GLO-7027

RAPPORT 1 : ANALYSE ET PRÉTRAITEMENT DES DONNÉES

February 24, 2021

Adrien TURCHINI (cycle 2)
Marc HENRIOT (cycle 2)

Université Laval

1 Analyse des propriétés statistiques des données

1.1 Jeu de données analytics

Afin de se faire une première idée sur les données, il nous faut pouvoir les ouvrir dans python3. Pour se faire, nous utilisons Pandas afin de convertir les json en DataFrame. Pour cette étude statistique, il a été choisi d'ouvrir 10 json pour chaque journal ce qui fait un total de 60 json soit environ 4.7% des données. Sachant que tous les fichiers ne font pas la même taille, cela peut différer.

1.1.1 Étude du score

Le but de cette partie est de voir comment les articles sont consommés et à quelle fréquence. Afin de mieux comprendre les habitudes des internautes.

En utilisant les systèmes de notation de CN2i, on peut attribuer a chaque article une note.

	score
count	4.771000e+04
mean	9.482500e+02
std	1.283816e+04
min	1.000000e+00
25%	4.000000e+00
50%	1.900000e+01
75%	4.200000e+01
max	1.963752e+06

Figure 1: Detail des scores

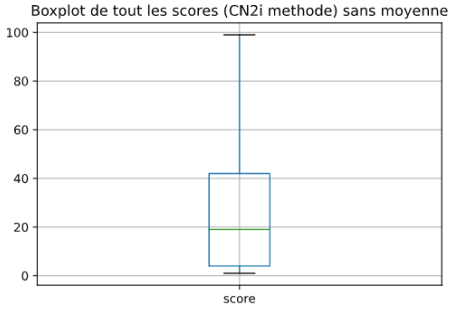


Figure 2: Boxplot sur score

Ici, nous avons accès aux statiques des scores. Un problème flagrant et que la variance est de 12 838 pour une moyenne de 948. Cela montre que les valeurs sont très inégalement reparties, avec des valeurs très grandes (une valeur maximum de 1 963 752).

Étant donné cette disparité dans les données et que seul les pires 10% des article nous intéressent nous choisissons de regarder les premier 87% des donnée. Ce 87% est obtenu en ne gardant que les valeurs inférieures à un maximum, c'est la borne supérieur de la *Figure 2*. Ou en calculant tel que,

$$max = Q3 + 1.5 * (Q3 - Q1)$$

Cela nous permet donc de tracer de nouvelle figure sur ces 87% restant.

Au final, la médiane n'a pas beaucoup baissé ce qui montre que la répartition des données reste quasi-inchangée. En revanche la moyenne et passé de 948 à 20. Nous avons donc des données plus correctement repartie et donc plus lisible.

La distribution *Figure 4* nous permet d'avoir une meilleure idée de la répartition des scores.

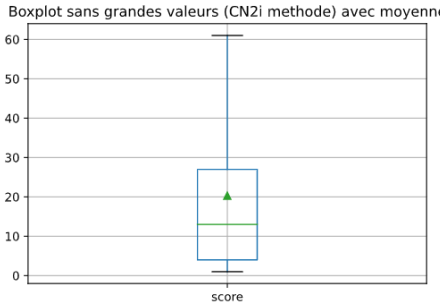


Figure 3: Box plot sur 87% de score

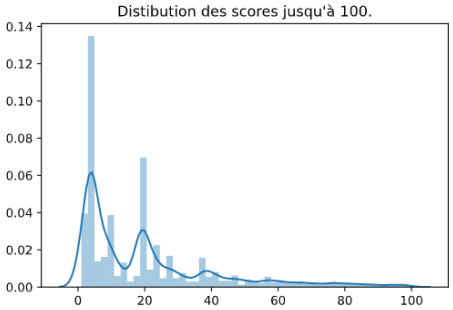


Figure 4: Distribution des score

En conclusion, il est évident que ces statistiques ne sont pas représentatives de l'ensemble des données. Mais elles permettent de se faire une idée globale de la distribution des données. Nous pourrons surtout retenir qu'un faible pourcentage représente la majorité des vues et que tout le reste se répartie un nombre de vues assez équitable. Il sera donc intéressant dans la suite de regarder pourquoi ces articles sont aussi bien placés.

