REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

Ministère de l'enseignement supérieur et de la recherche scientifique

UNIVERSITE SAAD DAHLED BLIDA

Faculté des sciences

Département d'informatique.

Mémoire Présenté par :

- Lounas Ismail
- Ould Zmirli Zakaria

Pour l'obtention du diplôme Master

Domaine : Mathématique et Informatique.

Filière : Informatique.

Spécialité : Informatique.

Option : Génie des Systèmes Informatique

Sujet :

DÉTECTION DES ÉVÈNEMENTS À PARTIR DE TWITTER

Examiné par :

Président jury : Mme ZERF

Examinateur 1 : Monsieur NAHAL

Examinateur 2 :Mme TOBALINE

Encadré par :
Madame Madani Amina

Année universitaire 2013/2014

TABLE DES MATIERES

CHAPITRE I :TWITTER

CHAPITRE I I : DETECTION DES EVENEMENT A PARTIR DES TWEETS : L'ETAT DE L'ART :

TABLE DES MATIERES

CHAPITRE I I I : CONCEPTION ET IMPLEMENTATION

TABLE DES MATIERES

LISTE DES TABLES

LISTE DES FIGURES

1. Introduction

Aujourd'hui les réseaux sociaux produisent d'énormes de données sur le web ainsi que dans les entreprises. Les réseaux sociaux sont de différentstypes. Certains sont très connus tels que **Facebook**[1], **Twitter**[2] et **LinkedLn**[3], qui comptent des millions de membres .D'autres sont moins connues et peuvent passer relativement inaperçus ou rester confidentiels, tels les réseaux d'entreprise.

Twitter est un espace et lieu de rencontre pour les individus d'expression libre qui, rendent compte aujourd'hui de la richesse des échanges d'idées et de partage d'opinions et d'information.

Les tweets ou les gazouillis, sont des messages textuels brefs partagés et diffusés sur Twitter. Un tweet ne doit pas plus de 140 caractères. Les tweets sont des flux d'information qui permettent a utilisateur de communiquer des informations sur leurs statuts , activités, pensées et opinion d'une manière plus démocratique , Au-delà de ces aspects d'utilisateur personnels a des fins de divertissement, ces flux offrent aux entreprise et aux communautés virtuelles un moyen de collaboration rapide et pratique.

Les tweets ont récemment posé de nouveaux défis tels que la fouille des tweets (tweetsmining) et plus spécialement la détection d'événement. L'analyse des tweets en temps réel peut permettre la détection des événements nouveaux qui font l'actualité sur Twitter.

2. Problématique

Nous vivons aujourd'hui dans un monde qui change chaque minute, un monde où les événements s' accélèrent de façon terrible, pour cette raison ,le besoin de rester informé sur ce qui se passe à chaque instant a augmenté , et en raison du grand nombre d'informations qui circule dans les réseaux sociaux et la diversité de ces informations, l'utilisateur a besoin des moyens pour l'aider à atteindre les nouvelles appropriées pour lui au bon moment .

---

[1]https://www.facebook.com

[2] https://twitter.com

[3] https://www.linkedin.com

A travers ce travail, nous nous essayons de répondre à la question : Qu'est-ce qui se passe, et sur quel événement se concentre l'information sur Twitter. Notre travail consiste à la détection des événements àpartir des tweets dans le réseauTwitter afin d'enquêter sur ce qui ce passe en temps réel sur ce réseau.

3. Objectifs

Quelques travaux ont été réalisés dans le domaine de ladétection des événements à partir des messages tweets.
Dans la première étape de notre travail, nous avons exposé l'évolution de ces tentatives en faisant une analyse approfondie de l'état de l'art et en parcourant un nombre important des travaux réalisés dans ce domaine. Aussi, nous avons arrêté un certain nombre de critères pour pouvoir faire une comparaison entre ces travaux.
Nous allons proposer une approche capable de détecter des événements à partir des messages tweets en temps réel.
Notre objectif est de fournir de bons résultats (avec une précision importante) en utilisant une collection des tweets puis de faire un prétraitement sur tous les tweets ensuite un calcul des valeurs TF-IDF pour chaque terme, l'utilisation de ces valeurs pour extraire des tendances afin de construire le clustering.

4. Organisation du mémoire

Dans le premier chapitre nous allons essayer de comprendre Twitter en général ensuite dans le deuxième chapitre nous proposons une étude approfondie sur l'ensemble des travaux réalisés dans le domainede la détection des événements. Dans le troisième chapitre nous proposons une nouvelle approche de détection des événements ainsi que l'implémentation du système.

Chapitre I : Twitter

1. Introduction

Au cours des dernières années, les réseaux sociaux comme **Facebook**, **Myspace** et **Twitter** ont transformé la façon dont les individus interagissent et communiquent les uns avec les autres à travers le monde.

Twitter, en particulier, fournit un moyen par lequel les utilisateurs peuvent créer du contenu et l'échanger avec un potentiel public plus large que **Facebook** ou **Myspace** [1].

Dans ce chapitre, nous allons présenter Twitter, son historique, son lexique...etc.

2. Historique

Twitter a commencé comme une idée que Twitter co-fondateur **Jack Dorsey** avait en 2006.**Dorsey** avait initialement imaginé Twitter comme une plate-forme de communication par SMS. Groupes d'amis peuvent garder un œil sur ce que les autres faisaient en fonction de leurs mises à jour de statut. Lors d'une séance de remue-méninges à l'entreprise de **podcasting Odeo** .**Jack Dorsey** a proposé cette plate-forme de SMS à co-fondateur d'**Odeo Evan Williams**. **Evan** et son co-fondateur **Biz** Stone par extension, a donné Jack le feu vert pour passer plus de temps sur le projet et le développer davantage. A ses débuts, Twitter a été appelé «twttr».

Développeur de logiciels **Noah Glass** est crédité de venir avec le nom twttr d'origine ainsi que son incarnation finale Twitter. Pour rappel, certains des principaux joueurs au début de l'histoire de Twitter sont: **Jack Dorsey**, **Noah verre**, **Biz Stone** et **Evan Williams**. Beaucoup seraient d'accord que c'est aussi l'ordre approprié de participation.

Le premier Tweet

Jack a envoyé le premier message sur Twitter le 21 Mars, 2006 21:50 PM On peut y lire, «juste d'installer mon twttr» [2].

Grandir, grandir et encore grandir

Twitter est désormais sur le point de sa plus grande poussée de croissance. La conférence South by Southwest Interactive 2007 a vu une énorme explosion de l'utilisation Twitter. Plus de 60 mille tweets ont été envoyés par jour lors de l'événement. L'équipe de Twitter avait une forte présence à l'événement et a profité de la nature virale de conférence et de ses participants [2].

3. Twitter

Twitter est un site de micro-blogging très populaire qui pose la question « What are you doing ? » (Que faites-vous ?).Où les utilisateurs recherchent des informations en temps opportun et sociales tels que les dernières nouvelles, messages sur les célébrités, et sujets tendance.

Les utilisateurs postent des messages texte courts appelés tweets, qui sont limités à 140 caractères et peuvent être consultés par les partisans de l'utilisateur.
Twitter a été utilisé comme un moyen pour obtenir des informations en temps réel et il a été utilisé dans diverses campagnes de marque, élections, et en tant que média de nouvelles.

Depuis son lancement en 2006, la popularité de son utilisation a été considérablement croissante. En 2014, environ 700 millions de tweets sont générés chaque jour [2].

Quand un nouveau sujet devient populaire sur Twitter, il est répertorié comme un sujet tendance, qui peut prendre la forme de phrases courtes ou hashtags.

Figure 1 : Capture d'écran de l'interface utilisateur de Twitter.

4. Les Followers

Twitter a mis en œuvre un concept de soi-disant disciples. Si un utilisateur met à jour son statut, tous les disciples sont informés du nouveau statut. Ce résultat est obtenu par l'ajout de la nouvelle entrée de leur page personnelle aperçu Twitter [3].

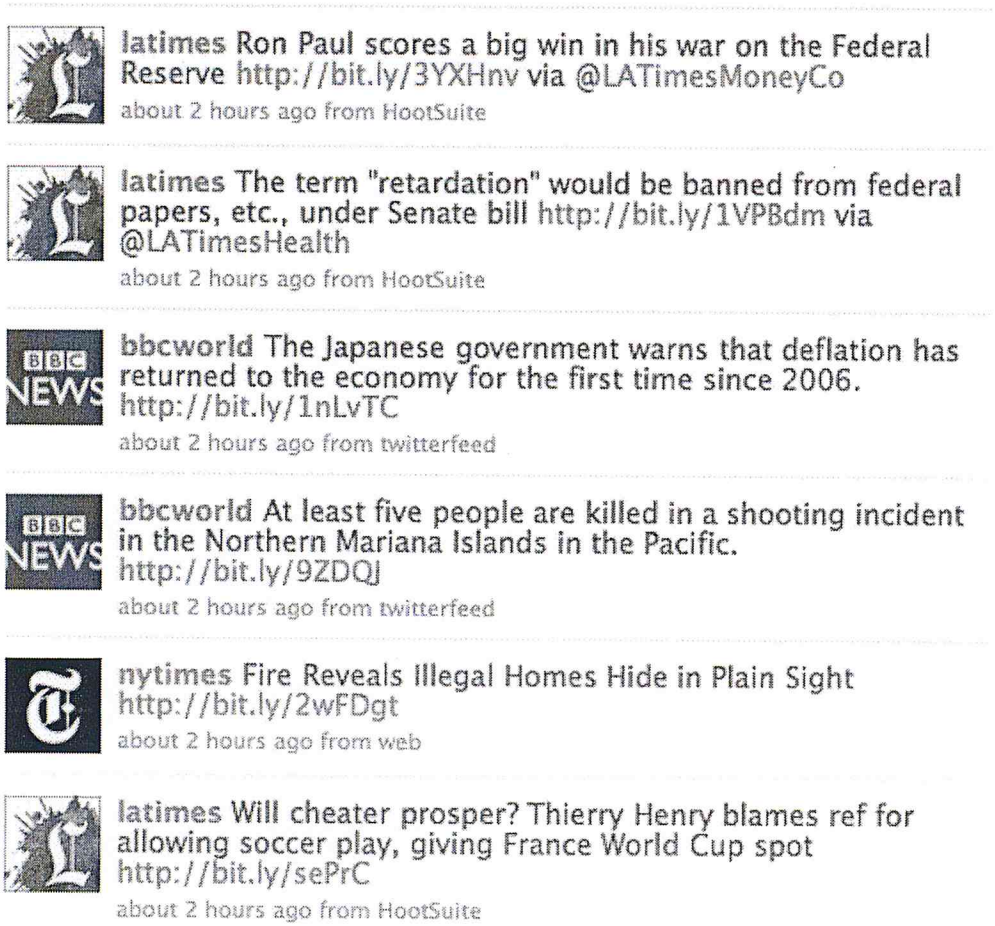


Figure 2 : Capture d'écran de la page personnelle d'ensemble Twitter [4].

