

6-2017

Exploratory Content Analysis Using Text Data Mining: Corporate Citizenship Reports of Seven US Companies from 2004 to 2012


Carlos M. Parra
Florida International University

Monica Tremblay
William & Mary, monica.tremblay@mason.wm.edu

Karen Paul
Florida International University

Arturo Castellanos
CUNY Bernard M Baruch College

Follow this and additional works at: <https://scholarworks.wm.edu/businesspubs>

 Part of the [Business Administration, Management, and Operations Commons](#), [Business Law, Public Responsibility, and Ethics Commons](#), [Organizational Behavior and Theory Commons](#), and the [Technology and Innovation Commons](#)

Recommended Citation

Parra, Carlos M.; Tremblay, Monica; Paul, Karen; and Castellanos, Arturo, Exploratory Content Analysis Using Text Data Mining: Corporate Citizenship Reports of Seven US Companies from 2004 to 2012 (2017). *Journal of Corporate Citizenship*, 66.
10.9774/T&F.4700.2017.ju.00007

This Article is brought to you for free and open access by the Mason School of Business at W&M ScholarWorks. It has been accepted for inclusion in Mason School of Business Articles by an authorized administrator of W&M ScholarWorks. For more information, please contact scholarworks@wm.edu.

Exploratory Content Analysis using Text Data Mining: Corporate Citizenship Reports of Seven U.S. Companies from 2004 – 2012

Carlos M. Parra* – cmaparra@fiu.edu

Monica Chiarini Tremblay** – monica.tremblay@mason.wm.edu

Karen Paul*** – paulk@fiu.edu

Arturo Castellanos**** – arturo.castellanos@baruch.cuny.edu

* Clinical Professor, Department of Information Systems and Business Analytics, College of Business, Florida International University, Miami, Florida

** Associate Professor, Raymond A. Mason School of Business, College of William & Mary, Williamsburg, Virginia

*** Professor, Department of Management and International Business, College of Business, Florida International University, Miami, Florida

**** Assistant Professor, Department of Information Systems and Statistics, Zicklin School of Business, Baruch College, CUNY, New York, New York

Abstract This study demonstrates the use of Text Data Mining (TDM) for exploring the content of a collection of Corporate Citizenship (CC) reports. The collection analyzed comprises CC reports produced by seven Dow Jones companies (Citi, Coca-Cola, ExxonMobil, General Motors, Intel, McDonalds and Microsoft) in 2004, 2008 and 2012. Exploratory content analysis using TDM enables insights for CC professionals and analysts, in less time using fewer resources, which in turn could help them explore collaboration opportunities around supply chains, re-training programs, and alternative risk mitigation strategies in terms of governance and compliance. In addition, TDM, using supervised machine learning on the whole collection (or corpus) as well as unsupervised machine learning on document collections by year, suggests the integration of CC considerations related to environmental sustainability in CC report components discussing the core business of some firms. This method has been used in many contexts in which a collection of documents needs to be categorized and/or analyzed to uncover new patterns and relationships.

Key Words: Corporate Citizenship Reports; Exploratory Content Analysis; Text Data Mining; CC Insights; CSR Integration

Introduction

The techniques used for exploring the content of Corporate Citizenship (CC) report collections, or large volumes of CC related documents, range from simple monitoring of data (such as investment amounts), categorizing and analyzing themes by manual coding, to word counting for making conceptual inferences. Professionals and practitioners tend to monitor data in order to establish competitive comparisons and benchmarks, since it is fairly easy to track key performance indicators associated with different CC programs. These performance indicators tend to be summarized in one-page yearly performance tables (i.e., tabulated numbers, or structured data) included in CC reports. This allows CC professionals to track a firm's performance (as well as that of its competitors) in terms of, for example, footprint reduction ef-

forts. Similarly, financial industry CC professionals may track amounts invested in financial education and asset building programs, or enterprise development initiatives by year and by country in order to justify current investment levels or argue for new ones.

Many academic studies have followed the second approach, which is manually coding CC reports or documents. For example, Moreno and Capriotti (2009) analyzed corporate websites of publicly traded Spanish firms by developing ten content categories and information hierarchies verified by independent coders. They found that the web has become a prominent medium for communicating CC issues, but that these sites lack external validation for guaranteeing the trustworthiness of claims posted. They further suggest that richer programming codes would allow better access and more two-way dialogue. Tang (2012) conducted a similar exercise by coding Chinese newspaper articles on Corporate Social Responsibility (CSR) in order to characterize the role of media in facilitating social dialogue. Another study focused on numerical coding by counting beneficiaries mentioned in CC reports to analyze the impact of CC initiatives on various quality of life dimensions including income, health, education, market capabilities, and democracy (Parra, 2008). Finally, Barkemeyer et al. (2014) used rhetoric analysis to develop sentiment metrics and a readability score to analyze CEO statements from CC and financial reports of 34 automobile, oil and gas, and mining companies. They found that the rhetoric of CEO statements in CC reports was more indicative of impression management than of accountability.

A different group of studies has focused on making inferences from word frequency counts in websites or other CC sources of information. Paul (2008) used frequency analysis on 100 websites, finding that social responsibility was the most frequent first-used term, but that, without order considerations for terms, sustainability was the most frequently used term. Meyskens and Paul (2010) followed a more comprehensive version of this approach. They analyzed the evolution of CC reporting practices in Mexico by dividing websites into first generation or early adopters of CC reporting practices and second-generation companies or recent adopters. They found that first-generation companies referred to stakeholders, citizenship, human rights, and codes of conduct more often than second-generation companies. From the corporate communications perspective, studies have related word frequency counts to affective management practices (Saito et al., 2012), brand differentiation (Gill et al., 2008), and have been performed in the context of a particular industry, such as oil and gas (Dickinson et al., 2008).

In an effort to automate the task of exploring the content of large CC report collections, researchers have started to use Text Data Mining (TDM) techniques to identify themes in large bodies of text or CC document collections. For example, Barkemeyer et al. (2009) analyzed 20 million newspaper articles published from January 1990 to July 2008. They found evidence that media made more references to sustainable development and CSR than to CC or Corporate Sustainability. Our work takes a similar but more elaborate approach in order to classify documents (using supervised machine learning) and to group them based on similarities (with unsupervised machine learning).

The studies summarized in this section evidence a progression in the use of technology for analyzing large bodies of text (specifically, collections of CC documents). Arguably, the objective has been to find ways to turn words into quantities that can be analyzed using novel statistical methods (some of which are described below and in Appendix A), in an effort to

help academic researchers, CC professionals, and analysts squeeze inferences and insights out of the numbers without having to actually read all the text. It would be prohibitively expensive and lengthy to actually read the quantity of articles (20 million) analyzed in one of the studies described above. Given the increasing amounts of information and documents produced and faster access to larger amounts of information, it seems reasonable to expect researchers to leverage technology for analyzing and exploring the content of increasing amounts of documents and bodies of text.

To clarify, whatever the tool utilized to analyze CC document collections (i.e., TDM, the actual critical reading of CC documents, or hiring consultancies to carefully dissect them), CC documents may include manufactured stories about their CC practices unsupported by actual facts on the ground. Our intention is not to verify the accuracy of the information included in the CC documents analyzed. We assume their contents to be true based on external audits commonly attached at the end of CC reports (including those produced by Dow Jones companies), but rather to demonstrate how exploratory content analysis could be performed using TDM on a collection of CC reports to gain CC insights.

We believe supervised and unsupervised TDM can help simplify and augment the job of CC professionals and analysts, by allowing them to obtain timely and pertinent CC insights based on how firms have treated and approached CC issues in their reports over time, even making comparisons with other firms. CC professionals and analysts having access to these CC insights, in less time with fewer resources, may become more efficient at identifying new partnership opportunities as well as alternative risk mitigation strategies. Our task is to utilize TDM to analyze CC reports of a sample of large, publicly traded Dow Jones companies (Citi, Coca-Cola, ExxonMobil, General Motors, Intel, McDonalds and Microsoft) in 2004, 2008 and 2012. We do this to showcase how TDM can be used to automate the task of conducting exploratory content analysis on a collection of CC reports. The exploratory analysis conducted here helps characterize the way firms have treated CC issues through time. We relate our findings to theoretical propositions that help validate the accuracy of our method. Moreover, our method makes available CC insights that may have been overlooked otherwise, or prohibitively expensive to find. In order to demonstrate how one CC analyst could perform this analysis in a matter of weeks, we begin by explaining what *machine learning and text analytics* is and how TDM works. We follow with a description of our methodology, which others are encouraged to replicate. Finally, we present our results, conclusions, and recommendations for future research.

Machine Learning

Machine learning has revolutionized many fields. Specifically, supervised machine learning provides advantages in being the “most widely used variety [...] can be used to train a classification system with the aid of a labeled set of examples” (Standage, 2016). As such, supervised machine learning is used to filter out spam email (Sasaki and Shinnou, 2005), classify images (as Facebook or Google Photos do), recognize speech (as automated customer service representatives and smart phones do), and to identify fraudulent credit card transactions or insurance claims (Fawcett and Provost, 2002, Bolton and Hand, 2002). Meanwhile, unsupervised machine learning is “used to search for things when you do not know what they look like” for example, data flow patterns indicative of cyber-attacks, or new kinds of insurance claim fraud (Economist, 2016). *TDM is a machine learning application designed to perform content analysis on large document collections.*

TDM is used in its unsupervised form, for example, by manufacturing firms to identify new product features that users value by analyzing product reviews (i.e., customer satisfaction surveys). TDM is also used by law firms, in its supervised version, to identify older but relevant judicial outcomes and build defense/prosecution strategies for current cases. Another use is by hospitals to augment doctor’s capabilities while diagnosing patients based on their symptoms. TDM has been used to predict the likelihood of Veterans Affairs (VA) hospital patients’ falling using the contents of their clinical histories and progress notes (Tremblay et al., 2009). Finally, supervised TDM is also regularly used in authorship attribution studies (Juola, 2006).

TDM can be used to explore the content of publicly available documents produced by an organization, or a group of organizations (e.g., a collection of CC reports). This involves the assignment of natural language texts to one or more predefined categories based on the collection’s content (Dumais et al., 1998). This text categorization process is then used to describe, infer or predict associations between terms in documents contained in the collection. In Table 1, we describe common techniques used to categorize text and perform exploratory content analysis on a document collection.

Table 1. Content Analysis Methods

Method	Description
Content Analysis Done by humans assisted by computers	Systematic examination of large quantities of textual data. This can include frequency analysis. Treats text as discourses to be understood and interpreted (Laver et al., 2003) Three approaches from (Hsieh and Shannon, 2005): 1. Conventional: coding categories derived directly from text by independent coders 2. Directed: initial codes guided by theories 3. Summative: involves counting and comparisons, usually of keywords or content, followed by the interpretation of the underlying context
Text Data Mining Done by computers guided by humans (enables knowledge workers to perform more tasks more efficiently)	Text mining is the discovery and extraction of interesting, non-trivial knowledge from free or unstructured text (Kao and Poteet, 2007) and includes: Information retrieval Text classification and clustering Entity, relation, and event extraction Utilizes Natural Language Processing (NLP). NLP extracts meaning and representation from text using linguistic concepts such as part of speech and grammatical structure.

As illustrated in Table 1, different approaches can be used to summarize themes found in documents. In the realm of CC, most content analysis studies and related tasks have been performed by humans (assisted by computers), as discussed above. In this study, we show how exploratory content analysis can be performed by computers guided by humans (using TDM). This involves automating a task that has traditionally been completed by trained humans, usually knowledge workers. As with most automation, competitive advantages may accrue to organizations that are quick to realize that a labor intensive task, which used to require sever-

al people and months to complete (e.g., directed content analysis, where text categorization codes guided by theory are applied on large document collections) can now be handled by one trained analyst in a matter of weeks, enabling organization to analyze relevant data quickly and economically.

