# TEXT MINING FOR GLOBAL REPORTING INITIATIVE (GRI) STANDARDS: STUDY OF NORDIC LISTED COMPANIES

Marcelo G. Gutierrez B.

Master's Thesis
Master of Engineering - Big Data Analytics
May 15, 2020

| MASTER'S THESIS | |
|---|---|
| Arcada University of Applied Sciences | |
| | |
| Degree Programme: | Master of Engineering - Big Data Analytics |
| | |
| Identification number: | 7752 |
| Author: | Marcelo G. Gutierrez B. |
| Title: | Text Mining for Global Reporting Initiatives (GRI) Standards: Study of Nordic Listed Companies. |
| Supervisor (Arcada): | Leonardo Espinosa Leal |
| | |
| Commissioned by: | |
| | |


Abstract:

This thesis investigates using text mining methods, if Corporate Social Responsibility reports have been published following the guidelines of the Global Reporting Initiative Standards. To achieve this goal, several text mining techniques were implemented such as Latent Dirichlet Allocation (LDA), Transformation into inverse frequency (TF-idf), FastText, Global Vectors (gloVE), Latent Semantic Index (LSA), wor2vec and doc2vec. These results are analysed from an Unsupervised Learning perspective. We extract string, corpus and hybrid semantic similarities and we evaluated the models trough the intrinsic assessment methodology. Index matching was develop to complement the semantic valuation. The final results show us that LSA and gloVE as the best option for our study of Nordic listed companies.



| Keywords: | Text mining, sustainability, Semantic similiarity, GRI reports, Corporate Social Responsability, Machine Learning. |
|---|---|
| Number of pages: | 79 |
| Language: | English |
| Date of acceptance: | |

# CONTENTS

## Figures

## Tables

# 1 INTRODUCTION

The objective of this research is to evaluate the degree of affinity that Nordics companies' report published under the Global Reporting Initiatives (GRI) framework have. The grade of affinity will be the result of the combination of two text mining methods. Therefore this work implement Natural Language Processing (NLP) and Information Retrieval (IR) methods in order to obtain the semantic similarity and matching disclosures respectively. The combination of this methods can give us somewhat clues to know which documents are following GRI guidelines and to what degree.

Currently, the Corporate Social Responsibility reports (CSR), whose most important referent is the GRI standards, are considered as a decision investment factor comparable to the company's financial statements. The CSR not only represent companies' commitment to Environmental, Social and Governance (ESG) practices, or engagement to the UN 2030 agenda, the CSR is a benchmark of the real economic health of a company in long-term. (Servaes et al., 2017). Even in many stock markets in emerging countries[1], the submission of these reports is periodic and mandatory. It is estimated that the lack of regulation and consensus of these frameworks creates a gap of USD 12 billion of direct investment on sustainability[2]. Complying with these frameworks is voluntary and does not require much detail, the majority of reports are presented in an unstructured way and therefore there is no other alternative than to use text mining techniques for extract some knowledge of them.

Since the GRI framework removed the rating that was weighted on these documents from the G4 versions, we ran into an Unsupervised Learning challenge. In this way we begin to implement text mining methods to extract the degree of semantic similarity that the texts published by the companies have under the guidelines published by the GRI institution. GRI is the entity in charge of promoting, maintain and modify these standards. New models were trained and pretrained models were implemented, a methodology for evaluating words, sentences from different abstractions of the text, corpus and hybrids was developed.

## 1.1 Motivation and theoretical framework

Over recent years, corporations have begun to focus on the corporate social responsibility (CSR) concept, particularly on one of its central platforms – the notion of sustainability and sustainable development. Despite the fact that several researchers have discovered conflicting results between Corporate Social Investments and Corporate Financial Performance (CFP) (Griffin et al, 1997), there is increasing proof that Environmental, Social and Governance (ESG) factors, may deliver significant long-term performance advantages when incorporated into portfolio investment analysis (Wang, et al, 2016).

It is, therefore, important for CSR companies to effectively communicate their economic and ESG performance to their stakeholders. There are several guidelines issued by different organizations for CSR reporting; the most important are Global Reporting

[1] In October of 2019 a coalition of asset managers, public pension funds, and responsible investment organizations filed a petition (https://www.sec.gov/comments/4-711/4-711.htm) with the Securities Exchange Commission (USA) to request that it develop a compressive ESG disclosure framework.
[2] http://www3.weforum.org/docs/WEF_Global_Risks_Report_2019.pdf

Initiative (GRI), Global Compact Issued by UN and ISO 2600. For this study, we selected the GRI versions G3, G4 and GRI Standards. Also in several countries GRI are linked to local regulatory reporting requirements[3] (KPMG, 2017). Since the number of companies and organization reporting their CSR activities is increasing, the current manual process of analyzing the reports demands a lot of effort (Aryal and Nabin, 2014) and is rapidly becoming obsolete.

According to Shahi et al., (2015) the automated CSR report analysis system has been overlooked by the research community, despite the fact that its text categorization and Machine Learning (ML) approach have been the subject of research since their early introduction, with the aim of solving various document analysis problems. Shahi et al., (2015) have produced the only work in this area, using the GRI G3 version. This version used a score grade ranging from A+ to C to measure the effectiveness of the Level Check which was removed from the framework for the GRI G4 version. Presently, a company has two options, or levels, for reporting "in accordance with" the GRI guideline – *core* and "comprehensive" reports. The most substantial difference between a core and a comprehensive report are the number of governances and strategy disclosures. Due to this development, comparing the accuracy of classification is now more difficult. Nevertheless, we can choose to conduct our study in a qualitative method (Wilson and Rayson, 1993). This includes the compilation and classification of quantitative and qualitative data into the GRI guidelines in order to discover similarities within the scope selected (Guthrie & Abeysereka, 2006).

To the best of our knowledge, previous work has never included characteristics of CSR reports in G4 and new standards. This implementation or adaption could increase the value of the evidence that is used to demonstrate the importance that the market places on EGS activities that are captured in a non-systematic way. Therefore, due to the current state of the literature review regarding the implementation of ML to valuate ESG activities, we believe it is important to produce a work that has the ability to discover and analyze the relationship of the GRI reports published by the companies with the GRI official guidelines through text mining.

### 1.2 Limitations

This work is limited only to the context of Nordic Companies that have published on the GRI reports Database using English language.

The final solution designed is mostly an implementation of the tools already built for the application of machines learning techniques, as they are. We do not make an improved or deep review of the available tools.

### 1.3 Ethical considerations

About the data, the data we used in this work is in the public domain. Companies submitted this report voluntarily.

---

[3] KPMG Survey of Corporate Responsibility Reporting provides a very useful insight into the recent trends in CSR reporting. KPMG started publishing such report from 1993. [https://home.kpmg.com/content/dam/kpmg/campaigns/csr/pdf/CSR_Reporting_2017.pdf].

About the results, the models that are created for this proposal will be available for later revisions in order to confirm the veracity of the results.

About the models, we will ensure that Artificial Intelligence model will behave in a way that prioritizes human safety above their assigned tasks. And their own safety and that are also in accordance with accepted precepts of human morality. See Figure 1.


*Figure 1. Ethics guidelines for trustworthy AI by the European Union. (Reprinted: https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines/1)*

### 1.4 The structure

After this introductory section, we will describe the fundamentals of the related tools and their state of the art in Chapter 2. In the third chapter, the environment of the problem is examined in more detail, we will apply Exploratory Data Analysis to obtain a more adjusted vision for the development of the models to implement. The technical and design details of the models, both parameterization, architecture, and execution capacity of the system are revealed in chapter 4. The results of the executions are examined in section 5. Finally, the conclusions and certain suggestions for improvement of this work are made in Chapter 6.

# 2 FUNDAMENTALS AND LITERATURE REVIEW

## 2.1 Text Mining

The concept of *Text Mining (TM)* is still under discussion. But for our research we can define as the process of extracting knowledge or patterns, previously unknown, non-trivial and interesting (potentially 'useful') and understandable by humans from unstructured text sources. Text Mining or knowledge discovery text (KDT), first introduced by Feltham et al., (1995), is an extension of Data Mining (DM) that seeks to extract useful and important information from text. We have to be careful not to think that the classical DM techniques can be directly applied to textual information. DM works with databases with a known scheme e.g. Relational DataBase. Each text document is an ordered collection of words and separators with meaning associated, whose the location in the text is determined by synthetic and semantic constraints. There are semi-structured texts such as documents written in XML or JSON. In TM the data is:

- Inherently unstructured
    - Implicit structure
    - Much greater richness than in structures cases
- Ambiguous
- Multilingual

This absence of structure is the biggest problem of TM and implies the need to pre-process the texts, to move them to an *intermediate form* (Allahyari et al., (2017).

## 2.2 Text representation and encoding

In order to apply text mining it is necessary to represent the content of the documents by a model. One of the most simple but effective and commonly used ways to represent text is using the bag of words (BOW), which considers the number of occurrences of each term (word/phrase) but ignores the order. When using this representation, we discard most of the structure of the input text, like chapters, paragraphs, sentences, and formatting, and only count how often each word appears in each text in the corpus.

This representation leads to a vector representation that can be analyzed with dimension reduction algorithms from machine learning and statistics. Three of the main dimension reduction techniques used in text mining are:

- Latent Semantic Indexing (LSI) (Dumais et al., 1995),
- Probabilistic Latent Semantic Indexing (PLSA) (Hofmann, 1999) and
- Topic models LDA (Blei and Jordan, 2003).

Over a long time, most techniques used to research text mining problems used shallow machine learning models and hand-crafted features (high-dimensional features). However, with the recent popularity and success of word embedding's implemented in Word2vec (Mikolov et al., 2013) low dimensional, distributed representations and deep learning methods (Socher et al., 2013), have achieved superior results on various language-related tasks as compared to traditional machine learning models.

Collbert et al., (2011) as cited in Young et al., (2018) demonstrated how a simple deep learning framework outperforms most approaches in several Natural Language Processing (NLP) tasks such as semantic role labeling (SRL), named-entity recognition (NER), and POS tagging. Since then, numerous complex deep learning based algorithms have been proposed to solve difficult NLP tasks.

### 2.2.1 Text Preprocessing.

The steps involved in a traditional text mining preprocessing comprises:

*Tokenization:* Tokenization is the process of dividing a text document (or a collection of them) into the list of words that make it up, by identifying, for example, blank spaces or punctuation marks on. This step is essential in the subsequent analysis since we are going to use vector space models, in which the words are the basic elements.

*Filtering:* This stage consists of the elimination of the words that do not have interest for the mining of texts, such as articles, pronouns, prepositions, conjunctions, and even words that are used very frequently and do not help the differentiation from one document to another.

*Lemmatization and Stemming:* At these stages, it is assumed that the meaning of the words is not influenced by the grammatical form it presents. Stemming is the process of heuristically removing parts of a word to reduce it to a common form. For its part, lemmatization refers to a more complex process that uses a syntactic analysis, as well as morphological analysis of structures. Therefore, one of the effects of these stages is to reduce the number of words that will appear in our dictionary.

### 2.2.2 Text-Transformation.

The Intermediate Form (See Table 1) is a model of knowledge representation capable of expressing the implicit content of the text in a computable form by means of an algorithm or a program. The technique for obtaining the intermediate form determines the type of information to be obtained in the discovery process (de la Torre, 2005).

*Table 1. The intermediate form*

| **Pre-processing** | **Representation** | **Discovery** |
|---|---|---|
| Categorization | Vector of representative terms | Relationship between terms |
| Full Text Analysis | Sequences of words | Language Patterns |
| Information Extraction | Database table | Relationships between entities |

#### 2.2.2.1 Bag of Words (BOW)

The representation of the text as words is, in the end, a feature engineering or feature engineering task. According to Domingos (2012) and the general literature, in this area,

each individual property of an observed phenomenon is called a characteristic, that is, each input variable in a machine learning algorithm. Many learning models obtain good results and others, on the contrary, do not, and the difference between them is the most important factor: the extraction of these characteristics.

Feature extraction is an essential process in a machine learning model that involves creating the input features for a something-rhythm to work. The main idea of this model is obtained from its own name, "word bag", and it is to transform each word into a number so that the input of the classification algorithm is a vector of numbers, a bag of words, in which each position is a word of the text to classify.

There are two possible ways to carry out this transformation: transform into a vector of occurrences or transform into a vector of frequencies. The reason it is called a bag is that it does not take into account the meaning of each word or its context, that is, it takes each word as an independent number and only takes into account whether or not the word appears in the text. This is one of the main disadvantages of the model, since considering the semantics of words can benefit the result. Another possibility, widely used and perhaps the most powerful, is the use of n-grams, an n-gram being a sequence of elements.

#### 2.2.2.2 Skip-gram

Mikolov (2013) introduced another variant of BOW, Skip-gram, which uses the same architecture but in a contrary way, tries to predict the context based on the word that is to be represented, producing this representation in the process (See Figure 6).

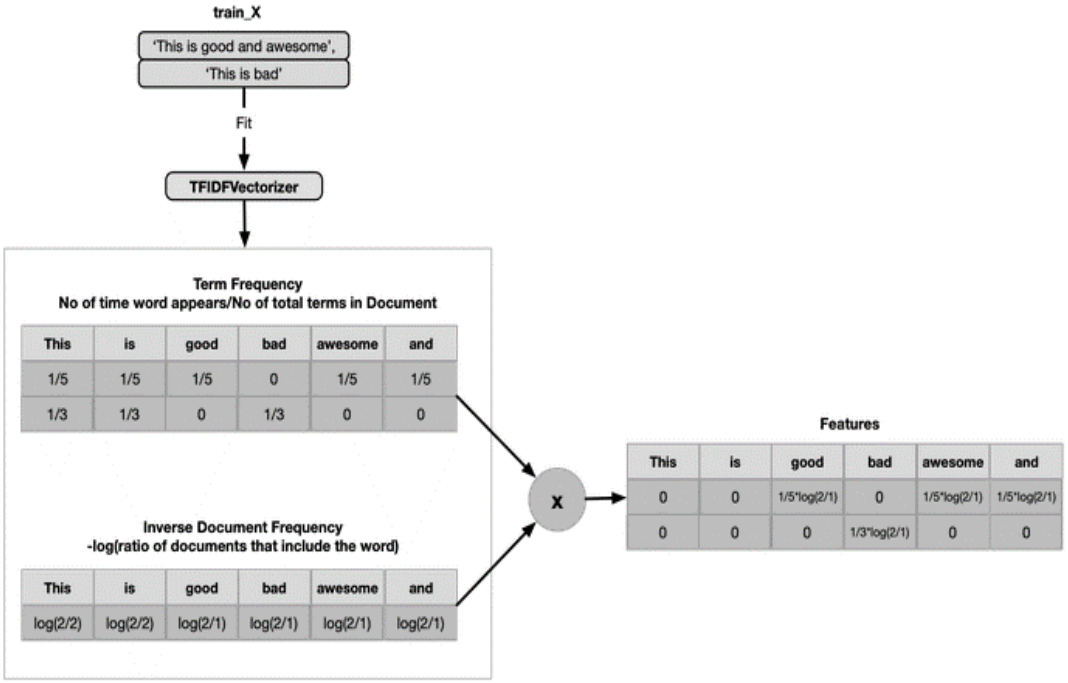
A priori it may seem like you have scattered vectors again, but this is only at the input. The strength of this model is in the weight matrix. When multiplying the input vector by the weight matrix, the result is the vector of the element that has a one, since the rest are zeros. Thus, you have completely dense vectors.

#### 2.2.2.3 Transformation into inverse frequency vector TF-Idf

Inverse document frequency TF-Idf is a numerical measure found from a set of documents $D$ for a term $t$ in a document $d \in D$, which expresses how relevant $t$ is to document $d$ depending on the set of documents. $t$ may consist of one or more consecutive words (*n-gram*) of $d$.

The TF-Idf value increases proportionally to the number of times a word appears in the document, but is compensated for by the frequency of the word in the document collection, which allows for handling the fact that some words are generally more common than others. During the text we call "model", with the same name, referring the group of these measures calculated and ordered from the data and with certain training criteria, along with other attributes or meta-data generated.

TF-Idf is a simple technique to find features from sentences. In the Figure 2, let's consider two things about any word from a document:




- **Term Frequency:** How important is the word in the document?

$$TF(word\ in\ a\ document) = \frac{No\ of\ occurances\ of\ that\ word\ in\ document}{No\ of\ words\ in\ document}$$

- **Inverse Document Frequency:** How important the term is in the whole corpus?

$$IDF(word\ in\ a\ corpus) = -log(ratio\ of\ documents\ that\ include\ the\ word)$$

*Figure 2. TF-IDF Architecture.* *(Reprinted: https://mlwhiz.com/blog/2019/02/08/deeplearning_nlp_conventional_methods/)*

As we can see in the Figure 2, TF-IDF then is just multiplication of these two scores. Intuitively, One can understand that a word is important if it occurs many times in a document. But that creates a problem. Words like "a", "the" occur many times in sentence. Their TF score will always be high. We solve that by using Inverse Document frequency, which is high if the word is rare, and low if the word is common across the corpus.

### 2.2.2.4 Hashing Features

When we are manipulating a large number of sentences or words in a document corpus. One way to resolve to mitigate risks of memory collapse is by utilizing the Hashing representations.

The hash feature is a string of numbers and letters of fixed length and in a unique and unrepeatable order that represent a series of data. In the Figure 3, this string is created by a unique cryptographic function known as a hash function.

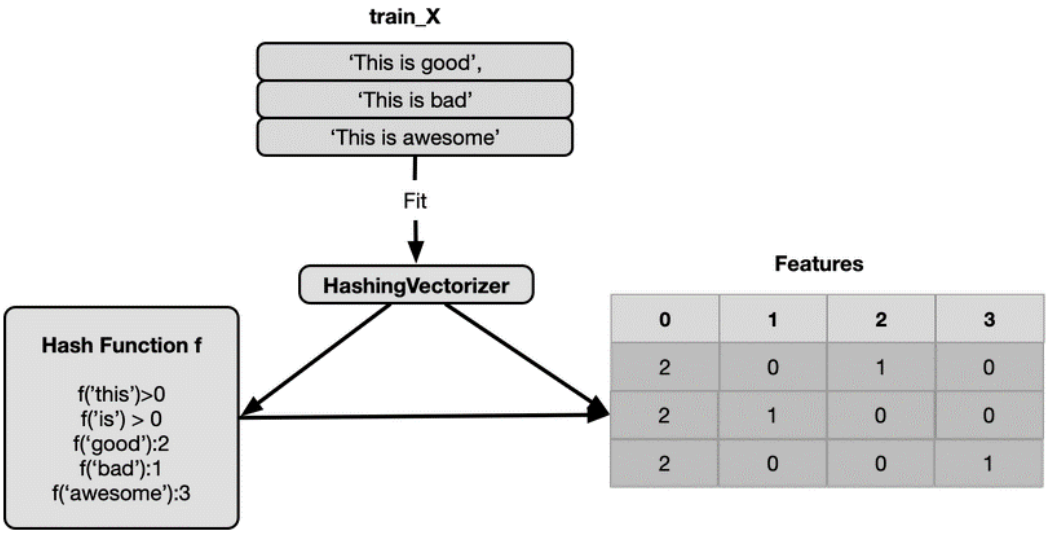



*Figure 3. Hashing Architecture.* *(Reprinted: https://mlwhiz.com/blog/2019/02/08/deeplearning_nlp_conventional_methods/)*

With the hashing function we can get the index of any word, rather than getting the index from a dictionary.

### 2.2.2.5 Word embedding

Word embedding is an approach to distribution semantics that represents words as real number vectors (See Figure 4). Such a representation has useful grouping properties, since it groups words that are semantically and syntactically similar. For example, we hope that the words "Samsung" and "Xiaomi" are close, but "Xiaomi" and "dolphin" are not close since there is no strong relationship between them. Therefore, the words are represented as vectors of real values, where each value captures a dimension of the meaning of the word. This causes semantically similar words to have similar vectors. In a simplified way, each dimension of the vectors represents a meaning and the numerical value in each dimension captures the closeness of the association of the word with that meaning. Word embedding are a series of approaches that seek to represent text taking into account the context of words through "dense" vectors, that is, they solve the problem of the bag-of-words model that used scattered vectors, since each vector represented all the vocabulary words and most of their elements are zeros, since the texts only include some words.

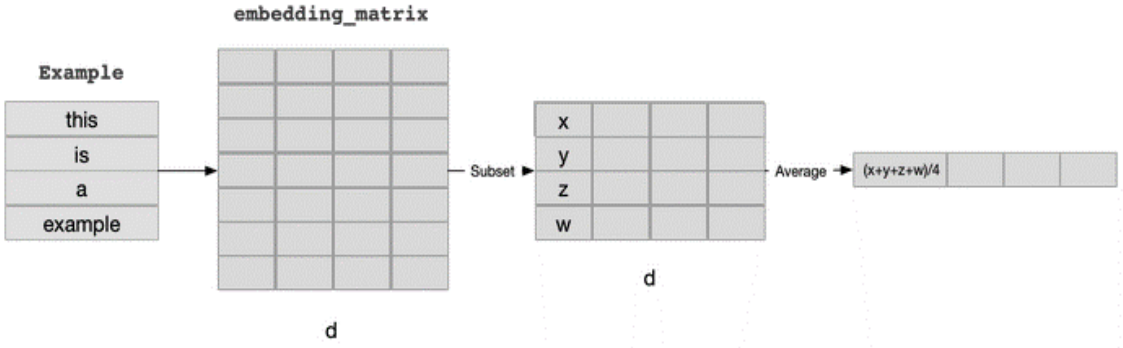



*Figure 4. Representation of word embedding.* *(Reprinted: https://mlwhiz.com/blog/2019/02/08/deeplearning_nlp_conventional_methods/)*

The term word embedding was originally conceived by Bengio et al., (2003) who trained this type of vectors in a probabilistic neuronal model. However, Collobert and Weston were possibly the first to demonstrate the power of word embeddings in their paper *A unified architecture for natural language processing* published in 2008, in which they establish word embeddings as a highly effective tool for different types of tasks, and present a neural network architecture on which many of the current approaches are based.

### 2.2.2.6 Sentence Embedding

Words combine in order to produce units of discourse: an utterance. Words do not 'carry' or encode meaning. Rather, meaning is a property associated with a complete utterance. Utterances do not exist in written language, only their representations do. For written language, the closest concept to utterance is sentence, knowing that they are not the same thing (Evans, 2006). Many successful models have been developed for sentence semantic embedding or sentence dense vector representation. However, most of such techniques use deep learning techniques on very large text corpus; and, in many cases, reuse the word vectors as input to such deep learning models. Example of these deep learning techniques are convolutional neural network (Collobert et al., 2011), recurrent neural networks (Mikolov et al., 2017) using many architectures like long-short term memory (LSTM) (Gers et al., 2000), bidirectional LSTM (Graves and Schmidhuber, 2005) and Gated Recurrent Units (GRU) (Kiros et al., 2015). All of such neural networks are mainly used to learn the dense vector representation or the semantic features of the sentence in an unsupervised learning approach.

Sent2Vec is one of the recent practical open-source models that has performed very well in semantic similarity tasks (Pagliardini et al., 2018). Sent2Vec, was used by Microsoft Research for one of their sentence embedding models that performs the mapping using the Deep Structured Semantic Model (DSSM) proposed in (Huang et al., 2013), or the DSSM with convolutional-pooling structure (CDSSM) (Shen et al., 2014) (Gao et al., 2014). But is Doc2Vec from Mikolov's (Mikolov et al., 2017) who extends the idea of

Sent2Vec. Doc2Vec learns a randomly initialized vector for the document along with the words, (document could be a sentence).

### 2.3 Methods of generating meta-embeddings

Since word2vec in 2013 started to popularize word embeddings, a lot of different methods to generate word embeddings have emerged. In this section we will present the word embeddings that we will later use in the comparisons. We have chosen the best known word embeddings for analysis, as well as some that have been interesting to us, either because of their performance or because they are very different from the rest of the word embedding methods. These word embedding methods are usually accompanied by pre-calculated vectors that we will use for evaluation and will also describe during the section.

