CÉGEP DU VIEUX MONTRÉAL

**Projet Final**

Clustering - kmeans

Réalisé par

Caroline Emond-Serret,

François Lapierre,

Dany Viens

Présenté à

Pierre Paul Monty

Dans le cadre du cours 420-C62-VM

« Données, mégadonnées et intelligence artificielle II»

Le 18 mai 2021

Table des matières

[INTRODUCTION 2](#_Toc72239689)

[RÉSULTATS 3](#_Toc72239690)

[DISCUSSION 8](#_Toc72239691)

[Observations générales. 8](#_Toc72239692)

[Hypothèse 1 : Plus le score sera élevé, plus le mot sera commun. 9](#_Toc72239693)

[Hypothèse 2 : La densité d’un cluster influence sa qualité. Plus c’est dense plus c’est précis. 10](#_Toc72239694)

[Hypothèse 3 : Les catégories syntaxiques (mot, verbe, adjectif, adverbe, etc.) auront prédominances sur les champs lexicaux dans les résultats. 11](#_Toc72239695)

[CONCLUSION 12](#_Toc72239696)



# INTRODUCTION

Le but du travail était d’implanter l’algorithme k--means pour tester différentes configurations sur la qualité des résultats obtenus. L’algorithme utilise les données des matrices de cooccurrences générées avec le fichier Entrainement.py.

Nous avons procédé à l’entraînement des trois textes suivant :

* Les Trois Mousquetaires
* Germinal
* Le Ventre de Paris

Le total des mots distincts obtenus est de 25 402. Nous avons procédé à l’entraînement de ces trois textes avec des fenêtres de : 5, 7 et 9 mots. Les résultats de l’entraînement sont stockés dans le fichier "synonyms dict.db"

Nous avons formulé trois hypothèses en nous basant sur les résultats des deux précédents projets :

1. Plus le score sera élevé, plus le mot sera commun.
2. La densité d’un cluster influence sa qualité. Plus c’est dense plus c’est précis.
3. Les catégories syntaxiques (mot, verbe, adjectif, adverbe, etc.) auront prédominances sur les champs lexicaux dans les résultats.

# RÉSULTATS

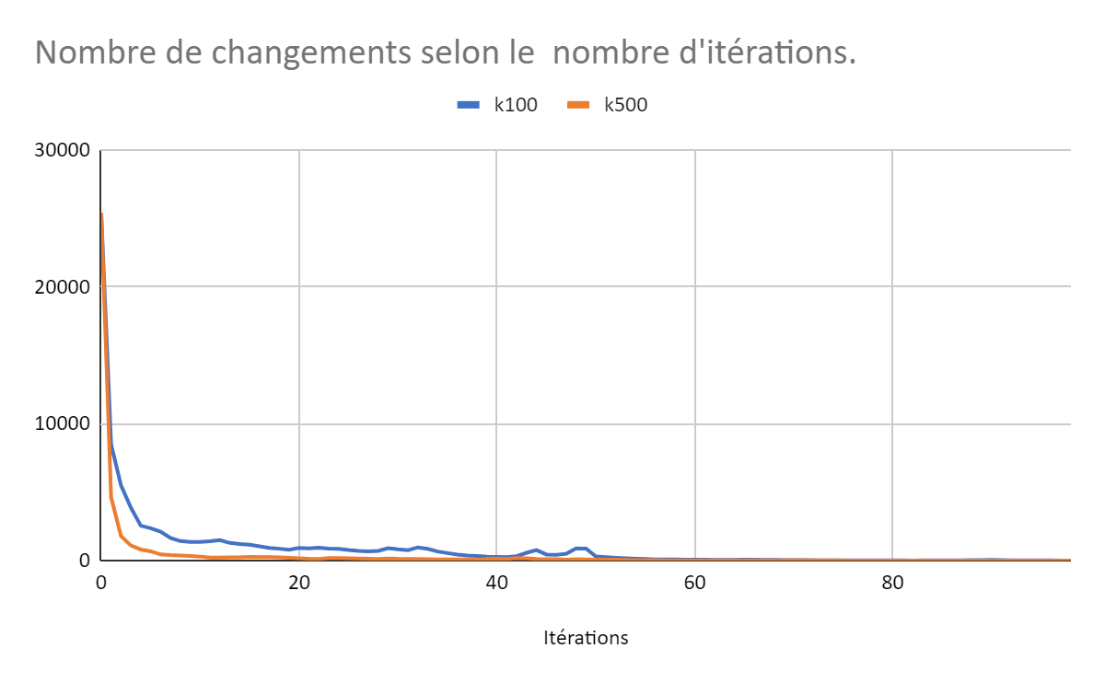
Les résultats ont été obtenus en effectuant les calculs à partir du même ordinateur, nous avons décidé d’utiliser l’ordinateur offrant les meilleures performances pour nos calculs. Nous avons choisi six tailles de clusters (k : 10, 20, 50, 100, 250, 500) différentes pour obtenir l’ensemble des résultats et nous avons testé les calculs sur trois tailles de fenêtre différentes (5,7 et 9).

