Rapport de stage d'initiation

Effectué du 12 Juin 2023 au 30 Juillet 2023

Filière : Cycle d'ingénieur en DATA SCIENCE

Niveau : 1ère Année

Entreprise d'accueil :

Elaboré par : ZITOUNI Hajer

Année Universitaire : 2022/2023

Rapport de stage d'initiation

Effectué du 12 Juin 2023 au 30 Juillet 2023

Filière : Cycle d'ingénieur en DATA SCIENCE

Niveau : 1ère Année

Entreprise d'accueil :

Elaboré par : ZITOUNI Hajer

Responsable à l'entreprise :	*Avis de la commission des stages*

Année Universitaire : 2022/2023

Remerciements

Au terme de ce stage, je tiens à exprimer mes connaissances et mes sincères remerciements au mon encadrant ***Mr. Mohamed Marrouchi*** pour m'avoir encadrée et orientée durant mon projet d'initiation ainsi que pour ses remarques judicieuses, sa sympathie et son amitié.

Mes remerciements s'étendent également à tout le personnel de **Hexastack** pour leur chaleureux accueil, leur aide précieuse et leurs conseils.

A tous ceux qui ont contribué de près ou de loin à la réalisation de ce travail modeste trouvent ici l'expression de mes sincères gratitudes.

TABLE DES MATIERES

LISTE DES FIGURES

LISTE DES ACRONYMES

IA : Intelligence Artificielle.

NLP : Natural Language Processing.

ML : Machine Learning.

TF-IDF : Term Frequency-Inverse Document Frequency.

Bi-classe: Binary classification

Multi-classe : Multiple classification

Introduction Générale

Titre : Analyse Comparative des Algorithmes de Regroupement de Phrases ou de Textes Non Supervisés et des Techniques de Prétraitement NLP pour Comparer l'Exactitude entre les Différents Modèles

Introduction :

Le Traitement Automatique du Langage Naturel (NLP) est un domaine de recherche essentiel qui vise à permettre aux machines de comprendre, interpréter et générer le langage humain de manière intelligente. L'une des tâches cruciales en NLP est le regroupement de phrases ou de textes, qui consiste à classer automatiquement des documents ou des segments textuels similaires dans des groupes distincts sans étiquettes préalables. Cette tâche d'apprentissage non supervisée revêt une grande importance dans divers domaines tels que la recherche d'informations, la catégorisation de documents, la recommandation de contenu et l'analyse de sentiments.

Dans le cadre de ce stage d'été, nous avons concentré nos efforts sur l'exploration et la comparaison de plusieurs algorithmes de regroupement de phrases ou de textes non supervisés. Nous avons également étudié l'impact de différentes techniques de prétraitement du langage naturel sur les performances de ces modèles de regroupement. Ces techniques de prétraitement permettent de préparer les données textuelles brutes avant de les soumettre aux algorithmes de regroupement.

L'objectif principal de ce travail était de déterminer quel algorithme de regroupement et quelles techniques de prétraitement NLP sont les plus appropriés pour obtenir des clusters cohérents et informatifs à partir de données textuelles non supervisées. Pour atteindre cet objectif, nous avons évalué

quantitativement et qualitativement chaque modèle de regroupement en utilisant des mesures appropriées, telles que l'évaluation de la similarité intra-cluster et la dissimilarité inter-cluster.

Dans les sections suivantes de ce rapport, nous décrirons en détail la méthodologie utilisée pour mener ces expérimentations, y compris la description de l'ensemble de données utilisé, les différentes étapes de prétraitement NLP mises en œuvre, ainsi que la mise en place des algorithmes de regroupement. Nous présenterons ensuite les résultats obtenus et discuterons des conclusions tirées de cette étude comparative. Enfin, nous soulignerons les implications pratiques de nos découvertes et discuterons des perspectives futures pour améliorer davantage les performances des modèles de regroupement de phrases ou de textes non supervisés dans des applications réelles de NLP.

Il convient de noter que les résultats et les recommandations présentés dans ce rapport sont basés sur les spécificités de notre stage d'été, mais les méthodes et les enseignements pourraient être extrapolés à d'autres contextes et fournir une base solide pour des recherches ultérieures dans le domaine du regroupement de textes en NLP.

Chapitre 1:
Contexte Général

Introduction :

Dans ce chapitre, nous allons évoquer brièvement la présentation de l'organisme d'accueil, ses missions et son organisation. Par la suite, nous allons introduire le contexte du projet et déduire la problématique. Enfin, nous allons conclure ce chapitre par la présentation de la solution proposée.

1.1 Présentation de l'organisme d'accueil :

1.1.1 Présentation de l'entreprise :

***Hexastack* :** est une entreprise informatique qui explore constamment de nouvelles technologies, en particulier dans le développement web et mobile. Fondée en 2018, Hexastack couvre toutes les étapes et les domaines du développement logiciel, comprenant le frontend, le backend ainsi que le devops et la sécurité. L'objectif principal de l'entreprise est de fournir un bon service aux clients, en particulier avec un solide support technique et une maintenance.

Figure 1: Hexastack Logo

1.1.2 Produits et services :

Hexastack vise à introduire plusieurs produits logiciels sur le marché, tels que Hexabot.io et HexaMaps. Hexabot.io est un outil multiplateforme alimenté par l'IA pour créer des chatbots robustes. Hexabot peut actuellement fonctionner sur Facebook, Messenger, Twitter et même être intégré dans des sites web. HexaMaps est un outil de création de cartes. Avec cette application, l'utilisateur peut personnaliser et enregistrer différentes cartes, ajouter des extensions et consulter des statistiques.

