Département Informatique

Projet MINING INFORMATION FROM SOCIAL MEDIA NETWORKS DURING CRISIS EVENTS

Elhadj Mamadou SOW

Niveau d'étude
2ème année Master Informatique - IAFA_DC

Responsable du Module
Mme Lynda Tamine-Lechani

Année Universitaire
2024-2025

Plan du rapport

1. Introduction
    a. Contexte
    b. Objectifs
    c. Outils
2. Méthodologie
    a. Création du graphe
    b. Analyse de données
    c. Dashboard
    d. Générations des Embeddings
    e. Clustering des noeuds
    f. Système de Recommandation
3. Conclusion

I- Introduction

Dans un monde où les réseaux sociaux occupent une place importante dans nos vies, comprendre les interactions sociales et les structures sous-jacentes est essentiel. Ce projet a été réalisé pour analyser un graphe social complexe basé sur des données représentant des utilisateurs, des tweets, des hashtags, et des événements.

Contexte

Le projet s'inscrit dans le cadre de l'étude des graphes sociaux pour répondre à des questions telles que :

- Quels utilisateurs sont les plus influents ?
- Quels tweets ont un impact majeur ?
- Comment regrouper les nœuds ayant des propriétés similaires ?
- Comment recommander des tweets à un utilisateur spécifique ?

Objectifs

L'objectif principal de ce projet est de :

- Explorer la structure du graphe en analysant les nœuds et les relations.
- Générer des embeddings pour représenter les nœuds sous forme vectorielle.
- Cluster les nœuds pour identifier des groupes ayant des propriétés similaires.
- Mettre en place un système de recommandation de tweets basé sur des similarités calculées.

Outils

Pour atteindre ces objectifs, les outils suivants ont été utilisés :

- Python pour l'ensemble des analyses.
- NetworkX pour manipuler et analyser le graphe.
- Node2Vec pour générer des embeddings à partir des nœuds.
- Scikit-learn pour le clustering et les calculs de similarités.
- Matplotlib pour visualiser les résultats.

Pour atteindre ces objectifs, nous avons suivi une méthodologie rigoureuse décrite dans la section suivante.

II - Méthodologie

Nous avons choisi de travailler sur la partie 1 du projet, qui consiste à analyser le réseau social, pour les raisons suivantes :

- Accessibilité des données: Les données disponibles incluent une variété de nœuds et de relations, comme les utilisateurs, les tweets, les hashtags, et les relations associées comme POSTED, IS_ABOUT. Ces informations sont suffisantes pour analyser la structure du graphe.

- L'importance des analyses structurales: Comprendre les nœuds influents utilisateurs ayant de nombreux followers ou tweets retweetés et les relations (comme les mentions ou les hashtags) est crucial pour identifier les dynamiques sociales dans le graphe.

- Applications directes: Les résultats de l'analyse structurale peuvent être utilisés pour :
    - Identifier les utilisateurs influents (leaders d'opinion).
    - Recommander des tweets pertinents pour un utilisateur donné.
    - Mieux comprendre l'impact social de certains événements.

Ces analyses permettent d'extraire des informations significatives pour la compréhension globale des réseaux sociaux.

Création du graphe

Pour analyser les interactions sociales, nous avons construit un graphe dirigé où chaque nœud et chaque relation représente une entité ou une interaction spécifique. Le graphe a été créé à partir des fichiers JSON fournis.

Nœuds
Les types de nœuds présents dans le graphe sont :

- User : Représente un utilisateur des réseaux sociaux, identifié par son ID unique. Les propriétés incluent :
    - Nombre de followers.
    - Nombre de tweets postés.

- Tweet : Représente un tweet, identifié par son ID.
  Les propriétés incluent :
    - Le texte du tweet.
    - Le nombre de retweets.

