REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITÉ SAAD DAHLEB
BLIDA

FACULTÉ DES SCIENCES

DÉPARTEMENT D'INFORMATIQUE

MEMOIRE DE MASTER

Spécialité : Informatique
Option : Génie des systèmes informatiques

Thème :

RAT^{TR} : Résumé automatique des tweets liés aux tendances de Twitter

Présenté par : ZERIRI Nassima

Promotrice : Mme MADANI Amina

Année universitaire : 2015-2016

REMERCIEMENTS

En tout premier lieu, je remercie le bon Dieu, tout puissant, de m'avoir donné la force pour survivre, ainsi que l'audace pour dépasser toutes les difficultés.

J'adresse tous mes remerciements et ma sincère reconnaissance a ma promotrice **Mme MADANI AMINA**, qui a bien prodiguer ses précieux conseils et son aide et pour m'avoir soutenue tout au long de cette tâche. Je tiens à la remercier de la qualité de son suivi et de la confiance qu'elle a bien voulu m'accorder.

Je tiens à remercier tous les enseignants du département d'informatique qui ont su nous donner une formation didactique et appréciable durant tout notre cursus, et aussi pour leurs conseils et orientations.

Mes plus profonds remerciements vont à mes parents. Tout au long de mon cursus, ils m'ont toujours soutenu, encouragé et aidé. Ils ont su me donner toutes les chances pour réussir. Qu'ils trouvent, dans la réalisation de ce travail, l'aboutissement de leurs efforts ainsi que l'expression de ma plus affectueuse gratitude.

Je tiens à remercier également KERROUR Hadjer pour toute l'aide qu'elle m'a apporté.

Je remercie BOUATMANE Abdessamed pour son aide, son écoute et son soutien.

Je remercie toutes mes sœurs : Sonia, Karima, Nadjet et Imene pour leurs disponibilité et encouragement.

RÉSUMÉ

Titre : Résumé automatique des tweets liés aux sujets tendances sur Twitter

Twitter est devenu extrêmement populaire, avec des centaines de millions de tweets postés chaque jour sur une grande variété des sujets. La richesse du contenu de ces messages offre une opportunité dans le domaine de la recherche. Ces messages courts avec plusieurs langues, sont généralement mal orthographié, et bruité composent des sujets tendances qui couvrent l'actualité dans le monde entier.

L'évolution de l'information sur ce réseau social nous a aidés à s'introduire dans le domaine du résumé automatique des tweets de ces sujets tendances en temps réel. Le but est d'éviter la lecture manuelle des milliers des messages, et de fournir aux utilisateurs une information compréhensible et correcte syntaxiquement et sémantiquement et résumant ces tweets.

Dans ce mémoire, nous proposons une nouvelle approche pour le résumé automatique des sujets tendances en temps réel. Nous avons récupéré les tweets afin de répondre aux requêtes des utilisateurs en fournissant la liste des sujets tendances et ses messages, pour l'analyse et le traitement. Enfin, nous avons appliqué une méthode de résumé automatique avec catégorisation en se basant sur les entités nommées, pour avoir un résultat en temps réel.

Mots-clefs : Tweets, Traitement Automatique des Langues, Recherche d'Information, Résumé automatique, sujets tendance.

ABSTRACT

Title: Automatic summarization of trending topics in real time

Twitter has become extremely popular, with hundreds of millions of tweets posted each day on a variety of subjects. The richness of the content of these messages is an opportunity in the field of research. These short messages with multiple languages, are usually misspelled, and noisy consist of trending topics news coverage worldwide.

The evolution of information on this social network helped us get into the field of automatic summary of tweets of trending topics in real time. In order to avoid manual reading thousands of messages, and provide users with understandable and correct information syntactically and semantically and summarizing these tweets.

In this paper, we propose a new approach for the automatic summary of real-time trend topics. We retrieve tweets to meet users' requests by providing the list of trends and messages subjects for analysis and treatment. Finally, we applied a method of automatic summarization by the sub categorization, to have a result in real time.

Keywords: Tweets, Natural Language Processing, Information Retrieval, Automatic summarization, trending topics.

الملخص

العنوان: التلخيص الأوتوماتيكي للموضوعات الرائجة في الوقت الحالي

أصبح لتويتر شعبية للغاية، مع مئات الملايين من التغريدات المحتواة كل يوم في مجموعة متنوعة من الموضوعات. الثراء المحتوى في هذه الرسائل هو فرصة في مجال للبحوث. هذه الرسائل القصيرة ذات لغات متعددة، وعادة ما تحتوي على أخطاء إملائية، و تكون داخل الموضوعات الرائجة لتغطية الأخبار في جميع أنحاء العالم.

ساعد تطور معلومات هذه الشبكة الاجتماعية لنا للتدخل في مجال الملخص التلقائي من هذه التغريدات في مواضيع في الوقت الحالي. من أجل تجنب قراءات هذه الرسائل،و تزويد المستخدمين بمعلومات مفهومة و صحيحة نحويا ودلاليا لخصنا هذه التغريدات.

في هذا العمل، نقترح نهجا جديدا للملخص التلقائي من الموضوعات الرائجة في الوقت الحالي. قمنا باسترداد التغريدات لتلبية طلبات المستخدمين من خلال توفير قائمة من الرسائل لتحليلها ومعالجتها. وأخيرا، فإننا نطبق أسلوب التلخيص التلقائي عن طريق تصنيف فرعي، والنتيجة تكون في الوقت الحالي.

الكلمات الرئيسية: التويت ،المعالجة التلقائية للغات، التلخيص التلقائي، استخراج المعلومات، استرجاع المعلومات،الموضوعات الرائجة.

SOMMAIRE

SOMMAIRE

SOMMAIRE

LISTE DES FIGURES

LISTE DES FIGURES

LISTE DES TABLEAUX

INTRODUCTION GENERALE

1. Introduction générale :

Les réseaux sociaux se sont installés petit à petit dans notre quotidien, bouleversant les méthodes de communication et le partage d'informations. Ces plateformes sociales ne cessent d'évoluer, offrant toujours plus de nouvelles fonctionnalités.

Plusieurs recherches ont montré que les données publiées par les internautes sur les sites de médias sociaux, notamment Twitter[1], le réseau social de l'information instantanée reflètent presque en temps réel l'intérêt du public. Twitter fait partie des réseaux sociaux qui prennent de plus en plus de place dans notre quotidien. Il participe activement à la vie économique, politique, et aux débats sociétaux du monde entier.

Les utilisateurs inscrits sur ce réseau social sont en mesure d'établir des relations entre eux. Un utilisateur pouvant s'abonner à d'autres, ce qui lui permet de consulter leurs messages au moment de sa connexion. Ces messages courts appelés « tweets » sont limités à 140 caractères.

Le contenu d'un tweet peut être un avis, une information ou un témoignage. La vaste communauté de Twitter, le fort taux d'utilisation, plus de 500 millions **[Houssem Eddine Dridi et al., 2013]** de tweets par jour. Il offre la possibilité de communiquer avec la communauté entière des internautes, et garde une place à part dans le paysage des réseaux sociaux, ainsi qu'il est une ressource importante d'annonces et de diffusion de nouvelles actualités.

Les tweets sont publiés par rapport à la variété des intérêts des utilisateurs conduisant à une accumulation d'informations sur des évènements ou locaux à l'échelle internationale. Ce grand volume d'information peut nous guider vers le domaine de l'indexation, la recherche d'informations et la fouille de données afin de concevoir de nouvelles applications de détection et d'analyse de ces messages.

[1] https://twitter.com/

INTRODUCTION GENERALE

Les évènements pertinents sur Twitter (les sujets tendances[2]) composés des millions de tweets[3] et échangés par de multiples utilisateurs. Ces événements sont une source d'information précieuse, mais ce grand flux de messages courts qui composent les tendances est parfois difficile à comprendre le contenu de chaque sujet. Par conséquent, il est important de réaliser un système pour résumer ces tendances et comprendre le contenu de chaque sujet.

Cette grande évolution de l'information a donné le besoin d'avoir des systèmes conçus pour *les résumés automatiques*. Les résumés automatiques sont utilisés dans toutes les informations présentes sur les sites web, et les réseaux sociaux. Dans Twitter, faire un résumé des sujets émergents est une action très importante et demandée par tous ces utilisateurs, afin d'éclaircir l'information et avoir un document plus condensé et compréhensible.

Le résumé automatique consiste à extraire les mots les plus dominants et qui apportent un sens au sujet en question pour mieux comprendre l'actualité du moment.

Dans ce contexte, notre étude sera basée sur ces sujets tendances de Twitter, plus précisément, le résumé automatique de ces évènements.

2. Problématique :

La recherche et l'extraction de l'information (RI) à partir d'un ou plusieurs documents est l'un des domaines les plus utilisés en informatique. Le domaine de recherche d'informations remonte au début des années 1950, peu après l'invention des ordinateurs. La « RI » fut donné par Calvin N. Mooers **[Mooers, 1948]**. En 1948 pour la première fois, quelques années après, le domaine est devenu plus actif. Plusieurs groupes de recherche ont identifié la RI comme un des thèmes importants.

---

[2] https://en.wikipedia.org/wiki/Twitter#Trending_topics

[3] **Le tweet** : message court de 140 caractères envoyé entre les utilisateurs sur Twitter (https://fr.wikipedia.org/wiki/Twitter#Le_tweet_et_le_retweet)

INTRODUCTION GENERALE

L'utilisation de la recherche d'informations dans les réseaux sociaux, notamment pour Twitter est très essentielle. Elle est particulièrement limitée par la taille courte de ses messages qui augmente à son tour la difficulté de la recherche textuelle par mots-clés.

Le besoin d'avoir l'information voulue et demandée par les utilisateurs nous guide vers l'étude de ce domaine pour rechercher ces mots-clés qui sont les sujets tendances concernant un événement et comportant des milliers de tweets.

L'information demandée extraite est composée par un grand volume de messages nous mène au développement de nouveaux systèmes du résumé automatique, d'où ces tweets seront résumés ce qui permet de fournir une information minimisée et compréhensible peu bruitée.

Dans le cadre des systèmes de résumés automatiques, le résumé automatique des sujets tendances a besoin de plusieurs étapes pour fournir une réponse correcte à l'utilisateur : en premier lieu, l'extraction des tweets les plus pertinents et qui appartient à un des sujets d'actualités sur Twitter, ensuite, l'analyse et le traitement du contenu, enfin, la catégorisation en utilisant les stratégies de la recherche d'informations pour arriver à un résumé cohérent.

Les problématiques abordées dans ce mémoire, se présentent comme suit :

- La compréhension du contenu de chaque sujet tendance.
- Eviter la lecture manuelle des millions des tweets.
- La perte de temps.

3. Cadre de travail :

Notre travail se base sur le domaine de traitement automatique de langage (TAL). Ce dernier est devenu une source incontournable. L'explosion de ces différentes applications a rendu nécessaire la description et la formalisation de phénomènes linguistiques, mais aussi la création de nouveaux outils manipulant de vastes ensembles de données textuelles.

INTRODUCTION GENERALE

Son histoire a commencé au début des années 1950, ou **[Alan Turing, 1950]** a édité un article célèbre sous le titre « *Computing machinery and intelligence* » qui propose ce qu'on appelle à présent le test de Turing comme critère d'intelligence. Ce domaine issu de l'intelligence informatique s'appuie sur le domaine de linguistique fondamentale (lexique, sémantique, syntaxe, morphologie, et analyse du discours).

Parmi les tâches les plus importantes auquelles s'intéresse le TAL : la traduction automatique, le résumé automatique, la reconnaissance d'entités nommées, la classification …etc.

Le résumé automatique de textes est une discipline TAL, qui a pour objectif de compresser des documents textuels. Ce processus de compression implique une perte d'informations. Déterminer la pertinence de l'information est l'une des difficultés majeures du procédé.

4. Objectifs :

Plusieurs travaux s'intéressent au domaine du résumé automatique. Dans notre travail, nous avons réalisé une étude approfondie de l'état de l'art sur ces recherches qui ont déjà exposé des résultats dans ce domaine. Ensuite, à partir de cette analyse, nous avons fait une comparaison entre ces différentes recherches en se basant sur des critères.

Cette comparaison nous a permis de proposer une nouvelle approche pour extraire les tweets des sujets tendances et les analyser afin de les catégoriser dans trois catégories différentes. L'approche permet de fournir un résumé automatique cohérent, compréhensible et en temps réel.

5. Organisation du mémoire :

Notre mémoire s'organise en quatre chapitres :

Chapitre I : Les généralités de Twitter, et l'utilité de son usage.

Chapitre II : Présente une étude approfondie de l'état de l'art, nous présentons l'ensemble des travaux importants dans le domaine du résumé automatique.

Chapitre III : Présente une nouvelle approche du résumé automatique en se basant sur les travaux cités dans le chapitre 2.

Chapitre IV : Présente la partie test de notre approche afin d'évaluer notre travail.

Chapitre I : GÉNERALITÉS : TWITTER

CHAPITRE I : GENERALITES : TWITTER

Introduction :

Twitter est un réseau social très populaire, connu par le partage d'informations avec des messages courts composés de 140 caractères appelés « tweets ». Ces derniers se rapportent à des événements produisant en temps réel.

Les tweets constituent un grand corpus de données explorables par différents domaines de recherches. Ils sont utilisés aussi pour l'extraction des sujets émergents et le résumé automatique de ces sujets.

Dans ce chapitre, nous allons présenter Twitter et ses différentes fonctionnalités, ensuite, les différentes APIs et leur importance dans la collecte des tweets. Enfin, nous explorons les statistiques récentes de ce réseau social.

2. Historique :

Twitter a été créé à San Francisco au sein de la startup Odeo fondée par Noah Glass et Evan Williams autour de 2005**[Stéphane Manet, 2013]**. L'idée de départ lancée par Jack Dorsey était de permettre aux utilisateurs de partager facilement leurs petits moments de vie avec leurs amis.

Ouverte au public le 13 juillet 2006, la première version s'intitulait*Stat.us* puis *Twittr*, en référence au site de partage de photosFlickr puis *Twitter*, son nom actuel. Le 21 mars 2006, M. Dorsey envoyait son premier tweet : « Just setting up mytwttr » (« Suis en train d'installer mon twttr »).

Au début de la création, il n'y avait pas de limite au nombre de caractères permis. Twitter compte alors une centaine d'abonnés. En avril 2007[4], Twitter devient une véritable entreprise et Jack Dorsey en prend les commandes.

Le 4 octobre 2010, Evan Williams, le cofondateur, annonce qu'il passe la main à Dick Costolo, ancien directeur d'exploitation.

En juin 2012, les mots « Twitter » (nom propre), « twitt » ou « tweet », « twitteur » ou « twitteuse », ainsi que « tweeter », font leur apparition dans Le Petit Larousse édition 2013

Twitter connaîtra son véritable envol un an plus tard, au Festival South by Southwest (SXSW) à Austin, au Texas. Lieu de rencontre de l'avant-garde techno, SXSW récompense Twitter en lui accordant un Web Award. Les participants à la conférence, eux, adoptent Twitter sur-le-champ. C'est le coup de baguette magique qu'il fallait. En l'espace d'un week-end, le nombre de gazouillis envoyés passe de 20 000 à 60 000.

Avec les années, les gens ont développé le réflexe de se tourner vers Twitter en temps de crise (séisme en Haïti, révolution en Égypte) pour y trouver des informations en premier plan.

3. Twitter :

Le site Twitter est un service de micro-blogging[5]. Cet outil de réseau social permet d'envoyer des messages (appelés tweets) de 140 caractères maximum à partir de son espace membre. C'est un outil de micro-blogage géré par l'entreprise Twitter Inc, et donne la possibilité aux

[4] http://www.guichetdusavoir.org/viewtopic.php?f=2&t=53297

utilisateurs de savoir tout ce qui se passe autour de nous en temps réel, ainsi que raconter ce qu'on fait au moment où l'on fait en répondant à la question qu'il pose « what are youdoing ? » ou en français « que faites-vous ? ».

Twitter est utilisé par différentes personnes comme les passionnés d'informatique pour effectuer de la veille technologique, par des entreprises pour communiquer, par des célébrités pour informer leurs fans ...

Le principal avantage de Twitter est le fait qu'une information peut circuler et informer des millions de personnes du réseau en très peu de temps. Cet avantage est à double tranchant : Il est si facile pour une personne de faire circuler une information, une désinformation plus ou moins large sera faite si l'actualité est fausse ou partiellement erronée.

La figure 1.1 présente l'interface de l'utilisateur sur Twitter.

Figure 1.1 : Page d'accueil de Twitter le 15/01/2016.

3.1 Les tweets : Un tweet est un petit message de 140 caractères maximum diffusé sur la plateforme Twitter.

Les tweets d'un auteur sont diffusés auprès de ses followers ou abonnés, c'est à dire les individus ayant choisi de suivre la publication de ses petits messages. Ils sont généralement composés de :

[5] http://cyberchemille.org/spip.php?article72

- **Les # (hashtag)**[6] placés devant un mot permettent de signaler des mots clés. Il est ensuite possible de lister tous les tweets contenant un hashtag
- **Les @ (mention)** permet de désigner un compte utilisateur. Chacun suit la liste des tweets dans lesquels il est mentionné.
- **URL**[7] courte grâce à des services de codage permettant de réduire le nombre de caractères utilisés pour un lien
- Association de photos grâce à des services de dépôts en ligne.

Figure 1.2 : Exemple de quatre tweets

3.2Le fil d'actualités (Timeline) :

C'est le flux de messages qu'un utilisateur reçoit en provenance des personnes qu'il suit. Ce fil d'actualité est mis à jour en temps réel.

3.3Retweet (RT) :

Un retweet est le message d'une personne republié par une autre. Un message retweeté se compose comme suit : RT @personne message. Certains logiciels utilisent le mot « via » pour signifier qu'il s'agit d'un retweet.

---

[6]**Hashtag :** un marqueur de métadonnées permet de marquer un contenu avec un mot-clé plus ou moins partagé.

[7]**URL (*Uniform Resource Locator*) :** une chaîne de caractères utilisée pour adresser les ressources du World Wide Web

On retweet parce qu'il y a des informations importantes qui sont publiées par des personnes et qui pourraient intéresser d'autres abonnés.

Il y a plusieurs raisons d'utiliser le symbole RT, la plus importante étant de mentionner aux lecteurs que ce n'est pas l'auteur original du message. Ensuite dans les autres raisons, il y a la question de politesse en accordant le crédit de l'information à son auteur, mais également de lui notifier qu'on a trouvé son tweet pertinent. De plus les droits d'auteur ou de propriété intellectuelle sur les tweets n'étant pas très clairs, vaut mieux citer et identifier l'auteur d'un tweet qu'on reposte. L'avantage est qu'il permet également de découvrir de nouveaux comptes Twitter à suivre.

3.4 Les followers :

Twitter a mis en place un concept de *followers* (suivre les gens). Donc, on a des followers (des personnes qui nous suivent) et ont suit les gens (on est leur follower), c'est-à-dire que l'on suit les informations qu'ils postent et dès qu'un utilisateur met à jour son statut, tous les followers sont informés.

Cette opération est réalisée en cliquant sur le bouton suivre ou (Follow) sur une page Twitter.On peut suivre tous les autres utilisateurs à moins qu'un utilisateur ait mis son profil en mode privée. Dans ce cas, une demande d'approbation doit être envoyée en premier.

Figure 1.3 : Figure représentant l'action « suivre »

3.5 Les hashtags :

Les HashTags sont un moyen pour ajouter des informations additionnelles aux tweets pour les catégoriser selon un contexte. Ainsi qu'une information permettant de les lier à un groupe de

tweets décrivant un évènement ou un lieu. Il est créé en ajoutant le symbole # avant le mot clé.

3.6 Le Direct Message (DM) :

Un DM ou Direct Message est un message envoyé directement à la personne et qui n'est visible que par celle-ci. Il n'est pas publié publiquement et n'apparait pas dans le Timeline. Un Direct Message peut être assimilé à un email interne dans Twitter. Cependant pour pouvoir envoyer un DM à une personne il faut que celle-ci vous suive et réciproquement recevoir un DM d'une personne, il faut être abonné à son compte.

3.7 Les sujets tendances #TT :

Les sujets tendances (en anglais trendingtopics) peuventêtreréférencé dans un tweets avec #TT, cela veut dire que cette actualité est populaire et plusieurs personnes en parlent.

3.8 Les mentions (@) :

Le « @ » est toujours accolé au pseudo d'un compte Twitter et permet de faire savoir à son destinataire qu'un message lui était adressé.

