Metadata mining of large collections of historical newspapers

Supervisors : Maud Ehrmann and Matteo Romanello

10 avril 2019

Bachelor Project from Christian Gasser - student in Electrical and Electronics Engineering

Introduction

The Impresso Project explore the past through the newspapers. My goal is to carry out an analysis of varoious types of metadata describing the newspapers in our corpus.

In this perspective, I spend the time I have to build the most general tool to achieve this behaviour

Objective

The goal of the project is to monitoring newspaper processing in order to detect where potential problems are and make a large scale newspaper analysis.

To follow this behaviour, I chose to build a variation detector. In order to do that, I defined a list of variable to make the tool the more generic as possible. Then, the idea is to made this first on the publication date variation.

Milestones

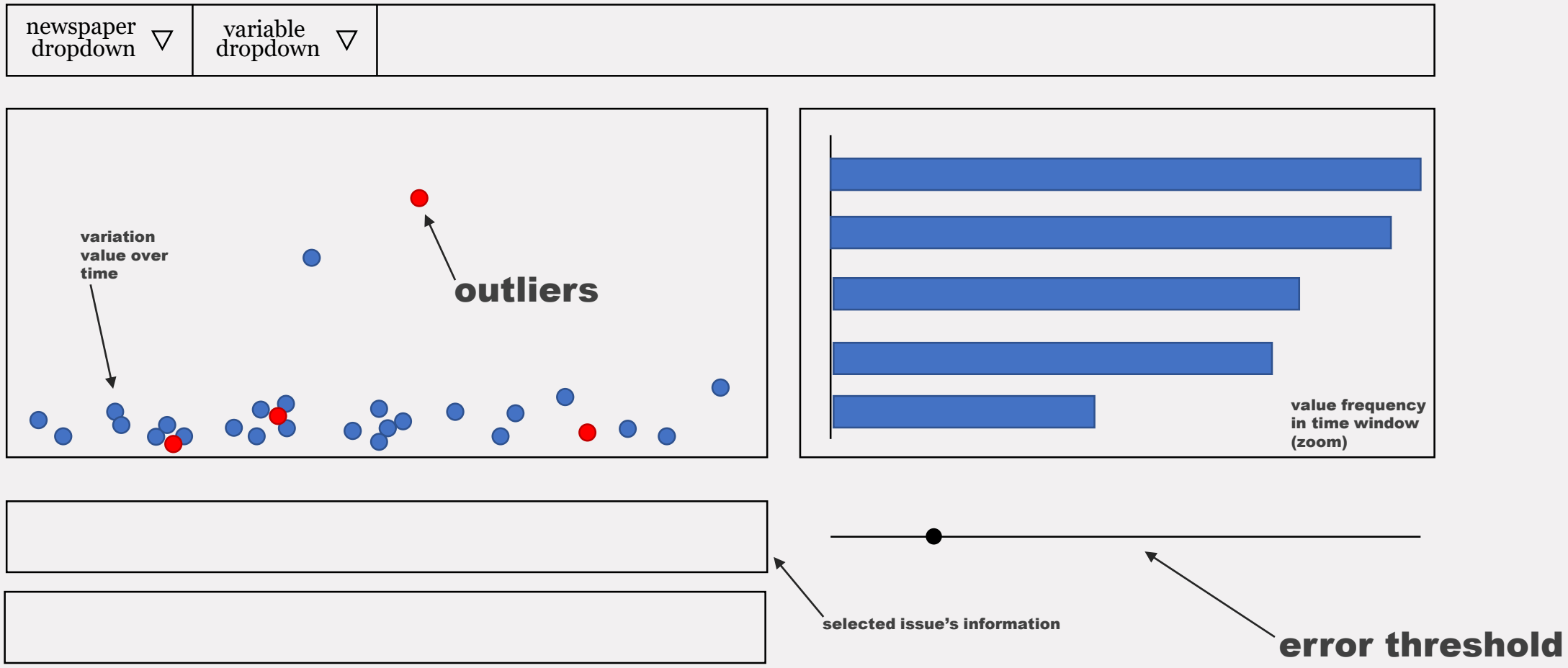
- definition of the project
  decide about the frame and what we want to observe
- material configuration,
  install anaconda environment
  jupyter remote and configuration for data access
- technology choices and learn how it works and how I can use that stuff
- first prototyping and results
- mid term presentation
- make prototype complete and better
- generalize prototype for all variables and make it finished

