Ministère de l'Enseignement supérieur et de la

Recherche scientifique

Faculté des sciences économiques et gestion de Nabeul

Analyse Réseaux Sociaux: StackOverFlow case

1 ére année mastére de recherche "Business Computing"

Travail élaboré par : waad bouzidi

Partie 1:

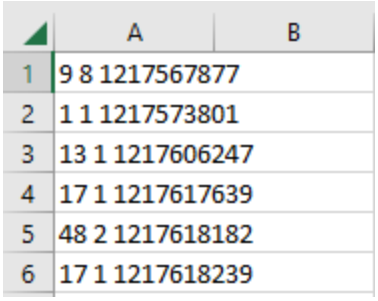
1. **Source de données en ligne** :
La source de données en ligne sélectionnée est le site web Stack Overflow, qui est une plateforme de questions-réponses pour les développeur(https://stackoverflow.com/).
Le lien de téléchargement du dataset est :
https://snap.stanford.edu/data/sx-stackoverflow.html

2. Entités et relations :
Les entités principales dans le réseau sont les utilisateurs de Stack Overflow, représentés par des nœuds. Les relations entre ces utilisateurs sont définies par trois types d'interactions :
- L'utilisateur U a répondu à la question de l'utilisateur V à un moment donné t (représenté par un lien de u à v dans le graphique sx-stackoverflow-a2q).
- L'utilisateur U a commenté la question de l'utilisateur V à un moment donné t (représenté par un lien de u à v dans le graphique sx-stackoverflow-c2q).
- L'utilisateur U a commenté la réponse de l'utilisateur V à un moment donné T (représenté par un lien de U à V dans le graphique sx-stackoverflow-c2a).

3.Informations additionnelles :
Outre les relations entre les utilisateurs, d'autres informations valables peuvent inclure des attributs pour chaque nœud, tels que le nombre de questions posées par chaque utilisateur, le nombre de réponses fournies par chaque utilisateur, etc.(ces caractéristiques ont été utilisés dans la partie IV:prédictions)

4. Obtention des données :
Les données ont été extraites à partir du Stack Exchange Data Dump, qui est une archive publique contenant des données de divers sites Stack Exchange, y compris Stack Overflow.

5. Construction du réseau :
Les données extraites ont été utilisées pour construire un réseau à partir des interactions entre les utilisateurs de Stack Overflow, en représentant les utilisateurs comme des nœuds et les interactions comme des liens dirigés entre ces nœuds. Le réseau a été construit en utilisant les informations sur les réponses aux questions, les commentaires sur les questions et les commentaires sur les réponses.



	A	B
1	9 8 1217567877	
2	1 1 1217573801	
3	13 1 1217606247	
4	17 1 1217617639	
5	48 2 1217618182	
6	17 1 1217618239	

Fig 1 : exemple du fichier xlsx du dataset.

Dataset statistics (sx-stackoverflow)	
Nodes	2601977
Temporal Edges	63497050
Edges in static graph	36233450
Time span	2774 days

Fig 2 : détails sur le graph.

Le graphe construit à partir des données de Stack Overflow présente une grande complexité, comprenant un total de 2 601 977 nœuds représentant les utilisateurs de la plateforme. Ces utilisateurs sont connectés par un total impressionnant de 63 497 050 arêtes, reflétant les interactions dynamiques entre eux au fil du temps. Dans le graphe statique, qui représente l'ensemble des interactions sans tenir compte du moment où elles se produisent, il y a 36 233 450 arêtes. Ces chiffres témoignent de l'ampleur de l'activité sur Stack Overflow, avec un ensemble de données capturant un large éventail d'interactions entre les utilisateurs au cours d'une période de 2774 jours, mettant en lumière la richesse des échanges et des collaborations au sein de la communauté des développeurs.