Un utilisateur peut suivre tous les autres utilisateurs sauf si cet utilisateur a mis son profil à privé.

5. Le Lexique Twitter

@ : Le « @ » est toujours accolé au pseudo d'un compte Twitter (ne jamais mettre d'espace au risque de casser le lien) et permet de faire savoir à son destinataire que vous lui adressez un message. Par exemple si vous tapez « @BioPourdemain Bonjour », le message apparaitra dans la liste de @BioPourDemain sur sa page d'accueil (colonne de droite).

Hashtag ou # : Son utilisation et sa présence peuvent paraitre un peu énigmatiques mais en fait c'est tout simple.

Le « # » suivi d'un mot (sans espace et éviter les accents et autres caractères spéciaux) fonctionne un peu comme un mot clé ou un tag. Il permet de définir de manière générale le sujet principal du tweet. Lors d'un événement, il permet de suivre toutes les conversations sur Twitter relatives à cet événement. Ce qui est intéressant avec les hashtags, ils permettent de découvrir de nouvelles personnes qui parlent ou s'intéressent aux mêmes sujets que vous.

ReTweet ou RT : Un message contenant « RT » est un message déjà publié par une première personne et republié par une autre personne. Le message est constitué comme suit : RT @auteurdutweet message.

Vous pourrez notamment retweeter des tweets d'autres personnes en fonction Automatique (retweet auto) ou manuel, cette seconde option vous permet d'ajouter un commentaire.

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. Un DM n'est pas publié publiquement et n'apparait pas dans vos tweets.

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 si vous recevez un DM d'une personne c'est que vous êtes abonné à son compte. En bref, c'est un SMS via Twitter.

FollowFriday(FF) : Un autre mot créé par les utilisateurs de Twitter, le FollowFridayou (ViveVendredi) fréquemment trouvé avec les hashtags #FF ou #VV est un moyen de faire découvrir aux personnes qui vous suivent de nouveaux membres que vous appréciez et dont vous aimez suivre les tweets.

Abonnements (Following) : Les Abonnements ou Following correspondent aux comptes Twitter que vous suivez. Pour connaitre le nombre d'abonnements, allez sur votre page d'accueil Twitter le nombre se trouve dans la colonne de droite tout en haut. Et pour voir tous vos following (personnes que vous suivez) cliquez sur le nombre ou « Abonnements ».

Abonnés (Followers): Les Abonnés ou Followers sont les personnes qui suivent votre actualité. Tout comme pour les abonnements, le nombre se situe sur la page d'accueil dans la colonne de droite et vous pouvez voir qui vous suit en cliquant sur le nombre ou « Abonnés ».

LiveTweetting (#LT) : Il s'agit tout simplement d'écrire des tweets en direct d'un événement : un concert, une manifestation, une conférence de presse...

Twittosphère: La Twitosphère ou Twittersphère également appelée Twitterworld ou Twitterland correspond à l'univers de twitter et comprends l'ensemble des utilisateurs de ce réseau social. C'est un mot généralement utilisé pour faire un bonjour général, mais pas exclusivement.

Twittos, Twitteux ou Tweeterers : Un twittos, twitteur ou Twitterer est un utilisateur de Twitter et lorsqu'il y a plusieurs personnes ou obtient Tweeple ou Tweople (contraction de Twitter et People).

Timeline : La Timeline correspond à l'ensemble des tweets postés et classés anté-chronologiquement (du plus récent au plus ancien). Si vous parlez de votre Timeline cela correspondra au fil d'actualité des tweets postés par vos abonnements. Si vous parlez de la Timeline en général cela référera à l'ensemble des tweets publiés sur Twitter.

#TT : Cela signifie **TrendingTopics** qui en français veut dire Tendances, cela fait référence à une actualité qui se retrouve propulsée à la une de Twitter.

Si en page principale (page d'accueil de Twitter après connexion) vos tendances sont paramétrées correctement sur France (colonne de droite), vous verrez de quels sont les principaux sujets abordés par les twittos français [5].

Figure 3 : Exemple de Lexique Twitter

6. Type des tweets

- **Tweets générales** : Un tweet envoyé publiquement à tous ceux qui vous suivent.
- **Réponses :** Un tweet envoyé publiquement à une personne spécifique sur Twitter. Vous pouvez envoyer une @ réponse (réponse à) dans réponse à un bip émis par une personne. Ou vous pouvez utiliser @ réponse à envoyer un message à n'importe Twitter.

- **Mentions :** très semblable a une réponse mais le nom d'utilisateur n'est pas au début de Tweet.
- **Message direct :** Un tweet privé envoyé à une personne qui vous suit, vous ne pouvez pas envoyer un message direct à quelqu'un qui vous ne suit pas.

7. API de Twitter

API (Application Programming Interface).C'est un peu comme une interface utilisateur, sauf qu'au lieu de la livraison de contenu que les humains peuvent lire et utiliser, une API fournit un contenu que le logiciel peut lire et utiliser. Par exemple, un site Web peut offrir de beaux graphismes qui sont bien agencés, avec de grandes polices lisibles, de sorte que l'utilisateur peut facilement trouver et lire les informations.

Ce type de conception orientée humaine est difficile pour un programme à lire, car il repose sur le contexte. Un programme peut accéder au même site Web à l'aide d'une API.L'API renvoie un fichier de données XML ou JSON qui peuvent ensuite être analysés et traités facilement.

Une API fait plus que permettre à votre programme de lire facilement les données. Il vous permet également d'effectuer des actions sur le système distant. Avec l'API de Twitter, en demandant simplement une URL avec quelques paramètres HTTP POST, vous pouvez poster un tweet ou envoyer un message direct.

L'API de Twitter est conçu pour être RESTful. REST (Representational State Transfer) est un modèle de conception de logiciel pour créer des API. Cela signifie que les API est conçu pour tirer parti de requêtes HTTP, tels que GET, POST, DELETE et PUT.Et cela signifie que la demande de données de l'API est aussi simple que de demander une page Web [6].

8. The Open API

Twitter fournit toutes les données et les fonctionnalités gratuitement comme une open API .Cela signifie que vous pouvez inventer et construire de nouvelles applications autour de la fonctionnalité de Twitter. Vous pouvez même créer une toute nouvelle interface Twitter [6].

9. Statistiques

Dans le tableau ci-dessous, nous présentons les statistiques de Juillet 2014 sur Twitter.

Description des Statistiques	**Montant**
Nombre d'utilisateurs enregistrés actifs mensuels	**271 millions**
Nombre d'utilisateurs actifs mensuels qui publient des tweets	**117 millions**
Nombre d'utilisateurs actifs mensuels sur mobile	**211 millions**
Nombre de tweets publié par jour	**500 millions**
nouveaux comptes sont créés chaque minute.	**320 000**
Nombre de tweets envoyés depuis le 21 mars 2006.	**300 milliards**
Le temps moyen passé chaque mois sur Twitter.	**170 minutes**
Le prix d'un TrendingTopic sponsorisé (tendances mondiales) durant 24h.	**200.000 Dollars**
Pourcentage de membres utilisent leur mobile pour accéder à Twitter.	**80%**
Nombre sites intègrent des tweets.	**1 millions**
Le nombre moyen de followers d'un compte Twitter.	**208**

Tableau 1 : Statiques générales sur Twitter.[1]

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

10. Conclusion

Depuis son lancement en 2006, **Twitter** est un outil social qui trouve un vrai succès dans l'univers brutal du web 2.0. Avec plus de 200 millions d'utilisateurs et 500 millions de tweets quotidiens. **Twitter** a grignoté une place confortable dans l'ombre de Facebook, rendant des services insoupçonnables à son lancement.

Twitter regroupe maintenant tous les utilisateurs directement dans son écosystème, il va être intéressant de suivre quelle direction les dirigeants de **Twitter** vont finalement privilégier dans le futur :

1. **Twitter hyper social:** un site social cherchant à concurrencer Facebook sur son propre terrain : la vie de tous les jours et les échanges du quotidien pour le grand public avec plus de façons simples de regrouper ses amis.
2. **Twitter nouveau média :** un site de diffusion et d'analyse de l'information mondiale orienté vers le data-mining et la visualisation des données, pour concurrencer la recherche de **Google.**

Dans les deux cas **Twitter** semble avoir une place à prendre. Mais les modèles économiques seraient probablement très différents.

1. Introduction

Récemment, Twitter, est devenu un nouveau canal d'information pour les utilisateurs de recevoir et d'échanger des informations. Chaque jour, près de 500 millions de tweets sont créés et redistribués par des millions d'utilisateurs actifs.

Chaque utilisateur peut rapporter les nouvelles qui se passent autour de lui. Ainsi, les tweets couvrent presque tous les aspects de la vie quotidienne.

Les tweets ne sont pas isolés, elles sont associées à des informations riches. Par exemple, pour chaque tweet, nous pouvons trouver un horodatage explicite, le nom de l'utilisateur, le réseau social auquel appartient l'utilisateur, ou même les coordonnées GPS si le tweet est créé avec un appareil mobile compatible GPS. Grâce à ces fonctionnalités, twitter est, par nature, une bonne ressource pour la détection et l'analyse des événements, qui sont les principaux concepts.

La plus part des recherches dans le tweet mining se focalise sur la classification des utilisateurs de ce réseau à travers leurs publications ou la détection des évènements.

Dans ce chapitre nous allons présenter une étude approfondie sur l'état de l'art dans le fouille des messages tweets et plus précisément la détection des **événements.**

2. La fouille de donnée (Data mining)

La fouille de donnée[1] désigne l'ensemble de Technique et méthodes dans les domaines des statistiques, des mathématiques et de l'informatique qui permettent de sortir d'un grand volume de données, des connaissances précises sur les éléments inconnus auparavant [7].

