

Analyseur d'intervalles de dates de publication

1 Introduction

Mon projet se nomme « Metadata mining of large collections of historical newspapers » pour lequel j'ai accès à un certain nombre de métadonnées d'une collection de journaux d'Impresso, comme par exemple le nombre de caractères dans un article, le nombre et la taille des illustrations ou bien la date de publication des journaux.

Dans un premier temps, à l'aide de Maud Ehrmann et Matteo Romanello, ma réflexion s'est portée sur la création d'un outil assez général qui permet à son utilisateur d'analyser des variations. Dans un deuxième temps, j'ai orienté mon projet vers l'analyse des intervalles de publication entre deux parutions d'un même journal.

1.1 Présentation de l'outil

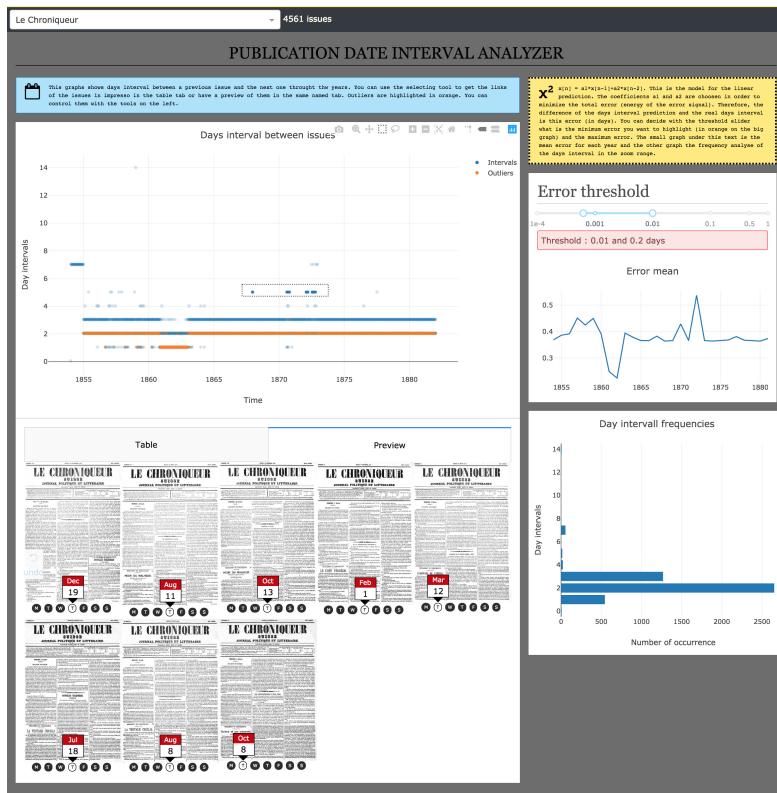


Figure 1: Analyseur de variations

Description (visuelle) de l'interface

L'interface de l'analyseur est composée de plusieurs espaces. La barre du haut permet de sélectionner le journal ainsi que de voir le nombre de parutions du titre. Ensuite, la partie grande à gauche contient le *graphe des apparitions* et un *prévisualisateur*. La partie droite contient la commande du *surveillage*, le *graphe de l'erreur moyenne de la prédition* et le *graphe de l'analyseur de fréquences*.

Description fonctionnelle de l'outil

Outil de vue d'ensemble : La vue d'ensemble se fait sur le *graphe des apparitions*. Sur ce graphe on voit toutes les revues publiées d'un journal sélectionné. En axe des abscisses représente la date de publication de la revue et en axe des ordonnées le nombre de jours qui séparent la parution d'une édition et la parution de l'édition précédente. Les points surlignés en orange sont ceux que l'on considère comme « outliers » du point de vue de notre *outil de prévision*. Le *graphe de l'analyseur de fréquences* va automatiquement se mettre à jour en fonction du niveau de zoom du *graphe des apparitions*.