what are the variables

To think about our future tool I'm going to build, let's have a look on what kinds of variables can we manipulate

Characteristic	Description
Surface	Physical size of the page over time
Nb pages	Number of pages per issue
Nb articles	Average number of articles per page per issue
Nb illustration	Number of pictures and drawings for 1000 pages
Publication date	Variation of publication dates
Front page ill.	Nb of front page illustration per issue
Nb words	Number of words per page. To reduce the noise in our graphic we can compute a quantification of 100 words (packets of 100 words)
Words density	The density of words can be computed by dividing the number of words through the surface
Surface of article	Surface that an article need in the page

generic variation analyser

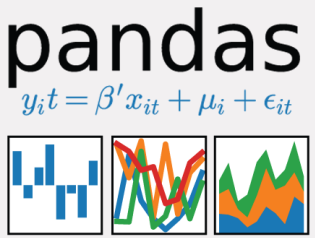


technology choices

python

juypter

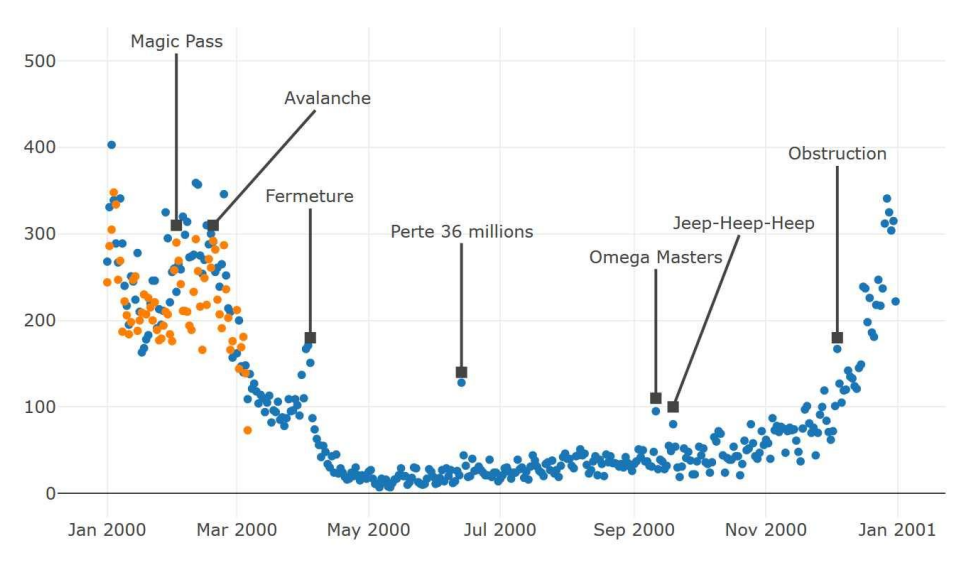
dash



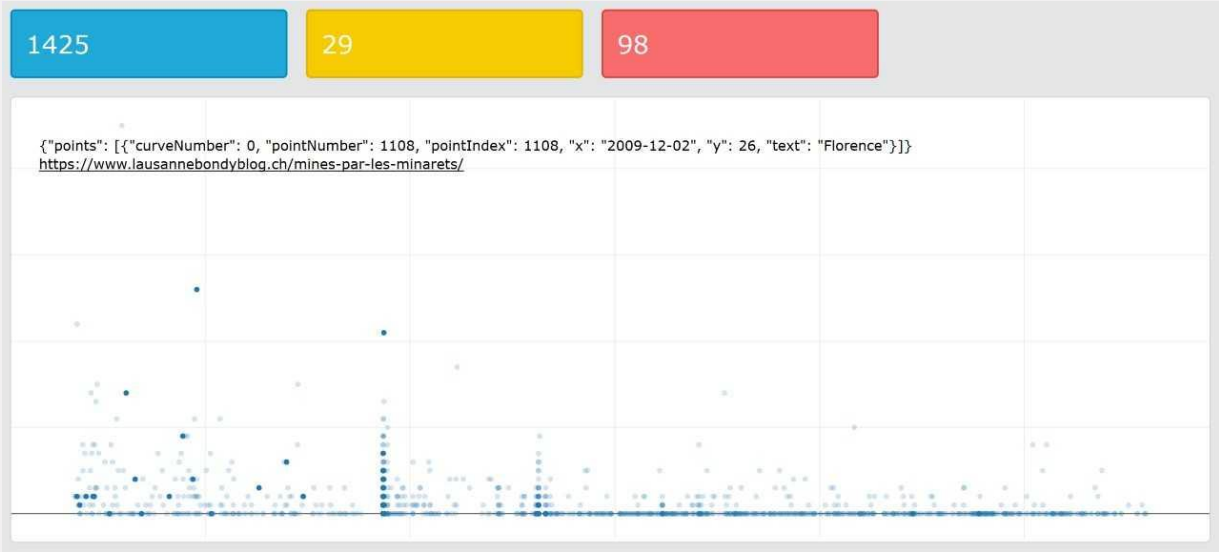
pandas

Let's have a look on the differents technologies I use for the project. It's quite obvious to use the python, jupyter and pandas. But I also use jupyter lab, numpy for matrix calculation and the combination of Dash with Plotly

quick look at plotly+dash



Plotly is able to plot various datas types (time for example) and add some interactive behaviours (zoom, selecting) and we can choose to plot with the webgl technology



With Dash, it's possible to build some dashboards with live update in response of user events

variation analyser

purpose : detect regular patterns and outliers

The idea of this tool is to analyse the variation between the previous issue and the next one. In order to do that, we compute for a given variable the variation. Then we plot it on the big graph on the left. For the first prototype, we're looking for the variation of publication frequency