[1]http://www.experian.fr/ressources/glossaire/datamining.html

3. La fouille de texte

La fouille de textes (Text mining) ou l'extraction de connaissances à partir d'un texte est une spécialisation de la fouille de données. C'est un ensemble de traitements informatiques consistant à extraire des connaissances selon un critère de nouveauté ou de similarité dans des textes produits par des humains pour des humains. Dans la pratique, cela revient à mettre en algorithmes un modèle simplifié des théories linguistiques dans des systèmes informatiques d'apprentissage et de statistiques [7].

4. La fouille dans les tweets

Le tweet mining est une technique permettant d'automatiser le traitement de gros volumes de contenus des massages tweets pour en extraire des connaissances comme les principales tendances et répertorier de manière statistique les différents sujets évoqués. Les techniques de tweet mining sont surtout utilisées pour des données déjà disponibles au format numérique.

Beaucoup de recherches dans le tweet mining se focalise sur la classification des tweets.

5. La classification de données

Méthodes d'analyse de données [8], leur objectif est d'obtenir une représentation schématique simple d'un tableau de données complexe à partir d'une typologie (segmentation), c'est à dire d'une partition des n individus dans des classes, définies par l'observation de p variables. Ils existent deux types de classification : Supervisée et non supervisée

5.1. Classification supervisée

L'objectif de la classification supervisée est principalement de définir des règles permettant de classer des objets dans des classes à partir de variables qualitatives ou quantitatives caractérisant ces objets [9].

Méthodes

- Les k plus proches voisins
- Le classifieur Bayésien naïf
- Arbres de décision
- Réseaux de neurones
- SVM

5.2. Classification non supervisée

Il s'agit pour un système de diviser un groupe hétérogène de données, en sous-groupes de manière à ce que les données considérées comme les plus similaires soient associées au sein d'un groupe homogène et qu'au contraire les données considérées comme différentes se retrouvent dans d'autres groupes distincts ; l'objectif étant de permettre une extraction de connaissance organisée à partir de ces données [10].

Méthodes

- K-means.
- Les réseaux de neurones
- Les algorithmes génétiques

6. Classification des tweets

La classification est le processus d'attribution des tweets, sous une forme ou une autre, à des groupes ou classes à partir d'un ensemble prédéfini, ce processus fait partie de la fouille dans les tweets, La classification qu'elle soit supervisée ou non supervisée, permet d'analyser les tweets, de les classifier dans une, plusieurs ou aucune catégorie afin d'extraire des connaissances afin d'établir automatiquement qui participe à la production de l'information dans twitter ou de la conversation autour des événements ou des produits afin d'améliorer la consommation de contenu de l'événement pour aider à exposer les parties prenantes de cet événement et leurs intérêts variés, et même aider à orienter la couverture d'un événement par les médias.

7. Détection des évènements à partir des tweets : travaux réalisés

Un évènement est généralement représenté par un burst de mots-clés basés sur le nombre d'occurrences (Kleinberg, 2003).

En tirant parti à la fois de la vitesse et de la couverture de Twitter, on peut détecter des événements en temps opportun en écoutant les tweets entrants. Comme un tweet est souvent associé à l'information spatiale et temporelle, nous pouvons détecter quand et où un événement se passe.

Dans cette partie, nous allons décrire un certain nombre d'études récentes sur la détection des événements à partir des tweets.

7.1. Les travaux de [Siriam et al.2010] [12]

Ils proposent une approche intuitive pour déterminer les étiquettes de classe et l'ensemble de dispositifs avec un foyer sur des intentions d'utilisateur sur le Twitter comme le broutement quotidien, les conversations, partageant l'information/URL, le reportage. Leur approche est plus générale en comparaison avec le TweetStand. Ilsclassifient les tweets entrants dans des catégories telles que les nouvelles (n), les événements (e), les avis (o), les affaires (d), et messages privé (P.M.) basés sur

l'information et les dispositifs d'auteur dans les tweets. Les résultats expérimentaux prouvent que l'exactitude de classification est haute même sans méta-information et approche proposéeSurpasse la stratégie traditionnelle de « Sac-De-Mots ».
Les résultats empiriques prouvent que la profession d'auteur joue un rôle crucial dans la classification. Les auteurs adhèrent généralement à un modèle de gazouillement (Tweetting) spécifique c.-à-d., qu'une majorité de tweets du même auteur tendent à être dans un ensemble limité de catégories.
Ils définissent un événement en tant que « quelque chose qui se produit à un endroit et à un temps donnés », la présence du participant, endroit, et l'information de temps pourrait déterminer l'existence d'un événement dans le texte. Par conséquent, ils ont extrait l'information de date/heure et les expressions de temps d'événement qui sont rassemblées d'un ensemble de tweets ont basé sur l'observation générale des utilisateurs et ont placé la présence d'eux comme dispositif.
Participantl'information est également saisie par l'intermédiaire de la présence caractère de `@' suivi d'un username dans les tweets.
Ils ont téléchargé une collection de tweets récents des utilisateurs aléatoires et ont éliminé ceux qui n'ont pas en anglais, avec trop peu de mots (seuil réglé en tant que trois), avec trop peu de mots indépendamment des mots de salutation, avec juste un URL, et avec trop peu de mots indépendamment de l'URL.

Leur collection finale se compose de 5407 tweets de 684 auteurs. Ces tweets ont été manuellement marqués avec le meilleur assortimentcatégorie (c.-à-d., 2107 N, 625 O, E 1100 D, 1057, et 518 P.M.). Après élimination des mots d'arrêt (stop words), il y a 6747 mots uniques.
Des expériences sont entreprises avec l'exécution disponible du classificateur de Naïve Bayes dans WEKA[2] utiliser la contre-vérification de 5 fois.
Ils ont proposé une approche pour classifier des tweets dans le général mais des catégories importantes en employant l'information et les dispositifs d'auteur dans les

[2]http://www.cs.waikato.ac.nz/ml/weka/

tweets. Avec un tel système, les utilisateurs peuvent souscrire à ou regarder seulement certains types de tweets basés sur tweets intérêt.

7.2. Les travaux de [Sakaki et al. 2010][13]

Leur travail est basé sur l'un des principes de Twitter qui est : Un utilisateur peut suivre d'autres utilisateurs, et ses disciples peuvent lire ses tweets.Un utilisateur qui est suivi par un autre utilisateurne doit pas nécessairement rendre la pareille en les Suits, ce qui rend les liens du réseau comme indiqué.

Ils appliquent une analyse sémantique d'un tweet, par exemple, les utilisateurs peuvent faire des tweets comme «Tremblement de terre! «ou "Maintenant, il tremble"ainsi tremblement de terre ou secousses pourraient être des mots clés,mais les utilisateurs peuvent également faire des tweets tels«Je participe à une conférence tremblement de terre", ils préparent les données d'apprentissage et ilsélaborentun classificateur à l'aide d'un vecteurSVM(Support Vector Machine) basé sur des fonctionnalités telles que les mots clésdans un tweet, le nombre de mots et le contexte des mots de l'événement cible.

Ils font un modèle spatio-temporel probabiliste d'un événement. Ils font une hypothèse cruciale: chaque utilisateur Twitter est considéré comme unsensor (capteur), et chaque tweet qui contient cette évènement estconsidéré comme sensory information, Ces capteurs virtuels, qu'ilsappellentsocial sensors (capteurs sociale).

Comme application, ils ont développé un système d'information de tremblement de terre, qui est une nouvelle approche pour informer les gens rapidement d'un tremblement de terre.

7.3. Les travaux de [Dridi, 2011][14]

Ilse base principalement sur les différents termes (textes, hashtags, hyper textes...) trouvés dans les tweets et il adapte leur méthodesur les tweetsécrits en dialecte Tunisien.

La première étape de son travail consisteà regrouper les termes qui représentent le même évènement puis de calculer le nombre quotidien de tweets portant sur chaque sujet.

Il traite des tweets qui portent sur la Tunisie d'où la plupart des tweets sont écrits en dialecte et très influencé par la langue Française, ce qui fait que les Tunisiens utilisent souvent des vocabulaires et des expressions françaises dans leur communication.

Il a utilisé lestreamingAPI qui permet d'extraire des tweets en temps réel et de façon continue.

Leur processus de détection de l'événement consiste à regrouper les termes qui représentent un même événement. Ensuite il calculele nombre de tweets portant sur chaque évènement par jour. Enfin il détecte les dates saillantes de chaque évènement.

Il a vérifié la pertinence de sa méthode par des experts et des médias traditionnels numériques. Les résultats ont montré que presque tous les évènements retournés par leurs méthodes ont été cité dans les médias, et que 80% d'entre eux ont été jugés importants par les experts.

7.4. Les travaux de [Ozdikis et al. 2012] [15]

Ils présentent une méthode de détection d'événement sur Twitter basée sur la classification des hashtags en se basant sur l'étude des similarités sémantiques entre

les hashtags. Dans ce but, ils ont conçu deux méthodes pour générer le vecteur des tweets et ils ont évalué leur effet sur la classification et la performance de la détection d'évènement par rapport au vecteur basé sur les mots et les méthodes de génération. En analysant des contextes de hashtags et leurs cooccurrences statistiques avec d'autres mots, ils identifient leurs relations paradigmatiques et les similitudes. Ils font usage de cette information lors de l'application d'une expansion lexico-sémantique sur le contenu des tweets avant le regroupement des tweets basé sur de leurs similitudes.

Leur objectif est de tolérer les fautes d'orthographe et les déclarations de capture qui se réfèrent en fait aux mêmes concepts. Ils évaluent leur solution d'amélioration sur un ensemble de données de tweets de trois jours avec un contenu turc.

8. Comparaison

Dans cette partie nous allons comparer les travaux de ladétectiondesévènements à partir des tweetsétudiés précédemment en utilisant un tableau comparatif et en se basant sur les critères de comparaison décrits ci-dessous :

- **Le type de l'évènement** : le nom d'évènement détecté dans chaque approche
- **Classification** : ils ont utilisé la classification supervisée dans leurs travaux.
- **Temps réel** : ces travaux sont en temps réel.

Approche	Type d'évènement détecté
Ozdikis et al 2012	Informations (news)
Sakaki et al 2010	Evénements de la nature (tremblements de terre)
Dridi 2011	Sujets émergeants au Tunisie
Siriam et al	Derniers nouvelles

Tableau 2: comparaison de quelques approches récentes

9. Conclusion

Bien que les tweets soient très échangés sur le web, nous avons constaté qu'il y a peu de travaux qui s'intéressent à la fouille des tweets. Le problème majeur dans ce domaine consiste à déterminer, les informations à extraire à partir des tweets pour servir dans différents domaines.

