

Université de Paris

UFR Maths-Info

Rapport de Projet Pluridisciplinaire

Intitulé du projet :

Logiciel d'e-réputation et d'aide à la décision : application au marketing

Encadrants :

- Séverine AFFELDT
- Lazhar LABIOD

Étudiants :

- Hacene ISSELNANE
- Ayale HADDAD

Sommaire

Introduction.....	3
1. Contexte et motivations	4
1.1. Réalisations attendues.....	5
1.2. Description des tâches techniques.....	5
1.3. Planification	5
1.4. Déroulement général des événements.....	6
1.5. Répartition des tâches.....	6
2. Solutions existantes.....	7
2.1. Brand24.....	7
2.2. Mention.....	7
2.3. HubSpot.....	9
2.4. Bilan de l'analyse comparative	9
3. Solution proposée.....	10
3.1. Statistiques	11
3.2. Description des fonctionnalités.....	11
3.2.1. Extraction de tweets.....	11
3.2.2. Volumétrie des tweets	12
3.2.3. Histogramme des fréquences	12
3.2.4. Répartition par pays.....	13
3.2.5. Top utilisateurs	13
3.3. Analyses	14
3.4. Description des fonctionnalités.....	15
3.4.1. Répartition des sentiments	15
3.4.2. Nuage de mots	16
3.4.3. Co-clustering (topic modeling)	16
3.4.4. Représentation en graphe	17
3.5. Recommandations	18
4. Technologies utilisées.....	19
5. Conclusion.....	20

Table des figures

Figure 1 - Exemple de Dashboard e-réputation	4
Figure 2 - Exemple de dashboard proposé par Brand24.....	7
Figure 3 - Exemple de dashboard proposé par Mention	8
Figure 4 - Exemple de Dashboard proposé par Hubspot	9
Figure 5 - Dashboard e-réputation proposé.....	10
Figure 6 - Fenêtre de lancement de la pipeline d'extraction de tweets.....	11
Figure 7 - Graphique représentant la volumétrie des tweets.....	12
Figure 8 - Histogramme des termes les plus fréquents	12
Figure 9 - Distribution des tweets dans le monde.....	13
Figure 10 - Tableau représentant les utilisateurs les plus influents.....	13
Figure 11 - Onglet Analyse.....	14
Figure 12 - Graphique représentant l'analyse de sentiment	15
Figure 13 - Graphique représentant le nuage de mots	16
Figure 14 - Topic modeling avec Co-clust	16
Figure 15 - Graph représentant la similarité des termes d'un cluster.....	17

Introduction

Aujourd'hui, l'avènement de l'intelligence artificielle, en particulier l'apprentissage automatique et l'apprentissage profond, ouvre un nouvel horizon pour la résolution de problèmes multifactoriels à grande échelle.

L'apprentissage automatique est encore une technologie émergente mais polyvalente, qui est par nature théoriquement capable d'accélérer le rythme de l'automatisation et de s'auto-apprendre. Combiné à l'émergence de nouveaux moyens de production, de stockage et de circulation de la donnée. L'apprentissage a le potentiel de révolutionner la technologie et la société (comme l'ont fait la machine à vapeur et l'électricité, puis le pétrole et les ordinateurs lors des précédentes révolutions industrielles). L'apprentissage automatique pourrait générer des innovations et des capacités inattendues. À l'ère de la science des données, il est plus que jamais nécessaire de relever plusieurs défis grâce aux techniques d'apprentissage automatique.

Dans ce contexte, et dans le cadre d'un projet pluridisciplinaire, nous avons travaillé sur un sujet lié à la réputation numérique (ou e-réputation).

Pour ce projet, il est demandé d'élaborer une interface permettant la présentation des résultats d'analyse de e-réputation, à partir des informations disponibles en ligne, pour les différents produits d'une même marque. Cette interface devra également fournir des recommandations pour l'entreprise afin que celle-ci mette sur le marché des produits vendeurs.

Le lecteur trouvera dans ce rapport tous les éléments qui composent notre approche pour résoudre le problème.

1.Contexte et motivations

La e-réputation correspond à la notoriété numérique d'une personne, d'une entreprise ou d'une marque. Elle prend aujourd'hui une place importante dans notre société qui exploite toujours plus les nouvelles technologies. La e-réputation peut être analysée à partir de plusieurs sources (*scraping* d'articles de presse) et via les nombreux échanges que l'on peut trouver sur les différents réseaux sociaux (tweeter, facebook). La donnée ici exploitée est principalement le texte.

La gestion, le suivi et l'optimisation d'une e-réputation sont devenus très importants. En effet, le web occupe une place de plus en plus importante dans la vie quotidienne, que ce soit dans le cadre privé ou professionnel. Par exemple, le recrutement et l'employabilité sont étroitement liés à cette image en ligne.

De cette viralité de l'information, qui peut dégrader ou, a contrario, améliorer l'identité numérique d'une marque ou d'une personne, naissent certains concepts et expressions comme l'effet Streisand.

