JADI – Brazil – v. 2 n.2 – 2016

Visualizing Term Eigenvector Prominence in a Corporate Social Responsibility Context

Carlos M Parra
Florida International University
United States of America

Monica Tremblay
Florida International University
United States of America

Arturo Castellanos
City University of New York
United States of America

Abstract— **In this study we develop a simplified technique for helping researchers and analysts visualize the alternative prominence of term eigenvectors obtained after exploring term associations (Term Clusters) while conducting Text Data Mining on a collection to Corporate Social Responsibility (CSR) reports. The collection analyzed is comprised of CSR reports produced by 7 US firms (Citi, Coca-Cola, Exxon-Mobil, General Motors, Intel, McDonald's and Microsoft) in 2004, 2008 and 2012. The analysis is performed by year in order to discern how the prominence of term eigenvectors has evolved for each firm and for different CSR topics. Results indicate that term eigenvectors maintain their prominence when CSR topics are related to the core business of the firm in question.**

Index Terms—**Clustering Methods, Text Mining, Corporate Social Responsibility**

I. Introduction

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 [1]. First, a term-by-document matrix (*A*) is built, which contains all the terms *t* in the document collection *d*, *A = t x 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 [2]. 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.

The creation of the stop list is iterative, and is conducted on the corpus (or document collection) by parsing it and filtering it until topics are obtained that exclude non-value added terms. For this study the following terms were included in the Stop List: McDonald, Roland, Restaurant, Intel, Microprocessor, Processor, Wafer, Chipset, Exxonmobil, oil, Coca-cola, Beverage, Citi, Citigroup, Citibank, Bank, GM, Opel, Saab, Automotive, Vauxhall, Microsoft, Windows, Outlook and Bottler. Again, TDM algorithms need Stop Lists to prevent SVDs from considering terms that do not add value. This is done not only to reduce computational complexity, but also to reduce spurious language patterns [3] and to minimize the degree to which the term space is distorted [4].

Deerwester et al. developed a way to improve document similarity called Latent Semantic Indexing (LSI) [4]. 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 [5]. 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 [6]. LSA has been widely studied in the information retrieval literature to improve indexing and search query performance [5, 7, 8]. 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 weighing 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 [9, 10]. Researchers tend to prefer having few documents that contain the term of interest (e.g., Corporate Social Responsibility - CSR) 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 [11, 12]. 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. 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 [13].

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

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 vectors allows researchers to perform operations such as scoring documents on a query, document classification, as well as document and term clustering [14]. 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 [3, 13, 15]. Last, post-LSA may include comparing and classifying documents using either cosine similarity technique or by clustering or factor analysis. Evangelopoulos, Zhang and Prybutok [13] makes some recommendation on LSA extension and argue that researchers should use clustering techniques such as K-means [16, 17] or the expectation-maximization algorithm [18] for document summarization.

The SVD loadings represent the term loadings and/or document loadings [13] depending on whether term clusters or document clusters are being explored. We then applied traditional centroid clustering on the SVDs (term eigenvectors) obtained while exploring *term associations* (or, Term Clusters) and mapped them using radar graphs to identify prominent term eigenvectors around specific CSR topics. Please note that our intention is not to propose an alternative method for identifying topics in natural discourse, but rather to use existing methods for doing this along with centroid clustering to identify prominent term eigenvectors around different CSR topics. In essence, our aim is to help visualize and understand more of the information that SVDs can convey.

To do this, we will use the expression *Data Cluster analysis* when referring to the application of traditional centroid cluster analysis on data, in an effort to differentiate the procedure from Term Cluster analysis described above. Data Cluster analysis allow us to identify prominent term eigenvectors (or centroid-guiding-SVDs) by year, which in turn allows us to discern *the firm with the most prominent voice around a specific CSR topic*. This matters as firms see business value in determining whether or not the CSR topics they discuss in their reports have more or less prominence compared to other firms as well as to other CSR topics discussed, through time. Firms may also see value in determining whether they were able to maintain prominence around CSR issues at different points in time. For each of the time periods (2004, 2008, and 2012) we visualize (using radar graphs) and discuss the characteristics of every Data Cluster obtained.

In this study we are interested visualizing term clusters or term associations obtained after analyzing a collection of CSR reports by year and in applying centroid clustering to the SVDs obtained while exploring these term associations. In order to demonstrate how this can be done, the next section will describe our methodology, we will then present our results, and finalize with conclusions and recommendations.

II. METHODOLOGY

In order to build a context in which to apply our technique, we obtained text by downloading complete CSR reports in PDF format from the corresponding official corporate websites. In total we download 20 reports from 2004, 2008 and 2012. These reports were then manually scrubbed in order to obtain the main text.

We then divided the main text of each report into 5 components associated to 5 different CSR dimensions using the Sustainability Accounting Standards Board (SASB's) framework [19]. This framework contains 5 CSR dimensions: business and innovation, governance, environment, human capital and social capital. We did this by asking a subject matter expert to divide the main text of each report into the 5 sustainability components recommended by SASB (a business component, a governance component, an environmental component, a human capital component, and a social capital component). In this study, this is analogous to asking the subject matter experts to label different report sections or components according to a predefined taxonomy (i.e., SASB's guidelines). Not all of the 20 reports downloaded were used in the analysis. Citi's 2008 CSR report had security settings that did not allow us to obtain main text, while Microsoft's 2008 and McDonalds' 2012 CSR reports were too short to obtain meaningful components. Thus, out of the 17 reports used in the analysis a total of 85 components were obtained - 35 components from 7 reports for 2004; 20 components from 4 reports for 2008; and 30 components from 6 reports for 2012.

