

**MINISTERE ALGERIENNE DE L'ENSEIGNEMENT SUPERIEUR DE LA
RECHERCHE SCIENTIFIQUE ET DE LA TECHNOLOGIE**

Université Saad Dahleb Blida

Faculté des Sciences - Département d'Informatique



MEMOIRE DE MASTER

PRESENTE EN VUE D'OBTENIR LE DIPLOME DE MASTER EN INFORMATIQUE

OPTION : INGENIERIE DE LOGICIEL

Thème

CLUSTERING AN ENTERPRISE SOCIAL NETWORKS

Réalisé par :

**HAMMOUCHE Aicha
HAOUCHE Baraa**

Encadré par :

**M^{me} EL ALLIA Nadia (ESI)
M^{me} BOUTOUMI Bachira**

Jury :

**Présidente : M^{me} ARKAM
Examinatrice : M^{me} ZAHRA Fatima Zahra**

Promotion : 2019/2020

Remerciements

On tient à remercier toutes les personnes qui ont contribué de près ou de loin au succès de notre projet, qui nous ont aidés lors de sa réalisation et lors de la rédaction de ce travail. Tout d'abord, on adresse nos remerciements à nos chers parents qui sans eux rien de cela ne serait possible. Ils nous ont soutenu tout au long de notre parcours et nous ont offert un environnement de travail idéal.

Ensuite, on tient à remercier vivement nos encadreurs **M^{me} BOUTOUMI Bachira** et **M^{me} EL ALLIA Nadia (ESI)**, qui nous a beaucoup aidées dans nos recherches, pour le temps passé ensemble et pour le partage de son expertise.

Enfin, on tient à remercier toutes les personnes qui nous ont soutenu, aidé et ont contribué même de loin à la création de notre projet.

Dédicace

Je dédie ce modeste travail à :

À MES CHERS PARENTS

Aucune dédicace ne saurait exprimer mon respect, mon amour éternel

Et ma considération pour les sacrifices que vous avez consenti pour mon

Instruction et mon bien être.

Je vous remercie pour tout le soutien et l'amour que vous me portez depuis mon enfance et
j'espère que votre bénédiction m'accompagne toujours.

À mon cher frère mon deuxième père « Zaki » qui m'a soutenu toute ma vie

À mes très chers frères et mes belles sœurs

À toute ma famille : mes grands-parents, mes tantes, mes oncles et mes cousins et cousines

À tous mes amis et mes collègues

Enfin, Toute personne qui m'aime et que j'aime ...

Baraa

Je dédie ce modeste travail accompagné

d'un profond amour :

à celle qui ma arrosé de tendresse et d'espoirs,

à la source d'amour incessible

à ma bénie par ces prières ... ma mère

à mon support dans ma vie,

qui ma appris supporté et ma dirigé

vers la gloire.....mon père

à mes chères frères et sœurs

à toutes les personnes de ma grande famille

à tous les amis de ma vie

Aicha

Résumé

A l'heure actuelle, les organisations développent et mettent en place des environnements collaboratifs de travail pouvant donc être les vecteurs des informations, des savoirs, des connaissances portées par des collaborateurs. Ils peuvent ainsi générer une activité plus ciblée, plus réactive et donc plus performante, avec des prises de décisions efficaces et pertinentes au regard des objectifs que l'organisation s'est donnée. En effet, la mise en œuvre de tels dispositifs dans les organisations de travail permet, à l'instar du Knowledge Management, d'élaborer peu à peu une véritable intelligence collective.

Notre objectif à travers ce projet est d'analyser les réseaux sociaux d'entreprise considérés comme des plateformes collaboratives pour rehausser l'intelligence collective au sein d'une organisation. Pour ce faire, on utilise des algorithmes de classification non supervisée plus précisément le Clustering. Les algorithmes de clustering utilisés sont le K-Means et le clustering hiérarchique agglomérative. Les résultats obtenus (clusters) appliqués sur le dataset du Workplace de l'école nationale supérieure d'informatique serviront, peut-être, à une analyse et un enrichissement des critères de l'intelligence collective.

Mots-clés : Réseau social d'entreprise, K-Means, Clustering hiérarchique Agglomérative, Workplace

Abstract

Today, organizations are developing and implementing collaborative work environments that can be the vectors of information, knowledge and knowledge carried by employees. They can thus generate a more targeted, more reactive and therefore more efficient activity, with efficient and relevant decision-making with regard to the objectives that the organization has set itself. Indeed, the implementation of such systems in work organizations allows, like Knowledge Management, to gradually develop a real collective intelligence.

Our objective through this project is to analyze corporate social networks as collaborative platforms to enhance collective intelligence within an organization. To do so, we use unsupervised classification algorithms more precisely clustering. The clustering algorithms used are K-means and hierarchical agglomerative clustering. The results obtained (clusters) applied to the Workplace dataset of the National School of Computer Science will perhaps be used to analyze and enrich the criteria of collective intelligence.

Keywords: Corporate social network, K-means, Clustering hierarchical Agglomerative, Workplace

ملخص

في الوقت الحالي، تعمل المؤسسات على تطوير وإعداد بيانات عمل تعاونية يمكن أن تكون بالتالي ناقلات للمعلومات والمعرفة التي يحملها الموظفون. وبالتالي يمكنهم إنشاء نشاط أكثر استهدافاً وأكثر استجابة وبالتالي أكثر كفاءة، مع اتخاذ قرارات فعالة وذات صلة فيما يتعلق بالأهداف التي حددتها المنظمة لنفسها. في الواقع، يتيح تطبيق مثل هذه الأجهزة في مؤسسات العمل، مثل إدارة المعرفة، تطوير ذكاء جماعي حقيقي تدريجياً.

هدفنا من خلال هذا المشروع هو تحليل الشبكات الاجتماعية للشركات التي تعتبر منصات تعاونية لتعزيز الذكاء الجماعي داخل المنظمة. للقيام بذلك، نستخدم خوارزميات تصنيف غير خاضعة للرقابة، وبشكل أكثر دقة التجميع. الخوارزميات العنقودية المستخدمة هي تجميع K-Means والتجميع التكتلي الهرمي. النتائج التي تم الحصول عليها (المجموعات) المطبقة على مجموعة البيانات الخاصة بمكان العمل بالمدرسة الوطنية العليا للإعلام الآلي تستخدم، ربما، لتحليل وإثراء معايير الذكاء الجماعي.

الكلمات المفتاحية: الشبكة الاجتماعية للشركات، K-الوسائل، التجميع الهرمي التكتلي، مكان العمل

Table des matières

Remerciements.....	1
Dédicace.....	2
Résumé.....	3
Abstract.....	4
ملخص.....	5
Liste des Figures	8
Liste des Tableaux	9
Liste d'abréviations.....	10
Introduction Générale	12
PARTIE I – ETAT DE L'ART	15
Chapitre I : Réseaux Sociaux d'Entreprise	15
I.1. Introduction	15
I.2. Définition.....	15
I.3. Avantages	15
I.4. Objectifs.....	16
I.5. Types des RSE.....	16
I.6. Synthèse sur les types des RSE	18
I.6.1. RSE internes	18
I.6.2. RSE externes.....	19
I.7. Etude de cas « Workplace ESI »	21
I.8. Conclusion.....	25
Chapitre II : Apprentissage non supervisé – Le clustering.....	26
II.1. Introduction	26
II.2. Définition.....	26
II.3. Catégories (classes)	26
II.3.1. Clustering par partitionnement	26
II.3.2. Clustering basé sur la densité	27
II.3.3. Clustering hiérarchique	27
II.4. Algorithmes	28
II.4.1. Algorithme K_Means	28
II.4.2. Algorithme Agglomerative.....	29

II.4.3. Algorithme DBSCAN	30
II.5. Mesures de similarité.....	31
II.6. Synthèse des algorithmes.....	33
II.7. Conclusion.....	34
Chapitre III : Classification non supervisée dans les réseaux sociaux d'entreprise	35
III.1. Introduction	35
III.2. Travail 1 : Regroupement des personnes dans un réseau social basé sur la similitude textuelle.....	35
III.3. Travail 2 : Classification des postes sur des réseaux sociaux.....	36
III.4. Travail 3 : La détection des communautés dans les réseaux sociaux :.....	37
III.5. Travail 4 : Regroupement des audiences sociales dans les réseaux d'information d'entreprise :	38
III.6. Conclusion.....	39
PARTIE II – CONTRIBUTION.....	41
Chapitre IV: Clustering Workplace ESI.....	41
IV.1. Introduction :	41
IV.2. Présentation de dataset extraite du workplace ESI :.....	41
IV.3. Algorithmes de clustering :.....	42
IV.3.1. Algorithme K-Means	42
IV.3.2. Algorithme Hierarchical Agglomerative clustering.....	47
IV.3.3. Comparaison entre K_Means et le clustering hiérarchique	49
IV.3.4. Clustering spatial basé sur la densité des applications avec bruit (DBSCAN)	
49	
IV.4. Interprétation des résultats :.....	51
IV.5. Outils de développement :	51
• Configuration matérielle	51
• Configuration logicielle	52
IV.6. Conclusion :.....	54
Conclusion Générale et Perspectives	55
Références Bibliographiques:	56

Liste des Figures

Figure 1 : Nombre total de personnes de Workplace ESI.....	21
Figure 2 : Nombre total de profils complétés de Workplace ESI	22
Figure 3 : Répartition par types de groupes dans Workplace ESI	23
Figure 4 : Représentation 1 de l’algorithme agglomérative [13]	29
Figure 5 : Représentation 2 de l’algorithme agglomérative [13]	30
Figure 6 : Les différents types de liens[19]	32
Figure 7 : Représentation matricielle de dataset Workplace ESI.....	42
Figure 8 :Meilleur K dans K_Means de Workplace ESI	43
Figure 9 : Nombre optimal de clusters dans K_Means de Workplace ESI.....	44
Figure 10 : Représentation graphique de l’analyse de silhouette de K-Means de Workplace ESI	44
Figure 11 : Score de l’analyse de silhouette de K_Means de Workplace ESI.....	45
Figure 12 : Visualisation des données en 3D (K_Means).....	46
Figure 13 : Visualisation des données en 3 D (K_Means).....	46
Figure 14 : Représentation du dendrogramme de l’algorithme de clustering hiérarchique agglomérative clustering de Workplace ESI.....	48
Figure 15 : Représentation graphique des résultats de l’algorithme de clustering hiérarchique agglomérative de Workplace ESI.....	48
Figure 16 : la distance utiliser et le nombre du cluster.....	50
Figure 17 : visualisation des clusters en 2D	50
Figure 18 : Visualisation des clusters en 3D	50
Figure 19 : Représentation graphique d'un cluster (publication, commentaire, réaction, membres actifs par date)	51

Liste des Tableaux

Tableau I : Synthèse sur les RSE internes	18
Tableau II : Synthèse sur les RSE externes	19
Tableau III : Algorithme K_Means	28
Tableau IV : Algorithme DBSCAN	31
Tableau V : Synthèse algorithmes du clustering	33
Tableau VI : Résultats d'analyse utilisant l'algorithme KNN et la variation des techniques de représentation (K=1 et distance cosinus).....	Erreur ! Signet non défini.
Tableau VII : Résultats d'analyse utilisant l'algorithme naïve bayes et la variation des techniques de représentation	Erreur ! Signet non défini.
Tableau VIII : Résultats d'analyse utilisant l'algorithme c4.5 et la variation des techniques de représentation.....	Erreur ! Signet non défini.
Tableau IX : Comparaisons entre les algorithmes d'apprentissage supervisés classiques et les algorithmes bio-inspirés	Erreur ! Signet non défini.
Tableau X : Avantages et Inconvénients des algorithmes [14]	Erreur ! Signet non défini.

Liste d'abréviations

RSE : Réseau social d'entreprise

BSN: Business social networking

ESN: Enterprise social network

ESI: Ecole nationale Supérieure d'Informatique

ERP: Enterprise Resource Planning

CRM: Customer Relationship Management

DBSCAN: Density-Based Spatial Clustering of Applications with Noise

OPTICS: Ordering Points To Identify the Clustering Structure

EBIRI : Environnement Bio-Inspiré pour la Recherche d'Information

Introduction Générale

Un réseau social d'entreprise RSE (Business Social Networking BSN ou Enterprise Social Network ESN) est un outil collaboratif permettant aux membres d'une entreprise de centraliser l'information sur une même plateforme et de mieux communiquer. Les utilisateurs peuvent notamment y travailler plus efficacement en groupe et mener des projets communs plus facilement afin de gagner du temps. En d'autres termes, les RSE sont des plateformes technologiques permettant aux employés d'une organisation de travailler ensemble tout en profitant des fonctionnalités du social média [1]. A l'heure actuelle, d'après [2], les organisations développent et mettent en place des environnements collaboratifs de travail pouvant donc être les vecteurs des informations, des savoirs, des connaissances portées par des collaborateurs. Ils peuvent ainsi générer une activité plus ciblée, plus réactive et donc plus performante, avec des prises de décisions efficaces et pertinentes au regard des objectifs que l'organisation s'est donnée. En effet, la mise en œuvre de tels dispositifs dans les organisations de travail permet, à l'instar du Knowledge Management, d'élaborer peu à peu une véritable intelligence collective. On désire élaborer le corpus d'un RSE choisi (par exemple Workplace de Facebook) et à partir des algorithmes de la classification non supervisée (ceux du Clustering), on génère des classes de ce corpus conçu. Ces classes obtenues serviront pour une analyse et un enrichissement des critères de mesure de l'intelligence collective [3].

Notre problématique est d'analyser comment les réseaux sociaux d'entreprise peuvent rehausser l'intelligence collective au sein d'une entreprise. L'objectif donc de ce travail est d'extraire des connaissances à partir du corpus réalisé du réseau social d'entreprise afin de les classifier.