Étant donné le temps nécessaire pour compléter le calcul pour le k500 et k250 et que nous désirions garder un seul ordinateur pour les calculs, nous n’avons pu faire plusieurs calculs avec les mêmes paramètres. Ainsi, nous n’avons pas la possibilité d’analyser la variance des résultats avec les mêmes paramètres, c’est pourquoi nous avons décidé de présenter les résultats sous forme de la moyenne de tous nos résultats.

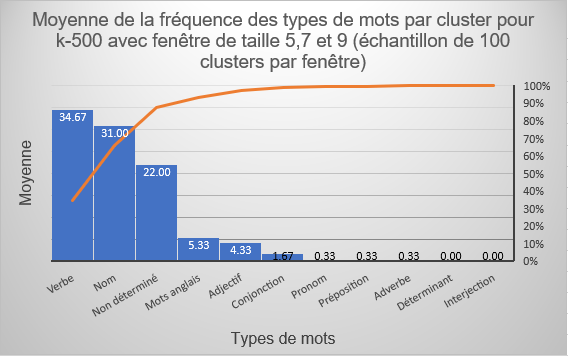
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Nombre de clusters** | **500** | **250** | **100** | **50** | **20** | **10** |
| Nombre d’itérations | 59,67 | 107,00 | 118,67 | 159,67 | 88,00 | 68,00 |
| Temps moyen par itération (s) | 382,80 | 180,53 | 67,31 | 29,53 | 12,69 | 7,35 |
| Temps Total (s) | 22841,14 | 19261,11 | 8022,37 | 4730,20 | 1116,62 | 500,40 |

**Tableau 1 : Moyenne des résultats des fenêtres 5 à 9 selon le nombre de clusters**

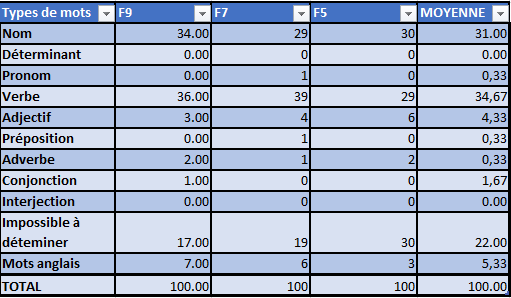
**Figure 1 : Nombre moyen d’itérations selon le nombre de clusters**



**Figure 2 : Graphique démontrant le nombre de changements par itérations pour les calculs effectués avec une fenêtre de 5.**



**Figure 8 : Graphique démontrant la moyenne de la fréquence des types de mots par cluster pour k-500 avec la fenêtre de taille 5, 7 et 9 (échantillon de 100 clusters par fenêtre)**



**Tableau 2 : Tableau des mots d’un échantillon de 100 clusters par taille de fenêtre (5,7 et 9) divisé selon la catégorie de mots.**

# DISCUSSION

## Observations générales.

En analysant les données figurant dans le Tableau 1, nous observons que le temps moyen de calcul par itération suit une augmentation constante et linéaire. Quant au temps total (en secondes) requis pour effectuer l’ensemble des calculs, l’augmentation ne suit pas une augmentation linéaire compte tenu que le nombre d’itérations varie grandement d’un exercice à l’autre.