Unsupervised TDM was applied to each year separately in order to perform a longitudinal analysis of term eigenvector evolution over the three time periods. For each of the time periods (2004, 2008, and 2012) we show the document groupings and a table that includes the cluster name, the descriptive terms of the cluster, and which documents fall under each of the clusters [20]. In addition, for each of the time periods, we explore SVDs generated by Text Topic analysis through traditional Data Clustering in order to find prominent term eigenvectors or (centroid guiding components). That is, the firms that advanced a particular CSR topic more prominently than others in a specific year as well as the terms used to do so.

A. 2004 Results

For 2004, we obtained three Data Clusters, and plotted the Euclidian distances from the centroid of the components grouped in each one using radar graphs (see Figure 1 for the components in Data Cluster 1). There are two components closer to the centroid and guiding Data Cluster 1, namely: General Motors business (gm2004biz) component and General Motors environmental (gm2004env) component. So General Motors was the prominent voice for the first Data Cluster.

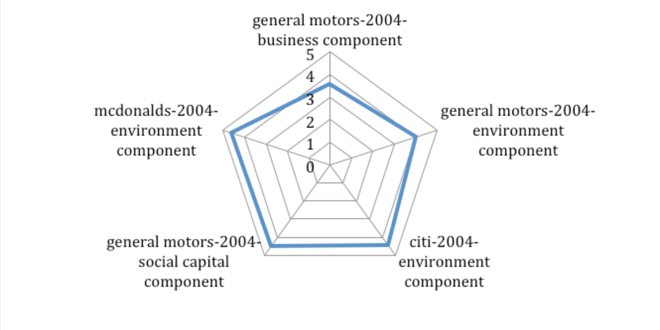


Fig. 1. Prominent Term Eigenvectors for Data Cluster 1 – 2004 CSR Reports

Table 1 shows the SVDs (term cluster) loadings of the 2004 CSR report components closest to the centroid and highlights the topics with the highest loadings in order to characterize the CSR issue in question.

TABLE I.
Prominent term eigenvectors loadings for Data Cluster 1 – 2004 reports

Topic NAME	Term cluster weights for general motors-2004-business component	Term cluster weights for general motors-2004-environment component
+emission,+gas,ghg,+water,+waste	**0.67**	**1.172**
governance,+committee,conduct,+code,+board	0.372	0.348
+human right,discrimination,safety,+injury,hiv	**0.512**	0.352
animal,welfare,food,+operator,+chain	0.39	0.341
ing,credit,+housing,+client,nonprofit	0.366	0.333
+fiscal year,ict,fiscal,internet,+server	0.35	0.383
+motor,privacy,aids,iso,+vehicle	**0.437**	**0.428**
nutrition,+lifestyle,physical,fitness,food	0.418	0.321
+bottle,aids,hiv,+partner,+water	0.345	0.349
+fiscal year,ict,fiscal,software,+computer	0.379	0.241
wireless,+computer,+pc,+teacher,education	0.423	0.344
+diesel,fish,+package,+source,+saving	0.407	**0.662**
+vehicle,+motor,gri,economy,sullivan	**0.604**	**0.593**
+pipeline,+land,indigenous,construction,+farm	0.381	0.418
+operator,+disability,+owner,+wage,inclusion	0.32	0.291
ireland,ehs,israel,retention,epa	0.354	0.365
+transaction,principles,+database,forestry,tions	0.361	0.348
hydrogen,+fuel,+vehicle,mobility,+dividend	**0.66**	0.356
medical,on-site,+hire,+career,family	0.28	0.23

The topic with the highest load for General Motors 2004 business component (0.67) also has the highest loading for General Motors 2004 environment component (1.172) and thus the most prominent term eigenvector for this Data Cluster alluded to: "emissions, Green House Gasses (GHG), water, waste." Because of this, Data Cluster 1 refers to environmental considerations and thus in 2004's Text universe environmental issues were guided by General Motors. Please note, other topics also have high loadings in both business and environmental components, and this indicates integration of environmental considerations in General Motors' business component. Citi's and McDonald's environmental contributions were also grouped in this first Data Cluster but their term eigenvectors were not as prominent.

The second Data Cluster obtained had three components closer to the centroid and thus guiding Data Cluster 2, namely: General Motors human capital component, Microsoft social capital component and Intel social capital component (see Figure 2 for the components in Data Cluster 2).

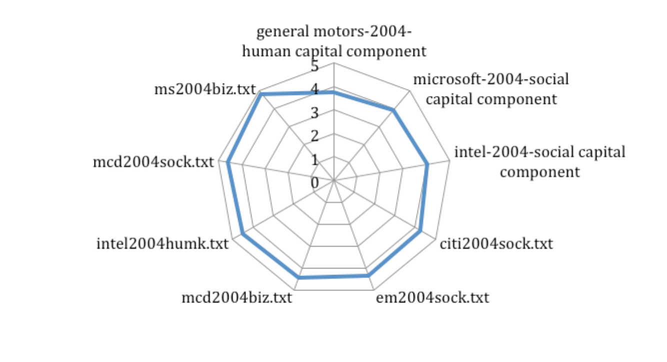


Fig 2. Prominent Term Eigenvectors for Data Cluster 2 – 2004 CSR Reports

From Table 2, the topic with the highest loading for General Motors 2004 human capital component (0.922) is combined with terms in the topic with the highest loading for Microsoft 2004 social capital component (0.918) and also with the ones in the topic with the highest loading for Intel 2004 social capital component (0.891) to establish that in 2004's Text Universe human capital issues were guided by General Motors, Microsoft and Intel and the terms used alluded to: "human rights, discrimination, safety, injury, HIV, Information and Communications Technology (ICT), software, computer, wireless, teacher, education."