3.9 Liste :

Permet de classer les membres qu'un utilisateur suit ou non dans des listes (20 maximum) en fonction de critères comme (famille, collègues, chanteurs…). Ces listes peuvent être privées ou publiques. Réciproquement, les autres membres peuvent aussi inscrire d'autres utilisateurs dans une de leurs listes.

4. Statistiques :

Twitter[8] fait désormais partie intégrante des principaux réseaux sociaux. Il se démarque des autres services par de nombreux aspects. Ainsi qu'il dénombre aujourd'hui pas moins de 200 millions d'utilisateurs, qui passent majoritairement par des plateformes tierces pour accéder au réseau (TweetDeck[9], HootSuite[10]…). Il est plus utilisé aux États-Unis, l'Argentine, la

[8] http://www.blogdumoderateur.com/statistiques-twitter-entree-en-bourse/

[9] **TweetDeck :** est une application logicielle qui permet de consulter et gérer un ou plusieurs comptes Twitter, via une interface graphique conviviale.

[10] **HootSuite :** est un outil de gestion, prend la forme d'un tableau de bord et intègre les flux de différents réseaux sociaux

France, le Japon, la Russie, l'Arabie Saoudite et l'Afrique du Sud mais il n'est pas loin pour se propager dans tous les pays du monde. Pour cela, une étude a été faite pour analyser la progression de ce réseau, et les statistiques les plus récentes se basent sur le nombre de tweets envoyés chaque jour, le nombre des utilisateurs actifs, le prix d'une tendance sponsorisé, le nombre de followers d'un compte Twitter…. Etc.

Voici les principaux chiffres à retenir concernant Twitter pour l'année 2015[11] :

- Nombre d'utilisateurs actifs mensuels sur mobile : environ 80%.
- Nombre d'utilisateurs actifs quotidiens (DAU) : environ 100 millions.
- Nombre de comptes certifiés : 136 000.
- Les tweets contenant une photo sont deux fois plus partagés que la moyenne.
- 500 millions de tweets sont envoyés chaque jour.
- 300 milliards de tweets ont été envoyés depuis le 21 mars 2006.
- 170 minutes, c'est le temps moyen passé chaque mois sur Twitter.
- 20 millions : c'est le nombre de « faux comptes » sur le réseau social.
- $200.000 : c'est le prix d'un TrendingTopic sponsorisé (tendances mondiales) durant 24h.
- 80% des membres utilisent leur mobile pour accéder à Twitter.
- 208, c'est le nombre moyen de followers d'un compte Twitter.
- 40 millions d'internautes utilisent Vine.
- 1 million de sites intègrent des tweets.
- 181 milliards d'impressions de timelines ont été enregistrées au troisième trimestre 2014.
- 44% des utilisateurs n'ont jamais tweeté.

[11] http://www.blogdumoderateur.com/chiffres-twitter/

5. Conclusion :

Dans ce chapitre, nous avons présenté Twitter, la structure des tweets qui constitue nos données à traiter et analyser dans cette étude, ensuite, les APIs-Twitter qui permettent de récupérer ces tweets.

Nous pouvons conclure que Twitter a des caractéristiques très importantes qu'ils le rendent explorable par le domaine du résumé automatique surtout pour les sujets émergents :

- ✓ **La diversité des sujets :** Twitter est utilisé par un large public, ce qui permet de fournir un grand corpus d'informations de divers sujets qui se rapportent en temps réel.

Chapitre II : RÉSUMÉ AUTOMATIQUE DES TWEETS LIÉS AUX SUJETS TENDANCES : ÉTAT DE L'ART

CHAPITRE II : RESUME AUTOMATIQUE DES TWEETS LIES AUX SUJETS TENDANCES : ETAT DE L'ART

1. Introduction :

Twitter est devenu une source importante pour l'échange d'informations en temps réel. Un grand nombre de tweets se produit chaque minute avec des sujets différents.

Pour aider les utilisateurs à s'ouvrir aux évènements qui sont des sujets tendances, Twitter a ajouté un nouvel outil qui sert à voir les nouveautés du moment, ces sujets tendances sont soit des mots ou encore des phrases comme : « Coupe du monde », « élections présidentielles » etc…

Afin de comprendre pourquoi un sujet est tendance, il est obligé de lire tous les tweets connexes. Ces tweets courts sont pleins d'erreurs, hiérarchisées, en plusieurs langues et qui est susceptible de rencontrer des Spams.

Dans l'intention de fournir des définitions des sujets tendances, Twitter a un partenariat avec le site tiers **WhatTheTrend**[12]qui permet aux utilisateurs de saisir manuellement des descriptions du sujet pour lequel il est tendance.

L'inconvénient majeur de ce site est que les définitions sont entrées à la main et non automatisées et qu'il y a souvent un temps de latence avant qu'une tendance soit définie par un utilisateur.

Dans le but de générer une explication compréhensible par des humains et avec moins de volume, un processus automatique est crucial pour extraire et résumer les nouveautés qui se produisent sur Twitter en temps réel.

Dans ce chapitre, nous présentons les sujets tendances sur Twitter, le résumé automatique et son histoire, ensuite nous allons explorer les différents travaux connexes existant dans ce domaine en présentant les caractéristiques de comparaison.

[12] http://www.whatthetrend.com

2. Les sujets tendances sur Twitter :

Les sujets tendances[13] (ou en anglais Trending topics), abrégés « TT » sur Twitter, sont les sujets tendances. Ce sont des mots, des hashtags ou des phrases qui ont été tweetés et retweetés multiples fois par des utilisateurs différents durant une période. Twitter permet d'afficher les tendances par pays, par ville ou encore dans le monde entier.

Les tendances[14] sont des sujets à la une qui sont déterminés en temps réel par un algorithme en fonction de la localisation et des abonnements d'un utilisateur. Affichées dans la colonne de gauche via les onglets « Accueil », « Notifications », « Découvrir » et « Moi ».

La figure 2.1 montre un exemple de dix tendances, pour accéder à une de ces tendances, en cliquant, des millions de tweets apparaissent avec le mot-clé de ce sujet tendance.

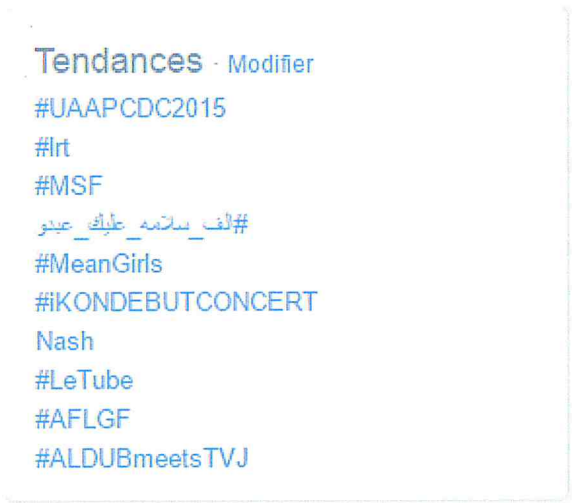


Figure 2.1 : Exemple des tendances sur Twitter, Date du 03/10/2015

Pour qu'un sujet devienne une tendance, ce n'est pas le nombre de tweets qui compte, mais le nombre de personnes qui tweetent[15].

Les tendances sont déterminées par un algorithme qui vérifie tous les tweets (il y a maintenant plus de 500 millions par jour) **[Houssem Eddine DRIDI1et al., 2013].**

[13] https://fr.wikipedia.org/wiki/Twitter

[14] http://www.emarketinglicious.fr/social-media/comment-creer-tendance-twitter

[15] http://www.agence-indigo.com/communication-blog/comment-fonctionnent-et-prevoir-les-trending-topics-sur-twitter/

Cet algorithme est découvert par Devavrat Shah (chercheur au Massachusetts Institute of Technology (MIT)), dont il pourrait s'appliquer à d'autres domaines comme la Bourse, explique-t-il. Ainsi qu'Il est Consultable au niveau global, par régions, pays ou même villes.

Cet algorithme peut prédire les sujets qui seront tendance sur Twitter 1h30 à l' avance. : Il fait le calcul interne des mots-clés par Twitter en fonction du nombre de personnes qui parlent d'un même sujet à un moment donné sur le réseau social.A partir d'une base de données de 200 sujets populaires et 200 autres peu relayés à un instant T, l'algorithme compare chaque nouveau sujet et lui attribue un taux de probabilité de devenir populaire[16].

Il est applicable à n'importe quel type de données qui varient dans le temps, comme la durée d'un trajet en bus, les ventes d'entrées de cinéma et même le marché boursier.

3. L'importance des sujets tendances :

Les sujets tendances sont une fonction importante qui aide les utilisateurs à s'orienter vers un domaine précis sur Twitter et englobent une bibliothèque assez importante dans tous les domaines pour toute personne travaillant dans le journalisme, le marketing, la médecine, l'économie …etc. ils permettent l'accès aux derniers sujets en discussion publique est essentielle. On peut voir quels sont les sujets actuellement en cours de discussion, en les parcourant, on peut aussi avoir une idée sur les réactions et les différentes opinions à ce sujet.

4. Résumé automatique d'un texte :

Selon **[Mohamed Hedi Maaloul, 2013]**, un résumé est une forme abrégée de texte, d'un discours..etc. Et d'après **[Mani 1999],** un résumé consiste à condenser l'information la plus importante provenant d'un document (ou de plusieurs documents) afin d'en produire une version abrégée pour un utilisateur (ou plusieurs utilisateurs) et une tâche (ou plusieurs tâches). De même, un résumé de texte peut être défini en tant qu'objet dont la taille est inférieure au texte source et dans lequel on retrouve certaines idées essentielles présentes du texte d'origine **[Masson 1998].**

[16] http://www.01net.com/actualites/un-algorithme-pour-identifier-les-buzz-a-venir-sur-twitter-579563.html

CHAPITRE II : RESUME AUTOMATIQUE DES TWEETS LIES AUX SUJETS TENDANCES : ETAT DE L'ART

Le résumé automatique est une version condensée d'un document textuel obtenu au moyen de techniques informatiques avec une représentation abrégée et exacte du contenu d'un document.

4.1 Historique du résumé automatique :

L'origine de la tâche du résumé automatique de texte remonte aux années 50 **[Luhn, 1958]** quand il s'intéressait à réaliser un système automatique pour les grands corpus et faire des résumés synthétiques. Jusqu'aux années 80, les investigations sur ce domaine se sont traitées d'une façon ininterrompue.

A partir des années 90, la croissance exponentielle de la masse d'information disponible en format électronique n'as cessé d'augmenter, ces formats sont généralement des phrases mais peuvent être aussi des paragraphes **[Mitra et al. 1997, Strzalkowski et al. 1998].**

Afin de progresser ces recherches, les travaux de **[Mahesh, 1997]**, **[Zhou et Hovy, 2006]** et **[Hu et al, 2007]** ont également réalisés des résumés automatiques pour des documents avec de nouvelles formes telles que les pages web, et les blogs.

Ainsi, différentes méthodes ont été développées au cours des trente dernières années pour produire automatiquement un résumé à partir d'un texte d'origine **[Minel 2002a]**.
Tous ces systèmes réalisés s'intéressent au résumé d'un seul document. L'objectif est d'améliorer ces systèmes dans le but est de résumer plusieurs documents avec un grand corpus d'informations non structurés.

4.2 Méthodes existantes du résumé automatique :

Le domaine de la recherche qui s'intéresse au résumé automatique est très vaste, ce qui engendre la naissance de deux types de méthodes **[Ihab Mallak, 2011]** :

- **Par extraction :** L'approche par extraction consiste en la sélection des unités (mots, phrases, paragraphes, etc.) censées contenir l'essentiel des informations que contienne le document et produire un extrait par assemblage de ces dernières.

L'avantage de cette méthode est de ne pas passer par une analyse en profondeur du texte, et de pouvoir fournir un résumé de façon plus simple sans devoir générer du texte. Les inconvénients portent souvent sur le manque de cohérence du résumé, et sur les mauvaises liaisons entre les différents segments textuels extraits juxtaposés qui peuvent modifier l'interprétation. Néanmoins, c'est une méthode très utilisée dans ce domaine.

- **Par abstraction :** Les systèmes produisant des résumés par abstraction sont fondés sur la compréhension du document et la génération d'un véritable texte grammatical et cohérent.

L'avantage de cette méthode se repose sur le fait que les résumés produits ressemblent à ceux des humains c'est-à-dire qu'ils sont compréhensibles et corrects. L'inconvénient majeur est de parcourir tout le texte ou la phrase pour faire le résumé.

5. Résumé automatique dessujets tendances de Twitter :

Le résumé automatique des sujets tendances nécessite l'accès aux données en temps réel sur Twitter. Ce dernier a favorisé des applications tierces en fournissant des interfaces de programmation pour son site Web depuis sa création. Grâce à une API basée sur HTTP, les utilisateurs peuvent effectuer une tâche par l'intermédiaire de l'interface utilisateur du site. Les recherches déjà faites dans ce domaine sont limitées à retourner 1500 tweets par requête. Par contre, s'ils veulent avoir plus de 1500 tweets, ils peuvent demander à être ajoutés à la liste blanche (whitelist) qui permet de recevoir un flux très important des mises à jour envoyé àTwitter en temps réel **[Kalucki 2009]**.

6. Travaux connexes :

Dans cette partie, nous allons parler des travaux qui s'intéressent au résumé automatique des tweets avec les différentes méthodes qu'ils utilisent.

6.1 L'approche de [Sharifi et al. 2010] :

Cette approche propose une nouvelle technique pour extraire les tweets liées aux sujets tendances de Twitter afin de faire un résumé automatique de ces messages courts dont le sens est plus proche du résumé humain.

Pour éviter la lecture manuelle des tweets et afin de comprendre le résumé et le contenu de chaque sujet, un nouvel algorithme a été développé, nommé PR (Phase Reinforcement). Ce dernier commence par les sujets tendances qui sont sous forme de phrases sur Twitter.

Au début, une requête sera envoyée à Twitter.com qui répond avec 1500 tweets qui contiennent la phrase recherchée.

De petites erreurs peuvent se produire durant le filtrage car au lieu d'extraire un contenu souhaité, des Spams seront pris et résumés, pour cela le Spam sera filtré à l'aide d'un classifieur Bayésien [**Kalita, 2002**] qui sert à supprimer les messages qui ne sont pas en anglais et les messages en double, car cette technique s'intéresse qu'aux messages en langue anglaise. Cet ensemble de messages s'appelle les tweets d'apprentissage.

L'idée principale est de construire un graphe acyclique qui ordonne tous les mots filtrés de l'ensemble de ces tweets .Le nœud racine du graphe contient la phrase de départ qui sera résumé. Les nœuds adjacents au nœud racine sont les mots qui se produisent immédiatement avant ou après la phrase de départ à l'intérieur de chaque tweet d'apprentissage. Ces mots adjacents sont placés soit avant ou après le nœud racine respectant l'ordre trouvé dans les tweets.

Une fois le graphe est construit, les nœuds sont pondérés en fonction de la fréquence d'apparitions respective des mots à partir de la racine. Un poids d'occurrence sera calculé pour chaque mot qui se produit M fois après la phrase de départ, et par conséquent son poids sera proportionnel à M.

Après la construction du graphe, l'algorithme est prêt pour générer des résumés, et commence à rechercher le meilleur résumé partiel en additionnant le poids de chaque chemin unique en partant du nœud racine à chaque nœud feuille. Le chemin du poids le plus

élevé est considéré comme le meilleur chemin de résumé partiel à partir du nœud racine, Pour le chemin du poids le plus élevé l'algorithme crée un nouveau graphe avec une nouvelle phrase racine qui contient tous les mots dans ce chemin. L'ensemble de tweets est alors filtré pour garder que les tweets contenants la nouvelle phrase racine. Ce processus est répété jusqu'à trouver le résumé final qui est les mots du chemin avec le poids le plus élevé.

Discussion : [Sharifi et al. 2010] comparent les résumés automatiques générés avec ceux produits par des humains pour chaque sujet tendance. L'algorithme PR apporte de meilleurs résultats quand un sujet a un motif de phrase dominante autour du thème central. Il est capable aussi d'isoler ces phrases dominantes de l'ensemble des tweets d'entrées afin de faire un bon résumé. Cela est particulièrement vrai pour les sujets avec un hashtag (#) qui rend la recherche facile pour l'utilisateur, mais si le sujet ne comporte pas le hashtag, l'algorithme PR n'est pas en mesure de générer une expression dominante autour du thème. Le résumé n'est pas fait en temps réel et uniquement en langue anglaise.

6.2 L'approche de [Chakrabarti et Punera, 2011]:

Cette approche appelée SUMMHMM se base sur le modèle de Markov caché (Hidden Markov model HMM) de façon que les paramètres du modèle sont accordés avec le type d'événement, et l'applique sur des évènements simples comme le séisme, le sport,…etc.

L'objectif est de faire un résumé pour les tweets d'un évènement qui est le football américain. Les messages choisis se répètent beaucoup et chacun se produit dans une période de temps quelconque et avec de langues différentes.

Ce travail complète les travaux de **[Petrovic, Osborne, et Lavrenko 2010]**, le processus réalisé se compose de deux étapes :

- Faire une conception modifiée en utilisant le modèle HMM qui segmente la chronologie d'évènement (des tweets courants) par rapport aux mots les plus utilisés, et chaque segment (ensemble de tweets) représente un sous évènement.

- Supprimer les segments en double et qui n'ont pas une relation avec le football américain, puis relier les autres pour avoir un résumé compréhensible.

Pour appliquer cette technique, une extraction des tweets ayant des mots hashtags comme les noms des joueurs ou les noms des équipes dans le même jour où le tournoi aura lieu, et si ce n'est pas le cas, le problème du vocabulaire est pris en compte.

Pour avoir de bons résultats, **[Chakrabarti et Punera, 2011]** proposent l'utilisation de trois algorithmes :

- **Premier algorithme** : appelé SUMMALLTEXT, considère le tweet comme un document, et applique une compression sur tout le corpus. Chaque tweet est associé avec un vecteur TF-logIDF (Term Frequency-Inverse Document Frequency) avec les mots qui le contiennent **[Baeza-Yates and Ribeiro-Neto 1999]**, ensuite, calcule la distance entre les tweets et sélectionne les plus proches.

- **Deuxième algorithme :** appelé SUMMTIMEINT, choisi des tweets dont chacun peut avoir une durée (la somme des durées est le temps de l'évènement). Cet intervalle de la somme est divisé en tailles égales, pour sélectionner dans chaque intervalle les mots clés afin de savoir quels sont les tweets intéressants.

- **Troisième algorithme :** appelé SUMMHMM, s'appuie sur l'algorithme de viterbri **[Rabiner ,1989]**. Chaque partie des segments précédents sera résumée puis les rassembler pour avoir un seul résumé.

Discussion : [Chakrabarti et Punera, 2011] sont concentrés sur le problème de la production des résumés automatiques en temps réel pour des évènements sur Twitter. Ils ont proposé une approche basée sur l'apprentissage d'une représentation sous-jacente de l'état caché d'un événement spécifique qui est le football américain. Néanmoins, cette approche ne peut pas s'appliquer sur les événements récurrents et structurés qui prennent beaucoup de temps pour expirer.

6.3 L'approche de [Liu et al. 2011] :

Dans cette approche, les auteurs ont exploré une variété de textes sources pour résumer les sujets Twitter y compris les tweets normalisés via un système dédié pour la normalisation, ainsi que les tweets qui contiennent des pages web pour l'intégration des différentes sources de textes. L'objectif est de générer un résumé textuel court qui peut bien introduire chaque tendance.

Au début, 5537 phrases de sujets différents ont été collectées durant une période de 70 jours. Ensuite, comme ils s'intéressent qu'aux tweets en langue anglaise, un filtrage a été fait pour supprimer tous les tweets des autres langues. Si un tweet intègre une URL d'une page web, le contenu de cette page sera récupéré. Pour chaque sujet le nombre de tweets est limité à 5000 et les pages web liées à 100. Les sujets seront divisés en deux groupes : les sujets généraux, et les sujets contenant des hashtags.

A l'aide de l'optimisation qui a été introduite par **[Gillick et al., 2009; Xie et al., 2009; Murrayet al., 2010]**, une minimisation de la redondance des tweets qui se répètent est importante.

Les tweets sont passés à travers un ensemble de prétraitements pour supprimer les caractères non ASCII, les caractères spéciaux, les émoticônes, les signes de ponctuation, les @, et les # hashtags et sont prétraités, classé par date. Les tweets sont normalisés en se basant sur le modèle de **[Liu et al., 2011]**afin de les mettre en langue anglaise standard.