Il s'agit dans ce travail de recherche de :

- faire un état de l'art sur les concepts :
 - les réseaux sociaux d'entreprise (définitions, les plateformes les plus répandues, les critères d'évaluation,...)
 - le clustering (définition, les algorithmes,...)
 - la classification des réseaux sociaux d'entreprises
- L'utilisation d'un corpus du RSE (par exemple workplace de Facebook)
- classification (en utilisant au moins deux algorithmes) des connaissances obtenues du RSE.

Le présent document est divisé en deux grandes parties. La première partie concerne l'état de l'art alors que la seconde présente nos contributions. Chacune de ces parties contient des chapitres. L'organisation des parties et les points abordés dans les chapitres sont présentés ci-dessous.

Partie I : Etat de l'art

Dans cette première partie, nous commencerons par présenter les réseaux sociaux et plus particulièrement les réseaux sociaux d'entreprise. Ensuite, nous introduirons les notions d'apprentissage non supervisés à savoir le clustering. Puis, nous présenterons les travaux sur la classification des réseaux sociaux d'entreprises.

Chapitre 1 : Réseaux sociaux : Ce premier chapitre concerne une étude sur les réseaux sociaux à savoir leur définition, ... puis nous présenterons en détails les réseaux sociaux d'entreprise et spécifiquement le workplace d'ESI.

Chapitre 2 : Apprentissage non supervisé – Clustering : Dans ce deuxième chapitre, nous allons introduire les notions et les algorithmes d'apprentissage non supervisé.

Chapitre 3 : Classification non supervisée dans les réseaux sociaux d'entreprise : Dans ce chapitre, nous abordons les travaux de recherche qui traitent la classification des réseaux sociaux.

Partie II : Contribution

Dans cette partie, nous présenterons la solution proposée à la problématique soulevée. Nous nous intéresserons au Dataset du réseau RSE workplace de l'Ecole nationale Supérieure d'Informatique (ESI) dont nous allons présenter, puis on classifie d'après les algorithmes du clustering ce Dataset pour générer un ensemble de classe que nous analyserons.

Chapitre 4 : Clustering Workplace ESI : Dans ce chapitre, nous allons aborder la classification à partir des algorithmes vus dans le chapitre 2 tout en exploitant le jeu de donnée de workplace ESI.

Nous terminerons ce mémoire par une conclusion générale et des perspectives

PARTIE I

ETAT DE L'ART

PARTIE I – ETAT DE L'ART

Chapitre I : Réseaux Sociaux d'Entreprise

I.1. Introduction

Les réseaux sociaux ont envahi notre quotidien depuis déjà plusieurs années et il n'y pas un jour sans que l'on n'entende parler de l'un d'entre eux (Facebook, Twitter, LinkedIn, Viadeo, etc.). Après la sphère personnelle, ils s'immiscent désormais dans le monde de l'entreprise : après l'engouement pour les ERP (Enterprise Resource Planning) dans les années 90 et les CRM (Customer Relationship Management) dans les années 2000, voici venu le temps des réseaux sociaux d'entreprise, souvent appelés RSE.[4]

I.2. Définition

Le réseau social d'entreprise est une plateforme numérique propre à une entreprise donnée. Il permet la mise en relation des employés pour échanger de l'information, des savoirs et des pratiques. L'échange peut être professionnel ou extra professionnel, même si le RSE est conçu avant tout comme un outil de travail. L'origine du réseau social d'entreprise est à attribuer au réseau social grand public de type Facebook[5].

I.3. Avantages

Parmi les avantages des réseaux sociaux d'entreprise, on peut citer [6]:

- **Donner plus de sens au travail** : une vision élargie de l'entreprise permet à ses salariés de mieux se positionner et ainsi mieux cerner les enjeux stratégiques de l'entité toute entière. Toute tâche prend alors un sens.
- **Gommer la verticalité avec un management transversal** : le manager ne se limite plus à donner des ordres à ses collaborateurs, mais plutôt à les accompagner, les responsabiliser. On glisse vers un mode de management plus participatif.
- **Stimuler l'intelligence collective** : la facilité du partage de l'information et des connaissances permet une véritable émulation collective.
- **Favoriser l'agilité et l'innovation** : via des ruches à idées dans lesquelles des collaborateurs de différents services parfois géographiquement très éloignés participent entre eux. Des idées et des solutions innovantes et novatrices émergent alors.
- **Booster la culture d'entreprise et la cohésion** : transparence de la communication et rapprochement entre les services ou les différentes entités d'une même entreprise engendrent un sentiment d'appartenance et renforcent la cohésion.

- **Rendre les salariés plus autonomes et plus impliqués dans leur entreprise** : en fluidifiant la circulation de l'information notamment, les collaborateurs ont une meilleure compréhension de l'entreprise dans sa globalité. Il leur est ainsi plus aisé de prendre des initiatives.

I.4. Objectifs

Les objectifs d'un RSE sont multiples, on peut citer[7] :

- Créer et développer une intelligence collective partagée, renforcer la participation des salariés et archiver les connaissances.
- Valoriser dans l'entreprise les compétences, les individus et l'estime de soi et découvrir des experts et de nouveaux talents.
- Créer des communautés de pratiques et des réseaux d'entraide et de support au sein de l'entreprise.
- Favoriser la phase d'intégration de nouveaux collaborateurs.
- Créer des liens entre employés éparpillés dans le monde entier.

I.5. Types des RSE

On retrouve deux types de réseaux sociaux d'entreprise en termes de localisation des données et de positionnement des utilisateurs : les RSE internes et les RSE externes.

I.5.1. Les RSE internes

Voici quelques exemples des RSE internes[8] :

I.5.1.1. Yammer

Possédé par Microsoft, Yammer est un RSE dont l'interface n'est pas sans rappeler Facebook. Yammer vous permet d'améliorer la transmission des informations au sein de votre organisation, de réaliser vos tâches plus vite grâce à une meilleure communication (groupes, messagerie...). La suite Office est également intégrée dans la plateforme, permettant de visualiser et modifier simplement ses documents.

I.5.1.2. Sharepoint

SharePoint 2013[9] a amélioré très nettement ses fonctionnalités, intégrant notamment un fil de microblogging très complet, avec possibilité de mentionner et de tagger à la volée. L'outil facilite à l'utilisateur la construction de son réseau et la personnalisation de son fil d'activités, via des suggestions plus présentes et plus pertinentes. Enfin, la partie sociale de SharePoint a été pensée pour l'usage en mobilité, grâce à la sortie d'une application mobile,

fait savoir Guillaume Guérin, Analyste chez Lecko, cabinet de conseil en organisation et nouvelles technologies.

I.5.2. Les RSE externes

Parmi les RSE externes, on peut citer :

I.5.2.1. Workplace by Facebook

Facebook, le géant des médias sociaux, possède sa propre plateforme dédiée aux entreprises appelée Workplace. Avec son interface similaire à celle de Facebook, Workplace a l'avantage d'être familier et accessible. Il permet de créer des groupes pour organiser vos différents projets en équipes réduites, de diffuser des vidéos live en un clin d'œil pour vous adresser à toute votre équipe, ou encore de travailler avec d'autres entreprises au sein d'un groupe. En bref, un outil très complet pour communiquer et travailler en équipe. Il pourrait devenir leader du marché par sa qualité et la pertinence de sa cible.

I.5.2.2. TalkSpirit

Une pépite française du RSE, TalkSpirit est une plateforme collaborative ergonomique et pratique pour rassembler tous vos contenus et partager les connaissances. C'est une version française de Slack optimisée grâce à une nouvelle interface graphique. Elle facilite aussi la diffusion de l'information grâce à des conversations structurées et organisées.

I.6. Synthèse sur les types des RSE

I.6.1. RSE internes

Tableau I : Synthèse sur les RSE internes

Nom	Localisation des données	Gestion Documentaire	Sécurité des données	Contenus	Développement	Recherche	Evolutivité
Yammer	Disponible dans le Cloud	Pas d'archivage et d'extraction. Pas de notification par mail sur l'ajout de document	Tous les contenus sont visibles dès lors que je suis membre à l'intérieur d'un groupe privé, un utilisateur peut démarrer une discussion avec une personne qui n'est pas membre de la communauté	Les documents ou les communautés supprimées ne peuvent pas être restaurées	Difficile	Moins de customisations	Des mises à jour développées régulièrement par Microsoft
SharePoint	Conserver les données sur une infrastructure hébergée dans les locaux	Synchroniser, via OneDrive for Business, des documents dans les communautés avec un répertoire sur le poste local de l'utilisateur	Seuls les membres de la communauté peuvent consulter les données d'une communauté privée	Restauration Dans corbeille	Facile	Offre plus de possibilité de customisations d'affiner les résultats de recherche indexée des contenus en dehors du RSE	Intégrer des outils tiers, ce qui permet de garder la main sur la fonctionnalité que l'on souhaite apporter au RSE.

I.6.2. RSE externes

Tableau II : Synthèse sur les RSE externes

Nom	Localisation des données	Gestion documentaire	Sécurité des données	Intégration	Recherche	Interface	Fonctionnalité
Workplace	Disponible sur Cloud, le stockage illimité des fichiers et des appels vidéo acceptant jusqu'à 50 collaborateurs	Facebook embarque un chat puissant similaire à Slack qui permet de communiquer de manière instantanée avec ses collaborateurs en one to one ou en groupes	Chaque compte d'entreprise de Workplace est isolé sur une architecture hébergée	Autorise des intégrations de logiciels. Le nombre d'intégrations de services tiers est plus réduit	L'outil de recherche moins performant (une recherche par conversation seulement).	Il s'administre à la manière de G Suite de Google : une interface simple et intuitive qui permet d'ajouter/supprimer des collaborateurs facilement et d'activer/désactiver des fonctionnalités.	Plateforme de travail (workplace) complète, avec toutes les fonctionnalités requises pour travailler à plusieurs et pour communiquer en interne (top-down, bottom-up et transverse), ouverte sur l'écosystème digital (intégrations).

TalkSpirit	Le stockage des données sur le continent européen de stockage (illimité),	Ils peuvent centraliser tous les contenus issus d'applications tierces, qui sont intégrés grâce à des connecteurs. Tous les contenus sont aisément retrouvés avec la fonction de recherche avancée, et certaines tâches sont automatisables grâce aux bots intelligents de TalkSpirit.	haut niveau de sécurité	L'intégration des logiciels de l'entreprise. Plus de 1000 intégrations (natives et via Zapier) avec des apps tierces : Trello, Pipedrive, Box, Dropbox, Chargebee, Twitter, Facebook, ...Github.	Performante	Sa plateforme de blogs nommée BlogSpirit, le haut niveau de sécurité.	Flux d'activité, pages communautés et individus, partage et fonctions conversationnelles
-------------------	---	--	-------------------------	--	-------------	---	--

I.7. Etude de cas « Workplace ESI »

L'Ecole nationale Supérieure d'Informatique d'Alger (ESI) a mis en place le réseau social Workplace de Facebook pour améliorer la collaboration entre ses différents membres : enseignants, étudiants et personnel. On y retrouve l'ensemble des fonctionnalités habituelles de Facebook (profils, groupes, fil d'actualité, messagerie instantanée, événements), mais focalisées cette fois sous un angle professionnel.

La figure suivante donne un aperçu sur le nombre total de personnes de la plateforme.

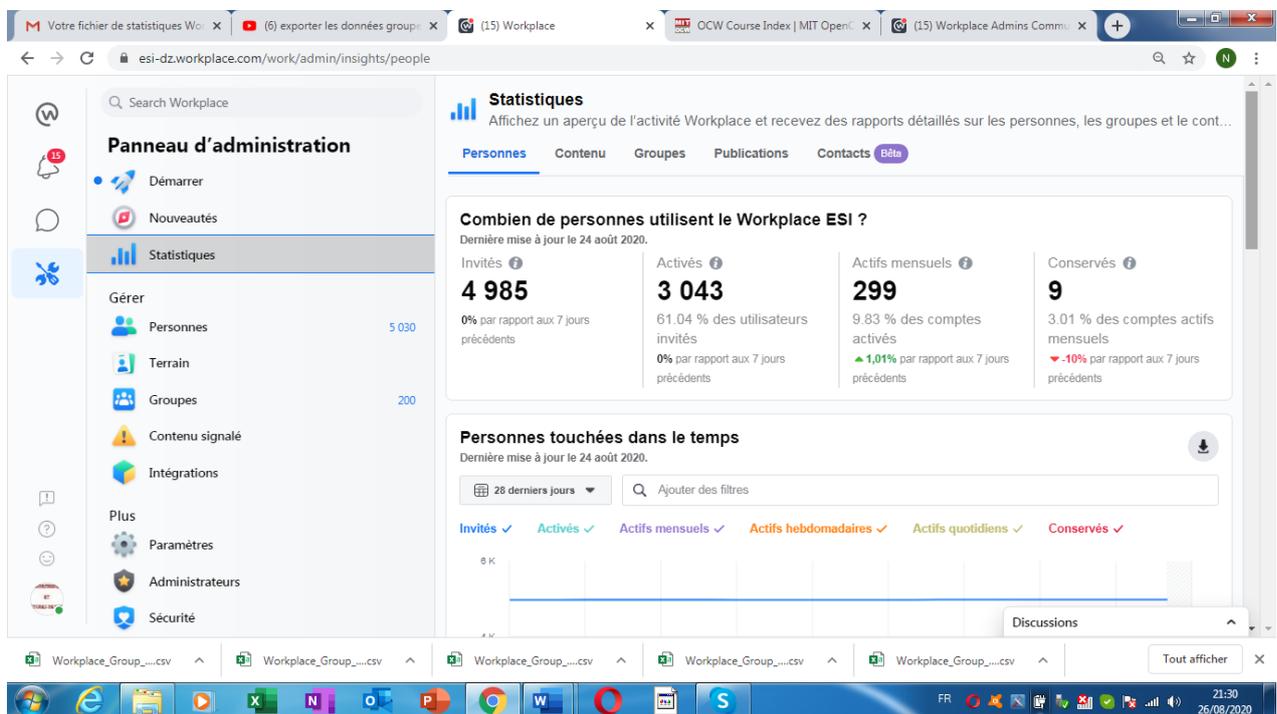


Figure 1 : Nombre total de personnes de Workplace ESI

Pour accéder à la plateforme, il suffit d'avoir un compte de messagerie propre à l'école à savoir prénom_nom@esi.dz (exemple n_benahmed@esi.dz). Chaque utilisateur doit remplir son profil afin d'avoir une base de données assez riches en information.