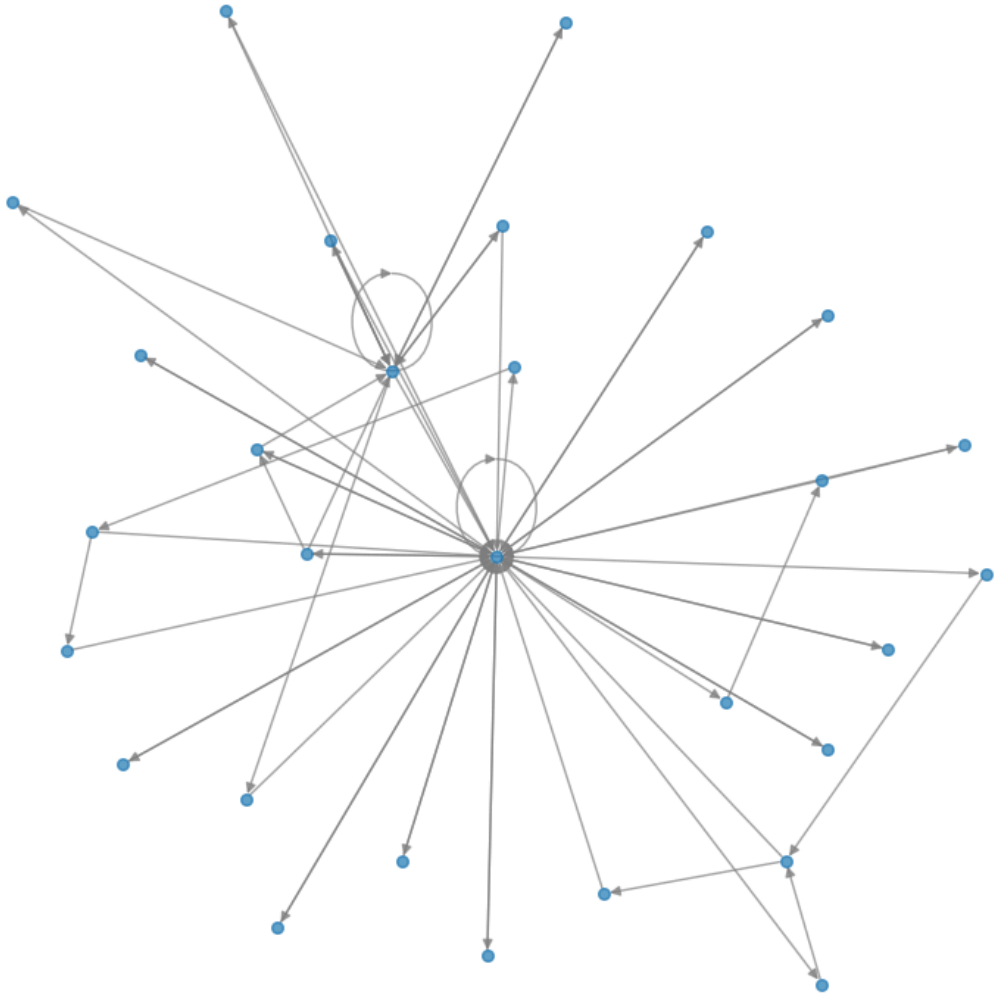
- Hashtag : Représente un hashtag utilisé dans les tweets, identifié par son ID.
- Event : Représente un événement spécifique (par exemple, une catastrophe naturelle).
- PostCategory : Représente la catégorie du tweet (par exemple, information, opinion).

Relations
Les types de relations ajoutées au graphe sont :
- IS_ABOUT : Relie un tweet à un événement.
- POSTED : Relie un utilisateur au tweet qu'il a posté.
- HAS_HASHTAG : Relie un tweet au hashtag utilisé.
- MENTIONS : Relie un tweet à un utilisateur mentionné.
- RETWEETED : Relie un utilisateur au tweet qu'il a retweeté.

Statistiques globales
Le graphe construit contient :
- 219247 nœuds.
- 304612 relations réparties sur plusieurs types.

Une visualisation du graphe global permet de comprendre la structure générale. Nous avons aussi exploré des sous-graphes pour des événements spécifiques ou des relations clés.

Sous-graphe fortement connecté



Analyse des données

Nous avons exploré les données pour mieux comprendre la structure du graphe et identifier les éléments clés.

Répartition des nœuds
Les nœuds sont répartis en plusieurs types, chacun ayant des propriétés spécifiques. Par exemple :

- Le graphe contient 43141 utilisateurs, 55986 tweets, et 10441 hashtags.
- Les utilisateurs les plus influents ont été identifiés en fonction de leur nombre de followers.

```
Seuil des influenceurs (99eme percentile) : 1206131.399999997
Nombre d'influenceurs : 432
       n.properties.id  n.properties.followers_count
59           454313925                       2552239
419             972651                       9375849
1124           1367531                      16159871
1730         170965705                       1432171
1749         202890266                       5619485
...                ...                           ...
42931          5577902                       2116204
42949         16389180                       3813438
42950         71297990                       1671906
42966         64038747                       6989214
42988         40076725                       5156169

[432 rows x 2 columns]
```

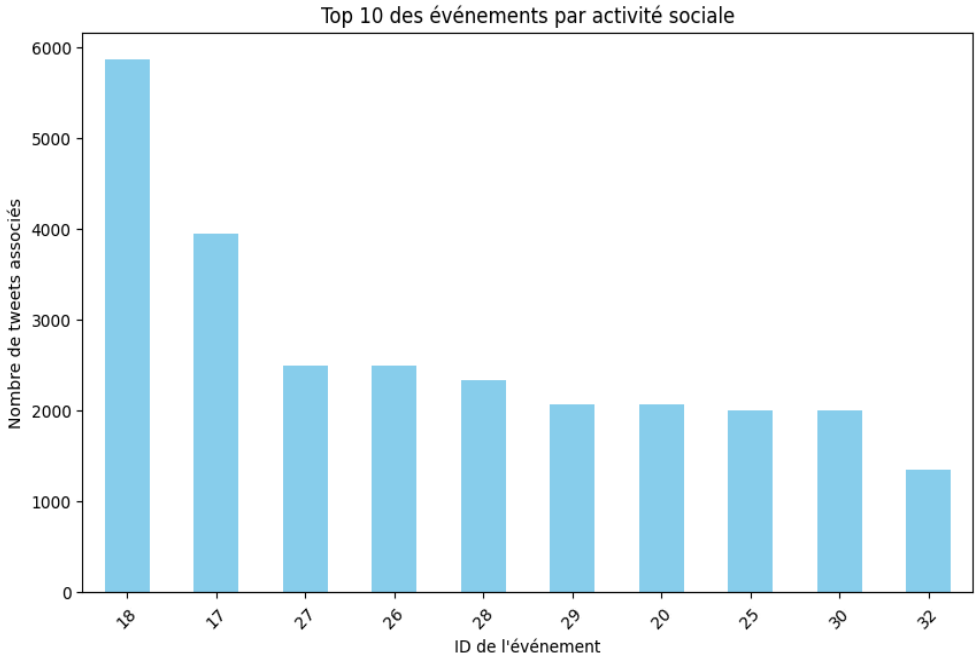
Relations fréquentes

Les relations les plus fréquentes dans le graphe sont :