#### 2.3.1 Word2vec

The Word2Vec model developed by Mikolov et al., (2013b), a group of Google engineers, which caused a drastic change from previous models thanks to the increase of the efficiency of training with neural networks. It is a predictive model for generating word embeddings. The Distributional Hypothesis is the main idea behind Word2Vec

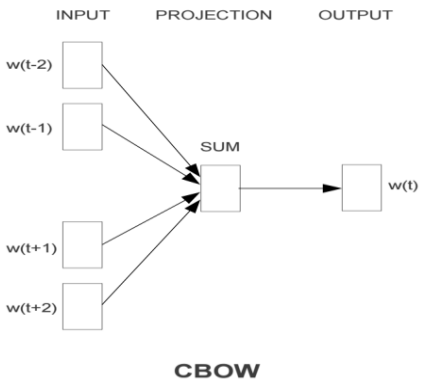



*Figure 5. CBOW Neuronal Network*

(*Reprinted:* (Gupta et al., 2016))

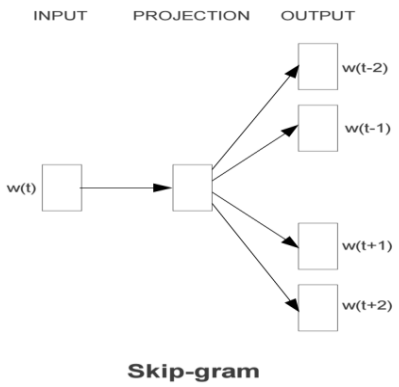



*Figure 6. Skip-gram Neuronal Network*

(*Reprinted:* (Gupta et al., 2016))

Word2Vec implements two neural models: CBOW and Skip-gram (See Figure 5 and 6). In the first one, given the context of the target word, it tries to predict it. In the second, given the word, it tries to predict the context. The internal layers of the neural network encode representation of the target word, i.e. the word embeddings. In this study we will use the vectors trained in the Google News dataset with about 100 billion words. The model contains 300 dimensional vectors for 3 million words and phrases. The vectors used are available at the following address: https://code.google.com/archive/p/word2vec/.

### 2.3.2 Doc2Vec

Paragraph vector or Doc2Vec (Le and Mikolov, 2014) applies very similar methodology of Word2Vec (Skip-gram and CBOW models) using the frequent neighbouring words to predict the document features and vice versa. Doc2Vec therefore has two algorithms to obtain the embeddings: PV-DM (Paragraph Vector - Distributed Memory) and PV-DBOW (Paragraph Vector - Distributed Bag of Words). Each one arises from the extension of the above mentioned wor2vec algorithms, respectively. In other words, PV-DM is an adaptation from word2vec's CBOW, and PV-DBOW is from Skip-gram. (See Figure 7 and 8)

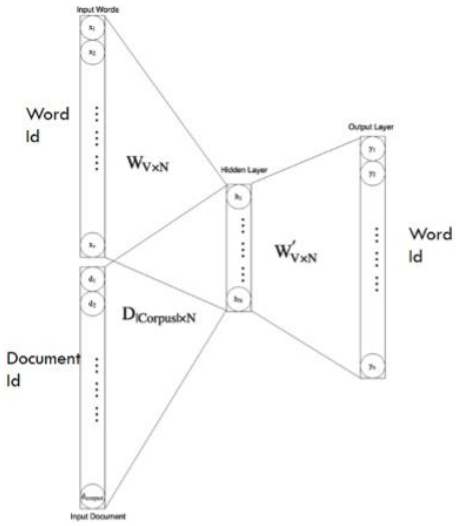



*Figure 7. PV-DM Neuronal Network*

(*Reprinted:* (Gupta et al., 2016))

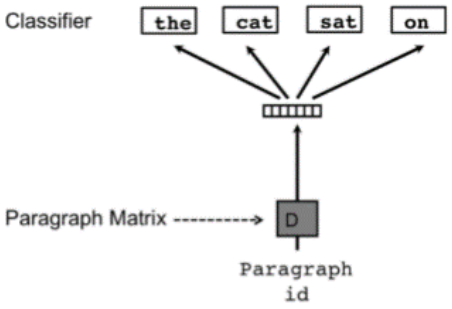



*Figure 8. PV-DBOW simplified*

(*Reprinted:* (Gupta et al., 2016))

Once the process of training word contexts together with the document *id* is finished, they end up obtaining in the matrix *D* document embeddings and in the matrix *W* word embeddings. By means of similarity measures, such as that of cosine, the vectors most similar to one defined are found, both in *W* and *D*.

### 2.3.3 FastText

This model (Boja-nowski and Mikolov, 2016) is an extension of Word2Vec that takes into account the morphology of words. The Word2Vec model typically ignores the morphological structure of each word and considers a word as a single entity. Each word is treated as the sum of its character compositions called ngrams. The vector for a word is made up of the sum of its ngrams. For example, the vector for the word “apple” is made up of the sum of the vectors for the ngrams “<ap, app, appl, apple, apple>, ppl, pple,

pple>, ple, ple>, le>". In this way, it is expected to obtain better representations for "rare" words, which have very few appearances in corpus of texts, and thus be able to generate vectors for words that are not in the vocabulary of word embeddings.

### 2.3.4 gloVE

The gloVE (Global Vectors) model, an unsupervised learning algorithm that obtains representations of words in vectors through statistics of co-occurrence. gloVE unlike Word2Vec is a count based model. gloVE generates a large matrix where the information of the concurrency between words and contexts is stored (See Figure 9). That is, for each word we count how many times that word appears in some context. The training objective of this matrix is to learn vectors so that the scalar product between the words is equal to the logarithm of the probability of co-occurrence between the words. The number of contexts is very high, therefore a factorization of said matrix is performed to obtain a smaller one. Thus obtaining a vector that represents each of the words. The advantage of gloVE over Word2Vec is that it is easier to parallelize the training, therefore it is possible to use more information during the training. Therefore it is possible to use more data during training.

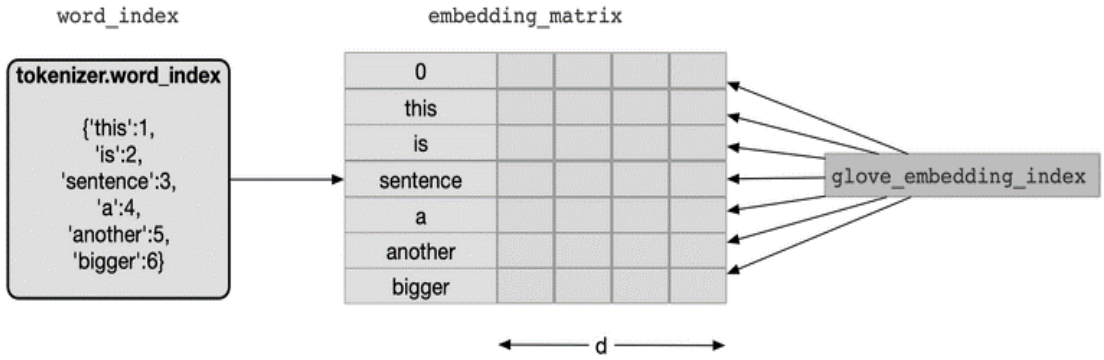



*Figure 9. gloVE storage of the information concurrency*

(*Reprinted:* *https://nlp.stanford.edu/projects/glove/*)

## 2.4 Text-Mining Methods

### 2.4.1 Text Mining Approaches

Information Retrieval (IR): Information Retrieval is the activity of finding information resources (usually documents) from a collection of unstructured data sets that satisfies the information need (Manning et al., 2008).

Natural Language Processing (NLP): Natural Language Processing is sub-field of computer science, artificial intelligence and linguistics which aims at understanding of natural language using computers (Manning et al., 1999).

Information Extraction from text (IE): Information Extraction is the task of automatically extracting information or facts from unstructured or semi-structured documents (Cowie and Lehnert, 1996).

Text Summarization: Many text mining applications need to summarize the text documents in order to get a concise overview of a large document or a collection of documents on a topic. (Radev et al., 2002).

Unsupervised Learning Methods: Unsupervised learning methods are those in which we do not have a pool of previously classified examples, but only from the properties of the examples we try to give a grouping (classification, clustering) of the examples according to their similarity. They are techniques trying to find hidden structure out of unlabeled data. They do not need any training phase, therefore can be applied to any text data without manual effort.

Supervised Learning Methods: Supervised classification systems are those in which, from a set of classified examples (training set), we try to assign a classification to a second set of examples. Supervised learning methods are machine learning techniques pertaining to infer a function or learn a classifier from the training data in order to perform predictions on unseen data. (Sebastiani, 2002).

Probabilistic Methods for Text Mining: There are various probabilistic techniques including unsupervised topic models such as probabilistic Latent semantic analysis (pLSA) (Hofmann, 1999) and Latent Dirichlet Allocation (LDA) (Blei and Jordan, 2003), and supervised learning methods such as conditional random fields that can be used regularly in the context of text mining.

Sentiment Analysis: also known as *opinion mining*, It is a great challenge for language technologies, as obtaining good results is much more difficult than many believe. The task of automatically classifying a text written in a natural language into a positive or negative feeling, opinion or subjectivity (Bo and Lee, 2008), is sometimes so complicated that it is even difficult to agree on different human notebooks about the classification to assign to a given text. With the advent of e-commerce and online shopping, a huge amount of text is created and continues to grow about different product reviews or users opinions.

## 2.5 Evaluation of word embeddings

The methods of evaluation of recordings can be grouped into large groups, extrinsic methods and intrinsic methods.

### 2.5.1 Extrinsic evaluation methods

Extrinsic evaluation methods are based on the ability of a word embedding to be used as vectors of characteristics of supervised self-learning algorithms used in various NLP tasks. The performance of the supervised method (usually measured in a dataset for NLP tasks) is taken as a measure of the quality of the word embedding. Some of the most common tasks in which word embeddings are evaluated are

1. Phrase name extraction. The objective is recognize nominal phrases and their limits within a sentence.

2. Entity name recognition. Recognize name of entities as names of organizations, persons, brands... within a sentence and its limits.
3. Sentimental analysis. A particular case of classification of texts, where a fragment must be marked with a binary tag reporting whether the text has positive or negative feeling towards something.
4. Syntax analysis superficial. Breakdown of sentences in groups (nominal sentences, verbal sentences, adjective sentences…).
5. Scope of denial. This is a text classification task. It is about identifying whether a specific action in a sentence determines denial or not.

### 2.5.2 Intrinsic evaluation methods

Intrinsic evaluation methods are experiments in which word embeddings are compared with human judgments about word relationships. Often manually created word sets are used, first the human evaluations are obtained and then these are compared with the word embeddings. Most intrinsic evaluation methods are designed to collect evaluations that are the result of conscious processes in the human brain. There are a large number of evaluation methods that fall under the heading of intrinsic methods, so we will focus on the most widely used, so-called conscious intrinsic evaluation methods. There are different tasks that are included within them:

1. Analogy of words. It is based on the idea that arithmetic operations in word vector space could be predicted by humans: given a set of three words, a, a* and b, the task is recognize such a word b* such that the relation b: b * is the same as the relation a: a *. "Paris is to France as Moscow is to Russia." The main criticism of this method is the lack of precise evaluation metrics.
2. Thematic adjustment. The method evaluates the ability of a model to separate different thematic roles from the arguments of a predicate. The idea is to find out how well word embeddings can find the most semantically similar noun for a certain verb used in a certain role. For humans, a certain verb might make a person expect that a certain role should be filled with a certain noun (for example, for the argument "I'm going to cut" the most expected argument in the object role is "cake").
3. Synonym detection. The objective is to evaluate the ability of a word embedding by stopping a word W and a series of words K=a1,a2,a3... to find the word K most similar to W.
4. Given a list of words the objective is to detect the anomalous word within the group. For example, given the words "pineapple, apple, cherry, orange, book, banana" the anomalous word is "book" because it is unrelated to the rest.
5. Semantic similarity between words: This is the popular method of evaluating word embeddings and is the method we are going to use to evaluate the word embeddings in the next chapters. The method is based on the idea that the distances between words in a word embedding can be evaluated by human heuristic judgments about the actual distances between words. (For example, the distance between "cop" and "cup" could be defined in a range of 0.1 to 0.) The evaluator is given a series of words and is asked to evaluate the degree of similarity for each one. The more similar they are, the better the word embedding. This method dates back to 1965, when the first experience with human judgments on the semantic similarity of words was made to test the distribution hypothesis.

However, this method also receives some criticism, since there are conditions, linguistic, psychological and social, that can affect human judges, even fatigue after scoring a large number of word pairs can affect the score given. Another criticism of this method is that different experiments tend to give different semantic similarity definitions, for example some describe it as hyperonymy/hyponymy ("machine", "car") and others as synonymy ("car", "vehicle"). However, this method is the most popular for evaluating word embeddings.

There are several different data sets for evaluating word similarity. These can be grouped into two large groups, those that measure the semantic similarity of words, and those that measure the semantic relationship or association. These two concepts are different. Datasets that study word similarity generally capture synonymic relationships between words, sometimes also hyperonymic and hyponymic relationships. For example, the words "coast" and "coastline" are synonymous, so both in datasets where semantic similarity is studied and in those where semantic relationships are studied, they will appear with a very high score. On the other hand, the words "clothes" and "wardrobe" do not have any kind of synonymy or hyperonomy/hyponymy relationship, therefore, in datasets that study semantic similarity, they will have a very low score. However, both words have a great relationship with each other, since clothes are kept in closets, therefore in datasets that analyze the semantic relationship they will have a very high score.

## 2.6 Semantic Text Similarity

### 2.6.1 Overview

In the Figure 10 is a summary of the seminal work of Wael and Fhamy, (2013) and Dwi et. al, (2018) which make an exhaustive and detailed study of the different approaches have been promoted to measure the similarity between texts. Given the limited scope of our work, we will only mention the measures of similarity implemented in our research.

<table>
<tr><td rowspan="19">Text Similarity Measure</td><td rowspan="10">String based<br>A string metric is a metric that measures similarity or dissimilarity (distance) between two text strings for approximate string matching or comparison</td><td rowspan="6">Character-based</td><td colspan="2">Longest Common Subsequence</td></tr>
<tr><td colspan="2">Damerau-Levenshtein</td></tr>
<tr><td colspan="2">Jaro</td></tr>
<tr><td colspan="2">Jaro-Winkler</td></tr>
<tr><td colspan="2">Needleman-Wunsch</td></tr>
<tr><td colspan="2">Smith-Waterman</td></tr>
<tr><td rowspan="4">Term-based</td><td colspan="2">n-gram</td></tr>
<tr><td colspan="2">Block Distance</td></tr>
<tr><td colspan="2">Cosine Similarity</td></tr>
<tr><td colspan="2">Dice's Coefficient</td></tr>
<tr><td colspan="3"></td></tr>
<tr><td rowspan="6">Corpus-based<br>Similarity between words according to information gained from large corpora.</td><td colspan="2">Hyperspace Analogue to Language (HAL)</td><td></td></tr>
<tr><td colspan="2">Latent Semantic Analysis (LSA) [27]</td><td>CL-LSA</td></tr>
<tr><td colspan="2">Explicit Semantic Analysis (ESA)</td><td>SOC-PMI</td></tr>
<tr><td colspan="2">Pointwise Mutual Information (PMI),</td><td></td></tr>
<tr><td colspan="2">Normalized Google Distance (NGD)</td><td>DISCO1</td></tr>
<tr><td colspan="2">DIstributionally Similar words using CO-occurrence</td><td>DISCO2</td></tr>
<tr><td colspan="4"></td></tr>
<tr><td rowspan="7">Knowledge-based<br>Is one of semantic similarity measures that bases on identifying the degree of similarity between words using information derived from semantic networks</td><td rowspan="5">Similarity</td><td rowspan="3">Information Content</td><td>res</td></tr>
<tr><td>lin</td></tr>
<tr><td>jcn</td></tr>
<tr><td rowspan="3">Path Length</td><td>lch</td></tr>
<tr><td>wup</td></tr>
<tr><td>path</td></tr>
<tr><td rowspan="3">Relatedness</td><td colspan="2">hso</td></tr>
<tr><td colspan="2">lesk</td></tr>
<tr><td colspan="2">vector</td></tr>
<tr><td colspan="4"></td></tr>
<tr><td>Hybrid similarities<br>this approach is to combine the previously described approaches, including string-based, corpus-based, and knowledge-based similarity to reach a better metric by adopt their advantages.</td><td colspan="3">Soft Cosine similarity</td></tr>
</table>

*Figure 10. Overview of Text Similarity Measure*

As can be seen in the table above, similarity measures are generally classified into four main groups. Of these, hybrid similarities are the ones that currently receive the most attention and popularity.

We will highlight three measures of similarity, which correspond to the approach of string, corpus and hybrid based, in order to provide robustness to our comparison that we will see in the following chapters.

### 2.6.2 Cosine similarity

Cosine similarity is a measure of similarity that is calculated between two non-zero vectors within the internal space of the product that measures the cosine of the angle

between them. The similarity between two word vectors can be obtained through the angle they form, specifically by means of the cosine of the angle. It is considered that the smaller the angle, and consequently the cosine of the angle, the greater similarity there will be between them.

With the following equation we obtain the cosine measure between the documents $d_i$ and $d_j$, where $d_{ik}$ is the weight of the semantic trait $k$ in the document $d_i$.

$$sim(d_i, d_j) = \cos(\alpha) = \frac{\sum_{k=1}^{m} d_{ik} * d_{jk}}{\sqrt{(\sum_{k=1}^{m} d_{ik}^2) * (\sum_{k=1}^{m} d_{jk}^2)}} = \frac{d_i}{|d_i|} * \frac{d_j}{|d_j|}$$

*Figure 11. Cosine similarity equation*

With the following formula of the Euclidean distance it is measured how far two vectors are in the vector space. This formula will only be useful when dealing with two vectors with not very large dimensions.

$$dist(d_i, d_j) = \sqrt{\sum_{k=1}^{m} (d_{ik} - d_{jk})^2}$$

*Figure 12. Cosine similarity with Euclidean distance equation*

It is important to understand the concept of cosine similarity to understand its usefulness for our projects. It is used particularly in a positive space where the result is clearly delimited in $[-1,1]$.

### 2.6.3 Soft cosine similarity

The soft cosine allows to take into account the similarity between features in a vector space model. For the calculation of the soft cosine, the matrix containing the similarity between the characteristics is entered. It can be calculated using the Levenshtein distance or other similarity measures, for example, various WordNet[4] similarity measures. It is then only multiplied by this matrix.

Given two vectors $a$ and $b$ of dimension *N*, the soft cosine is calculated as follows:

[4] WordNet is a lexical database of the English language that groups English words into sets of synonyms called synsets

$$soft\_cosine(a,b) = \frac{\sum_{i,j}^{N} s_{ij} a_i b_j}{\sqrt{\sum_{i,j}^{N} s_{ij} a_i a_j} \sqrt{\sum_{i,j}^{N} s_{ij} b_i b_j}}$$

*Figure 13. Soft-Cosine similarity equation*

where $s_{ij}$ = similarity(feature $i$, feature $j$).

If there is no similarity between characteristics ($s_{ij}$= 1, $s_{ij}$= 0 for $i \neq j$), the given equation is equivalent to the conventional cosine similarity formula.

The complexity of this measure is quadratic, which makes it fully applicable to real-world problems. The complexity can even be transformed to linear.

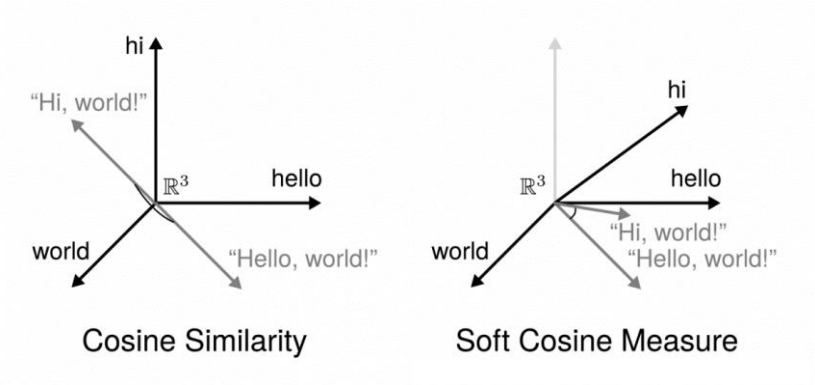



*Figure 14. Sample Cosine and Soft-Cosine similarities*

To compute soft cosines, we need a dictionary (a map of word to unique id), a corpus (word counts) for each sentence and the similarity matrix.

### 2.6.4 Latent Semantic Analysis (LSA)

The LSA (Latent Semantic Analysis) also known as LSI (Latent Semantic Index) is a mathematical tool that analyzes semantic relationships between different linguistic units in a fully automated way (Landauer and Dumais, 1997). It was originally presented in 1990 as a method of information retrieval, to overcome the great limitations presented by search engines in databases, although later it has also been considered a model of acquisition and representation of knowledge.

The operation of this technique is as follows. The LSA processes the documents we are handling, which can be large, containing a large number of paragraphs and, ultimately,

information, giving rise to the linguistic corpus. Next, this corpus is represented by a matrix whose rows and columns contain, respectively, the different terms that appear in the corpus and the documents that we are considering. This matrix reflects the number of times each term appears in each document.

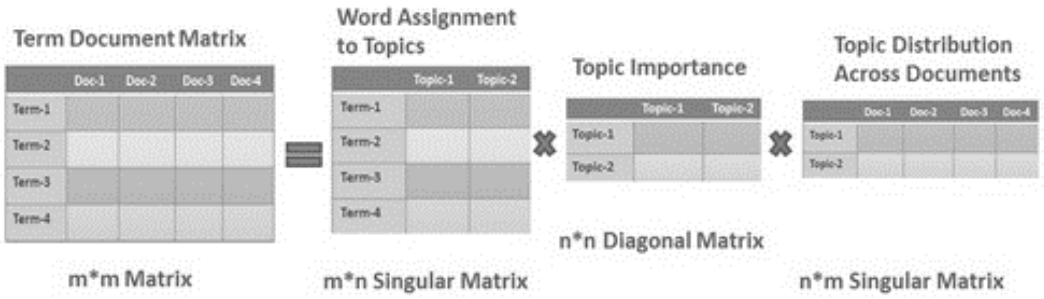



*Figure 15. LSA workflow*

Since we can assume that excessively frequent words do not discriminate the information in each document, the LSA weighs down the importance of notably more frequent words and increases the importance of moderately infrequent words. Next, apply to the obtained matrix the singular value decomposition algorithm (SVD)[5] with the aim of reducing the dimension of the matrix with which we are working to a more manageable number (approximately 300), and without losing relevant information. The objective of applying the SVD is to weight each term based on its ability to represent a document. In addition, by means of this algorithm, the vector space with which we will work in the following will have been created and with all the advantages of working with a vector space (for example, the comparison of vectors using distances).

Whereas before applying the algorithm, the vectors were hollow vectors, the new vectors would not be so in general. This makes it possible to detect significant relationships between pairs of documents, even if those documents did not have common terms. The idea is that terms that have a similar meaning will be oriented in approximately the same direction in latent space.

On the other hand, the LSA presents the possibility of introducing into the space new vectors that represent texts that do not appear in the linguistic corpus that the LSA had analyzed, and that we call pseudo-documents (Landauer et al., 1998). The incorporation of the pseudo documents does not require recalculating the vector space, which is a great advantage. These pseudo documents will later be used to categorize terms and texts.

[5] The SVD is a specific form of factor analysis. In the starting matrix, terms and documents are mutually dependent on each other, while, after applying the algorithm, these relationships appear broken down. The SDV decomposes the original matrix into the product of three new matrices T, D, and S, which contain the eigenvectors and eigenvalues. The eigenvalues, in turn, contain the variability information, in terms of terms and documents, explained by each dimension. The matrix T contains the information of the factors that have been determined by the analysis carried out.