Figure 2: Hexabot.io Logo

1.2 Présentation du projet :

1.2.1 Cadre du projet :

Ce projet s'inscrit dans le cadre d'un stage d'été en première année du cycle d'ingénieur en Data Science à la faculté des sciences de Tunis. Ce stage dont la durée est d'un mois et demi, est réalisé au sein de la société Hexastack. Son objectif principal est l'analyse Comparative des Algorithmes de Regroupement de Phrases ou de Textes Non Supervisés et des Techniques de Prétraitement NLP pour Comparer l'Exactitude entre les Différents Modèles.

1.2.2 Description du problème :

Plusieurs méthodes de classification de texte existent, et le choix entre elles dépend à la fois du résultat souhaité et de l'ensemble de données disponible. Une approche pour classer un ensemble de documents en tant que positifs ou négatifs ou en tant que plusieurs clusters consiste à étiqueter un petit ensemble de documents d'entraînement sélectionnés au hasard. Les méthodes d'apprentissage automatique supervisées peuvent ensuite s'entraîner sur l'ensemble étiqueté et utiliser ces informations pour étiqueter les données non étiquetées restantes comme positives ou négatives. Cependant, ce processus est long car les documents doivent être étiquetés manuellement et comporte également le risque d'un étiquetage biaisé. Trouver des regroupements naturels au sein de l'ensemble de données, sans étiquetage explicite par un humain, simplifierait le processus.

Cette thèse tente d'imiter les catégories qui ont été étiquetées manuellement pour un ensemble de documents textuels en regroupant les documents. Nous souhaitions savoir comment une approche non supervisée se comporte dans l'étiquetage de documents textuels. Lors de la différenciation des clusters, il y a un risque que les clusters diffèrent en fonction du choix des représentations de données. Cela a motivé l'examen et la validation de différentes représentations de données afin d'étudier l'effet des représentations sur le résultat du regroupement. Pour évaluer les résultats, nous nous sommes appuyés sur une partie des données étiquetées pour la validation croisée.

1.2.3 Portée et Objectifs :

Le rapport a pour objectif d'effectuer une analyse comparative des algorithmes de regroupement de phrases ou de textes non supervisés, ainsi que des techniques de prétraitement NLP, afin de comparer leur exactitude et leurs performances pour différentes représentations de données. L'accent sera mis sur l'utilisation de méthodes d'apprentissage non supervisé pour regrouper des ensembles de données textuelles sans étiquettes préalables.

Les objectifs spécifiques du rapport sont les suivants :

1. Explorer et comparer plusieurs algorithmes de regroupement de phrases ou de textes non supervisés, tels que K-means, le regroupement hiérarchique : Agglomertive, DBSCAN et Gaussien Mixture.

2. Mettre en œuvre différentes techniques de prétraitement NLP, notamment la tokenisation, la suppression des mots vides, et la vectorisation TF-IDF, pour améliorer les performances de regroupement.

3. Évaluer quantitativement et qualitativement l'exactitude de chaque modèle de regroupement en utilisant des mesures appropriées, telles que le score de silhouette, la cohésion et la séparation, afin de déterminer leur efficacité dans la formation de groupes cohérents.

4. Analyser l'impact des différentes représentations de données et des méthodes de prétraitement sur les résultats de regroupement pour identifier les configurations optimales pour chaque algorithme.

5. Discuter des implications pratiques et des applications potentielles du modèle de regroupement le plus performant, démontrant ainsi son potentiel pour fournir des informations significatives dans des scénarios réels de traitement de texte non supervisé.

L'ensemble de données utilisé dans cette étude sera spécifique au domaine du traitement de texte non supervisé, et les conclusions tirées ne seront pas généralisées à d'autres domaines. Cependant, les résultats devraient servir de base pour des recherches futures et une meilleure compréhension des méthodes de regroupement de textes non supervisées et des techniques de prétraitement NLP pour des applications pratiques.

1.2.4 Aperçu du plan du rapport :

Dans cette section, nous présenterons un aperçu de la structure globale du rapport.

Nous commencerons par introduire le contexte de l'étude et les objectifs de la recherche dans la section "**Introduction**".

Ensuite, nous examinerons l'état de l'art en matière de regroupement de textes non supervisés et de techniques de prétraitement NLP dans la section "**État de l'art**".

La méthodologie utilisée pour cette étude sera détaillée dans la section "**Méthodologie**", où nous présenterons l'ensemble de données utilisé, les techniques de prétraitement NLP appliquées et les algorithmes de regroupement étudiés.

Les résultats expérimentaux de l'analyse comparative des algorithmes de regroupement et des techniques de

prétraitement seront présentés dans la section "**Résultats expérimentaux**". Nous discuterons des performances de chaque modèle de regroupement, de l'impact des différentes représentations de données et des conclusions tirées de l'analyse qualitative des clusters obtenus.

Dans la section "**Discussion**", nous interpréterons les résultats, identifierons les limites de l'étude et proposerons des pistes d'amélioration. Nous examinerons également les applications potentielles des modèles de regroupement dans des scénarios réels de traitement de texte non supervisé.

Enfin, nous résumerons les principales conclusions de l'étude dans la section **"Conclusion Générale"** et mettrons en évidence les contributions de cette recherche. Nous discuterons également des perspectives pour des recherches futures dans le domaine de l'analyse comparative des algorithmes de regroupement de textes non supervisés et des techniques de prétraitement NLP.

