Apprentissage et classification automatiques pour améliorer la pertinence d'un corpus d'articles

Julien Gobeill,
julien.gobeill@hesge.ch
Haute École de Gestion, Genève, Institut Suisse de Bioinformatique (SIB), Genève, Suisse
https://orcid.org/0000-0001-9809-7741

Matthias van den Heuvel,
matthias.vandenheuvel@epfl.ch
École Polytechnique Fédérale de Lausanne (EPFL)

Laura Minu Nowzohour,
laura.nowzohour@graduateinstitute.ch
Institut de Hautes Études Internationales et du Développement (IHEID)

Joëlle Noailly,
joelle.noailly@graduateinstitute.ch
Institut de Hautes Études Internationales et du Développement (IHEID)

Gaétan de Rassenfosse,
gaetan.derassenfosse@epfl.ch
École Polytechnique Fédérale de Lausanne (EPFL)

Patrick Ruch,
patrick.ruch@hesge.ch
Haute École de Gestion, Genève, Institut Suisse de Bioinformatique (SIB), Genève, Suisse

Résumé

Dans le cadre d'un projet étudiant le développement des politiques environnementales et climatiques sur les quatre dernières décennies, l'un des moyens envisagés par des chercheurs en sciences économiques est de construire puis exploiter un corpus d'articles de presse relatifs à cette thématique. La première année du projet s'est concentrée sur les seules archives du New York Times. Ce sont néanmoins 2,6 millions d'articles qui étaient à traiter – une masse trop importante pour l'homme. Des chercheurs en sciences de l'information et en fouille de texte ont donc été associés à cette tâche de recherche d'information. Dans un premier temps,

les 2,6 millions d'articles ont été moissonnés depuis le Web, puis indexés dans un moteur de recherche. La conception d'une équation de recherche complexe a permis de sélectionner un corpus intermédiaire de 170 000 articles, dont la précision (taux d'articles pertinents) a été évaluée à 14%. Dans un deuxième temps, un algorithme d'apprentissage automatique a donc été entraîné et utilisé pour prédire la pertinence ou non d'un article. Pour nourrir l'algorithme, un échantillon de 700 articles a été manuellement étiqueté par les chercheurs en sciences économiques. L'application du classifieur à l'ensemble du corpus intermédiaire a produit un corpus final de 15 000 articles, dont la précision a été évaluée à 83%. Nos résultats montrent qu'une centaine d'articles étiquetés semble ici une quantité suffisante pour maximiser les performances du classifieur, et obtenir un corpus final de qualité proche de celle obtenue par des experts humains. La fouille de texte n'est plus une discipline émergente, ni extérieure aux sciences de l'information ; c'est une discipline mature qui peut dès à présent être utilisée pour assister le spécialiste de recherche documentaire dans une tâche de construction de corpus ou de classification de documents, tout spécialement avec des masses d'informations importantes.

La construction d'un corpus de documents pertinents pour un besoin d'information donné, à l'aide d'un outil de recherche, est une tâche qui relève de la recherche d'information (RI). La RI peut être considérée selon deux points de vue en sciences de l'information : celui du spécialiste en recherche documentaire, et celui du spécialiste en sciences informatiques. Selon le point de vue adopté, les interactions entre l'outil de recherche et son utilisateur sont complémentaires. Le spécialiste de recherche documentaire s'attache à construire la meilleure équation de recherche possible pour un besoin d'information donné ; dans son travail, l'outil reste inchangé et la requête évolue afin d'obtenir un maximum de documents pertinents. L'informaticien spécialiste en fouille de texte (de l'anglais text mining) s'attache quant à lui à construire le meilleur outil de recherche possible ; dans son travail, la requête reste inchangée et l'outil évolue afin de prédire correctement et renvoyer un maximum de documents pertinents (Weiss, 2015). Les travaux que nous allons présenter ici combinent les deux approches.