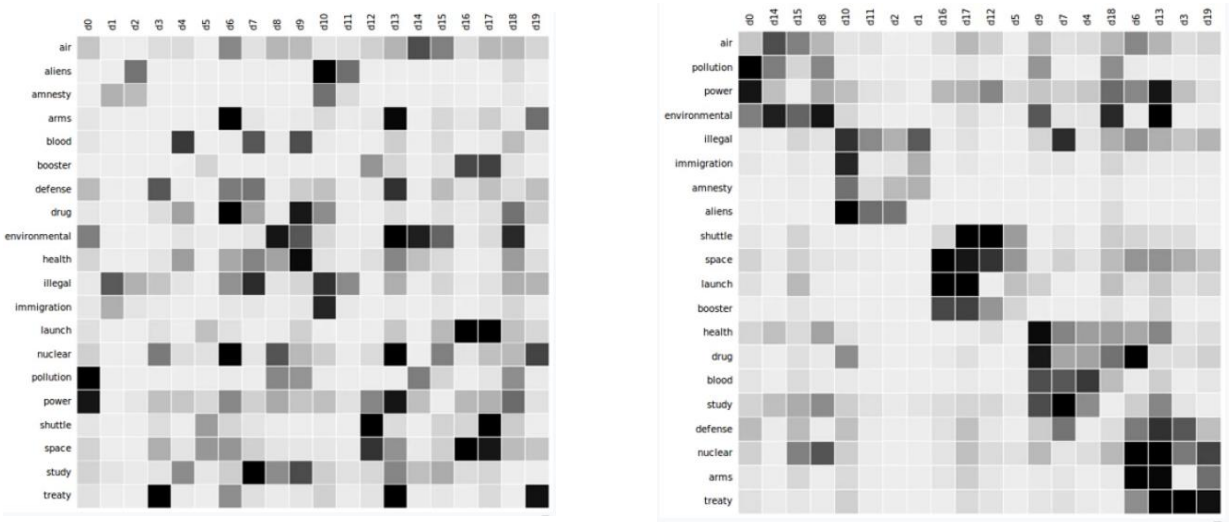



*Figure 16. Example: Matrix data Transformation using LSA, left side raw values, right side after applied LSA*

## 2.7 GRI reports

The Corporative Social Responsibility (CSR) s is part of the Environmental Social and Governance (ESG) initiatives. This reports are created in order to satisfied the stakeholders demands and have to contain both qualitative and quantitative information to the extent which reveals how the company has improved its own economic, environmental and social effectiveness and efficiency in the reporting period and how the company has integrated these aspects into its sustainability management system. (KPMG, 2017) highlights "the necessity of balance between qualitative and quantitative information in sustainability reports when providing an overview of the company's financial/economic, social/ethical, and environmental performance". One of the most popular reporting and considering as the most excellent and worldwide acknowledged framework is the Global Reporting Initiative (GRI) (Issakson and Steimle, 2009); (Knebel and Seele, 2015). Currently, 93% of the 250 biggest companies report on their sustainability based on the GRI Guidelines (KPMG, 2017).

### 2.7.1 GRI versions

The development of GRI guideline generations is constantly in progress. From July 2018, a new generation called "GRI Standards" will replace GRI G4. One of the main differences is that now the GRI standard is going through to simplify the framework and avoid labelling the ESG commitment of the companies.

In GRI G3, the chapters on company profile and management approach were followed by the section of non-financial performance indicators, including 84 indicators in total. The 56 core and 28 additional indicators were further classified into economic indicators (7 core, 2 additional), environmental indicators (18 core, 2 additional), and social indicators (31 core, 14 additional). In the case of social indicators, four subcategories. Were identified: human rights, labour, product responsibility, and society. In the G3

system, companies could decide on different levels (A, B, or C), containing different amounts of core and additional indicators. The + sign indicated the independent third party assurance of the report (Knebel and Seele, 2015). This standard was observed for the use of an excessive amount of indicators (Knebel and Seele, 2015b) and that the guideline did not consider the synergies among different dimensions (Lozano–Huisingh, 2011).

In the case of GRI G4, there is no separation of core and additional indicators, while indicators have been further extended in number. This may cause problems in internal comparison with previous reports of the same company, when switching from G3 to G4 (Global Report Initiative, 2013). In addition, G4 includes further differences compared to G3. One of the central elements of G4 is materiality assessment–the function of which is to serve as an input for preparing the report –since its aim is to explore the main environmental, social and economic aspects relating to the activities of the company from the points of view of stakeholders and the company itself. The boundaries of reporting were redefined as well, resulting in a replacement of A, B, C classification by "in accordance" levels.

For the GRI Standards, an update of GRI G4, new requirements have been introduced in terms of corporate governance and impacts along the supply chain (Global Report Initiative, 2016). It is a change of format from GRI G4 which is made up of two documents to a compendium of 36 independent but interrelated documents. This new, more flexible structure aims to make is easier to use and to update (it will possible to update only one of the documents, without modifying the rest). The GRI standards do not include new aspects, but they do include certain changes in the way of reporting e.g. the difference between what is mandatory and what is a recommendation or orientation is now clearer, in the location of the aspects and in the indicators. The GRI standards is mandatory since July 2018.

### 2.7.2 Text Mining on GRI reports

The Corporative Sustainability Reports (CSR) are becoming increasingly important for the scientific community, especially in the study of methodology, definition and frequency (Kolk, 2004), (Kolk, 2003), (Bjørn et al., 2004). Also in the comparison of the different techniques used by companies from a qualitative point of view (Freundlieb and Teuteberg, 2013).

We will examine the content of CRS reports, focus on the GRI reports, in a more quantitative way through text mining techniques, e g. Liew et al., (2014) try to identify sustainability trends and practices in the chemical process industry by analyzing published sustainability reports. Székely et al., (2017) confirms previous research on a more widely with 9514 sustainability reports Yamamoto et al., (2017) develops a method that can automatically estimate the security metrics of documents written in natural language. This paper also extends the algorithm to increase the accuracy of the estimate. Chae and Park, (2018) study adopts computational content analysis for understanding themes or topics from CSR-related conversations in the Twitter-sphere and Benites-Lazaro, (2018) identify companies' commitment to sustainability and business-led governance.

The default technique used in previous investigations is LDA, but other techniques were also implemented such as unsupervised learning using the expectation-maximization algorithm for identify clusters and patterns. Tremblay and Gonzales (2015) that use an attractor network to learn a sequence series with the goal to predict the GRI scoring.

Extensive attention has been paid to this topic for the works by Modapothala starting from statistical techniques (Modapothala, 2014), Bayesian (Modapothala, 2009), or multidiscriminatory analysis (Modapothala et al., 2013), for analysis of corporate environment reports.

Shahi and Modapothala, (2015) have produced a specific work in this area, using the GRI G3 version. This version used a score grade ranging from A+ to C to measure the effectiveness of the Level Check which was removed from the framework for the GRI G4 version. As such as Liu et al., (2017) make use of TF-idf method to obtain important and specific terms for further analytical algorithm and test with many shallow machine learning models.

# 3 METHODOLOGY

## 3.1 Overview

Since the last GRI framework was implemented, there is no record of the level of compliance that published reports have with current standards, therefore, there is no test information that we can use to validate text mining techniques. Henceforth, we are faced with an Unsupervised Learning Problem. In an unsupervised learning problems it is not possible to know which model or algorithm gives the best result on a data set without having previously experimented, so when choosing a model for a certain problem the only thing that can be done it is trial and error, that is, testing with different representations of the data set, different algorithms and different parameters of each algorithm, which is why a procedure must be followed.

In this instance, we need understand our data set. We will apply Exploratory Data Analysis (EDA) in order to design which algorithms would best suit our needs and environment. Carrying out a methodology allows planning and estimating the work to be done, preparing a development plan and focusing the focus on each of the phases independently.

## 3.2 Data Collection

The GRI standards database is publicly accessible, this database has more than sixty thousand reports stored, for our study we initially decided to focus on Finnish companies, but given the small volume that we had we decided to expand to all Nordic countries, that is, we download all reports using only country as filter parameter. In total, we have 450 reports where some were discarded because they were written in another language than English, leaving a total of 424 reports. Of which, as can be seen in figure 5, most correspond to Sweden and Finland.


*Figure 17. Distribution of collecting data by country and standard Guideline*

We can appreciate the volume of reports that belong to the last standard and that will be the framework studied, because as we will see later, it will be an important feature when evaluating these reports.

### 3.3 Pre-processing

To normalize the extracted data, the texts were subjected to a default process of debugging and transformation of the texts. The task at hand is focused on cleaning the text of the data set to eliminate all irrelevant aspects or those that do not facilitate the performance of the model. This task is complicated and it is not possible to know if a modification to the text may affect the result of the model better or worse, so we will try to make it as simple as possible.

1. Normalization: the first step in this preprocessing is to normalize the text, that is, that different representations of the same word become a common representation. In this case, all words will be normalized to lowercase.

2. Elimination of repeated characters: it is common in texts that are not formal (such as in social networks) the repetition of the same characters in order to produce emphasis. In English there are no words that repeat the same character more than 2 times, so we correct all the words in which this situation occurs.

3. Correction of spelling errors: even eliminating repeated characters, in English there are many words that contain the same character twice, so it is not possible to eliminate all the characters that are repeated these times. If we remove the ones that are repeated the most, it could still happen that the word was spelled incorrectly with the same character repeated 2 times. For this reason, a spell checker will be used that, in addition to solving this problem, will correct misspelled words. Automatic spelling correction is a task still under study in the field of language processing, due to its complexity (if the word differs much from the correct one it is very difficult to correct it), which makes this correction not perfect. Therefore, a basic spell checker implementation is used to correct words in the simplest way possible.

4. Elimination of punctuation marks and non-alphanumeric characters: once all the words have been normalized, the text content that is not relevant to the problem is eliminated. Punctuation marks and other non-alphanumeric characters do not provide any information on this problem, so the following characters are removed from our documents:! "# $% & '() * +, -. / :; < =>? @ [\] ˆ _ '{|}.

5. Elimination of empty words: empty words are those common words that appear very frequently in the text and do not have any meaning or impact within it. These can be articles, prepositions, etc. In English, for example, the most common would be "the" or "is".

6. Stemming: stemming consists of reducing a word to an English common root form. The utility of this modification is that it allows to reduce the size of the vocabulary, identifying in the same way the different variations of the words. For example, if it is a review, the words “fishes” and “fishing” are found, both would be represented as “fish”. However, performing this step may cause the meaning of some words to be lost, as the different variations of the same words may cause different meanings. Furthermore, in the case of experimentation with the model Word2Vec, the

representation in the vector space of the words may have made the pre-trained model learn the different morphological variations of them. For these reasons, it will be experimented previously with the clean set without lemmatization and with the stemmed set to determine which one produces better results and use it in experimentation.

Because the reports we are using are generally unstructured, so it has been necessary to apply specific tasks to solve problems such as the absence of fields or the existence of incomplete data. Unfortunately the text preprocessing is not perfect and always can be improved, but it is considered that the reports have been well cleaned and that going deeper into it would remove the focus from the objective of the work.

### 3.4 EDA

It has already been explained previously that the problem that we face using text mining methods is to represent the text in such a way that an algorithm can interpret it, e.g. in all machine learning models one of the main tasks prior to experimentation is the preparation of the data. As it is a UL problem, we need to build our methodology by experimenting a little, to identify the limits and best options that could be adjusted to our problem, which is why we will apply the following method:

#### 3.4.1 Exploratory Data Analysis (EDA)

The Exploratory Data Analysis (EDA) is an approach to the study of data collections, mostly utilizing visual methods, to summarize their key characteristics. Instead of jut apply statistical descriptive functions, EDA can help us for seeing what the data can tell us beyond the formal modelling or hypothesis testing task. EDA can show us hidden relationships and attributes present in our data even before we throw it at a text mining model.

*Dataset description.*

We have two important tables, that of the companies (see Fig. 5) and that of the guidelines. The guidelines database are the result of the extraction of the embedded text in pdf documents (see chapter 4).

These GRI guidelines are distributed as follows[6]:

[6] For more details: https://www.globalreporting.org/standards/

GRI 102: General Disclosures 2016

Organizational profile

**102-1**: Name of the organization
**102-2**: Activities, brands, products, and services
**102-3**: Location of headquarters
**102-4**: Location of operations
**102-5**: Ownership and legal form
**102-6**: Markets served
**102-7**: Scale of the organization
**102-8**: Information on employees and other workers
**102-9**: Supply chain
**102-10**: Significant changes to the organization and its supply chain
**102-11**: Precautionary Principle or approach
**102-12**: External initiatives
**102-13**: Membership of associations

Strategy

**102-14**: Statement from senior decision-maker
**102-15**: Key impacts, risks, and opportunities

Ethics and integrity

**102-16**: Values, principles, standards, and norms of behavior

Governance

**102-18**: Governance structure
**102-22**: Composition of the highest governance body and its committees
**102-23**: Chair of the highest governance body
**102-24**: Nominating and selecting the highest governance body
**102-32**: Highest governance body's role in sustainability reporting
**102-38**: Annual total compensation ratio
**102-39**: Percentage increase in annual total compensation ratio

Stakeholder engagement

**102-40**: List of stakeholder groups
**102-41**: Collective bargaining agreements
**102-42**: Identifying and selecting stakeholders
**102-43**: Approach to stakeholder engagement
**102-44**: Key topics and concerns raised

Reporting practice

**102-45**: Entities included in the consolidated financial statements
**102-46**: Defining report content and topic Boundaries
**102-47**: List of material topics
**102-48**: Restatements of information
**102-49**: Changes in reporting
**102-50**: Reporting period
**102-51**: Date of most recent report
**102-52**: Reporting cycle
**102-53**: Contact point for questions regarding the report
**102-54**: Claims of reporting in accordance with the GRI Standards
**102-55**: GRI content index
**102-56**: External assurance

Series 200: Economic Topics

Economic Performance

**GRI 103: Management Approach 2016**

**103-1**: Explanation of the material topic and its Boundary
**103-2**: The management approach and its components
**103-3**: Evaluation of the management approach

**GRI 201: Economic Performance 2016**

**201-1**: Direct economic value generated and distributed

**User defined disclosures**

G4 NGO Sector Disclosure: Ethical Fundraising

Series 400: Social Topics

Employment

**GRI 103: Management Approach 2016**

**103-1**: Explanation of the material topic and its Boundary
**103-2**: The management approach and its components
**103-3**: Evaluation of the management approach

**GRI 401: Employment 2016**

**401-1**: New employee hires and employee turnover

Training and Education

**GRI 103: Management Approach 2016**

**103-1**: Explanation of the material topic and its Boundary
**103-2**: The management approach and its components
**103-3**: Evaluation of the management approach

**GRI 404: Training and Education 2016**

**404-1**: Average hours of training per year per employee
**404-3**: Percentage of employees receiving regular performance and career development reviews

**User defined disclosures**

G4 NGO Sector Disclosures: Mechanisms for workforce feedback and complaints and their resolutions

Diversity and Equal Opportunity

**GRI 103: Management Approach 2016**

**103-1**: Explanation of the material topic and its Boundary
**103-2**: The management approach and its components
**103-3**: Evaluation of the management approach

**GRI 405: Diversity and Equal Opportunity 2016**

**405-1**: Diversity of governance bodies and employees

Other Topics

Fostering Effective Collaboration with other Organizations

**Management Approach**

**103-3**: Evaluation of the management approach
**103-1**: Explanation of the material topic and its Boundary
**103-2**: The management approach and its components

**Custom Disclosures**

Driving Better Sustainability Reporting

**Management Approach**

**103-3**: Evaluation of the management approach
**103-2**: The management approach and its components
**103-1**: Explanation of the material topic and its Boundary

**Custom Disclosures**

Improving Performance through Sustainability Reporting

**Management Approach**

**103-3**: Evaluation of the management approach
**103-2**: The management approach and its components
**103-1**: Explanation of the material topic and its Boundary

**Custom Disclosures**

Harmonizing the Sustainability Reporting Landscape

**Management Approach**

**103-3**: Evaluation of the management approach
**103-2**: The management approach and its components
**103-1**: Explanation of the material topic and its Boundary

**Custom Disclosures**

*Figure 18. GRI Guidelines distribution*

The GRI guidelines consist on 169 disclosures grouping in 37 Standards (See Appendix A). This guidelines contain information about the minimal technical information that need to be provide by the companies. The companies itself determine if they accomplish or not this requirements.

### 3.4.2 Descriptive analysis of the dataset

Next we will show some characteristics of the dataset that we are manipulating.

#### 3.4.2.1 N-Grams distribution

Distribution of the text by length and number of words:

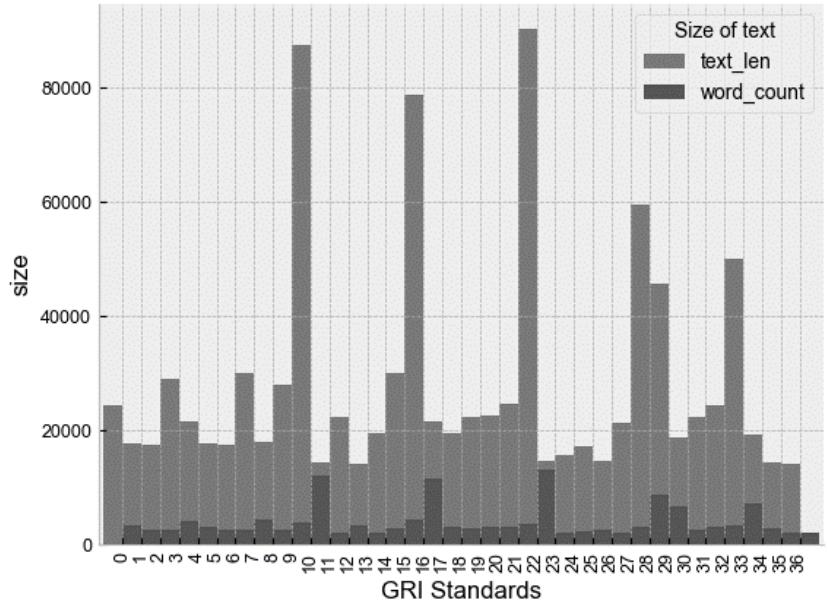



*Figure 19. Distribution of Guidelines by length and number of words*

Standard 22, 10 and 16 stand out as greater containers of words and therefore with greater length of text. Now let's see which words are the most frequent by n-grams, implemented stop words

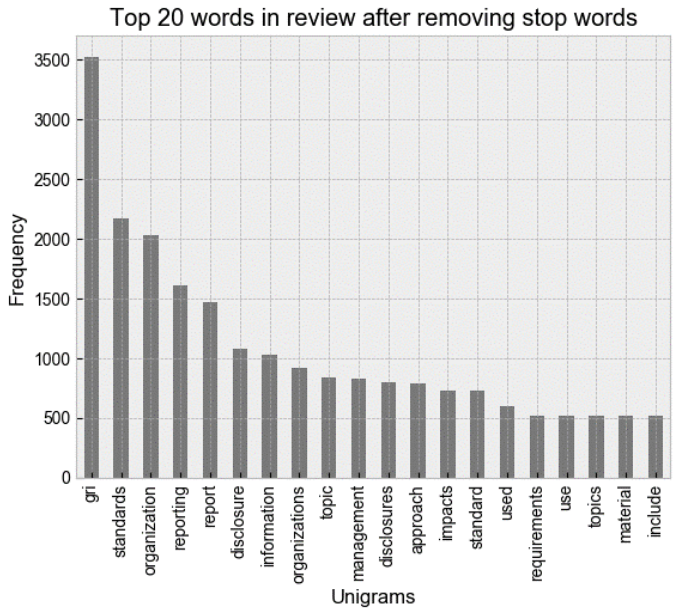



*Figure 20. Top frequently words on Guidelines*

It becomes evident, the necessity that the use of stopwords makes necessary for the interpretation.

The distribution of top bigrams before removing stop words

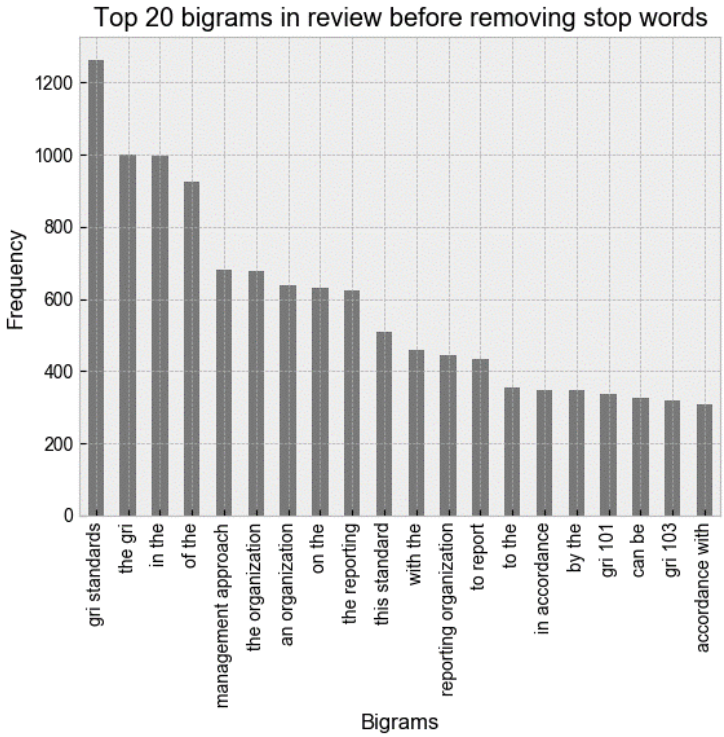



*Figure 21. Top frequently bigrams on Guidelines*

The distribution of Top trigrams before removing stop words

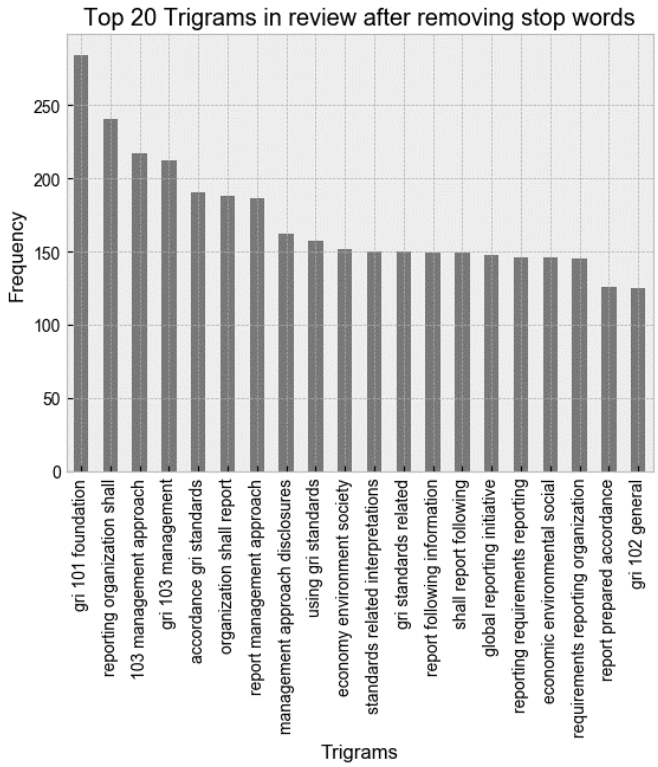



*Figure 22. Top frequently trigrams on Guidelines*

For us, as humans interpreters, the last graph provides more information than the previous ones, highlighting among all the previous graphs words such as: “GRI” that refers to the initials of the organization, “the standards”, and surprisingly, to “standard 101”, one might expect that the most mentioned or most frequent standard would be those that belong to standards 22, 10 and 16. This information is telling us that the size of the text does not imply that allusion is made to a specific standard for its text length, but more well to its content itself. Standards 101 has a general and fundamental character for the guidelines, therefore it is not surprising that it is one of the most widely used terms throughout all the documents explored.

The distribution of top part-of-speech tags of review corpus

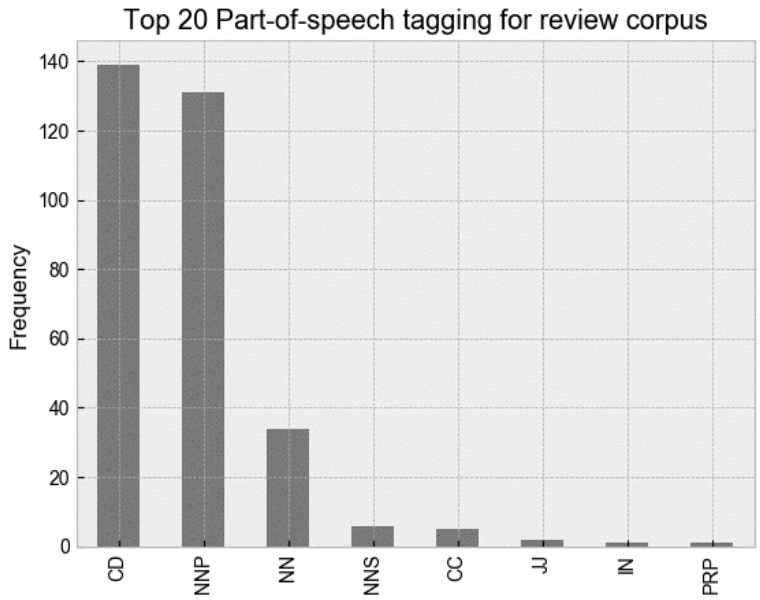



*Figure 23. Frequently of part of speech on Guidelines*

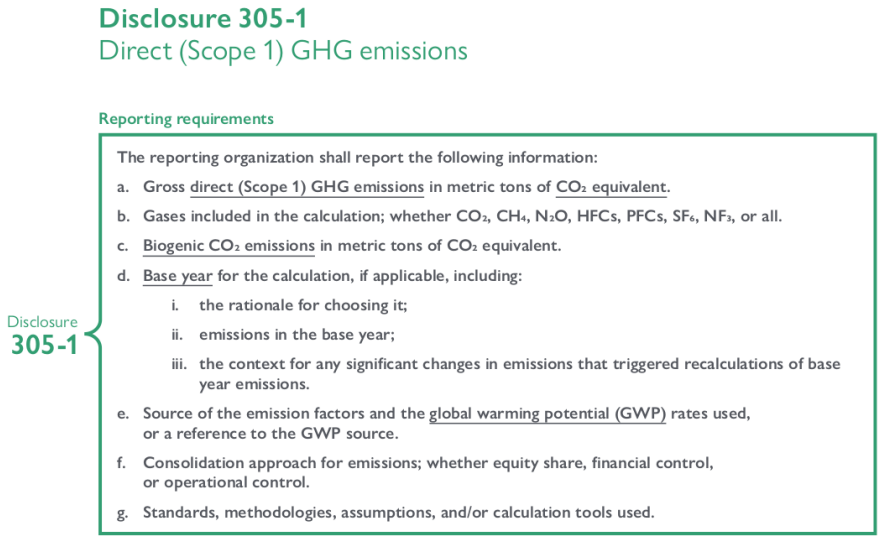
The CD (cardinal digit) stands out, because atypically we do not remove number representations at the processing stage, because in our context, numbers are used in order to represent the membership of a disclosure. Furthermore, the high difference between NNP (noun plural), NN (noun) and NNS (proper noun) vs. CC (coordination conjunction), may suggest the formality of how the text is written.