Chapitre 2:
État de l'art

2.1 Analyse de texte avec des méthodes d'apprentissage automatique (ML) :

La capacité d'analyser des textes est un aspect important dans de nombreux domaines différents. Une grande partie des informations pertinentes pour les entreprises est présentée sous forme non structurée, principalement sous forme de textes. Des exemples courants incluent les rapports, les e-mails, les discussions en ligne, les tweets, les mises à jour sur les réseaux sociaux ou tout autre document contenant principalement du texte en langage naturel.

En général, l'idée de l'analyse de texte est de transformer des données non structurées en leur imposant une structure pour faciliter l'extraction d'informations. Cela peut être réalisé avec différentes approches. Les méthodes d'apprentissage automatique (ML) s'efforcent d'automatiser la transformation des données en données structurées et de trouver des motifs dans l'ensemble de données sans programmation explicite.

L'application de l'apprentissage automatique pour la détection de sujets dans un document s'est avérée être une réussite et un avantage computationnel par rapport aux processus manuels. Les algorithmes de ML peuvent généralement être divisés en deux sous-domaines différents : supervisés et non supervisés.

La principale différence réside dans la formulation du problème.

Pour une approche supervisée, les étiquettes ou les catégories sont connues dès le départ et sont utilisées pour générer une hypothèse lorsqu'un point de données inconnu est introduit. Les algorithmes non supervisés, tels que les algorithmes de regroupement, n'utilisent aucune étiquette préalable pour différencier les différents regroupements naturels dans un ensemble de données. Un exemple d'application de la méthode non supervisée est la détection de sujets au sein d'un corpus de documents.

2.2 Traitement automatique du langage naturel (NLP) :

Lorsqu'on travaille avec des méthodes d'apprentissage automatique pour le regroupement de texte, un aspect important à prendre en compte est la façon dont les données doivent être traitées et représentées. Le domaine du traitement automatique du langage naturel (NLP) tourne autour de cette interaction entre les ordinateurs et les langues humaines.

Dans ce domaine, plusieurs problèmes intéressants peuvent être retracés jusqu'en 1950, lorsque Alan Turing a publié l'article "Computational Machinery and Intelligence" (Turing, 1950). Il a proposé ce qui est maintenant appelé le "test de Turing". Ce n'était pas la première formulation liée au NLP, mais c'était un point de rupture majeur dans le domaine. Depuis lors, la recherche autour du NLP s'est principalement basée sur l'apprentissage automatique, en particulier depuis la "révolution statistique" à la fin des années 1980.

L'objectif principal du NLP est de trouver des structures et des modèles pour qu'un ordinateur puisse comprendre, générer et gérer un langage naturel. Lors de la représentation d'un texte, une approche couramment utilisée consiste à représenter les mots sous

forme de symboles discrets et à vectoriser le document dans le but de conserver les informations les plus importantes. Chaque mot est remplacé par un poids et chaque document peut ensuite être quantifié sous forme de vecteur. Si la corrélation sémantique entre les mots est négligée, la méthode qui classe chaque mot comme unique pose problème lors de la classification contextuelle. En incluant des relations de sens, l'étiquetage des parties du discours et en combinant des séquences de mots plutôt qu'une analyse mot par mot, les informations fournies peuvent être améliorées car davantage d'aspects du sens émotionnel du langage sont pris en compte.

2.3 La sélection de caractéristiques :

Dans les problèmes de sélection de caractéristiques, l'objectif est de trouver les caractéristiques les plus pertinentes qui maximisent l'information sur l'entrée. Il a été démontré qu'il existe une dépendance entre la précision de la classification et les dimensions des données. Le taux de précision peut diminuer de manière significative si un nombre redondant de caractéristiques est utilisé, ce qui rend la réduction des caractéristiques une partie cruciale d'un problème de regroupement.

Les avantages de la sélection de caractéristiques ont été démontrés en comparant le taux de précision et l'utilisation de la puissance de traitement des données. Les comparaisons entre différentes méthodes de sélection de caractéristiques, ne peuvent être effectuées que par des tests sur un ensemble de données spécifique. Il montre qu'un algorithme donné peut se comporter différemment sur différents ensembles de données, ce qui rend plus difficile la prédiction d'un modèle optimal.

En particulier, la précision de la classification et le nombre de

caractéristiques réduites sont utilisés comme mesures comparatives entre deux modèles de sélection de caractéristiques différents.

2.4 Exploration des méthodes de regroupement de textes non supervisés :

Les méthodes de regroupement de textes non supervisés sont des approches essentielles pour analyser et organiser de grands ensembles de données textuelles sans étiquettes préalables. Parmi ces méthodes, on trouve le célèbre algorithme K-means, qui vise à partitionner les données en k clusters en minimisant la variance intra-cluster. Le regroupement hiérarchique est une autre approche, qui construit une hiérarchie de clusters en fusionnant ou divisant les groupes successivement. DBSCAN est une méthode basée sur la densité des données, qui peut identifier les groupes de forme arbitraire dans les données.

2.5 Techniques de prétraitement NLP pour la préparation des données :

Le prétraitement NLP joue un rôle crucial dans le nettoyage et la préparation des données textuelles avant l'application des méthodes de regroupement. Les techniques de prétraitement courantes comprennent la tokenisation, qui divise le texte en mots ou en phrases, la suppression des mots vides pour éliminer les mots fréquents mais peu informatifs, la racinisation pour ramener les mots à leur forme de base, et la vectorisation TF-IDF qui convertit les mots en vecteurs numériques pondérés en fonction de leur fréquence d'occurrence dans les documents.