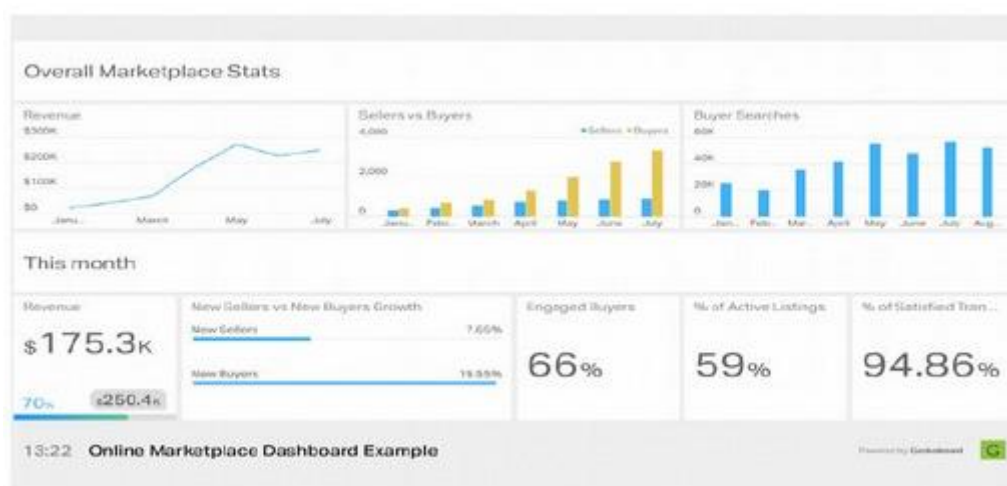


Figure 1 - Exemple de Dashboard e-réputation

1.1. Réalisations attendues

La première réalisation attendue est l'identification des sources à analyser pour cet outil ainsi que les méthodes d'analyse de texte.

- Extraction des tweets
- Nettoyage et préparation des tweets
- Exploration et analyse des tweets en utilisant des approches de visualisation, "Topics modeling" (LDA, NMF), et d'analyse de sentiment
- Implémentation des méthodes de nettoyage et d'analyse des tweets et intégration à l'interface web

La seconde réalisation est la mise au point d'un logiciel qui permettra l'exploitation de l'analyse et des recommandations dans le cadre d'une étude de marché.

1.2. Description des tâches techniques

- Identification des indicateurs et des sources les plus importants
- Identification des recommandations d'intérêt
- Implémentation d'un outil d'analyse de textes
- Design d'une interface de visualisation des résultats

1.3. Planification

Le projet s'est déroulé sur une période de 8 semaines, qui a été divisé en trois phases distinctes organisées comme suit :

Phase 1 : Conception générale du projet

Cette phase a commencé dès la première semaine du processus de planification et s'est poursuivie jusqu'à la deuxième semaine, au cours de laquelle diverses réunions ont eu lieu afin de cadrer le projet et valider le besoin client (utilisateur).

Phase 2 : Développement

Cette phase s'est déroulée en aval de la phase de conception, elle constitue la période allouée au développement du projet lui-même, l'écriture du code source ainsi que l'implémentation des fonctionnalités discutées.

Phase 3 : Intégration et réception

Ayant eu lieu dans les deux dernières semaines restantes du projet. C'est durant cette période qu'a eu lieu l'intégration c'est-à-dire le rassemblement de tous les fichiers sources édités par les développeurs.

1.4. Déroulement général des événements

Durant ces huit semaines de travail laborieux et méticuleux, nous avons su nous organiser et garder un esprit d'équipe en nous adaptant à chacun afin de réussir au mieux le projet qui nous a été confié.

La phase de conception du projet a commencé par la compréhension des objectifs attendus. Les encadrants nous ont expliqué en détail leurs principales attentes et ont partagé avec nous leurs expériences en nous guidant tout au long du processus de conception.

Des réunions hebdomadaires ont été mises en place pour suivre l'avancement du projet.

1.5. Répartition des tâches

Grâce aux nombreuses réunions de travail, nous avons pu définir une répartition des tâches permettant à chacun de contribuer au mieux à la bonne réalisation du projet. Si bien que nous avons choisi une rotation hebdomadaire des chefs de projet permettant à chacun d'intervenir tout au long du processus de développement du projet.

De manière générale, tous les membres du groupe ont contribué à la conception des différents algorithmes qui composent le projet.

2. Solutions existantes

2.1. Brand24

Brand24 est une solution de veille en ligne utilisée par des entreprises de toutes tailles pour identifier et analyser les conversations en ligne sur leurs marques, produits et concurrents.

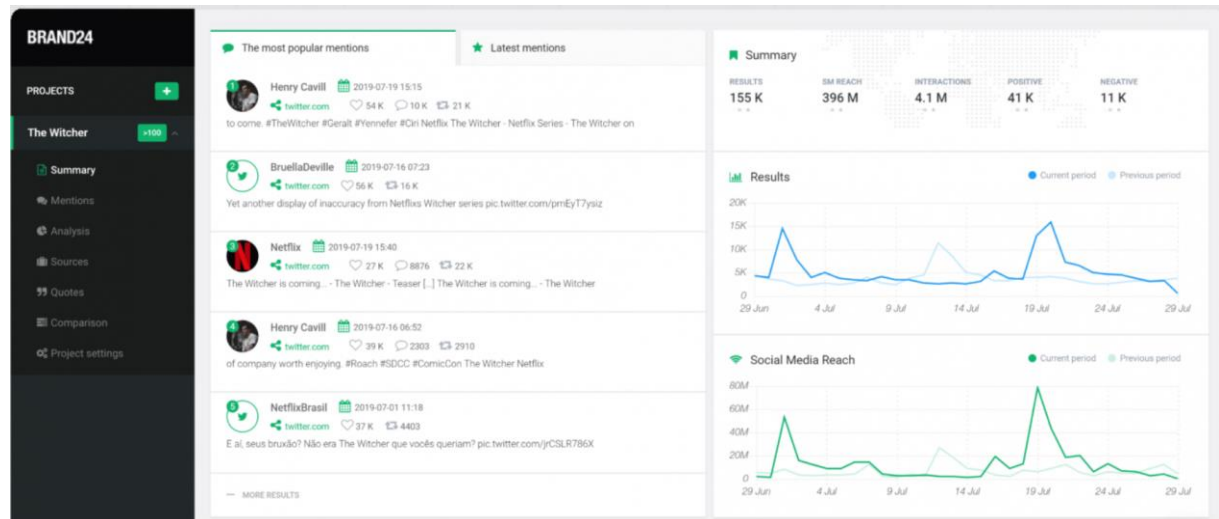


Figure 2 - Exemple de dashboard proposé par Brand24

- Avantages :

Brand24 offre une solution assez simple d'utilisation et permet à la fois de suivre et d'engager des conversations en ligne, découvrir ce que les gens disent en ligne sur la marque et donne un accès instantané aux mentions sur le Web, les réseaux sociaux et éditeurs influents.