### 3.4.3 Example Analysis Bottom-up

Our objective is to evaluate the degree of affinity of the CSR reports of the companies with the GRI guidelines. Therefore, we will perform a bottom-up evaluation, to obtain enough information to facilitate the modelling process. To obtain an idea of how text mining can be implemented later, now we will explore how a descriptive comparison would be made between an official standard and a real report of a company. In this case we selected randomly the Emissions standard GRI-305 as a guideline example and a Skatkraft[7] as a company example.

The selection of the company, has not been random, we select the company that has the greatest semantic variation in the results of the test sets when we filtered by standard 33, that includes the disclosure GRI-305 (See chapter 5).

[7] Statkraft AS is a hydropower company, fully owned by the Norwegian state. The Statkraft Group is a generator of renewable energy, as well as Norway's largest and the Nordic region's third largest energy producer. (Source: Wikipedia)



**Disclosure 305-1**
Direct (Scope 1) GHG emissions

**Reporting requirements**

Disclosure **305-1**

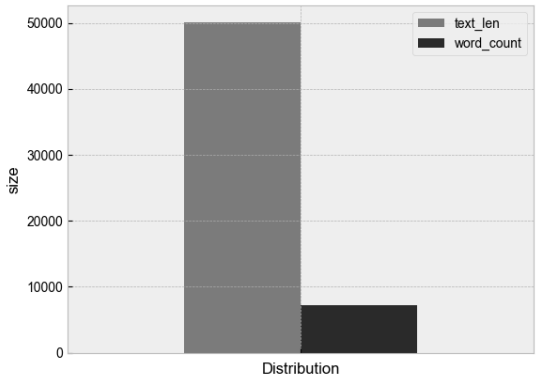
The reporting organization shall report the following information:

a. Gross direct (Scope 1) GHG emissions in metric tons of $CO_2$ equivalent.
b. Gases included in the calculation; whether $CO_2$, $CH_4$, $N_2O$, HFCs, PFCs, $SF_6$, $NF_3$, or all.
c. Biogenic $CO_2$ emissions in metric tons of $CO_2$ equivalent.
d. Base year for the calculation, if applicable, including:
    i. the rationale for choosing it;
    ii. emissions in the base year;
    iii. the context for any significant changes in emissions that triggered recalculations of base year emissions.
e. Source of the emission factors and the global warming potential (GWP) rates used, or a reference to the GWP source.
f. Consolidation approach for emissions; whether equity share, financial control, or operational control.
g. Standards, methodologies, assumptions, and/or calculation tools used.

*Figure 24 Text sample of Standard 305_1*

Next we put the descriptive results in parallel to have a better idea of what we are facing:

Standard 305




Stackcraft

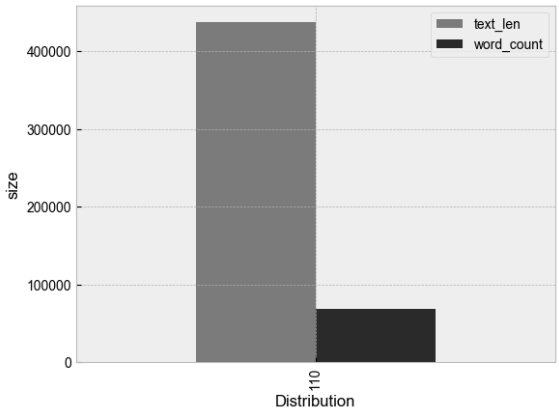



*Figure 25. Distribution of the text by length and number of words of GRI-305 and Skatkraft*

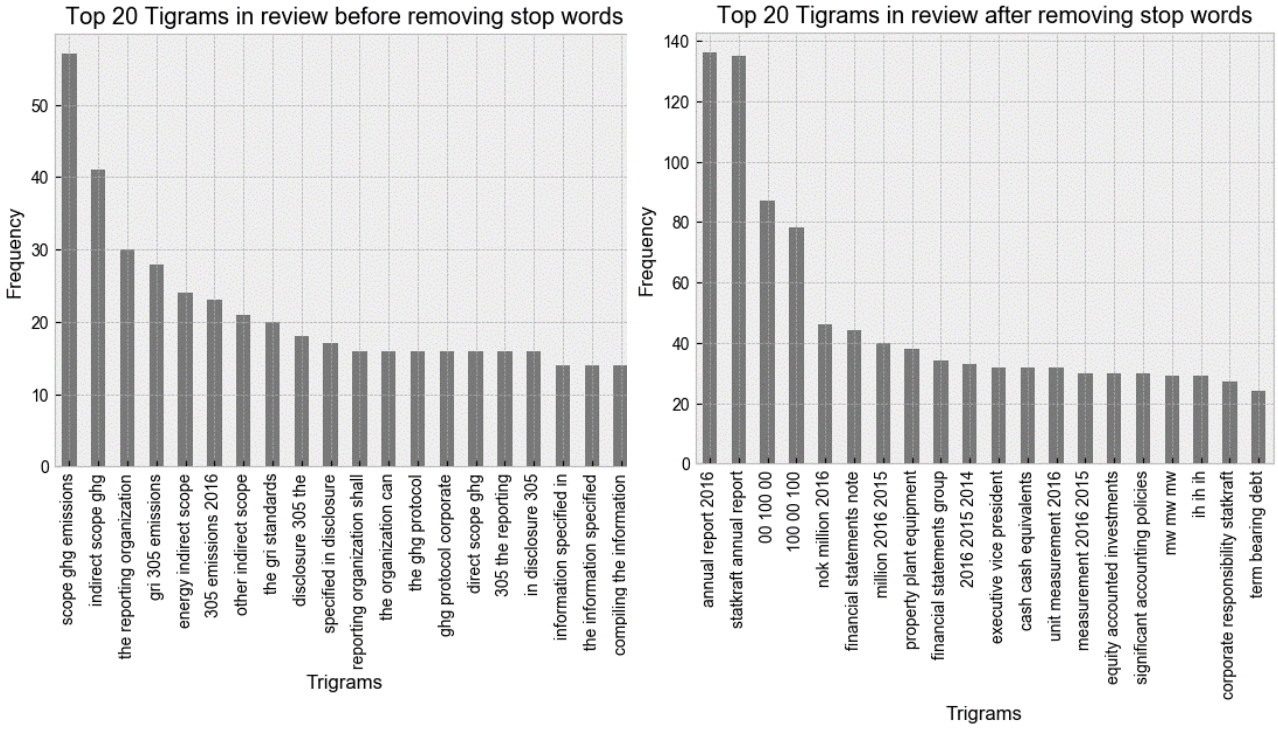



*Figure 26. Distribution of Top trigrams of GRI-305 and Skatkraft*

* The objective is to have a snapshot about raw values. We implemented other variants using stemming and lemmatization, but the differences were not significant.

* The numbers have not been eliminated, because they are very important for these documents if they are correctly associated.

Both tables, tell us immediately, that they have a relationship with the business environment, reports and energy. being the most frequent terms "scope gri reporting" and "indirect scope ghg" for the emissions side and "annual report 2016" and "statkraft annual report" for Skatcraft. Apparently, very little knowledge can be extracted directly from word strings.

Now is important, and despite the we have very little text, we should check if the creation of a word embedding is feasible:

We use a classical projection methods to reduce the high-dimensional word vectors to two-dimensional plots and plot them on a graph. The visualizations can provide a qualitative diagnostic for our learned model.

This is the representation for example of only emissions (building our own corpus using the standard 33) and implementing a default Word2Vec model.

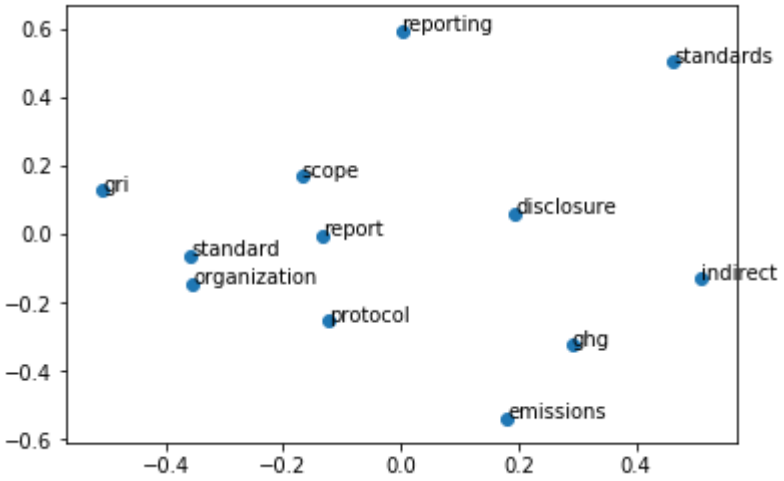



*Figure 27. Applying word2vec to a Standard 33 Corpus*

It is clear that the creation of a corpus for each standard will not be feasible for the assessment of semantic similarity between the documents.

But even so, we must continue trying to get to know our texts, that's why we are going to use LDA in order to extract the most relevant terms or topics in all our dataset text.

#### 3.4.3.1 LDA

The Latent Dirichlet Allocation (Blei, Ng,and Jordan, 2003) is a way to group semantically similar documents under a topic. The document could however belong to more than one topic but with different degree of membership. So, topic modelling could be seems as a text fuzzy clustering method. It is based on a simple exchange ability assumption for the topics and terms in a document where the topics are distributions over words and this discrete distribution generates observations (words in documents) (Blei, 2012). Tagging a document with a ranked list of semantic topics could be observed as a semantic information extraction. That is to say, the grouped documents per topic are semantically similar as they share common semantically related terms over the text corpus of what can be generally called discrete data collection where the probabilistic topic model was built on. For this model, both word order and document order do not matter. Knowing the terms that are used in each document and their frequencies already provides a good enough result to make decisions about which topic each one belongs to. Instead of working with the document-term matrix, you change to a subject-document matrix and thereby reduce the dimension. In this way, we would like to find some similarities between our documents.

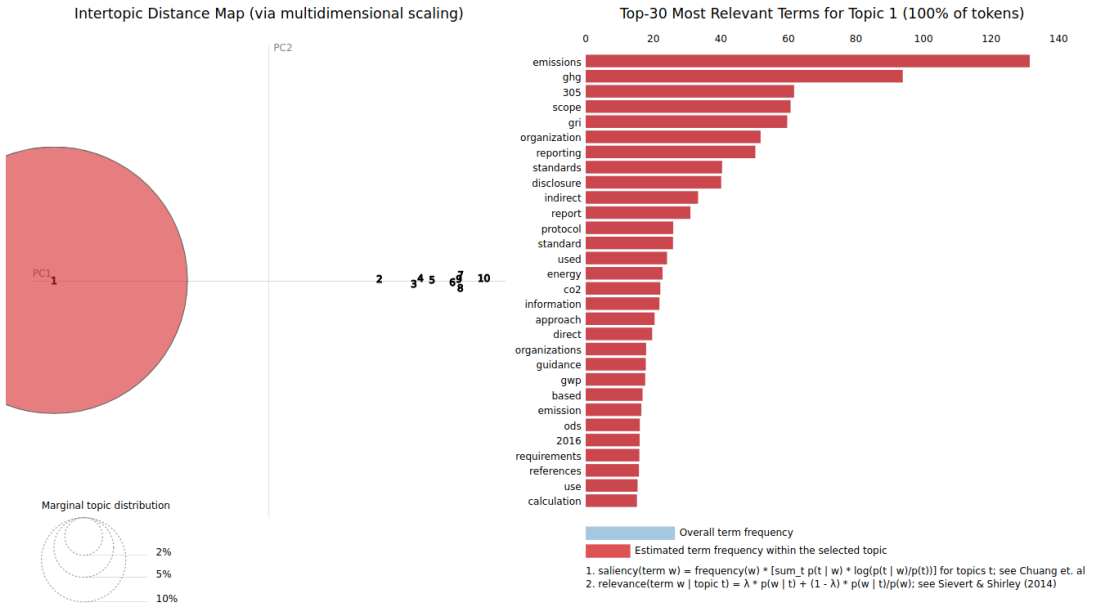



*Figure 28. LDA for GRI - 305*

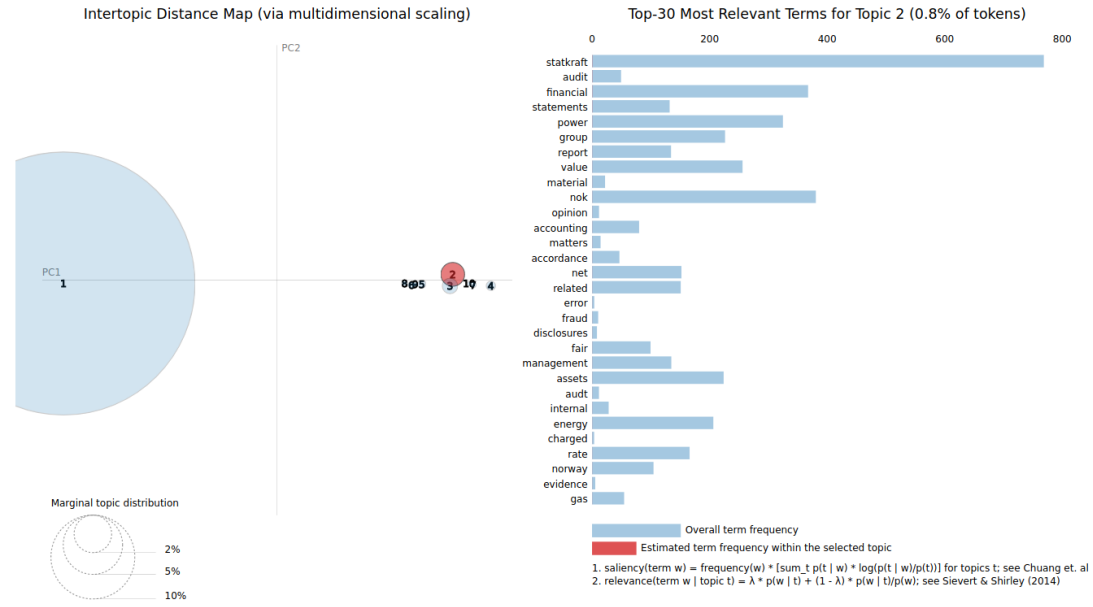



*Figure 29. LDA for Statkraft*

We can clearly see on Figure 27 and Figure 28 how LDA shows us the topics that have the most predominance in both texts, on the Emissions and Skatkraft:

Topics in LDA model:

| | |
|---|---|
| Emissions GRI-305 | |
| Topic #0 | emissions, starting, described, gri, transport, given, reports, scheme, base, |
| decreases, forcing, lead, classification, 16, reporting, bold, incentive, ghg, wants, activities. | |
| | |
| Skatfkraft | |
| Topic #0 | statkraft, nok, financial, power, million, value, group, assets, energy, note, |
| tax, total, rate, cash, market, risk, net, related, corporate, term | |

*Table 2. Topic #0 for GRI-305 and Skatkraft*

Both represent more than 70% of their marginal topic distribution in both texts. Only one term "energy" appears in Skaftkraft topic #6 (See Appendix B for more detail).

A comparison by topic cannot be made. Due to the total imposition of one topic over the others, as in the case of the standard emissions and the example company.

The topics are very similar and difficult to catalog.

#### 3.4.3.2 Visualizing how Corpora Differ

Now we would like to understand the terms association between their corpora. To carry out this task we will use the Scattertext tool[8].

In Figure 29, we can use the Scattertext plot for search terms that may be useful for GRI searching similarities through scaled f-score[9]. The most associated terms in each category make some sense as we saw with LDA, with “skatkfraft” and “emissions” as the most frequent terms.

Developing and using bespoke word representations Scattertext can interface with a Word2Vec model. Note the similarities produced reflect quirks of the corpus, e.g., "Climate" tends to be a one of the most frequent terms in both documents.

[8] Scattertext is a tool that's intended for visualizing what words and phrases are more characteristic of a category than others.

[9] While a term may appear frequently in both categories (High and Low rating), the scaled f-score determines whether the term is more characteristic of a category than others (High or Low rating).

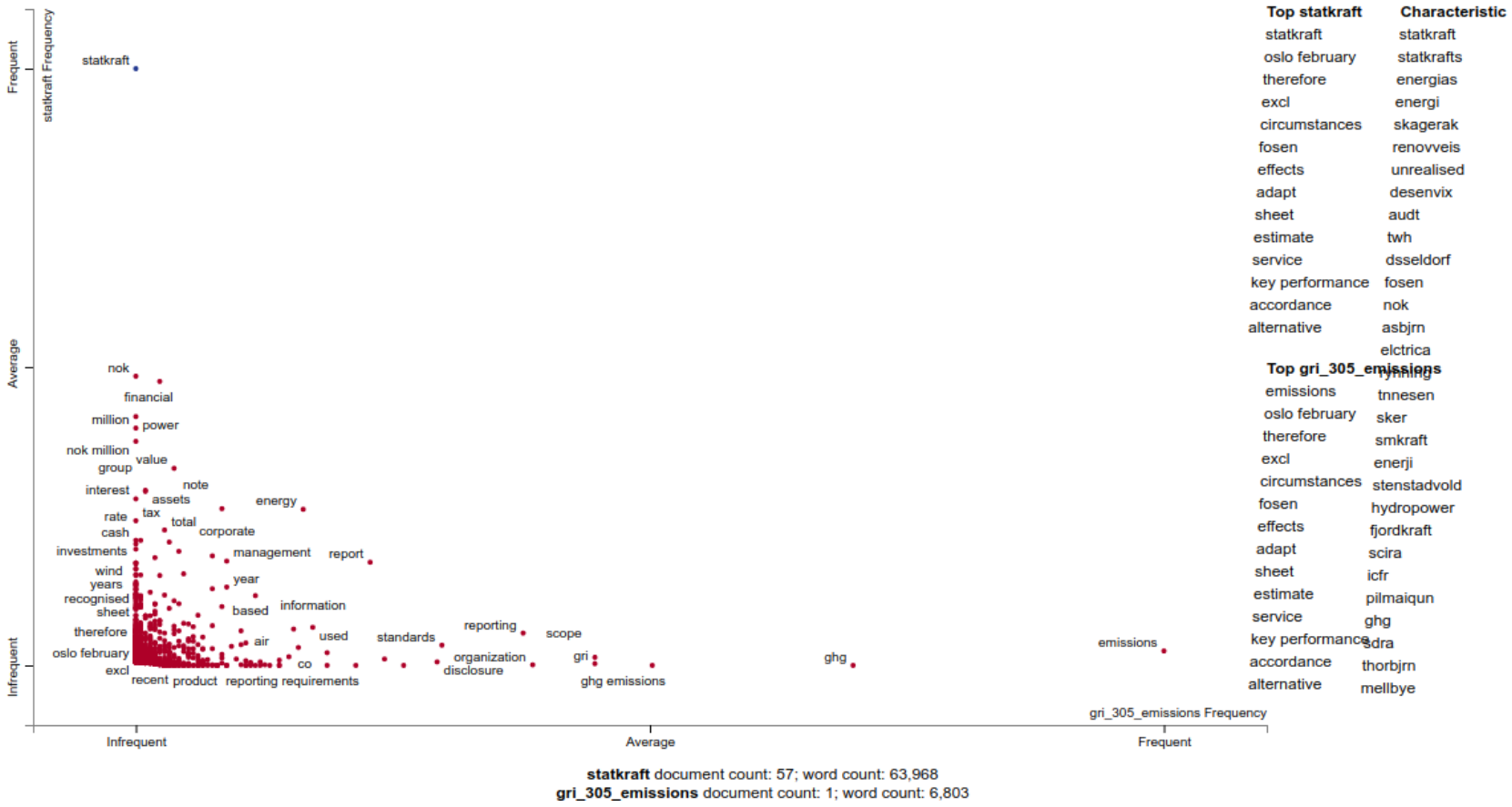



*Figure 30. Terms associations between GRI-305 and Skatkraft*


*Figure 31. Terms associations between GRI-305 and Skatkraft trough F-score*

We see that it would not be enough to implement models to calculate the semantic similarity of documents, because the information is not very descriptive and does not necessarily share the same technical terms. Therefore, we will have to reinforce this analysis with the help of Information Retrieval.

### 3.5 Matching the reports by Guidelines (IR)

Regardless of the degree of similarity, or the topics that can be associated between documents to be studied. We have to be able to do a search matching, and check what terms or standards are mentioned in the reports of the companies that coincide with the guidelines.

Therefore, we must design a strategy linked to controlled vocabularies and the definition of descriptors will be listed in a vocabulary of a closed and normalized domain, called controlled. In this vocabulary, there may even be interrelationships between these terms. How could it be the association of the standard number with the title or the description of it. The objective of this vocabulary control would be to try to solve two of the main problems of information retrieval: polysemy, homonymy and synonymy.

The relationship of these vocabularies will have to be of a hierarchical type, of relationship and equivalence.

#### 3.5.1 Recovery Measures

The performance of an information retrieval system can be measured by analysing the data (or documents) recovered from a query. There are two main measures:

- Precision: volume of relevant data among the total data recovered
- Completeness: volume of relevant data among the total of relevant data in the repository or the DB

Both measures tend to evolve in reverse (Cleverdon's Law). The more the precision increases the more the exhaustivity decreases, and vice versa. This is because they measure different factors, noise and silence:

- Noise: non-relevant information retrieved
- Silence: unrecovered information that is relevant

Given that in order to calculate these measures, it is necessary to know how many relevant elements exist, it is necessary to list the relevance of the documents before a set of queries. These listings are called test collections.

### 3.5.2 Recovery Models

Recovery models try to calculate the degree to which a certain element of information responds to a certain query. In general, this is achieved by calculating the coefficients of similarity (Cosine, Phi, etc.). The three most used models are:

- Boolean: one set is created with the elements of the query and another with the documents, and the correspondence is measured.

- Vectorial: in which the query and the terms of the document are represented by two vectors, and the degree to which both vectors diverge is measured.

- Probabilistic: the probability that the document responds to the query is calculated. Frequently uses feedback. The feedback is based on the user indicating which documents are more similar to their ideal response, in order to reformulate the query.

The application of IR we can see on the next chapter and its results on chapter 5.

## 3.6 Conclusions

As we saw, the implementation of similarities by topic modelling are discard, as well the creation of a own corpus, after this short evaluation process of our problem is clear that we need to test the models with more popularity based on word, sentence and hybrid measures. These we will see in the next chapter. And left for the final chapter the evaluation of this process of experimentation.

It will also be necessary to implement solutions with pre-trained algorithms. Discard the ideas of the Modapothala (2014) because a text classification with supervised learning, could not be possible, due to the change in the GRI methodology.

Therefore, calculating the degree of semantic similarity that the documents have with the guidelines, and more precisely abstracting the terms that coincide with keywords of the guidelines themselves, will be the basis to be able to extract some information on the affinity of the reports to the general and specific requirements described in the GRI standards.

