Evolution of direct democracy in Switzerland

Youssef Janjar
youssef.janjar@epfl.ch

Pierre Mbanga Ndjock
pierre.mbangandjock@epfl.ch

Robin Denhardt-Eriksson
robin.denhardteriksson@epfl.ch

Abstract

This document contains the instructions for preparing a report for ADA 2017. The document itself conforms to its own specifications, and is therefore an example of what your manuscript should look like. This document is based on the ACL 2014 paper format.

1 Credits

In the scope of this project our focus lies on written political articles in newspapers. We are willing to assess the diversity of subjects submitted for votations to Swiss residents over the last 200 years. We essentially want to classify political articles in order to identify trends, distributions, densities or patterns among others over the decades. Therefore, we seek to analyse "Le Temps digital archives and data". This dataset consists of articles representing two centuries of informations provided by two newspapers, namely 'Le journal de Genve' and 'La Gazette de Lausanne'. These are ancestors of today well-known Swiss newspaper 'Le Temps' whose publications are written in the French language.

2 Introduction

Due to its federal constitution Switzerland Confederation has forged itself a solid reputation as one of the world most democratic country. The land is praised by observers and often regarded as a model to follow. One key specificity of this state relies undoubtedly in the voting frequency. Indeed Swiss citizens are regularly called to the polls. This aspect among others constitutes one fundamental argument showing the 'democraticity' of the country.
Votations, as they are named in the country, can be initiate by any citizen. Those votations can take place at three different layers of the federal confederation, namely communal, cantonal or federal. Communal votations are reserved for swiss residents inside a given commune. The same apply for cantonal votations where eligible citizen must reside in the canton. Finally, federal votations concern all citizen in the country. This system, often depicted as highly democratic, allow any citizen under some conditions to submit a votation to its fellow citizens.
We want to challenge this polished reputation acquired over the years by gaining some insights into the various subjects that have been addressed.

3 Project Goal

In the scope of this project our focus lies on written political articles in newspapers. We are willing to assess the diversity of subjects proposed to votations to Swiss residents over the last 200 years. We essentially want to classify political articles in order to identify trends, distributions, densities or patterns among others over the decades. This approach should allow us to gain valuable insight into concerns that underlie Swiss political flow.

4 Dataset

We seek to analyze "Le Temps digital archives and data". This dataset consists of articles representing two centuries of informations provided by two newspapers, namely 'Le journal de Geneve' and 'La Gazette de Lausanne'. These are ancestors of today well-known Swiss newspaper 'Le Temps' whose publications are written in the French language.

The dataset at hand is very well organized. It is easy to navigate and find articles for any given time period as publications are gathered with respect to their release month. Hence for each newspaper all articles published during a specific month of a given year are all stored in the same file. The data structure used in each of those files is xml.

Unfortunately some texts are unreadable as they contain a bunch of symbols or special characters. This observation mostly arises in old publications. This may be caused by the scanner, used to create the digital dataset, not adapted for those fonts. Or the archive being in a seriously damaged state. As an example, the following articles issued on February 1798:

*[...] JLEfauteuil c $ t vacc'ant, et l'Assemble s'ccp * de la nomination d'un nouveau Prsident. Une immense majoritS y rappelle le C. Glayre " J'a * vois" besoin de forces [...]*

Specifically the dataset consists of 4'335 xml files for a total of () articles. The period of time covered by those articles ranges from February 1798 to February 1998.

5 Data preparation

Naturally we apply tools related to the field of natural language processing to acquire our results. More precisely we are interested in topic modeling using unsupervised learning. It has been essential to develop a sensitive pipeline to improve the performance of the algorithm. Therefore we have modeled this pipeline into x different stages that we detail in subsequent section: data retrieval, data reduction, data cleaning and data processing. Due to its performances and the variety of implementations available through libraries we decide to use Latent Dirichlet Allocation (LDA).

5.1 Data retrieval

Our first task is to extract the articles from the xml files to store them in an array. The array data structure makes it easier to manipulate whether for access, write or process using the different libraries. For each article in the files we save only the text and the date of publication.
It is important to mention here that articles inside the xml structure do not consistently respond the natural classification by theme as we may except. As an example let's consider the following publication issued on 04 September 1993:

QUOTIDIENNES JUSTICE *Cadres du btiment acquitts Le Tribunal correctionnel de Lausanne a acquitt jeudi deux anciens cadres [...]* ***CONJONCTURE*** *Vaud toujours malade La marche des affaires de l'industrie vaudoise est anmique. L'entre des commandes [...]****VOTATIONS*** *Deux oui et trois non libraux Runis Lausanne, les dlgus du Parti libral vaudois ont pris position sur les cinq objets soumis la votation populaire des 26 et 27 septembre prochains. [...]* ***PAYSANS DE MONTAGNE*** *Aide du gouvernement Face la grogne suscite chez les paysans de montagne [...]*

The entire text is considered a single entity inside the corresponding xml file altough we can clearly discriminate four different topics merely related to each others, namely 'Quotidiennes de Justice', 'Conjoncture', 'Votation' and 'Paysan de montagne'. This remark motivates a choice we explain in the next subsection.