How Text Data Mining (TDM) works

Eighty percent of business-relevant information originates in unstructured form (i.e., not in tabulated format), and primarily from text (Grimes, 2008). Consequently, researchers and practitioners have taken interest in deriving high quality information and business insights from text through the use of *text analytics and machine learning* (or TDM). TDM is a process of knowledge discovery which allows the extraction of implicit and potentially useful information from textual data using statistical methods (Feldman and Dagan, 1995). These algorithms take a statistical approach, calculating word frequencies and term weights to discriminate among, or uncover associations between, terms and documents in a collection using similarity detection techniques. High quality, in TDM, usually refers to some combination of relevance, novelty, and adding material of interest (Tan, 1999b). Typical TDM tasks include text categorization, text clustering, concept extraction, sentiment analysis, and document summarization (Dörre et al., 1999). Thus, TDM is the process of *extracting interesting and non-trivial patterns from collections of text documents* by combining machine learning, Natural Language Processing (NLP), information retrieval, and knowledge management (Feldman and Sanger, 2007, Tan, 1999a).

The documents analyzed can contain any type of text (e.g., claim files, reports, emails, progress notes, etc.). Unsupervised TDM helps reveal patterns and relationships among these documents as well as between the terms contained in the documents. Supervised TDM allows for content categorization to facilitate document classification, sentiment analysis, author attribution, and ontology management by interpreting nuances of human language using NLP. TDM algorithms start by turning text into numerical representations to which researchers and practitioners can apply multivariate statistical techniques for exploring and analyzing the content of documents in a collection. In particular, Latent Semantic Analysis (LSA) focuses on mathematically representing a document's contents through weighted vectors of terms that summarize that document's latent "concepts" (Deerwester et al., 1990a). In essence, TDM is a statistical approach to sift through large document collections of unstructured text (i.e., not tabulated), identify concepts or topics, and characterize the way they are treated.

First, TDM counts occurrences of words or the number of terms in the documents analyzed. Second, it generates a matrix that has all the terms across columns and the documents across rows. Each cell in the matrix contains the number of times a term appears in each document (or term frequencies). As the number of documents in a collection increases, the number of columns increases as new terms are added to the matrix. Third, TDM reduces the size of this matrix through parsing, in order to improve computational performance, by removing words that appear either too frequently or too sparsely. Zipf's law states that terms that occur in many documents do not act as good discriminators and thus should weigh less during statistical analysis than terms appearing in fewer documents. Associated technical details are discussed in Appendix A. Parsing also involves stemming which removes words that have a common root, for example: number, numbers, numbered, numbering, etc. Finally, parsing involves removing words that according to the task at hand do not add value to the analysis.

This last group of words is referred to as a “stop list” and the methodology section outlines the one developed for this study.

Fourth, term frequency entries in this rectangular matrix are transformed into weights using weighting schemes (Salton and Buckley, 1988). *Frequency weights*, often called *local weights* represent the first step in quantifying documents and relate to the importance of terms contained in those documents. Unfortunately, absolute term frequency counts can be influenced by documents that have high variability with respect to size. For this reason, *term weights*, often called *global weights*, modify frequency weights to adjust for document size and term distribution in the whole document collection (or corpus). The aim is simply to grasp both the number of times a word appears in a document and the number of times it appears in the corpus. There are different frequency-to-weight transformation techniques. Selection depends on the length of the documents being analyzed as well as the type of machine learning application. Inverse Document Frequency (IDF) is used for analyzing longer documents in unsupervised machine learning applications, while Mutual Information (MI) must be used for supervised machine learning tasks. Research has shown that good results are often obtained using entropy transformations for short documents and IDF for longer documents like CC reports (Woodfield, 2011). Fifth, TDM transposes this rectangular matrix of transformed frequencies (or weights) to obtain Singular Vector Decompositions (SVDs), which are similar to the Eigen-vectors obtained through exploratory principal component factor analysis.

SVDs are a reduced set of vectors that summarize the contents of the original document-by-term matrix through a dense, low-dimensional representation of the corpus, and this in turn allows for the exploration of *document associations* (Text Clusters) or *term associations* (also called Term Clusters or Text Topics) (Dempster et al., 1977; Do and Batzoglou, 2008, Zhang et al., 2005). In this study, *document associations* are explored by year to identify, through document groupings or clusters, non-intuitive relationships between CC report components from different firms at different points in time in order to obtain CC insights. Elsewhere, SVDs obtained while exploring *term associations* were used to help visualize prominent voices around specific CC issues and to characterize the way they were treated (Parra et al., 2016a, Parra et al., 2016b). Findings suggest that it is difficult for firms to maintain a prominent voice around a CC issue through time, but, when firms manage to do so, it is because the issue in question has direct core business implications. Finally, SVDs can also be used to train algorithms to automate text classifications tasks for supervised machine learning applications. The rest of the manuscript describes how TDM was used to perform exploratory content analysis on CC reports produced by seven U.S. firms in 2004, 2008 and 2012.

Methodology

After CC reports were obtained from official corporate websites, two analyses were conducted. First, we performed supervised machine learning on the whole document collection (or corpus). Second, we performed unsupervised machine learning on document collections by year.

We performed supervised machine learning on the corpus to compare and validate the labeling of different parts of each report done by a subject matter expert against the labeling

performed by a trained classification algorithm. Each CC report was divided into parts, or CC report components, in order to explore how components from different firms would group or cluster based on their similarities (which was the second analysis performed). As explained above, non-intuitive relationships between CC report components, from different firms at different points in time, are important to indicate potential partnerships or risk mitigation strategies.

The subject matter expert classified and labeled each CC report component. Elaboration of the composition and nature of these components and how they are obtained is given in the Data Formatting Framework section. The subject matter expert is a CC executive with seven years of experience advancing CC strategies in different industries and overseeing the production of CC reports following varied guidelines. We decided to use only one coder because the effective automation of a task ought to involve frugality as well as the capacity to augment an analyst's abilities to code CC reports, verify coding consistency, and perform exploratory content analysis on large document collections to gain insights. Second, we performed unsupervised machine learning on document collections by year, using CC report components from 2004, 2008, and 2012. For each year we explored the way documents grouped together based on their similarities (i.e., document associations) in order to gain insights and identify potential areas for collaboration.

Sample

This sample included representative firms from diverse industries in the Dow Jones Industrial Average (also known as DJIA, Industrial Average, the Dow Jones, the Dow Jones Industrial, the Dow 30, or the Dow). The DJIA was established in 1896 and shows trading patterns for the U.S. stock market. Since the aim of this study is to explore general CC report tendencies and associations, the DJIA was considered an appropriate source because of its wide scope, the representativeness of the sectors included, and its relative stability over time. The DJIA encompasses a wide diversity of economic sectors using only thirty firms, allowing parsimonious sampling with the expectation that the data obtained might be replicable in future analyses. Table 2 contains basic company information for the time period of interest.

Table 2. Firm Demographics for 2004, 2008 and 2012*

Company	2004			
	Number of Employees	Revenues (Billion Dollars)	Mkt. Cap. (Billion Dollars)	EPS (Diluted Quarterly Dec 31)
Citi	157,812	86.2		10.16
Coca-Cola	50,000	22.0	2,410,089,440 (Shares Outstanding)	0.2488
Exxon-Mobile	85,900	291.3	328.13	1.304
General Motors		193.5	479.6	1.37
Intel	85,000	34.2	48.14	0.3342
McDonalds	398,000	17.1	27.84	0.31
Microsoft	61,000 (2005)	36.8	92.39	0.32
2008				
Citi	176,003	52.8	36	-33.97
Coca-Cola	92,400	31.9	2,314,658,162 (Shares Outstanding)	0.2143
Exxon-Mobile	79,900	459.6	397.24	1.553
General Motors	243,000	149.0	91.05	-52.38
Intel	83,900	37.6	50.47	0.0416
McDonalds	390,000	23.5	28.46	0.8698
Microsoft	91,000	60.4	72.79	0.47
2012				
Citi	192,244	70.2	120	0.3934
Coca-Cola	150,900	48.02 (Net Operating Revenues)	4,456,717,996 (Shares Outstanding)	0.4095
Exxon-Mobile	76,900	453.1	389.68	2.191
General Motors	213,000	152.3	149.42	12.73
Intel	105,000	53.3	84.35	0.4842
McDonalds	440,000	27.6	35.39	1.381
Microsoft	94,000	73.7	121.27	0.76*

Sources for data in Appendix B

CC reports for each firm for the year indicated were downloaded from the corresponding official corporate websites. Table 3 shows the CC reports downloaded and their number of PDF pages by company and by year.

Table 3. Length of CC Reports (in PDF pages) for 7 U.S. Dow Jones Companies in 2004, 2008 and 2012

Company	2004	2008	2012	Average by company
Citi	56	95**	82	77.7
Coca-Cola	44	65	91	66.7
ExxonMobil	62	48	67	59
General Motors	172	Chapter 11	57	76.3
Intel	40	108	126	91.3
McDonalds	88	70	8*	55.3
Microsoft	80	5*	89	85
Average by year	77.4	55.9	74.3	

* Document too small to be divided into components

** Un-editable file could not be scrubbed or divided into components

The highest average number of pages occurred in 2004, followed by 2012. The average in 2008 was low because General Motors did not publish a CC report and Microsoft issued only a five-page update. Intel had the highest average number of PDF pages in the three years analyzed, closely followed by Microsoft. Citi and General Motors were similar, averaging 77 pages. ExxonMobil and McDonalds had the lowest number of PDF pages in the years analyzed, fewer than 60 pages. These descriptive statistics are for illustrative purposes only, since document length is inconsequential for the TDM settings used.

In total, we downloaded seven 2004 reports, six reports from 2008, and seven from 2012. In 2008, General Motors filed for Chapter 11 bankruptcy and did not issue a CC report. Microsoft in 2008 and McDonalds in 2012 issued only short updates which could not be divided into components and analyzed using the TDM settings used for longer documents (i.e., using

IDF instead of entropy frequency to weight transformations, as explained in the TDM section above). Citi’s 2008 CC report had settings that prevented its division into components suitable for analysis. Thus, we analyzed seven 2004 reports, four reports from 2008, and six reports from 2012. Only the main text of each CC report was analyzed. We excluded pictures, tables, footnotes, hyperlinks, and diagrams from the analysis in order to obtain main text files.

Data Formatting Framework and CC Report Components

The subject matter expert divided the main text files emerging from each CC report downloaded into smaller text files called components. We considered three main CC practitioner frameworks which act as reporting guidelines: the Global Reporting Initiative (GRI), the International Standards Organization (ISO) 26000, and the Sustainability Accounting Standards Board (SASB). Taken together, the GRI and ISO 26000 include the following seven sustainability dimensions:

1. **Organizational:** related to strategy (mission/vision), governance, ethics, leadership, stakeholder engagement (shareholders), compensation, regulatory and legal challenges
2. **Economic:** related to financial performance, market presence, long term viability, accounting for externalities (indirect economic impacts), procurement and fair operating practices
3. **Environmental:** related to materials and footprint (energy, water, biodiversity, emissions, waste, etc.)
4. **Labor:** related to employees, occupational safety, working conditions, training and development, recruitment, retention, union practices, diversity, equal opportunity/remuneration, grievance mechanisms
5. **Human Rights:** related to child labor, forced/compulsory labor, indigenous rights
6. **Society:** related to local community development and engagement, access to services/products
7. **Product/Business:** related to consumer safety and welfare, quality, packaging, labeling, ethical advertising, privacy, pricing, research/development and innovation

We compare these frameworks in Table 4. The SASB framework only refers to five dimensions, because economic performance is included in Business Model and Innovation, and Human Rights issues are included in Human Capital.