Ils ont utilisé aussi un parseur HTML pour extraire le contenu de chaque page web et effectuer une segmentation aux phrases qui les contient **[Reynar and Ratnaparkhi, 1997]**, et les grouper dans des groupes similaires s'ils traitent le même sujet, et appliquer un système de compression.

Les IDF sont calculés à partir d'un grand corpus de textes pour éliminer les mots à faible IDF, les mots vides….etc.

Le but est d'extraire les n-grammes qui apparaissent fréquemment dans chaque sujet, ceux qui ont un poids lourd apportent plus d'informations importantes donc seront extraites.

Ce travail se base sur l'extraction d'un ensemble de concepts importants pour chaque sujet, et sélectionne une collection de phrases qui peuvent couvrir le nombre maximum de ces derniers tout en respectant la longueur spécifiée ; cela est réalisé avec l'ILP (Integer Linear Programming).

Ces concepts sont extraits de n-gramme (n=1, 2,3) à partir des documents d'entrée correspondant à chaque sujet, puis ils suppriment les :

1- N-grammes qui apparaissent qu'une seule fois dans ces documents.
2- N-grammes qui ont un mot composé avec la valeur d'IDF inférieur au seuil.
3- N-grammes qui sont inclus avec un ordre des n-grammes mais utilisent la même fréquence.

Ces filtres sont conçus pour exclure les n-grammes qui ne sont pas significatifs à l'ensemble de concepts.

A la fin, les auteurs concatènent les tweets originaux et les tweets normalisés avec leurs pages web liées comme entrée pour le système de synthèse à base de concept.

Discussion : [Liu et al.,2011] ont proposé d'explorer une variété de textes sources pour résumer les différents sujets tendances sur Twitter. Leur travail se repose sur l'optimisation basée sur le concept avec de multiples sources de saisie de texte pour générer les résumés. Ils ont comparé ce travail avec des résumés produits par des humains ainsi qu'avec la méthode automatique ROUGE. De bons résultats ont été portés lorsque la normalisation se fait sur les tweets en entrée, en plus, le contenu Web lié peut fournir des informations supplémentaires sur le sujet traité.

6.4 L'approche de [Wei et al., 2012]:

Dans cette approche, les auteurs ont fait un résumé pour des sujets par sous-thème en temps réel pour démontrer l'évolution rapide des tendances sur Twitter. Pour cela, ils ont classé et sélectionné les tweets saillants et diversifiés comme un résumé de chaque sous-thème

Cette approche consiste à modéliser et formuler le classement des tweets dans un modèle graphique de renforcement mutuel unifié, où l'influence sociale des utilisateurs et la qualité

du contenu des tweets sont prises en considération. Le traitement se fait pour des tweets de différentes langues. Cette approche se décompose en trois étapes :

- Ils ont effectué une segmentation thématique qui segmente le flux de tweets sur le sujet en sous-groupes thématiques en termes de temps d'affichage, dans lequel chaque groupe décrit un sous-thème.

- Les tweets dans chaque groupe sous-thème sont classés selon la saillance du tweet par renforcement du modèle de classement en profitant de la qualité du contenu des tweets et l'influence sociale des auteurs.

- Ils ont généré le résumé pour chaque sous-thème sur les résultats du classement des tweets en enlevant les tweets redondants au niveau de toute la question.

Discussion : [Wei et al., 2012] ont proposé de résumer un grand corpus de tweets qui s'évolue en temps réel. Ces derniers sont groupés en sous thèmes dans l'ordre chronologique et les classer pour produire un résumé selon leurs saillances. Dans les documents traditionnels, les tweets souffrent beaucoup de l'information inutile et le style d'écriture irrégulier, donc, la solution était d'utiliser un modèle graphique de renforcement mutuel unifié qui permet d'intégrer l'influence sociale des utilisateurs ainsi que la qualité du contenu des tweets qui montre une grande efficacité dans la mesure de la saillance du tweet. Par conséquent, l'approche proposée réalise des améliorations par rapport à LexRank et phrase graph. Les tweets qui sont retweetés ne sont pas pris en compte par ce modèle pour générer des résumés automatiques.

6.5 L'approche de [Jeffrey Nichols et al. 2012]:

Les auteurs ont réalisé un algorithme qui génère un résumé automatique d'un événement en utilisant uniquement des mises à jour de statut de Twitter comme une source.

[Jeffrey Nichols et al. 2012] se concentrent de résumer les événements sportifs, en particulier la Coupe du Monde (matchs de football), parce que chaque événement se déroule

sur une courte période de temps définie. Il existe une quantité importante de tweets sur chaque événement, et il y a la couverture de presse de chaque événement.

Les événements sportifs sont constitués d'une séquence d'instants, chacun d'eux peut contenir des mesures prises par les joueurs, l'arbitre, les fans,…etc. l'algorithme repose sur les utilisateurs de Twitter collectivement et s'intéresse aux moments importants dans l'événement et aussi les décrire.

Les spams, les tweets qui sont des réponses a d'autres messages, et ceux qui contiennent des URLs sont supprimés, un filtrage est réalisé pour récupérer que les tweets en langue anglaises, et les tweets qui contiennent des mots clés.

Ils ont utilisé un algorithme de détection de moments importants de cet événement qui repose sur le volume des messages par minute, la récupération des tweets commence au début du match et se termine à la fin.

Ils ont utilisé deux versions pour cet algorithme :

- **La version hors ligne**, il calcule le seuil pour l'ensemble des tweets de l'évènement. Par exemple, pour un match de football particulier, le seuil est calculé à partir des tweets enregistrés.
- **La version en ligne**, le seuil pour l'ensemble de tweets est calculé au fur et à mesure que les messages seront postés.

Puisqu'ils s'intéressent aux moments importants dans cet évènement, ils ont calculé le seuil et tous les points qui dépassent le seuil sont les tweets recherchés.

Discussion : [Jeffrey Nichols et al. 2012] ont proposé un algorithme pour le résumé d'un évènement important qui est la coupe du monde en temps réel sous forme de phrases, ces résumés peuvent être enchaînés pour produire un résumé de l'événement en quelques paragraphes. Cette approche peut s'appliquer aussi sur les événements à long terme tel que **TwitInfo**[17]pour générer des descriptions journalistiques qui ne sont pas couvertes par les journalistes. L'algorithme marche bien s'il n'y a pas une répétition dans les mis à jour des tweets, ce grand volume de tweets ne permet pas de produire un résumé efficace.

[17]Une plateforme pour l'exploration de Twitter en temps réel, et la recherche d'un événement à partir d'une requête.

6.6 L'approche de [Yosef Ardhito Winatmokoet Masayu Leylia Khodra, 2013] :

Cette approche propose une nouvelle méthode pour l'extraction et le résumé automatique des sujets tendance de Twitter.

Ils s'intéressent à l'extraction des tweets de 8 sujets différents en langue indonésienne chacune avec 300 messages et de régions différentes en Indonésie.

Après la récupération des tweets, ils utilisent une compression pour optimiser le corpus et le normaliser afin de réduire les erreurs orthographiques et syntaxiques. Ainsi, les tweets comportant des Urls, des émoticônes et les retweets ne seront pas pris en compte.

Comme les sujets tendances sont des fois des phrases, ces dernières ont besoin aussi d'un traitement comme ils sont présents dans tous les tweets. Pour cela, ils utilisent une tokénisation en remplaçant toute la phrase (du sujet tendance) par un jeton appelé « TOPIC ». Cela permet à l'algorithme de traiter le sujet comme un jeton.

Ils utilisent la méthode TF-IDF pour calculer la fréquence des termes comportant les messages filtrés. Ils modifient aussi l'algorithme de POS tagger[18] qui ne peut pas être appliqué sur les tweets à cause de l'informalité des mots abrégés, afin de garder des mots avec un sens compréhensible. Ensuite, faire un étiquetage pour chaque tweets afin de détecter tous les verbes et les noms ayant une relation avec le sujet tendance.

Discussion : [Yosef Ardhito Winatmoko et Masayu Leylia Khodra, 2013] ont proposé une méthode pour la production des résumés automatiques en langue indonésienne. Deux parties sont importantes, qui sont la catégorisation du sujet et la génération des explications pour ces sujets. Les tweets ont été compressés pour générer des résumés sous forme de phrases puis les classifier en sous thèmes. Ce qui pose un problème pour l'algorithme utilisé est que les tweets sont bruyants. Ils ont conclu que les sujets avec un grand corpus de messages donnent de mauvais résultats par rapport aux sujets avec un nombre de messages restreint.

[18] https://en.wikipedia.org/wiki/Part-of-speech_tagging

7. Etude comparative :

Dans cette partie, nous allons comparer les travaux du résumé automatique étudiés précédemment en utilisant un tableau comparatif, en se basant sur des critères de comparaison.

7.1 Critères de comparaison :

- **Données en entrée :** Chaque approche présente le nombre de tweets qui sont extraits et le type de sujet tendance ainsi que les informations extraites à partir des tweets. Ces informations peuvent représenter une phrase, un mot, des hashtags, des URLs et des liens vers d'autres pages web.
- **La langue** : la langue des tweets extraits.

- **Le prétraitement :** Afin de faire le résumé automatique des tweets, une étape de prétraitement est nécessaire. Cette étape permet de nettoyer les tweets pour ne garder que les mots importants.

- **La méthode :** Pour chaque approche, nous allons citer la méthode utilisée dans la section 4.2.

- **Algorithme utilisé :** Pour chaque approche nous allons spécifier l'algorithme utilisé.

- **Méthode statistique :** Calcule la fréquence des mots qui contiennent les tweets et évalue l'importance des mots clés. La méthode TF-IDF est la plus courante : c'est une méthode qui est utilisée pour déterminer dans les tweets les mots les plus favorables dans l'expression recherché.

Le calcul de la valeur TF-IDF s'obtient en multipliant la fréquence du mot qui est son nombre d'occurrences en la fréquence inverse du corpus qui est le nombre d'occurrences de tous les autres mots moins fréquents. Un résultat avec une grande valeur implique une forte relation avec le sujet tendance.

- **Temps réel :** Pour approche nous vérifions si elle collecte les tweets à l'instant de la publication et donc résume les tendances en temps réel ou non.

- **Mesure d'évaluation :** L'évaluation des travaux précédents se fait à partir d'une comparaison entre le résumé automatique de chaque approche, le résumé humain, et la méthode automatique ROUGE[19].

- **Données en sortie :** Le nombre de tweets représentant le résumé automatique.

7.2 Tableau comparatif :

[19]**Méthode automatique ROUGE :** compare l'occurrence des mots entre les résumés produits automatiquement et les résumés humains.

Approches	Données en entrée			Langue	Prétraitements	Méthode	Algorithme utilisé	Méthodes statistiques	Temps réel	Mesures d'évaluation	Données en sortie
	Nombre de tweets	Informations extraites à partir des tweets	type								
Sharifi et al. 2010	1500 tweets	phrases	Tous les sujets tendance	anglais	Non specifié ou non utilisé	Extraction	PR (Phase Reinforcement)	Calcul de poids d'occurrences	Non	Méthode automatique ROUGE et humaine	1 tweet
Chakrabarti et al. 2011	1,8K tweets	Tweets avec @ et mots clés	Football américain	Différentes langues	HMM Hidden node	Extraction	- SUMMALLTEXT - SUMMTIMEINT - SUMMHMM	TF-logTDF	Non	Non spécifié	10 jusqu'à 70 tweets
Liu et al. 2011	1,7K tweets	Tweets avec des URLs, hashtags et @	Sujets différents	anglais	Modèle d'optimisation et de transformation	Extraction	Version récente de PR	TF-TDF	Non	Méthode automatique ROUGE et humaine	2 à 3 tweets
Wei et al., 2012	10K tweets	Les tweets et les retweets	Tous les sujets tendance	anglais	HMM hiden node modifié	Extraction	Algorithme de classement basé sur des graphes	Non spécifié	Oui	Méthode automatique ROUGE	10 tweets
Jeffrey Nichols et al. 2012	4K tweets	Les tweets avec des mots clés qui sont postés entre début et fin du match	Coupe de monde	Différentes langues	Non specifié ou non utilisé	Extraction	Algorithmes en ligne/ hors ligne	TF-IDF	Oui/Non	Méthode automatique ROUGE	3 tweets
Yosef Ardhito Winatmoko et al., 2013	2400 tweets	Des phrases qui ont relation avec le sujet tendance	8 sujets différents	Indonésien	Modèle de compression et normalisation	Extraction	Algorithme POS tagger modifié	TF-IDF	Non	Non spécifié	Chaque sujet a un nombre de phrases

Tableau 2.1 : Tableau de comparaison entre les approches

7.3 Synthèse :

A partir de la comparaison des différents travaux présentés dans le tableau précédent, nous constatons que :

- L'approche de **[Chakrabarti et al. 2011]** et **[Jeffrey Nichols et al. 2012]** ne peut pas s'appliquer sur tous les sujets tendances car les auteurs ont fait l'extraction que pour les tweets de la coupe du monde et le football américain.

- La langue courante est la langue anglaise sauf pour les approches de **[Chakrabarti et al. 2011] et [Jeffrey Nichols et al. 2012]** que les auteurs utilisent de différentes langues, et **[Yosef Ardhito Winatmoko et al., 2013]** utilisent la langue indonésienne.

- Toutes les approches utilisent soit un algorithme déjà utilisé comme l'approche de **[Liu et al. 2011] et [Yosef Ardhito Winatmoko et al., 2013],** ou un nouvel algorithme développé par les auteurs.

- La méthode statistique utilisée dans toutes les approches est la méthode TF-IDF.

- La méthode du résumé automatique utilisée dans toutes les approches et celle par extraction.

- L'extraction des tweets ne se fait pas en temps réel pour toutes les approches sauf pour l'approche de **[Wei et al., 2012].**

- Le résumé automatique produit est souvent des tweets sous forme de phrases pour toutes les approches.

- La méthode d'évaluation utilisée est soit le résumé humain ou la méthode automatique ROUGE.

8. Conclusion :

Dans ce chapitre, nous avons parlé de l'importance des sujets tendances sur Twitter, du résumé automatique en général, ensuite, l'utilisation de ce dernier pour ces messages courts afin d'avoir une définition du sujet. Nous avons exploré aussi quelques travaux dans le domaine de l'extraction et le résumé automatique des tweets, et les méthodes utilisées.

Bien que toutes les approches ont relevé le défi qui est le résumé automatique mais plusieurs problèmes ont été trouvés comme :

- Les tweets non structurés par cause des mots abrégés, des émoticônes et des URLs sont hiérarchisés et courts.
- L'existence de plusieurs langues dans les sujets tendances engendre une difficulté pour générer un résumé automatique
- Les messages courts sont pleins d'erreurs orthographiques et syntaxiques et avec un volume très grand d'informations ce qui rend difficile la compréhension du sens des tweets.

Dans le prochain chapitre, nous allons présenter notre approche de résumé automatique des tweets liés aux sujets tendances en se basant sur les travaux déjà présentés.

Chapitre III :

RATTR : RÉSUMÉ AUTOMATIQUE DES TWEETS LIÉS AUX TENDANCES DE TWITTER

CHAPITRE III : RATTR : RESUME AUTOMATIQUE DES TWEETS LIES AUX TENDANCES DE TWITTER

1. Introduction :

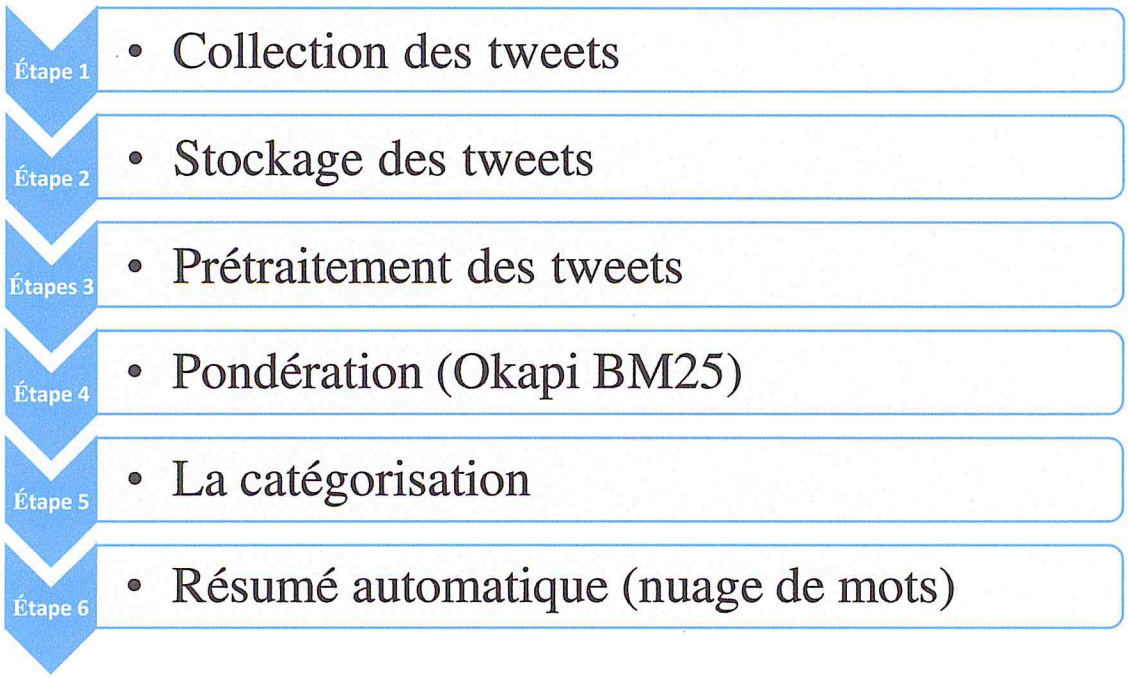
Dans ce chapitre, nous allons proposer notre approche de résumé automatique des tweets des dix tendances affichées par Twitter en temps réel. En se basant sur les travaux cités dans le chapitre précédent par rapport aux critères d'évaluation, nous avons collecté les tweets saillants de tous les sujets tendances d'un pays spécifié, et avec trois langues différentes afin de les analyser, et les classifier pour faciliter la tâche du résumé ainsi qu'utiliser une méthode statistique qui définit les mots les plus importants. Nous avons extrait ces messages à l'aide des informations qui les contiennent et avoir en sortie un résumé simple compréhensible pour les humains de tous les évènements sera présents en temps réel.

2. L'approche du résumé automatique des tweets :

Notre approche appelée RAT^{TR} (Résumé automatique des tweets en temps réel) consiste à extraire les tweets des dix sujets tendances de Twitter en temps réel afin de les traiter et éviter la lecture manuelle, pour cela, effectuer un résumé automatique sur le contenu des tweets qui compose chaque tendance.

Tout d'abord, il faut collecter les tweets et les stocker. Ensuite, effectuer un prétraitement sur tout le corpus extrait, en utilisant les entités nommées, en les classifions selon trois entités. Enfin, le calcul du poids de chaque mot est nécessaire pour savoir où sont les plus saillants dans le but de présenter un résumé sous forme d'un nuage de mots.

Notre approche procède en six étapes importantes qui sont :



3. La collecte des tweets :

Cette phase consiste à extraire les tweets des évènements de Twitter en temps réel.

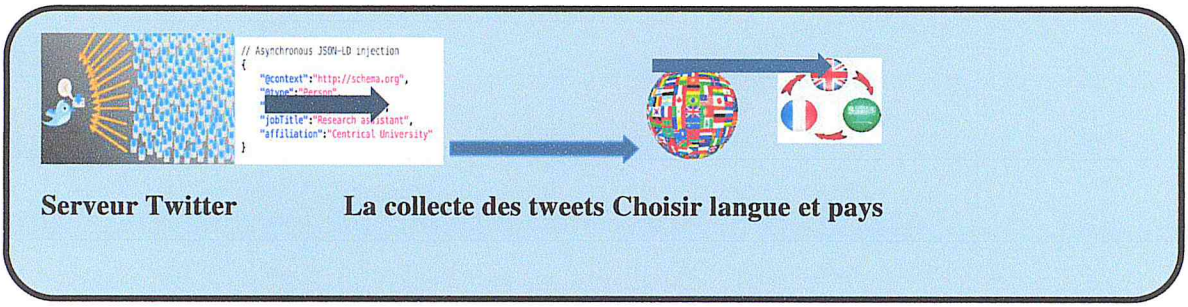


Figure 3.1 : La phase de la collection des tweets

3.1 L'accès àTwitter :

Cette première étape consiste à collecter les tweets afin de les faire passer par une analyse approfondie, nous nous intéressons aux tweets qui font partie des dix sujets tendances de Twitter. Pour cela nous avons créé une application sur le réseau socialpour récupérer les clés développeurs (*consumer key, consumer secret, accesstoken et access secret*)qui nous permettent d'extraire ces messages courts via les noms les tendances présentes sur Twitter.