2.6 Travaux antérieurs liés à l'analyse comparative des modèles de regroupement :

Plusieurs recherches antérieures ont abordé l'analyse comparative des modèles de regroupement de textes non supervisés et des techniques de prétraitement NLP. Certains travaux se sont concentrés sur la comparaison des performances de différents algorithmes de regroupement sur des ensembles de données spécifiques, tandis que d'autres ont exploré l'effet des différentes techniques de prétraitement sur les résultats de regroupement. Certaines études ont également proposé des approches hybrides combinant plusieurs méthodes de regroupement et de prétraitement pour améliorer la qualité du regroupement.

Chapitre 3:
Méthodologie

3.1 Description de l'ensemble de données utilisé :

Pour cette étude, nous avons travaillé avec deux ensembles de données distincts, chacun présentant une structure spécifique et des informations essentielles.

Le premier ensemble de données est de type **bi-classe**, composé de deux clusters distincts. Le Cluster 1 : rassemble des phrases exprimant des sentiments positifs ou neutres concernant divers sujets. Les réponses dans ce cluster sont généralement favorables ou neutres, reflétant des opinions positives ou des points de vue équilibrés. En revanche, le Cluster 2 : englobe des expériences et des sujets plus négatifs ou complexes. Les réponses présentes dans ce cluster sont souvent liées à des expériences difficiles ou à des sujets délicats, reflétant des points de vue plus critiques ou des situations problématiques.

	id	label	tweet
0	1	0	@user when a father is dysfunctional and is s...
1	2	0	@user @user thanks for #lyft credit i can't us...
2	3	0	bihday your majesty
3	4	0	#model i love u take with u all the time in ...
4	5	0	factsguide: society now #motivation

Figure 3: Extrait de 5 premiers lignes de la dataset1

Le deuxième ensemble de données est de type **multi-classe** et se compose de phrases provenant de quatre sujets distincts : mathématiques, sciences, biologie et chimie. Chacun de ces sujets représente une classe spécifique dans l'ensemble de données. Cette structure multi-classe permet d'explorer comment les algorithmes de regroupement différencient et organisent ces sujets variés et spécialisés.

eng	Subject
An anti-forest measure is\nA. Afforestation\nB...	Biology
Among the following organic acids, the acid pr...	Chemistry
If the area of two similar triangles are equal...	Maths
In recent year, there has been a growing\nconc...	Biology
Which of the following statement\nregarding tr...	Physics

Figure 4: Extrait de 5 premiers lignes de la dataset2

Dans chacun de ces ensembles de données, les informations sont organisées en deux colonnes. La première colonne contient les phrases (corpus) elles-mêmes, tandis que la deuxième colonne attribue à chaque phrase un numéro de cluster auquel elle appartient. Cette étiquette de cluster fournit des informations fondamentales pour évaluer les performances des algorithmes de regroupement et pour comparer les regroupements prédits aux regroupements réels.

3.2 Techniques de prétraitement NLP appliquées aux données textuelles :

Après avoir nettoyé les ensembles de données en éliminant les caractères non alphabétiques, notamment la ponctuation, et en réduisant les espaces excessifs, ainsi qu'en supprimant les URL, les noms d'utilisateur, les hashtags et autres éléments

indésirables des textes, nous avons appliqué une série de méthodes de prétraitement distinctes.

label	tweet	clean tweet
0	@user when a father is dysfunctional and is s...	when a father is dysfunctional and is so selfi...
0	@user @user thanks for #lyft credit i can't us...	thanks for credit i cant use cause they dont o...
0	bihday your majesty	bihday your majesty
0	#model i love u take with u all the time in ...	i love u take with u all the time in ur
0	factsguide: society now #motivation	factsguide society now

Figure 5: Extrait de 5 premiers lignes de la dataset1 avant et après le prétraitement NLP appliqué

Pour chaque ensemble de données, nous avons exploré quatre approches de prétraitement différentes :

Sentences-Transformer : Cette méthode nous a permis d'encoder un corpus de phrases ou de textes en embeddings correspondants en utilisant un modèle pré-entraîné. Cela a permis de capturer des informations sémantiques riches pour chaque texte.

Tokenizer : Nous avons utilisé un tokenizer pour diviser les textes en mots individuels, créant ainsi une représentation structurée des données textuelles. Pour ce faire, nous avons créé et ajusté un tokenizer à l'aide de la bibliothèque Keras. Le tokenizer a été ajusté sur le corpus en utilisant une limite de 2000 mots, puis il a été utilisé pour convertir les séquences de mots en séquences numériques. Enfin, nous avons utilisé la méthode de padding pour uniformiser la longueur des séquences, les ajustant à une longueur maximale de 1000.

TF-IDF (Term Frequency-Inverse Document Frequency) : La vectorisation TF-IDF a été appliquée pour représenter les mots sous forme de vecteurs numériques, en mettant l'accent sur l'importance relative des mots dans chaque texte par rapport à l'ensemble du corpus.

GloVe (Global Vectors for Word Representation) : En utilisant les embeddings pré-entraînés GloVe, nous avons attribué des vecteurs de similarité à chaque mot en fonction de son utilisation dans le langage naturel.