Table 4. Summary of Dimensions in Main CC Reporting Guidelines

CC Reporting Guidelines	Components / Dimensions						
GRI	Strategy, organization engagement governance, ethics	Economic Performance Indicators	Environmental Performance Indicators	Labor Practices and Decent Work Performance Indicators	Human Rights Performance Indicators	Society Performance Indicators	Product Responsibility Performance Indicators
ISO 26000	Organization and governance	Fair Operating Practices	The Environment	Labor Practices	Human Rights	Community Involvement and Development	Consumer Issues
SASB	Leadership and governance		Environment	Human Capital		Social Capital	Business Model and

							Innovation
--	--	--	--	--	--	--	------------

The SASB framework was used in this analysis because it is overarching and simpler, and the more parsimonious framework permitted a more efficient analysis, which also helped advance our frugal automation methodology. SASB’s framework was slightly adjusted to facilitate exploratory content analysis based on subject matter expert recommendations about the way CC reports in this collection were put together and the way information was commonly found in them. The following modifications were made:

- Raw material demand issues, which in SASB’s framework were classified in Government and Ethics, were added to the Environmental component.
- Marketing and ethical advertising issues, which in SASB’s framework were classified in Social Capital, were considered part of the Business component.
- Supply chain issues, which in SASB’s framework are classified in Government and Ethics dimension, were added to the Social Capital component.

Table 5. Modified SASB Five-Dimensional Framework Used to Divide Contents Files

Environment	Social capital	Human capital	Business model and innovation	Leadership and governance
<ul style="list-style-type: none"> • Climate Change risk • Environmental remediation • Water use and management • Energy management • Fuel management and transportation • Green House Gas emissions and air pollution • Waste management and effluents • Biodiversity impacts • Natural resource and raw material demand* 	<ul style="list-style-type: none"> • Communications and stakeholder engagement • Community development • Impact from facilities • Customer satisfaction • Customer health and safety • Customer privacy • Disclosure and labeling • Access to products or services provided • New market development • Disaster relief efforts • Employee volunteering • Supply chain standards and selection* • Supply chain engagement and transparency* 	<ul style="list-style-type: none"> • Employee diversity and equal opportunity • Employee training and development • Recruitment and retention • Compensation and benefits • Labor relations and union practices • Employee health, safety and wellness • Human rights (child and forced labor policies) 	<ul style="list-style-type: none"> • Long term viability of core business • Economic and financial performance • Sustainability/Citizenship strategy • Research, development and product innovation • Product quality and safety • Product societal value • Pricing • Product life cycle • Accounting for externalities • Packaging • Marketing and ethical advertising* 	<ul style="list-style-type: none"> • Regulatory and legal issues • Policies, standards and codes of conduct • Decision making instances, structure, independence and transparency • Business ethics and competitive behavior • Shareholder engagement • Executive compensation • Lobbying and political contributions

* These issues belong to a different sustainability dimension in SASB’s original classification (More information can be found at <http://www.sasb.org/materiality/determining-materiality/>)

The subject matter expert manually divided the main text files from each CC report into five separate components. This was done based on the modified SASB framework presented above, to produce the following elements for each CC report: a business component, a governance component, an environmental component, a human capital component, and a social capital component. We analyzed seven 2004 reports (resulting in 35 separate components), four reports from 2008 (resulting in 20 separate components) and six from 2012 (resulting in 30 separate components). In total, 85 components constitute this study’s corpus. The resulting component text files contain all of the main text in that year’s CC report for each firm. Figure 1 summarizes our approach for obtaining CC report components.

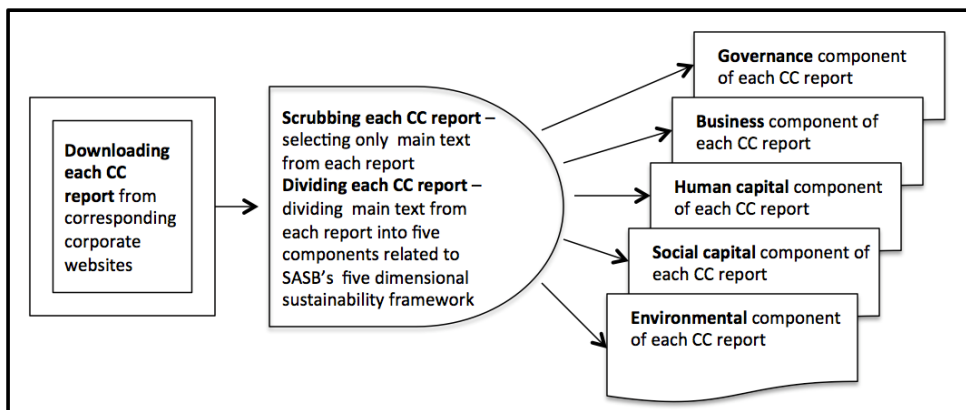


Figure 1. Obtaining CC Report Components

Labels were assigned to each component by naming the corresponding file using the following nomenclature: [company name] [year] [sustainability dimension assigned by the subject matter expert]. Thus, a text file named “Citi2004business” refers to a CC report component obtained from Citi’s 2004 report that mainly discusses business and innovation issues.

Developing a Stop List

As explained above, Stop Lists prevent TDM algorithms from considering terms that do not add value. The creation of the stop list is iterative, and is conducted on the corpus (85 components) as shown in Figure 2.

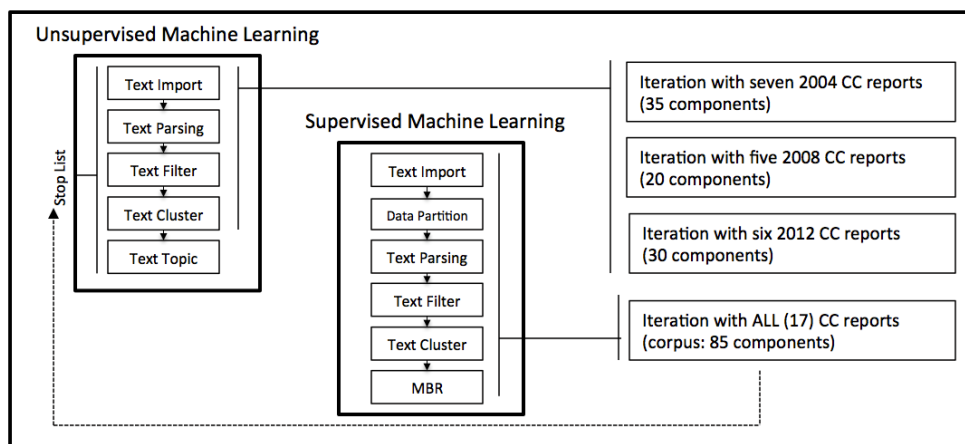


Figure 2. Analyzing Contents of CC Report Components

Frequently occurring terms in a document tend to be assigned high weights by the TDM algorithm, but often these terms do not add value or are redundant to the analysis. For example, analyzing Intel’s 2004 social capital component without a stop list would assign a higher weight to “Intel” and “microprocessor” than to more meaningful terms such as “wireless”, “teacher”, “computer”, and “education.” Thus, we used a stop list to exclude non-value-added terms.

Table 6. List of Words Included in “Stop List”

Term
mcdonald
ronald
restaurant
intel
microprocessor
processor
wafer
chipset
exxonmobil
oil
coca-cola
beverage
citi
citigroup
citibank
bank
gm
opel
saab
automotive
vauxhall
microsoft
windows
outlook
bottler

Results

In this section, we describe the results of both supervised machine learning TDM and unsupervised machine learning TDM approaches. First, we compare the components labeled by the subject matter expert to the classification done by supervised machine learning, in order to validate the consistency of divisions produced by the subject matter expert. Second, using unsupervised machine learning, we explore the way CC report components group through the years in order to characterize the way firms have treated and approached CC issues over time. Figure 2 shows how the first approach (supervised machine learning) is conducted on the corpus, while the second approach (unsupervised machine learning) is applied on CC report component collections by year.

Supervised Machine Learning – Document Classification Exercise

Supervised machine learning requires model building and model validation. The original dataset (the corpus of 85 components) is split by the algorithm into a training set with 45 components and a validation set containing the remaining 40 components. Components are randomly assigned by the algorithm to one of two mutually exclusive sets: training or validation. Partitioning data for classification purposes helps evaluate the performance of classification models by benchmarking the accuracy of the model classification on the test data. In addition, data and text mining tools offer many possible algorithms for building classification models, e.g., decision trees, logistic regression, neural networks, and memory based reasoning (MBR). We considered several models, then selected the most accurate one. We were interested in validating the subject matter expert labels described in Table 7 for each CC report component with the classification produced by supervised machine learning. Since previous

research successfully used MBR techniques when the dependent variable was categorical and had several possible values (Masand et al., 1992), it was judged to be an appropriate choice.

Table 7. CC Report Components and Labels per Company for 2004, 2008 and 2012 (Corpus - 85 Components)

Year	2004		2008		2012	
Company	Component file name	Label assigned by expert	Component file name	Label assigned by expert	Component file name	Label assigned by expert
Citi	Citi2004business	Business component	Uneditable file		Citi2012business	Business component
	Citi2004environment	Environmental component			Citi2012environment	Environmental component
	Citi2004governance	Governance component			Citi2012governance	Governance component
	Citi2004human capital	Human capital component			Citi2012human capital	Human capital component
	Citi2004social capital	Social capital component			Citi2012social capital	Social capital component
Coca-Cola	Coca-Cola2004business	Business component	Coca-Cola2008business	Business component	Coca-Cola2012business	Business component
	Coca-Cola2004environment	Environmental component	Coca-Cola2008environment	Environmental component	Coca-Cola2012environment	Environmental component
	Coca-Cola2004governance	Governance component	Coca-Cola2008governance	Governance component	Coca-Cola2012governance	Governance component
	Coca-Cola2004human capital	Human capital component	Coca-Cola2008human capital	Human capital component	Coca-Cola2012human capital	Human capital component
	Coca-Cola2004social capital	Social capital component	Coca-Cola2008social capital	Social capital component	Coca-Cola2012social capital	Social capital component
Exxon-Mobil	Exxon-Mobil2004business	Business component	Exxon-Mobil2008business	Business component	Exxon-Mobil2012business	Business component
	Exxon-Mobil2004environment	Environmental component	Exxon-Mobil2008environment	Environmental component	Exxon-Mobil2012environment	Environmental component
	Exxon-Mobil2004governance	Governance component	Exxon-Mobil2008governance	Governance component	Exxon-Mobil2012governance	Governance component
	Exxon-Mobil2004human capital	Human capital component	Exxon-Mobil2008human capital	Human capital component	Exxon-Mobil2012human capital	Human capital component
	Exxon-Mobil2004social capital	Social capital component	Exxon-Mobil2008social capital	Social capital component	Exxon-Mobil2012social capital	Social capital component
General Motors	General Motors004business	Business component	Chapter 11		General Motors012business	Business component
	General Motors004environment	Environmental component			General Motors012environment	Environmental component
	General Motors004governance	Governance component			General Motors012governance	Governance component
	General Motors004human capital	Human capital component			General Motors012human capital	Human capital component
	General Motors004social capital	Social capital component			General Motors012social capital	Social capital component
Intel	Intel2004business	Business component	Intel2008business	Business component	Intel2012business	Business component
	Intel2004environment	Environmental component	Intel2008environment	Environmental component	Intel2012environment	Environmental component
	Intel2004governance	Governance component	Intel2008governance	Governance component	Intel2012governance	Governance component
	Intel2004human capital	Human capital component	Intel2008human capital	Human capital component	Intel2012human capital	Human capital component
	Intel2004social capital	Social capital component	Intel2008social capital	Social capital component	Intel2012social capital	Social capital component
McDonalds	McDonalds004business	Business component	McDonalds008business	Business component	8-page update, document too short to meaningfully create components	
	McDonalds004environment	Environmental component	McDonalds008environment	Environmental component		
	McDonalds004governance	Governance component	McDonalds008governance	Governance component		
	McDonalds004human capital	Human capital component	McDonalds008human capital	Human capital component		
	McDonalds004social capital	Social capital component	McDonalds008social capital	Social capital component		
Microsoft	Microsoft2004business	Business component	5-page update, document too short to meaningfully create components		Microsoft2012business	Business component
	Microsoft2004environment	Environmental component			Microsoft2012environment	Environmental component
	Microsoft2004governance	Governance component			Microsoft2012governance	Governance component
	Microsoft2004human capital	Human capital component			Microsoft2012human capital	Human capital component
	Microsoft2004social capital	Social capital component			Microsoft2012social capital	Social capital component

Table 8 shows a comparison of validation set classification done by the subject matter expert to that produced by the supervised machine learning TDM approach.

