ALGORITHME K-MEANS

Le clustering au temps du mousquet et des menottes

RAPPORT DE LABORATOIRE

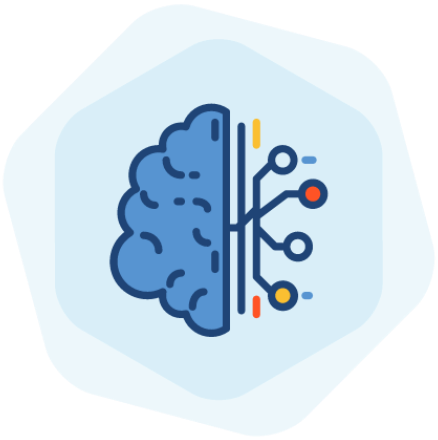
PAR

Déric Marchand

Karl Robillard-Marchand

PRÉSENTÉ À

M. Pierre-Paul Monty



Technique de l'informatique || Veille technologique – C62

Cégep du Vieux Montréal

15 décembre 2022

TABLE DES MATIERES

RÉSULTATS

Curieux d'analyser l'impact de la nature des données soumises à l'algorithme *k-means*, nous avons opté pour la génération de deux ensembles de données. L'un d'eux demeure celui avec laquelle nous avons eu à conjuguer depuis le premier projet, c'est-à-dire constitué de textes littéraires classiques[1], entraînés chacun une seule fois. L'autre est composé de deux textes juridiques d'importance, à savoir le *Code Civil* du Québec et le *Code Criminel* du Canada. Afin de comparer ces deux corpus, nous avons exécuté l'algorithme selon deux configurations avec la méthode de calcul des moindres-carrés. La somme des centroïdes reportés dans les tableaux ne correspond pas nécessairement au nombre de centroïdes totaux : nous faisons fi de ceux que nous considérons comme du bruit (comme deux centroïdes qui contiendraient uniquement des chiffres ou ceux représentant les mentions légales en anglais des textes sources).

Enfin, nous avons tenté pour chaque centroïde d'extraire une identité commune. Évidemment, nous sommes conscients du caractère arbitraire de cette proposition. Pour modérer cet aspect, nous tentons d'indiquer la pertinence de cette identité liant les meilleurs résultats via un code de couleur[2].

PREMIÈRE EXPÉRIENCE

TAILLE DE FENÊTRE 7 / 30 CENTROÏDES / 12 RÉSULTATS PAR PARTITION

CORPUS LITTÉRAIRE

Identité(s)	*Nb de centroïdes*	*No. centroïdes*	*Meilleurs scores*	*Force de l'identité*
Stop words (pronoms, marqueurs de relation, verbes être et avoir)	6	2, 4, 8, 18, 24, 27	7904168, 315556, 386809, 7312, 12717417, 6780	

[1] *Les Trois Mousquetaires*, *Germinal*, *Don Quichotte*, *Le Ventre de Paris*.

[2] Vert : identité forte / Jaune : identité partiellement pertinente / Rouge : identité très faible.

Verbes d'action	4	6, 17, 23, 28	1004, 14524, 9873, 17459	
Noms	1	0	3.54	
Soi	1	7	221444	
Faim, établissement	1	3	3287	
Intérieur, intimité	2	5, 19	22020, 1066	
Relations, économie	1	9	337	
Passion, nature	1	10	221	
Statut social	1	12	38019	
Liesse, déchéance	1	14	304	
Anatomie, humains	2	15, 29	23157, 61951	
Marqueurs de temps	1	20	3360	
Espace	1	22	5137	
Cadre officiel	1	25	1467	
Nombre de centroïdes jugés significatifs			14 / 30 -> Environ 47 %	

CORPUS JURIDIQUE

Identité(s)	*Nb de centroïdes*	*No. centroïdes*	*Meilleurs scores*	*Force de l'identité*
Stop words (pronoms, marqueurs de relation, verbes être et avoir)	5	2, 4, 11, 21, 25	131496, 11552170, 1070107, 4853824, 328143	
Convocation	1	0	3.54	
Loi, règlement	1	1	86044	
Obligation, preuves	3	5, 24, 26	550, 6917, 2631	
Action (noms), autorité	3	6, 9, 29	4816, 11221, 167	
Bureaucratie	1	7	27957	

