The ultimate goal was to define data extraction methods to be applied to supplementary notes in order to identify peculiarities and business behaviors not immediately evident in other traditional financial documents but capable of providing clarity and revealing new patterns related to business risk.

As the consortium company of Italian Chambers of Commerce for informatics, InfoCamere S.C.p.A. plays a crucial role in the country's digital landscape [1], its main mission is oriented towards the digital transformation of businesses, providing innovative and reliable solutions for the management of corporate information. Originally founded as Cerved in Padova in 1974 by Professor Mario Volpato, who was then President of the Chamber of Commerce of Padova and Professor of Probability Theory at the University of Padova, the company has always overseen and managed the Business Register, becoming the official registry of Italian companies. This register is available to citizens, businesses, professionals and public administration, contributing significantly to the analysis of the value of Italian companies and to the modernization and efficiency of the economic system.

This introduction aims to introduce the reader to the context in which modern Data Science methods, particularly Text Mining, are employed in the financial sector. Additionally, it outlines the general objectives that guided the activity carried out at InfoCamere, highlighting the questions that were sought to be answered through the application of these advanced methodologies, and presents a summary of the results achieved.

To facilitate the reader's understanding of the thesis structure a detailed overview of the subsequent chapters is provided. Each section has been carefully designed to create a logical thread connecting the central theme of the thesis, a summary of the tools used with their competencies and specific challenges, the applied methodologies, and the results obtained.

1.1 Data Science Methods for Finance

In recent years with the advancement and evolution of technology we have witnessed an increasingly complete digitization of every sector of society. This transformation has led to an exponential increase in the amount of data available, making the economic and financial context particularly rich in opportunities and challenges. In this scenario the management and interpretation of such information have become crucial activities for making correct business decisions and for developing services that are increasingly efficient and flexible, capable of adapting to the changing needs of the market.

In this situation, Data Science [2] emerges as a fundamental tool. Originally born as a set of multidisciplinary techniques combining knowledge from various fields including computer science, statistics and mathematics, its main mission is to extract and interpret insights from large amounts of data. If used appropriately, Data Science allows for the optimal utilization of gathered information, offering significant competitive advantages. Indeed, thanks to advanced data analysis techniques it enables the identification of hidden patterns and correlations, thus refining existing services or creating new ones. This predictive capability enables companies to anticipate future trends and adapt their business strategies accordingly, offering a significant advantage in an ever-changing economic environment.

Machine Learning algorithms [3], a subset of Artificial Intelligence, empower Data Science by enabling systems to automatically learn and improve from experience without being explicitly programmed. By exploiting these methods financial institutions can delve deeper into data analysis, uncovering intricate patterns and insights that might have otherwise remained hidden. This amalgamation of Data Science and Machine Learning facilitates more accurate predictions, enhances risk management strategies, automates decision-making processes, and ultimately allows more informed and agile financial decision-making.

As a result, the synergy between Data Science and Machine Learning can prove decisive in the financial context, with its potential applications constantly evolving, in line with the advancement of technology and the increase in available data and its variation.

A first demonstration of how the adoption of Data Science methods can prove fundamental is in managing risks associated with a company's operations, as predictive algorithms can identify potential risks and the most effective strategies to mitigate them [4]. One of the most relevant examples of this is the management of credit risk through credit scoring models within the field of Credit Risk Management. These models exploit advanced data analysis techniques to assess the creditworthiness of individuals or companies, providing financial institutions with a valuable tool for making credit decisions quickly, accurately and efficiently.

The innovation provided by Machine Learning models lies in their ability to exploit not only a vast array of financial and behavioral data but also historical data, thus identifying patterns and correlations not easily detectable through traditional models. This methodology allows for more informed and timely decisions, thereby contributing to the reduction of financial risks and optimization of credit management strategies.

Another significant example of the utility of Data Science is in fraud detection through advanced Machine Learning models. The increasing complexity of financial transactions and the rise of digital threats have made it essential to employ new sophisticated techniques to identify and prevent these fraudulent activities [5], which pose a constant threat to banks, credit card companies and other financial operators. Fraud detection models based on Data Science utilize Machine Learning algorithms to analyze large volumes of data in real-time and learn from historical patterns of legitimate transactions, incorporating a wide range of information such as geolocation, spending habits and credit card usage patterns. This allows them to identify significant deviations or anomalous behaviors that may indicate fraudulent activity. Thanks to their real-time analysis capability and continuous learning the models are key tools for countering increasingly sophisticated threats in the landscape of digital transactions.

A widely adopted method of Data Science especially in the sectors of video or music streaming and finance is customer segmentation, which allows for optimizing customer experience and maximizing service profitability [6]. The approach to customer segmentation is based on in-depth analysis of demographic, behavioral and transactional data of users, enabling the identification of homogeneous groups or clusters. The study of these clusters allows to adapt marketing and sales strategies in a targeted way, thereby improving customer satisfaction. The collected data is analyzed and leveraged by complex Machine Learning algorithms to identify patterns that can divide customers into homogeneous groups based on various criteria such as age, income and investment preferences. The goal is to enable companies to better understand customer needs and personalize products and service offerings accordingly. Therefore customer segmentation represents a strategic tool that goes beyond mere data analysis, serving as a guide for the formulation of effective customer-oriented strategies in the current competitive landscape.

The success of the reported Data Science methods and techniques would not have been possible without an adequate preliminary phase of analysis, cleaning and preprocessing of the initial data. In fact, rarely the available information are in a format already suitable for immediate use, but instead require careful study and appropriate modifications, which may also vary depending on the type of subsequent application adopted.

Even the supplementary notes subject to the thesis were not exempt from these considerations, with cleaning and preprocessing activities requiring approximately 80% of the total time spent in the entire work process. These activities included thoroughly analyzing the content of the supplementary notes to better understand their structure for extracting information, cleaning the retrieved data from unwanted elements, studying the obtained results and in case repeating the process to refine it. Only after various methods of recovery and cleaning of the supplementary notes a level of data quality was achieved that allowed for valuable Text Mining analyses and applications.

The last example of Data Science methods is also the main theme of the thesis: Text Mining techniques. The recent proliferation of textual data such as corporate disclosures, informal internal communications related to business management and online news articles regarding companies and their future plans has posed a significant challenge in taking full advantage of the information contained in this vast amount of data. To address this complexity it becomes essential to adopt efficient techniques for organizing, managing and analyzing texts, taking into account the intrinsic challenges associated with linguistic characteristics.

In this context the role of Text Mining proves to be fundamental, as its primary purpose is to extract knowledge from textual data that is not immediately accessible [7]. Born in the early 1990s as a result of the intersection of various disciplines such as Data Mining, Knowledge Discovery, Information Retrieval, Statistics and Natural Language Processing, Text Mining has seen a significant increase in its utility thanks to the widespread adoption of Big Data and the increase in available computational power. Its applications are multiple and span across various sectors and industries. In the context of accounting the use of Text Mining can be particularly valuable as its capabilities can help clarify, integrate and interpret the numerical quantitative data present in financial documents, thus providing a deeper understanding of the information contained.

The use of advanced text analysis techniques to process financial data can lead to new perspectives and revelations, contributing to improve the understanding of the business landscape and supporting more informed decisions.

1.2 The purpose of the thesis

This thesis aims to apply data mining and text-analytics techniques to the financial statements of Italian companies as presented within the official balance sheets stored in the Italian Business Registry. In this thesis, the candidate will present state-of-art methods for data extraction, cleaning and analysis applied in the context of administrative data. The type of documents to be analyzed are a subset of the individual supplementary notes deposited in the year 2021. The purpose of the analysis presented within this thesis is the extraction of information concerning the drivers of financial decisions of cooperative societies. The results of this analysis, together with providing insights on the decision making processes of such companies, will lay the foundation for a working prototype of a novel solution for the exploration of supplementary notes data to be developed within InfoCamere.

In summary, the most relevant results which are further elaborated in the subsequent chapters include:

- having acquired a suitable understanding of data structure for their manipulation;

- having defined an effective cleaning procedure to retrieve and to structure supplementary notes data into a more easily usable format than the original;

- having conducted valuable exploratory analysis of the contents of the supplementary notes for the Chambers of Commerce of Veneto;

- having created Text Mining methods to address credible business cases related to cooperative information, such as verifying the declaration of having a mutual purpose, being registered in a register or complying with articles of the civil code;

- having compared own methods with the performances of two pre-trained Large Language Models [8][9], that are different fine-tuned versions of the same model DeBERTa V3 [10];

- having defined a prototype of a web interface that utilizes the devised methods or a Large Language Model to answer user questions on any section of the supplementary notes of available documents.

1.3 The structure of the thesis

This thesis is organized as follows.

Chapter 2 provides a detailed analysis of the supplementary notes, their context and role, deepening the understanding of these documents and the importance of unstructured data contained within them.

Chapter 3 presents the softwares provided for the study of supplementary notes, Elasticsearch [11] and Kibana [12], along with the XBRL taxonomy [13] that regulates the structure of the financial documents, including supplementary notes.

Chapter 4 describes the procedure applied to extract data from the supplementary notes and the exploratory analyses conducted to verify the valuation of the fields in the supplementary note.

Chapter 5 outlines the Text Mining applications adopted to solve specific tasks of interest simulating real business cases.

Chapter 6 describes how two pre-trained Large Language Models [8][9], fine-tuned versions of DeBERTa V3 [10], are utilized to address the same tasks introduced in Chapter 5 in order to make a performance comparison of the two approaches.

Chapter 7 presents a prototype of a web interface that allows users to pose questions on the textual data of the supplementary notes.

Chapter 8 reports the conclusions drawn from the application of Text Mining techniques on supplementary notes.

Finally, there is an Appendix 9 with the algorithms created and the Bibliography.

2 Supplementary notes: understanding their content

This Chapter introduces the supplementary note, the document upon which the analyses and Text Mining techniques were focused during the internship. This Chapter of the thesis aims to familiarize the reader with the context of the supplementary note and its strategic role in corporate financial management by explaining its various aspects and concepts that characterize it. Additionally, this chapter can be seen as an introductory guide for non-specialized readers in the field, as it illustrates and explains the key terms and concepts of the supplementary note that have been studied and analyzed.

In the following sections the overall framework in which the supplementary note is situated is outlined, emphasizing its importance in the comprehensive interpretation of the financial statement.

2.1 The financial statement and its different formats

According to Article 2423 of the Italian Civil Code (Drafting of the financial statement) [14], it is required that at the end of each administrative year every company must prepare the annual financial statement. This act, composed of all the accounting documents, plays a fundamental role in defining the health status of a company by pursuing the principle of truth and allowing the assessment of its financial and asset situation.

The financial statement data must be presented according to the XBRL taxonomy format, which is explained in detail in Chapter 3.

Through the financial statement investors, creditors and various stakeholders have the opportunity to evaluate the performance and financial stability of a company, benefiting from a comprehensive and clear overview of its economic activities.

The responsibility for preparing the financial statement falls on the directors and it is articulated in four fundamental components: the balance sheet, the revenue account, the cash flow statement and the supplementary note.

As reported by the Business Register of Commerce [15], in Article 2423 of the Civil Code there are various types of financial statements to be deposited depending on the performances and characteristics of the company.

Since 2016 the administrators responsible for its compilation must include in the financial statement the balance sheet, the revenue account, the cash flow statement and the supplementary note. This is the ordinary form of the financial statement and companies obliged to prepare it in this structure include publicly traded companies listed on stock exchanges that have issued securities traded on regulated markets or exceed the limits for preparing the abbreviated financial statement.

The option to prepare the financial statements in abbreviated form is reserved for capital companies that in the first fiscal year or subsequently for two consecutive fiscal years have not exceeded two of the following limits to be identified as small enterprises: a total of assets in the balance sheet of €4,400,000, total revenue from sales and services of €8,800,000, an average number of employees during the fiscal year of 50 units. For companies meeting these conditions it is possible to prepare the financial statement in either ordinary or abbreviated form, excluding the cash flow Statement.

A similar scenario applies to micro-enterprises, which are companies that in the first fiscal year or subsequently for two consecutive fiscal years have not exceeded two of the following limits: a total of assets in the balance sheet of €175,000, total revenue from sales and services of €350,000, an average number of employees during the fiscal year of 5 units. For micro-enterprises the financial statement may consist only of the balance sheet and the revenue account, with the form, structure, and content identical to those of the financial statement in abbreviated form. However, there is the possibility for this type of enterprises to present the complete financial statement with supplementary note and if applicable the cash flow statement.

2.2 The components of the financial statement

The documents that form the financial statement are the balance sheet, the revenue account, the cash flow statement and the supplementary note.

Financial Statement			
Balance Sheet	Revenue account	Cash Flow Statement	Supplementary Note
- Assets and properties	- Value of production	- Summarized cash flows	- Valuation criteria for financial statements
- Sources of financing	- Production costs		- Information about the balance sheet
	- Income and expenses from financial activities		- Details about the revenue account
	- Adjustments to the value of financial assets and liabilities		- Other aspects of the company

The primary role of the balance sheet is to provide an overview of the financial situation of the company reporting the value of assets and capitals available. The document highlights the components and activities representing the active and passive side of the company according to the schema defined by Article 2424 of the Civil Code [16].

The active side indicates how company resources have been employed, distinguishing between elements that are liquid or can be liquidated in the short term within the financial year and those that can be liquidated in the medium-long term. The items comprising the active side of the balance sheet are:

1. credits towards shareholders for payments still due;
2. intangible, tangible, and financial assets;
3. current assets consisting of inventories, receivables, other financial assets that do not constitute fixed assets and cash and cash equivalents;
4. accruals and deferrals.

In summary, the active side include the assets and properties owned by the company such as liquid assets in cash and current accounts, receivables from customers, tangible assets such as machinery and equipment, and intangible assets such as patents and trademarks.

The passive side on the other hand is made of the sources of financing for the company and consists of:

1. the net worth, including the share capital and reserves;
2. provisions for risks and charges;
3. severance pay for employees;
4. debts;
5. accruals and deferrals.

The revenue account represents the economic dynamics of the company during a reporting period, showing how the revenues generated from business activities are transformed into net profits or losses through costs and expenses. Article 2425 of the Civil Code [17] explains in detail the items it should include:

1. value of production, including revenues from sales and services, changes in work in progress and finished goods inventories, and increases in fixed assets due to internal work;
2. production costs for raw materials, services, personnel, depreciation and devaluations, provisions for risks, and changes in raw material inventories;
3. income and expenses from financial activities including investments and others;
4. adjustments to the value of financial assets and liabilities, including revaluations of investments and fixed assets, income taxes, and net profit (loss) for the period.

In the cash flow statement as established by Article 2425-ter of the Civil Code [18] all cash flows that occurred during the fiscal year are summarized. Specifically, it outlines the sources that have increased available cash funds and the uses that, conversely, have led to their reduction.

Finally the supplementary note plays a crucial role as it serves as a collection of supplemental information to the other financial statements, providing a more comprehensive and truthful representation of the company's financial status.

This document offers more detailed explanations of the information already present in the other statements, introducing new information if necessary. In doing so, it provides essential depth for a comprehensive and informed evaluation by stakeholders.

In summary the explanatory notes serves a dual purpose: complementing the data provided by the balance sheet and revenue account, which by their nature are concise and quantitative, by offering clarifications and details on specific accounting choices and justifying certain business behaviors, especially in cases where changes have been made to the mandatory schema defined by the Civil Code; providing information about the company not necessarily related to the financial statement but contributing to a clearer and broader understanding of the entity. These additional details highlight the importance of the supplementary notes, as they cannot be included in other documents.

Article 2427 of the Italian Civil Code [19] provides a detailed list of the mandatory minimum contents to be included in the supplementary notes, both quantitative and qualitative elements. The quantitative elements allow for a more in-depth analysis of the balance sheet and revenue account, while the qualitative elements serve to describe specific aspects of business management. In general, the various data contained in the supplementary notes can be grouped into four sections: the valuation criteria applied in estimating certain items in the financial statements, information about the balance sheet, details about the revenue account and other aspects of the company such as commitments made, the number of employees, and compensation for administrators and auditors.

The mandatory contents of the supplementary notes are as follows:

1. the criteria applied in valuing items in the financial statements;

2. movements of assets;

3. composition of the items “establishment and expansion costs” and “development costs”;

4. measurements and justifications for changes of value of tangible and intangible assets;

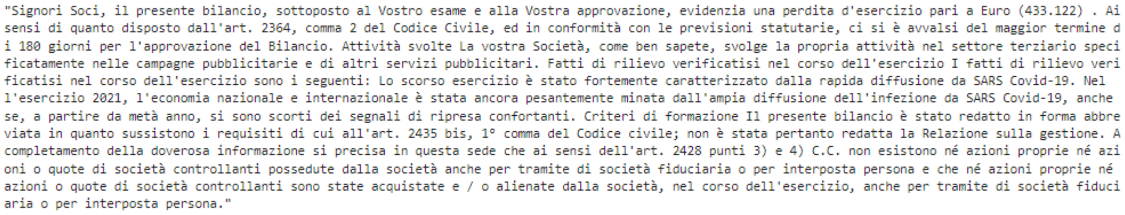
5. changes occurred in the composition of other items in the active and passive side;

6. list of investments in controlled and affiliated companies;

7. amount of credits and debts with a remaining duration exceeding five years, and debts secured by real guarantees on company assets;

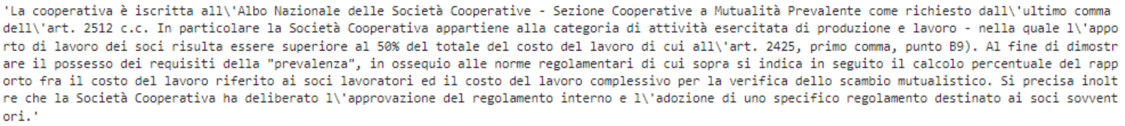
8. changes in currency exchange rates subsequent to the end of the financial year;

9. amount of credits and debts for transactions involving the obligation for the buyer to resell;
10. composition of the items “accruals and deferrals” and “other funds” in the balance sheet;
11. origin, availability and distributability of equity reserves;
12. capitalized financial charges;
13. commitments not reflected in the balance sheet;
14. distribution of revenues from sales and services by activity and geographical area;
15. gains from investments other than dividends;
16. composition of the item relating to interests and other financial charges;
17. amount and nature of individual revenue or cost elements of exceptional size or impact;
18. prepaid and deferred taxes;
19. average number of employees, categorized;
20. compensation for administrators and auditors;
21. fees due to the statutory auditor or the statutory auditing company;
22. number and nominal value of each category of shares;
23. other financial instruments issued by the company;
24. shareholders’ loans;
25. assets allocated to a specific transaction;
26. revenues from financing allocated to a specific transaction;
27. financial leasing operations;
28. off-balance sheet transactions;
29. name and registered office of the company preparing the consolidated financial statements;
30. proposal for the allocation of profits or coverage of losses.



```
"Signori Soci, il presente bilancio, sottoposto al Vostro esame e alla Vostra approvazione, evidenzia una perdita d'esercizio pari a Euro (433.122) . Ai
sensi di quanto disposto dall'art. 2364, comma 2 del Codice Civile, ed in conformità con le previsioni statutarie, ci si è avvalsi del maggior termine d
i 180 giorni per l'approvazione del Bilancio. Attività svolte La vostra Società, come ben sapete, svolge la propria attività nel settore terziario speci
ficatamente nelle campagne pubblicitarie e di altri servizi pubblicitari. Fatti di rilievo verificatisi nel corso dell'esercizio I fatti di rilievo veri
ficatisi nel corso dell'esercizio sono i seguenti: Lo scorso esercizio è stato fortemente caratterizzato dalla rapida diffusione da SARS Covid-19. Nel
l'esercizio 2021, l'economia nazionale e internazionale è stata ancora pesantemente minata dall'ampia diffusione dell'infezione da SARS Covid-19, anche
se, a partire da metà anno, si sono scorti dei segnali di ripresa confortanti. Criteri di formazione Il presente bilancio è stato redatto in forma abbre
viata in quanto sussistono i requisiti di cui all'art. 2435 bis, 1° comma del Codice civile; non è stata pertanto redatta la Relazione sulla gestione. A
completamento della doverosa informazione si precisa in questa sede che ai sensi dell'art. 2428 punti 3) e 4) C.C. non esistono né azioni proprie né azi
oni o quote di società controllanti possedute dalla società anche per tramite di società fiduciaria o per interposta persona e che né azioni proprie né
azioni o quote di società controllanti sono state acquistate e / o alienate dalla società, nel corso dell'esercizio, anche per tramite di società fiduci
aria o per interposta persona."
```

Figure 2.1: Example of the descriptive field "Introduction to the financial statement".



```
'La cooperativa è iscritta all\'Albo Nazionale delle Società Cooperative - Sezione Cooperative a Mutualità Prevalente come richiesto dall\'ultimo comma
dell\'art. 2512 c.c. In particolare la Società Cooperativa appartiene alla categoria di attività esercitata di produzione e lavoro - nella quale l\'appo
rto di lavoro dei soci risulta essere superiore al 50% del totale del costo del lavoro di cui all\'art. 2425, primo comma, punto B9). Al fine di dimostr
are il possesso dei requisiti della "prevalenza", in ossequio alle norme regolamentari di cui sopra si indica in seguito il calcolo percentuale del rapp
orto fra il costo del lavoro riferito ai soci lavoratori ed il costo del lavoro complessivo per la verifica dello scambio mutualistico. Si precisa inolt
re che la Società Cooperativa ha deliberato l\'approvazione del regolamento interno e l\'adozione di uno specifico regolamento destinato ai soci sovvent
ori.'
```

Figure 2.2: Example of the descriptive field "Introduction to the cooperative societies".

For small businesses that can prepare the abbreviated form of the supplementary notes it still must include changes in equity, fixed assets, shareholders' loans, issued shares and bonds and investments in other companies.

In conclusion, it's important to emphasize that all companies obligated to prepare the financial statement must also compile the supplementary notes. To provide a glimpse of the complexity of the task at hand, Figure 2.1 reports an example of the descriptive field "Introduction to the financial statement" which is required to all companies. As we can see, this text presents an important presence of administrative jargon that needs to be taken into consideration within the analysis.

To add additional complexity, the form and content of the supplementary notes might vary across the individual forms. More specifically, corporations must adhere to the prescribed financial statement formats outlined in the Civil Code and are required to publish the financial statements to provide information on management performance to stakeholders.

To give a reader with a sense of complexity on the type of text, Figure 2.2 reports the descriptive text reported within the field "Introduction for cooperative societies". We can see that even for an introductory text to a single section of the supplementary note, a specific vocabulary is still used, both for the declaration of compliance with certain articles of the Civil Code and for the approval of internal company regulations.

On the other hand, partnerships and individual businesses, while not having mandatory formats and not being obliged to publish financial statement, prepare it mainly for internal needs and only occasionally for tax compliance or external financing requests.

3

Analytical infrastructure: using ElasticSearch for textual analysis

This Chapter introduces the data analysis system ElasticSearch along with its accompanying data visualization platform Kibana and the XBRL standard used for storing financial statements. Specifically, it explains how ElasticSearch works for textual analysis and information retrieval, the concept of indexing, the advantages and challenges of using this tool in the financial domain and how it interacts with Kibana. Then, the Chapter introduces the XBRL standard in which the financial statements are presented, its context and how it is utilized.

3.1 ElasticSearch and Kibana

In the digital age of the financial sector efficient management, analysis, and visualization of data have become crucial components. This is particularly evident when dealing with large and diverse datasets that require operations such as extraction and textual analysis of targeted information. These tasks can be challenging both in terms of configuration and computational resources. In this context, ElasticSearch and Kibana can prove to be valuable tools for addressing these challenges associated with data management and analysis.

3.1.1 ElasticSearch

ElasticSearch [11] is a distributed search server and open-source data analysis system based on Apache Lucene, specifically designed for efficient indexing and real-time searching of large volumes of textual data managed as JSON documents [20]. It is used in particular for performance monitoring, full-text search and analysis of large datasets, leveraging the elasticity of its system, making it suitable for distributed environments.