5.2 data reduction

In order to discriminate articles related to votations from others we take a very simple yet sensitive approach. We argue that it is very unlikely that any publication related to Swiss votations does not contain at least one word in a set of predefined words. We mention here those words we consider as strongly related to votations: (). Those words are obtained based on our personal intuition. Of course, One can argue about the accuracy of such a method. We may certainly generate false positive or discard relevant articles. Yet the sample size we obtain with this first process appears to be large enough to capture the information we are interested in. For the unique year 1990 this approach allows us to extract more than 3000 publications.
Furthermore, we make an assumption that reduces the length of any given publication related to votations. As mentionned earlier the goal is to identify and classify votation subjects. By visual inspection, we notice this information is usually closed to one of the keywords showed earlier. Let's consider the following extract of a publication issued on 04 September 1993:

VOTATIONS Deux oui et trois non libraux Runis Lausanne, les dlgus du Parti libral vaudois

ont pris position sur les cinq objets soumis la ***votation*** *populaire des 26 et 27 septembre prochains. C'est ainsi qu'ils ont accept le rattachement du district de Laufon Ble-Campagne et l'arrłt fdral urgent en matire* ***d'assurance-maladie****. En revanche, l'arrłt fdral contre l'usage abusif* ***d'armes*** *et l'initative* ***Pour un jour de Fłte national fri*** *n'ont pas trouv grce leurs yeux. Pas plus d'ailleurs que l'arrłt fdral urgent en matire* ***d'assurance chmage****, jug trop coteux pour l'conomie.*

This extract is part of the article presented in the previous subsection. It represents approximately twenty percent of the original in terms of number of characters. Yet it is more than enough to visually capture the subjects, highlighted in bold, of the votations. More importantly, the remaining part of the article does not provide any insight for the information we are looking for. We can even argue that it is pure noise that needs to be discarded. As a consequence we retain only sentences containing one of our defined keywords together with its closest neighboors, namely the preceding and following sentences. Again we are aware that this filtering method may discard relevant information. More importantly, this assumption may reveal disastruous for our results if it appears to be erroneous.

5.3 Data cleaning

At this stage we have stored in an array articles related to votations. Our objective is now to pre-process those publication before running the unsupervised learning method. The pipeline we use for this task can be summarized by the following schema

remove non French words $\rightarrow$ *lemmatize* $\rightarrow$ *remove stop words* $\rightarrow$ *remove digits*

To remove non French words we use a dictionary proposed by the NLTK library. The lemmatization step is resolved through Spacy lemmatizer. Spacy library also proposes functionalities to easily detect and remove stop words and digits. All those actions put together return a cleaner corpus that can be processed more efficiently at later stages.

6 Data processing

The pre-processing stages described in the previous section can now be exploited to derive results. As mentioned earlier the task we are interested in is named topic modeling. The problem constraints forces us to use an unsupervised learning method for clustering known as Latent Dirichlet Allocation (LDA). This choice is motivated by discussions with pairs and practical experiment results observed in the literature. Based on observations, other clustering methods such as K-means or DBSCAN don't appear to be interesting enough alternatives.

6.1 Latent Dirichlet allocation

Latent Dirichlet allocation (LDA) is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. From a practical point of view, LDA represents documents as mixtures of topics that spit out words with certain probabilities. Documents are interpreted using the bag-of-words representation. LDA then tries to backtrack from the documents to find a set of topics that are likely to have generated the collection.

The generative process is as follows. Documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words. LDA assumes the following generative process for a corpus $\boldsymbol{D}$ consisting of $\boldsymbol{M}$ documents each of length $\boldsymbol{N}_i$:

1.We choose $\theta_i \sim$, where $\boldsymbol{i} \in \{\boldsymbol{1},...,\boldsymbol{M}\}$ and Dir(α) is a Dirichlet distribution with a symmetric parameter α which typically is a spares (α ¡ 1).