Personne, rôle	3	12, 19, 22	21492, 4.46, 5147	
Transition, cadre	3	14, 16, 20	7757, 1028, 2257.62	
Pénitencier, culpabilité	1	15	387762	
Principes	1	23	34460	
Nombre de centroïdes jugés significatifs			14 / 30 -> Environ 47 %	

SECONDE EXPÉRIENCE

TAILLE DE FENÊTRE 5 / 40 CENTROÏDES / 10 RÉSULTATS PAR PARTITION

CORPUS LITTÉRAIRE

Identité(s)	*Nb de centroïdes*	*No. centroïdes*	*Meilleurs scores*	*Force de l'identité*
Stop words (pronoms, marqueurs de relation, verbes être et avoir)	16	0, 1, 7, 8, 11, 16, 18, 19, 24, 25, 28, 29, 30, 31, 33, 35	16160, 2461593, 2185992, 217991, 302883, 7542, 5554309, 38601, 1056798, 30428, 65973, 458531, 5145, 279436, 416806, 18279	
Verbes d'action	6	9, 17, 21, 22, 23, 32	1580, 335, 2846, 0, 38	
Principes	1	6	6679	
Enfance	1	3	2.45	
Espace social, vie urbaine	4	2, 4, 34, 36	1084, 3595, 14839, 88	
Lieux	1	15	96	
Vulnérabilité, recueillement	1	12	7481	
Émancipation, travail	1	20	255	
Métier, rôles	3	10, 13, 26	401, 28266, 843	

Obligation, droit	1	14	915	
Relations, sentiments	1	27	248	
Marqueurs de temps	1	5	8120	
Anatomie, humains	1	38	12573	
Nombre de centroïdes jugés significatifs			30 / 40 -> 75 %	

CORPUS JURIDIQUE

Identité(s)	*Nb de centroïdes*	*No. centroïdes*	*Meilleurs scores*	*Force de l'identité*
Stop words (pronoms, marqueurs de relation, verbes être et avoir)	7	8, 14, 30, 34, 35, 36, 38	6501, 3708950, 976236, 7427737, 331458, 1029902	
Cour	5	0, 21, 26, 29, 37	339756, 58299, 32478, 197377, 612935	
Paix sociale	1	2	3.25	
Arrestation	2	6, 39	83925, 4802	
Personne, rôle	2	10, 17	3903, 2291	
Obligation, preuves	2	16, 32	38884, 411	
Action (noms), tierce partie	2	22, 33	1170, 1138	
Bureaucratie	1	24	20574	
Contrat	2	25, 31	161, 32748	
Marqueurs de temps	2	23, 28	5388, 5065	
Nombre de centroïdes jugés significatifs			21 / 40 -> Environ 53 %	

Discussion

Nous avions quelques aprioris sur les résultats attendues, et nous sommes forcés d'avouer qu'ils ont été renversés. Examinons de plus près les résultats et tentons d'en tirer certaines observations. Mais d'abord, quelques précisions sur la méthodologie employée. Seules les données rattachées à la couleur verte indiquent une pertinence satisfaisante. Ainsi, celles colorées de jaune ne sont pas comptabilisées dans le score final d'un corpus, mais indiquent quand même un lien minimalement défendable entre les résultats. Les données en rouge dénotent un effort de gymnastique mentale à ne pas considérer. Et certains centroïdes sont simplement absents, car considérés comme du pur bruit. Enfin, nous avons opté pour l'agrégation possibles des centroïdes sous une même identité. À l'occasion, certains des centroïdes agrégés était plus ou moins pertinent (**jaune**) alors que d'autres l'étaient davantage (**vert**). Il a fallu trancher, et nous avons préféré conserver un biais « optimiste » lorsque la représentation des deux cas était semblable.