Chaque méthode a été appliquée séparément à chaque ensemble de données, générant ainsi quatre ensembles de données traités distincts. Ces ensembles de données ont ensuite été utilisés comme entrées pour les algorithmes de regroupement. L'objectif était de comparer les résultats de clustering obtenus avec chaque méthode de prétraitement et d'évaluer leur impact sur la qualité des clusters formés. Cette approche comparative a permis d'identifier quelle méthode de prétraitement fonctionne le mieux pour chaque algorithme de regroupement et chaque ensemble de données spécifique.

3.3 Présentation des algorithmes de regroupement étudiés (K-means, regroupement hiérarchique, DBSCAN, Gaussian Mixture Model - GMM):

Nous avons examiné en détail quatre algorithmes de regroupement non supervisés dans le cadre de cette étude. Chacun de ces algorithmes présente des caractéristiques uniques pour la formation de clusters à partir des données textuelles.

1. K-means : L'algorithme K-means vise à partitionner les données en k clusters en minimisant la variance intra-cluster. Dans notre étude, nous avons utilisé l'implémentation K-means standard pour regrouper les données textuelles. Il commence par initialiser k centres de cluster, puis il attribue chaque point de données au centre de cluster le plus proche, suivant une alternance entre la mise à jour des centres de cluster et la réattribution des points de données jusqu'à la convergence.

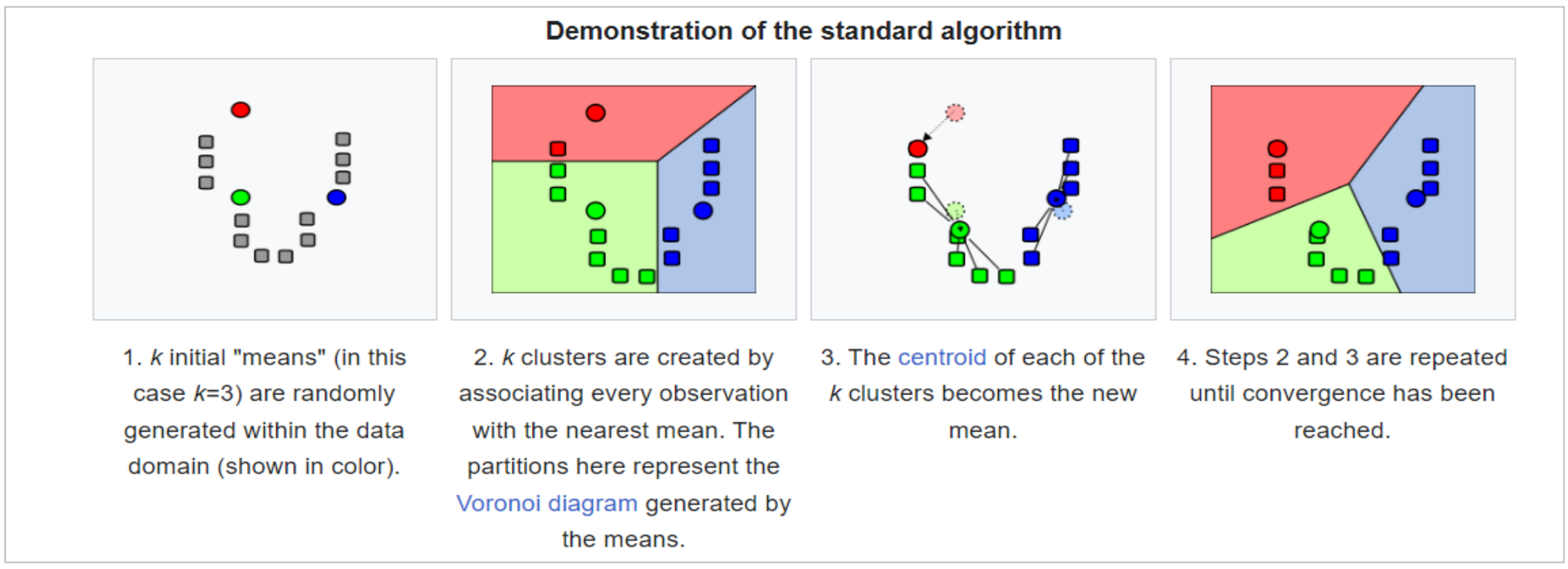

Figure 6: La démarche de l' algorithme K-Means

2. Regroupement hiérarchique : L'algorithme de regroupement hiérarchique construit une hiérarchie de clusters en fusionnant ou en divisant les groupes successivement. Nous avons utilisé la méthode d'agglomération, qui commence par traiter chaque point de données comme un cluster individuel, puis fusionne les clusters les plus proches jusqu'à ce qu'un seul cluster global soit formé.