Table 8. Performance Supervised Learning TDM Approach

Item	Component file name	Subject matter expert classification	Supervised Machine Learning TDM Classification
1	Citi 2004 environment	Environment	Environment
2	Citi 2004 governance	Governance	Governance
3	Coca-Cola 2004 business	Business	Business
4	Coca-Cola 2004 governance	Governance	Governance
5	Coca-Cola 2004 human capital	Human Capital	Human Capital
6	Coca-Cola 2004 social capital	Social Capital	Social Capital
7	General-Motors 2004 business	Business	Business
8	Intel 2004 environment	Environment	Environment
9	Intel 2004 human capital	Human Capital	Human Capital
10	Intel 2004 social capital	Social Capital	Social Capital
11	McDonalds 2004 human capital	Human Capital	Human Capital
12	Microsoft 2004 business	Business	Business
13	Microsoft 2004 governance	Governance	Governance
14	Microsoft 2004 human capital	Human Capital	Human Capital
15	Coca-Cola 2008 human capital	Human Capital	Human Capital
16	ExxonMobil 2008 business	Business	Environment
17	ExxonMobil 2008 environment	Environment	Environment
18	ExxonMobil 2008 governance	Governance	Governance
19	Intel 2008 business	Business	Business
20	Intel 2008 environment	Environment	Environment
21	Intel 2008 governance	Governance	Governance
22	Intel 2008 social capital	Social Capital	Social Capital
23	McDonalds 2008 environment	Environment	Environment
24	Citi 2012 social capital	Social Capital	Social Capital
25	Coca-Cola 2012 business	Business	Environment
26	Coca-Cola 2012 environment	Environment	Environment
27	Coca-Cola 2012 human capital	Human Capital	Human Capital
28	ExxonMobil 2012 business	Business	Environment
29	ExxonMobil 2012 environment	Environment	Environment
30	ExxonMobil 2012 governance	Governance	Governance
31	ExxonMobil 2012 human capital	Human Capital	Human Capital
32	ExxonMobil 2012 social capital	Social Capital	Human Capital
33	General-Motors 2012 governance	Governance	Governance
34	General-Motors 2012 social capital	Social Capital	Environment
35	Intel 2012 environment	Environment	Environment
36	Intel 2012 human capital	Human Capital	Human Capital
37	Intel 2012 social capital	Social Capital	Social Capital
38	Microsoft 2012 business	Business	Business
39	Microsoft 2012 governance	Governance	Governance
40	Microsoft 2012 social capital	Social Capital	Social Capital

Out of 40 components in the validation set, five components, which appear in *italic* and **bold** in table 8, were classified differently by supervised machine learning. This represents a 12.5% discrepant classification ratio, indicating that the subject matter expert and the supervised machine learning TDM algorithm coincided 87.5% of the time. The subject matter expert classified three components as business components, while the supervised machine learning technique classified them as environmental components, suggesting *overlapping classifications*. We further explored the nature of these overlapping classifications through Table 9. Here we provide excerpts from two sample documents to showcase both discrepant and coincident labels in terms of business and environmental components.

File	Excerpts from text	Classification by subject matter expert	Classification by TDM
ExxonMobil 2012business	“After decades of growth, energy related GHG emissions are expected to plateau around 2030, despite a steady rise in overall energy demand. As global demand increases, advanced technologies to boost energy supplies are becoming more important. Thirty years from now, oil and natural gas are expected to meet about 60 percent of global demand, and an increasing share of this supply will be produced from unconventional oil and gas resources and deepwater fields. ExxonMobil is developing new technologies to support the safe and economical development of these resources, which are not always located where energy demand is highest. International trade plays an important role in ensuring the wide distribution of energy around the world. Around 2025, we expect North America will transition to a net exporter of energy, which will help grow the U.S. economy while providing much-needed energy to other regions of the world.”	Business Text contains references to: -Long term viability of core business -Research, development and product innovation	Environment
Intel 2004environment	“In 2004, Intel Massachusetts awarded more than \$220,000 in grants to four model projects with the potential to recharge more than 40 million gallons of water to local aquifers that replenish the Assabet River and its tributaries. The \$1.5 million Intel Assabet River Aquifer Recharge Fund remains in place to award grants to support such projects. For the 11th year in a row, Intel Ireland funded a comprehensive limnological survey of the nearby Rye, a tributary of the River Liffy and an important salmon spawning ground. Extensive ecological information is now available, enabling individuals to study even minute changes in the river’s long term health.”	Environment Text contains references to: - Environmental remediation	Environment

Table 9. Classification Comparison

Table 9 shows that the subject matter expert recognized references to environmental sustainability in the context of long-term business viability or product innovation as pertaining to a firm’s business component, while supervised machine learning TDM identified these same references as pertaining to a firm’s environmental component. Consequently, supervised ma-

chine learning TDM approach results show that ExxonMobil's 2008 and 2012 business components and Coca-Cola 2012 business components contain enough environmental sustainability considerations to consider them as being environmental components (see Table 8). This was the first indication that environmental sustainability considerations were permeating core business components for some firms, suggesting how CSR integration could be taking place in the document collection analyzed here.

CSR Integration

The idea that CSR and corporate strategy may be intertwined, integrated, or overlapping, is a somewhat recent conceptualization (McWilliams et al., 2006). From this line of thought it follows that, rather than detracting from financial performance (Friedman, 2007), being antagonistic to capitalistic enterprise (Marcoux, 2000), or even being a disguise for socialist ideologies (Direction, 2003), CSR may contribute to or even become an essential element of corporate strategy to create competitive advantage (Porter and Kramer, 2006). The idea of strategic CSR is a logical extension. Dating from early formulation of the CSR imperative, Carroll conceptualized strategic CSR as that which will help the firm accomplish strategic business goals (Carroll, 1979). Since the first development of stakeholder theory, Freeman has consistently maintained that the most promising opportunities for managers come from areas where the interests of different stakeholders are aligned rather than in opposition (Freeman, 2010).

Studies have explored levels of CSR integration using different approaches. Ihlen and Roper (2011) used grounded theory to investigate how non-financial corporate 2006 and 2008 reports from thirty Fortune 500 companies, communicated sustainability and sustainable development. They found that “sustainability and sustainable development are part of common business language” (p. 48), which is a qualitatively obtained empirical finding in support of CSR integration. They also found that environmental issues were increasingly addressed in these reports, but that this was done in a corporate-centric manner. Meanwhile, Yadava and Sinha (2015) assigned scores to each Global Reporting Initiative (GRI) indicator included in 2012 CC reports produced by five Indian firms in an effort to quantify the level of sustainability inclusiveness. They found that reporting on the economic dimension was more comprehensive than reporting on social and environmental dimensions, but that “environmental dimension is becoming comprehensive because of increased environmental awareness” (p. 9). Results obtained from supervised machine learning TDM help corroborate the above findings and exemplify how CSR integration may occur in environmental sustainability terms, especially when considering the fact that ExxonMobil and Coca-Cola have environmental sustainability considerations in their core business components insofar as they rely on natural resource extraction (oil and water, respectively).

In addition, the inclusion of environmental sustainability considerations in core business components leads to overlapping classifications while conducting supervised machine learning TDM. These overlaps helped us expose the nature of three out of the five discrepancies highlighted in Table 8. In general, the similarity between the subject matter expert and supervised machine learning TDM classifications attests to the consistency with which CC reports were divided into components. This method could be used by a CC analyst wishing to check her own consistency while dividing CC reports into components. As emphasized above, dividing CC reports into components is an essential step in order to be able to explore unintui-

tive associations between CC report components, from different firms at different points in time, and gain CC insights.

Table 8 shows the extent to which two coders (i.e., subject matter expert and machine learning algorithm) agreed and disagreed while classifying CC report components into the five categories: business, environment, government, social capital, or human capital. In particular, we used the validation set to calculate the Cohen’s Kappa (κ) coefficient and assess inter-coder reliability. There were five instances in which the two coders did not agree or had discrepant classifications. Cohen’s κ demonstrated very good agreement between both coders, $\kappa = .844$, $p < .001$ (Altman, 1990, Landis and Koch, 1977). We acknowledge the fact that using only one coder may have introduced a bias into the way the classification algorithm (or supervised machine learning TDM approach) was trained, however, for frugal automation purposes and the CC insights that can emerge from exploratory content analysis, what matters is that any classification bias is consistently applied. We continued our exploratory content analysis using unsupervised machine learning TDM.

Unsupervised Machine Learning – Exploring Document Groupings by Year

The exploration of document clusters (or groupings) allowed us to examine the way in which CC report components from different firms group together based on content similarities for each year in the analysis. By doing so, we were able to explore how each firm approached and treated CC issues in their reports over time, while uncovering unintuitive associations between CC report components out of which CC insights can be gained. As shown in Figure 2, for 2004, we had seven reports and a total of 35 components. For 2008, we had five reports and a total of 20 components. For 2012, we had six reports and a total of 30 components. For each year’s analysis, we provide depictions of the Euclidian (i.e., spatial) distance between clusters (or groupings of documents), indicating that the clusters are sufficiently different from one another, along with a description of each document grouping obtained.

2004 Cluster Analysis

We obtained a total of six clusters for the collection of 2004 CC report components. Figure 3 shows the six clusters and how they differed from each other in Euclidean (i.e., geometric) distance terms. Table 10 uses the name of the clusters graphed in Figure 3, to provide cluster descriptions and to detail the file names grouped in each cluster for the 2004 iteration.

