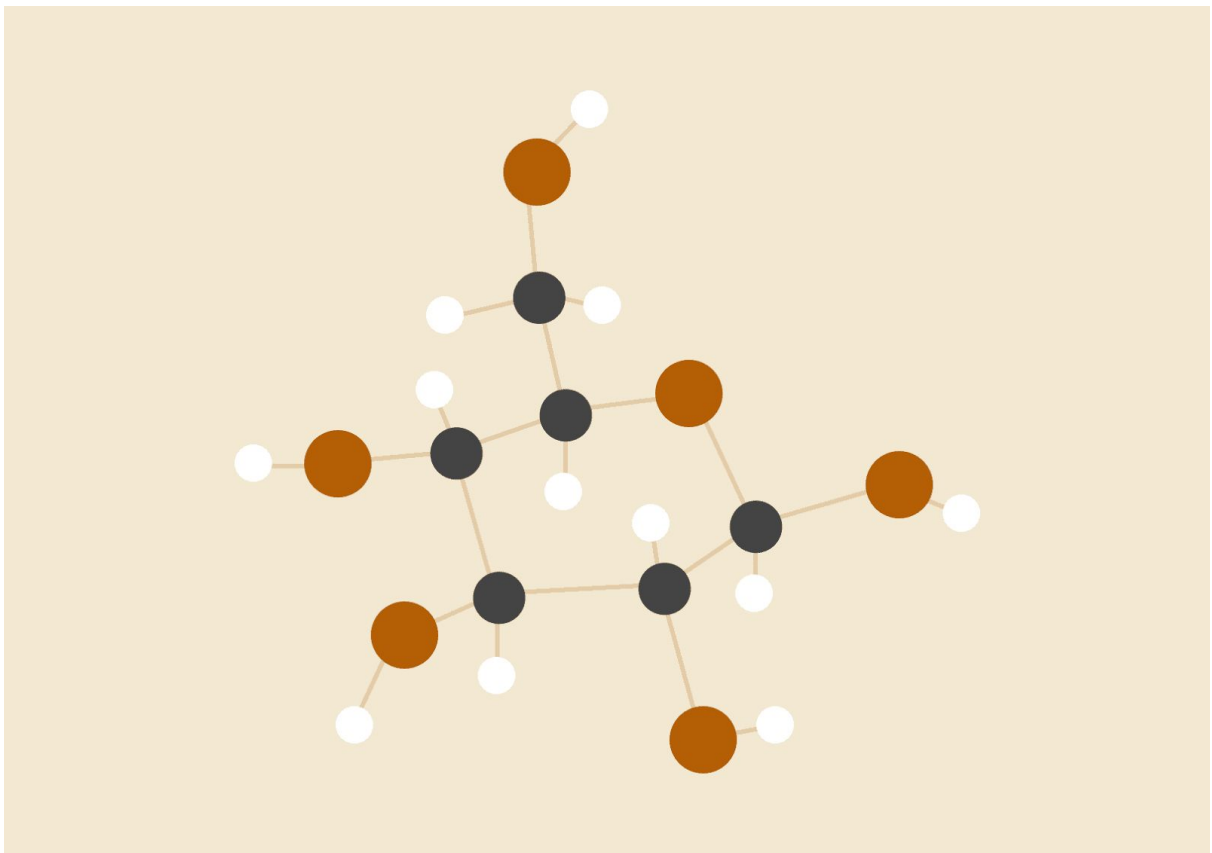


Projet industriel

BRGM: Analyse de tweets pour profiling via les réseaux sociaux



Sylvain courtin - Jimmy Vogel - Bryce Marbois

M2 INIS

Analyse des besoins	2
Contexte	3
Existant	4
Connaissances	5
Etat de l'art	5
Documentation sur les réseaux sociaux	5
Twitter	5
LinkedIn	7
Facebook	8
Analyse	10
Analyse des besoins d'implémentations	10
Choix du langage de programmation	11
Premières pistes de solutions	13
Résoudre le problème de scrapping	13
Google	13
Facebook	13
Linkedin	14
Résoudre le problème d'homonymie	16
Facebook	16
LinkedIn	16
Différencier une entreprise d'un utilisateur	17
Recoupement d'informations	17
Conclusion et planning	17
Références	18

I. Analyse des besoins

Problématique:

Identifier l'auteur d'un tweet à partir des informations fournis par twitter, notamment déterminer si c'est une entreprise ou un individu.

Déterminer l'entreprise affilié lorsqu'il s'agit d'un individu et des informations sur l'entreprise dans les deux cas.

Service:

Produire un programme répondant à la problématique, c'est-à-dire:

- prenant en entrée un format standardisée représentant un ou des tweets
- sortant une entreprise, ou/et un individu, auteur du tweet et ses informations.

Évoluer éventuellement la solution vers une interface utilisateur ergonomique.

Composantes techniques:

Les bibliothèques importées et le langage utilisé par la solution.