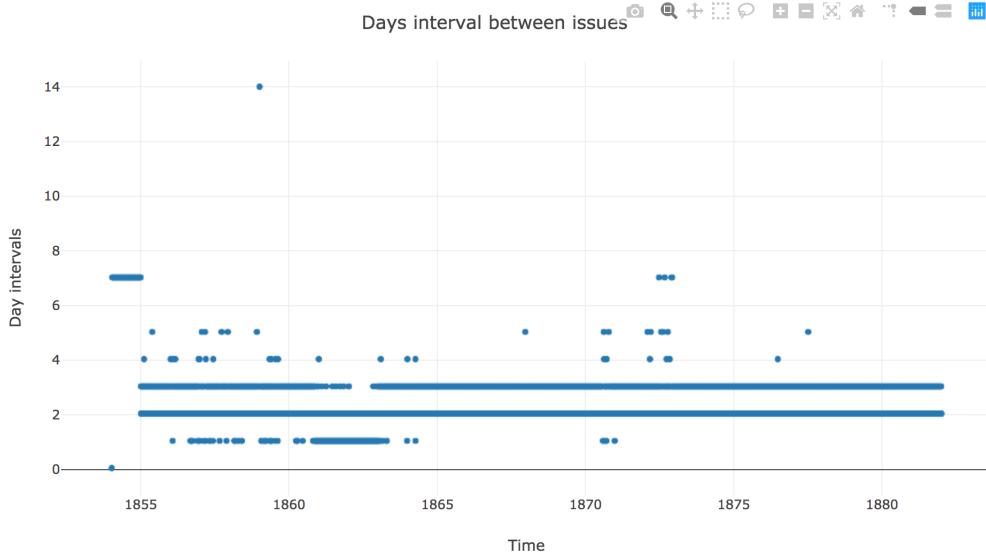


Figure 2: Outil vue d'ensemble

Outil de prévisualisation : À partir du graphe des apparitions on peut sélectionner des points pour les afficher dans le prévisualisateur. On peut voir les revues, soit dans un tableau, soit dans le prévisualisateur graphique. Celui-ci affiche une miniature de la première page ainsi que la date positionnée au-dessus du jour de semaine de publication.



Figure 3: Zoom sur la prévisualisation graphique

Outil de prédition : C'est l'instrument le plus abstrait de l'analyseur. Il s'agit de surligner des points qui semblent être les plus intéressants à l'aide des équations de prédition linéaire. Sur le graphe des apparitions on remarque rapidement les points dont l'intervalle de publication est très anormal, c'est-à-dire dont la valeur du point est particulièrement haute ou basse par rapport à la plupart des autres. Il paraît donc pertinent d'utiliser un modèle qui permet de détecter les variations anormales mais plus ou moins cachées dans des accumulations de points sur le graphe. L'idée est de calculer la moyenne des deux points précédents. Soit x_{n-1} et x_{n-2} deux valeurs d'intervalles de temps consécutives :

$$m_n = \frac{1}{2}x_{n-1} + \frac{1}{2}x_{n-2}$$

On définit m_n comme étant la prédition. Puis, on soustrait m_n à la valeur d'intervalle du point qui suit les deux autres publications pour ainsi calculer l'erreur :

$$|e_n| = |m_n - x_n|$$

Pour améliorer le modèle, j'ai décidé de rajouter des coefficients afin de faire une moyenne pondérée que l'on choisit de manière à minimiser la somme des carrés de l'erreur. On peut les obtenir par un simple calcul de dérivation.

L'erreur moyenne par année est affichée dans le graphe d'erreur moyenne de prédition afin de voir la stabilité du rythme de publication à travers les ans. Ensuite, on peut surligner les points dont la prédition m_n est trop éloignée de sa valeur (nombre de jours entre la publication de deux revues) avec le réglage d'un minimum de distance (avec l'erreur) et un maximum. La valeur minimale et maximale seront deux « Threshold ».