Partie II:

Pour approfondir notre compréhension du réseau construit à partir des données de Stack Overflow, nous entreprenons dans cette partie une analyse approfondie à l'aide du langage de programmation Python. Pour cela, nous utiliserons la bibliothèque NetworkX, une ressource bien établie dans le domaine de l'analyse des réseaux. Notre analyse du réseau comprend plusieurs aspects essentiels, notamment *la distribution des degrés, les composants connectés, les chemins, le coefficient de clustering, la densité du réseau, et la centralité*. Ces analyses nous permettront d'appréhender la structure et les caractéristiques essentielles du réseau, offrant ainsi des perspectives éclairantes sur la dynamique et l'organisation de la communauté des développeurs sur Stack Overflow. Les résultats obtenus seront interprétés et présentés de manière détaillée par la suite, fournissant ainsi des insights significatifs pour une meilleure compréhension du réseau et de son fonctionnement.

⚡La taille du dataset est trop volumineuse, l'exécution des codes prend beaucoup du temps pour cela j'ai travaillé sur 1000 nœuds.
Commençant par **la représentation du graph:**

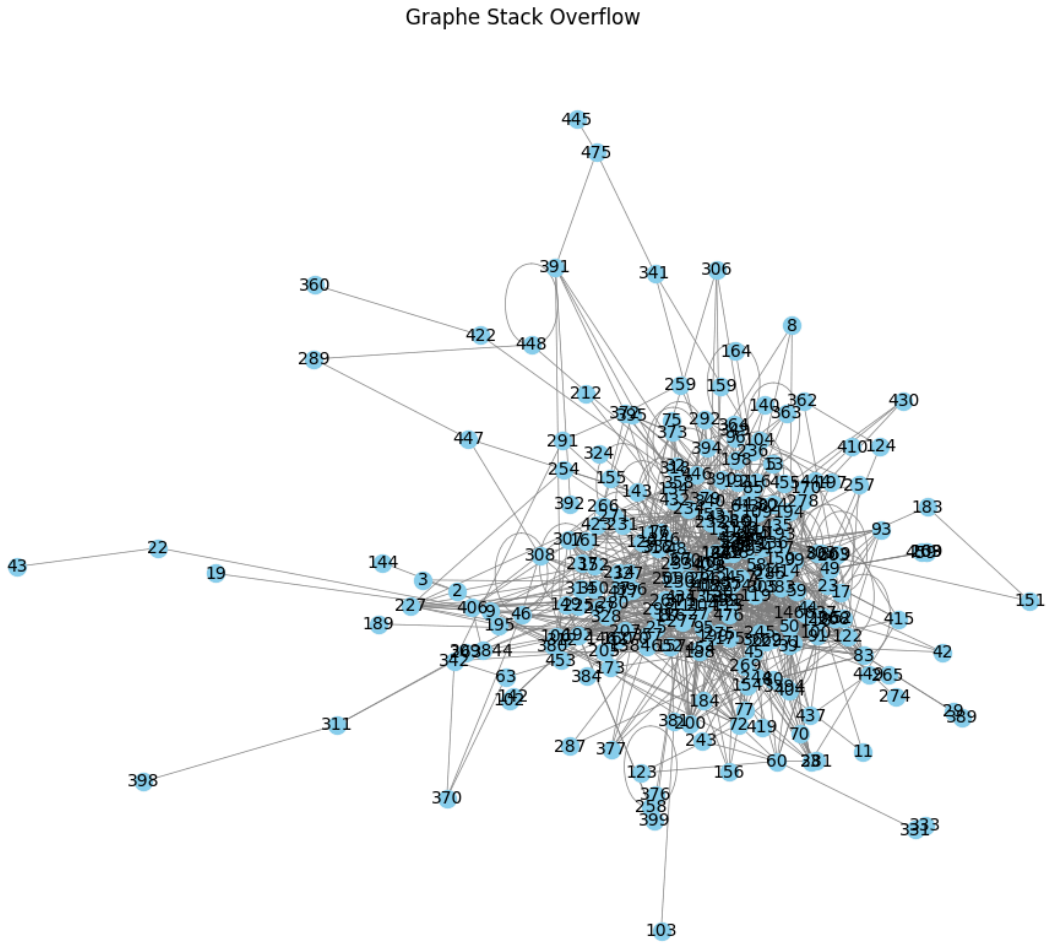


Fig 3 : Graph StackOverFlow