If we zoom in this graph, the graph on the right which shows the frequencies of the difference values will upload itselfs in order to analyze the new time window.

Finaly we have a highlight system. It indicates on the main graph what are the outliers points to make the anomalies visible. To do that, we compute a linear prediction and calculate the error which is the difference between our calculation and the real value. The slider allow us to chose which range of error we want to see.

On the bottom, we have a table that show us the issues that we select on the main graph with a link on impresso.

linear prediction

similar to linear regression!

model : $$x[n] = a_1 x[n-1] + a_2 x[n-2] + \xi[n]$$

estimator : $$\hat{x}[n] = a_1 x[n-1] + a_2 x[n-2]$$

minimal error condition : $$\frac{\partial}{\partial a_i} \sum_n (\hat{x}[n] - x[n])^2 = 2 \sum_n x[n-i](x[n] - \hat{x}[n]) = 0$$

The linear prediction is a way to estimate the next value depends of one or more previous values. The calculations are done in a similar way than with the linear regression.

linear prediction

similar to linear regression!

what’s next ?

there is a lot of work!

Since we have a lot of outcomes it needs a bit time to finish. A first prototype is already done but this following tasks must be achieved to conclude my project :

- ~~merge the main graph with the graph with outliers~~
- add an error histogram in years/months
- ~~make the timline progress from today to the past~~
- ~~add the newspaper choice dropdown~~
- create computing methods for the others variables
- add the variable choice dropdown
- add the issue information (including link of issue) table with support of the selecting tool
- add the recomputing fonctionality of the linear prediction
- add checkbox to freeze the linear prediction

Questions ?

Thanks for listening