ALMA MATER STUDIORUM – UNIVERSITY OF BOLOGNA
CAMPUS OF CESENA

Department of Computer Science and Engineering
Master's degree in Computer Science and Engineering

A text mining approach to materiality assessment

Thesis in the subject of
Big Data

Supervisor
Dott. Enrico Gallinucci

Presented by
Marta Luffarelli

Academic Year 2019-2020

Abstract

Worldwide companies currently make a significant effort in performing the materiality analysis, whose aim is to explain corporate sustainability in an annual report. Materiality reflects what are the most important social, economic and environmental issues for a company and its stakeholders. Many studies and standards have been proposed to establish what are the main steps to follow to identify the specific topics to be included in a sustainability report. However, few existing quantitative and structured approaches help understanding how to deal with the identified topics and how to prioritise them to effectively show the most valuable ones. Moreover, the use of traditional approaches involves a long-lasting and complex procedure where a lot of people have to be reached and interviewed and several companies' reports have to be read to extrapolate the material topics to be discussed in the sustainability report. This dissertation aims to propose an automated mechanism to gather stakeholders and the company's opinions identifying relevant issues. To accomplish this purpose, text mining techniques are exploited to analyse textual documents written by either a stakeholder or the reporting company. It is then extracted a measure of how much a document deals with some defined topics. This kind of information is finally manipulated to prioritise topics based on how the author's opinion matters. The entire work is based upon a real case study in the domain of telecommunications.

"You think you own whatever land you land on,
the Earth is just a dead thing you can claim,
but I know every rock and tree and creature
has a life, has a spirit, has a name.
You think the only people who are people,
are the people who look and think like you
but if you walk the footsteps of a stranger
you'll learn things you never knew.."

Contents

List of Figures

List of Tables

Introduction

Climate change, global warming, human rights violations, gender equality and circular economy are only a few examples of the main discussed topics in the last few decades. Billions of people are involved and so there is a growing interest in how this kind of topics influence people's daily life and future. In this context, companies increasingly appear as entities detached from the reality, whose interests concern only mere economical profit and that pursue not very noble values. This is why worldwide firms are trying to overcome this decreasing trust by actively working to understand what are the third-parties requirements and needs. Indeed, they concentrate their investments to respond to both external and internal necessities, always taking care of a more sustainable future. *Materiality analysis* is the name given to the analysis process conducted by many organisations to understand what matters (i.e., what is material) for the company itself and how it can meet stakeholders interests. Stakeholders are all those actors involved in the company entourage, starting from employees, suppliers, investors to customers, governments, people. On the one side, materiality analysis helps the company conducting it in identifying the main sectors or topics on which they should focus and invest resources; on the other side, it helps stakeholders understand what is the economical, social, and environmental philosophy of the company and where it is going to concentrate its efforts in the next years. The result of the materiality analysis is a sustainability report where are detailed what are the highest-priority topics or sectors in which the company would invest and how this would be done. Since economic resources are limited, it is necessary to give a priority to each topic trying to understand what are the most material ones. An intuitive and very used way to present prioritised results is to organise them in a materiality matrix, that consists of a Cartesian plan where axes represent how much a topic is important to the company and how much it is important to stakeholders. Topics are placed in this matrix so that it is visually represented their materiality. Conducting a materiality analysis can be complex and burdensome under many aspects: it is not easy to identify people among the various stakeholders to ask for an opinion and, moreover, it is hard to give a differentiated weight to their opinions; once stakeholders have been identified, it is necessary to find an easy and effective way to obtain their opinions about well-defined topics; lastly, it is hard to aggregate different opinions to obtain a unique explanatory result useful for the company to understand where to address their investments. In the last years, guidelines have been proposed that help companies in the first phase of the analysis, where it is requested to understand who interrogate and what kind of topics should be asked about. Few practical approaches have been proposed for the second phase and so many companies, especially the smallest ones with less economic availability, encounter difficulties. To the best of our knowledge only three approaches [15], [4], [2] practically guide companies in gathering opinions and obtaining unique results. Anyway, all these approaches need long and boring interviews where stakeholders are asked to express an opinion for each of the identified topics.

The scope of this dissertation is to propose a more efficient and objective method for materiality analysis that simplifies and automates opinions gathering and formulation of results. More in details, this thesis aims to exploit text mining techniques to analyse textual documents searching for topics occurrences. The topics to be searched for are the typical materiality analysis topics retrieved from many well-defined standards in this sector, e.g. the Global Reporting Initiative (GRI) standard. The analysed documents are divided into i) documents to understand what matters to the company conducting the analysis, and ii) documents to understand what is material to the company's stakeholders. Data are manipulated and prioritised drawing inspiration from [4] and [2]. To be mentioned, these papers organised a concrete approach proposing some formulas useful to mathematically evaluate the opinion of each involved actor upon each defined topic. They also suggest to differently weight opinions based on the importance of the actor expressing them. However, they manually gather opinions, i.e. scores, from the involved stakeholders and this is the most valuable dissimilarity among our approaches. Indeed, the idea proposed in this thesis is to get stakeholders' scores by manipulating topics occurrences rescued from the text mining procedure. Once occurrences have been manipulated to represent a stakeholder's opinion about a certain topic, they are aggregated based on stakeholders importance and so the main score for each topic is computed. These scores are suitable for prioritising topics and organising them either in a materiality matrix or in a rank. Lastly, this thesis deals with a first-step evaluation of the obtained results that should be further explored.

The structure of this research is hereby presented:

Chapter 1 focuses on the *State of the Art* giving a theoretical background useful to have a comprehensive understanding of the problem and the proposed solution.

Chapter 2 presents the *Text Mining Approach* created to analyse unstructured text searching for topics occurrences.

Chapter 3 concerns the *Materiality Results Extraction* conducted, starting from the formalisation of the created approach, through data description and ending with some implementation notes.

Chapter 4 presents the *Evaluation* of the obtained results. It is focused on the data visualisation obtained and on performance tests.

Chapter 1

State of the Art

This chapter aims at providing an overview of sustainability reporting and materiality analysis development in the enterprises' environment. It is also useful to first get the basic knowledge to understand what are the main problems in addressing sustainability report and then to learn about solutions in the literature.

1.1 Materiality Analysis

Since the last decades of the 20th century, worldwide companies have started reporting about their corporate sustainability. Corporate sustainability, following Wilson explanation [26], can be viewed as a new and evolving way for corporations to continue planning and pursuing economic growth, but focusing more on their commitment into the societal sphere, specifically that relating to sustainable development, e.g. environmental safety, social justice and equity, and economic development. More in details, corporate sustainability is a concept that encompasses four other previously defined terms described below.

- *Sustainable development* was firstly used in 1987 in the "Our Common Future" [7] book published by the World Commission for Environment and Development (WCED) where they recognised the corporations' obligations to look after their impact, both in the present and future days, on society, environment and economy. Indeed, in the book the sustainable development is described as "a process of change in which the exploitation of resources, the direction of investments, the orientation of technological development, and institutional change are all in harmony and enhance both current and future potential to meet human needs and aspirations". Ultimately, the contribution of sustainable development to corporate sustainability is to underline the role of companies in society: they have work to reach an ecological, social and economic sustainability.

- *Corporate Social Responsibility (CSR)* was largely used since 1953 with the publication of the book “Social Responsibilities of the Businessman” [3] and it deals with the ethical responsibility that managers have in understanding and satisfying the needs of society. It can be considered a philosophical contribution to corporate sustainability to the extent that it explains *why* corporate managers should work toward sustainable development.

- *Stakeholder theory* was firstly provided in 1984 in the “Strategic Management: A stakeholder Approach” book [13] where the author defines a stakeholder as “any group or individual who can affect or is affected by the achievement of the organisation’s objectives”. Following this theory, the stronger are the relationships between an organisation and its stakeholders, the easier it will be for that company to reach its programmed business objectives. Therefore, when firms organise their investments, they have to consider and mediate among their stakeholders’ interests in order to meet both stakeholders and their own needs. This theory is more practical than philosophical and it helps firms understanding that most of the common goals across different stakeholders deal with economic stability, environmental safeguard and social equity. Stakeholder theory contributes to corporate sustainability in suggesting that it is in the company’s economic interest to work towards sustainability because it will strengthen its relationship with stakeholders.

- *Corporate accountability* is a term used since 1997 by John Elkington who also referred to it as *triple bottom line*[1] reporting, where the number refers to the three main topics of a sustainability report, namely social, economic and environmental. Generally, the word “accountability” deals with the legal or ethical responsibility to provide an explanation, a justification, and a report about the actions for which it is considered responsible. In the corporate world, there are many different accountability relationships, but the one relevant for this context is the relationship between an organisation and its stakeholders: a firm must inform stakeholders about its intentions and effective actions, dealing with both financial and non-financial topics. So corporate accountability contributes to corporate sustainability in describing why companies should report to society about their performance in the economic, social and environmental areas.

In the context of corporate sustainability, another term has grown up in importance since its first mention in 2011 [19]: Creating Shared Value (CSV). This term refers to the companies’ ability and intention to create new business opportunities

[1] The “bottom line” traditionally refers to the monetary profits that a company has made. The “triple bottom line” adds two more “bottom line”: social and environmental (ecological) concerns, around which develop the company’s reports

and new business value without being forced neither by pressure from outside nor by external factors [27]. Firstly, they have to think about what has an economic value for them and then, they will concentrate their efforts on the maximisation of its sustainability, with consequent business and social benefits. The concept of CSV emerged because many companies were focused on optimising short-term performance while underestimating both stakeholders needs and elements that could bring long-term success. Consequently, CSV has become a popular way to address the decline of social trust in firms and to connect society with the business world [22] and so it is necessary for firms to report and be accountable for what they do or they are going to do for society.

Sustainability report, i.e. an annual report firms draw up about their corporate sustainability involvement, is nowadays seen as a crucial element to explain firms' intentions and performance under the three aforementioned aspects: social, environmental and economic, often referred to as a *triple bottom line.* To compile such a report, *materiality analysis* has to be made, and both materiality analysis and sustainability report have a role to play in producing CSV because they can be considered as a tool for prioritising issues and strategic planning [22].

The term *material* was traditionally derived from 1867 when the English Court used it to refer to "a relevant, not negligible fact" that emerged in the judgements in a case concerning the Central Railways of Venezuela[2]. In the financial reporting environment, *materiality* is considered as a threshold for influencing economic decisions of those people, mainly investors, exploiting an organisation's report to make decisions [15]. Following the US SEC (Securities and Exchange Commission, 1999[3]) definition, an item should be considered *material* when "in the light of surrounding circumstances, the magnitude of the item is such that it is probable that the judgement of a reasonable person relying upon the report would have been changed or influenced by the inclusion or correction of the item". The most commonly accepted definition of what material means can be found in the Global Reporting Initiative (GRI) Guidelines, described in 1.2.1, stating that "materiality reflects an organisation's significant social, economic and environmental impacts, together with their influence on stakeholders' assessments and decisions. Thus, the concept of materiality for sustainability reporting is complex because it is concerned with a wider range of impacts and stakeholders"[4].

Following this meaning, the purpose of the materiality analysis is to evaluate what kind of information is the most important for companies and their stakeholders (e.g. employees, clients, pressure groups, communities, etc.) and to what

[2] Definition taken from `https://www.datamaran.com/materiality-definition/`

[3] `https://www.sec.gov/interps/account/sab99.htm`

[4] `https://www.globalreporting.org/`

extent [5]. Materiality analysis is useful to show investments and objectives of the reporting company, i.e. the companies in charge to draw up a sustainability report, but it gives them also a chance to evaluate risks and business opportunities. The ultimate purpose of this process is to create a simple and effective result able to summarise the main aspects that a company has to focus on. GRI proposes a *materiality matrix* that is a graphical representation of the identified aspects, it is widely described in section 1.2.1.

1.2 Standard

Every company uses a different approach to define what is material, mostly because reporting and sustainability practise are influenced by companies organisational characteristics, such as business models, size, social context, etc. However, some common elements could be found and should be taken into account. Indeed:

- every company has to identify a number of environmental, social and economic issues around which develop the sustainability report;

- once relevant topics are found, each company has to evaluate them by considering both their and stakeholders' concerns on each of the identified issues;

- finally, the reporting company must prioritise the analysed issues in order to inform stakeholders and society about its sustainability strategy.

It is clear that, despite local differences, there is the need for a standard from which start to develop a more structured approach. The following subsections (1.2.1, 1.2.2 and 1.2.3) contain some of the main standards developed during the last decades and widely recognised.

1.2.1 Global Reporting Initiative

The GRI is an international organisation born in Boston in 1997 by the collaboration of the Coalition for Environmentally Responsible EconomieS (CERES) and the Tellus Institute, a not-for-profit organisation that wants to promote the transition to a more sustainable future. Even though it was born to simplify accounting reporting for a limited group of investors, GRI suddenly became a widespread reference to anyone interested in sustainability reporting [16].

The main aspects of the GRI are presented below.

Continuous improvement The GRI launched its first reporting guidelines, GRI G1, in 2000. These guidelines were adapted from the original for reporting on economic, environmental, and social performance. The G2 guidelines, published in 2002, provide a significant advancement in rigour and quality with reference to the G1. G2 guidelines deal with a revised set of principles which included transparency, inclusiveness, auditability and clarity. In 2006, the GRI published the third generation (G3) of sustainability reporting guidelines that identified three sets of standard disclosures that organisations were encouraged to adopt in a flexible and incremental manner to facilitate transparency in the reporting process [16]. The GRI G4 guidelines, launched in 2013, offer a complete manual on how to standardise issues, risks and opportunities prioritisation, according to both company and stakeholders views [22].

In 2016 GRI published a series of documents that comprehend a reformulation of G4 Guidelines principles and provide new standards, namely GRI STANDARD, divided into [9]:

- *universal standards*, whose objective is to guide organisations through the reporting process and to provide them detailed guidance on how to use the other Standards. Universal Standards include:
  - *GRI 101: Foundation*, which represents the entry point for using the GRI sets of Standards. Indeed, it helps firms defining sustainability report content and quality. Moreover, it describes how the GRI Standards have to be used and referenced;
  - *GRI 102: General Disclosures*, which defines what kind of information a company should use to report about its sustainability reporting practises, e.g. organisational profile, strategy;
  - *GRI 103: Management approach*, which is used to report about how an organisation deals with a material topic. It is designed to be used also for the topics covered by the *topic-specific standards*, described below;
- *topic-specific standards*, whose objective is to inform organisations about social, environmental and economic disclosures in order to make them able to understand what they need to report for every topic. Topic-specific standards include *GRI 200: Economic Topics*, *GRI 300: Environmental Topics*, *GRI 400: Social Topics.*

The above-described standards, ultimately, contain all the useful *disclosures* about many different topics that a company should be able to provide in their sustainability report, depending on the materiality of those topics.

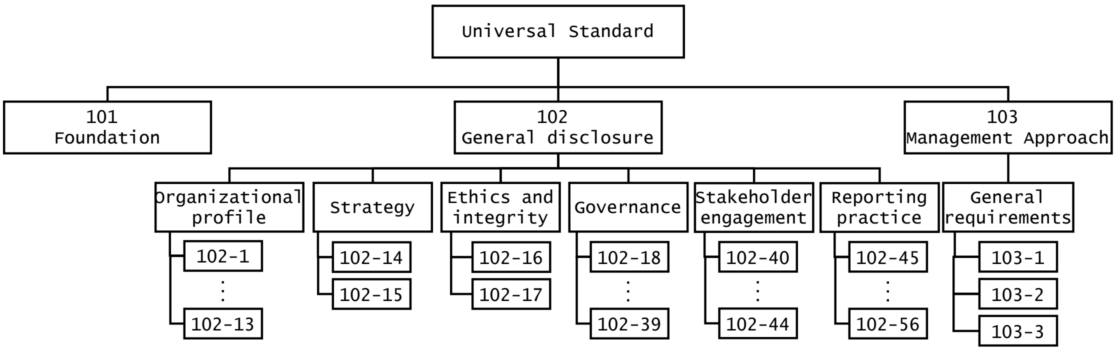


Figure 1.1: Hierarchy representation of the GRI Universal standard

The standards and their contents can be visually summarised in hierarchies. Elements at the same level of a hierarchy[5] can be mentioned with a specific word: every Standard, e.g. "101 Foundation" or "102 General disclosure", is a "category"; every child of a category is an "aspect" and every child of an aspect is an "indicator". Figure 1.1 illustrates a hierarchy representation of the *Universal Standard* (the root of the hierarchy). Due to space constraints, the last level of the hierarchy, the one of the indicators, shows the first part of the indicator's name and a node for the first and the last indicator only, if they are more than three. For example, in the "102 General disclosure" branch, following the "Organisational profile" node, we find only two indicators represented instead of the 13 defined for that aspect and their name is truncated: the complete name of the indicator labelled as "102-1" is "102-1 Name of the organisation"; the complete name of the indicator labelled as "102-13" is "102-13 Membership of associations". Figure 1.2, fig. 1.3 and fig. 1.4 represents the same structure for the *GRI 200: Economic Topics* standard, *GRI 300: Environmental Topics* standard and *GRI 400: Social Topics* standard respectively.

Stakeholders engagement stakeholders' opinion is a key component in writing an effective report. It is necessary, then, to understand who they are, to what extent they are important for the reporting company and how it is possible to understand their needs and then proceed to integrate their needs with those of the reporting company.

Materiality driven approach The GRI framework advises companies to:

1. identify triple bottom line aspects and topics (both internal and external) actively involving stakeholders;

[5] The complete list can be found at `https://www.globalreporting.org/standards/media/2594/gri-standard-glossary-2020.pdf`