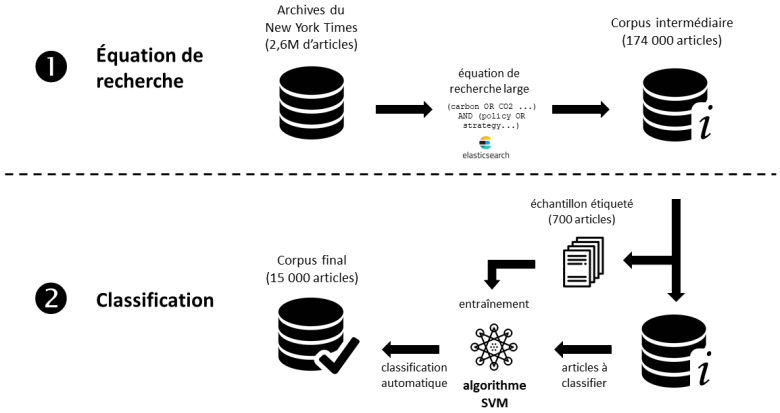
Cette étude se déroule dans le cadre du Projet National de Recherche 73 sur l'économie durable, financé par le Fonds National Suisse de la Recherche Scientifique. Un groupe de chercheurs en sciences économiques, provenant de l' Institut de Hautes Études Internationales et du Développement (IHEID) et de l'École Polytechnique Fédérale de Lausanne (EPFL), travaille sur l'investissement dans les technologies propres, et cherche à quantifier le développement des politiques environnementales et climatiques sur les quatre dernières décennies. L'un des moyens envisagés est l'exploitation d'articles de presse relatifs à cette thématique. A terme le but est de construire un indice mesurant le degré d'incertitude dans les politiques économiques liées au changement climatique. Cet indicateur devrait pouvoir refléter des pics et changements brusques dans les politiques climatiques, comme par exemple la sortie des Etats-Unis de l'accord de Paris sur le climat sous la présidence de Donald Trump.

1. Description générale de la tâche et de l'approche

Une tâche fondamentale du projet consiste donc à construire un corpus d'articles, issus d'une dizaine de revues généralistes de différents pays, relatifs aux politiques économiques liées au changement climatique. C'est pour cette tâche que l'expertise de chercheurs de la filière Information Documentaire de la Haute École de Gestion (HEG) a été mobilisée. Nous avons décomposé cette tâche en deux sous-tâches successives :

- **Sous-tâche 1, filtrage par équation de recherche :** construire une équation de recherche large dans une base de données documentaire, pour ramener un corpus intermédiaire contenant le plus grand nombre d'articles pertinents possible ;

- **Sous-tâche 2, classification automatique :** filtrer ce corpus intermédiaire grâce à un classifieur, utilisant des techniques de d'apprentissage et de classification automatiques de documents, afin d'obtenir le corpus final.



Enchaînement des deux sous-tâches de construction du corpus final

Le classifieur que nous avons développé pour la sous-tâche (2) a donc pour fonction de prédire automatiquement si un article est pertinent ou non pour un besoin d'information donné. Ce classifieur s'appuie sur l'apprentissage automatique. Dans les sciences informatiques, l'apprentissage automatique est un champ de l'intelligence artificielle dont le principe est d'entraîner un algorithme avec des exemples déjà étiquetés (pertinents ou non), pour lui donner la capacité d'étiqueter lui-même de nouveaux exemples. Contrairement à un système expert dans lequel le comportement de l'algorithme est régi par des règles explicites spécifiées par l'humain (comme la présence dans l'article de certains termes), un algorithme d'apprentissage artificiel généralise à partir de données d'entraînement pour adopter un comportement non explicitement programmé.

2. Études précédentes

L'étude (Baker et al., 2016) est fondatrice, car elle est la première de grande ampleur à essayer de mesurer l'incertitude de politique économique (EPU) avec des outils bibliométriques. Contrairement à notre projet NRP, l'étude de Baker et al. s'intéresse à tous les domaines économiques (pas seulement liés au réchauffement climatique) et remonte jusqu'au début du 19ème siècle, étudiant des faits comme la première guerre mondiale ou la crise de 1929. Une dizaine de revues américaines majeures sont considérées dans cette étude. Un premier corpus d'articles – appelé par les auteurs « computer-generated » – est construit grâce à une équation de recherche relativement simple : ("economic" OR "economy") AND ("uncertain" OR "uncertainty") AND ("congress" OR "deficit" OR "Federal Reserve" OR "legislation" OR "regulation" OR "White House"). Un second corpus d'articles (appelé human-generated) est construit après lecture et évaluation de 12 000 articles. La construction manuelle de ce second corpus a nécessité un effort conséquent : des équipes d'étudiants de l'université de Chicago ont lu et sélectionné des articles pendant dix-huit mois, s'appuyant sur un guide de 65 pages, et étant supervisés toutes les semaines par les auteurs de l'étude. Les deux corpus ainsi construits montrent une forte corrélation. Quant à l'indice d'incertitude de politique économique ainsi généré, il présente des corrélations avec plusieurs autres indicateurs économiques (comme la volatilité des marchés). De plus, l'indicateur généré présente des pics lors d'événements comme le 11 septembre, ou la faillite de Lehman Brothers.