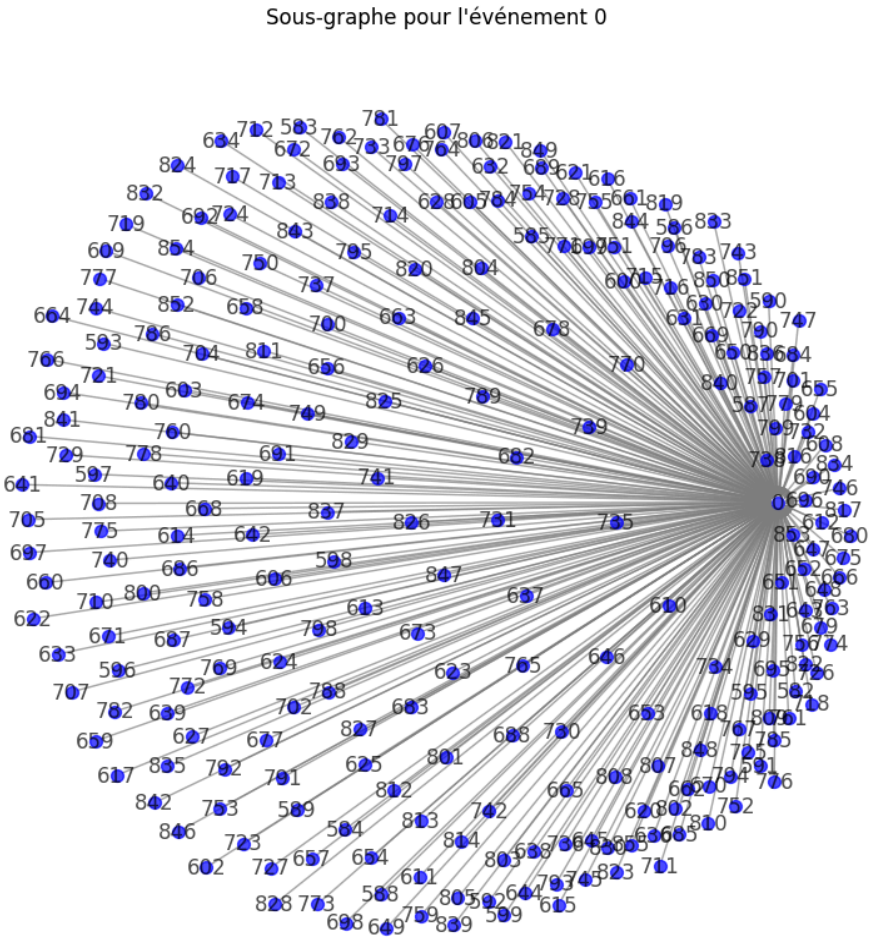
- POSTED: 55986 relations.
- HAS_HASHTAG: 96566 relations.
- IS_ABOUT: 36668 relations.

Nous avons fait une analyse détaillée des données en nous basant sur plusieurs visualisations générées à partir du graphe.