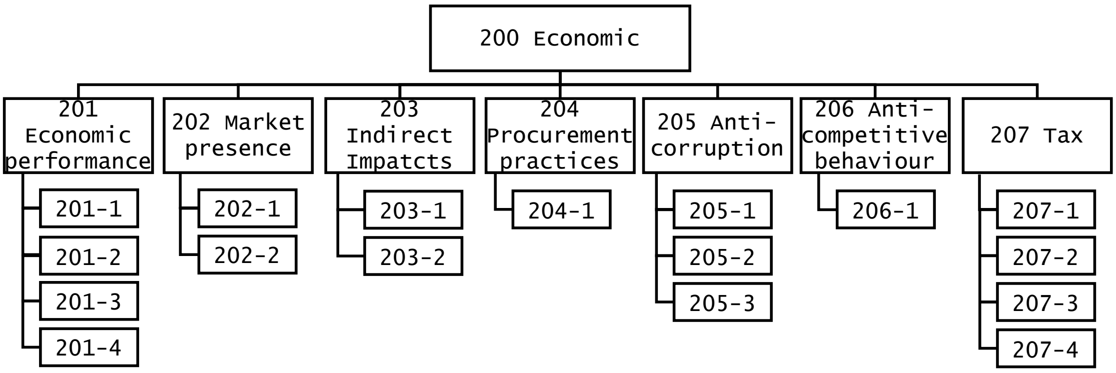


Figure 1.2: Hierarchy representation of the GRI Economic topics standard

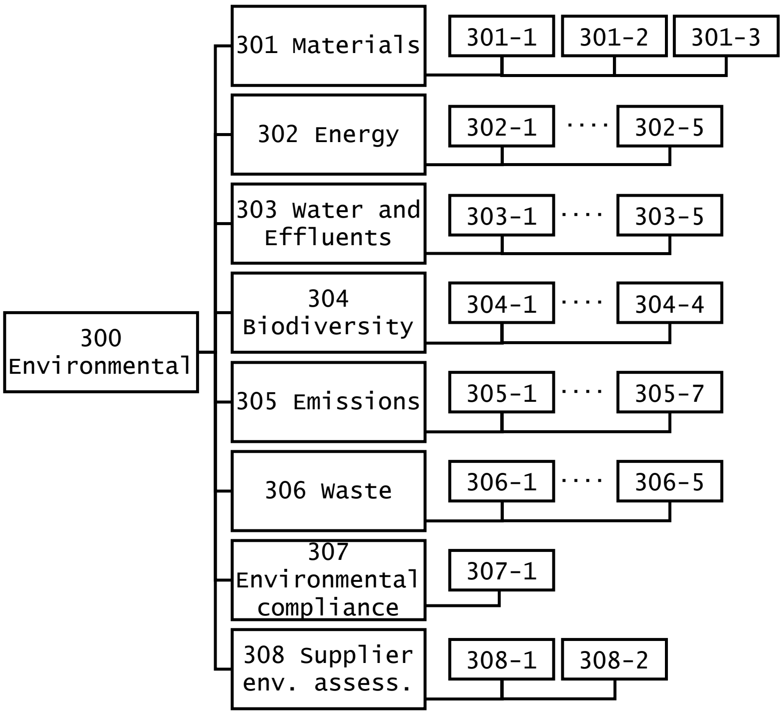


Figure 1.3: Hierarchy representation of the GRI Environmental topics standard

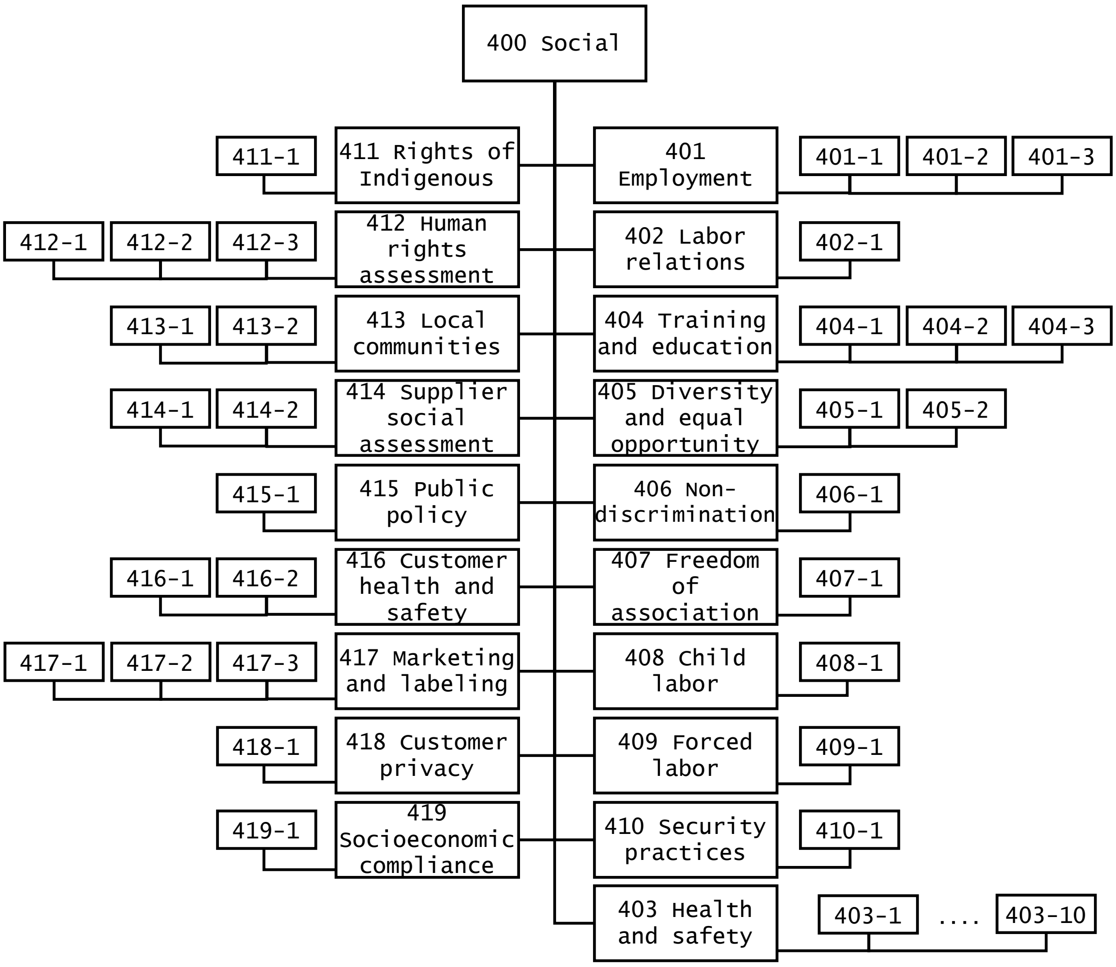


Figure 1.4: Hierarchy representation of the GRI Social topics standard

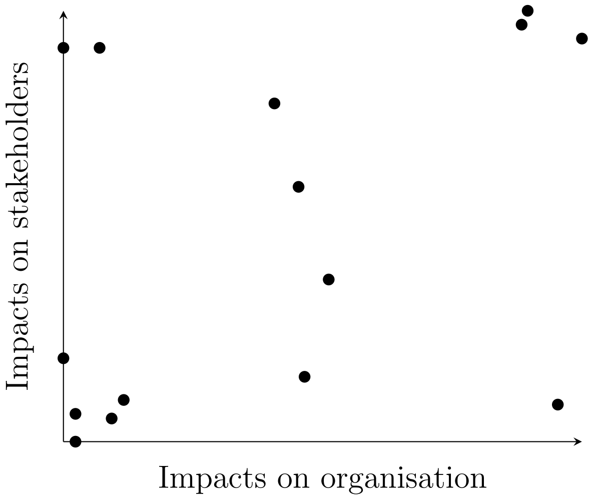


Figure 1.5: An example of a materiality matrix described by GRI

2. prioritise those aspects following materiality principles and implementing stakeholder inclusiveness. Material aspects should be visually represented by a materiality matrix;

3. validate results against scope, boundaries and time [22].

The materiality matrix is a visual representation of the most relevant issues rescued during the analysis process. It is important to find the right threshold at which these issues could be considered sufficiently important to be reported.

Different firms and different standards (not only GRI) could lead to different matrices, but approximately with the same meaning. This materiality matrix (fig. 1.5), generally consists of two axes: the horizontal axis usually identify what is material according to the company insights. Specifically, whatever placed on the right is more material than what is positioned on the left; the vertical axis, instead, represents what is material according to involved stakeholders. In detail, whatever is represented in the upper part is more material than what is in the lower part. Points in the material matrix represent issues retrieved in the materiality analysis conducted by the organisation. Material issues are those positioned in the right-upper part of the material matrix. Indeed, they resulted as material for both company and stakeholders.

1.2.2 AccountAbility 1000

AccountAbility, or Institute of Social and Ethical Accountability (ISEA), is an independent, global, not-for-profit organisation founded in 1995 in London and promoting accountability and sustainable business in accordance with corporate responsibility[6]. AccountAbility 1000 (AA1000) Series of Standards were developed by the ISEA in 1999. Their purpose is to guide organisations through the identification and prioritisation of sustainability issues in order to improve long-term performance. AA1000 Standards are founded upon four principles[7].

- *Inclusivity*: people should know what are the issues that would have an impact on their decisions and, moreover, they must have the chance to express their opinion;
- *Materiality*: decision-makers (people who organise sustainability reports) should identify what are the sustainability topics that matter;
- *Responsiveness*: companies should be clear on material issues and their related impacts;
- *Impact*: firms should monitor and be accountable for how their actions affect their entourage.

AA1000 Series of Standards is a set of principle-based standards aimed at helping organisations better understand what issues they should be concerned with and how they can measure their performance with these issues. It is composed of three Standards, namely:

- AA1000 AccountAbility Principles (APS), whose objective is to drive companies to look at issues as value drivers. This framework is the basis for the other two standards in the series;
- AA1000 Stakeholder Engagement Standard (SES), that explores in more details the inclusivity principle;
- AA1000 Assurance Standard (AS), that allows organisations to have their approach, reporting and performance verified.

AA1000 vs GRI AA1000 and GRI G4 Guidelines are two of the most used standards in the sustainability reporting environment and *materiality* is a core point for both of them. Even if they are not mutually exclusive, as stated by the director of AccountAbility AA1000 in [25], their focus is slightly different. The main distinction is that GRI offers a guide to understand what should be the main contents of responsibility reports, while AA1000 concentrates on outlining the main steps for stakeholders engagement and results verification.

[6] `https://en.wikipedia.org/wiki/AccountAbility#AA1000_Series_of_Standards`

[7] `https://www.accountability.org/standards/aa1000-accountability-principles/`

1.2.3 Other standards

Social responsibility reporting has been increasingly adopted by companies of any size and from every country and so the development of many standards, whose main objective is helping those organisations in reporting what really matters, has started. Follows an incomplete list of other important standards in the sustainability world:

- *ISO26000*, is an international standard launched in 2010 and providing guidelines for CSR[8];

- *DJSI*, launched in 1999, they evaluate the sustainability performance of thousands of companies. They are considered the longest-lived global sustainability benchmarks worldwide[9];

- *Integrated Reporting (IR)*, is a framework published in 2013 that put together information about organisational strategy and performance regarding the economic, social and environmental context in which the firm operates.

1.3 Practical Approaches

All the above-mentioned standards on sustainability reporting offer their own guidelines. However, none of them provides a structured approach to conduct the materiality analysis and to extrapolate a sort of materiality matrix. In fact, sustainability teams are free to operate with personal opinions, experiences and expectations.

That said, the major material issues identified by different companies vary significantly and without a fixed strategy. Following Morgan [18] ideas, "a major risk in non-financial reporting is that corporate managers publish only what they consider important information" or, even worse, they tend to spread and underline only what is convenient for the company reputation and not their real objectives and investments. stakeholders engagement is rarely conducted properly: since it is an expensive and long procedure, many firms (especially Small and Medium Enterprises (SMEs) with low budget reserved to sustainability reporting) barely involve the stakeholders.

The three works presented below ([15], [4], [2]) deal with the formulation of a structured approach and propose different solutions. They address the materiality analysis problem using different versions of a Multi-Criteria Decision Making

[8] `https://en.wikipedia.org/wiki/ISO_26000`

[9] `https://en.wikipedia.org/wiki/Dow_Jones_Sustainability_Indices`

(MCDM) method that can contribute to structure a transparent and reliable approach [6]. MCDM methods are decision support tools allowing decision-makers (in this case, those who perform the materiality analysis) to compare and rank different alternatives (i.e. sustainability issues) following one or multiple criteria. As an example in our everyday life, we generally make decisions about everything, evaluating different criteria, approximately weighting pros and cons for every criterion considered and then taking our decisions. For example, consider John Smith has to buy a car to get to work, but he also wants a dishwasher to help him fastening his house chores; he has a limited budget (cost criterion) and he cannot afford both of them, which should he choose and why? A car is much more expensive than a dishwasher, but he can go to work neither on foot nor by public transportation because it is too far (necessity criterion), while he can do dish on his own. So the car is the choice, but what car? And so on, evaluating more criteria.

When the problem is simple as the aforementioned, the decision is easy to come, but when more than one decision-maker has to find an agreement about a huge number of not always independent issues, considering a lot of criteria, it becomes much more complicated. Materiality analysis needs decision-makers to evaluate different issues identifying the most material ones and they have to rank them with respect to different criteria, both qualitative and quantitative. For example, in [17] are presented different issues in the renewable energy context and their related criteria by which extrapolate the significance of each issue, e.g. "to identify and prioritise the barriers existing in the developmental path of solar power in Indian perspective" it is necessary to estimate and evaluate: institutional barrier, technical barrier, political and regulatory barrier, market barrier, social-cultural and behavioural barrier, finance barrier and high cost of capital.

Hsu approach Hsu et al. [15] propose an assessment framework to identify what issues in sustainability reporting are material. They employ Failure Modes and Effects Analysis (FMEA) to establish what would be the evaluation criteria. FMEA is an approach mainly used in aerospace, automotive and electronic industries to identify, prioritise and eventually eliminate potential failures in their products at the design phase. An easy way to measure risk and severity of the identified failures is to adopt Risk Priority Numbers (RPNs), an index obtained by multiplying three indicators, namely *occurrence* (O), *detection* (D) and *severity* (S), where O is the probability of the failure; D is the probability of not detecting the failure and S is the severity of the failure. The approach described is divided into three phases as described below.

1. In the first phase three managers coming from human resources, public relations and social responsibility sectors, identify and formulate the three criteria of *detection, occurrence* and *severity* in the following way:

  - *Occurrence*, i.e. the probability of the failure, is seen as the percentage of concerned stakeholders. Indeed, the higher is the number of involved stakeholders the higher will be the probability of failure for information disclosure in the report: it is not easy to respond to a large number of stakeholders' needs correctly;
  - *Severity*, i.e. the severity of the failure, corresponds to the influence of an issue on strategic engagement objective[10] because, when information disclosure in sustainability reporting does not accomplishstakeholderneeds, strategic engagement objectives are not fulfilled. A significantly high influence of issues on strategic engagement objectives will result in serious effects. Its value is estimated by the reporting company's internal members (ten members of the sustainability reporting committee);
  - *Detection*, i.e. the probability of not detecting the failure, derives from the level of stakeholders' concern for a specific issue. The authors state that detection probability is higher when stakeholders are very concerned about an issue.

2. The second phase determines the relative importance, i.e. a weight, of each criterion to each of the three managers. This is achieved by exploiting the Analytic Network Process (ANP) [21], an MCDM method. It is a more general form of the Analytic Hierarchy Process (AHP), an MCDM technique defined by Saaty in 1980 [20], that structures a decision problem as a hierarchy with a specific goal to reach, deciding among different decision criteria and alternatives. The ANP, instead, structures the problem as a network. Both of these approaches use a pair-wise comparison procedure to measure the weights of the alternatives in the structure and to finally rank them to provide a choice. The main difference between these two approaches is that in the AHP elements in the hierarchy (decision criteria and alternatives) are considered as *independent* from the others, while in the ANP it is not a requirement. For example, in the previous example of the car choice, the AHP considers cost, colour and autonomy criteria as independent from each other, while ANP can consider their interdependencies giving a weight to each different criterion. In the case of this paper, the ANP fits perfectly the authors' needs and intentions.

3. The third phase illustrates a concrete example conducted in a company in Taiwan.

[10] "Strategic Engagement is a method of finding, attracting, and keeping the best customers for your business or organisation. Strategic Engagement utilises science and technology combined with creativity and psychology to achieve efficient and sustainable results" `https://musemarketinggroup.ca/strategy/what-is-strategic-engagement/`

To be more accurate, there are now detailed the steps drawn in fig. 1.6:

- the three internal managers rate each of the three criteria to find the most significant one. Their rates are aggregated and, in the end, every criterion has its own weight. In their study case: *occurrence* weights 0.071, *severity* weights 0.220 and *detection* weights 0.708;
- the three internal managers consider stakeholders which are more influenced by the company engagement. In their study case, seven categories of stakeholders are found, i.e. employees, customers, community, investors, suppliers, non-government organisations and media;
- the three internal managers generate a list of 23 relevant issues to be ranked. They choose among sustainability former reports of the company, internal and external sources of the DJSI questionnaire and also GRI guidelines;
- the three internal managers analyse both the number of participants and their opinions to compute values for the three criteria for each issue. To get the stakeholders' opinions, it is created and distributed a questionnaire where it is asked stakeholders to give a rank from 2 (no concern) to 10 (extreme concern) for each issue based on their interest in that issue;
- the three internal managers compute a score for each issue by summing the products of the values of three criteria for their related weights. To give an example extracted from [15]: the occurrence index shows that 326 stakeholders (78.7% of the total number of stakeholders) are concerned with the issue of corporate governance. So, following their tables, it has a score of 8, which is considered high. The mean of 326 stakeholders concerned with the issue of corporate governance (detection criterion) is 6.28. The issue of corporate governance, with respect to the severity index, leads to a mean score of 6.74. The materiality issue RPN for corporate governance is $8 \times 0.071 + 6.28 \times 0.708 + 6.74 \times 0.22 = 6.496$.

Calabrese approach Calabrese et al. [4] propose a hierarchically structured approach to prioritise the issues retrieved in the GRI G4 hierarchy and to consequently organise them in a rank. The hierarchy of the GRI is divided into three levels, namely categories, aspects and indicators in such a way that each indicator refers to a single aspect which in turn refers to a single category. To fulfil their intention to rank issues, authors exploit an AHP method, detailed in the previous paragraph, that is used to overcome decisions problems like the one of the materiality analysis where both qualitative and quantitative criteria have to be evaluated. AHP is integrated with fuzzy logic in order to address the linguistic

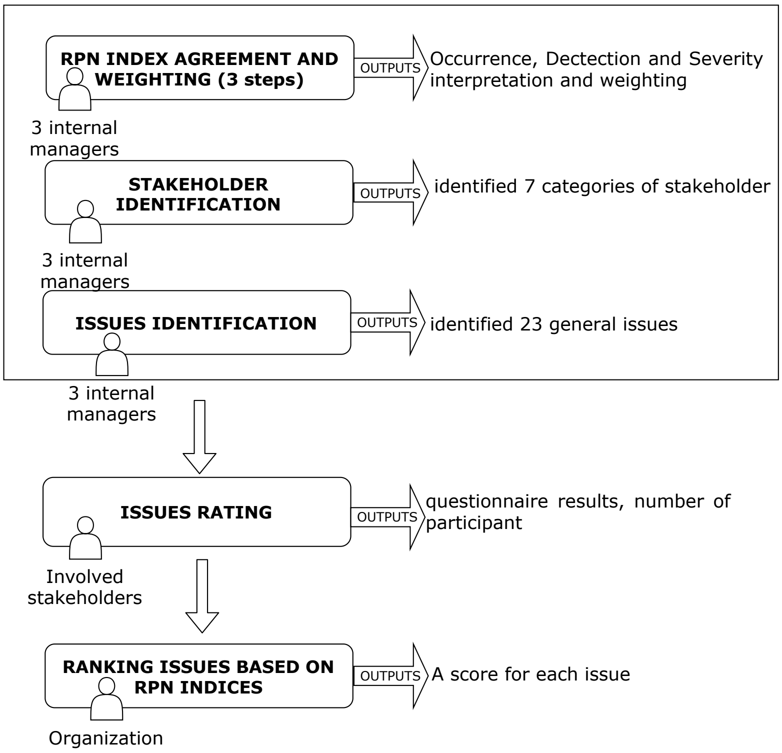

RPN INDEX AGREEMENT AND WEIGHTING (3 steps)
OUTPUTS
Occurrence, Dectection and Severity interpretation and weighting
3 internal managers
STAKEHOLDER IDENTIFICATION
OUTPUTS
identified 7 categories of stakeholder
3 internal managers
ISSUES IDENTIFICATION
OUTPUTS
identified 23 general issues
3 internal managers
ISSUES RATING
OUTPUTS
questionnaire results, number of participant
Involved stakeholders
RANKING ISSUES BASED ON RPN INDICES
OUTPUTS
A score for each issue
Organization

Figure 1.6: Schematisation of the approach described by Hsu et al.

judgements used to vote for each issue and that has to be converted to a number. To effectively integrate fuzzy logic in the AHP, fuzzy numbers, and in particular Triangular Fuzzy Numbers (TFNs), are used. Fuzzy numbers were introduced by Zadeh in 1975 to create a link between numbers and linguistic labels which are variables whose values are words or sentences, e.g. *age* is a linguistic label if its values are something like (*young, quite young, old*, rather than (16, 27, 70) [29]. Therefore, fuzzy numbers deal with questions like "How concretely old is a man defined as *not so old*?". Moreover, for every kind of fuzzy number, there is a set of well-defined operations that allow to manipulate and combine them. TFNs are the most popular kind of fuzzy number (other examples are trapezoidal or Gaussian) and they can generally be denoted by three values: TFN = (L, M, U), where the M indicates the modal or median value, L stands for the lower value and U indicates the upper value [1].

In the context of this paper, these numbers are exploited to convert the linguistic labels, used to express the importance of an issue with respect to another one of the same level in the hierarchy, into a mathematical form. To be more accurate, there are now detailed the steps drawn in fig. 1.7.

- Authors make the company and stakeholders vote for each issue with respect to the others at the same level in the hierarchy using a set of verbal labels. Each label is then converted to a specific value represented as a TFN. They obtain a different comparison matrix for every aspect in a category and indicator in an aspect;

- if many stakeholders are involved, their opinions, i.e. different comparison matrices, are mediated;

- starting from the rates expressed in the form of a TFN, they compute a single rate (crisp) for each issue exploiting the centroid defuzzification method called *centre of gravity* [28] which consists of arithmetic mean among the three values of each TFN in the matrix;

- the consistency of the matrices is evaluated and, if a matrix is not consistent, the approach is reiterated;

- the *local score* of each criterion, sub-criterion and alternative part of the hierarchical structures, i.e. categories, aspects and indicators, is determined summing up elements in each row of a crisp matrix and then normalising that sum with respect to the sum of all rows sum of that matrix. The *local score* represents how an issue is important with respect to the others at the same level in the hierarchy;

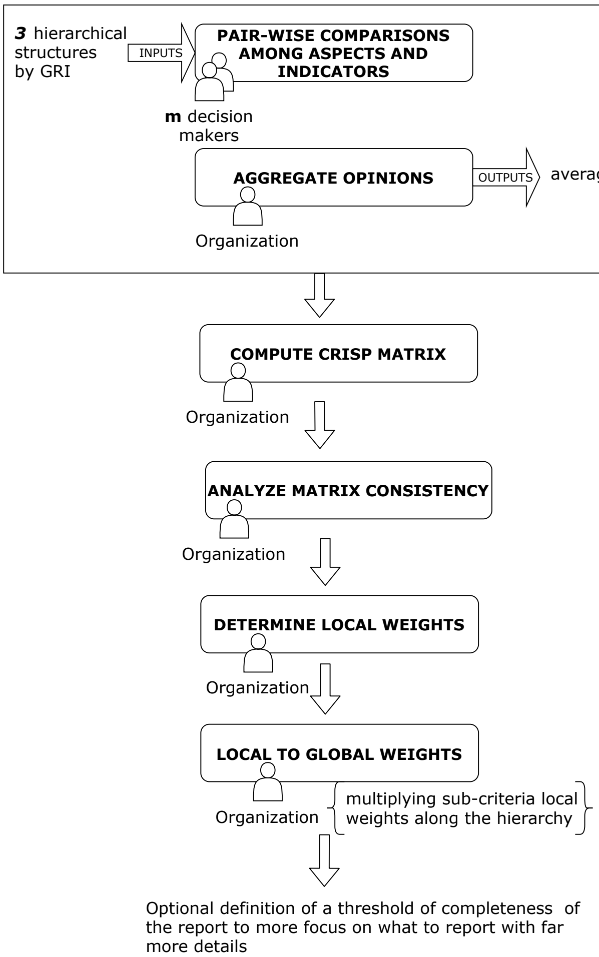

3 hierarchical structures by GRI
INPUTS
PAIR-WISE COMPARISONS AMONG ASPECTS AND INDICATORS
m decision makers
AGGREGATE OPINIONS
OUTPUTS
average of different views
Organization
COMPUTE CRISP MATRIX
Organization
ANALYZE MATRIX CONSISTENCY
Organization
DETERMINE LOCAL WEIGHTS
Organization
LOCAL TO GLOBAL WEIGHTS
Organization
multiplying sub-criteria local weights along the hierarchy
Optional definition of a threshold of completeness of the report to more focus on what to report with far more details

Figure 1.7: Schematisation of the approach described by Calabrese et al.