Nous avons décrit des travaux qui nous intéressent à l'analyse de tweets. Ceux-ci nous ont apporté des idées pour la détection des évènements.

Dans le chapitre suivant, nous proposons une nouvelle approche qui rentre dans le domaine du tweet mining pour le but de détecter des événements nouveaux qui font l'actualité sur Twitter.

Chapitre 3 : Conception et Implémentation

1. Introduction

Ce chapitre présente une nouvelle approche de détection des événements à partir des tweets. Nous avons faire une étude sur les trois travaux sur la détection des événementsà partir des twitter en temps réel de **[12],[13], [14]**et **[15]**étudiés dans le chapitre précédant.

Ici nous avons proposé une nouvelle approche qui a comme but la détection des événementsà partir des tweets.

2. L'approche de détection des événements à partir destweets

Globalement, l'approche est constituée de trois grandes phases dont chacune est composée de différentes étapes. La figure 4 présente ces trois phases qui sont :

Phase 1 : La Collection des données :

1- Extraction des données à partir des documents XML.

2- Prétraitement des données.

Phase 2 : pré-clustering

3- Calcul du poids avec TF-IDF

4- Pré-clustering

Phase 3 : détection des événements

5- Construction des vecteurs / tweets

6- Clustering avec l'algorithme k-means.

Chapitre 3 : Conception et Implémentation

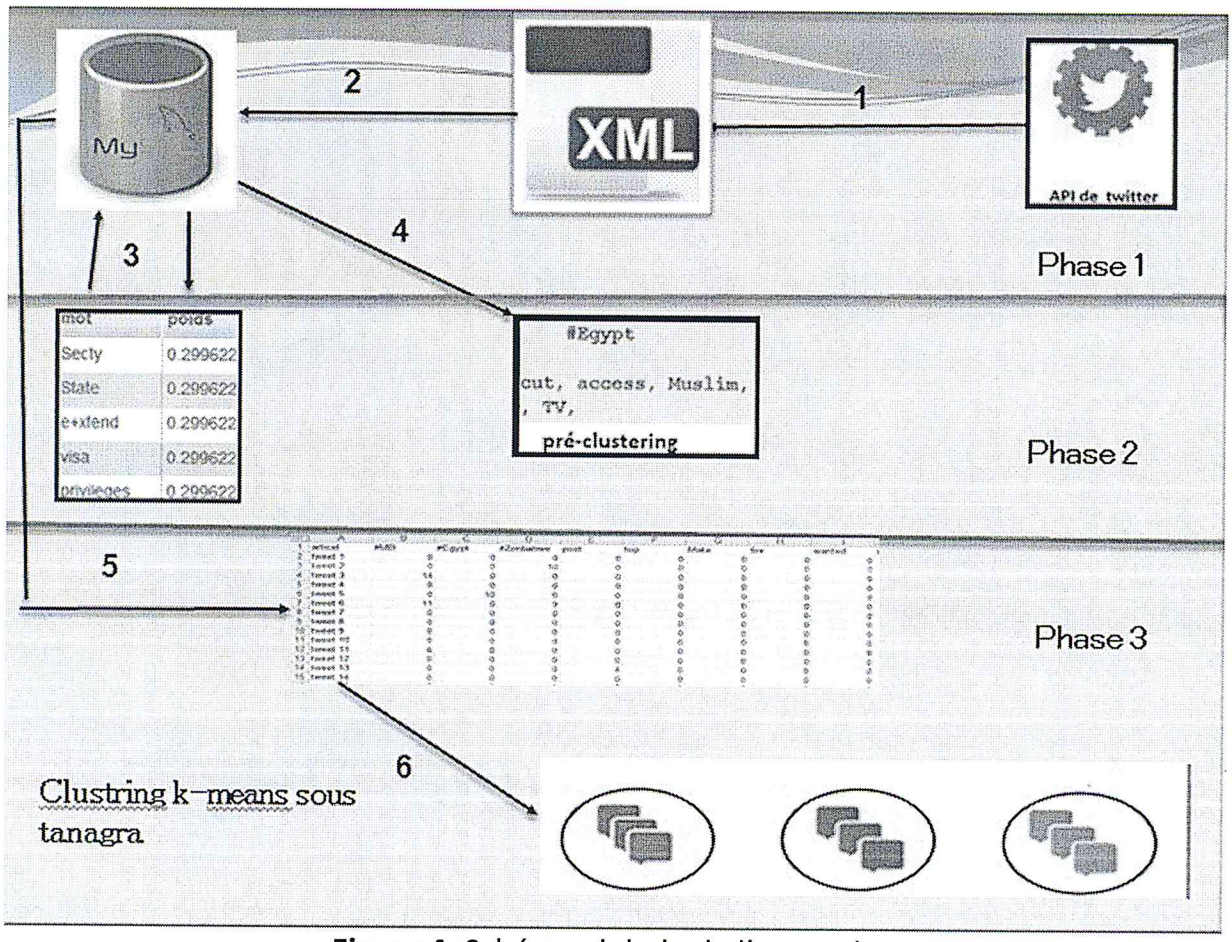


Figure 4 : Schéma globale de l'approche

Phase 1 :La collection des tweets

Cette phase consiste à collectionner l'ensemble des données à traiter dans notre approche. Ces données sont le résultat des requêtes HTTP envoyées à **l'API** de twitter.

On commence par enregistrer l'application dans twitter, les applications sont connues en tant que consommateur. Après l'enregistrement, **l'API** fournit 4 codes : (consumer key) et (consumer secret) pour l'authentification, (accesstoken) et (access secret) pour la vérification de l'authentification. Ce protocole fournit une alternative plus sure au niveau de la sécurité des mots de passe. Ces 4 codes par la suite sont utilisés via une bibliothèque spéciale pour l'authentification et la récupération des données. La recherche est effectuée par mot clé ou hashtag et à chaque requête on récupère les 200 *tweets* les plus populaires contenant le mot de recherche.

Chapitre 3 : Conception et Implémentation

Le fichier XML à traiter dans notre approche contient 100 des publications sur twitter des chaines télévisées anglo-saxonne : BBC, CNN, ABC.

1. Extraction des données à partir d'un fichier XML

Cette étape consiste à extraire l'ensemble des données importantes pour notre approche à partir d'un fichier XML locale ou en ligne, les documents XML comporte plusieurs informations sur un tweet, les données à extraire sont : le contenu de tweet (texte), date de publication, localisation (Pays). Ces données sont stockées dans la base de données pour les exploiter ensuite.

2. Prétraitement des données

Cette étape consiste à préparer l'ensemble des tweets pour la classification à fin de gagner le temps et l'espace de stockage de notre système.

Le prétraitement se compose de toutes les tâches de traitement automatique de langue, ces tâches sont :

- **Elimination des mots vides**

  Un mot vide est un mot non significatif. Ce mot apparaît avec une fréquence semblable dans chacun des textes de la collection n'est pas discriminant, ne permet pas de distinguer les textes les uns par rapport aux autres. Il existe des déférents collection des mots vides de la langue anglaise sur le web, nous avons utilisé un fichier texte qui contient les mots vide et puis aétape de prétraitement on faire le filtrage des mots vides

- **Elimination des ponctuations**: cette étape consiste à filtrertoutes lesponctuations dans les tweets (. , ! ? ...etc.).

Phase 2 : Pré-clustering

3. Calcul du poids avec TF-IDF (TermFrequency-Inverse Document Frequency)

Cette étape consiste à calculer le poids de chaque mot ou terme en utilisant la méthode TF-IDF[16], ce poids permet d'évaluer l'importance d'un terme contenu dans un tweet, relativement à une collection.

TF-IDF (Term Frequency-inverse Document Frequency)

TF-IDF est une méthode de pondération souvent utilisée en recherche d'information et en particulier dans la fouille de textes. Cette mesure statistique permet d'évaluer l'importance d'un terme contenu dans un document, relativement à une collection ou un corpus. Le poids augmente proportionnellement au nombre d'occurrences du mot dans le document. Il varie également en fonction de la fréquence du mot dans le corpus. Des variantes de la formule originale sont souvent utilisées dans des moteurs de recherche pour apprécier la pertinence d'un document en fonction des critères de recherche de l'utilisateur.[1]

En règle générale, le poids de TF-IDF est composé de deux termes: le premier calcule la fréquence normalisée terme (TF), le nombre de fois où un mot apparaît dans le document, divisé par le nombre total de mots dansledocument.

TF(t) = (Nombre de fois où le terme t apparaît dans un document) / (Nombre total de termes dans le document).

Le second terme est la fréquence inverse de document (IDF), calculé comme le logarithme du nombre de documents dans le corpus divisé par le nombre de documents dans lesquels apparaît le terme spécifique.

[1]http://fr.wikipedia.org/wiki/TF-IDF

IDF (t) = log (nombre total de documents / Nombre de documents dont la durée t en elle) [16].

Le score est le plus élevé lorsque le terme apparaît souvent dans un petit sous-ensemble des documents, et sera plus bas lorsqu'il apparait plusieurs fois dans d'autres documents. TF-IDF est largement utilisé pour comparer la similarité entre les documents, fournissant une liste triée des documents les plus pertinents.
Les Tweets sont des messages courts (limité à 140 caractères), alors la fréquence (TF) d'un terme est généralement 1. Cela signifie que la fréquence inverse (IDF) prend plus d'importance, mais la pondération globale perd une puissance due à l'absence de richesses locale. Pour éviter ce problème, on a décidé d'augmenter le poids des termes qui sont précédés par le dièse #, par exemple : #Ghaza.

4. pré-clustering

Cette étape consiste à utiliser les termes avec leurs poids (TF-IDF) pour déterminer les sujets émergents, nous organisons les termes par ordre décroissant du poids, les termes avec les poids élevés ont plus de chance d'être un sujet car les sujets sont des termes qui reviennent souvent dans les tweets mais plus d'une fois dans le même tweet. Pour notre approche .on va commencer par mettre le 1er mot au centre de premier cluster ensuite mettre tous les mots qui appartient aux tweets ou contient ce mots ensuite on passe sur le deuxième mot et on teste s'il existe dans le premier cluster sinon on crée un nouveau cluster et ainsi des suit.