Parmi les études inspirées par ces travaux, (Tobback et al., 2016) est tout à fait intéressante. Le point de départ est que l'équation de recherche utilisée par Baker et al. induit probablement

des erreurs de type 1 et 2, qui doivent fausser l'indice. Les erreurs de type 1 sont des faux positifs : des articles qui satisfont l'équation de recherche mais ne sont pas pertinents. Les erreurs de type 2 sont des faux négatifs : des articles pertinents mais non ramenés car ils traitent du sujet en d'autres termes que ceux spécifiés dans l'équation de recherche. Les exemples suivants illustrent les deux types d'erreur.

"As always there are risks to the outlook, not least Brexit uncertainty," Dr O'Sullivan said. "This is the big issue for the UK and is also a cloud on the horizon for Ireland, along with **uncertainties** related to the external **policy environment** and exchange rates.

Exemple de faux positif : retourné par la recherche mais non pertinent

Australia news
Feed-in tariffs could be cut back due to high take-up of solar power

7 March 2018
The Guardian

Exemple de faux négatif : pertinent mais non retourné par la recherche

Dans leur publication, Tobback et al. proposent d'améliorer la qualité du corpus d'articles en le filtrant grâce à un outil de classification automatique. La collection initiale comprend 210 000 articles issus de cinq revues flamandes, collectées gratuitement sur le Web. Un échantillon de 400 articles est d'abord étiqueté (pertinent ou non-pertinent) manuellement, puis utilisé pour entraîner l'algorithme d'apprentissage. L'algorithme utilisé est une Machine à Support de Vecteurs, ou SVM (Joachims, 1998). Comparée à l'approche naïve de Baker et al., l'approche de Tobback et al. montre une précision (ou spécificité) comparable (97% contre 99%), mais un rappel (ou sensitivité) beaucoup plus élevé (68% contre 21%). En d'autres termes, l'équation de recherche ramène seulement 21% des articles pertinents, contre 68% pour le classifieur. Quant à l'indice d'incertitude de politique économique ainsi généré, les auteurs le présentent comme meilleur, et lui accordent même un pouvoir prédictif sur certains indicateurs économiques. Des fortes limites sont toutefois que l'étude de Tobback et al. se limite à l'économie belge, et ne remonte pas avant 2000.

3. Revues étudiées, Factiva, et New York Times

Comme nous devions accéder à des articles de revues de différents pays du monde, nous nous sommes naturellement tournés vers une base de données documentaires. Le consortium des bibliothèques universitaires suisses détient notamment des licences pour Factiva. Factiva est un outil de recherche d'information professionnel, détenu par Dow Jones & Company, et agrégeant plus de 32 000 sources différentes de 200 pays.

Les experts en économie de notre équipe ont donc construit méthodiquement une équation de recherche complexe dans l'interface de Factiva, ayant pour but de ramener le plus de documents possible traitant de politiques économiques liées au changement climatique. L'équation de recherche finale contient près de 6 000 caractères. Elle se compose de 297 opérateurs OR, 7 opérateurs AND, 150 opérateurs de proximité NEAR, 19 troncatures, et 215 paires de guillemets pour chercher des expressions exactes. C'est une équation de recherche largement plus complexe que celle utilisée par Baker et al., ce qui reflète aussi le fait que les politiques environnementales et climatiques couvrent des domaines plus variés que la politique monétaire économique. Le tableau suivant donne une indication des concepts considérés comme pertinents.

