Evolution of sustainable development interests in society during the last century

First Author
Mathilde Guillaumot

Second Author
Priscille Guerrier de Dumast

Third Author
Hippolyte Lefebvre

Abstract

In a time where ressources are diminishing and population is growing, the theme of sustainable development is naturally becoming central in topical issues. When we see the state of our planet it rises a lot of questions on our lack of interest on these issues and the difficulty to integrate them in the social most important matters. Our main goals are to understand how sustainable development related thematics are now interfering with topical society interests and how it is received by its people. Also, it is important to understand in which context we talk about these issues and how we talk about them. Do authorities and media like newspapers act in favor of giving a bigger place for this theme on the politics plate or not? In order to draw up an analysis on this trend, we are going to use the archives of the Gazette de Lausanne and the Journal de Genève that gather 200 years of journalistic work and have followed society main interests over the time.

1 Introduction

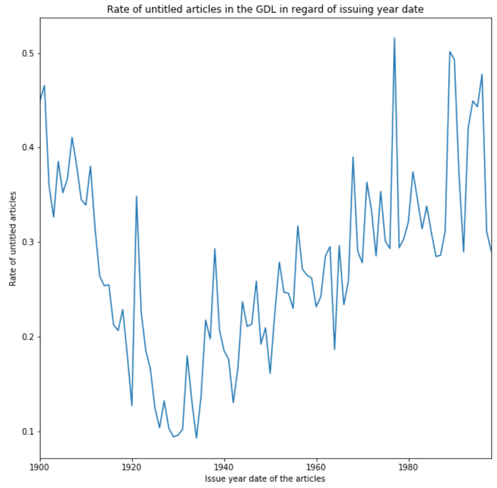
The dataset provided by Le Temps gathering 200 years of newspapers archives is a mine of information on society interests trends over the time. We want to explore it with our prism regarding sustainable development and provide an analysis on this topic's vision and evolution over the time. When did we start to talk about sustainable development? It was certainly not mentioned in those terms at the time, so what did set the stage to this matter? This dataset is wide and gives us the opportunity to explore a lot of data. Nevertheless, we can not process that much data with our ressources, that is the reason why we decided to only look into the last 100 years of archives from 1900 to 2000. This choice is also motivated by the awareness that environnemental topics were less treated prior to that period. Another important assumption is that regarding the age of some archives there are a lot of untitled articles in the journals before and at the beginning of the 20th century, so it is more meaningful to focus on the late years of the 20th century rather than considering the dataset in its entirety.

2 Pre-processing of the data

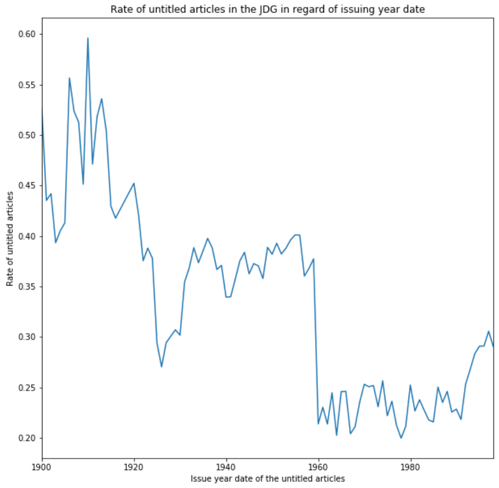
As mentioned above, we only took care of the digital archives from year 1900 to 2000 for both journal: La Gazette de Lausanne (GDL) and Le Journal de Genève (JDG). All these archives were under XML format so we used the beautiful soup html parser to extract the data. All the articles title were embedded in the tag ”name” while the content of the articles were inserted in the tag ”full_text”. For the whole 20th century, the GDL archives gather **2184812** articles and the JDG: **2912098**. Given this huge amount of data, we have undertaken the following pre-processing strategy which consists in retrieving only the articles' title for both journals, then selecting the articles which title matches words related to ”green” topics and finally retrieving the full text of these selected articles. In this first approach, we have treated the GDL and the JDG separately.

2.1 Handling of Untitled Articles

An important part of the pre-processing involved the handling of untitled articles. Indeed, we have counted **947051** untitled articles for the JDG and **636655** for the GDL. Which respectively represents 0.32% and 0.29% of the JDG and GDL datasets. From this counts, the following main question has arised: How many untitled articles over time and total per year are there?
From the above plots, you can infer the following:



(a) Rate representing the # of untitled articles in the GDL over total per issuing date year over time



(b) Rate representing the # of untitled articles in the JDG over total per issuing date year over time

- For the GDL, there is no clear pattern as the rates decrease around 1930 before rising back to an approximative 40% rate as it was for the beginning of the century.