Première expérience : une pertinence équitable

À notre grande surprise, les deux corpus offrent un indice de pertinence identique ***(47 %)***. L'idée de choisir un corpus juridique reposait sur l'intuition suivante : les textes de loi investissent l'écriture avec un souci de précision et une construction sémantique élaborée. Les textes littéraires peuvent, quant à eux, épouser des tangentes radicalement différentes les unes des autres, en misant sur des associations d'idées plus près de l'expérience poétique. Dans cette optique, nous nous attendions à un indice de pertinence plus élevé des textes juridiques relativement à celui des textes littéraires.

Toutefois, des différences sont à noter. Le corpus à partir duquel nous avons travaillé depuis le début du cours offre moins de centroïdes ayant une identité minimale (excluant **rouge** **ou** **absents**), c'est-à-dire ***6 / 13 (20 %)*** contre ***9 / 13 (30 %)*** pour les textes juridiques. Il y a certes plus de résultats colorés rouges dans le premier, mais plus de centroïdes absents dans le second. De plus, on relève une plus grande diversité d'identités chez le premier ***(14 contre 10).*** On ne sera pas étonné que les textes littéraires offrent davantage matière à interprétation (centroïdes rouges plutôt qu'absents et plus d'identités différentes) et que les identités du corpus juridique soient voisines (le droit en général).

SECONDE EXPÉRIENCE : LA LITTÉRATURE COMME VALEUR DE VÉRITÉ

Les deux corpus prennent une tangente opposée dans le cadre de cette seconde expérience. Une augmentation de ***25%*** du nombre de centroïdes avec une taille de fenêtre réduite augmente l'indice de pertinence du corpus juridique de ***6%***, générant énormément de bruit au passage. Du côté des textes littéraires, on parle plutôt d'une augmentation foudroyante de ***28%***! Toutefois, le nombre de centroïdes listant des *stopwords* a considérablement augmenté. Une redondance marquée au sein des identités est donc à noter et une analyse plus fine permettrait d'en éclairer les nuances.

La taille de la fenêtre a semblé diminuer grandement la pertinence des résultats des textes juridiques : beaucoup de centroïdes se sont révélés dénués d'intérêt, car seulement remplis de nombres. Ce n'était pas le cas avec moins de centroïdes et, surtout, une taille de fenêtre plus élevée. Ainsi, ***15 / 40 (38 %)*** centroïdes sont dénués d'identité dans ce corpus et avec cette configuration comparativement à ***2 / 40 (0, 05%)*** pour le corpus littéraire. La faute en est probablement au fait qu'il s'agit d'articles de loi. Les insertions récurrentes de nombres à des fins de référence exigent probablement une taille de fenêtre plus élevée si l'on souhaite obtenir des résultats intéressants. Ainsi, la seconde expérience accorde, selon les paramètres de l'expérience, une plus grande valeur de vérité au corpus littéraire. C'est dire que pour obtenir des données probantes, il demeure essentiel de garder à l'esprit la nature des données à partir desquelles on travaille et s'y adapter.

De manière idéale, il n'y aurait pas d'agrégation à réaliser, puisque chaque centroïde devrait représenter une idée, un concept délimité et exclusif. Toutefois, les langues naturelles sont poreuses, organiques. Et le matériau brut qu'est le langage échappe aux considérations simplistes. Ainsi, il est de la responsabilité des chercheurs d'identifier des identités assez précises et englobantes à la fois. Les résultats manquaient de qualité et de quantité, et notre expérience en la matière est trop réduite pour avoir la prétention de rendre une analyse impeccable. Avec des centaines, voire des milliers de centroïdes, une taille de fenêtre parfaitement adaptée au corpus, et bien plus de données, l'exercice serait sans doute plus probant. Mais nos ordinateurs seraient encore en train d'effectuer leurs calculs à l'heure actuelle... Plutôt que de se présenter à la cour pour faire sens du monde, on préférera pour l'instant se couler une tisane et ouvrir un roman.