- the *global score* of each issue is computed by multiplying its local score with the local score of its father in the hierarchy. The *global score* represents how an issue is important with respect to the entire hierarchy considered;

- finally, the paper proposes a method to define a threshold which helps companies in determining the completeness that their sustainability report has to reach. Indeed, especially for SMEs, it is not easy and too expensive to address a complete report and so, with this approach, they can express at what level they are accurate.

Bellantuono approach Bellantuono et al. [2] aim to develop a quantitative structured approach useful to conduct a sustainability report and to organise material issues in a materiality matrix. Their approach is based upon Multi-Attribute Group Decision (MAGDM) techniques to effectively support stakeholder engagement and deal with multi-criteria and multi-decision makers problem. The MAGDM technique, part of the MCDM methods, is used in this paper because, when evaluating criteria and alternatives, it takes into account the different opinions of many independent decision-makers, each having his own importance.

Firstly, authors in the suggest a method to evaluate who are the stakeholders to be involved and what is their importance (i.e. salience). Then, the company is asked to identify some material issues, i.e. it is suggested to take them from GRI G4. Finally, both stakeholders and the company give a score to each issue which is then plotted in a materiality matrix. Specifically, the authors focus on the importance of stakeholders engagement and prioritisation. To better perform in the sustainability reporting, it is necessary to understand who has to be consulted, for which issues and how much it is important his voice over the others. It is stressed that the same stakeholder might have a high level of salience with respect to a specific issue, whereas he should not be as important for another topic. Therefore, their approach can consider a different value of salience for a certain stakeholder according to the type of the issue considered. To be more accurate, there are now detailed the six steps drawn in fig. 1.8.

- The reporting company needs to identify all the aspects that have to be analysed in the materiality analysis. In their approach, the authors suggest the GRI guidelines to derive those issues;

- the organisation has to identify its stakeholders and prioritise them. Indeed, stakeholders could be differentiated based on their importance, or salience, which may vary with respect to the considered sustainability aspect. They suggest providing a different salience to eachstakeholderbased on three main categories, namely social, economic and environmental;

- both stakeholders and organisation have to agree upon the linguistic labels that they will use when voting for an issue. They have to identify: i) the *labels*, e.g. (excellent, very good, good, fair, poor) or (crucial, substantial, valuable, marginal, negligible) and ii) the *relative importance of each label compared to another*, i.e. each of them has to establish the numerical distance between the chosen labels, or in other words, their weights. To extrapolate these weights, each involved person has to express a score from 1 (less important) to 9 (extremely more important) for each label with respect to the others of the set;

- both stakeholders and organisation give a rate to each issue by choosing a label in the agreed set. These rates, for each voter, are then converted into

Criteria	*Reference*		
	Hsu et al. (2013)	*Calabrese et al. (2015)*	*Bellantuono et al. (2016)*
MCDM method	ANP + FMEA	Fuzzy AHP	MAGDM
Sustainability issues considered	DJSI, EICC, GRI	GRI	GRI
Issues hierarchy considered	no	yes	no
Weighted mean of stakeholders' views	no	no	yes
Prior agreement on rating labels	no	no	yes
Modality of issues rating	direct	pair-wise comparisons	pair-wise comparisons
Output	ranking	ranking	matrix

Table 1.1: Comparisons among three different approaches

quantitative values picking up the label related weight. At this step, we have the importance of each issue for the company, i.e. the x-axis value, and as many values for each issue as the number of stakeholders involved, i.e. a set of y-axis values;

- since there are multiple stakeholders, their opinion about each issue has to be unified and this is achieved by multiplying their scores for a specific issue with their salience;
- issues are ready to be drawn in the materiality matrix, i.e. they have both coordinates computed. The authors of this paper, in addition, identify a method to filter among values that do not have to be considered material. Indeed, they define a threshold and compute the Euclidean norm of each aspect: all those aspects with values under the identified threshold are not important and there is no need to widely detail them in the company's sustainability report.

These three papers share some similarities and differ in a few details; table 1.1 sums up a comparison among the three works.

Valuable notes are:

- strict use of a single standard in [4] and [2] and not in [15];
- structured multi-stakeholders engagement and involvement in [2] and not in the others;
- both [2] and [4] aims to develop an approach to simplify materiality reporting for SMEs.

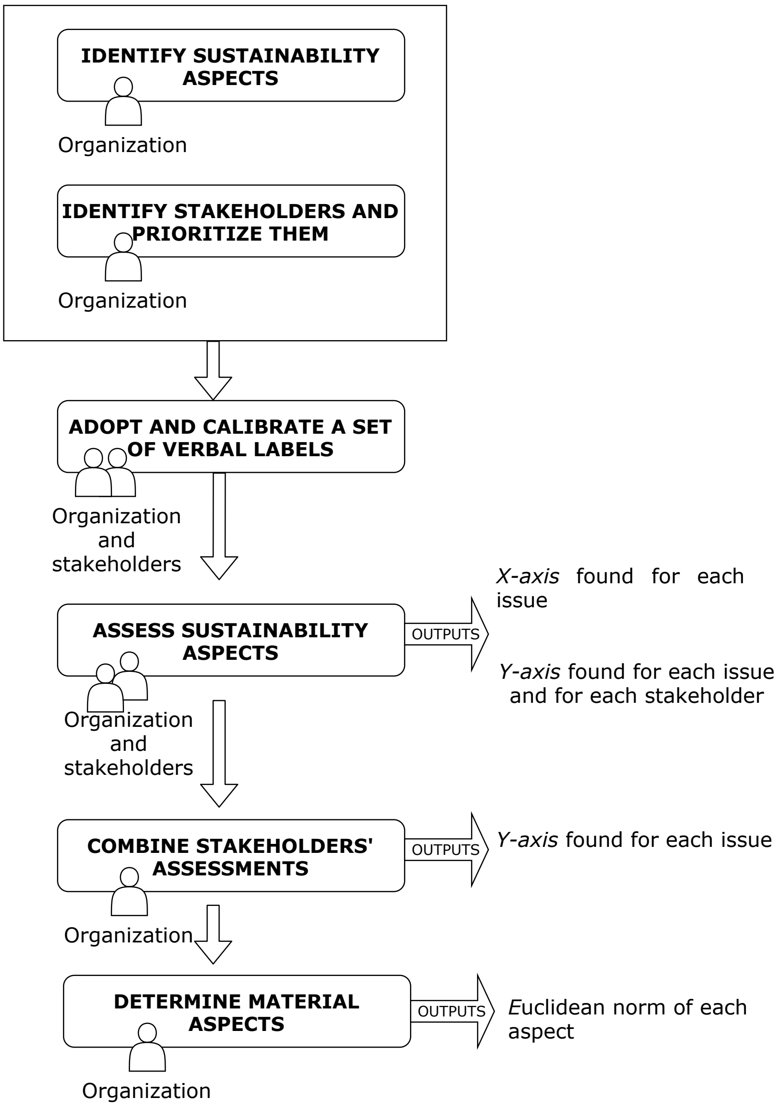

IDENTIFY SUSTAINABILITY ASPECTS
Organization
IDENTIFY STAKEHOLDERS AND PRIORITIZE THEM
Organization
ADOPT AND CALIBRATE A SET OF VERBAL LABELS
Organization and stakeholders
ASSESS SUSTAINABILITY ASPECTS
Organization and stakeholders
OUTPUTS
X-axis found for each issue
Y-axis found for each issue and for each stakeholder
COMBINE STAKEHOLDERS' ASSESSMENTS
Organization
OUTPUTS
Y-axis found for each issue
DETERMINE MATERIAL ASPECTS
Organization
OUTPUTS
Euclidean norm of each aspect

Figure 1.8: Schematisation of the approach described by Bellantuono et al.

Chapter 2

Text Mining Approach

Chapter 1 explains what are the main problems to be faced when conducting a materiality analysis. One of the most important difficulty concerns the necessity to involve a lot of people to define how much a sustainability topic matters to each of them. A promising approach is to analyse reports written by all the involved actors in order to extrapolate the needed opinions from them. This chapter describes in detail the text mining approach used in this thesis to analyse unstructured data from various documents to understand whether or not these documents deal with certain sustainability topics and to what extent. It is first described what is the problem and why a text mining approach is necessary to overcome it; then, it is described the idea of the main procedure to execute and finally, it is presented a short overview of the implementation.

2.1 Context

Since the first Sumerian writing system was born, there has been the necessity to understand and manipulate text in order to better summarise its semantic. Humans have always performed this job and, since their childhood, they are involved in tasks like: text summarisation, personal interpretation and re-elaboration of a text, and so on. Nowadays, however, we are flooded by a lot of enormous documents and so our task has become harder, or even impossible but, thanks to computers and new technologies we are now able to analyse a text and extrapolate semantic from it in an automatic and accurate fashion, deleting all the unnecessary information. Considering the computer science environment, words in a written text or a speech are referred to as *unstructured data* and so, before a computer can effectively manage them, it is necessary to make them somehow *structured.* This is why *text mining* approaches have been developed. Text mining is an Artificial Intelligence discipline that can be defined as a “knowledge-intensive process in which a user interacts with a document by using a suite of analysis tools. It seeks to extract useful information from data sources through the identification and exploration of interesting patterns" [10]. Its objective is definitely to stress facts, relationships and assertions that can be further analysed both from humans and computers. Text mining discipline can include different tasks such as classification, text clustering, document summarisation, sentiment analysis and, once text is analysed, results can be presented either in a graphical form or stored into databases.

2.2 Approach Description

The problem analysed in this chapter is to deal with a lot of natural language documents of various kinds (e.g. business report, legal memorandum, e-mail, social media posts, articles) and of various dimension. More in details, it is asked to understand whether or not a document deals with some topics and how much these topics are mentioned within it, specifically how many times a topic *occurs* in a document. This kind of analysis is performed in the context of the materiality analysis and the development of a sustainability report and its aim is to evaluate documents (written by stakeholders or the reporting company) and their content to find matches against sustainability topics. Indeed, to perform a materiality analysis it is necessary to gather opinions from different *actors* (i.e. the reporting company and its stakeholders). This is traditionally achieved through interviews and questionnaires with structured questions given to a lot of involved actors, but this is a tedious and expensive procedure that is often non-exploitable by many SMEs. A more powerful approach consists of inferring the required information through the analysis of the various previously written reports of the various actors. This approach can be automatised and managed by exploiting text mining techniques able to scan unstructured documents searching for sustainability topic mentioning. The proposed approach draws inspiration from techniques adopted in the Social Business Intelligence (SBI) environment [11, 12, 14]. Even if this kind of procedures operates in a different context, its main objective fulfil a similar goal, i.e. to identify relevant topics discussed over social networks and the Web by analysing the user-generated content. To pursue the above-mentioned methodology is necessary to:

1. define what are the sustainability topics to search for. The topics are defined in collaboration with domain specialists;

2. define the research strategy taking into account all the linguistic variation of the same topic, i.e. feminine/masculine, singular/plural, synonymous and more complex sentences. This chapter deals with the definition of a smart

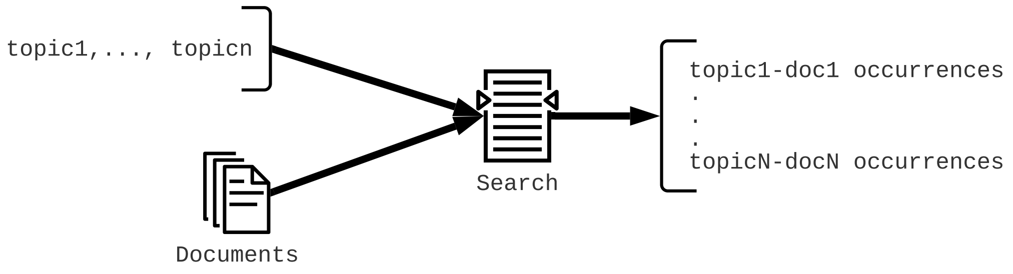


Figure 2.1: Sketch of the text mining approach of this thesis

research strategy based upon wildcards and it is described below in this section.

2.2.1 Modelling Alias to Recognise Topics

This thesis work operates on a previous-defined solution in order to improve its effectiveness. The existing method firstly defines what are the main topics that the procedure has to search for and then develops a procedure to analyse a plethora of documents, either provided by a reporting company or rescued from the Internet, searching for the defined topics. Figure 2.1 shows an informal representation of the idea behind this approach.

It is important to notice that the target topics are retrieved among the different materiality analysis standards described in chapter 1 and so, for example, considering the GRI standard some of the chosen topics are "social" or "indirect economic impacts", i.e. every topic described in the GRI hierarchy. Moreover, it is evident that searching only for the specific topics would give partial results because some topics may occur even if the specific words are not written: think about the "environmental" topic, even if this specific word is not found in the document, the latter could ultimately deal with "environmental" when other words or phrases like "aquifer pollution" or "deforestation" are present. Finally, there is the necessity to define a sort of semantic synonym which aims is to help in recognising the main topics. A semantic synonym is defined as an *alias* that can be either a single word or a sentence and can be linked to more than one topic. It is possible to link an alias with more than one topic because topics from different standards may represent the same concept and so an alias can refer to both of them, e.g. the alias "air pollution" can be linked to both "environmental" in the GRI and "Environmental Dimension" in the Dow Jones Sustainability Indices (DJSI). Some topics need context-aware aliases and so aliases have to be defined according to a domain expert's indications and under his control. Furthermore, to avoid a wasteful listing process of a lot of similar aliases wildcards are used to match more than one word with a single alias. In software, a wildcard character is a kind of placeholder represented by either a single or multiple special characters, such as an asterisk, that can substitute the full word that need not to be typed[1]. Three kinds of wildcards are defined in the approach, the first two wildcards are used to refer to characters in a word, whereas the last one refers to entire words.

- *asterisk (*)* can be interpreted as a number of literal characters or an empty string. For example, the alias “doc*” matches both the word “doc” and “document” but not “dodo”;

- *question mark (?)* can be interpreted as 0 or 1 literal character. For example, the alias “digit?” matches both “digit” and “digits” but not “digital”;

- *underscore and a number (_n)* indicates the presence of 0 or n words at most. For example, the alias “conflict _2 interest” matches both “conflict interest”, “conflict of interest” and “conflict of international interest”.

Every above-presented wildcard can be used in combination with another one and so, as an example, we can have an alias like “conserv* _3 ocean?” that matches sentences like “conservation of the oceans” or “conserving biodiversity in oceans” or “conserv ocean”.

This approach simplifies and accelerates the redaction of the needed aliases.

2.2.2 Conceptual View of the Approach

The proposed approach provides the reification of the previous-mentioned concepts, such as documents to analyse or topics/alias to search for and expands them to develop an effective solution. The main objective is to understand and synthesise what topics a document deals with by searching for aliases. To reach this purpose, it is necessary a preprocessing phase where the input document is analysed and deprived of useless words, called stopwords, such as articles or prepositions and then it is transformed reducing its words to their base form, i.e. it is performed a stemming. So, for example, a document containing this sentence:

“The big values of your original actions will be recognised by the whole”

is reduced to

“big valu your origin action will be recogn whole”

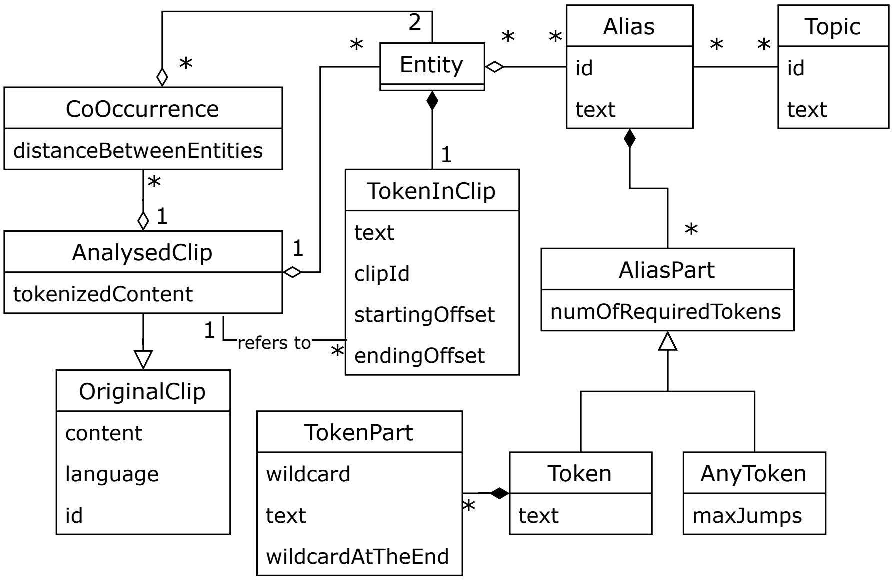


Figure 2.2: Class diagram

This phase facilitates the successive step that consists in the research of topics. The high-level solution is presented in the diagram class in fig. 2.2 where we find different related concepts that can be divided into two groups:

- the first group deals with concepts used before or during the text mining algorithm. More in details, there are concepts to achieve a topics/alias understanding as well as the representation of the documents to analyse and contains `Topic`, `Alias AliasPart`, `Token`, `AnyToken`, `TokenPart` and `OriginalClip`;

- the second group deals with concepts representing the result of the mining procedure. It contains `Entity`, `CoOccurrence`, `TokenInClip` and `AnalysedClip`.

These concepts and their relations are discussed below.

Topic This concept represents the reification of the topic discussed in the previous section. A topic is made of text e.g. the topic "environmental", and it has an id to uniquely identify it. Many aliases could refer to a specific topic and, vice versa, a topic could be described by different aliases.

[1] https://en.wikipedia.org/wiki/Wildcard_character

Alias This is the main concept. An alias can be made of simple text or a mix of words and wildcards so the idea is to see it as a composition of different parts (`AliasPart`) which in turn can be i) a `Token` containing simple text or those wildcards referring to characters (i.e. asterisks or question mark) or ii) `AnyToken`, containing only the wildcard referring to words. So, in this context, *token* is synonymous with *word*. Moreover, a `Token` can be further split into different parts, `TokenPart`, dividing the possible wildcards in order to analyse them separately. To have a match, every `AliasPart` of an `Alias` needs a corresponding counterpart in the given text and so an `AliasPart` indicates how many tokens/words are necessary to have a match. The number of required tokens corresponds to 1 when the `AliasPart` is a `Token` and so it has only asterisks and question marks wildcards or no wildcards, otherwise it corresponds to an interval that goes from 0 to the number specified in the wildcard. To give an example, matching a `Token` like "parrot?" needs one and only one word to be checked; whilst matching an `AnyToken` like _2 requires at most two words to be checked.

OriginalClip This is the representation of the document (also called clip) to be analysed in the text mining procedure. Its importance is mainly due to its content that is firstly manipulated in order to extract structured data upon which match the given aliases.

TokenInClip The first phase of the approach establishes that the content of a document has to be pre-processed in order to both delete stopwords and perform a stemming. The result of this phase consists of a set of more structured tokens, i.e. `TokenInClip`, that will be used to check whether an alias (and consequently its related topics) has a match or not.

Entity Every `TokenInClip`, either grouped with others or alone depending on the alias structure, is checked against every `Alias` and the result is saved in an `Entity` that declares whether a `TopicInClip`, or a group of them, has a matching alias or not.

CoOccurrence For some kinds of problem it can be useful and meaningful to analyse not only what kind of topics a document deals with, but also what is the context within which a certain topic is mentioned. For example, the firm *A* operates in the field of habitat preservation and fauna protection. This firm wants to know whether one of its stakeholders, e.g. B, is also interested or not and so analyse the B's documents. During the analysis it is discovered at some point that a document deals with "fauna preservation" and this can indicate that for stakeholder B the animals' protection is somehow important. But, by the successive five words, the document contains a sentence like “hunting allowed in summer” and this makes the firm understand that B does not care about animals life. This is why the concept of the co-occurrence is defined. It concerns two entities both present in the same document within a maximum distance.

AnalysedClip This concept is to represent a document after its analysis and so decorated with other information like the entities and the co-occurrences rescued and also the content after the tokenisation process.

2.3 Main Steps of the Approach

The solution proposed deals with different kind of entities and follows precise steps to achieve the desired result. The algorithm is presented in fig. 2.3, fig. 2.4 and fig. 2.5 that synthesise the main phases detailed below.

2.3.1 High-level Steps

Figure 2.3 describes the entire approach from the initial steps to the final ones. To start the analysis it is necessary to provide a text in a certain language and language-aware aliases to be matched against it. Then, there is the first phase where the provided text is analysed in order to remove stopwords and where the remained words are stemmed. To perform this step it is necessary to have a tool able to understand language-specific stopwords and rules for stemming that are different among different languages. Once the tokenisation phase is completed, it is time to search for aliases occurrences and, ultimately, for co-occurrences.