Malgré le nombre de personnes s'élevant à 5030 invités dont 3043 ont activé leurs comptes, Workplace ESI n'est pas vraiment exploités par la communauté de l'école. Malheureusement que 2% des utilisateurs ont complété leurs profils (voir figure ci-dessous).

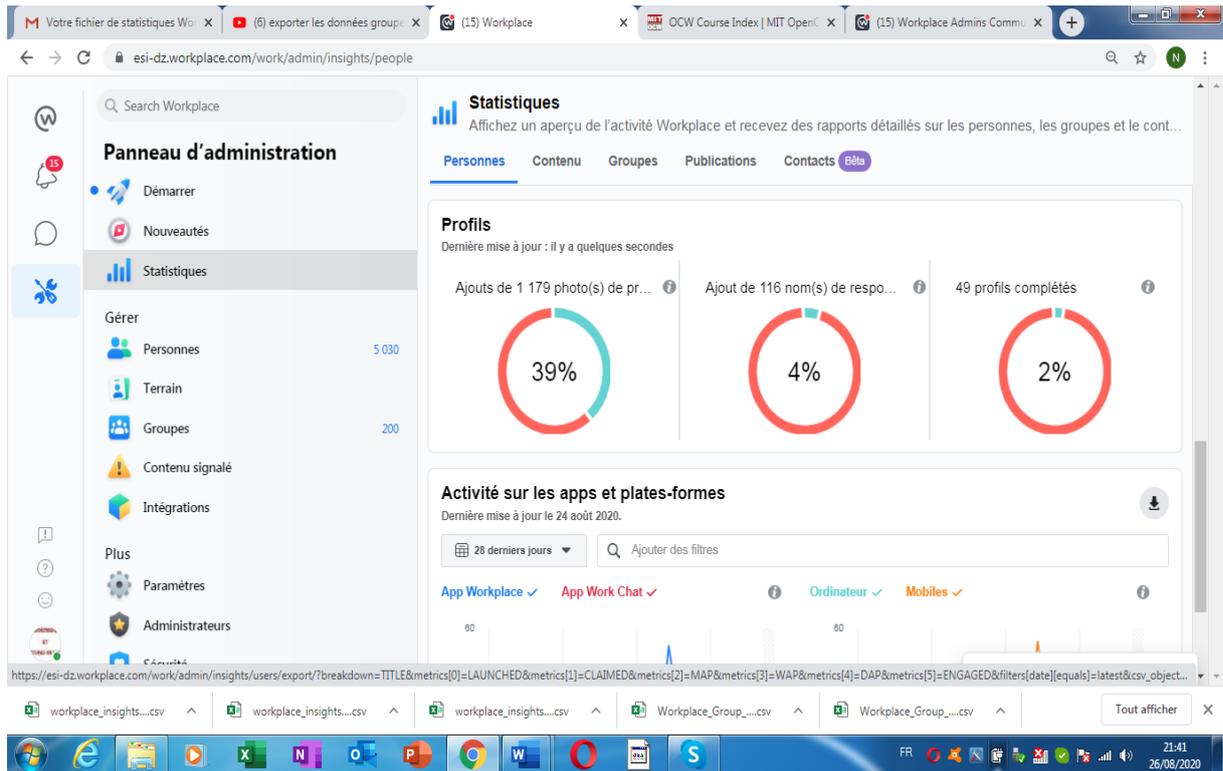


Figure 2 : Nombre total de profils complétés de Workplace ESI

En dehors de l'aspect le nombre d'utilisateurs, un second aspect assez important à prendre en dans le cadre de notre projet est la notion de groupes. Sachant que notre objectif est d'analyser et d'enrichir les critères de mesure de l'intelligence collective, la notion d'interaction dans des groupes est très importante. La figure suivante montre que la plateforme Workplace ESI compte plus de 200 groupes dont 52% est de types « Equipes et projets ».

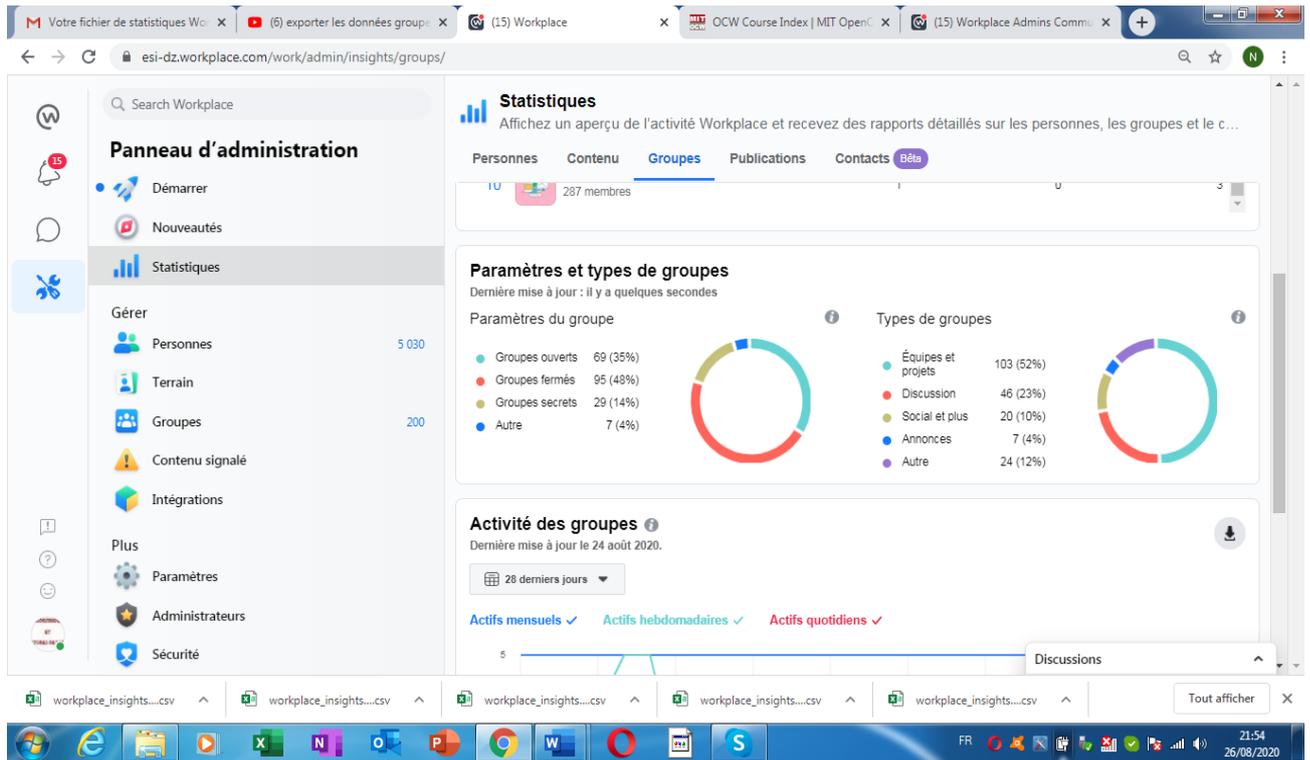


Figure 3 : Répartition par types de groupes dans Workplace ESI

Le fait d'avoir un taux de participation et d'utilisation assez faible, une étude d'audit a été effectuée par des étudiants de 2^{ème} année du cycle supérieur durant le mois d'avril 2020. Dans cette étude, un certain nombre de constatations a été soulevé :

- **Mauvaise communication et promotion insuffisante de la plateforme**
 - L'Utilisation du RSE pour le partage de thématique ou commenter des sujets d'intérêt.
 - Une promotion annuelle (début de l'année) existe mais pas suffisante pour inciter les utilisateurs à utiliser plus souvent le RSE. Des efforts ont été mis dans le sens de faciliter l'accès (à travers le SSO), des liens vers la documentation, formation des employés (module conduite de changement), et finalement l'incitation des employés à utiliser le RSE.
 - Le but était de mettre à la disposition des acteurs de l'ESI tous les outils nécessaires qu'aurait une école de classe mondiale.
 - La culture de RSE est inconnue pour une grande partie usagers.
- **Non transparence sur les prérogatives de l'administrateur**
 - Peur de l'exposition car le nom est public, peut-être dû à la peur de la critique et l'absence de la culture d'accepter la critique, chose qui peut être n'a pas été traité lors de la conduite du changement.
 - Un contrôle sur le RSE peut décourager les initiateurs.

- Un peu moins de la moitié des utilisateurs n'utilisent pas ou moins Workplace car ils se sentent "surveillé" et limité dans leur liberté, ils ne savent pas ce qu'ils peuvent dire ou poster et où le faire.
- **Barrière Enseignant-employé-étudiant**
 - Les employés trouvent que c'est "une autre chose à faire de plus au boulot", et évitent de s'impliquer plus dans le réseau si ce n'est récolter des informations.
 - Différence dans le langage utilisé des différents acteurs que ce soit le français, anglais, arabe, et encore "Darja", ce qui crée de silo de personne utilisant un certain langage et minimise certaines interactions.
 - La mise en place du projet n'avait pas pour but de briser la barrière enseignant-étudiant, mais plutôt l'incitation à l'utilisation d'outil propre à l'école pour toute autre communication. Sachant que l'utilisation principale qui avait été imaginé pour le RSE était de remplacer Messenger (de Facebook) dans les communications inter-école.
- **Objectifs inexistant ou pas assez mis en avant, conduite de changement lors de la mise en place du projet et implications de la plateforme dans la prise de décision**
 - Le projet s'aligne avec les objectifs stratégiques de l'école en renforçant l'esprit d'appartenance, cela s'accroît du fait que les alumnis y ont toujours accès, mais l'alignement n'est pas très clair.
 - Le projet n'a pas été accompagné par un changement dans la culture de l'école (partage, communication, barrière étudiant/enseignant, gestion de connaissance (alumni))
 - L'initiation du projet RSE ne répondait pas à un objectif précis, mais plutôt suivait un effet de mode, de plus, du fait que ce soit fait par Facebook, la transition des utilisateurs devaient se faire naturellement
 - L'objectif du Workplace de l'ESI n'a pas été communiqué de façon explicite aux utilisateurs.
 - Le RSE regroupe certes la totalité des acteurs de l'ESI, mais il n'intervient à aucun moment dans un processus de prise de décision à cause du faible/inexistant taux de participation de ces derniers.
 - Le besoin d'un RSE ne se ressentait pas.
- **Gestion de la plateforme et modération du contenu et des groupes**
 - A part les postes dans le groupe "Everybody at ESI" toute modération se fait à posteriori.
 - Toute modération est faite par une seule personne.
 - Il n'existe pas non plus de politique de gestion de contenu si ce n'est que les postes doivent être un tant soit peu professionnel.
 - Le comité éthique de l'école ne s'occupe pas de Workplace.
- **Contenu existant insuffisant**
 - Les usagers sont plus intéressés par des postes informatifs, tel que des renseignements (actualités) sur les activités de l'école, les activités des clubs de l'école, des informations sur le monde professionnel, études et de la recherche dans le domaine
 - Le manque d'implication de la communauté (club, étudiant, travailleur, doctorant, pour la diversité).

- Les commentaires peuvent faire resurfer un sujet d'intérêt si beaucoup de personne commentent sur un post le citant.
 - Il n'existe pas de groupe fait pour étudier et régler les problèmes de l'école impliquant ainsi une partie de l'administration, des enseignants et des étudiants.
 - L'un des résultats de ce confinement est qu'il y ait plus d'utilisateur maintenant et plus d'interactivité au sein de Workplace.
 - Le contenu de Workplace n'est assez intéressant ou pas assez suffisant pour susciter l'intérêt des utilisateurs, ce qui serait l'une des raisons de leurs absences.
 - Workplace n'est pas exploité comme il devrait l'être.
 - Les utilisateurs ne sont pas motivés d'utiliser un Workplace par catégorie de profil.
 - Les usagers sont moyennement satisfaits du contenu, dû à la quantité limitée, ou à des sujets qu'ils jugent inintéressant.
- **L'utilisation d'autres outils de par les usagers, concurrence et ergonomie remise en cause.**
 - Concurrence des autres outils de partage (Messagerie, Intranet enseignants, Facebook, LinkedIn, ...).
 - Habitude des différents acteurs au réseau Facebook.
 - Les principales raisons pour lesquelles Workplace a été choisi et pas un autre outil est la familiarité avec ce que tout le monde utilise déjà (Facebook), le SSO, et la possibilité d'importer les utilisateurs depuis g-suite.
 - Ouvrir le RSE aux ESI-partenaire n'est pas chose impossible mais éventualité a étudié.
 - L'utilisation de Workplace n'est pas aussi intuitive qu'il n'y paraît pour ceux qui ne sont pas déjà sur Facebook par exemple. De plus, les utilisateurs n'ont pas été formé au préalable pour, ni ne savent comment contacter l'administrateur en cas de problème.

I.8. Conclusion

Dans ce chapitre, nous avons abordé la définition, les avantages, les objectifs ainsi que les types des RSE. A partir de ces types des RSE, nous avons cité quelques exemples des RSE que par la suite, nous avons effectué une synthèse sur ces RSE. Cette synthèse a été faite sur la localisation des données, la gestion documentaire, la sécurité des données, l'intégration, la recherche, l'interface et les fonctionnalités.

Le prochain chapitre sera consacré sur la classification non supervisée et plus précisément sur le Clustering.

Chapitre II : Apprentissage non supervisé – Le clustering

II.1. Introduction

L'apprentissage automatique (Machine Learning) est un domaine de l'intelligence artificielle qui s'intéresse à l'étude et à la conception d'algorithmes qui extraient de la connaissance à partir de données. Cet apprentissage automatique peut être devisé en 4 types :

- apprentissage supervisé,
- apprentissage non-supervise,
- apprentissage semi-supervise
- et apprentissage par renforcement.