3.2 La langue :

Les trois langues les plus dominantes dans le monde et aussi les plus utilisées dans Twitter sont l'anglais, le français et l'arabe. Nous avons filtré les tweets selon ces trois langues pour les traiter.

3.3 Le pays :

Le Woeid[20]c'est des identifiants de 32 bits, Twitter définit chaque pays avec un Woeid, nous avons collecté les tweets des sujets tendances selon un pays bien spécifié tout en reposant sur les woeid de chacun de ces pays.

[20] https://en.wikipedia.org/wiki/GeoPlanet

4. Le prétraitement :

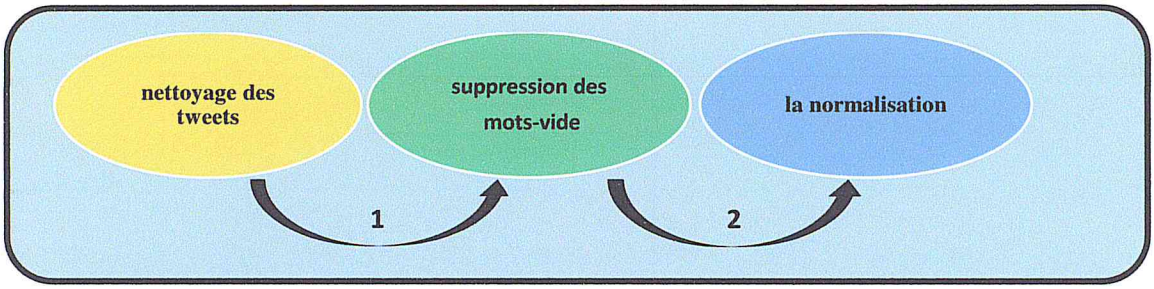


Figure 3.2 : la phase du prétraitement

Les tweets filtrés contiennent des informations incohérentes, mal orthographiées, et bruyantes. Il est nécessaire d'analyser le contenu de chaque tweet en supprimant les caractères spéciaux, les messages en double et les mots vides.Nous avons appliqué un processus de nettoyage et filtrage sur ce corpus appelé le prétraitement qui est essentiel afin d'obtenir un document bien structuré et compréhensible.

Pour cette phase il faut passer par le découpage du texte,la suppression des mots vides et enfin la normalisation.

Ce processus contient les étapes suivantes :

4.1 Nettoyage des tweets :

Cette partie comprend plusieurs étapes pour supprimer :

- **Les émoticônes :** Les émoticons, ce sont une suite de caractères qui représentent un visage penché à 90 degrés. Par exemple : ☺ ☹ … etc.

Nous avons supprimé toutes les émoticônes qui figurent dans les tweets collectés.

- **La ponctuation et les caractères spéciaux :** la ponctuation précise le sens de la phrase. Elle sert à fixer les rapports entre les propositions et les idées.

Les principaux signes de ponctuation sont : le point (.), le point d'interrogation (?), le point d'exclamation (!), le point-virgule (;), les points de suspension (…), les deux points (:), la virgule (,), les guillemets (« »), le tiret (-) et les parenthèses [()].

Les caractères spéciaux suivants : ",#,$,%,&, ',(,), *, +, -, /,<,=,>, @, [,\,], ^, _, {,|,}, ~.
Nous avons supprimé tous les caractères précédents pour avoir un document seulement avec des mots.

- **Les Urls :** C'est des liens qui guide à d'autres pages ou des sites web, qui a leurs tours sont plein d'autres informations. Twitter raccourcit ces urls à 19 caractères pour minimiser leur longueur. Nous avons éliminé ces urls pour ne pas traiter les données qui les contiennent.

4.2 **Les mots vides :** Est un mot qui ne doit pas être indexé, qu'il soit mot grammatical ou mot lexical. Ils sont alors souvent regroupés dans un « anti-dictionnaire » ou une « stop-list » ou une liste de « stopwords ». Il est généralement admis que ces mots très fréquents (environ la moitié des occurrences d'un texte) ne sont pas à indexer, car ils ne sont pas informatifs, et ils augmentent énormément la taille de l'index.

Exemple :« le », « la », « les », « de », « pour », « en », « avec ».

4.3 **La normalisation :** Cetteétape est très importante et consiste àsimplifier et analyser tous le document et son contenu, et appliquer ce concept pour avoir un lexique harmonisé ayant la meilleure couverture possible en diminuant le nombre d'erreurs de reconnaissance dues aux mots hors vocabulaire.Nous avons traité aussi les problèmes des abréviations, toutes ces opérations sont expliquées comme suit :

4.3.1 La casse des tweets : Nous avons converti tous les mots en majuscule qui composent les tweets vers le minuscule pour les rendre uniforme.

4.3.2 Les abréviations : L'écriture courante dans les tweets est celle avec les abréviations, Afin d'éviter les problèmes des abréviations, nous avons prédéfini manuellement une liste des

abréviations qui puisse être écrite par les utilisateurs sur Twitter afin de les remplacer par les mots corrects syntaxiquement.

4.3.3 La lemmatisation : La lemmatisation est une analyse lexicale qui permet de regrouper les mots d'une même famille ensemble : c'est un regroupement par lemme. Chaque mot à une forme canonique (forme racine) et des formes fléchies (différentes occurrences possibles). Ces dernières sont toutes les déclinaisons qu'une entité peut prendre : verbes à l'infinitif / conjugué, mots au singulier / pluriel, déclinaisons masculines / féminines, etc....

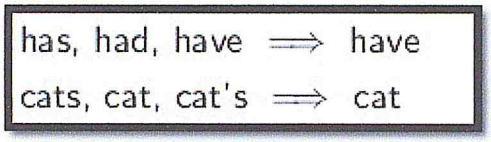
Exemple : Cet exemple montre le lemme des différents mots en français présents ci-dessous :

Lemme principal retenu	Variantes de mot détecté
Analyser	analyser
	analysez
	analyses

Ou encore pour la langue arabe :

بكرته	ليكتبها
بِكُرَتِه	لِيَكْتُبَها
بـ/ كرة/ ه	لـ/ يكتب/ ها

Pour la langue anglaise :



Nous avons utilisé l'algorithme de **[Ahmet Aker, 2010]** pour la langue française, cet algorithme de lemmatisation sert à définir le lemme des mots français. La bibliothèque qu'il utilise se base sur la bibliothèque OpenNLP et comporte un dictionnaire avec tous les mots existants.

4.3.4 La stemmatisation (la racination) :

La stemmatisation (en anglais stemming) est le processus d'élimination de suffixes des mots afin d'obtenir leurs racines communes. Cela permet de générer leurs formes de base.

Exemple : computers, computing, computation ⟶ comput.

Nous avons utilisé l'algorithme de **[Khoja and Garside, 1999]** pour la langue arabe, le principe de cet algorithme consiste à supprimer le plus long suffixe et préfixe. Ensuite, le résultat sera comparé avec des motifs verbaux et nominaux pour l'extraction de la racine, et il fait appel à plusieurs fichiers de données linguistiques. Il est basé sur un dictionnaire de 200.000 mots.

5. Le stockage dans la base de données :

Après avoir passé par toutes les étapes du prétraitement, les tweets seront stockés dans une base de données. Nous avons implémenté cette base de données pour stocker les tweets analysés et les utiliser comme une source de classification dans la prochaine phase.

CHAPITRE III : RATTR : RESUME AUTOMATIQUE DES TWEETS LIES AUX TENDANCES DE TWITTER

5.1 Les règles de gestion :

Notre application comporte cinq classes principales permettant d'identifier les tweets des tendances récupérés.

- Un utilisateur est identifié par un « Idutilisateur », et possède un « Nomutilisateur », et un « mdp ».
- Un utilisateur possède 4 clés d'authentification sont : Cosumer Key, Consumer Secret, Access Token et Access Secret. Il peut récupérer le nom de la tendance et sa localisation à partir de ses clés.
- Un utilisateur peut récupérer un ou plusieurs tweets.
- Une tendance est identifiée par un « Idtendance », et le « nomtendance », et se compose de 100 tweets au maximum.
- Un tweet est identifié par un « Idtweet », le « contenu » et la « langue ». La récupération des tweets se fait par rapport à la langue de ces messages.
- Un pays est identifié par un « Woeid », et le « nompays ». Chaque pays comporte 10 tendances.
- Un terme est identifié par un « IdTerme », un « poids » et une « categorie ». Une tendance contient un ou plusieurs termes.

Le diagramme des cas d'utilisations représentant notre application est le suivant :

CHAPITRE III : RATTR : RESUME AUTOMATIQUE DES TWEETS LIES AUX TENDANCES DE TWITTER

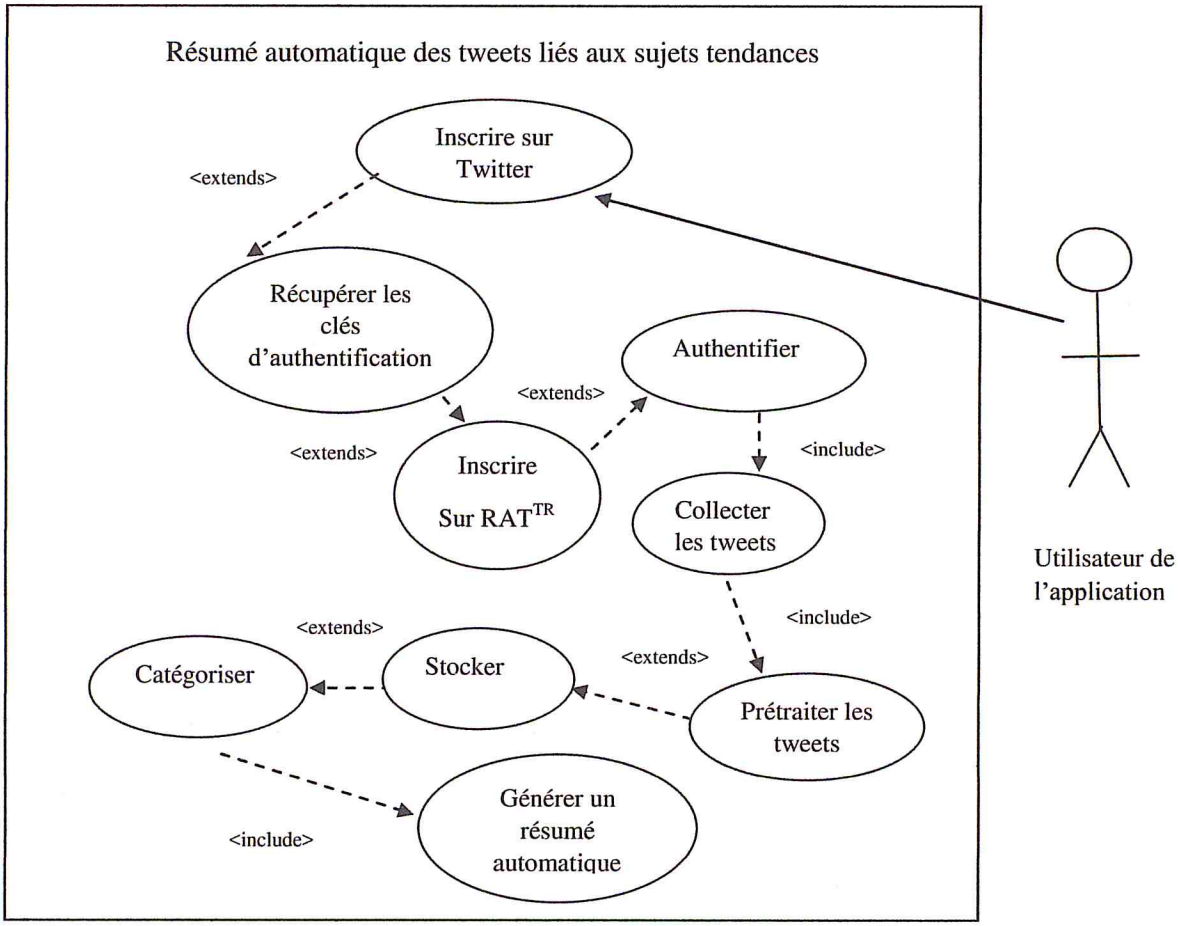


Figure 3.3 : Diagramme de cas d'utilisation

Le diagramme de classes représentant notre application est le suivant :

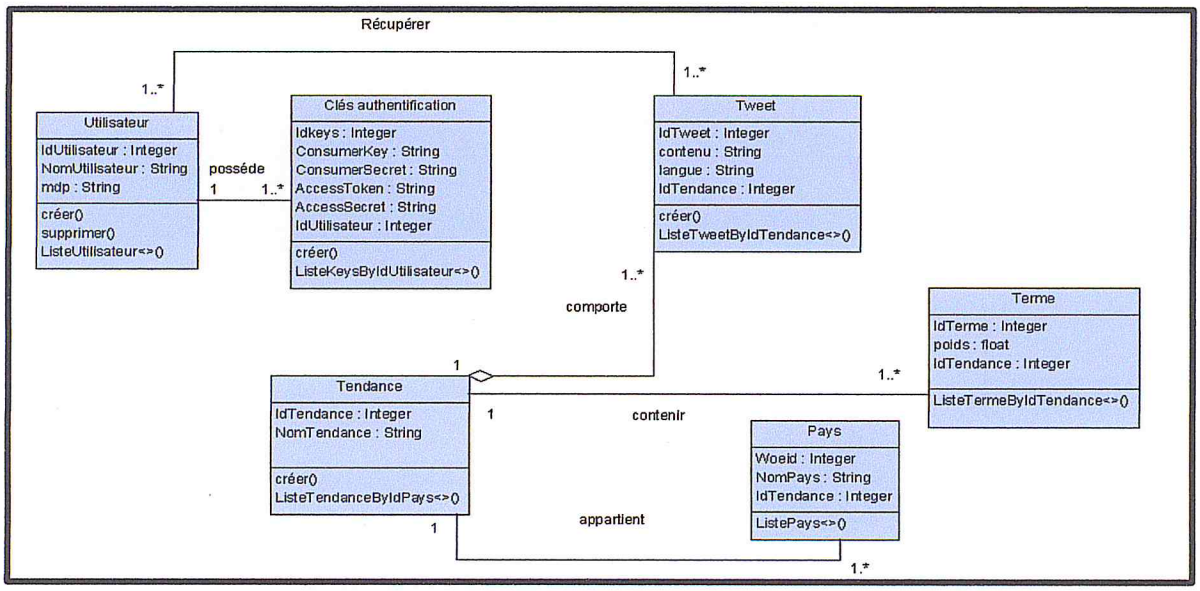


Figure 3.4 : Diagramme de classes

5.2 Les tables relationnelles :

Utilisateur (IdUtilisateur, NomUtilisateur, mdp)

Tendance (IdTendance, NomTendance)

Clés authentification (IdKeys, ConsumerKey, ConsumerSecret, AccessToken, AccessSecret, IdUtilisateur*)

Tweet (IdTweet, contenu, langue, IdTendance*)

Terme (IdTerme, poids, IdTendance*)

Pays (Woeid, NomPays, IdTendance*)

Récupérer (Idutilisateur*, Idtweets*)

6. La pondération :

Dans cette phase, nous avons calculé la pondération de chaque mot contenant les tweets. Les messages de chaque tendance sont considérés comme un seul document.

La pondération permet de caractériser non seulement la présence ou l'absence de termes dans le document, mais également leur importance relative pour décrire le contenu de chaque tendance.

Pour cela, nous avons choisi Okapi BM25[21] **[Robertson et al., 1998]** qui est une méthode de pondération utilisée en recherche d'informations. Cette pondération a initialement été proposée comme modèle de similarité dans un cadre probabiliste **[Robertson et al., 1998].**Elle repose sur le principe de classement probabiliste*(PRP, Probability Ranking Principle)*. Le modèle Okapi peut ainsi être vu comme un TF-IDF prenant mieux en compte la longueur des documents[22].

[21] https://fr.wikipedia.org/wiki/Okapi_BM25
[22] http://www.aclweb.org/anthology/F12-2007

La définition du modèle est dans l'équation (1) qui indique le poids du terme t dans le document d (l'ensemble des tweets d'une tendance) :

$$w_{BM25}(t,d) = TF_{BM25}(t,d) * IDF_{BM25}(t) = \frac{tf(t,d)*(k_1+1)}{tf(t,d)+k_1*(1-b+b*dl(d)/dl_{avg})} * \log\frac{N-df(t)+0.5}{df(t)+0.5}$$

Où :

Tf(t,d) : est le nombre d'occurrence du terme t dans le document d.

Df(t) : la fréquence du document d.

k1 = 2, b = 0.75 sont des constantes.

dl : la longueur du document représentée par le nombre de ses termes.

dl_{avg} *: la longueur moyenne des documents.*

N : le nombre de documents.

La partie T FBM25 est dérivée d'un modèle probabiliste de la fréquence des termes dans les documents, le modèle 2-Poisson de Harter**[Spärck Jones et al., 2000].**

La partie IDFBM25 est une simplification d'une formule dérivée du PRP **[Spärck Jones et al., 2000]**, théoriquement optimale, mais nécessitant des données d'apprentissage.

Après avoir calculé la pondération de chaque document, les mots ordonnés selon leurs importances.

Les mots ayant une valeur supérieure, ils apportent plus d'informations que ceux avec une petite valeur.

Pour notre cas nous avons considéré chaque tweet comme un document afin d'avoir des résultats de poids plus précis. Nous avons calculé le total de poids de chaque mot dans les documents ou il apparaisse pour avoir le résultat de son poids.

7. La catégorisation par entité nommée :

Les entités nommées constituent un champ de recherche très actif depuis de nombreuses années. Elles sont depuis longtemps considérées comme un point central dans de nombreuses applications mettant en jeu des notions comme la compréhension, la recherche sémantique, etc.

7.1 Les entités nommées :

Les entités nommées sont des unités textuelles particulières, « saillantes » sur le plan sémantique : noms de personnes, de lieux, d'organisations, dates, unités monétaires, pourcentages [Maud Ehrmann, 2008].

7.2 La reconnaissance des entités nommées :

La REN (reconnaissance des entités nommées) est une sous-tâche de l'activité d'extraction d'information dans des ensembles de documents. Elle consiste à rechercher des objets textuels (c'est-à-dire un mot, ou un groupe de mots) catégorisables dans des classes telles que noms de personnes, noms d'organisations ou d'entreprises, noms de lieux, quantités, distances, valeurs, dates, etc[23].

Après l'étape de la pondération, nous avons classifié ces mots en quatre sous classes. Ces classes sont les suivantes : les faits, les noms et les lieux.

- **Les faits :** C'est les verbes qui expriment des actions faites par une personne quelconque.
- **Les noms :** Cette classe regroupe tous les noms propres, les noms d'organisations qui peuvent se rapporter à des notions plus techniques comme les noms de maladie, ou de phénomènes naturels, etc…
- **Les lieux :** C'est les noms des endroits d'une ville, région, ou pays qui sont mentionnés par les utilisateurs dans un tweet.
- **Les noms d'utilisateurs :** C'est les noms des personnes qui ont publié les tweets.

[23] https://fr.wikipedia.org/wiki/Reconnaissance_d%27entit%C3%A9s_nomm%C3%A9es

8. Résumé automatique (nuage de mots) :

Les *nuages de mots*, appelés également *nuages de tags* ou *tag clouds*, sont de plus en plus utilisés sur les sites web actuels. Il s'agit de figures composées de mots placés de façon à obtenir une forme.

Nous avons utilisé le nuage de mots pour présenter le résumé automatique de chaque tendance.

Les mots composant le nuage sont obtenus en analysant le contenu de chaque document (les tweets de chaque tendance) : le nuage de mots contient en fait les mots les plus utilisés dans la tendance, et la taille des mots dans le nuage donne une indication de leur importance dans le contenu (nombre d'apparitions de chaque mot dans le contenu). Les nuages de mots permettent ainsi aux utilisateurs de connaitre en un coup d'œil les mots clés du sujet tendance.

