## 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) 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 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 |

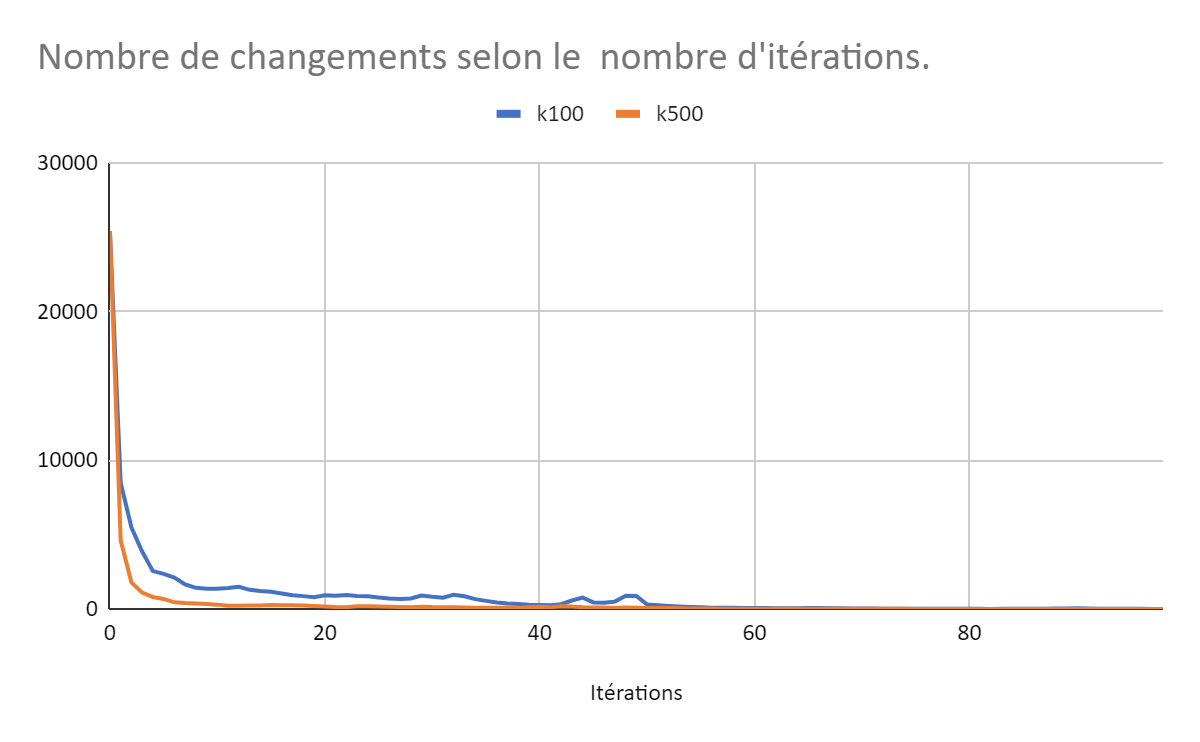
**Tableau 1 : Moyenne des résultats des fenêtres 5 à 9 selon le nombre de clusters**

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.

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

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 k10, 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.



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

En analysant la **Figure 2**, le nombre de changements par itérations suit une distribution expentionelle 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.

## Discussion :

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

|  |  |
| --- | --- |
| T9 - K500 | |
| **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 2 : Exemple de résultats avec score haut et score bas**

En prenant l’exemple du **Tableau 2,** nous constatons que lorsque les mots réunis dans un groupe possèdent un score élevé, ces mots sont généralement des mots plus fréquemment utilisés dans le langage français. Tandis que, lorsqu’un groupe de mots possède un score plus bas, ces mots ont tendance à être moins commun dans le langage commun.

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.