2.3.2 Alias Research Steps

The task performing the alias research is further expanded in fig. 2.4. The idea is to analyse token by token (words coming from the tokenisation phase) in order to find whether it can match an alias or not. So, taken a token, taken an alias, are also taken as many consecutive tokens (starting from the one already taken) as those required by the alias and so it is performed a check to find a match between the alias and the token(s) considered (this checking is detailed in section 2.3.3). Even if there is a match, the procedure is repeated for the next alias in the list because there is the possibility that a token would match more than one alias and then the more specific one will have to be considered as the valid one. For example, if a token matches both “rights” and “rights of the workers” the latter is more specific and then maintained. Once a token is checked against all the aliases, it has to be saved as an entity that can be either related to an alias if the checking

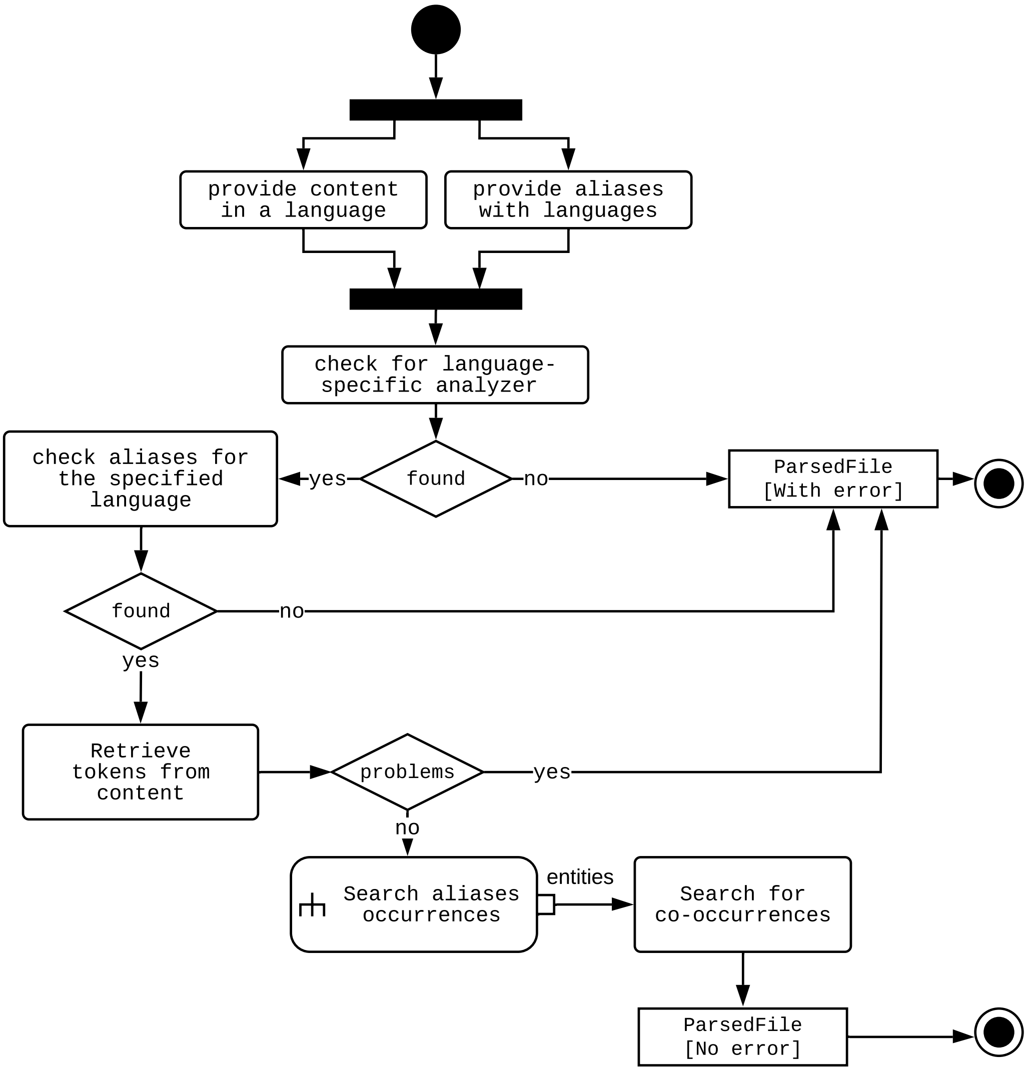

provide content in a language
provide aliases with languages
check for language-specific analyzer
found
yes
no
check aliases for the specified language
ParsedFile [With error]
found
no
yes
Retrieve tokens from content
problems
yes
no
Search aliases occurrences
entities
Search for co-occurrences
ParsedFile [No error]

Figure 2.3: Activity diagram complete

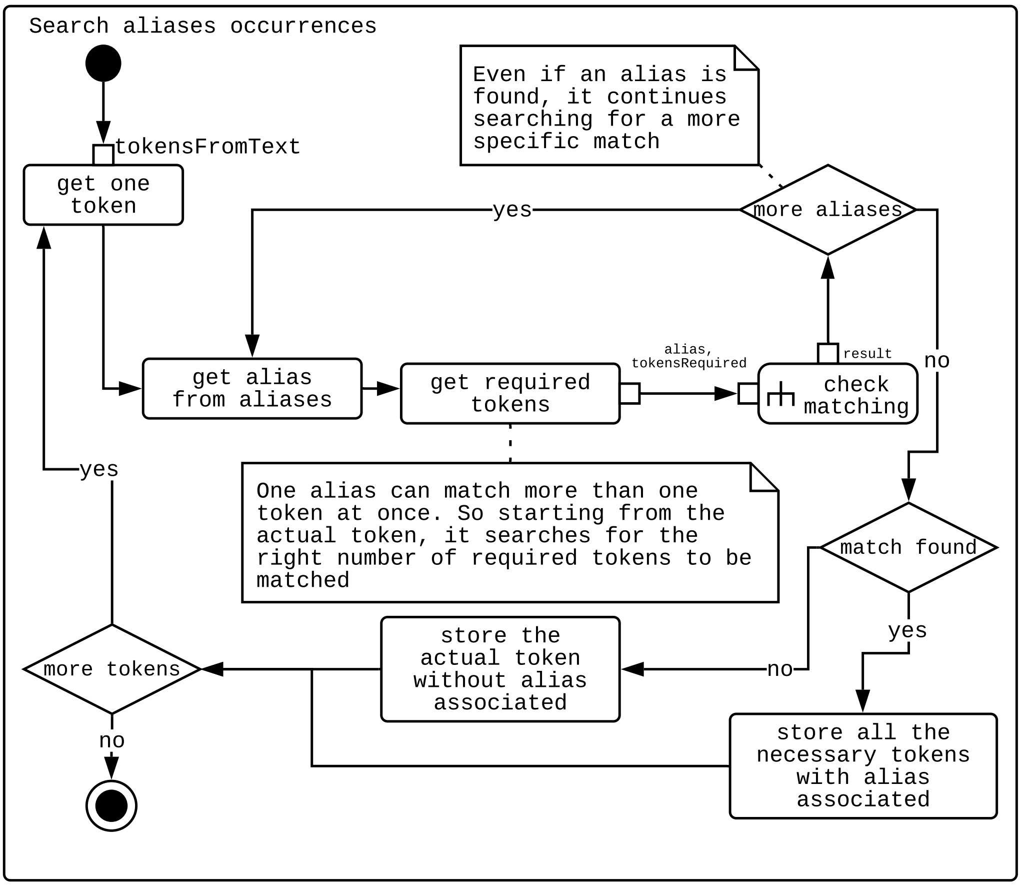

Search aliases occurrences
tokensFromText
get one token
Even if an alias is found, it continues searching for a more specific match
yes
more aliases
no
get alias from aliases
get required tokens
alias, tokensRequired
result
check matching
yes
One alias can match more than one token at once. So starting from the actual token, it searches for the right number of required tokens to be matched
match found
yes
more tokens
store the actual token without alias associated
no
no
store all the necessary tokens with alias associated

Figure 2.4: Activity diagram - search aliases occurrences sub-activity

phase has given a positive match, or not. It is important to understand that, even if it is considered one token at a time, when there is a match against an alias that requires more than one token, then all the necessary tokens are saved as one entity and so not again re-analysed. When there are no more tokens to analyse the occurrences phase is terminated.

2.3.3 Token Research Steps

To conclude, fig. 2.5 describes the deepest part of the algorithm where the possible matching between tokens and alias is checked. Here it is possible to see the relationships between the above-introduced entities like `Token` and `AnyToken`. It is important to notice that the `AnyToken` lets the algorithm loop searching for

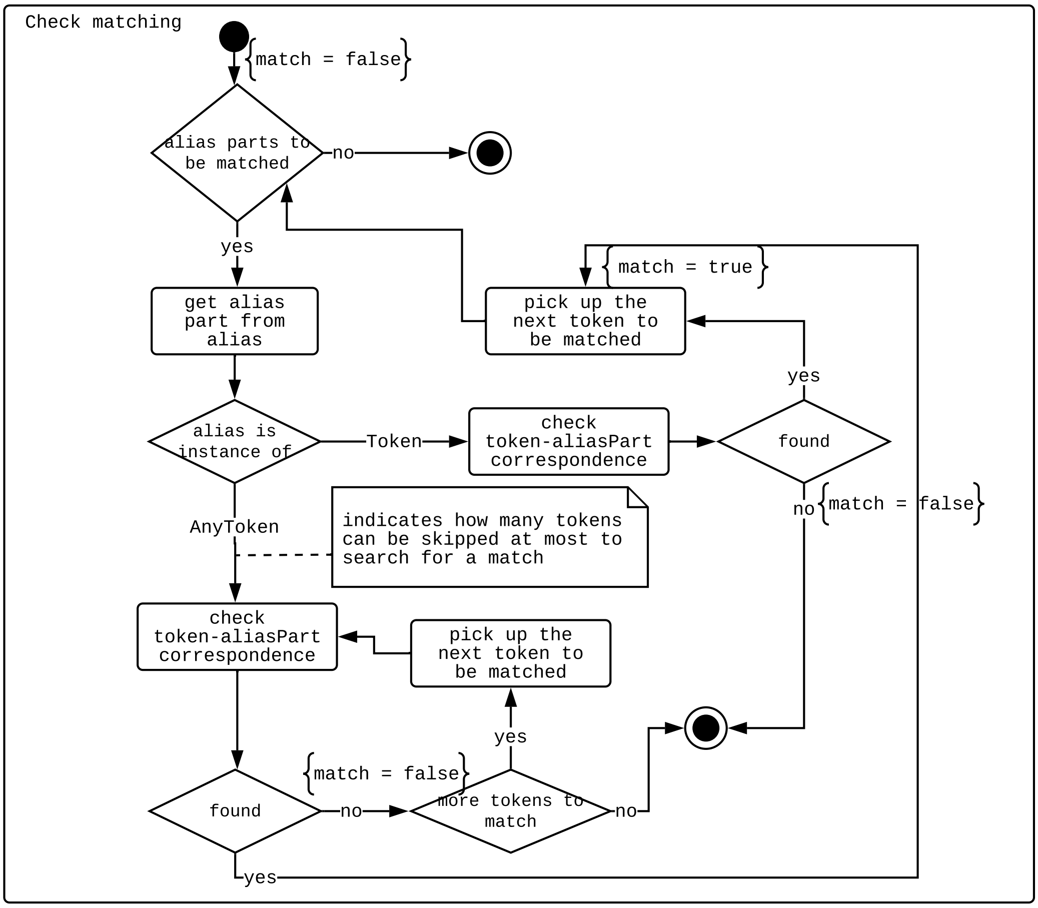

Check matching
match = false
alias parts to be matched
no
yes
get alias part from alias
pick up the next token to be matched
match = true
yes
alias is instance of
Token
check token-aliasPart correspondence
found
no
match = false
AnyToken
indicates how many tokens can be skipped at most to search for a match
check token-aliasPart correspondence
pick up the next token to be matched
yes
match = false
found
no
more tokens to match
no
yes

Figure 2.5: Activity diagram - check matching sub-activity

correspondence for at most n steps, where n indicates the maximum number of whichever words that can be encountered to continue searching for a match.

2.4 Implementation

The idea discussed in section 2.1 and section 2.2 is implemented using Java as the main programming language and in particular, it exploits some of the functionalities provided by a famous search-text library called Lucene. Lucene[2] is an open source Java library originally wrote by Doug Cutting in 1999. It provides indexing and search features with also the possibility to perform spellchecking,

[2] https://lucene.apache.org/

advanced analysis/tokenization capabilities, and hit highlighting, where the hits are the matching words/sentences against a query. During the years many other projects have joined Lucene and there are a lot of libraries built upon it that extend its power, e.g. Apache Solr[3] and Elasticsearch[4]. This library is used to perform the tokenization phase. Indeed it offers a series of language-specific `Analyzers` that are able to analyse a text and transform its words into an exploitable shape. Words are converted to lower case and it is performed a stemming on them, stop-words are removed and text is split into tokens, i.e. words, based on white spaces. So, starting from an unstructured stream of text, a series of tokens is extracted. Once tokenization is done, tokens are analysed checking for a match with tokens. This phase is implemented by performing a manual scan of every single token in order to understand whether it can match an alias or not in all its parts. All the resulting entities and aliases are stored into an Oracle database and are used for further analysis described in the following chapters.

To be noticed, Lucene supports research on text based on wildcard queries (alias in our context) and so we tried to exploit its power also in the second phase of the approach. A query is a word/sentence that Lucene can search for in a document or, more specifically, in an index. Indeed, before Lucene can be able to perform research on a text, the latter has to be converted in a more structured form (tokenization phase) and also indexed: Lucene uses *inverted indexing* for data and so, instead of mapping documents to tokens, it maps tokens to documents like a glossary at the end of any book. Once data are tokenized and indexed, the search phase can start and Lucene offers different ways to write a query to be searched for. As said, we tried to integrate Lucene capabilities in the approach discussed in the previous sections but there are some limitations and search performance does not have an improvement as discussed in chapter 4. Lucene integration comes into play when the `AliasPart` to be checked is a `Token` and, regardless of how many sub-parts it contains, Lucene tries to match the whole. The base approach, on the contrary, continues to split `Token` in `TokenPart` based on asterisks or question mark and then analyse them separately letter by letter to find a match. To give an example, the `Token` "economic* matter?" is split into two `TokenPart` that are singularly matched in the base approach, while in Lucene it is analysed as a whole. One of the main problems encountered trying to integrate Lucene in the base approach is that Lucene does not understand the wildcard *_n* that is a wildcard specifically defined in the base approach. To overcome this problem, we continue analysing token by token and let the Lucene query-engine matches only the other two kind of wildcards. The second problem concerns the different interpretation given to the wildcard *?* by the two procedures. More precisely,

[3] `https://solr.apache.org/`

[4] `https://www.elastic.co/elasticsearch/`

when Lucene finds the question mark wildcard it searches for one and only one literal character, whilst the other approach expects at most one character. To understand why this is a problem, pretend we have the alias “question?” and that the token to match it against is “questions” then, Lucene does not find a match, whereas the other method does. Since stemming is performed, many words would be truncated and many aliases will encounter this problem when Lucene is used. So, if we want to use Lucene, at least we do not have to perform any stemming on words or we have to change our wildcard semantic to meet the one proposed by Lucene. A final consideration concerns the way Lucene checks for matches. Indeed, even if we have an alias with no wildcards, like “demand”, Lucene considers as a match also words that are not exactly the same, for example “video-on-demand”. This is due to the fact that Lucene does not perform an exact match research, it uses a similarity method to search for matches and so the two words are very similar and then matched.

Chapter 3

Materiality Results Extraction

This chapter deals with the formalisation of our own approach underlining how it is developed starting from the approaches described in [2] and [4]. It also exposes the implementation phase, detailing the data used and the manipulations required.

3.1 Formalisation

This section describes and formalises the approach undertaken in this thesis work. One of the main problems in conducting a sustainability report is that it has to take into account different opinions and actively involve a lot of people (both internal and external). Our approach, instead, tends to save time and effort gaining in efficiency, effectiveness and objectivity. Drawing inspiration from [2] and [4], we have re-implemented them in a way that the score given by an actor to a certain topic corresponds to the weighted number of occurrences retrieved for that actor. Furthermore, they only consider the GRI standard practically, whereas our approach is able to explore every hierarchy in our database.

The following formulas synthesise what are the main values we compute to achieve our results. Indeed, we want to calculate a score for every topic in order to prioritise them.

Known that:

- $\mathbf{A_c} = \{a_{|A_c|}\}$, is a set containing all the authors considered as *company*;

- $\mathbf{A_s} = \{a_{|A_s|}\}$, is a set containing all the authors considered as *stakeholders*;

- $\mathbf{A} = A_s \cup A_c$, is a set containing all the authors, e.g. company, stakeholder, social;

- $\mathbf{T} = \{t_{|T|}\}$, is a set containing all the topics in the considered hierarchy;

- $\mathbf{D} = \{d_{|D|}\}$, is a set containing all the documents analysed in our database;

let:

- **children** : $T \rightarrow B_t$, be a function that, given a topic, returns its children in the considered hierarchy, where B_t, is a set containing all the partitions of T;
- **author** : $D \rightarrow A$, be a function that, given a document, returns its author;
- **words** : $D \rightarrow \mathbb{N}$, be a function that, given a document, returns how many words it contains;

To compute all the necessary scores we have to manage some preliminary results. In particular:

$$\mathbf{D(a_i)} \subseteq D \ s.t. \ author(d) = a_i \ \forall d \in D(a_i)$$

represents a set containing all the documents of a specified author.

Our reference papers, [2] and [4], make every involved actor giving a rate to each topic in order to establish a priority. In particular, Bellantuono et al. design a procedure in which both stakeholders and reporting company choose a label indicating the importance they give to a certain topic, see section 3.2.2 for more details. Calabrese et al., instead, make the actors voting for each topic with respect to the others at the same level in the hierarchy and belonging to the same *father*, more details in section 3.2.1. Our approach neither has a direct rate as a basis nor any stakeholders to ask about. We only have reports and documents authored by both the reporting company and the stakeholders and by which we extract what would be considered the importance of a topic to an actor. More in details, the key idea is that the more a topic *occurs* in the documents of an author the more it should be considered important for that author. Starting from this idea, we formalise the concept of direct occurrences as follows in definition 3.1.

Definition 3.1 (Direct occurrence)**.** A direct occurrence represents how many times an author has directly mentioned a specific topic in his documents, i.e. the occurrences retrieved by the text mining approach described in chapter 2. The computation of the direct occurrences can be seen as a function that, given an author a_i and a topic t_j, returns the occurrences of t_j retrieved in documents written by a_i. This is shown in the following formula where **do** stands for *direct occurrences.*

$$\mathbf{do(a_i, t_j)} : (A, T) \rightarrow \mathbb{N}$$

Since we work with hierarchies, we can link the *occurrences* of the topics belonging to the same hierarchy and, definitely, consider every topic both as a standalone topic and as a topic that aggregates other topics (its children in the hierarchy, if any). This is why the concept of recursive occurrences has emerged and it is defined in definition 3.2.

Definition 3.2 (Recursive occurrence)**.** A recursive occurrence represents how many times an author has mentioned a specific topic and all of its children in his documents. The computation of the recursive occurrences can be seen as a function that, given an author a_i and a topic t_j, returns the occurrences of that topic t_j and its children retrieved in the documents written by a_i. This is shown in the following formula where **ro** stands for *recursive occurrences.*

$$\mathbf{ro(a_i, t_j)} = do(a_i, t_j) + \sum_{t \in children(t_j)} ro(a_i, t)$$

Our main objective is to extrapolate a valid vote representing the real interest of an actor for a topic. The previously mentioned concepts, i.e. direct and recursive occurrences, are not enough because they do not take into account an important parameter presented below. Indeed, every author can write many documents with different size and in these documents he can mention a specific topic more than once. It is important to evaluate how many times an author mentions a topic, i.e. occurrences, but with reference to how much that author says in general so that we can estimate the real importance of a certain topic to an author, i.e. his score. This is evident considering the following situation: an author publishes hefty documents dealing with a lot of different topics. Inside these big documents the topic *1* has 100 occurrences, while an other author deals with the same topic only 10 times in a not very big document. Considering only the occurrences, the topic *1* should be far more important for the first author than for the latter but this is not necessarily true: even if 100 occurrences is a big number, it has to be put in relation to how much that specific author has to say in general because he could deal with other topics with the same occurrences and so the topic *1* should not be considered so important. Definitely, we have to compute what we refer to as the speech volume of an author. This concept is defined in definition 3.3.

Definition 3.3 (Speech volume)**.** The speech volume represents how many words an author has written in his documents. The computation of the speech volume can be seen as a function that, given an author a_i, sums up all the words written by that author. This is shown in the following formula where **sv** stands for *speech volume.*

$$\mathbf{sv(a_i)} = \sum_{d \in D(a_i)} words(d)$$

Thanks to the speech volume and the recursive occurrences we can define a coherent vote, namely score, given by an author to a topic that is representative to the author's opinion about that topic. This concept is defined in definition 3.4.

Definition 3.4 (Score)**.** The score represents how much it is important a certain topic (and all of its sub-topics) to a given author. The computation of the score can be seen as a function that, taking an author a_i and a topic t_j, computes how much a topic is mentioned by that author ($\mathbf{ro(a_i, t_j)}$) with respect of how much that author says in general ($\mathbf{sv(a_i)}$). The result is presented in the following formula.

$$\mathbf{score(a_i, t_j)} = \frac{ro(a_i, t_j)}{sv(a_i)}$$

We now have, for the same topic, different scores computed for every author. It is necessary somehow to aggregate these values in order to retrieve a unique score for every topic that is useful to prioritise them. Following Bellantuono et al. idea, we search for a valid way to weight different authors' opinions defining their salience and then computing a weighted score for every topic by aggregating different views. The main idea from which derives our definition of salience is: consider, for example, author *A* publishing 100 hefty documents and author **B** publishing only 2 small documents. Author *A* considers topic *1* not so important, while according to author *B* this is a very important topic. Since author *A* has a lot to say his opinion has to be considered much more than *B*'s opinion. Starting from this reasoning, we exploit the speech volume to derive the importance of the different authors and we define our concept of salience, presented in definition 3.5.