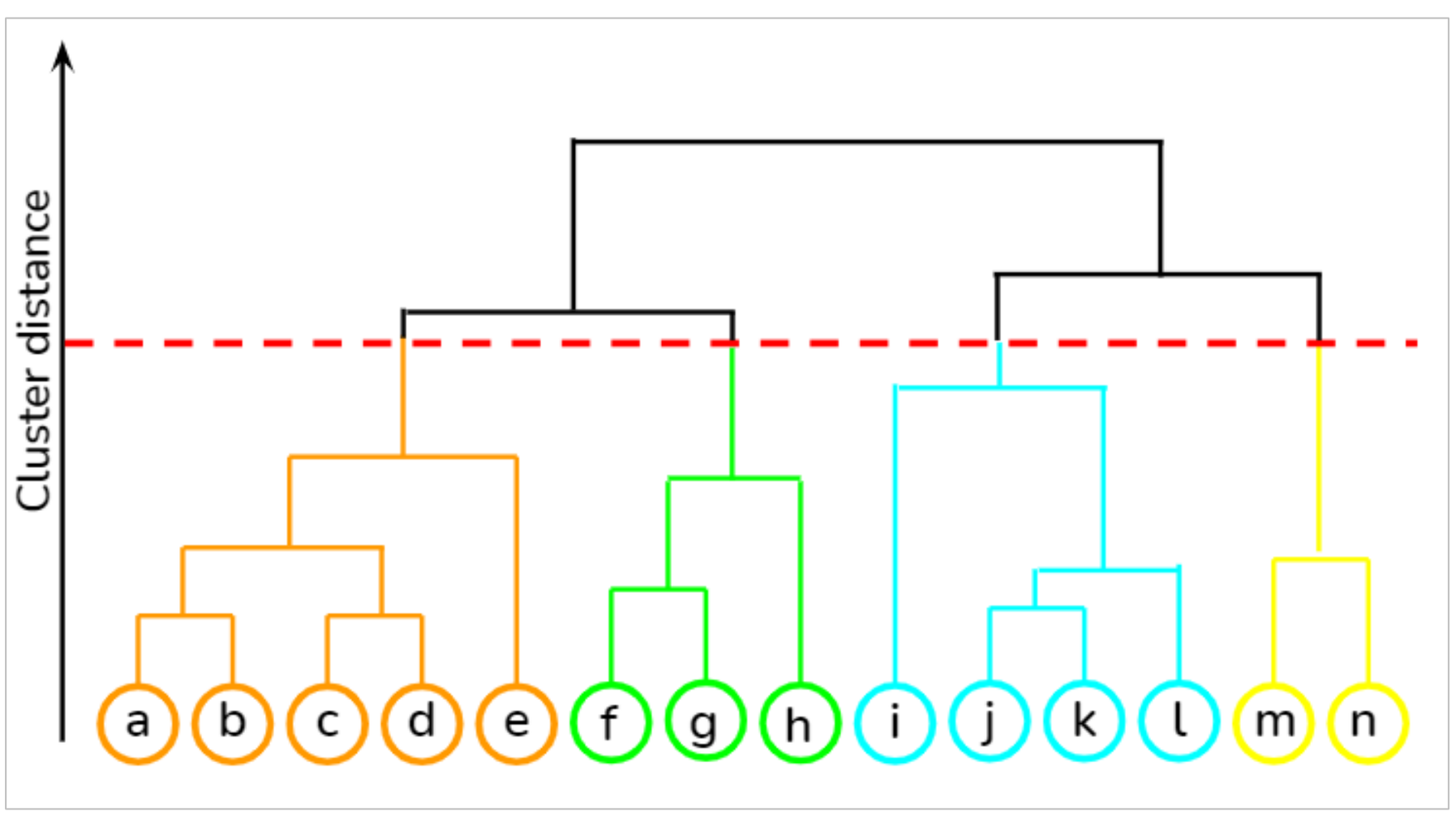

Figure 7: La démarche de l' algorithme hiérarchique

3. DBSCAN (Density-Based Spatial Clustering of Applications with Noise) : L'algorithme DBSCAN repose sur la densité des données pour identifier des groupes de forme arbitraire. Il est capable de détecter des clusters de différentes formes et de gérer les points de données considérés comme du bruit. Les points de données sont regroupés en

fonction de leur proximité et de leur densité.

4. Gaussian Mixture Model (GMM) : Le modèle de mélange gaussien (GMM) est un algorithme de regroupement qui suppose que les données dans chaque cluster sont générées à partir d'une distribution gaussienne. Chaque cluster est modélisé comme une combinaison de plusieurs distributions gaussiennes. GMM est capable de modéliser des clusters de différentes formes et d'ajuster des distributions plus complexes que K-means.

Chaque algorithme a été appliqué individuellement à chaque ensemble de données traité à l'aide des méthodes de prétraitement mentionnées précédemment. Ces algorithmes de regroupement ont été choisis pour leur popularité et leurs capacités à former des clusters significatifs à partir de données textuelles. La comparaison des résultats de ces algorithmes nous a permis de déterminer lequel d'entre eux offre les performances les plus cohérentes et informatives pour chaque ensemble de données spécifique.

3.4 Évaluation des performances à l'aide de métriques appropriées :

Pour évaluer les performances des algorithmes de regroupement pour chaque ensemble de données et méthode de prétraitement, nous avons utilisé une variété de métriques appropriées. Ces métriques nous ont permis de quantifier la qualité et la cohérence des clusters formés. Nous avons utilisé les métriques suivantes :

Accuracy : Cette métrique mesure la proportion de points de données correctement attribués à leur cluster d'origine. Elle nous

permet de quantifier la justesse des clusters formés par rapport aux étiquettes de cluster réelles.

F-mesure : La F-mesure est une métrique qui combine la précision et le rappel en un seul score. Elle permet de prendre en compte à la fois les faux positifs et les faux négatifs lors de l'évaluation des clusters.

Score de silhouette : Le score de silhouette évalue à quel point les points d'un cluster sont similaires entre eux et distincts des autres clusters. Un score de silhouette élevé indique que les clusters sont bien définis et distincts.

