Partitionnement (clustering) d'articles de nouvelles

Israel Akobi, Samuel Brin-Marquis, Charles Matte-Breton

Université Laval, Québec (Québec), Canada

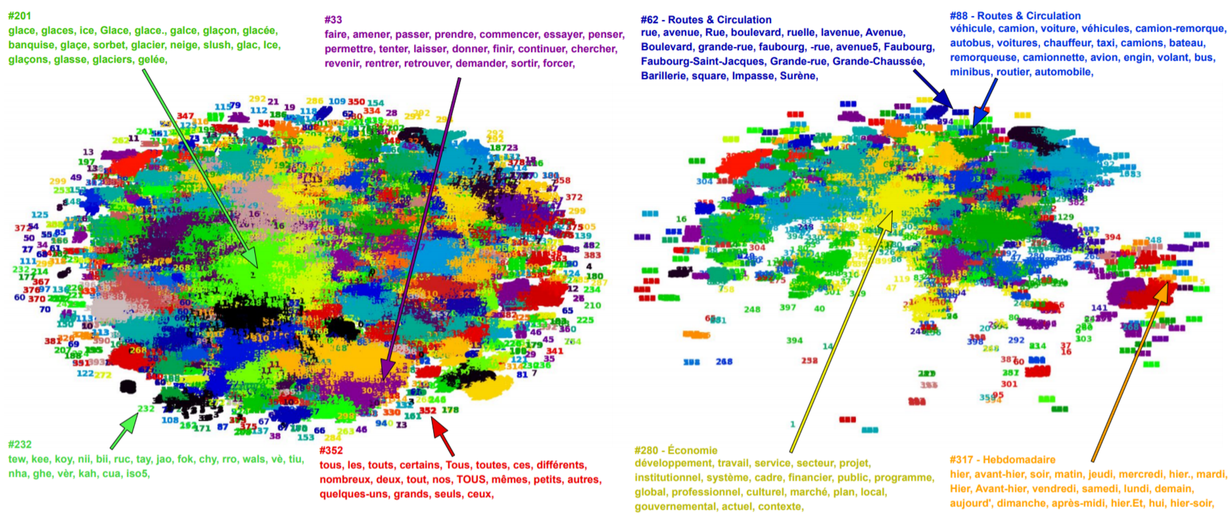
Problématique

- Depuis l'époque de la presse écrite, les nouvelles dans les journaux sont classées manuellement, générant ainsi peu de constance avec des catégories allant de très générales (Justice et faits divers, Actualité, Actualités, etc...) à trop spécifiques (Journaliste en classe, EMPLOIS D'AVENIR, Le Soleil, Le fruit de ma passion, Pierre Jury, LA VOIX DE ST-ALPHONSE-DE-GRANBY, Être LGBT dans l'Est ontarien, Branché - L'Ami Junior Nissan, Article commandité 2 de 4, etc...) pour un total de plus de 1200 catégories différentes. De plus, chaque nouvelle n'est assignée qu'à une seule catégorie bien qu'elle puisse toucher à des sujets variés.

Objectifs

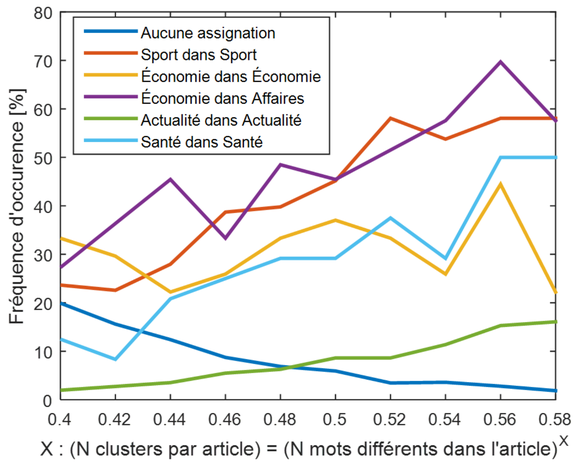
- De nos jours, avec la presse en-ligne et les outils qu'offrent l'apprentissage machine, il est possible de créer un système de tri plus performant.
- Le défi à relever est donc l'élaboration d'un nouvel algorithme basé sur le traitement automatique des langues naturelles afin de catégoriser les articles de façon plus juste, en assignant parfois plusieurs étiquettes à un article, et plus uniforme, en ayant un plus petit nombre de catégories différentes.

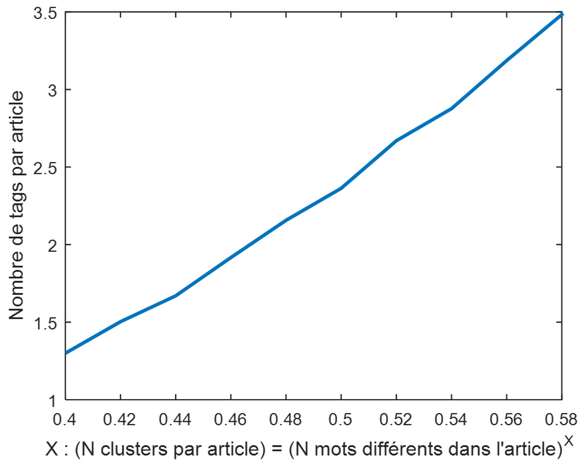
Regroupement des thèmes similaires



Représentation 2D des regroupements par thèmes des articles en 300 dimensions

Optimisation des paramètres d'étiquetage





Analyse de l'impact de la variation du nombre de clusters à conserver par article sur les étiquettes attribuées et comparaison avec les étiquettes initiales (tests effectués sur 500 échantillons)

Informations d'intérêt supplémentaires

- La base de données utilisée pour entraîner l'algorithme est de 25000 articles.
- Parmis les 400 groupes, 274 sont jugés inutiles (assignation manuelle)
- 62 catégories assignées manuellement contenant chacune un groupe ou plus
- 3959 articles sans classement se font assigner la catégorie 'Variété'
- Le nombre moyen d'étiquettes attribuées par article est d'environ 2,1.