Dans notre projet, il nous a été demandé d'utiliser l'apprentissage non supervisé et plus précisément le « **Clustering** ». Son but est de regrouper en classes les éléments formant la base de données étudiée et ce sans connaissance au préalable des classes qu'elle contient.

II.2. Définition

Le clustering (la segmentation) ou le regroupement en classes homogènes consiste à réduire un nuage des points d'un espace quelconque en un ensemble de représentants moins nombreux permettant une représentation simplifiée des données initiales. Ainsi, comme les méthodes d'analyse factorielle, la classification automatique est une méthode de réduction des données. Il s'agit d'une démarche très courante qui permet de mieux comprendre l'ensemble analysé. C'est donc une méthode de réduction des données. Ces applications sont nombreuses en statistique, en traitement d'image, en intelligence artificielle, en reconnaissance des formes ou encore la compression de données [10].

II.3. Catégories (classes)

On peut distinguer deux grandes familles de clustering **par partitionnement** et **par hiérarchie** selon [10] et [11]. Les algorithmes de clustering peuvent être classés dans les catégories suivantes : Clustering par partitionnement, Clustering basé sur la densité et Clustering hiérarchique.

II.3.1. Clustering par partitionnement

Le regroupement par partitionnement est considéré comme la plus populaire catégorie de l'algorithme de clustering. Cet algorithme divise les points de données en « k » partitions, où chaque partition représente un cluster. La partition est basée sur une certaine fonction objective.

Les grappes sont formées de telle sorte que les objets de données d'une grappe sont « similaires », et les objets de données dans différents clusters sont « différents ». Les méthodes de regroupement par partitionnement sont utiles dans les applications où le nombre de grappes requises est statique. K_Means, PAM (Partition around médoids) et CLARA sont des algorithmes de partitionnement [11].

II.3.2. Clustering basé sur la densité

Les algorithmes de clustering basés sur la densité créent des clusters de forme arbitraire. Dans ce type d'approche de clustering, un cluster est considéré comme une région dans laquelle la densité des objets de données dépasse une valeur seuil particulière. L'algorithme DBSCAN est un exemple célèbre de l'approche de clustering basée sur la densité [11].

II.3.3. Clustering hiérarchique

Les algorithmes de clustering hiérarchique fonctionnent pour diviser ou fusionner un ensemble de données particulier en une séquence de partitions imbriquées. La hiérarchie de ces cloisons imbriquées peut être de deux types : agglomération (ascendante) et divisive (descendante).

- Dans la méthode d'agglomération, le regroupement commence par un seul objet de données dans un seul cluster et continue à regrouper les paires les plus proches de clusters jusqu'à ce que tous les objets de données soient regroupés en un seul cluster.
- Le regroupement hiérarchique divisive, par contre, commence avec tous les objets de données dans un seul cluster et continue de diviser les plus grands clusters en plus petits jusqu'à ce que tous les objets de données soient divisés en clusters unitaires.

BIRCH (Balance Iterative Reducing and Clustering using Hiérarchies), CURE (Cluster Using Representatives) sont des exemples d'approche de regroupement hiérarchique [11].

II.4. Algorithmes

II.4.1. Algorithme K_Means

La technique K_Means est un moyen d'organiser les éléments de données reposant sur certaines fonctionnalités du groupe K [12] ; où K est un entier positif. Le regroupement est réalisé en réduisant la quantité totale de carrés de distances entre le centre de gravité du cluster et les données.

K_Means est largement utilisé en raison de sa mise en œuvre facile et de sa complexité linéaire en $O(n)$. Il possède plusieurs variantes comme : C-means flou (Fuzzy c-means) et clustering de K_Means du noyau (kernel K_Means clustering) [13].

Le nombre de clusters est difficile à déterminer et le résultat du clustering est sensible à l'initialisation des centres (les nombres de clusters k).

Parmi les domaines d'application de cet algorithme :

- Segmentation d'images
- L'organisation d'images
- Les graphiques regroupements théoriques

Tableau III : Algorithme K_Means

Algorithme 1 : K_Means

Entrée :

$D = \{d_1, d_2, \dots, d_n\}$ // ensemble de n éléments de données.
K // Nombre de clusters souhaités

Sortie :

Un ensemble de k clusters.

Etapas :

1. Choisir arbitrairement k éléments de données de D comme centres de gravité initiaux ;
 2. Répéter
 - Attribuez chaque élément d_i au cluster qui a le centre de gravité le plus proche ;
 - Calculer la nouvelle moyenne pour chaque cluster
 Jusqu'à ce que les critères de convergence soient satisfaits
-

II.4.2. Algorithme Agglomerative

Les algorithmes de clustering hiérarchique tentent de générer une hiérarchie de clusters pour les données objets. En règle générale, les techniques de clustering hiérarchique se divisent en deux types [13] :

- **Clustering agglomératif** : chaque objet de données est un cluster feuille de la hiérarchie, et les paires de grappes sont fusionnées de manière itérative selon certaines mesures de similitude. Il utilise une combinaison de mesure de distance pour objet (Euclidienne /Manhattan...) et pour cluster (single-linkage /average /complete).
- **Clustering diviseur** : tous les objets de données commencent dans un cluster, et fractionnement de cluster est effectué de manière récursive selon certaines mesures de dissimilarité. Il utilise un algorithme de clustering (K_Means / arbre de décision /ART).

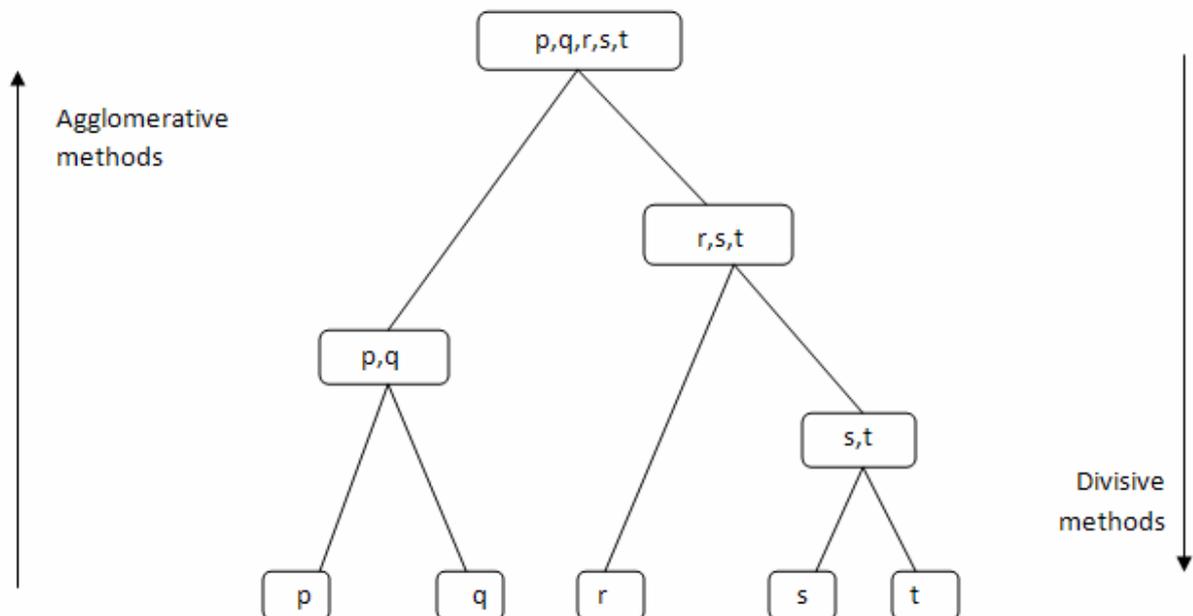


Figure 4 : Représentation 1 de l'algorithme agglomératif [13]

Trois problèmes majeurs demeurent :

- 1- Complexité temporelle élevée, généralement de $O(n^3)$, limite son évolutivité pour les grands ensembles de données.
- 2- La hiérarchie générée peut être très complexe pour un ensemble de données contenant divers contenus.
- 3- Il est difficile de décider des critères d'arrêt.

Le clustering hiérarchique a été largement utilisé dans les domaines de l'image et du texte.

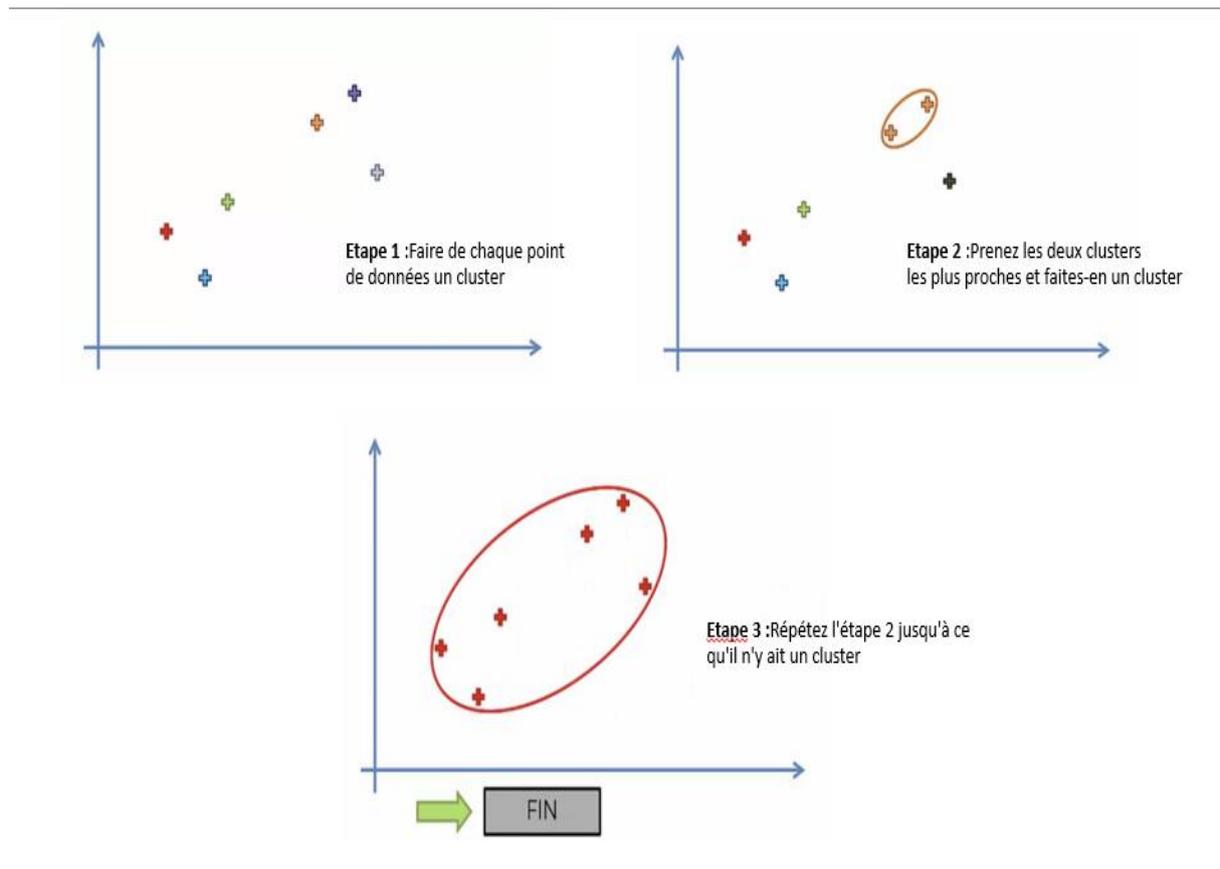


Figure 5 : Représentation 2 de l'algorithme agglomérative [13]

II.4.3. Algorithme DBSCAN

DBSCAN c'est le premier algorithme de clustering basé sur la densité. Il forme le degré de densité en utilisant deux paramètres [14], à savoir : la distance maximale pour la recherche de voisins ϵ et du nombre minimum de voisins minPts . Les objets de données et leurs voisins qui satisfont aux exigences ci-dessus sont appelés points de base et sont réputés à appartenir au même cluster.

La complexité de l'ensemble des algorithmes est $O(n^2)$, où n est le total des points de données. Si des arrangements actifs sont utilisés et la dimension des points est faible ($d \leq 5$), le calcul de la complexité de DBSCAN peut être réduite à $O(n \log)$ [15].

[12] ont proposé un algorithme de clustering basé sur la densité, KNNCLUST, dans lequel la densité est mesurée par une table de noyau KNN.

Tableau IV : Algorithme DBSCAN**Algorithme 3 : DBSCAN****Entrée :**

N objet à regrouper et paramètres globaux Eps, MinPts.

Sortie :

Groupes d'objets

Etapas :

- 1- Sélection arbitraire d'un point P
- 2- Récupérer tous les points de densité accessibles depuis P wrt Eps et MinPts
- 3- Si P est un point central, un cluster s'il est formé
- 4- Si P est un point frontière, aucun point n'est accessible en densité à partir de P et DBSCAN visite le point suivant de la base de données
- 5- Continuez le processus jusqu'à ce que tous les points aient été traités

Remarque :

Besoin de spécifier les paramètres globaux Eps, MinPts à l'avance de l'utilisateur

II.5. Mesures de similarité

Une mesure de similarité peut être définie comme la distance entre différents points de données. Bien que la similarité soit un montant qui reflète la force de la relation entre deux éléments de données, la dissimilarité traite de la mesure de la divergence entre deux éléments de données [16]. En fait, la performance de nombreux algorithmes dépend de la sélection d'une bonne fonction de distance sur l'ensemble de données d'entrée [16].

Une distance entre deux objets x, y est une application positive, symétrique, obéissant à l'inégalité triangulaire et telle que $d(x, y) = 0$ si et seulement si $x = y$

La distance entre deux objets x, y , définis chacun par ses attributs $x = (x_1, x_2, \dots, x_n), y = (y_1, y_2, \dots, y_n)$. Exemple de distance :

- Minkowski Distance $D_{ij} = \sum_{l=1}^d (|x_{il} - x_{jl}|^{1/n})^n$ [17]
- Euclidienne Distance $D_{ij} = \sum_{l=1}^d (|x_{il} - x_{jl}|^{1/2})^2$ [17]

La mesure de la distance utilisée entre deux classes sont : le lien simple (single-link), lien complet (complete-link) et le lien moyen (average-link) [18].