Les outils informatiques existants utilisés en internes dans la solution.

Contraintes et maintenance:

Il existe une contrainte légale; on ne s'autorise aucune incorporation des données récoltés en base de donnée durant la production de la solution.

Le langage de programmation doit être un langage standard tel que Java, Python, C#.

La récolte de donnée demande d'exploiter des données externes qui peuvent subir des mises à jour ou des modifications structurelles; il faut donc une documentation sur les points de lecture et comment mettre en place une éventuelle mise à jour.

II. Contexte

Un projet plus globale:

Ce travail d'analyse de tweets et de recherche identitaire s'incorpore dans un projet plus vaste du BRGM, l'écoute mondial d'échanges spécifiques au domaine géologique et plus particulièrement de la prospection minière.

L'idée de base de ce projet est d'utiliser le flux de données de twitter - les tweets qui viennent d'être postés - pour récolter ces échanges.

Ces données permettraient d'exploiter une nouvelle source d'information, complétant les outils de veilles existants.

Il s'agit aussi d'automatiser cette récolte, de faciliter l'accès à la donnée et de normaliser ces données pour implémenter des analyses statistiques & machine learning sur les résultats obtenus.

Filtrage:

La récolte de donnée via twitter est sous la forme d'un flux qui n'est que très peu paramétrable.

Il y a donc un besoin de filtrer les données pour n'obtenir que des informations spécifiques au domaine géologique.

Le filtrage pour certifier que les tweets appartiennent au domaine géologique n'est cependant pas suffisant car il faudra aussi analyser qui est l'auteur du message.

Incorporation du projet industriel:

Le but de ce projet industriel est donc de s'incorporer à cette recherche et de résoudre qui est l'auteur de l'action minière spécifié dans le tweet. Cela nécessite un processus d'identification de l'auteur, une affiliation à une entreprise si il s'agit d'une personne et une certification qu'il s'agit d'une entreprise de prospection minière et non de news dans le cas où l'entreprise est directement l'auteur du tweet.

III. Existant

Le projet plus global:

La résolution du filtrage des tweets pour certifier une appartenance à l'ontologie de prospection minière expliqué précédemment est déjà mis en place au démarrage de ce projet industriel.

Il existe donc déjà une base de donnée de tweets pertinents, filtrés par des algorithmes statistiques et/ou analysé par un expert du domaine géologique.

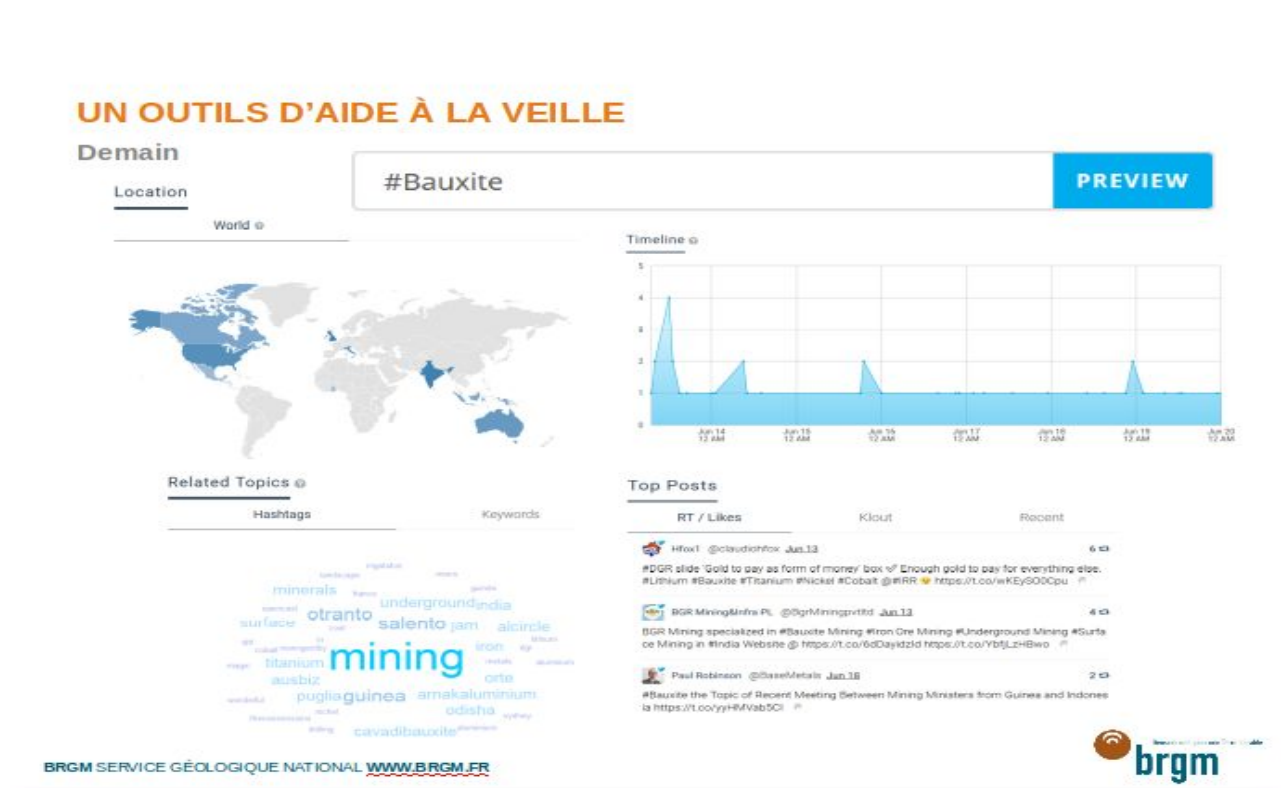
Cette base de donnée, sous format csv, contient une centaine de tweets considérés pertinents, et certains arguments propres à un tweet, tel que le nom de l'auteur du tweet et l'id correspondant.