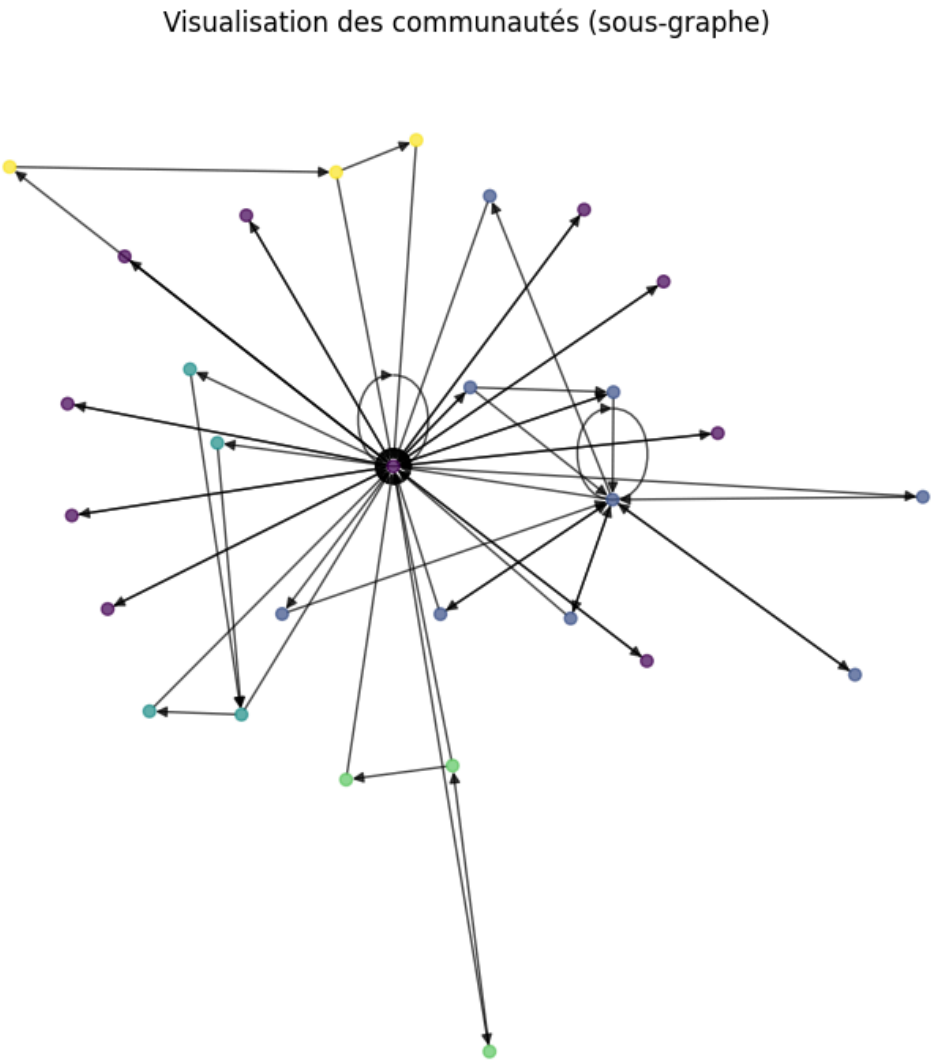
- Répartition des événements par activité sociale
  Ce graphique montre les 10 événements ayant généré le plus d'interactions sociales: tweets, mentions, etc... Cela nous permet d'identifier les événements les plus influents dans le réseau social étudié.



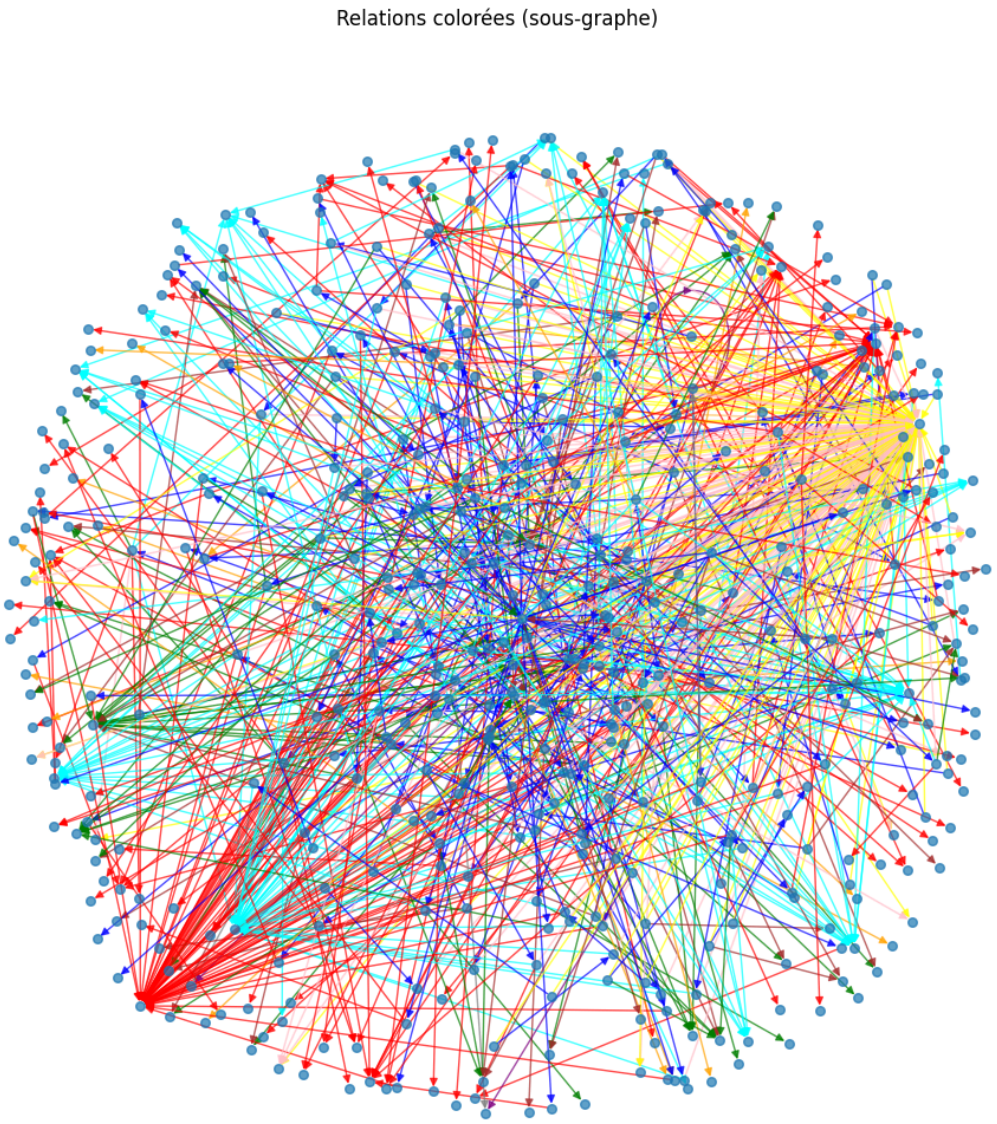
- Visualisation des sous-graphes
  La structure de ce sous-graphe met en évidence les interactions sociales pour l'événement 0, avec un focus sur les relations clés IS_ABOUT, HAS_HASHTAG, etc... Cela aide à comprendre comment les utilisateurs interagissent autour d'un événement spécifique.