- For the JDG, most of the untitled articles go back to the first decades of the 20th century from 1900 to 1920. The rates decrease considerately afterwards.

Just dropping the untitled articles without further data exploration could seem a bit harsh and may insert bias in our further analysis. Thus, we decided to retrieve the first sentences of all the untitled articles and apply the same treatment that will be applied to the articles titles previously extracted; we have set up a taxonomy of words and we are going to loop over them and see if they appear in the first sentences of the untitled articles. In this manner, we will be able to account if many environnemental topics related articles were dropped if we make the choice of excluding the untitled articles from out datasets. This method has enabled us to retrieve the few articles mentioning environmental topics but deprived of a title: **78 for the GDL and 69 for the JDG**. For the rest of these articles, we have simply excluded them from our datasets just like the articles which title did not relate to any topic of our wanted kind. It is also important to add that the few articles rescued were globally more dated from the second part of the 20th century.

2.2 Retrieving of ”green” articles full text

We have briefly mentioned the process involved to retrieve ”green” articles full content in the previous section but we will develop it here. Prior to extracting any of it, we have set up a list of words that we estimate could be prone to relate to articles talking about environmental topics. It is a list of 58 words. For each journal, while looping over all the articles titles extracted before, we have looked into this list to see if some of these specific words were contained in the articles titles. It resulted in the creation of a dataframe where each row mentioned the title of the article, the issue date of the article and the ”green” word found in this article. Thanks to that dataframe, we then had all the information needed to retrieve the full text of these articles that we will need for further and more consistent analysis.

3 Visualization of the pre-processed data

From both our datasets (GDL and JDG) combined, we retrieve from our original list **21 ”green” words** (see notebook). Below is a plot here to visualize the repartition of the number of occurences of these terms in the articles titles or beginning of content (for untitled articles) over time in both journals. Separated plots for each journal are available for visualization in the notebook but this plot is enough to infer the major trends in the eclosion of environmental topics in the press. Indeed, we can easily see that it is not before the 70s that the press started to include theses topics in their editorial line and the main peaks are gathered around the 90s before decreasing again a little. When looking at the separated plots, it also shows that the peaks of occurences of ”green” words occur at the same time in both journals which reinforces our observations.

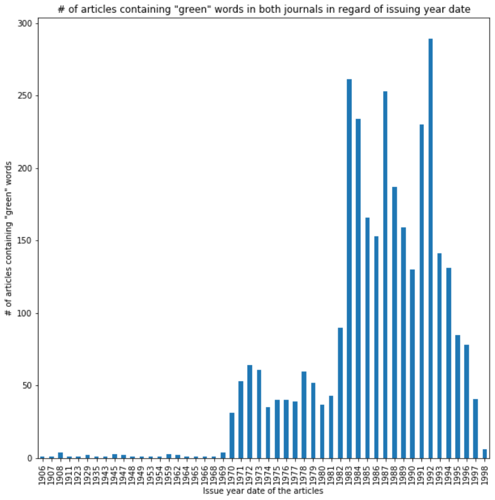


Figure 2: Number of occurences of ”green” words in articles titles in both journals in regard of issuing year date

Above, we have considered the occurences of words over time, now we are going to get into which words occur the most over the century. And the pattern is quite clear as the word ’environnement’ is clearly predominant and outranks all the other ones except for the word ’écologie’ which is the only one to show up a little (see histograms in notebook). This gives us insight on precautions to take for our further analysis. The word ’environnement’ in French is not totally dedicated to the matters of environmental issues and can be employed in several other contexts. This is the reason why we have randomly verified that most of the retrieved ”green” articles were effectively talking about environmental topics and not just mentioning the word in another context. Nevertheless, our next task being performing topic modeling on the data will enable us to observe wether these ”non green” articles have a strong impact on topics extraction.

4 Topic Modeling

4.1 What is topic modeling?

Topic modeling provides hidden thematic structure in a collection of texts. From a dedicated input of texts, it can learn a set of ”topics” defined by a certain amount of words which constitute recurring themes in the text collection. The purpose of topic modeling is to provide the degree of probability of each text element in the collection to relate to some learnt topic. To perform topic modeling on our processed set of ”green” articles, we have used a technique called non-negative matrix factorization (NMF), it is highly similar to the very popular Latent Dirichlet Allocation (LDA). What distinguish the two is that LDA is a probabilistic model while NMF is a deterministic algorithm. The latter is closer to a Machine Learning Algorithm.

4.2 Pre-processing of articles content

- **Cleaning articles text:** In this first part of the process, we are getting the tokens for each article. The next step is to get rid of punctuation, numbers and stop words in articles. In french, small words no longer than 4 characters often are conjonctions or such words that do not have an important meaning, it is the reason why we get rid of them too.