- Inconvénients

- Uniquement disponible en Anglais
- Peu d'analyses à valeurs ajoutées proposées
- Licence propriétaire associée à chaque compte (1 entreprise par compte)

2.2. Mention

Mention est décrit comme "l'outil tout-en-un qui vous permet d'écouter votre public, de publier des messages remarquables et de répondre à vos clients".

Mention permet aux marques et aux agences de tirer parti de la surveillance des médias et des médias sociaux afin d'accroître la notoriété des marques. En suivant une marque, un concurrent ou un sujet de l'industrie, la plateforme permet aux clients de comparer et

d'analyser les conversations en ligne pour créer du contenu basé sur d'importantes informations sociales et web. Mention se targue d'avoir aidé plus de 4 000 entreprises clientes telles que Spotify, Airbnb, MIT et Microsoft à améliorer leurs stratégies de communication et de marketing.

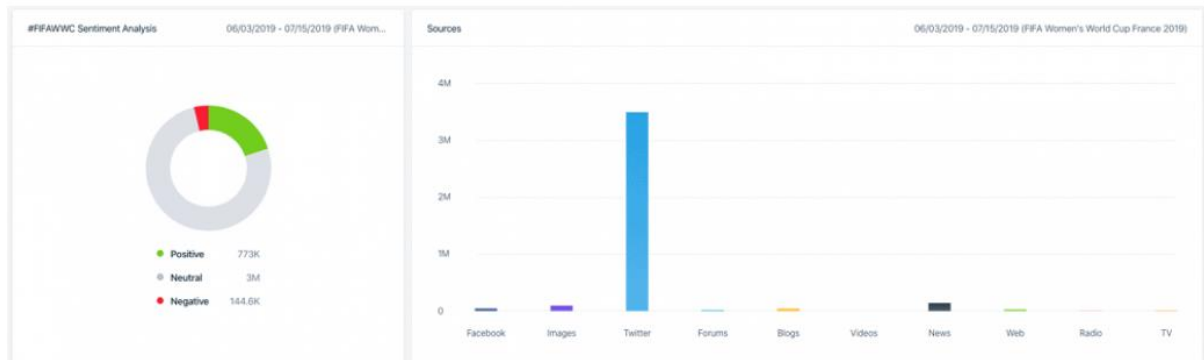


Figure 3 - Exemple de dashboard proposé par Mention

- Avantages :
 - Facile à utiliser
 - S'intègre facilement aux réseaux sociaux
 - Comprend des outils de référencement
 - Trouve des mentions que d'autres outils ne trouvent pas
- Inconvénients :
 - Payant (offres à bas prix très limitées)
 - Les analyses proposées sont très orientées vers les influenceurs au détriment des entreprises en elles-mêmes.
 - La recherche personnalisée d'entreprise est très limitée.

2.3. HubSpot

HubSpot Marketing est l'un des plus grands acteurs du secteur. Il s'agit d'une plateforme tout-en-un qui fournit tout, du marketing aux ventes et à l'acquisition de nouveaux clients. Elle comprend également un CRM gratuit qui aide à mieux concevoir le contenu pour les réseaux sociaux et le Web.