Les idées existantes pour le QUI:

La résolution du QUI avait déjà soulevé quelques réflexions avant le démarrage du projet industriel et notamment l'idée d'utiliser les réseaux sociaux et plus précisément Facebook et LinkedIn pour trouver l'affiliation des individus. On pourrait donc relier les différents comptes de l'auteur d'un tweet pertinent et ainsi combler le manque d'information disponible sur l'auteur dans le tweet.

Interface ergonomique:

Une idée d'interface finale existait également:



IV. Connaissances

A. Etat de l'art

- Thème : connaissance pratique des réseaux sociaux Twitter et Facebook
- Programmation: connaissance de bases en python
- Data mining: connaissance de base en analyse sémantique

B. Documentation sur les réseaux sociaux

Une documentation sur Twitter, Facebook et LinkedIn, leurs vocabulaires et leurs apis:

a. Twitter

Description :

Twitter est un réseau social de microblogage géré par l'entreprise Twitter Inc. Il permet à un utilisateur d'envoyer gratuitement de brefs messages, appelés tweets, sur internet, par messagerie instantanée ou par SMS. Ces messages sont limités à 280 caractères.

Il existe deux manières de « retweeter » : soit comme initialement en copiant/collant intégralement le tweet lu en le précédant de la mention « RT @Bob », soit comme depuis fin 2009 en le « re-tweetant » automatiquement pour l'afficher à ses abonnés tel qu'on l'a vu soi-même.

Vocabulaire :

Les utilisateurs écrivent souvent un tweet en s'adressant à un utilisateur spécifique.

Ces tweets sont appelés "mention". La réponse et la mention contiennent @ suivi de l'id de l'user.

Chaque utilisateur peut consulter les mentions qu'il a reçues dans l'onglet @Notifications. Si un tweet débute par une mention, seuls les followers suivant le compte mentionné verront le tweet dans leur fil d'actualité.

Un mot précédé du signe # est un hashtag. Il s'agit d'un sujet attribué au message, Twitter peut afficher tous les tweets comportant un hashtag précis, et établit un classement des mots ou bien des hashtags du moment les plus utilisés.

Un utilisateur peut établir des listes parmi ses abonnements. Depuis fin 2009 il est possible de suivre une liste établie par un autre utilisateur. Il est également possible de rendre privées les listes (pour s'en servir comme répertoire de contacts d'un secteur, ou de concurrents, par exemple).

Les utilisateurs peuvent s'échanger des messages privés à travers des « messages directs », « MD » (« Direct Message » en anglais, abrégé « DM »). Cependant, on ne peut envoyer de DM que si l'on est abonné à un compte et que ce compte est lui-même abonné en retour (abonnements réciproques).

Après s'être connecté à Twitter, on accède aux tweets postés par ses propres abonnements, c'est-à-dire par les comptes d'utilisateurs que l'on a choisi de « suivre »(follower <-> followee).

Etudes générales :

Des études sur les réseaux sociaux très pertinentes existent et une en particulier pour Twitter est particulièrement intéressante, 'What is Twitter, a social Network or a News Media'*

Cette étude nous indique que Twitter contrairement aux autres réseaux sociaux montre un niveau de réciprocité plutôt bas: 77.9% des liens sont unilatéral et donc 22.1% sont à double sens.

De plus, 67.6% des utilisateurs ne sont pas suivis par aucun de leurs followees.

On peut en déduire que Twitter n'est pas utilisé comme un réseau social par la plupart des utilisateurs mais comme un réseau d'information.

Un tweet repris par des personnes ayant beaucoup de followers peut créer de nombreuses opinions sur le web et amener selon l'opinion sur un objet/un sujet amené à une grande publicité ou à l'inverse mettre à mal la réputation des entités qui lui sont liés.

API tweeter :

L'API REST 1.1 retourne des données historiques. La réponse à une requête ressemble plus ou moins à ce que l'on obtiendrait en tapant un mot clé dans le moteur de recherche de Twitter.

