The FRENCH BIG NATIONAL DEBATE and big data

12 août 2019

This is an extended abstract of the french language preprint describing our research. Since the analysed corpus is written in French we used French to describe the analysis. But the technique and the software that we used are applicable to any language.

We use a simple technique for information retrieval applicable to large corpuses based on the detection of relevant terms in the texts.

We are now familiar with the practice of internet search engines. The user searches for documents containing a set of terms. It is this practice that we generalize for an quantitative analysis of the answers to the French big debate (Le grand débat National during spring 2019).

We here propose to use a generalization which is the one used for so-called advanced research.

Suppose we want to analyze the answers to the question of the Grand Débat about the organization of the state : "Whom do you trust most to represent you in society and why ?"

After examining some hundred replies among 89,266, terms like 'mayor', 'elected', 'deputy', 'nobody' etc. seem relevant. But in addition to counting each of these exact terms appearing in the answers we would like :

— Take into account answers containing approximate terms : some participants may have written 'deputies', others 'elected' or 'president', 'President', 'deputy'. Instead of a single term, we will test the presence of 'elected' OR 'Elected' OR 'deputy' OR 'deputies' ... The OR is here a logical OR, corresponding to the presence of one or more terms related to the first. We will use the word 'lemme' to describe these terms that we consider equivalent in the context of the asked question.

— It is interesting to count the presence of a lemma among other possible, but we would also like to count the simultaneous presence of several lemmas in the same answer : for example, how many answers contain the lemma 'elected' but also the lemma 'mayor' in the examined answer. The test then involves a logical AND : the response to the test is positive only if both lemmas are present. The 'nobody' lemma, and the lemma ('elected', 'elected', 'deputy', 'deputies', 'president', 'president') represented

by the term 'elected'.

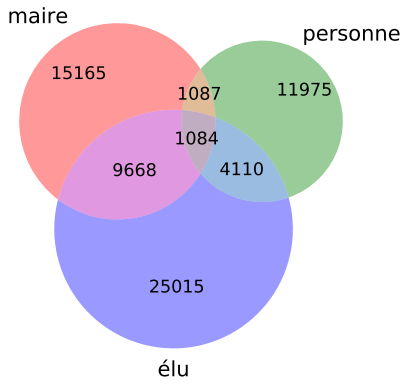
The procedure therefore consists of :

— Search the lemmas used in the answers by reading some of the answers, say a hundred. The issue is not only finding the relevant terms, but also of grouping them into lemmas. We also apply the regrouping procedure to lemmisation in the strict sense : we consider as equivalent terms written in capital letters or in lower case, the terms masculine or feminine and those in singular or plural.

— Once the lemmas are defined, their presence or absence is recorded in the the set of answers.

Several graphical representations of the results are possible. We used Venn diagrams at 2 and 3 dimensions, so for 2 or 3 lemmas tested, because they retain the proportionality of the domain area to the number of responses.

In the case of the question taken as an example : "Whom do you trust most to represent you in society and why ? " We chose the following three lemmas whose representative is the first element. The lemma ('mayor', 'mayor', 'mayors') represented by the term 'mayor', the lemma ('nobody', 'me', 'people') represented by the term 'person', and the lemma ('elected', 'elected', 'deputy', 'deputies', 'president', 'President') represented by the term 'elected'. The analysis of 89,266 responses is to test the presence in the responses of one or more lemmas. By testing the presence of only one lemma independently of two others, we observe that 27,004 answers mention one of the terms of the lemma mayor, 39,877 the lemma elected and and 18,256 the lemma person. 21,162 responses do not mention any of the three.



The Venn diagram is the graphical representation of the presence in the responses of one or more lemmas. The area of each portion of a circle is proportional to the number of replies mentioning lemmas. The three complete circles

of color brick, green and blue respectively correspond to lemmas Mayor, nobody and elected. Intersections with composite colors correspond mixed answers mentioning two or three lemmas. The numbers are the corresponding number of responses.

We here see that a majority of respondents trust the elected officials, and more specifically the mayors