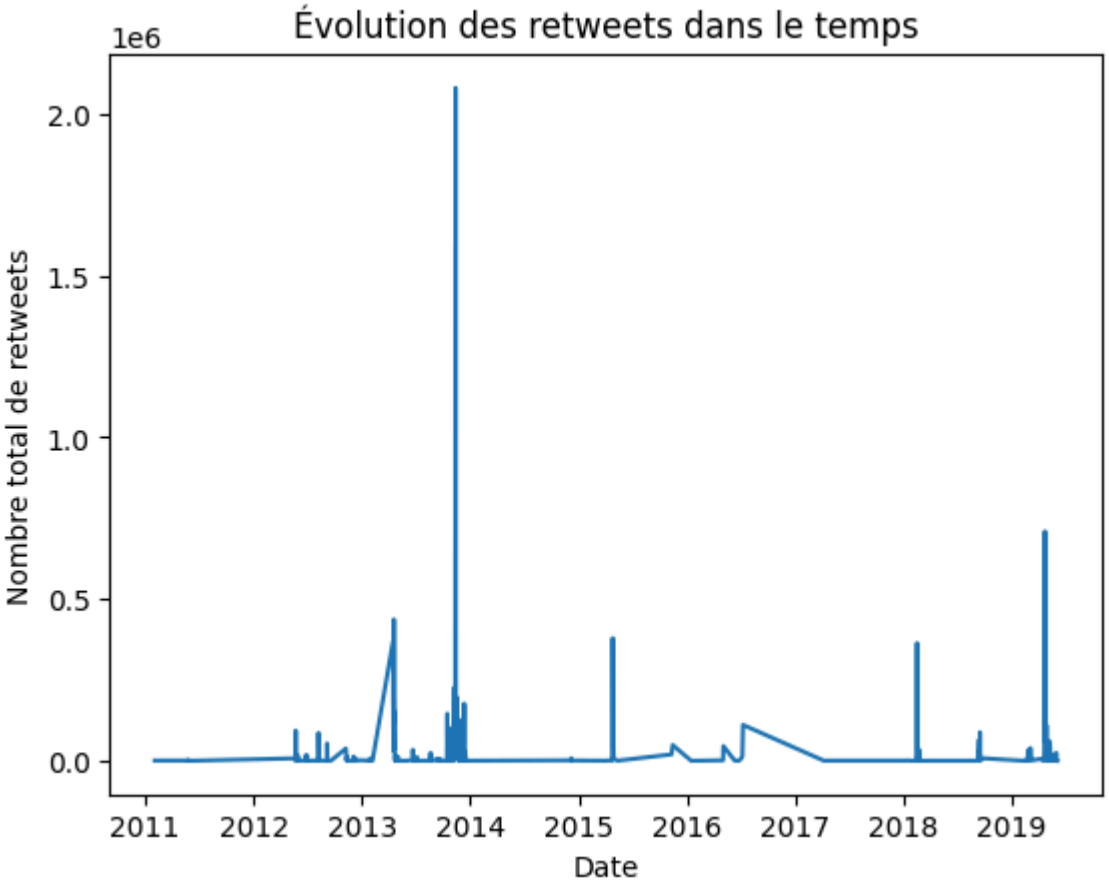
- Visualisation des communautés détectées
  Nous avons utilisé des algorithmes de détection de communautés pour identifier des groupes denses dans le graphe. Les couleurs des nœuds représentent les différentes communautés.



- Relations colorées dans le graphe
  Les relations clés du graphe (POSTED, MENTIONS, RETWEETED, etc.) sont représentées par des couleurs distinctes. Cela permet de mieux comprendre la distribution et l'importance relative de chaque type de relation.



- Évolution des retweets dans le temps
  Ce graphique illustre l'évolution du nombre total de retweets sur la période d'analyse. On remarque une augmentation significative de l'activité sociale durant certaines années, notamment en 2014, où un pic important est observé. Ces périodes pourraient correspondre à des événements majeurs ayant suscité un grand intérêt sur les réseaux sociaux.