Phase 3 : Détection des événements

5. Construction des vecteurs / tweets

La dernière phase consiste à créer des vecteurs,chaque vecteur concerne un tweet, la taille du vecteur est déterminée par le nombre de clusterstrouvésdansl'étape de pré-clustering.

La valeur de chaque case du vecteur représente le nombre de termes d'un sujet dans le tweetpar exemple siona un vecteur qui contient les valeurs 7,8,7,9,.....ect il existe 7 termes dans le tweet concerné qui appartient à la collection du centre de premier cluster , et 8 termes du deuxième centre ainsi de suite.

6. Clustering avec l'algorithme k-means

L'algorithme k-meansest utilisé avec *k* un paramètre défini arbitrairement en entrée qui indique le nombre de clusters à construire. La distance euclidienne est Utilisée pour mesurer la distance entre paires de vecteurs où *n* est le nombredes tweets. Elle est calculée par la formule suivante :

Soit t_1 , t_2 deux tweets différents dans notre collection est:

$t_1 = (x_1,x_2,x_3,.....,x_n)$ et $t_2= (y_1,y_2,y_3,.....,y_n)$

alors : Distance $(t_1,t_2)= \sqrt{\sum_{i=1}^{n}(xi,yi)^2}$

Les étapes de *k-means*sont les suivantes :

1. Choisir aléatoirement *k* documents qui formeront l'ensemble des centroïdes Initiaux représentant les *k* clusters à construire.
2. Assigner chaque document au cluster dont le centroïde le plus proche selon la distance *d* (si un minimum de *d* est trouvé entre deux objets).
3. Si aucun document ne change de cluster d'une itération à l'autre alors arrêt et sortir les clusters. Sinon, mettre à jour les centroïdes des clusters en fonction Des objets qui leur sont associés.
4. Aller à 2[17].

3. Implémentation

3. 1 Technologie et outils de développement

3.1.1 Java

Pour la réalisation de notre application, nous avons utilisé le langage de programmation JAVA.

Java est un langage de programmation et une plate-forme informatique qui ont été créés par Sun Microsystems en 1995. Beaucoup d'applications et de sites Web ne fonctionnent pas si Java n'est pas installé et leur nombre ne cesse de croître chaque jour. Java est rapide, sécurisé et fiable. Des ordinateurs portables aux centres de données, des consoles de jeux aux superordinateurs scientifiques, des téléphones portables à Internet, la technologie Java est présente sur tous les fronts! [18]

3.1.2. Netbeans

NetBeans est un environnement de développement intégré (EDI), placé en *open source* par Sun en juin 2000 sous licence CDDL et GPLv2 (Common Development and Distribution License). En plus de Java, NetBeans permet également de supporter différents autres langages, comme Python, C, C++, JavaScript, XML, Ruby,PHP et HTML , Il comprend toutes les caractéristiques d'un IDE moderne (éditeur en couleur, projets multi-langage, refactoring, éditeur graphique d'interfaces et de pages Web).Conçu en Java, NetBeans est disponible sous Windows, Linux, Solaris, Mac OS X ou sous une version indépendante des systemes d'exploitation (requérant une machine virtuelle Java)

Un environnement Java Développements Kit JDKest requis pour les développements en Java.NetBeans constitue par ailleurs une plate forme qui permet le développement d'applications spécifiques (bibliothèque Swing (Java)). L'IDE NetBeans s'appuie sur cette plateforme.L'IDE Netbeans s'enrichit à l'aide de greffons.[2]

3.1.3 MySQL

MySQL est un système de gestion de base de données(SGBD).

Il est distribué sous une double licence GPL. Il fait partie des logiciels de gestion de base de données les plus utilisés au monde, autant par le grand public (applications web principalement) que par des professionnels, en concurrence avec Oracle, Informix et Microsoft SQL Server [19].

3.1.4 Tanagra

Tanagra est un logiciel gratuit de *data mining* destiné à l'enseignement et à la recherche. Il implémente une série de méthodes de fouilles de données issues du domaine de la statistique exploratoire, de l'analyse de données, de l'apprentissage automatique et des bases de données.

Nous avons utilisé la version 1.4.50 de Tanagra pour appliquer l'algorithme de *clusteringk-means*.

4. Présentation de l'application

4.1 L'architecture globale de l'application

[2]http://fr.wikipedia.org/wiki/NetBeans

Chapitre 3 : Conception et Implémentation

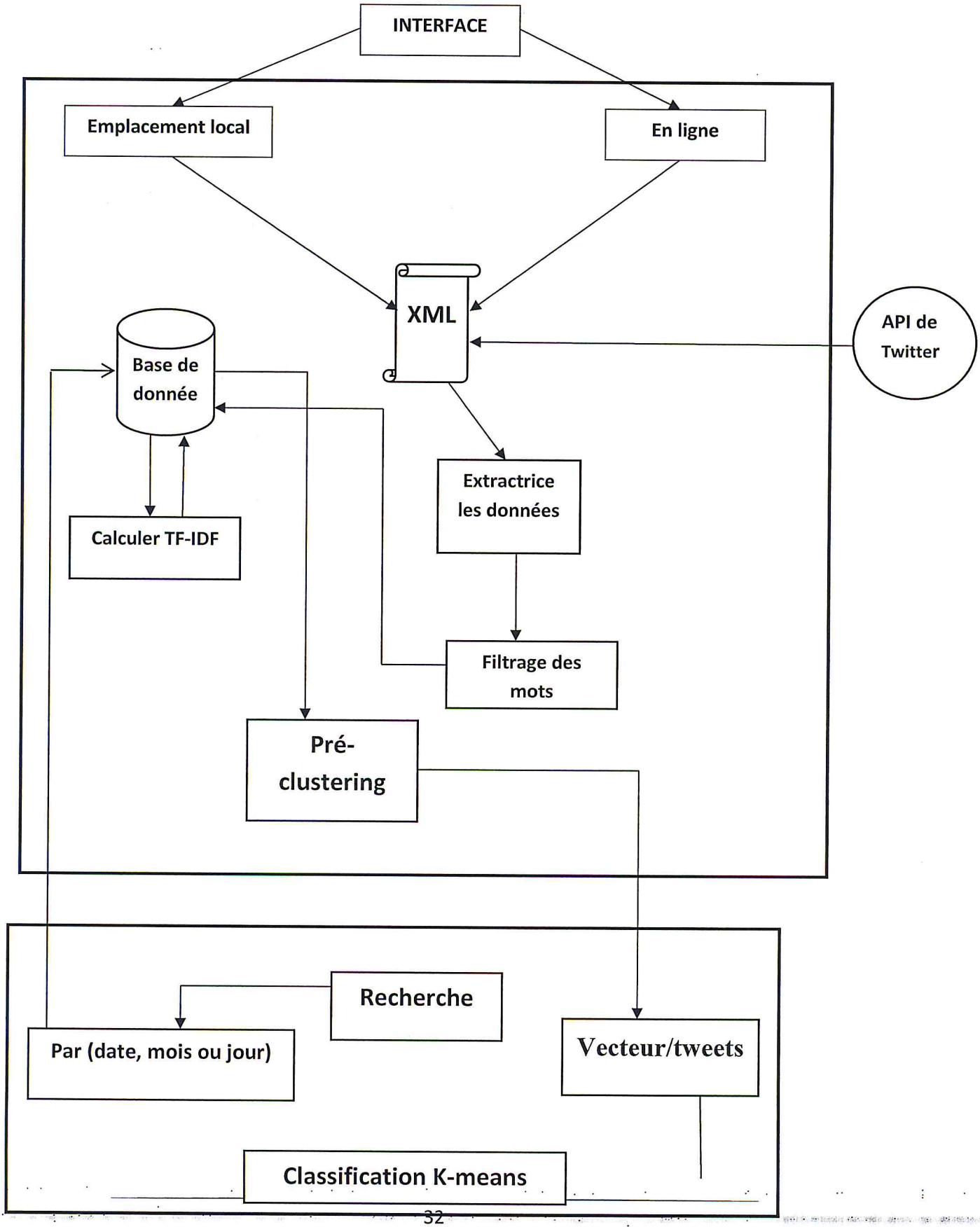


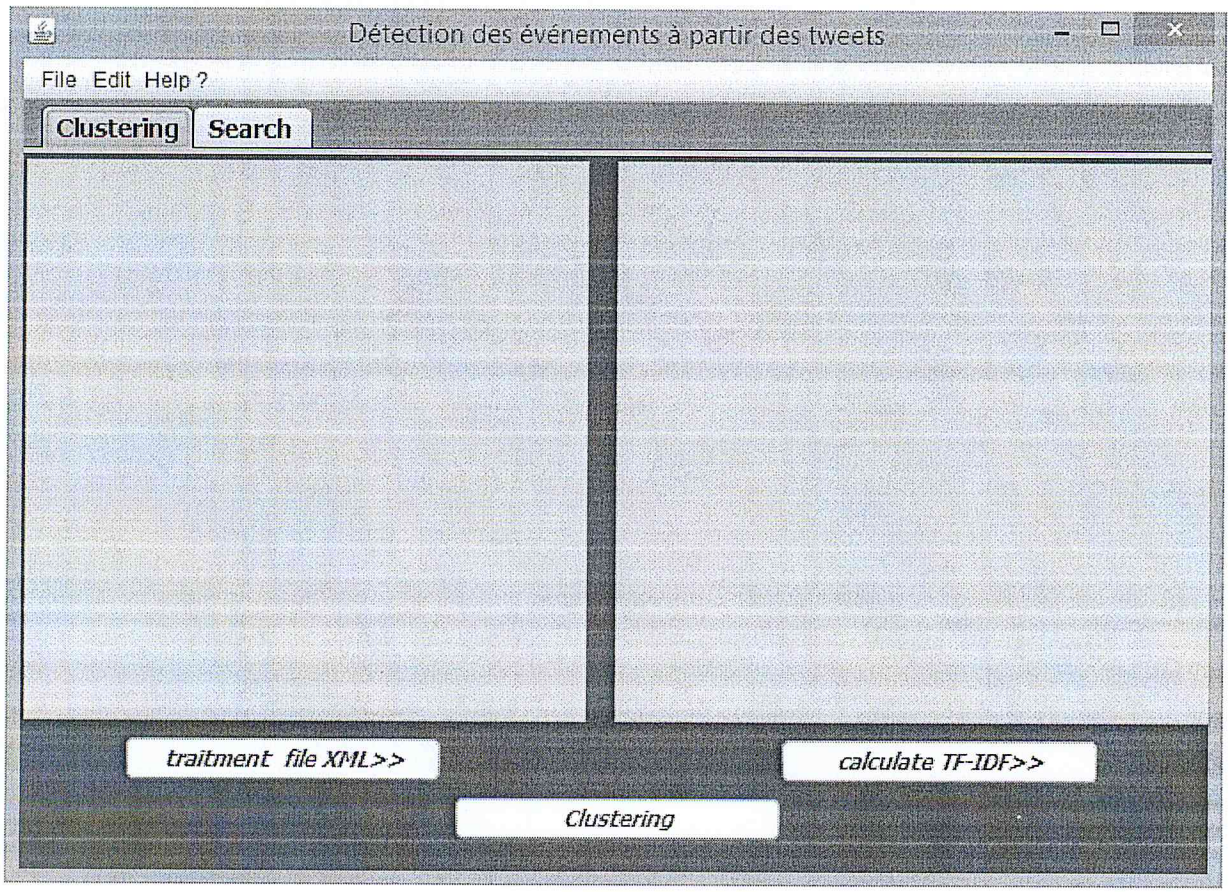
Figure 5: Architecture globale de l'application