L'API REST permet de demander ou modifier des informations sur un compte user. Il n'y a pas de permission pour voir les informations, mais elle est nécessaire pour la modification. Il donne la permission par (OAuth) authentication.

Limite:Le nombre de requêtes est limité à 180 Requests demandes toutes les 15 min.

Par requête on peut demander un maximum de 100 tweets, soit un total de 18000 tweets / 15 mins, soit 1200 tweets / minute.

Une solution pour augmenter le nombre de requêtes est d'utiliser l'authentification par application. Elle permet jusqu'à 450 requête par 15 min et toujours 100 tweets au maximum. Soit 45000 tweets/15min, soit 3000 tweets / minute.

EXEMPLE EN PYTHON

```
import tweepy
auth = tweepy.AppAuthHandler(API_KEY, API_SECRET)
api = tweepy.API(auth, wait_on_rate_limit=True, wait_on_rate_limit_notify=True)
```

Ensuite un processus de pagination est nécessaire pour ne pas récupérer les mêmes 100 tweets. Possible en précisant 'max_id' et 'since_id' lesquels représentant l'id min et max pour un tweet. (voir tweepy/exemple pagination.py. 1)

API tweeter - STREAMING :

L'API Streaming fonctionne comme un enregistreur. L'API ne renvoie pas de données historiques, mais une fois la requête mise en place on récupérera l'ensemble des messages répondant à cette requête.

L'API Streaming délivre des tweets en se basant sur des termes ou des utilisateurs précis. Des informations sur l'auteur sont aussi accessible en temps réel. La permission de l'auteur n'est pas nécessaire. Vous devez être logué avec un compte tweeter pour utiliser le streaming.

La limite de l'API est difficilement atteignable, il faudrait atteindre un volume équivalent à 1% des messages publiés sur Twitter à un instant t .

b. LinkedIn



Description :

LinkedIn est un réseau social professionnel en ligne, LinkedIn fonctionne sur le principe de la connexion (pour entrer en contact avec un professionnel, il faut le connaître auparavant ou qu'une de nos connexions intervienne) et du réseautage (mise en relation professionnelle). Ainsi, il existe 3 degrés de connexions :

- le premier degré, ou nos contacts directs
- le deuxième degré, ou les contacts de nos contacts
- le troisième degré, ou les contacts de nos contacts de deuxième degré.

LinkedIn peut être utilisé pour tout ce qui concerne la vie professionnelle : trouver du travail, des employeurs, des prestataires, développer les affaires, etc.

L'utilisation du réseau LinkedIn d'un grand nombre de ses membres est assez réduite.

Pourtant, les données de LinkedIn montrent que le temps d'utilisation augmente avec le temps de souscription. En d'autres termes, plus on utilise LinkedIn et plus on va l'utiliser.

LinkedIn constitue aujourd'hui un moyen efficace pour construire, développer et enrichir son capital social. Il vise à créer une relation de confiance entre des professionnels, des étudiants et des entreprises afin que chacun puisse mobiliser ces ressources en ligne pour acquérir ou développer de nouvelles idées, obtenir des opportunités d'emploi, bénéficier des

communautés d'experts qui existent sur le réseau, faire du crowdsourcing (pour les entreprises principalement), etc.

Cependant, il appartient à chacun de construire son identité virtuelle en ligne et de gérer à sa guise son e-réputation (réputation en ligne).

Le renouvellement de l'abonnement peut s'annuler à tout moment, tout comme la suppression ou la désactivation d'un compte.

Particulièrement usité par les cabinets RH et les employeurs en recherche de profils d'exception, LinkedIn permet aux chercheurs d'emploi une visibilité auprès des "chasseurs de tête".

API LinkedIn :

LinkedIn autorise le libre accès à son API pour uniquement les 4 types d'utilisation ci-dessous :

- Pour permettre aux membres LinkedIn d'afficher leur identité professionnelle via leur profil LinkedIn en utilisant l'API de profil.
- Pour permettre aux membres de poster des recommandations directement à leur profil LinkedIn à l'aide des outils d'ajout au profil de l'API.
- Pour permettre aux membres de partager du contenu professionnel à leur réseau LinkedIn à partir du web en s'appuyant sur l'API de partage.
- Pour permettre aux entreprises de partager du contenu professionnel sur LinkedIn avec l' API Entreprise.

L'API mis à disposition par LinkedIn possède certaines limites, l'API répond très bien à la gestion de son propre profil mais ne permet pas, par exemple, d'effectuer du scraping ou de nombreuses recherches de profils (limité à 10 par jours avec une clé développeur classique).

En effet, pour tout autre accès et utilisation de l'API de LinkedIn, les développeurs tiers devront s'inscrire auparavant au programme de partenariat de LinkedIn, permettant ainsi d'étendre les limites d'utilisation de l'API.

c. Facebook



Description :

Facebook est un réseau social en ligne qui permet à ses utilisateurs de publier des images, des photos, des vidéos, des fichiers et documents, d'échanger des messages, joindre et créer des groupes et d'utiliser une variété d'applications.

