Figure 1

Figure 2

Figure 3

Figure 4

Figure 5

Hypothèse : 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 premier résultats. Plus le résultat 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 ou l’on a observé une forte corrélation des mots entre eux.

D’abord en observant les figure 1, 2 et 3, les groupes les plus dense sont le 3 (fig.1, em = 0,06), le 0 (fig.2, em = 0,16) et le 15 (fig.3, em = 0,13). Dans le groupe 3 des résultats de la fenêtre 5, on peut déjà 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, il est n’est pas possible de vraiment trouver de lien qui uni 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 que l’on peut faire 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 différent soit ceux des figures 4 et 5. Les groupes les plus denses sont le 2 (fig. 4, em = 0,034) et le 18 (fig. 5, em = 0,05). Le groupe 3 de fenêtre 7 contient d’abord une majorité de verbe (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 nom masculin pluriel. De plus, certain 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 récoltées, 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 observer des clusters dans lesquels nous avons observé des corrélations évidente et vérifier si on obtient un écart moyen bas. Ces examens se ferons sur les recherches faites sur 500 clusters de la fenêtre de taille 7. D’abord avec le cluster numéro 43, ou 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, ou l’on retrouve des mots qui sont soit des synonymes ou des antonyme, on obtient un écart moyen de 4,00. Clairement, les groupe 43 et 259 on une densité élevée, mais même s’il est un peu plus grand le groupe 197 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 cluster est élevé plus la qualité des résultats sera bonne la majorité du temps. Une recherche plus exhaustive pourrait nous le démontrer hors de tout doutes.