TABLE II.
Prominent term eigenvectors loadings for Data Cluster 2 – 2004 reports

Topic NAME	Term cluster weights general motors-2004-human capital component	Term cluster weights microsoft-2004-social capital component	Term cluster weights intel-2004-social capital component
+emission,+gas,ghg,+water,+waste	0.291	0.175	0.289
governance,+committee,conduct,+code,+board	0.35	0.273	0.224
+human right,discrimination,safety,+injury,hiv	**0.922**	0.274	0.234
animal,welfare,food,+operator,+chain	0.327	0.272	0.211
ing,credit,+housing,+client,nonprofit	0.286	0.374	0.396
+fiscal year,ict,fiscal,internet,+server	0.232	**0.401**	0.248
+motor,privacy,aids,iso,+vehicle	0.323	0.29	0.27
nutrition,+lifestyle,physical,fitness,food	0.355	0.182	0.192
+bottle,aids,hiv,+partner,+water	0.288	0.302	0.242
+fiscal year,ict,fiscal,software,+computer	0.321	**0.918**	0.362
wireless,+computer,+pc,+teacher,education	0.232	**0.413**	**0.891**
+diesel,fish,+package,+source,+saving	0.168	0.121	0.146
+vehicle,+motor,gri,economy,sullivan	0.219	0.08	0.114
+pipeline,+land,indigenous,construction,+farm	0.387	0.359	0.32
+operator,+disability,+owner,+wage,inclusion	**0.387**	0.29	0.252
ireland,ehs,israel,retention,epa	**0.444**	0.23	0.265
+transaction,principles,+database,forestry,tions	0.259	0.163	0.175
hydrogen,+fuel,+vehicle,mobility,+dividend	0.162	0.119	0.115
medical,on-site,+hire,+career,family	0.301	0.271	0.245

The third Data Cluster obtained with 2004 data had only one component closest to the centroid and guiding Data Cluster 3, namely: Intel governance component (see Figure 3 for the components in Data Cluster 3).

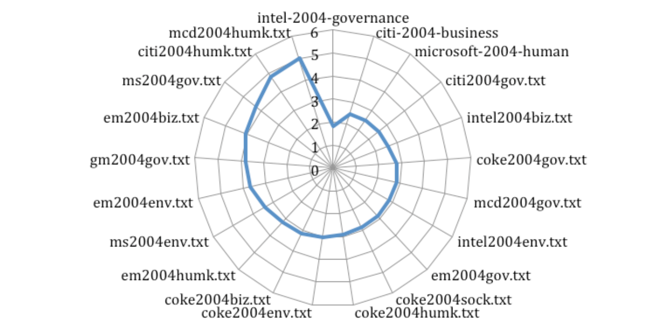
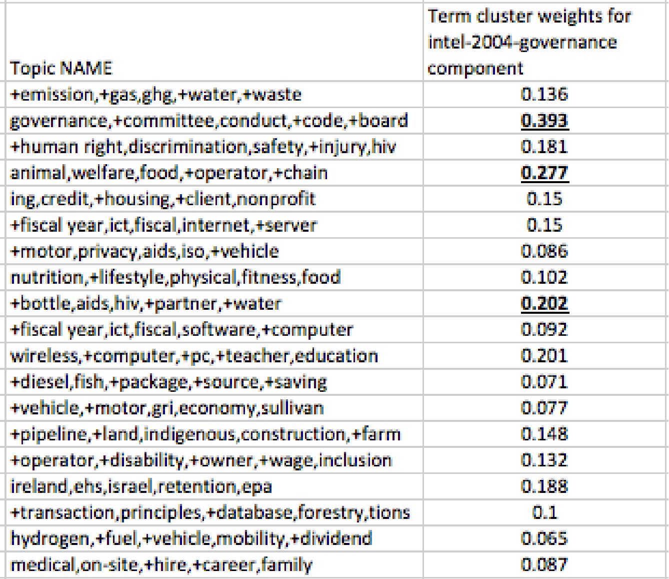


Fig 3. Prominent Term Eigenvectors for Data Cluster 3 – 2004 CSR Reports

From Table 3, the topic with the highest loading for Intel 2004 governance component (0.393) helped us determine that in 2004's Text Universe governance issues were guided by Intel by referring to: "governance, committee, conduct, code, board."

TABLE III.
Prominent term eigenvectors loadings for Data Cluster 3 – 2004 reports



Topic NAME	Term cluster weights for intel-2004-governance component
+emission,+gas,ghg,+water,+waste	0.136
governance,+committee,conduct,+code,+board	**0.393**
+human right,discrimination,safety,+injury,hiv	0.181
animal,welfare,food,+operator,+chain	**0.277**
ing,credit,+housing,+client,nonprofit	0.15
+fiscal year,ict,fiscal,internet,+server	0.15
+motor,privacy,aids,iso,+vehicle	0.086
nutrition,+lifestyle,physical,fitness,food	0.102
+bottle,aids,hiv,+partner,+water	**0.202**
+fiscal year,ict,fiscal,software,+computer	0.092
wireless,+computer,+pc,+teacher,education	0.201
+diesel,fish,+package,+source,+saving	0.071
+vehicle,+motor,gri,economy,sullivan	0.077
+pipeline,+land,indigenous,construction,+farm	0.148
+operator,+disability,+owner,+wage,inclusion	0.132
ireland,ehs,israel,retention,epa	0.188
+transaction,principles,+database,forestry,tions	0.1
hydrogen,+fuel,+vehicle,mobility,+dividend	0.065
medical,on-site,+hire,+career,family	0.087

B. 2008 Results

While identifying prominent term eigenvectors (or centroid-guiding components) for 2008, three Data Clusters were obtained and figure 4 below depicts a radar graph of the Euclidian distances from the centroid of the components grouped in the first Data Cluster (see Figure 4 for the components in Data Cluster 1 for 2008).