Dans le cas de Stack Overflow, où les interactions se manifestent sous forme de questions, réponses et commentaires entre les utilisateurs, il est raisonnable que les nœuds du réseau soient relativement proches les uns des autres. Cela s'explique par le fait que les utilisateurs interagissent souvent entre eux sur des sujets similaires ou complémentaires. Par exemple, un utilisateur posant une question peut recevoir des réponses de plusieurs autres utilisateurs, ce qui crée des liens directs entre ces nœuds dans le réseau. De plus, les commentaires peuvent être laissés sur les questions ou les réponses, ajoutant ainsi de nouvelles connexions entre les utilisateurs. Cette interconnexion est renforcée par le fait que les utilisateurs sont souvent actifs dans des communautés spécifiques liées à leurs domaines d'intérêt ou à leurs compétences. Ainsi, la proximité des nœuds dans le réseau de Stack Overflow reflète la nature collaborative et interconnectée de la plateforme, où les utilisateurs s'engagent activement les uns avec les

autres pour échanger des connaissances, résoudre des problèmes et contribuer à la communauté dans son ensemble.

Distribution des degrés:

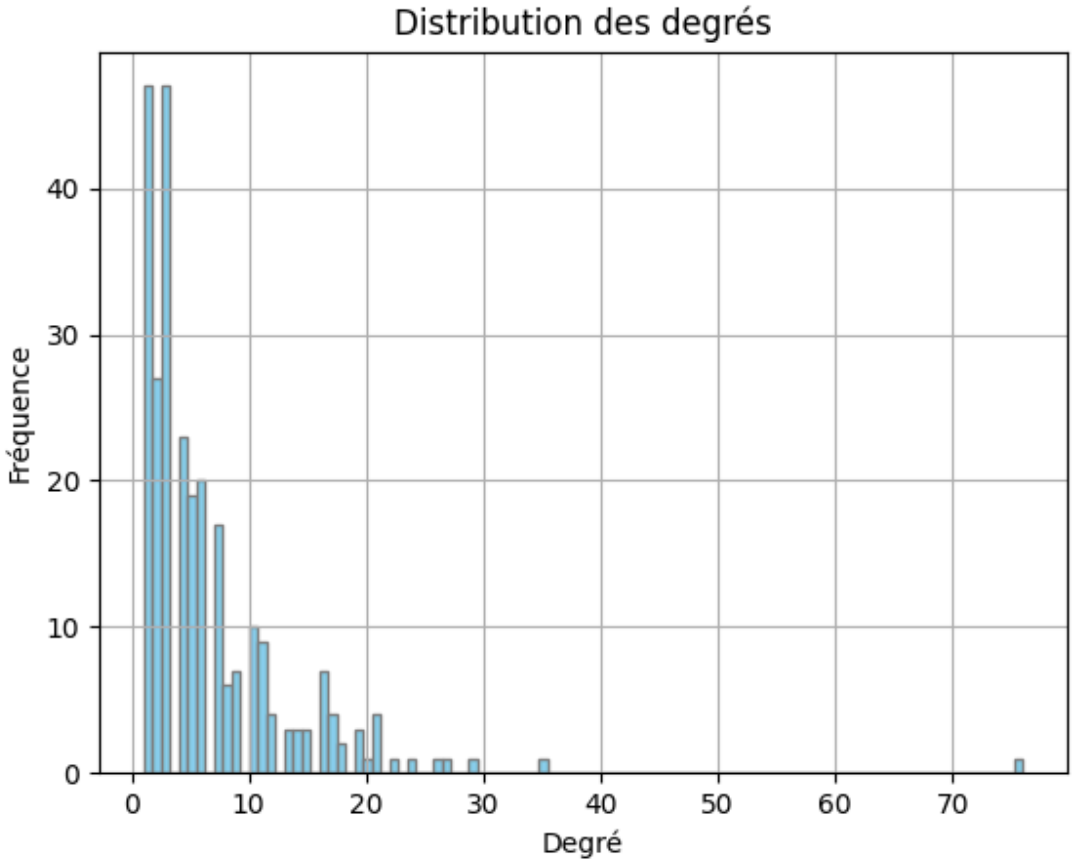


Fig 4: Graphique de la distribution des degrés.

