Project title:	*Exploratory analysis of the use of the concept of Sustainable Mobility by automakers in their annual reports using text mining tools*
Main author:	**DA SILVA, Ricardo H. (Polytechnique Montréal)**
Other authors:	BEAUDRY, Catherine, Polytechnique Montréal ARMELLINI, Fabiano, Polytechnique Montréal KAMINSKI, Paulo C., University of Sao Paulo (USP) Brazil

Purpose

The main objective of this research work is to analyze in an exploratory way, using text mining tools with support from R-Studio, if the major vehicle manufacturers in the world are adopting initiatives that favor the growth of the 'sustainable mobility' paradigm as part of their business strategies. Therefore, the main research question that this research will seek to answer is whether the terms that identify strategies that focus on sustainable mobility appear in the annual reports that are published annually by these companies. Other information that will also be analyzed are the terms most used among these automakers to identify sustainable mobility initiatives and how these terms have evolved over the years. This analysis is expected to bring new insights into whether automakers are committed to sustainable mobility.

Design

This research objective will be achieved using Text Mining techniques through the R-Studio software and the packages available in its library that serve this purpose. This process will be broken down into the following macro-steps: (1) text extraction; (2) cleaning; (3) partitioning, annotation and lemmatization; (4) univariate analysis; (5) vectorization; (6) co-occurrence networks; (7) concordance analysis; (8) clustering using K-means. The corpus that will be used for the research consists of the annual report of the twelve largest vehicle manufacturers in the world, responsible to produce more than 70% of the total light vehicles produced in the world annually during the period of 2008 to 2018. The three first steps of our methodology belongs to the data pre-processing phase of the text mining process and their main objective is to prepare the data for our first quantitative analysis – the univariate analysis – which is part of the content analysis phase. As for the text extraction, its main objective is to remove all the text as a single string of characters of the PDF files that will constitute our corpus and insert it into a column of our main data-frame. The cleaning stage was made up of at least two operations. The first operation has the main purpose of removing initial and last pages of the annual reports, as these pages usually contain useless information for our research. The second operation had as its mains objective to regular expressions (also called Regex) to remove undesirable parts of the text such as numbers, special characters, page indicators, first and last lines of each page among others. The main objective of the third step (partitioning, annotation and lemmatization) is to partition the text of each file into units of words and its classes, lemmas, phrases, paragraphs, going down to the document level and giving to each of these units, unique identifiers. All this information will be stored within a data-frame and associated with the respective metadata. This is the data-frame that will be used throughout all the analysis that will follow such as the univariate, the co-occurrences, and the concordance analysis. The objective of univariate analysis is to explore the corpus trying to identify relevant statistical trends, analyzing one variable at a time. In the case of this project, this analysis will be made using the frequency of the lemmas 'Sustainable' an 'Mobility' within the corpus in order to try to identify if there is any type of trend that can be observed among the companies surveyed (metadata). As for co-occurrence networks and concordance analysis, these are multivariate analysis with the main purpose of graphically composing a network of terms where each node represents a term and an arc between two nodes represents the co-occurrence relationship between these two terms. By doing so we can identify the terms that often appear close to each other (not necessary together). Finally, we will pursue with a clustering analysis of our corpus. Cluster analysis is one of the most important text mining methods. Its goal is the automatic partitioning of several objects into a finite set of homogeneous groups (clusters). The objects should be as similar as possible within a group.

Findings

Looking at the univariate analysis, we can infer that some important information regarding the growth and consolidation of the concept of sustainable mobility begins to emerge throughout the annual reports and over the years. We were able to conclude from the analysis carried out that the terms 'mobility' and 'sustainable' appear more frequently within the annual reports issued by the world's largest automakers (exception only to Tesla). We can also verify the use of other relevant terms that seem to indicate a possible sustainable mobility ecosystem that is being formed once terms that point to economic, social, and sustainability aspects always appear connected to the theme of mobility.