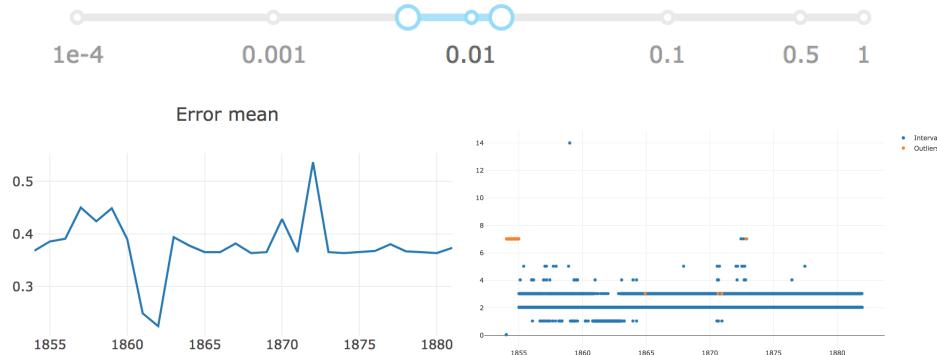


Figure 4: En haut, le sélecteur de threshold (logarithmique), à gauche le graphe de l'erreur moyenne, à droite les outliers sur le graphe des apparitions

2 Analyse qualitative

2.1 Analyse de l'erreur moyenne

À partir des graphes de l'erreur moyenne de la prédiction par année, on peut s'aventurer à interpréter la stabilité du modèle et donc si la prochaine date de publication est prévisible ou non. Ainsi, voici différentes typologies de ces graphes.

Point unique : il s'agit de journaux dont la durée de vie était de moins d'un an. Ce sont « Le Bulletin de la Séance Constituante », « Le Journal du Valais » et « La Tribune de Fribourg »

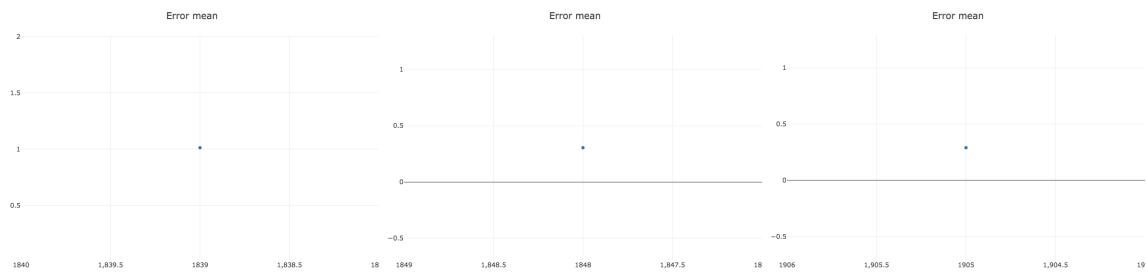


Figure 5: Mean error du Bulletin de la Séance Constituante, le Journal du Valais et la Tribune de Fribourg

Moins de 1000 points : Les titres ayant moins de mille publications sont « l'Écho des Alpes », « Le Journal de Fribourg », « Le Narrateur Fribourgeois », « Le Véridique » et « Der Landbote des freiburgischen Seebbezirks ». Sur les graphes ci-dessous on devine quelques caractéristiques comme les graphes en « U ». On peut émettre l'hypothèse que la date de publication était irrégulière au début et à la fin. Mais aussi les graphes convergents (publications plus régulières avec le temps), les graphes divergents (publications de plus en plus irrégulières) et un dernier graphe qui montre deux pics d'irrégularités.