Dans ce graphe Stack Overflow peut fournir des informations sur la structure du réseau et les interactions entre les différents types d'entités: les questions, les réponses et les commentaires.

Dans cet histogramme, l'axe des abscisses (X) représente les degrés, c'est-à-dire le nombre de connexions qu'a chaque nœud dans le réseau. L'axe des ordonnées (Y) représente la fréquence, c'est-à-dire le nombre de nœuds dans le réseau qui ont un certain degré.

En observant l'histogramme, nous pouvons remarquer que la majorité des nœuds ont un faible degré, ce qui signifie qu'ils ont peu de connexions avec d'autres nœuds. Cela pourrait correspondre aux utilisateurs qui posent des questions ou donnent des réponses, car ils peuvent ne pas être très actifs dans la communauté ou ne pas avoir beaucoup de commentaires sur leurs publications.

Les premiers barreaux montrent des valeurs relativement élevées, ce qui suggère qu'il y a quelques nœuds très connectés dans le réseau. Ces nœuds pourraient être des questions ou des réponses très populaires qui ont reçu un grand nombre de commentaires.

La diminution progressive de la fréquence à mesure que le degré augmente indique qu'il y a de moins en moins de nœuds avec un degré élevé. Cela est cohérent avec l'idée que dans Stackoverflow, il y a quelques utilisateurs très actifs qui interagissent beaucoup, mais la plupart des utilisateurs ont un niveau d'activité plus faible.

***les composants connectés*:**

Dans le contexte de Stack Overflow, où les utilisateurs interagissent en posant des questions, fournissant des réponses et laissant des commentaires, il est probable que la plupart, voire tous les composants du réseau, soient connectés. Cette connectivité est due à la nature de la plateforme, où les utilisateurs peuvent naviguer librement entre les questions, les réponses et les commentaires, et où chaque interaction crée un lien direct entre les utilisateurs impliqués.

Puisque chaque interaction crée un lien dans le réseau, et étant donné que les utilisateurs sont souvent engagés dans plusieurs interactions différentes, il est probable que même les composants les plus petits soient connectés au reste du réseau. Par exemple, même si un sous-ensemble de questions et de réponses peut sembler isolé à première vue, il est fort probable qu'il y ait des utilisateurs qui ont interagi avec ces éléments et qui sont également connectés à d'autres parties du réseau.

En outre, les utilisateurs de Stack Overflow sont encouragés à explorer et à participer à diverses discussions, ce qui contribue également à la connectivité globale du réseau. La diversité des sujets traités sur la plateforme et la large base d'utilisateurs actifs favorisent ainsi la connexion de tous les composants du réseau, assurant une interactivité fluide et une circulation efficace de l'information au sein de la communauté.

Les chemins plus courts:

Les chemins les plus courts représentent les itinéraires les plus courts entre différents nœuds dans le graphe représentant les interactions sur Stack Overflow. Chaque ligne indique un chemin spécifique entre deux nœuds, où chaque nœud dans le chemin représente une interaction spécifique. Par exemple, "Chemin le plus court de 376 à 160: ['376', '277', '269', '72', '160']" indique le chemin le plus court de l'utilisateur 376 à l'utilisateur 160, passant par les interactions suivantes : 376 a interagi avec 277, puis 277 avec 269, ensuite 269 avec 72, et enfin 72 avec 160.

Ces interactions peuvent être interprétées comme suit par rapport aux interactions sur Stack Overflow :

- Poser une réponse à une question : L'utilisateur qui pose une réponse (source) interagit avec la question (target).
- Commenter à une question : L'utilisateur qui commente (source) interagit avec la question (target).
- Commenter à une réponse : L'utilisateur qui commente (source) interagit avec la réponse (target).

