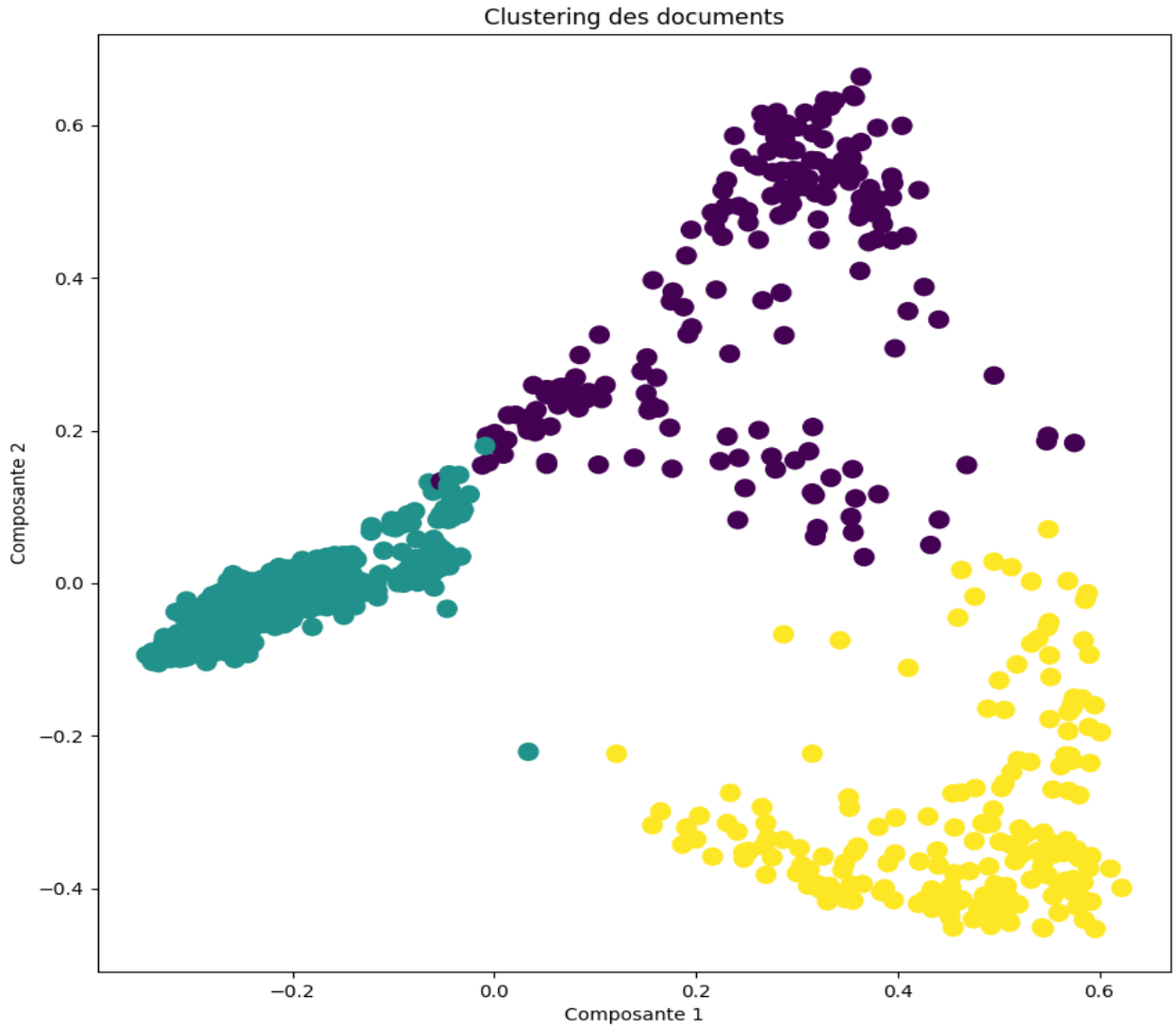
TP3 : Clustering et word2vec
Taleb Ahmed 000575310
https://github.com/00atal/tac

1.1 Clustering

Suite à l'étape de clustering demandée dans le TP3, j'ai choisi d'adopter une division en **trois clusters**. Ce choix n'a pas été fixé arbitrairement : il découle directement d'une observation empirique du nuage de points obtenu après réduction de dimension (PCA ou t-SNE).
En effet, le graphique révèle clairement **trois agglomérations distinctes**, bien séparées les unes des autres. Chaque groupe forme une zone dense et relativement cohérente, ce qui suggère l'existence naturelle de trois ensembles de documents partageant un vocabulaire ou un style similaire. Bien qu'elle ne soit pas parfaite, cette séparation visuelle justifie donc l'utilisation de $k = 3$ pour l'algorithme de clustering.

1.2 Résultats

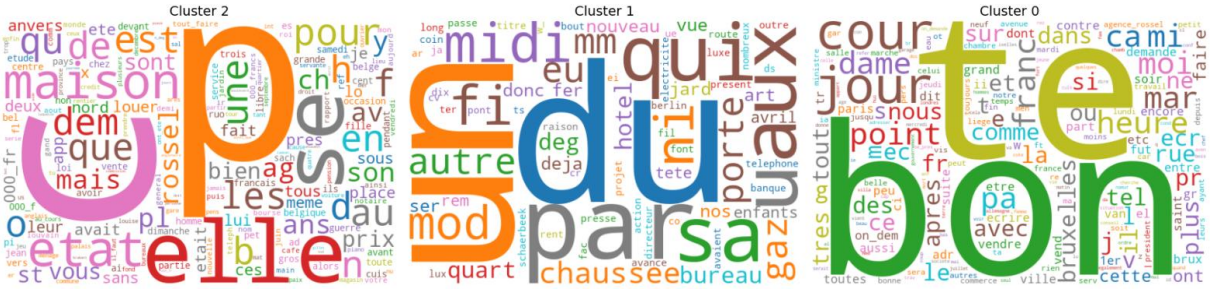


Les mots-clés TF-IDF se regroupent principalement autour de trois axes thématiques :

1. **Vocabulaire administratif et annonces**
   On retrouve des termes liés à la correspondance professionnelle ou institutionnelle, comme *rossel*, *écrire*, *agence*, *bilingue*, *curriculum*, *manuscrite*. Ces mots reflètent un contexte d'organisation, de communication formelle ou d'offres publiées.