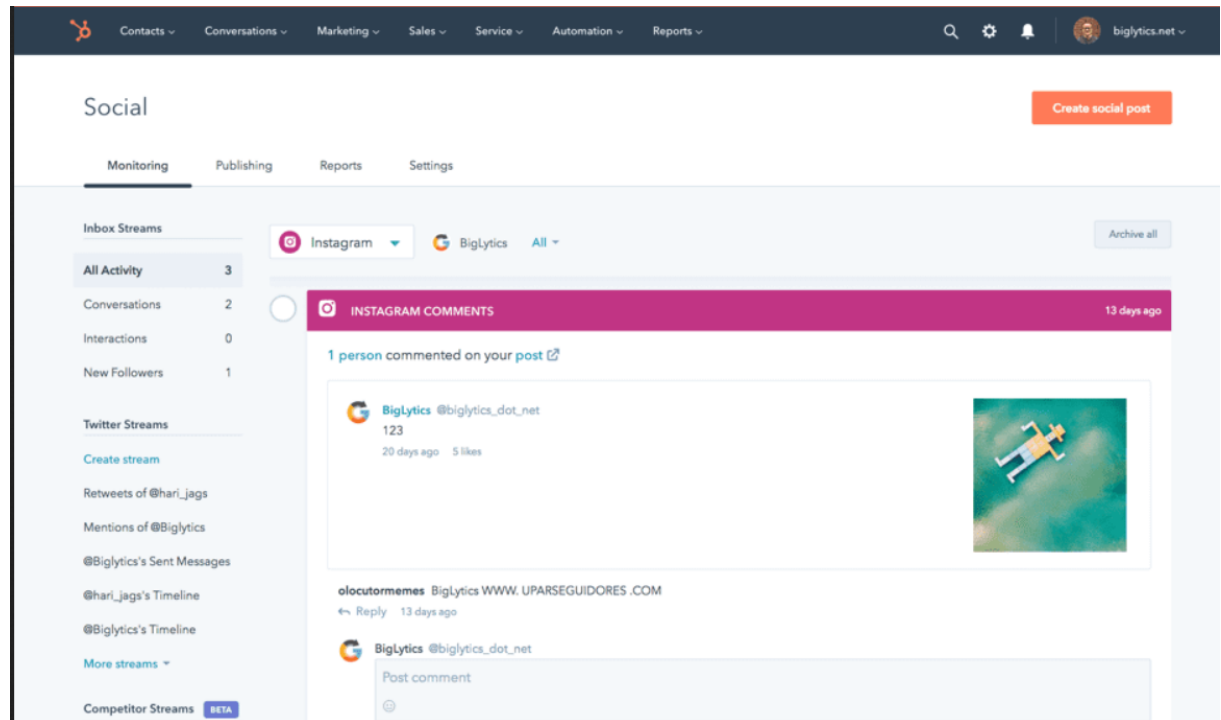


Figure 4 - Exemple de Dashboard proposé par Hubspot

Bien que HubSpot soit une solution complète, comme on peut l'imaginer, elle reste tout de même très coûteuse mais aussi trop complexe pour le suivi des informations sociales. Et de ce fait s'adresse à clientèle limitée.

2.4. Bilan de l'analyse comparative

Les trois solutions présentées s'inscrivent dans l'optique de « continuous monitoring » des réseaux sociaux en donnant à leurs utilisateurs un accès en temps réel aux messages, posts, billets de blog que les internautes publient/partagent sur les différents canaux existants (twitter, instagram, facebook etc.). Certains outils comme Marketing IQ proposent d'extraire de la connaissance utile à partir de données textuelles mais est encore au stade expérimental. Cependant, les principales analyses à valeur ajoutées concernent l'analyse de sentiments liée aux messages (texte) des utilisateurs.

3. Solution proposée

Dans le cadre de ce projet, nous avons, avec l'aide de nos encadrants, conçu une application web sous forme de Dashboard (interface de représentation interactive) permettant à ses utilisateurs de lancer des requêtes d'extraction et d'analyse de données textuelles issues de tweets ayant un rapport avec une ou plusieurs entreprise(s).

Le but étant de pouvoir offrir une solution avancée de monitoring de réseaux sociaux mais également de mettre à disposition des analyses à forte valeur ajoutée comme le topic modeling, l'analyse de sentiment ou encore la détection de relation/entités dans le corpus de texte extrait.

Ce projet a pour intérêt de mettre en application nos acquis et enseignements couvrant la fouille de texte, l'apprentissage non supervisé et l'extraction de connaissance à un cas d'usage issu d'une problématique réelle que représente le marketing digitale et la transformation des métiers du marketing.

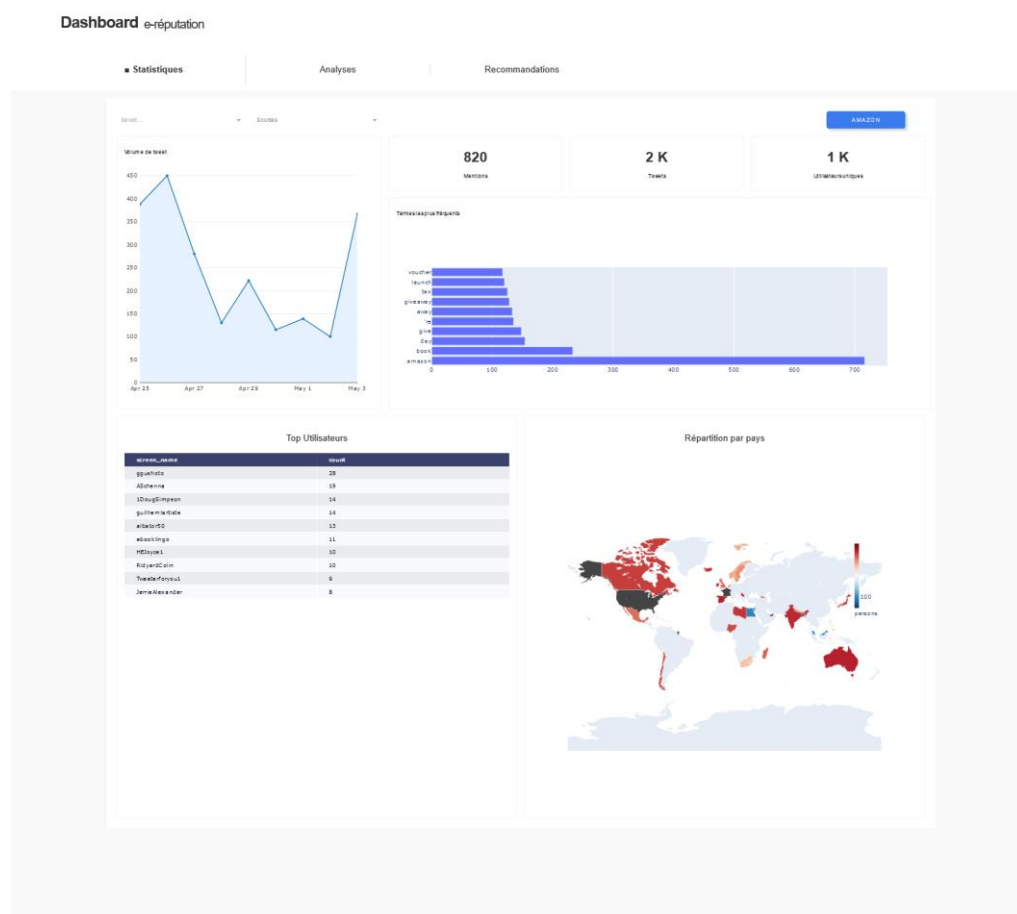


Figure 5 - Dashboard e-réputation proposé

Notre solution s'articule autour de trois onglets :

- Statistiques
- Analyses
- Recommandations

Chaque onglet couvre un volet distinct du projet afin de répondre au critère d'ergonomie et de facilité d'utilisation fortement apprécié par le public cible.