Ainsi, ces chemins les plus courts décrivent les parcours les plus courts entre différentes actions sur Stack Overflow, ce qui peut aider à comprendre la dynamique des interactions entre les utilisateurs sur la plateforme.

Coefficient de clustering moyen:

Le coefficient de clustering moyen mesure à quel point les nœuds dans le réseau tendent à se regrouper en cliques ou en clusters.
Le coefficient de clustering moyen est 0.10140071974482612 ce qui suggère que les nœuds du réseau ont tendance à former des clusters locaux dans une certaine mesure. Cela peut être interprété comme une indication que les utilisateurs de Stack Overflow ont tendance à interagir avec des groupes restreints d'autres utilisateurs autour de sujets spécifiques ou de domaines d'intérêt communs. Ces clusters locaux peuvent représenter des communautés spécialisées ou des groupes de collaboration sur des sujets particuliers.

Densité du réseau:

La densité du graphe indique à quel point le réseau est connecté, c'est-à-dire le nombre réel de liens par rapport au nombre maximum possible de liens.
La densité du graphe est 0.023809523809523808 ce qui signifie que seulement environ 2.38% des connexions potentielles entre les noeuds sont réellement présentes dans le graphe.une densité relativement faible peut être attendue en raison de la diversité des sujets et des utilisateurs présents sur la plateforme. Bien que les interactions soient fréquentes entre certains groupes d'utilisateurs, il peut y avoir également de nombreuses parties du réseau où les liens sont moins courants. Cela peut refléter la nature spécifique des questions posées, des domaines d'expertise des utilisateurs et des flux d'activité sur la plateforme.

La centralité des nœuds :

La centralité des nœuds peut être interprétée comme une mesure de l'importance de chaque nœud dans le réseau. Dans le contexte de Stack Overflow, cela pourrait représenter l'importance ou la visibilité des utilisateurs ou des questions/réponses.
Les nœuds avec une centralité plus élevée sont potentiellement plus influents ou actifs dans la communauté.
En examinant les valeurs de centralité des nœuds, on constate une variation dans les niveaux d'importance des nœuds dans le réseau. Certains nœuds ont des valeurs de centralité

relativement élevées, tandis que d'autres ont des valeurs plus faibles. Les nœuds avec des valeurs de centralité plus élevées sont susceptibles d'être des acteurs clés dans le réseau, jouant un rôle central dans la transmission d'informations ou l'interconnexion d'autres parties du réseau. En revanche, les nœuds avec des valeurs de centralité plus faibles peuvent être moins influents ou moins connectés dans le réseau.

Partie III:

Dans cette partie, nous abordons la découverte de la communauté, une tâche cruciale dans l'analyse des réseaux sociaux. Nous nous concentrons sur la décomposition d'une topologie de réseau complexe en clusters de nœuds significatifs. Pour ce faire, nous évaluons et validons la structure modulaire de l'échantillon de réseau sélectionné. Nous comparons les résultats de différents algorithmes de détection des communautés, dont ***le K-clique, Louvain,et Demon/Angel***.

Pour analyser les résultats des algorithmes k-clique, Louvain, et démon/angel dans le contexte des interactions sur Stack Overflow, nous devons d'abord comprendre comment ces algorithmes identifient les groupes ou les communautés dans les données.

1. k-clique : Cet algorithme identifie les groupes des nœuds (utilisateurs dans ce contexte) qui sont complètement connectés les uns aux autres, c'est-à-dire qu'ils forment un sous-graphe où chaque nœud est connecté à chaque autre nœud. Les cliques dans ce contexte pourraient représenter des groupes d'utilisateurs très actifs ou interconnectés d'une manière spécifique sur Stack Overflow.

