

Challenge 1 - Find influencers to promote a restaurant in the Bay Area

I°/ Business understanding

Introduction :

Nous sommes une équipe de data scientist chez LinkedIn. Le département marketing souhaite organiser une campagne de marketing en ligne pour l'un de leur client, un restaurant situé dans la région de San Francisco. Le département marketing nous demande de trouver les 5 personnes les plus influentes sur le réseau qui seraient les mieux placées pour promouvoir le restaurant.

Qu'est ce qu'une campagne de marketing / communication ?

Une campagne de communication désigne l'ensemble des actions mises en œuvre par un organisme pour faire passer un message à un public ciblé. Les entreprises l'utilisent pour faire connaître leur marque et leurs produits ([Hubspot](#)).

Qu'est ce qu'un influenceur ?

Dans les réseaux sociaux, les influenceurs sont des utilisateurs qui possèdent une communauté, c'est-à-dire qu'ils sont au centre d'un groupe qui les suit et les écoute particulièrement. Les influenceurs génèrent de l'activité dans leur communauté respective. Il reste à savoir pourquoi les influenceurs fédèrent autant. (<https://theses.hal.science/tel-03640442/document>, P11).

Solution apportée

Par rapport à ces définitions et aux données dont nous disposons, nous allons, en tant que data scientists, tenter de cibler les 5 individus les plus influents du réseau pour faire la promotion du restaurant auprès d'un public ciblé. Autrement dit, cela consiste à déterminer les individus (ou sommets) les plus centraux du réseau. La centralité d'un sommet peut être mesurée de différentes manières, selon les objectifs de notre problème.

Dans notre contexte, nous pouvons d'abord identifier les personnes ayant un grand nombre de contacts directs (**centralité de degré**). Cela peut être utile si notre objectif est de maximiser la diffusion rapide de l'information.

Nous pourrions également tenter d'identifier les personnes qui ne sont pas seulement bien connectées, mais qui le sont aussi à d'autres individus "influent" (**centralité pagerank**). Cela garantit que l'information sur le restaurant atteint non seulement un large public, mais un public qualitatif, augmentant ainsi les chances d'attirer des clients potentiels.

Enfin, il serait intéressant aussi de détecter les individus qui agissent comme des points de passage (**betweenness centrality**) clés dans le réseau, facilitant ainsi la diffusion d'informations entre des groupes autrement non connectés. Ces individus peuvent être stratégiques pour atteindre des communautés isolées.

Traduction du problème métier en un problème data sciences

Nous allons représenter notre base de données sous forme de graphe pondéré tel que :

- **noeuds** : profils utilisateurs
- **arêtes** : connexion entre utilisateurs
- **attributs des noeuds** : localisation, employeur, université

- **poids** : (entre 1 et 3) représente le nombre d'attributs en commun entre deux utilisateurs connectés

Pour mesurer la centralité dans le réseau, nous regardons le degré de chaque nœud. Pour mesurer le nombre d'attributs en commun, nous utiliserons le poids des arêtes dans le graphe afin d'identifier les personnes qui ont des connexions avec un grand nombre d'attributs en commun.

II°/ Data Understanding

Nous disposons d'un ensemble de données comprenant un graphe G ainsi que trois dataframes : df_l, df_e et df_c. Le graphe représente les relations entre les individus, tandis que les dataframes fournissent des informations supplémentaires sur ces individus, telles que leurs emplacements, leurs employeurs et leurs universités.

Pour mieux comprendre la nature de nos données, nous avons entrepris une analyse approfondie, débutant par l'examen des premières lignes de chaque dataframe afin de saisir leur structure et leurs colonnes.

Structure du graphe :

- 811 nœuds
- 1597 arêtes
- non orienté
- non pondéré

Structures des dataframes :