1.1.2 Problème de dimensionnalité

L'analyse d'un tel jeu de données pose un problème majeur qui reste pour l'instant irrésolu, c'est le souci de la mémoire mise en jeu lors de l'ouverture des json dans analytic qui représente 100Gb il faut donc impérativement trouver une solution pour pouvoir ouvrir 50% voir 100% du jeu de données. Une piste envisagée serait de réduire le nombre de lignes des documents en appliquant un groupement par hash en sommant les autres colonnes afin qu'il ne reste que des hash unique, mais cela mènerait à une perte d'information de la colonne 'name' et 'product'. La colonne 'name' et facilement transformable en entier, mais ce n'est pas le cas de la colonnes 'product'.

1.1.3 Données manquantes

On retrouve certaines données manquantes principalement dans le jeu de données des publications ou la source peut être manquante par exemple. Cependant la plupart des données sont complètes. Nous remplacerons les valeurs manquantes en NaN. Selon les algorithmes et bibliothèques utilisés comme scikit learn par exemple, il est possible de prendre en compte ces valeurs, de les remplacer par la moyenne ou en prédisant leur valeurs en réalisant une régression. Selon la proportion de valeur manquantes pour des attributs, il peut être nécessaire de drop ce dernier si il y a une proportion trop importante de données manquantes.

1.2 Jeu de données des publications

Nous avons par la suite réalisé une analyse des données des articles. Pour des soucis de taille de fichiers et de mémoire disponibles sur nos ordinateurs nous avons réalisé cette analyse sur le jeu de test qui, malgré un nombre moindre de données comporte tout de même 5230 articles.

1.2.1 visual

Nous retrouvons donc 3 types de visuel au sein des articles 5210 articles contenant une photo ainsi que 10 articles contenant une vidéos et 10 articles contenant un slideshow. Tous les articles publiés ont donc un visuel et 99,6% d'entre eux ont une photo comme visuel. De même pour le contentType, on retrouve hormis les vidéos qui sont au format youtube seulement des images. Parmi les 5230 articles, 4122 ont une caption sur leur visuel.

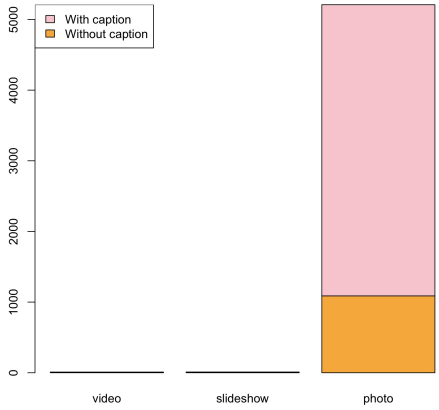


Figure 5: Détail des données

1.2.2 templateName

Il y a 5 différents types de template name, donc d'affichage du contenu selon ce dernier. On retrouve "article3", "article_columnist", "article1", "cartoon1", "sponsored1". On a environ 46% d'article3, 5% d'article_columnist, 46% également d'article1, 2% de cartoon1 et 1% de sponsored1. Les template d'articles (articles1 et articles3) représentent donc près de 90% des articles de nos données.

1.2.3 availableInPreview

Tous les articles contiennent la variable false pour ces attributs, leur contenu n'est pas disponible gratuitement.

1.2.4 creationDate et modificationDate

La date d'ajout de l'article peut-être intéressante. Nous avons des articles depuis 2013 mais seulement les données analytiques de consultations des articles de janvier à juillet 2019. Il est important de prendre cela en compte par la suite afin de ne pas fausser notre modèle car le score est calculé en fonction du nombres de vues et un article ancien sera moins bien référencé et en sera pénalisé.

1.2.5 authors

L'attribut auteur est intéressant et fourni une bonne quantité d'informations. Cependant cet attribut n'est absolument pas normalisé. On retrouve 535 auteurs différents par leur nom, une moyenne d'environ 10 articles par auteur pour le jeu de test mais une médiane de 1. Aussi les noms des auteurs peuvent comprendre des valeurs manquantes, comprendre plusieurs auteurs pour une même publication, ou encore la fonction de l'auteur comme professeur ou député. Il est a noter que certains articles sont écrits par plusieurs auteurs en même temps.