Dans le lien simple (resp complet, moyen), la distance entre deux classes est la valeur minimum (resp maximum, moyenne) des distances entre toutes les paires d'objets, l'un de la première classe, l'autre de la deuxième. Une illustration de chacun de ces types est présentée dans la figure 3.

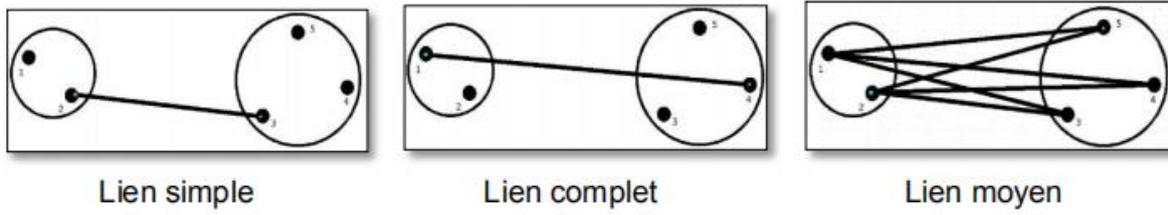


Figure 6 : Les différents types de liens[19] .

II.6. Synthèse des algorithmes

Tableau V : Synthèse algorithmes du clustering

Méthode	Paramètre	Evolutivité	Cas d'utilisation	Géométrie (distance)	Complexité
K_Means	Nombre de cluster	N-échantillons très grands N_clusters moyens	- Usage normale. - même taille de clusters. - géométrie plate - Pas trop de clusters.	Distance entres les points	$O(n)$
Agglomerative clustering (hierarchical clustering)	nombre de grappes ou seuil de distance, type de liaison, distance	Grands n_échantillons et N_clusters	- Nombreux clusters, - éventuellement contraintes de connectivité.	N'importe quelle distance par paire	$O(n^3)$
DBSCAN	taille du quartier (neighborhood size)	N-échantillons très grands, N_clusters moyens	- Géométrie non plate, - tailles de grappes inégales	Distances entre les points les plus proches	$O(n^2)$ $O(n \log n)$

II.7. Conclusion

Dans ce chapitre, nous avons d'une façon particulière sur le clustering qui fait partie de l'apprentissage non supervisé. Nous avons défini ces catégories et pour chaque catégorie nous avons présenté leurs domaines d'application. Sachant que chaque catégorie renferme une multitude d'algorithmes, nous avons préféré de représenter uniquement trois d'entre eux que nous avons jugé de les utiliser dans notre projet.

Le prochain chapitre sera consacré sur l'utilisation de la classification des réseaux sociaux d'entreprise dans la recherche scientifique.

Chapitre III : Classification non supervisée dans les réseaux sociaux d'entreprise

III.1. Introduction

Nous abordons dans ce chapitre les travaux de recherche qui traitent la classification des réseaux sociaux. En faisant une recherche dans ce contexte, nous avons trouvé des milliers d'articles. Il nous a été difficile de sélectionner les articles qui se rapprochent à notre contexte. Nous allons choisir les travaux de recherche que nous jugeons intéressants.

III.2. Travail 1 : Regroupement des personnes dans un réseau social basé sur la similitude textuelle.

[20] La similitude textuelle entre différentes personnes dans un réseau social est basée sur les mots utilisés par ces personnes. L'exploration de données est un processus clé qui utilise divers types de techniques pour découvrir des modèles ou des connaissances à partir de données (Han et Kamber, 2012). Le réseau social utilisé est twitter ils ont donc extrait les tweets des utilisateurs et effectué une analyse sur eux. Les tweets extraites sont très approximatifs, un pré-traitement est nécessaire. Le pré-traitement se fait en cinq étapes (Extraction de données, Suppression des mots d'arrêt, Racine du texte, Analyse lexicale, Calcul de la matrice de résistance).

Les algorithmes utilisés sont : k-moyennes simples et l'algorithme spectral avec des jeux de données factices et réels. Le résultat montre que les deux algorithmes donnent des sorties presque similaires. La matrice de force joue un rôle important dans la recherche de similitudes entre les personnes. Les k-moyennes simples sont basées sur la compacité, donc elles donnent toujours des résultats plus proches d'approximation pour des ensembles de données numériques généraux. Spectral k-means est basé sur une approche de connectivité si efficacement appliquée au réseau social pour trouver la similitude textuelle et largement utilisée pour la formule de matrice de force. Le clustering spectral donne des résultats relativement rapides pour les ensembles de données élémentaires clairs semés et plus élevés, mais le coût de calcul du clustering spectral pour un grand ensemble de données est très élevé.

III.3.Travail 2 : Classification des postes sur des réseaux sociaux

[21]Les différentes stratégies de représentation du texte et le grand nombre de paramètres dans la phase de prétraitement, il était nécessaire de mettre en place un protocole d'expérimentation qui permet de varier ces paramètres et de trouver l'influence sur les résultats finaux. Pour concevoir ce protocole d'expérimentation, les auteurs ont défini les valeurs possibles de chaque paramètre dans des vecteurs pour l'exploitation automatique. Aussi ils ont utilisé une base de données qui contient la configuration de chaque expérimentation exécuté. La base d'expérimentation permet d'éviter les redondances des tests avec les mêmes paramètres.

Lors de premières expérimentations effectuées sans nettoyage, ils ont constaté que leur regroupement donne des clusters non équilibrés pour les quatre algorithmes et on n'arrive pas à identifier les fils des discussions. Après il ont utilisé la combinaison des types de nettoyage parmi : la suppression des liens, la suppression des identifiants, la conversion en minuscule et la suppression des caractères seuls ne conduit que vers un impact minimal sur les résultats finaux de clustering et cela donne aussi des partitions non-équilibrées .La combinaison de bons paramètres de clustering avec le paramètre le plus important de prétraitement (trouvés par expérimentation) qui est :

- Clustering avec les 2 types de nettoyage qui sont la mise à jour des hashtags par des dièses et la suppression du reste des mots.
- Transformation en vecteurs par conversion vers des vecteurs numériques Zscore.
- Clustering avec l'algorithme Kmeans. Cela a permet d'avoir un clustering équilibré, et d'améliorer considérablement l'identification des fils de discussions et aussi les termes les plus importants dans chaque cluster.

A partir d'une requête de collecte qui contient les coordonnées géographiques de la région Lyonnaise et les mots : "Lyon, FeteDesLumieres2015, FDL2015 et lumignon", les auteurs de cet article ont pu identifier des fils de discussions qui parlent de cet événement culturel sur Twitter, ce qui représente leur objectif principal. Aussi à la fin de leurs processus, ils arrivent à suggérer automatiquement d'autres termes liés à cet événement qui peuvent enrichir la requête de collecte. Leur contribution consiste à effectuer un regroupement avec les quatre algorithmes de clustering (K_Means, Clustering Hiérarchique, DBSCAN, OPTICS), sur le flux de Twitter et Instagram en passant par une phase de nettoyage et une phase de conversion du texte vers des vecteurs numériques, en fonction de la présence des termes statistiquement.

III.4.Travail 3 : La détection des communautés dans les réseaux sociaux :

La recherche des partitions en communautés est l'un des principaux problèmes liés à l'analyse des réseaux sociaux. Les Réseaux Sociaux sont parmi les plus gros producteurs de données. Donc, l'identification des communautés dans les réseaux sociaux s'avère un problème complexe. Les techniques de détection de communautés d'intérêt se réfèrent à une classification des nœuds du réseau plus densément connectés que d'autres, pour construire des classes connexes d'utilisateurs ayant les mêmes caractéristiques au regard d'une mesure de similarité se référant à des intérêts communs. D'après le travail de [22] Plusieurs classifications des méthodes proposées pour la détection des communautés d'intérêt dans les Réseaux Sociaux. Notons que, aujourd'hui, détection de communautés, partitionnement de graphe sont souvent utilisés indifféremment. Les techniques proposées dans ce travail sont :

- **Girvan et Newman** : C'est la méthode séparative la plus classique qui introduit une mesure de centralité appelée centralité d'intermédiarité des liens. Elle repose sur l'idée suivante : si un lien se trouve fréquemment sur les plus courts chemins entre les nœuds du graphe, alors il est naturel de penser qu'il ne se trouve pas « au sein d'une communauté donnée », Mais plutôt qu'il relie des portions distantes du graphe, des communautés distinctes. En retirant progressivement le lien qui a la plus forte « centralité », on obtient un découpage des blocs de notre réseau.

- **Louvain** : En2008, trois chercheurs de l'université de Louvain ont proposé une autre méthode "gloutonne", plus rapide que la majorité des autres approches. Sa particularité est de se fonder sur une approche locale de la modularité Situation initiale : Chaque nœud est une petite communauté à lui tout seul. Ensuite la passe 1 est constituée de deux phases : la **Migrations et la Réduction**

Ce travail s'intéresse à la détection de communauté dans un réseau social « Facebook ». Le résultat obtenu c'est été 16 différentes communautés.

La détection de communautés est un domaine qui est encore dans une phase d'exploration, et pour lequel il faudra encore attendre quelques années avant d'arriver à un stade de maturation. Cette relative jeunesse du domaine a, d'une part, représenté un challenge et, d'autre part, a été un facteur de motivation important.

III.5. Travail 4 : Regroupement des audiences sociales dans les réseaux d'information d'entreprise :

Une étape clé dans le processus d'intelligence économique des organisations consiste à regrouper les utilisateurs ayant des intérêts similaires en publics sociaux et à découvrir les rôles qu'ils jouent au sein d'un réseau d'entreprises. une nouvelle approche proposée par [23] appelée CBIN (Co-Clustering Business Information Network), qui regroupe les réseaux d'information commerciale pour découvrir et comprendre ces publics. Le cadre du CBIN est basé sur la Co-factorisation. Les groupes de publics sont découverts à partir d'une combinaison de structures de réseau et d'informations contextuelles riches, telles que les interactions entre les nœuds et les corrélations entre les nœuds et le contenu. La détermination préalable du nombre de groupes est généralement très difficile. C'est pourquoi CBIN basée sur un paradigme de chevauchement des grappes avec une stratégie d'attente pour découvrir le nombre optimal de grappes compte tenu des données sous-jacentes. Des expériences valident une performance exceptionnelle du CBIN par rapport à d'autres algorithmes sur 13 ensembles de données d'entreprise du monde réel.

Pour valider les performances des CBIN, les auteurs ont comparé les neuf algorithmes pertinents. Ces algorithmes comprenaient des approches qui exploitent uniquement des informations textuelles ou des informations topologiques, ainsi que des approches qui prennent en compte les deux. Les algorithmes qui ne prennent en compte que les informations textuelles ou topologiques sont : BigClam ,AgmFit ,K_Means ,Censa ,Circle , Block-LDA , FNMTF, DRCC , GNMF . Les expériences valident une performance exceptionnelle du CBIN par rapport à d'autres algorithmes sur 13 ensembles de données d'entreprise du monde réel.

Les principaux avantages de cette approche, par rapport à des algorithmes similaires, sont que le CBIN est une avancée dans la recherche actuelle de regroupement sur les réseaux d'information, tels que l'analyse de réseau personnel ou académique. Plus loin, CBIN Co-factorise simultanément les informations sous plusieurs angles : structure topologique (audiences), caractéristiques textuelles (données utilisateur) et corrélations entre les caractéristiques. Un principe de consensus intègre ces trois différents ensembles de résultats pour surmonter efficacement les incohérences dans le BIN pour un regroupement optimal performance. Le résultat final est la prise en charge d'un outil de business intelligence qui peut fournir une compréhension fonctionnelle approfondie des clients d'une entreprise et autres publics sociaux.

III.6. Conclusion

Dans ce chapitre, nous avons étudié et synthétiser 4 travaux de recherche qui traitent la classification non supervisée des réseaux sociaux tout en utilisant un ensemble d'algorithmes. Vu la diversité de ces recherches dans des contextes très différents, il nous a été difficile d'effectuer une analyse et une synthèse car il fallait trouver des critères et des paramètres de comparaisons que nous n'avons pas eu le temps de trouver.

Ces trois premiers chapitres de la partie une concernant l'état de l'art nous ont permis d'avoir un aperçu et une meilleure compréhension des réseaux sociaux et plus particulièrement les RSE, les algorithmes du clustering et enfin la classification de ces réseaux sociaux. La seconde partie sera consacré sur la classification d'un réseau social d'entreprise Workplace de Facebook au sein d'une école.

PARTIE II

CONTRIBUTION

PARTIE II – CONTRIBUTION

Chapitre IV: Clustering Workplace ESI

IV.1. Introduction :

Malgré la non exploitation de la plateforme Workplace ESI et vue que nous n'avons pas eu assez de temps pour chercher un dataset assez peuplé de données d'un autre réseau social d'entreprise, nous allons quand même effectuer du clustering sur les données récupérées de Workplace ESI. Nous allons utiliser l'algorithme K_Means et celui du clustering hiérarchique agglomérative.

Pour développer tous ces algorithmes, nous avons utilisé un certain nombre d'outils que nous allons aussi les décrire.

IV.2. Présentation de dataset extraite du workplace ESI :

La base qu'on a adopté pour nos tests de clustering afin d'établir notre algorithmes, prises de workplace ESI. Nous avons choisir la table du groupe « Everybody ». Les données extraites sont sous format CSV (Comma-Separated Values).

Le CSV est un fichier tableur qui a la structure suivante :

- La première ligne indique les noms des attributs séparés par des virgules «,», ('Date', 'Membres actifs', 'Publications', 'Commentaires', 'Réactions')
- Dans le reste du fichier, chaque ligne contient les valeurs des attributs prises par un objet, séparés par des virgules et mises dans le même ordre que celui des attributs correspondants.