# 4 EXPERIMENTATION

In this chapter we describe the details of the implementation of the system. All the solution was based on Python 2.7 and 3.6 versions (Van Rossum and Drake, 1995). The libraries have been used to carry out the project, which we list below. Also tools open source, among others mentioned above.

## 4.1 Tools for Data collection, preprocessing and transformation

All data for the GRI reports are obtained from official GRI database[10]. We focus on the latest reports, which quoted from the all Nordic companies. In total 450 reports that correspond to G3 and G4 versions of which 424 are in English (See Figure 17). We select only the reports written in English. All GRI reports were released in PDF format. GRI reports have no predefined format and structure therefore reporting entities have full flexibility on how, where and to what extent to disclose information. It is therefore safe to believe that this input is completely unstructured when it comes to searching for a particular data.

Nowadays the reports use more visualizations in order to facilitate the explanation on the state of health of the company. This means that the methodology that consist of convert a pdf format to text format in an attempt to define a hierarchical structure of data, used in previous works such as (Shabi et al., 2015), would be obsolete.

In this way we created two modules for extract and save the text in a data base (See Figure 32). One module export the PDF file to html format using several tools creating two folders: one for html files and other for images. An overall view of modular architecture is illustrated in Figure 31.

### 4.1.1 Collecting

*OCRopus OCR Library* (https://github.com/tmbarchive/ocropy), is a collection of document analysis programs that provide good tools for extracting text from many digital sources. I our case was the best solution available for extract text from images embedding in pdf documents.

*Textract* (https://github.com/deanmalmgren/textract/), is a pack-age that provides a single interface for extracting content from any type of file, without any irrelevant markup.

### 4.1.2 Encoding

*spaCy. (*https://spacy.io/*),* is a specialized NLP tool that introduce a novel tokenization algorithm that we confirm gives a better balance between ease of definition and ease of alignment into the original string.

*NLTK. (*https://www.nltk.org/*)* The RegexpTokenizer help us to splits a string into substring using regular expressions.

[10] https://database.globalreporting.org/

### 4.1.3 Vectorization of text and calculation of similarities

*Scikit-Learn (Fabian et al. 2011).* The TfidfVectorizer class has been used from this library. Converts a collection of documents into an array of TF features IDF, used as the standards vectorizer for based engines in TF-IDF of the system.

*Gensim (Rehurek and Sojka 2010).* Gensim is an open source platform in Python for modeling vector side of texts and thematic modelling. It is specifically designed mind to handle large collections of texts, using streaming data and efficient incremental algorithms. We use the genism class to vectorization engines of the system, which implements the algorithm Doc2Vec and pre-trained models as Glove, fastText and Word2vec.

*Tensorflow (*https://www.tensorflow.org/*). tf.Hub.* TensorFlow Hub is a library of reusable machine learning modules. The example solution uses the Universal Sentence Encoder pre-trained text-embedding module to convert each title to an embedding vector.

*sparse dot topn (*https://pypi.org/project/sparse-dot-topn/*).* To calculate the similarity between two vectors of TF-IDF values, Cosine similarities are usually used, which can be seen as the normalized dot product between vectors.

### 4.1.4 Graphics

*seaborn, matplotlib (Hunter, 2007).* They have served to make word clouds, graphics and visualize statistics during the project.

### 4.1.5 Storage

*MySQL.* To store data from both the corpus, validations and recommendations MySQL relational database engine has been used. The tool allowed to create databases, tables and insert values and quickly query from python with SQL statements.

*Pickle (*https://github.com/python/cpython/blob/3.8/Lib/pickle.py*).* To store the trained engines, a pickle. Pickle is a utility that allows python objects to be saved to disk.

### 4.1.6 Text preprocessing

*re (Friedl, 2009).* Library to use regular expressions in python. It was used in the preprocess, on the text of the standards in the definition of filters individuals.

*NLTK*. English stopwords were used to exclude these words of each standard and from the same library we used WordNetLemmatizer and PorterStemmer classes.

*Bs4 (*https://pypi.org/project/beautifulsoup4/*).* Beautiful Soup is a Python package for parsing HTML and XML documents. It creates a parse tree for parsed pages that can be used to extract data from HTML, which was useful for web scraping

*Textblob (*https://textblob.readthedocs.io/en/dev/*).* It is a library for processing textual data. It was used for part-of-speech tagging and noun phrase extraction.

### 4.1.7 Multithreading

*joblib, threading (*https://joblib.readthedocs.io*)*. After completing the sequential version of the project, added currence in order to improve computing times. They were identified calculations independent of each other in intermediate stages of training ment, assigning to each task a process and a CPU core.

### 4.1.8 Information Extraction

*Fuzzywuzzy (*https://github.com/seatgeek/fuzzywuzzy*)*. Based on Fuzzy Logic this tool was used for string matching process.

### 4.1.9 Others

*docopt*: To set the python scripting interface online from bash commands.
*lxml*: To load, save xml files and extract their data.
*Collections:* for dictionaries and counters, time for control time, os for directory management, csv for stored data in .csv format.
*Scattertext:* for visualizations o F-score with embeddings
*GCP:* shell for monitorizing and execution of the routines.

## 4.2 Hardware

This project has been implemented on the hardware provided by the Google Cloud Platform, used for tests and for deployment.

We build the following instances:

Instance 1:
24 vCPUs, 105 GB memory
SO
Ubuntu 18.04

Instance 2:
72 vCPUs, 240 GB memory
SO
Ubuntu 18.04

Instance 3:
10 vCPUs, 37.5 GB memory
GPU's
1 x NVIDIA Tesla V100
SO
ubuntu-1804-bionic-v20200317

Instance 4:
12 vCPUs, 16 GB memory
GPU's
1 x NVIDIA GTX 1060
SO
Ubuntu 18.04

Instance 5:
10 vCPUs, 37.5 GB memory
GPU's
1 x NVIDIA Tesla V100
SO
ubuntu-1804-bionic-v20200317

### 4.3 Architecture

The Figure 31 shows an overview of the final solution architecture. The system extracts and process the text, validate and build an approximate similarity matching index, finally serves the build index for semantic search and retrieval.

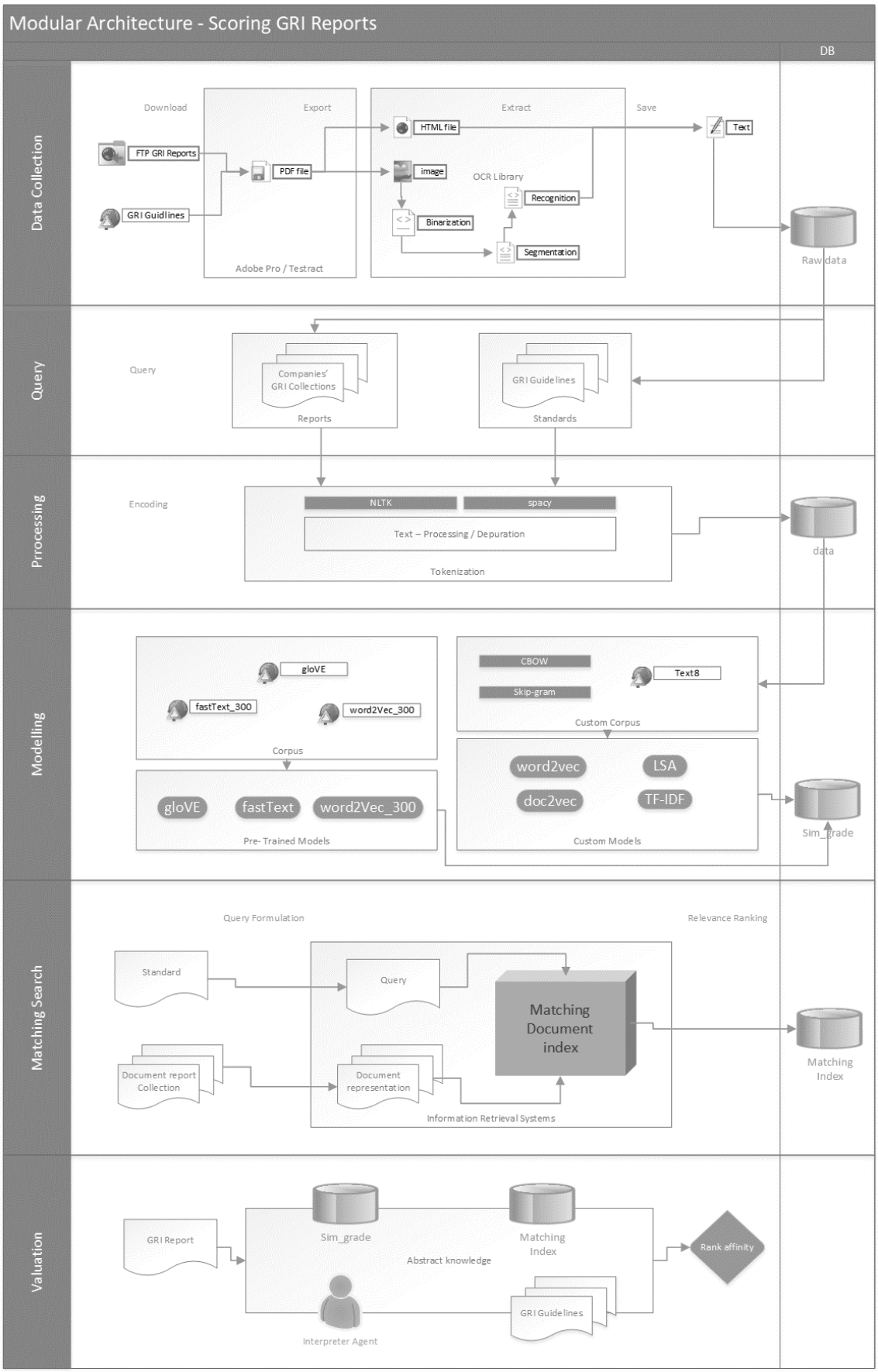



*Figure 32. High-level solution architecture for the text semantic search system*

### 4.3.1 Key modules of the architecture

#### 4.3.1.1 Corpus creation: scraping and preprocessing

To obtain the corpus, we combined Testract and Ocropus on order to extract the text from pdf files, despite being a routine process, we had to adjust many parameters for the task of extraction of text embedded in the images of the pdf themselves. (See Figure 31).

Then scratch clean up routines were applied i.e. loaded to obtain standards in XML format where the data extracts are neatly stored in labels. Of each of these pdf metadata are extracted, such as type of title and standard number among others they are saved on a table for further manipulation.

#### 4.3.1.2 Text preprocessing

For each tokenization process, the sentence is filtered by a depuration process, where we defined the politics of treatment of the manipulation of the text. e.g. the boundaries of the minimum number of words that can build a sentence.

Stages:

1. Pre-processing of the standard plain text (input text, output token list) with preprocess string from gensim.parsing. preprocessing using the text as filters:

    Remove tags of the form <w *> (appearance of possible tags)
    Remove excess spaces and \ n

2. Then, to each token:

    a) convert it to lowercase
    b) ignore it if it is inside the set of specific wordsto ignore (not for example)
    c) ignore it if it is a English stopword.
    d) remove unwanted characters like quotes ", apostrophe' ,or the or symbol, using regular expressions
    e) selectively treat the case that the token contains any subword containing the characters in the set. \ / and it is surrounded by numbers

#### 4.3.1.3 Training and saving the models

In this section we configure the parameters to develop an environment where the models that are going to be executed, to be able to be compared later. About architecture for embeddings we found that bag of words was very slightly faster and produced better results than skip-gram.

1. **Training algorithm**: tf-idf, LSA, word2vec custom (see next section), word2vec (300), fastText (300), and GloVE (3).
2. **Downsampling of frequent words**. We set a values around 0.001.
3. **Word vector dimensionality**: We used 300.
4. **Context/window size**: We define as 5 as minimum on window size.

5. **Worker threads**: We define as top capacity for take all the power of our hardware available.
6. **Minimum word count**: For sentence similarities the value was setup to 4.
7. **Similarity measure**: Soft Cosine and cosine similarity.

#### 4.3.1.3 Database

In order to store the results of the models, a database is created at the beginning of the training stage, with the following tables that they will be filled at runtime.


*Figure 33. Database*

We split the parameters and results of each model on a different tables, in order to mitigate the risk of running exceptions.

### 4.4 Overall workflow

In order to extract the similarities from documents against the GRI standards reports we need to design a similarity matching system, this means that: in the first instance we need to represent items as numeric vectors. There vectors in turn represent semantic embeddings of the item discovered through the models mentioned.

Later we need to organize and store these embeddings for apply cosine distance in order to find similar to the embedding vector of the standard query.

The solution described in this research illustrates an application of embeddings similarity matching in text semantic search. The goal of the solution is to retrieve semantically relevant documents compare with the standards query.

The workflow of the semantic search system proposed illustrated in Figure 11 can be divided into the following steps:

1. **Extract embeddings using Module 1 and 2**
    1. Read the pdf files from GRI Database.
    2. Extract the text embeddings using our set of algorithms in module 2.
    3. Store the extracted embeddings in the Database.
    4. Store the original text and their identifiers in Datastore.

2. **Build the index using AI Platform using module 3**
    1. Load the embeddings from the files in Database into the GRI index.
    2. Build the index in memory.
    3. Save the index to disk.
    4. Upload the saved index again to Database.

3. **Serve the scoring**
    1. Download the Guideline index from Database.
    2. Extract the query embedding using module 2.
    3. Using the GRI index, find embeddings that are similar to the query embedding.
    4. Get the item IDs of the similar embeddings.
    5. Retrieve the GRI reports titles using the identifiers from Datastore.
    6. Return the results.

#### 4.4.1 Vectorization models

To carry out the vectorization of the tokens of each standard, implemented the two aforementioned engines in the system. Both have undergone experiments to analyse their performance and quality of recommendations.

Doc2Vec

Gensim implementation of the Doc2Vec method. The great capacity of this library allowed to make the core task of the project, which links the mathematical calculations to obtain the order of the recommendations. The following parameters have been used in training the model:

- vector size = 300. Vector size.
- dm = 0. Vectorization algorithm. PV-DBOW is used.
- alpha = 0.01. Initial learning rate.
- min alpha = 0.0001. The learning rate drops freely. neally at this value as the training process.
- window = 10. Window size.
- min count = 1. Words frequently in the corpus less than this value are ignored in training.
- workers = 16. Number of sub-processes.
- epochs = 50. Number of training iterations of model.
- dbow words = 0. value 1 to construct vectors of words, 0 otherwise.

Internally, textual queries are obtained through inferences to a vector in the model based on the text tokens. The similarities in Doc2Vec are found with the most similar function that internally computes the cosine similarity calculation.

Default values from the library were used for the rest of the model parameters that do not appear in the list.

Tf-Idf

Engine using Scikit-Learn TF-IDF vectorizer. We have used the parameters:

- analyzer = 'word'. The features are composed of words.
- ngram range = (1, 3). Sets the length of n-grams (1, 2 and 3) to take into account during training. That is to say, in the matrix columns we will also have bigrams and trigrams.
- min df = 0. Also called a cut-off value, to ignore the terms that have a strict document frequency-mind below this threshold.

### 4.4.2 Pre-trained models

The gensim package has nice wrappers providing us interfaces to leverage pre-trained models available under the gensim.models module.

### 4.4.3 Information Retrieval

For the extraction we implement fuzzywuzzy determining at least the ratio 95, but we also apply a coverage to the similarity between neighbouring sentences. This, because it was found that the terms for matching are often tokenized in different sentences.

```
df_sentences = extract_sectences_fuzzy(row[4])
   similar_sentences = process.extractBests(title,df_sentences[0])
   for line in similar_sentences:
       if line[1] > 85:
           good_rate.append(line[2])
           ....
   for i in good_rate:
       for j in df_sentences[df_sentences[0].str.contains(gri_standard)].index:
           if i == j or abs(i-j) < 4:
               column_std = "c" + str(gri_standard_number) + "_" + str(gri_standard_spec)
               sql = "UPDATE " + table_name + " SET " + column_std + " = (%s) WHERE name_company = (%s)"
        #if the search back empty or is not necesarry to make the test for all vector
               break
```

*Figure 34. Snippet of index matching*

### 4.5 Search and Semantic systems in practice

Below we present the table of executions, the results of which are analysed in the next chapter.

*Table 3. Design of executions*

| Pre-Process | Feature Extraction | Similarity Measure | Text Similarity based |
|---|---|---|---|
| Stop words removal | TF-IDF | word | String |
| Punctuation removal | word2vec (Custom Trained) | cosine | Corpus |
| Lemmatization | LSA | cosine by sentence | Corpus |
| Spell Correction | doc2vec (Custom Trained) | cosine | Hybrid |
| Abbreviation | fastText | softcosine | Hybrid |
| Accept numbers | word2vec_300 | softcosine | Hybrid |
| | gloVE | softcosine | Hybrid |
| | USE | cosine | Hybrid |
| | IE | matching | String |
| text_8 corpus | word2vec (Custom Trained) | Cosine | Corpus |
| | doc2vec (Custom Trained) | cosine | Hybrid |

In practice, search and retrieval systems often combine semantic-based search techniques with token-based (inverted index) techniques.

### 4.5.1 Volumetrics and load testing

After the example solution was created, a sample run was performed to get performance information. The following tables show the settings that were used to run the end-to-end example using the dataset.

#### 4.5.1.1 Extracting embeddings

The following table shows the configurations of the gloVE job used to extract the embeddings and the resulting execution time.

| | |
|---|---|
| Configuration | • Record limit: **5 million**<br>• Embedding vector size: **512**<br>• vCPUs: **64** (32 worker)<br>• Worker machine type: **instance-1** |

Results: Job time : **32 minutes**

#### 4.5.1.2 Building the index, time by launch

The following table shows configuration and results information for the task of building the index using an GCP job.

| | |
|---|---|
| Configuration | • GRI index number of trees: **100**<br>• GCP platform scale tier: **large_model (gloVE.01)** |
| Results | • Job time: **17 hours, 56 minutes**<br>• Index file size: **2.12 GB**<br>• Mapping file size: **263.21 MB** |
| Configuration | • vCPUs: **6**<br>• Memory: **24 GB**<br>• Disk: **50 GB**<br>• Scaling: **manual 4 instances** |
| Results | • Deployment time (up and running): ~**19 minutes**<br>• Building and uploading the container image: **~6 minutes**<br>• Concurrency level: **1500**<br>• Requests per seconds: **~2500** |

# 5 RESULTS

## 5.1 Results

As we had previously commented, the evaluation of the results provided by the implemented models are only a complementary part to a deeper analysis that is required to know the true degree of affinity that the reports presented can have under the latest framework applied by GRI. This is not due to the bias that the models themselves present or due to lack of adjustments, but rather to the natural subjectivity and associated with the difference of opinions about which is the semantic relationship between texts.

While dealing then with a problem, as subjective as the assessment of texts, for which it has been necessary to implement UL tools, we are going to introduce the results in an intrinsic way of assessment. Intrinsic assessment (Barakrov Amir, 2018) are experiments in which the results are compared by human judgments on words relations (See Chapter 2 for more details).

To proceed with this assessment, we will then use the standard mime and report selected in Chapter 3. That is, the GRI_305 standard that corresponds to the emissions section and the CSR of the Norwegian company Statkraft.

For this effect, we will prepare the following control data set for the evaluation of the models (See Chapter 4 for more details):

For the analysis of words, the terms will be used: {"emissions", "sustainability", "gri"}

> *Emissions:* It is a specific term used in GRI_305 standard, which is related to control measures regarding the level of emissions produced by the company.
>
> *Sustainability*: For using a generic term, related to ESG practices.
>
> *GRI*: It is a not relevant term for general use or pretrained corpus, but has a particular definition for our context.

For the analysis of the sentences, we extract some sentences from the Statkraft GRI report. The following sentences will be used:

> a = "Statkraft’s power plants have low variable costs, long lifespans and low carbon emissions".
>
> b = "Statkrafts high level Climate Roundtable gathered scientists, business leaders and politicians to explore new business solutions to the climate challenge".
>
> c = "However, 233 minor environmental incidents were registered (228 for 2015)".

In the same way as for the analysis of words, the phrases are selected, based on their specificity to our studied topic, in this case GRI_305.

The sentence a, is an example of a specific phrase that determines an objective pursued by our standard: the reduction of emissions. It is a specific sentence in which it is clearly

indicated that one of the disclosures of the GRI-305 standard has been achieved. The sentence b, does not become so specific in its semantic meaning, however it does have many words related to emissions. Finally, phrase c is a very common sentence in this type of corporate report, leaving its interpretability open and with very little relation to our topic.

## 5.2 Conclusions

### About similarity of words

In the following table we can see a summary of which terms are extracted as similar from our models:

*Table 4. Similarity by words*

| model | Word2VecCustom | | word2vec model 300 | | FASTTEXT | | Doc2vec | | Glove | |
|---|---|---|---|---|---|---|---|---|---|---|
| | word | degree | word | degree | word | degree | word | degree | word | degree |
| emissions | u'greenhouse' | 0.8671912 | nn | | emission-control | 0.7493376 | bulk | 0.966668 | sustainable | 0.68704343 |
| | u'depletion' | 0.7348825 | nn | | emissions-related | 0.7408717 | pollutants | 0.95382 | governance | 0.56540704 |
| | u'dioxide' | 0.7332385 | nn | | emission-reduction | 0.7376918 | fuel | 0.945321 | environmental | 0.54441196 |
| | | | nn | | *for positions later:* | | | | | |
| | | | | | greenhouse-gas | 0.7204161 | | | | |
| sustainability | developed | 0.9711098 | environmental _sustainability | 0.8592561 | self-sustainability | 0.7990537 | scrutinise | 0.965396 | environmental | 0.85044056 |
| | director | 0.9612654 | sustainable | 0.7534031 | sustainable | 0.7828761 | seignorage | 0.961584 | initiative | 0.85044056 |
| | encourage | 0.9591941 | environmental _stewardship | 0.7027169 | non-sustainability | 0.7822891 | part-time | 0.955912 | responsibility | 0.85044056 |
| gri | standard | 0.973475 | nn | | hali | | part-time | 0.984044 | global | 0.8322165 |
| | 102_general | 0.9631332 | nn | | gra | | disclosure | 0.961802 | board | 0.82135171 |
| | foundation_gri | 0.9613792 | nn | | dri | | country-by-country | 0.954183 | dri | 0.81995618 |

Now, we can see how the doc2vec model, despite of making an interesting connection with the term "emissions", we can see that the doc2vec models does not seem to be the case for the term "sustainability". The domain management also stands out, when the specified corpus models are executed, as in the case of word2vecCustom, which was the only one, logically, was able to extract the similarity that we expected about the term "gri" relating it to "standard" or "102_general ” (Which 102 corresponds to the Standard that describes general aspects of the companies). And not so for fastText, which we can see how lexical similarity helps define its results. Instead, it is interesting to see the strong relationship presented by the word “emission” together with “governance” and “sustainable” for gloVE, which is closer to the guidelines determined by the GRI Framework.