Definition 3.5 (Salience)**.** The salience represents how much it is important an author's opinion with respect to the others. The computation of the salience can be seen as a function that, taking an author a_i, computes how much that author has to say ($\mathbf{sv(a_i)}$) with reference to the maximum value of speech volume computed among all the authors. It is also used a correction factor $weight(a_i)$ in order to adjust the right importance of an author: even if an author has few things to say, he would have to be considered a bit more important. The result is presented in the following formula.

$$\mathbf{salience(a_i)} = \frac{sv(a_i)}{\max\limits_{a \in A} sv(a)} \cdot weight(a_i)$$

In conclusion, we have all the necessary results to compute a single score for every topic, considering and weighting every author's score. In definition 3.6 we define our average score.

Definition 3.6 (Average score)**.** The average score represents the final score of a topic weighting the different scores given by the authors that have mentioned it. The computation of the average score can be seen as a function that, taking a topic t_j and a set of authors A', computes the final score of that topic by multiplying the score given to that topic from authors and the authors salience. The result is then normalised upon the authors' cardinality. The result is presented in the following formula.

$$\mathbf{avgScore}(\mathbf{t_j}, \mathbf{A'}) = \frac{\sum_{a \in A'} salience_a \cdot score(a, t_j)}{|A'|}, A' \subseteq A$$

3.2 Matrix and rank extraction

This section aims at describing how to use the results presented in section 3.1 to build up our materiality rank and materiality matrix drawing inspiration from what is exposed in [2] and [4].

3.2.1 Calabrese

Calabrese et al. consider the GRI hierarchy at different levels in order to draw up a rank in which local scores (those for topics at sub-levels) are mapped into global scores (every local score is seen in terms of the whole hierarchy). In the paper they did not stress the importance of assign a different salience to the actors demanded to score each topic. Our approach, instead, uses the salience to compute the local and the global scores, indeed we use the average score, definition 3.6, that is computed exploiting the actors' salience.

To obtain a rank, we complete the formulas presented in the previous section with the following ones.

Let:

- **parent** : $T \rightarrow T$, be a function that, given a topic, returns its father in the hierarchy;
- **siblings** : $T \rightarrow B_t$, be a function that, given a topic, returns its siblings in the hierarchy, where B_t, is a set containing all the partitions of T.

We can finally define what the *local score* of a topic represents. It is presented in definition 3.7

Definition 3.7 (Local score)**.** The local score represents the score computed for a certain topic considering a set of authors with respect to all the topics at the same level of the hierarchy to which the considered topic belongs. The computation of the local score can be seen as a function that, taking a topic t_j, computes its average score and divides it by the sum of the total average score of both t_j and its siblings. The result is shown in the following formula.

$$\mathbf{LS(t_j)} = \frac{avgScore(t_j, A)}{\sum_{t \in \{siblings(t_j)\ \cup\ t_j\}} avgScore(t, A)}$$

Once the local score has been defined we can see the global score as described in definition 3.8.

Definition 3.8 (Global score)**.** The global score represents the local score computed for a certain topic considering a set of authors with respect to the whole hierarchy to which the considered topic belongs. The computation of the global score can be seen as a function that, taking a topic t_j, computes its local score and multiplies it by the global score of its father. By this recursive approach, every topic obtains a score representative of its importance among the other topics at the same level in the hierarchy, even if they are not siblings. The result is shown in the following formulas.

Assuming that

$$\mathbf{GS(root)} = 1$$

where the *root* is the top-level topic in the hierarchy, we can compute the *global score* for a topic as follows:

$$\mathbf{GS(t_j)} = LS(t_j) \cdot GS(parent(t_j))$$

3.2.2 Bellantuono

Bellantuono et al. use the GRI hierarchy to achieve a result. They draw up a Cartesian plane (i.e. materiality matrix) in which every point is an aspect extracted from the GRI hierarchy and considered as material (i.e. important for both stakeholders and company). The x-axis represents what is material for the company, whereas the y-axis is to see what is material for all the considered stakeholders. So, in this paper different levels of the hierarchy are not taken into account. On the contrary, we implement a distinct materiality matrix for every level of the hierarchy, summing up all the potential sub-topics: if you want to compute a materiality matrix for the root level for a certain standard, you will obtain a materiality matrix with only one point (representing the root), but coordinates for that point are computed considering also the scores of its children.

The following formulas show how we obtain coordinates for every issue:

$$\mathbf{xScore(t_j)} = score(a, t_j),\ a \in A_c$$

represents the x-axis component for a certain topic (i.e. issue) t_j. It only uses the score because there is only one kind of actor considered, i.e. the reporting company.

$$\mathbf{yScore(t_j)} = avgScore(t_j, A_s)$$

represents the y-axis component for a certain topic. It uses the average score in order to consider the different salience values for each of the involved stakeholders in the set A_s.

3.3 Implementation of the approach

This section is designed to detail what are the data used in this thesis and how they are manipulated in order to obtain more effective results. It also deals with the requirement analysis, the design phase of the approach described in the previous sections, section 3.1 and section 3.2. The implementation is widely discussed in 4.3.3.

3.3.1 Database structure

The text mining approach described in chapter 2 works with relational tables saved in an Oracle database and organises the extracted results in the same way. Figure 3.1 shows the logical data model of the database. It is represented exploiting the Crow's Foot notation [8]. The original tables, those coming from the text mining approach, were manipulated to obtain more useful derived data and to structure them in a simpler and more effective fashion. The following three tables are those directly created starting from the original ones.

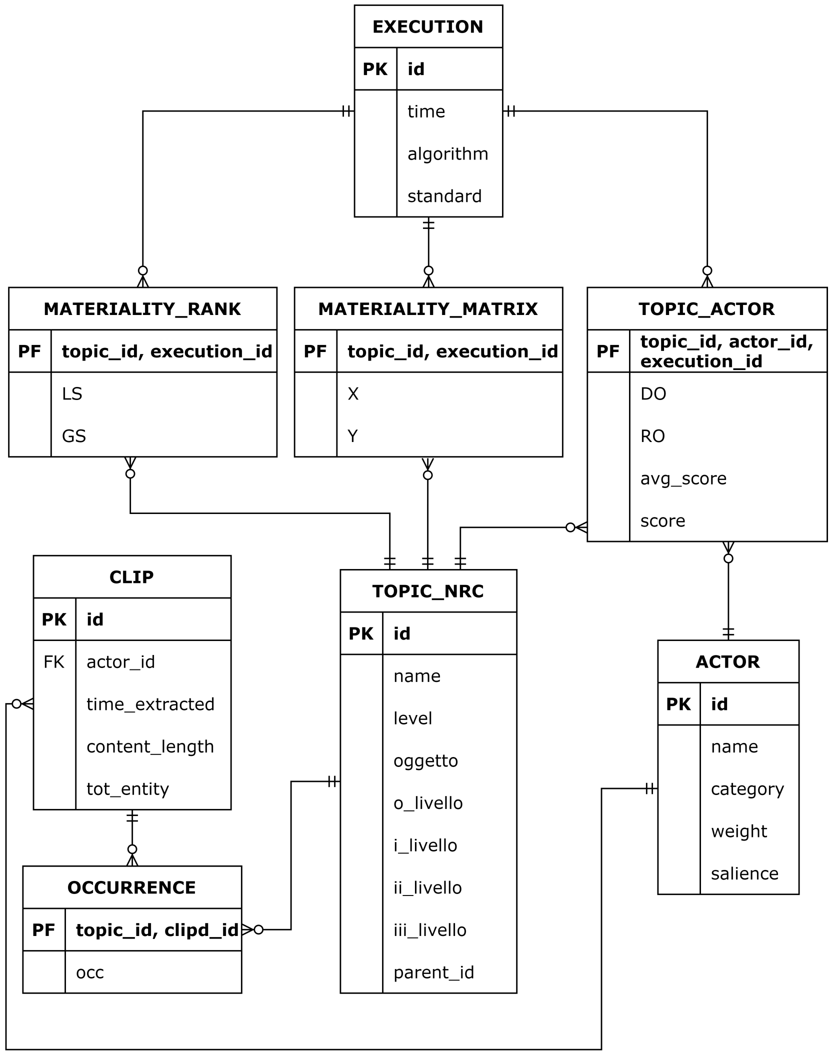


Figure 3.1: Logical data model of the database using the Crow's Foot notation

Topic The original table contains all the topics that the algorithm has to find in the documents. More in details, it lists different standards hierarchies (e.g. GRI hierarchy with categories, aspects and indicators) with a different row for every different level-topic pair in the hierarchy, so we have a flattened vision of the hierarchies. `Topic` is obtained from the previous one by adding two columns, i.e. `o_livello` and `parent_id`, whose objective, together with the other columns presented below, is to help in reconstructing the flattened hierarchy.

- `level`, indicates for each row which column, between `i_livello`, `ii_livello` and `iii_livello`, is the first column with not relevant data. Identify the *worthless* column is useful to get the level of a certain topic in the hierarchy. For example, a row where the column `level` shows the value "L0" indicates a root topic, whilst a row with the value "L2" represents a topic where only the columns `oggetto`, `o_livello` and `i_livello` contain significant values;

- `oggetto`, is a column always filled with meaningful values and indicates the standard (or main category) to which a topic belongs, e.g. *GRI standard* or Dow Jones Sustainability Indices (DJSI);

- `o_livello`, may contain either the repetition of the standard if that hierarchy does not contain any subdivision at the first level below the root, or names of topics otherwise. To give an example, for the standard GRI we have four categories, namely *social*, *economic*, *general_report* and *environmental*, while for the DJSI we do not have any subdivision at that level so the DJSI word is repeated in the `o_livello` column;

- `i_livello`, `ii_livello` and `iii_livello` as for the previous level, show names of the topics at different levels;

- `parent_id`, is an ad-hoc computed column used to link a topic (every topic is uniquely identified by an `id`) with its father in the hierarchy. This column makes computation easier and recursive-prone.

In table 3.1 we can see a comparison between the schema of the original table `Topic` and the new created for the table `Topic`. It can be clearly noticed that there are two new columns, i.e. `o_livello` and `parent_id`. Table 3.2 shows an example of the table `Topic` where it is possible to understand what is the role of every column in reconstructing the hierarchy. It represents, indeed, a flattened extract of the GRI hierarchy where it is possible to understand, for example, which row stands for the root node of the hierarchy, i.e. it is the row with id = 1 because its `level` value is "L0".

Topic	*Topic_nrc*
id	id
name	name
level	level
oggetto	oggetto
	0_livello
I_livello	I_livello
II_livello	II_livello
III_livello	III_livello
	parent_id

Table 3.1: Comparison between `Topic` and `Topic_nrc` schema

id	*name*	*level*	*oggetto*	*0_livello*	*i_livello*	*ii_livello*	*iii_livello*	*parent_id*
1	gri standard	L0	GRI	null	null	null	null	null
2	economic	L1	GRI	economic	null	null	null	1
3	perf. economica	L2	GRI	economic	perf. economica	null	null	2
4	201-1 Valore economico	L3	GRI	economic	perf. economica	201-1..	null	3

Table 3.2: An extract of the flattened GRI hierarchy

Clip This table lists all the documents, or clips, analysed by the text mining algorithm. Every clip is identified by an `id`, is written by an author (i.e. `actor_id` taken from `Actor` table described below), contains a number of characters, i.e. `content_length`, and a number of entities, i.e. `tot_entity`, retrieved using the table `Entity`.

Occurrence This table is obtained starting from other tables which are `Alias`, `Alias_topic`, `Entity` and `Entity_contains` and it simply puts in correlation a `Topic_id-Clip_id` pair with the occurrences (i.e. how many times) retrieved by the text mining algorithm for that topic in that clip.

The following table is created to better represent authors' data.

Actor This table contains information about clips' authors (also called actors in our context). Indeed, every author (e.g. the company preparing the sustainability report or another firm involved) has a dedicated `id`, a `name` and a `category` (i.e. an author can be considered as a part of the stakeholders or of the reporting company). Furthermore, drawing inspiration from Bellantuono et al., it is possible to define a weight and a salience for each author. Section 3.1 describes what are these values and how they are computed and utilised in the approach pursued in this thesis work.

Finally, the following tables are created to store results for each execution of the program.

TopicActor This table links a `Topic_id-Actor_id` pair with values computed in a certain execution. These values, i.e. `score`, `average_score`, `direct_occurrence` and `recursive_occurrence`, are discussed in section 3.1.

Execution This table describes meta-information about different executions of the implemented approach, namely Bellantuono and Calabrese. Indeed, there is an `id` for every execution, a timestamp, the indication of what algorithm has been executed and for what standard(s) (e.g. GRI, DJSI, all,..). It is used to compare different results obtained varying some parameters (e.g. it is possible to explore results changing the importance of an actor).

MaterialityRank This table describes results obtained executing the implementation of Calabrese et al. approach. In detail, it offers a global score and a local score (both described in section 3.1) for every topic in a particular hierarchy.

MaterialityMatrix This table describes results obtained executing the implementation of Bellantuono et al. approach. Specifically, it contains the axes values useful to draw a materiality matrix for every topic in a particular hierarchy; more details are presented in section 3.1.

3.3.2 Requirements

This section describes what are the requirements given to implement our version of Bellantuono and Calabrese's approaches, computing data for a matrix and a rank respectively. This project can be viewed as a script that, given some parameters, computes some results. Requirements can be divided into i) functional , ii) non functional, iii) implementation.

Functional

Functional requirements concern the functionalities a software has to provide. Figure 3.2 represents the functional requirements of this project in the form of a flowchart[1]. Starting from retrieving data from the database, this diagram represents the possibility to compute and upload actors' salience. Furthermore, it

[1] Reference symbols used in this diagrams can be retrieved at `https://www.smartdraw.com/flowchart/flowchart-symbols.htm`

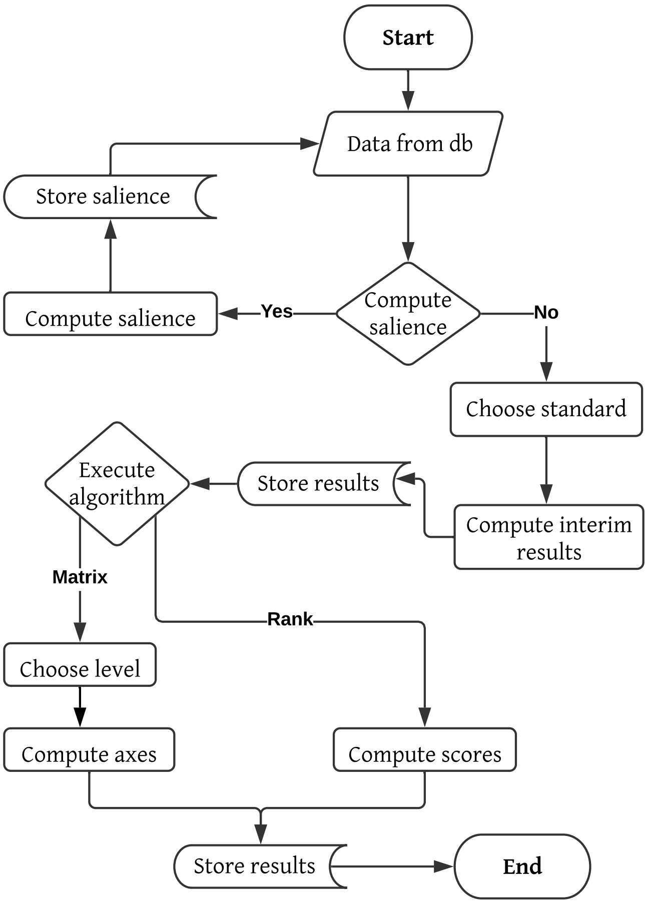


Figure 3.2: Functional requirements flowchart

deals with the choice of the standard(s) to execute the algorithm for; it expresses the possibility to compute interim results necessary to both the algorithms (e.g. recursive occurrences, score, ...). It is possible to choose whether to compute the rank (and so compute local and global scores) or a materiality matrix (computing x-axis and y-axis for the topics of the chosen level). Finally, it is possible to notice that all the obtained results can be stored in the database.

Functional requirements are also listed below.

1. Take updated data from the Oracle database

2. allow the execution of the Calabrese's algorithm

   2.1 allow the execution considering one or more different standards

2.2 compute local and global scores

3. allow the execution of the Bellantuono's algorithm

    3.1 allow the execution considering one or more different standards

    3.2 allow the execution considering a certain level in the hierarchy

    3.3 compute the x-axis and the y-axis for each specified topic

4. compute the salience for each author in the database

5. compute interim results, which are necessary to both the algorithms (e.g. recursive occurrences, score, ...)

6. load obtained results in the Oracle database

    6.1 allow to load computed salience

    6.2 allow to load computed interim results (e.g. average score, direct occurrences)

    6.3 allow to load computed final results of the algorithm executed

Non functional

Usability It should be easy to run both rank and matrix results and for different standards

Implementation

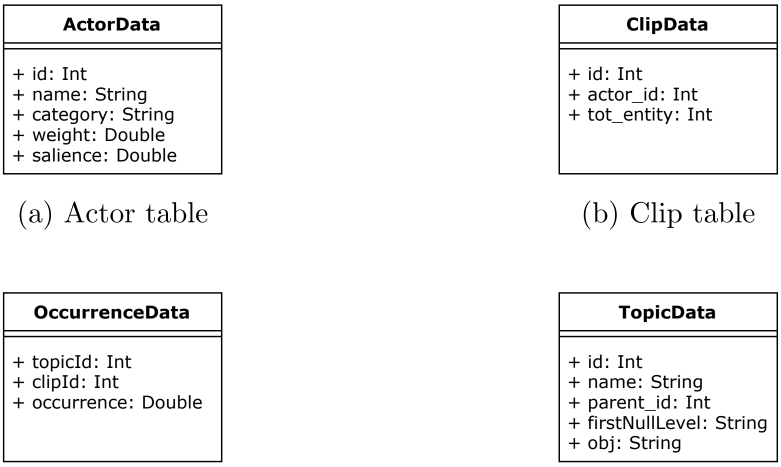
This work will be implemented using Scala[2] as the main programming language. Moreover, we interact with an Oracle database.

3.3.3 Design

Starting from the problem formalisation and the requirements analysis, it is necessary to design a software able to interact with an Oracle database, req. 1., and capable to compute the described formulas, req. 2. and 3.. Therefore, this section aims at describing what are the main concepts formulated to respond to the above mentioned requirements, section 3.3.2.

Scala, more than a programming language, is a good tool to support the design phase. Therefore, since it was considered as a requirement, section 3.3.2, it has been actively used to make some concepts easier to understand and conceptualise.

[2] https://www.scala-lang.org/



(a) Actor table

(b) Clip table

(c) Occurrence table

(d) Topic table

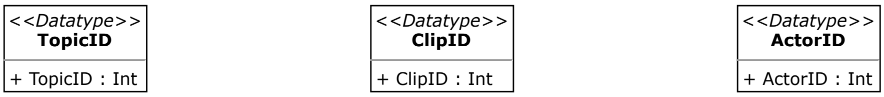
Figure 3.3: Database tables views

More in details, it is necessary to understand data in the database and to organise them. Therefore, every table is thought as a specific class in which the table columns are the class properties. Figure 3.3 shows the results, in particular:

- `ActorData`, fig. 3.3a, represents data coming from the `Actor` table and it is necessary to retrieve authors' information such as their salience and category;
- `ClipData`, fig. 3.3b, represents data coming from the `Clip` table and it is necessary to know how much content a document has;
- `OccurrenceData`, fig. 3.3c, represents data coming from the `Occurrence` table and it is necessary to elaborate scores;
- `TopicData`, fig. 3.3d, represents data coming from the `Topic` table and it offers all the valuable information to reconstruct the flattened hierarchy.

The focus of this phase, once data are defined and ready to be used, is on defining the main classes useful to design the context of the algorithms execution. Scala, once again, supports design process and helps the schema readability with the introduction of type alias. Figure 3.4 shows how common data types, like Integer, can be wrapped to become more meaningful and context-aware entities. In particular, concepts as ids, which are simple Integer numbers, can be wrapped to become entities with a more powerful meaning. Even if there is no semantic difference among a `TopicID` and an `Int`, the first one results in a more effective reference.