N°DOI : 10.55790/journals/ressi.2018.e1906

Environment	Policy
Renewable Energy Generation	Regulation
Energy Storage	Standards & Certification
Energy Infrastructure & Efficiency	Feed-in tariffs & premiums
Transportation	Taxes & Subsidies
Water & Wastewater	Emissions trading schemes
Air & Environment	International agreements
Recycling & Waste	Loan guarantees
Clean Manufacturing	Green & Climate bonds

Concepts considérés pour l'élaboration la requête

Dans la dizaine de revues étudiées, cette équation ramène dans Factiva environ un million d'articles. Malheureusement, il nous est vite apparu que si l'interface de Factiva permettait à un humain de consulter et télécharger autant d'articles qu'il le voulait, elle l'interdisait techniquement à une machine. En fait, la licence du consortium interdit explicitement le text mining, et plus généralement toute lecture des articles par une machine. Ce genre de clauses peut être vu comme abusif, voire illégal ; toutefois, dans le cadre d'un projet, il est difficile d'entrer en confrontation avec une telle source de données. Nous sommes donc entrés en négociation avec Dow Jones pour une licence nous permettant de télécharger ce million d'articles et de les traiter avec du text mining.

Parallèlement aux négociations, nous avons décidé pour la première année du projet de télécharger les archives du New York Times, en libre accès, pour faire une première étude sur la classification d'articles. En juin 2018, nous avons donc téléchargé 2,6 millions de pages Web pour récupérer les archives du New York Times. 5% des pages n'ont pas pu être téléchargées à cause de dysfonctionnements techniques du côté de la revue. Le texte des articles a ensuite été extrait de ces pages Web.

4. Première sous-tâche : Filtrage par équation de recherche

Pour construire notre corpus intermédiaire, nous avons déployé un moteur de recherche dans les 2,6 millions d'articles du New York Times, en utilisant la solution open source Elasticsearch Lucene. Nous avons ensuite appliqué l'équation de recherche construite dans l'interface de Factiva. Les langages de requêtes n'étant pas parfaitement identiques (notamment pour l'opérateur de proximité NEAR qui est plus puissant dans l'interface de Factiva), nous avons dû légèrement modifier l'équation de recherche initiale. Nous verrons que cela aura une forte incidence sur le ratio de documents pertinents dans le corpus intermédiaire. Nous avons finalement obtenu notre corpus intermédiaire de 174 000 articles potentiellement pertinents.

N°DOI : 10.55790/journals/ressi.2018.e1906

Pour préparer la deuxième étape (élaboration du corpus final par classification automatique), nous avons étiqueté – c'est-à-dire jugé chaque élément comme pertinent ou non – un échantillon représentatif d'articles. En effet, les méthodes d'apprentissage automatique requièrent des données d'entraînement. Trois membres de l'équipe experts en économie ont donc parcouru un échantillon représentatif de 700 articles présents dans notre corpus intermédiaire. Sur ces 700 articles, 94 étaient pertinents. Ce ratio de 14% est très inférieur au ratio de 50% observé empiriquement dans les résultats de Factiva. Cette différence s'explique par la puissance de l'opérateur de proximité NEAR dans Factiva, qui peut être utilisé entre des expressions complexes, là où il ne peut être utilisé qu'entre deux termes simples dans le langage Lucene Elastisearch.

5. Prétraitement des articles

Avant d'être utilisés par le classifieur, les articles doivent être prétraités par des méthodes de fouille de texte. En effet, les algorithmes de classification reposent sur des méthodes statistiques, et prennent généralement en entrée des exemples décrits par des attributs qui ont des valeurs numériques ou catégorielles. Ce type de données structurées peut typiquement se représenter dans un tableau, où les attributs sont des colonnes et les exemples sont des lignes. Par exemple, pour des données médicales d'un dossier patient, des attributs peuvent être la tension artérielle ou le poids (des nombres), ou bien le sexe ou la présence de codes maladies (des catégories). L'algorithme va alors apprendre à établir des seuils et des liens entre chaque attribut pour inférer ses décisions.

De leur côté, les articles de revues sont du texte : par nature, ils ne peuvent pas être représentés tels quels dans un tableau. Pour avoir une représentation compatible, la première chose à faire est d'obtenir l'ensemble des mots apparaissant dans le corpus d'articles. Cet ensemble de mots est appelé le dictionnaire, et chaque mot du dictionnaire va devenir un attribut. Dans les 700 articles que nous avons étiquetés, il y a ainsi 38 300 mots différents, donc potentiellement 38 300 colonnes dans un tableau représentant nos données. Pour remplir les lignes, la représentation la plus simple consiste à donner, pour chaque article et chaque mot, la valeur 1 si le mot apparaît dans l'article, ou la valeur 0 dans le cas contraire. Une représentation plus évoluée, que nous nommerons « count » dans les résultats présentés, consiste à donner le nombre de fois où le mot apparaît dans l'article ; l'hypothèse sous-jacente est qu'un mot répété plusieurs fois dans un article a plus de chances d'être représentatif de l'article. Enfin, une représentation pondérée et habituellement utilisée en fouille de texte est « term frequency inverse document frequency » (tfidf). Cette pondération prend en compte la fréquence d'un mot dans un article (tf), mais aussi l'inverse de la fréquence du mot dans tout le corpus (df). L'hypothèse sous-jacente est ici que plus un mot est présent dans tout le corpus, moins il est représentatif d'un article.