2. Louvain : Cet algorithme est une méthode de détection de communauté qui maximise la modularité dans un graphe en attribuant des nœuds à des communautés. Il détecte des structures de communauté dans les graphes en trouvant des partitions qui maximisent le nombre de connexions à l'intérieur des communautés et minimisent le nombre de connexions entre les communautés. Dans le contexte de Stack Overflow, cela peut être interprété comme la détection de groupes d'utilisateurs qui sont très actifs ensemble ou qui partagent des intérêts similaires.

3. Démon/Angel: Cet algorithme est utilisé pour identifier les structures hiérarchiques dans les graphes. Les nœuds sont classés en démons et anges selon leur rôle dans la hiérarchie. Les démons sont des nœuds qui sont des points de passage obligés entre les autres nœuds, tandis que les anges sont des nœuds qui se connectent à ces démons. Dans le contexte de Stack Overflow, cela est interprété comme l'identification des utilisateurs qui ont une influence significative sur les interactions entre autres utilisateurs (démons) et ceux qui sont fortement influencés ou dépendent de ces interactions (anges).

Maintenant, pour interpréter les résultats :

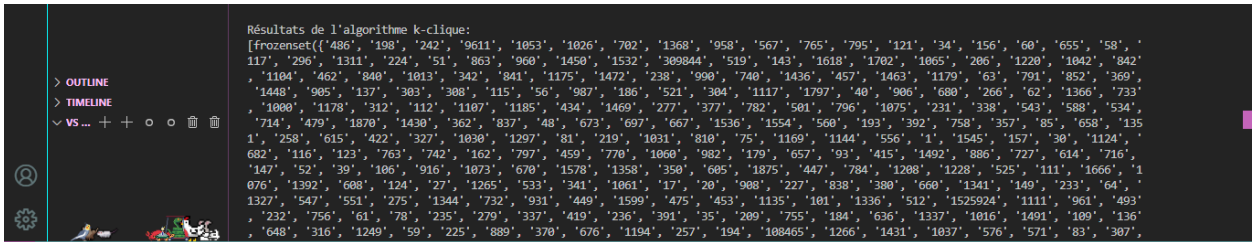


Fig 5: résultat de l'algo K-clique.

- Les cliques identifiées par l'algorithme k-clique pourraient représenter des groupes d'utilisateurs très actifs ou interconnectés dans des discussions spécifiques sur Stack Overflow.

- Les communautés détectées par l'algorithme Louvain pourraient représenter des groupes d'utilisateurs partageant des intérêts similaires ou travaillant sur des projets similaires, ce qui les amène à interagir fréquemment sur la plateforme.

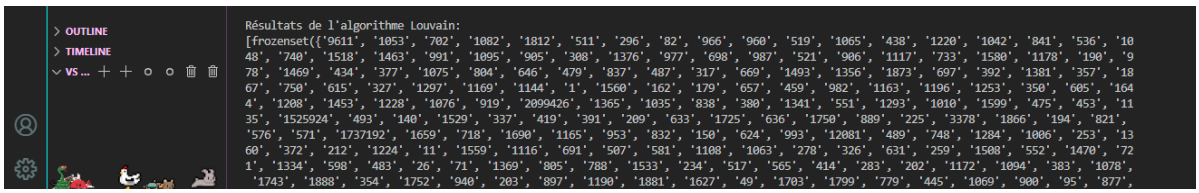


Fig 6: résultat de l'algo louvain

- Les structures hiérarchiques détectées par l'algorithme démon/ange pourraient révéler des utilisateurs qui sont des leaders d'opinion ou des contributeurs majeurs (démons) et ceux qui sont plus réceptifs ou qui suivent ces leaders (anges).

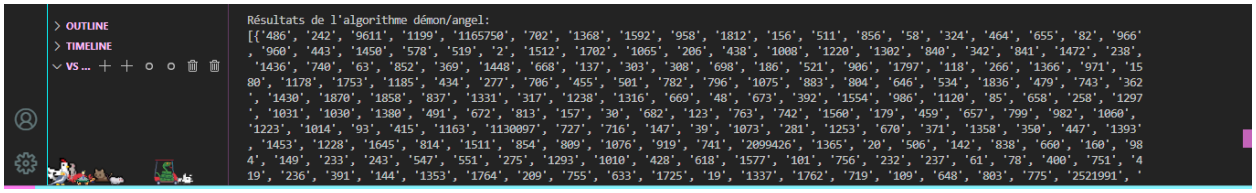


Fig 7: résultat de l'algo demon/angel

En combinant ces informations, on peut avoir une vision plus complète des dynamiques d'interaction et des sous-communautés présentes sur Stack Overflow, ce qui peut être utile pour comprendre comment l'information circule et comment les utilisateurs interagissent sur la plateforme.

