Analyse clustering :

Nous avons analysé les résultats en comparant la nature du groupe majoritaire des mots d’un cluster. Une valeur de 0.1 est attribuée à chaque mot faisant partie du groupe majoritaire. Les mots pouvant appartenir a plus d’une nature dont au groupe majoritaire se sont vu assigner a celui-ci. Cette méthode nous a permis d’évaluer la pertinence du tirage d’un clustering.

Résultat obtenu via le clustering vs la matrice de cooccurrence

­Les groupes de mots obtenus grâce à l’algorithme nous offre une facette très différente des synonymes de la matrice de cooccurrence. En effet, chaque cluster retourne une série de points lui étant spatialement proche. Ces mots ne sont pas des synonymes, mais ils sont liés grâce à leur classe ou fonction grammaticale, ce qui veut dire qu’un cluster représente un élément linguistique unique si on le met en relation avec les autres clusters originaires de la même exécution de l’algorithme. Par exemple, un cluster représente généralement une classe grammaticale, c’est-à-dire, les noms communs, les verbes, les adjectifs, etc. Cependant, il est possible qu’il représente une caractéristique commune comme des mots au pluriel ou au féminin.

Impact de taille de fenêtre sur la précision des clusters et la variété des clusters :

La production d’une série de test par bond de 5 de 5 à 20 de taille de fenêtre inclusivement permet d’obtenir un aperçu de l’impact de la taille de fenêtre sur la précision des clusters et sur leur variété. À l’aide des graphiques ci-dessus, on peut déterminer que la taille de fenêtre semble diminuer graduellement la précision de l’algorithme. Cependant, un moins grand lot de donnée est nécessaire pour détecter une plus grande variété de classe grammaticale. C’est résultat contraste avec ceux obtenu dans la précédente itération de ce travail où la taille de fenêtre obtenant les meilleurs synonymes était 15, dans le cas du clustering, il est préférable de maintenir la taille de fenêtre vers 5.

Impact du nombre de clusters sur la précision des clusters et la variété des clusters :

Le concept même de clustering défini qu’un nombre plus élevé de cluster permettra de découvrir plus de minimums locaux, ou en d’autres mots, plus de zone de densité d’information. Cette caractéristique se maintient dans notre contexte, mais son impact n’est pas seulement souhaitable. L’augmentation du nombre de cluster dans l’algorithme provoque une augmentation plus marquée du nombre de cluster indésirable, ceux retournant des mots incohérents ou un nombre trop peu nombreux. En effet, un trop grand nombre de cluster subdivisera des classes grammaticales sous plusieurs clusters permettant d’isoler leur fonction, leur genre ou leur nombre linguistique, malheureusement au risque de diminuer la précision de chacun.

Information extrapolée par le nombre de points dans un cluster :

Les clusters indésirables représente soit des données aberrantes ou démontre une imprécision dans le lot de mots initial. Il est important d’écarter ces clusters pour augmenter la qualité des informations produite par l’algorithme. Pour identifier un cluster indésirable, il faut identifier les clusters de basse qualité. Les tests démontrent qu’un cluster ayant un nombre de mots très bas, soit en dessous de 10 mots, ou une distance moyenne très haute, dans notre contexte en haut d’un million sont rarement précis dans les informations qu’ils contiennent.

Hypothèse d’une constante selon la qualité du texte et la classification des résultats :

Les résultats de notre échantillon de données offre en moyenne une précision au alentours de 60%, ce qui est relativement insatisfaisant. Cette tendance autour du 60% reste spécifique aux textes analysés et à la classification manuelle des données dans leur classe grammaticale. Ces deux aspects mettent la moyenne de précision sous le risque d’être affecté par un nombre insuffisant de donnée et par des erreurs cléricales. Il est fort probable qu’un plus grand échantillon de données et une classification automatique offre une moyenne plus proche des 75% comme le démontre le nom commun qui est la classe grammaticale la plus facile à isoler.

I

Information extrapolée par la moyenne de pertinence des natures majoritaire :

Les résultats de notre échantillon se sont avérer avoir des scores de pertinence dramatiquement plus important pour certaines natures de mots. En effet, les score des centroïde majoritairement composé de noms ou de déterminant se sont avérer dégager des scores moyens dans les 70 à 80% là ou les verbes et conjonctions ne récoltes qu’un maigre 40%. Nous expliquons ces différences marquées par l’aspect polymorphe de certains mots par rapport à leur nature. Prenons en exemple le mot « Tout », pouvant être utiliser comme adjectif, comme adverbe, comme pronom et nom; un centroïde balancer autour de « tout » risquerait alors de retourner des mots de toute ces natures. Les nom ont eu une tendance à retourner des valeurs de pertinence plus importante considérant qu’ils ont moins fréquemment tendance à pouvoir faire partie de plus d’une nature. Les centroïde ayant tendance a ce diriger vers les maximum locaux, nous avons donc déduis que les retour plus fréquent de noms comme groupe majoritaire était attribuable a cette tendance.