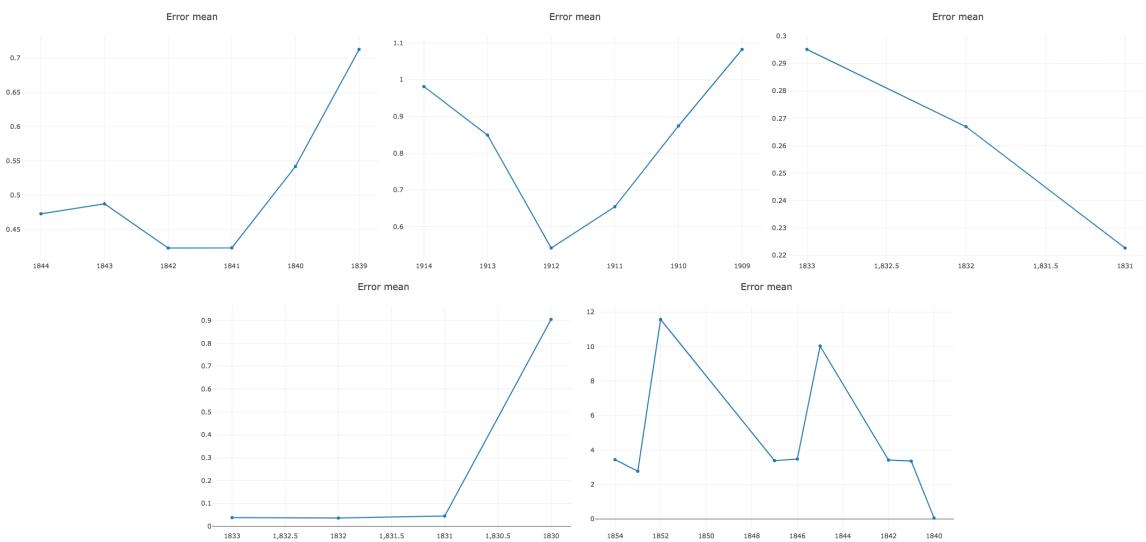


Figure 6: Mean error de l'Écho des Alpes, le Narrateur Fribourgeois, der Landbote des freiburgischen Seebbezirks, le Journal de Fribourg et le Véridique

Graphes convergents : Ici il s'agit de graphes ayant une convergence au cours du temps. Je les interprète comme des journaux dont la publication devient de plus en plus régulière. La « Gazette de Lausanne », « Le Journal de Genève », « l'Express », la « Neue Zürcher Zeitung » et « le Peuple, la Sentinelle » sont dans ce cas-ci. Comme la période d'instabilité est plus ou moins commune à tous les journaux on peut invoquer des raisons technologiques. À contrario des pics d'instabilité vers le milieu comme à la NZZ en 1880, Le Peuple, La Sentinelle en 1910 et le Journal de Genève en 1920 pourraient avoir des causes politiques.

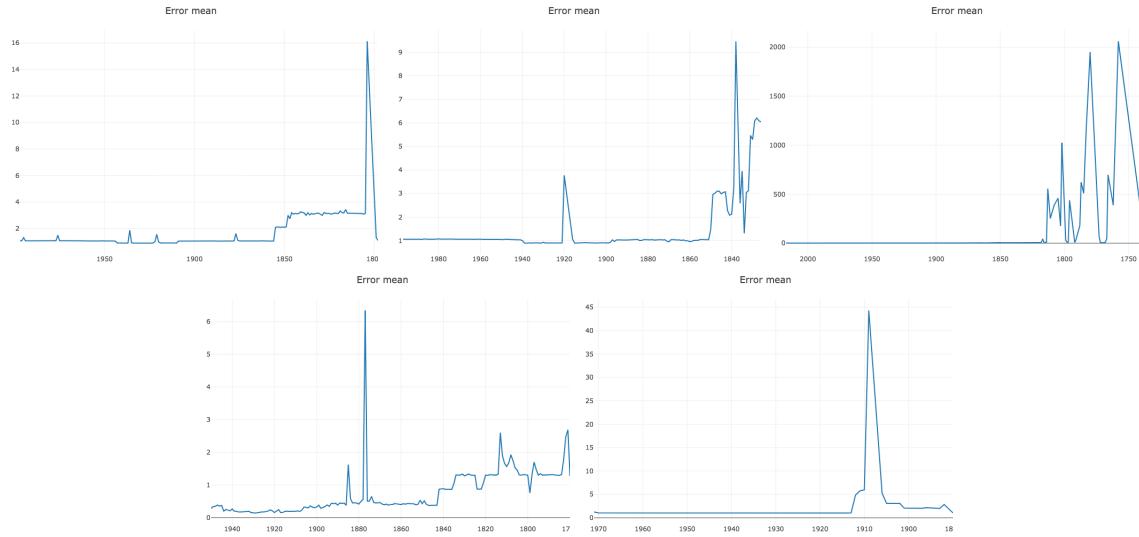


Figure 7: Error mean du GDL, du JDG, de l'Express, de la NZZ et le Peuple, La Sentinelle

Graphes divergents : « Le Confédéré », « L'Essor » et « L'Impartial » ont des graphes divergents. Au contraire des graphes convergents, leur graphe va de la stabilité à l'instabilité. Notons que l'échelle des temps est différente pour chaque graphe. On remarque que pour celui de l'Impartial, l'échelle est très petite. Donc notre prédiction est très précise.