ElasticSearch differs from relational databases for using a document-oriented database instead of predefined tables and relationships. This means it organizes and stores data in the form of JSON documents, which can vary in structure from document to document as there is no fixed and predefined data model, allowing greater flexibility in data management and adaptation to evolving application needs. This document-oriented architecture is fundamental to the flexibility and efficiency of ElasticSearch, enabling easy handling of large volumes of unstructured and semi-structured data and conducting complex searches and analyses on them quickly and efficiently.

Overall the functioning of ElasticSearch is based on key concepts such as inverted indexing and distribution across clusters. This allows complex searches and efficient and scalable analysis of large volumes of information.

The first process involves the insertion of raw data which is then analyzed, normalized, and indexed in ElasticSearch. Once this process is completed, users can execute more or less sophisticated queries on their data and obtain complex summaries of it. Additionally, by using Kibana users can also obtain effective visualizations of the collected information.

The foundation of ElasticSearch's power lies in its indices, which are collections of logically related documents due to their similar structure, and they are used to facilitate the search and analysis of raw information.

Before data can be inserted into ElasticSearch a general index that will contain them must be defined. Indices are defined with a mapping, which specifies how the fields of documents are structured and indexed, correlating field names and their data types with corresponding values, which can be numeric, text strings, boolean values, dates, or others. This mapping can be implicit if ElasticSearch automatically infers it, or explicit if specified by the user.

Once a document is inserted inverted indexing occurs, a fundamental technique used in various search engines to improve the efficiency of query execution. It essentially involves creating an index that maps each term present in all available documents to a list of documents containing that term.

In more detail, when a document is inserted into a general index ElasticSearch analyzes its entire content and breaks it down into individual terms or tokens (this process is known as tokenization): for example, a sentence is decomposed into the individual words that compose it, ignoring punctuation and non-significant characters. Subsequently, additional techniques for text analysis and cleaning may be applied to ensure the most efficient and accurate search possible.

Then comes the creation of the inverted index, which maps each term to the list of documents in which that term appears. The inverted index typically consists of three main components: the term itself, a list of document IDs in which it appears, and additional information such as its frequency in the document or position.

To further optimize query execution on document terms ElasticSearch also employs shards and clusters. Each index can be divided into one or more shards, which are the basic units of storage and search in ElasticSearch. The primary shard contains the actual data, while replica shards are duplicates created to ensure redundancy and fault tolerance, allowing for workload distribution to improve ElasticSearch's scalability and resilience. Additionally, shards can be distributed across nodes in clusters, which are interconnected sets designed to leverage parallelization and enhance the overall performance of the system.

Once all received documents have been processed and inverted indexing has been performed for the terms they contain, ElasticSearch can effectively resolve received queries without having to scan the texts but by directly returning the relevant ones. This allows for quick and efficient full-text search of the requested information.

ElasticSearch not only returns the list of documents that match a given query but also provides a score for each document. This value, known as a relevance score, is obtained through a scoring algorithm based on Term Frequency-Inverse Document Frequency (TF-IDF) [21] along with other factors:

$$TF = \frac{\text{number of times the term appears in a document}}{\text{total number of terms of that document}}$$

$$IDF = log(\frac{\text{number of the documents in the corpus}}{\text{number of the documents in the corpus that contain that term}})$$

$$TF - IDF = TF * IDF.$$

The calculation of TF-IDF takes into account the frequency of a term in a document and its rarity in the entire index, assigning a higher score to terms that appear frequently in a document but rarely in the entire index. Additionally, the score is normalized considering the length of the document, ensuring that a short text is evaluated in the same way as a lengthy one where term frequency would naturally be higher simply due to its size. Other factors that ElasticSearch can utilize include assigning greater weight to certain fields over others or considering the proximity of terms within the document. Once ElasticSearch resolves a query it returns the results ordered by the total score, displaying the most relevant documents first. This score can be further utilized to filter the obtained results more effectively.

In Figure 3.1 it is shown an example of query execution with the aim of obtaining the number of documents that meet certain requirements. The first two queries return the number of documents related to the Chamber of Commerce of Padova that contain the Introduction to the Explanatory Notes taxonomy tag, while the last query considers the documents of all provinces. In the example provided two peculiarities can be noticed:

1. when numerous documents are available if the duration of the search call is not increased approximate results may be obtained; in fact, the first query returns that there are at least 10,000 documents that satisfy it ("greater than or equal"), while the second query reports the exact number;

2. despite the size of the search field the search times in all three queries are very fast (about 0.01 seconds), highlighting the efficiency of ElasticSearch.

Figure 3.2 shows query results in more detail:

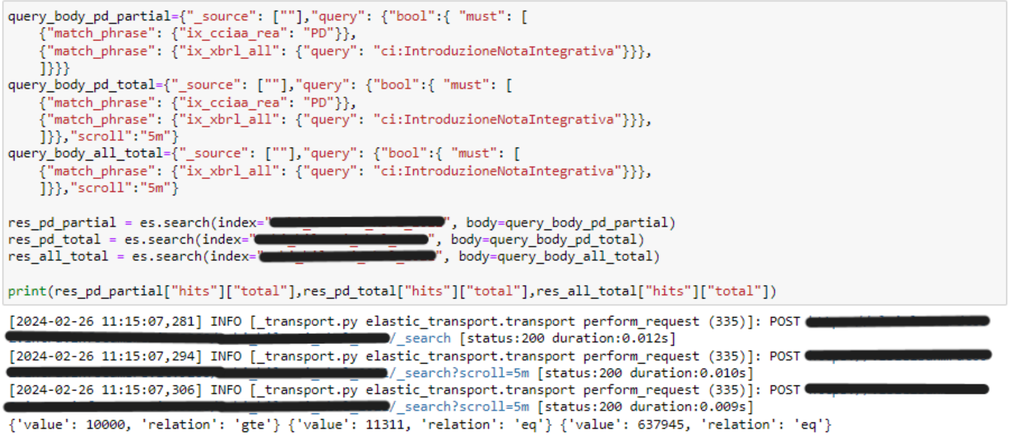
1. the identifiers of the first ten documents with the highest scores are reported in order to allow further checks, but it is possible to modify the queries to return a complete list of identifiers and not limiting to the first ones;

2. regardless of the duration of the call the documents reported are the same.

3.1.2 Kibana

Once the documents have been indexed the data can be analyzed and visualized through Kibana. Originally developed as an open-source data visualization and analysis platform designed to work in synergy with Elasticsearch [12], it is often used for advanced analysis and searching of large volumes of data.

While Elasticsearch manages data indexing and searching, Kibana provides tools for visualizing and analyzing this data. This combination offers a comprehensive solution for real-time analysis and the creation of interactive dashboards.



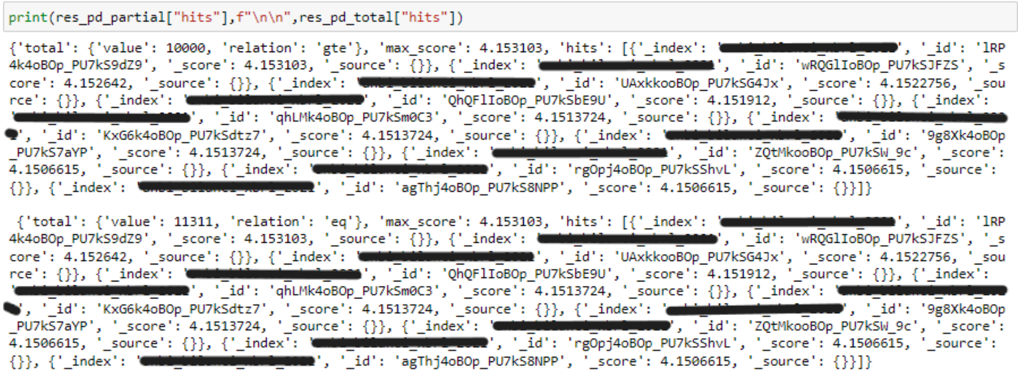
```
query_body_pd_partial={"_source": [""],"query": {"bool":{ "must": [
    {"match_phrase": {"ix_cciaa_rea": "PD"}},
    {"match_phrase": {"ix_xbrl_all": {"query": "ci:IntroduzioneNotaIntegrativa"}}},
    ]}}}
query_body_pd_total={"_source": [""],"query": {"bool":{ "must": [
    {"match_phrase": {"ix_cciaa_rea": "PD"}},
    {"match_phrase": {"ix_xbrl_all": {"query": "ci:IntroduzioneNotaIntegrativa"}}},
    ]}},"scroll":"5m"}
query_body_all_total={"_source": [""],"query": {"bool":{ "must": [
    {"match_phrase": {"ix_xbrl_all": {"query": "ci:IntroduzioneNotaIntegrativa"}}},
    ]}},"scroll":"5m"}

res_pd_partial = es.search(index="[redacted]", body=query_body_pd_partial)
res_pd_total = es.search(index="[redacted]", body=query_body_pd_total)
res_all_total = es.search(index="[redacted]", body=query_body_all_total)

print(res_pd_partial["hits"]["total"],res_pd_total["hits"]["total"],res_all_total["hits"]["total"])
```

```
[2024-02-26 11:15:07,281] INFO [_transport.py elastic_transport.transport perform_request (335)]: POST [redacted]
/_search [status:200 duration:0.012s]
[2024-02-26 11:15:07,294] INFO [_transport.py elastic_transport.transport perform_request (335)]: POST [redacted]
/_search?scroll=5m [status:200 duration:0.010s]
[2024-02-26 11:15:07,306] INFO [_transport.py elastic_transport.transport perform_request (335)]: POST [redacted]
/_search?scroll=5m [status:200 duration:0.009s]
{'value': 10000, 'relation': 'gte'} {'value': 11311, 'relation': 'eq'} {'value': 637945, 'relation': 'eq'}
```

Figure 3.1: Compilation time of queries and number of documents retrieved; references to the InfoCamere general index, which contains the documents, are obscured for privacy and security reasons.

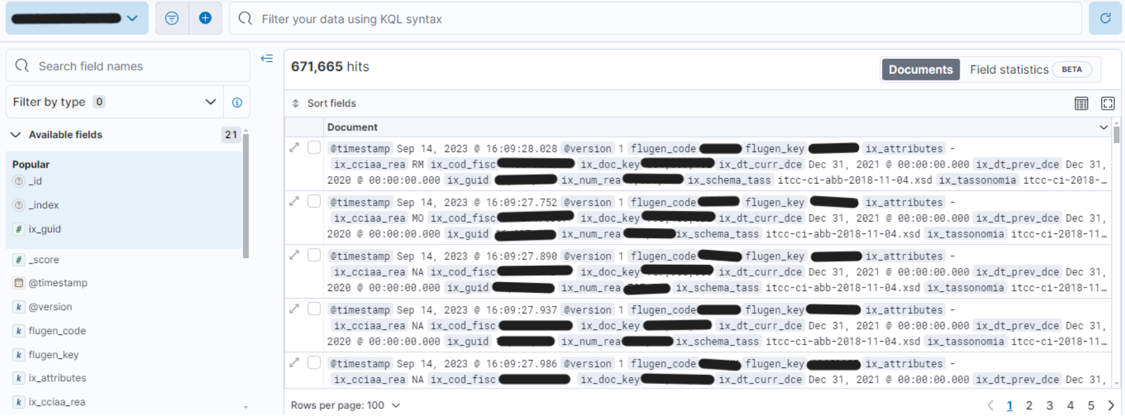


```
print(res_pd_partial["hits"],f"\n\n",res_pd_total["hits"])
```

```
{'total': {'value': 10000, 'relation': 'gte'}, 'max_score': 4.153103, 'hits': [{'_index': '[redacted]', '_id': '1RP
4k4oBOp_PU7kS9dZ9', '_score': 4.153103, '_source': {}}, {'_index': '[redacted]', '_id': 'wRQGlIoBOp_PU7kSJFZS', '_s
core': 4.152642, '_source': {}}, {'_index': '[redacted]', '_id': 'UAxkkooBOp_PU7kSG4Jx', '_score': 4.1522756, '_sou
rce': {}}, {'_index': '[redacted]', '_id': 'QhQFlIoBOp_PU7kSbE9U', '_score': 4.151912, '_source': {}}, {'_index':
'[redacted]', '_id': 'qhLMk4oBOp_PU7kSm0C3', '_score': 4.1513724, '_source': {}}, {'_index': [redacted]
', '_id': 'KxG6k4oBOp_PU7kSdtz7', '_score': 4.1513724, '_source': {}}, {'_index': '[redacted]', '_id': '9g8Xk4oBOp
_PU7kS7aYP', '_score': 4.1513724, '_source': {}}, {'_index': '[redacted]', '_id': 'ZQtMkooBOp_PU7kSW_9c', '_score':
4.1506615, '_source': {}}, {'_index': '[redacted]', '_id': 'rgOpj4oBOp_PU7kSShvL', '_score': 4.1506615, '_source':
{}}, {'_index': '[redacted]', '_id': 'agThj4oBOp_PU7kS8NPP', '_score': 4.1506615, '_source': {}}]}

 {'total': {'value': 11311, 'relation': 'eq'}, 'max_score': 4.153103, 'hits': [{'_index': '[redacted]', '_id': '1RP
4k4oBOp_PU7kS9dZ9', '_score': 4.153103, '_source': {}}, {'_index': '[redacted]', '_id': 'wRQGlIoBOp_PU7kSJFZS', '_s
core': 4.152642, '_source': {}}, {'_index': '[redacted]', '_id': 'UAxkkooBOp_PU7kSG4Jx', '_score': 4.1522756, '_sou
rce': {}}, {'_index': '[redacted]', '_id': 'QhQFlIoBOp_PU7kSbE9U', '_score': 4.151912, '_source': {}}, {'_index':
'[redacted]', '_id': 'qhLMk4oBOp_PU7kSm0C3', '_score': 4.1513724, '_source': {}}, {'_index': [redacted]
', '_id': 'KxG6k4oBOp_PU7kSdtz7', '_score': 4.1513724, '_source': {}}, {'_index': '[redacted]', '_id': '9g8Xk4oBOp
_PU7kS7aYP', '_score': 4.1513724, '_source': {}}, {'_index': '[redacted]', '_id': 'ZQtMkooBOp_PU7kSW_9c', '_score':
4.1506615, '_source': {}}, {'_index': '[redacted]', '_id': 'rgOpj4oBOp_PU7kSShvL', '_score': 4.1506615, '_source':
{}}, {'_index': '[redacted]', '_id': 'agThj4oBOp_PU7kS8NPP', '_score': 4.1506615, '_source': {}}]}
```

Figure 3.2: Details of the first results of the queries; the indexes of individual documents can be traced back to companies and are therefore obscured for privacy reasons.



```
@timestamp Sep 14, 2023 @ 16:09:28.028 @version 1 flugen_code [redacted] flugen_key [redacted] ix_attributes -
 ix_cciaa_rea RM ix_cod_fisc [redacted] ix_doc_key [redacted] ix_dt_curr_dce Dec 31, 2021 @ 00:00:00.000 ix_dt_prev_dce Dec 31,
2020 @ 00:00:00.000 ix_guid [redacted] ix_num_rea [redacted] ix_schema_tass itcc-ci-abb-2018-11-04.xsd ix_tassonomia itcc-ci-2018-…
@timestamp Sep 14, 2023 @ 16:09:27.752 @version 1 flugen_code [redacted] flugen_key [redacted] ix_attributes -
 ix_cciaa_rea MO ix_cod_fisc [redacted] ix_doc_key [redacted] ix_dt_curr_dce Dec 31, 2021 @ 00:00:00.000 ix_dt_prev_dce Dec 31,
2020 @ 00:00:00.000 ix_guid [redacted] ix_num_rea [redacted] ix_schema_tass itcc-ci-abb-2018-11-04.xsd ix_tassonomia itcc-ci-2018-11…
@timestamp Sep 14, 2023 @ 16:09:27.890 @version 1 flugen_code [redacted] flugen_key [redacted] ix_attributes -
 ix_cciaa_rea NA ix_cod_fisc [redacted] ix_doc_key [redacted] ix_dt_curr_dce Dec 31, 2021 @ 00:00:00.000 ix_dt_prev_dce Dec 31,
2020 @ 00:00:00.000 ix_guid [redacted] ix_num_rea [redacted] ix_schema_tass itcc-ci-abb-2018-11-04.xsd ix_tassonomia itcc-ci-2018-11…
@timestamp Sep 14, 2023 @ 16:09:27.937 @version 1 flugen_code [redacted] flugen_key [redacted] ix_attributes -
 ix_cciaa_rea NA ix_cod_fisc [redacted] ix_doc_key [redacted] ix_dt_curr_dce Dec 31, 2021 @ 00:00:00.000 ix_dt_prev_dce Dec 31,
2020 @ 00:00:00.000 ix_guid [redacted] ix_num_rea [redacted] ix_schema_tass itcc-ci-abb-2018-11-04.xsd ix_tassonomia itcc-ci-2018-11…
@timestamp Sep 14, 2023 @ 16:09:27.986 @version 1 flugen_code [redacted] flugen_key [redacted] ix_attributes -
 ix_cciaa_rea NA ix_cod_fisc [redacted] ix_doc_key [redacted] ix_dt_curr_dce Dec 31, 2021 @ 00:00:00.000 ix_dt_prev_dce Dec 31,
```

Figure 3.3: Kibana Initial Discover Interface; the indexes of individual documents can be traced back to companies and are therefore obscured for privacy reasons.

Kibana provides users with a web interface with various sections each offering different functionalities, including creating customized dashboards, exploring data through queries and creating visualizations such as bar charts, pie charts, geographic maps, and more.

Figure 3.3 shows how the general index from which the analysis work started is presented. Here users can read the general index containing all the data (obscured for privacy reasons), view the initial number of documents (671,665) or those that meet a defined query in the above filter, browse and expand documents and available fields such as the identifier, the referring Chamber of Commerce ("ix_cciaa_rea"). The number of documents available refers to only those uploaded on ElasticSearch, it does not represent all the documents filed by companies.

One of the candidate's initial tasks was to utilize the Discover interface of Kibana, a tool for easier navigation and exploration of the documents, to become familiar with the structure of the uploaded documents and how various items of interest such as the introduction to the explanatory notes are defined . Additionally, the second task parallel to the first was to understand the most effective technique for conducting general data analysis on all documents and how to extract information.

In Section 3.2 a more comprehensive overview of the content of the documents uploaded to ElasticSearch and how the explanatory notes are structured is provided.

To conclude the section related to ElasticSearch and Kibana it is important to highlight a drawback of these tools, namely resource consumption. Despite the qualities and versatility of Elasticsearch in textual analysis and the ease of use of Kibana in relation to the amount of data to be analyzed, these tools can require significant resources in terms of memory, CPU, and disk space, particularly when managing large quantities of data and using complex queries.

Indeed, it is not uncommon to encounter errors related to excessive memory usage and discrepancies in presented results when executing queries that return several thousand documents. This issue can be addressed by interfacing with ElasticSearch through a Python script but memory shortage issues may still arise frequently with complex queries, leading to the interruption of running code.

3.2 XBRL Taxonomy

In this section, the functioning and structure of the taxonomy of the financial statements is introduced.

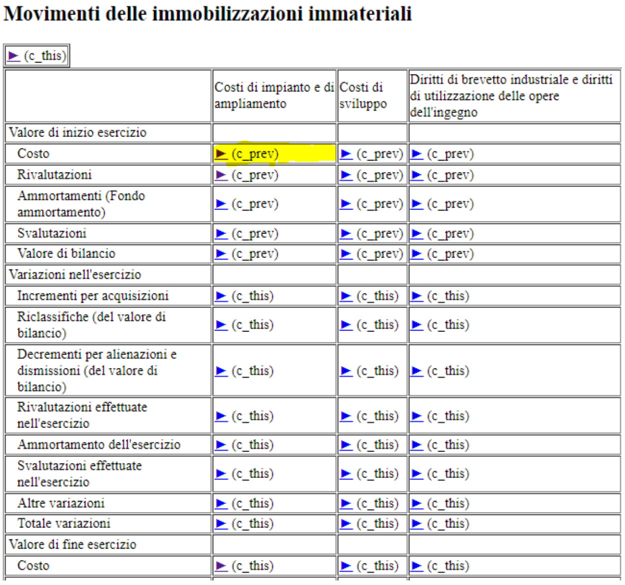
The XBRL taxonomy, which stands for eXtensible Business Reporting Language, is a key element of the standard designed to facilitate the exchange of financial and accounting information between data processing systems [13]. It plays a crucial role in organizing and classifying financial statements standardizing the presentation of information.

It is an XML-based standard [22] designed as a hierarchical tree structure, with the root element at the top and financial concepts branching out as sub-elements, reflecting the logical hierarchy of financial information.

XBRL represents an international computer standard that allows companies to deposit their financial statements with the Business Register [23] in a format that makes the data immediately accessible, guaranteeing their official status derived from the direct responsibility of the company that deposited them. With the Decree of the President of the Council of Ministers of December 10, 2008 [24] the XBRL language was recognized as the mandatory format for depositing financial statements with the Business Register and for presenting economic/financial reporting starting from 2010. Some companies are exempted by law from this obligation such as banks and other financial institutions.

The Business Register provides companies with a basic tool for preparing financial statements in XBRL format, validating them and representing them in PDF [25] or HTML [26] format, both for ordinary and abbreviated forms or for micro-enterprises. The taxonomy and related documentation are available on the website of the "Agenzia per l'Italia Digitale" [27] and on the "XBRL Italia" website [13].

To assist the candidate in consulting the numerous entries of the taxonomy related to various documents of the financial statements two files have been provided: the first, "bilanci.html", was prepared by InfoCamere for internal use and divides the various fields of the taxonomy based on the reference section such as the introduction to the supplementary notes or intangi-

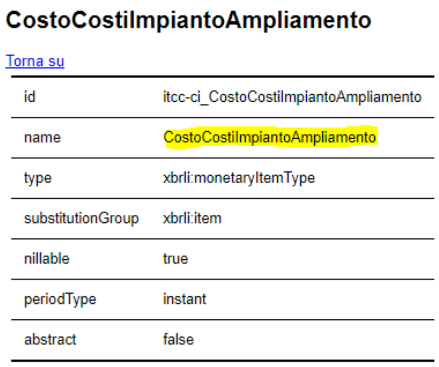


Movimenti delle immobilizzazioni immateriali

► (c_this)

	Costi di impianto e di ampliamento	Costi di sviluppo	Diritti di brevetto industriale e diritti di utilizzazione delle opere dell'ingegno
Valore di inizio esercizio			
Costo	► (c_prev)	► (c_prev)	► (c_prev)
Rivalutazioni	► (c_prev)	► (c_prev)	► (c_prev)
Ammortamenti (Fondo ammortamento)	► (c_prev)	► (c_prev)	► (c_prev)
Svalutazioni	► (c_prev)	► (c_prev)	► (c_prev)
Valore di bilancio	► (c_prev)	► (c_prev)	► (c_prev)
Variazioni nell'esercizio			
Incrementi per acquisizioni	► (c_this)	► (c_this)	► (c_this)
Riclassifiche (del valore di bilancio)	► (c_this)	► (c_this)	► (c_this)
Decrementi per alienazioni e dismissioni (del valore di bilancio)	► (c_this)	► (c_this)	► (c_this)
Rivalutazioni effettuate nell'esercizio	► (c_this)	► (c_this)	► (c_this)
Ammortamento dell'esercizio	► (c_this)	► (c_this)	► (c_this)
Svalutazioni effettuate nell'esercizio	► (c_this)	► (c_this)	► (c_this)
Altre variazioni	► (c_this)	► (c_this)	► (c_this)
Totale variazioni	► (c_this)	► (c_this)	► (c_this)
Valore di fine esercizio			
Costo	► (c_this)	► (c_this)	► (c_this)

Figure 3.4: Section from "bilanci.html"; a randomly selected tag is highlighted which, if clicked, refers to "taxo.html" for more details.



CostoCostiImpiantoAmpliamento

Torna su

id	itcc-ci_CostoCostiImpiantoAmpliamento
name	CostoCostiImpiantoAmpliamento
type	xbrli:monetaryItemType
substitutionGroup	xbrli:item
nillable	true
periodType	instant
abstract	false

Figure 3.5: Details of tag from "taxo.html"; in particular, the name of the tag used to execute search queries on ElasticSearch is highlighted.

ble assets; the second, "taxo.html", contains the details of the taxonomy fields obtained from the XBRL website, stable since 2018.

