Université Libre de Bruxelles
Sciences et Technologies de l'Information et de la Communication

Traitement Automatique de Corpus

Rapport TP3

Wala Zerelli

Année académique

2024-2025

1. Clustering

La décennie choisie pour ce TP est 1960-1969. Le clustering s'est fait avec le *s2_clustering* divisé en 5 clusters, représenté dans le tableau suivant :

N° Cluster	Année	N° Documents	N° total de documents
1	1960	15 documents	110
	1961	8 documents	
	1962	17 documents	
	1963	10 documents	
	1964	11 documents	
	1965	11 documents	
	1966	8 documents	
	1967	8 documents	
	1968	10 documents	
	1969	12 documents	
2	1960	10 documents	125
	1961	13 documents	
	1962	15 documents	
	1963	11 documents	
	1964	11 documents	
	1965	10 documents	
	1966	12 documents	
	1967	10 documents	
	1968	16 documents	
	1969	17 documents	
3	1960	9 documents	99
	1961	16 documents	
	1962	12 documents	
	1963	12 documents	
	1964	11 documents	
	1965	7 documents	
	1966	13 documents	
	1967	6 documents	
	1968	9 documents	
	1969	4 documents	
4	1960	22 documents	192
	1961	14 documents	
	1962	12 documents	
	1963	17 documents	
	1964	11 documents	
	1965	21 documents	
	1966	23 documents	

	1967	21 documents	
	1968	23 documents	
	1969	28 documents	
5	1960	44 documents	476
	1961	49 documents	
	1962	43 documents	
	1963	50 documents	
	1964	56 documents	
	1965	51 documents	
	1966	44 documents	
	1967	55 documents	
	1968	45 documents	
	1969	39 documents	

L'analyse de la distribution des documents par année au sein des cinq clusters révèle des tendances intéressantes. Bien que les clusters 1, 2 et 3 présentent des volumes de documents relativement similaires, le cluster 4 se distingue par un nombre significativement inférieur. En revanche, le cluster 5 se démarque nettement en regroupant près de la moitié de l'ensemble des documents.

Cette répartition inégale suggère que les documents ont été regroupés en fonction de critères distinctifs, créant ainsi des clusters aux caractéristiques bien définies. Il est à noter que, malgré ces différences en termes de volume, la répartition des documents par année au sein de chaque cluster est relativement homogène sur la période 1960-1969, suggérant une cohérence temporelle au sein de chaque groupe

Partie sur la similarité j’ai pas su le faire.