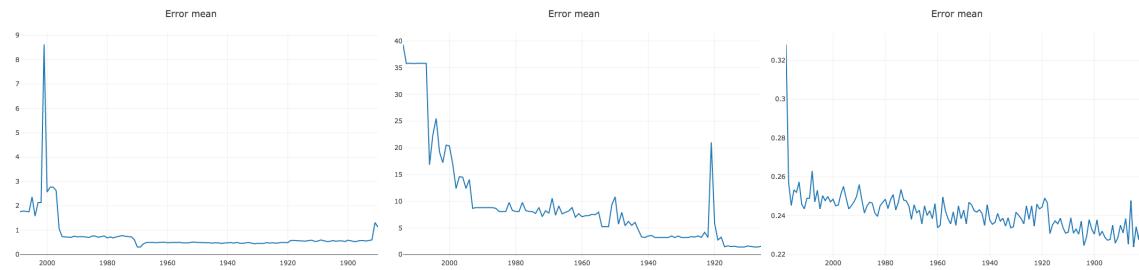


Figure 8: Error mean du Confédéré, de l'Essor et de l'Impartial

2.2 Analyse des intervalles d'apparition

Ici, comme analyse je propose de lier trois dates aux changements de fréquence d'apparition des journaux :

Année	Événement
1830	Invention du télégraphe
1876	Invention du téléphone
1895	Création de l'agence télégraphique suisse, ATS/SDA

Table 1: Événement clés du 19^e siècle

Voici différents graphes d'intervalles d'apparition avec les trois dates indiquées par des marqueurs verticaux puis le tableau qui recense les changements significatifs de fréquence suite à ces événements :

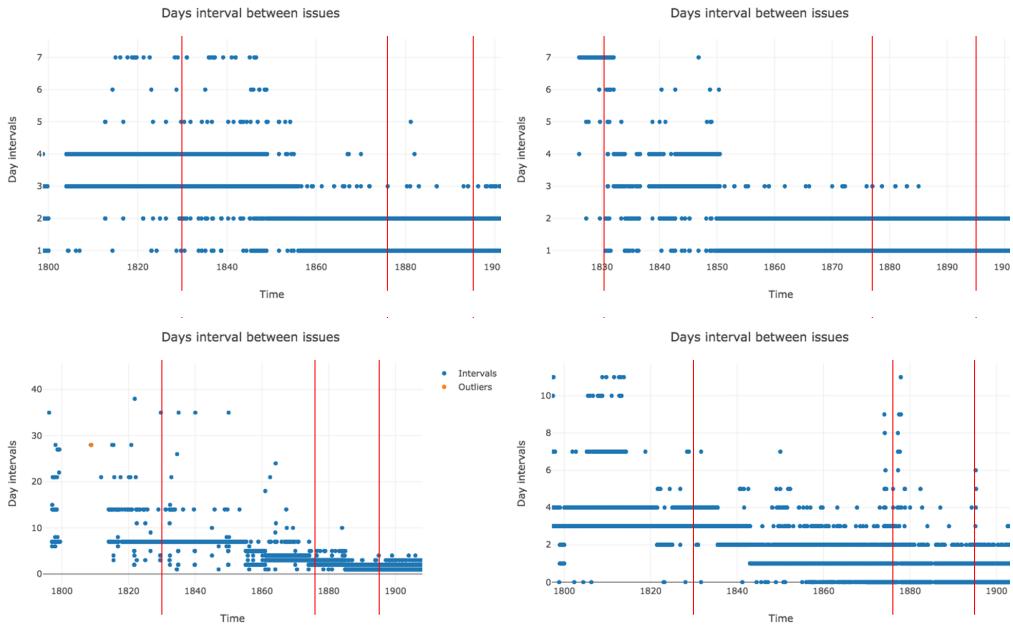


Figure 9: Intervalle de parution du GDL, du JDG, de l'Express et de la NZZ

Journal	$\Delta t_{telegraph}$	Δt_{phone}	Δt_{ats}	$t_{telegraph}$	t_{phone}	t_{ats}
GDL	19	34	15	1849	1910	1910
JDG	2	22	3	1832	1898	1898
EXP	25	8	-	1855	1884	-
NZZ	5	5	-	1835	1881	-