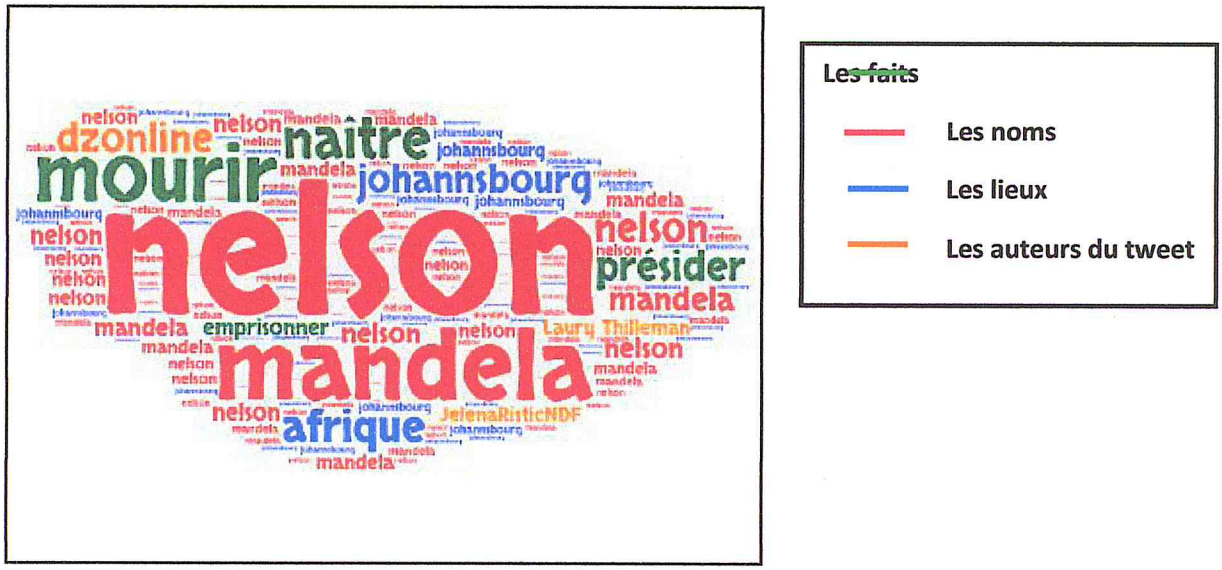


Figure 3.5 : Exemple d'un nuage de mots

Le résumé automatique de chaque tendance se fait par rapport aux quatre classes qu'on a citées dans la figure 3.5.

CHAPITRE III : RATTR : RESUME AUTOMATIQUE DES TWEETS LIES AUX TENDANCES DE TWITTER

Nous avons généré un résumé automatique de chaque tendance en utilisant les quatre classes (les faits, les lieux,les noms, et les noms d'utilisateurs).

Chaque classe est représentée par une couleur différente, ainsi que les mots qui se répètent beaucoup et qui apportent une information de plus au résumé seront présentés avec des caractères plus grands.

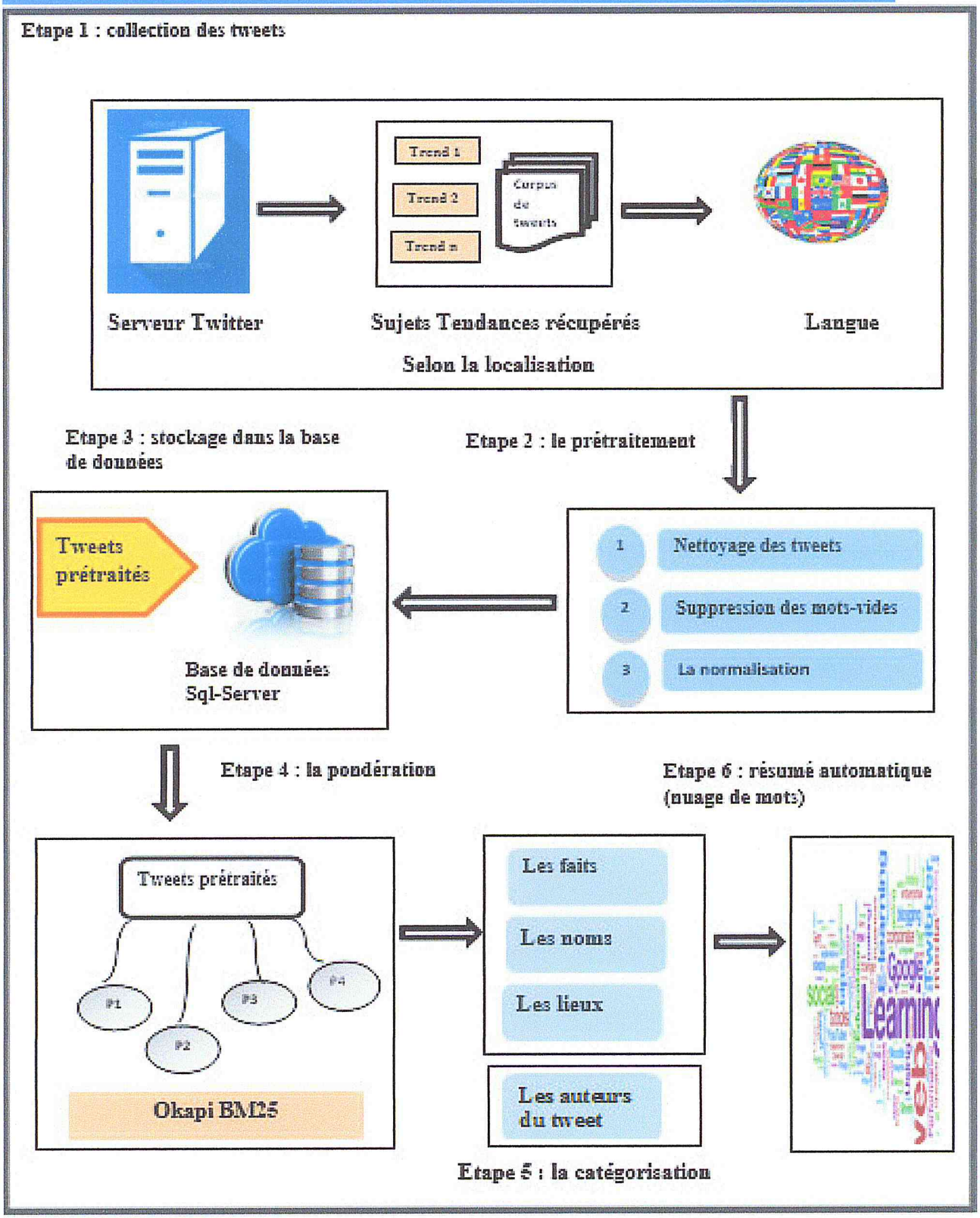


Figure 3.6 : Le schéma général du résumé automatique des tweets

9. Conclusion :

Dans ce chapitre, nous avons présenté une nouvelle approche de résumé automatique des sujets tendances sous forme d'un nuage de mots en se basant sur les faits, les noms propres et les lieux.

Les tweets extraits sont de différentes langues : français, anglais et arabe. Cette approche se fait en temps réel et constitue de plusieurs étapes. La première consiste à collecter les tweets des sujets émergents, les nettoyer pour garder les mots essentiels qui aident à faire un bon résumé. Ensuite, calculer la fréquence d'apparition de chaque mot, et les classifier en construisant un résumé automatique.

Dans le chapitre suivant, nous allons présenter l'implémentation et les tests de cette approche.

Chapitre IV : IMPLÉMENTATION ET TESTS

CHAPITRE VI : IMPLEMENTATION ET TESTS

1. Introduction :

Dans ce chapitre, nous allons présenter la partie pratique qui constitue une mise en œuvre d'une plateforme pour notre approche qui est le résumé automatique des sujets tendances de Twitter en temps réel.

Nous commençons par introduire les outils de développement utilisés pendant la création de notre application, ensuite nous allons comparer notre approche avec d'autres déjà vues dans la partie état de l'art.

2. Environnement de développement :

Pour la réalisation de notre application, nous avons adapté cet environnement de développement : le langage de programmation Java 1.7, l'éditeur de texte Eclipse Mars 4.5.2 et le système de gestion de base de données SQL server. Nous avons utilisé plusieurs bibliothèques qui sont définies ci-dessous :

- **Twitter4J 4.0.4**[24] **:** Twitter4J est une librairie Java permettant d'intégrer facilement l'API Twitter dans toute application Java. La librairie propose différentes classes et méthodes permettant de manipuler les méthodes qu'offre l'API Twitter. Pour utiliser cette librairie, il suffit de télécharger un fichier au format "jar" et de l'ajouter au classpath de l'application JAVA. La JavaDoc de la librairie permet une prise en main rapide et facile de cette librairie.

- **Apache lucene 5.4.0**[25]**:** Est une librairie qui se concentre surtout sur l'indexation et la recherche. Elle comporte plusieurs classes importantes, parmi celles qu'on a utilisées : Query, Analyzer…etc.

- **Opennlp v3.0.1**[26] **:** La bibliothèque Apache OpenNLP est une boîte à outils qui se base sur l'apprentissage et le traitement de texte en langage naturel.Elle prend en charge les taches suivantes : la segmentation, la tokenisation des phrases, l'extraction des entités nommées.
- **JAWS v1.2 (Java WordNet Search)**[27] : L'API Java de recherche pour WordNet (JAWS) est une API qui fournit des applications Java avec la possibilité de récupérer des données à partir de la base de données WordNet[28].
- **JWNL v1.1 (Java WordNet Library)**[29] **:** Est une API Java pour accéder au dictionnaire naturel WordNet. WordNet est largement utilisé pour développer des applications de la NLP (Natural Language Processing).
- **Sql Server Management Studio** (SSMS) **2014**[30] **:**Est un environnement intégré pour l'accès, la configuration, la gestion, l'administration et le développement de tous les composants de SQL Server. SSMS combine un large groupe d'outils graphiques avec un certain nombre de riches

---

[24] http://twitter4j.org/ko/
[25] https://lucene.apache.org/core/downloads.html
[26] https://opennlp.apache.org/
[27] http://wordnet.princeton.edu/wordnet/download/
[28] **WordNet :**est une base de données lexicale, Son but est de répertorier, classifier et mettre en relation de diverses manières le contenu sémantique et lexical de la langue anglaise
[29] http://jwordnet.sourceforge.net/
[30] https://msdn.microsoft.com/fr-fr/library/mt238290.aspx

éditeurs de script pour fournir aux développeurs et aux administrateurs de tous les niveaux de compétence accès à SQL Server.

- **Microsoft Windows Server 2012 r2**[31] : Est un système d'exploitation orienté service, anciennement connu sous le nom de code Windows Server 8.
- **Vmware Workstation 12**[32]: Ce logiciel est un virtualisateur de machine. Il permet de faire fonctionner un système d'exploitation virtuel sur une machine (en plus du système présent) mais non de le créer.
- **OpenCloud v0.3**[33] : OpenCloud est une bibliothèque Java qui peut être utilisé pour générer un nuage de tags sur un site web ou application.
- **JRouge**[34]: est un outil d'évaluation développé en java pour les résumés de textes.

- **Architecture de l'application :**

La figure suivante montre l'architecture de notre application :

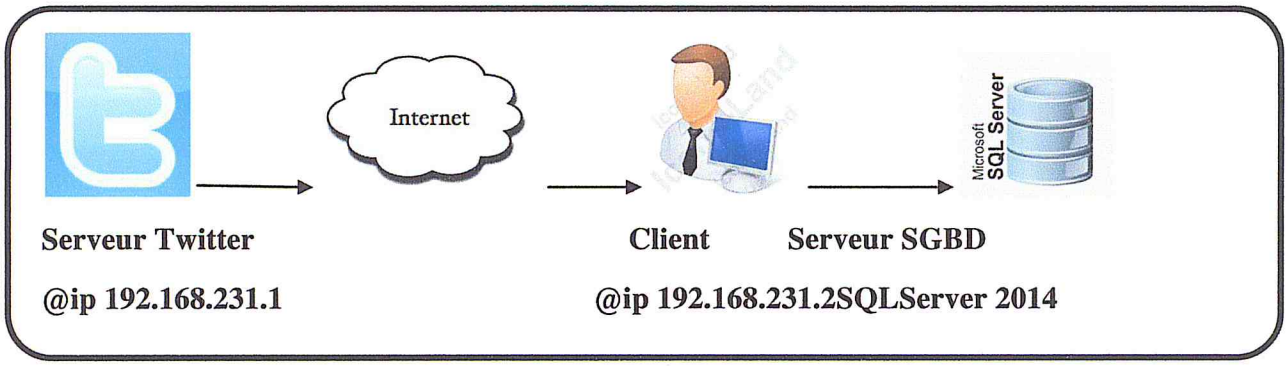


Figure 4.1 : L'architecture de l'application

- **API Twitter**[35] **:**

Une API (Application Programming Interface) est une série de méthodes mise à disposition par un site à des développeurs tiers, leur permettant d'utiliser certaines fonctionnalités ou d'accéder à des données du site.

[31] https://msdn.microsoft.com/fr-fr/windowsserver2012r2.aspx
[32] https://msdn.microsoft.com/fr-fr/windowsserver2012r2.aspx
[33] http://www.java2s.com/Code/Jar/o/Downloadopencloud03sourcesjar.htm
[34] https://bitbucket.org/nocgod/jrouge/wiki/Home
[35] http://www.erwanlenagard.com/general/tutoriel-utiliser-lapi-twitter-pour-collecter-des-tweets-sans-coder-avec-talend-1029

Twitter dispose de plusieurs APIs permettant de requêter sa base de données, mais aussi de construire des services au-dessus de sa plateforme. Ces APIs sont particulièrement riches en retournant presque une centaine de variables par requête. Les données concernent les tweets (date de publication, le texte du message, etc.), l'auteur (date de création du compte, pseudo…), les entités contenues dans les messages (hashtags, mentions, urls…) et des informations de localisation (pays, timezone, longitude / latitude).

Pour accéder aux données de Twitter, les APIs peuvent être classés en deux types en fonction de leur méthode de conception et d'accès :

❖ **API REST :** sont basés sur l'architecture REST maintenant couramment utilisés pour la conception des API Web. Ces API utilisent la stratégie d'attraction pour la récupération de données. Pour recueillir des informations d'un utilisateur doit explicitement la demande. Le nombre de requêtes est limité à 450 demandes toutes les 15 min.

❖ **API STREAMING :** fournit un flux continu de l'information publique de Twitter. Ces API utilisent la stratégie de pression pour la récupération de données. Une fois la demande de renseignements est faite, l'API streaming fournit un flux continu de mises à jour sans autre intervention de l'utilisateur.

Ils ont de différentes capacités et limites à l'égard de ce qui est de combien d'informations peuvent être récupérées. Le Streaming API a trois types de paramètres :

a. **flux public (Public streams) :** Ce sont des flux contenant les tweets publics sur Twitter.
b. **Les flux de l'utilisateur (User streams) :** Ce sont les flux mono-utilisateur, avec pour tous les tweets d'un utilisateur.
c. **Site flux (Site streams) :** Ce sont des flux multi-utilisateurs destinés à des applications qui accèdent aux tweets de plusieurs utilisateurs.

3. Présentation du prototype :

La première étape de RAT^{TR} consiste à s'authentifier en fournissant le nom d'utilisateur et le mot de passe, qui seront identifiés par les clés développeurs stockées dans la base de données.

La figure suivante montre l'interface graphique de l'authentification :

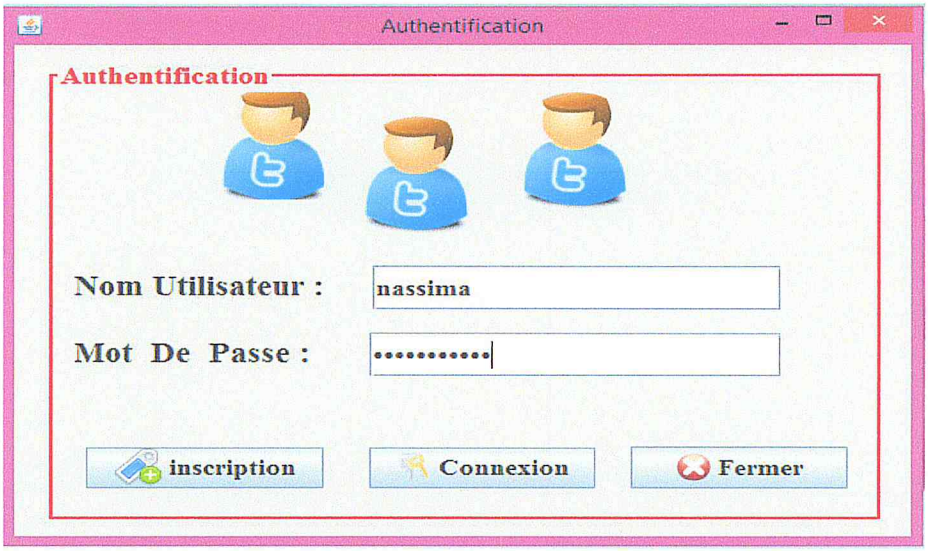


Figure 4.2 : L'interface de l'authentification

Si l'utilisateur ne possède pas de clés d'authentification, il doit s'inscrire pour les stocker dans la base de données.

3.1 La collection des tweets : Après la phase de l'authentification, la récupération des tweets des sujets tendance se fait selon trois paramètres : la langue (français, anglais et arabe), le pays et le nombre de tweets à récupérer pour chaque tendance en temps réel.

3.1.1 L'accèsà l'API Twitter :

Nous avons déjà montré dans la section 4 du chapitre 1 les deux types d'API existants de Twitter. Nous avons utilisé l'**API**[36] **v1.1** du type **API STREAMING** afin de collecter les tweets des sujets tendances. Notre utilisation s'intéresse qu'aux tweets publiques, donc le paramètre utilisé est le flux public (stream public).

L'API de twitter repose sur le protocole OAuth 1.0a[537], celui-ci permet à une application tierce d'obtenir un accès limité à un service http. Elle est accessible via des requêtes qui sont des actions effectuées via un utilisateur à l'accès du site twitter. Ces requêtes d'authentification sont limitées à 350 requêtes par heure[238].

Nous avons créé une application sur Twitter, le panneau de contrôle de l'application de **dev.twitter.com** offre la possibilité de générer un jeton d'accès OAuth pour le propriétaire de

[36] http://dev.twitter.com/docs/api/1.1
[37] http://oauth.net/core/1.0a/
[38] http://kianti.fr/twitter-limites.htm

l'application. Après la création de l'application, l'API fournit 4 clés : (consumer key), (consumer secret) pour l'authentification, (access token) et (access secret) pour la vérification de l'authentification.

Les données sont extraites tout en protégeant les informations personnelles de l'utilisateur. Ces clés sont utilisées via une bibliothèque conçue pour l'authentification et la récupération des données. La collecte se fait pour les tweets des dix sujets tendances de twitter, la limite de la récupération est de 50 requêtes par 15 minutes.

La figure suivante montre la création de l'application sur Twitter et la récupération des clés d'authentification :

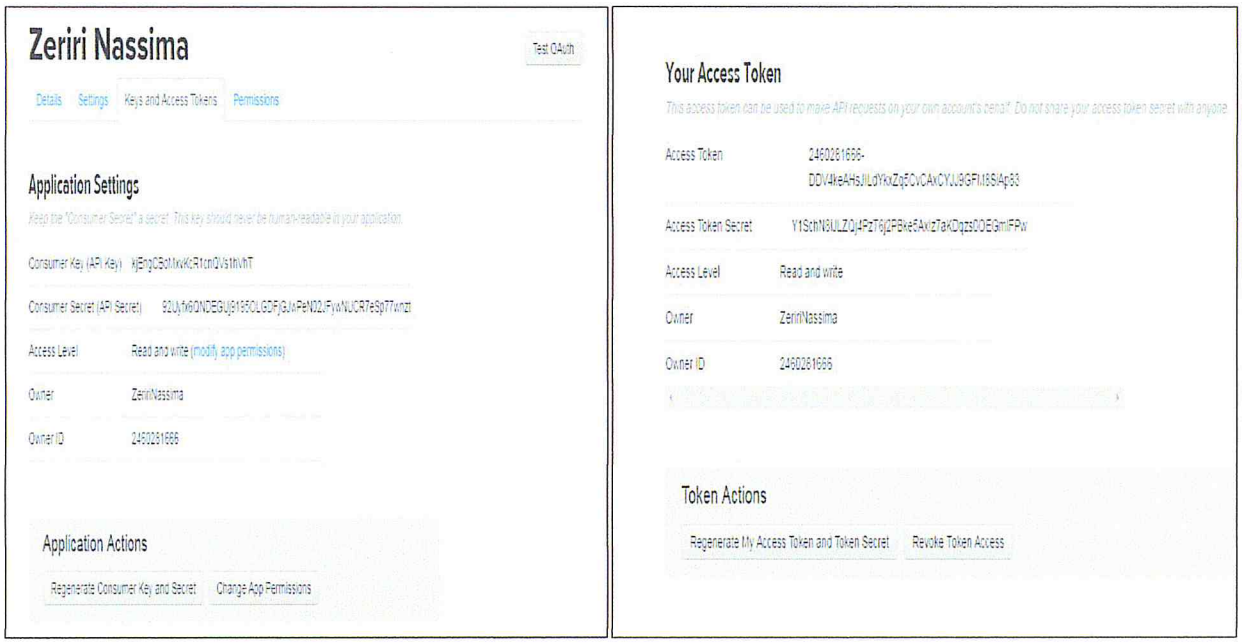


Figure 4.3 : La récupération des clés d'authentification

3.1.2 La langue :

Nous avons essayé de filtrer les tweets selon la langue. La préférence est dans le choix de trois langues : français, anglais et arabe. Pour cela nous avons utilisé une bibliothèque Apache lucene[39]5.1.0 qui contient des outils de recherche et d'analyse de texte.