2004 Document Groupings

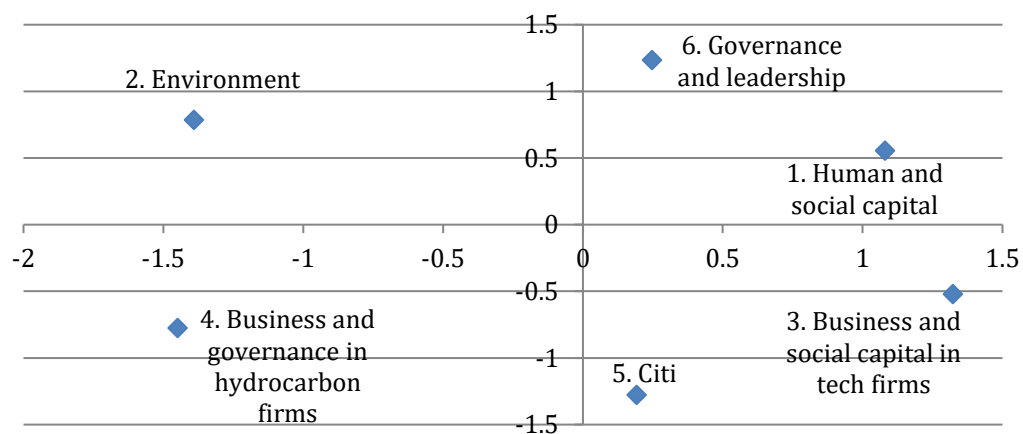


Figure 3. 2004 Text Clusters

Table 10. 2004 Clusters Descriptions

Cluster Number	Cluster Name	Cluster Description	Components Grouped
1	Human and social capital	Cluster points to <i>human capital</i> themes (e.g., “diversity, skills, training, etc.”) - only Microsoft’s human capital text file did not group here but in document cluster 3 - and also to <i>social capital</i> issues (e.g., “health, safety, education, etc.”)	Citi2004governance Citi2004human capital Coca-Cola2004human capital Coca-Cola2004social capital ExxonMobil2004human capital ExxonMobil2004social capital General Motors2004human capital General Motors2004social capital Intel2004human capital McDonalds2004human capital
2	Environment	Cluster refers to <i>environmental</i> issues (e.g., “conservation, climate, waste, packaging, biodiversity, water, air, etc.”) Citi’s 2004 environmental text file grouped with Citi’s business and social capital components in cluster 5.	Coca-Cola2004environment ExxonMobil2004environment General Motors2004environment Intel2004environment McDonalds2004environment Microsoft2004environment
3	Business and social capital in tech firms	Cluster refers to <i>business and social capital considerations in tech firms</i> as it only groups Microsoft and Intel components with descriptive terms that refer to technology supply chain issues (e.g., “China, stock, skills, tools, values, service, technical, etc.”)	Intel2004business Intel2004social capital Microsoft2004business Microsoft2004governance Microsoft2004human capital Microsoft2004social capital
4	Business and governance in hydrocarbon firms	Cluster brings up <i>business and governance considerations for the hydrocarbon firms</i> insofar as it includes terms such as: “engine, fuel, demand, trends, air, costs, etc.”, which are common in extractive and automotive industries.	ExxonMobil2004business General Motors2004business General Motors2004governance
5	Citi	Cluster includes most of <i>Citi’s components</i> (business, environment and social capital) –except governance and human capital components, which appeared in Cluster 1, which includes the following terms: “credit services, stock, share, potential, future, leading markets, etc.” Thus, this Text Cluster raises <i>financial industry considerations</i> .	Citi2004business Citi2004environment Citi2004social capital
6	Governance and leadership	Cluster points to <i>governance considerations</i> through: “guidelines, conduct, directors, effective standards, compliance, etc.”	Coca-Cola2004business Coca-Cola2004governance ExxonMobil2004governance Intel2004governance

			McDonalds2004business McDonalds2004governance McDonalds2004social capital
--	--	--	---

In 2004, and for all other years analyzed, we first focused on describing where firms’ business components grouped. No business components grouped in the environmental cluster, and only one environmental component grouped outside the “Environment” cluster (No. 2), Citi’s 2004 environment component (Citi2004environment) grouped in the Citi cluster. We believe this may have happened because financial firms, by having low carbon footprint operations, included environmental considerations in their banking products and thus treated environmental CC issues differently from other firms (e.g., most firms in the 2004 collection used terms such as waste, packaging, biodiversity, water, air, etc.). Intel and Microsoft business components grouped with the tech firms cluster, while General Motors and ExxonMobil business components grouped with the Hydrocarbon firms cluster, indicating a similar treatment of CC issues based on firm types instead of individual company characteristics (i.e., evidencing firm-type convergence in the treatment of CC issues). Coca-Cola and McDonalds business components grouped in the governance and leadership cluster.

Human and social capital CC issues were treated in very similar ways by most firms in 2004, using terms such as diversity, skills, training, health, safety, education, etc. The fact that Microsoft’s 2004 human capital component did not group in cluster No. 1 (Human and social capital) but in cluster No. 3 (Business and social capital in tech firms) may be indicative of Microsoft’s treatment of human capital CC issues in a manner that is more aligned with its treatment of business and governance issues than with traditional human resource practices preferred by other firms in 2004. However, a CC professional or analyst could use these results to realize that cluster No. 3 (Business and social capital in tech firms) shows similarities in supply chain approaches for tech firms. In both cases there may be a focus on providing skills to ensure not only appropriate technical and service levels, but also a preferred set of values to be upheld. This intersection of interests and approaches could have lead Microsoft’s and Intel’s CC teams to realize there was an opportunity to leverage common resources in support of *similar supply chain management practices*. In particular, this CC insight could lead to a strategic alliance for innovation, scale and cost reductions, as well as to risk mitigation through more resilient and efficient supply chains. This does not necessarily mean that Microsoft and Intel should have a shared or unique supply chain, which is probably very difficult given related legal issues involving licensing, intellectual property, patents, etc. But they could join forces, for example, to help educational institutions in China develop curricula to ensure desired technical and service levels, as well as the adoption of appropriate values. A similar CC insight could be obtained from cluster No. 6 (Governance and leadership), in which an alliance between Coca-Cola and McDonalds around the importance of governance issues (regarding compliance with guidelines and standards as well as conduct) for the core business of food and beverage firms could be developed for risk mitigation purposes.

2008 Cluster Analysis

We obtained three clusters for the 2008 document collection, which are shown in Figure 4 along with the Euclidean distance that helps differentiate clusters from one another. Table 11

provides descriptions for each of the clusters depicted in Figure 4, and also details the components grouped in each cluster.

2008 Document Groupings

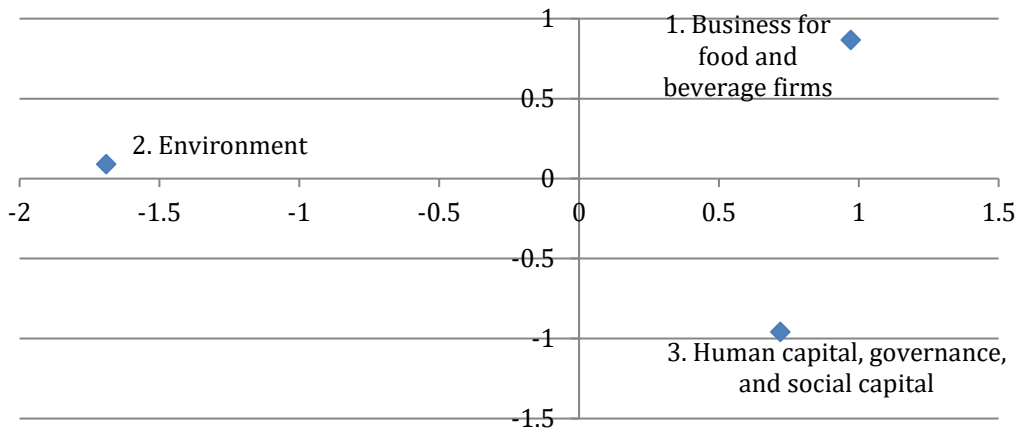


Figure 4. 2008 Text Clusters

Table 11. 2008 Cluster Descriptions

Cluster Number	Cluster Name	Description	Components Grouped
1	Business for food and beverage firms	Cluster points to <i>business for food and beverage firms</i> (e.g., children, brand, good conduct, values) as Coca-Cola’s business text file also groups here, these words are not only relevant for food firms—also in governance and social capital terms).	Coca-Cola2008business McDonalds2008business McDonalds2008governance McDonalds2008social capital
2	Environment	Cluster refers to <i>environmental</i> issues (e.g., waste, fuel, gas, energy) and this cluster also included ExxonMobil’s business text file, evidencing that in the oil extraction business these environmental terms are important.	Coca-Cola2008environment ExxonMobil2008business ExxonMobil2008environment Intel2008environment McDonalds2008environment
3	Human capital, governance, and social capital	Cluster refers to <i>human capital, governance, and social capital considerations</i> (e.g., training, plans, access, benefits).	Coca-Cola2008governance Coca-Cola2008human capital Coca-Cola2008social capital ExxonMobil2008governance ExxonMobil2008human capital ExxonMobil2008social capital Intel2008business Intel2008governance Intel2008human capital Intel2008social capital

			McDonalds2008human capital
--	--	--	----------------------------

ExxonMobil’s 2008 business component is in **bold** and *italics* because it was classified as an environmental component by supervised machine learning TDM. This component grouped with all other environmental components. In fact, it is the only business component in the environmental grouping, corroborating the fact that this business component (as determined by the consistency with which reports were divided) had environmental sustainability considerations that weigh heavily. Once again, Coca-Cola and McDonalds business components grouped together, while Intel’s business component grouped in the largest cluster No. 3 (Human capital, governance and social capital). Cluster No. 3 evidences similarities in the ways firms from different industries treated human and social capital issues as well as governance. In particular, all firms analyzed used terms such as plans, access, benefits, training, etc. This overlap in treatment of human capital, social capital, and governance CC issues may have been brought about by the economic factors contributing to the Great Recession. The downturn in the business cycle required firms to deal with laying off workers, hence it seems reasonable for them to have focused on describing access to benefits, as well as on *re-training programs*. Here, the CC insight, once again, suggests partnership opportunities between firms and educational institutions around re-training programs.

Another CC insight becomes apparent while examining the terms used in cluster No. 1 (Business for food and beverage firms), namely: good conduct, children, brand, and values. Both Coca-Cola and McDonalds had been advertising directly to children in spite of the fact that their products were not necessarily the best nutritional options for children. When the role of these firms in contributing to childhood obesity started to be observed and discussed, they responded by adopting new *ethical advertising practices*, revealed by this cluster. An alliance between the CC and Public Relations teams of these two companies to change that perception probably became apparent as they started to jointly sponsor sporting events (e.g., the Olympics, professional soccer games, etc.). Our data also point to that potential alliance, not necessarily in terms of just attempting to alter perceptions, but around the ability of these two firms to work together on *governance and compliance* issues related to not targeting children through advertising, educating consumers on the importance of considering the caloric content of their products or developing new healthier drink/menu options (initiatives that may have already transpired or could be underway).

2012 Cluster Analysis

Figure 5 shows the six clusters obtained for the 2012 document collection and how clusters differed from each other in Euclidean distance terms. Table 12 describes the each clusters depicted in the figure along with the components grouped in each one.

2012 Document Groupings

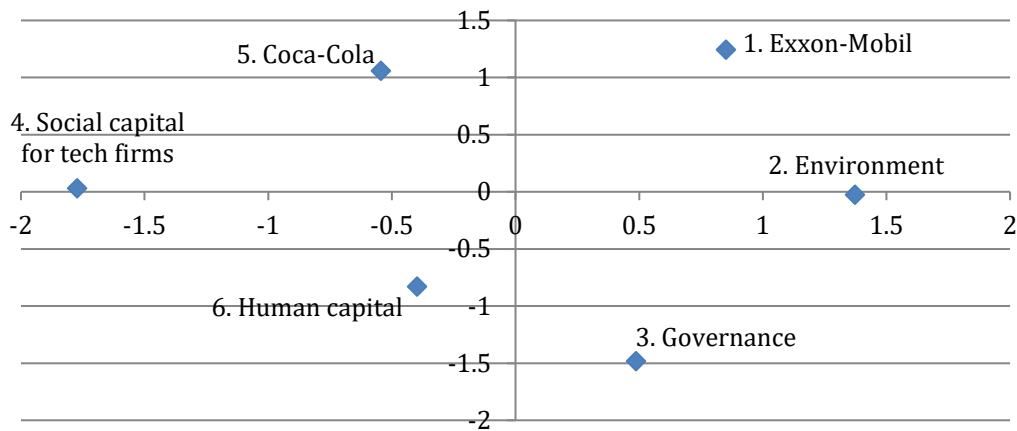


Figure 5. 2012 Text Clusters

Table 12. 2012 Cluster Descriptions

Cluster Number	Cluster Name	Description	Components Grouped
1	Exxon-Mobil	Cluster points to <i>ExxonMobil considerations</i> , this cluster groups all ExxonMobil components except governance and points to firm specific issues such as: upstream lines, gas, plans, projects, operations, conditions, international, design.	<i>ExxonMobil2012business</i> ExxonMobil2012environment ExxonMobil2012human capital ExxonMobil2012social capital
2	Environment	Cluster refers to <i>environmental</i> issues (e.g., air, carbon, power, waste, drive, water, cost, energy). It is important to note that General Motors and Intel business components also grouped here, indicating that the descriptive terms are increasingly relevant for these firms' business activities. This also happened to ExxonMobil in 2008.	Citi2012environment General Motors2012business General Motors2012environment Intel2012business Intel2012environment Microsoft2012environment
3	Governance	Cluster includes most <i>governance</i> components and mentions: conduct, directors, compensation, independent, legal, principles, review, compliance, risk, etc. Citi's business and social capital components also grouped here, evidencing that governance topics were a priority for Citi in 2012.	Citi2012business Citi2012governance Citi2012social capital Coca-Cola2012governance ExxonMobil2012governance General Motors2012governance Microsoft2012governance
4	Social capital for tech firms	Cluster brings up <i>social capital considerations for tech firms</i> insofar as it groups Microsoft and Intel components referring to supply chain issues (e.g., cash, central, campaign, audits, future, serve, conditions, china, centers, etc.).	Intel2012social capital Microsoft2012social capital
5	Coca-Cola	Cluster includes most of Coca-Cola's components (business, environment and social capital) – as well as General Motors social capital text file - and includes the following terms: “grant, Brazil, water, fund,	<i>Coca-Cola2012business</i> Coca-Cola2012environment