En observant les résultats de la Figure 1, nous constatons que les données sont réparties sous forme de cloche s’approchant d’une courbe de distribution de la loi normale. En effet, le nombre d’itérations est nettement plus élevé pour le k50 et le k100, contrairement au nombre plus faible du k10 et k500.

Néanmoins, compte tenu le faible échantillon disponible pour chaque catégorie (3), il est difficile de s’avancer plus loin que de la simple observation. Il aurait été intéressant d’acquérir un échantillon de données plus significatif avec plus de temps disponible.

En analysant la Figure 2, le nombre de changements par itérations suit une distribution exponentielle négative. En effet, comme l’algorithme Kmeans tend à stabiliser les clusters, il est normal de constater que le nombre de changements se rapprochant invariablement de zéro.

## Hypothèse 1 : Plus le score sera élevé, plus le mot sera commun.

|  |  |
| --- | --- |
| Deux groupes résultats de fenêtre9 et cluster 500 | |
| **Groupe 62** | **Groupe 106** |
| **Score bas** | **Score haut** |
| confondît --> 5.14 | moi --> 46182.33 |
| massacrât --> 5.30 | monsieur --> 51899.50 |
| foute --> 5.34 | donc --> 59870.66 |
| ouvrît --> 5.37 | votre --> 83933.16 |
| grondât --> 5.39 | mon --> 94099.33 |
| fichât --> 5.44 | fait --> 95692.16 |
| beignets --> 5.45 | me --> 96515.33 |
| jase --> 5.48 | athos --> 134285.16 |
| outrée --> 5.48 | nous --> 238999.33 |
| emportât --> 5.48 | bien --> 265233.33 |

**Tableau 3 : Exemple de résultats avec score haut et score bas**

En prenant l’exemple du **Tableau 3,** nous constatons que lorsque les mots réunis dans un groupe possèdent un score élevé, ces mots sont généralement des mots plus fréquemment utilisés dans la langue française. Tandis que, lorsqu’un groupe de mots possède un score plus bas, ces mots ont tendance à être moins communs.

Par exemple, six mots du groupe 62 emploient le passé simple comme temps de verbe (fichât, emportât, confondît, massacrât, ouvrît, grondît). Il s’agit ici de mots devant être utilisés dans un contexte plus précis et moins commun. Le groupe 106 quant à lui, utilise des déterminants possessifs (moi, votre, mon, me, nous) ou des allocations de titre ou de personnage (monsieur, Athos). Il s’agit ici de mots pouvant être utilisés très régulièrement dans un roman.

Compte tenu la taille du corpus, des mots non reliés en apparences pourraient être associés parce qu’ils n’ont pas été utilisés à de nombreuses reprises dans les textes. Nous avons remarqué que plus les scores étaient élevés, plus les mots peuplant le groupe risquaient d’être des mots que nous avions insérés dans notre liste de *stop-words.*

En ce sens, nous croyons que notre hypothèse est valide.

## Hypothèse 2 : La densité d’un cluster influence sa qualité. Plus c’est dense plus c’est précis.

Méthode : Calculer l’écart moyen d’un cluster à partir de ses dix premiers résultats.

Densité : Plus l’écart moyen sera petit plus le cluster sera considéré comme dense.

Échantillon :

* 20 clusters avec les fenêtres de taille 5, 7 et 9
* 100 clusters avec les fenêtres de taille 7 et 9
* À partir des résultats de 500 clusters avec les fenêtres 5, 7 et 9 : Des clusters où l’on a observé une forte corrélation des mots entre eux.

