Laurent Lavoie

Répartition des auteurs selon trois médias en 2018

Mon projet de données consistait à évaluer la proportion d'articles rédigés par des agences par rapport la proportion d'articles rédigés « maisons » au sein de trois journaux écrits différents : La Presse, Le Devoir et Le Journal de Montréal. Je me suis intéressé à cette analyse puisqu'on nous répète depuis maintenant plusieurs années que les médias sont aux prises avec une crise financière, mais aussi aux prises avec leur adaptation à la montée des réseaux sociaux. À l'ère de l'instantanéité et de la vitesse, les médias se tournent régulièrement vers des agences de presse telles que l'Agence France-Presse, La Presse canadienne et l'Associated Press. C'est donc pourquoi j'ai désiré extraire l'entièreté des articles de toutes les sections, y compris l'opinion, des trois journaux et d'analyser la composition de l'ensemble des auteurs.

Pour chacun des médias sélectionnés, j'ai décortiqué une façon d'extraire tous les articles de la période établie, et ce, à l'aide du logiciel Sublime Text. La construction des sites Web varie selon la mise en page, j'ai donc dû adapter trois codes différents qui iraient chercher les données recherchées. Dans le cas du Devoir, ma première tentative consistait à passer au travers de chacune des pages de la section « Recherche » qui me permettait de choisir précisément la période que je désirais analyser. Cependant, le code rencontrait plusieurs problématiques, puisqu'il m'était difficile d'éviter certaines informations qui n'étaient pas pertinentes aux données recherchées. Par exemple, certains liens étaient dédiés au regroupement de textes d'un auteur précis, ou encore, le code accrochait sur l'URL des articles « La parole à nos lecteurs ». J'ai donc pris tenter d'utiliser un autre lien pour faire mon scrapping. Le lien était le suivant : http://m.ledevoir.com/article-{numero}. Chaque article du Devoir a un numéro qui l'identifie. Pour cibler les bonnes dates, j'ai trouvé les premier et dernier articles publiés au cours de l'année 2018 et j'ai adapté un code qui analysait tous les articles. En ciblant les sections « meta » et les boîtes qui contenaient l'information désirée, j'ai pu extraire les données recherchées.

Pour ce qui est du Journal de Montréal, chaque section du journal contenait une section archives de chaque jour de l'année. J'ai construit une boucle qui passait dans chaque section d'archives de tous les jours de l'année et dans tous les articles. Le site du Journal étant bien construit, je n'ai eu qu'à extraire les données « meta ».

Finalement, pour La Presse, la tâche s'est avérée plus complexe. D'abord, les archives étaient très bien rangées par journée, et les articles n'étaient séparés selon leur section, ce qui m'a facilité la tâche, puisque la construction du code est plus simple, ne nécessitant pas une autre boucle. La tâche fut légèrement ardue pour extraire les auteurs. Les pages HTML des articles présentent des boîtes d'auteurs très différentes si c'est une agence de presse, un collaborateur ou un journaliste « maison ». J'ai alors créé plusieurs variantes dans mon code qui allaient chercher les informations nécessaires dépendamment de la page HTML.

De façon générale, j'ai dû faire certains choix dans la sélection de mes articles. Les cas de figure variaient selon les médias. Du côté du Devoir, le code plantait sur les longs formats présentant des galeries photos, La Presse avait de nombreux textes qui étaient retirés ou encore on comptait les pages sommaires des matchs de hockey du Canadien qui n'étaient pas des articles en soi. Au Journal de Montréal, quelques URL étaient problématiques et des articles étaient également supprimés. Bref, les nombreuses variantes ont fait en sorte que plusieurs liens étaient écartés, mais au volume ça n'avait pas d'impacts concrets sur les données.

Du côté des auteurs, certains textes présentaient des collaborations entre un auteur maison et une agence ou encore des collaborations entre deux agences. À l'aide d'un tableau croisé sur Google Sheet, je pouvais faire le calcul des auteurs, et pour avoir la méthodologie la plus tranchée possible, j'ai décidé que je comptabilisais seulement le premier auteur pour faire les répartitions. Par exemple, une signature « Marie-Michèle

Sioui, La Presse canadienne » comptait comme un texte du Devoir, ou encore « AFP, Associated Press » comptait comme un texte de l'AFP.

L'ensemble des URL consultés :

******Dossier Google comportant le reste des mes fichiers :**

https://drive.google.com/drive/folders/1Vte3lSG6ULBcqNsmXvSK1YtdBXnlnAOV?usp=sharing

Cas de figure :

[Grand format] **https://www.ledevoir.com/monde/544582/photos-de-l-annee**

Page de recherche du *Devoir* :

https://www.ledevoir.com/recherche?expression=&rechercher=

Exemple de page d'archives du JDM :

https://www.journaldemontreal.com/actualite/archives/2019/01/31

Exemple de page d'archives de *La Presse* :

https://www.lapresse.ca/archives/2018/1/1.php

Exemples de boîtes d'auteurs de *La Presse* :

https://www.lapresse.ca/sports/tennis/201801/01/01-5148727-sharapova-et-halep-accedent-au-tour-suivant-en-chine.php

https://www.lapresse.ca/actualites/national/201801/01/01-5148720-deux-petites-filles-sont-les-1ers-bebes-de-lannee-a-montreal-et-quebec.php

Site utilisé pour la visualisation :

https://www.datawrapper.de/