		partner, china, campaign, waste, partners, etc.” Thus, this Text Cluster raises <i>Coca-Cola considerations</i> .	Coca-Cola2012social capital General Motors2012social capital
6	Human capital	Finally, this cluster points to <i>human capital considerations</i> across industries through: “top, culture, women, workplace, safety, directors, people, human rights, etc.”	Citi2012human capital Coca-Cola2012human capital General Motors2012human capital Intel2012governance Intel2012human capital Microsoft2012business Microsoft2012human capital

The 2012 ExxonMobil’s and Coca-Cola’s business components (classified as environmental by supervised machine learning TDM) grouped with other firm-specific components. This indicates that these firms started to use overarching CC strategies with a clear company-guided conducting thread that differentiated them from other components in the analysis. Meanwhile, the business component for General Motors and Intel grouped with all other environmental components. This observation supports the idea that business components appear to be increasingly permeated by environmental sustainability considerations during the time period studied. These two components were not reclassified by the supervised machine learning TDM approach, or identified as potentials for overlapping classification, because they were part of the training set (e.g., they were used to train the supervised machine learning TDM algorithm). It is also because of this that these two components were not included in Table 8, which only lists components used in the validation set.

Cluster No. 4 (Social capital for tech firms) reiterates the CC insight identified in 2004 suggesting the possibility for an alliance between Microsoft and Intel to work together on building more *resilient and efficient supply chains* but this time through campaigns and audits (hopefully along with the capacity building programs proposed above). Finally, as in 2004 Microsoft’s human capital component grouped together with Microsoft’s business component, in 2012 Microsoft’s business component was the only one grouping with all other human capital components in the collection (see cluster No. 6). This corroborates findings reported by Parra et al. (2016a) where it was reported that after exploring term associations on collections by year, a consistent association between Microsoft’s business and human capital components was found. The terms characterizing this relationship were “software, donations, and teachers,” which begets considering Microsoft’s business strategy in human capital terms. The more software the firm donates to schools and teachers, the more captive users it will have in the future (Parra et al., 2016b).

Consolidated unsupervised findings and CC insights

Table 13 below shows the document groupings obtained using unsupervised machine learning on document collections per year. In 2004 there were six clusters: human capital and social capital cluster, environmental cluster, business and social capital for *tech firms*, business and governance for *hydrocarbon firms* (oil extraction and automotive), Citi cluster, and governance cluster. In 2008 three clusters were obtained: business for *food and beverage firms*, environmental cluster including one business component (ExxonMobil’s as classified by supervised machine learning), and a human capital, governance, and social capital cluster. In 2012, six clusters were obtained: ExxonMobil cluster, environmental cluster with business

components from General Motors and Intel, governance cluster with Citi’s business and social capital components, social capital for *tech firms* cluster, Coca-Cola cluster, and human capital cluster.

Table 13. Clusters Obtained in 2004, 2008 and 2012 with Machine Learning TDM

2004	2008	2012
Cluster No. 1. Human capital and social capital	Cluster No. 1. Business for food and beverage firms	Cluster No. 1. ExxonMobil
Cluster No. 2. Environment		Cluster No. 2. Environment
Cluster No. 3. Business and social capital for tech firms	Cluster No. 2. Environment	Cluster No. 3. Governance
Cluster No. 4. Business and governance for hydrocarbon firms		Cluster No. 4. Social capital for tech firms
Cluster No. 5. Citi	Cluster No. 3. Human capital, governance, and social capital	Cluster No. 5. Coca-Cola
Cluster No. 6. Governance		Cluster No. 6. Human Capital

In terms of CC insights, in 2004 CC professionals from Microsoft and Intel could have leveraged similarities in the way they approached business and social capital issues, specifically in relation to supply chains, and thus explored partnership opportunities around building more *resilient and efficient supply chains*. In particular, there was room for joint initiatives to partner with Chinese educational institutions to create programs ensuring desired technical and service levels. In fact, supply chain collaboration has been strongly advocated since the 1990’s (Holweg et al., 2005). Firms that strive to achieve greater supply chain collaboration to leverage the resources and knowledge of their suppliers and customers develop competitive advantages that enhance performance (Cao and Zhang, 2011). In addition, CC professionals from McDonalds and Coca-Cola could have explored partnering on *risk mitigation strategies focused on good governance for compliance* with food and beverage industry guidelines and standards. Corporate governance issues have traditionally focused on compliance aspects related to business ethics with legal implications that can also be understood as enablers of appropriate behavior, for example while managing product quality and transactions with integrity (Wieland, 2001).

In 2008, and considering the prevailing economic trends, firms from all industries could have worked together on providing *better benefits including re-training programs* (Edelman et al., 2011). Please note that the terms around which CC report components grouped (in cluster No. 3) could also refer to programs to attract and retain employees (Mayo, 2001). Once again, in 2008, CC professionals from McDonalds and Coca-Cola could have explored partnering on initiatives around educating children on nutrition and health as well as on offering *healthier drink/menu options*. This implies a values oriented approach to governance and business operations beyond compliance (Wieland, 2005). Finally, in 2012, CC professionals from Microsoft and Intel could have again used outcomes to explore partnership opportunities around building more resilient and efficient supply chains through campaigns and audits (as well as trainings, as noted while discussing 2004 outcomes).

Regarding CSR integration, in 2004, the environmental cluster did not include any business components. In 2008, it included one business component (that of ExxonMobil) as indicated by supervised machine learning findings. In 2012, the environmental grouping encompassed business components from General Motors as well as Intel. Thus, business

components in this document collection tended to increasingly group in environmental clusters through time, suggesting increasing levels of CSR integration in environmental sustainability terms for the firms considered. Finally, in 2004 and 2012, there were standalone governance clusters, but in 2008 governance components grouped in the human and social capital cluster, possibly because of the political and economic juncture that suggested heightened legal and regulatory scrutiny.

Conclusions

In this study, we attempted to showcase the potential of TDM techniques to help uncover patterns in large bodies of text and the possibility to streamline the production of CC insights. We did this in the context of exploring and analyzing the content of a collection of CC reports. In particular, we applied supervised machine learning and unsupervised machine learning techniques to CC reports produced by seven American firms. The analysis was done over three time periods (2004, 2008 and 2012) in order to characterize the way in which CC issues have been treated over time, as well as the way they grouped in order to uncover CC insights.

The outcome of our supervised machine learning TDM exercise exemplifies how CSR integration may occur in environmental sustainability terms (corroborating findings from studies that used different approaches), especially when considering the fact that ExxonMobil's core business relies on hydrocarbons and natural resource extraction, and Coca-Cola relies on water. Thus, it is intuitive that ExxonMobil and Coca-Cola include environmental sustainability considerations in their core business components. It could be argued that it also makes sense for all firms to highlight environmental sustainability considerations in their CC reports. However, this type of CSR integration around environmental sustainability did not happen for all firms considered, and was not consistently observed during the years considered by the firms in which it was found. Similarly, company specific effects, such as Citi's 2004 cluster, or ExxonMobil's or Coca-Cola's 2012 clusters, are more indicative of evolving and dynamic CC terms and topics used by firms to differentiate themselves than of industry specific jargon. This is the case because, once again, these effects did not happen for every firm considered and were not repeated or maintained throughout the years analyzed.

Core business descriptions, discussions, and prospects provided by firms in their CC reports may differ from those included in other public corporate manuscripts. It is beyond the scope of this paper to check whether there is a significant difference between core business descriptions included in CC reports and those included in other corporate documents (e.g. annual financial reports). This is an interesting question that merits a study of its own. Also, the labels assigned to CC report components by the subject matter expert were validated by supervised machine learning, except for a few overlapping classifications that could have also been caused by the limits of statistical content analysis of unstructured data in terms of context recognition (Lee et al., 2010).

Our analysis also helps characterize a practitioner-oriented and TDM-based perspective on the CSR integration discussion, specifically, in terms of environmental sustainability considerations. Unsupervised machine learning TDM performed on 2004, 2008, and 2012 components produced document groupings exhibiting a tendency for business components to increasingly group with environmental components through time for the firms considered. In

2004, only environmental components grouped in the environmental cluster, whereas in 2008 ExxonMobil's business component grouped within the environmental cluster, and in 2012 General Motors and Intel business components grouped with the environmental cluster, suggesting how environmental sustainability considerations have increasingly permeated core business components.

The automated method outlined for exploring the content of a document collection may help analysts find patterns and identify similarities/differences in CC reports or other publicly available corporate manuscripts. The main advantages of the method used here revolve around efficiency (in terms of resource utilization, which cannot be overemphasized in light of larger amounts of documents and information available) as well as effectiveness. Outcomes were corroborated by theoretical propositions from the CC literature in relation to CSR integration. In addition, the actual launch of joint CC initiatives by two of the firms analyzed (e.g., Coca-Cola and McDonalds working together in sponsoring sports events and offering healthier drink/menu options) corroborates the validity of the CC insights uncovered. Other CC insights produced here could be used to explore partnership opportunities, and to design alternative risk mitigation strategies. In particular, there is room for alliances between tech firms on building more *resilient and efficient supply chains*, between food and beverage firms around *governance and compliance*, and among firms from any and all industries on better *re-training programs*. Finally, we set out to extend the use of TDM in the CC field by analyzing the evolution of firms' treatment of CC issues using a novel method that could be followed by researchers and practitioners alike to uncover unintuitive relationships and gain CC insights.

Appendix A – Technical Details of Text Data Mining

In Text Data Mining (TDM) a document collection (or corpus) is transformed into a vector space model, which is reduced to obtain a numerical representation of the corpus (Salton et al., 1975). First, a term-by-document matrix (A) is built, which contains all the terms t in the document collection d , $A = t \times d$. This typically rectangular matrix (A) is reduced by removing terms that do not add value to the analysis being performed (e.g., using a Stop List), by stemming, and by using Zipf's law. Zipf's law ranks words (in a large body of text) in order of decreasing frequency, and plots a graph of the log of frequency against the log of rank to obtain a harmonic function. Then these terms are divided into equal intervals (Robertson, 2004). This helps quantify the importance of a term in a document collection by avoiding extremes (terms that appear too frequently as well as those that do not appear very often) and instead focusing on those terms in between as most likely to provide meaning to an analyst. This is done not only to reduce computational complexity, but also to reduce spurious language patterns (Evangelopoulos and Visinescu, 2012) and to minimize the degree to which the term space is distorted (Deerwester et al., 1990a).