- Identification des valeurs nulles
- Vérification de l'unicité des valeurs et voici les conclusions :
 - Aucune personne n'a été associée à deux emplacements différents
 - 11 individus ont fréquenté plus qu'une université
 - Plusieurs personnes ont changé leur lieu de travail au moins une fois
- Comptage des valeurs uniques :

employer		college	
university of illinois at urbana-champaign	76	university of illinois at urbana-champaign	66
google	15	shanghai jiao tong university	16
microsoft	15	bangladesh university of engineering and technology	10
university of texas at austin	5	tsinghua university	9
amazon	5	peking university	6
..
new jersey department of education	1	punjab engineering college	1
brookings institution	1	northern illinois university dekalb il	1
syrian emergency task force (setf)	1	indian institute of technology kharagpur	1
muslim public affairs council	1	chaitanya bharathi institute of technology.	1
yuhuan taijie hardware co. ltd.	1	huazhong university of science and technology	1
Name: count, Length: 723, dtype: int64		Name: count, Length: 109, dtype: int64	

figure 1,2 : occurrence de chaque employer et college dans les datasets df_e et df_c

```

location
urbana-champaign illinois area    92
greater chicago area              33
san francisco bay area            24
greater new york city area        16
greater boston area               16
..
iraq                              1
finland                          1
wichita kansas area               1
indianapolis indiana area         1
kuala lumpur malaysia            1
Name: count, Length: 89, dtype: int64

```

figure 3 : occurrence des locations dans df_l

VI°/ Data Preparation

Pour préparer nos données, nous avons mis à jour le graph en l'alimentant avec les attributs connus (location, employer, college) de la manière suivante :

```

def ajouter_attributs(graph, df_l, df_c, df_e):
    # Ajouter les attributs pour chaque nœud du graphe
    for node in graph.nodes():
        # Vérifier si le nœud a des données dans les dataframes
        if node in df_l['name'].values:
            graph.nodes[node]['location'] = df_l.loc[df_l['name'] == node, 'location'].values[0]
        if node in df_c['name'].values:
            graph.nodes[node]['college'] = df_c.loc[df_c['name'] == node, 'college'].values[0]
        if node in df_e['name'].values:
            graph.nodes[node]['employer'] = df_e.loc[df_e['name'] == node, 'employer'].values[0]

```

figure 4 : alimentation du graph G avec les attributs

V°/ Modeling

Le projet que nous entreprenons se décompose en deux étapes distinctes :

- Compléter les données manquantes relatives aux emplacements, aux employeurs et aux universités.
- Identifier les cinq influenceurs de la région de San Francisco afin de promouvoir le restaurant.

Etape 1 : Remplissage des données manquantes

Dans le cadre de notre problématique de régression visant à prédire des informations telles que les emplacements, les employeurs et les universités des individus pour lesquels nous disposons de données incomplètes, nous avons mis en œuvre deux approches distinctes : une approche fondée sur le principe d'[homophilie](#) et une approche basée sur l'[apprentissage automatique](#) utilisant l'algorithme Random Forest.

1°/ Approche basée sur l'homophilie :

Pour commencer, nous avons choisi d'exploiter les communautés de Louvain, une méthode de détection de communautés dans les réseaux complexes. Voici le déroulement de cette approche :

- Détection des communautés : Nous utilisons l'algorithme de Louvain pour détecter les communautés au sein de notre réseau d'individus.
- Attribution des nœuds à une communauté de Louvain : Chaque individu est assigné à une communauté en fonction de sa similarité avec les autres individus de cette communauté, selon le principe d'homophilie.
- Sélection des attributs (par exemple, l'emplacement) en fonction de leur fréquence d'occurrence dans la communauté : Nous identifions les attributs les plus pertinents pour chaque communauté en analysant leur fréquence dans celle-ci.
- Remplissage des valeurs manquantes : Une fois les attributs sélectionnés, nous comblons les valeurs manquantes en utilisant les informations disponibles au sein de chaque communauté, exploitant ainsi les similarités entre les individus.

Cette approche permet de tirer parti des structures communautaires présentes dans le réseau social des individus pour prédire les informations manquantes de manière plus précise et contextuelle.

Implémentation :

```
from networkx.algorithms import community

# Identifier les communautés à l'aide de l'algorithme Louvain
communities = community.greedy_modularity_communities(G)

# Attribuer les nœuds aux différentes communautés
community_membership = {}
for idx, comm in enumerate(communities):
    for node in comm:
        community_membership[node] = idx

community_membership
```