The initial navigation adopted for retrieving the fields followed these steps:

1. given a topic of interest such as movements of intangible assets, it was searched among the various sections of the "bilanci.html" file;

2. once the section was identified, a specific tag, such as the cost for expanding facilities at the beginning of the exercise (Figure 3.4), was clicked;

3. then the user was redirected to the "taxo.html" file which automatically presents the details related to the selected tag, particularly the name (Figure 3.5) necessary for subsequently executing a query on Kibana or via a Python script connected to ElasticSearch;

4. the notes "c_this" and "c_prev" do not make a difference in obtaining the tags, but it is necessary to be aware of their difference in meaning when reading the results of queries or when wanting to obtain structured data for the section of interest. Specifically, the value "c_this" refers to the year of deposit of the current financial statement while "c_prev" refers to the previous year.

This procedure may initially be convenient for searching for a single tag within a specific section but it is not feasible for studying the 1,418 tags of the supplementary notes across the hundreds of thousands of available documents. Chapter 4 illustrates the process that was applied to facilitate valuable analysis.

4
Supplementary notes: an exploratory analysis

The primary scope of the internship carried out at InfoCamere is the exploration of the data from the supplementary notes stored on ElasticSearch with the purpose of defining textual analysis methods for efficient data extraction, thus enabling the formulation of questions related to the textual sections of the supplementary notes.

This chapter presents the results of the initial analyses conducted for the available supplementary notes, the encountered challenges and how they were addressed.

4.1 Planning for Information Retrieval

The data uploaded to ElasticSearch consists of 671,665 documents for the year 2021, each containing tags from the XBRL taxonomy stable since 2018. These are the documents uploaded, they do not represent all of those filed by the companies. The information present is related to company demographic data, balance sheet, revenue account, cash flow statement in various forms depending on the size of the company as specified in Chapter 2, and finally the supplementary note.

However, the taxonomy tags do not correspond to separate fields callable via a query. In fact, the 21 fields available for search operations pertain to the document identifier (“_id”), the province of the company's Chamber of Commerce reference ("ix_cciaa_rea"), the company's tax code ("ix_cod_fisc"), and others. In particular, the "ix_xbrl_all" field contains all the content of the company's financial statement and not just the supplementary note tags. Furthermore, the field is a flattened string type field and this particularity has proven to be the major obstacle to tag analysis and text retrieval through ElasticSearch. This challenging structure is due to the fact that the index originally available only has the function of storage and it was not designed for effective mass use, for which it would be necessary to create a dedicated index.

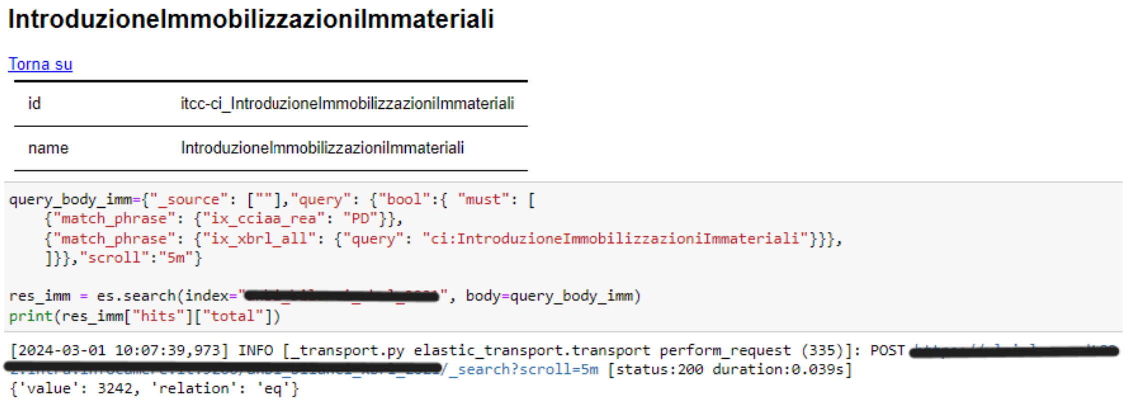
In Algorithm 9.1 in the Appendix 9 a partial example of a document from ElasticSearch is provided, where various fields related to the general index, document identifier, deadline for deposition, reference Chamber of Commerce, company's tax code, and the text of the financial statement in XML format can be observed. In this field, after a preamble not subject to study, the declarations of various tags from the taxonomy begin, such as "TotaleImmobilizzazioniImmateriali" with a value of 25445 or the "CommentoNotaIntegrativa" where the text needs to be cleaned up to obtain a coherent string "Il Bilancio è vero e reale e corrisponde alle scritture contabili. PADOVA 30/06/2023."

To explore how much a topic of interest is valorised one must first obtain the name of the tag from the XBRL taxonomy and then execute a query that returns the documents containing that particular tag in the "ix_xbrl_all" field. This way the number of documents ("hits") containing that tag and which ones are obtained. Figure 4.1 illustrates an example of this procedure to retrieve the occurrence count of the introduction to intangible assets for companies related to the Chamber of Commerce of Padova. The upper half of the figure shows an extract of the taxonomy from the file "taxo.html", while the lower half demonstrates the execution of the query.

It's worth noting that to obtain correct results the prefix "ci:" must be added to the tag name. This requirement in query composition is counterintuitive and took some time to discover, but without it the results are invalid. In fact, in Figure 4.2 it can be observed that two queries executed with the same conditions but one without the "ci:" prefix yield completely different results. If the first query returns a certain number of documents, the second one returns none.

This particular potentially harmful behavior emphasizes the Data Scientist's task of fully understanding how data is interpreted by the tools used and their characteristics in order to effectively exploit them.

To more easily filter the content of documents a Python script has been developed to retrieve from the available file the tags related to supplementary notes from the taxonomy file "taxo.html", thus excluding the balance sheet, revenue account and cash flow statement. In



IntroduzioneImmobilizzazioniImmateriali

Torna su

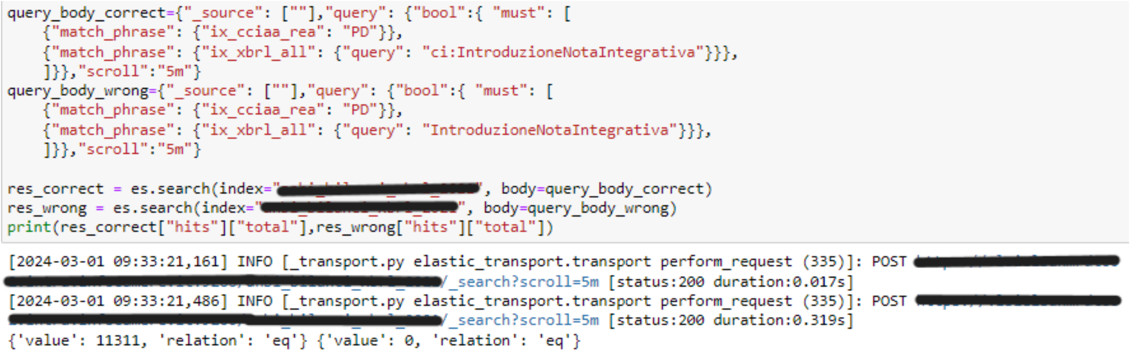
id	itcc-ci_IntroduzioneImmobilizzazioniImmateriali
name	IntroduzioneImmobilizzazioniImmateriali

```
query_body_imm={"_source": [""],"query": {"bool":{ "must": [
    {"match_phrase": {"ix_cciaa_rea": "PD"}},
    {"match_phrase": {"ix_xbrl_all": {"query": "ci:IntroduzioneImmobilizzazioniImmateriali"}}},
    ]}},"scroll":"5m"}

res_imm = es.search(index="[redacted]", body=query_body_imm)
print(res_imm["hits"]["total"])
```

```
[2024-03-01 10:07:39,973] INFO [_transport.py elastic_transport.transport perform_request (335)]: POST [redacted]
[redacted]/_search?scroll=5m [status:200 duration:0.039s]
{'value': 3242, 'relation': 'eq'}
```

Figure 4.1: Example of retrieval of tag and execution of query; references to the InfoCamere general index, which contains the documents, are obscured for privacy and security reasons.

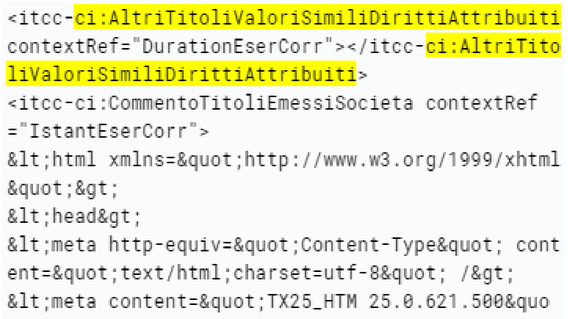


```
query_body_correct={"_source": [""],"query": {"bool":{ "must": [
    {"match_phrase": {"ix_cciaa_rea": "PD"}},
    {"match_phrase": {"ix_xbrl_all": {"query": "ci:IntroduzioneNotaIntegrativa"}}},
    ]}},"scroll":"5m"}
query_body_wrong={"_source": [""],"query": {"bool":{ "must": [
    {"match_phrase": {"ix_cciaa_rea": "PD"}},
    {"match_phrase": {"ix_xbrl_all": {"query": "IntroduzioneNotaIntegrativa"}}},
    ]}},"scroll":"5m"}

res_correct = es.search(index="[redacted]", body=query_body_correct)
res_wrong = es.search(index="[redacted]", body=query_body_wrong)
print(res_correct["hits"]["total"],res_wrong["hits"]["total"])
```

```
[2024-03-01 09:33:21,161] INFO [_transport.py elastic_transport.transport perform_request (335)]: POST [redacted]
[redacted]/_search?scroll=5m [status:200 duration:0.017s]
[2024-03-01 09:33:21,486] INFO [_transport.py elastic_transport.transport perform_request (335)]: POST [redacted]
[redacted]/_search?scroll=5m [status:200 duration:0.319s]
{'value': 11311, 'relation': 'eq'} {'value': 0, 'relation': 'eq'}
```

Figure 4.2: Example of misleading query results; references to the InfoCamere general index, which contains the documents, are obscured for privacy and security reasons.



```
<itcc-ci:AltriTitoliValoriSimiliDirittiAttribuiti
contextRef="DurationEserCorr"></itcc-ci:AltriTito
liValoriSimiliDirittiAttribuiti>
<itcc-ci:CommentoTitoliEmessiSocieta contextRef
="IstantEserCorr">
&lt;html xmlns=&quot;http://www.w3.org/1999/xhtml
&quot;&gt;
&lt;head&gt;
&lt;meta http-equiv=&quot;Content-Type&quot; cont
ent=&quot;text/html;charset=utf-8&quot; /&gt;
&lt;meta content=&quot;TX25_HTM 25.0.621.500&quo
```

Figure 4.3: Example of a tag with no value; occurrences of the tag name with the prefix "ci:" are highlighted to indicate the beginning and end of the tag.

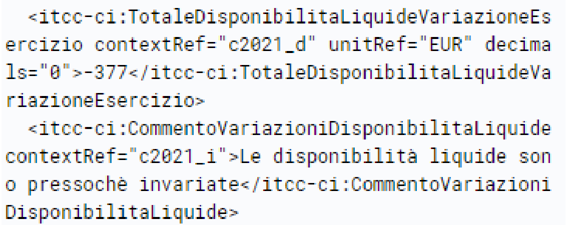
fact, the candidate's task was to focus on the supplementary notes to obtain information that could complement what is already obtainable from other documents.

However, directly executing a query that returns the number of hits for each tag in the taxonomy and therefore the number of documents in which it is present is a procedure that leads to incorrect results. This is because a tag may appear within the "ix_xbrl_all" field of a document but not be populated, having no text or number associated with it. An example of this event is shown in Figure 4.3 where following the context of the tag "AltriTitoliValoriSimiliDiritti-Attribuiti" no number or string is defined between the ">" and "<" characters, while there is text present after the context of the underlying tag "CommentoTitoliEmessiSocieta". The example provided is taken from one of the documents returned by a query that simply searches for the tag "AltriTitoliValoriSimiliDirittiAttribuiti", demonstrating that a behavior that may seem logically correct can indeed occur but it could be potentially harmful for future processes.

Therefore proceeding without further filtering operations would result in analyses contaminated by an unspecified number of false positives and would provide an incorrect overall picture of tags compilation, leading to invalid future clustering studies, text mining processes, and machine learning processes.

In order to continue exploiting the speed of query searches on ElasticSearch several approaches have been attempted to ensure that for each tag of interest only the number of documents actually populating that particular tag is returned. However, due to the representation of the "ix_xbrl_all" field as a single text string and the fact that the XBRL language in which the data is saved is based on XML this has not been possible.

The first approach attempted involved verifying that there was text between the ">" and "<" characters between a tag and its repetition, which indicates the end of population. However, due to the XBRL representation, ElasticSearch is unable to read these specific characters and



```
  <itcc-ci:TotaleDisponibilitaLiquideVariazioneEs
ercizio contextRef="c2021_d" unitRef="EUR" decima
ls="0">-377</itcc-ci:TotaleDisponibilitaLiquideVa
riazioneEsercizio>
  <itcc-ci:CommentoVariazioniDisponibilitaLiquide
contextRef="c2021_i">Le disponibilità liquide son
o pressochè invariate</itcc-ci:CommentoVariazioni
DisponibilitaLiquide>
```

Figure 4.4: Different parameters following different tags.

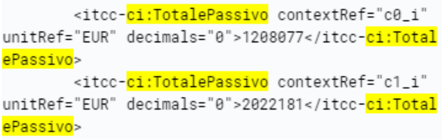
therefore cannot retrieve any useful information for the filtering operation.

Another approach leveraged the distance between words, based on searching for text between a tag and its subsequent repetition. This method also proved unusable due to the non-standard structure of tag definitions, which would make this process valid for some tags and not for others. In fact, as shown in Figure 4.4, following the tag name there are additional parameters such as "contextRef", "unitRef" and "decimals" with a different arrangement depending on the type of value of the tag. Furthermore, depending on the reference tag each of these parameters can have a different value. These factors make the distance between the name of a tag and the start of its value with ">" variable, thus preventing the development of a general procedure that utilizes it to verify the population of tags.

Finally it was decided to download locally the documents from the Chambers of Commerce of the Veneto region into separate files according to the province of reference, saving for each one the content of the "_id" and "ix_xbrl_all" fields. This region was chosen for future data analysis and processing because it offered a good balance between the execution times required to process the information and the quantity of documents available, amounting to 59,154 documents (8.8% of the total files using one of the 20 Italian regions).

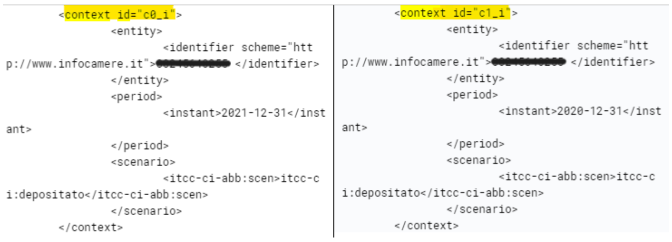
After downloading the data efforts began to define how to structure a procedure for retrieving tags and their values. By comparing the documents with the taxonomy it was realized that some tags can repeat with the same name but different context and values as shown in Figure 4.5. This occurs for numeric tags that report both the measures of the current fiscal year and the previous one, as regularly provided by the taxonomy to allow for comparison.

So the need arose to establish which year each tag refers to based on the "contextRef", in order to properly organize the collected data. The difficulty of this task lies in the fact that the alphanumeric string of the context does not always clearly and unequivocally declare which year it refers to. Additionally, a context defined for example as "c2020_i" is easily understood to refer to the previous year but this is only valid because there is knowledge that the available documents relate to the 2021 deposition. A user unaware of this fact would not be able to deter-



```
        <itcc-ci:TotalePassivo contextRef="c0_i"
unitRef="EUR" decimals="0">1208077</itcc-ci:Total
ePassivo>
        <itcc-ci:TotalePassivo contextRef="c1_i"
unitRef="EUR" decimals="0">2022181</itcc-ci:Total
ePassivo>
```

Figure 4.5: Example of tags with same names but different contexts and values. The occurrences of the tag name with the prefix "ci:" are highlighted to indicate the beginning and end of the two occurrences of the tag; in the middle there are the identifiers of the time contexts, the type of value and the actual associated value.



```
        <context id="c0_i">
                <entity>
                        <identifier scheme="htt
p://www.infocamere.it">[redacted] </identifier>
                </entity>
                <period>
                        <instant>2021-12-31</inst
ant>
                </period>
                <scenario>
                        <itcc-ci-abb:scen>itcc-c
i:depositato</itcc-ci-abb:scen>
                </scenario>
        </context>
```

```
        <context id="c1_i">
                <entity>
                        <identifier scheme="htt
p://www.infocamere.it">[redacted] </identifier>
                </entity>
                <period>
                        <instant>2020-12-31</inst
ant>
                </period>
                <scenario>
                        <itcc-ci-abb:scen>itcc-c
i:depositato</itcc-ci-abb:scen>
                </scenario>
        </context>
```

Figure 4.6: Example of definition of two contexts; the names of the two contexts are highlighted, followed by details such as the reference period, while references to the InfoCamere general index, which contains the documents, are obscured for privacy and security reasons.

mine with certainty whether the context in question is contemporaneous with the deposition of the document or previous to it. Moreover, it is important that the procedure for correctly retrieving the tags and their values be applicable to any document regardless of the reference year.

It has been established that before proceeding with data extraction one must consult the legend of contexts declared at the beginning of each document, just before the definition of the personal data tags. As seen in Figure 4.6 each context is followed by a complete date including the year; based on this, it can be inferred for each context whether it refers to the current year 2021 ("_this"), or the previous year 2020 ("_prev"). In the future to generalize the procedure for application to a document from any year one will have to consult the "ix_dt_curr_dce" field (visible at the beginning in 9.1) to define the year of the balance deposition and thus correctly establish the references of the contexts.

The procedure outlined so far has been tested to retrieve for each document a dictionary with keys and values structured as follows: {"id": document identifier from the "_id" field, "cciaa": Chamber of Commerce from "ix_cciaa_rea", tag_name_1_this: value of tag_1 for the year 2021, tag_name_1_prev: value of tag_1 for the year 2020, tag_name_2_this: value of tag_2 for the year 2021, tag_name_3_this: value of tag_3 for the year 2021, ...}. For readers unfamiliar with this data structure, each key is unique and can be associated with a variable of any type such as a string, a number, a list, etc.

Upon reviewing the results it was noticed that the values of some tags were being omitted. This is primarily due to the fact that certain sections may allow the repetition of certain tags with the same context if they refer, for example, to different geographical areas of belonging, or if they are distinguishable by other descriptions. Additionally, some sections may repeat in case of compilation errors by the compilers of the note. Since it is desired to save the tags as a dictionary where the key is unique, logically if overlaps occur only the value of the last identically named tag is preserved. Therefore to avoid losing any information and continue operating with a dictionary it was decided to add a number to each key to make it truly unique. When updating the dictionary with a new tag it is checked if an identical tag has already been retrieved; in that case, this key is copied and the number after the context is increased by one. In this way, not only are all the data correctly retrieved but the tags belonging to the same repetition of the section receive the same number, so that they can then be grouped together while preserving their initial order.

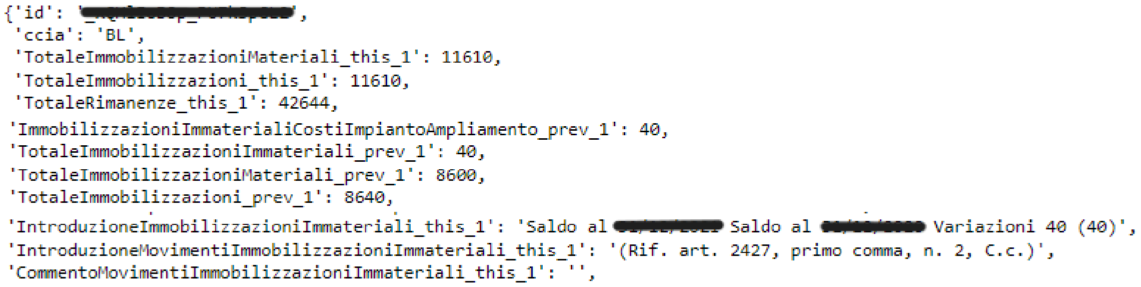
The procedure for retrieving all the tags and their values for a document is outlined as follows:

1. obtain the lists of contexts;
2. iteratively retrieve the names of the tags present in the document if they belong to the explanatory note;
3. create a new key using the tag name;
4. append either "_this" or "_prev" to the key as a suffix, depending on the context's membership in the lists;
5. add an incrementing numerical value if there are already other identically named keys with the same suffix;
6. associate the corresponding value of the tag under examination with the key.

An executable Python script 9.2 viewable in the Appendix 9 was developed to extract for each document of a province a dictionary of the tags found within the "ix_xbrl_all" field. This was done based on the anticipated future need to obtain structured data in tabular form for various sections of interest within the explanatory notes.

Through the developed Python script 9.2 it is possible to:

1. retrieve all the tags of the explanatory note present in a document;
2. associate each tag with its correct context and value;
3. avoid losing any tags due to potential repetitions;
4. obtain for each document a dictionary suitable for further analysis and processing.



```
{'id': [redacted]',
 'ccia': 'BL',
 'TotaleImmobilizzazioniMateriali_this_1': 11610,
 'TotaleImmobilizzazioni_this_1': 11610,
 'TotaleRimanenze_this_1': 42644,
'ImmobilizzazioniImmaterialiCostiImpiantoAmpliamento_prev_1': 40,
'TotaleImmobilizzazioniImmateriali_prev_1': 40,
'TotaleImmobilizzazioniMateriali_prev_1': 8600,
'TotaleImmobilizzazioni_prev_1': 8640,
'IntroduzioneImmobilizzazioniImmateriali_this_1': 'Saldo al [redacted] Saldo al [redacted] Variazioni 40 (40)',
'IntroduzioneMovimentiImmobilizzazioniImmateriali_this_1': '(Rif. art. 2427, primo comma, n. 2, C.c.)',
'CommentoMovimentiImmobilizzazioniImmateriali_this_1': '',
```

Figure 4.7: Example of final dictionary; details that could lead to the company that filed the sample document are obscured for privacy reasons.

In Figure 4.7, a partial example of a dictionary obtained for a document is provided. It includes the tags "TotaleImmobilizzazioniMateriali" and "TotaleImmobilizzazioni" with two different contexts and their corresponding numerical values. Additionally, the text content of the textual tags "IntroduzioneImmoblizzazioniImmateriali" and "IntroduzioneMovimenti-ImmoblizzazioniImmateriali" has been cleaned from XML content, and the tag "Commento-MovimentiImmoblizzazioniImmateriali" has an empty string as its value.

Therefore, compared to the original data present on Elasticsearch with the obtained dictionaries it is easier to perform statistical analysis on numerical tags, text mining operations on textual ones and check how many and which tags have an empty string as a value. These operations can be executed for all sections of the explanatory notes or for specific ones in order to extract both general and targeted information on a particular topic.

It is worth emphasizing that despite transitioning to local data processing and analysis ElasticSearch was still utilized through the Kibana interface to verify that the data was correctly retrieved and to visualize its original organization. Additionally, in Chapter 6 it is illustrated how ElasticSearch is used in a prototype web interface to retrieve a document requested by a user for the purpose of formulating questions about the sections of the explanatory note.

4.2 Exploratory Analysis of companies in Veneto

Once it was confirmed that the retrieved data from the explanatory notes of Veneto were valid and appropriately structured, the extraction of general information commenced.

The first analysis carried out was the exploration of textual and numerical tags. The main objective was to understand how thoroughly the various fields were filled with a value or not and how prevalent each of them was among the various documents. Therefore, a Python procedure was developed to extract from each Veneto dictionary the tags of the explanatory note, check if the assigned value matched the expected type according to the taxonomy and verify if they were effectively populated. Thus, if a textual tag had an empty string as its value it was considered present and with the correct type but not populated. A numerical tag with a value of 0 was considered populated as it cannot be determined whether 0 is the default value or the actual one.