En premier lieu, en observant les figures 3, 4 et 5, les groupes les plus denses sont le 3 (fig.3, écart moyen(em) = 0,06), le 0 (fig.4, em = 0,16) et le 15 (fig.5, em = 0,13). Dans le groupe 3 des résultats de la fenêtre 5, nous pouvons voir que les mots n’ont pas une signification similaire, mais que ceux-ci sont majoritairement des noms communs. Ensuite dans le groupe 0 des résultats de la fenêtre 7, nous n’avons pas trouvé de lien unissant les résultats. Dans le groupe 15 des résultats de la fenêtre 9, il y a peu ou pas de lien entre les mots. La conclusion préliminaire est qu’il n’y a pas de corrélation entre la densité et la qualité des clusters lorsque l’expérience est faite sur 20 clusters.

En second lieu, faisons les mêmes observations sur des échantillons différents soit ceux des figures 6 et 7. Les groupes les plus denses sont le 2 (fig. 6, em = 0,034) et le 18 (fig. 7, em = 0,05). Le groupe 3 de fenêtre 7 contient d’abord une majorité de verbes (6/10) et presque l’ensemble des mots (8/10) sont liés par un thème commun. Ensuite, dans le groupe 18, on retrouve une similitude dans les mots. Ils sont en forte majorité des noms masculins pluriels. De plus, certains mots ont un lien sous le thème de la nourriture : maquereaux, lards, prunes et jambons. Il semblerait, en examinant ces résultats, qu’il y est un rapport cette fois entre la densité et la qualité d’un cluster. En ajoutant ces données à ceux précédemment récoltés, on peut ajouter à notre conclusion préliminaire que si la quantité de clusters augmente alors on observe une corrélation plus importante entre la densité et la qualité.

Finalement, les dernières observations se feront dans le sens contraire, c’est-à-dire observé des clusters dans lesquels nous avons observé des corrélations évidentes et vérifier si on obtient un écart moyen bas. Ces examens se feront sur les recherches faites sur 500 clusters de la fenêtre de taille 7[[1]](#footnote-1). D’abord avec le cluster numéro 43, où l’on obtient des verbes à la troisième personne du pluriel, on a un écart moyen de 1,97. Ensuite avec le cluster numéro 197, ou l’on obtient des verbes à la deuxième personne du pluriel, on a un écart moyen de 58,44. Finalement avec le cluster numéro 259, où l’on retrouve des mots qui sont soit des synonymes ou des antonymes, on obtient un écart moyen de 4,00. Clairement, les groupes 43 et 259 ont une densité élevée, mais même si l’écart-moyen du groupe 197 est plus haut, ce groupe a aussi une densité haute, car la moyenne des écarts moyens des données recueillies se situe entre 89 059,76 et 643 982,02.

Conclusion : Plus la densité est haute ET que le nombre de clusters est élevé plus la qualité des résultats sera bonne la majorité du temps. L’hypothèse tend à être confirmée, néanmoins une recherche plus exhaustive pourrait nous le démontrer hors de tout doute.

## Hypothèse 3 : Les catégories syntaxiques (mot, verbe, adjectif, adverbe, etc.) auront prédominances sur les champs lexicaux dans les résultats.

Trois K-500 ont été effectués, soit un clustering pour chaque taille de fenêtre 5, 7 et 9. Parmi les résultats illustrés par le **tableau 2**, les 100 groupes (clusters) de chacun ont ensuite été sélectionnés aléatoirement comme échantillon, dans le but d’analyser l’homogénéité des types de mots regroupés ensemble. Si 60% des mots d’un même groupe appartenaient au même type mots (déterminants, verbes, adjectifs, etc.), l’occurrence du type de mots en question était incrémentée de 1. Lorsqu’un cluster était trop mixte pour déterminer une tendance, la catégorie « Impossible à déterminer » était incrémentée.

La **figure 8** illustre la moyenne d’occurrence de chaque type de mots à l’intérieur des trois échantillons. Visuellement, les résultats sont assez clairs : les clusters les plus homogènes contenaient plus souvent des verbes (34.67%) et des noms (31%). Il était également impossible à déterminer une tendance précise dans 22% des groupes. La courbe orange du graphique, illustrant le total des moyennes sur 100%, montre que ces trois catégories à elles seules constituent presque 90% des résultats. Notre hypothèse est que cela est dû à la structure de la langue française elle-même, les phrases étant construites et ordonnées la forte majorité du temps en sujet, verbe et complément. Les verbes et les noms se retrouvent ainsi souvent à la même position dans une phrase. La présence, quantité et position d’adjectifs, adverbes, etc., peut, quant à eux, varier selon la phrase ou la prose de l’auteur, ce qui pourrait sans doute expliquer le nombre de groupe mixte « Impossible à déterminer ».