Table 2: Délais de réaction

On voit que le temps de réaction est à l'invention du téléphone du Journal de Genève et de la NZZ est particulièrement rapide. Suite à l'invention du téléphone ce seront l'Express et la NZZ. Ces deux mêmes journaux seront en revanche insensible à la création de l'agence télégraphique suisse. Contrairement au Journal de Genève et à la Gazette de Lausanne.

Cette analyse pourrait aussi être faite avec le nombre d'articles par journal, le nombre d'illustrations ou de pages par journal.

2.3 Analyse des miniatures

On remarque qu'il y a une certaine quantification des valeurs en Y sur le graphique des apparitions. Les étages supérieurs ont moins de publications. Ce qui est pratique pour sélectionner un nombre raisonnable de journaux et ainsi par dichotomie voir les changements graphiques sur la une. Voici un exemple avec l'Impartial :



Figure 10: Liste des 1ères pages de l'Impartial

Voici la liste des événements liés au changement graphique du journal :

Nom	Date	Événement
image 1	4 janvier 1881	Première publication
image 2	29 mars 1912	Photos de plus en plus fréquentes dans les journaux
image 3	4 octobre 1912	Dernière apparition des horaires de train en-dessous du titre
image 4	28 février 1948	Nouvelle charte suite au centenaire du canton de Neuchâtel
image 5	3 avril 1967	Nouvelle charte après fusion avec la Feuille d'Avis des Montagnes
image 6	23 septembre 1981	Nouvelles formules (après compte à rebours)
image 7	31 aout 1987	Passage à la couleur
image 8	20 mars 2001	Nouvelle formule pour distinguer l'Express et l'Impartial
image 9	6 février 2007	Création de la marque Arc Presse

Table 3: Changements de maquette de l'Impartial

2.4 Analyse des outliers

Les marqueurs verticaux rouges montrent que certains points supérieurs sont en quelques sortes projetés sur un segment horizontal inférieur (ou supérieur). On va prendre ceci comme une propriété de mon modèle mathématique qui dit que si x_{n-1} et/ou x_{n-2} a une grande valeur alors on prédit que le point suivant a également une valeur plus élevée de part mon modèle $m_n = \frac{1}{2}(x_{n-1} + x_{n-2})$. Donc le point suivant est indiqué comme anormal.

Les points surlignés qui sont à la position où j'ai mis les marqueurs verticaux ne sont pas très utiles car on voit très clairement qu'il y a des points anormaux qui dépassent la valeur moyenne. Donc on peut baisser la valeur maximale du threshold error sur le curseur latéral.

En revanche, le point que j'ai encerclé en rouge, par exemple, est beaucoup plus utile car à ce niveau de zoom on ne voit pas pourquoi il est spécial : s'agit-il d'un jour férié ? D'une grève des journalistes ? D'un manque de papier ? D'un événement spécial ?