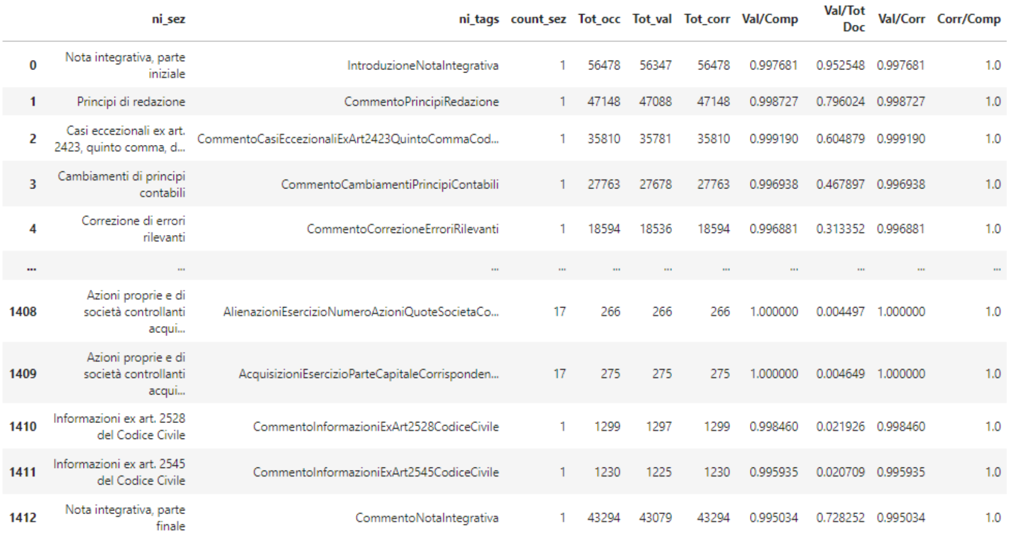
Combining the obtained data with the dataset of tags and their sections the dataset shown in Figure 4.8 was derived. From this dataset the following information can be obtained for each tag defined in the "ni_tags" column:

1. its section of belonging in the explanatory note "ni_sez";
2. how many tags compose each section "count_sez";
3. the total occurrences of the tag across documents "Tot_occ";
4. the total number of documents in which the tag is effectively populated "Tot_val";
5. the total number of documents in which the tag has a value of the corresponding type as defined in the taxonomy "Tot_corr";
6. the ratio between the occurrences when the tag is populated and when it appears "Val/-Comp";
7. the ratio between the occurrences when the tag is populated and the total number of documents "Val/Tot Doc";
8. the ratio between the occurrences when the tag is populated and has the correct type "Val/Corr";
9. the ratio between the occurrences when the tag has a value of the correct type and when it appears "Corr/Comp".

From the dataset statistics on the compilation of tags can be derived both using the obtained data (Figure 4.9) and considering median values by grouping tags based on their section of belonging (Figure 4.10). This approach provides information that is less influenced by extreme values.

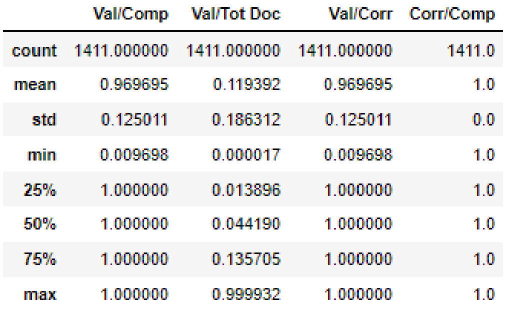
The most valuable results obtained are:

1. when tags appear they are mostly effectively populated and have the correct type;



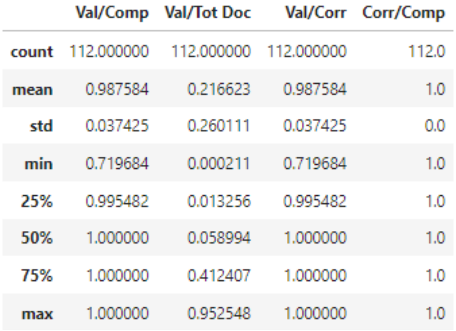
	ni_sez	ni_tags	count_sez	Tot_occ	Tot_val	Tot_corr	Val/Comp	Val/Tot Doc	Val/Corr	Corr/Comp
0	Nota integrativa, parte iniziale	IntroduzioneNotaIntegrativa	1	56478	56347	56478	0.997681	0.952548	0.997681	1.0
1	Principi di redazione	CommentoPrincipiRedazione	1	47148	47088	47148	0.998727	0.796024	0.998727	1.0
2	Casi eccezionali ex art. 2423, quinto comma, d...	CommentoCasiEccezionaliExArt2423QuintoCommaCod...	1	35810	35781	35810	0.999190	0.604879	0.999190	1.0
3	Cambiamenti di principi contabili	CommentoCambiamentiPrincipiContabili	1	27763	27678	27763	0.996938	0.467897	0.996938	1.0
4	Correzione di errori rilevanti	CommentoCorrezioneErroriRilevanti	1	18594	18536	18594	0.996881	0.313352	0.996881	1.0
...	...	...	...	...	...	...	...	...	...	...
1408	Azioni proprie e di società controllanti acqui...	AlienazioniEsercizioNumeroAzioniQuoteSocietaCo...	17	266	266	266	1.000000	0.004497	1.000000	1.0
1409	Azioni proprie e di società controllanti acqui...	AcquisizioniEsercizioParteCapitaleCorrisponden...	17	275	275	275	1.000000	0.004649	1.000000	1.0
1410	Informazioni ex art. 2528 del Codice Civile	CommentoInformazioniExArt2528CodiceCivile	1	1299	1297	1299	0.998460	0.021926	0.998460	1.0
1411	Informazioni ex art. 2545 del Codice Civile	CommentoInformazioniExArt2545CodiceCivile	1	1230	1225	1230	0.995935	0.020709	0.995935	1.0
1412	Nota integrativa, parte finale	CommentoNotaIntegrativa	1	43294	43079	43294	0.995034	0.728252	0.995034	1.0

Figure 4.8: Tags valorization dataset.



	Val/Comp	Val/Tot Doc	Val/Corr	Corr/Comp
count	1411.000000	1411.000000	1411.000000	1411.0
mean	0.969695	0.119392	0.969695	1.0
std	0.125011	0.186312	0.125011	0.0
min	0.009698	0.000017	0.009698	1.0
25%	1.000000	0.013896	1.000000	1.0
50%	1.000000	0.044190	1.000000	1.0
75%	1.000000	0.135705	1.000000	1.0
max	1.000000	0.999932	1.000000	1.0

Figure 4.9: Statistics of compilation of tags.



	Val/Comp	Val/Tot Doc	Val/Corr	Corr/Comp
count	112.000000	112.000000	112.000000	112.0
mean	0.987584	0.216623	0.987584	1.0
std	0.037425	0.260111	0.037425	0.0
min	0.719684	0.000211	0.719684	1.0
25%	0.995482	0.013256	0.995482	1.0
50%	1.000000	0.058994	1.000000	1.0
75%	1.000000	0.412407	1.000000	1.0
max	1.000000	0.952548	1.000000	1.0

Figure 4.10: Statistics of median compilation of tags.

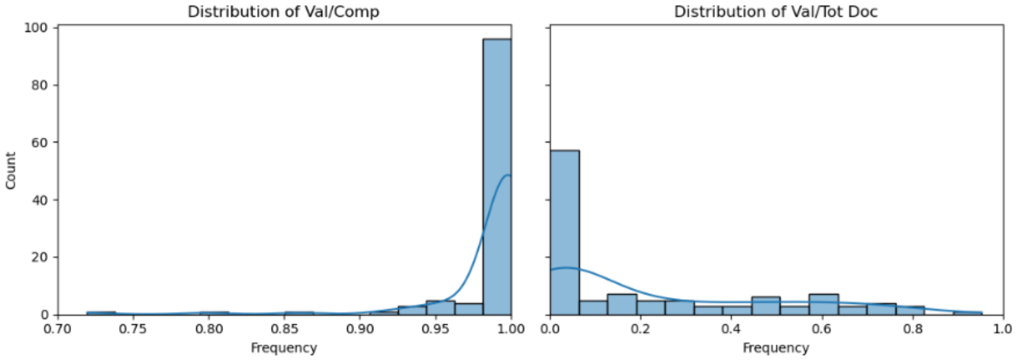


Figure 4.11: Comparison of valorization and appearance of tags; the X axis ("Frequency") shows the percentage of times that a tag is valued out of those that appears (image on the left) or the total number of documents (image on the right), the Y axis ("Count") shows how many tags contain that percentage.

2. despite containing actual information tags are generally not widespread as at least half of them appear in only 5% of the documents.

Figure 4.11 highlights the difference between the statistics of compilation regarding population and the total number of documents, where the graphs represent practically opposite situations. While the first graph emphasizes the effective population of tags, the second graph reaffirms how, despite this, they are generally not widespread.

Therefore attempts were made to obtain more specific data regarding the completion of individual sections and by filtering for the type of data contained in the tags.

Analyzing the individual sections did not yield significant insights regarding particular differences in completion among the various parts of the explanatory note. As observed in Figure 4.12 while the ratio between population and presence of tags remains very high for almost all sections, the ratio between completion and the total number of documents is heavily dependent on the reference division, providing no significant data.

Next, the tags were examined based on the type of corresponding value, whether numerical or textual.

If we consider only the tags of numerical type they account for 1,155 out of the total 1,418 tags, representing 81%. Figure 4.13 highlights how the tags are practically always populated when they appear, but most of them are not popular when considering all documents from Veneto. Specifically, only 19 tags (1.6% of the numerical tags and 1.3% of the total) have a completion percentage higher than 80%. These are highly prevalent items due to the structure of the XBRL taxonomy, such as "TotaleCrediti" and "TotaleImmobilizzazioni".

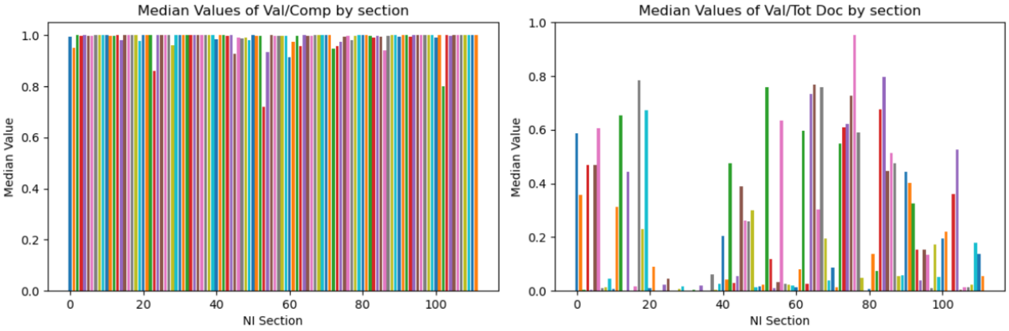


Figure 4.12: Comparison of valorization and appearance of tags by section; the X axis represents the single sections, the Y axis the percentage of valorization for each section using the median value of its tags.

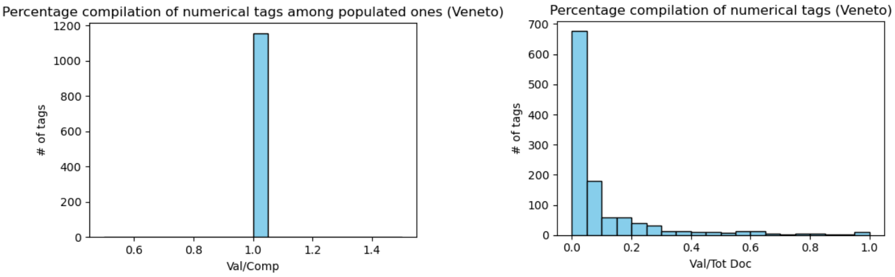


Figure 4.13: Compilation of numerical tags; the X axis shows the percentage of times that a tag is valued out of those that appear ("Val/Comp") or the total number of documents ("Val/Tot Doc"), the Y axis ("# of tags") shows how many tags contain that percentage.

Textual tags amount to 258, representing 18% of the total tags. Out of these 187 (72%) exceed the 80% completion threshold considering only the documents in which they appear, while only 2 (0.7%) do so when considering all documents. These are "IntroduzioneNotaIntegrativa" and "CommentoPropostaDestinazioneUtiliCoperturaPerdite". Figure 4.14 illustrates the obtained data regarding the completion and presence of textual tags.

The graphs in question consider the documents of all the Chambers of Commerce in Veneto, but extremely similar results are also obtainable by evaluating the supplementary notes of only specific Chambers.

4.3 Key points and results of the Exploratory Analysis

Due to how the documents on ElasticSearch were stored the XBRL taxonomy tags are not freely explorable as separate fields. In fact, the entire content of the financial statement is saved

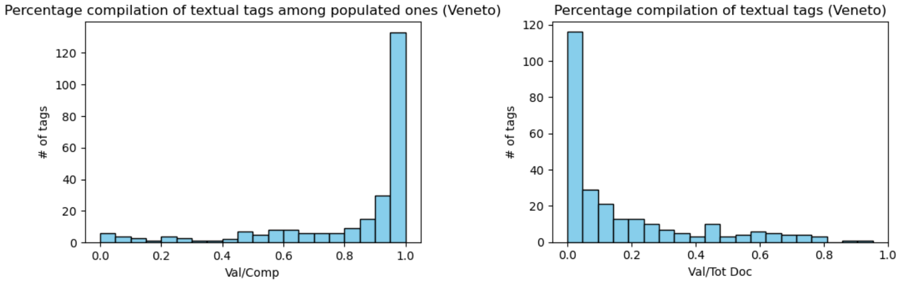


Figure 4.14: Compilation of textual tags; the X axis shows the percentage of times that a tag is valued out of those that appear ("Val/Comp") or the total number of documents ("Val/Tot Doc"), the Y axis ("# of tags") shows how many tags contain that percentage.

as a single text string representing an XML file, making it challenging to retrieve the content of a tag and apply necessary filters, such as ignoring it if it contains an empty string.

Several approaches were attempted to continue leveraging ElasticSearch but ultimately the data for the documents related to the Chambers of Commerce of Veneto was downloaded locally. A procedure was then defined and executed to extract and organize the tags from each document in order to obtain comprehensive and suitable information for subsequent structured data extraction.

Upon examining the completion percentages of the tags it was found that, while they are effectively populated when they appear, they are not prevalent when compared to the available documents.

In summary, the most important results include the development and testing of a valid procedure to extract comprehensive and suitable data for further analysis, as well as obtaining valuable insights. Due to the initial data formatting, only analyses on individual tags were possible and they were corrupted by tags not effectively populated, resulting in incomplete or inaccurate tag overviews. Instead, with the defined method a broader and more precise analysis is now possible, leading to more reliable results.

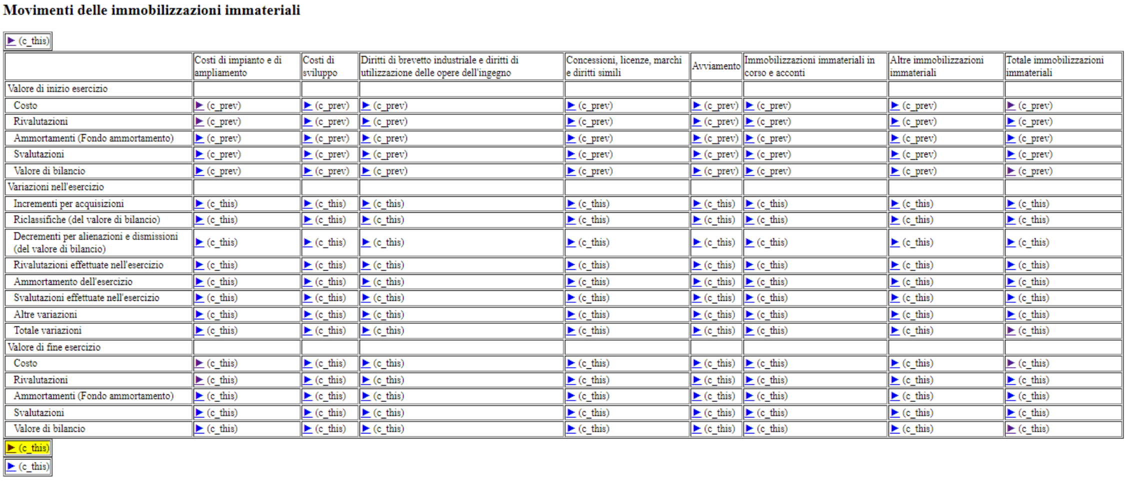
5
Supplementary notes: text mining application

This chapter presents the methodology applied to obtain structured data from the different sections of the explanatory note's taxonomy, starting from the data retrieved as described in Chapter 4. Subsequently, specific sections ("Cooperative societies" 5.2 and "Intangible Assets" 5.3) subject to text mining activities are discussed with the aim of discovering new information not obtainable from other financial statement documents.

5.1 Obtaining structured data

Once the exploratory phase of tag valuation was concluded the next step was to outline how to proceed with text mining activities on textual tags. This capability to manipulate and extract information on a massive scale is particularly relevant considering that the data obtainable in this way cannot be derived from any other financial statement document. Being able to easily integrate the already available information in the balance sheet and revenue account can lead to new insights and interpretation possibilities not previously available.

The sections of interest for these text mining investigations are related to cooperative societies and movements of intangible assets. The involved tags are the introduction to cooperative societies information and the comment on movements of intangible assets, which is closely re-



Movimenti delle immobilizzazioni immateriali

▶ (c_this)

	Costi di impianto e di ampliamento	Costi di sviluppo	Diritti di brevetto industriale e diritti di utilizzazione delle opere dell'ingegno	Concessioni, licenze, marchi e diritti simili	Avviamento	Immobilizzazioni immateriali in corso e acconti	Altre immobilizzazioni immateriali	Totale immobilizzazioni immateriali
Valore di inizio esercizio								
Costo	▶ (c_prev)	▶ (c_prev)	▶ (c_prev)	▶ (c_prev)	▶ (c_prev)	▶ (c_prev)	▶ (c_prev)	▶ (c_prev)
Rivalutazioni	▶ (c_prev)	▶ (c_prev)	▶ (c_prev)	▶ (c_prev)	▶ (c_prev)	▶ (c_prev)	▶ (c_prev)	▶ (c_prev)
Ammortamenti (Fondo ammortamento)	▶ (c_prev)	▶ (c_prev)	▶ (c_prev)	▶ (c_prev)	▶ (c_prev)	▶ (c_prev)	▶ (c_prev)	▶ (c_prev)
Svalutazioni	▶ (c_prev)	▶ (c_prev)	▶ (c_prev)	▶ (c_prev)	▶ (c_prev)	▶ (c_prev)	▶ (c_prev)	▶ (c_prev)
Valore di bilancio	▶ (c_prev)	▶ (c_prev)	▶ (c_prev)	▶ (c_prev)	▶ (c_prev)	▶ (c_prev)	▶ (c_prev)	▶ (c_prev)
Variazioni nell'esercizio								
Incrementi per acquisizioni	▶ (c_this)	▶ (c_this)	▶ (c_this)	▶ (c_this)	▶ (c_this)	▶ (c_this)	▶ (c_this)	▶ (c_this)
Riclassifiche (del valore di bilancio)	▶ (c_this)	▶ (c_this)	▶ (c_this)	▶ (c_this)	▶ (c_this)	▶ (c_this)	▶ (c_this)	▶ (c_this)
Decrementi per alienazioni e dismissioni (del valore di bilancio)	▶ (c_this)	▶ (c_this)	▶ (c_this)	▶ (c_this)	▶ (c_this)	▶ (c_this)	▶ (c_this)	▶ (c_this)
Rivalutazioni effettuate nell'esercizio	▶ (c_this)	▶ (c_this)	▶ (c_this)	▶ (c_this)	▶ (c_this)	▶ (c_this)	▶ (c_this)	▶ (c_this)
Ammortamento dell'esercizio	▶ (c_this)	▶ (c_this)	▶ (c_this)	▶ (c_this)	▶ (c_this)	▶ (c_this)	▶ (c_this)	▶ (c_this)
Svalutazioni effettuate nell'esercizio	▶ (c_this)	▶ (c_this)	▶ (c_this)	▶ (c_this)	▶ (c_this)	▶ (c_this)	▶ (c_this)	▶ (c_this)
Altre variazioni	▶ (c_this)	▶ (c_this)	▶ (c_this)	▶ (c_this)	▶ (c_this)	▶ (c_this)	▶ (c_this)	▶ (c_this)
Totale variazioni	▶ (c_this)	▶ (c_this)	▶ (c_this)	▶ (c_this)	▶ (c_this)	▶ (c_this)	▶ (c_this)	▶ (c_this)
Valore di fine esercizio								
Costo	▶ (c_this)	▶ (c_this)	▶ (c_this)	▶ (c_this)	▶ (c_this)	▶ (c_this)	▶ (c_this)	▶ (c_this)
Rivalutazioni	▶ (c_this)	▶ (c_this)	▶ (c_this)	▶ (c_this)	▶ (c_this)	▶ (c_this)	▶ (c_this)	▶ (c_this)
Ammortamenti (Fondo ammortamento)	▶ (c_this)	▶ (c_this)	▶ (c_this)	▶ (c_this)	▶ (c_this)	▶ (c_this)	▶ (c_this)	▶ (c_this)
Svalutazioni	▶ (c_this)	▶ (c_this)	▶ (c_this)	▶ (c_this)	▶ (c_this)	▶ (c_this)	▶ (c_this)	▶ (c_this)
Valore di bilancio	▶ (c_this)	▶ (c_this)	▶ (c_this)	▶ (c_this)	▶ (c_this)	▶ (c_this)	▶ (c_this)	▶ (c_this)

▶ (c_this)

▶ (c_this)

Figure 5.1: It is highlighted how the single tag of interest, referring to the comments on the movements of fixed assets, is closely linked to all those of the section to which it belongs.

lated to surrounding numerical tags as seen in Figure 5.1. Rather than retrieving the individual comment tag for each document, in line with the already present intention of obtaining structured data, it was decided to develop a system to retrieve for each section a dataframe of all the tags contained within it before initiating text mining investigations. In this way each comment can be linked to the numerical entries of the same document, allowing for the discovery of any correlations between textual content and numerical data.

Therefore a Python procedure was outlined which, given the title of a section of interest, a list of dictionaries from a chamber of commerce and an optional index returns a dataframe structured similarly to that accessible in the "bilancio.html" file, retrieving all the tags of that division of the explanatory note for all the indicated documents. The main difficulties encountered in writing the code were:

1. obtaining a unique row for each index by transforming the row indices into new features to facilitate future processes;
2. managing nested indices in both rows and columns;
3. sections containing tags that can repeat by definition.

The Algorithm 9.3 viewable in the Appendix 9 takes as input the title of a section of the explanatory note from "bilancio.html", an optional document ID and the list of dictionaries the document is part of. If an index is not provided all documents are processed. Then it exploits Algorithms 9.4 and 9.5 to obtain a dataframe with a unique row of all the tag names contained in the section of interest, in order to filter which tags to retrieve from each document, and to return a dataframe organized with the same structure.

With the created algorithms structured data with the same textual references as the taxonomy can be obtained and organized in a way that facilitates subsequent operations. In paragraphs 5.2 and 5.3 the results of the algorithms are shown, which were then used as a starting point for text mining analyses.

5.2 Cooperative Societies

A cooperative company according to Italian law [28] is an organization established to jointly manage a business aimed at providing its members (mutualistic purpose) with goods or services for which the cooperative was created. The truly characteristic and unifying element of any type of cooperative, regardless of the sector and category of work of its members, is the fact that unlike capital companies that aim to create and distribute profits cooperatives instead have a mutualistic purpose, which consists of ensuring to the members work, consumer goods, or services under conditions better than those obtainable in the free market.