3.1.3 Le pays :

Les pays sont identifiés par les Woeid qui sont stockés dans la base de données.Le Woeid[40] est une partie intégrante de GeoPlanet ayant des identifiants de 32 bits qui sont « uniques et non

[39] https://lucene.apache.org/

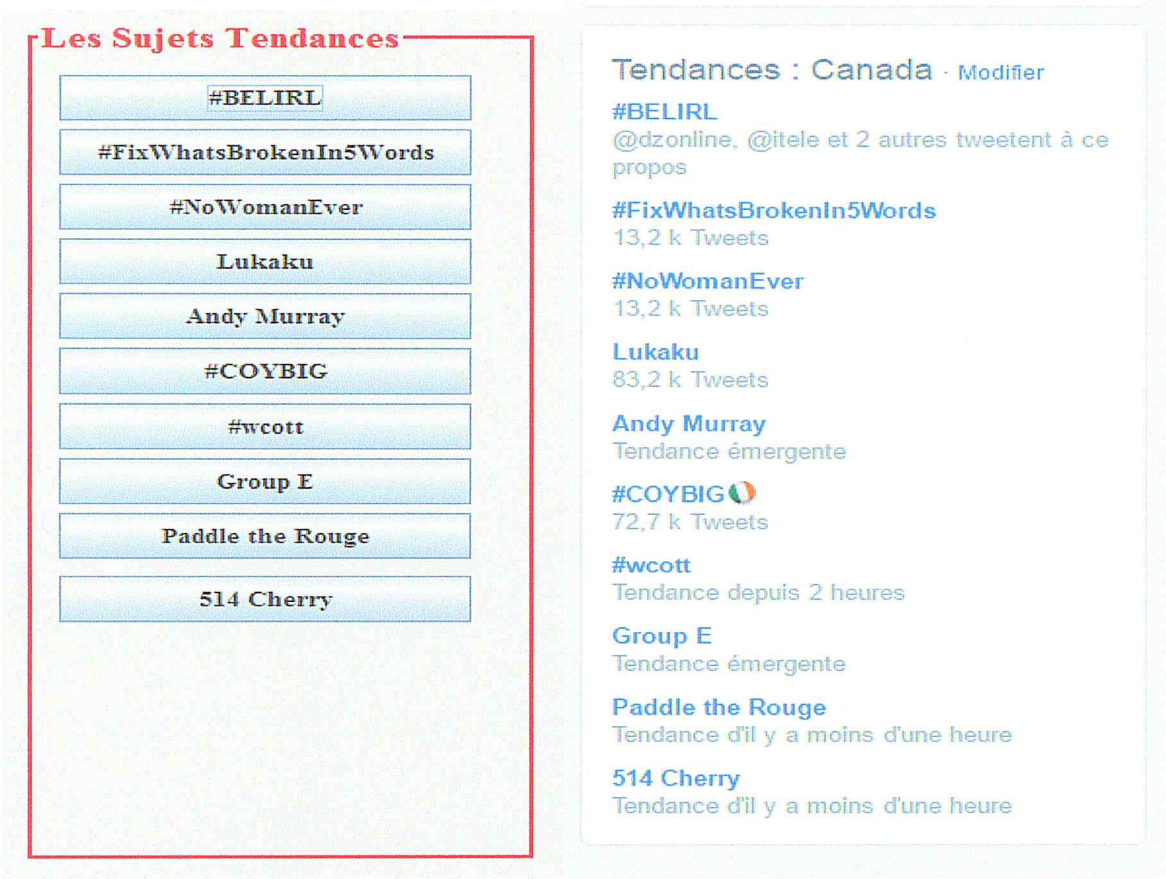
répétitives», maintenant attribués par Yahoo !, qui identifie une caractéristique sur la Terre, sur twitter chaque pays est défini par un woeid. Nous avons collecté les tweets des sujets tendances selon un pays bien spécifié tout en reposant sur les woeid de chacun de ces pays en utilisant l'API Trends.

La figure suivante présente la collecte des tweets en insérant les paramètres cités précédemment :

Figure 4.4 : La récupération des tweets

L'affichage de tweets se fait par tendance, selon le pays et la langue choisis précédemment. Les dix tendances affichées se composent de plusieurs tweets, les tweets seront visibles en cliquant sur une tendance voulue.

[40] https://en.wikipedia.org/wiki/GeoPlanet



(a) (b)

Figure 4.5 : Les tendances dans RAT^{TR}(a)et Twitter (b) (date 18/06/2016)

Les figures 4.6, 4.7, 4.8 montrent l'étape de la collecte des tweets avec les trois langues pour la tendance « Spain » :

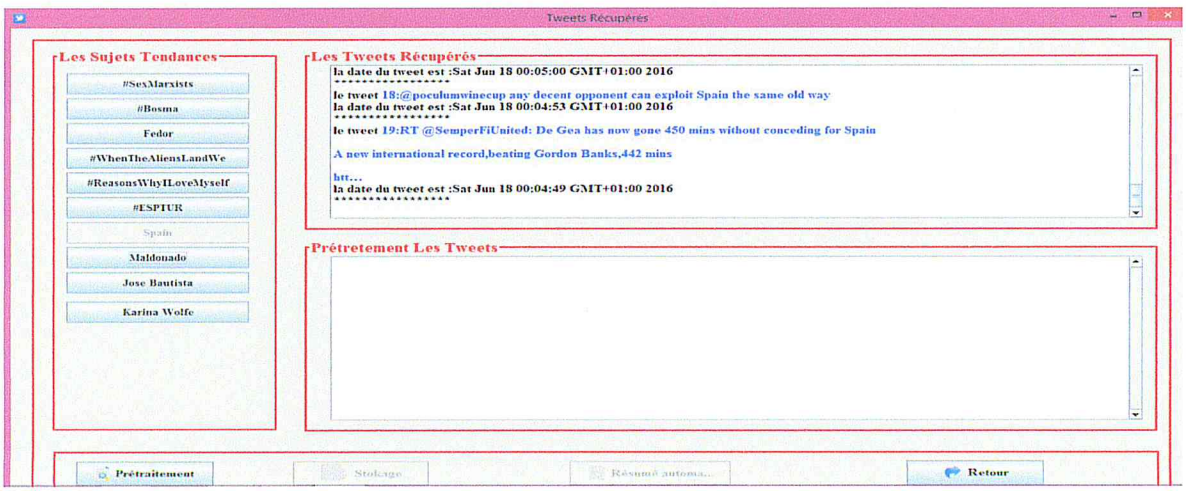


Figure 4.6 : La collecte de tweets en langue anglaise

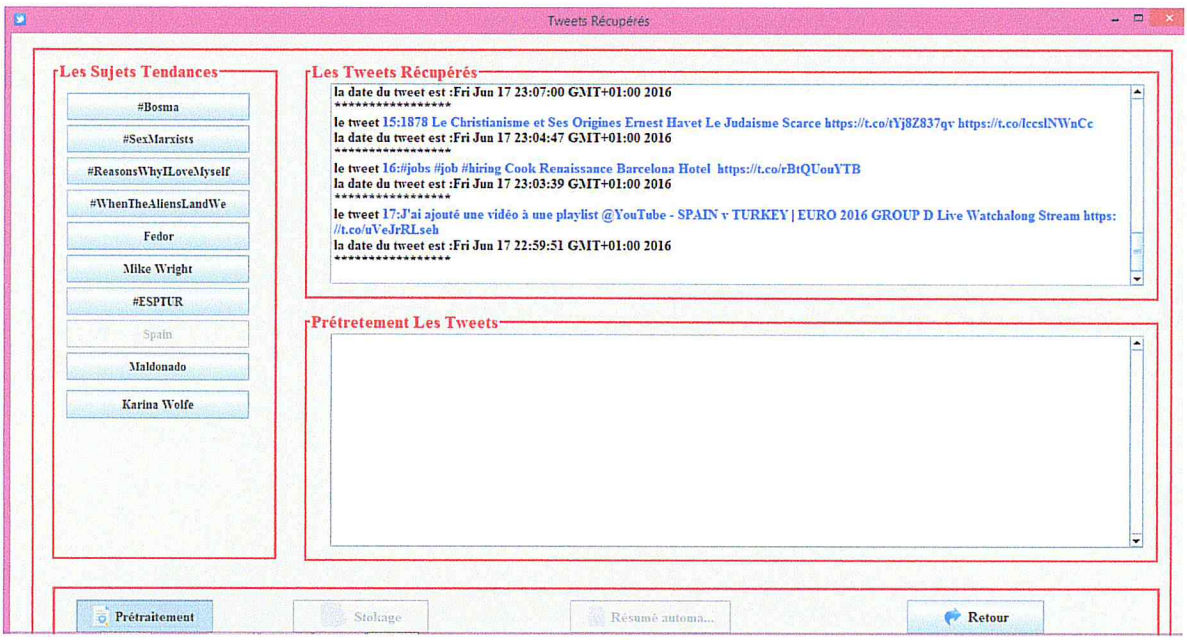


Figure 4.7 : La collecte des tweets en langue française

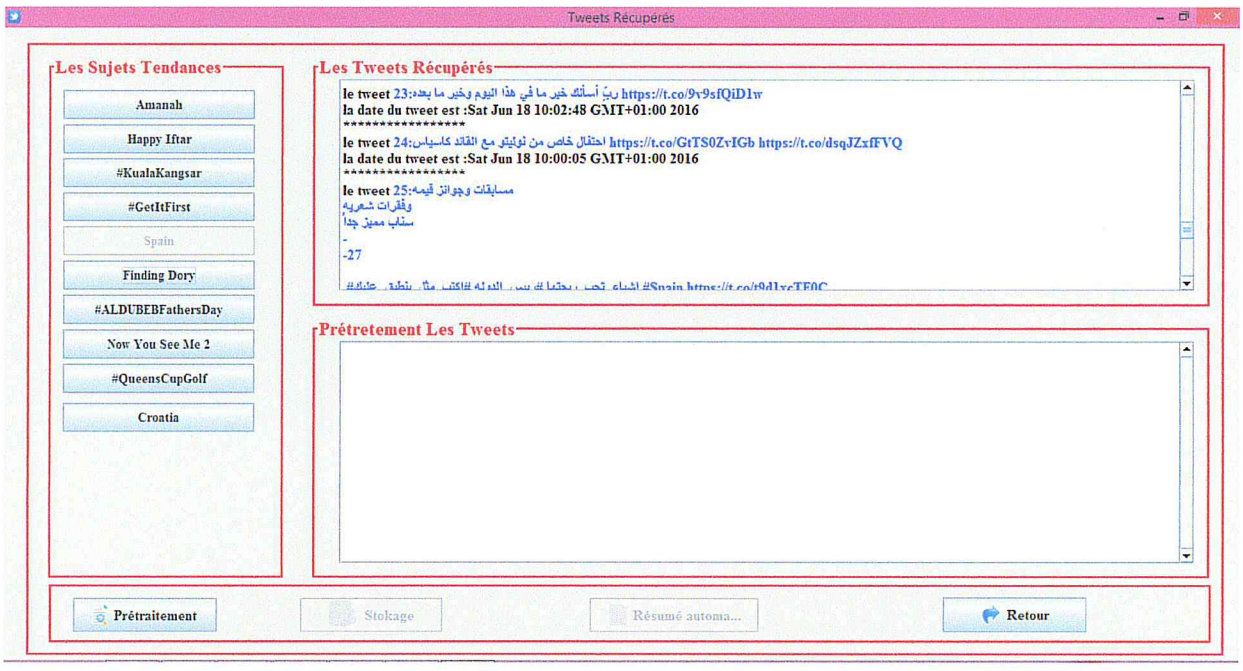


Figure 4.8 : La collecte des tweets en langue arabe

3.2 Prétraitement des tweets :

Les messages courts contiennent des informations inutiles, incohérentes, pour cela nous avons effectué un prétraitement sur tout le corpus extrait.

- La première étape du prétraitement consiste à faire un nettoyage des tweets, c'est-à-dire, effectuer plusieurs opérations comme la suppression des émoticônes, les ponctuations, les urls, les messages retwettés.Nous avons créé des fonctionsà l'aide des expressions régulières (regex).

- La deuxième étape du prétraitement consiste à convertir tous les mots en majuscule au minuscule. Nous avons créé trois listes manuellement avec les trois langues pour les problèmes des abréviations.
- La troisième étape du prétraitement consiste àéliminer les mots-vides, ces mots n'ajoutent aucune valeur à notre résumé. Nous avons utilisé l'API Apache Lucene v5.3 en utilisant l'outil StandardAnalyzer.

- La quatrième étape du prétraitement consiste à analyser lexicalement tous les mots et extraire leurs lemmes. Nous avons basésur le projet d'Ahmed Aker qui utilise l'API OpenNLP et POS Tagger la lemmatisation de la langue française et anglaise.

Les figures 4.9, 4.10 et 4.11, montrent les étapes du prétraitement pour les trois langues :

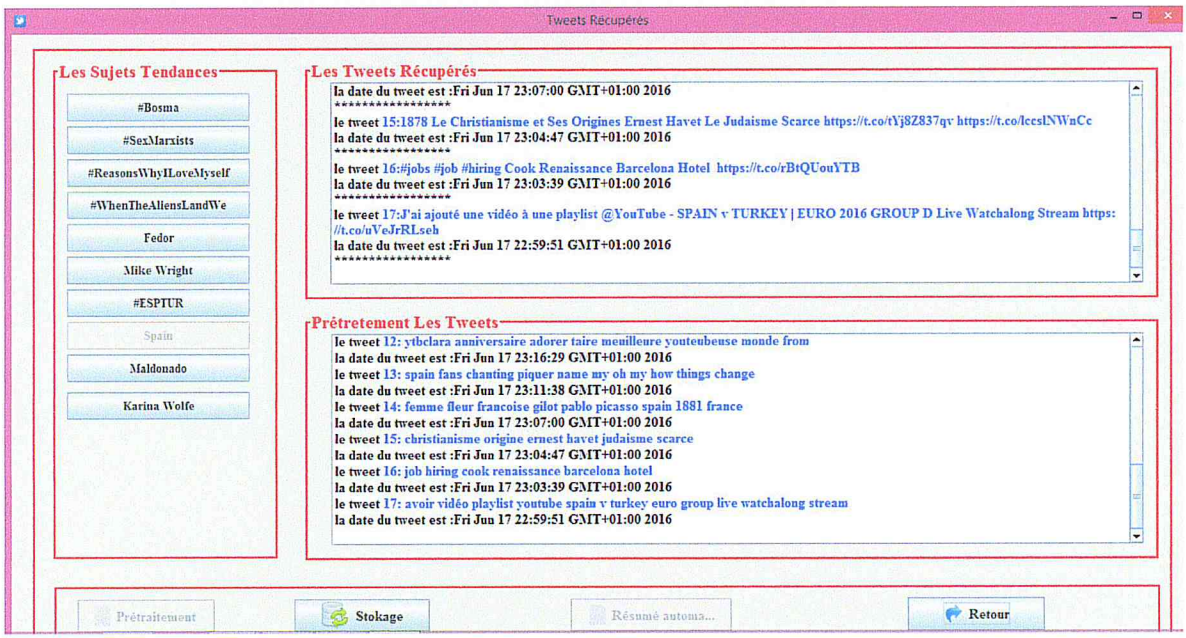


Figure 4.9 : Prétraitement des tweets en langue française

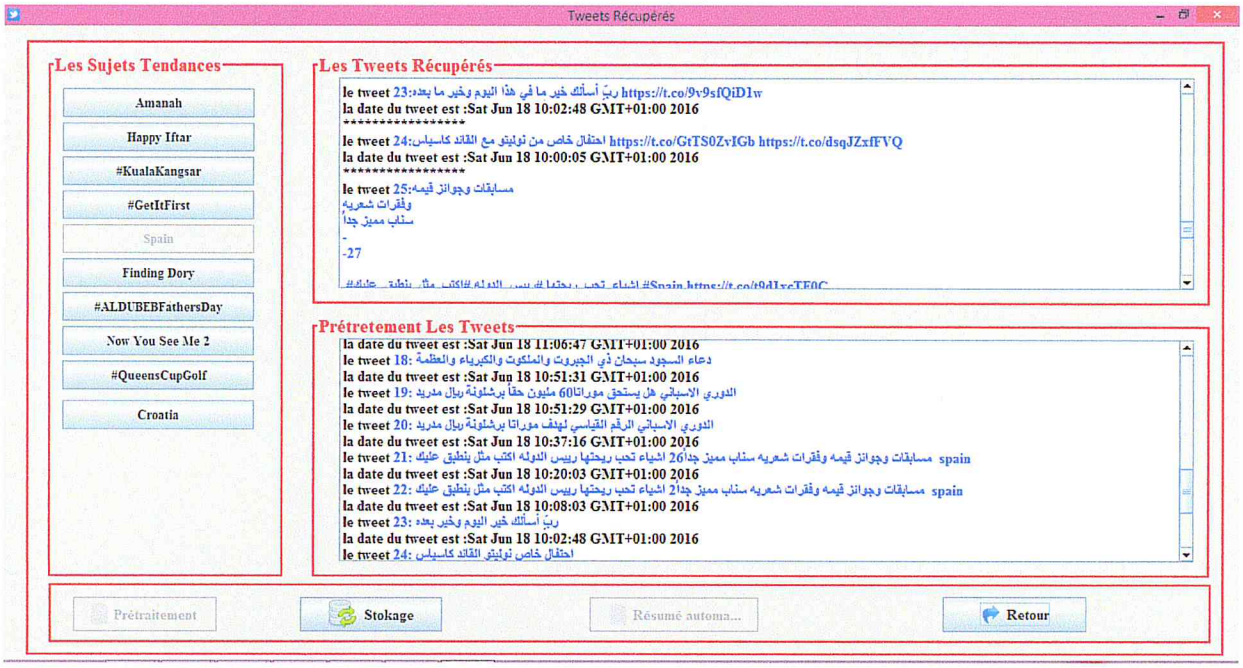


Figure 4.10 : Prétraitement des tweets en langue arabe

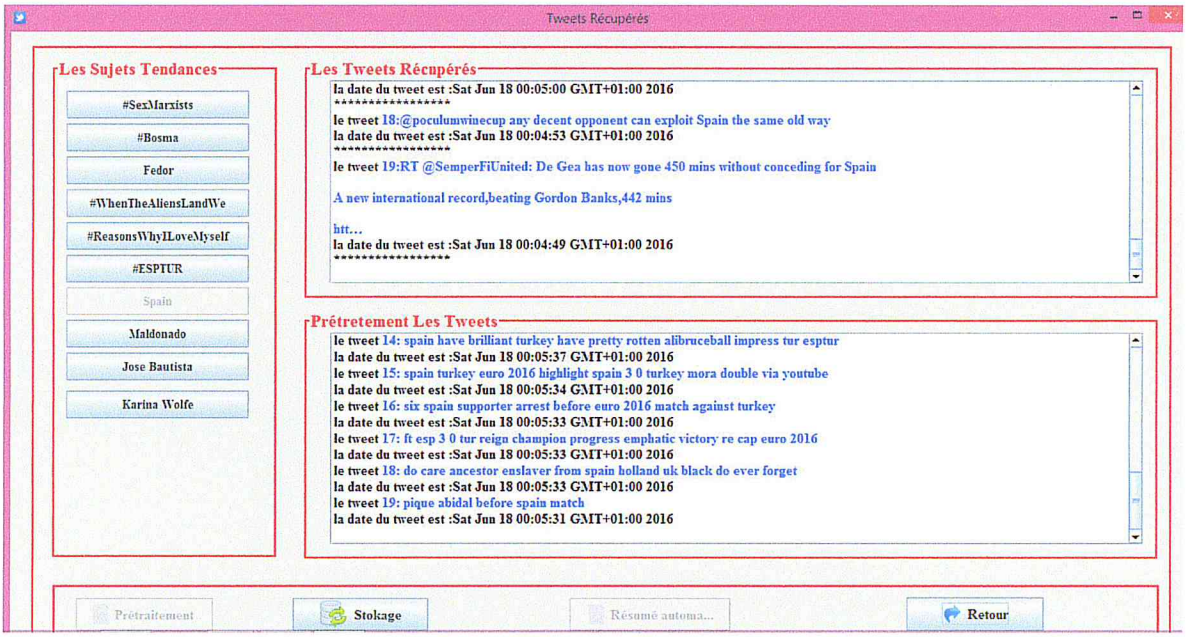


Figure 4.11 : Prétraitement des tweets en langues anglaise

En ce qui concerne la langue arabe, une stemmatisation est plus importante, nous avons basé sur le projet de ShereenKhoja en utilisant l'API Stanford corenlp.