Deerwester et al. developed a way to improve document similarity called Latent Semantic Indexing (LSI) (Deerwester et al., 1990a). LSI, which when applied becomes Latent Semantic Analysis (LSA), assumes a "latent" semantic structure to further reduce A 's dimensionality by producing a Singular Vector Decomposition (SVD) –a technique related to eigenvector decomposition and principal component factor analysis (Dumais, 2004). LSA is used to analyze large volumes of unstructured data (i.e., not presented in tables) including large document collections in order to extract key latent vectors of terms. LSA allows us to discover common themes across different documents and identify important terms that describe concepts or topics across documents (Konchady, 2006). LSA has been widely studied in the information retrieval literature to improve indexing and search query performance (Dumais, 2004, Dumais, 2007, Deerwester et al., 1990b). LSA does text quantification by developing a vector space model and obtaining SVDs from it.

After reducing the size of term-by-document matrix (A), each term in a document is assigned its frequency count, or term frequency ($tf_{t,d}$), which is simply a *local weight* that reflects the number of times term t appears in document d . This does not consider the order in which the words appear in the document and because of this it is typically referred to as a *bag of words*. To attenuate the effect of terms occurring too often the document frequency (df_t) is also considered, which reflects the number of documents that contain term t . Term weighting techniques provide a greater degree of discrimination among terms by adjusting local weights for document size and term distribution, thus distinguishing individual documents from a collection of documents (Salton and Buckley, 1988, Sparck Jones, 1974). Researchers tend to prefer having few documents that contain the term of interest (e.g., Corporate Citizenship) to get a higher relevance than many documents containing more common words (e.g., car). To achieve this, the Inverse Document Frequency (idf) of term t is used to assign a *global weight* represented by the formula below (Sparck Jones, 1972, Singhal et al., 1996). The idf of a rare term (low document frequency) would be high, whereas the idf of a frequent term (high document frequency) would be low.

$$idf_t = \log \frac{N}{df_t}$$

A widely used weighing technique is the Term Frequency-Inverse Document Frequencies (TF-IDF), which produces a composite weight for every term in a document that increases proportionally to the term frequency ($tf_{t,d}$) or number of times a word appears in a document, but is compensated by the frequency of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general (as stated by Zipf's Law).

$$tf - idf_{t,d} = tf_{t,d} \times idf_t$$

TF-IDF is commonly used to assign weights to longer documents while performing unsupervised machine learning in order to explore associations between terms or between documents. Before explaining how these associations are explored, the weighing scheme used for supervised learning purposes is presented below.

For supervised machine learning purposes, a more objective term weighting scheme is used, namely: mutual information $MI(t_1, t_2)$, which compares the joint probability of observing t_1 and t_2 together with the probabilities of observing t_1 and t_2 independently.

$$MI(t_1, t_2) = \log_2 \frac{P(t_1, t_2)}{P(t_1) \cdot P(t_2)}$$

The term probabilities $P(t_1)$ and $P(t_2)$ are estimated by counting the number of observations of t_1 and t_2 in the corpus and normalizing by the size of the corpus. If t_1 and t_2 are associated, $P(t_1, t_2) > P(t_1) P(t_2)$, and $MI(t_1, t_2) > 0$. If t_1 and t_2 are not associated, $P(t_1, t_2) = P(t_1) P(t_2)$, and $MI(t_1, t_2) = 0$ (Church and Hanks, 1990). MI measures the reduction in entropy that is achieved when one variable is conditioned on another one. Since MI does not take the term frequency into account, it is common to adjust the term frequency $tf_{t,d}$ with $MI(t_1, t_2)$ (Jing et al., 2002):

$$\omega_{t,d} = tf_{t,d} \cdot MI(t_1, t_2)$$

Independent of the weighing scheme utilized, LSA is used to obtain SVDs out of the reduced and transformed rectangular matrix (A), which is an extension of exploratory principal component factor analysis for rectangular matrices that decomposes variables (e.g. terms or documents) to obtain a set of vectors that represent the corpus.

SVDs include the *term eigenvectors* U , the *document eigenvectors* V , and the diagonal matrix of singular values Σ . The term T denotes transposition. The factor loadings obtained from transposing matrices $U\Sigma$ for terms and $V\Sigma$ for documents represent term clusters or document clusters, respectively (Evangelopoulos et al., 2012).

$$A = U\Sigma V^T$$

The document collection summarized in matrix (A) is represented by SVDs that capture the relative importance of terms in each document. Representing a document collection with vec-

tors allows researchers to perform operations such as scoring documents on a query, document classification, as well as document and term clustering (Manning et al., 2008). These SVDs can then be rotated to alternatively model the data's behavior and facilitate interpretation in an unsupervised setting as well as labeling in supervised approaches (Evangelopoulos et al., 2012, Sidorova et al., 2008, Evangelopoulos and Visinescu, 2012). Last, post-LSA may include comparing and classifying documents using either cosine similarity technique or by clustering or factor analysis. Evangelopoulos et al. (2012) makes some recommendation on LSA extension and argue that researchers should use clustering techniques such as K-means (Jain, 2010, Hartigan and Wong, 1979) or the expectation-maximization algorithm (Do and Batzoglou, 2008) for document summarization.

The SVD loadings represent the term loadings and/or document loadings (Evangelopoulos et al., 2012) depending on whether term clusters or document clusters are being explored. These components can be thought as artificial concepts. Each term or document is then characterized by a vector of weights indicating the strength of association with the underlying concepts and overcomes the problem of multiple terms referring to the same topic.

Appendix B

Sources for data (by company and year) presented in Table 2 (with basic firm demographics).

Company	2004	2008	2012
	References	References	References
Citi	Citigroup (2005)	Citigroup (2009)	Citigroup (2013)
Coca-Cola	The Coca Cola Company (2005)	The Coca Cola Company (2009)	The Coca Cola Company (2013)
ExxonMobil	ExxonMobil (2005)	ExxonMobil (2009)	ExxonMobil (2013)
General Motors	General Motors Corporation (2008)		General Motors Corporation (2012)
Intel	Intel (2005)	Intel (2009)	Intel (2013)
McDonalds	Corporation (2005)	Corporation (2009)	Corporation (2013)
Microsoft	Microsoft Corporation (2005)	Microsoft Corporation (2009)	Microsoft Corporation (2013)

References

- Altman, D. G., (1990). *Practical statistics for medical research*, CRC press.
- Barkemeyer, R., Comyns, B., Figge, F. & Napolitano, G., (2014), "CEO statements in sustainability reports: substantive information or background noise?", *Accounting Forum*, 241-257.
- Barkemeyer, R., Figge, F., Holt, D. & Wettstein, B., (2009), "What the papers say: trends in sustainability. A comparative analysis of 115 leading national newspapers worldwide", *Journal of Corporate Citizenship*, 2009, 68-86.
- Bolton, R. J. & Hand, D. J., (2002), "Statistical fraud detection: A review", *Statistical science*, 235-249.
- Cao, M. & Zhang, Q., (2011), "Supply chain collaboration: Impact on collaborative advantage and firm performance", *Journal of Operations Management*, 29, 163-180.
- Carroll, A. B., (1979), "A Three-Dimensional Conceptual Model of Corporate Performance", *The Academy of Management Review*, 4, 497-505.
- Church, K. W. & Hanks, P., (1990), "Word association norms, mutual information, and lexicography", *Computational linguistics*, 16, 22-29.
- Citigroup. 2005. *Citigroup: Annual Report 2004* [Online]. Author. Available: http://www.citigroup.com/citi/investor/quarterly/2006/ar051c_en.pdf [Accessed January 14 2015].
- Citigroup. 2009. *Citigroup: Annual Report 2008* [Online]. Author. Available: http://www.citigroup.com/citi/investor/quarterly/2009/ar08c_en.pdf?ieNocache=319 [Accessed January 14 2015].
- Citigroup. 2013. *Citigroup: Annual Report 2012* [Online]. Author. Available: http://www.citigroup.com/citi/investor/quarterly/2013/ar12c_en.pdf?ieNocache=319 [Accessed January 14 2015].
- Corporation, M. s. 2005. *McDonald's Corporation: Corporate Responsibility Report 2004* [Online]. Available: https://www.aboutmcdonalds.com/content/dam/AboutMcDonalds/Sustainability/Sustainability_Library/2004_Report_%28English%29.pdf [Accessed January 14 2015].
- Corporation, M. s. 2009. *McDonald's Corporation: 2008 Annual Report* [Online]. Available: <http://www.aboutmcdonalds.com/content/dam/AboutMcDonalds/Investors/C-%5Cfakepath%5Cinvestors-2008-annual-report.pdf> [Accessed January 14 2015].
- Corporation, M. s. 2013. *McDonald's Corporation: 2012 Annual Report* [Online]. Available: http://www.aboutmcdonalds.com/content/dam/AboutMcDonalds/Investors/Investor_2013/2012_Annual_Report_Final.pdf [Accessed January 14 2015].
- Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. & Harshman, R., (1990a), "Indexing by latent semantic analysis", *Journal of the American society for information science*, 41, 391.
- Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W. & Harshman, R. A., (1990b), "Indexing by latent semantic analysis", *JASIS*, 41, 391-407.
- Dempster, A. P., Laird, N. M. & Rubin, D. B., (1977), "Maximum likelihood from incomplete data via the EM algorithm", *Journal of the royal statistical society. Series B (methodological)*, 1-38.
- Dickinson, S. J., Gill, D. L., Purushothaman, M. & Scharl, A., (2008), "A web analysis of sustainability reporting: an oil and gas perspective", *Journal of Website Promotion*, 3, 161-182.
- Direction, S., (2003), "Corporate socialism unethically masquerades as "CSR": The difference between being ethical, altruistic and strategic in business", *Strategic Direction*, 19.
- Do, C. B. & Batzoglou, S., (2008), "What is the expectation maximization algorithm?", *Nature biotechnology*, 26, 897-899.
- Dörre, J., Gerstl, P. & Seiffert, R. "Text mining: finding nuggets in mountains of textual data", Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, 1999. ACM, 398-401.