figure 5 : identification des communautés de Louvain

En résultat, notre méthode a identifié 20 communautés de Louvain distinctes dans notre ensemble de données de la manière suivante :

```
{'U27617': 0,
 'U27585': 0,
 'U4638': 0,
 'U27634': 0,
 'U27476': 0,
 'U27283': 0,
 'U27608': 0,
 'U7256': 0,
 'U11597': 0,
 'U16174': 0,
 'U16078': 0,
 'U27473': 0,
 'U2649': 0,
 'U11609': 0,
 'U27676': 0,
 'U27541': 0,
 'U27279': 0,
 'U25559': 0,
 'U27801': 0,
 'U27635': 0,
 'U4635': 0,
 'U27742': 0,
 'U28772': 0,
 'U4574': 0,
 'U4628': 0,
 ...
 'U24087': 20,
 'U24091': 20,
 'U2734': 20,
 'U24113': 20,
 'U24095': 20}
```

figure 6 : output

Après avoir attribué chaque nœud à une communauté, nous entamons la sélection des informations les plus fréquentes. À cet effet, nous avons créé les dataframes 'communities_df_l', 'communities_df_e' et 'communities_df_c'. Chacun de ces dataframes contient pour chaque nœud du graphe G, la communauté à laquelle il appartient ainsi que l'attribut le plus fréquent qui sera attribué après aux individus du dataframe 'empty'.

	name	community_index	freq_location
0	U11566	0	urbana-champaign illinois area
1	U27759	0	urbana-champaign illinois area
2	U16115	0	urbana-champaign illinois area
3	U4665	0	urbana-champaign illinois area
4	U27498	0	urbana-champaign illinois area
...
806	U24113	20	san francisco bay area
807	U24091	20	san francisco bay area
808	U2734	20	san francisco bay area
809	U24095	20	san francisco bay area
810	U24087	20	san francisco bay area
811 rows × 3 columns			

figure 7 : Exemple de communities_df_l

Ainsi, cette approche nous permet de récupérer toutes les informations requises.

	name	location
0	U27476	urbana-champaign illinois area
1	U4665	urbana-champaign illinois area
2	U14078	urbana-champaign illinois area
3	U9628	urbana-champaign illinois area
4	U9721	urbana-champaign illinois area
...
470	U14564	urbana-champaign illinois area
471	U14112	urbana-champaign illinois area
472	U4586	urbana-champaign illinois area
473	U18520	miami fort lauderdale area
474	U22044	greater chicago area

475 rows x 4 columns

Nombre d'occurrences de chaque localisation :

location	
urbana-champaign illinois area	293
greater chicago area	55
greater boston area	48
san francisco bay area	18
bangladesh	17
hyderabad area india	14
miami fort lauderdale area	13
china	11
greater new york city area	6

Name: count, dtype: int64

figure 8 : Nombre d'occurrences de chaque localisation pour les individus du dataframe 'empty'

Ainsi, nous avons réussi à identifier les employeurs et les universités de tous les individus du graphe. À ce stade, nous pouvons entamer notre recherche pour trouver les cinq influenceurs localisés dans 'bay San Fransisco' qui contribueront à promouvoir le restaurant.

2°/Approche 2 :

Dans cette méthodologie, nous avons développé un modèle basé sur l'apprentissage automatique appelé "Random Forest". Avant de construire ce modèle, nous avons effectué une modélisation du problème, qui s'est déroulée comme suit :

1. Nous avons extrait la matrice d'adjacence du graphe, une représentation de ses connexions entre les nœuds.
2. Chaque ligne de cette matrice représente une relation entre les nœuds et constitue une observation (ou échantillon) pour un nœud spécifique.
3. Nous avons ensuite attribué à chaque ligne d'adjacence une étiquette, qui correspond aux valeurs des caractéristiques "college", "employer" ou "location". Dans notre cas, puisque nous cherchions à prédire la localisation des individus (nœuds), les étiquettes se trouvent dans l'ensemble des valeurs prises par la caractéristique "location".
4. Enfin, avec chaque ligne d'adjacence (nœud) associée à une étiquette (localisation), nous avons formulé un problème bien défini pouvant être résolu en utilisant l'apprentissage supervisé à travers l'algorithme "Random Forest".

Cette approche nous a permis de capitaliser sur la structure des relations entre les nœuds du graphe, grâce à la matrice d'adjacence qui contient une grande partie des informations structurelles du graphe.

Etape 2 : Identification des influenceurs :

Cette partie vise à présenter les techniques d'identification des acteurs influents de notre graphe résidant dans la ville de San Francisco. En règle générale, dans un réseau, les personnes par qui l'information circule le plus souvent sont les plus influentes. En théorie des graphes, on parle de personnes centrales et on cherche donc à calculer la "centralité" des différentes personnes résidant à San Francisco.