Cohésion et séparation des clusters : La cohésion mesure à quel point les points d'un même cluster sont proches les uns des autres, tandis que la séparation mesure à quel point les points de différents clusters sont éloignés les uns des autres. Ces métriques permettent d'évaluer la compacité et la séparation des clusters formés.

Calinski-Harabasz Score : Cette métrique évalue la cohérence des clusters en calculant le rapport entre la variance inter-cluster et la variance intra-cluster. Un score Calinski-Harabasz plus élevé indique des clusters plus compacts et distincts.

Davies-Bouldin Score : Le score Davies-Bouldin mesure la similarité moyenne entre chaque cluster et son cluster voisin le plus proche. Un score plus faible indique des clusters plus cohérents et séparés.

En utilisant ces métriques, nous avons pu évaluer la performance de chaque algorithme de regroupement en fonction de différentes méthodes de prétraitement et ensembles de données. Cela nous a permis de déterminer quelles combinaisons produisent les clusters les plus cohérents, informatifs et conformes aux étiquettes de cluster réelles.

Chapitre 4:
Résultats expérimentaux

Introduction :

Cette section présente en détail les résultats obtenus à partir de nos expérimentations exhaustives sur les deux ensembles de données, à savoir la dataset multi-classes et la dataset bi-classe.

4.1 Comparaison des performances des différents algorithmes de regroupement :

L'évaluation des performances des différents algorithmes de regroupement dans les deux ensembles de données, à savoir la dataset multi-classes et la dataset bi-classe, a fourni des aperçus intrigants sur l'efficacité de chaque approche.

Pour la **dataset multi-classes**, nous avons soumis les textes à une variété d'algorithmes de regroupement, tels que KMeans, Agglomerative Clustering, et Gaussian Mixture, en utilisant différentes techniques de prétraitement. Les résultats ont démontré des variations significatives des performances en fonction de l'algorithme utilisé. Par exemple, en utilisant la méthode Sentence Transformers en conjonction avec KMeans, nous avons atteint un taux de précision remarquable de **0.745**. En revanche, les méthodes traditionnelles telles que TF-IDF, Tokenizer et GloVe ont généré des taux de précision nettement plus bas, oscillant entre **0.35 et 0.38**. Parmi les métriques de qualité du regroupement, le score de silhouette a également révélé que la méthode Sentence Transformers était la plus cohérente en termes de structure des clusters, tandis que les autres méthodes ont montré une dispersion plus prononcée.

La **dataset bi-classe** a produit des résultats similaires. Par exemple, l'approche Sentence Transformers associée à l'algorithme Agglomerative Clustering a généré des performances exceptionnelles, atteignant une précision de **0.893**. En revanche, l'approche Tokenizer avec KMeans a montré des performances relativement plus faibles, avec une précision de **0.576**.

Ces résultats confirment que l'algorithme de regroupement choisi joue un rôle critique dans la formation de clusters significatifs, et que certaines méthodes, telles que Sentence Transformers, se distinguent par des performances plus élevées. Cependant, il est important de noter que les performances peuvent varier en fonction de la nature des données et du contexte d'application.

4.2 Impact des techniques de prétraitement sur les résultats de regroupement :

Une étape cruciale dans l'analyse de texte non supervisée est le prétraitement des données, qui vise à nettoyer et à préparer les textes bruts avant l'application des algorithmes de regroupement. Nous avons minutieusement évalué comment différentes techniques de prétraitement affectent les performances des algorithmes de regroupement, en particulier en utilisant les méthodes Sentence Transformers, TF-IDF, Tokenizer et GloVe.

Nos observations montrent que les techniques de prétraitement jouent un rôle majeur dans la qualité des clusters formés. Par exemple, dans la dataset multi-classes, nous avons constaté que les scores de précision obtenus avec chaque méthode de prétraitement étaient légèrement différents. Cependant, quel que soit le prétraitement, la méthode Sentence Transformers est restée la plus performante, avec une précision élevée de **0.745**. En ce qui concerne la dataset bi-classe, nous avons rencontré

des résultats similaires. Même si les techniques de prétraitement ont légèrement modifié les scores de précision, l'approche Sentence Transformers en conjonction avec Agglomerative est restée la plus performante, avec une précision de **0.893**. Cette constance dans les performances de Sentence Transformers suggère que cette méthode est robuste et adaptable, indépendamment des variations dans le prétraitement.

Ces résultats soulignent l'importance d'une préparation minutieuse des données textuelles. Il est important de noter que les techniques de prétraitement peuvent avoir un impact plus significatif sur d'autres types de données ou domaines d'application.

4.3 Analyse qualitative des clusters obtenus :

L'analyse qualitative des clusters formés a permis de générer des informations instructives sur la nature et la pertinence des regroupements obtenus dans les deux ensembles de données.

Pour **la dataset multi-classes**, l'évaluation des clusters formés a été réalisée en utilisant un ensemble de métriques, dont le score de silhouette, la cohésion et la séparation des clusters. En considérant ces mesures de qualité du regroupement dans leur ensemble, nous pouvons noter que les clusters présentaient des variations en termes de densité et de séparation. Alors que certaines configurations de clusters ont montré une bonne cohésion interne, d'autres ont révélé des chevauchements et des frontières moins claires.
Par exemple, la méthode Sentence Transformers en conjonction avec l'algorithme KMeans a obtenu un score de silhouette de **0.0269**, indiquant une séparation relativement adéquate des clusters. Cependant, d'autres approches, telles que la méthode GloVe avec l'algorithme Agglomerative, ont enregistré un score de silhouette de **0.1029**, indiquant une meilleure séparation. Cette observation souligne la complexité inhérente à la formation de clusters dans un contexte où les

textes peuvent être polyvalents et aborder plusieurs sujets en même temps.