Partie IV:

Dans cette quatrième et dernière partie, nous aborderons la prédiction des liens, une étape cruciale dans l'analyse des réseaux. Pour ce faire, nous adopterons une approche basée sur l'apprentissage non supervisé.

L'algorithme utilisé est: K-means pour regrouper les utilisateurs en clusters en fonction du nombre de questions qu'ils publient, du nombre de réponses qu'ils fournissent et du nombre de commentaires qu'ils écrivent sur une plateforme de questions-réponses comme Stack Overflow. Voici une explication détaillée de chaque étape :

1. Lecture des données : Le code commence par lire un fichier texte "stackoverflow.txt", où chaque ligne représente une action d'un utilisateur sur la plateforme, avec les valeurs séparées par des espaces.

2. Calcul des statistiques : Trois types d'actions sont considérés : publier une question (Action = 9), fournir une réponse (Action = 1) et écrire un commentaire (Action = 2). Le code calcule le nombre de chaque type d'action pour chaque utilisateur, regroupé par leur identifiant.

3. Fusion des statistiques : Les statistiques calculées sont fusionnées en un seul DataFrame pour chaque utilisateur, contenant le nombre de questions publiées, le nombre de réponses fournies et le nombre de commentaires écrits.

4. Remplacement des valeurs manquantes :Les valeurs NaN (qui représentent les utilisateurs qui n'ont pas effectué une certaine action) sont remplacées par 0.

5. Préparation des données pour le clustering : Les caractéristiques utilisées pour le clustering sont extraites du DataFrame en enlevant l'identifiant de l'utilisateur.

6.Application de l'algorithme K-means : L'algorithme K-means est appliqué pour regrouper les utilisateurs en un nombre donné de clusters (dans cet exemple, 3 clusters sont utilisés).

7.Ajout des clusters au DataFrame : Les clusters assignés à chaque utilisateur sont ajoutés au DataFrame.

8. Affichage des résultats : Le DataFrame avec les clusters assignés est affiché, montrant à quel cluster chaque utilisateur appartient. Ensuite, une visualisation des clusters est réalisée en utilisant un diagramme de dispersion où chaque point représente un utilisateur, avec le nombre de questions publiées sur l'axe des x et le nombre de réponses fournies sur l'axe des y, et la couleur du point indique le cluster auquel il appartient.

	1	Nombre_questions_publiees	Nombre_reponses_fournies	Nombre_commentaires_ecrits	Cluster
0	1	0.0	22.0	1.0	1
1	2	0.0	0.0	1.0	2
2	3	0.0	1.0	0.0	0
3	4	0.0	2.0	0.0	0
4	5	2.0	1.0	0.0	2
..	...	...	...	...	...
536	1567513	0.0	1.0	0.0	0
537	1663528	0.0	0.0	1.0	2
538	1670022	0.0	1.0	0.0	0
539	2068301	2.0	0.0	0.0	2
540	2486915	0.0	1.0	0.0	0

[541 rows x 5 columns]

Fig 8: affichage des calculs .