Dans cette étude, nous avons considéré les représentations count et tfidf.

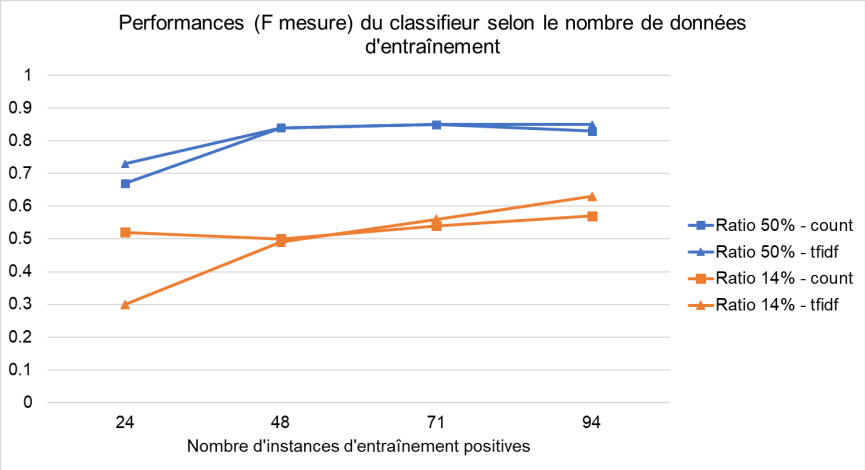
6. Deuxième sous-tâche : classification automatique

L'algorithme de classification que nous avons choisi pour cette sous-tâche est une Machine à Vecteurs de Supports (SVM). L'approche par SVM est largement répandue dans la classification automatique de documents, et obtient généralement de très bons résultats (Sun et al., 2009). Une SVM cherche des régularités dans des articles déjà étiquetés, et sélectionne des combinaisons de mots qui ont les plus grands pouvoirs de discrimination. C'est l'algorithme utilisé par Tobback et al.

N°DOI : 10.55790/journals/ressi.2018.e1906

Nous avons entraîné et évalué le classifieur avec le jeu de 700 articles étiquetés : 94 positifs (pertinents) et 606 négatifs (non pertinents), soit un ratio de 14%, comme dans le corpus intermédiaire obtenu avec Elasticsearch. Mais nous avons aussi construit un deuxième jeu de données avec un ratio de 50%, pour étudier les performances du classifieur dans le futur corpus intermédiaire obtenu avec Factiva : les 94 positifs, et 94 négatifs. Le classifieur a donc été évalué avec deux jeux de données présentant des ratios différents. Pour chaque jeu de données, nous avons constitué quatre sous-jeux de taille croissante, pour étudier comment les performances du classifieur évoluaient avec la taille des données d'entraînement. L'évaluation s'est faite selon la méthode de « leave one out crossvalidation ». Pour chaque article étiqueté, un classifieur est entraîné avec tous les autres ; ensuite, le classifieur est évalué sur sa capacité à prédire correctement la classe (pertinent ou non) de l'article retiré des données d'entraînement. Le classifieur est ainsi entraîné sur un maximum de données ; d'autre part, il est évalué sur tous les articles du jeu de données, ce qui maximise la significativité de l'évaluation. Des valeurs de précision et de rappel peuvent ainsi être calculées pour tous les jeux de données.

La figure suivante indique les performances du classifieur pour deux ratios d'articles pertinents (50% et 14%), en utilisant deux représentations de documents (count ou tfidf). Les valeurs reportées sont les F-mesures du classifieur ; la F-mesure est calculée en effectuant la moyenne harmonique de la précision et du rappel.