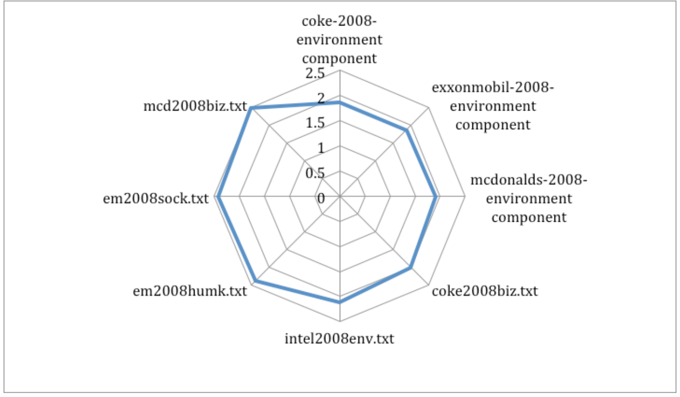
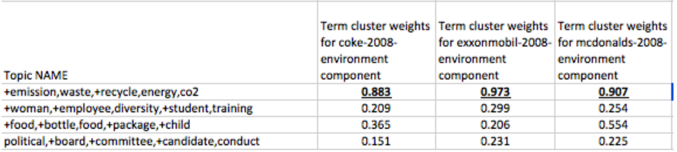


Fig 4. Prominent Term Eigenvectors for Data Cluster 1 – 2008 CSR Reports

There are three components closest to the centroid and guiding Data Cluster 1: Coca Cola's, ExxonMobil's and McDonalds' environmental components. Because of this, Data Cluster 1 refers to environmental considerations and in 2008's Text universe there were three prominent term eigenvectors guiding this environmental cluster: Coca-Cola closely followed by ExxonMobil and McDonalds. Table 4 shows the SVD loadings for topics in 2008's Data Cluster 1 and highlights the ones with the highest loadings.

TABLE IV.
Prominent term eigenvectors loadings for Data Cluster 1 – 2008 reports

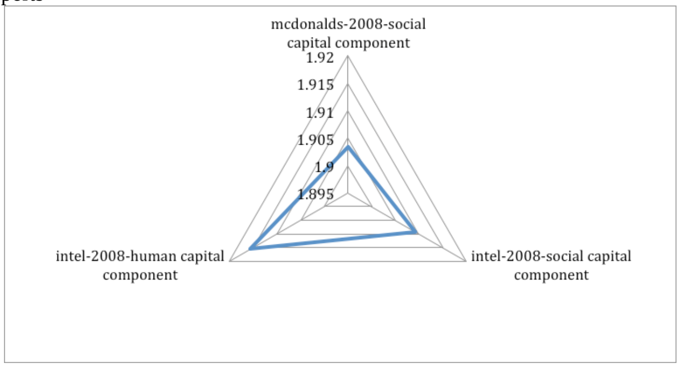


Topic NAME	Term cluster weights for coke-2008-environment component	Term cluster weights for exxonmobil-2008-environment component	Term cluster weights for mcdonalds-2008-environment component
+emission,waste,+recycle,energy,co2	**0.883**	**0.973**	**0.907**
+woman,+employee,diversity,+student,training	0.209	0.299	0.254
+food,+bottle,food,+package,+child	0.365	0.206	0.554
political,+board,+committee,+candidate,conduct	0.151	0.231	0.225

The topic with the highest loading for all three centroid guiding components alludes to: "emission, waste, recycle, energy, Carbon Dioxide (CO2)," and characterizes the prevalent terms utilized around this CSR topic (i.e., environment).

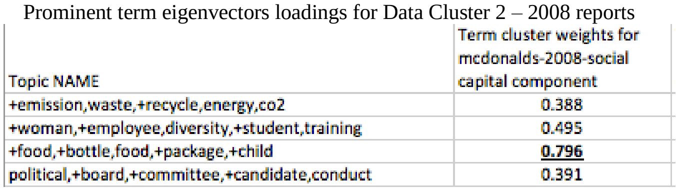
In the second Data Cluster obtained, there was one CSR component closer to the centroid and guiding Data Cluster 2, namely: McDonald's social capital component (see Figure 5 for the components in Data Cluster 2).

Fig 5. Prominent Term Eigenvectors for Data Cluster 2 – 2008 CSR Reports



From Table 5, the topic with the highest loading for McDonald's 2008 social capital component (0.796) refers to: "food, bottle, package, child" such that in 2008's Text Universe the *social capital issue was guided by McDonald's*.

TABLE V.
Prominent term eigenvectors loadings for Data Cluster 2 – 2008 reports



Topic NAME	Term cluster weights for mcdonalds-2008-social capital component
+emission,waste,+recycle,energy,co2	0.388
+woman,+employee,diversity,+student,training	0.495
+food,+bottle,food,+package,+child	**0.796**
political,+board,+committee,+candidate,conduct	0.391

For the third Data Cluster obtained from 2008 SVDs only one CSR component was closest to the centroid and thus guided Data Cluster 3: Coca-Cola human capital component (see Figure 6 for the components in Data Cluster 3).