2. **Actualité politique et internationale**
   Plusieurs mots indiquent un intérêt pour les affaires politiques et géopolitiques, par exemple *gouvernement*, *parlement*, *réformes*, *Guerre froide*, *URSS*, *États-Unis*, *conflits*. Cela suggère une couverture de sujets politiques, diplomatiques ou historiques.

3. **Société, culture et faits divers**
   Enfin, le vocabulaire touche la vie sociale, culturelle et quotidienne : *cinéma*, *théâtre*, *littérature*, *jeunesse*, *école*, *travail*, *famille*. Ces mots mettent en avant les activités culturelles et les questions sociales, ainsi que des récits ou informations de la vie quotidienne.



2.1 Word2vec

Dans cette section, j'ai d'abord extrait les bigrammes et trigrammes à partir du fichier sents.txt, afin de compléter les unigrammes déjà présents et de constituer un corpus d'n-grams unifié.
Une fois le corpus finalisé, plusieurs modèles Word2Vec ont été entraînés pour comparer leurs performances.
Le quatrième modèle s'est révélé être le plus efficace en termes de temps de calcul. Il a été configuré avec les paramètres suivants :

- Dimension des vecteurs : 32
- Taille du contexte autour du mot cible : 3
- Fréquence minimale d'apparition des mots : 10
- Nombre de threads pour l'entraînement : 5
- Nombre d'itérations sur le corpus (epochs) : 5

Après l'entraînement, le modèle a été enregistré pour un usage ultérieur.

Une inspection rapide du corpus a toutefois montré la nécessité d'un nettoyage supplémentaire : suppression des stopwords, de la ponctuation, des chiffres et des caractères isolés avant de poursuivre l'analyse.

2.2 Résultats

Dans cette partie, j'ai utilisé similarity pour comparer des termes liés à des faits historiques des années 1960 et évaluer si le modèle capture leurs relations contextuelles et géopolitiques.

MOT 1	MOT 2	SIMILARITÉ	HYPOTHÈSE / INTERPRÉTATION
république	gaulle	0.725	Très forte similarité, ce qui reflète le lien historique entre Charles de Gaulle et la République française.
paris	urss	0.502	Similarité modérée, probablement due à des contextes géopolitiques ou historiques où Paris et l'URSS apparaissent ensemble (ex. guerre froide, relations internationales).
mur	automobile	0.288	Similarité faible, peut refléter des contextes où les mots apparaissent ensemble mais sans lien sémantique fort (ex. mention d'obstacles dans des textes techniques ou historiques).
anvers	francs	-0.034	Similarité négative, ce qui suggère que ces mots apparaissent dans des contextes opposés ou non liés dans le corpus (Anvers = ville, francs = monnaie, peu de co-occurrences).

Dans la même logique, j'ai utilisé most_similar pour comparer des termes liés aux faits historiques marquants des années 1960.
J'ai choisi ces mots car cette période est caractérisée par des événements et phénomènes majeurs :

- **Mur de Berlin et Guerre froide**, représentés par le mot *Berlin*.
- **Indépendances africaines**, représentées par le mot *Afrique* ;
- **Cinquième République en France**, symbolisée par le mot *république* ;
- **Pop culture**, illustrée par le mot *pop* ;

Ces termes permettent de vérifier si le modèle capture correctement les relations contextuelles et historiques de cette période.

Terme cible	*Top 3 mots les plus similaires*	*Interprétation / Hypothèse*
berlin	londres, moscou, rome, geneve, washington, budapest, vienne, berne, teheran, tokio	Les villes les plus similaires reflètent les grandes capitales impliquées dans la géopolitique de la guerre froide, ce qui montre que le modèle capture le contexte international de Berlin.
afrique	algerie, amerique, europe, asie, ukraine, indochine, europe_centrale, irak, abyssinie, indonesie	Les associations reflètent des régions et pays souvent mentionnés dans le contexte colonial, décolonisation et affaires internationales des années 1960.
republique	republique_francaise, cour_supreme, diete, famille_royale, municipalite, republique_argentine, delegation, douma, reichswehr	Le modèle capture des entités politiques et institutions, montrant une forte cohérence contextuelle autour des systèmes républicains et gouvernements.
pop	popul, eleg, fraric, farn, jpet, tranqu, euv, pct, bpn, aerv	Les mots proches semblent issus d'abréviations ou de termes liés aux médias/pop culture, montrant que le modèle capture des co-occurrences fréquentes même si le sens précis est plus bruité.

Hypothèse générale : Les résultats montrent que le modèle identifie correctement les relations contextuelles pour des termes historiques ou géopolitiques (berlin, afrique, republique), tandis que pour des termes plus génériques ou bruités (pop), la similarité reflète surtout des co-occurrences fréquentes dans le corpus.

Enfin j'ai essayé un calcule entre vecteurs « France : Paris = Belgique : ? » avec Word2Vec mais ça ne marche pas car mon modèle connaît mal les relations pays-capitale et renvoie juste les villes proches dans le vecteur global (Londres, New York, Leipzig...).