La figure 5 visualise comment l'application fonctionne en interne. L'utilisateur communique avec l'ouverture du fichier XML de l'origine de API Twitter , un traitement est effectué au fichier XML afin d'extraire les données, en suite l'ensemble des données subit des filtrages avant de se charger dans la base de données .Le système calcul le TF-IDF de chaque mot des tweets déjà traité et stocké dans la base de données, la liste des mots avec leurs poids (calculé par TF-IDF) sera sauvegardé dans la base de données.

Et puis faire le pré-clusteringpour classifier l'ensemble des termes

Un utilisateur peut faire la recherche locale par la date de publication les tweets sauvegardé dans la base de donnée et faire le pré-clustering.

Après l'exécution de l'application, une interface est affichée

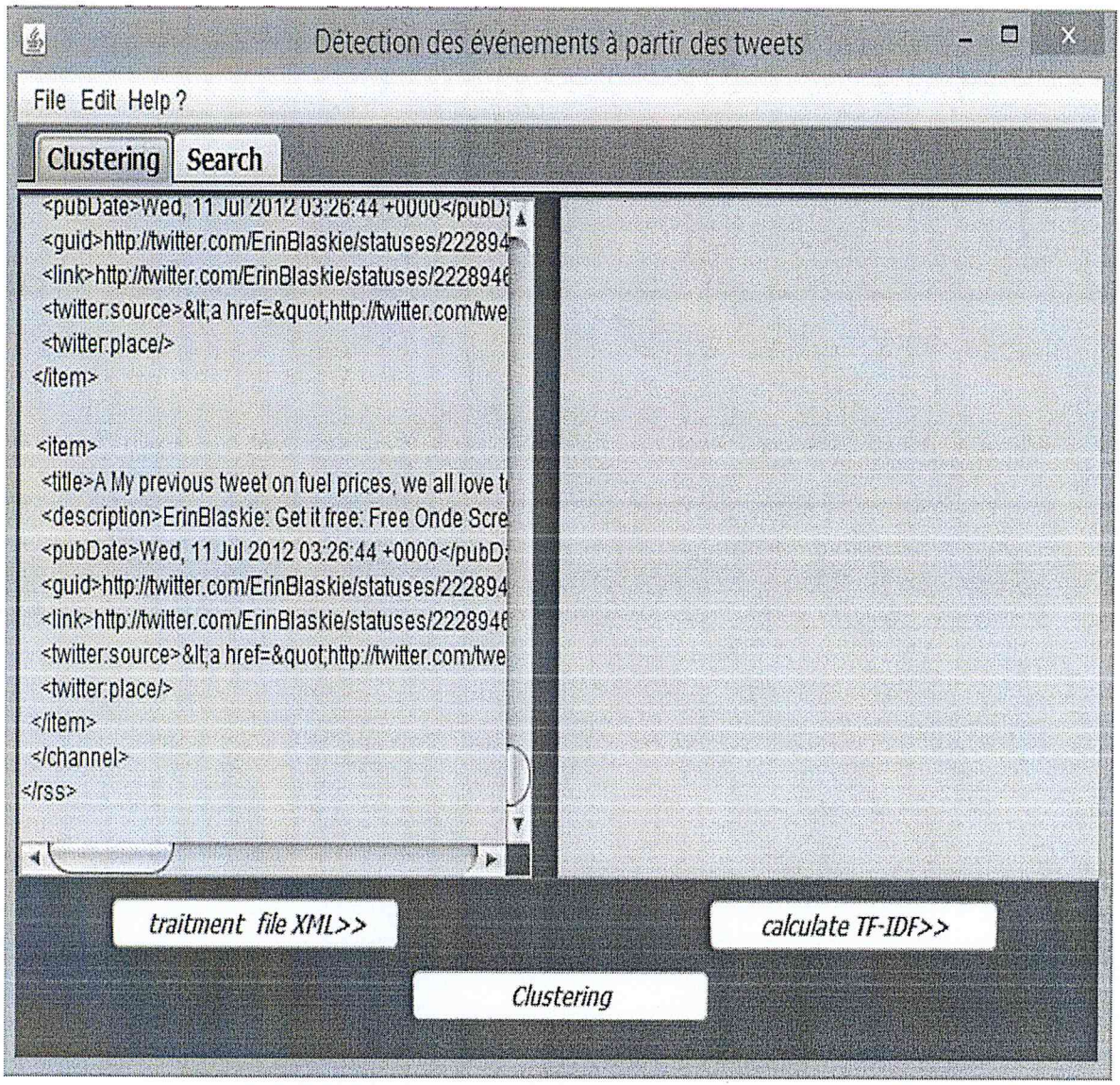


Le menu se compose de trois boutons « File» ouvre le menu indique à l'utilisateur qu'il doit importer le fichier contenant l'ensemble de données à analyser, « Edit» éditer un fichier XML avant de l'analyser et « help» Le bouton HELP pour aider les utilisateurs à comprendre les étapes du système.

I. Le pré-Clustering

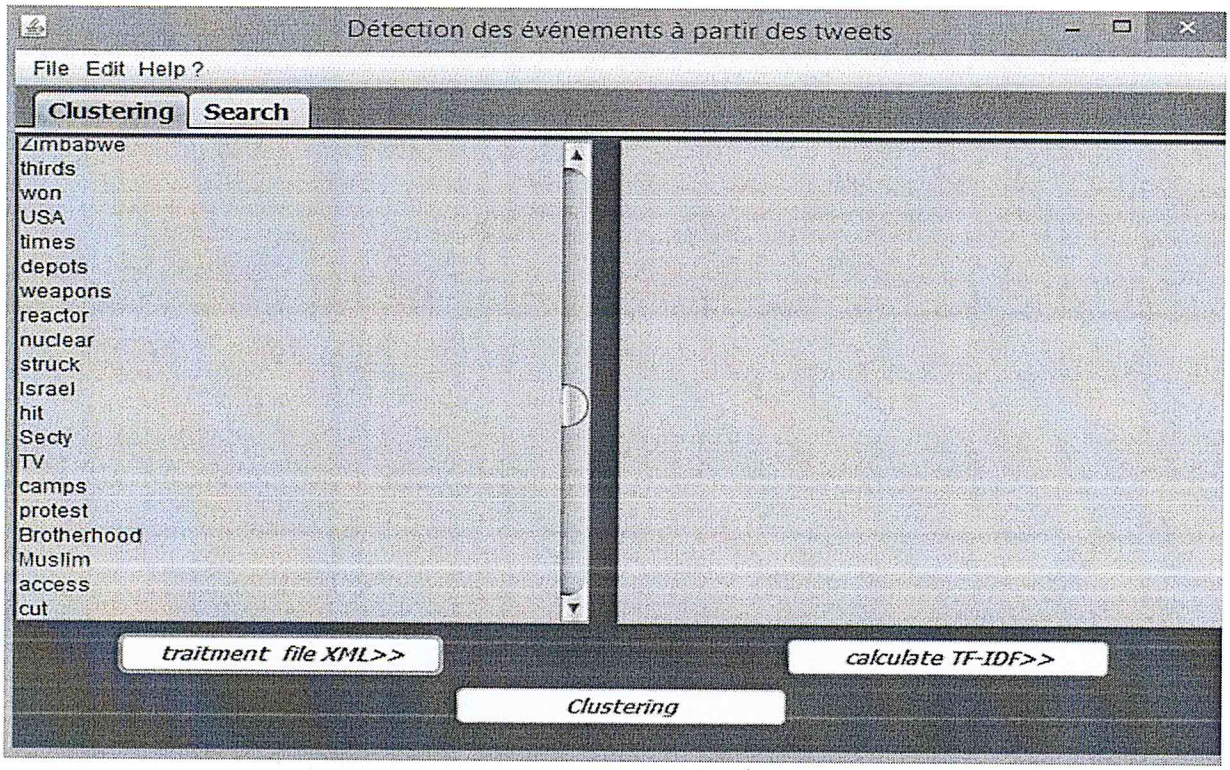
Quand on clique sur le boutons «file » puis sur « open » qui faire importation locale de fichier XML .et aussi un bouton « en ligne » qui permet d'importer le fichier xml en ligne.