- **Stemming:** We wanted to include a stemming step which purpose is to remove or replace word suffixes to arrive at a common root form of a word so that it prevents from having words of the same root recurring in one article and thus in the top words of learnt topics after learning. However, despite having tried several techniques such as the Snowball language which provides a stemmer for french language, the results from stemming were not at all pertinent. In spite of knowing that it may affect topic modeling we decided to skip this step in a first time.

4.3 Topic modeling learning algoirthm

We use the CountVectorizer from the sklearn library to convert our collection of ”green” articles to a matrix of token counts. We provide a maximum frequency value of 0.7 and a minimum frequency of 0.01 above and under which words are ignored because often it is common words which are not relevant to our analysis. It extracts a number of features from our provided dataset. The NMF will help us retrieve an article-topic matrix associated with a list of top words for each topic. We have chosen to learn 10 topics each composed of 10 top words. They are displayed at the end of the notebook.

4.4 Visualization of Topic Modeling

4.4.1 Strength association with a topic

One important thing to think of after having done topic modeling is the degree of association between topic top words and the topic itself. In some cases, a word despite being representative of a topic can be found in an article which main associated topic is not the one of the top word. It shows the influence of this top word on the topic association of the article. Therefore, we have counted the occurences of each top word for each topic and compared it to the number of articles which main associated topic is the one from which the top word comes from. It results in a series of rates that represents the shares of these top words in topic association. To have a visual approach on this matter, we display each topic top words sized with their strength association to their related topic. Following is an example of this representation with one of the topics found.

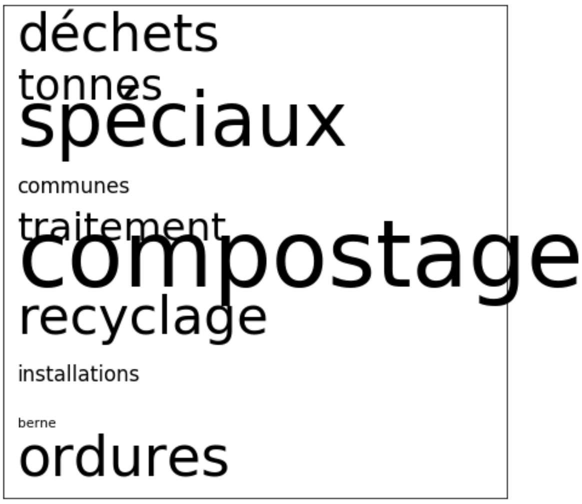


Figure 3: Displaying of top words weighted by their degree of relation with their associated topic

4.5 Heatmap

Another way to account for topic modeling in articles is the heatmap, it gives a quick understanding on which topics really matter in the article and it also gives us the opportunity to see if a topic is mainly associated with a ”green” word. For example, let’s look at the 6 ”reboisement” related articles of the GDL. Below is the associated heatmap, we can observe that for all of them the topic 8 is the dominant topic and when looking at the topics table we see that the topic 8 is associated with words like forest or tree which totally relate to the meaning of ”reboisement”. Looking at other words in the topic’s top words can give us insight on which topics are associated with this kind of articles. In this case, it is associated with ”dépérissement” or ”dégât” which can make us think that the ”reboisement” theme is treated when nature is in danger and it makes sense since ”reboisement” involves former destruction of nature.

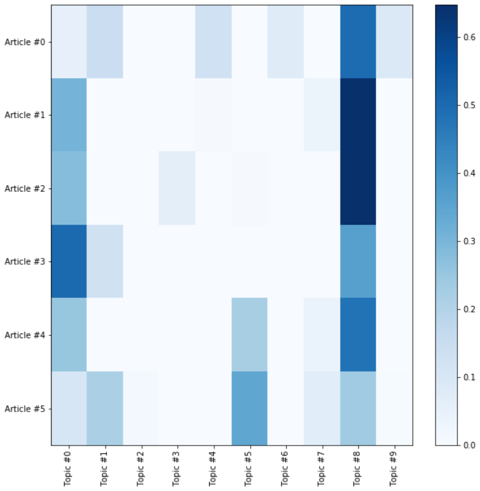


Figure 4: Displaying of top words weighted by their degree of relation with their associated topic

5 Conclusion

This projet has lead to many reflexions on data analysis, the means to handle data and especially how to process it in the way that it can answer our analytic needs. Having to deal with a huge dataset was a constraint that taught us a lot as it has forced us to set things upstream before proceeding to proper data analysis. Regarding the project itself, there are still improvements to be made such as finding a proper way of stemming the articles content before running topic modeling. Also a better organization of the retrieved data could ease our operations when needing to run our analysis. Nevertheless, it was very rewarding to be able to draw up consistent trends and analysis related to the research questions we had chosen to deal with in the beginning.