### About similarity of sentences

Following the previous structure, we present the results table below, with the sentences with the highest degree of similarity:

*Table 5. Similarity by sentences*

| Sentence Control | model | Similaritie sentence | sim grade |
|---|---|---|---|
| a | word2vecCustom | emission set reporting requirement topic emission | 0.495452 |
| | | emission | 0.492307 |
| | | region emission cap volume emission also direct cost implication | 0.45344 |
| | doc2vec | detail locationbased market-based method available ghg protocol scope 2 guidance | 0.245192 |
| | | chosen emission factor originate mandatory reporting requirement voluntary reporting framework industry group | 0.160205 |
| | | thus rate used disclosing ghg emission conflict national regional reporting requirement | 0.157251 |
| | fasttext | biogenic carbon dioxide co2 emission emission co2 combustion biodegradation biomass carbon dioxide co2 equivalent measure used compare emission various type greenhouse gas ghg based global warming potential | 0.307128 |
| | | region emission cap volume emission also direct cost implication | 0.301845 |
| | | emission 2016 calculation based published criterion emission factor gwp rate direct measurement ghg emission continuous online analyzer estimation | 0.259396 |
| | GLOVE | region emission cap volume emission also direct cost implication | 0.359467 |
| | | biogenic carbon dioxide co2 emission emission co2 combustion biodegradation biomass carbon dioxide co2 equivalent measure used compare emission various type greenhouse gas ghg based global warming potential gwp note co2 equivalent gas determined multiplying metric ton gas associated gwp | 0.354669 |
| | | primary effect element activity designed reduce ghg emission carbon storage | 0.323098 |
| | Word2vec_300 | biogenic carbon dioxide co2 emission emission co2 combustion biodegradation biomass carbon dioxide co2 equivalent measure used compare emission various type greenhouse gas ghg based global warming potential | 0.252893 |
| | | ghg emission include co2 emission fuel consumption | 0.214474 |
| | | emission 2016 | 0.212684 |
| | TF.IDF | In regions with emission caps, the volume of emissions also has direct cost implications.) | 0.35 |
| | | This Standard covers the following GHGs: Carbon dioxide () | 0.26 |
| | | iogenic carbon dioxide (CO2) emission emission of CO2 from the combustion or biodegradation of biomass carbon dioxide (CO | 0.25 |
| | LSA | other significant air emissions Pollutants such as NOX and SOX have adverse effects on climate, ecosystems, air quality, habit | 0.9 |
| | | e.g., from coal mines) and venting; HFC emissions from refrigeration and air conditioning equipment; and methane leakages (e. | 0.9 |
| | | significant air emission air emission regulated under international conventions and/or national laws or regulations | 0.999994 |
| b | word2vecCustom | climate change 2007 | 0.476296 |
| | | business travel 7 | 0.458803 |
| | | intergovernmental panel climate change ipcc climate change 1995 | 0.430374 |
| | doc2vec | chosen emission factor originate mandatory reporting requirement voluntary reporting framework industry group | 0.17624 |
| | | organization-specific metric denominator chosen calculate ratio | 0.160024 |
| | | chosen emission factor originate mandatory reporting requirement voluntary reporting framework industry group | 0.135743 |
| | fasttext | intergovernmental panel climate change ipcc climate change 1995 | 0.272688 |
| | | management approach | 0.25423 |
| | | business travel 7 | 0.248638 |
| | GLOVE | intergovernmental panel climate change ipcc climate change 1995 science climate change contribution working group second assessment report intergovernmental panel climate change 1995 | 0.359562 |
| | | intergovernmental panel climate change ipcc climate change 2007 physical science basis contribution working group fourth assessment report intergovernmental panel climate change 2007 | 0.325229 |
| | | example impact economy environment society lead consequence organization business model reputation ability achieve objective | 0.24639 |
| | Word2vec_300 | business travel 7 | 0.228121 |
| | | intergovernmental panel climate change ipcc climate change 1995 | 0.175295 |
| | | climate change 2007 | 0.175295 |
| | TF.IDF | Intergovernmental Panel on Climate Change (IPCC), Climate Change 1995:) | 0.32 |
| | | Climate Change 2007:) | 0.29 |
| | | The Science of Climate Change, Contribution of Working Group I to the Second Assessment Report of the Intergovernmental Panel on Climate Change, 1995. ) | 0.24 |
| | LSA | b. If applicable, gross market-based energy indirect (Scope 2)) | 0.99983764 |
| | | GHG emissions by: 2.4.5.1 business unit or facility; 2.4.5.2 country; ) | 0.99978733 |
| | | a. Gross location-based energy indirect (Scope 2)) | 0.9996985 |
| c | word2vecCustom | emission topic-specific gri standard 300 series environmental topic | 0.381007 |
| | | amendment ghg protocol corporate standard 2015 | 0.367676 |
| | | topic-specic disclosure disclosure 305-1 direct | 0.33721 |
| | doc2vec | united nation environment programme unep convention stockholm convention persistent organic pollutant pop annex b c 2009 | 0.273063 |
| | | detail locationbased market-based method available ghg protocol scope 2 guidance | 0.165613 |
| | | reporting recommendation 2 10 compiling information specified disclosure 305-5 reporting organization subject different standard methodology describe approach selecting | 0.157341 |
| | fasttext | gri 305 emission 2016 2 | 0.236812 |
| | | note significant air emission include listed environmental permit organization operation | 0.221944 |
| | | emission 2016 | 0.21918 |
| | GLOVE | many organization track environmental performance intensity ratio often called normalized environmental impact data | 0.273903 |
| | | significant air emission include example persistent organic pollutant particulate matter well air emission regulated international convention national law regulation including listed organization environmental permit | 0.193188 |
| | | however neither document extract may reproduced stored translated transferred form mean electronic mechanical photocopied recorded otherwise purpose without prior written permission gri | 0.187834 |
| | Word2vec_300 | many organization track environmental performance intensity ratio often called normalized environmental impact data | 0.171785 |
| | | note significant air emission include listed environmental permit organization operation | 0.136536 |
| | | emission topic-specific gri standard 300 series environmental topic | 0.111803 |
| | TF.IDF | An amendment to the GHG Protocol Corporate Standard, 2015.) | 0.43 |
| | | Many organizations track environmental performance with intensity ratios, which are often called normalized environmental impa | 0.2 |
| | | Emissions is a topic-specific GRI Standard in the 300 series (Environmental topics).) | 0.17 |
| | LSA | United Nations (UN)) | 0.97 |
| | | Base year or baseline, including the rationale for choosing it.) | 0.97 |
| | | All defined terms are underlined.) | 0.95 |

The first 3 sentences with the highest degree of similarity have been selected and they are presented on the Table 5. From table 5 we need to determine which sentences have greater accuracy when comparing them with our control sentences. LSA, for our appreciation stands out above the others in the control statement a; but declines quite a lot with the control statement c. gloVE instead seems to handle better in generalist statements such as b and c, but not so much in more specific as in a, fasText continues to demonstrate that the lexicon is one of the most important points to value as well as tf-idf. With the other models, we find it difficult to abstract a more homogeneous conclusion, due to the diversity of its results.

Therefore, we decided to combine the results provided by LSA and gloVE, because we believe that both are complementary to our problem environment. In this way we would try to balance the lack of text with gloVE and the specificity of the documents with LSA.

Below we present the first ten reports with the average of cosine valuations by LSA and gloVE as total in descendent order:

General:

*Table 6. Top 10 semantic similarity*

| rank | name_company | country | year | total |
|---|---|---|---|---|
| 1 | nsb_group | Norway | 2018 | 0.782332 |
| 2 | ikea_(uk,ireland) | Sweden | 2016 | 0.768476 |
| 3 | x_op | Finland | 2016 | 0.751637 |
| 4 | vestas_wind_systems | Denmark | 2018 | 0.747473 |
| 5 | tdc | Denmark | 2016 | 0.744967 |
| 6 | x_pohjolan_voima | Finland | 2016 | 0.743451 |
| 7 | sydbank | Denmark | 2016 | 0.738819 |
| 8 | united_nation_office_for_project_services_(unops) | Denmark | 2018 | 0.736289 |
| 9 | tgs-nopec | Norway | 2018 | 0.735754 |
| 10 | x_stora_enso | Finland | 2016 | 0.720677 |

According to Table 6, the report prepared by the Norwegain company nsb_group in 2018 is the one with the highest assessment of semantic similarity to the guidelines proposed by GRI standards.

By Countries:

*Table 7. Top 10 semantic similarity by countries*

Sweden:

| rank | name_company | year | total |
|---|---|---|---|
| 2 | ikea_(uk,ireland) | 2016 | 0.768476 |
| 23 | advania | 2018 | 0.668866 |
| 48 | epiroc | 2018 | 0.601492 |
| 49 | sas_group_(sweden) | 2016 | 0.598343 |
| 57 | h&m_group | 2018 | 0.571747 |
| 61 | lundin | 2016 | 0.558097 |
| 62 | amf | 2016 | 0.557889 |
| 70 | boliden | 2016 | 0.538131 |
| 78 | seb | 2018 | 0.501366 |
| 79 | if_p&c_insurance | 2018 | 0.500537 |

Norway:

| rank | name_company | year | total |
|---|---|---|---|
| 1 | nsb_group | 2018 | 0.782332 |
| 9 | tgs-nopec | 2018 | 0.735754 |
| 25 | rica_hotels_(scandic) | 2017 | 0.665714 |
| 29 | bank_1_oslo_akershus_as | 2015 | 0.663055 |
| 36 | dno | 2018 | 0.645690 |
| 38 | agder_energi | 2018 | 0.644552 |
| 41 | avinor | 2018 | 0.635915 |
| 47 | eltek_power_systems | 2014 | 0.604682 |
| 51 | intex_resources | 2016 | 0.592722 |
| 56 | odfjell | 2018 | 0.573586 |

Denmark:

| rank | name_company | year | total |
|---|---|---|---|
| 4 | vestas_wind_systems | 2018 | 0.747473 |
| 5 | tdc | 2016 | 0.744967 |
| 7 | sydbank | 2016 | 0.738819 |
| 8 | united_nation_office_for_project_services_(unops) | 2018 | 0.736289 |
| 12 | aco | 2018 | 0.711786 |
| 15 | tivoli | 2017 | 0.698833 |
| 17 | bavarian_nordic | 2016 | 0.686570 |
| 20 | lm_group_holding | 2016 | 0.675025 |
| 22 | simcorp | 2014 | 0.671496 |
| 33 | egetaepper | 2017 | 0.657794 |

Finland:

| rank | name_company | year | total |
|---|---|---|---|
| 3 | x_op | 2016 | 0.751637 |
| 6 | x_pohjolan_voima | 2016 | 0.743451 |
| 10 | x_stora_enso | 2016 | 0.720677 |
| 11 | kesko | 2018 | 0.713064 |
| 13 | rahapaja | 2016 | 0.710043 |
| 14 | x_alko | 2016 | 0.707655 |
| 16 | sampo | 2018 | 0.694905 |
| 18 | varma | 2018 | 0.685204 |
| 19 | x_nordkalk | 2016 | 0.684067 |
| 21 | x_oriola-kd | 2016 | 0.673749 |

Iceland:

| rank | name_company | year | total |
|---|---|---|---|
| 67 | isavia | 2017 | 0.540004 |
| 77 | ossur | 2018 | 0.504857 |

It should be noted that Finnish reports have the best rankings in the overall picture, as its top 10 reports are in the top 21 of the total. It is followed by Denmark, putting its top 10 reports in the top 33, Sweden is below the top 79 and Iceland is in the top 77.

**About similarity matching by index**

Capturing the semantic similarity that a document can have is not a guarantee to know whether a report mentions compliance with a specific standard. Since June of 2018 the GRI standards, which is currently the last in force with respect to its predecessors G4 and G3. They suggest that a summary of what standards are being complied with in core or comprehensive be attached to reports where possible. Therefore, we will look for reports that match the guidelines described in chapter 4. The "total" field provides the number of disclosures that a report match with the guidelines. The "total_E", "total_S" and "total_G" fields provide the total amount that the reports match with the ESG Metrics of the World Federation Exchanges guidelines[11] mapped with the GRI Standards.

Next, as we did previously, we will present the 10 first reports with the highest matches according to the index guidelines.

General:

*Table 8. Top 10 index matching*

| rank | name_company | country | year | total_E | total_S | total_G | total |
|---|---|---|---|---|---|---|---|
| 1 | stora | Finland | 2018 | 12.0 | 6.0 | 6.0 | 127.0 |
| 2 | upm | Finland | 2018 | 12.0 | 5.0 | 6.0 | 115.0 |
| 3 | kesko | Finland | 2018 | 13.0 | 5.0 | 5.0 | 102.0 |
| 4 | palsgaard | Denmark | 2018 | 11.0 | 6.0 | 3.0 | 91.0 |
| 5 | posti | Finland | 2018 | 12.0 | 5.0 | 6.0 | 89.0 |
| 6 | kemira | Finland | 2017 | 8.0 | 5.0 | 3.0 | 85.0 |
| 7 | dna | Finland | 2018 | 11.0 | 4.0 | 6.0 | 83.0 |
| 8 | tokmanni | Finland | 2017 | 7.0 | 5.0 | 3.0 | 77.0 |
| 9 | aco | Denmark | 2018 | 2.0 | 4.0 | 3.0 | 74.0 |
| 10 | telenor_group | Norway | 2018 | 7.0 | 3.0 | 3.0 | 70.0 |

Stora of Finland has 127 disclosure matches, out of 166, which is relatively very high, with a distance of more than 35% to the report in tenth position. It can be noted that the top 10 positions have been virtually monopolized by the Finnish reports.

[11] The World Federation of Exchanges, formerly the Federation Internationale des Bourses de Valeurs, or International Federation of Stock Exchanges, is the trade association of publicly regulated stock, futures, and options exchanges, as well as central counterparties

By countries:

*Table 9. Top 10 index matching by countries*

Sweden:

| rank | name_company | year | total_E | total_S | total_G | total |
|---|---|---|---|---|---|---|
| 12 | epiroc | 2018 | 7.0 | 4.0 | 3.0 | 66.0 |
| 19 | advania | 2018 | 3.0 | 3.0 | 2.0 | 56.0 |
| 22 | okq8_scandinavia | 2018 | 7.0 | 3.0 | 3.0 | 55.0 |
| 23 | seb | 2018 | 2.0 | 4.0 | 3.0 | 52.0 |
| 24 | beckers_group | 2018 | 6.0 | 2.0 | 2.0 | 51.0 |
| 27 | transcom_worldwide | 2018 | 6.0 | 4.0 | 3.0 | 48.0 |
| 31 | whistleb_whistleblowing_center | 2017 | 1.0 | 3.0 | 3.0 | 41.0 |
| 32 | diab | 2017 | 5.0 | 0.0 | 2.0 | 39.0 |
| 35 | h&m_group | 2018 | 5.0 | 0.0 | 0.0 | 16.0 |
| 56 | foodtankers | 2015 | 0.0 | 1.0 | 0.0 | 6.0 |

Norway:

| rank | name_company | year | total_E | total_S | total_G | total |
|---|---|---|---|---|---|---|
| 10 | telenor_group | 2018 | 7.0 | 3.0 | 3.0 | 70.0 |
| 20 | rica_hotels_(scandic) | 2017 | 6.0 | 4.0 | 3.0 | 55.0 |
| 29 | avinor | 2018 | 3.0 | 3.0 | 2.0 | 44.0 |
| 30 | plantasjen | 2018 | 2.0 | 4.0 | 3.0 | 43.0 |
| 36 | agder_energi | 2018 | 1.0 | 0.0 | 0.0 | 16.0 |
| 38 | odfjell | 2018 | 1.0 | 0.0 | 0.0 | 16.0 |
| 49 | petroleum_geo-services | 2016 | 0.0 | 0.0 | 0.0 | 7.0 |
| 50 | borregaard | 2018 | 4.0 | 0.0 | 0.0 | 7.0 |
| 51 | klp | 2016 | 0.0 | 0.0 | 0.0 | 7.0 |
| 59 | nsb_group | 2018 | 0.0 | 0.0 | 0.0 | 6.0 |

Denmark:

| rank | name_company | year | total_E | total_S | total_G | total |
|---|---|---|---|---|---|---|
| 4 | palsgaard | 2018 | 11.0 | 6.0 | 3.0 | 91.0 |
| 9 | aco | 2018 | 2.0 | 4.0 | 3.0 | 74.0 |
| 33 | brdr_moller | 2016 | 0.0 | 1.0 | 2.0 | 31.0 |
| 37 | greentech_energy_systems | 2016 | 1.0 | 2.0 | 1.0 | 16.0 |
| 39 | novo_nordisk | 2018 | 1.0 | 0.0 | 0.0 | 11.0 |
| 41 | vestas_wind_systems | 2018 | 0.0 | 0.0 | 0.0 | 10.0 |
| 45 | royal_unibrew | 2017 | 0.0 | 1.0 | 0.0 | 9.0 |
| 52 | alm._brand | 2016 | 0.0 | 0.0 | 0.0 | 7.0 |
| 53 | bang_&_olufsen | 2016 | 0.0 | 1.0 | 0.0 | 7.0 |
| 54 | a.p_moller_maersk | 2016 | 0.0 | 0.0 | 0.0 | 7.0 |

Finland:

| rank | name_company | year | total_E | total_S | total_G | total |
|---|---|---|---|---|---|---|
| 1 | stora | 2018 | 12.0 | 6.0 | 6.0 | 127.0 |
| 2 | upm | 2018 | 12.0 | 5.0 | 6.0 | 115.0 |
| 3 | kesko | 2018 | 13.0 | 5.0 | 5.0 | 102.0 |
| 5 | posti | 2018 | 12.0 | 5.0 | 6.0 | 89.0 |
| 6 | kemira | 2017 | 8.0 | 5.0 | 3.0 | 85.0 |
| 7 | dna | 2018 | 11.0 | 4.0 | 6.0 | 83.0 |
| 8 | tokmanni | 2017 | 7.0 | 5.0 | 3.0 | 77.0 |
| 11 | nib | 2018 | 2.0 | 5.0 | 3.0 | 68.0 |
| 13 | yit | 2019 | 7.0 | 5.0 | 6.0 | 63.0 |
| 14 | fennovoima | 2018 | 0.0 | 2.0 | 2.0 | 60.0 |

Iceland:

| rank | name_company | year | total_E | total_S | total_G | total |
|---|---|---|---|---|---|---|
| 42 | isavia | 2017 | 0.0 | 0.0 | 0.0 | 10.0 |
| 121 | ossur | 2018 | 0.0 | 0.0 | 0.0 | 3.0 |

The Danish and Swedish reports are in the top 60, and Iceland, in the ratings for not applying the latest GRI standards.

Finally, we would like to combine the semantic similarity obtained by LSA with gloVE and the marching index, as these values belongs to a different ranges, we need to apply the standardization method in order to normalize them. The results are the following:

*Table 10. Top 10 companies with more cosine similarity and index matching*

| rank | name_company | country_x | year_x | total |
|---|---|---|---|---|
| 1 | stora | Finland | 2018 | 8.387491 |
| 2 | kesko | Finland | 2018 | 7.504839 |
| 3 | upm | Finland | 2018 | 7.047673 |
| 4 | posti | Finland | 2018 | 6.394485 |
| 5 | palsgaard | Denmark | 2018 | 6.057551 |
| 6 | aco | Denmark | 2018 | 5.937491 |
| 7 | tokmanni | Finland | 2017 | 5.112981 |
| 8 | sampo | Finland | 2018 | 4.949607 |
| 9 | epiroc | Sweden | 2018 | 4.861749 |
| 10 | nordea | Finland | 2018 | 4.758798 |

Although the Finnish company Stora, is not even in the top 10 best reports according to their semantic affinity, their excellent rating according to disclosures were addressed in their management, may be an indication why they are in the first position in the overall table.

# 6 CONCLUSIONS

The objective of this research was to discover how can help us the implementation of text mining technics, if we would like to know how the reports published by Nordic companies are alienated to the GRI standards. Intrinsic valuation was implemented in chapter 5, to find out the degree to which these reports are in line with the latest version of Global Reporting Initiative guidelines. Different techniques were implemented in order to cover the different forms that exist for semantic evaluation. LSA and gloVE were the best models in terms of congruence.

*Regarding the quality of the data.* It has been evident that the creation of corpus or the training new models is not feasible for the volume of data, which we have. And although they can offer good results in the part of similarity by strings, by sentences, which is what interested us the most.

The enrichment of the text was discarded, as not to break the framework of the official guidelines provided by GRI.

*Regarding the models.* Despite the drawback of the amount of text to train an LSA model, it has proven to confirm it's popularity for handling small volumes of text well. Also it's docility when updating the training it is a more than feasible solution for this type of study. fastText, it was not very forceful when presenting the results in terms of clarity, word2vec pretrained, it was too slow compared to others. Doc2vec, is an interesting model, but not enough robust for our problem. gloVE proved to be very robust and consistent with it's results.

*Regarding the results.* The reports that have obtained a higher semantic similarity rating may not necessarily obtain a good index-matching rating. The reasons may be of a different nature, starting from the extraction of the text, which is often not 100% reliable when the text is embedded in images. Another cause may be that the reports were simply not up to date with the new standards, or simply omitted the GRI index in their reports. Also, another reason is the size of the text in the reports, sometimes it can help to get a better semantic assessment, but in the long run, if the document contains many generalist phrases it tends to penalize it's own assessment.

The combination of different methods of text mining certainly can provide us some insights about the degree of affinity that the Nordics Corporate Social Reports have with the latest Global Reporting Initiative Standards. But, the bias of the valuation process is a non-systematic risk to be considered. Therefore, the results of this work is not a guide about the actions of companies in matters of Environment, Social and Governance, for the points outlined above, but it does give some guidance on how information on their achievements should be presented. A clear, concise text and without many textual or media decorations, will enjoy a greater probability of positive evaluation, independent of which or how many CSR Frameworks they are using.

## 6.1 Future Work

Fortunately, text mining is a wide field where several actions can be taken in order to improve the accuracy of these results. For example, wold be necessary extend the valuation process to experts and non-experts to reduce the bias criteria. Also, we would like to incorporate more information about the different available guidelines, to enrich a corpus, and make it more specific not only in Environmental Social and Governance objectives but also in Sustainable Development Goals (SDGs). In addition to incorporating Sustainability Accounting Standards Board (SASB) indicators, which will help us to have an accurate view of the size of the *Social Capital* of a company.

**Bibliography**


Aryal, Nabin (2014), Comparative Study of CSR reporting in Finnish and UK listed Companies, University of Arcada.

A. Wilson and P. Rayson (1993), Automatic content analysis of spoken discourse: a report on work in progress, Corpus based computational linguistics, pp. 215-226, 1993.

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). A brief survey of text mining: Classification, clustering and extraction techniques.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. Foundations and trends in information retrieval 2, 1-2 (2008), 1–135.

Benites-Lazaro, L. L., Giatti, L., & Giarolla, A. (2018). Sustainability and governance of sugarcane ethanol companies in Brazil: Topic modeling analysis of CSR reporting. Journal of Cleaner Production, 197, 583-591.

Bjørn A, Bey N, Georg S, Røpke I, Hauschild MZ (2016). Is Earth recognized as a finite system in corporate responsibility reporting? Journal of Cleaner Production.

Christopher D Manning, Hinrich Schütze, et al.1999. Foundations of statistical natural language processing. Vol. 999. MIT Press.

Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. Introduction to information retrieval. Vol. 1. Cambridge university press Cambridge.

Chae, B., & Park, E. (2018). Corporate Social Responsibility (CSR): A Survey of Topics and Trends Using Twitter Data and Topic Modeling. Sustainability, 10(7), 2231.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. The Journal of machine Learning research 3 (2003), 993–1022.

de la Torre, Marıa del Consuelo Justicia (2005). New text mining techniques: Applications. AWord at a Time: Computing Word Relatedness using Temporal Semantic Analysis.

Domingos Pablo, (2012). A few useful things to know about machine learning. Communications of the ACM, 2012.

Dragomir R Radev, Eduard Hovy, and Kathleen McKeown. 2002. Introduction to the special issue on summarization. Computational linguistics 28, 4 (2002), 399–408.

Dwi Prasetya, Aji Prasetya Wibawa, Tsukasa Hirashima (2018), “The performance of text similarity algorithms”, International Journal of Advances in Intelligent Informatics Vol. 4, No. 1, March 2018, pp. 63-69.

Fabrizio Sebastiani. 2002. Machine learning in automated text categorization. ACM computing surveys (CSUR) 34, 1 (2002), 1–47.

Feltham G, Ohlson J (1995), Valuation and clean surplus accounting for operating and financial activities, Contemp Acc Res 11(2):689–731.

Freundlieb M, Teuteberg F (2013). Corporate social responsibility reporting-a transnational analysis of online corporate social responsibility reports by market–listed companies: contents and their evolution. International Journal of Innovation and Sustainable Development 7: 1–26.

Global Reporting Initiative (2013). G4 Sustainability Report Guidelines –Reporting Principles and Standard Disclosures. Amsterdam: GRI.

Global Reporting Initiative (2016). First Global Sustainability Reporting Standards Set to Transform Business. Amsterdam: GRI.

Griffin, J.J., Mahon, J.F. (1997), The corporate social performance and corporate financial performance debate: Twenty-five years of incomparable research, Business and Society, 36 (1), pp. 5-31. Cited 890 times.

González, M., del Mar Alonso-Almeida, M., Avila, C., & Dominguez, D. (2015). Modeling sustainability report scoring sequences using an attractor network. Neurocomputing, 168, 1181-1187.

Gao Jianfeng, Yih, Wen-tau & He, Xiaodong &,. (2015). Deep Learning and Continuous Representations for Natural Language Processing. 6-8. 10.3115/v1/N15-4004.

Guthrie, J., & Abeysekera, I. (2006). Content analysis of social, environmental reporting: What is new? Journal of Human Resource Costing & Accounting, 10(2), 114-126.

Graves and J. Schmidhuber (2005). Framewise Phoneme Classification with Bidirectional LSTM Networks. In Proceedings of the 2005 International Joint Conference on Neural Networks, 20

Gers, F.A., Schmidhuber, J. (2000) Recurrent nets that time and count. In: Neural Networks: Como, vol 3, pp. 189-194.