3.3 Stockage dans la base de données :

Après le prétraitement, le stockage de la tendance ainsi que les tweets seront dans une base de données SQLServer. Une fenêtre s'affiche après chaque opération du stockage.

La figure suivante montre la fenêtre qui s'affiche au stockage des données :

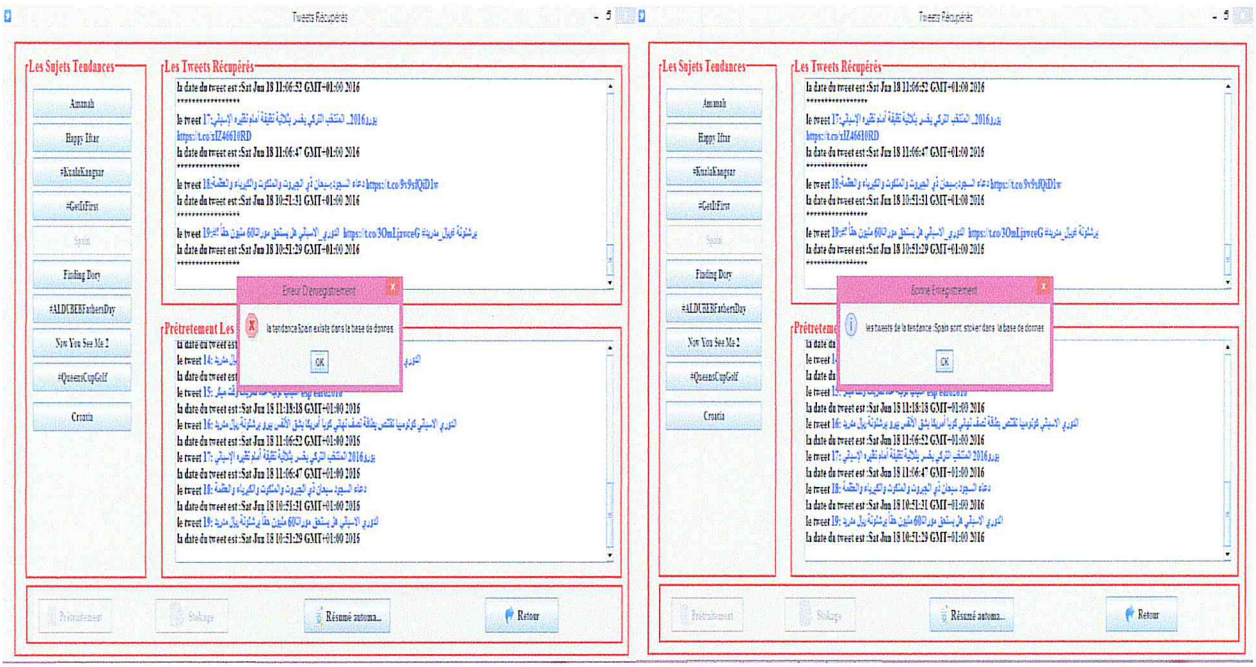


Figure 4.12 : Le stockage de la tendance et des tweets

La figure suivante montre le stockage d'une tendance et ses tweets dans la base de données :

	idtweet	contenu	langue	idtendance	nom
▶	1	voto spain aich...	fr	2	AmirFansES ...
	2	pari aerienne t...	fr	2	extremely sensa...
	3	palmer article f...	fr	2	Articles Spain ...
	4	hahaha degea...	fr	2	à¸§à¸´à¸¨à¸§Fc ...
	5	europe packag...	en	2	BriaGoodman ...
	6	europe packag...	en	2	IreneWise ...
	7	today spain ju...	en	2	Deep'a & Biri ...
	8	europe packag...	en	2	RavenPerkins ...
	9	europe packag...	en	2	KyleeSalazar ...
	10	orange produc...	en	2	Mpilo Enoch ...
	11	wish my mom ...	en	2	SRK UNIVERSE ...
	12	europe packag...	en	2	Rebecca Lee ...
	13	sterling spoon ...	en	2	extremely won...
	14	europe packag...	en	2	Erica Morgan ...
	15	spain now unb...	en	2	chasing thewin...
	16	europe packag...	en	2	Danica Bernard ...
	17	today spain ju...	en	2	Marta Reverte ...
	18	europe packag...	en	2	Leila Lindsay ...
	19	greece 1905c vi...	en	2	extremely won...
	20	nigerium spain...	en	2	ProudlyCule...!! ...
	21	europe packag...	en	2	Skye Sweeney ...

Figure 4.13 : Le stockage des tweets dans la base de données

	idtendance	nomtendance	idpays
	1	Alger ...	23424740
	2	Spain	23424901
	3	Taylor Swift ...	23424901
	4	#MUMUSCARV...	23424901
	5	Lions ...	23424775
	6	#Monster2ndW...	23424901
	7	#FirstSE ...	23424901

Figure 4.14 : Le stockage des tendances dans la base de données

3.4 La pondération :

Les tweets prétraités vont passer par une étape de calcul, qui permet d'affecter un poids à chaque mot. Nous avons utilisé l'API Apache Lucene v5.4.0 qui se compose de l'outil BM25Similarity afin de calculer la similarité de chaque terme.

Tous les tweets de la tendance voulue seront récupérés y compris ceux de la base de données, nous avons considéré chaque tweet comme un seul document qui se compose de plusieurs mots. Ainsi, l'extraction de chaque auteur de tweet est importante pour savoir qui sont les utilisateurs qui s'intéressent plus au sujet tendance en question.

La figure suivante montre la liste des documents (tweets) et le poids de chaque mot en utilisant la méthode BM25 :

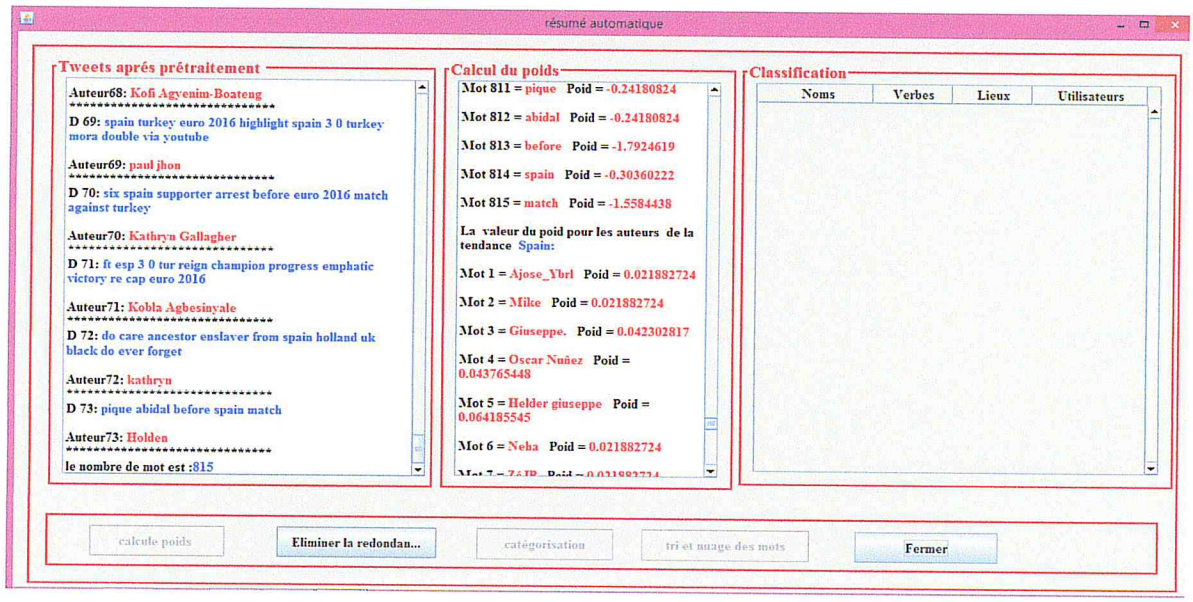


Figure 4.15 : Le calcul du poids pour l'anglais

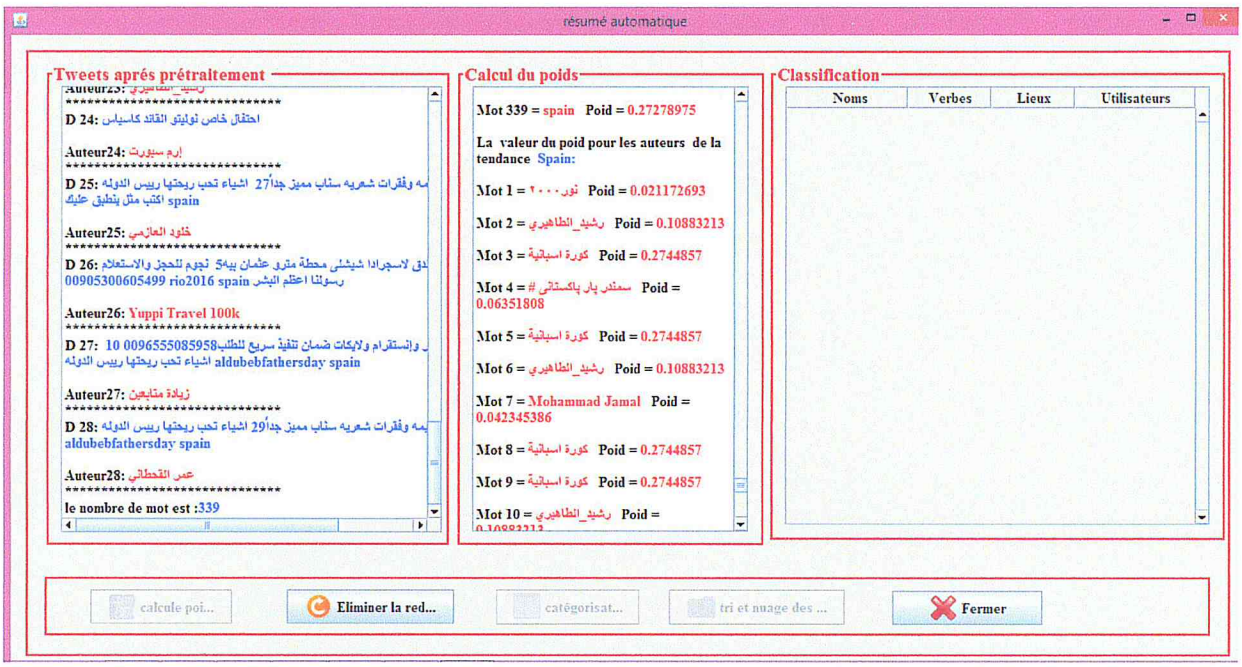


Figure 4.16 : Le calcul du poids pour l'arabe

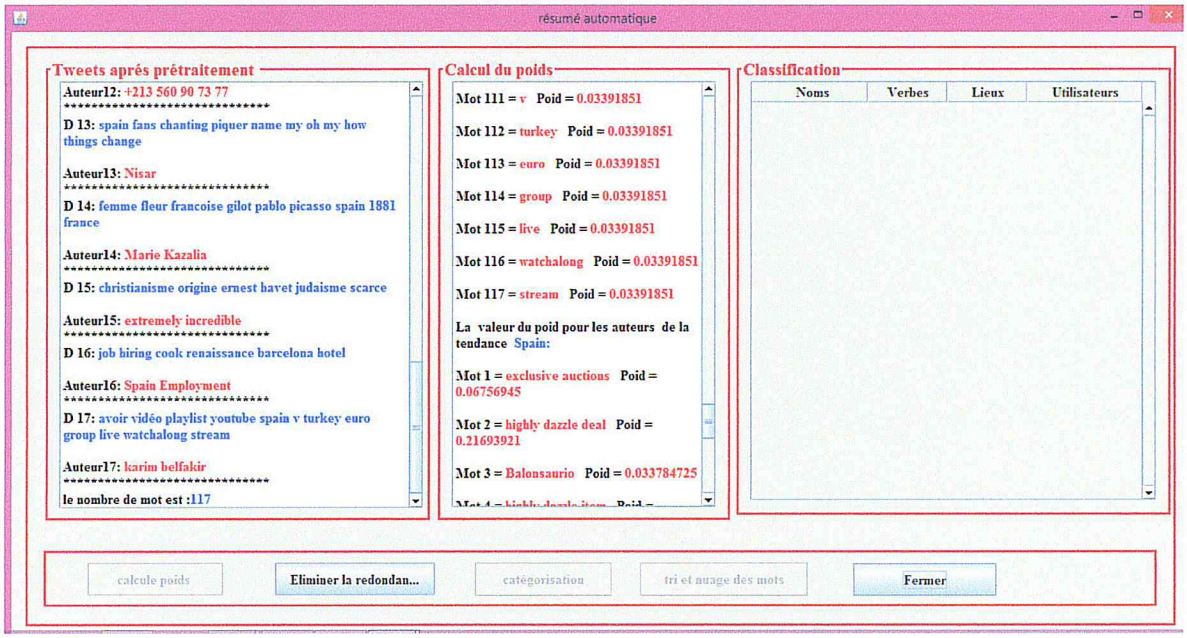


Figure 4.17 : Le calcul du poids pour le français

Dans les tweets un mot important peut être répétés plusieurs fois, afin d'éliminer la redondance, on a créé une fonction qui somme le poids de chaque mot répété, la figure suivante montre le nombre de mots qui est 339 et réduit à 216 on supprimant les mots qui sont présents dans plusieurs documents :

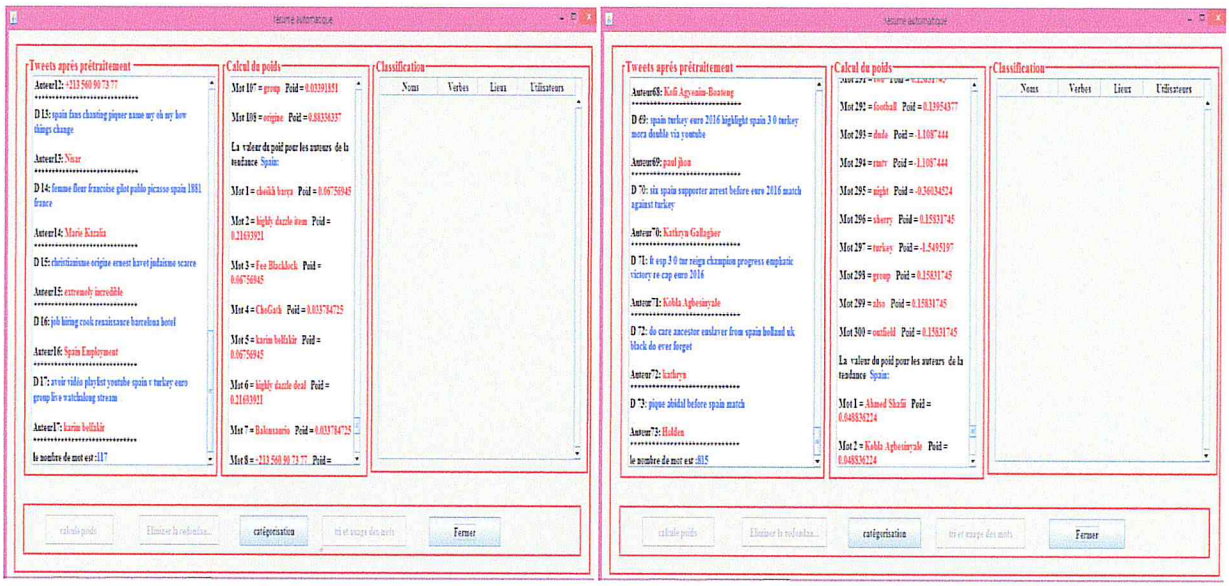


Figure 4.18 : Elimination des redondancespour l'anglais et le français

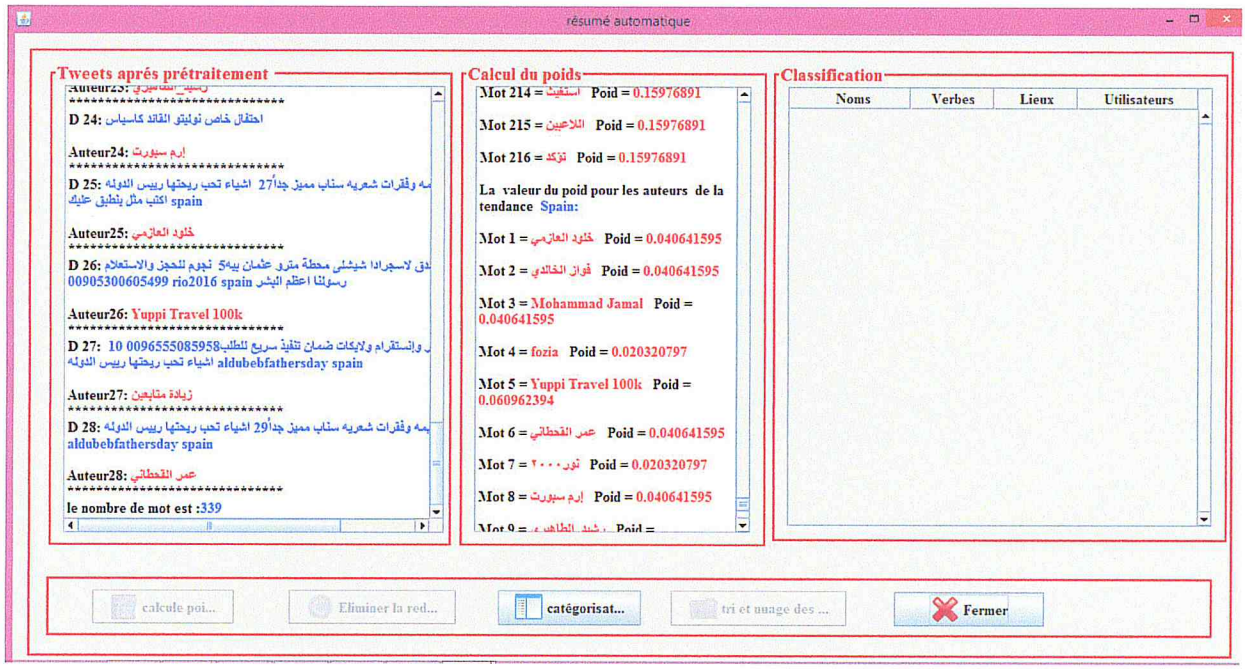


Figure 4.19 : Elimination des redondances pour l'arabe

Cette figure montre le calcul de poids pour chaque auteur de tweet afin de savoir qui s'intéresse le plus à la tendance en question.

3.5 La catégorisation :

La catégorisation nous permet de classer les poids dans quatre classes qui sont : les noms, les lieux, les faits et les auteurs du tweet.

Nous avons utilisé l'API Standford v3.5.0 avec l'outil Part-Of-Speech Tagger qui permet de détecter les entités nommées cités précédemment dans les tweets en langue anglaise.

En ce qui concerne la liste des lieux en langue française et arabe, nous avons rempli deux fichiers avec tous les pays, les villes, régions et lieux et en les intégrant dans notre projet afin de les extraire de nos documents.