Sur le mur Facebook sont affichés les notifications et événements publiés par le titulaire du compte ou profil Facebook, ainsi que les commentaires et messages des fans et amis.

Le profil Facebook est le lieu où chacun peut renseigner les informations qui le concernent (photo de profil, intérêts, écoles fréquentées, profession...). La Timeline permet de consulter

les publications, contenus et événements marquants publiés par vos amis ou par vous-même.

L'interface de publication : L'interface de publication du profil Facebook permet de publier :

- Du texte seul (un "statut")
- Un lien (vers une page Internet ou un vidéo)
- Une vidéo (téléchargée depuis votre disque dur ou publiée en live depuis votre mobile)
- Une photo seule
- Plusieurs photos réunies dans un même post

Les pages Facebook (anciennement "Pages fan" ou "Fanpages") sont des pages conçues pour permettre aux marques, aux entreprises, aux associations et aux artistes de créer une communauté et de communiquer auprès de leurs fans. Les pages sont dotées d'outils de publication pour engager la conversation avec vos abonnés, de rapports statistiques pour piloter vos actions et permet d'accéder à la plateforme de publicité Facebook pour amplifier vos messages.

Les pages Facebook permettent de publier en plus des posts classiques :

- Des offres
- Des événements
- Des moments-clé
- Des articles
- Des carrousels photo
- Des diaporamas
- Des canevas

Les groupes Facebook sont des pages dédiées à la création de communautés plus restreintes et privées (famille, groupes de travail, de passionnés) et dotés d'outils facilitant l'échange et la collaboration.

Les groupes Facebook permettent de publier en plus des posts classiques :

- Des sondages
- Des petites annonces
- Des fichiers Word, Excel...
- Des documents en ligne
- Événements

Les utilisateurs de Facebook mais aussi les pages et les groupes ont la possibilité de créer des événements au sein desquels ils peuvent inviter leurs amis, communiquer avec eux, organiser les préparatifs puis poster des photos, vidéos ou commentaires après qu'il ait eu lieu.

Apparu en 2009, le bouton "j'aime" permet d'exprimer son appréciation vis à vis d'un contenu. En 2016 un bouton "Reaction" vient enrichir les options, permettant d'exprimer 5 émotions prédéfinies : "Love", "Haha", "Wow", "Sad", "Angry".

Facebook propose aussi tout un ensemble de fonctionnalités, dont certaines pour mobile, comme Nearby Friends permettant de savoir si un amis est proche ou Facebook Hello permettant d'obtenir plus d'informations sur les personnes qui vous appellent en utilisant les données de Facebook.

API Facebook :

Le nom de l'API Graph s'inspire de l'idée d'un « graphe social » : une représentation des informations sur Facebook composée des éléments suivants :

- nœuds : représentent essentiellement des « éléments », comme un utilisateur, une photo, une Page ou un commentaire ;
- arêtes : représentent les liens entre ces « éléments », comme les photos d'une Page ou les commentaires d'une photo ;
- champs : informations relatives à ces « éléments », comme la date de naissance d'une personne ou le nom d'une Page.

Puisque l'API Graph est basée sur le protocole HTTP, elle est compatible avec tous les langages qui ont une bibliothèque HTTP, comme cURL ou urllib.

Lorsqu'une personne se connecte avec une application à l'aide de Facebook Login, l'application peut obtenir un token d'accès qui fournit un accès provisoire et sécurisé aux API Facebook.

Un token est une chaîne cryptée qui identifie l'application, comme une clef, lui permettant d'avoir accès à l'API. Il en existe plusieurs sorte mais tous contiennent leur date d'expiration ainsi que le nom de l'application en ayant fait la demande.

Les différents types de token sont:

- Token d'accès utilisateur
- Token d'accès d'app
- Token d'accès de page
- Token client

Le type de token que nous aurions pu utiliser dans notre situation aurait été un token d'accès de page.

Ces tokens d'accès donnent l'autorisation aux API pour lire, écrire ou modifier les données appartenant à une Page Facebook. Pour obtenir un token d'accès de Page, vous devez d'abord obtenir un token d'accès utilisateur et demander une autorisation.

Presque toutes les requêtes sont transmises à l'API sur graph.facebook.com. La seule exception concerne les téléchargements de vidéos qui utilisent graph-video.facebook.com.

Chaque nœud possède un ID unique qui vous permet d'y accéder par le biais de l'API Graph. En particulier, nous ne documentons pas la structure ou le format d'un ID de nœud/d'objet, car il est fort probable qu'ils changent au fil du temps et les apps ne doivent pas formuler d'hypothèses à partir de la structure actuelle.