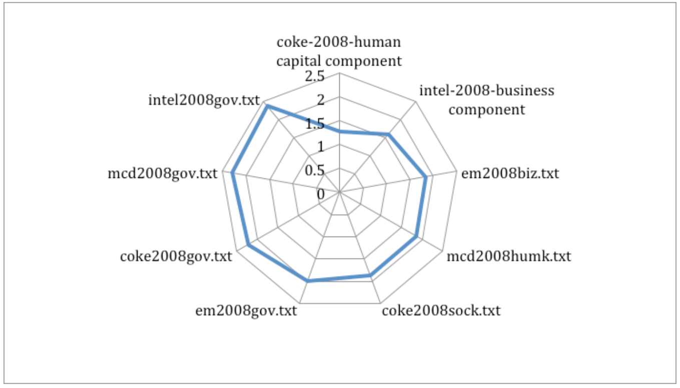


Fig 6. Prominent Term Eigenvectors for Data Cluster 3 – 2008 CSR Reports

C. 2012 Results

Four Data Clusters were obtained in 2012. Figure 7 depicts a radar graph of the Euclidian distances from the centroid of the CSR components grouped in Data Cluster 1. There is one CSR component closest to the centroid and guiding Data Cluster 1, namely: ExxonMobil business component (see Figure 7 for the components in Data Cluster 1 for 2012 CSR reports)

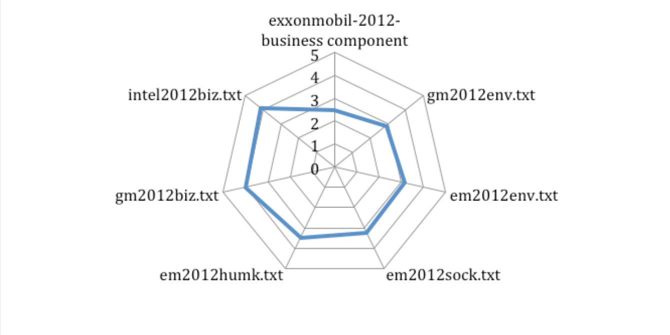


Fig 7. Prominent Term Eigenvectors for Data Cluster 1 – 2012 CSR Reports

Table 7 shows the SVD loadings for all the topics grouped in 2012's Data Cluster 1, and highlights the highest loading for this centroid-guiding component.

TABLE VII.
Prominent term eigenvectors loadings for Data Cluster 1 – 2012 reports

Topic NAME	Term cluster weights for exxonmobil-2012-business component
emission,water,energy,waste,carbon	0.508
+board,political,code,+committee,+director	0.264
teacher,software,+nonprofit,+donation,+young	0.163
+teacher,software,+nonprofit,+donation,+young	0.155
oims,socioeconomic,upstream,upstream,+business line	**0.643**
+farmer,+woman,hiv,red,water	0.166
+package,+ingredient,+bottle,+bottle,+food	0.405
+client,esrm,+loan,credit,+transaction	0.174
+compute,privacy,software,+pc,corporate responsibility	0.22
bottle,sugar,guiding,mutual,human	0.21
+vehicle,traffic,mobility,+driver,+park	0.291
+vehicle,+supplier,chevrolet,+foundation,diversity	0.147
fy12,software,+compute,+developer,citizenship	0.076
+land,+incident,+pipeline,+response,preparedness	-0.051

The topic with the highest loading for ExxonMobil 2012 business component (0.643) refers to ExxonMobil considerations, in particular its "Operations Integrity Management System (OIMS)," which was designed to guide the firm's commitment to excellence in Safety, Security, Health and Environmental (SSH&E) performance, as well as to: "socioeconomic, upstream, business line." Thus, in 2012's there is a very unique and innovative business issue that evidences CSR integration to daily business operations guided by ExxonMobil.

The second Data Cluster obtained also had one component closest to the centroid and guiding Data Cluster 2, namely: Coca-Cola environment component (see Figure 8 for the components in Data Cluster 2).

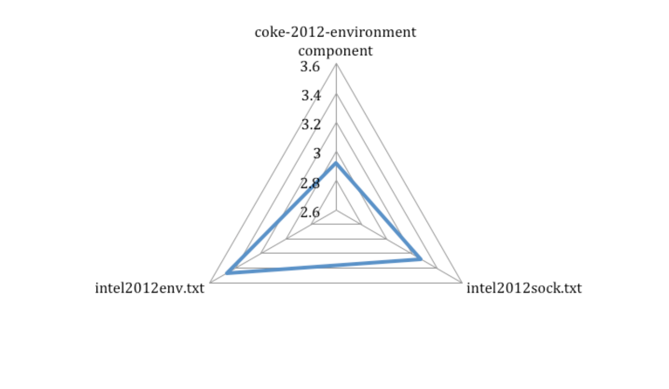


Fig 8. Prominent Term Eigenvectors for Data Cluster 2 – 2012 CSR Reports

The topic with the highest loading refers to: "emission, water, energy, waste, carbon." But, as shown in Table 8, there were two more topics with very high loadings in this centroid-guiding-component with terms such as: "farmer, woman, HIV, red, water, package, ingredient, bottle, food." So in 2012 environment and social capital CSR issues were guided by Coca-Cola.