La figure suivante, montre un tableau avec les quatre classes extraites (lieux, noms, verbes et auteurs de tweet) de la tendance « Spain » :

Figure 4.20 : La catégorisation des mots prétraités pour l'anglais et l'arabe

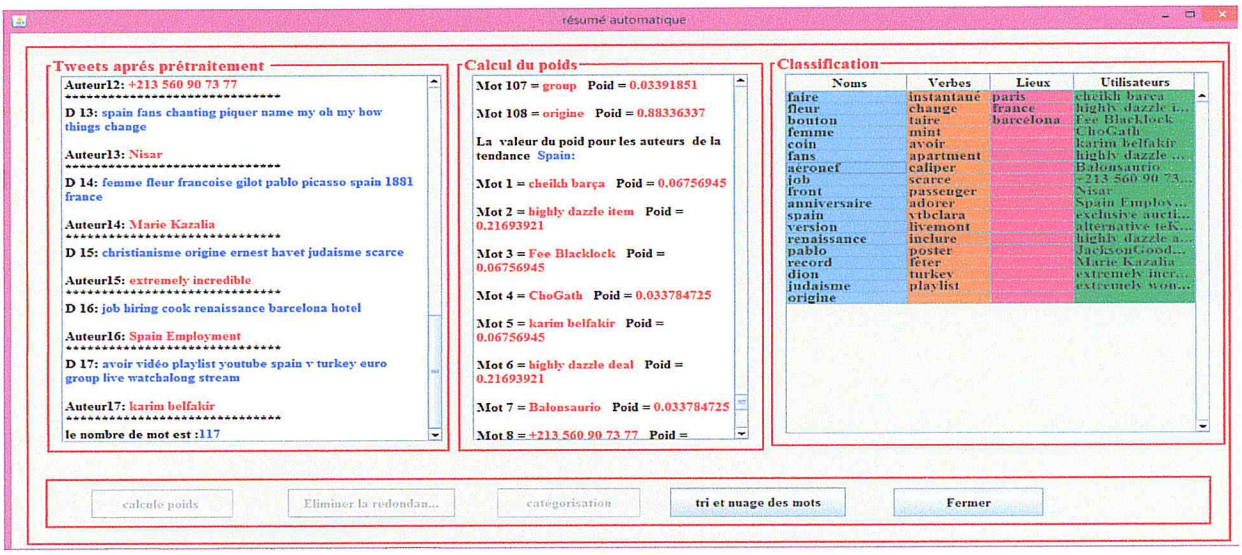


Figure 4.21 : La catégorisation des mots prétraités pour le français

3.6 Le tri :

Après avoir classifié tous les mots dans le tableau, nous avons affecté chaque mot à son poids.

Ensuite, les trié selon un ordre décroissant, cela nous permettra de savoir où sont les mots important par rapport à la tendance dans le but de les afficher dans un nuage de mots.

La figure suivante présente le tri de chaque classe du poids le plus élevé au poids inférieur c'est-à-dire des plus importants au moins.

Figure 4.22 : Le tri des poids par ordre décroissant pour l'anglais et le français

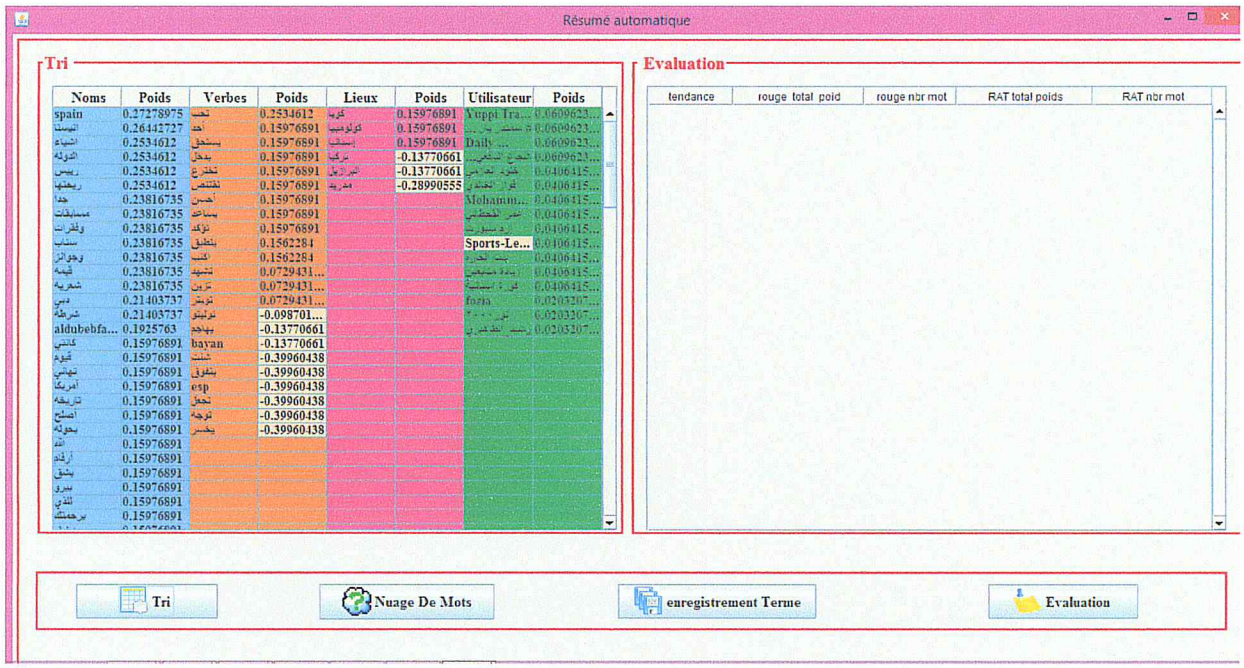


Figure 4.23 : Le tri des poids par ordre décroissant pour l'arabe

Les termes (les noms, les verbes, les lieux et les noms d'utilisateurs) seront enregistrés dans la base de données, et chaque tendance contient un terme avec un poids différent.

	idterme	nomterme	poids	classe	idtendance
▶	1	article ...	0,496174037456...	nom ...	2
	2	condition ...	0,416453987360...	nom ...	2
	3	jambon ...	0,259379029273...	nom ...	2
	4	concours ...	0,259379029273...	nom ...	2
	5	flowers ...	0,259379029273...	nom ...	2

Figure 4.24 : L'enregistrement des termes dans la base de données

3.7 Résumé automatique (le nuage de mots) :

Nous avons utilisé l'API OpenCloud afin de générer un résumé automatique d'une tendance sous forme d'un nuage de mots. Chaque classe est représentée par une couleur, dans notre cas, les faits en orange, les noms en bleu, les lieux en rose et les auteurs de tweets en vert.

Les mots ayant un grand poids s'affichent en taille plus grande par rapport aux autres. Ainsi, que les mots avec un poids négatif ne s'affichent pas sur le nuage.

Figure 4.25 : Résumé automatique (nuage de mots) pour l'anglais

Figure 4.26 : Résumé automatique (nuage de mots) pour le français

Figure 4.27 : Résumé automatique (nuage de mots) pour l'arabe

3.8 Evaluation et tests :

3.8.1 La méthode automatique ROUGE :

ROUGE signifie « Recall-Oriented understudy for Gisting Evaluation »est une méthode automatique qui, à partir des mesures « P,R,F », elle détermine la qualité du résumé automatique en la comparant à d'autres idéaux résumés faits par des humains. L'évaluation se fait par rapport à deux résumés : le résumé référence, et le résumé candidat.

La formule du N-ROUGE est la suivante :

$$= \frac{\sum_{S \in \{Reference\ Summaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{Reference\ Summaries\}} \sum_{gram_n \in S} Count(gram_n)}$$

Où :

N : le nombre des N-Gram (N-Gram c'est une séquence de mots)

S : l'ensemble des séquences

$Gram_n$, et $Count_{match}(gram_n)$: le nombre maximum des mots qui apparaissent dans le résumé candidat et le résumé référence.

Le dominateur de l'équation est le nombre total de mots des N-gram

Cette méthode calcule trois mesures importantes pour déterminer la qualité du résumé qui sont :

$$R_{lcs} = \frac{LCS(X,Y)}{m} \quad P_{lcs} = \frac{LCS(X,Y)}{n} \quad F_{lcs} = \frac{(1+\beta^2)R_{lcs}P_{lcs}}{R_{lcs} + \beta^2 P_{lcs}}$$

Où :

X : le résumé référence avec une longueur m

Y : le résumé candidat avec une longueur n

LCS(X,Y) : la longueur de la plus grande séquence commune entre X et Y

β *: P_{LCS}/R_{LCS}.*

Résultats : nous avons évalué notre approche sur cinq tendances et voilà les résultats :

La méthode rouge évalue les résumés automatiques par rapport aux phrases, pour cela, en entrée nous avons considéré chaque tweet comme une phrase et calculer le nombre de phrases avant et après le résumé.

Nous remarquons qu'il y a des tendances ou les valeurs de F-Mesure étaient proches de la valeur 1, donc le résumé est compréhensible.

Evaluation

tendance	F	Total F	R	Total R	P	Total P
Spain	0.26666666...	6.0	0.15789473...	38.0	0.85714285...	7.0
#Monster5thWin	0.72524752...	293.0	0.87462686...	335.0	0.61945031...	473.0
Imam Muda	0.53619302...	100.0	0.8	125.0	0.40322580...	248.0
Snapchat	0.56716417...	114.0	0.69938650...	163.0	0.47698744...	239.0
Deepika Paduko...	0.44444444...	16.0	0.64	25.0	0.34042553...	47.0
#HayatınSesiSus...	0.32786885...	10.0	0.66666666...	15.0	0.21739130...	46.0
#birişaretistiyor...	0.37837837...	7.0	0.38888888...	18.0	0.36842105...	19.0
#MaliyedenÖgre...	0.0	0.0	0.0	0.0	0.0	14.0
#quyinciqIleFan...	0.48000000...	18.0	0.5	36.0	0.46153846...	39.0
#İçSesimDiyorKi	0.0	0.0	0.0	5.0	0.0	4.0
Stairway to Hea...	0.50773993...	82.0	0.75925925...	108.0	0.38139534...	215.0
Toronto Life	0.60606060...	20.0	0.625	32.0	0.58823529...	34.0
#OnwardYYC	0.0	0.0	0.0	0.0	0.0	8.0
#PastTenseTV	0.11764705...	1.0	0.14285714...	7.0	0.1	10.0
#vidcon2016	0.0	0.0	0.0	0.0	0.0	3.0
#Brexit	0.78709005...	756.0	0.79328436...	953.0	0.78099173...	968.0
#IIFA2016	0.70410367...	163.0	0.79512195...	205.0	0.63178294...	258.0
#AntiVax	0.50793650...	32.0	0.50793650...	63.0	0.50793650...	63.0
London	0.67924528...	108.0	0.69677419...	155.0	0.66257668...	163.0
#AIMHvideo	0.49315068...	54.0	0.84375	64.0	0.34838709...	155.0
Abu Kassim	0.0	0.0	0.0	0.0	0.0	3.0
Britain	0.67272727...	37.0	0.74	50.0	0.61666666...	60.0
#SawasdeekaDara	0.0	0.0	0.0	0.0	0.0	3.0
Remain	0.0	0.0	0.0	0.0	0.0	3.0
Scotland	0.0	0.0	0.0	0.0	0.0	3.0
European Union	0.0	0.0	0.0	0.0	0.0	3.0
Leave	0.0	0.0	0.0	0.0	0.0	3.0
David Cameron	0.0	0.0	0.0	0.0	0.0	3.0
الاتحاد الأوروبي	0.0	0.0	0.0	0.0	0.0	3.0

Figure 4.28 : Evaluation avec la méthode automatique ROUGE

4. Conclusion :

Dans ce chapitre, nous avons présenté l'implémentation de notre approche RAT^{TR}, en l'appliquant sur une collection de tweets.

Nous avons démontré les outils de développement utilisés, ainsi que le fonctionnement de notre application.

CONCLUSION GENERALE ET PERSPECTIVES

Le domaine de la recherche et de l'extraction d'informations nous a portait à s'intéresser aux systèmes du résumé automatique, plus précisément des tweets qui composent plusieurs sujets tendances sur Twitter.

Ces millions de tweets parlent d'une information qui n'est pas toujours comprise par tous les utilisateurs, pour cela, il faut effectuer une lecture manuelle de tous les tweets qui composent un sujet tendance, ce qui engendre une perte de temps, et parfois une incompréhension du contenu. Ce problème nous a conduits à la tâche du résumé automatique. Cette dernière peut aider chaque utilisateur à s'ouvrir à l'actualité du moment, tout en ayant un résumé simple compréhensible par toute catégorie, et en un petit moment.

Nous avons réalisé une nouvelle approche tout en se basant sur plusieurs travaux cités dans l'état de l'art en prenons en compte tous les critères d'évaluation. Cette approche s'intéresse à l'extraction des tweets en temps réel, en en trois langues et pour plusieurs pays. Les mots qui composent les tweets sont des noms propres, des pronoms personnels, des verbes, des adjectifs, des adverbes etc…, parmi ces unités textuelles, on trouve celles qui donnent plus de sens qu'aux autres, ces mots nécessaires seront regroupés afin de générer un résumé automatique et les afficher dans un nuage avec plusieurs couleurs.

Nous avons extrait aussi les auteurs de chaque tweet qui sont une source importante d'information, afin de savoir qui a posté le message ou qui s'intéresse le plus au sujet tendance en question. Cette information est très utilisée dans le domaine du journalisme, de la politique, la criminologie etc…

Tout d'abord, nous avons constaté qu'une étape du prétraitement est indispensable après la collecte des tweets. Cette dernière permet de rendre le contenu de chaque tweet lisible et important, nous avons effectué un nettoyage pour chaque tweet afin de garder que les mots essentiels.

Ensuite, nous avons appliqué une méthode de pondération (Okapi BM25) qui se base sur le nombre de documents et l'occurrence de chaque terme dans son document. Cette méthode nous a permis de savoir quels sont les mots les plus importants par rapport à la tendance en question, ces derniers s'affichent avec un poids plus lourd.

Enfin, nous avons remarqué que parmi la collection des mots des tweets prétraités, certains enrichi notre résumé et le rend plus cohérent. Nous avons utilisé une technique de catégorisation afin d'extraire que les noms, les faits, les lieux et les noms des utilisateurs, et présenter ces classes dans un nuage de mots, chaque classe sera représentée avec une couleur différente, et la taille des mots varie selon le poids calculé avec la méthode Okapi BM25.

Afin d'évaluer notre approche, nous avons calculé les trois mesures : la précision (P), le rappel (R), et le F-mesure (F), à partir des tweets en entrée et le résumé fourni, et trouvé les valeurs des trois mesures pour chaque tendance. Le résumé était bon pour les tendances ayant des valeurs proches de la valeur 1.

Le travail de recherche que nous avons mené n'a pas été aisé. Nous avons rencontré plusieurs difficultés comme :

- Pour certains pays comme l'Algérie, les tweets en langue arabe sont généralement en arabe dialectes.
- Il faut avoir toujours une mise à jour des Woeids des pays.
- Les noms d'utilisateurs avec des caractères spéciaux mentionnés dans les tweets posent un problème lors de la récupération des tweets.

Ce travail s'ouvre à plusieurs perspectives :

- Mettre notre application en ligne.
- Comparer notre approche avec d'autres approches existantes.
- Traiter les problèmes de l'arabe dialecte.
- Traiter les tweets qui se composent de plusieurs langues en même temps.

CONCLUSION GENERALE ET PERSPECTIVES

- Récupérer le contenu des pages web mentionné dans les tweets pour améliorer le résumé automatique.
- Convertir les émoticônes en texte pour traiter les émotions des utilisateurs.

BIBLIOGRAPHIE

[Ahmet Aker, 2010] Ahmet Aker, 2010, Expertise in automatic text summarization, image captioning, comparable corpora collection, parallel phrase extraction and term alignment, Department of Computer Science/ University of Sheffield,211 Portobello, Sheffield S1 4DP,UK.

[Chakrabartiet al. 2011] Chakrabarti D., et Punera K., 2011. Event summarization using tweets. In Proceedings of theFifth International AAAI Conference on Weblogs and Social Media, pages 66–73.

[Houssem Eddine DRIDI1et al., 2013] Houssem Eddine DRIDI, Guy LAPALME, Détection d'évènements à partir de Twitter, Université de Montréal, 2013, pages 41-43.

[Houssem Eddine Dridi et al., 2013] Atefeh Farzindar, Mathieu Roche, Houssem Eddine Dridi, Guy LapalmeAmitava Das. Traitement automatique des langues, Réseaux sociaux, Université Paris VII et de l'Université de Provence, France.

[Hu et al., 2007] Hu M., Sun A., et Lim E. P., 2007. Comments-oriented blog summarization by sentence extraction. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 901–904, New York, NY, USA. ACM Press.

[Ihab Mallak, 2011] Ihab Mallak, De nouveaux facteurs pour l'exploitation de la s´emantique d'un texte en Recherche d'Information, Universit´e Paul Sabatier - Toulouse III, 2011, pages 57-59.

[Jeffrey Nichols et al. 2012] Jeffrey Nichols, Jalal Mahmud, Clemens Drews, Summarizing Sporting Events Using Twitter, 2012.

[Kalita, 2002] Kalita J. K., 2002. Naïve Bayes Classifiers for Spam Detection. MXLogic, Inc. Colorado Springs, CO.

[Kalucki 2009] Kalucki, J. Editor. Streaming API Documentation, https://twitterapi.pbworks.com/Streaming-APIDocumentation, accessed 11/21/2009

[Khoja and Garside, 1999] Khoja, S. and Garside, R. (1999). Stemming Arabic Text. Lancaster, UK, ComputingDepartment, Lancaster University. http://zeus.cs.pacificu.edu/shereen/research.htm.

[Liu et al. 2011] Liu F., Liu Y., et Weng F., 2011. Why is"SXSW" trending? Exploring multiple text sources for twitter topic summarization. In Proceedings of the ACL Workshop on Language in Social Media (LSM).

BIBLIOGRAPHIE

[Luhn, 1958] Luhn H. P., 1958. The Automatic Creation of Literature Abstracts. IBM Journal of Research and Development 2(2), 159–165

[Mahesh, 1997] Mahesh K., 1997. Hypertext Summary Extraction for Fast Document Browsing, Working Notes of the AAAI Spring Symposium for the WWW, pages 95-103.

[Mani et Maybury 1999] Mani I., Maybury M. T., 1999, *Advances in automatic text summarization*, Cambridge, MIT Press.

[Masson 1998] Masson N., 1998, *Méthodes pour une génération variable de résumé automatique : vers un système de réduction de textes*, thèse de doctorat, Paris, université d'Orsay.

[Maud Ehrmann, 2008]. Maud Ehrmann, Les entités nommées, de la linguistique au TAL : statut théorique et méthodes de désambiguïsation, Joint Research Centre, Ispra, Italie.

[Mohamed Hedi Maaloul, 2013] Mohamed Hedi Maaloul, Approche hybride pour le résumé automatique de textes. Application à la langue arabe, Universit´e de Provence, 2013, pages 8-9.

[Robertson et al., 1998] ROBERTSON, S. E., WALKER, S. et HANCOCK-BEAULIEU, M. (1998). Okapi at TREC-7 : Automatic Ad Hoc, Filtering, VLC and Interactive. In Proceedings of the 7th Text Retrieval Conference, TREC-7, pages 199–210.

[Sharifi et al., 2010] Sharifi B., Hutton M. A., et Kalita J., 2010. "Automatic Summarization of Twitter Topics," in National Workshop on Design and Analysis of Algorithm, Tezpur, India.

[Spärck Jones et al., 2000] SPÄRCK JONES, K., WALKER, S. G. et ROBERTSON, S. E. (2000). Probabilistic model of information retrieval : Development and comparative experiments. Information Processing and Management, 36(6).

[Stéphane Manet, 2013] **Stéphane Manet,** Découverte de Twitter, L'@nnexe Espace Public Numérique, Centre social Relais 59 - Paris 12.

[Wei et al., 2012] Furu Wei F., Liu X., Zhou M., et Shum H. Y., 2012. Quickview : Nlp-based tweet search. In Proceedings of the ACL System Demonstrations, pages 13–18. Association for Computational Linguistics.

BIBLIOGRAPHIE

[Yosef Ardhito Winatmoko et al., 2013] Yosef Ardhito Winatmoko, Masayu Leylia Khodra, Automatic Summarization of Tweets in Providing Indonesian Trending Topic Explanation, School of Electrical Engineering and Informatics, Institut Teknologi Bandung Bandung 40132, West Java, Indonesia.

[Zhou et Hovy, 2006] Zhou L., et Hovy E., 2006. On the summarization of dynamically introduced information: Online discussions and blogs, AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs.

WEBOGRAPHIE

- www.twitter.com : le site officiel de Twitter (page 1).