On retrouve 96 sources différentes pour les auteurs donc pour une moyenne de 65 articles par sources. Cependant certaines sources n'ont qu'un seul article comme "Université du Québec à Trois-Rivières", ces sources avec moins de 100 articles représentes 8% des articles, d'autres sources sont manquantes, cela représente 9% des articles. On peut voir sur la *Figure 6* la distribution des sources.

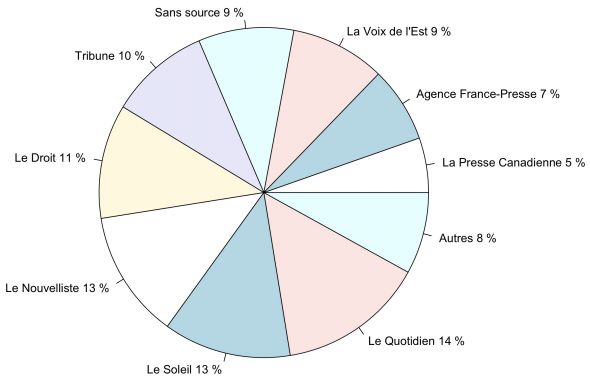


Figure 6: Source des articles

1.2.6 channel

On retrouve 334 channel différents, par exemple "Justice et faits divers" on retrouve 308 articles, ou encore "F1" avec 7 articles et "Patrick Duquette" avec 15 articles. On remarque qu'il y a un grand nombre de catégories d'articles pour le nombre total d'articles. Lorsqu'on analyse les données on remarque qu'on retrouve des catégories très proches, par exemple "Justice et faits divers" et "Faits Divers". Evidemment les deux catégories sont intimement lié mais selon le journal, le mode du publication, la personne qui le publie, on se retrouve avec des noms différents. On pourrait donc regrouper certaines catégories en une seule afin de permettre à notre modèle de mieux généraliser.

1.2.7 lead et chapter

Le lead représente un résumé de l'article qui apparaît au début de ce dernier. On retrouve 4940 lead différents pour 5230 articles. La plupart des articles en ont donc un mais un peu plus de 5% s'en passent.

Pour les chapitres on retrouve des articles avec un grand nombre de chapitres comme d'autres avec un seul chapitre ou pas de chapitre du tout. Ci-dessous on peut retrouver le nombre de chapitres pour 20 articles tirés aléatoirement.

```
{ "_id" : ObjectId("6035215535f43b7d8468f39e"), "nbItems" : 1 }
{ "_id" : ObjectId("603521554a2c9a6c367dde3c"), "nbItems" : 1 }
{ "_id" : ObjectId("60352155280c0e3a1c4f21ab"), "nbItems" : 1 }
{ "_id" : ObjectId("603521558f43426bb99802ad"), "nbItems" : 7 }
{ "_id" : ObjectId("60352155732b99814107f750"), "nbItems" : 24 }
{ "_id" : ObjectId("6035215502ecddefbe403c7b"), "nbItems" : 1 }
{ "_id" : ObjectId("6035215555222b6050b3617a"), "nbItems" : 1 }
{ "_id" : ObjectId("603521550a8e3286040b46aa"), "nbItems" : 2 }
{ "_id" : ObjectId("60352155fbe2bb597c08d613"), "nbItems" : 1 }
{ "_id" : ObjectId("60352155be41c922ca17f527"), "nbItems" : 18 }
{ "_id" : ObjectId("6035215555fe8a713dbec9f0"), "nbItems" : 1 }
{ "_id" : ObjectId("603521559668c28a50a51919"), "nbItems" : 1 }
{ "_id" : ObjectId("60352155f7fb81cc51ad5d94"), "nbItems" : 2 }
{ "_id" : ObjectId("60352155b06eee4204112dc2"), "nbItems" : 9 }
{ "_id" : ObjectId("603521558e9a9d0e610f6a39"), "nbItems" : 10 }
{ "_id" : ObjectId("60352155ab92821871594938"), "nbItems" : 10 }
{ "_id" : ObjectId("6035215583b00934e505315e"), "nbItems" : 6 }
{ "_id" : ObjectId("60352155bfde3b24af19911d"), "nbItems" : 7 }
{ "_id" : ObjectId("60352155d45bc0a36131fdec"), "nbItems" : 19 }
{ "_id" : ObjectId("603521557dcecc4d68b71140"), "nbItems" : 9 }
```