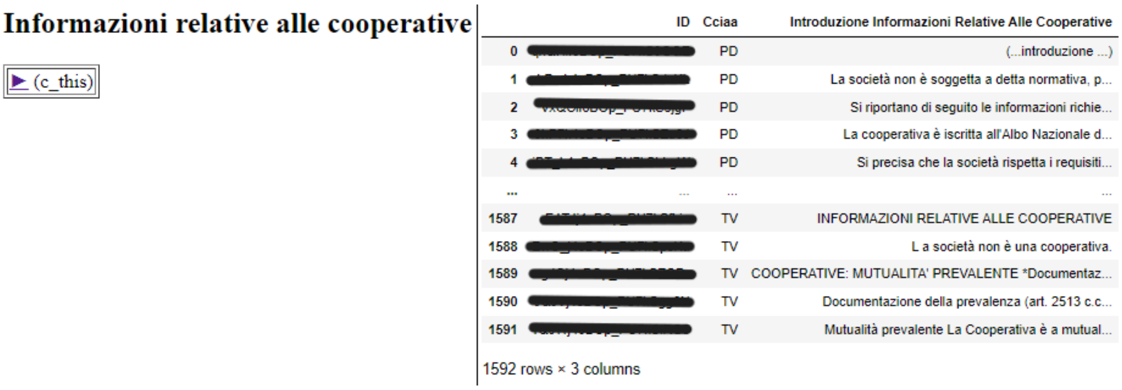
Therefore the focal point of a cooperative is the satisfaction of the needs of its members protecting their interests and pursuing the principles of mutuality, solidarity, and democracy.

A particular type of cooperative is the one with prevalent mutualism such as social cooperatives, which primarily carry out activities for the benefit of the members mainly using their labor, goods, and services.

The Algorithm 9.3 was then used to obtain a dataframe of the section related to information for cooperative societies. Figure 5.2 shows the dataframe obtained by inputting the title of the section "Informazioni relative alle cooperative", no specific index and all the dictionaries of companies in Veneto. The dataset includes all the documents containing the single tag of the section of interest, their indices, and the corresponding province, thus making it possible to perform general analyses on the entire region or for specific provinces or documents.

5.2.1 Cooperative societies tasks

Text mining tasks to be performed on the dataset obtained from cooperative societes were drafted based on real needs that InfoCamere's clients might have. This allowed testing the candidate's work through a realistic simulation of the actual utilization of methods and processed data. The various tasks to be carried out were:



Informazioni relative alle cooperative

	ID	Cciaa	Introduzione Informazioni Relative Alle Cooperative
0	[redacted]	PD	(...introduzione ...)
1	[redacted]	PD	La società non è soggetta a detta normativa, p...
2	[redacted]	PD	Si riportano di seguito le informazioni richie...
3	[redacted]	PD	La cooperativa è iscritta all'Albo Nazionale d...
4	[redacted]	PD	Si precisa che la società rispetta i requisiti...
...	...	...	...
1587	[redacted]	TV	INFORMAZIONI RELATIVE ALLE COOPERATIVE
1588	[redacted]	TV	L a società non è una cooperativa.
1589	[redacted]	TV	COOPERATIVE: MUTUALITA' PREVALENTE *Documentaz...
1590	[redacted]	TV	Documentazione della prevalenza (art. 2513 c.c...
1591	[redacted]	TV	Mutualità prevalente La Cooperativa è a mutual...

1592 rows × 3 columns

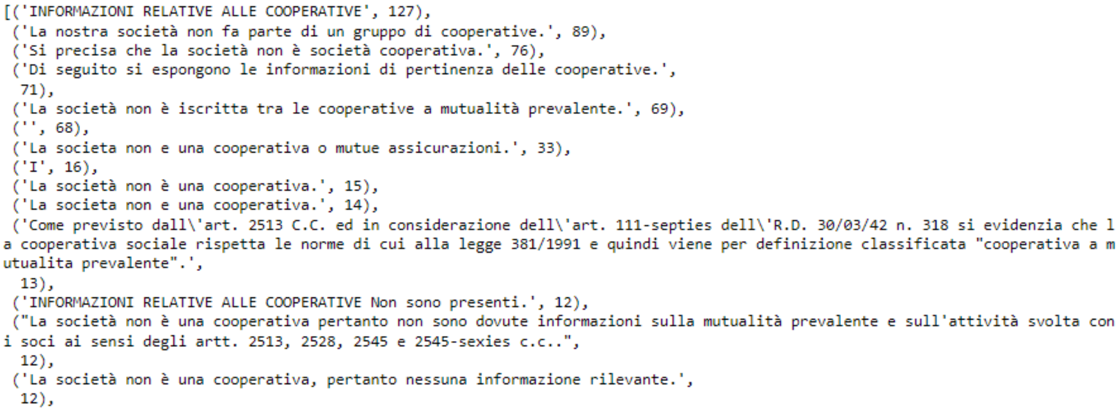
Figure 5.2: Structured data for cooperatives; on the left there is the section with a single tag, on the right the dataframe obtained with the identifiers of company documents obscured for privacy reasons, the relevant Chamber of Commerce and the textual content of the tag.

1. checking which texts use the word "mutualistico", "mutualità", or similar, with the aim of focusing on companies that are mutualistic cooperatives;
2. checking how many companies declare to meet Article 2513 of the Civil Code regarding the condition of prevalence and how many meet Article 2514 regarding the condition of mutuality;
3. checking how many companies mention having undergone an annual cooperative review and extract the authors;
4. checking if registration with a register is mentioned;
5. checking if the social composition is mentioned.

Most of these tasks are similar, varying only by the items to search for. Therefore, it was decided to adopt a common strategy based on searching for task-dependent keywords in a text.

To define an optimal method sample texts were examined to outline the general aspects and other characteristics. By inspecting the most common texts it was observed that they are short and usually their message, explicit or implicit, is that the respective company is not a cooperative. In Figure 5.3 it can be noted that companies, if they are not a cooperative society, affirm this fact with very similar and brief texts or leave a default text unchanged. Only one of the reported cases positively attests to the nature of the company. Therefore the length of this particular text was defined as the minimum number of characters for a content to be considered for a more in-depth verification.

To manipulate and extract valuable information from a text where words naturally deviate from their base lemma, Spacy, an open-source Python library for Natural Language Processing



```
[('INFORMAZIONI RELATIVE ALLE COOPERATIVE', 127),
 ('La nostra società non fa parte di un gruppo di cooperative.', 89),
 ('Si precisa che la società non è società cooperativa.', 76),
 ('Di seguito si espongono le informazioni di pertinenza delle cooperative.',
  71),
 ('La società non è iscritta tra le cooperative a mutualità prevalente.', 69),
 ('', 68),
 ('La societa non e una cooperativa o mutue assicurazioni.', 33),
 ('I', 16),
 ('La società non è una cooperativa.', 15),
 ('La societa non e una cooperativa.', 14),
 ('Come previsto dall\'art. 2513 C.C. ed in considerazione dell\'art. 111-septies dell\'R.D. 30/03/42 n. 318 si evidenzia che l
a cooperativa sociale rispetta le norme di cui alla legge 381/1991 e quindi viene per definizione classificata "cooperativa a m
utualita prevalente".',
  13),
 ('INFORMAZIONI RELATIVE ALLE COOPERATIVE Non sono presenti.', 12),
 ("La società non è una cooperativa pertanto non sono dovute informazioni sulla mutualità prevalente e sull'attività svolta con
i soci ai sensi degli artt. 2513, 2528, 2545 e 2545-sexies c.c..",
  12),
 ('La società non è una cooperativa, pertanto nessuna informazione rilevante.',
  12),
```

Figure 5.3: The most common texts for Cooperative; each text is associated with the number of times it appears in the documents under examination.

(NLP) [29], was leveraged. Designed to be fast, efficient, and usable for various languages including Italian, it is used to perform various natural language processing tasks such as tokenization [30], lemmatization [31], part-of-speech tagging (POS) [32], named entity recognition (NER) [33] and more. For stemming, the Python library SnowBallStemmer [34] was used to reduce a word to its root through a different process from lemmatization.

The following method was then defined to answer the question of whether certain keywords are present in a text:

1. if the text is empty or shorter than the minimum length required for verification, return False;

2. replace any relevant abbreviations;

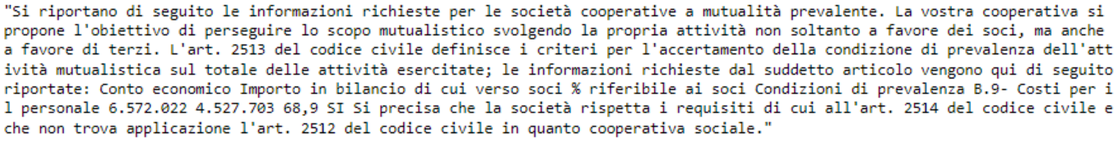
3. perform tokenization, stemming, or lemmatization on the keywords and the text to obtain substrings;

4. search for occurrences of the keywords in the substrings;

5. for each substring, based on the presence and position of the keywords and any negations, return True or False;

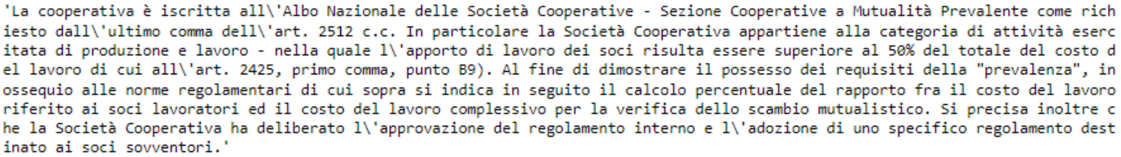
6. return the mode of the True/False values.

Algorithm 9.6 viewable in the Appendix 9 presents the code for the defined method. Invoking this function for each text in the dataset of cooperatives yields a subset of documents containing any of the keywords relevant to the task.



```
"Si riportano di seguito le informazioni richieste per le società cooperative a mutualità prevalente. La vostra cooperativa si
propone l'obiettivo di perseguire lo scopo mutualistico svolgendo la propria attività non soltanto a favore dei soci, ma anche
a favore di terzi. L'art. 2513 del codice civile definisce i criteri per l'accertamento della condizione di prevalenza dell'att
ività mutualistica sul totale delle attività esercitate; le informazioni richieste dal suddetto articolo vengono qui di seguito
riportate: Conto economico Importo in bilancio di cui verso soci % riferibile ai soci Condizioni di prevalenza B.9- Costi per i
l personale 6.572.022 4.527.703 68,9 SI Si precisa che la società rispetta i requisiti di cui all'art. 2514 del codice civile e
che non trova applicazione l'art. 2512 del codice civile in quanto cooperativa sociale."
```

Figure 5.4: Example of a document from a company that the Algorithm 9.6 flags as declaring it has a "scopo mutualistico" and respects the articles "2513","2514"



```
'La cooperativa è iscritta all\'Albo Nazionale delle Società Cooperative - Sezione Cooperative a Mutualità Prevalente come rich
iesto dall\'ultimo comma dell\'art. 2512 c.c. In particolare la Società Cooperativa appartiene alla categoria di attività eserc
itata di produzione e lavoro - nella quale l\'apporto di lavoro dei soci risulta essere superiore al 50% del totale del costo d
el lavoro di cui all\'art. 2425, primo comma, punto B9). Al fine di dimostrare il possesso dei requisiti della "prevalenza", in
ossequio alle norme regolamentari di cui sopra si indica in seguito il calcolo percentuale del rapporto fra il costo del lavoro
riferito ai soci lavoratori ed il costo del lavoro complessivo per la verifica dello scambio mutualistico. Si precisa inoltre c
he la Società Cooperativa ha deliberato l\'approvazione del regolamento interno e l\'adozione di uno specifico regolamento dest
inato ai soci sovventori.'
```

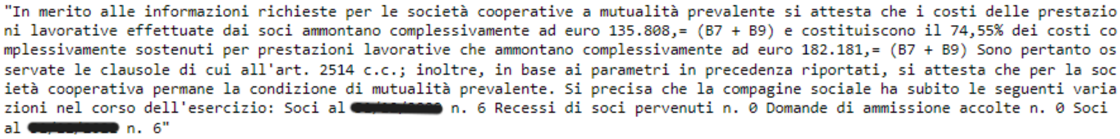
Figure 5.5: Example of a document from a company that the Algorithm 9.6 flags as declaring it is registered with a "albo".

The Algorithm 9.6 was tested in three different combinations using stemming, lemmatization or a combination of both. The results show minimal differences in the outcomes. The texts identified as containing the searched keywords were practically the same across all three methods. Further details on the general results are available in Chapter 6.

Figure 5.4 represents one of the texts that respond positively to the question of whether the company has a mutualistic purpose or complies with articles 2513 and 2514 on mutualism and prevalence conditions. It can be observed that the company indeed declares to pursue the mutualistic purpose, citing articles 2513 and 2514 and stating compliance with them. Figure 5.5 represents one of the texts that respond positively to the question of whether the company is registered in a register, while figure 5.6 indicates whether a social composition is mentioned.

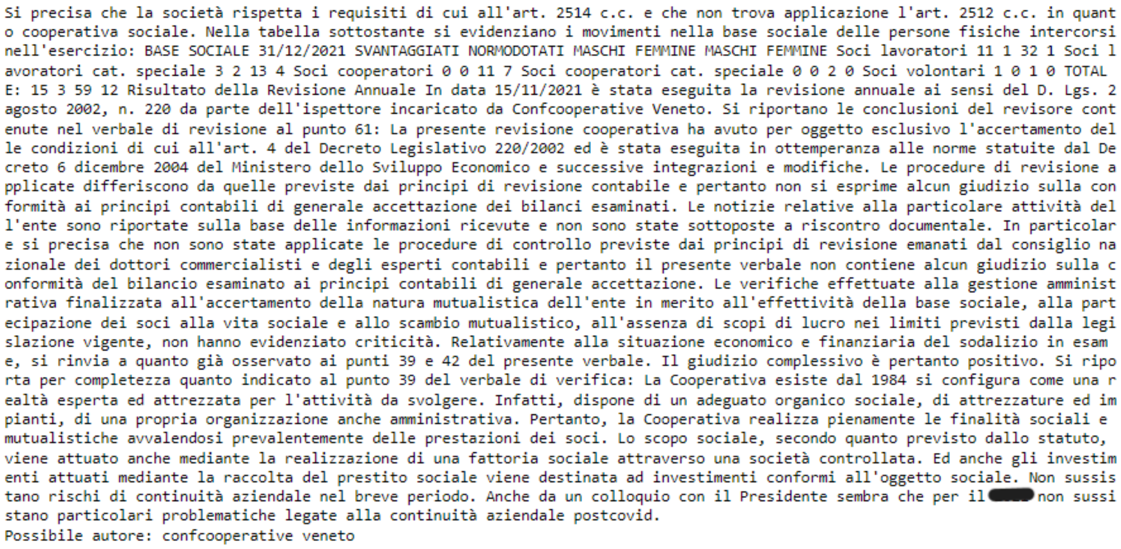
The last text mining task involved was verifying if an annual review is mentioned and, if so, extracting the authors. For this purpose the Named Entity Recognition (NER) and Part-of-Speech (POS) Tagging functionalities of Spacy were utilized.

NER identifies and assigns labels to each named entity in a text such as person, organization, location, date, money, and others. Similarly, POS assigns a grammatical category to each word in a text such as verb, noun, adjective, pronoun, etc.



```
"In merito alle informazioni richieste per le società cooperative a mutualità prevalente si attesta che i costi delle prestazio
ni lavorative effettuate dai soci ammontano complessivamente ad euro 135.808,= (B7 + B9) e costituiscono il 74,55% dei costi co
mplessivamente sostenuti per prestazioni lavorative che ammontano complessivamente ad euro 182.181,= (B7 + B9) Sono pertanto os
servate le clausole di cui all'art. 2514 c.c.; inoltre, in base ai parametri in precedenza riportati, si attesta che per la soc
ietà cooperativa permane la condizione di mutualità prevalente. Si precisa che la compagine sociale ha subito le seguenti varia
zioni nel corso dell'esercizio: Soci al [redacted] n. 6 Recessi di soci pervenuti n. 0 Domande di ammissione accolte n. 0 Soci
al [redacted] n. 6"
```

Figure 5.6: Example of a document from a company that the Algorithm 9.6 flags as declaring it belongs to a "compagine"; details that could lead to the company that owns the document are obscured.



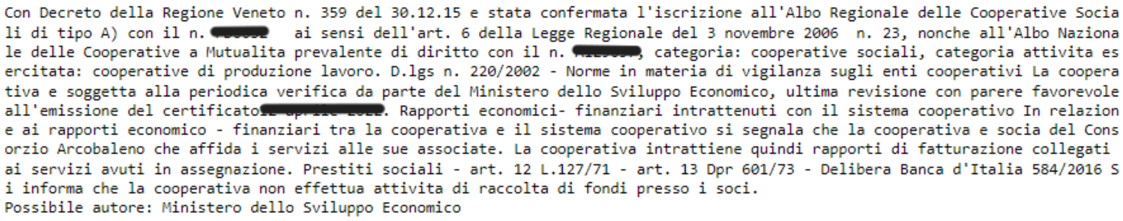
```
Si precisa che la società rispetta i requisiti di cui all'art. 2514 c.c. e che non trova applicazione l'art. 2512 c.c. in quant
o cooperativa sociale. Nella tabella sottostante si evidenziano i movimenti nella base sociale delle persone fisiche intercorsi
nell'esercizio: BASE SOCIALE 31/12/2021 SVANTAGGIATI NORMODOTATI MASCHI FEMMINE MASCHI FEMMINE Soci lavoratori 11 1 32 1 Soci l
avoratori cat. speciale 3 2 13 4 Soci cooperatori 0 0 11 7 Soci cooperatori cat. speciale 0 0 2 0 Soci volontari 1 0 1 0 TOTAL
E: 15 3 59 12 Risultato della Revisione Annuale In data 15/11/2021 è stata eseguita la revisione annuale ai sensi del D. Lgs. 2
agosto 2002, n. 220 da parte dell'ispettore incaricato da Confcooperative Veneto. Si riportano le conclusioni del revisore cont
enute nel verbale di revisione al punto 61: La presente revisione cooperativa ha avuto per oggetto esclusivo l'accertamento del
le condizioni di cui all'art. 4 del Decreto Legislativo 220/2002 ed è stata eseguita in ottemperanza alle norme statuite dal De
creto 6 dicembre 2004 del Ministero dello Sviluppo Economico e successive integrazioni e modifiche. Le procedure di revisione a
pplicate differiscono da quelle previste dai principi di revisione contabile e pertanto non si esprime alcun giudizio sulla con
formità ai principi contabili di generale accettazione dei bilanci esaminati. Le notizie relative alla particolare attività del
l'ente sono riportate sulla base delle informazioni ricevute e non sono state sottoposte a riscontro documentale. In particolar
e si precisa che non sono state applicate le procedure di controllo previste dai principi di revisione emanati dal consiglio na
zionale dei dottori commercialisti e degli esperti contabili e pertanto il presente verbale non contiene alcun giudizio sulla c
onformità del bilancio esaminato ai principi contabili di generale accettazione. Le verifiche effettuate alla gestione amminist
rativa finalizzata all'accertamento della natura mutualistica dell'ente in merito all'effettività della base sociale, alla part
ecipazione dei soci alla vita sociale e allo scambio mutualistico, all'assenza di scopi di lucro nei limiti previsti dalla legi
slazione vigente, non hanno evidenziato criticità. Relativamente alla situazione economico e finanziaria del sodalizio in esam
e, si rinvia a quanto già osservato ai punti 39 e 42 del presente verbale. Il giudizio complessivo è pertanto positivo. Si ripo
rta per completezza quanto indicato al punto 39 del verbale di verifica: La Cooperativa esiste dal 1984 si configura come una r
ealtà esperta ed attrezzata per l'attività da svolgere. Infatti, dispone di un adeguato organico sociale, di attrezzature ed im
pianti, di una propria organizzazione anche amministrativa. Pertanto, la Cooperativa realizza pienamente le finalità sociali e
mutualistiche avvalendosi prevalentemente delle prestazioni dei soci. Lo scopo sociale, secondo quanto previsto dallo statuto,
viene attuato anche mediante la realizzazione di una fattoria sociale attraverso una società controllata. Ed anche gli investim
enti attuati mediante la raccolta del prestito sociale viene destinata ad investimenti conformi all'oggetto sociale. Non sussis
tano rischi di continuità aziendale nel breve periodo. Anche da un colloquio con il Presidente sembra che per il[redacted]non sussi
stano particolari problematiche legate alla continuità aziendale postcovid.
Possibile autore: confcooperative veneto
```

Figure 5.7: First example of a text and the author of the revision identified by the Algorithm 9.8; details that could lead to the company that owns the document are obscured.

The underlying idea of the developed algorithm is to use both these functionalities to extract the name of the most probable author of the annual review if mentioned in the text. First, Algorithm 9.7, viewable in the Appendix 9, was executed with all the documents as input, intending to create three lists of entities: two obtained through POS and NER containing possible author names for documents where the review is mentioned, and one obtained through NER with entity names for documents where the review is not mentioned. The idea is that if a name of a possible author is extracted from a text it will be relevant if it is present in the first two lists and less relevant if it is in the third list.

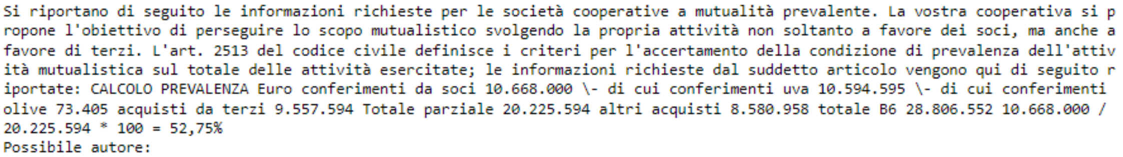
In Algorithm 9.8, viewable in the Appendix 9, from each of these lists the number of occurrences for each entity is obtained and the search is then rerun using NER. Then, a score is associated with each word obtained, corresponding to the sum of its occurrences in the first two lists minus the value of the third list: this gives more value to entities that actually appear when a review is mentioned and decreases the value of words present when the review is not mentioned. Finally, the entity with the highest score is returned as the probable name of the review author.