3.1. Statistiques

Dans cet onglet, l'utilisateur aura à disposition les statistiques globales.

3.2. Description des fonctionnalités

3.2.1. Extraction de tweets

Nous avons conçu et développé une pipeline d'extraction et de traitement automatique de tweet depuis l'api twitter (version gratuite). Son intégration à l'application se fait d'une manière intuitive via une fenêtre (formulaire) dans laquelle l'utilisateur peut remplir les différents champs correspondant aux critères de sa requête.

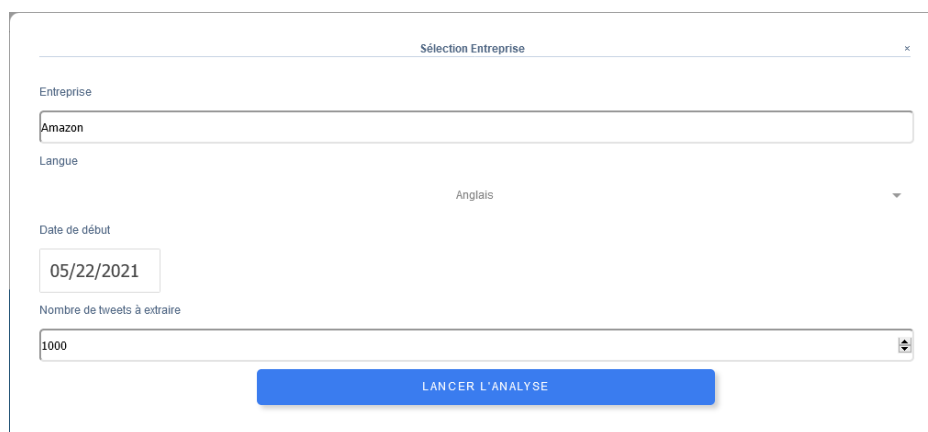
The image shows a web application window titled "Sélection Entreprise" with a close button (X) in the top right corner. The form contains several input fields: "Entreprise" with the text "Amazon" entered; "Langue" with a dropdown menu showing "Anglais"; "Date de début" with a date picker showing "05/22/2021"; and "Nombre de tweets à extraire" with a numeric input field showing "1000". At the bottom of the form is a blue button labeled "LANCER L'ANALYSE".

Figure 6 - Fenêtre de lancement de la pipeline d'extraction de tweets

L'analyse se lance automatiquement après le déclenchement du processus d'extraction.

3.2.2. Volumétrie des tweets

Ce composant illustre l'évolution de la volumétrie des tweets en fonction du temps passé. Nous avons par ailleurs mis en place quelques filtres qui permettent : (i) de pouvoir sélectionner la durée maximale d'évolution du graphique qui peut aller d'un mois à un an et (ii) l'utilisateur a également la possibilité de sélectionner la plateforme à partir de laquelle le tweet a été posté (Android, iPhone, WebApp).

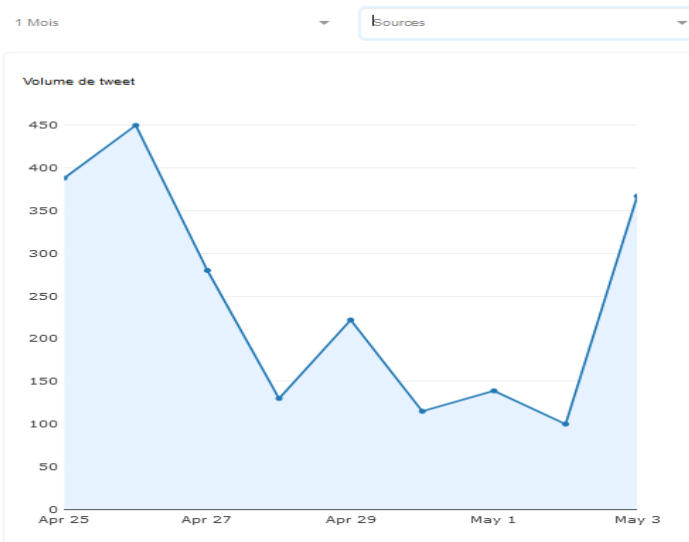


Figure 7 - Graphique représentant la volumétrie des tweets

3.2.3. Histogramme des fréquences

Ce graphique représente les termes les plus fréquents dans le corpus de tweet extrait. Il est construit en aval d'un processus classique de nettoyage appliqué aux données textuelles, à savoir : suppression des mots à faible impact lexical ex : préposition, articles définis/indéfinis, conjonctions, etc. L'algorithme ne retient finalement que les mots qui ont un caractère significatif en précisant également leurs fréquence/redondance respectives :

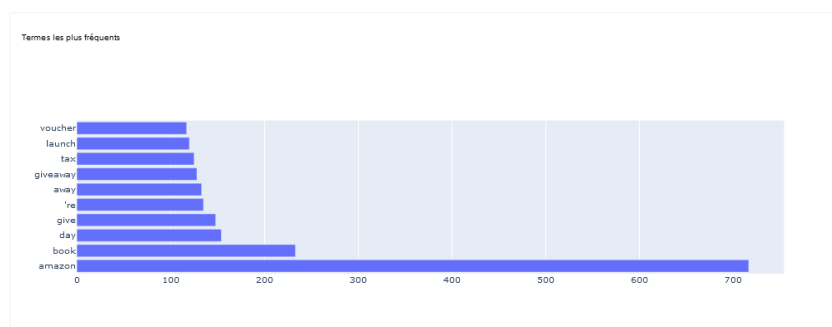


Figure 8 - Histogramme des termes les plus fréquents

3.2.4. Répartition par pays

Ce composant a pour objectif d'avoir une représentation illustrative de la distribution des mots clés par pays d'origination. L'algorithme afférant se base essentiellement sur une approche de détection d'entités nommées en analysant essentiellement les métas donnés accompagnant le tweet. En l'occurrence, il s'agit du traitement de la localisation indiquée par l'utilisateur associé au tweet.