Figure 7: Detail des données

1.3 Jeu de données des informations de publications

Nous avons à nouveau utilisé les fichiers situés dans le dossier de test afin d'analyser les données. Après importation dans une base de données mongo, nous nous retrouvons avec 23813 données d'emplacement sur les publications pour 5320 articles. On a donc une moyenne de 4,47 publications par articles. Certains articles sont publiés sous un editionId différent dans des organisations différentes mais aussi plusieurs fois dans les mêmes organisations. La moyenne

représente bien les données, on ne retrouve pas de valeur aberrantes, aucun article n'est publié un très grand nombre de fois.

1.3.1 type

On retrouve 3 types différents: mobile, site et external. Le type nous indique sous quelle forme l'article a été publié. On à 12385 pulications sur mobile, 11398 sur site et 30 externes.

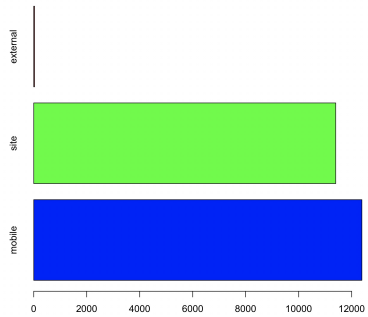


Figure 8: Formats des publications

1.3.2 organizationKey

L'organisation key nous indique dans quel journal la publication apparaît. On retrouve donc le journal la Tribune, le Droit, le Soleil, la Voix de l'est, la Nouvelliste et le Quotidien. On peut voir dans la *Figure 12* la répartition de la publication des articles. Chaque journal publie à 2% près le même nombre d'articles.

2 Prévision des attributs gardés pour l'algorithme de traitement de données

Nous choisissons de garder les attributs suivants pour la suite : *pulications.date, publication.slug, type* pour les informations de publications, *channel, authors, lead, chapters, visual, templateName* pour les publications et *hash, name, product* les évènements des journaux.

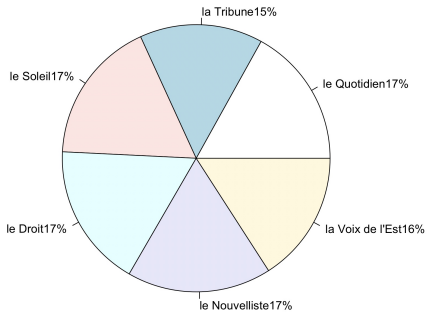


Figure 9: Répartition de la pulication des articles parmi les journaux

2.1 Choix des attributs pour les informations de publications

2.1.1 publications.date

La date de chaque publication d'article est importante, elle peut nous permettre d'analyser l'impact d'une publication sur le nombre de vues, et l'important de publier un article plusieurs fois. Aussi il sera possible de prendre en compte le jour, le mois et l'heure de la publication afin de pouvoir prédire des périodes de temps propices à une bonne visibilité.

2.1.2 publications.type

Dans notre jeu de donnée utilisé pour l'analyse nous n'avons que des articles, cependant il est possible que cela ne soit pas le cas pour toutes les données que nous traiterons par la suite. Nous choisissons donc de mettre cet attribut de côté mais de manière non définitive selon les données totales qui seront traitées par la suite.

2.1.3 publications.slug

Le slug, formaté afin de récuperer la catégorie de la publication en question nous donne ensemble des informations sur la catégorie de l'article et nous permet donc de définir certains critères avec ce denier. En effet certaines catégories sont plus attirantes que d'autres, afin de ne pas biaiser notre prédiction, il est important de prendre en compte la catégorie de l'article lorsque nous cherchons à prédire son succès. Un article scientifique sur les commutateurs optiques serait très intéressant pour un grand nombre de chercheurs et d'industriels sans pour autant attirer le grand public qui est plus important en terme d'audience.

2.1.4 type

La format ou type de publication de chaque article est important. En effet une minorité d'articles sont postés en external, cela pourrait avoir un impact sur leur popularité. Aussi certains articles sont publiés plusieurs fois via le site, ou via mobile. Un des types pourrait avoir un impact sur le nombre de vues d'une publication et est donc à prendre en compte.