Il existe de nombreuses mesures de la centralité, mais ici nous présenterons quatre qui sont le plus communément utilisés.

-Degree centrality

La centralité de degré attribue un score d'importance basé simplement sur le nombre de liens détenus par chaque nœud. Dans cette analyse, cela signifie que plus la centralité de degré d'un nœud est élevée, plus ce nœud est connecté à un plus grand nombre d'arêtes et donc à un plus grand nombre de nœuds voisins (amis LinkedIn). En fait, la centralité de degré d'un nœud est la fraction de nœuds auxquels il est connecté. En d'autres termes, c'est le pourcentage du réseau auquel le nœud particulier est connecté, ce qui signifie être ami avec.

Pour commencer, nous trouvons les nœuds ayant les centralités de degré les plus élevées. Plus précisément, les nœuds avec les 5 plus hautes centralités de degré vivant à San Francisco sont indiqués ci-dessous, avec leur centralité de degré correspondante :

```
('U8670', 0.056790123456790124, 'san francisco bay area')
('U15267', 0.03950617283950617, 'san francisco bay area')
('U24045', 0.02962962962962963, 'san francisco bay area')
('U4568', 0.01728395061728395, 'san francisco bay area')
('U27661', 0.011111111111111112, 'san francisco bay area')
```

figure 9 : degré de centralité des individus

-Closeness centrality

La centralité de proximité attribue à chaque nœud un score basé sur sa "proximité" avec tous les autres nœuds du réseau. Pour un nœud donné, sa centralité de proximité mesure la distance moyenne à tous les autres nœuds. En d'autres termes, plus la centralité de proximité d'un nœud est élevée, plus il est proche du centre du réseau.

La mesure de centralité de proximité est très importante pour la surveillance de la propagation de fausses informations (par exemple, les fausses nouvelles) ou de virus (par exemple, les liens malveillants qui prennent le contrôle du compte LinkedIn dans ce cas). Prenons l'exemple des fausses nouvelles. Si l'utilisateur avec la mesure de centralité de proximité la plus élevée commençait à propager des fausses informations (en partageant ou en créant une publication), tout le réseau serait désinformé le plus rapidement possible. Cependant, si un utilisateur avec une centralité de proximité très faible essayait la même chose, la propagation de la désinformation à l'ensemble du réseau serait beaucoup plus lente. Cela

s'explique par le fait que les fausses informations devraient d'abord atteindre un utilisateur avec une centralité de proximité élevée qui les propagerait à de nombreuses parties différentes du réseau.

Pour commencer, nous trouvons les nœuds ayant les centralités de closeness les plus élevées. Plus précisément, les nœuds avec les 5 plus hautes centralités de closeness vivant à San Francisco sont indiqués ci-dessous, avec leur centralité de degré correspondante :

```
('U4568', 0.23189235614085313, 'san francisco bay area')  
( 'U7202', 0.22959183673469388, 'san francisco bay area')  
( 'U24045', 0.22810475922275417, 'san francisco bay area')  
( 'U8670', 0.2162883845126836, 'san francisco bay area')  
( 'U2627', 0.2131578947368421, 'san francisco bay area')
```

figure 10

-Eingenvector centrality

La centralité de vecteur propre est la mesure qui montre à quel point un nœud est connecté à d'autres nœuds importants dans le réseau. Elle mesure l'influence d'un nœud en fonction de sa connectivité au sein du réseau et du nombre de liens que ses connexions ont, et ainsi de suite. Cette mesure peut identifier les nœuds ayant le plus d'influence sur l'ensemble du réseau. Une centralité de vecteur propre élevée signifie que le nœud est connecté à d'autres nœuds qui ont eux-mêmes des centralités de vecteur propre élevées. Dans cette analyse Facebook, cette mesure est associée à la capacité des utilisateurs à influencer l'ensemble du graphe, et donc les utilisateurs avec les centralités de vecteur propre les plus élevées sont les nœuds les plus importants dans ce réseau.