Hongzhao Huang, Zhen Wen, Dian Yu, Heng Ji,Yizhou Sun, Jiawei Han, and He Li. 2013. Re-solving entity morphs in censored data. In Proceed-ings of the 51st Annual Meeting of the Associationfor Computational Linguistics (Volume 1: Long Pa-pers), pages 1083–1093. Association for Computa-tional Linguistics.

Isaksson, R. Steimle, U. (2009). What does GRI reporting tell us about corporate sustainability? The TQM Journal 21 (2): 168-181.

Jim Cowie and Wendy Lehnert. 1996. Information extraction. Commun. ACM 39, 1 (1996), 80–91.

Jung, S., Nam, C., Yang, D. H., & Kim, S. (2017). Does corporate sustainability performance increase corporate financial performance? Focusing on the information and communication technology industry in Korea. Sustainable Development.

Kiros, Ryan & Zhu, Yukun & Salakhutdinov, Ruslan & Zemel, Richard & Torralba, Antonio & Urtasun, Raquel & Fidler, Sanja. (2015). Skip-Thought Vectors. Advances in Neural Information Processing Systems. 28.

Knebel, S. –Seele, P. (2015). A (critical) assessment of GRI G3.1 A+ non-financial reports and implications for credibility and standardisation. Corporate Communications: An international Journal 20(2): 196-212.

Kolk A (2004). A decade of sustainability reporting: developments and significance. International Journal of Environment and Sustainable Development 3: 51–64.

Kolk A (2003). Trends in sustainability reporting by the Fortune Global 250. Business Strategy and the Environment 12: 279–291.

KPMG (2017), The KPMG Survey of Corporate Responsibility Reporting 2017.

Knebel, Sebastian & Seele, Peter. (2015). Quo vadis GRI? A (critical) assessment of GRI 3.1 A+ non-financial reports and implications for credibility and standardization. Corporate Communications: An International Journal. 20. 10.1108/CCIJ-11-2013-0101.

Liew WT, Adhitya A, Srinivasan R (2014). Sustainability trends in the process industries: A text mining-based analysis. Computers in Industry 65: 393–400.

Lozano, R. – Huisingh, D. (2011). Inter- linking issues and dimensions in sustainability reporting. Journal of Cleaner Production 19(2–3): 99-107.

Liu, S. H., Chen, S. Y., & Li, S. T. (2017, July). Text-Mining Application on CSR Report Analytics: A Study of Petrochemical Industry. In Advanced Applied Informatics (IIAI-AAI), 2017 6th IIAI International Congress on (pp. 76-81). IEEE.

Miralles-Quiros JL, Arraiano IG (2017). Are firms that contribute to sustainable development valued by investors? Corporate Social Responsibility and Environmental Management 24: 71–84.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. & Dean, J. (2013b). Distributed Representations of Words and Phrases and their Compositionality. CoRR, abs/1310.4546. Recuperado desde http://arxiv.org/abs/1310.4546.

Mikolov, T., P Bojanowski, E Grave (2017). Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics 5, 135-146

Mitsuzuka, K., Ling, F., Ohwada, H.(2017), Analysis of CSR activities Affecting Corporate Value Using Machine Learning, ACM International Conference Proceeding Series, Part F128357, pp. 11-14.

Modapothala, J. R., & Issac, B. (2014). Analysis of corporate environmental reports using statistical techniques and data mining. arXiv preprint arXiv:1410.4182.

Modapothala, J. R., & Issac, B. (2009). Study of economic, environmental and social factors in sustainability reports using text mining and Bayesian analysis. In Industrial Electronics & Applications, 2009. ISIEA 2009. IEEE Symposium on (Vol. 1, pp. 209-214). IEEE.

Modapothala, J. R., Issac, B., & Jayamani, E. (2010). Appraising the corporate sustainability reports–text mining and multi-discriminatory analysis. In Innovations in Computing Sciences and Software Engineering (pp. 489-494). Springer, Dordrecht.

Pagliardini, Matteo & Gupta, Prakhar & Jaggi, Martin. (2017). Unsupervised Learning of Sentence Embeddings using Compositional n-Gram Features.

Rehurek Radim and Petr Sojka (2010). Software framework for topic modelling with large corpora. - In Proceedings of the LREC 2010 Workshop on New CHALLENGES FOR NLP FRAMEWORKS.

Ronen Feldman and Ido Dagan.(1995). Knowledge Discovery in Textual Databases (KDT). In KDD, Vol. 95. 112–117.

R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, Natural language processing (almost) from scratch. Journal of Machine Learning Research, vol. 12, no. Aug, pp. 2493–2537, 2011.

Servaes, H., and Tamayo, A. (2017), 'The role of social capital in corporations: a review', Oxford Review of Economic Policy, Volume 33, Number 2, 2017, pp. 201–220.

Székely, N., & vom Brocke, J. (2017). What can we learn from corporate sustainability reporting? Deriving propositions for research and practice from over 9,500 corporate sustainability reports published between 1999 and 2015 using topic modelling technique. PloS one, 12(4), e0174807.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the 2013 conference on empirical methods in natural language processing (pp. 1631-1642).

Shahi, A.M., Issac, B., Modapothala, J.R. (2015), Reliability assessment of an intelligent approach to corporate sustainability report analysis. Lecture Notes in Electrical Engineering, 313, pp. 233-240.

S Dumais, G Furnas, T Landauer, S Deerwester, S Deerwester, et al 1995. Latent semantic indexing. In Proceedings of the Text Retrieval Conference.

Tremblay, M. C., Parra, C., & Castellanos, A. (2015). Analyzing Corporate Social Responsibility Reports Using Unsupervised and Supervised Text Data Mining. In International Conference on Design Science Research in Information Systems (pp. 439-446). Springer, Cham.

Thomas Hofmann. 1999. Probabilistic latent semantic indexing. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 50–57.

Tom M Mitchell. 1997. Machine learning. 1997. Burr Ridge, IL: McGraw Hill 45 (1997).

Yamamoto, Y., Miyamoto, D., & Nakayama, M. (2015, November). Text-Mining Approach for Estimating Vulnerability Score. In Building Analysis Datasets and Gathering Experience Returns for Security (BADGERS), 2015 4th International Workshop on (pp. 67-73). IEEE.

Young, T., Hazarika, D., Poria, S., & Cambria, E. (2018). Recent trends in deep learning based natural language processing. ieee Computational intelligenCe magazine, 13(3), 55-75.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, Christian Jauvin (2003). A New probabilistic Language Model. Journal of Machine Learning Research 3 (2003) 1137–115

Wael H. Gomaa and Aly Fhamy (2013), "A Survey of Text Similarity", Approaches, International Journal of Computer Applications (0975 – 8887) Volume 68– No.13, April 2013.

Wang, Q., Dou, J., Jia, S. (2016), A Meta-Analytic Review of Corporate Social Responsibility and Corporate Financial Performance: The Moderating Effect of Contextual Factors, Business and Society, 55 (8), pp. 1083-1121.

| APPENDIX A<br>GRI Standards | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| # | Required for CORE | GRI Standard Number | GRI Standard Title | Publication Year | Disclosure Number | Disclosure Title<br>Individual disclosure items ('a', 'b', 'c', etc.) are not listed here | Page Number | 2016 and 2018 Updates<br>See key above |
| 15 | Core | **GRI 102** | **General Disclosures** | **2016** | **102-14** | Statement from senior decision-maker | p. 14 | Revised disclosure |
| 16 | | **GRI 102** | **General Disclosures** | **2016** | **102-15** | Key impacts, risks, and opportunities | p. 15 | Revised disclosure |
| 1 | Core | **GRI 102** | **General Disclosures** | **2016** | **102-1** | Name of the organization | p. 7 | No revision |
| 2 | Core | **GRI 102** | **General Disclosures** | **2016** | **102-2** | Activities, brands, products, and services | p. 7 | Revised disclosure |
| 4 | Core | **GRI 102** | **General Disclosures** | **2016** | **102-3** | Location of headquarters | p. 8 | No revision |
| 5 | Core | **GRI 102** | **General Disclosures** | **2016** | **102-4** | Location of operations | p. 8 | No revision |
| 6 | Core | **GRI 102** | **General Disclosures** | **2016** | **102-5** | Ownership and legal form | p. 8 | No revision |
| 7 | Core | **GRI 102** | **General Disclosures** | **2016** | **102-6** | Markets served | p. 8 | Minor clarification |
| 8 | Core | **GRI 102** | **General Disclosures** | **2016** | **102-7** | Scale of the organization | p. 9 | No revision |
| 9 | Core | **GRI 102** | **General Disclosures** | **2016** | **102-8** | Information on employees and other workers | p. 10 | Revised disclosure |
| 43 | Core | **GRI 102** | **General Disclosures** | **2016** | **102-41** | Collective bargaining agreements | p. 30 | No revision |
| 10 | Core | **GRI 102** | **General Disclosures** | **2016** | **102-9** | Supply chain | p. 11 | Revised disclosure |
| 11 | Core | **GRI 102** | **General Disclosures** | **2016** | **102-10** | Significant changes to the organization and its supply chain | p. 12 | No revision |
| 12 | Core | **GRI 102** | **General Disclosures** | **2016** | **102-11** | Precautionary Principle or approach | p. 12 | No revision |
| 13 | Core | **GRI 102** | **General Disclosures** | **2016** | **102-12** | External initiatives | p. 13 | No revision |
| 14 | Core | **GRI 102** | **General Disclosures** | **2016** | **102-13** | Membership of associations | p. 13 | Revised disclosure |
| 48 | Core | **GRI 102** | **General Disclosures** | **2016** | **102-45** | Entities included in the consolidated financial statements | p. 33 | No revision |
| 49 | Core | **GRI 102** | **General Disclosures** | **2016** | **102-46** | Defining report content and topic Boundaries | p. 34 | No revision |
| 50 | Core | **GRI 102** | **General Disclosures** | **2016** | **102-47** | List of material topics | p. 35 | Revised disclosure |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 62 | Core | **GRI 103** | **Management Approach** | **2016** | **103-1** | Explanation of the material topic and its Boundary | p. 6-7 | Revised disclosure |
| 63 | Core | **GRI 103** | **Management Approach** | **2016** | **103-1** | Explanation of the material topic and its Boundary | pp. 6-7 | Revised disclosure |
| 51 | Core | **GRI 102** | **General Disclosures** | **2016** | **102-48** | Restatements of information | p. 35 | No revision |
| 52 | Core | **GRI 102** | **General Disclosures** | **2016** | **102-49** | Changes in reporting | p. 36 | Minor clarification |
| 42 | Core | **GRI 102** | **General Disclosures** | **2016** | **102-40** | List of stakeholder groups | p. 29 | No revision |
| 44 | Core | **GRI 102** | **General Disclosures** | **2016** | **102-42** | Identifying and selecting stakeholders | p. 31 | No revision |
| 45 | Core | **GRI 102** | **General Disclosures** | **2016** | **102-43** | Approach to stakeholder engagement | p. 31 | No revision |
| 47 | Core | **GRI 102** | **General Disclosures** | **2016** | **102-44** | Key topics and concerns raised | p. 32 | No revision |
| 53 | Core | **GRI 102** | **General Disclosures** | **2016** | **102-50** | Reporting period | p. 36 | No revision |
| 54 | Core | **GRI 102** | **General Disclosures** | **2016** | **102-51** | Date of most recent report | p. 36 | No revision |
| 55 | Core | **GRI 102** | **General Disclosures** | **2016** | **102-52** | Reporting cycle | p. 37 | No revision |
| 56 | Core | **GRI 102** | **General Disclosures** | **2016** | **102-53** | Contact point for questions regarding the report | p. 37 | No revision |
| 57 | Core | **GRI 102** | **General Disclosures** | **2016** | **102-54** | Claims of reporting in accordance with the GRI Standards | p. 37 | Revised disclosure |
| 58 | Core | **GRI 102** | **General Disclosures** | **2016** | **102-55** | GRI content index | pp. 38-39 | Revised disclosure |
| 59 | Core | **GRI 102** | **General Disclosures** | **2016** | **102-56** | External assurance | pp. 41-42 | Revised disclosure |
| 60 | Core | **GRI 102** | **General Disclosures** | **2016** | **102-56** | External assurance | pp. 41-42 | Revised disclosure |
| 20 | Core | **GRI 102** | **General Disclosures** | **2016** | **102-18** | Governance structure | p. 18 | Minor clarification |
| 21 | | **GRI 102** | **General Disclosures** | **2016** | **102-19** | Delegating authority | p. 18 | No revision |
| 22 | | **GRI 102** | **General Disclosures** | **2016** | **102-20** | Executive-level responsibility for economic, environmental, and social topics | p. 19 | No revision |
| 23 | | **GRI 102** | **General Disclosures** | **2016** | **102-21** | Consulting stakeholders on economic, environmental, and social topics | p. 19 | No revision |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 24 | | **GRI 102** | **General Disclosures** | **2016** | **102-22** | Composition of the highest governance body and its committees | p. 19 | Minor clarification |
| 25 | | **GRI 102** | **General Disclosures** | **2016** | **102-23** | Chair of the highest governance body | p. 20 | No revision |
| 26 | | **GRI 102** | **General Disclosures** | **2016** | **102-24** | Nominating and selecting the highest governance body | p. 20 | No revision |
| 27 | | **GRI 102** | **General Disclosures** | **2016** | **102-25** | Conflicts of interest | p. 21 | No revision |
| 28 | | **GRI 102** | **General Disclosures** | **2016** | **102-26** | Role of highest governance body in setting purpose, values, and strategy | p. 21 | Minor clarification |
| 29 | | **GRI 102** | **General Disclosures** | **2016** | **102-27** | Collective knowledge of highest governance body | p. 21 | No revision |
| 30 | | **GRI 102** | **General Disclosures** | **2016** | **102-28** | Evaluating the highest governance body's performance | p. 22 | No revision |
| 31 | | **GRI 102** | **General Disclosures** | **2016** | **102-29** | Identifying and managing economic, environmental, and social impacts | p. 22 | Minor clarification |
| 32 | | **GRI 102** | **General Disclosures** | **2016** | **102-30** | Effectiveness of risk management processes | p. 22 | No revision |
| 33 | | **GRI 102** | **General Disclosures** | **2016** | **102-31** | Review of economic, environmental, and social topics | p. 23 | Minor clarification |
| 34 | | **GRI 102** | **General Disclosures** | **2016** | **102-32** | Highest governance body's role in sustainability reporting | p. 23 | No revision |
| 35 | | **GRI 102** | **General Disclosures** | **2016** | **102-33** | Communicating critical concerns | p. 23 | No revision |
| 36 | | **GRI 102** | **General Disclosures** | **2016** | **102-34** | Nature and total number of critical concerns | p. 24 | No revision |
| 37 | | **GRI 102** | **General Disclosures** | **2016** | **102-35** | Remuneration policies | p. 25 | Minor clarification |
| 38 | | **GRI 102** | **General Disclosures** | **2016** | **102-36** | Process for determining remuneration | p. 26 | No revision |
| 39 | | **GRI 102** | **General Disclosures** | **2016** | **102-37** | Stakeholders' involvement in remuneration | p. 26 | No revision |
| 40 | | **GRI 102** | **General Disclosures** | **2016** | **102-38** | Annual total compensation ratio | p. 27 | No revision |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 41 | | **GRI 102** | **General Disclosures** | **2016** | **102-39** | Percentage increase in annual total compensation ratio | p. 28 | No revision |
| 17 | Core | **GRI 102** | **General Disclosures** | **2016** | **102-16** | Values, principles, standards, and norms of behavior | p. 16 | No revision |
| 18 | | **GRI 102** | **General Disclosures** | **2016** | **102-17** | Mechanisms for advice and concerns about ethics | p. 17 | No revision |
| 19 | | **GRI 102** | **General Disclosures** | **2016** | **102-17** | Mechanisms for advice and concerns about ethics | p. 17 | No revision |
| 61 | Core | **GRI 103** | **Management Approach** | **2016** | **103-1** | Explanation of the material topic and its Boundary | pp. 6-7 | Revised disclosure |
| 64 | Core | **GRI 103** | **Management Approach** | **2016** | **103-2** | The management approach and its components | pp. 8-10 | Revised disclosure |
| 69 | Core | **GRI 103** | **Management Approach** | **2016** | **103-3** | Evaluation of the management approach | p. 11 | No revision |
| 70 | | **GRI 201** | **Economic Performance** | **2016** | **201-1** | Direct economic value generated and distributed | pp. 6-8 | No revision |
| 71 | | **GRI 201** | **Economic Performance** | **2016** | **201-2** | Financial implications and other risks and opportunities due to climate change | pp. 9-10 | No revision |
| 72 | | **GRI 201** | **Economic Performance** | **2016** | **201-3** | Defined benefit plan obligations and other retirement plans | p. 11 | No revision |
| 73 | | **GRI 201** | **Economic Performance** | **2016** | **201-4** | Financial assistance received from government | p. 12 | No revision |
| 74 | | **GRI 202** | **Market Presence** | **2016** | **202-1** | Ratios of standard entry level wage by gender compared to local minimum wage | pp. 6-7 | Revised disclosure |
| 75 | | **GRI 202** | **Market Presence** | **2016** | **202-2** | Proportion of senior management hired from the local community | p. 8 | No revision |
| 76 | | **GRI 203** | **Indirect Economic Impacts** | **2016** | **203-1** | Infrastructure investments and services supported | p. 6 | No revision |
| 77 | | **GRI 203** | **Indirect Economic Impacts** | **2016** | **203-2** | Significant indirect economic impacts | p. 7 | No revision |
| 78 | | **GRI 204** | **Procurement Practices** | **2016** | **204-1** | Proportion of spending on local suppliers | p. 7 | No revision |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 83 | | **GRI 301** | **Materials** | **2016** | **301-1** | Materials used by weight or volume | p. 6 | No revision |
| 84 | | **GRI 301** | **Materials** | **2016** | **301-2** | Recycled input materials used | p. 7 | No revision |
| 86 | | **GRI 302** | **Energy** | **2016** | **302-1** | Energy consumption within the organization | pp. 6-7 | Minor clarification |
| 87 | | **GRI 302** | **Energy** | **2016** | **302-2** | Energy consumption outside of the organization | pp. 8-9 | Minor clarification |
| 88 | | **GRI 302** | **Energy** | **2016** | **302-3** | Energy intensity | p. 10 | No revision |
| 89 | | **GRI 302** | **Energy** | **2016** | **302-4** | Reduction of energy consumption | p. 11 | Minor clarification |
| 90 | | **GRI 302** | **Energy** | **2016** | **302-5** | Reductions in energy requirements of products and services | p. 12 | Minor clarification |
| 91 | | - | - | **2016** | - | - | - | New disclosure available |
| 92 | | - | - | **2016** | - | - | - | New disclosure available |
| 93 | | - | - | **2016** | - | - | - | New disclosure available |
| 94 | | **GRI 303** | **Water and Efflu** | **2018** | **303-1** | Interactions with water as a shared resource | p. 6 | New disclosure |
| 95 | | **GRI 303** | **Water and Efflu** | **2018** | **303-2** | Management of water discharge-related impacts | p. 8 | New disclosure |
| 96 | | **GRI 303** | **Water and Efflu** | **2018** | **303-3** | Water withdrawal | p. 9 | New disclosure |
| 97 | | **GRI 303** | **Water and Efflu** | **2018** | **303-4** | Water discharge | p. 12 | New disclosure |
| 98 | | **GRI 303** | **Water and Efflu** | **2018** | **303-5** | Water consumption | p. 15 | New disclosure |
| 99 | | **GRI 304** | **Biodiversity** | **2016** | **304-1** | Operational sites owned, leased, managed in, or adjacent to, protected areas and areas of high biodiversity value outside protected areas | p. 7 | Minor clarification |
| 100 | | **GRI 304** | **Biodiversity** | **2016** | **304-2** | Significant impacts of activities, products, and services on biodiversity | p. 8 | No revision |
| 101 | | **GRI 304** | **Biodiversity** | **2016** | **304-3** | Habitats protected or restored | p. 9 | No revision |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 102 | | **GRI 304** | **Biodiversity** | **2016** | **304-4** | IUCN Red List species and national conservation list species with habitats in areas affected by operations | p. 10 | No revision |
| 103 | | **GRI 305** | **Emissions** | **2016** | **305-1** | Direct (Scope 1) GHG emissions | pp. 7-8 | Minor clarification |
| 104 | | **GRI 305** | **Emissions** | **2016** | **305-2** | Energy indirect (Scope 2) GHG emissions | pp. 9-10 | Revised disclosure |
| 105 | | **GRI 305** | **Emissions** | **2016** | **305-3** | Other indirect (Scope 3) GHG emissions | pp. 11-12 | Minor clarification |
| 106 | | **GRI 305** | **Emissions** | **2016** | **305-4** | GHG emissions intensity | p. 13 | No revision |
| 107 | | **GRI 305** | **Emissions** | **2016** | **305-5** | Reduction of GHG emissions | p. 14 | Minor clarification |
| 108 | | **GRI 305** | **Emissions** | **2016** | **305-6** | Emissions of ozone-depleting substances (ODS) | pp. 15-16 | Minor clarification |
| 109 | | **GRI 305** | **Emissions** | **2016** | **305-7** | Nitrogen oxides ($NO_X$), sulfur oxides ($SO_X$), and other significant air emissions | p. 17 | Minor clarification |
| 110 | | - | - | **2016** | - | - | - | New disclosure available |
| 111 | | **GRI 306** | **Effluents and Waste** | **2016** | **306-2** | Waste by type and disposal method | p. 7-8 | Revised disclosure |
| 112 | | **GRI 306** | **Effluents and Waste** | **2016** | **306-3** | Significant spills | p. 9 | No revision |
| 113 | | **GRI 306** | **Effluents and Waste** | **2016** | **306-4** | Transport of hazardous waste | p. 10 | Revised disclosure |
| 114 | | - | - | **2016** | - | - | - | New disclosure available |
| 167 | | NA | NA | **2016** | NA | NA | NA | Discontinued |
| 85 | | **GRI 301** | **Materials** | **2016** | **301-3** | Reclaimed products and their packaging materials | p. 8 | No revision |
| 115 | | **GRI 307** | **Environmental Compliance** | **2016** | **307-1** | Non-compliance with environmental laws and regulations | p. 6 | Minor clarification |
| 168 | | NA | NA | **2016** | NA | NA | NA | Discontinued |
| 169 | | Several | Several | **2016** | NA | NA | NA | Revised disclosure |
| 116 | | **GRI 308** | **Supplier Environmental Assessment** | **2016** | **308-1** | New suppliers that were screened using environmental criteria | p. 7 | No revision |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 117 | | **GRI 308** | **Supplier Environmental Assessment** | **2016** | **308-2** | Negative environmental impacts in the supply chain and actions taken | p. 8 | No revision |
| 65 | Core | **GRI 103** | **Management Approach** | **2016** | **103-2** | The management approach and its components | pp. 8-10 | Revised disclosure |
| 118 | | **GRI 401** | **Employment** | **2016** | **401-1** | New employee hires and employee turnover [This Standard includes a Standard Interpretation on how to calculate the rates of new employee hires and employee turnover.] | p. 7 | No revision |
| 119 | | **GRI 401** | **Employment** | **2016** | **401-2** | Benefits provided to full-time employees that are not provided to temporary or part-time employees | p. 8 | No revision |
| 120 | | **GRI 401** | **Employment** | **2016** | **401-3** | Parental leave | p. 9 | No revision |
| 121 | | **GRI 402** | **Labor/Management Relations** | **2016** | **402-1** | Minimum notice periods regarding operational changes | p. 6 | No revision |
| 122 | | - | - | **2016** | - | - | - | New disclosure available |
| 123 | | - | - | **2016** | - | - | - | New disclosure available |
| 124 | | - | - | **2016** | - | - | - | New disclosure available |
| 125 | | - | - | **2016** | - | - | - | New disclosure available |
| 126 | | **GRI 403** | **Occupational H** | **2018** | **403-1** | Occupational health and safety management system | p. 9 | New disclosure |
| 127 | | **GRI 403** | **Occupational H** | **2018** | **403-2** | Hazard identification, risk assessment, and incident investigation | p. 10 | New disclosure |
| 128 | | **GRI 403** | **Occupational H** | **2018** | **403-3** | Occupational health services | p. 11 | New disclosure |
| 129 | | **GRI 403** | **Occupational H** | **2018** | **403-4** | Worker participation, consultation, and communication on occupational health and safety | p. 12 | New disclosure |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 130 | | **GRI 403** | **Occupational H** | **2018** | **403-5** | Worker training on occupational health and safety | p. 13 | New disclosure |
| 131 | | **GRI 403** | **Occupational H** | **2018** | **403-6** | Promotion of worker health | p. 14 | New disclosure |
| 132 | | **GRI 403** | **Occupational H** | **2018** | **403-7** | Prevention and mitigation of occupational health and safety impacts directly linked by business relationships | p. 16 | New disclosure |
| 133 | | **GRI 403** | **Occupational H** | **2018** | **403-8** | Workers covered by an occupational health and safety management system | p. 17 | New disclosure |
| 134 | | **GRI 403** | **Occupational H** | **2018** | **403-9** | Work-related injuries | p. 19 | New disclosure |
| 135 | | **GRI 403** | **Occupational H** | **2018** | **403-10** | Work-related ill health | p. 23 | New disclosure |
| 136 | | **GRI 404** | **Training and Education** | **2016** | **404-1** | Average hours of training per year per employee | pp. 6-7 | No revision |
| 137 | | **GRI 404** | **Training and Education** | **2016** | **404-2** | Programs for upgrading employee skills and transition assistance programs | p. 8 | No revision |
| 138 | | **GRI 404** | **Training and Education** | **2016** | **404-3** | Percentage of employees receiving regular performance and career development reviews | p. 9 | No revision |
| 139 | | **GRI 405** | **Diversity and Equal Opportunity** | **2016** | **405-1** | Diversity of governance bodies and employees | p. 6 | Revised disclosure |
| 140 | | **GRI 405** | **Diversity and Equal Opportunity** | **2016** | **405-2** | Ratio of basic salary and remuneration of women to men | p. 7 | No revision |
| 152 | | **GRI 414** | **Supplier Social Assessment** | **2016** | **414-1** | New suppliers that were screened using social criteria | p. 7 | Revised disclosure |
| 155 | | **GRI 414** | **Supplier Social Assessment** | **2016** | **414-2** | Negative social impacts in the supply chain and actions taken | p. 8 | Revised disclosure |
| 66 | Core | **GRI 103** | **Management Approach** | **2016** | **103-2** | The management approach and its components | pp. 8-10 | Revised disclosure |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 149 | | **GRI 412** | **Human Rights Assessment** | **2016** | **412-3** | Significant investment agreements and contracts that include human rights clauses or that underwent human rights screening | p. 9 | No revision |
| 148 | | **GRI 412** | **Human Rights Assessment** | **2016** | **412-2** | Employee training on human rights policies or procedures | p. 8 | No revision |
| 141 | | **GRI 406** | **Non-discrimination** | **2016** | **406-1** | Incidents of discrimination and corrective actions taken | p. 6 | No revision |
| 142 | | **GRI 407** | **Freedom of Association and Collective Bargaining** | **2016** | **407-1** | Operations and suppliers in which the right to freedom of association and collective bargaining may be at risk | p. 6 | Revised disclosure |
| 143 | | **GRI 408** | **Child Labor** | **2016** | **408-1** | Operations and suppliers at significant risk for incidents of child labor | pp. 6-7 | No revision |
| 144 | | **GRI 409** | **Forced or Compulsory Labor** | **2016** | **409-1** | Operations and suppliers at significant risk for incidents of forced or compulsory labor | p. 6 | No revision |
| 145 | | **GRI 410** | **Security Practices** | **2016** | **410-1** | Security personnel trained in human rights policies or procedures | p. 6 | No revision |
| 146 | | **GRI 411** | **Rights of Indigenous Peoples** | **2016** | **411-1** | Incidents of violations involving rights of indigenous peoples | p. 7 | No revision |
| 147 | | **GRI 412** | **Human Rights Assessment** | **2016** | **412-1** | Operations that have been subject to human rights reviews or impact assessments | p. 7 | No revision |
| 153 | | **GRI 414** | **Supplier Social Assessment** | **2016** | **414-1** | New suppliers that were screened using social criteria | p. 7 | Revised disclosure |
| 156 | | **GRI 414** | **Supplier Social Assessment** | **2016** | **414-2** | Negative social impacts in the supply chain and actions taken | p. 8 | Revised disclosure |
| 67 | Core | **GRI 103** | **Management Approach** | **2016** | **103-2** | The management approach and its components | pp. 8-10 | Revised disclosure |