TABLE VIII.
Prominent term eigenvectors loadings for Data Cluster 2 – 2012 reports

NAME	Term cluster weights for coke-2012-environment component
emission,water,energy,waste,carbon	**0.891**
+board,political,code,+committee,+director	0.302
teacher,software,+nonprofit,+donation,+young	0.237
+teacher,software,+nonprofit,+donation,+young	0.322
oims,socioeconomic,upstream,upstream,+business line	0.338
+farmer,+woman,hiv,red,water	**0.721**
+package,+ingredient,+bottle,+bottle,+food	**0.657**
+client,esrm,+loan,credit,+transaction	0.313
+compute,privacy,software,+pc,corporate responsibility	0.382
bottle,sugar,guiding,mutual,human	**0.463**
+vehicle,traffic,mobility,+driver,+park	0.271
+vehicle,+supplier,chevrolet,+foundation,diversity	0.283
fy12,software,+compute,+developer,citizenship	-0.002
+land,+incident,+pipeline,+response,preparedness	0.025

The third Data Cluster obtained had two CSR components are closest to the centroid and guiding Data Cluster 3, namely: Microsoft human capital and governance components. (see Figure 9 for the components in Data Cluster 3).

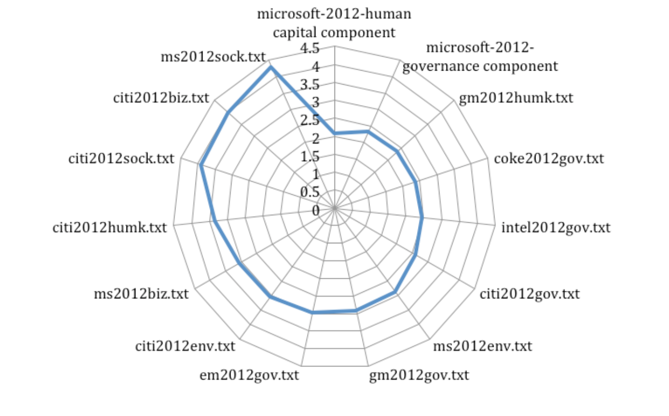
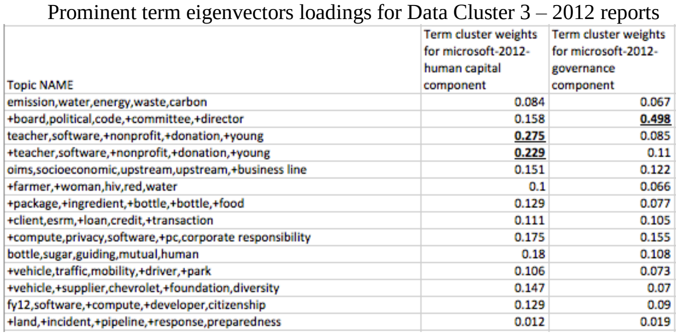


Fig 9. Prominent Term Eigenvectors for Data Cluster 3 – 2012 CSR Reports

As per Table 9, the topic with the highest loadings for Microsoft 2012 human capital component (.229) is combined with the topic with the highest loading for Microsoft 2012 governance component (0.498) to establish that in 2012's Text

Universe governance and human capital CSR issues were guided by Microsoft by referring to: "board, political, code, committee, director, teacher, software, nonprofit, donation, young."

TABLE IX.
Prominent term eigenvectors loadings for Data Cluster 3 – 2012 reports



Topic NAME	Term cluster weights for microsoft-2012-human capital component	Term cluster weights for microsoft-2012-governance component
emission,water,energy,waste,carbon	0.084	0.067
+board,political,code,+committee,+director	0.158	**0.498**
teacher,software,+nonprofit,+donation,+young	**0.275**	0.085
+teacher,software,+nonprofit,+donation,+young	**0.229**	0.11
oims,socioeconomic,upstream,upstream,+business line	0.151	0.122
+farmer,+woman,hiv,red,water	0.1	0.066
+package,+ingredient,+bottle,+bottle,+food	0.129	0.077
+client,esrm,+loan,credit,+transaction	0.111	0.105
+compute,privacy,software,+pc,corporate responsibility	0.175	0.155
bottle,sugar,guiding,mutual,human	0.18	0.108
+vehicle,traffic,mobility,+driver,+park	0.106	0.073
+vehicle,+supplier,chevrolet,+foundation,diversity	0.147	0.07
fy12,software,+compute,+developer,citizenship	0.129	0.09
+land,+incident,+pipeline,+response,preparedness	0.012	0.019

The fourth Data Cluster had Coca-Cola's social capital and business components guiding it (see Figure 10 for the components in Data Cluster 4).

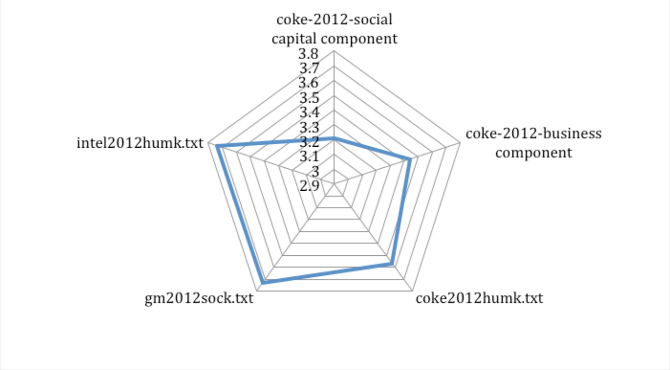
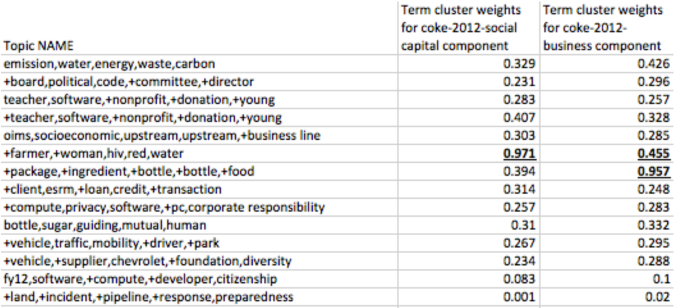


Fig 10. Prominent Term Eigenvectors for Data Cluster 4 – 2012 CSR Reports

There was one topic with the highest loading for Coca Cola 2012 social capital component (.971), as shown in Table 10. As well as, one topic with the highest loading for Coca Cola 2012 business component (.957). Thus, in 2012's Text Universe *social capital and beverage business issues were guided by Coca-Cola* by referring to: "farmer, women, HIV, red, water, package, ingredient, bottle, food."