Si cette méta-donnée est jugée non pertinente (n'étant associée à aucune localisation connue). Celle-ci est ignorée.

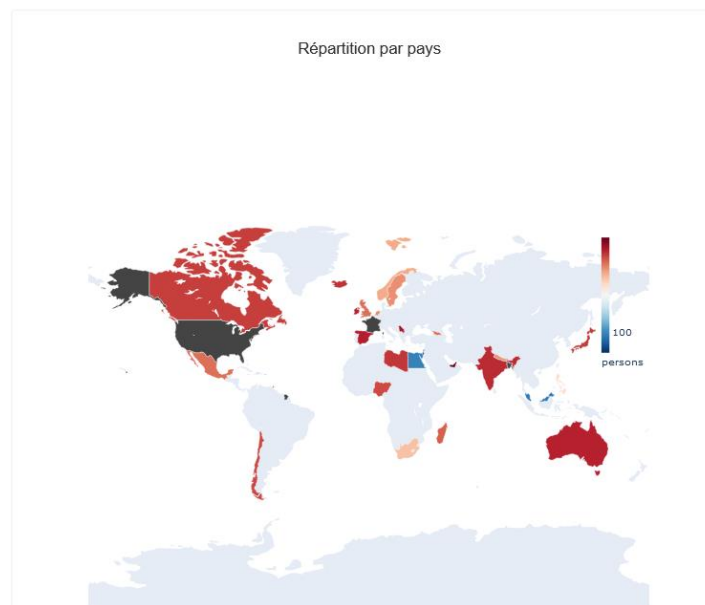


Figure 9 - Distribution des tweets dans le monde

3.2.5. Top utilisateurs

Ce composant illustre les utilisateurs les plus influents à travers le nombre de tweets associés au mot clé recherché. Ces utilisateurs sont ensuite classés par ordre décroissant en fonction du nombre d'abonnés qu'ils possèdent.

Top Utilisateurs	
screen_name	count
ggushoto	28
ASchenna	19
1DougSimpson	14
guilhemlartiste	14
albator50	13
ebooklingo	11
HEJoyce1	10
RidyardColin	10
Tweeterforyou1	9
JerrieAlexander	8

Figure 10 - Tableau représentant les utilisateurs les plus influents

3.3. Analyses

Dans cet onglet, l'utilisateur retrouvera l'essentiel des analyses des tweets sous différents angles. Ces analyses reposent sur des méthodes d'apprentissage non supervisés tels que : (i) l'analyse de sentiments, (ii) le co-clustering et (iii) la reconstruction d'un graph de similarité décrivant chaque sujets (cluster).