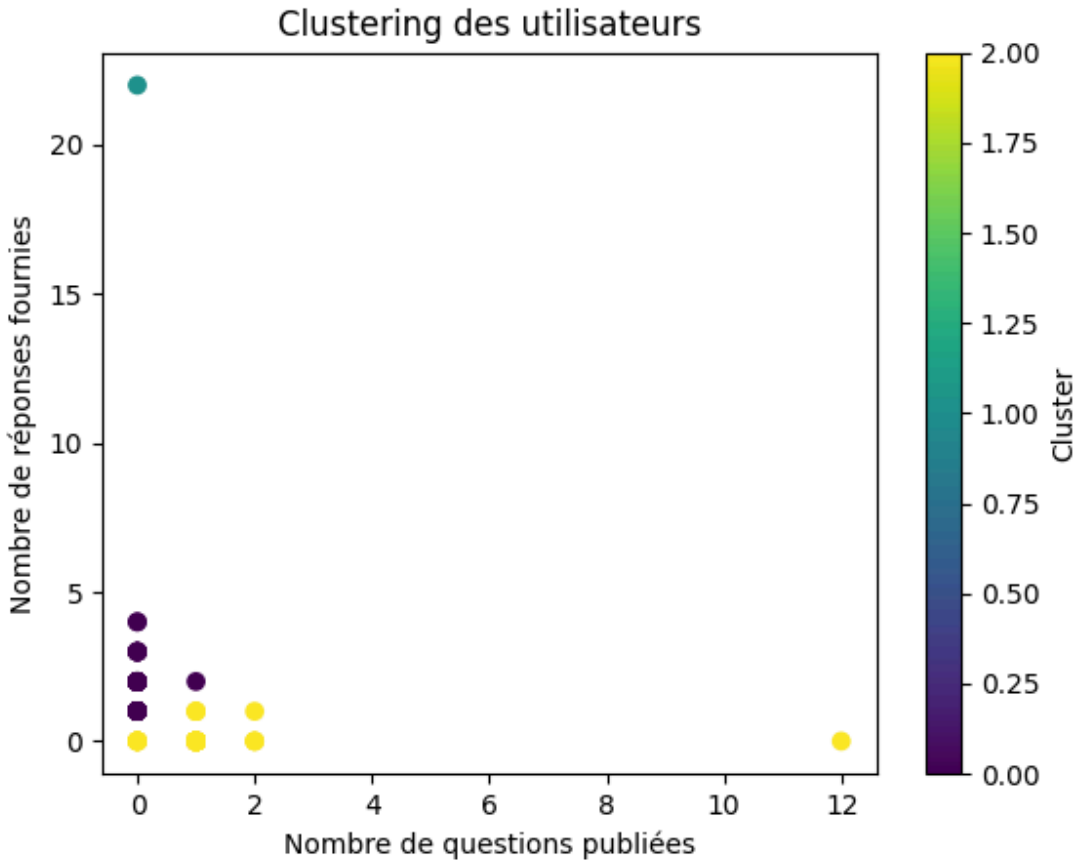


Fig 9: affichage des clusters.

En analysant les résultats du clustering, nous pouvons observer les caractéristiques suivantes :

1. Cluster 0 Ce cluster regroupe les utilisateurs qui ont une faible activité sur la plateforme. Ils ont publié peu de questions, fourni peu de réponses et écrit peu de commentaires. Ces utilisateurs pourraient être considérés comme passifs ou moins

actifs, peut-être simplement des observateurs de la communauté plutôt que des contributeurs actifs.

2. Cluster 1 Ce cluster est composé d'utilisateurs qui se distinguent par leur engagement dans la fourniture de réponses. Ils ont fourni un nombre relativement élevé de réponses par rapport aux autres actions, mais ils n'ont pas publié beaucoup de questions ni écrit beaucoup de commentaires. Ces utilisateurs pourraient être des contributeurs spécialisés dans l'apport de réponses aux questions des autres membres de la communauté.

3. Cluster 2 Les utilisateurs de ce cluster se distinguent par leur engagement équilibré dans différentes actions sur la plateforme. Ils ont publié un nombre modéré de questions et fourni un nombre modéré de réponses, mais ils ont écrit plus de commentaires par rapport aux autres clusters. Ces utilisateurs semblent être impliqués dans l'interaction avec d'autres membres de la communauté à travers des commentaires, ce qui suggère un niveau d'engagement plus élevé dans les discussions et les interactions sociales.