2.2 Choix des attributs sur les publications

2.2.1 channel

Tout comme le slug dans publication-info, le channel nous donne de l'information sur la catégorie de l'article et nous permet de classer celui-ci avec des articles de sujets similaires. Les slug et les channel n'étant pas forcément les mêmes il sera important de comparer les deux afin de classer les articles dans les catégories les plus globales afin de réduire le nombre de catégories possibles et de bien prédire les score des articles.

2.2.2 availableInPreview

Comme énoncé précedemment, tous les articles contiennent la variable false, cependant il est possible que les données de test ne représentent pas la totalité des données, certains moins auraient pu être gratuit mais ne sont pas présents dans le jeu de test. Selon nos observations nous ne garderons pas cet attribut, cependant il ne sera pas oublié pour autant et selon la totalité des données traités par la suite, il deviendra peut-être nécessaire de l'intégrer au modèle.

2.2.3 authors

Certains auteurs ont écrits plusieurs articles, d'autres moins. La popularité d'un article peut découler de la popularité de son auteur (authors.name), et un auteur productif et très publié pourrait écrire des articles plus populaires qu'un autre moins publié. Il est donc important de prendre cette variable en compte. De plus selon les sources des auteurs (authors.source), certains articles pourraient également être plus ou moins populaires, que d'autres selon leur provenance (une certaine université ou un certain pays ou sans source).

2.2.4 lead

Nous avons vu que 95% des articles ont un résumé, cependant cela peut-être un critère pour le lecteur d'avoir aperçu de ce qu'il va lire afin de choisir ou pas de continuer sa lecture. Cela peut avoir un imapact significatif sur la durée passée sur l'article.

2.2.5 chapters

Les articles peuvent avoir 0 chapitres comment 20, la taille de l'article peut également déprendre du nombre de chapitres selon la taille de ces derniers et un article avec beaucoup de chapitres pourra traiter des thèmes différents ou pour un article scientifique par exemple analyser sous plusieurs angles un thème. Nous choisissons donc de garder cet attribut car le nombre de chapitres peut jouer sur la popularité d'un article.

2.2.6 visual

Au sein d'un article, le visuel peut faire partie intégrante de ce dernier, on peut par exemple penser à la photo de Pete Souza, "The Situation Room" durant l'opération contre Oussama ben Laden. Cette dernière montre des émotions qui ne peuvent être retranscrites par texte. Dans d'autres cas, un visuel peut simplement accompagner un article et le rendre ainsi plus vendeur, inciter les gens à cliquer dessus.
Aussi on a vu que certains visuels possédaient une légende et d'autres non. Cela pourrait inciter les gens à lire la totalité de l'article que d'avoir un visuel décrit. Il est donc nécessaire de garder cet attribut.

2.2.7 templateName

Le template est important, en effet selon l'utilisation de ce dernier, les publications peuvent-être plus ou moins visibles aux lecteurs et plus facilement lisibles et entraînantes. De plus l'attribut sponsored1 est porté par peu de publications mais ce dernier pourrait avoir un rôle à jouer dans la popularité des articles car peut-être plus vu que les autres mais aussi moins intriguant si le fait que le lien sponsorisé est écrit.

2.3 Choix des colonnes sur analytics

	hash	name	product	createdAt
count	18795935	18795935	18795935	18795935
unique	47710	5	12	17521147
top	23bd9902e6daddfb903151ec265284a6	View	lenouvelliste/mobile	2019-01-08T12:10:27.408Z
freq	573864	7320933	2761816	7

Figure 10: Détail des données

Un première aperçu des données nous permet de voir quelles variables sont importantes. Ainsi seulement les variables hash, name et product sont intéressante. La date de consultation n'étant pas prédictive sur la qualité de l'article.

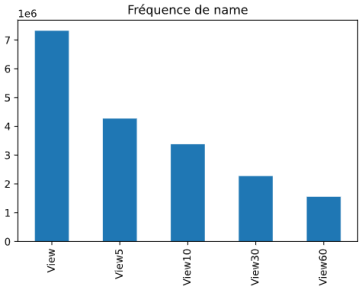


Figure 11: Fréquence sur name

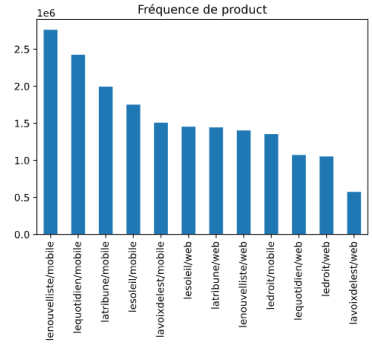


Figure 12: Fréquence sur product

Nous pouvons déjà voir quel journal performe le mieux en terme de lecture, et quelle sont les habitudes de lecture des internautes.