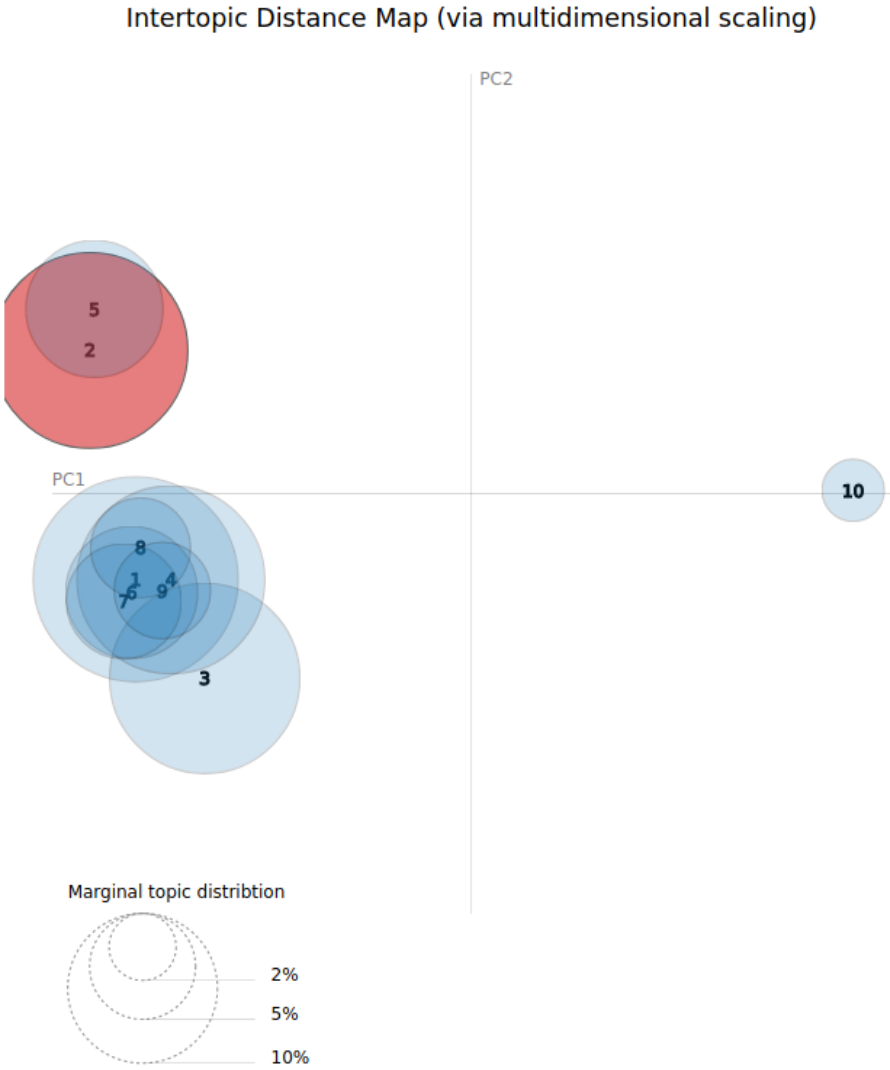
| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 150 | | **GRI 413** | **Local Communities** | **2016** | **413-1** | Operations with local community engagement, impact assessments, and development programs | pp. 7-8 | Revised disclosure |
| 151 | | **GRI 413** | **Local Communities** | **2016** | **413-2** | Operations with significant actual and potential negative impacts on local communities | pp. 9-10 | No revision |
| 79 | | **GRI 205** | **Anti-corruption** | **2016** | **205-1** | Operations assessed for risks related to corruption | p. 7 | No revision |
| 80 | | **GRI 205** | **Anti-corruption** | **2016** | **205-2** | Communication and training about anti-corruption policies and procedures | p. 8 | Revised disclosure |
| 81 | | **GRI 205** | **Anti-corruption** | **2016** | **205-3** | Confirmed incidents of corruption and actions taken | p. 9 | No revision |
| 158 | | **GRI 415** | **Public Policy** | **2016** | **415-1** | Political contributions | p. 6 | No revision |
| 82 | | **GRI 206** | **Anti-competitive Behavior** | **2016** | **206-1** | Legal actions for anti-competitive behavior, anti-trust, and monopoly practices | p. 6 | No revision |
| 165 | | **GRI 419** | **Socioeconomic Compliance** | **2016** | **419-1** | Non-compliance with laws and regulations in the social and economic area | p. 6 | Revised disclosure |
| 154 | | **GRI 414** | **Supplier Social Assessment** | **2016** | **414-1** | New suppliers that were screened using social criteria | p. 7 | Revised disclosure |
| 157 | | **GRI 414** | **Supplier Social Assessment** | **2016** | **414-2** | Negative social impacts in the supply chain and actions taken | p. 8 | Revised disclosure |
| 68 | Core | **GRI 103** | **Management Approach** | **2016** | **103-2** | The management approach and its components | pp. 8-10 | Revised disclosure |
| 159 | | **GRI 416** | **Customer Health and Safety** | **2016** | **416-1** | Assessment of the health and safety impacts of product and service categories | p. 7 | No revision |
| 160 | | **GRI 416** | **Customer Health and Safety** | **2016** | **416-2** | Incidents of non-compliance concerning the health and safety impacts of products and services | p. 8 | Minor clarification |

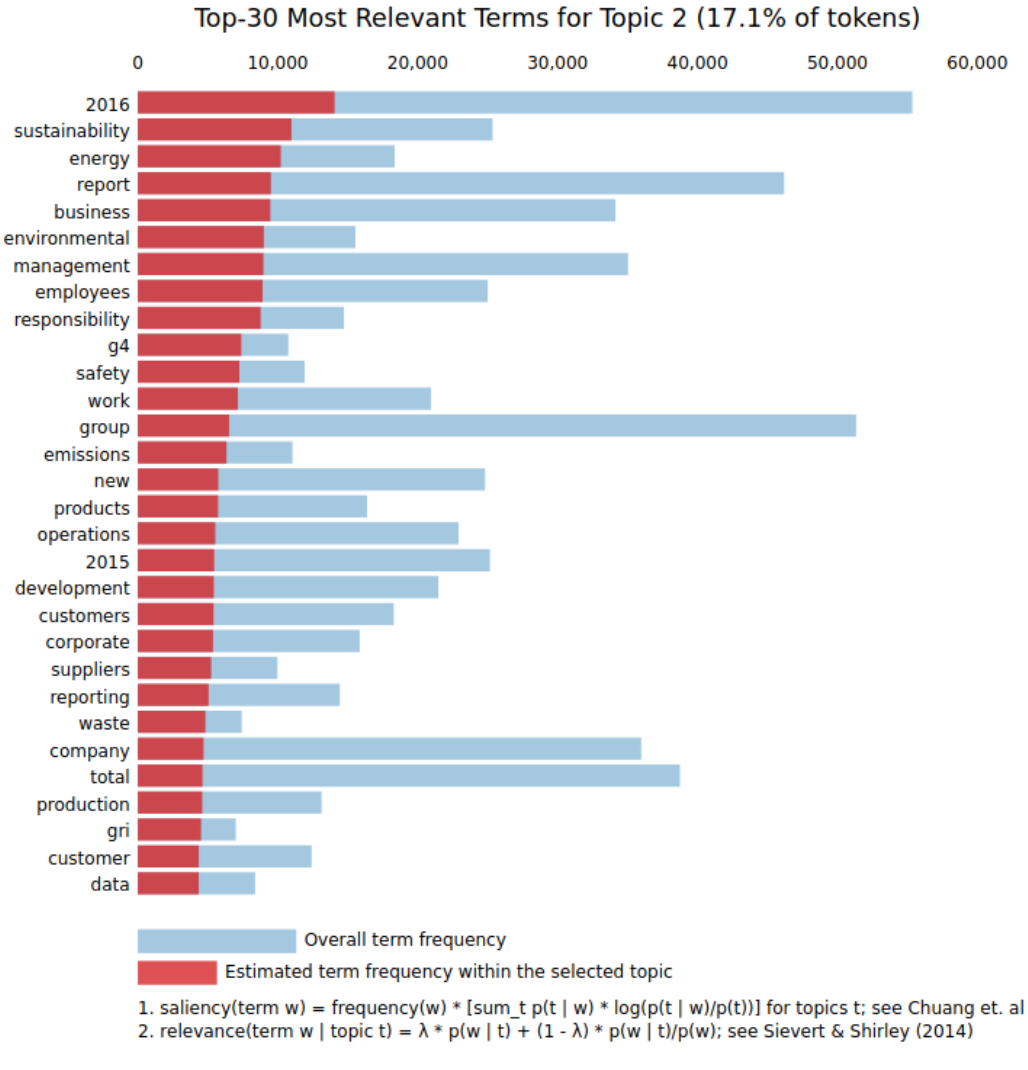
| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 161 | | **GRI 417** | **Marketing and Labeling** | **2016** | **417-1** | Requirements for product and service information and labeling | p. 6 | No revision |
| 162 | | **GRI 417** | **Marketing and Labeling** | **2016** | **417-2** | Incidents of non-compliance concerning product and service information and labeling | p. 7 | Minor clarification |
| 46 | Core | **GRI 102** | **General Disclosures** | **2016** | **102-43**<br>**102-44** | Approach to stakeholder engagement<br>Key topics and concerns raised | pp. 31-32 | Revised disclosure |
| 3 | Core | **GRI 102** | **General Disclosures** | **2016** | **102-2** | Activities, brands, products, and services | p. 7 | Revised disclosure |
| 163 | | **GRI 417** | **Marketing and Labeling** | **2016** | **417-3** | Incidents of non-compliance concerning marketing communications | p. 8 | Minor clarification |
| 164 | | **GRI 418** | **Customer Privacy** | **2016** | **418-1** | Substantiated complaints concerning breaches of customer privacy and losses of customer data | p. 6 | No revision |
| 166 | | **GRI 419** | **Socioeconomic Compliance** | **2016** | **419-1** | Non-compliance with laws and regulations in the social and economic area | p. 6 | Revised disclosure |

# Appendix B

**Topics LDA Model - Skatfraft**

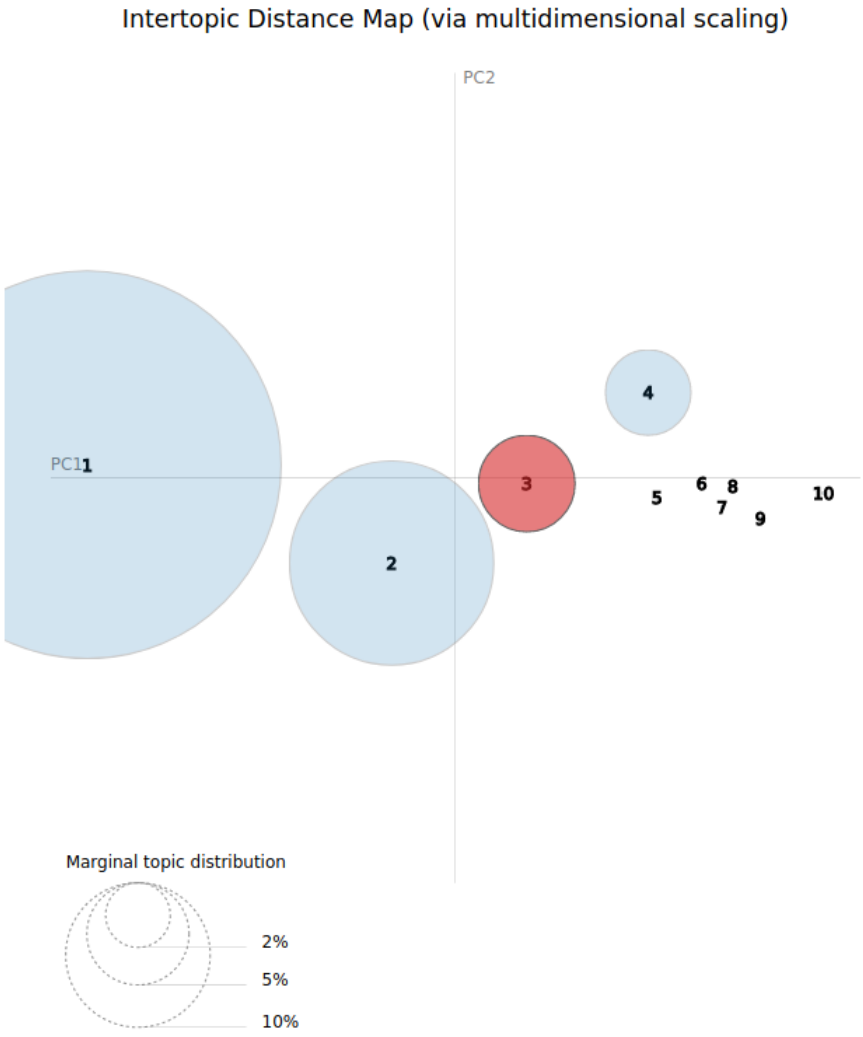
Topic #0:
2016, report, annual, financial, eur, board, management, company, 2015, services, year, million, operations, finland, total, group, directors, personnel, risk, wrtsil
Topic #1:
2018, 2017, sustainability, employees, work, group, report, new, sustainable, global, development, business, people, environment, year, products, health, fish, use, management
Topic #2:
2016, financial, annual, report, group, eur, board, business, company, value, statements, 2015, year, assets, total, million, management, finland, upm, shares
Topic #3:
2016, sustainability, energy, report, business, environmental, management, employees, responsibility, g4, safety, work, group, emissions, new, products, operations, 2015, development, customers
Topic #4:
2017, 2018, financial, group, assets, board, cash, year, annual, note, report, tax, total, net, income, company, business, sales, value, share
Topic #5:
2018, risk, insurance, financial, group, customers, capital, value, credit, report, total, assets, annual, pension, income, market, bank, management, board, 2017
Topic #6:
og, er, af, til, en, av, med, som, 44, ss, aa, es, och, har, det, den, 2015, fr, a4, ee
Topic #7:
financial, group, assets, 2016, value, total, income, 31, liabilities, cash, company, net, report, board, risk, tax, shares, annual, fair, note
Topic #8:
2016, financial, 2015, group, nok, board, million, value, eur, power, management, 2014, statements, total, finnair, annual, report, op, company, energy
Topic #9:
sek, year, value, financial, company, ab, board, 2017, total, annual, property, 2016, report, properties, sustainability, development, income, group, swedish, 000

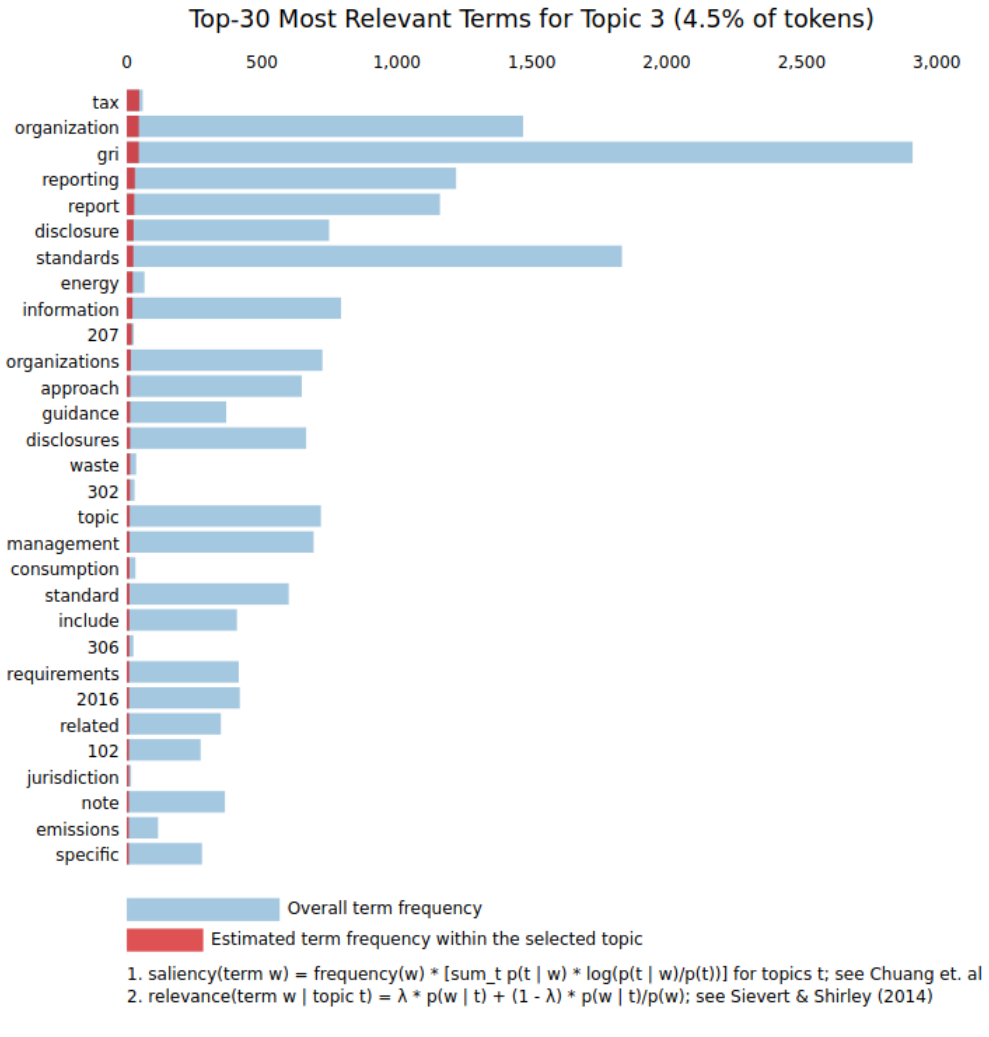

Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
5
2
10
8
1
4
6
9
7
3
Marginal topic distribtion
2%
5%
10%


Top-30 Most Relevant Terms for Topic 2 (17.1% of tokens)
0
10,000
20,000
30,000
40,000
50,000
60,000
2016
sustainability
energy
report
business
environmental
management
employees
responsibility
g4
safety
work
group
emissions
new
products
operations
2015
development
customers
corporate
suppliers
reporting
waste
company
total
production
gri
customer
data
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

**Topics LDA Model - GRI Standards**

Topics in LDA model:
Topic #0:
gri, organization, standards, reporting, report, disclosure, information, standard, 102, disclosures, used, impacts, topic, using, management, organizations, approach, use, 2016, include
Topic #1:
water, 303, organization, gri, ml, discharge, total, reporting, standards, report, information, impacts, stress, effluents, disclosure, organizations, related, areas, quality, 2018
Topic #2:
gri, standards, organization, report, disclosure, reporting, management, information, approach, organizations, health, workers, standard, topic, rights, disclosures, related, used, impacts, requirements
Topic #3:
gri, organization, standards, reporting, rights, disclosure, organizations, report, human, emissions, management, information, standard, approach, ghg, health, used, requirements, note, include
Topic #4:
tax, organization, gri, reporting, report, disclosure, standards, energy, information, 207, organizations, approach, guidance, disclosures, waste, 302, topic, management, consumption, standard
Topic #5:
organization, gri, reporting, disclosure, standards, report, health, 102, work, information, workers, organizations, emissions, related, used, occupational, management, ghg, disclosures, safety
Topic #6:
gri, standards, organization, report, reporting, information, topic, organizations, management, disclosures, approach, impacts, standard, disclosure, used, material, use, topics, 2016, sustainability
Topic #7:
gri, reporting, organization, standards, disclosure, report, information, water, impacts, 102, approach, economic, significant, requirements, governance, organizations, management, disclosures, used, labor
Topic #8:
gri, organization, report, standards, reporting, topic, impacts, organizations, disclosures, approach, standard, use, information, management, used, disclosure, material, topics, sustainability, note
Topic #9:gri, organization, reporting, standards, information, report, emissions, disclosure, organizations, disclosures, requirements, approach, standard, work, topic, used, impacts, ghg, health, management


Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
4
5
6
7
8
9
10
Marginal topic distribution
2%
5%
10%


Top-30 Most Relevant Terms for Topic 3 (4.5% of tokens)
0
500
1,000
1,500
2,000
2,500
3,000
tax
organization
gri
reporting
report
disclosure
standards
energy
information
207
organizations
approach
guidance
disclosures
waste
302
topic
management
consumption
standard
include
306
requirements
2016
related
102
jurisdiction
note
emissions
specific
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)