Selon les données recueillies, l’hypothèse semble être confirmée.

Au-delà de l’homogénéité des groupes de mots selon leur type, nous réalisons que l’algorithme du k-means réussit également à grouper des mots selon une expression ou un champ lexical rapproché. Dans les résultats du K-500 avec taille de fenêtre 5 par exemple, le groupe 222 contient [six, vingt, cinq, huit, dix, ans], possiblement dû à l’énumération du temps (ex. : « six ans plus tard ») ou lorsque l’âge des personnages est mentionné dans les textes. Le groupe 332 [toutes, hommes, mains, tous, autres, yeux, trois, deux] est également intéressant. À première vue, il semble un peu hétéroclite, même si quelques pairs de mots se ressemblent lexicalement (toutes/tous, mains/yeux, trois/deux). Si l’on considère toutefois la façon dont les mots sont agencés dans une phrase, ce groupe de mots prend alors tout son sens : « tous les autres », « tous les hommes », « toutes les trois », « tous les deux », « tous les yeux », etc.

L’algorithme regroupait également parfois plusieurs mots évoquant ensemble un champ lexical commun ou un certain lieu. Par exemple, le groupe 312 [caisses, plats, ventres, fromages, énormes, vagues, carottes, travailleurs, roses, bruits] pourrait faire référence à un marché ou la place publique, où des marchands exhibent d’ « énormes caisses » de « fromages », de « carottes » ou de « roses », des restaurants prêts à servir leurs « plats » à des « vagues » de « travailleurs » affamés aux « ventres » vides. Il en est de même pour le groupe 234 [port, jardin, canal, salon, bastion, bureau, mercier, corridor, sol, marché], dont les mots représentent tous des lieux (sauf le mot « sol », qui aurait pu toutefois décrire l’état des lieux). Ce sont tous des mots qui auraient pu décrire une ville/village, ou préciser les déplacements d’un personnage, tels un « corridor » menant au « bureau » ou « salon », un « canal » menant au « port », se rendre chez le « mercier » au « marché », etc.

# CONCLUSION

Tel que mentionné dans la majorité de nos observations, bien que nous commulons un nombre important de résultats, notamment avec les taille de fenêtre 500, il est difficile de généraliser nos observations étant donné que nous n’avons pu répété chaque calcul avec les mêmes paramètres pour s’assurer que nous obtiendrions les mêmes résultats.

Il aurait été intéressant d’ajouter d’autres paramètres dans l’algorithme k-means pour réduire le côté aléatoire des résultats. Par exemple, en assignant une étiquette à chaque corpus de mots (nom, verbe, adverbe, adjectif, etc.) pour s’assurer que chaque groupe contiennent uniquement des termes ayant le potentiel d’être un synonyme. Néanmoins, nous avons été surpris de la qualité de certains groupes parmis les résultats obtenus.

Nous sommes curieux de voir quel impact aurait pu avoir un corpus de textes plus grand ou encore l’utilisation d’un corpus de textes contemporains. Toutefois, cela nous a permis d’approfondir nos connaissances en intelligence artificielle en travaillant sur un même projet et en nous amenant à complexifier une application tout au long de la session et de mieux comprendre les enjeux reliés à la gestion d’un grand nombre de données. Bien que nous avons œuvré avec seulement des dizaines de milliers de mots, nous comprenons mieux les enjeux reliés aux calculs complexes impliquant des millions, voire de milliards, de données.

1. Aucune figure n’a été générée pour ces données. Le tableau à 500 colonnes n’était pas lisible, néanmoins une trace des calculs est disponible à côté de chaque groupe énuméré dans le fichier « resultats\_t7\_n10\_k500.csv » [↑](#footnote-ref-1)