V. Analyse

A. Analyse des besoins d'implémentations

- Le projet demande une gestion de fichier csv pour la récupération des tweets pertinents.
- Une connexion à des urls et un processus de scraping.
- Une analyse sémantique de certains arguments d'un tweet pour mettre en corrélation les paramètres des différents comptes de réseaux sociaux.

B. Choix du langage de programmation

Le langage python a été retenu en vue des besoins du projet car il :

- incorpore des librairies d'analyses sémantiques documentés
- répond à la contrainte d'un langage de programmation standard
- permet une implémentation rapide
- comprend une large documentation et de nombreuses librairies pour la résolution du scraping.

C. Analyse d'un tweet et des champs

Les deux fichiers csv contenant les tweets obtenus, l'un contenant 200000 tweets et l'autre contenant une centaine de tweets pertinents, ont été utilisés pour analyser et sélectionner les champs intéressants: la position et description de l'utilisateur texte du tweet et hashtags.

Analyse des champs user location et user description (des 155 tweets pertinents) :
(82,5 % entreprises)

Entreprises	Avec description	Sans description	Total
Avec localisation	92(72%)	2(1.5%)	94
Sans localisation	27(21%)	7(5%)	34

Total :	119	9	128
---------	-----	---	-----

(18,5 % personnes)

Personnes	Avec description	Sans description	Total
Avec localisation	18(66%)	3(1%)	21
Sans localisation	5(18.5%)	1(4%)	6
Total :	23	4	27

L'analyse précédente démontre que les tweets pertinents sont plus souvent liés à des entreprises (82.5%), ce qui facilitera la recherche car l'analyse par réseaux sociaux est plutôt complexe, il faudra cependant certifier que les entreprises ne sont pas des entreprises de news et si c'est le cas, possiblement faire une analyse du tweet pour trouver l'auteur de l'action liée à la prospection minière.

Lorsque la localisation n'est pas présente (18.5%), les arguments disponibles seront importants pour faire du matching et certifier que les comptes facebook et linkedin (si existant) trouvés par le nom de l'auteur du tweet, sont bien liés à ce compte Twitter.

Pour les 3.7% pour lesquels ni la description ni la localisation ne sont renseignées, ou autres cas d'échecs, une analyse des liens twitter de l'auteur ou de ces anciens tweets sont envisageables.

VI. Premières pistes de solutions

A. Résoudre le problème de scrapping

Le problème le plus complexe est lorsque l'auteur du tweet se trouve être une personne, nous avons alors besoin d'extraire des informations de manière automatique sur les pages des réseaux sociaux (scraping).

Le scraping est généralement bloqué ou tout moins limité car les informations personnelles sont généralement le cœur de métiers des sites que nous visitons. Ceci nous oblige à user de moyens détournés pour tout de même y accéder.

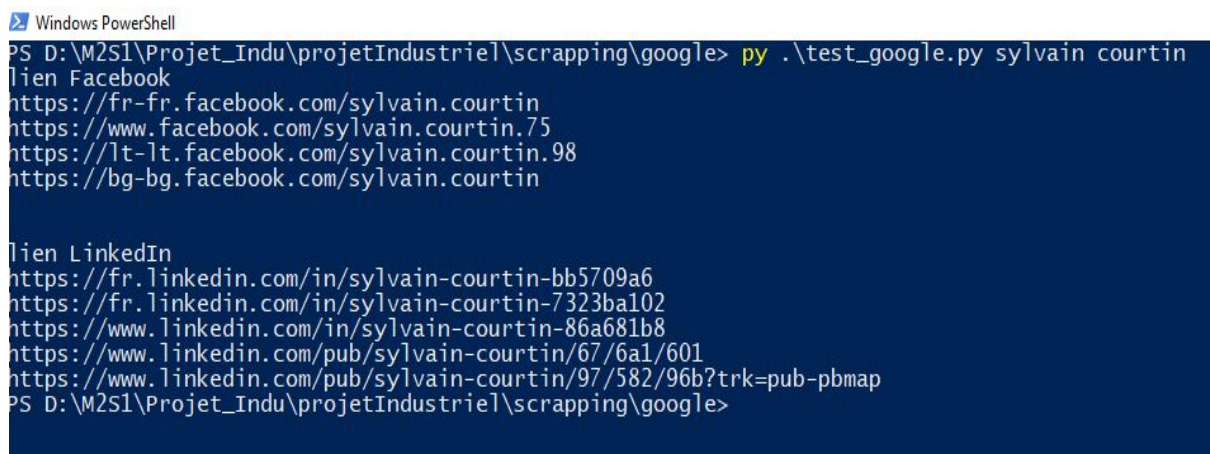
a. Google

La partie de notre travail suivant la récolte d'informations sur le tweet se déroule dans la recherche Google. En effet, ce moteur de recherche est particulièrement performant et nous nous en servons donc pour accéder directement aux pages de profils Facebook et LinkedIn. Pour pouvoir automatiser ces recherches, nous utilisons un script python exécutant une requête Google et renvoyant les sites correspondant à une expression régulière permettant de filtrer les pages privés de Facebook et LinkedIn.