La table qu'on a choisie possède 5 attributs (caractéristiques) de type numérique, et 238 objets (tuples) de données (**Figure 07**).

	Date	Membres actifs	Publications	Commentaires	Reactions
0	01/01/2020	26	0	0	6
1	02/01/2020	27	0	0	0
2	03/01/2020	12	0	0	0
3	04/01/2020	24	0	0	2
4	05/01/2020	8	0	0	1
...
233	21/08/2020	8	0	0	3
234	22/08/2020	5	0	0	2
235	23/08/2020	35	0	1	9
236	24/08/2020	13	0	0	0
237	25/08/2020	20	0	0	1

Figure 7 : Représentation matricielle de dataset Workplace ESI

IV.3. Algorithmes de clustering :

IV.3.1. Algorithme K-Means

Nous avons implémenté le clustering K-Means pour fournir des visualisations de tous les clusters connexes que nous trouvons dans l'ensemble de données. Nous avons donc au préalable filtré et nettoyé les données avant de les injectées dans l'algorithme.

Filtrage des valeurs numériques uniquement : Nos données sur le Workplace de ESI ne contiennent que des nombres, donc pas besoin de filtrage pour extraire uniquement les colonnes numérotées et non celles avec des mots (dans notre cas).

Suppression des valeurs nulles : Dans notre base, il n'existe pas des valeurs nulles

Standardisation : Enfin, nous devons trouver une sorte de moyen de standardiser les données et pour cela, nous utilisons le Standard Scaler de sklearn.

- **Méthodes d'évaluation**

Contrairement à l'apprentissage supervisé, l'analyse de clustering n'a pas une mesure d'évaluation solide que nous pouvons utiliser pour évaluer les résultats de différents algorithmes de clustering. En outre, puisque K-Means nécessite k comme une entrée et ne l'apprend pas à partir de données, il n'y a pas de bonne réponse en termes de nombre de clusters que nous devrions avoir dans n'importe quel problème. Parfois, la connaissance du domaine et l'intuition peut aider, mais généralement ce n'est pas le cas. Dans la méthodologie de prévision des clusters, nous pouvons

évaluer la performance des modèles en fonction de différents clusters K puisque les clusters sont utilisés dans la modélisation en aval.

Dans notre projet, nous allons couvrir deux mesures qui peuvent nous donner une certaine intuition sur k : - Méthode elbow et - Analyse de silhouette.

- La méthode elbow

La méthode idéale pour trouver le bon nombre de clusters serait d'essayer la mesure : Dans la somme des carrés du cluster (WCSS), qui mesure la distance moyenne au carré de tous les points d'un cluster au centre de gravité du cluster méthode ELBOW. (La distance euclidienne entre un point donné et le centre de gravité auquel il est affecté).

Code statique pour obtenir le nombre maximum de clusters :

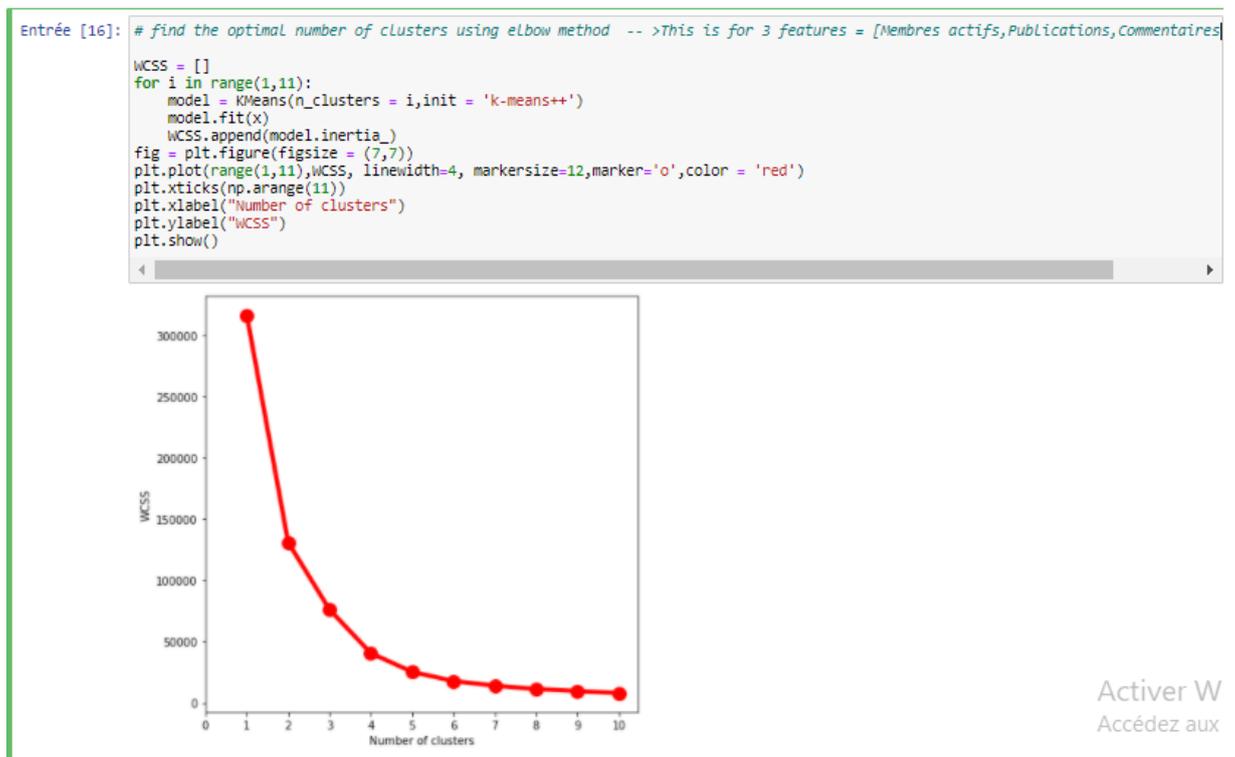


Figure 8 : Meilleur K dans K_Means de Workplace ESI

Si vous effectuez un zoom arrière sur cette courbe, vous verrez que le dernier coude arrive à $k = 4$ peu importe la plage que nous sélectionnons ex- (1,21) aussi nous verrons le même comportement mais si nous choisissons une plage plus élevée, il est peu difficile de visualiser le COUDE, c'est pourquoi nous préférons généralement la gamme (1,11).

Pour l'apprentissage non supervisé, nous utilisons "fit_predict ()", dans le cas de l'apprentissage supervisé, on utilise "fit_tranform ()".

Dans ce cas, le nombre optimal de clusters obtenu est de 4.

```
# since elbow occurred at 4, hence the optimal number of clusters for the mall data is 4
model = KMeans(n_clusters = 4, init = "k-means++", max_iter = 300, n_init = 10, random_state = 0)
y_clusters = model.fit_predict(x)
y_clusters
```

Figure 9 : Nombre optimal de clusters dans K_Means de Workplace ESI

- Analyse de silhouette

L'analyse de la silhouette peut être utilisée pour déterminer le degré de séparation entre les grappes. Pour chaque échantillon :

Calculer la distance moyenne de tous les points de données dans le même cluster (a_i).

Calculer la distance moyenne de tous les points de données dans le cluster le plus proche (b_i).

Calculer le coefficient :

$$\frac{b^i - a^i}{\max(a^i, b^i)}$$

Le coefficient peut prendre des valeurs dans l'intervalle $[-1, 1]$.

S'il est 0 \rightarrow l'échantillon est très proche des clusters voisins.

Il est 1 \rightarrow l'échantillon est loin des clusters voisins.

Il est -1 \rightarrow l'échantillon est attribué aux mauvais clusters.

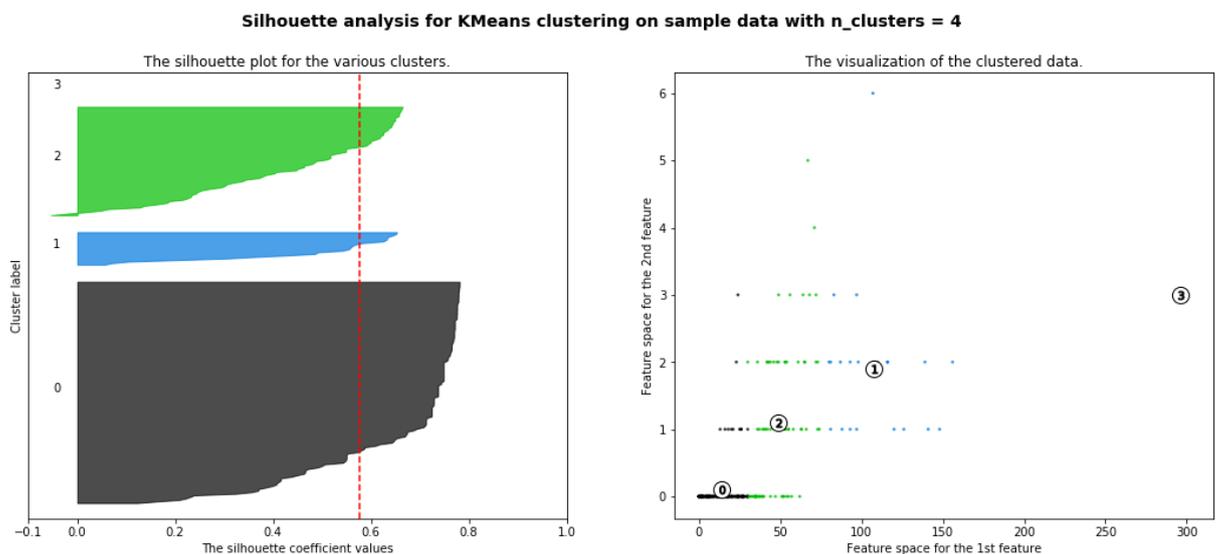


Figure 10 : Représentation graphique de l'analyse de silhouette de K-Means de Workplace ESI

```

# The vertical line for average silhouette score of all the values
ax1.axvline(x=silhouette_avg, color="red", linestyle="--")

ax1.set_yticks([]) # Clear the yaxis labels / ticks
ax1.set_xticks([-0.1, 0, 0.2, 0.4, 0.6, 0.8, 1])

# 2nd Plot showing the actual clusters formed
colors = cm.nipy_spectral(cluster_labels.astype(float) / n_clusters)
ax2.scatter(x[:, 0], x[:, 1], marker='.', s=30, lw=0, alpha=0.7,
           c=colors, edgecolor='k')

# Labeling the clusters
centers = clusterer.cluster_centers_
# Draw white circles at cluster centers
ax2.scatter(centers[:, 0], centers[:, 1], marker='o',
           c="white", alpha=1, s=200, edgecolor='k')

for i, c in enumerate(centers):
    ax2.scatter(c[0], c[1], marker='%s%s' % i, alpha=1,
               s=50, edgecolor='k')

ax2.set_title("The visualization of the clustered data.")
ax2.set_xlabel("Feature space for the 1st feature")
ax2.set_ylabel("Feature space for the 2nd feature")

plt.suptitle(("Silhouette analysis for KMeans clustering on sample data "
            "with n_clusters = %d" % n_clusters),
            fontsize=14, fontweight='bold')

plt.show()

For n_clusters = 2 The average silhouette_score is : 0.6627500978681343
For n_clusters = 3 The average silhouette_score is : 0.5819143770786638
For n_clusters = 4 The average silhouette_score is : 0.5754329860153755
For n_clusters = 5 The average silhouette_score is : 0.5278136377919975
For n_clusters = 6 The average silhouette_score is : 0.5173254276151044
For n_clusters = 7 The average silhouette_score is : 0.507699440467782
For n_clusters = 8 The average silhouette_score is : 0.48899596197494505
For n_clusters = 9 The average silhouette_score is : 0.4728046162982342
For n_clusters = 10 The average silhouette_score is : 0.47825172221234913

```

Figure 11 : Score de l'analyse de silhouette de K_Means de Workplace ESI

- **Visualisations des données**

Un simple K_Means va maintenant être appliqué aux données. Chaque cluster sera visualisé avec une couleur différente, donc j'espère que nous serons en mesure de choisir les clusters à l'œil nu. Nous avons utilisé les couleurs suivantes pour la représentation (figure 12) :

- noire : représente les centroides des clusters
- cluster0 (blue) : moins de publication et réaction
- cluster1 (rose) : publication et réaction moyenne
- cluster2 (jaune) : beaucoup de publication et réaction (membres sont actifs)
- cluster3 (vert) : il ya quelques membres qui sont très actifs

- La représentation en 3D

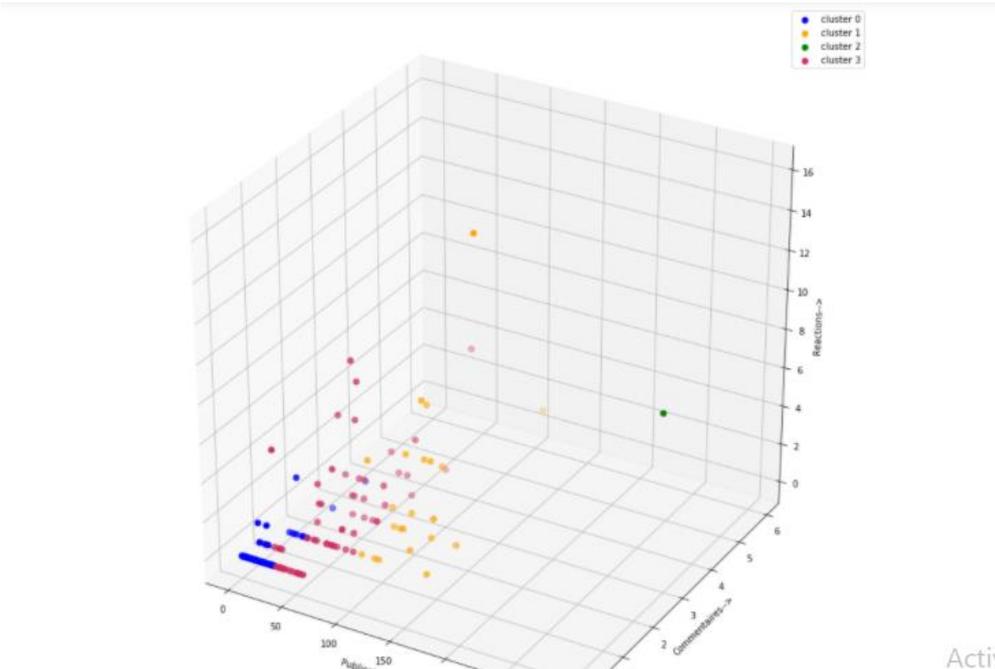


Figure 12 : Visualisation des données en 3D (K_Means)