Figure 3.5 shows how the formalisation described in 3.1 is reified in a class diagram made up of three main concepts:



(a) Topic id data type (b) Clip id data type (c) Actor id data type

Figure 3.4: New data types created exploiting type alias mechanism

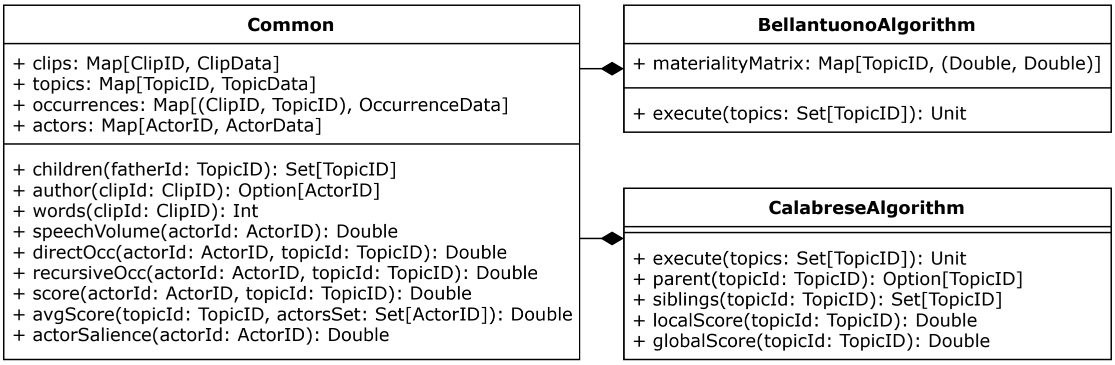

Common
+ clips: Map[ClipID, ClipData]
+ topics: Map[TopicID, TopicData]
+ occurrences: Map[(ClipID, TopicID), OccurrenceData]
+ actors: Map[ActorID, ActorData]
+ children(fatherId: TopicID): Set[TopicID]
+ author(clipId: ClipID): Option[ActorID]
+ words(clipId: ClipID): Int
+ speechVolume(actorId: ActorID): Double
+ directOcc(actorId: ActorID, topicId: TopicID): Double
+ recursiveOcc(actorId: ActorID, topicId: TopicID): Double
+ score(actorId: ActorID, topicId: TopicID): Double
+ avgScore(topicId: TopicID, actorsSet: Set[ActorID]): Double
+ actorSalience(actorId: ActorID): Double
BellantuonoAlgorithm
+ materialityMatrix: Map[TopicID, (Double, Double)]
+ execute(topics: Set[TopicID]): Unit
CalabreseAlgorithm
+ execute(topics: Set[TopicID]): Unit
+ parent(topicId: TopicID): Option[TopicID]
+ siblings(topicId: TopicID): Set[TopicID]
+ localScore(topicId: TopicID): Double
+ globalScore(topicId: TopicID): Double

Figure 3.5: Class diagram of the main entities

- `Common`, is the fundamental entity. It acts as the entry point of the main application and it is used firstly, to interact with database to retrieve data, req. 1.; then, to compute intermediate results useful to the reification of the two algorithms, req. 5.). Indeed, `Common` exposes all the functions described in section 3.1.

- `BellantuonoAlgorithm`, represents the specific Bellantuono's algorithm, req. 3., so it has the concept of materiality matrix and knows how to compute it exploiting methods exposed by `Common`, req. 3.3;

- `CalabreseAlgorithm`, represents the specific Calabrese's algorithm, req. 2., so it has the concept of local and global scores and knows how to compute them, req. 2.2, exploiting methods defined in `Common` as well as other methods, like `siblings`, which are useful to go back up along the hierarchy.

In order to fulfil req. 6. class `Uploader` is presented in fig. 3.6. More in details, it is necessary for it to know how to interact with the Oracle database and how to upload data retrieved from `Common`, `CalabreseAlgorithm` and `BellantuonoAlgorithm`. It has to manage i) salience, req. 4. and 6.1; ii) intermediate results coming from `Common`, req. 6.2, and final results coming from the algorithm executed, either Bellantuono or Calabrese, req. 6.3.

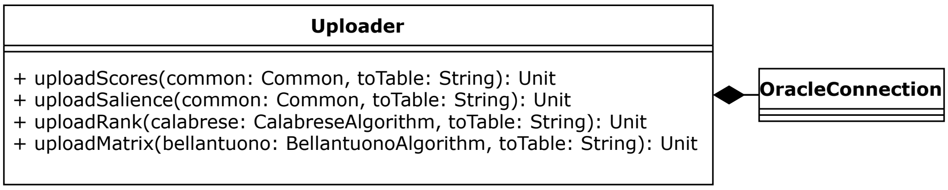

Uploader
+ uploadScores(common: Common, toTable: String): Unit
+ uploadSalience(common: Common, toTable: String): Unit
+ uploadRank(calabrese: CalabreseAlgorithm, toTable: String): Unit
+ uploadMatrix(bellantuono: BellantuonoAlgorithm, toTable: String): Unit
OracleConnection

Figure 3.6: Class diagram of the uploader concept

Chapter 4

Evaluation

This chapter discusses the results obtained by performing the approaches described in chapter 2 and chapter 3. Section 4.1 provides an overview of the data analysed in this dissertation. The results of the text mining approach described in chapter 2 are evaluated in section 4.2. The results obtained from the materiality extraction described in chapter 3 are discussed in section 4.3 where they are visually represented to evaluate and compare the two algorithms discussed in chapter 3. Effective visualisation is successfully obtained thanks to the *Tableau* software, described in section 4.3, and the reading of [24], [23]. Execution times are registered by exploiting a Java Profiler, namely YourKit[1].

Performance tests are made on a computer with the following technical specification:

- CPU: AMD Ryzen 3 2200G with Radeon Vega Graphics 3.50 GHz
- RAM: 8GB 2400 MHz

4.1 Dataset

Data analysed and manipulated in this thesis work come from a series of documents provided by a telecommunications company during a past research collaboration on the subject of materiality reporting. The retrieved documents come from different sources, i.e. textual documents in different formats, social posts. There are both English and Italian documents either referring to the reporting company or its stakeholders and with different dimensions. A more structured overview of the available data is presented in table 4.1. There it is possible to observe how many documents (column *Count*) are provided for each language (column *Language*) and either for the company or stakeholders (column *Category*). It is also detailed

[1] `https://www.yourkit.com/java/profiler/features/`

Language	*Category*	*Type*	*Count*	*Avg Length*
English	Company	docx	1	130,759
		pdf	3	67,761
	Stakeholder	html	12	11,003
		pdf	106	201,312
		txt	1	10,420
Italian	Company	pdf	19	245,420
		txt	2	9,045
	Stakeholder	docx	1	21,652
		html	1	34,118
		pdf	62	341,867
		social	57,664	203
		txt	21	12,741

Table 4.1: Overview of the available data

the documents average length in term of characters (column *Avg Length*) and the kind of source it is (column *Type*).

4.2 Text Mining Approach Evaluation

The text mining approach described in chapter 2 aims at analysing a text in order to understand what topics it deals with and to what extent. The approach follows two main phases: the first one is the tokenisation phase and it is based upon the Lucene library, whilst the second one practically executes the research of an alias upon a text and is implemented with no Lucene support. In this section are described the tests made to evaluate the algorithm performance. Both the custom implementation and the integration with the Lucene query engine discussed in chapter 2 are tested. The showed times refer to how long it takes for both approaches to perform only the research of a match upon a `Token`. Indeed, times to build up the index necessary for Lucene to perform research are not shown but they further slow down the Lucene approach by $\approx 80\%$. Figure 4.1 compares how long it takes for both approaches to execute the research by varying the number of aliases they have to search for. Data are presented for $n, 2n, 4n, 8n$ aliases, where n is the actual number of Italian aliases stored in the reference dataset and it is equal to 2194. The results show that the custom implementation performs better than the one exploiting Lucene, and both of them substantially scale up linearly with the number of aliases they have to analyse.

Figure 4.1 highlights how long it takes for both approaches to execute the research by varying the content length of the documents to be analysed. Times are registered on a sample of 5 documents within each of the indicated ranges.

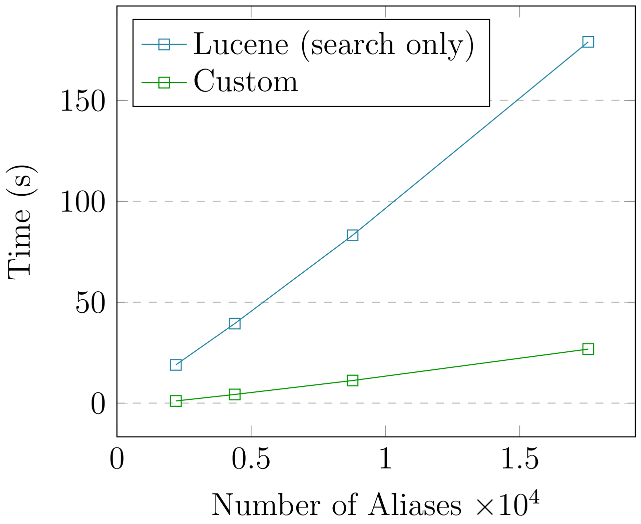


Figure 4.1: Lucene and Base approaches varying aliases

The results show that the custom implementation is far more efficient than the one exploiting Lucene.

In conclusion, the custom implementation remains the fastest. This is mainly due to the fact that the custom approach stops at the first non-matching character and has no infrastructure to build. Besides this, Lucene strength is to analyse and query a lot of big documents at once, without scanning them token by token as in our approach. This Lucene feature could not be directly exploited for two main reasons:

- the wildcard _n could not be used anymore because Lucene does not recognise it;
- Lucene is fast and optimised to recognise whether a query, i.e. alias, matches a text or not but when you want to retrieve what is the matched text, what is its offset or even how many times a query matches a text, it is not straightforward. Since our approach aims at retrieving this kind of information, entirely use Lucene to scan the entire file at once, would bring no gain.

4.3 Materiality Results Visualisation

This section visually details the results obtained when computing the materiality matrix and rank discussed in chapter 3. It also deals with the main manipulations exploitable during the data visualisation thanks to the *Tableau* software described below. Furthermore, a comparison among the results obtained looking at topics from different standards is accomplished in 4.3.3.

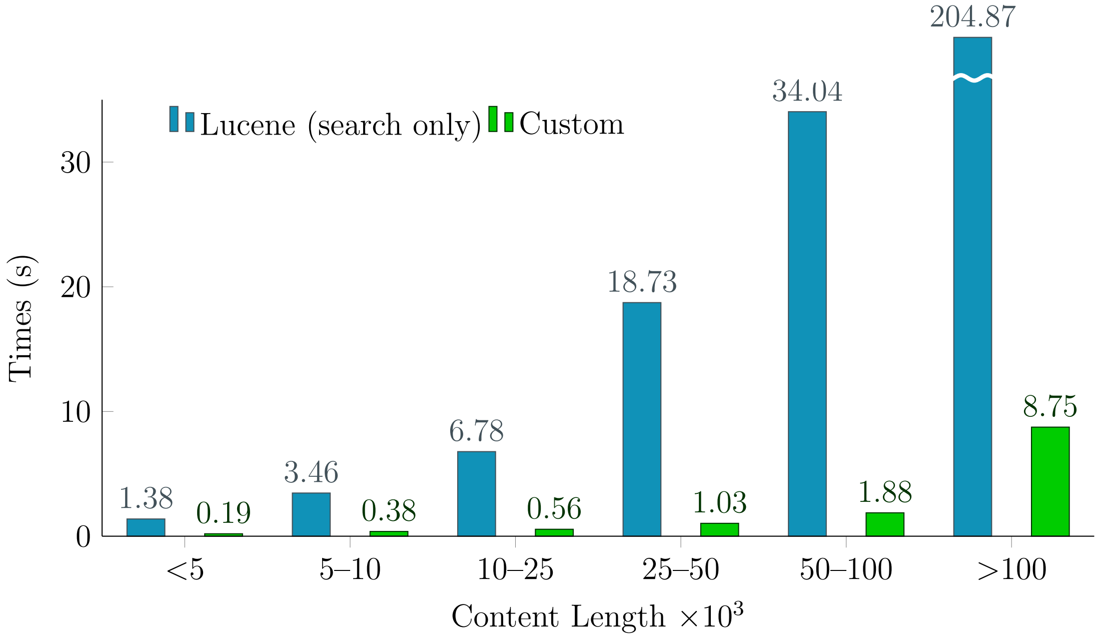


Figure 4.2: Lucene and Base approaches varying content length

Tableau

Tableau[2] is an analytic platform which aims is to efficiently translate simple drag-and-drop actions into data queries through a simple but powerful graphical interface. To better represent data and to better explore them, some of the most recent Tableau features have been used. More in details:

- *calculated fields*, allow you to filter and manipulate fields you need to use to better understand and represent your data. In this thesis they are widely used, especially to either hide or show data depending on certain conditions. For example, the local score (or global score as well) of a certain topic at a determined level in the hierarchy must be shown only if the user wants to visualise data from that specific level;

- *Level Of Detail (LOD)* expressions, a kind of transformation you can apply to data that allows you to compute and then show values at different levels of granularity. In this context, they are used to normalise the x-value and the y-value in the matrix with respect to the maximum x-value and the maximum y-value;

- *parameters*, they can be seen as variables such as numbers, dates or strings,

[2] https://www.tableau.com/why-tableau/what-is-tableau

that can be used to replace a constant value in a calculation or a filter. Once you have created a parameter you can dynamically change the value in your calculation using the parameter control. In this project, parameters are exploited to let the user decide what is the *execution* he wants to observe data for;

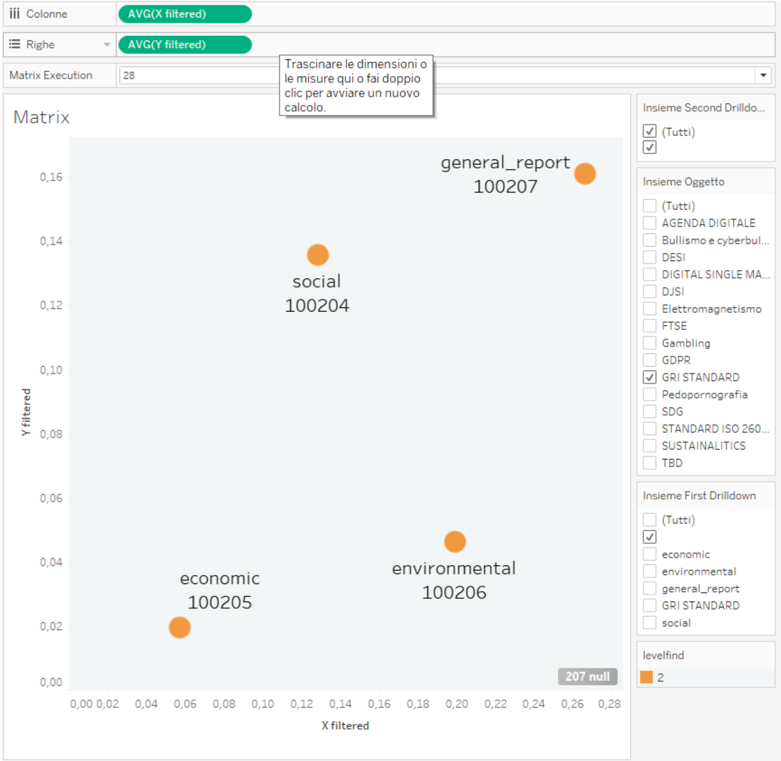
– *sets and set actions*, are a powerful way to dynamically change what the user sees. Sets are custom fields that include a subset of data based on some conditions. To make sets more interactive, *set actions* can be exploited. Indeed, they allow you to dynamically change what elements have to be included in the set and so showed, basically filtering on some conditions. In this thesis, they are used to create a dynamic and asymmetric drill down along the hierarchies levels, i.e. sets are dynamically filled with the topics to be shown.

4.3.1 Matrix results

The implemented script, described in chapter 3, can compute the coordinates of specified topics. Tests on performance have shown that it takes ≈ 9 minutes to compute values for topics from all the standards in the database (15) and for every level in the standards related hierarchies. The obtained values can be visually represented in a materiality matrix via Tableau. Both x and y values are normalised with respect to the maximum value along x and y respectively. Figure 4.3 shows a part of the Tableau interface with the plotted materiality matrix. X-values are on the columns, while y-values are on the rows as can be observed in the upper part of the image.

Users can interact with Tableau to analyse different executions through the parameter "Matrix Execution", in the upper part of the image, below the axes values. On the right side, there are multiple values to be chosen to make it possible to drill down through the different levels of the hierarchy considered (the image shows the second level of the GRI hierarchy). This drill down interaction can also be achieved by left-clicking on the single bullets. Figure 4.4 shows how the visualisation changes as the interaction proceeds: 4.4b is obtained by left-clicking on the blue bullet in 4.4a; 4.4c is obtained by left-clicking on the "economic" bullet in 4.4b and 4.4d is obtained by left-clicking on the "impatti economici indiretti" dot in 4.4c.

Different executions, with different parameters and settings, could lead to different matrices. As an example, table 4.2 illustrates two executions of the matrix algorithm and focuses on topics that have had an evident change by varying three actors' weights. The only coordinate varying is along the y-axis because salience, and consequently the weight, is considered only when computing values for the



Colonne: AVG(X filtered)

Righe: AVG(Y filtered)

Matrix Execution: 28

Trascinare le dimensioni o le misure qui o fai doppio clic per avviare un nuovo calcolo.

Insieme Second Drilldo...
- ☑ (Tutti)
- ☑

Insieme Oggetto
- ☐ (Tutti)
- ☐ AGENDA DIGITALE
- ☐ Bullismo e cyberbul...
- ☐ DESI
- ☐ DIGITAL SINGLE MA...
- ☐ DJSI
- ☐ Elettromagnetismo
- ☐ FTSE
- ☐ Gambling
- ☐ GDPR
- ☑ GRI STANDARD
- ☐ Pedopornografia
- ☐ SDG
- ☐ STANDARD ISO 260...
- ☐ SUSTAINALITICS
- ☐ TBD

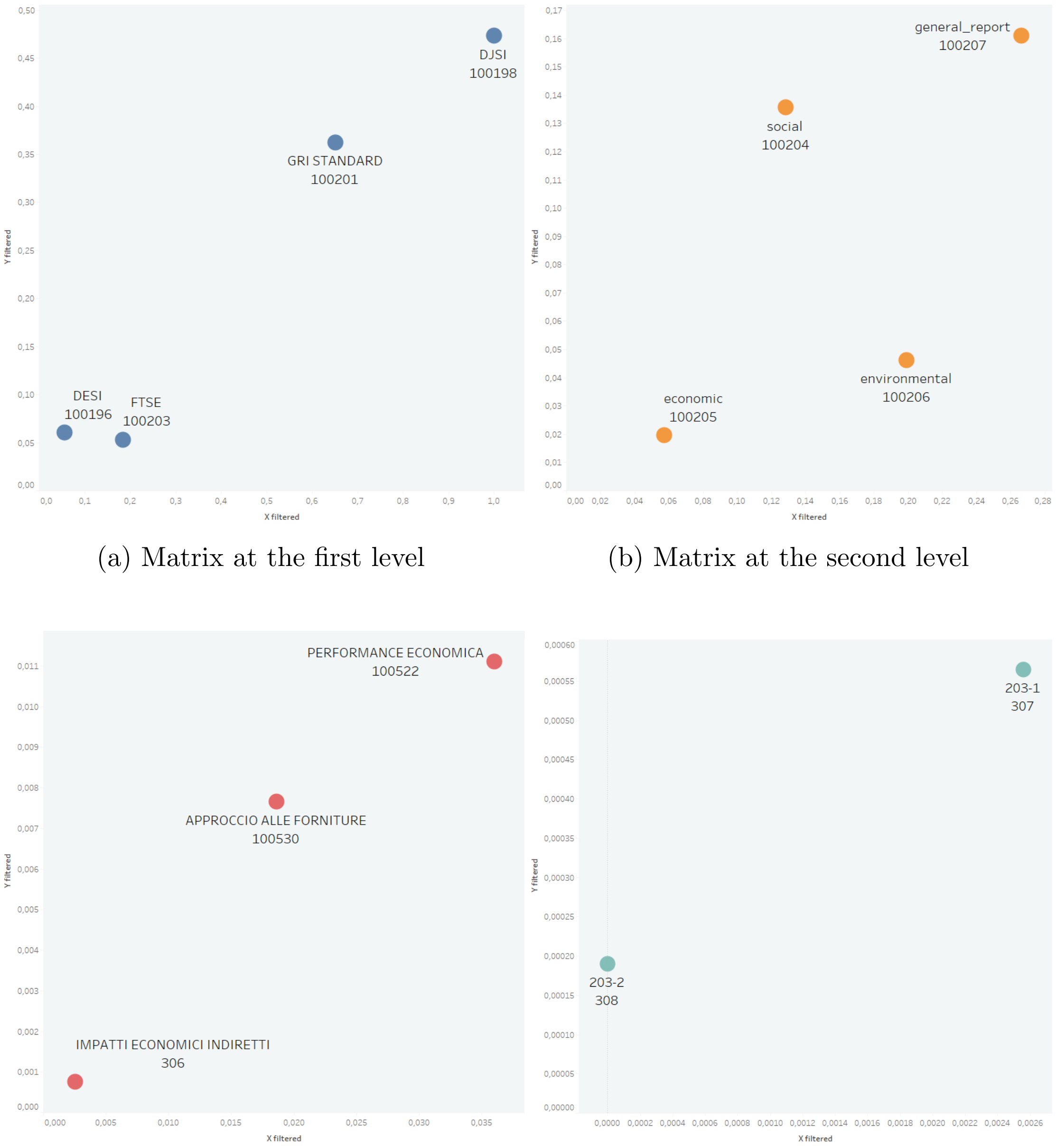
Insieme First Drilldown
- ☐ (Tutti)
- ☑
- ☐ economic
- ☐ environmental
- ☐ general_report
- ☐ GRI STANDARD
- ☐ social

levelfind
- 🟧 2

Figure 4.3: An example of the Tableau interface with a materiality matrix plotted

topic_name	exec_id	*weight*			*y_norm*
		social	stakeholder_17	stakeholder_1	
Corp. Govern	20	0.5	1	0.25	0.030
	23	0.8	0.2	1	0.029
Human Rights	20	0.5	1	0.25	0.037
	23	0.8	0.2	1	0.024