Pour se faire le script prend en argument le nom, le prénom ainsi que des informations complémentaire comme la localisation ou d'autre précision que l'on voudrait apporter à la recherche google.

Puis le script utilise ces arguments transformé en chaîne de caractère pour faire deux recherches, l'une visant à trouver une page Facebook d'une personne homonyme à la personne recherché, l'autre étant pour trouver une page LinkedIn.

Les expressions régulières nous permettent d'exclure des résultats les sites autres, ainsi que les pages /public/ pour Facebook, et de nous concentrer sur les pages pertinentes en rapport avec un homonyme de la personne que nous recherchons.



```
PS D:\M2S1\Projet_Indu\projetIndustriel\scrapping\google> py .\test_google.py sylvain courtin
lien Facebook
https://fr-fr.facebook.com/sylvain.courtin
https://www.facebook.com/sylvain.courtin.75
https://lt-lt.facebook.com/sylvain.courtin.98
https://bg-bg.facebook.com/sylvain.courtin

lien LinkedIn
https://fr.linkedin.com/in/sylvain-courtin-bb5709a6
https://fr.linkedin.com/in/sylvain-courtin-7323ba102
https://www.linkedin.com/in/sylvain-courtin-86a681b8
https://www.linkedin.com/pub/sylvain-courtin/67/6a1/601
https://www.linkedin.com/pub/sylvain-courtin/97/582/96b?trk=pub-pbmap
PS D:\M2S1\Projet_Indu\projetIndustriel\scrapping\google>
```

b. Facebook

Le passage par une API pour Facebook avait été essayé initialement avant le début du projet industriel mais n'avait pas donné de résultat.

La solution a été de ne pas passer par l'Api mais directement par la library 'request' de python qui ne demande pas de processus d'identification.

On accède directement à un profil facebook par une url obtenue comme expliquer précédemment via Google. Il doit cependant exister une limite temporel ou quantitatif pour les requêtes sur un site comme facebook, qu'il faudra prendre en compte mais qui n'a pas encore été déterminé.

données récupérés via BeautifulSoup sur facebook (ecole,entreprise,favoris):

```
DONNEES METIERS ET ECOLES:
Scientific Games International : https://www.facebook.com/pages/Scientific-Games-International/317890511690434
Oberthur Jeux et Technologies aka Scientific Games : https://www.facebook.com/pages/Oberthur-Jeux-et-Technologies-aka-Scientifi
Lake Superior State University : https://www.facebook.com/pages/Lake-Superior-State-University/103798479658299
Lake Superior State University : https://www.facebook.com/pages/Lake-Superior-State-University/103798479658299
Algoma University : https://www.facebook.com/pages/Algoma-University/105526149480903
Korah Collegiate And Vocational School : https://www.facebook.com/pages/Korah-Collegiate-And-Vocational-School/114797728535917

DONNEES GEOGRAPHIQUES:
Cumming : https://www.facebook.com/pages/Cumming/104123022957689
Sault-Sainte-Marie (Ontario) : https://www.facebook.com/pages/Sault-Sainte-Marie-Ontario/106087732763236

Favoris:
Bob Seger : https://www.facebook.com/bobseger/
Spurs Live : https://www.facebook.com/sanantoniospurslive/
San Antonio Spurs : https://www.facebook.com/Spurs/
```

c. LinkedIn

Dans le cas de notre étude, il n'était pas possible d'utiliser l'api fourni par LinkedIn. Nous cherchons à scraper des informations sur des personnes, alors que l'api elle ne nous propose que la gestion de notre propre profil. Il a donc fallu se tourner vers une autre approche.

Pendant notre étude, nous avons pu remarquer que la recherche d'un utilisateur sur LinkedIn est assez limitée en tant qu'anonyme.

Dans la suite nous avons créé un compte LinkedIn pour nous aider dans notre développement.

Les applications pour LinkedIn utilisent 'selenium' de Python pour fonctionner, un driver permettant d'avoir un bot qui contrôle votre navigateur internet. Pour se faire l'application va en premier, se connecter à un compte LinkedIn créer spécialement pour notre application. Ensuite, directement avec le lien de la page, se rendre sur la page principal de la personne qu'on cherche à scraper. Et grâce à une librairie de Python, [BeautifulSoup](#), scraper les informations utiles comme les expériences professionnels ou les diplômes. Sans l'utilisation d'un compte LinkedIn et d'un bot, ils nous auraient été impossible d'effectuer du scraping sur la page d'une personne.

Scraping d'un profil LinkedIn avec l'utilisation d'un Bot via Selenium:



B. Résoudre le problème d'homonymie

a. Facebook

Pour répondre à cette problématique, nous utilisons le fait que Facebook propose une liste d'homonymes sur la page privée d'une personne.

