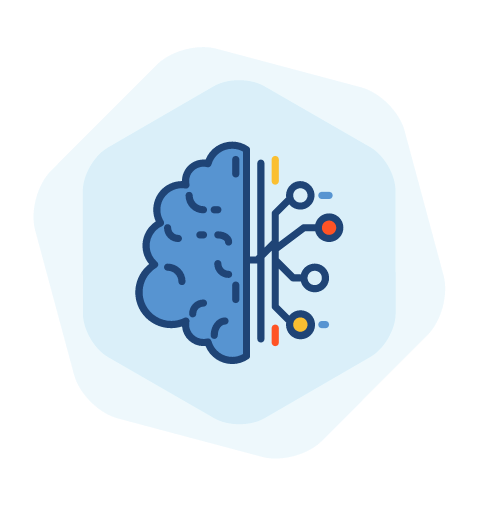
Algorithme K-means

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 matières**

[Résultats 3](#_Toc121081983)

[Première expérience 3](#_Toc121081984)

[Taille de fenêtre 7 / 30 centroïdes / 12 résultats par partition 3](#_Toc121081985)

[Seconde expérience 5](#_Toc121081986)

[Taille de fenêtre 5 / 40 centroïdes / 10 résultats par partition 5](#_Toc121081987)

[Discussion 7](#_Toc121081988)

[Conclusion 7](#_Toc121081989)

## Résultats

C

urieux 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, 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[[1]](#footnote-1).

### 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 |  |
| Verbes d’action | 4 | 6, 17, 23, 28 | 1004, 14524, 9873, 17459 |  |
| Noms | 0 | 1 | 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 -> Environ 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

Précision sur les résultats (on garde seulement les résultats jugés réellement pertinents, c’est-à-dire notés avec la couleur verte.

Première expérience

* Texte littéraire ratisse plus large, texte juridique plus ciblé (attendu). S’exprime par un plus grand nombre d’identités différentes du côté du texte littéraire et par davantage d’identités jugées moyennement convaincantes.
* Même performance des deux côtés! Une valeur de vérité semblable ? Attente : texte juridique avec un meilleur score.

Seconde expérience

Les deux corpus prennent une tangente très différente dans ce cas. Une augmentation de 25% centroïde avec une taille de fenêtre réduite augmente les performances de pertinence du corpus juridique de 6%, en générant énormément de bruit (des centroïdes pas intéressants). Du côté des textes littéraires, on parle plutôt d’une augmentation foudroyante de 28%! Toutefois, le nombre de centroïdes listant des stopwords a considérablement augmentée. Une sorte de redondance dans l’identité, par conséquent, des centroïdes.

La taille de la fenêtre a semblé diminué grandement la pertinence des résultats des textes juridiques : beaucoup de centroïdes se sont révélés complètement vides de sens, car seulement remplies de chiffres. Ce n’était pas le cas avec moins de centroïdes et, surtout, une taille de fenêtre relativement plus élevée. 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 imposent probablement une taille de fenêtre plus élevée pour des résultats plus intéressants.

Conclusion générale : Que peut-on dire à la lumière des résultats des deux corpus lorsque mis en comparaison ?

1. Vert : identité forte / Jaune : identité partiellement pertinente / Rouge : identité très faible. [↑](#footnote-ref-1)