3 Prévision du traitement des données d'un point de vue pratique

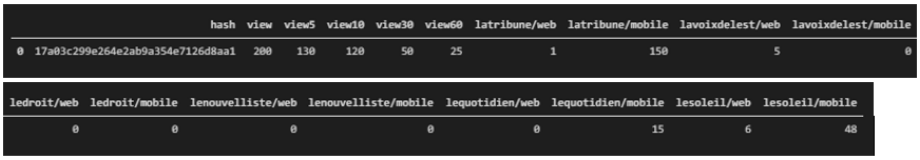
3.1 Traitement des données

Une première approche serait de réduire les dimension des données, sur analytic comme discuté plus haut et sur les articles en ne gardant que les clés qui nous intéressent.

3.1.1 Analytic

Nous envisageaon d'ouvrir un à un les json pour réécrire dans un autre json ce qui nous intéresse, de se fait nous préservons l'intégrité des données au cas où un mauvais traitement est fait. Avec les 18 795 935 de lignes ouvertes qui représentent 4,7% des json nous estimation 400 millions de lignes au total. Nous savons qu'il y a 5 230 articles dans le jeu de données, nous pouvons donc ramener analytic à ce même nombre de lignes. Nous pouvons changer la forme du tableau analytic (400 M x 4) en (5230 x 18) comme *Figure 13* ci-dessous. De cette façon, nous aurons un tableau uniquement composé d'entier très facile à utilisé et très léger à manipuler. De cette façon nous pourrons utiliser des méthode tels que l'ACP ou la sélection de variable (stepwise) pour faire un deuxième tris dans les données.

(Les données de la *Figure 13* sont uniquement là à titre d'exemple, elles ne sont pas représentatives des données.)



	hash	view	view5	view10	view30	view60	latribune/web	latribune/mobile	lavoixdelest/web	lavoixdelest/mobile
0	17a03c299e264e2ab9a354e7126d8aa1	200	130	120	50	25	1	150	5	0

ledroit/web	ledroit/mobile	lenouvelliste/web	lenouvelliste/mobile	lequotidien/web	lequotidien/mobile	lesoleil/web	lesoleil/mobile
0	0	0	0	0	15	6	48

Figure 13: Nouvelle forme d'analytic

3.2 Discussion sur les algorithmes de prévision des scores

Une première approche envisagé est de réaliser un clustering sur les données des articles afin de repérer des groupe d'article similaire. Un premier modèle pourrait être créé sur cette base.

Par la suite, un modèle à l'aide d'un SMV pourrait prédire le score potentiel d'un article en fonction des colonnes retenues.

Enfin, un modèle plus complexe de traitement du langage naturel pourrait arriver à fournir les mêmes prédictions. En ayant l'avantage d'être plus flexible. Ce qui en fait un modèle plus aventureux étant donné que les données sont essentiellement du texte.

4 Discussion de la procédure de tests

Afin de pouvoir entraîner notre modèle, il sera important d'avoir des données permettant de refléter la qualité de l'article. Dans notre cas, ces données se trouvent dans analytics où pour chaque article vue, nous pouvons calculer une note, comme fait dans la section de l'étude des scores.

Dans un premier temps, nous pourrons reprendre la méthode CN2i afin d'attribuer une note à chaque article. Une fois la note calculée, nous pouvons retrouver l'article en question grâce à son hash. Par la suite, il pourra aussi être intéressant d'établir notre propre méthode de notation basée sur le taux de lecture, par exemple en prenant en compte la fréquence de lecture journalière. C'est-à-dire mettre une notre plus haute à un article consulter beaucoup de fois d'affiler et au contraire pénaliser les articles ou la lecture se fait de façons plus dispersées dans le temps. Bien sure, rajouter à cela le nombre de vues totales et le temps de lecture moyen plutôt que l'absolu.