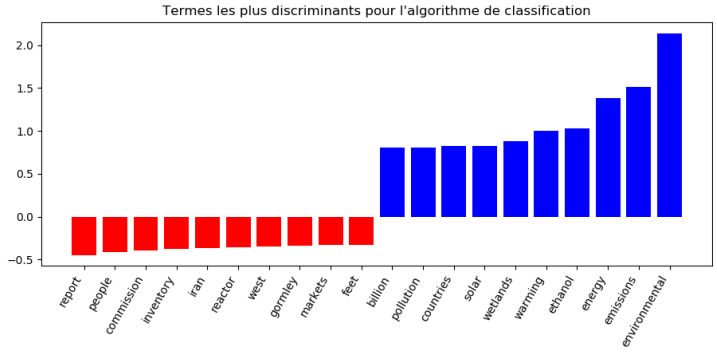


Concernant le jeu de données avec un ratio de 50% (courbes bleues), les courbes montrent un plafonnement rapide des performances. Une centaine d'articles étiquetés semble ici une quantité suffisante pour entraîner de manière optimale le classifieur. La différence entre les deux représentations (count et tfidf) apparaît comme minime. La meilleure F-mesure reportée est de 84%, ce qui est comparable à celle reportée par Tobback et al. (80%). Dans cette configuration, la précision du classifieur est de 80% et le rappel de 90%.

En revanche, concernant le jeu de données avec un ratio de 14% (courbes oranges), les courbes ne plafonnent pas encore. Ici, les articles positifs sont noyés dans le bruit, et obtenir une classification correcte est plus difficile, et demande plus de données d'entraînement. La différence entre les deux représentations est toujours minime, mais la représentation tfidf (marqueurs triangulaires) pourrait se montrer plus efficace avec plus de données. La meilleure

F-mesure reportée est de 63%. Dans cette configuration, la précision du classifieur est de 83% mais le rappel seulement de 51%.

La figure suivante indique les mots les plus discriminants selon le classifieur, que ce soit pour prédire qu'un article est non-pertinent (en rouge) ou pertinent (en bleu). Les termes reportés sont majoritairement cohérents.



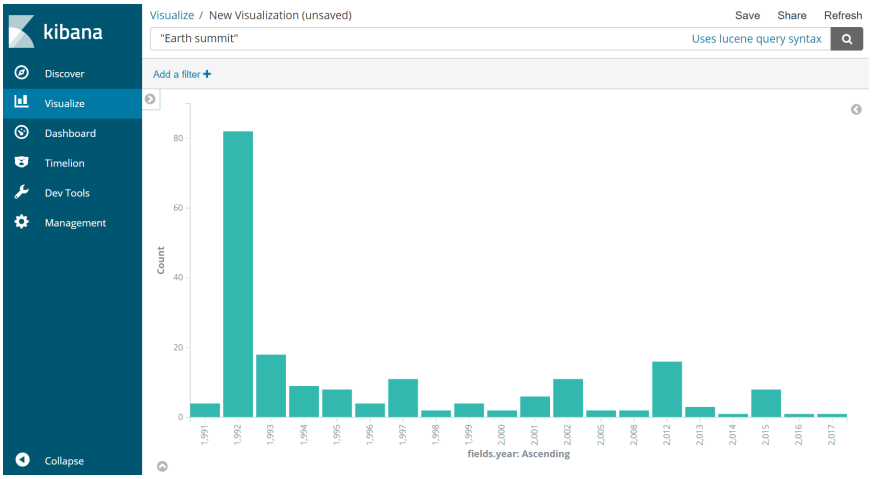
Enfin, pour interpréter les résultats, il est nécessaire de prendre en compte l'accord inter-annotateur. La pertinence d'un article est une notion subjective, qui dépend de l'expert jugeant l'article ; le même article peut ainsi être jugé comme pertinent ou non selon l'expert. Trois cents articles ont donc été jugés en parallèle par deux experts différents. Sur la totalité des 300 articles, les experts montrent un accord de 89%. Le calcul du kappa (mesure statistique pour mesurer l'accord inter-annotateur) donne une valeur de 0,63, ce qui est interprété comme un accord assez fort (Baeza-Yates et al., 2011). Ces valeurs sont à mettre en perspective avec les résultats du classifieur : si les experts s'accordent à 89% sur la pertinence d'un document, la performance du classifieur évalué sur leurs jugements ne peut théoriquement dépasser ce plafond. La F-mesure obtenue (84%) pour le jeu de données avec un ratio de 50% indique donc que le classifieur montre des performances assez proches de celles d'un expert humain, tout en étant bien sûr incomparablement plus rapide pour juger des milliers d'articles.