Dashboard

Pour chaque type d'événement (exemple inondation, incendie), un tableau de bord a été conçu pour fournir des informations clés :

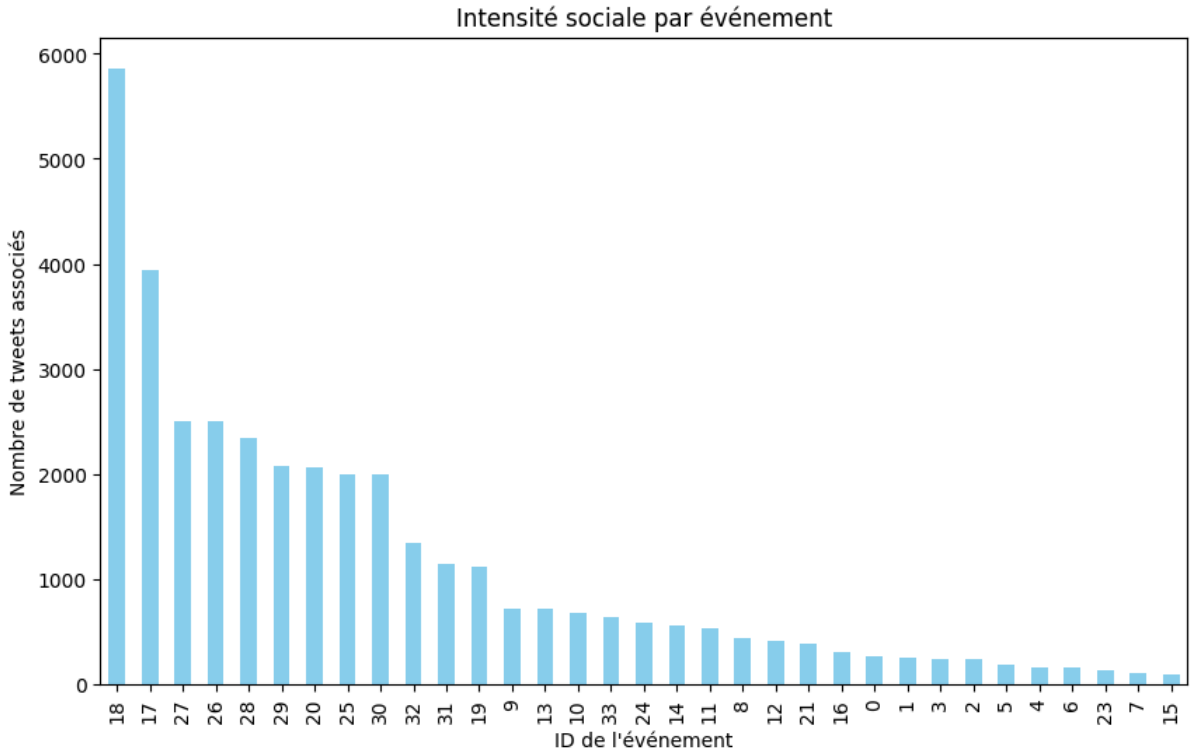
- Intensité de l'activité sociale :
    - Mesurée par le nombre de tweets associés à un événement spécifique.
    - Visualisation sous forme de graphiques en barres.

- Utilisateurs centraux :
    - Identifiés selon plusieurs critères :
        - Capacité à connecter d'autres utilisateurs.
        - Capacité à diffuser ou recueillir des informations.
    - Présentés dans un tableau avec les métriques correspondantes.

- Clustering des événements :
    - Les tweets ont été regroupés en clusters pour analyser les sous-graphes d'événements.

Méthodologie pour construire le dashboard

Le tableau de bord a été construit en utilisant les bibliothèques matplotlib et pandas pour visualiser et analyser les données. Les visualisations incluent :
- Graphique des tweets par événement.
- Répartition des clusters.
- Liste des utilisateurs influents.



Génération des embeddings

Pour représenter les nœuds sous forme vectorielle, nous avons utilisé la bibliothèque Node2Vec. Cette technique transforme les relations entre nœuds en vecteurs dans un espace de dimensions fixes, permettant ainsi d'analyser et de comparer les nœuds de manière mathématique.

Paramètres utilisés
Les paramètres pour Node2Vec sont les suivants :

- dimensions=32 : Chaque nœud est représenté par un vecteur de 32 dimensions.
- walk_length=20 : La longueur de chaque chemin aléatoire est de 20.
- num_walks=50 : Chaque nœud est exploré à travers 50 chemins aléatoires.
- workers=2 : Deux threads ont été utilisés pour accélérer la génération.

Résultat

Un total de **264 embeddings** a été généré pour un sous-graphe contenant des nœuds significatifs liés aux relations principales IS_ABOUT, POSTED, HAS_HASHTAG

```
# Extraire les nœuds et relations associés à l'événement
relations_of_interest = ['IS_ABOUT', 'POSTED', 'HAS_HASHTAG']
event_id = '0'

# Filtrer les arêtes correspondant à l'événement et aux relations d'intérêt
subgraph_edges = [
    (u, v) for u, v, d in G.edges(data=True)
    if d['relation'] in relations_of_interest and d.get('relation') == 'IS_ABOUT' and v == event_id
]

# Créer le sous-graphe basé sur les arêtes
subgraph = G.edge_subgraph(subgraph_edges)

print(f"Nombre de noeuds dans le sous-graphe : {subgraph.number_of_nodes()}")
print(f"Nombre de relations dans le sous-graphe : {subgraph.number_of_edges()}")
```

```
Nombre de noeuds dans le sous-graphe : 264
Nombre de relations dans le sous-graphe : 263
```

.