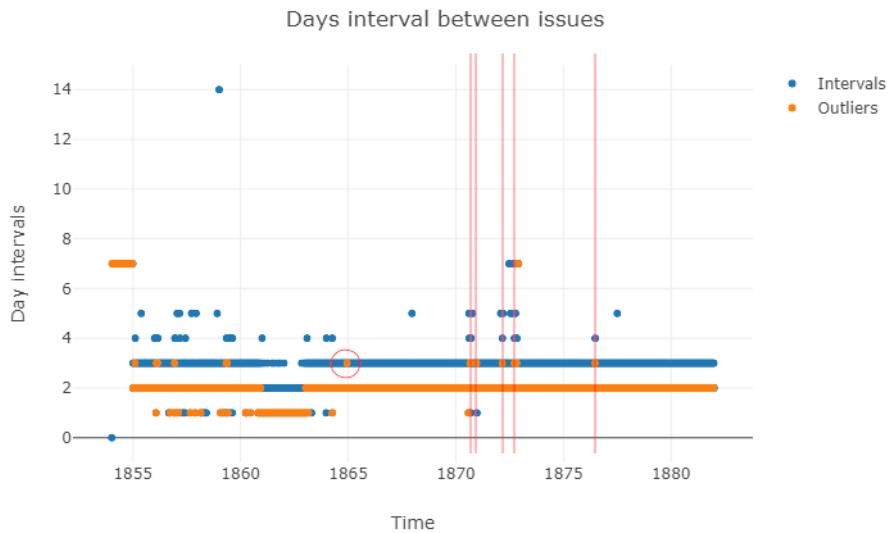


Figure 11: Graphe des apparitions du Chroniqueur

Cas de figure : l'Impartial

Le journal « L'impartial » a ceci de particulier qu'il a été publié sur une longue durée (1881 à 2018), que l'erreur de prédiction moyenne est très petite (moins 0.25 jours) et que ses jours de publication sont très réguliers (on voit très bien sur le graphe des apparition trois lignes horizontales qui apparaissent).

Pour illustrer l'utilisation du curseur de threshold on va le régler à 0.1 jours pour le minimum et 0.16 jours pour le maximum de l'erreur entre notre modèle et la vraie courbe. On remarque qu'il y a un motif régulier qui se forme. Pour avoir une meilleure impression j'ai fait un zoom sur les années 1900-1920 et j'ai encerclé les points en question. Il s'agit, la plupart du temps, les jours fériés de Noël et du 1er mars (création du canton de Neuchâtel). Beaucoup de 25 décembre et 1er mars sont détectés, mais pas tous.

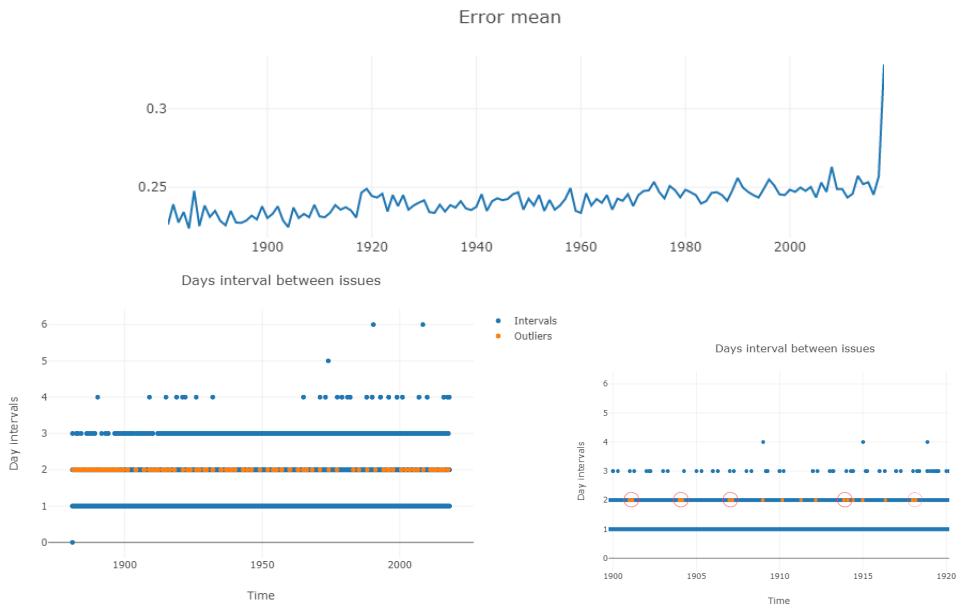


Figure 12: Error mean, graphe des apparitions complet et de 1900 à 1920

En revanche, si on a un peu l'œil on s'approche en 1922 d'un outliers un peu spécial. Il s'agit du début de la grève des typographes du 1er au 15 décembre 1922 qui constraint trois journaux (*L'impartial*, *L'effort* et *La Feuille d'Avis des Montagnes*) de publier sous une maquette commune leur journal. Sur les images ci-dessous on voit la maquette commune et le mot de fin de la grève ainsi que le point, un peu caché parmi les autres.