- La représentation en 3D

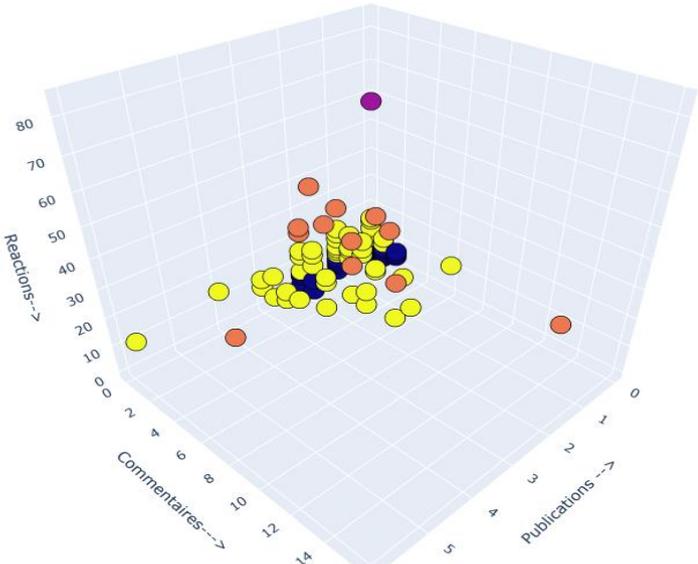


Figure 13 : Visualisation des données en 3 D (K_Means)

IV.3.2. Algorithme Hierarchical Agglomerative clustering

```
#Étape 1: importation des bibliothèques requises

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import scipy.cluster.hierarchy as sch
from sklearn.cluster import AgglomerativeClustering

%matplotlib inline
```

Nous créons une instance d'Agglomérative Clustering en utilisant la distance euclidienne comme mesure de la distance entre les points et la liaison de Ward pour calculer la proximité des grappes.

```
import scipy.cluster.hierarchy as sch
from sklearn.cluster import AgglomerativeClustering
dendrogram = sch.dendrogram(sch.linkage(data, method='ward'))
```

```
model = AgglomerativeClustering(n_clusters=3, affinity='euclidean', linkage='ward')
model.fit(data)
labels = model.labels_
```

labels

```
array([1, 1, 1, 1, 1, 1, 2, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 2, 0, 2, 1, 0, 2, 0, 2, 2, 1, 1, 1, 2, 1, 1,
       1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1,
       0, 1, 1, 2, 2, 1, 1, 1, 1, 2, 1, 1, 0, 2, 2, 0, 2, 2, 2, 2, 2, 2,
       2, 2, 0, 2, 2, 1, 2, 0, 0, 2, 2, 2, 2, 2, 1, 2, 2, 2, 2, 0, 2, 2,
       2, 2, 2, 2, 1, 2, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 1, 1, 1,
       1, 2, 2, 1, 2, 2, 2, 0, 2, 2, 1, 1, 1, 2, 1, 1, 1, 1, 2, 1, 1, 1,
       1, 1, 2, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1, 1, 2, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1, 1], dtype=int64)
```

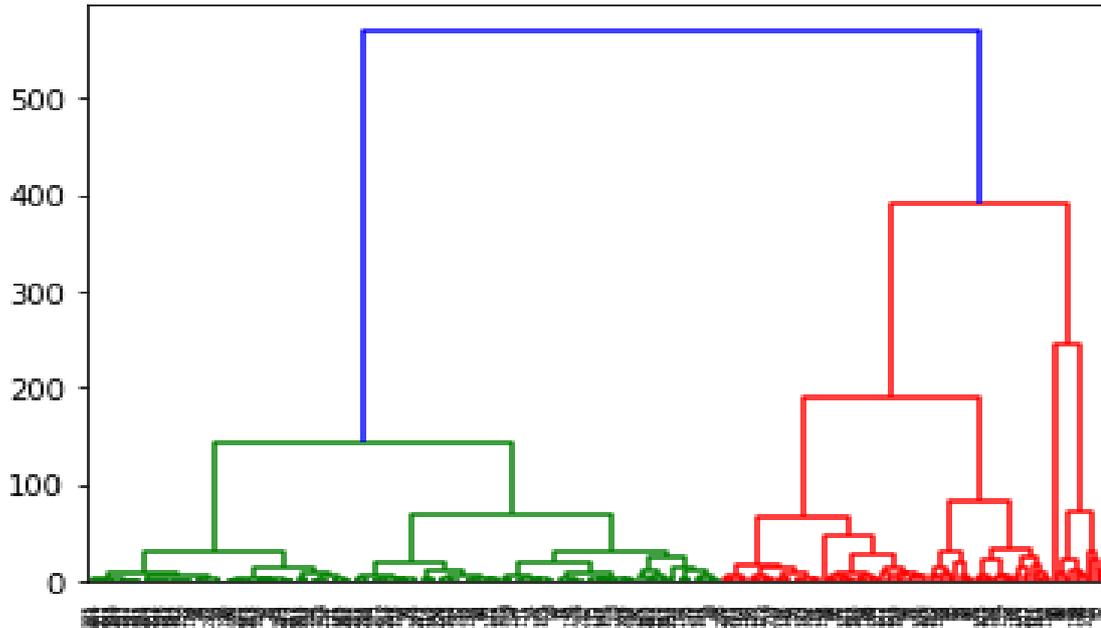


Figure 14 : Représentation du dendrogramme de l'algorithme de clustering hiérarchique agglomérative clustering de Workplace ESI

En regardant le dendrogramme, la distance verticale la plus élevée qui ne croise aucun cluster. Étant donné que 3 lignes verticales franchissent le seuil, le nombre optimal de clusters est de 3.

```
plt.scatter(X[labels==0, 0], X[labels==0, 1], s=50, marker='o', color='red')
plt.scatter(X[labels==1, 0], X[labels==1, 1], s=50, marker='o', color='blue')
plt.scatter(X[labels==2, 0], X[labels==2, 1], s=50, marker='o', color='green')
plt.show()
```

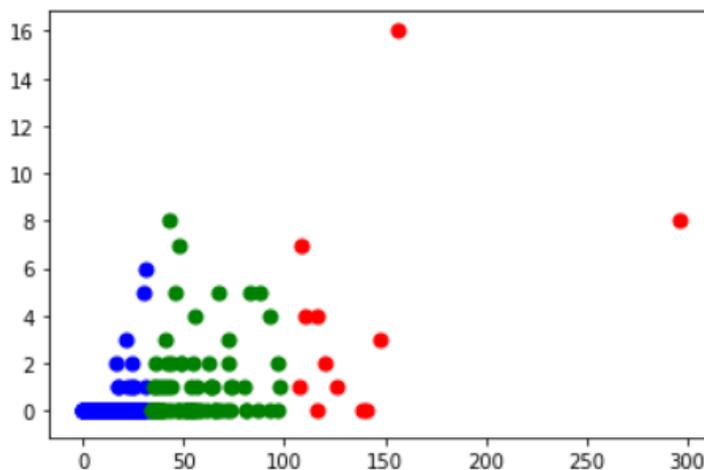


Figure 15 : Représentation graphique des résultats de l'algorithme de clustering hiérarchique agglomérative de Workplace ESI

IV.3.3. Comparaison entre K_Means et le clustering hiérarchique

Dans le clustering K_Means, nous essayons d'identifier la meilleure façon de diviser les données en k ensembles simultanément. Une bonne approche consiste à prendre k éléments de l'ensemble de données en tant que représentants du cluster initial, à affecter tous les éléments au cluster dont le représentant est le plus proche, puis à calculer la moyenne du cluster en tant que nouveau représentant, jusqu'à ce qu'il converge (tous les clusters restent les mêmes).

Dans le clustering agglomératif (clustering hiérarchique ascendant), nous commençons avec chaque élément de données ayant son propre cluster. Nous recherchons ensuite les deux éléments les plus similaires et les combinons dans un plus grand cluster. Nous continuons à répéter jusqu'à ce que tous les groupes qu'il nous reste soient trop différents pour être rassemblés, ou jusqu'à ce que nous atteignons un nombre prédéfini de groupes.

Les deux algorithmes ont des approches différentes mais sont censés aboutir aux mêmes résultats : Clusters précis.

Dans notre cas, le nombre optimal de clusters obtenu dans K_Means est de 4, par contre dans le clustering hiérarchique est de 3.

IV.3.4. Clustering spatial basé sur la densité des applications avec bruit (DBSCAN)

Nous allons utiliser le DBSCAN pour l'algorithme à des fins de clustering. Il s'agit d'un algorithme d'apprentissage automatique non supervisé. Il est utilisé pour les grappes de haute densité. Il prédit automatiquement les valeurs aberrantes et les supprime. C'est mieux que l'algorithme de clustering hiérarchique et K_Means. Il crée les clusters en fonction de paramètres tels que epsilon, points min et bruit. Il prédit séparément les points centraux, les points limites et les valeurs aberrantes de manière efficace.

Appliquez l'algorithme DBSCAN :

[24]L'algorithme DBSCAN considère les clusters comme des zones de haute densité séparées par des régions de faible densité. Pour cette raison, les clusters trouvés par DBSCAN peuvent être de n'importe quelle forme, par opposition aux K_Means, ce qui suppose que tous les clusters sont de forme convexe.

Sklearn avait un algorithme DBSCAN prédéfini ; tout ce que nous devons faire pour l'utiliser est de connaître trois paramètres :

eps: ce facteur indique la distance entre les différents points d'un même cluster.

min_samples: le nombre minimum d'échantillons dans le cluster. Dans le cas de grands ensembles de données bruyants, augmentez ce nombre.

metric: Ceci définit la métrique de distance entre les points de données. Sklearn a de nombreuses mesures de distance, telles que euclidienne, Manhattan et Minkowski. Pour notre cas, cependant, nous avons besoin d'une mesure de distance qui décrit la distance sur un chiffre . La métrique pour cela est appelée euclidien.

Nous pouvons maintenant appliquer notre modèle à l'ensemble de données :

```

Entrée [5]: from sklearn.cluster import DBSCAN
db=DBSCAN(eps=3,min_samples=4,metric='euclidean')

Entrée [6]: model=db.fit(x)

Entrée [7]: label=model.labels_
label

Entrée [9]: from sklearn import metrics

#identifying the points which makes up our core points
sample_cores=np.zeros_like(label,dtype=bool)

sample_cores[db.core_sample_indices_]=True

#Calculating the number of clusters

n_clusters=len(set(label))- (1 if -1 in label else 0)
print('No of clusters:',n_clusters)

No of clusters: 3
    
```

Figure 16 : la distance utiliser et le nombre du cluster

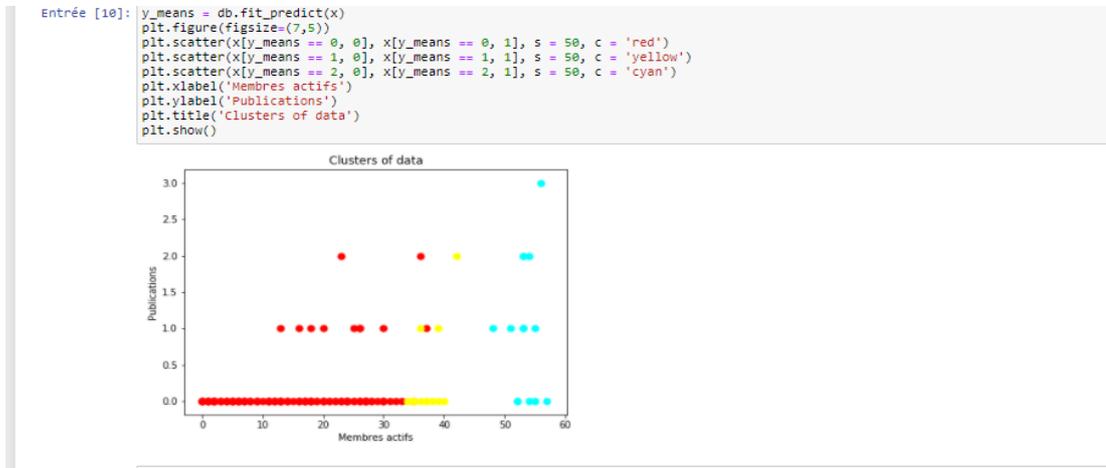


Figure 17 : visualisation des clusters en 2D

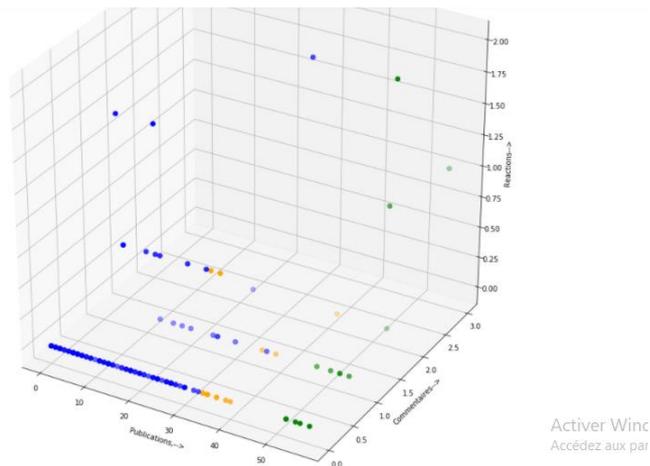


Figure 18 : Visualisation des clusters en 3D

IV.4. Interprétation des résultats :