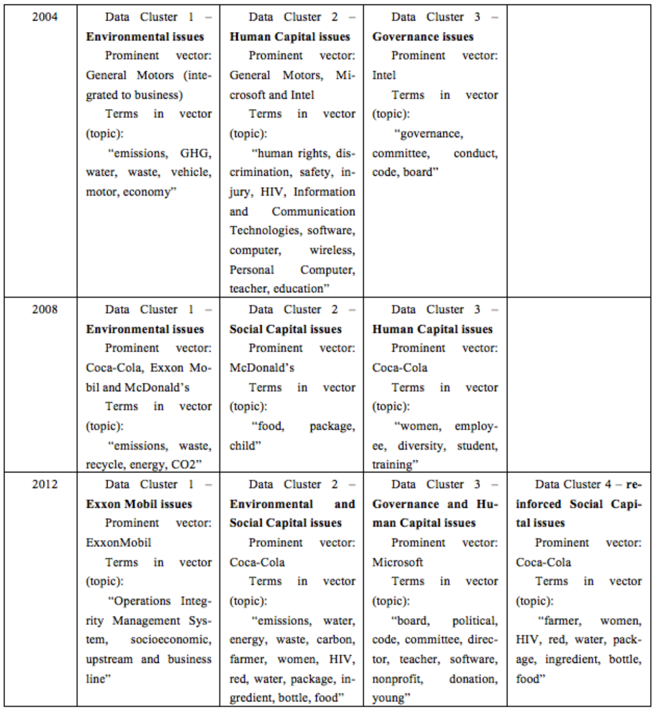
TABLE X.
Prominent term eigenvectors loadings for Data Cluster 5 – 2012 reports



Topic NAME	Term cluster weights for coke-2012-social capital component	Term cluster weights for coke-2012-business component
emission,water,energy,waste,carbon	0.329	0.426
+board,political,code,+committee,+director	0.231	0.296
teacher,software,+nonprofit,+donation,+young	0.283	0.257
+teacher,software,+nonprofit,+donation,+young	0.407	0.328
oims,socioeconomic,upstream,upstream,+business line	0.303	0.285
+farmer,+woman,hiv,red,water	**0.971**	**0.455**
+package,+ingredient,+bottle,+bottle,+food	0.394	**0.957**
+client,esrm,+loan,credit,+transaction	0.314	0.248
+compute,privacy,software,+pc,corporate responsibility	0.257	0.283
bottle,sugar,guiding,mutual,human	0.31	0.332
+vehicle,traffic,mobility,+driver,+park	0.267	0.295
+vehicle,+supplier,chevrolet,+foundation,diversity	0.234	0.288
fy12,software,+compute,+developer,citizenship	0.083	0.1
+land,+incident,+pipeline,+response,preparedness	0.001	0.02

Table 11 summarizes longitudinal results and shows that in 2004 there were three prominent term eigenvectors: General Motors', treating environmental issues with terms related to emissions, motors and vehicles that prevailed over others; Intel treating governance issues by focusing on codes of conduct; and Microsoft's (along with General Motors' and Intel's), with a prominent human capital eigenvector that alluded to human rights, safety, computers and education. These are non-counterintuitive results as it makes sense for firms dealing with products that pollute and run on non-renewable energy sources to focus on environmental considerations. Similarly, it makes sense for firms whose employees constitute their main asset (knowledge workers) to focus on human capital and good governance issues. In general 2004's CSR issues were treated in a traditional and clearcut Environmental, Social (mainly focused on human capital considerations) and Governance dimensions (evidencing that firms followed traditional Environmental Social and Governance – ESG reporting guidelines).

TABLE XI.
Prominent term eigenvectors loadings for Data Cluster 5



2004	Data Cluster 1 – **Environmental issues** Prominent vector: General Motors (integrated to business) Terms in vector (topic): "emissions, GHG, water, waste, vehicle, motor, economy"	Data Cluster 2 – **Human Capital issues** Prominent vector: General Motors, Microsoft and Intel Terms in vector (topic): "human rights, discrimination, safety, injury, HIV, Information and Communication Technologies, software, computer, wireless, Personal Computer, teacher, education"	Data Cluster 3 – **Governance issues** Prominent vector: Intel Terms in vector (topic): "governance, committee, conduct, code, board"	
2008	Data Cluster 1 – **Environmental issues** Prominent vector: Coca-Cola, Exxon Mobil and McDonald's Terms in vector (topic): "emissions, waste, recycle, energy, CO2"	Data Cluster 2 – **Social Capital issues** Prominent vector: McDonald's Terms in vector (topic): "food, package, child"	Data Cluster 3 – **Human Capital issues** Prominent vector: Coca-Cola Terms in vector (topic): "women, employee, diversity, student, training"	
2012	Data Cluster 1 – **Exxon Mobil issues** Prominent vector: ExxonMobil Terms in vector (topic): "Operations Integrity Management System, socioeconomic, upstream and business line"	Data Cluster 2 – **Environmental and Social Capital issues** Prominent vector: Coca-Cola Terms in vector (topic): "emissions, water, energy, waste, carbon, farmer, women, HIV, red, water, package, ingredient, bottle, food"	Data Cluster 3 – **Governance and Human Capital issues** Prominent vector: Microsoft Terms in vector (topic): "board, political, code, committee, director, teacher, software, nonprofit, donation, young"	Data Cluster 4 – **reinforced Social Capital issues** Prominent vector: Coca-Cola Terms in vector (topic): "farmer, women, HIV, red, water, package, ingredient, bottle, food"