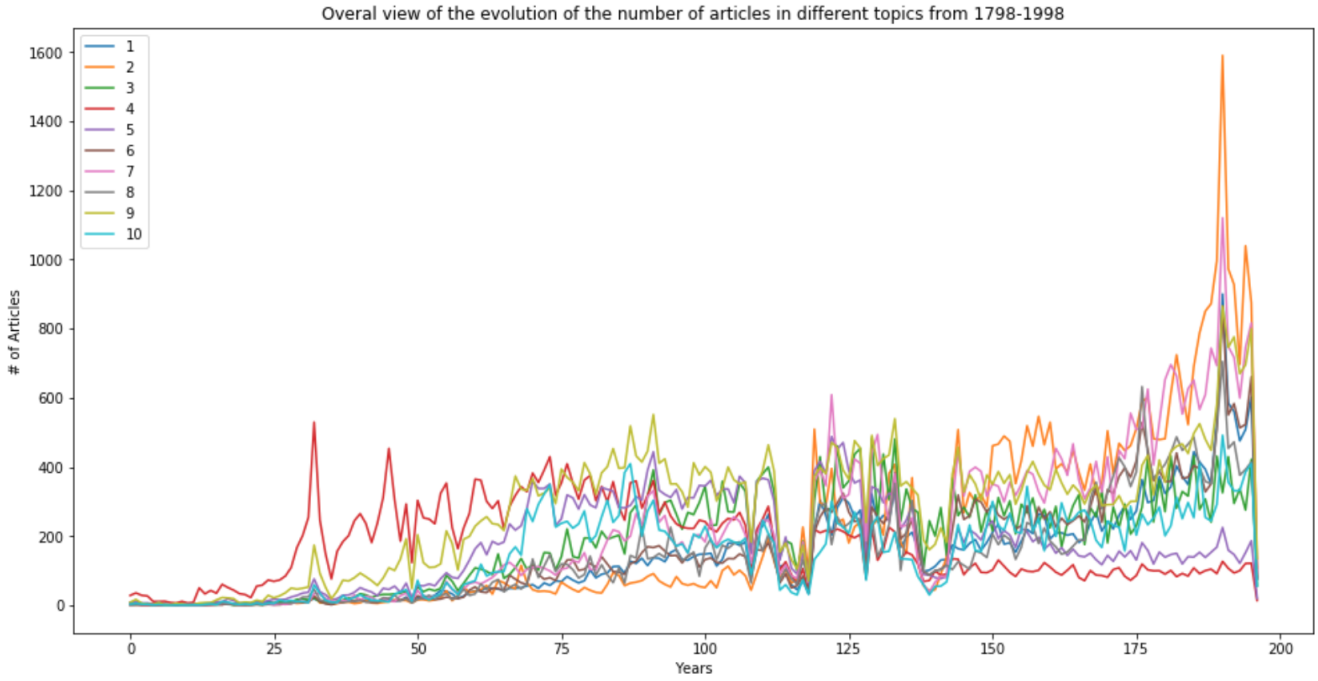
2.We choose $\varphi_k \sim$, where $\boldsymbol{k} \in \{\boldsymbol{1},...,\boldsymbol{M}\}$ and β is a sparse.

3.For each of the word position $\boldsymbol{i,j}$ where $\boldsymbol{j} \in \{\boldsymbol{1},...,\boldsymbol{N}_i\}$ and $\boldsymbol{i} \in \{\boldsymbol{1},...,\boldsymbol{M}\}$:
Choose a topic $\boldsymbol{z}_{i,j} \sim \boldsymbol{Multinomial}(\theta_i)$.
Choose a topic $\boldsymbol{w}_{i,j} \sim \boldsymbol{Multinomial}(\varphi_{z_{i,j}})$.

The lengths $\boldsymbol{N}_i$ are treated as independent of all the other data generating variables ($\boldsymbol{w}$ and $\boldsymbol{z}$) The subscript is often dropped, as in the plate diagrams shown here.

6.2 Results and Graphics

After processing all the articles using the LDA algorithm specify a classification over 10 topics and we can see figure 1 one the evolution of the differ-



ent topics occurrence on the newspapers over 200 years (1798-1998).

We choose not to give a formal name to each topic since the classification is done on recurrence of words basis, but looking at the most recurrent word of each topic we can have a clear idea on the subject addressed on the articles classified under each topic. We do a deeper analyze and visualization of our results in our notebooks.

7 Libraries

The two main libraries we used are Spacy and Gensim. We used Spacy for pre-processing our data we lemmatized the texts using the fonction 'lemma',we used it also to drop the stop words,the digits and the words that were not french words. We used Gensim to implemente of the LDA algorithm for the text classification.

References

Hamed Jelodar, Yongli Wang, Chi Yuan, Xia Feng Department of Computer Science and Engineering Nanjing University of Science and Technology. *Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey*.

Patrick Harrison. 2016. *Natural Language Processing (NLP)*. PyData DC 2016.