Here are some examples of the output of Algorithm 9.8. Figures 5.7 and 5.8 show how it successfully extracted the name of the actual review author from the analyzed texts, while in figure 5.9 for a text without the keyword nothing was returned. More details on the overall results are available in Chapter 8.



```
Con Decreto della Regione Veneto n. 359 del 30.12.15 e stata confermata l'iscrizione all'Albo Regionale delle Cooperative Socia
li di tipo A) con il n. [redacted]  ai sensi dell'art. 6 della Legge Regionale del 3 novembre 2006  n. 23, nonche all'Albo Naziona
le delle Cooperative a Mutualita prevalente di diritto con il n. [redacted], categoria: cooperative sociali, categoria attivita es
ercitata: cooperative di produzione lavoro. D.lgs n. 220/2002 - Norme in materia di vigilanza sugli enti cooperativi La coopera
tiva e soggetta alla periodica verifica da parte del Ministero dello Sviluppo Economico, ultima revisione con parere favorevole
all'emissione del certificato[redacted]. Rapporti economici- finanziari intrattenuti con il sistema cooperativo In relazion
e ai rapporti economico - finanziari tra la cooperativa e il sistema cooperativo si segnala che la cooperativa e socia del Cons
orzio Arcobaleno che affida i servizi alle sue associate. La cooperativa intrattiene quindi rapporti di fatturazione collegati
ai servizi avuti in assegnazione. Prestiti sociali - art. 12 L.127/71 - art. 13 Dpr 601/73 - Delibera Banca d'Italia 584/2016 S
i informa che la cooperativa non effettua attivita di raccolta di fondi presso i soci.
Possibile autore: Ministero dello Sviluppo Economico
```

Figure 5.8: Second example of a text and the author of the revision identified by the Algorithm 9.8; details that could lead to the company that owns the document are obscured.



```
Si riportano di seguito le informazioni richieste per le società cooperative a mutualità prevalente. La vostra cooperativa si p
ropone l'obiettivo di perseguire lo scopo mutualistico svolgendo la propria attività non soltanto a favore dei soci, ma anche a
favore di terzi. L'art. 2513 del codice civile definisce i criteri per l'accertamento della condizione di prevalenza dell'attiv
ità mutualistica sul totale delle attività esercitate; le informazioni richieste dal suddetto articolo vengono qui di seguito r
iportate: CALCOLO PREVALENZA Euro conferimenti da soci 10.668.000 \- di cui conferimenti uva 10.594.595 \- di cui conferimenti
olive 73.405 acquisti da terzi 9.557.594 Totale parziale 20.225.594 altri acquisti 8.580.958 totale B6 28.806.552 10.668.000 /
20.225.594 * 100 = 52,75%
Possibile autore:
```

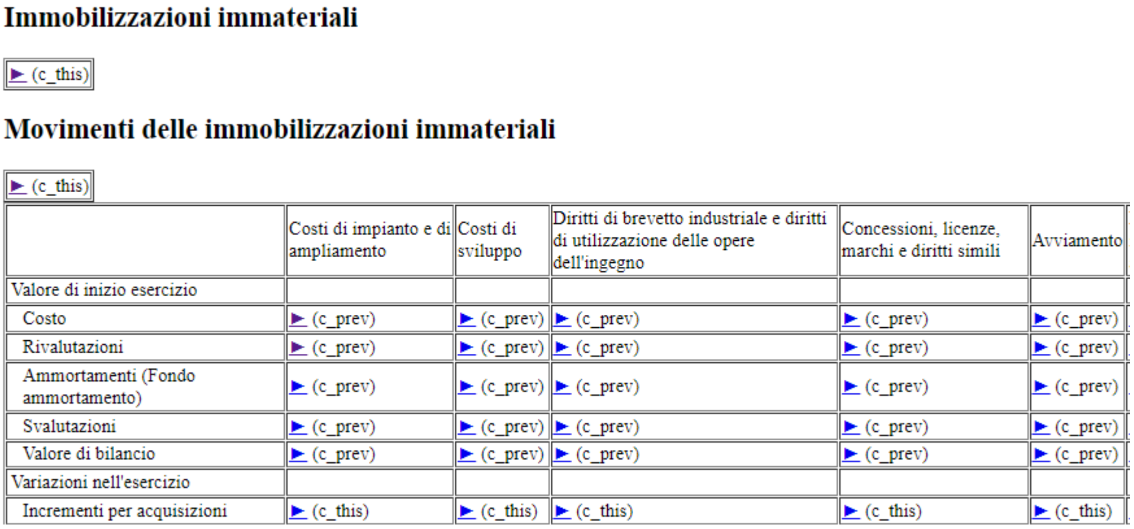
Figure 5.9: Third example of a text and the author of the revision identified by the Algorithm 9.8.

The outputs demonstrate that although the method defined on pre-trained Spacy functionalities is highly empirical and certainly requires modifications and refinements,it provides an initial valuable result that can serve as a starting point for further analysis and investigations, especially considering that internally at InfoCamere no initial procedure had been defined to extract the authors of the annual reviews without manually consulting the text.

5.3 Intangible Assets

The second section under study was related to intangible assets, understood as the part of invested capital concerning intangible corporate resources. A company not only requires equipment and monetary resources to continue its operations but also durable assets of an intangible nature, which often form the foundation of the entire business. The term "durable assets" indicates that investments of this kind are used not for a single financial period but over a longer period that can extend for several years. According to the Civil Code [16] the intangible resources to be considered in the balance sheet are diverse, such as establishment, expansion, and development costs, industrial patents, rights of use of intellectual works, concessions, licenses, trademarks, and startup costs, etc.

In particular, establishment, expansion, and development costs concern the three crucial phases in the life cycle of a company: its establishment, the expansion of its structure, and the development of production processes. These are unavoidable expenses without which a com-



Immobilizzazioni immateriali

► (c_this)

Movimenti delle immobilizzazioni immateriali

► (c_this)

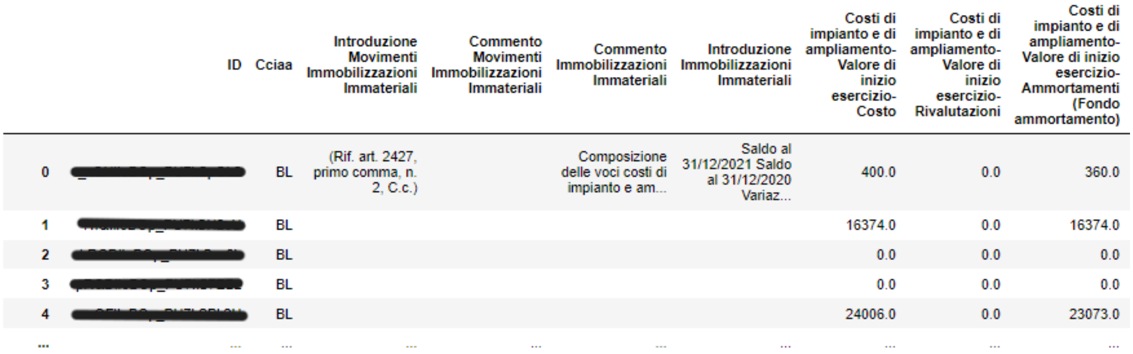
	Costi di impianto e di ampliamento	Costi di sviluppo	Diritti di brevetto industriale e diritti di utilizzazione delle opere dell'ingegno	Concessioni, licenze, marchi e diritti simili	Avviamento
Valore di inizio esercizio					
Costo	► (c_prev)	► (c_prev)	► (c_prev)	► (c_prev)	► (c_prev)
Rivalutazioni	► (c_prev)	► (c_prev)	► (c_prev)	► (c_prev)	► (c_prev)
Ammortamenti (Fondo ammortamento)	► (c_prev)	► (c_prev)	► (c_prev)	► (c_prev)	► (c_prev)
Svalutazioni	► (c_prev)	► (c_prev)	► (c_prev)	► (c_prev)	► (c_prev)
Valore di bilancio	► (c_prev)	► (c_prev)	► (c_prev)	► (c_prev)	► (c_prev)
Variazioni nell'esercizio					
Incrementi per acquisizioni	► (c_this)	► (c_this)	► (c_this)	► (c_this)	► (c_this)

Figure 5.10: Partial sections of the taxonomy for intangible assets.

pany cannot be founded and sustained in its function. Among these costs are included notarial fees, resources to be used for restructuring premises, initiating new production processes, and research and development expenses for the creation of new ones.

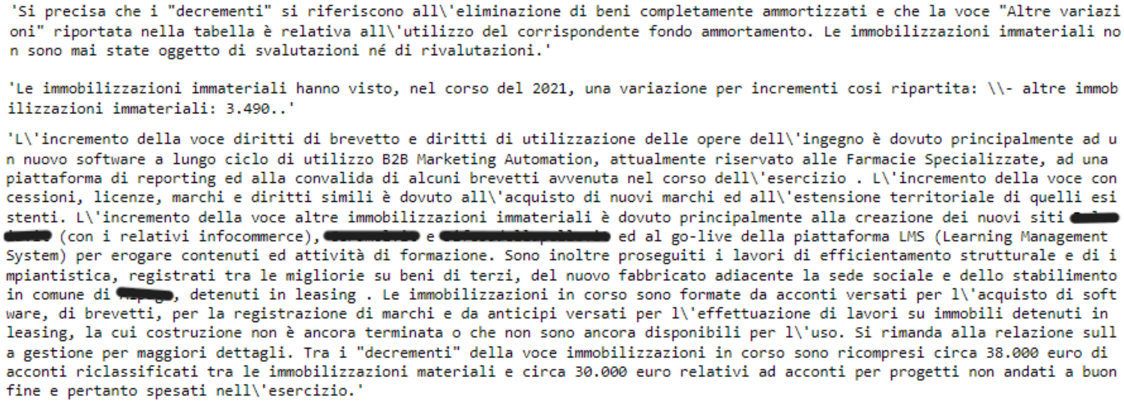
In Figure 5.10 a partial example of two sections of interest related to intangible assets is presented. The first section includes a textual tag introducing the intangible assets, while the second section presents numerous entries of numeric tags for different areas and time periods, along with final comments. Only one of the highlighted comment tags in Figure 5.1 would be the subject of study, but to obtain a structured dataset that includes all tags on the topic function 9.3 was invoked to retrieve and merge the two dataframes of the respective sections. Figure 5.11 shows only a partial example of the result, as it includes a total of 148 features. It can be observed how the overall structure of the taxonomy has been transformed so that each document corresponds to a single row. This result leads to a less readable view of the complete picture but is optimal for data analysis and processing workflows.

The previous Algorithm 9.6 was tested in all three different combinations to identify texts mentioning a revaluation or change in intangible assets using keywords such as “revaluation”, “change”, “increase”, and similar terms. Once again, very similar results were obtained for all three settings. Figure 5.12 provides examples of texts identified by the algorithm as containing a change in the value of intangible assets.



	ID	Cciaa	Introduzione Movimenti Immobilizzazioni Immateriali	Commento Movimenti Immobilizzazioni Immateriali	Commento Immobilizzazioni Immateriali	Introduzione Immobilizzazioni Immateriali	Costi di impianto e di ampliamento-Valore di inizio esercizio-Costo	Costi di impianto e di ampliamento-Valore di inizio esercizio-Rivalutazioni	Costi di impianto e di ampliamento-Valore di inizio esercizio-Ammortamenti (Fondo ammortamento)
0	[redacted]	BL	(Rif. art. 2427, primo comma, n. 2, C.c.)		Composizione delle voci costi di impianto e am...	Saldo al 31/12/2021 Saldo al 31/12/2020 Variaz...	400.0	0.0	360.0
1	[redacted]	BL					16374.0	0.0	16374.0
2	[redacted]	BL					0.0	0.0	0.0
3	[redacted]	BL					0.0	0.0	0.0
4	[redacted]	BL					24006.0	0.0	23073.0
...	...	...	...	...	...	...	...	...	...

Figure 5.11: Partial structured data for intangible assets; each row represents a document and it contains the identifier obscured for privacy reasons, the relevant chamber of commerce and the values associated with each tag in the section.



```
'Si precisa che i "decrementi" si riferiscono all\'eliminazione di beni completamente ammortizzati e che la voce "Altre variazi
oni" riportata nella tabella è relativa all\'utilizzo del corrispondente fondo ammortamento. Le immobilizzazioni immateriali no
n sono mai state oggetto di svalutazioni né di rivalutazioni.'

'Le immobilizzazioni immateriali hanno visto, nel corso del 2021, una variazione per incrementi cosi ripartita: \\- altre immob
ilizzazioni immateriali: 3.490..'

'L\'incremento della voce diritti di brevetto e diritti di utilizzazione delle opere dell\'ingegno è dovuto principalmente ad u
n nuovo software a lungo ciclo di utilizzo B2B Marketing Automation, attualmente riservato alle Farmacie Specializzate, ad una
piattaforma di reporting ed alla convalida di alcuni brevetti avvenuta nel corso dell\'esercizio . L\'incremento della voce con
cessioni, licenze, marchi e diritti simili è dovuto all\'acquisto di nuovi marchi ed all\'estensione territoriale di quelli esi
stenti. L\'incremento della voce altre immobilizzazioni immateriali è dovuto principalmente alla creazione dei nuovi siti [redacted]
 (con i relativi infocommerce), [redacted] e [redacted] ed al go-live della piattaforma LMS (Learning Management
System) per erogare contenuti ed attività di formazione. Sono inoltre proseguiti i lavori di efficientamento strutturale e di i
mpiantistica, registrati tra le migliorie su beni di terzi, del nuovo fabbricato adiacente la sede sociale e dello stabilimento
in comune di [redacted], detenuti in leasing . Le immobilizzazioni in corso sono formate da acconti versati per l\'acquisto di soft
ware, di brevetti, per la registrazione di marchi e da anticipi versati per l\'effettuazione di lavori su immobili detenuti in
leasing, la cui costruzione non è ancora terminata o che non sono ancora disponibili per l\'uso. Si rimanda alla relazione sull
a gestione per maggiori dettagli. Tra i "decrementi" della voce immobilizzazioni in corso sono ricompresi circa 38.000 euro di
acconti riclassificati tra le immobilizzazioni materiali e circa 30.000 euro relativi ad acconti per progetti non andati a buon
fine e pertanto spesati nell\'esercizio.'
```

Figure 5.12: Three texts that the Algorithm 9.6 flags as mentioning revaluations; details that could lead to the companies that own the documents are obscured.

5.4 Key points and results of the text mining applications

Starting from the data retrieved through the procedure described in Chapter 4, a valid method was defined to obtain structured data conforming to the representation in the taxonomy file and suitable for analysis and possible future machine learning processes.

Methods for keywords and entities search were devised and tested on the texts of introductions to cooperatives and intangible assets, aiming to answer simulated questions reflecting potential real needs of InfoCamere's clients. The methods were not configured for a simple check of term presence, but with the objective of finding actual fulfillment of the requests.

In particular, the method for retrieving the names of authors of the annual review of cooperatives using Spacy functions, although requiring refinement, already provides promising results. This output is highly relevant since this data is not directly extractable from other tags in the taxonomy of the supplementary notes and can lead to valuable new insights.

6
Large Language Models

This chapter focuses on the application of pre-trained large language models to expand the number and variety of possible questions applicable to the textual tags of the supplementary notes.

6.1 Exploiting a Large Language Model

The algorithms developed to solve the Text Mining tasks outlined in Chapter 5 prove suitable for searching for mentions of keywords within a text. The results obtained have been considered satisfactory for being the first attempt to define an automatable procedure, especially considering that previously such requests were fulfilled through manual searches in the available documents. However, their type of application is quite limited, and the methods may fail to return the correct outputs if suitable keywords are not provided or if synonyms of these keywords are used within the text. Additionally, one must consider the possible presence of grammatical errors, which could prevent a correct verification of the content of the text, even though some of these errors could be mitigated with an automatic corrector.

It would therefore be preferable to exploit a Large Language Model (LLM) [35] both to make the search for keywords more flexible and to expand the number and variety of solvable tasks.

LLMs are artificial neural networks designed to understand and generate text in an advanced manner. They are models known for their ability to interpret text and perform various Natural Language Processing tasks such as question-answering, text-generation from an initial prompt, classification, and more. They accomplish these tasks by learning the statistical relationships between tokens in the training documents through highly computationally intensive self-supervised and semi-supervised processes and reinforcement learning from human feedback.

At the moment creating a specialized LLM for analyzing the documents of supplementary notes is not feasible, especially due to the inability to perform semi-supervised learning since the texts lack true labels containing the correct answers to possible questions regarding the content of textual tags. Although semi-supervised learning is not strictly necessary, without true labels the only way to verify the correctness of an LLM's answers is to manually check the text under examination, and to assess the quality of a model one would need to examine several thousand responses. The inherent difficulty in this verification process is that each question may have multiple possible answers and one must possess adequate subject matter knowledge to determine which answers are truly correct and which are not. It would therefore be necessary to examine thousands of documents each with its textual tags, think of possible questions for each of these, and label the correct answers. One complication related to this task could be the distribution of textual tags. As seen in Chapter 4, when present these tags are often valued but are sparsely distributed, and only a few are truly common among documents. This applies to the data from Veneto but if the situation remains unchanged when considering all of Italy there is a risk of overfitting on highly valued tags. This could result in a model that is unable to correctly analyze and answer questions regarding less common tags.

In summary, the lack of true labels and thus the impossibility of conducting semi-supervised learning, combined with the sparse distribution of textual tags, leads to the risk of overfitting on the most common tags, the model's inability to correct its errors and the incapacity to adequately assess its performance.

Therefore, one possible alternative would be to process and train a LLM on Italian texts containing labels but it may underperform when applied to the supplementary notes due to the particular financial context. The other possibility ultimately adopted is to exploit a pretrained LLM.

To obtain additional support and take advantage of external information, a Retrieval Augmented Generation (RAG) model [36] could be used in the future. Basically, this model combines the capabilities of LLMs with the possibility of recovering information from external sources, in order to obtain texts more relevant and answers more accurate based on updated data and more specific for the sector. A RAG model can therefore obtain better performance than a LLM, but everything depends on the external information it manages to obtain, both in quantity and quality, and in this case it should be financial and in Italian. Despite the potential of RAG, it was decided to adopt a LLM, with the aim of verifying its ability to answer Text Mining tasks.

6.1.1 Searching for a Large Language Model

To search for a pre-trained LLM we relied on Hugging Face [37], an open-source platform that provides users with various pre-trained models for different tasks, datasets, and other development tools. Its strength lies in the ease of sharing these resources, allowing developers to quickly integrate them into their projects as was the case with the candidate.

Using the various filters provided by the search interface of available models, the candidate searched for a model suitable for question answering tasks on Italian text. The intrinsic difficulty of this search is the limited availability of models trained on Italian texts and with an acceptable number of parameters, indicative of model scalability.

During this phase it was noticed the large number of models suitable for processing English language. Therefore it was considered using them for question-answering tasks on the supplementary notes previously translated. For this purpose it was experimented with different text translation models to assess the quality of their results, but these highlighted the high variability in the level of translations, due to the presence of abbreviations, errors, and other peculiarities of the original text that make LLMs unable to provide adequate answers to different questions. For example, the term "Confcooperative" is correctly maintained unchanged, while "confcooperative" becomes "joint ventures", thus depriving the word of its initial value. The power of a highly performing model for English texts would therefore be nullified by an incorrect translation.

Therefore, two models for the Italian language were identified based on the results of some test cases: mdeberta-v3-base-squad2 by Tim Isbister [8] (referred as Timpal) and deberta-italian-question-answering by Francesco Russo [9] (referred as Osiria). Both are BERT-based models (Bidirectional Encoder Representations from Transformers) that use bidirectional training on large amounts of unlabeled text and transformer architecture to create deep and contextualized language representations.

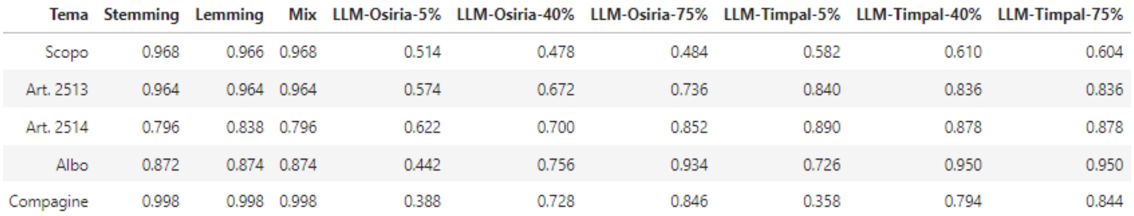
BERT [38] is an open source machine learning framework for NLP developed by researchers at Google in 2018 that stands out for its ability to be effective for a wide range of NLP tasks such as named entity recognition, question answering [39], and more. BERT's model architecture is a multi-layer bidirectional Transformer [40] encoder and its main charateristic is using bidrectional self-attention to incorporate context from both directions, a decisive factor in sentence-level tasks. Each sentence is divided into two substrings, and for each token its input representation is constructed by summing the corresponding token itself, the substring it belongs to and the position embeddings. Bidrectionality is obtained by two unsupervised learning tasks during pre-training: Masked Language Model (MLM) and Next Sentence Prediction (NSP). MLM randomly masks some percentage of the input tokens at random, and then predict those masked tokens based only on their context. The final hidden vectors corresponding to the mask tokens are fed into an output softmax over the vocabulary. Unlike left-to-right language model pre-training, the MLM objective enables the representation to fuse the left and the right context, which allows to pre-train a deep bidirectional Transformer. Many important tasks such as Question Answering are based on understanding the relationship between two sentences, which is not directly captured by language modeling. In order to train a model that understands sentence relationships, the binary learning task NSP picks two sentences A and B from the corpus, where 50% of the time B is the actual next sentence that follows A, and 50% of the time it is a random sentence. The pre-trained BERT model can be fine-tuned with just one additional output layer for a wide range of tasks, such as question answering and language inference, without changing the architecture. Compared to pre-training, fine-tuning is relatively inexpensive.

DeBERTa (Decoding-enhanced BERT with disentangled attention) model evolves BERT architecture by two novel techniques: a disentangled attention mechanism and an enhanced mask decoder [41]. While in BERT a single vector represents both the content and the position of each input word, the disentangled attention mechanism uses two separate vectors: one for the content and the other for the position. Then the attention weights among words are computed via disentangled matrices on both their contents and relative positions. That is, the attention weight of a word pair can be computed as a sum of four attention scores using disentangled matrices on their contents and positions as content-to-content, content-to-position, position-to-content, and position-to-position. This helps to better focus on relevant information and reduce interference from irrelevant context. Like BERT, DeBERTa is pre-trained using MLM. The disentangled attention mechanism already considers the contents and relative positions of the context words, but not the absolute positions of these words, which in many cases are crucial for the prediction. DeBERTa uses an enhanced mask decoder to improve MLM by adding absolute position information of the context words at the MLM decoding layer. This is done right after all the Transformer layers but before the softmax layer for masked token prediction.

In this way, DeBERTa uses the relative positions in all the Transformer layers and the absolute positions only as complementary information when decoding the masked words.

DeBERTa Version 3 [10] substitutes MLM with Replace Token Detection (RTD), a more efficient pre-training task, and adopts a new embedding sharing method, Gradient Disentangled Embedding Sharing (GDES). In the new version there are two transfomer encoders in Generative Adversarial Network style. One is a generator trained with MLM, the other is a discriminator trained with a token-level binary classifier. The generator creates ambiguous tokens to replace masked tokens in the input sequence, then the modified input sequence is fed to the discriminator. The binary classifier in the discriminator needs to determine if a corresponding token is either an original token or a token replaced by the generator. The training objective in the discriminator is called RTD. Sharing token embeddings between the generator and the discriminator allows the generator to provide informative inputs for the discriminator and reduces the number of parameters to learn, enabling the two models to learn from the same vocabulary and leverage the rich semantic information encoded in the embeddings. However, embedding sharing hurts training efficiency and model performance, since the training losses of the discriminator and the generator pull token embeddings into opposite directions, and this leads to conflicting signals and inefficient training. MLM encourages the embeddings of semantically similar tokens to be close to each other, while RTD tries to separate them to make the classification easier. The GDES method does not allow the RTD loss to affect the gradients of the generator, thus avoiding the interference and inefficiency caused by the conflicting objectives. Instead, GDES only updates the generator embeddings with the MLM loss, which ensures the consistency and coherence of the generator output.

Both the adopted models are based on DeBERTa with just some differences for the fine-tuning process. Timpal model was fine-tuned on Stanford Question Answering Dataset 2.0 [42], a reading comprehension dataset of questions posed by crowdworkers on a Wikipedia articles set, where the answer to every question is a segment of text from the corresponding reading passage, or the question might be unanswerable. Osiria model is first fine-tuned on the English SQuAD v2, then further fine-tuned on the italian subversion, SQuAD-it, and lastly fine-tuned on the lowercase SQuAD-it. This helps making the model generally more robust, but particularly in uncased settings. SQuAD-it is derived from the SQuAD dataset and it is obtained through semi-automatic translation of the SQuAD dataset into Italian, containing more than 60,000 question/answer pairs. This means that the quality of the training set is limited by the machine translation. Moreover, the model is meant to answer questions under the assumption that the required information is actually contained in the given context. If the



Tema	Stemming	Lemming	Mix	LLM-Osiria-5%	LLM-Osiria-40%	LLM-Osiria-75%	LLM-Timpal-5%	LLM-Timpal-40%	LLM-Timpal-75%
Scopo	0.968	0.966	0.968	0.514	0.478	0.484	0.582	0.610	0.604
Art. 2513	0.964	0.964	0.964	0.574	0.672	0.736	0.840	0.836	0.836
Art. 2514	0.796	0.838	0.796	0.622	0.700	0.852	0.890	0.878	0.878
Albo	0.872	0.874	0.874	0.442	0.756	0.934	0.726	0.950	0.950
Compagine	0.998	0.998	0.998	0.388	0.728	0.846	0.358	0.794	0.844

Figure 6.1: Comparison of accuracy of methods and LLMs; for each Text Mining task "Tema", it is reported the percentage of times the Algorithm 9.6 using stemming, lemmatization or a combination of both ("Mix") and the LLMs with different threshold levels for the answer score return the correct label True/False.