In 2008, there also were three prominent term eigenvectors: Coca-Cola treating human capital issues by pointing to women, diversity and training; McDonalds' handling of social capital issues through food, package and child (approached in terms of ethical advertising in its report); and ExxonMobil's (along with Coca-Cola's and McDonald's), with a prominent environmental eigenvector alluding to emissions, waste and recycling. In 2008 there was no prominent term eigenvector related to governance (a traditional CSR dimension), this could have happened because of the great recession, and American firms' efforts to move their CSR approaches beyond mere compliance. In 2012, we also found three prominent term eigenvectors: Coca-Cola's treatment of environmental and social capital issues focusing on water, waste, women and HIV; Microsoft's handling of human capital issues alluding to software donations and teachers; and ExxonMobil's, with CSR integration into core business that revolved around its their OIMS.

III. CONCLUSION

Our use of centroid clustering (Data Clusters) on SVDs obtained while exploring term associations (Term Clusters) in a CSR context helped us identify prominent term eigenvectors around specific CSR topics per year, and to describe the terms used to handle those CSR issues. Longitudinal analysis helped us determine that it is difficult for firms to maintain prominence around specific CSR issues through time. Except for Coca-Cola in relation to the environment, which may be explained by the effect its plants have on groundwater levels (see Table 11, 2008 Data Cluster 1 and 2012 Data Cluster 2). And except for Microsoft, in terms human capital, which has been part of its business strategy, namely: the more software it donates to schools, the more future captive users it has (see Table 11, 2004 Data Cluster 2 and 2012 Data Cluster 3).

Our exploration of term eigenvectors by year, and applying centroid clustering to the SVDs obtained while exploring term associations, helped us visualize and understand more of the information that SVDs convey. Especially, regarding the way firms can the evolutions and prominence of their term eigenvectors through time.

REFERENCES

[1] G. Salton, A. Wong, and C.-S. Yang, “A vector space model for automatic indexing,” *Communications of the ACM,* vol. 18, no. 11, pp. 613-620, 1975.

[2] S. Robertson, “Understanding inverse document frequency: on theoretical arguments for IDF,” *Journal of documentation,* vol. 60, no. 5, pp. 503-520, 2004.

[3] N. Evangelopoulos, and L. Visinescu, “Text-mining the voice of the people,” *Communications of the ACM,* vol. 55, no. 2, pp. 62-69, 2012.

[4] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, “Indexing by latent semantic analysis,” *Journal of the American society for information science,* vol. 41, no. 6, pp. 391, 1990.

[5] S. T. Dumais, “Latent semantic analysis,” *Annual review of information science and technology,* vol. 38, no. 1, pp. 188-230, 2004.

[6] M. Konchady, *Text Mining Application Programming (Programming Series)*: Charles River Media, Inc., 2006.

[7] S. T. Dumais, “LSA and information retrieval: Getting back to basics,” *Handbook of latent semantic analysis*, pp. 293-321, 2007.

[8] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman, “Indexing by latent semantic analysis,” *JASIS,* vol. 41, no. 6, pp. 391-407, 1990.

[9] G. Salton, and C. Buckley, “Term-weighting approaches in automatic text retrieval,” *Information processing & management,* vol. 24, no. 5, pp. 513-523, 1988.

[10] K. Sparck Jones, “Automatic indexing,” *Journal of documentation,* vol. 30, no. 4, pp. 393-432, 1974.

[11] K. Sparck Jones, “A statistical interpretation of term specificity and its application in retrieval,” *Journal of documentation*, vol. 28, no. 1, pp. 11-21, 1972.

[12] A. Singhal, C. Buckley, and M. Mitra, "Pivoted document length normalization." pp. 21-29.

[13] N. Evangelopoulos, X. Zhang, and V. R. Prybutok, “Latent semantic analysis: five methodological recommendations,” *European Journal of Information Systems,* vol. 21, no. 1, pp. 70-86, 2012.

[14] C. D. Manning, P. Raghavan, H. Sch\, \#252, and tze, *Introduction to Information Retrieval*: Cambridge University Press, 2008.

[15] A. Sidorova, N. Evangelopoulos, J. S. Valacich, and T. Ramakrishnan, “Uncovering the intellectual core of the information systems discipline,” *Mis Quarterly,* pp. 467-482, 2008.

[16] A. K. Jain, “Data clustering: 50 years beyond K-means,” *Pattern recognition letters,* vol. 31, no. 8, pp. 651-666, 2010.

[17] J. A. Hartigan, and M. A. Wong, “Algorithm AS 136: A k-means clustering algorithm,” *Applied statistics,* pp. 100-108, 1979.

[18] C. B. Do, and S. Batzoglou, “What is the expectation maximization algorithm?,” *Nature biotechnology,* vol. 26, no. 8, pp. 897-899, 2008.

[19] Sustainability Accounting Standards Board. "Sustainability Accounting Standards Board," http://www.sasb.org/.

[20] C. M. Parra, and M. C. Tremblay, “Analyzing US Corporate Social Responsibility (CSR) Reports using Text Data Mining,” in Ninth International Design Science Research in Information Systems and Technologies (DESRIST) Miami,FL, 2014.