4.2.1 Interface d'importation du fichier XML



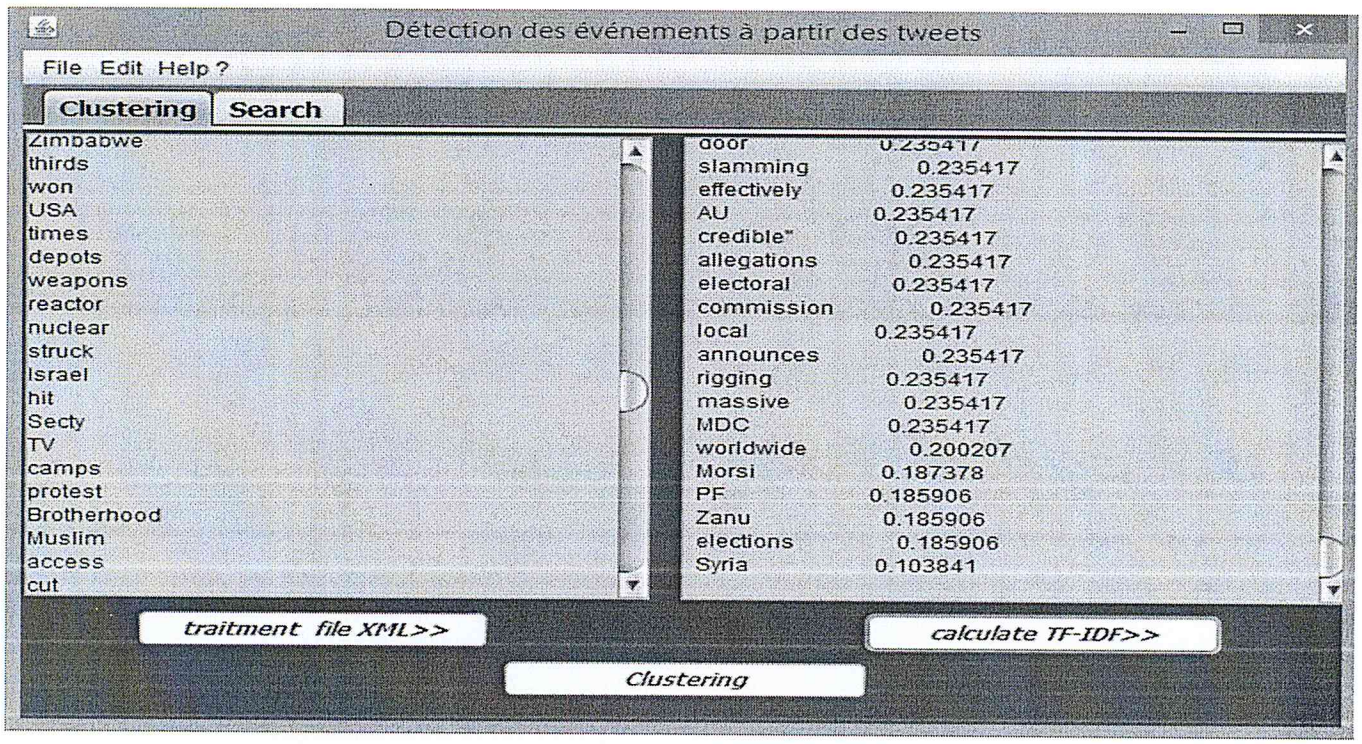
4.2.2 Le prétraitement

Le prétraitement se fait àpartir dubouton « traitement file XML ».



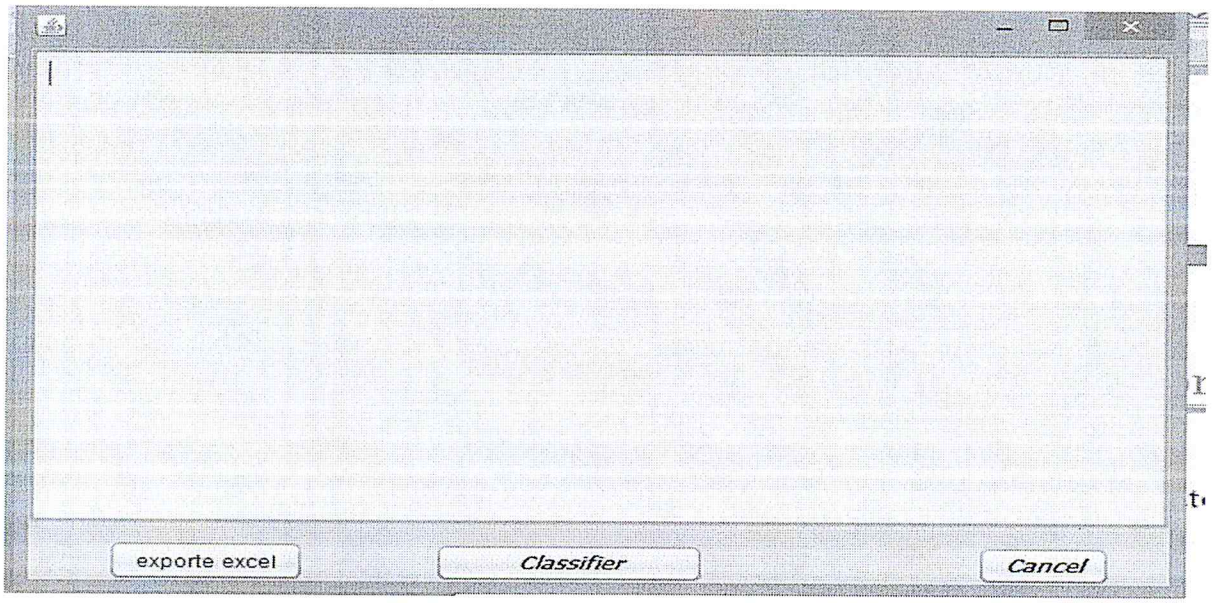
4.2.3 Calcul du poids

Une fois le traitement de données est terminé, on procède à l'étape du calcul du TF-IDF de chaque terme dans la collection pour permettre de distinguer les termes les plus lourds. Ces derniers ont plus de chance d'être un événement sur Twitter.

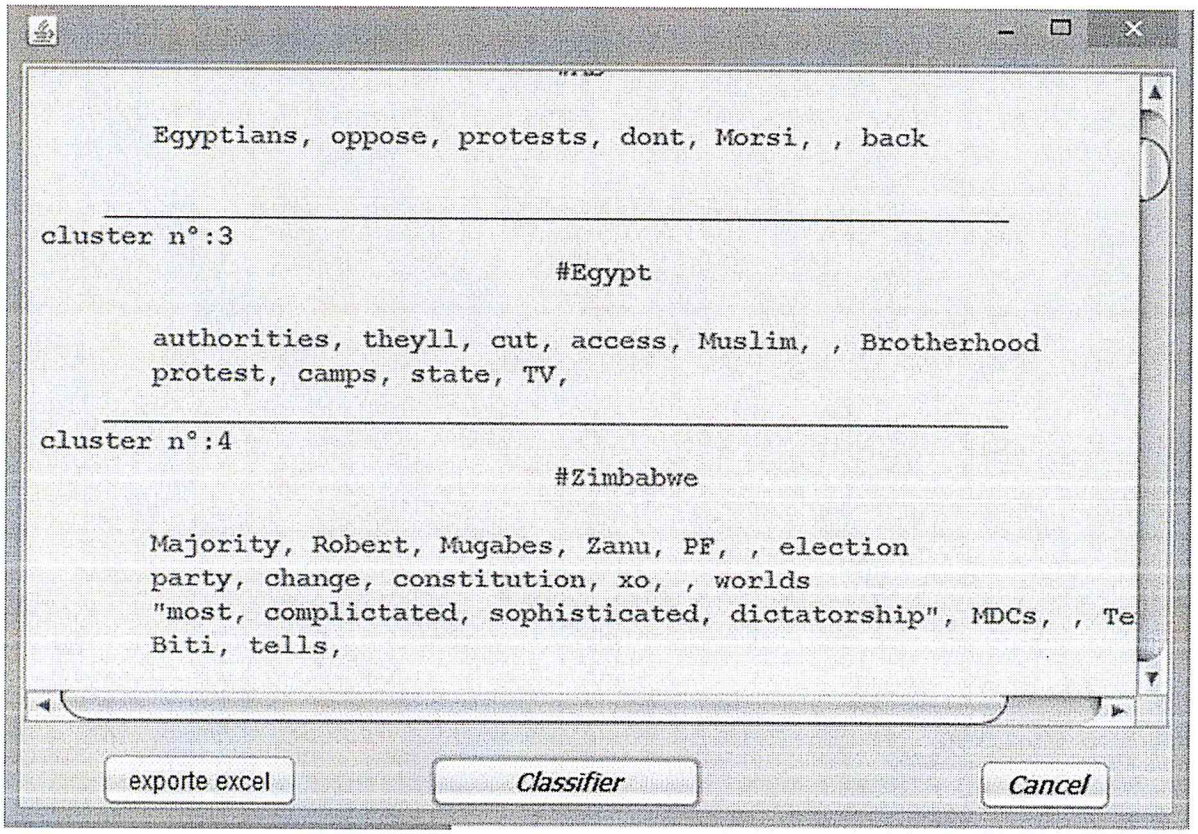


4.2.4 La classification

Quand on clique sur le bouton « Classifier » l'interface suivant sera affiché :

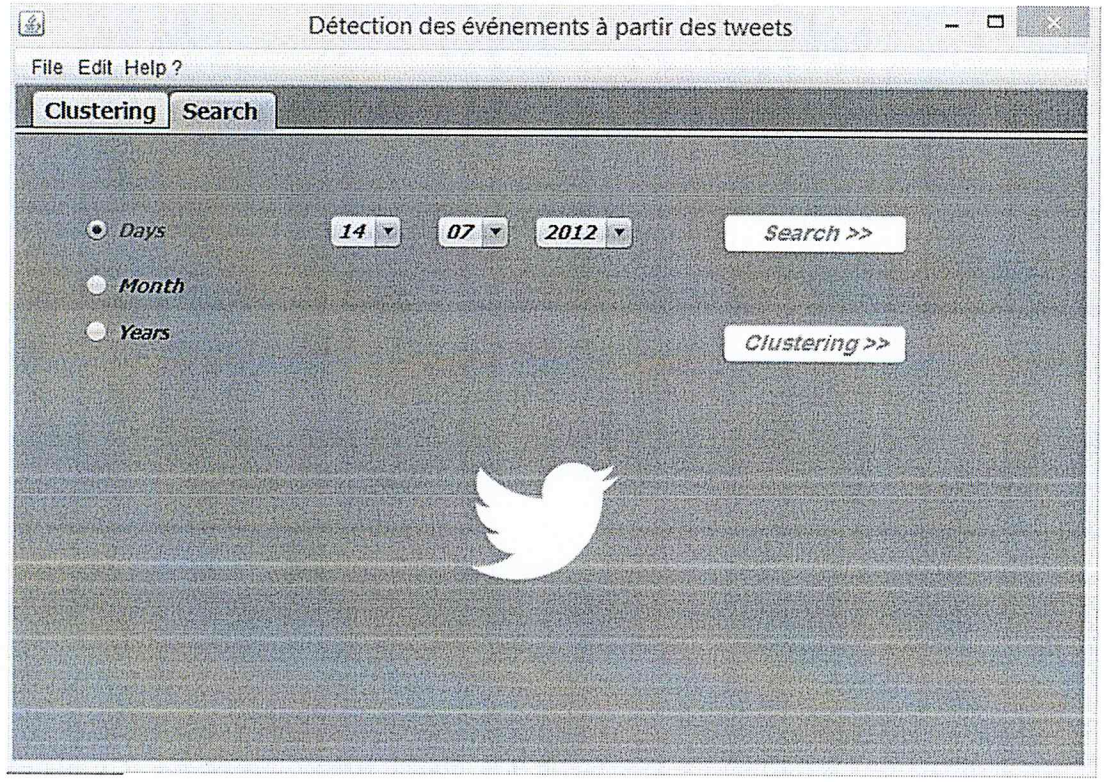


On clique sur le bouton « **Classifier** »