Pour commencer, nous trouvons les nœuds ayant les centralités de eigenvector les plus élevées. Plus précisément, les nœuds avec les 5 plus hautes centralités de eigenvector vivant à San Francisco sont indiqués ci-dessous, avec leur centralité de degré correspondante :

```
('U24045', 0.22040575754968936, 'san francisco bay area')  
( 'U4568', 0.09534485924225647, 'san francisco bay area')  
( 'U27661', 0.08269443436068134, 'san francisco bay area')  
( 'U24064', 0.06566753060774415, 'san francisco bay area')  
( 'U27460', 0.0638928512592222, 'san francisco bay area')
```

figure 11

-Betweenness centrality

La centralité d'intermédierité mesure le nombre de fois où un nœud se trouve sur le plus court chemin entre d'autres nœuds, ce qui signifie qu'il agit comme un pont. En détail, la centralité d'intermédierité d'un nœud est le pourcentage de tous les plus courts chemins entre deux nœuds (à l'exception du nœud lui-même), qui passent par ce nœud. Plus précisément, dans le graphe LinkedIn, cette mesure est associée à la capacité de l'utilisateur à influencer les autres. Un utilisateur avec une centralité d'intermédierité élevée agit comme un

pont vers de nombreux utilisateurs qui ne sont pas amis et a donc la capacité de les influencer en transmettant des informations (par exemple, en publiant quelque chose ou en partageant une publication) ou même en les connectant via le cercle de l'utilisateur (ce qui réduirait la centralité d'intermédiarité de l'utilisateur par la suite).

Pour commencer, nous trouvons les nœuds ayant les centralités de Betweenness les plus élevées. Plus précisément, les nœuds avec les 5 plus hautes centralités de betweenness vivant à San Francisco sont indiqués ci-dessous, avec leur centralité de degré correspondante :

```
('U8670', 0.1717989971049202, 'san francisco bay area')
('U4568', 0.12609941543882514, 'san francisco bay area')
('U14577', 0.10704019884151286, 'san francisco bay area')
('U7202', 0.0918985923399476, 'san francisco bay area')
('U15267', 0.05102962047337749, 'san francisco bay area')
```

figure 12

A la fin de cette partie nous sommes basés sur la fréquence des noeuds qui reviennent le plus pour dire que nos 5 influenceurs résidant dans la ville de San Francisco sont :

```
U8670
U15267
U24045
U4568
U27661
```

figure 13

IV°/ Evaluation

Pour évaluer le succès de notre solution, nous devons d'abord vérifier dans quelle mesure les données restaurées correspondent aux données réelles, notamment en termes de remplissage des valeurs manquantes. Ensuite, nous devons évaluer si les données que nous avons recueillies nous permettent d'identifier efficacement les cinq principaux influenceurs capables de promouvoir le restaurant sur LinkedIn.

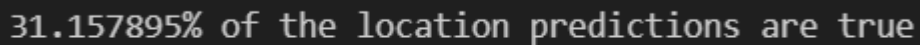
Comparaison de nos résultats avec ground truth :

Approche basée sur l'homophilie :

```
30.526316% of the location predictions are true
25.263158% of the employer predictions are true
24.421053% of the college predictions are true
```

figure 10 : Calcul des précisions en utilisant la méthode `evaluation_accuracy` présentée dans le notebook

Approche basée sur l'apprentissage automatique :



31.157895% of the location predictions are true

figure 11 : Calcul de la précision du modèle RF sur les locations

IIV°/ Conclusions

En conclusion, notre projet visant à identifier les cinq principaux influenceurs pour promouvoir un restaurant à San Francisco a abouti à des résultats significatifs. En utilisant une combinaison de techniques d'analyse de réseau et de traitement des données, nous avons pu cibler avec succès des individus clés au sein de la communauté de San Francisco.

Ce travail a permis d'identifier les influenceurs les plus pertinents en fonction de critères prédéfinis tels que leur degré de centralité, le betweenness, closeness. Ces influenceurs sont susceptibles de jouer un rôle crucial dans la promotion du restaurant en amplifiant sa visibilité et en suscitant l'intérêt de leur audience.

Cependant, malgré ces résultats, il reste des opportunités d'amélioration. Par exemple, dans la partie de remplissage des valeurs manquantes, essayer une méthode plus avancée pour détecter les communautés ou réformer le modèle de co-profiling mentionné sur l'étude.

En définitive, notre projet démontre le potentiel de l'analyse de réseau et de l'identification d'influenceurs pour renforcer la présence d'un restaurant dans une communauté locale. En continuant à explorer de nouvelles méthodes et à affiner nos approches, nous sommes confiants dans notre capacité à maximiser l'impact de la promotion du restaurant à San Francisco.