Validation qualitative à l'aide du jeu de données de test

Ce test utilise 100 articles tirés au hasard dans le jeu de tests. Son but est de regarder la pertinence des étiquettes attribuées par l'algorithme avec l'étiquette de base de l'article. La métrique utilisée est le jugement humain.

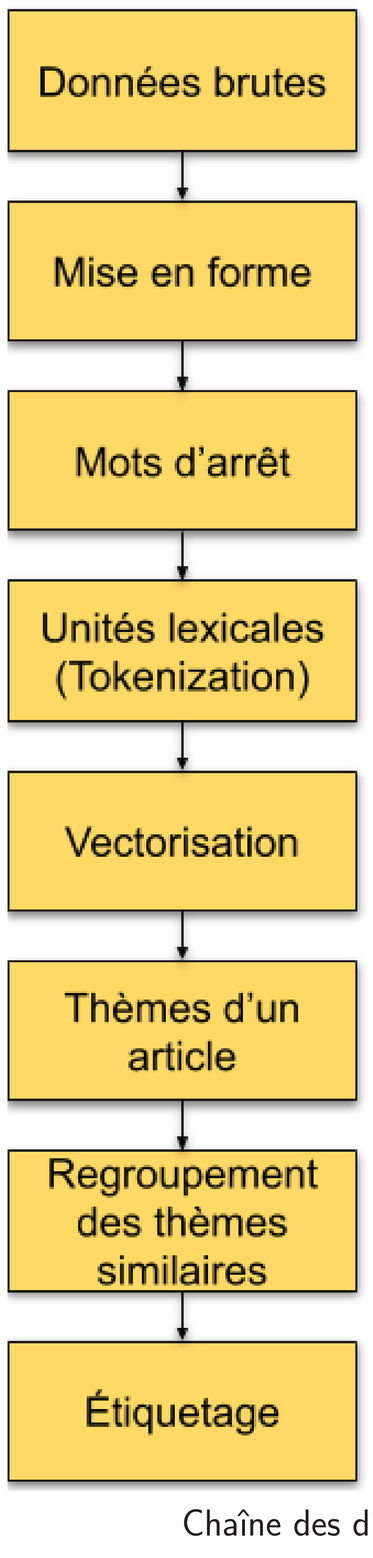
- 26 articles ayant une catégorisation moins bonne qu'initialement
- 39 articles ayant une catégorisation aussi bonne qu'initialement
- 35 articles ayant une catégorisation meilleure qu'initialement

Conclusion

Pistes d'amélioration :

- Augmenter le nombre de centres pour les catégories ($>$400) ;
- Ajouter des centres de façon manuelle (Innovation, Technologie, etc...) ;
- Utiliser une approche par convolution au lieu de l'algorithme Bag of Words.

Méthodologie



Chaîne des différentes étapes de l'élaboration de l'algorithme

Cet article se voit attribuer les étiquettes 'Justice', 'Local' et 'Maladies & Mortalité'. Son étiquette initiale était 'Justice et faits divers'.

Exemples de résultats

Bonne assignation : ARTICLE #36451

Menace de grève à la STTR

Le syndicat représentant les chauffeurs, les employés de bureau et les employés affectés à l'entretien des véhicules a déposé, mardi, sa demande d'application du droit de grève au ministère du Travail et au Tribunal administratif du travail. Le syndicat profite depuis le mois de juin d'un mandat de grève accordé par ses membres. «On ne sait pas encore si on va en grève, c'est le choix du comité de négociations. Il n'y a pas de date fixée pour une grève, mais c'est possible que ça arrive»...

Étiquette initiale : 'Actualités'

Étiquettes assignées : 'Économie', 'Actualité', 'Grèves', 'Politique'

Aucune étiquette assignée : ARTICLE #36460

L'Isle-Verte : enquête publique réclamée

«Quelles sont les applications qui pourraient être mises en place pour éviter que ça se reproduise ?» Le vice-président de l'Association, André Bonneau, sait pertinemment que sa question ratisse large, d'où, croit-il, la nécessité d'une enquête publique...

Étiquette initiale : 'Justice et faits divers'

Étiquettes assignées : 'Variété'

Mauvaise assignation : ARTICLE #36497

Les travaux reprennent sur le pont Dubuc

Les employés du ministère des Transports s'affairent à préparer le chantier. On procédera cet été au remplacement des dalles de béton qui se trouvent sous la voie de gauche en direction sud...

Étiquette initiale : 'Actualités'

Étiquettes assignées : 'Routes & Circulation', 'Blessures & Hospitalisation', 'Infrastructures'

L'équipe tient à remercier la compagnie **Capitales Médias** pour la base de données ayant été utilisée dans le cadre du projet.
L'équipe tient également à remercier Richard Khoury pour la définition de projet, ses conseils et son appui tout au long du projet.