Mining XIXth century Periodicals: an experimental toolbox

Pierre-Carl Langlais
GRIPIC-Paris-Sorbonne
pierrecarl.langlais@gmail.com

During the past ten years, libraries have digitized newspapers on a massive scale. Billions of pages in the public domain have morphed into textual data, giving researchers the ability to query effortlessly some of the most important testimonials of past social and intellectual lives. It has also opened up this corpora to new readers: algorithms.

For my doctoral thesis (Langlais 2015), I have applied "distant reading" techniques (Moretti 2013) to study the birth of a specific genre in the general press: the stock exchange section. I have developed a scraping app, Pyllica, in order to extract the 1500 weekly financial columns of one of the leading French newspaper of the XIXth century, the Journal des débats, from 1838 till 1870\(^1\).

While this corpora could be treated as a set of messy economic data, I privileged a stylistic approach. A combination of text mining techniques (correspondence analysis, topic modeling, syntax parsing, word embedding) helped me to reconstruct the defining patterns that transformed an experimental initiative into a regular feature of French news writing.

My current researches, notably within the Medias19 project\(^2\), have extended beyond financial journalism. Thanks to the recent ability to download in one time several major digitized newspaper collections of Europeana, I have begun to initiate extensive survey of the evolution of media writing and practices in the XIXth century.

This proposal aims to present an experimental toolbox for the algorithm-mediated reading of old periodicals. This field is specific enough to require the implementation of distinctive methods. It bears upon a text in progress: while the XIXth century marks the birth of modern media writing, this is not a sudden birth, but a complex condensation of tools, social networks and cultural representations of text. It cannot be easily cut off from the general flow of writing publications: much more than today, newspapers were then "intertextual" productions, that borrowed widely from external sources. Finally, we do not deal with a perfect text: given the frail nature of newspaper archives, OCR quality is usually lower than what we can expect from other forms of publication.

The three next subsections will present some major features of the specificities of the "distant reading" of XIXth century periodicals.

\(^1\)The scraping app can be downloaded at https://github.com/Dorialexander/Pyllica.
\(^2\)http://www.medias19.org/.
1 The structural transformation of the media sphere.

The distant reading of newspapers as texts needs calls for a closely correlated distant reading of newspapers as "entities", located in a specific time and place. In my doctoral thesis, I identified the transfer of French newsrooms toward the surroundings of the newly opened Stock Exchange as a major factor contributing to the advent and dissemination of financial journalism.

![Heat map of the localization of Newsrooms in Paris in 1837](image1.png)

**Figure 1:** Heat map of the localization of Newsrooms in Paris in 1837

I am currently extending this initial work into a wider "temporal map" of parisian media structures, that will track the development of information networks (optical and electrical telegraph, pneumatic tubes) and geographical-grounded communities (from the artistic reviews of Montmartre to the scholar journals of the "quartier latin"). Thanks to the experimental "geo-temporal" tools supported by QGIS (through the Time Manager plugin), I have created a dynamic map that generates the pattern of the streets of Montmartre at any given day since 1867.

![Map of Montmartre streets 1874](image2.png)

**Figure 2:** The state of the streets of Montmartre on 1874, November 20th
2 Modeling the topics: a distant reading of the making of journalistic genres.

Numerous regular forms of the daily press have first appeared at the turn of the XIXth century. While this chronology is frequently acknowledged, the stylistic process of "making" a genre remains fuzzy: how does an emerging genre create its own linguistic repertoire? How does it display its discursive specificities within the concrete geographical boundaries of a newspaper page and the indistinct circulations of the "public sphere"? On a lexical level, correspondence analysis and word embedding allow to investigate theses negotiations of semantics, by projecting the discourse as an evolving network of linked concepts.

Figure 3: Using word embedding, this discursive network of stock exchange sections maps its two main subgenre: the account of quotes (red) and the analytical commentary of financial tendencies (green).

While generally used to identify hidden classifications, topic modeling have proven quite effective to evaluate the "strength" of already known textual affiliations, such as literary standards (Underwood and Sellers 2015) or "authorship function" (Langlais 2015).

Figure 4: Attributions of stock exchange chronicles in the Journal des débats from 1838 to 1848 with topics modeling. The wide number of unattributed chronicles (in grey) suggests a composite multi-authorship system.
3 The archeology of "buzz": tracking the networks of a reprinting culture.

Given the loose application of copyrights and author rights to journalistic writing, periodicals heavily borrowed from previously published content. Building on the pioneering work of Ryan Cordell (2015), I am currently developing a set of tools to reconstruct this chain of borrowings in French newspapers.

By studying the virality of news, I am able to further establish the network of exchanges at a time when most of the dissemination of information worked beyond the structured networks of a news agency (such as the "transatlantic traffic of francophone news" identified by Pinson 2015). The use of reprinting within a media appears as a textual feature per se: it displays a rhetoric of recurrence and repetition, especially visible in the advertising section.

Figure 5: Preview of my Virality tool comparing the amount of reprinting between two newspapers.

A combination of virality programs with near-entity recognition techniques appears as a promising toolbox for the historical study of mediatization. For instance, through the occurrences of a celebrity (like 'Franz Liszt') in a wide newspapers corpora we could extract the correlative chain of reprints of formulas and articles surrounding this topic, therefore producing a valuable map of the way notoriety is built and disseminated.

References