Table 4.2: An extract of changes occurred after actors' weights modification



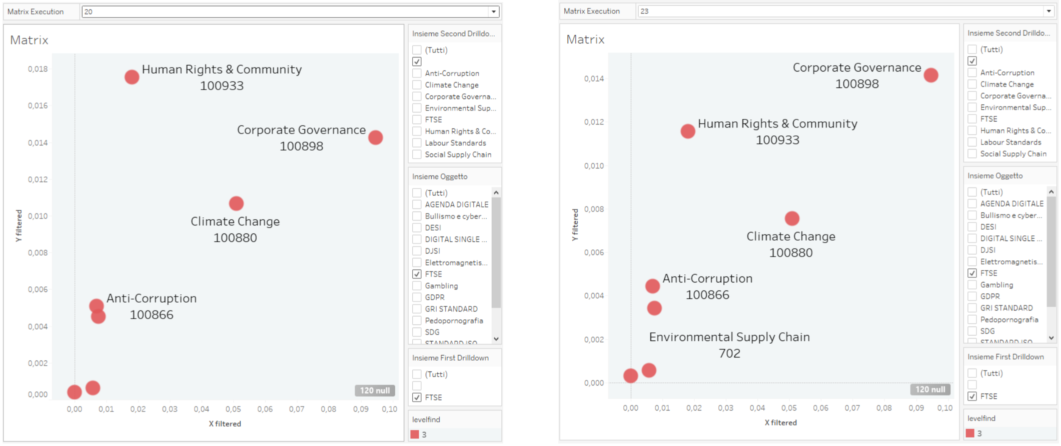
(a) Matrix at the first level

(b) Matrix at the second level

(c) Matrix at the third level

(d) Matrix at the fourth level

Figure 4.4: Four different matrices at four different levels of the hierarchy



Matrix Execution 20

Insieme Second Drilldo...
☐ (Tutti)
☑
☐ Anti-Corruption
☐ Climate Change
☐ Corporate Governa...
☐ Environmental Sup...
☐ FTSE
☐ Human Rights & Co...
☐ Labour Standards
☐ Social Supply Chain

Insieme Oggetto
☐ (Tutti)
☐ AGENDA DIGITALE
☐ Bullismo e cyber...
☐ DESI
☐ DIGITAL SINGLE ...
☐ DJSI
☐ Elettromagnetis...
☑ FTSE
☐ Gambling
☐ GDPR
☐ GRI STANDARD
☐ Pedopornografia
☐ SDG
☐ STANDARD ISO

Insieme First Drilldown
☐ (Tutti)
☐
☑ FTSE

levelfind
🟥 3

(a) Matrix at execution 20

Matrix Execution 23

Insieme Second Drilldo...
☐ (Tutti)
☑
☐ Anti-Corruption
☐ Climate Change
☐ Corporate Governa...
☐ Environmental Sup...
☐ FTSE
☐ Human Rights & Co...
☐ Labour Standards
☐ Social Supply Chain

Insieme Oggetto
☐ (Tutti)
☐ AGENDA DIGITALE
☐ Bullismo e cyber...
☐ DESI
☐ DIGITAL SINGLE ...
☐ DJSI
☐ Elettromagnetis...
☑ FTSE
☐ Gambling
☐ GDPR
☐ GRI STANDARD
☐ Pedopornografia
☐ SDG
☐ STANDARD ISO

Insieme First Drilldown
☐ (Tutti)
☐
☑ FTSE

levelfind
🟥 3

(b) Matrix at execution 23

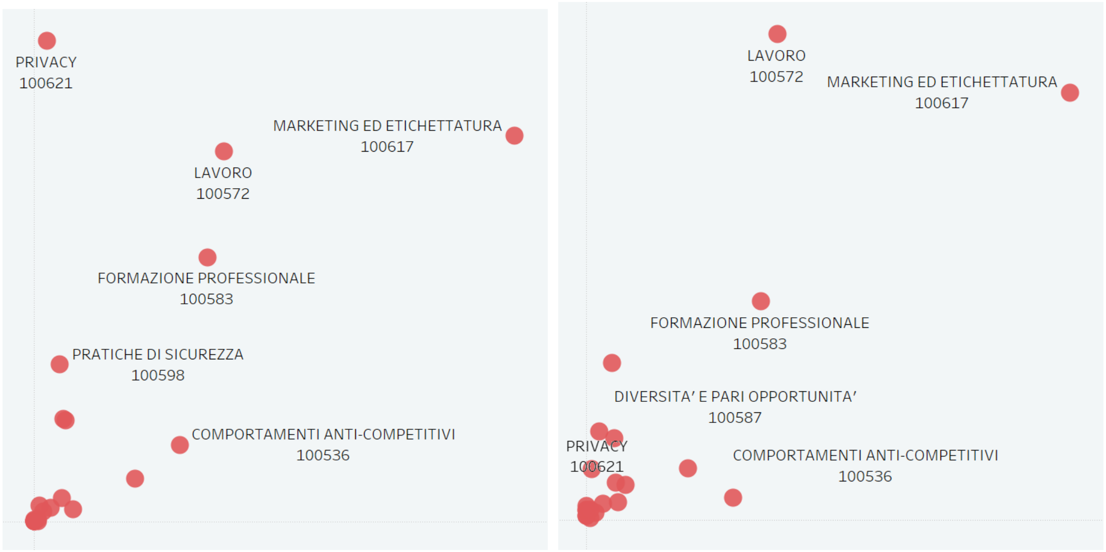
Figure 4.5: Two different matrices at two different executions

vertical axis. The considered hierarchy is the one belonging to the FTSE index and fig. 4.5 visually indicates those topics that have their value changed, i.e. "Human Rights & Community" and "Corporate Goverance". To be noticed, in this case topics values change a lot so that their importance for stakeholders is inverted and their y-position in fig. 4.5a and fig. 4.5b is strictly different.

Moreover, it is possible to observe that by changing the number of actors considered also results would change: for example, there can be a lot of stakeholders as in execution number 28 where we have multiple stakeholders, or a few ones as in execution number 27 where we have only three actors (two of the category stakeholder and one for the category company). Figure 4.6 shows an example of the GRI hierarchy at the second level represented in the materiality matrix. There, it is possible to notice that in execution 27, in fig. 4.6a, the topic "Privacy" is very important for the stakeholders, while in execution 28, in fig. 4.6b, it is far less considered. This is a meaningful result because it underlines that consider a few stakeholders' views lead to completely different and possibly wrong results.

4.3.2 Rank results

Rank computation requires two values for each topic, namely local score and global score. Indeed, every standard taken into account leads to an independent rank whose global score is equal to 1. Furthermore, every topic belonging to a specific rank, i.e. a specific standard, has a global score with reference to the whole hierarchy and a local score referring to the sub-hierarchy it belongs to. So, for



(a) Matrix at execution 27

(b) Matrix at execution 28

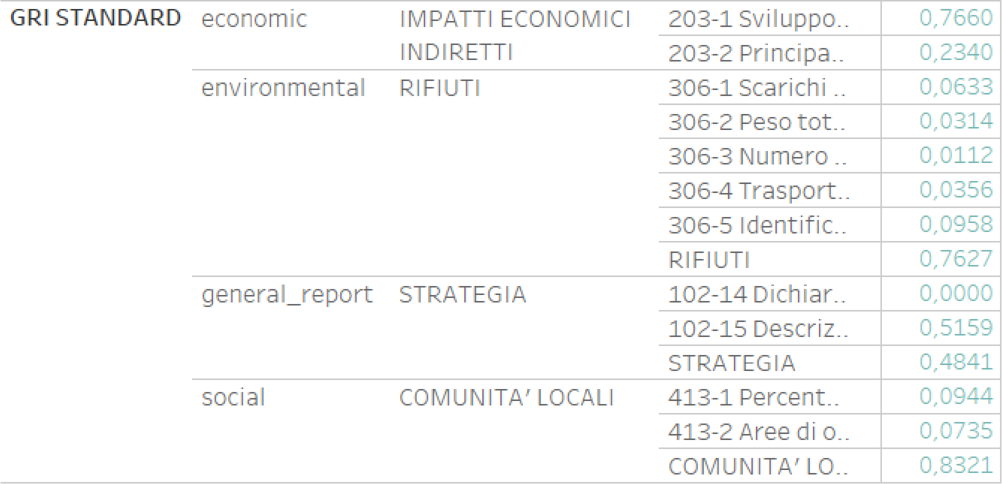
Figure 4.6: Two different matrices at two different executions with a different cardinality of the authors' set

example, the topic "GCG - 4 Number of women on the board" is part of the standard FTSE, is at the fourth level of the hierarchy and with reference to the whole hierarchy has a (global) score of $0,0019$, but within its sub-hierarchy, i.e. the one whose participants are its siblings, it has a (local) score of $0,0055$[3]. Summing up all the local scores of the topics of a specific sub-hierarchy we get 1. Tests on performance have shown that it takes ≈ 8.4 minutes to compute values for topics from all the standards in the database (15).

Tableau allows to represent ranks in a table and, exploiting the features described in 4.3, it is possible to dynamically and asymmetrically expand the hierarchy to analyse results. Figure 4.7 shows an extract of the GRI hierarchy partially and asymmetrically expanded. There, it is possible to see i) all the components of the first and second level of the hierarchy; ii) only one element for the third level and iii) all of the children of the third-level-topic selected with their related local score. This dynamic drill down, as for the matrix, is achievable by left-clicking on the single rows to expand or by left-clicking on the menu on the right (see fig. 4.3) and selecting the topics requested. Finally, the global score appears with other information with a hover upon each topic row.

To show an example of a drill down interaction, fig. 4.8 represents the GRI

[3] Data are taken from execution 25



GRI STANDARD	economic	IMPATTI ECONOMICI INDIRETTI	203-1 Sviluppo..	0,7660
			203-2 Principa..	0,2340
	environmental	RIFIUTI	306-1 Scarichi ..	0,0633
			306-2 Peso tot..	0,0314
			306-3 Numero ..	0,0112
			306-4 Trasport..	0,0356
			306-5 Identific..	0,0958
			RIFIUTI	0,7627
	general_report	STRATEGIA	102-14 Dichiar..	0,0000
			102-15 Descriz..	0,5159
			STRATEGIA	0,4841
	social	COMUNITA' LOCALI	413-1 Percent..	0,0944
			413-2 Aree di o..	0,0735
			COMUNITA' LO..	0,8321

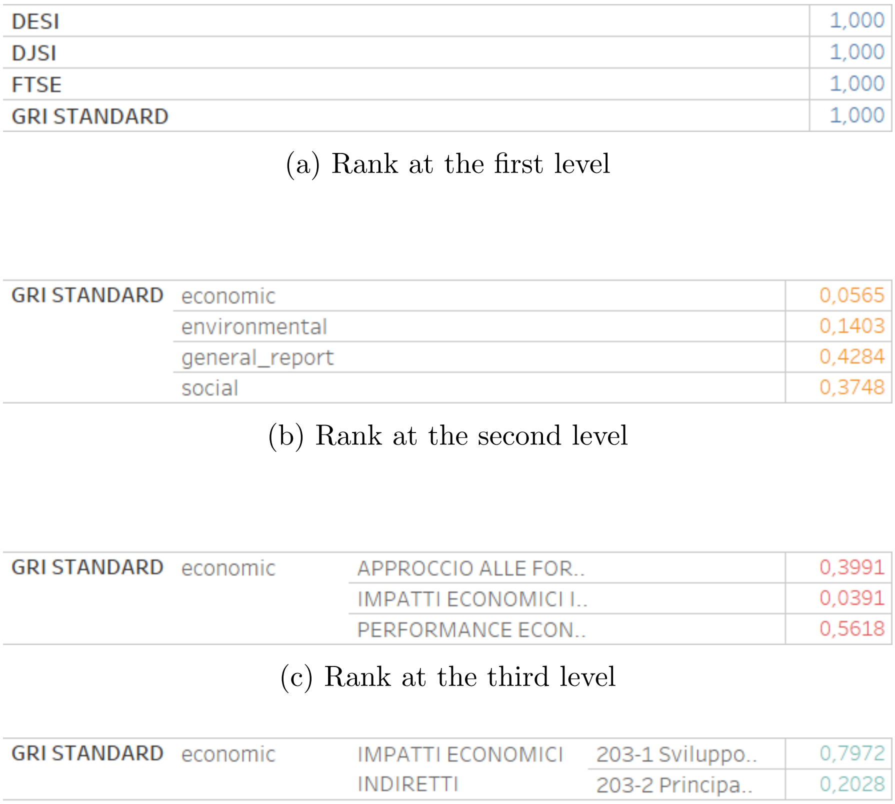
Figure 4.7: Asymmetrically expanded GRI hierarchy

topic_name	*exec_id*	weight			LS	*GS*
		social	stakeholder_17	stakeholder_1		
Comunicazioni	18	0.5	1	0.25	0.063	0.025
	25	0.8	0.2	1	0.065	0.029
Contenuto	18	0.5	1	0.25	0.845	0.336
	25	0.8	0.2	1	0.870	0.3844
Transazioni	18	0.5	1	0.25	0.031	0.012
	25	0.8	0.2	1	0.037	0.016
Uso di internet	18	0.5	1	0.25	0.059	0.023
	25	0.8	0.2	1	0.025	0.011

Table 4.3: An extract of changes occurred in DESI rank after actors' weights modification

hierarchy explored at different levels. Indeed, from fig. 4.8a by left-clicking on the "GRI standard" row, you can see fig. 4.8b with all the topics of the second level of the hierarchy and their related local score. By left-clicking on the "economic" row in fig. 4.8b, you can see all the topics of the third level belonging to the *economic* sphere and their related local score in fig. 4.8c. Finally, by left-clicking on the "impatti economici indiretti" row in fig. 4.8c you obtain its two topics and their own local score.

Different parameters could lead to very different results, and so, for example, a topic that is very important to an influential stakeholder will be evaluated with a high local score and global score, but changing the actor's weight (and salience as a consequence) can turn the situation around. To give an example, table 4.3 shows how, varying some actors' weights, the global score and the local score of topics belonging to the fourth level of the Digital Economy and Society Index (DESI)



DESI	1,000
DJSI	1,000
FTSE	1,000
GRI STANDARD	1,000

(a) Rank at the first level

GRI STANDARD	economic	0,0565
	environmental	0,1403
	general_report	0,4284
	social	0,3748

(b) Rank at the second level

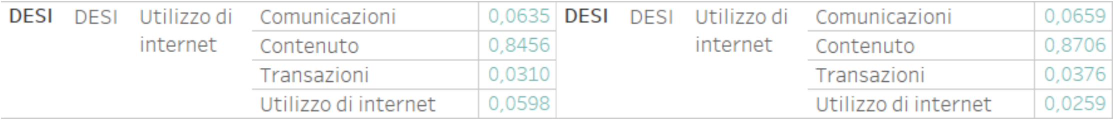
GRI STANDARD	economic	APPROCCIO ALLE FOR..	0,3991
		IMPATTI ECONOMICI I..	0,0391
		PERFORMANCE ECON..	0,5618

(c) Rank at the third level

GRI STANDARD	economic	IMPATTI ECONOMICI INDIRETTI	203-1 Sviluppo..	0,7972
			203-2 Principa..	0,2028

(d) Rank at the fourth level

Figure 4.8: Four different ranks at four different levels of the hierarchy



DESI	DESI	Utilizzo di internet	Comunicazioni	0,0635
			Contenuto	0,8456
			Transazioni	0,0310
			Utilizzo di internet	0,0598

(a) DESI rank at execution 18

DESI	DESI	Utilizzo di internet	Comunicazioni	0,0659
			Contenuto	0,8706
			Transazioni	0,0376
			Utilizzo di internet	0,0259

(b) DESI rank at execution 25

Figure 4.9: Two different matrices at two different executions

hierarchy (sub-category “Utilizzo di internet”) have clearly changed. Indeed, the topic “Transazioni” which was less important than “Uso di internet” in execution 18, became more important in execution 25.

Figure 4.9 shows the above-mentioned results in the Tableau interface.

4.3.3 Comparison of different standards

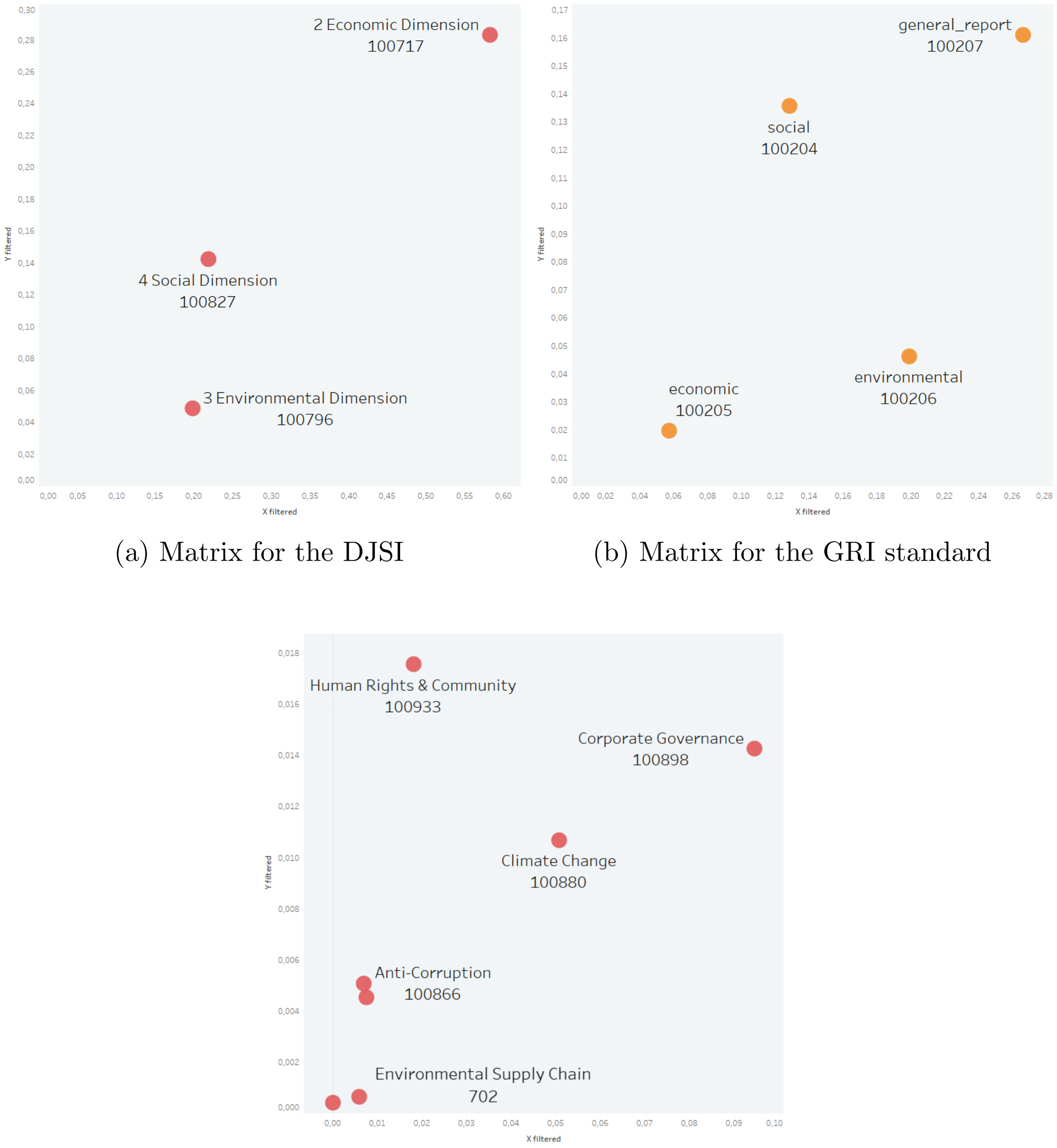
Unfortunately, for this thesis work it has not been possible to obtain the ground-truth data (i.e., the output of the manual approach to materiality assessment) that would have allowed a final validation of the proposed automatic approach. Despite this, some remarks can be done, especially comparing the results obtained for every standard to each other. Indeed, every standard has its own rank and materiality matrix where topics are organised and prioritised. Even if the standards are different and use different names to describe their topics, they often deal with data sharing a similar meaning. To give an example, the DJSI and the GRI share topics like “social”, “environmental and “economic”, whereas the FTSE has not such a division but it has topics like “Human Rights & Community” that can be seen as a social topic, or “Climate change” that can be attributed to an environmental topic, and so on. After a comparison between related topics belonging to different hierarchies, it has become evident that data are organised in the matrix and in the rank approximately in the same way: if a topic is important in a hierarchy it is important also in the others.

Figure 4.10 shows an example of three standards where it can be clearly seen that topics sharing the same semantic meaning are positioned in the same sector of the materiality matrix. Indeed, the topics related to *environmental* are more important for the company than for stakeholders, for which it is not so interesting. On the contrary, the social topic is far more interesting for stakeholders than for the reporting company. Looking at the most important topic for both stakeholders and company (upper-right side of the matrix) it seems to be inconsistent among the three considered standards but it is actually not. Indeed, what is considered “economic dimension” for the DJSI, fig. 4.10a, deals with topics like “2.1 Corporate Governance” (see fig. 4.11a) that is the argument highlighted in the other two

Standard	*Score*							
	economic		enviromental		social		corporate	
ISO STANDARD 26000		-	3°	0.201	2°	0.266	1°	0.533
SUSTAINALITICS		-	2°	0.344	3°	0.146	1°	0.510
SDG	3°	0.139	2°	0.300	1°	0.561		-
DJSI	4°	0.077	3°	0.106	2°	0.297	1°	0.520
FTSE		-	3°	0.220	2°	0.321	1°	0.459
GRI STANDARD	4°	0.055	3°	0.134	2°	0.368	1°	0.443

Table 4.4: Comparison among topics from different standards

standards, i.e. “corporate governance”, fig. 4.10c and “general_report”, fig. 4.10b, which in turn contains topics such as “governance”, see fig. 4.11b.