Utilité des embeddings
Les embeddings servent à :

- Regrouper les nœuds en clusters.
- Calculer des similarités pour les systèmes de recommandation.

Clustering des nœuds

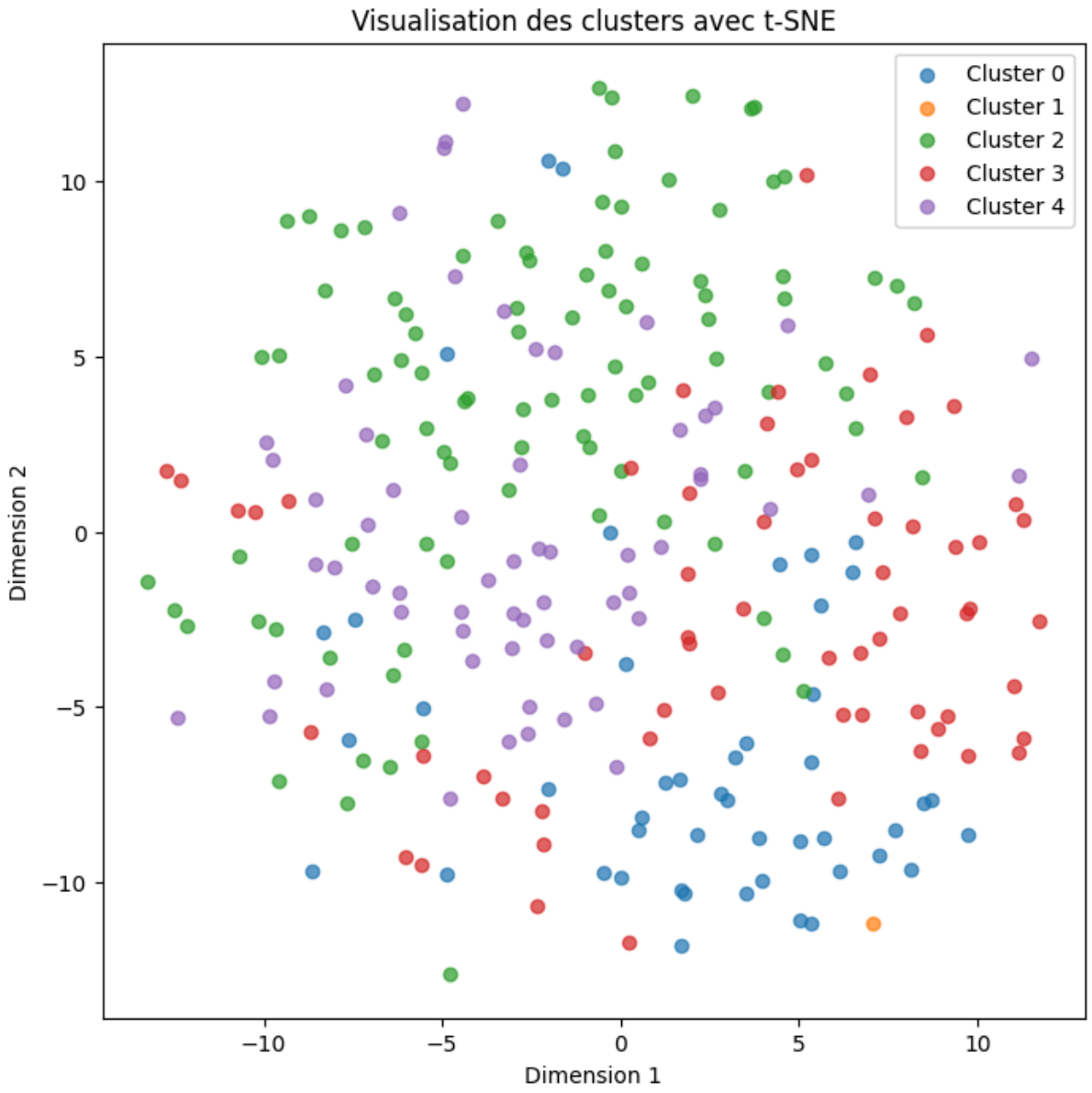
Une fois les embeddings générés, nous avons utilisé l'algorithme K-Means pour regrouper les nœuds en clusters. Les paramètres utilisés pour le clustering sont :

- n_clusters=5 : Les nœuds ont été regroupés en 5 clusters.
- random_state=42 : Pour assurer la reproductibilité des résultats.

Résultats obtenus

Le clustering a permis de regrouper les nœuds similaires en termes de connexions et de propriétés. Chaque cluster représente un groupe de nœuds ayant des propriétés similaires :
- Cluster 0 : Majoritairement des utilisateurs avec de nombreux followers.
- Cluster 1 : Tweets contenant des hashtags populaires.
- Cluster 2 : Nœuds représentant des événements.

Une visualisation des clusters montre la répartition des nœuds dans chaque groupe.



Système de recommandation

Après avoir généré les embeddings et regroupé les nœuds en clusters, nous avons conçu un système de recommandation de tweets. Ce système utilise la similarité cosinus entre les embeddings pour recommander des tweets à un utilisateur spécifique.