Dans cette partie nous avons utilisé l'excel pour la représentation graphique des données d'ESI (un cluster)

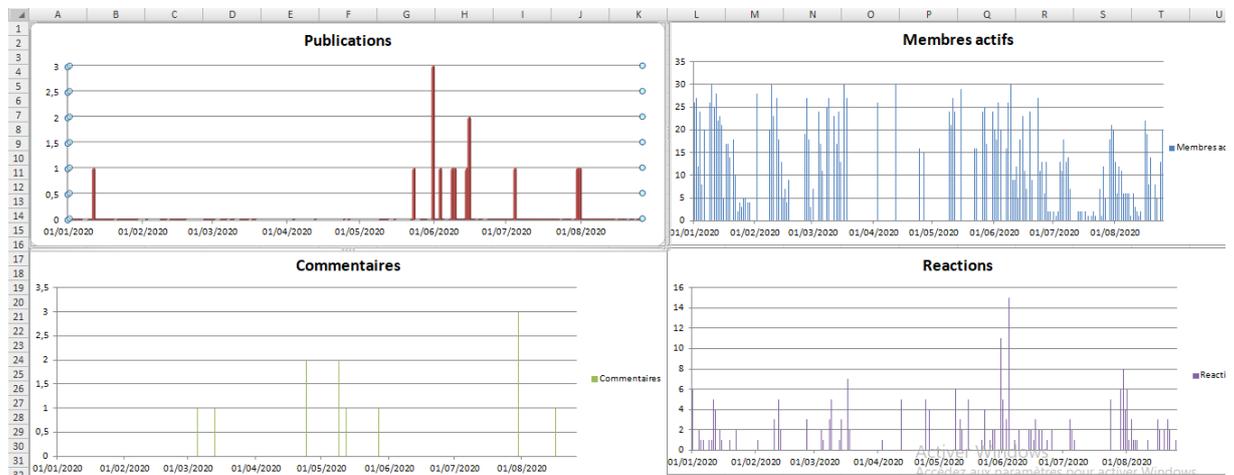


Figure 19 : Représentation graphique d'un cluster (publication, commentaire, réaction, membres actifs par date)

Après l'application de l'algorithme K_Means sur les données de Workplace ESI, nous avons obtenu des clusters, dans chaque cluster les données sont similaires, nous avons pris un seul cluster et nous l'avons traité pour extraire des connaissances.

D'après les quatre graphes (figure 19) :

- Au mois janvier, février, mars, avril, mai, juillet, août : les membres sont actifs mais peu ou pas de publication ni réaction ni commentaire.
- Par contre au mois de juin les membres sont plus actifs, plus de publications de commentaire et de réaction.
- Nous avons conclu que dans la période des examens les étudiants, enseignants interagissent plus.
- On peut extraire comme connaissance d'augmenter les publications dans le mois des examens (connaissance pas vraiment satisfaisante à cause des données insuffisantes).

IV.5. Outils de développement :

Avant tout, nous présentons en premier lieu la configuration matérielle que nous avons utilisée pour exécuter nos algorithmes et en second lieu l'environnement logiciel pour développer nos programmes. et matériel exploité dans notre projet.

• Configuration matérielle

- Ordinateur portable : LENOVO

- Système d'exploitation : Windows 10
- Processeur : Core I5
- Mémoire : 4 G RAM
- Disque dur : 512 Gb

• Configuration logicielle

• Anaconda

[25]Une distribution Python et R téléchargeable, gratuite, open source, haute performance et optimisée. Anaconda comprend conda, conda-build, Python et plus de 250 packages scientifiques open sources installées automatiquement et leurs dépendances qui ont été testés pour bien fonctionner ensemble, notamment SciPy, NumPy et bien d'autres. Utilisez la commande conda install pour installer facilement plus de 7500 packages open-source populaires pour la science des données, y compris des analyses avancées et scientifiques, - à partir du référentiel Anaconda. Utilisez la commande conda pour installer des milliers d'autres packages open source.

Étant donné qu'Anaconda est une distribution Python, il peut rendre l'installation de Python rapide et facile, même pour les nouveaux utilisateurs.

Disponible pour Windows, MacOS et Linux, toutes les versions d'Anaconda sont prises en charge par la communauté.

• Jupyter

[26]JupyterLab est un environnement de développement interactif basé sur le Web pour les blocs-notes, le code et les données Jupyter. JupyterLab est flexible : configurer et organiser l'interface utilisateur pour prendre en charge un large éventail de flux de travail en science des données, en informatique scientifique et en apprentissage automatique. JupyterLab est extensible et modulaire : écrivez des plugins qui ajoutent de nouveaux composants et s'intègrent aux composants existants.

Jupyter prend en charge plus de 40 langages de programmation, dont Python, R, Julia et Scala.

Les blocs-notes peuvent être partagés avec d'autres utilisateurs par e-mail, Dropbox, GitHub et Jupyter Notebook Viewer.

Votre code peut produire une sortie riche et interactive : HTML, images, vidéos, LaTeX et types MIME personnalisés.

Actuellement utilisé avec Google, Microsoft, IBM et Anaconda et on peut l'accéder à partir d'une plateforme en ligne Google Colab.

• Python :

[27]Python est un langage de programmation puissant et facile à apprendre. Il dispose de structures de données de haut niveau et permet une approche simple mais efficace de la

programmation orientée objet. Parce que sa syntaxe est élégante, que son typage est dynamique et qu'il est interprété, Python est un langage idéal pour l'écriture de scripts et le développement rapide d'applications dans de nombreux domaines et sur la plupart des plateformes.

- **Les bibliothèques**

Nous avons utilisé Python 3 qui est constitué de plusieurs bibliothèques. Parmi les packages que nous avons installé pour pouvoir travailler :

Matplotlib

Matplotlib est une bibliothèque python qui dessine des graphiques. Nul besoin de connaissances en interfaces graphiques pour créer un graphique dynamique avec possibilité de zoom et de sauvegarde par l'utilisateur. Il est d'ailleurs possible de sauvegarder les graphiques en format matriciels comme le PNG, JPEG, etc. et vectoriels comme le PDF et le SVG.

NumPy

Le module NumPy est incontournable en bio-informatique. Il permet d'effectuer des calculs sur des vecteurs ou des matrices, élément par élément, via un nouveau type d'objet appelé *array*.

Pandas

Le module pandas a été conçu pour l'analyse de données. Il est particulièrement puissant pour manipuler des données structurées sous forme de tableau.

Seaborn

Seaborn est une bibliothèque de visualisation de données Python basée sur Matplotlib. Il fournit une interface de haut niveau pour dessiner des graphiques statistiques attrayants et informatifs.

Plotly

La bibliothèque graphique Python de Plotly fait des graphiques interactifs de qualité de publication. Exemples de la façon de faire des parcelles de ligne, des parcelles de dispersion, des graphiques de zone, des graphiques à barres, des barres d'erreur, des parcelles de boîte, des histogrammes, des cartes thermiques, des sous-intrigues, des axes multiples, des graphiques polaires et des graphiques à bulles.

IV.6. Conclusion :

Dans ce chapitre, nous avons représenté les résultats obtenus à partir des algorithmes K_Means, le clustering hiérarchique agglomérative et le clustering basé sur la densité DBSCAN. On aurait souhaité de développer plus d'algorithmes mais malheureusement, on n'a pas eu assez de temps pour le faire. Les résultats obtenus sur le dataset de Workplace ESI ne sont pas intéressants pour effectuer une bonne analyse. Il faudra donc utiliser une autre plateforme d'un RSE pour pouvoir analyser et interpréter les clusters obtenus.

Conclusion Générale et Perspectives

Le but de ce mémoire est de présenter le travail que nous avons réalisé dans le cadre de notre projet de master. La mission qui nous a été confiée est d'effectuer une classification non supervisée utilisant les algorithmes du clustering sur le dataset du RSE Workplace de l'ESI. Pour ce faire, nous avons effectué un état de l'art pour comprendre ce que c'est qu'un RSE, les algorithmes du clustering et de voir quelques articles de recherche dans ce domaine. La seconde partie de ce travail consiste à appliquer le K_Means, le clustering hiérarchique agglomérative et l'algorithme DBSCAN à partir des données extraites de la base de données de Workplace de l'ESI. Les résultats obtenus ne sont pas satisfaisant le fait que cette plateforme n'est pas vraiment exploitée par ces collaborateurs. Ce qui nous a empêchés d'effectuer une bonne analyse et une interprétation des résultats obtenus.

Donc comme perspectives à ce travail, il serait souhaitable de tester ces algorithmes sur un dataset d'une autre plateforme vont servir à déterminer les bons critères de mesures de l'intelligence collective au sein d'une organisation.

Nous souhaitons que ce projet peut être amélioré en essayant de couvrir plus des domaines qui s'intéressent sur les RSE et l'intelligence collective et intégrer quelque approche

- Modération automatique des réseaux sociaux d'entreprise par le Deep Learning.
- Avoir recours à l'intelligence artificielle sur les médias sociaux d'entreprise pour personnaliser l'information en fonction des besoins de chaque utilisateur.

Références Bibliographiques:

- [1] M. Drahoov'a and P. Balco, "The benefits and Risks of Entreprise Social Networks," in *2016 International Conference on Intelligent Networking and Collaborative Systems (INCoS)*, IEEE, 2016, pp. 15–19.
- [2] I. Comtet, "Les environnements collaboratifs de travail au service de l'intelligence collective économique?," *Commun. Organ.*, pp. 61–72, 2012, doi: 10.4000/communicationorganisation.3858.
- [3] A. W. Woolley, Y. Kim, and T. W. Malone, "Measuring Collective Intelligence in Groups: A Reply to Credd and Howardson," *SSRN Electron. J.*, Jun. 2018, doi: 10.2139/ssrn.3187373.
- [4] G. Alain and G. Hervier, *Le réseau social d ' entreprise Le réseau social d ' entreprise*. 2011.
- [5] C. Imhoff, "Penser la collaboration dans les organisations à partir des communautés virtuelles sur le réseau social d'entreprise: rapports sociaux et modes de régulation émergents : continuités, contradictions et/ou ruptures," Conservatoire national des arts et metiers - CNAM, Mar. 2018.
- [6] "organisation-entreprise," 2020. [Online]. Available: <https://www.manager-go.com/organisation-entreprise/reseaux-sociaux-d-entreprise.htm>.
- [7] "e-marketing," 2020. [Online]. Available: <https://www.e-marketing.fr/Thematique/academie-1078/fiche-outils-10154/Les-reseaux-sociaux-d-entreprise-325694.htm>.
- [8] "codeur," 2020. [Online]. Available: <https://www.codeur.com/blog/reseau-social-d-entreprise/>.
- [9] "journaldunet," 2020. [Online]. Available: <https://www.journaldunet.com/solutions/reseau-social-d-entreprise/1124601-sharepoint-comme-reseau-social-d-entreprise-une-fausse-bonne-idee/>.
- [10] F.-X. Jollois, "Contribution de la classification automatique à la fouille de données .," 2003.
- [11] J. Irani, P. Nitin, and P. Madhura, "Clustering Techniques and the Similarity Measures used in Clustering : A Survey," vol. 134, no. 7, pp. 9–14, 2016.
- [12] S. P. S. Kushwah, K. Rawat, and P. Gupta, "Analysis and comparison of efficient techniques of clustering algorithms in data mining," *Int. J. Innov. Technol. Explor. Eng.*, vol. 1, no. 1, pp. 2278–3075, 2012.
- [13] L. and Meng, A.-H. Tan, and D. C. Wunsch II, "Clustering and Its Extensions in the Social Media Domain," in *Adaptive Resonance Theory in Social Media Data Clustering*, Springer, 2019, pp. 15–44.
- [14] M. Ester, H. Kriegel, J. Sander, X. X.- Kdd, and undefined 1996, "A density-based algorithm for discovering clusters in large spatial databases with noise.," *aaai.org*.
- [15] Y. Lv *et al.*, "An efficient and scalable density-based clustering algorithm for datasets with complex structures," *Neurocomputing*, vol. 171, pp. 9–22, 2016.
- [16] K. A.-S.-P. recognition and U. 1995, "A tabu search approach to the clustering problem," *Pattern Recognit.*, vol. 28, no. 9, pp. 1443–1451, 1995.
- [17] R. Xu and D. Wunsch, "Survey of clustering Algorithms," *IEEE Trans. neural networks*,

vol. 16(3), pp. 645–678, 2005.

- [18] I. H. Anh TUAN, “Réduction de base de données par la classification automatique(Doctoral dissertation),” 2004.
- [19] S. Tollari, “Indexation et recherche d’images par fusion d’informations textuelles et visuelles,” 2006.
- [20] K. Singh, H. K. Shakya, and B. Biswas, “Clustering of people in social network based on textual similarity,” *Perspect. Sci.*, vol. 8, pp. 570–573, Sep. 2016, doi: 10.1016/j.pisc.2016.06.023.
- [21] A. Remadna, “Classification des posts sur des réseaux sociaux,” 2015.
- [22] N. LOGRADA, “La détection de communautés dans les réseaux sociaux,” UNIVERSITE MOHAMED BOUDIAF-M’SILAFACULTE DES MATHEMATIQUES ET DE L’INFORMATIQUE, 2019.
- [23] Y. Zheng *et al.*, “Clustering social audiences in business information networks,” *Pattern Recognit.*, vol. 100, p. 107126, Apr. 2020, doi: 10.1016/j.patcog.2019.107126.
- [24] “datascience,” 2020. [Online]. Available: <https://towardsdatascience.com/contact-tracing-using-less-than-30-lines-of-python-code-6c5175f5385f>.
- [25] “Conda — conda 4.8.4.post45+6d1fea11 documentation.” .
- [26] U. S. P. & T. Office, “Project Jupyter | Home,” 2020. [Online]. Available: <https://jupyter.org/>. [Accessed: 25-Aug-2020].
- [27] “Le tutoriel Python — Documentation Python 3.8.5,” 19-Aug-2020. [Online]. Available: <https://docs.python.org/fr/3/tutorial/>. [Accessed: 25-Aug-2020].