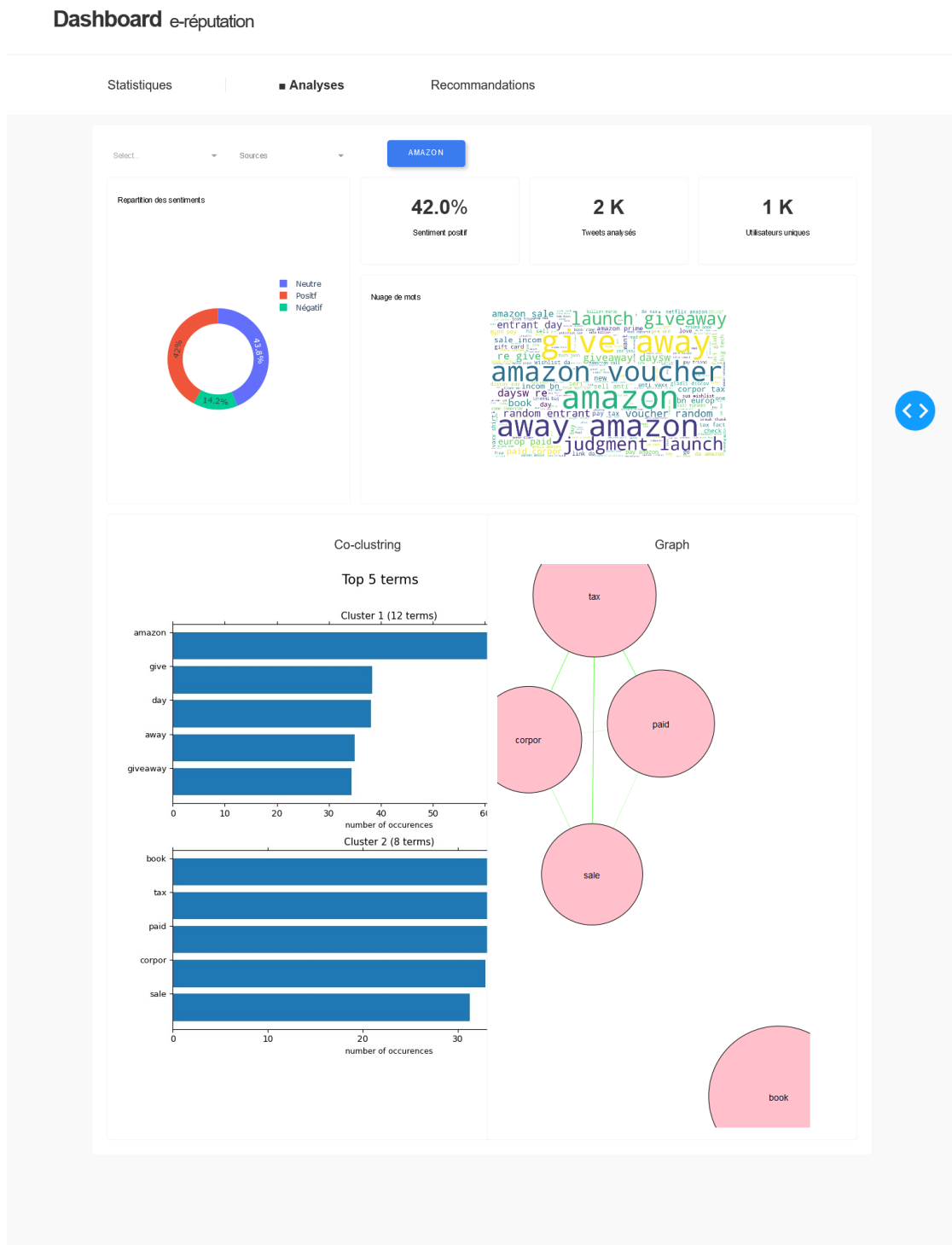


Figure 11 - Onglet Analyse

3.4. Description des fonctionnalités

3.4.1. Répartition des sentiments

Ce composant représente une répartition en doughnut des sentiments exprimés qui sont associés aux tweets analysés. Cette représentation se base essentiellement sur un algorithme issu de la librairie nltk et spicy.



Figure 12 - Graphique représentant l'analyse de sentiment

3.4.2. Nuage de mots

Nous avons choisi d'inclure une représentation en nuage de mots en faisant apparaître les mentions (hashtags) les plus redondantes en évidence (avec une police plus grande).



Figure 13 - Graphique représentant le nuage de mots

3.4.3. Co-clustering (topic modeling)

A partir de la base de tweets nous construisons un co-clustering représentant les sujets (topics) exprimés. Le co-clustering en question se base sur la maximisation de la modularité.

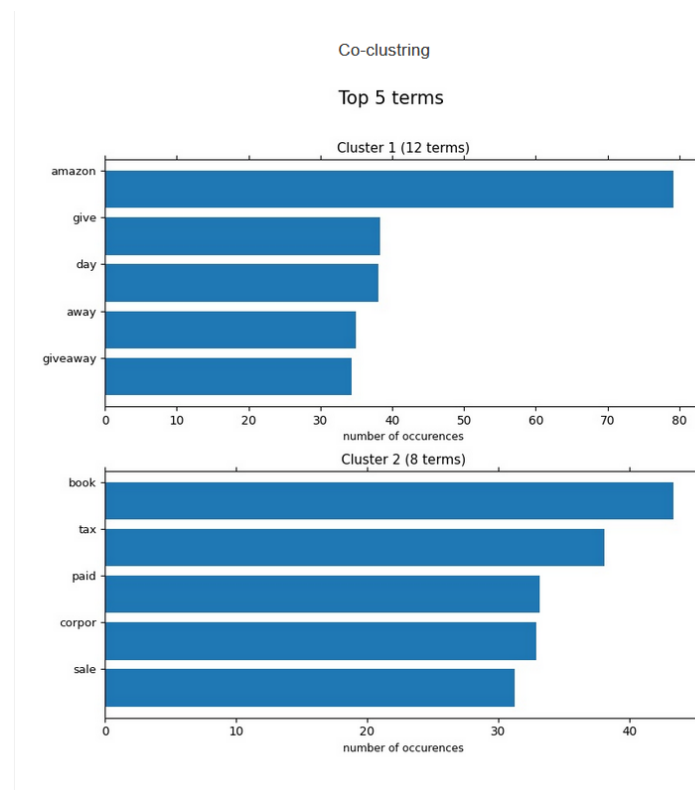


Figure 14 - Topic modeling avec Co-clust

3.4.4. Représentation en graphe

Nous avons également choisi de présenter les liens entre les termes composant chaque cluster en y associant une représentation en graphe en nous appuyant sur la métrique de « cosin similarity ».

