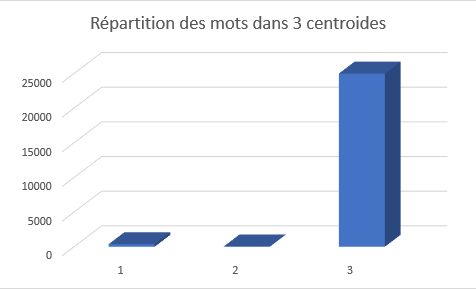
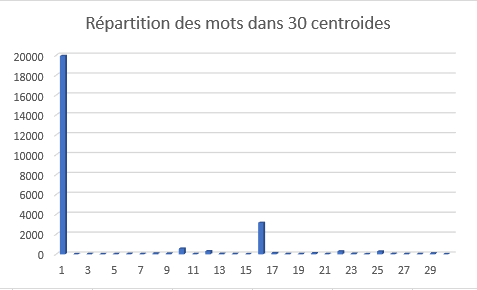
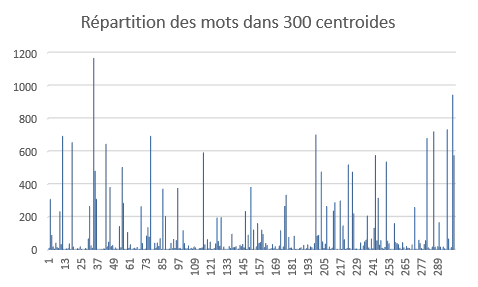
**Rapport sur les centroides**

L’analyse porte sur seulement trois textes (Le ventre de Paris, Germinal, Les trois mousquetaires) afin d’avoir un temps d’exécution raisonnable avec de nombreux centroïdes. Le total de mots différent est d’environ 25402.

Tous les jeux de données on était réalisé avec une taille de fenêtre de 5 et un nombre de mot à afficher de 5.

La première observation que nous avons pu constater est que peu importe le nombre de centroïde, il y avait forcément un centroïdes qui était beaucoup plus peuplé comme le montre les graphiques suivants :

Les mots sont placés dans un centroïde grâce a la distance qui les sépares de celui-ci, plus la distance est faible plus la similitude est grande. Afin de comparer la distance qui le sépare d’un centroïde on utilise un vecteur qui est composé d’une multitude de points, les points représentant tous la fréquence de cooccurrence d’un mot avec tous les autres mots. Le phénomène est donc explicable par la présence de nombreux 0 dans certains vecteurs de mot ce qui crée une similitude lorsqu’ils sont comparés ensemble.



On voit ci-dessus la repartions des mots par centroïde dans un jeu de donnée de 300 centroïdes. On perçoit clairement que la répartition est beaucoup plus repartie que dans les deux exemples précèdent, il n’y a plus un seul centroïde contenant tous les mots, mais plusieurs. Toutefois, la répartition n’est toujours pas parfaitement équitable.

Pour la suite, on va comparer la pertinence des mots entre diffèrent centroïde sur les jeux de données précèdent.

Pour ce faire, on va regarder si les mots appartiennent au même champ lexical, la même famille, si ce sont des verbes. Avec ça, on va calculer un ratio du nombre de pertinent dans un même centroïde.

|  |  |
| --- | --- |
| Pertinent |  |
| Plus ou moins pertinent |  |
| Non pertinent |  |

|  |  |  |
| --- | --- | --- |
| Jeu de donnée 1 ( 3 centroïdes) | | |
| Centroïde 1 | Centroïde 2 | Centroïde 3 |
| mal | c | coque |
| tréville | avec | continuité |
| passer | comme | usages |
| garde | sa | vigilance |
| catherine | au | abrutissement |
| Ratio de pertinence | | |
| 40% | 80% | 0% |

Dans ce premier jeu de donnée, le centroïde 1 a les mots « Treville » et « catherine »qui tout deux sont des noms propres.

|  |  |  |
| --- | --- | --- |
| Jeu de donnée 2 ( 30 centroïdes) | | |
| Centroïde 2 | Centroïde 4 | Centroïde 29 |
| mot | ouvriers | work |
| cri | enfants | and |
| geste | camarades | in |
| regard | gardes | ebook |
| rire | pieds | this |
| Ratio de pertinence | | |
| 85% | 80% | 100% |

Dans ce deuxième jeu de donnée, le centroïde 2 a les mots « cri », « geste », « regard » et « rire ». Tous ces mots sont des actions. Le mot « mot » est un cousin éloigné du mot « cri » car on cri des mots.

|  |  |  |
| --- | --- | --- |
| Jeu de donnée 3 ( 300 centroïdes) | | |
| Centroïde 211 | Centroïde 244 | Centroïde 41 |
| déjeunerons | Retourner | cinquante |
| craignons | remettre | quarante |
| mordicus | conduire | cents |
| morfondons | recevoir | pistoles |
| acharnons | garder | cent |
| Ratio de pertinence | | |
| 80% | 100% | 100% |

Dans ce troisième jeu de donnée, le centroïde 41 a les mots « cinquante », « quarante », « cent », « cent ». Tous ces mots sont des chiffres qui pourraient servir à quantifier « pistoles » qui était une monnaie espagnole en 1537.

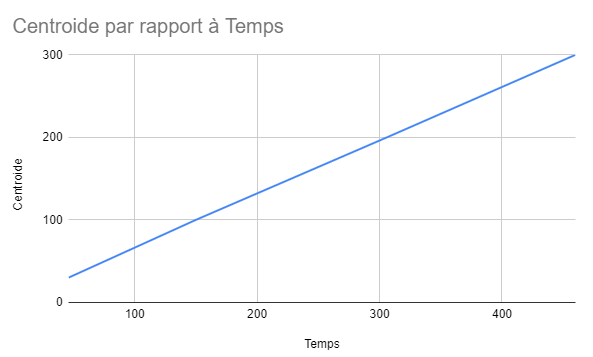
On peut en conclure que, plus le nombre de centroïde est élevée plus la pertinence des mots qui composent un même centroïde sera élevée.

Toutefois, plus le nombre de centroïde est élevée, plus le temps d’exécution est conséquent.

Par exemple, pour une exécution à 3 centroïdes, le temps moyen est de 4.6sec par itération.

Pour une exécution à 30 centroïdes, le temps moyen est de 46sec par itération.

Pour une exécution à 300 centroïdes, le temps moyen est de 460sec par itération.



Le temps est proportionnel au nombre de centroïde

On voit clairement que le temps d’exécution est calculable par une fonction linéaire de type y = ax+b

Pour conclure, les différentes configurations nous ont menées à des résultats tous différents. En effet, il est possible de constater une amélioration de la pertinence des mots dans les centroïdes lorsque le nombre de centroïde augmente. Cela s’explique par la diversité des centroïdes, un mot a donc plus de chance de trouver un groupe qui lui convient. La résultante est que les familles de mots au sein des centroïdes sont plus précises. Toutefois, les résultats obtenus n’excellent pas. Afin d’améliorer le produit, il serait nécessaire d’avoir un nombre de centroïde bien plus conséquent et une banque de texte bien plus riche. Néanmoins, le temps d’exécution ne serait que plus important. Les ressources demandées seraient astronomiques.