- Dumais, S., Platt, J., Heckerman, D. & Sahami, M. (1998), "Inductive learning algorithms and representations for text categorization", Proceedings of the seventh international conference on Information and knowledge management, 1998. ACM, 148-155.
- Dumais, S. T., (2004), "Latent semantic analysis", *Annual review of information science and technology*, 38, 188-230.
- Dumais, S. T., (2007), "LSA and information retrieval: Getting back to basics", *Handbook of latent semantic analysis*, 293-321.
- Economist, T. 2016. The return of the machinery question. *The Economist*. Special Report ed.
- Edelman, P., Holzer, H., Seleznow, E., Van Kleunen, A. & Watson, E., (2011), "State workforce policy: Recent innovations and an uncertain future", *Georgetown Center on Poverty, Inequality, and Public Policy, Washington, DC*.
- Evangelopoulos, N. & Visinescu, L., (2012), "Text-mining the voice of the people", *Communications of the ACM*, 55, 62-69.
- Evangelopoulos, N., Zhang, X. & Prybutok, V. R., (2012), "Latent semantic analysis: five methodological recommendations", *European Journal of Information Systems*, 21, 70-86.
- ExxonMobil. 2005. *ExxonMobil: 2004 Summary Annual Report* [Online]. Available: http://cdn.exxonmobil.com/~media/Reports/Summary_Report/2004/AR_2004.pdf [Accessed January 14 2015].
- ExxonMobil. 2009. *ExxonMobil: 2008 Summary Annual Report* [Online]. Available: http://cdn.exxonmobil.com/~media/Reports/Summary_Report/2008/news_pub_sar_2008.pdf [Accessed January 14 2015].
- ExxonMobil. 2013. *ExxonMobil: 2012 Summary Annual Report* [Online]. Available: http://cdn.exxonmobil.com/~media/Reports/Financial_Review/2012/news_pub_fo_2012.pdf [Accessed January 14 2015].
- Fawcett, T. E. & Provost, F. "Fraud detection", *Handbook of data mining and knowledge discovery*, 2002. Oxford University Press, Inc., 726-731.
- Feldman, R. & Dagan, I. "Knowledge Discovery in Textual Databases (KDT)", *KDD*, 1995. 112-117.
- Feldman, R. & Sanger, J., (2007). *The text mining handbook: advanced approaches in analyzing unstructured data*, Cambridge University Press.
- Freeman, R. E., (2010). *Strategic management: A stakeholder approach*, Cambridge University Press.
- Friedman, M. (2007), "The Social Responsibility of Business Is to Increase Its Profits", In: ZIMMERLI, W., HOLZINGER, M. & RICHTER, K. (eds.), *Corporate Ethics and Corporate Governance*, Springer Berlin Heidelberg.
- General Motors Corporation. 2008. *General Motors Corporation: Annual Report 2008 (Form 10K)* [Online]. Available: <http://www.sec.gov/Archives/edgar/data/40730/000119312509045144/d10k.htm> [Accessed January 14 2015].
- General Motors Corporation. 2012. *General Motors Corporation: Annual Report 2012 (Form 10K)* [Online]. Available: http://www.gm.com/content/dam/gmcom/COMPANY/Investors/Stockholder_Information/PDFs/2012_GM_Annual_Report.pdf [Accessed January 14 2015].
- Gill, D. L., Dickinson, S. J. & Scharl, A., (2008), "Communicating sustainability: a web content analysis of North American, Asian and European firms", *Journal of Communication Management*, 12, 243-262.
- Grimes, S. 2008. Unstructured Data and the 80 Percent Rule. Available from: <http://breakthroughanalysis.com/2008/08/01/unstructured-data-and-the-80-percent-rule/> 2014].
- Hartigan, J. A. & Wong, M. A., (1979), "Algorithm AS 136: A k-means clustering algorithm", *Applied statistics*, 100-108.
- Holweg, M., Disney, S., Holmström, J. & Småros, J., (2005), "Supply Chain Collaboration: Making Sense of the Strategy Continuum", *European management journal*, 23, 170-181.
- Hsieh, H.-F. & Shannon, S. E., (2005), "Three approaches to qualitative content analysis", *Qualitative Health Research*, 15, 1277-1288.

- Ihlen, Ø. & Roper, J., (2011), "Corporate reports on sustainability and sustainable development: 'we have arrived'", *Sust. Dev.*, 22, 42-51.
- Intel. 2005. *Intel: 2004 Annual Report* [Online]. Available: <http://www.intel.com/content/dam/doc/report/history-2004-annual-report.pdf> [Accessed January 14 2015].
- Intel. 2009. *Intel: 2008 Annual Report* [Online]. Available: <http://www.intel.com/content/dam/doc/report/history-2008-annual-report.pdf> [Accessed January 14 2015].
- Intel. 2013. *Intel: 2012 Annual Report* [Online]. Available: http://www.intc.com/intel-annual-report/2012/static/pdfs/Intel_2012_Annual_Report_and_Form_10-K.pdf [Accessed January 14 2015].
- Jain, A. K., (2010), "Data clustering: 50 years beyond K-means", *Pattern recognition letters*, 31, 651-666.
- Jing, L.-P., Huang, H.-K. & Shi, H.-B. "Improved feature selection approach TFIDF in text mining", *Machine Learning and Cybernetics*, 2002. Proceedings. 2002 International Conference on, 2002. IEEE, 944-946.
- Juola, P., (2006), "Authorship attribution", *Foundations and Trends in information Retrieval*, 1, 233-334.
- Kao, A. & Potet, S. R., (2007). *Natural Language Processing and Text Mining*, Springer Science and Business Media, London.
- Konchady, M., (2006). *Text Mining Application Programming (Programming Series)*, Charles River Media, Inc.
- Landis, J. R. & Koch, G. G., (1977), "The measurement of observer agreement for categorical data", *biometrics*, 159-174.
- Laver, M., Benoit, K. & Garry, J., (2003), "Extracting Policy Positions from Political Texts Using Words as Data", *American Political Science Review*, 97, 311-331.
- Lee, S., Baker, J., Song, J. & Wetherbe, J. C. "An empirical comparison of four text mining methods", *System Sciences (HICSS)*, 2010 43rd Hawaii International Conference on, 2010. IEEE, 1-10.
- Manning, C. D., Raghavan, P., Sch\, H., \#252 & tze, (2008). *Introduction to Information Retrieval*, Cambridge University Press.
- Marcoux, A. M., (2000), "Business ethics gone wrong", *Cato Policy Report*.
- Masand, B., Linoff, G. & Waltz, D. 1992. Classifying news stories using memory based reasoning. *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*. Copenhagen, Denmark: ACM.
- Mayo, A., (2001). *Human Value of the Enterprise*, Nicholas Brealey Publishing London.
- McWilliams, A., Siegel, D. S. & Wright, P. M., (2006), "Corporate Social Responsibility: Strategic Implications", *Journal of Management Studies*, 43, 1-18.
- Meyskens, M. & Paul, K., (2010), "The evolution of corporate social reporting practices in Mexico", *Journal of Business Ethics*, 91, 211-227.
- Microsoft Corporation. 2005. *Microsoft Corporation: 2004 Annual Report (Form 10K)* [Online]. Available: http://www.microsoft.com/investor/reports/ar04/nonflash/10k_dl_main.html [Accessed January 14 2015].
- Microsoft Corporation. 2009. *Microsoft Corporation: 2008 Annual Report (Form 10K)* [Online]. Available: http://www.microsoft.com/investor/reports/ar08/10k_dl_dow.html [Accessed January 14 2015].
- Microsoft Corporation. 2013. *Microsoft Corporation: 2012 Annual Report (Form 10K)* [Online]. Available: <http://www.microsoft.com/investor/reports/ar12/index.html> [Accessed January 14 2015].
- Moreno, A. & Capriotti, P., (2009), "Communicating CSR, citizenship and sustainability on the web", *Journal of Communication Management*, 13, 157-175.
- Parra, C. M., (2008), "Quality of Life Markets: Capabilities and Corporate Social Responsibility", *Journal of Human Development*, 9, 207-227.

- Parra, C. M., Tremblay, M. & Castellanos, A., (2016a), "Visualizing Term Eigenvector Prominence in a Corporate Social Responsibility Context", *JOURNAL ON ADVANCES IN THEORETICAL AND APPLIED INFORMATICS*, 2, 31-37.
- Parra, C. M., Tremblay, M. C. & Castellanos, A. "Prominent voices and prevalent discourses: A corporate social responsibility application", Eleventh International Conference on Digital Information Management (ICDIM), 2016b. IEEE, 74-78.
- Paul, K., (2008), "Corporate sustainability, citizenship and social responsibility reporting", *Journal of Corporate Citizenship*, 2008, 63-78.
- Porter, M. E. & Kramer, M. R., (2006), "The Link Between Competitive Advantage and Corporate Social Responsibility", *Harvard Business Review*, 84, 78-92.
- Reich, R. B., (1998), "The New Meaning of Corporate Social Responsibility", *California Management Review*, 40, 8-17.
- Robertson, S., (2004), "Understanding inverse document frequency: on theoretical arguments for IDF", *Journal of documentation*, 60, 503-520.
- Saito, M., Tang, Q. & Umemuro, H., (2012), "Text-Mining Approach for Evaluation of Affective Management Practices", *World Academy of Science, Engineering and Technology*, 72, 129-136.
- Salton, G. & Buckley, C., (1988), "Term-weighting approaches in automatic text retrieval", *Information processing & management*, 24, 513-523.
- Salton, G., Wong, A. & Yang, C.-S., (1975), "A vector space model for automatic indexing", *Communications of the ACM*, 18, 613-620.
- Sasaki, M. & Shinnou, H. "Spam detection using text clustering", 2005 International Conference on Cyberworlds (CW'05), 2005. IEEE, 4 pp.-319.
- Sidorova, A., Evangelopoulos, N., Valacich, J. S. & Ramakrishnan, T., (2008), "Uncovering the intellectual core of the information systems discipline", *Mis Quarterly*, 467-482.
- Singhal, A., Buckley, C. & Mitra, M. "Pivoted document length normalization", Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, 1996. ACM, 21-29.
- Sparck Jones, K., (1972), "A statistical interpretation of term specificity and its application in retrieval", *Journal of documentation*, 28, 11-21.
- Sparck Jones, K., (1974), "Automatic indexing", *Journal of documentation*, 30, 393-432.
- Standage, T. 2016. The return of the machinery question. *The Economist*.
- Tan, A.-H. "Text mining: The state of the art and the challenges", Proceedings of the PAKDD 1999 Workshop on Knowledge Discovery from Advanced Databases, 1999a. 65-70.
- Tan, A.-H. "Text mining: The state of the art and the challenges", Proceedings of the PAKDD 1999 Workshop on Knowledge Discovery from Advanced Databases, 1999b. 65-70.
- Tang, L., (2012), "Media discourse of corporate social responsibility in China: a content analysis of newspapers", *Asian Journal of Communication*, 22, 270-288.
- The Coca Cola Company. 2005. *The Coca Cola Company: Annual Report 2004 (Form 10K)* [Online]. Available: http://assets.coca-colacompany.com/0a/f8/95718fb545be878541ed90c8d1c4/form_10K_2004.pdf [Accessed].
- The Coca Cola Company. 2009. *The Coca Cola Company: Annual Report 2008 (Form 10K)* [Online]. Available: http://assets.coca-colacompany.com/21/8f/2ff9edb24bad8d6cc09264247ce4/form_10K_2008.pdf [Accessed January 14 2015].
- The Coca Cola Company. 2013. *The Coca Cola Company: Annual Report 2012 (Form 10K)* [Online]. Available: <http://assets.coca-colacompany.com/c4/28/d86e73434193975a768f3500ffae/2012-annual-report-on-form-10-k.pdf> [Accessed January 14 2015].
- Tremblay, M. C., Berndt, D. J., Luther, S. L., Foulis, P. R. & French, D. D., (2009), "Identifying fall-related injuries: Text mining the electronic medical record", *Information Technology & Management*, 10, 253-265.
- Wieland, J., (2001), "The ethics of governance", *Business Ethics Quarterly*, 11, 73-87.

- Wieland, J., (2005), "Corporate governance, values management, and standards: a European perspective", *Business & Society*, 44, 74-93.
- Woodfield, T., (2011). *Text Analytics Using SAS Text Miner Course Notes*, Cary, North Carolina.
- Yadava, R. N. & Sinha, B., (2015), "Scoring Sustainability Reports Using GRI 2011 Guidelines for Assessing Environmental, Economic, and Social Dimensions of Leading Public and Private Indian Companies", *Journal of Business Ethics*, 1-10.
- Zhang, S., Wang, W., Ford, J., Makedon, F. & Pearlman, J. "Using singular value decomposition approximation for collaborative filtering", *E-Commerce Technology, 2005. CEC 2005. Seventh IEEE International Conference on, 2005. IEEE*, 257-264.

Author Biographies

Dr. Carlos Parra has designed and executed corporate sustainability and business development strategies as well as overseen the continuous improvement of processes and metrics in the financial and manufacturing industries

Dr. Monica Tremblay has both research and consulting experience in the areas of electronic health records, health information exchanges, and data analytics

Dr. Karen Paul has served as the editor of several books and has published several articles on Business Ethics and Corporate Social Responsibility

Mr. Arturo Castellanos has a background in engineering and has a PhD in Business. His research interests include analytics, conceptual modeling, and data quality