Ainsi nous pouvons trouver une page Facebook privée d'une personne via une requête sur le moteur de recherche Google.

Puis une fois sur cette page privée, celle-ci nous offrant une liste aléatoire, on peut récupérer tous les homonymes par multiple requêtes.





```
PERSONNES MEME PRENOMS/NOMS:  
Frank Candido : https://pt-br.facebook.com/frank.candido.35  
Frank Manuel Candido Kick : https://pt-pt.facebook.com/people/Frank-Manuel-Candido-Kick/100008449993867  
Frank Sampayo Candido : https://es-es.facebook.com/francelin.sampayocandido  
Frank Reinaldo Candido : https://pt-br.facebook.com/frank.reinaldocandido  
Francisco Candido Frank : https://pt-pt.facebook.com/francisco.candidofrank.714  
Frank Victor Candido Cândido : https://pt-br.facebook.com/people/Frank-Victor-Candido-Cândido/100012656107211  
Frank Candido Peixoto : https://pt-br.facebook.com/frank.candido.7524  
Candido Frank : https://www.facebook.com/candido.frank
```

b. LinkedIn

On utilise encore 'selenium' et un bot pour pouvoir accéder à LinkedIn.

Cette fois on utilise une url qui mène vers une page de recherche. Dans cette url, on insère un nom, un prénom et éventuellement une école et/ou une entreprise. On extrait ensuite la liste résultant via 'BeautifulSoup'(library python) .

<https://www.linkedin.com/search/results/people/?firstName=sylvain&lastName=courtin>

 Sylvain Courtin Responsable encres chez Siegwerk Région de La Rochelle, France	Se connecter	MOTS CLÉS Prénom <input type="text" value="sylvain"/> Nom <input type="text" value="courtin"/> Titre <input type="text"/> Entreprise <input type="text"/> École <input type="text" value="orléans"/>
 sylvain courtin Magnétiseur.Energéticien Région de Lille, France	Se connecter	
 SYLVAIN COURTIN Magnétiseur Energéticien Région de Lille, France	Se connecter	
 Sylvain COURTIN Masseur-kinésithérapeute chez Libéral Région de Tours, France	Se connecter	

C. Différencier une entreprise d'un utilisateur

Déterminer si on est dans le cas où l'auteur du tweet est une entreprise ou un individu se fait par l'analyse du champ `user_name` fournit par Twitter pendant l'obtention des tweets. Cette différenciation en amont est importante car elle permet de ne pas attendre un échec sur une recherche d'un compte de type individu sur facebook ou/et linkedin (compte individu != compte entreprise) pour déterminer si le champ exprime l'auteur comme étant une personne ou une entreprise.

Elle a été réalisée à l'aide d'une base de donnée de prénoms

(<https://www.data.gouv.fr/fr/datasets/liste-de-prenoms/>) , et d'une liste de termes retrouvés généralement dans les noms d'entreprises, créée et à développer lors de ce projet.

(le résultat a été utilisé dans l'analyse des tweets précédemment)

D. Recoupement d'informations

Dans la suite du projet, nous espérons utiliser les informations récolté tout au long du processus (text, hashtags, location, description, favoris, écoles, entreprises) pour faire des corrélations (library nltk envisagé) et pouvoir relier les comptes de l'auteur du tweet.

VII. Conclusion et planning

On a réussi à passer outre les protections anti-scraping pour récupérer les données, et on a trouvé les champs de corrélations, ce qui va nous permettre de passer à la partie sémantique.

Il faut ensuite:

- Implémenter les algorithmes de matchings sémantiques,
- Analyser les échecs de matching avec les données obtenues à présent et voir si d'autres matching sont possibles avec d'autres données,
- Créer un script pour faire une recherche plus poussé sur une entreprise, son nom complet et sa localisation
- Vérifier si le tweet était un tweet de news et si oui, analyser le texte du tweet pour une analyse sémantique pour obtenir plus d'informations sur l'auteur de l'action liée à l'ontologie minière.
- Tester la surcharge des requêtes (voir un possible changement d'IP),
- Mettre en place un script lançant le processus complet,
- Documenter les accès externes pour les cas de mise à jour des réseaux sociaux,

VIII. Références

- BRGM, diaporama conférence projet innovation, Vincent labbe
- What is Twitter, a Social Network or a News Media
author = Haewoon Kwak Changhyun Lee, Hosung Park, and Sue Moon
OPTnote = <https://an.kaist.ac.kr/~haewoon/papers/2010-www-twitter.pdf>
- Etude API LinkedIn : http://www.xavierdupre.fr/blog/2013-09-28_nojs.html
- API Graph : <https://developers.facebook.com/docs/graph-api>
Outils:
- selenium : <http://www.seleniumhq.org/>
- nltk: <http://www.nltk.org/>
- bs4 : [Beautiful Soup](#)
- request: <http://docs.python-requests.org/en/master/user/quickstart/>