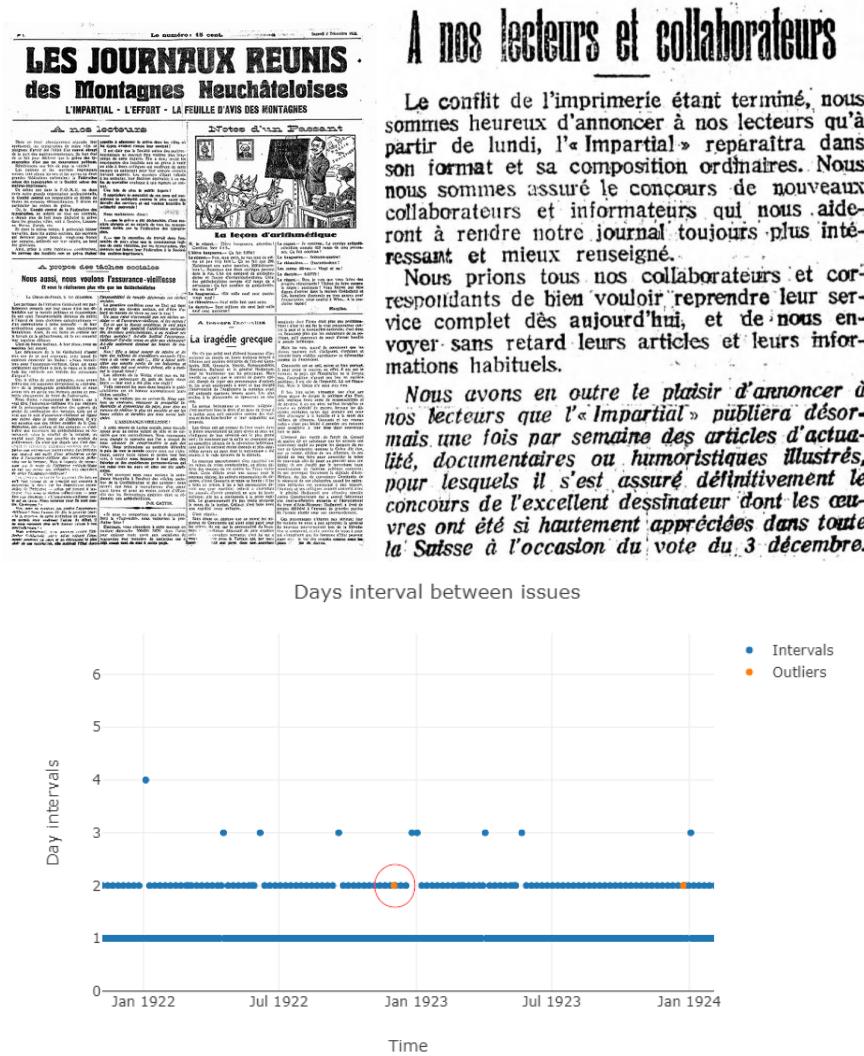


Figure 13: Maquette du journal lors de la grève, mot de fin et point particulier

3 Conclusion

Il serait intéressant d'utiliser le même outil avec d'autres métadonnées tel que le nombre d'illustrations, de caractères, de pages, etc. Toutefois, on voit que l'on peut déjà découvrir quelques informations à propos de chaque journal. On remarque qu'il est aisément de voir sa stabilité de publication à travers le temps, de mesurer l'effet d'une invention sur la fréquence de parution, de chercher les grandes étapes graphiques d'un journal ou avec un peu de doigté, repérer quelques points anormaux !

Pour finir, dans le cadre de ce travail que j'ai mené durant six mois pour mon projet de bachelor à l'EPFL, j'aimerais remercier tout chaleureusement Maud Ehrmann et Matteo Romanello pour leur soutien et la liberté dont j'ai pu bénéficier.