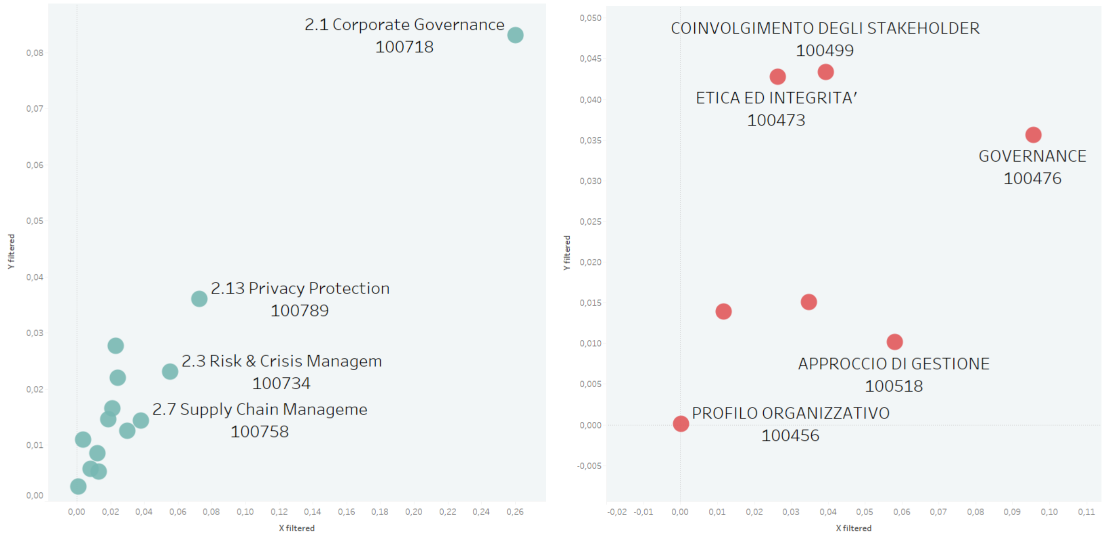


(a) Matrix for the DJSI

(b) Matrix for the GRI standard

(c) Matrix for the FTSE

Figure 4.10: Comparison among three different standards matrices



(a) Matrix for the DJSI

(b) Matrix for the GRI standard

Figure 4.11: Comparison among two standards at a deeper level of the hierarchy. In fig. 4.11a it is shown the level below the topic “2.1 Economic Dimension”; in fig. 4.11b it is shown the level below the topic “general_report”

Conclusions

The main objective of this thesis project was to develop a tool useful to simplify and automate the gathering process and the prioritisation of results coping with the definition of a complete and objective sustainability report. It was firstly conducted a deeper study and analysis of main research and papers at the state of the art; many well-defined standards in the sustainability environment were examined in order to comprehend what is the right way to draw up a sustainability report and to define what are the topics that should be discussed and prioritised. Later, it was necessary to search for practical approaches that propose a structured procedure to gather information from stakeholders and formulas to be used once opinions are retrieved and final results have to be computed. Then, it was conducted a re-organisation of the available data which were obtained by the application of a text mining approach. The main part of this dissertation involved the formalisation of the approach aimed at manipulating and prioritising data, and then its implementation. More in details, this thesis drew inspiration from two papers, [2] and [4], to propose a more automatised way to conduct a materiality analysis. During the final steps, tests are conducted to evaluate the text mining algorithm performance. Text mining techniques are used to extract structured knowledge from unstructured data and to organise it in a way that would be effectively used by the approach proposed for the materiality analysis. Lastly, visual results are presented and discussed highlighting how it is possible to further manipulate the visualisation to perform a more in-depth analysis. Accomplished results of this dissertation mainly concern an efficiently and detailed formalisation of the proposed approach to easily reproduce and reuse it. Moreover, it has been underlined how it is possible to make fruitful reasoning through simple data manipulations. Unfortunately, the limitation of this thesis work is the absence of a ground-truth to validate the results against the output of the standard manual approach. However, this thesis clearly illustrates the possibility to carry out materiality analysis by neither investing enormous sums nor wasting a lot of time and resources, thus encouraging Small and Medium Enterprises in pursuing this kind of approach. Indeed, following the proposed and illustrated approach, it is simply necessary to provide textual documents considered as representative of the interests and the opinions of both the reporting company and its stakeholders. Once these documents are retrieved, the implemented procedure provides fast results organised so that they are easily accessed and analysed. Lastly, even if this is not the first structured approach proposed to accomplish a productive materiality analysis and even if it has been inspired by the above-mentioned approaches, this thesis work is conceived to automatise, simplify and make impartial the entire process.

To better understand the implications of the achieved results, future studies could address more in-depth analysis and discussions. Specifically, the most important explorations are listed below.

- The main limitation of this thesis should be also the first concept to further develop. Indeed, it is necessary to obtain data against which match the results obtained through the approach described in this work. The comparison should concentrate on whether the most important topics for both stakeholders and the reporting company are identified as the most material by the implemented procedure. The idea is to analyse a sustainability report drawn up by the same reporting company that provided the textual documents processed and so searching for correspondences.

- The text mining approach realised may be re-organised to make it more efficient and complete. Indeed, starting from the aliases used to understand whether topics occur or not, it could be easier and more effective to use more sophisticated techniques for textual analysis. For example, a topic can be directly searched checking for its synonymous and hypernyms or by exploiting Latent Semantic Analysis techniques. However, a more advanced procedure must perform as well as the implemented one: the actual approach is very efficient and effective so this can be the lower bound for performance.

- It may be convenient and useful to exploit co-occurrence relationships among semantically linked topics to give a more accurate interpretation of how much a document deals with certain topics.

- It can be interesting to weight different occurrences based on the textual context where they are rescued. In particular, it could be analysed how a topic is used within a sentence through advanced Natural Language Processing techniques. Moreover, sometimes could happen that the same topic, rescued in a different positions in a document, could be more valuable than in another position. So, it is necessary to perform a more complex analysis by considering not only the occurrences of a topic in a document, but also its context.

Appendix

Implementation

Implementation starts with the technologies choice. More in details, it was necessary to interact with an Oracle database, implementative requirement in 3.3.2, and since Scala is the chosen programming language, the ojbc drivers[4] fitted perfectly. Indeed, they expose a simple and reliable tool to connect to an Oracle database and build a Java-based application.

In order to maintain a sort of clean organisation, the following packages are created:

- *bellantuono*, it contains all the code useful to compute the materiality matrix (e.g. `BellantuonoAlgorithm`);
- *calabrese*, it contains all the code useful to compute the materiality rank (e.g. `CalabreseAlgorithm`);
- *dbinteraction*, it contains all the code concerning interactions with the Oracle database. It is in turn subdivided in two sub-packages, namely *connection* and *schemas*. The former deals with database connection while the latter defines data schemas as described in fig. 3.3.

As a start, existing tables are manipulated to obtain the ones described in section 3.3.1. Most of the changes are done via Scala, so they are simply repeatable and embedded in the program. In particular, the `parent_id` value is filled for each topic analysing links among those columns that are used to reconstruct the hierarchy, e.g. `oggetto` or `i_livello`. Moreover, every topic that is a father and with some occurrences is pushed down through the hierarchy and it is created a duplicate with no occurrence and with its same values except for the id. This is done to simply compute the importance of that specific topic both considering and not considering the importance of its children.

[4] https://www.oracle.com/it/database/technologies/appdev/jdbc.html

The subsequent steps in the implementation concern how to connect to the Oracle database and how to retrieve and save data. All the necessary data are retrieved and stored in memory during the execution of the program. Scala maps are used to link an id with its related data, so, for example, topics retrieved from table `Topic` are saved as key-values entries where the key is the id of the topic and the value consists of the other columns structured as an element of the `TopicData` class described in section 3.3.3.

Once data retrieval and memorisation is completed, the common functions used to compute both rank and matrix are implemented. Section 3.1 stresses the recursive nature of the approach, and so the implementation follows the same steps and, exploiting Scala functional behaviour, it is done in a easy and readable way. Firstly, independent methods are implemented:

- `children`, to get the children of a specified topic, it is necessary to keep, among all the topics retrieved, only those topics whose `father_id` is equal to the specified topic;
- `author` and `words`, these are simple methods retrieving information directly from columns saved from the database, namely `actor_id` and `tot_entity`;
- `directOccurrences`, this method computes how many times an author has mentioned a specific topic. This result is accomplished counting the occurrences of all the clips whose author is the specified one and where the specified topic occurs;

After that, dependent functions are analysed and implemented:

- `recursiveOccurrences`, it uses `directOccurrences` and exploits `children` to call itself recursively on the children of the specified topic;
- `speechVolume`, simply exploits `words` aggregating all the clips written by the specified author.

The main methods of this phase are those representing the concepts of *score*, *salience* and, definitely, *average score*. Indeed, they are directly derived from the previous results, as shown in def. 3.4, 3.5 and 3.6, and they are used to compute algorithm-specific results. More in details:

- `salience` is computed using the `speechVolume` for an actor with respect to the maximum of speech volume between all actors and multiplying it by a `weight` factor stored in the database;
- `score` is computed dividing the recursive occurrences of a specified topic and actor by the speech volume of that actor;

- `avgScore` is directly derived from `score` and computed averaging the value of the score of a specified topic by the score given to that topic by a specified set of authors.

To compute the materiality matrix all the methods implemented are far enough: the x-axis is computed using the score of the selected topics according to the company actors; while to compute the y-axis it is sufficient to average the score of the selected topics by authors that are not *company*.

The materiality rank needs two other methods to be realised. More in details it is necessary to implement methods able to compute:

- `siblings` of a certain topic to understand the *local score* of that topic. The local score deals with the average score of a certain topic with respect to its siblings' average score in the hierarchy;

- `parent` of a certain topic to compute the *global score* of that topic. Global score is used to show the materiality of a topic with reference to all the other topics.

At this point, both `localScore` and `globalScore` methods can be implemented and so the approach is completed.

To respond to req. 6. it is created an `Uploader` object capable of uploading, at each execution, results such as salience, rank (i.e. local and global scores), matrix (i.e. x-values and y-values) and scores (i.e. direct occurrences, recursive occurrences, score and average score)

Launching

To launch the application it is necessary to specify different parameters which lead to different execution modalities:

1. *algorithm*, the first parameter indicates what algorithm you want to execute. Allowed values are “bellantuono” or “calabrese”. The default value is “calabrese”;

2. *standard*, the second parameter is used to choose a set of standards to execute the algorithm for. Recognised values are those standards which are in the database, or, if you want to execute an algorithm for all the standards in the database at once, you can specify a set with only the value “all”. The default value is a set of “GRI STANDARD”;

3. `level`, the third parameter is used only if the chosen algorithm is *Bellantuono* and indicates what level in the hierarchy should be considered when

computing axes values. Allowed values are those of the column *level* in the `Topic` table or, if you want to compute the matrix for all the levels in the hierarchy, you can specify the value “all”. The default value is “L3”;

4. `commonConfiguration`, the fourth parameter is to indicate what are the tables name to be considered in the execution. Default values are “clip”, “topic_nrc”, “actor” and “occurrence”;

5. `computeSalience`, the fifth parameter indicates whether or not compute the author’s salience. Allowed values are “true” or “false”; the default value is “false”.

Testing

Testing is conducted exploiting *ScalaTest*, an open source tool able to simplify and to improve effectiveness. In this thesis, it is used to test the correctness of the methods defined in `Common` and `CalabreseAlgorithm` which are necessary to compute results. Tests are made reading data from four *csv* files (simulating the database tables) created with an ad-hoc sample of data: data in these files are inserted trying to cover all the possible criticalities. To make testing possible, it is necessary to create a different instance of the program, where the fourth parameter indicates that data have to be retrieved from the specified files, instead that from tables.

Uploading phase is not tested since it can be directly checked looking at uploaded data in the target tables.

Acronyms

AA1000 AccountAbility 1000. 10

AHP Analytic Hierarchy Process. 13, 14, 16

ANP Analytic Network Process. 13

APS AccountAbility Principles. 10

AS Assurance Standard. 10

CERES Coalition for Environmentally Responsible EconomieS. 4

CSR Corporate Social Responsibility. 2, 11

CSV Creating Shared Value. 2, 3

DESI Digital Economy and Society Index. 58, 60

DJSI Dow Jones Sustainability Indices. 11, 14, 23, 41, 43, 60, 62, 63

FMEA Failure Modes and Effects Analysis. 12

FTSE Financial Times Stock Exchange. 56, 57, 60, 62

GRI Global Reporting Initiative. ii, ix, 3–10, 14, 18, 23, 33, 37, 38, 41, 43, 53, 56–58, 60, 62, 63

IR Integrated Reporting. 11

ISEA Institute of Social and Ethical Accountability. 9, 10

LOD Level Of Detail. 52

MAGDM Multi-Attribute Group Decision. 18

MCDM Multi-Criteria Decision Making. 11–13, 18

RPN Risk Priority Number. 12, 14

SBI Social Business Intelligence. 22

SES Stakeholder Engagement Standard. 10

SMEs Small and Medium Enterprises. 11, 17, 19, 22, 65

TFN Triangular Fuzzy Number. 16

WCED World Commission for Environment and Development. 1

Glossary

corporate sustainability A way for corporations to continue planning and pursuing an economic growth, but more focusing on their commitment into societal sphere, specifically that relating to sustainable development, e.g. environmental safety, social justice and equity, and economic development. 1–3

materiality analysis It is an analysis outlining what are the material aspects that a company has to take into account and largely describe during its sustainability report. i, ii, 1, 3, 4, 9, 11, 12, 14, 18, 21–23, 65, 66

materiality matrix A visual representation of the most relevant issues rescued during the materiality analysis. It generally consists of two axes: the horizontal axis identify what is material according to the company insights; the vertical axis, instead, represents what is material according to involved stakeholders.. ix, 4, 9, 11, 18, 19, 38, 39, 43, 44, 47, 53, 54, 67, 69

reporting company Companies in charge to draw up a sustainability-report. 4, 6, 13, 18, 22, 23, 34, 39, 42, 49, 60, 66

stakeholder It is any group or individual (e.g. employees, clients, pressure groups, communities, etc.) who can affect or is affected by the achievement of the organisation's objectives. i, ii, 2–6, 9–11, 13, 14, 16, 18, 19, 33, 38, 39, 42, 56

strategic engagement Strategic Engagement is a method of finding, attracting, and keeping the best customers for your business or organisation. Strategic Engagement utilises science and technology combined with creativity and psychology to achieve efficient and sustainable results. 13

sustainability report It is an annual report firms draw up about their corporate sustainability involvement. i, 1–5, 10–13, 17–19, 22, 33, 42, 65, 66

triple bottom line The “bottom line” traditionally refers to the monetary profits that a company has made. The “triple bottom line” adds two more “bottom line”: social and environmental (ecological) concerns, around which develop the company’s reports. 2, 3, 6

Bibliography

[1] I. Ahmad Sabri. Rating and ranking criteria for selected islands using fuzzy analytic hierarchy process (fahp). *Int. J. Appl. Math. Inform.*, 07 2019.

[2] N. Bellantuono, P. Pontrandolfo, and B. Scozzi. Capturing the stakeholders' view in sustainability reporting: a novel approach. *Sustainability*, 8(4):379, 2016.

[3] H. Bowen. *Social Responsibilities of the Businessman.* University of Iowa Press, 2013.

[4] A. Calabrese, R. Costa, N. Levialdi, and T. Menichini. A fuzzy analytic hierarchy process method to support materiality assessment in sustainability reporting. *Journal of Cleaner Production*, 121:248–264, 2016.

[5] A. Calabrese, R. Costa, N. Levialdi, and T. Menichini. Materiality analysis in sustainability reporting: a method for making it work in practice. *European Journal of Sustainable Development*, 6(3):439–439, 2017.

[6] M. Cinelli, S. R. Coles, and K. Kirwan. Analysis of the potentials of multi criteria decision analysis methods to conduct sustainability assessment. *Ecological indicators*, 46:138–148, 2014.

[7] UNWCED: United Nations World Commission and Development. *Our Common Future (Brundtland Report).* Butterworth-Heinemann, 1987.

[8] G. Everest. Basic data structure models explained with a common example. *Fifth Computing Conference, IEEE*, 10 1976.

[9] Fazio. Come cambia il mondo e come evolve la professione contabile. `https://www.odcec.mi.it/docs/default-source/default-document-library/materiale-didattico-a-cura-dei-relatoria4bd34714cc168548164ff0000ef0ce1.pdf?sfvrsn=0`, 2019. Accessed: March 16, 2021.

[10] R. Feldman and J. Sanger. *The text mining handbook: advanced approaches in analyzing unstructured data.* Cambridge university press, 2007.

[11] Matteo Francia, Matteo Golfarelli, and Stefano Rizzi. A methodology for social BI. In Bipin C. Desai, Ana Maria Almeida, Jorge Bernardino, and Elsa Ferreira Gomes, editors, *18th International Database Engineering & Applications Symposium, IDEAS 2014, Porto, Portugal, July 7-9, 2014*, pages 207–216. ACM, 2014. doi: 10.1145/2628194.2628250. URL `https://doi.org/10.1145/2628194.2628250`.

[12] Matteo Francia, Enrico Gallinucci, Matteo Golfarelli, and Stefano Rizzi. Social business intelligence in action. In Selmin Nurcan, Pnina Soffer, Marko Bajec, and Johann Eder, editors, *Advanced Information Systems Engineering - 28th International Conference, CAiSE 2016, Ljubljana, Slovenia, June 13-17, 2016. Proceedings*, volume 9694 of *Lecture Notes in Computer Science*, pages 33–48. Springer, 2016. doi: 10.1007/978-3-319-39696-5_3. URL `https://doi.org/10.1007/978-3-319-39696-5_3`.

[13] R. E. Freeman. *Strategic Management: A Stakeholder Approach.* Boston: Pitman, 1984.

[14] Enrico Gallinucci, Matteo Golfarelli, and Stefano Rizzi. Advanced topic modeling for social business intelligence. *Inf. Syst.*, 53:87–106, 2015. doi: 10.1016/j.is.2015.04.005. URL `https://doi.org/10.1016/j.is.2015.04.005`.

[15] C. Hsu, W. Lee, and W. Chao. Materiality analysis model in sustainability reporting: a case study at lite-on technology corporation. *Journal of cleaner production*, 57:142–151, 2013.

[16] P. Jones, D. Comfort, and D. Hillier. Managing materiality: a preliminary examination of the adoption of the new gri g4 guidelines on materiality within the business community. *Journal of Public Affairs*, 16(3):222–230, 2016. doi: 10.1002/pa.

[17] A. Kumar, B. Sah, A. R. Singh, Y. Deng, X. He, P. Kumar, and R.C. Bansal. A review of multi criteria decision making (mcdm) towards sustainable renewable energy development. *Renewable and Sustainable Energy Reviews*, 69:596–609, 2017. ISSN 1364-0321. doi: https://doi.org/10.1016/j.rser.2016.11.191. URL `https://www.sciencedirect.com/science/article/pii/S1364032116309479`.

[18] A. Morgan. *Eating the big fish: How challenger brands can compete against brand leaders.* John wiley & sons, 2009.

[19] M. Porter and M. Kramer. The big idea: Creating shared value. how to reinvent capitalism—and unleash a wave of innovation and growth. *Harvard Business Review*, 89:62–77, 01 2011.

[20] T. Saaty. Analytic hierarchy process. *Wiley statsRef: Statistics reference online*, 2014.

[21] T. Saaty and L. Vargas. *The Analytic Network Process*, pages 1–26. Springer, Boston, MA, 09 2006. ISBN 978-0-387-33859-0. doi: 10.1007/0-387-33987-6_1.

[22] C. Saenz. Creating shared value using materiality analysis: Strategies from the mining industry. *Corporate Social Responsibility and Environmental Management*, 26(6):1351–1360, 2019.

[23] R. Sleeper. Tableau tip: Single row drilldown "deluxe". `https://playfairdata.com/tableau-tip-single-row-drilldown-deluxe/`, 2019. Accessed: March 16, 2021.

[24] B. Wells. Scatter plot drill down. `https://www.thedataschool.co.uk/ben-wells/scatter-plot-drill-down`, 2019. Accessed: March 16, 2021.

[25] Willaert. Understanding accountability with aa1000 – an interview with claire hart. `https://dqs-cfs.com/2015/03/understanding-accountability-with-aa1000-an-interview-with-claire-hart/`, 2015. Accessed: March 16, 2021.

[26] M. Wilson. Corporate sustainability: What is it and where does it come from? `https://iveybusinessjournal.com/publication/corporate-sustainability-what-is-it-and-where-does-it-come-from/`, 2003. Accessed: March 16, 2021.

[27] P. Wójcik. How creating shared value differs from corporate social responsibility. *Journal of Management and Business Administracton. Central Europe*, 24:32–55, 06 2016. doi: 10.7206/jmba.ce.2450-7814.168.

[28] R. R. Yager. A procedure for ordering fuzzy subsets of the unit interval. *Information Sciences*, 24(2):143–161, 1981. ISSN 0020-0255. doi: https://doi.org/10.1016/0020-0255(81)90017-7. URL `https://www.sciencedirect.com/science/article/pii/0020025581900177`.

[29] L. A. Zadeh. The concept of a linguistic variable and its application to approximate reasoning—i. *Information Sciences*, 8(3):199–249, 1975. ISSN 0020-0255. doi: https://doi.org/10.1016/0020-0255(75)90036-5. URL `https://www.sciencedirect.com/science/article/pii/0020025575900365`.

[30] L. A. Zadeh. Fuzzy logic and approximate reasoning. *Springer Link*, pages 407–428, 1975. doi: https://doi.org/10.1007/BF00485052. URL `https://link.springer.com/article/10.1007/BF00485052`.