7. Corpus final

Nous avons entraîné un classifieur avec un échantillon de 700 articles, présentant un ratio de 14% d'articles positifs. Puis, nous l'avons appliqué sur un corpus intermédiaire de 174 000 articles potentiellement pertinents issus de la première sous-tâche (qui présente lui aussi un ratio de 14% d'articles pertinents). Nous avons ainsi généré le corpus final, qui ne contient plus que 15 000 articles. Ce corpus final affiche théoriquement une précision de 83%, mais un rappel de seulement 51% par rapport au corpus intermédiaire.

Le corpus final a été indexé dans un nouveau moteur de recherche Elasticsearch. Il est ainsi possible de faire des recherches dans le corpus final, comme le montre la figure suivante. Ici, une recherche est faite avec la requête « earth summit », qui ramène 195 articles. Les archives du site officiel du New York Times en comptent 438, mais on sait que le rappel de notre corpus final est d'environ 50%, et tous les articles du New York Times ne sont pas pertinents pour notre étude.

Il est aussi possible avec Elasticsearch de visualiser l'évolution du nombre d'articles mentionnant « earth summit », comme illustré dans la figure suivante. On observe ici un maximum pour 1992, année du sommet de la Terre à Rio, puis une diminution, hormis lors de pics en 1997 (année de signature du protocole de Kyoto déroulant de la conférence de Rio), 2002 (conférence de Johannesburg), 2012 (conférence de Rio+20) et 2015 (conférence de Paris).



8. Conclusion

Nous avons construit un corpus d'articles du New York Times relatifs aux politiques économiques liées au changement climatique, en enchaînant deux étapes : la construction d'une équation de recherche ramenant un maximum de documents potentiellement pertinents, et l'application d'un outil de classification automatique pour filtrer les résultats précédents. Pour un ratio de 50% d'articles pertinents issus de la première étape, une centaine d'articles étiquetés (en comparaison des 12 000 de Baker et al.) semble ici une quantité suffisante pour maximiser les performances du classifieur et obtenir un corpus final de qualité proche de celle obtenue par des experts humains. Dans la suite de l'étude, une fois obtenu de Factiva le million d'articles potentiellement pertinents issus d'une dizaine de revues, nous devrons ré-étiqueter de nouveaux articles pour entraîner et évaluer le classifieur avec les nouvelles revues.

Dans l'article de Baker et al., les auteurs écrivent qu'il est intéressant de noter que des archives de revues sont accessibles pour tous les pays du monde, et pour des dizaines d'années selon

les pays. Ils clament que cette ubiquité, couplée avec les outils informatiques, offre des possibilités gigantesques d'approfondir notre compréhension des développements économiques, politiques et historiques, à travers des démarches scientifiques empiriques. Malheureusement, ces possibilités ne s'offrent que si les données sont accessibles aux chercheurs. Si cette accessibilité paraît encouragée par quelques revues, elle ne l'est clairement pas par les agrégateurs de contenu, qui différencient encore dans leurs licences d'utilisation l'utilisateur humain de l'utilisateur machine.

La fouille de texte n'est plus une discipline émergente, ni extérieure aux Sciences de l'Information ; c'est une discipline mature qui peut dès à présent être utilisée pour assister le spécialiste de recherche documentaire dans une tâche de construction de corpus ou de classification de documents, tout spécialement avec des masses d'informations importantes.

BIBLIOGRAPHIE :

Baeza-Yates, R., & Ribeiro, B. D. A. N. (2011). Evaluation in Information Retrieval. In Modern information retrieval. New York: ACM Press; Harlow, England: Addison-Wesley.

Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring economic policy uncertainty. The Quarterly Journal of Economics, 131(4), 1593-1636.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In European conference on machine learning (pp. 137-142). Springer, Berlin, Heidelberg.

Sun, A., Lim, E. P., & Liu, Y. (2009). On strategies for imbalanced text classification using SVM: A comparative study. Decision Support Systems, 48(1), 191-201.

Tobback, E., Naudts, H., Daelemans, W., de Fortuny, E. J., & Martens, D. (2016). Belgian economic policy uncertainty index: Improvement through text mining. International journal of forecasting.

Weiss, S. M., Indurkhya, N., & Zhang, T. (2015). Fundamentals of predictive text mining. Springer.