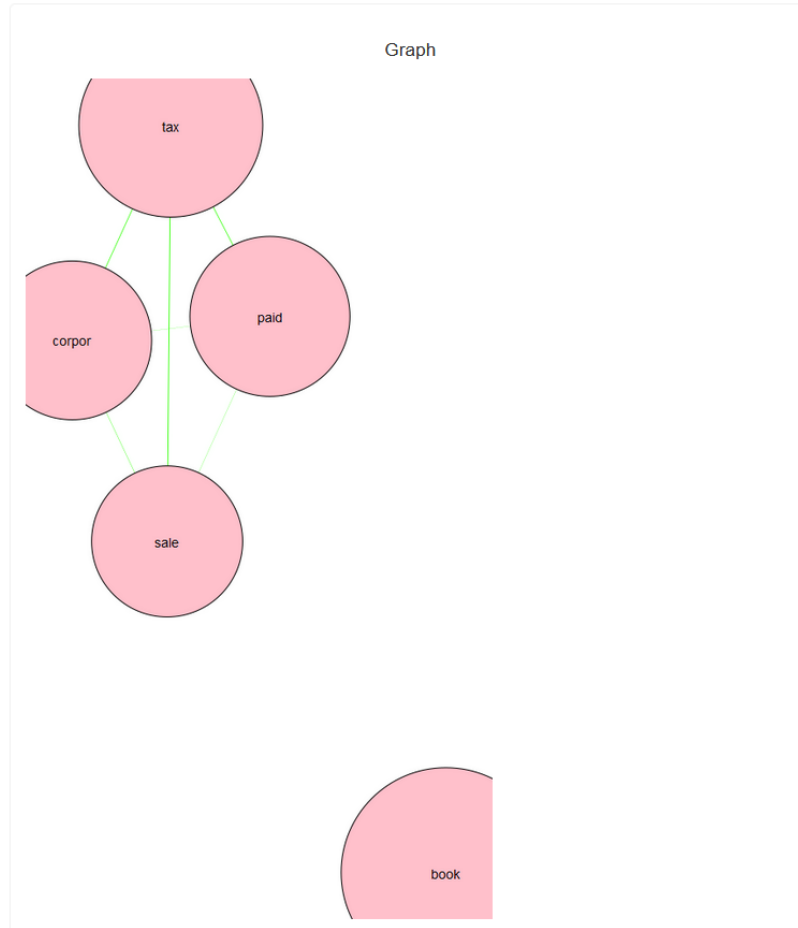
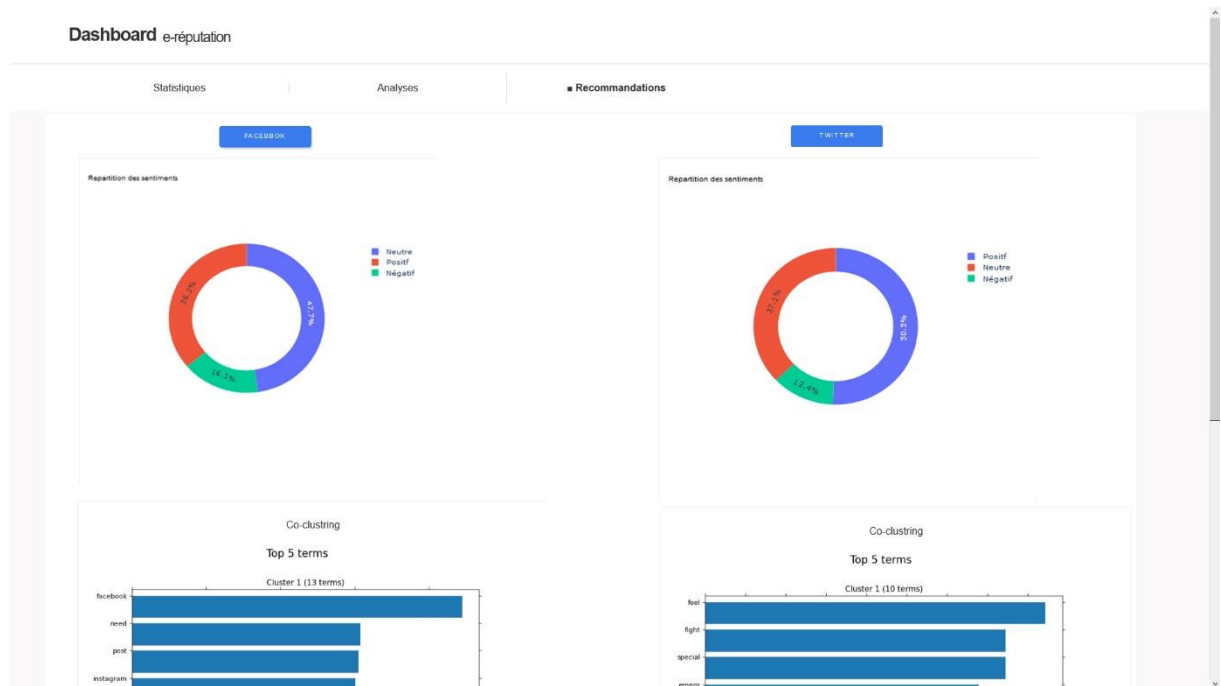


Figure 15 - Graphe représentant la similarité des termes d'un cluster

3.5. Recommandations

Le volet recommandations a pour principal objectif de comparer deux entreprises (indiquées par l'utilisateur) a fortiori concurrentes en déclinant les précédentes analyses présentées plus haut avec pour principale variation une vision illustrative des différences.



L'objectif étant de pouvoir aisément comparer et « benchmarker » ces deux entreprises d'un seul coup d'œil.

4. Technologies utilisées

Pour réaliser ce dashboard, nous nous sommes appuyés sur les technologies suivantes :

- ✓ Dash/Plotly : conception et design du dashboard
- ✓ Tweepy : intégration de l'api Tweeter
- ✓ NLTK : librairie permettant d'appliquer des approches NLP
- ✓ Co-clust : package implémentant les algorithmes de co-clustering sous ses différentes variantes.
- ✓ Textblob : package python permettant de simplifier le travail de pré-processing de données textuelles.
- ✓ Pandas/Numpy : librairie standard de traitement de données sous python.
- ✓ Scikit-learn : librairie dédiée à l'implémentation d'algorithmes de machine learning
- ✓ Igraph : module permettant de construire et manipuler des structures de réseaux/graph

L'implémentation du projet est mise à disposition dans notre repository github :

- https://github.com/catapult0/dashboard_ereputation/

5. Conclusion

Ce projet nous a permis de mettre en pratique nos connaissances en NLP et apprentissage non-supervisé. Nous avons également eu l'occasion de prendre en main un outil de visualisation pour illustrer les résultats des différents algorithmes.

L'objectif étant de restituer nos résultats dans une interface utilisateur qui soit simple et facile d'utilisation pour un utilisateur non averti.

Nous avons beaucoup appris de cette expérience et en retirons beaucoup d'enseignements à la fois sur le plan technique mais également sur le plan de la gestion de projet.