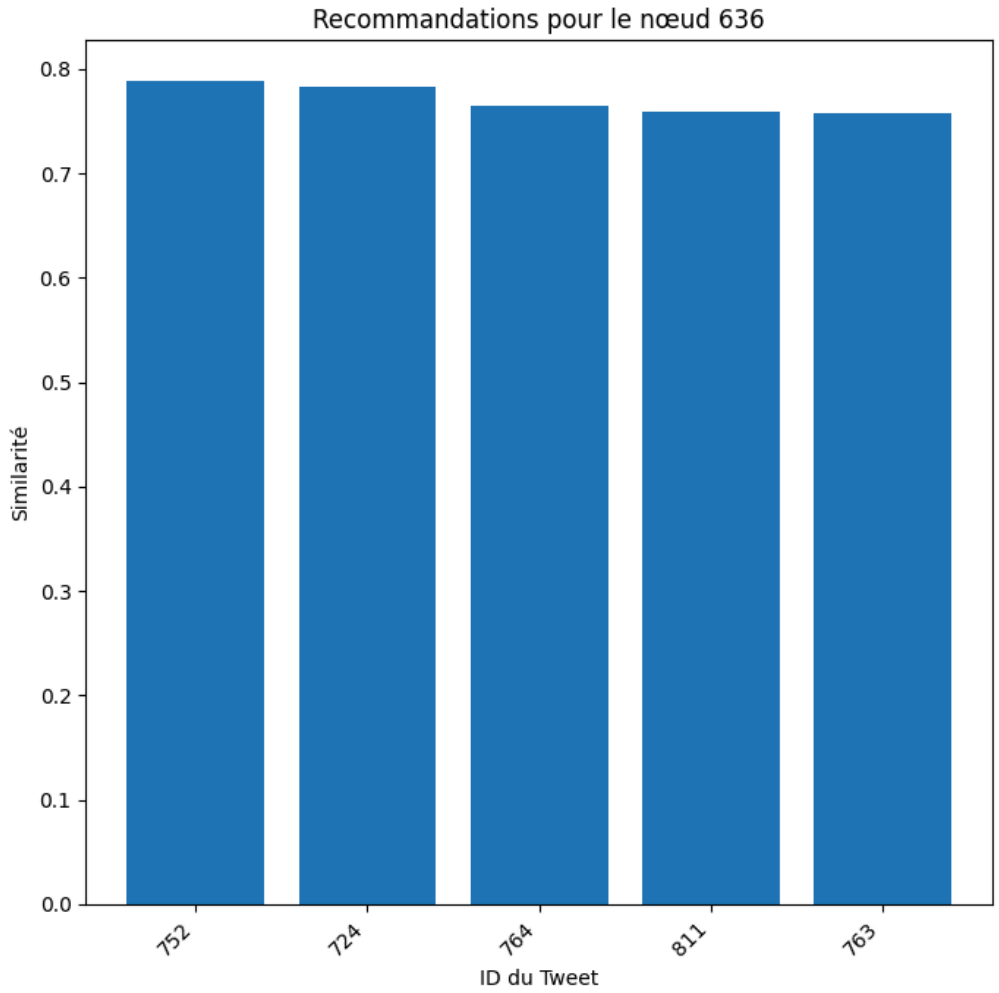
Méthodologie

- Sélection d'un utilisateur : Nous avons choisi un utilisateur dans le graphe (par exemple, User ID : 636).
- Calcul des similarités : La similarité cosinus a été calculée entre l'embedding de l'utilisateur et les embeddings des tweets.
- Classement des tweets : Les tweets les plus similaires à l'utilisateur ont été classés et affichés.

Résultat

Le système a recommandé les tweets les plus pertinents pour l'utilisateur sélectionné, basés sur leurs embeddings. Les résultats sont présentés sous la forme d'une liste avec les scores de similarité associés.

```
Résumé des clusters :
Cluster 2: 93 noeud
Cluster 3: 61 noeud
Cluster 4: 62 noeud
Cluster 0: 47 noeud
Cluster 1: 1 noeud
Recommandations pour le noeud 636:
Tweet ID : 752, Similarité : 0.7879
Tweet ID : 724, Similarité : 0.7831
Tweet ID : 764, Similarité : 0.7647
Tweet ID : 811, Similarité : 0.7586
Tweet ID : 763, Similarité : 0.7578
```



III- Conclusion

Ce projet a permis d'explorer et d'analyser un graphe social complexe. À travers les différentes étapes, nous avons pu :

- Construire un graphe contenant 219247 nœuds et 304612 relations.
- Identifier les utilisateurs influents et les relations clés.
- Générer des embeddings pour représenter les nœuds dans un espace vectoriel.
- Effectuer un clustering pour regrouper les nœuds similaires.
- Concevoir un système de recommandation de tweets basé sur la similarité cosinus.

Réussites

- L'analyse des données a permis de mieux comprendre les interactions sociales dans le graphe.
- Le système de recommandation a démontré l'utilité des embeddings pour des tâches pratiques.

Limites

- Le temps d'exécution pour certaines analyses (e.g., génération des embeddings pour de grands graphes) a été une contrainte.
- Les relations entre certains types de nœuds (e.g., événements) pourraient être explorées davantage.