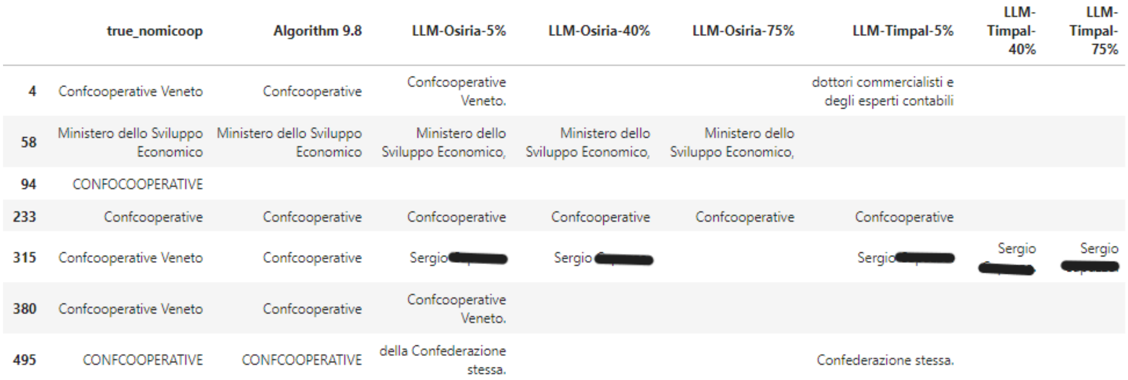
assumption is violated, the model will try to return an answer in any case, which is going to be incorrect.

6.1.2 Application of Large Language Models

These models were used to answer the questions about the contents of the information regarding cooperative societies. Each model receives in input a question as "La società è una cooperativa?" and a text where to find the answer, then it outputs a dictionary containing a substring of the text as the answer to the question asked and a corresponding score in the range [0,1] indicating the confidence in the reliability of the answer. Since all questions are posed with a positive connotation, the returned string is converted to False if it contains the word "non", otherwise it's considered True. In case the string that would be returned has a score lower than a certain threshold, False is returned.

To verify the accuracy of the output of the methods and LLMs the first 500 texts of the information of cooperative societies were manually read, and a true label (True/False) was assigned to each of them for the Text Mining questions. For the case of retrieving the names of the authors of the annual reviews, the substring containing the information is assigned if present, or an empty string.

Figure 6.1 shows the percentage of correctly predicted labels for the various questions from Algorithm 9.6 using stemming, lemmatization or a combination of both, and the LLMs with different threshold levels for the answer score. It can be noticed that using stemming, lemmatization or a combination leads to very similar results. It can be noticed how the percentage of correctly classified labels from the LLMs rise with higher thresholds; this can be explained by the fact that at low thresholds the models returns substrings that may be inconsistent with the search topic, while at higher thresholds the substrings suitable to be returned are fewer and fewer and the probability of returning an incorrect one decreases.



	true_nomicoop	Algorithm 9.8	LLM-Osiria-5%	LLM-Osiria-40%	LLM-Osiria-75%	LLM-Timpal-5%	LLM-Timpal-40%	LLM-Timpal-75%
4	Confcooperative Veneto	Confcooperative	Confcooperative Veneto.			dottori commercialisti e degli esperti contabili		
58	Ministero dello Sviluppo Economico	Ministero dello Sviluppo Economico	Ministero dello Sviluppo Economico,	Ministero dello Sviluppo Economico,	Ministero dello Sviluppo Economico,			
94	CONFOCOOPERATIVE							
233	Confcooperative	Confcooperative	Confcooperative	Confcooperative	Confcooperative	Confcooperative		
315	Confcooperative Veneto	Confcooperative	Sergio [redacted]	Sergio [redacted]		Sergi[redacted]	Sergio [redacted]	Sergio [redacted]
380	Confcooperative Veneto	Confcooperative	Confcooperative Veneto.					
495	CONFCOOPERATIVE	CONFCOOPERATIVE	della Confederazione stessa.			Confederazione stessa.		

Figure 6.2: Comparison of answers for authors of revision; for each row representing a text, there is the true author of the review "true_nomicoop" and the substring returned by the developed Algorithm 9.8 or by the LLMs with different confidence thresholds.

For the task of identifying authors of reviews Figure 6.2 shows the true authors and the ones identified by the Algorithm 9.8 and the LLMs. The developed method provides results more coherent with the true ones than the LLMs, with only Osiria returning some suitable substrings. This can be explained by the fact that by construction the method focuses only on entities of names, companies and similars, while the models pay attention to all the words present in the text and risk to not focus on the interesting tokens.

6.2 Key points and results of the adoption of Large Language Models

The methods defined in Chapter 5 despite performing well have a limited scope of application as they are focused on keywords search within a text. Particularly, they may yield incorrect results if appropriate keywords are not provided or if the text contains synonyms not accounted for, or if there are grammatical errors.

Hence, it is preferable to exploit a Large Language Model to expand the range and variety of solvable tasks.

The impossibility of conducting semi-supervised learning due to the absence of true labels and the limited dissemination of textual tags across the documents, coupled with the risk of overfitting on the most prevalent tags, lead to discarding the option of creating a custom model trained on the supplementary notes.

The adopted alternative is to use pre-trained Large Language Models on Italian texts from

HuggingFace, an open-source platform that facilitates the sharing of models, datasets and other development tools.

It is clear that to achieve an optimal level of efficiency it is essential to subject the models to intense testing and fine-tuning phases which, however, require considerable manual effort due to the absence of predefined true labels. A potential solution to mitigate this challenge could be the method defined for keyword research to create initial true labels. Considering the vast availability of documents with different sections of the explanatory note, each with various textual tags on which questions solvable by the method can be formulated, its results could be used to obtain initial true labels. Then they could be further refined comparing them with the outputs of LLMs at high threshold levels, until acceptable quality true labels are obtained. This strategy could not only reduce manual workload but also accelerate the optimization process of the web interface, allowing for more effective use of LLMs and overall facilitating the work of a data scientist to achieve valuable outcomes.

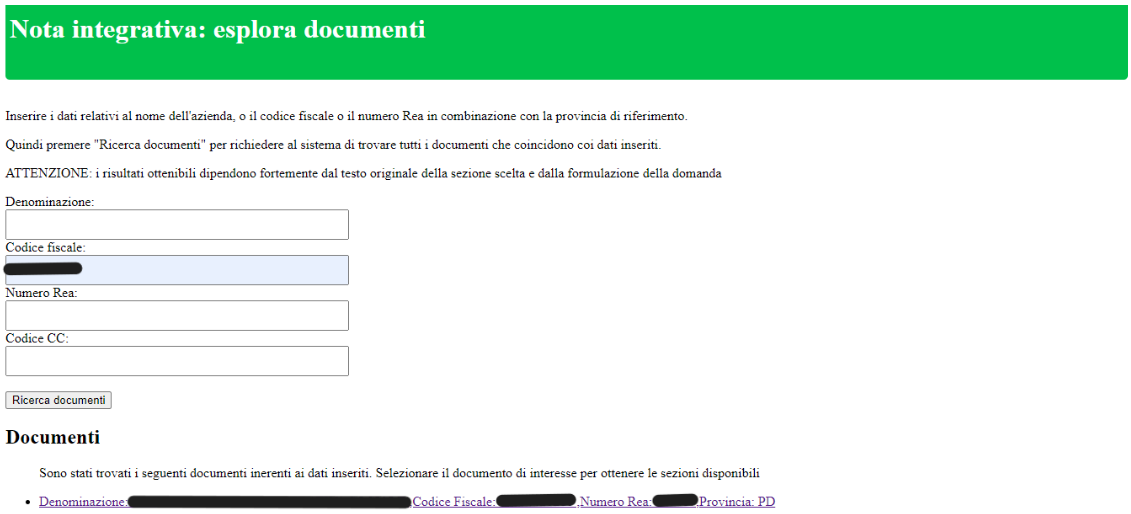
7
Pipeline prototype

This Chapter presents a first prototype of a web interface to allow user interaction with the textual sections of a searched document, enabling the execution of questions and receiving answers about their content using the defined data extraction and analysis methods.

The page appears as in Figure 7.1 with several text boxes where the user enters at least one of the company name, tax code, and the pair of REA number and reference province. Then the user presses the "Ricerca documenti" button to launch a background query to ElasticSearch which returns selectable strings of all possible documents corresponding to the entered data. To perform the search ElasticSearch checks the fields related to the tax code, REA number and chamber of commerce, while to filter by the company name it checks the documents where the distance between the "ci:DatiAnagraficiDenominazione" tag within the "ix_xbrl_all" field and the provided name is less than 20 tokens.

Once the user selects the document of interest the tags of the explanatory note are retrieved in the background using the Algorithm 9.2. Once the process is completed a list of the available sections of the explanatory note is displayed using the titles of the taxonomy. By selecting one of these sections the tags of the corresponding taxonomy are retrieved using the method 9.3. Then, it is possible to formulate a free question through the appropriate text box or choose one of the pre-set questions for that particular section to facilitate exploration. Figure 7.2 shows on the left the list of available sections and on the right a text box and a list of questions related to the selected section.

Once one of the preset questions is selected or a question is written in the text box, both



Nota integrativa: esplora documenti

Inserire i dati relativi al nome dell'azienda, o il codice fiscale o il numero Rea in combinazione con la provincia di riferimento.

Quindi premere "Ricerca documenti" per richiedere al sistema di trovare tutti i documenti che coincidono coi dati inseriti.

ATTENZIONE: i risultati ottenibili dipendono fortemente dal testo originale della sezione scelta e dalla formulazione della domanda

Denominazione:

Codice fiscale: [redacted]

Numero Rea:

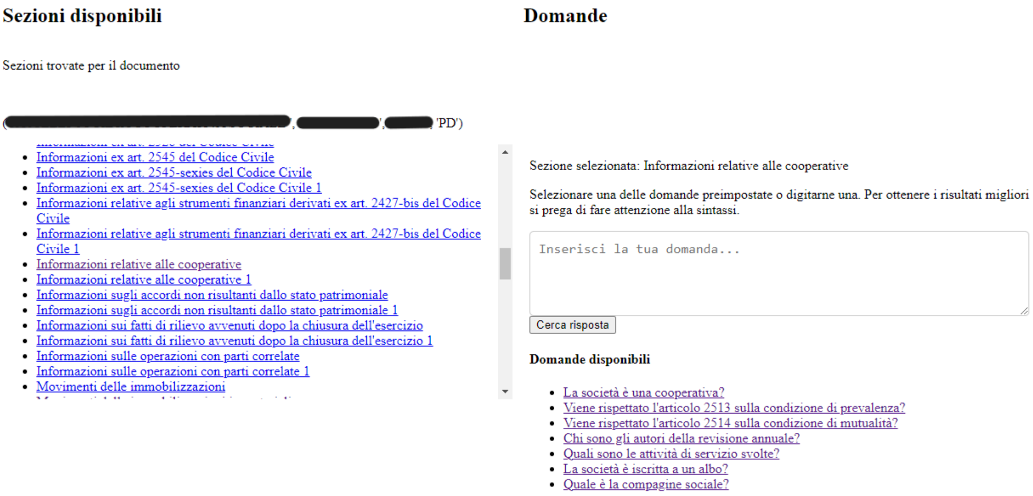
Codice CC:

Documenti

Sono stati trovati i seguenti documenti inerenti ai dati inseriti. Selezionare il documento di interesse per ottenere le sezioni disponibili

- Denominazione:[redacted] Codice Fiscale:[redacted] Numero Rea:[redacted] Provincia: PD

Figure 7.1: Initial search interface. The selectable link in "Documenti" refers to the document that contains the code specified in "Codice fiscale"; details that refer directly to the company are obscured for privacy reasons.



Sezioni disponibili

Sezioni trovate per il documento

([redacted], [redacted], [redacted] 'PD')

- Informazioni ex art. 2545 del Codice Civile
- Informazioni ex art. 2545-sexies del Codice Civile
- Informazioni ex art. 2545-sexies del Codice Civile 1
- Informazioni relative agli strumenti finanziari derivati ex art. 2427-bis del Codice Civile
- Informazioni relative agli strumenti finanziari derivati ex art. 2427-bis del Codice Civile 1
- Informazioni relative alle cooperative
- Informazioni relative alle cooperative 1
- Informazioni sugli accordi non risultanti dallo stato patrimoniale
- Informazioni sugli accordi non risultanti dallo stato patrimoniale 1
- Informazioni sui fatti di rilievo avvenuti dopo la chiusura dell'esercizio
- Informazioni sui fatti di rilievo avvenuti dopo la chiusura dell'esercizio 1
- Informazioni sulle operazioni con parti correlate
- Informazioni sulle operazioni con parti correlate 1
- Movimenti delle immobilizzazioni

Domande

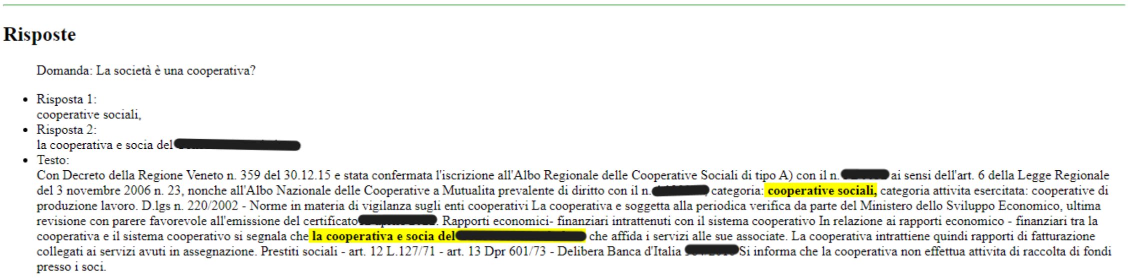
Sezione selezionata: Informazioni relative alle cooperative

Selezionare una delle domande preimpostate o digitarne una. Per ottenere i risultati migliori si prega di fare attenzione alla sintassi.

Domande disponibili

- La società è una cooperativa?
- Viene rispettato l'articolo 2513 sulla condizione di prevalenza?
- Viene rispettato l'articolo 2514 sulla condizione di mutualità?
- Chi sono gli autori della revisione annuale?
- Quali sono le attività di servizio svolte?
- La società è iscritta a un albo?
- Quale è la compagine sociale?

Figure 7.2: Sections retrieved for the document and possible answers. "Sezioni disponibili" contains a list of the sections of the selected document, shown above but obscured for privacy reasons. "Domande" displays a text box where the user can write a question or selects pre-set questions for the chosen section.

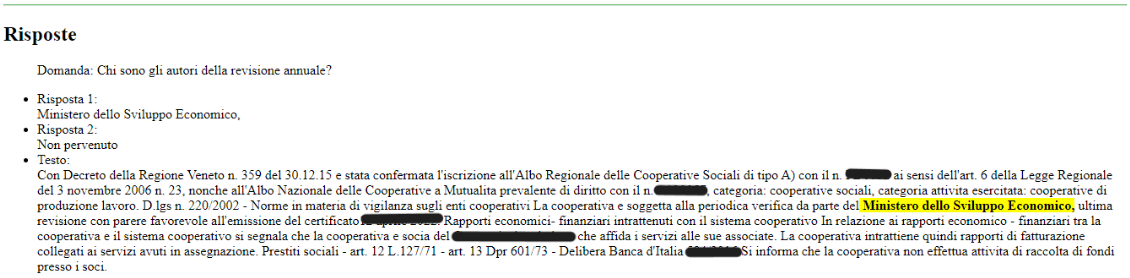


Risposte

Domanda: La società è una cooperativa?

- Risposta 1:
  cooperative sociali,
- Risposta 2:
  la cooperativa e socia del[redacted]
- Testo:
  Con Decreto della Regione Veneto n. 359 del 30.12.15 e stata confermata l'iscrizione all'Albo Regionale delle Cooperative Sociali di tipo A) con il n.[redacted]ai sensi dell'art. 6 della Legge Regionale del 3 novembre 2006 n. 23, nonche all'Albo Nazionale delle Cooperative a Mutualita prevalente di diritto con il n.[redacted] categoria: **cooperative sociali,** categoria attivita esercitata: cooperative di produzione lavoro. D.lgs n. 220/2002 - Norme in materia di vigilanza sugli enti cooperativi La cooperativa e soggetta alla periodica verifica da parte del Ministero dello Sviluppo Economico, ultima revisione con parere favorevole all'emissione del certificato[redacted]. Rapporti economici- finanziari intrattenuti con il sistema cooperativo In relazione ai rapporti economico - finanziari tra la cooperativa e il sistema cooperativo si segnala che **la cooperativa e socia del**[redacted] che affida i servizi alle sue associate. La cooperativa intrattiene quindi rapporti di fatturazione collegati ai servizi avuti in assegnazione. Prestiti sociali - art. 12 L.127/71 - art. 13 Dpr 601/73 - Delibera Banca d'Italia [redacted]Si informa che la cooperativa non effettua attivita di raccolta di fondi presso i soci.

Figure 7.3: To the reported question, the LLMs return the substring of the text understood as relevant to the answer and highlight it; details that refer directly to the company are obscured for privacy reasons.



Risposte

Domanda: Chi sono gli autori della revisione annuale?

- Risposta 1:
  Ministero dello Sviluppo Economico,
- Risposta 2:
  Non pervenuto
- Testo:
  Con Decreto della Regione Veneto n. 359 del 30.12.15 e stata confermata l'iscrizione all'Albo Regionale delle Cooperative Sociali di tipo A) con il n. [redacted]ai sensi dell'art. 6 della Legge Regionale del 3 novembre 2006 n. 23, nonche all'Albo Nazionale delle Cooperative a Mutualita prevalente di diritto con il n.[redacted], categoria: cooperative sociali, categoria attivita esercitata: cooperative di produzione lavoro. D.lgs n. 220/2002 - Norme in materia di vigilanza sugli enti cooperativi La cooperativa e soggetta alla periodica verifica da parte del **Ministero dello Sviluppo Economico,** ultima revisione con parere favorevole all'emissione del certificato[redacted]Rapporti economici- finanziari intrattenuti con il sistema cooperativo In relazione ai rapporti economico - finanziari tra la cooperativa e il sistema cooperativo si segnala che la cooperativa e socia del[redacted]che affida i servizi alle sue associate. La cooperativa intrattiene quindi rapporti di fatturazione collegati ai servizi avuti in assegnazione. Prestiti sociali - art. 12 L.127/71 - art. 13 Dpr 601/73 - Delibera Banca d'Italia[redacted]Si informa che la cooperativa non effettua attivita di raccolta di fondi presso i soci.

Figure 7.4: Answers of the LLM models for another question; for this text only the first model (Osiria) manages to find a relevant substring and it highlights it. Details that refer directly to the company are obscured for privacy reasons.

adopted LLMs are triggered and their responses are returned, highlighting the portions of text from which they extracted the information. Figure 7.3 shows how the models answered the question of whether the company is a cooperative by returning the most relevant portions of text considered for the positive response.

In case the associated score is below a predefined threshold level the model returns a default text related to the failure, as seen for the second response in Figure 7.4.

Thanks to the combination of the methods defined previously for data extraction and organization, along with the capability of LLM models to interpret free-text questions, the elaborated web interface prototype demonstrates promising potential. The methods for structuring the data can already render them in a format more suitable for search queries compared to the original texts, but there might still be difficulties for users who are not familiar with the necessary tools or if the specific data searched is located within a lengthy text. However, when combined with LLMs and a user-friendly interface it opens up the possibility for any user to formulate questions and receive more or less suitable answers. Despite its potential, the prototype highlights the need for significant improvements to optimize the performance of the LLMs.

8
Conclusions

The conclusions drawn from this research work reflect a thorough and critical analysis of the challenges and opportunities in the application of Text Mining techniques to financial data, focusing particularly on the supplementary notes stored in XBRL format. The investigations carried out in this thesis have brought to a series of fundamental considerations regarding the analysis of administrative data through Text Mining techniques.

First, the vastness of the available data, though inherently a valuable resource, requires appropriate infrastructure and domain expertise for its complete fruition. Despite the availability of the highly performing ElasticSearch software for query resolution, the initial resources, limited in terms of computing power, and the extremely detailed and context-specific information contained in administrative data considerably limited the application of Text Mining techniques, making the extraction and processing of significant and useful information challenging.

More specifically, this process of data cleaning and information retrieval required going beyond the acquired knowledge of the context of supplementary notes, delving into the specific schema of the XBRL format used for document storage. This study allowed for the delineation of an effective strategy for data extraction and structuring, ensuring the coherence and usability of the collected information for future processes. However, in order to apply the outlined procedure, it was necessary to focus on a subset of documents and to explore a subset of context-specific questions. This choice highlights how data must not only be plentiful but also be made available in formats suitable for one's purposes, or they may even prove unusable.

Analyzing the texts, it became apparent that the linguistic and terminological diversity makes them significantly different from other types of data commonly treated in the field of Text Mining. The presence of technicalities, abbreviations and grammatical errors can challenge the understanding of even the most sophisticated natural language models, preventing them from recovering valuable information that a human operator could identify. For these reasons, administrative and financial texts could be considered a specialized branch of Text Mining, with data vastly dissimilar to articles found on Wikipedia, Twitter and newspapers, which are often used to train LLMs.

A significant challenge in the process of fine-tuning to enhance the performance of models for this textual domain is the lack of true labels. Although solutions can be created to address this gap, it would remain extremely laborious to examine a sufficient number of supplementary notes and for each of these develop possible questions and answers for the hundreds of textual sections.

The main conclusion of this thesis is that, despite the availability of large quantities of data and high-performing software being undeniable advantages, the complexity and heterogeneity of the former can limit their practical utility. The presence of linguistic and structural characteristics can challenge even the most advanced Machine Learning models to the extent that access to valuable information remains closely tied to manual analysis and human interpretation processes. This remarks the importance of a flexible and adaptable approach in the field of Text Mining, capable of addressing the peculiarities and specific challenges of various application domains. While automation and Artificial Intelligence can offer powerful tools for data analysis, it is crucial to maintain a central role for human expertise to overcome the intrinsic limitations of Machine Learning models and fully exploit the potential of administrative data for analytical and decision-making purposes.

Despite the above limitation, this thesis has shown that Text Mining techniques can be applied to administrative data in order to answer detailed and context-specific questions. In this thesis this was done in the context of cooperative societies in the Veneto region. In addition, this thesis has shown that LLMs, although extremely advanced, need to be applied with caution given the potential differences between their training data and the specific wording and jargon of administrative data.

Further extensions of this work will include applying fine-tuning to pre-trained models and defining test prompts. Fine-tuning on exploratory notes can increase the models' ability to understand the type of content and therefore return more valuable and reliable results. To lighten the manual work of reading the texts and assigning true labels, essential for fine-tuning but absent for the data available, the defined methods for keywords search can be exploited to obtain at least some initial labels. The prompt tests serve to establish a standard to be able to apply valuable comparisons between the models used and any future ones and thus obtain reliable performance metrics.

9
Appendix

This appendix reports the algorithms and pseudocodes mentioned in the previous chapters. Each algorithm is associated with a brief description of its operation and purpose.