Dans le contexte de **la dataset bi-classes**, une évaluation similaire a été effectuée. Les résultats ont montré une distinction plus nette entre les deux classes en termes de sentiment positif/neutre et de sentiment négatif ou expériences difficiles. Par exemple, l'approche Sentence Transformers avec l'algorithme Agglomerative Clustering a obtenu un score de silhouette de **0.0279**, indiquant une séparation relativement bonne entre les clusters. Cependant, l'approche Sentences Transformers avec l'algorithme KMeans a montré un score de silhouette de **0.0238**, indiquant une séparation légèrement moins nette. Cependant, il est important de noter que même si les clusters semblaient distincts, leur densité et leur cohésion n'étaient pas uniformes dans l'ensemble.

En résumé, l'analyse qualitative des clusters a révélé des tendances et des distinctions dans la formation de clusters, en tenant compte de la variabilité des métriques de qualité du regroupement. Cette approche nuancée offre un aperçu plus complet des performances des algorithmes de regroupement et souligne la nécessité de considérer une gamme de métriques pour évaluer la pertinence et la qualité des regroupements obtenus.

4.4 Identification des configurations optimales pour chaque modèle :

Suite à une évaluation approfondie, il est recommandé d'utiliser la méthode Sentence Transformers avec l'algorithme Agglomerative Clustering pour la dataset multi-classes et KMeans pour la dataset bi-classe. Ces configurations ont démontré une cohérence et une pertinence élevées dans les résultats de regroupement.

Chapitre 5:
Discussion

Introduction :

La discussion explore en profondeur les implications des résultats obtenus. Nous avons observé que les performances varient en fonction des caractéristiques de l'ensemble de données et de la méthode de regroupement. Cela soulève des questions sur l'adaptabilité de ces méthodes à différents types de données.

5.1 Interprétation des résultats et des conclusions :

Nos conclusions soulignent l'efficacité de la méthode Sentence Transformers, en particulier lorsqu'elle est combinée avec Agglomerative Clustering et KMeans. Ces résultats indiquent que ces méthodes peuvent être appliquées avec succès à des problèmes de regroupement de textes non supervisés.

5.2 Limitations de l'étude et pistes d'amélioration :

Nous reconnaissons que nos résultats peuvent être influencés par les caractéristiques spécifiques des ensembles de données que nous avons utilisés. De plus, il existe d'autres algorithmes de regroupement et de prétraitement qui pourraient être explorés à l'avenir pour améliorer davantage les performances.

5.3 Applications potentielles dans des scénarios réels de traitement de texte non supervisé :

Nos résultats ont des implications pratiques, notamment pour l'amélioration des chatbots. Les approches performantes que nous avons identifiées, telles que Sentence Transformers avec KMeans, pourraient être exploitées pour :

- Analyse de sentiments améliorée: Les chatbots pourraient mieux comprendre les émotions des utilisateurs et fournir des réponses empathiques.
- Réponses contextuelles: Les chatbots pourraient adapter leurs réponses en fonction du contenu des questions.
- Détection de sujets: Les chatbots pourraient identifier les thèmes discutés et fournir des informations pertinentes.
- Tri automatique des requêtes: Les chatbots pourraient classer automatiquement les questions des utilisateurs en catégories.

En intégrant ces applications, les chatbots deviendraient des assistants plus intelligents et plus adaptés aux besoins des utilisateurs.

Conclusion Générale

Tout au long de mon stage en « HexaStack », j'ai eu l'opportunité de me plonger dans le domaine de l'IA et de l'NLP spécifiquement des algorithmes de regroupement non supervisés et du prétraitement des données pour l'amélioration d'un chatbot. Mon travail a consisté en la conception et la mise en place de méthodes visant à améliorer les performances de regroupement des données textuelles pour optimiser les réponses du chatbot.

Le choix d'utiliser des méthodes de regroupement non supervisés, telles que KMeans, Agglomerative Clustering et Gaussian Mixture, s'est avéré être une approche prometteuse pour organiser les données non étiquetées en clusters pertinents. Les techniques de prétraitement, notamment Sentence Transformers, TF-IDF, Tokenizer et GloVe, ont joué un rôle crucial dans la préparation des données en amont, ce qui a impacté directement les performances des algorithmes de regroupement. Ainsi, à travers la mise en œuvre de ces méthodes et techniques, j'ai pu observer des résultats variés.

L'analyse qualitative des clusters formés a révélé des regroupements cohérents et pertinents, reflétant les thèmes sous-jacents des données.

L'ensemble de ce stage m'a permis de mettre en pratique mes connaissances théoriques et techniques acquises au cours de mon parcours universitaire. Il a également renforcé mes compétences en organisation, en planification et en travail d'équipe puisque j'ai pu partager mes résultats pour une meilleure compréhension des améliorations potentielles.

En conclusion, ce stage a été une expérience enrichissante qui m'a permis d'explorer en profondeur les techniques de regroupement non supervisés et de prétraitement des données dans le contexte spécifique de l'amélioration d'un chatbot. Les résultats obtenus ouvrent la voie à de futures améliorations et développements dans le domaine des interfaces conversationnelles intelligentes. Je suis reconnaissant pour cette opportunité et j'espère que mes contributions pourront apporter une valeur ajoutée à l'équipe et au projet dans son ensemble.