Algorithm 9.1 Example of document from ElasticSearch, with its fields and the associated values.

```
”_index”: ”XXXX_XXXX_XXXX_0000”,
”_id”: ”XxXXxXxXXx_XX0xXXxXx”,
”ix_dt_curr_dce”: ”Dec 31, 2021 @ 00:00:00.000”,
”ix_cciaa_rea”: ”PD”,
”ix_cod_fisc”: ”000000000000,
......
”ix_xbrl_all”:”<?xml version=\”1.0\” encoding=\”UTF-8\”?>\r\n<xbrl
xmlns=\”http://www.xbrl.org/2003/instance\” xmlns:link=\”http://
www.xbrl.org/2003/linkbase\” xmlns:xlink=\”http://www.w3.org/1999/
xlink\” xmlns:iso4217=\”http://www.xbrl.org/2025/iso4217\”xmlns:
xbrli=\”http://www.xbrl.org/2003/instance\”xmlns:itcc-ci=\”http://
www.infocamere.it/itnn/fr/itcc/ci/2025-11-04\” xmlns:itcc-ciabb=
\”http://www.infocamere.it/itnn/fr/itcc/ci/abb/2025-11-04\”>\r\n
<link:schemaRef xlink:type=\”simple\” xlink:arcrole=\”http:/www.
w3.org/1999/xlink/properties/linkbase\” xlink:href=\”itcc-ci-abb-
2025-11-04.xsd\”/
>
\r\n <contextid=\”c2020_i\”>\r\n <entity>\r\n <identifier
scheme=\”http://www.infocamere.it\”>000000000000</identifier>\r\n
</entity>\r\n <period>\r\n <instant>2020-12-31</instant>\r
\n</period>\r\n <scenario>\r\n <itccci:scen>Depositato
</itcc-ci:scen>\r\n
.......
\r\n\t<itcc-ci:TotaleImmobilizzazioniImmateriali contextRef=\”
c2021_i\” unitRef=\”eur\” decimals=\”0\”>25445</itcc-ci:
TotaleImmobilizzazioniImmateriali>\r\n\t<itcc-ci:
TotaleImmobilizzazioniMateriali contextRef=\”c2021_i\”unitRef=
\”eur\” decimals=\”0\”>94883</itcc-ci:
TotaleImmobilizzazioniMateriali>\r\n\t
...
<itcc-ci:CommentoNotaIntegrativa contextRef=”c0_i”>&lt;div&gt;&lt;p
style=&quot;margin-top:4.0pt;margin-right:0cm;margin-bottom2.0pt;
margin-left:0cm;text-align:justify;line-height:11.0pt;font-size:10.0
pt;font-family:&apos;Times New Roman&apos;,serif;&quot;&gt;Il
Bilancio è vero e reale e corrisponde alle scritture contabili &lt;
/p&gt; &lt;p style=&quot;margin-top:4.0pt;margin-right:0cm;margin-
bottom:2.0pt;margin-left:0cm;text-align:justify;line-height11.0pt;
font-size:10.0pt;fontfamily:&apos;Times New Roman&apos;,serif;
&quot;&gt;PADOVA, 30/06/2025&lt;/p&gt; &lt;p style=&quot;margin-
top:4.0pt;margin-right:0cm;margin-bottom:2.0pt;margin-left:0cm;text-
align:justify;line-height:11.0pt;font-size:10.0pt;fontfamily:
&apos;Times New Roman&apos;,serif;&quot;&gt;Malvetta Giuseppe &lt;/
p&gt; &lt;p style=&quot;margin-top:4.0pt;margin-right:0cm;margin-
bottom:2.0pt;margin-left:0cm;text-align:justify;line-height11.0pt;
font-size:10.0pt;font-family:&apos;Times New Roman&apos;,serif;
&quot;&gt;&amp;#160;&lt;/p&gt;&lt;div&gt;
<itccci:CommentoNotaIntegrativa>
```

Algorithm 9.2 Pseudocode of function get_regione(data). It is used to extract from a dictionary of the documents of a Chamber all the tags and their values separately for each document.

```
Pseudo code of a function that receives a dictionary {id_doc:ix_xbrl_all} where id_doc
is the identifier of the document and ix_xbrl_all is a the associated string. It returns
a dictionary {id_doc:{dictionary of tags and values}}

#pattern to find all (tag t, context c, value v) from ix_xbrl_all
pattern = re.compile(r'<itcc-ci:([^>\s]+)(?:.*?contextRef=”([^”]*)”)?[^>]*?>(.*?)
<\/itcc-ci:\1>', re.DOTALL)

final_dict={} #dictionary to return

#iterate through the identifiers and their ix_xbrl_all strings
for key,value in data.items():

    matches = pattern.findall(value) #find all (t, c, v) in ix_xbrl_all

    #obtain the lists of xontexts _prev e _this from the beginning of the doc
    prev_list,curr_list = get_context_dates(value)

    result = [(t,c,v) if c else (t,v) for t,c,v in matches]

    nested_value=1 #to differentiate equal tags of repeated sections
    temp_dict=dict()
    nested_tags=dict()#to check homonymous tags

    for x in result:

        inner_text=x[-1] #take the value of the tag
        if(”itcc-ci” in inner_text):#if there are still nested tags extract them
            #use again the pattern to extract (t,c,v) and save them in matches2

            result2 = [(tag,c, v) for tag,c, v in matches2]
            for r in result2:
                #keep a tag only if of the supplementary note
                if r[0] in tags_nota_integrativa:
                    if r[1] in prev_list:
                        cont=”_prev”
                    else:
                        cont=”_this”
                    #clean up the text and convert it to a number if necessary
                    inner_text=clean_xml_content(r[-1])
                    inner_text=convert_numeric(inner_text)
                    #add the new key with its value
                    temp_dict[f'{r[0]}{cont}_{nested_value}']=inner_text
                nested_value+=1
        else:
            if x[0] in tags_nota_integrativa:
            #procedure as above to extract tag name and its context tag_cont
                value=1
                #if a tag of the same name has not already been retrieved,
                #assign a value of 1 in the dictionary nested_tags
                if tag_cont not in nested_tags.keys():
                    nested_tags[tag_cont]=value
                else:
                    nested_tags[tag_cont]+=1
                    value=nested_tags[tag_cont]
                inner_text=clean_html_content(v)
                inner_text=convert_numeric(inner_text)
                temp_dict[f'{tag_cont}_{value}']=inner_text
    #update the final dictionary with the new obtained one
    final_dict[key]=temp_dict
```

Algorithm 9.3 Code of function get_structured_section(section,index,provincia). It exploits 9.4 and 9.5 to return a dataframe of a section structured as it appears on “bilanci.html”.

```
#Code of a function that receives the title of a section from bilanci.html, a list
of dictionaries {id,ccia,tag:value,...} and an optional id of a dictionary and
#returns a structured data

res=get_structured_tags_xbrl_2(section)#obtain the dataframe of the section
total_tags=res.values.flatten().tolist()#obtain the list of the tags of the section
#from each dictionary pick the id, the province and the tags of the section
list_of_tuples=[(my_dict[”id”],my_dict[”ccia”],key, value) for my_dict in provincia
    for key, value in my_dict.items()
    if key.rstrip(’_1234567890’) in total_tags]

dis=res.copy()
dis.insert(0,”ID”,0)
dis.insert(1,”Cciaa”,0)
dis = dis.drop(index=dis.index)

if(len(list_of_tuples)>0):
    df = pd.DataFrame(list_of_tuples, columns=[’Index’, ’Cciaa’, ’Tag’, ’Value’])
    #extract the number associated to the tag in a separate feature and remove it
    df[’Tag_Number’] = df[’Tag’].str.extract(r’_(\d+)$’).astype(int)
    df[’Tag’] = df[’Tag’].apply(lambda x: x.rstrip(’_1234567890’))
    result_list=[]
    #if a specified id was not provided process all the documents
    if(index is None):
        indexes=df[”Index”].unique()
        for x in indexes:
            n=df[df[”Index”]==x][”Tag_Number”].max()
            if(math.isnan(n)):
                continue
            cciaa=list(df[df[”Index”]==x][”Cciaa”])[0]
            for i in range(1,n+1):
                df_temp = df[(df[”Index”]== x)& (df[”Tag_Number”]==i)]
                #create dictionary {tag:Value}
                tag_value_mapping = df_temp.set_index(’Tag’)[’Value’].to_dict()
                #replace the names of the section dataframe tags with the
                #corresponding ones
                df_temp = map_columns(res, tag_value_mapping)
                result_list.append([x,cciaa]+df_temp.iloc[0].tolist())
    else:
        x=index
        n=df[df[”Index”]==x][”Tag_Number”].max()
        if(math.isnan(n)):
            return ”No data”
        cciaa=list(df[df[”Index”]==x][”Cciaa”])[0]
        for i in range(1,n+1):
            df_temp = df[(df[”Index”]== x)& (df[”Tag_Number”]==i)]
            tag_value_mapping = df_temp.set_index(’Tag’)[’Value’].to_dict()
            df_temp = map_columns(res, tag_value_mapping)
            result_list.append([x,cciaa]+df_temp.iloc[0].tolist())
    #create the dataframe of the results
    columns=list(dis.columns)
    dis=pd.DataFrame(result_list, columns=columns)
    return dis
else:
    return ”No data”
```

Algorithm 9.4 Pseudocode of function get_structured_tags_xbrl_2(section). It supports 9.3 retrieving a dataframe of all the tags of the interested section.

```
#pseudocode of a function that returns a one row dataframe of the names of the tags
#of a section from the taxonomy
h2_tag=#html content of the taxonomy starting from the interested section
dataframes = []
total_tags=[]#all the tags of the section
df_temp=pd.DataFrame()
if(h2_tag is not None):
    #pick the id of the current and next sections to define when stop collecting info
    section_id =h2_tag.find_parent(’div’).get(’id’)
    next_section_id =next_h2_tag.find_parent(’div’).get(’id’)
    #pick html content of current section
    section_div=soup.find(’div’, {’id’: section_id})
    current_div = section_div.find_next([’table’])
    while(current_div.get(”qcode”)<next_section_id):
        #verify if there is the header of a table
        has_header = current_div.find(’thead’) is not None
        if(has_header):
            header=current_div.find(’thead’)
            #take the levels and the content of the first cell
            header_levels, content
            #for each cell of the header take name, level, width
            for i, row in enumerate(header_levels):
                columns_names.append((col_name,i, col_length))
            if(content is None):
                columns_names.insert(0,(”definizione”,0,1))
            #obtain the hierarchy of the cells of the header
            hierarchy=create_hierarchy(columns_names)
            #obtain a one-level list of the names
            columns_names=flatten_hierarchy(hierarchy)
            df_temp=pd.DataFrame(columns=columns_names)

            #iterate through the rows and columns of the section
            rows=current_div.find_all(’tr’)
            for r in rows:
                elements=[]
                cols = r.find_all(’td’)
                for c in cols:
                    #add the names and contexts of the tags
                    total_tags.append(href_value+”_”+span_content)
                    elements.append(href_value+”_”+span_content)
                    df_temp.loc[len(df_temp)]=elements
            if(df_temp.columns[0]==”definizione”):
                #remove empty rows and add their index names to the features
                #creating a new dataframe
                dataframes.append(new_df)
            else:
                #it is a section that can repeat itself, add the names of the
                #indexes to the features and create a new dataframe
                dataframes.append(new_df)
        else:
            #add tag name and context
            total_tags.append(href_value+”_”+span_content)
        current_div = current_div.find_next([’table’])
    #obtain a single dataframe of the results and clean it of empty columns
    return df
```

Algorithm 9.5 Code of function map_columns(df, mapping_dict). It supports 9.3 replacing tag names with their associated values if present.

```
#Code of a function that receives the structure of a dataframe with only tags names and
#replace them with the corresponding values of a dictionary {tag:value}
dfcopy=df.copy()
for tag in mapping_dict:
    dfcopy.replace(to_replace=tag, value=mapping_dict[tag], inplace=True)
#if there are still names of tags replace them with ””
dfcopy = dfcopy.applymap(lambda x: ”” if (’_prev’ in str(x) or
                                          ’_this’ in str(x)) else x)
return dfcopy
```

Algorithm 9.6 Code of function find_keywords(testo, parole_chiave, min_len). It uses stemming/lemming to discover if some keywords are present in a text if it has a minimum length. It also checks the presence of negative words and their position relative to the keywords to verify the positive meaning of the presence of the keywords.

```
#pseudocode of a function that returns True if it finds some keywords in a text, False
#otherwise or if the text is empty or shorter than the minimum length
if(testo==””):
    return False
if(min_len):
    if len(testo) < min_len:
        return False
#function to substitute abbreviations with complete words (art. -> articolo)
testo = sostituisci_abbreviazioni(testo.lower())
#obtain lists of tokens
lista_parole = tokenize(testo)
#perform stemming for each list of tokens and the keywords
stemming_list = [stemming(x) for x in lista_parole]
lemming_list = [lemming(x) for x in lista_parole]
key_stemm_list = stemming(parole_chiave)
key_lemm_list = lemming(parole_chiave)
negative_words = [”non”,”nè”]
res_stemm, res_lemm =[],[]
for sublist_stemm,sublist_lemm in zip(stemming_list,lemming_list):
    #obtain the positions of the keywords and negative words if present
    words_indexes_stemm=[sublist_stemm.index(word) for word in key_stemm_list
    if word in sublist_stemm]
    neg_indexes_stemm=[sublist_stemm.index(word) for word in negative_words
    if word in sublist_stemm]
    #if there are only keywords and no negative words add True
    if words_indexes_stemm and not neg_indexes_stemm:
        res_stemm.append(True)
    elif not words_indexes_stemm:
        pass
    #if negative words are after the keywords or too distant add True, else False
    else:
        for elem_first in words_indexes_stemm:
            for elem_second in neg_indexes_stemm:
                if(elem_first<elem_second):
                    res_stemm.append(True)
                else:
                    if (elem_first - elem_second) > 5:
                        res_stemm.append(True)
                    else:
                        res_stemm.append(False)
    #replicate using words_indexes_lemm, sublist_lemm, key_lemm_list, neg_indexes_lemm,
    #res_lemm
#if the lists are both empty return False, otherwise the mode
if ((not res_stemm) and (not res_lemm)):
    return False
elif ((not res_stemm) and res_lemm):
    return mode(res_lemm)
elif ((not res_lemm) and res_stemm):
    return mode(res_stemm)
else:
    return mode(res_stemm) and mode(res_lemm)
```

Algorithm 9.7 Code of function possible_cooperative_lists(data). It uses POS and NER to return 3 lists of retrieved entities in all the given texts. Two lists for when the review is mentioned, one for when it is not. These lists are then used by 9.8.

```
#Code of function that checks all texts of data to obtain 3 lists of words with NER
#and POS where ”revisione” appears and NER where it does not
#from NER consider only labels MISCellaneous, LOCation, ORGanization
all_ner, all_pos, avoid_names=[],[],[]
possible_names=[]
for ind, testo in enumerate(data):
    testo=sostituisci_abbreviazioni(testo)
    if ” revisione ” in testo.lower():
        sottostringhe = re.split(r’[.;]’, testo)
        possible_names_substring=[]
        for i, sottostringa in enumerate(sottostringhe):
            #if ”revisione” in substring do NER and POS in it and the neighbors
            if ” revisione ” in sottostringa.lower():
                index_before = max(0, i - 3)
                index_after = min(len(sottostringhe), i + 4)
                surrounding_strings = sottostringhe[index_before:index_after]
                testo=’ ’.join(surrounding_strings)

                ner_sottostringhe = riconosci_entita(testo)+
                riconosci_entita(testo.lower())
                all_ner.extend(list(set([x[0] for x in ner_sottostringhe if x[1]
                in [”MISC”,”LOC”,”ORG”]])))

                pos_sottostringhe=part_of_speech_tagging(testo)
                all_pos.extend(list(set([x[0] for x in pos_sottostringhe if x[1]
                in [”PROPN”]])))
    else:
        ner_sottostringhe=riconosci_entita(testo)+riconosci_entita(testo.lower())
        avoid_names.extend([x[0] for x in ner_sottostringhe if x[1]
        in [”MISC”,”LOC”,”ORG”]])
return all_ner, all_pos, avoid_names
```

Algorithm 9.8 Code of function cooperative_names(testo,ord_ner,ord_pos,ord_avo). It uses NER, POS and lists by 9.7 to discover the possible authors of the review based on the occurences of the entities in texts with and without the reviews.

```
#Code of function that for a text returns a the most probable name of the author of
#the revision using the number of occourences of the counters, otherwise ””
possible_names=[]
testo=sostituisci_abbreviazioni(testo)
if ” revisione ” in testo.lower():
    sottostringhe = re.split(r’[.;]’, testo)
    possible_names_substring=[]
    for i, sottostringa in enumerate(sottostringhe):
        if ” revisione ” in sottostringa.lower():
            index_before = max(0, i - 3)
            index_after = min(len(sottostringhe), i + 4)
            surrounding_strings=sottostringhe[index_before:index_after]
            testo=’ ’.join(surrounding_strings)
            ner_sottostringhe=riconosci_entita(testo)+riconosci_entita(testo.lower())
            possible_names_substring.extend(list(set([x[0] for x in ner_sottostringhe
            if x[1] in [”MISC”,”LOC”,”ORG”]])))
            pos_sottostringhe=part_of_speech_tagging(testo)
            possible_names_substring.extend(list(set([x[0] for x in pos_sottostringhe
            if x[1] in [”PROPN”]])))
    possible_names.extend(possible_names_substring)
else:
    return ””
coop_names=””
names=[]
#for each possible entity sum the number of its occurences for POS and NER and subtract
#the number of occurences for NER where ”revisione” does not appear
for word in set(possible_names):
    names.append((word,(ord_ner[word]+ord_pos[word]) -ord_avo[word]))
#return the entity with the highest score
coop_names=sorted(names, key=lambda tup: tup[1],reverse=True)[0][0]
return coop_names
```

References

[1] InfoCamere, “Infocamere,” https://www.infocamere.it/web/ic-home//profilo?

[2] Azure, “Data science,” https://azure.microsoft.com/en-us/resources/cloud-computing-dictionary/what-is-data-science.

[3] Oracle, “Machine learning,” https://www.oracle.com/uk/artificial-intelligence/machine-learning/what-is-machine-learning/.

[4] SAS, “Credit risk management,” https://www.sas.com/en_us/insights/risk-management/credit-risk-management.html.

[5] Stripe, “Fraudulent transactions,” https://stripe.com/it/resources/more/how-machine-learning-works-for-payment-fraud-detection-and-prevention.

[6] Forbes, “Customer segmentation,” https://www.forbes.com/advisor/business/customer-segmentation/.

[7] IBM, “Text mining,” https://www.ibm.com/topics/text-mining.

[8] T. Isbister, “Llm by tim isbister,” https://huggingface.co/timpalol/mdeberta-v3-base-squad2.

[9] F. Russo, “Llm by francesco russo,” https://huggingface.co/osiria/deberta-italian-question-answering.

[10] J. G. He, Pengcheng and W. Chen., “Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing.” *arXiv preprint arXiv:2111.09543*, 2021.

[11] Elastic, “Elasticsearch,” https://www.elastic.co/.

[12] ——, “Kibana,” https://www.elastic.co/kibana.

[13] XBRL, “Xbrl,” https://it.xbrl.org/.

[14] G. Ufficiale, “Gazzetta ufficiale art. 2423 bilancio di esercizi,” https://www.gazzettaufficiale.it/atto/serie_generale/caricaArticolo?art.progressivo=0&art.idArticolo=2423&art.versione=6&art.codiceRedazionale=042U0262&art.dataPubblicazioneGazzetta=1942-04-04&art.idGruppo=310&art.idSottoArticolo1=10&art.idSottoArticolo=1&art.flagTipoArticolo=2.

[15] R. delle Imprese, “Registro delle imprese bilancio di esercizio,” https://www.registroimprese.it/bilancio-d-esercizio.

[16] G. Ufficiale, “Gazzetta ufficiale art. 2424 stato patrimoniale,” https://www.gazzettaufficiale.it/atto/serie_generale/caricaArticolo?art.versione=7&art.idGruppo=310&art.flagTipoArticolo=2&art.codiceRedazionale=042U0262&art.idArticolo=2424&art.idSottoArticolo=1&art.idSottoArticolo1=10&art.dataPubblicazioneGazzetta=1942-04-04&art.progressivo=0.

[17] ——, “Gazzetta ufficiale art. 2425 conto economico,” https://www.gazzettaufficiale.it/atto/serie_generale/caricaArticolo?art.progressivo=0&art.idArticolo=2425&art.versione=5&art.codiceRedazionale=042U0262&art.dataPubblicazioneGazzetta=1942-04-04&art.idGruppo=310&art.idSottoArticolo1=10&art.idSottoArticolo=1&art.flagTipoArticolo=2.

[18] ——, “Gazzetta ufficiale art. 2425-ter rendiconto finanziario,” https://www.gazzettaufficiale.it/atto/serie_generale/caricaArticolo?art.versione=1&art.idGruppo=310&art.flagTipoArticolo=2&art.codiceRedazionale=042U0262&art.idArticolo=2425&art.idSottoArticolo=3&art.idSottoArticolo1=10&art.dataPubblicazioneGazzetta=1942-04-04&art.progressivo=0.

[19] ——, “Gazzetta ufficiale art. 2427 nota integrativa,” https://www.gazzettaufficiale.it/atto/serie_generale/caricaArticolo?art.progressivo=0&art.idArticolo=2427&art.versione=7&art.codiceRedazionale=042U0262&art.dataPubblicazioneGazzetta=1942-04-04&art.idGruppo=310&art.idSottoArticolo1=10&art.idSottoArticolo=1&art.flagTipoArticolo=2.

[20] json.org, “Json,” https://www.json.org/json-en.html.

[21] LearnDataSci.com, “Tf-idf,” https://www.learndatasci.com/glossary/tf-idf-term-frequency-inverse-document-frequency/.

[22] Microsoft, "Xml," https://support.microsoft.com/en-us/office/xml-for-the-uninitiated-a87d234d-4c2e-4409-9cbc-45e4eb857d44.

[23] R. delle Imprese, "Deposito bilanci," https://www.registroimprese.it/deposito-bilanci.

[24] G. Ufficiale, "Decreto xbrl," https://www.gazzettaufficiale.it/eli/id/2008/12/31/08A10127/sg%20.

[25] Adobe, "Pdf," https://www.adobe.com/acrobat/about-adobe-pdf.html.

[26] Wikipedia, "Html," https://en.wikipedia.org/wiki/HTML5.

[27] A. I. Digitale, "Standard xbrl per presentazione bilanci," https://www.agid.gov.it/it/dati/formati-aperti/xbrl-standard-formato-elettronico-editabile.

[28] G. Ufficiale, "Gazzetta ufficiale art. 2511 cooperative," https://www.gazzettaufficiale.it/atto/serie_generale/caricaArticolo?art.versione=3&art.idGruppo=332&art.flagTipoArticolo=2&art.codiceRedazionale=042U0262&art.idArticolo=2511&art.idSottoArticolo=1&art.idSottoArticolo1=10&art.dataPubblicazioneGazzetta=1942-04-04&art.progressivo=0.

[29] Spacy, "Spacy," https://spacy.io/.

[30] DataCamp, "Tokenization," https://www.datacamp.com/blog/what-is-tokenization.

[31] Stanford, "Stemming and lemmatization," https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html.

[32] Medium, "Part of speech tagging," https://towardsdatascience.com/part-of-speech-tagging-for-beginners-3a0754b2ebba.

[33] DataCamp, "Named entity recognition," https://www.datacamp.com/blog/what-is-named-entity-recognition-ner.

[34] Python, "Snowballstemmer," https://pypi.org/project/snowballstemmer/.

[35] A. W. Services, "Large language models," https://aws.amazon.com/what-is/large-language-model/?nc1=h_ls.

[36] ——, "Retrieval augmented generation," https://docs.aws.amazon.com/sagemaker/latest/dg/jumpstart-foundation-models-customize-rag.html.

[37] I. Hugging Face, “Hugging face,” https://huggingface.co/.

[38] e. a. Devlin, Jacob, “Bert: Pre-training of deep bidirectional transformers for language understanding.” *arXiv preprint arXiv:1810.04805*, 2018.

[39] H. Face, “Question answering,” https://huggingface.co/tasks/question-answering.

[40] e. a. Vaswani, Ashish, “Attention is all you need.” *Advances in neural information processing systems 30*, 2017.

[41] e. a. He, Pengcheng, “”deberta: Decoding-enhanced bert with disentangled attention.”,” *arXiv preprint arXiv:2006.03654*, 2020.

[42] e. a. Rajpurkar, Pranav, “Squad: 100,000+ questions for machine comprehension of text.” *arXiv preprint arXiv:1606.05250*, 2016.

Acknowledgments

Niccolò Stamboglis for the role of company supervisor
Professor Alessandro Sperduti for the role of supervisor
The Advanced Data Analysis Team of InfoCamere for the support