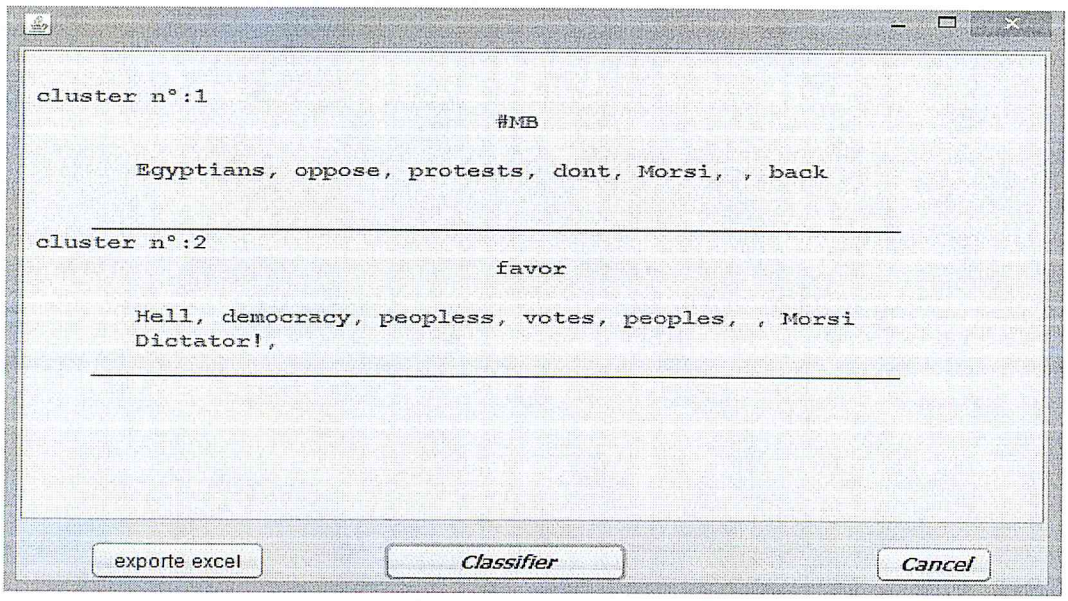


La recherche locale

L'utilisateur peut faire la recherche dans les tweetsstockésdans la base dedonnéesCette recherche permet de récupérer les tweets dansune date (jour. moi,année) etpréciser par utilisateur et refaire le pré-clustering.

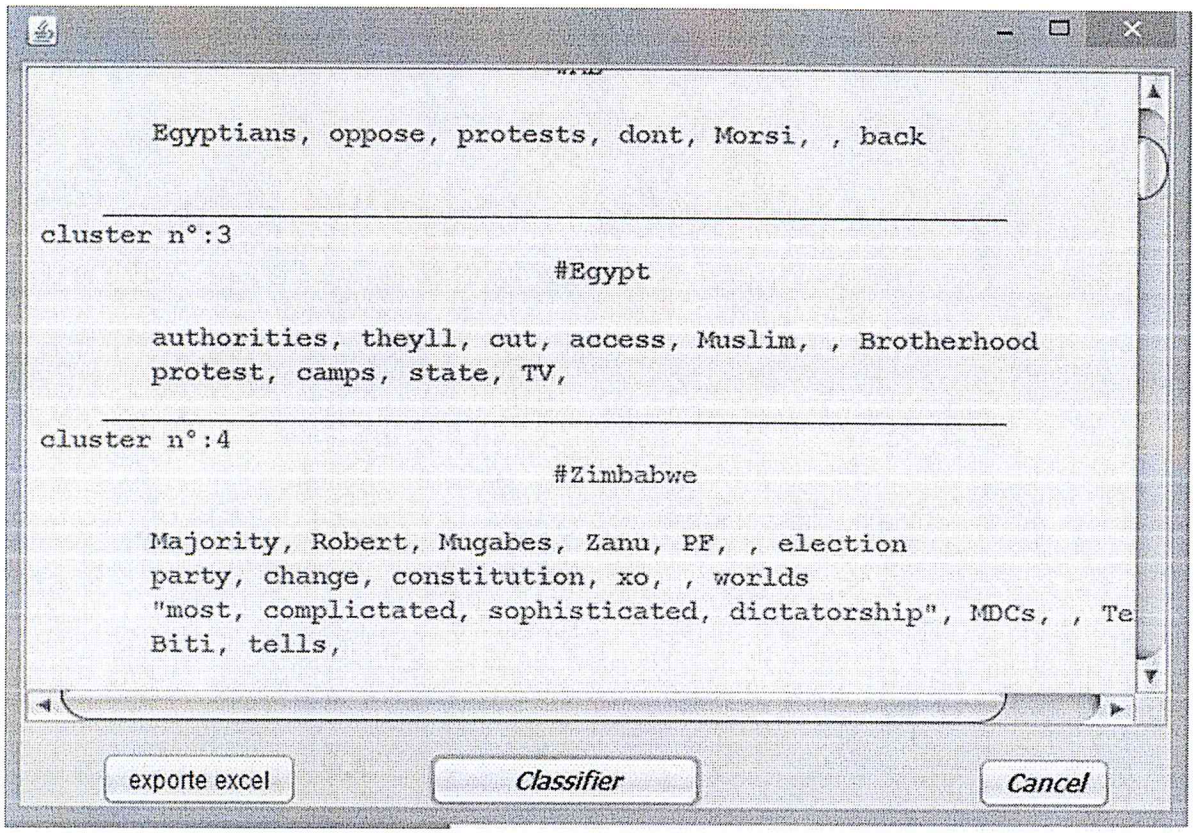


Alors on clique sur le bouton **« Search »** et puis sur **« Clustering »**



II. Clustering avec k-means sous TANAGRA

Dans cette étapeongénère un fichier Excel qui contient les vecteurs /tweet puis on applique le logiciel Tanagra pour faire la classification avec k-means



On clique sur le bouton « exporte Excel » le système faitl'exportation d'un fichier Excelquicontient les vecteur/tweet. Comme ceci :

	articel	#MB	#Egypt	#Zimbabwe	post	top	Mata	fire	wanted
2	tweet 1	0	0	0	0	0	0	0	0
3	tweet 2	0	0	10	0	0	0	0	0
4	tweet 3	14	0	0	0	0	0	0	0
5	tweet 4	0	0	0	0	0	0	0	0
6	tweet 5	0	10	0	0	0	0	0	0
7	tweet 6	11	0	9	0	0	0	0	0
8	tweet 7	0	0	0	0	0	0	0	0
9	tweet 8	0	0	0	0	0	0	0	0
10	tweet 9	0	0	0	0	0	0	0	0
11	tweet 10	0	0	0	0	0	0	6	0
12	tweet 11	6	0	0	0	0	0	0	0
13	tweet 12	0	0	0	0	0	0	0	0
14	tweet 13	0	0	0	4	0	0	0	0
15	tweet 14	0	0	0	0	0	0	0	0

Chapitre 3 : Conception et Implémentation

Apres on applique les différentes étapes dulogiciel tanagra pour exécuter l'algorithme k-means

En appliquant *k-means* à l'aide du Tanagra sur le fichier Excel et en mettant le nombre de clusters à 3, trois clusters sont construits. La figure ci-dessous montre les résultats du *clustering* à l'aide de Tanagra.

K-Means parameters	
Clusters	3
Max Iteration	15
Trials	10
Distance normalization	none
Average computation	Forgy
Seed random generator	Random

Global evaluation

Within Sum of Squares	1796,1600
Total Sum of Squares	2246,2963
R-Square	0,2004

Cluster size and WSS

Clusters 3

Cluster	Description	Size	WSS
cluster n°1	c_kmeans_1	1	0,0000
cluster n°2	c_kmeans_2	25	1796,1600
cluster n°3	c_kmeans_3	1	0,0000

Cluster centroids

Attribute	Cluster n°1	Cluster n°2	Cluster n°3
#MB	0,000000	0,680000	14,000000
#Egypt	0,000000	0,400000	0,000000
#Zimbabwe	0,000000	0,760000	0,000000
post	0,000000	0,160000	0,000000
top	0,000000	0,200000	0,000000
Mata	0,000000	0,200000	0,000000
fire	0,000000	0,240000	0,000000
wanted	0,000000	0,240000	0,000000
coupe	0,000000	0,280000	0,000000
PM	0,000000	0,280000	0,000000
Hell	0,000000	0,280000	0,000000
becuz	0,000000	0,280000	0,000000
kid	0,000000	0,280000	0,000000
blame	11,000000	0,000000	0,000000
bang	0,000000	0,320000	0,000000
president	0,000000	0,320000	0,000000
fair"	0,000000	0,360000	0,000000
partiament	0,000000	0,280000	13,000000
USA	0,000000	0,400000	0,000000
Secty	0,000000	0,400000	0,000000
gravity	0,000000	0,480000	0,000000
department	0,000000	0,480000	0,000000
endless	0,000000	0,480000	0,000000
observers	0,000000	0,520000	0,000000
Now	0,000000	0,080000	0,000000

v

R-Square for each attempt

Number of trials	10
Trial	R-square
1	0,100564
2	0,191416
3	0,093299
4	0,176030
5	0,169620
6	0,074494
7	0,200390
8	0,117974
9	0,059109
10	0,068084

Et voici une figure contient les détails du chaque cluster :

Cluster size and WSS

Clusters	3		
Cluster	Description	Size	WSS
cluster n°1	c_kmeans_1	1	0,0000
cluster n°2	c_kmeans_2	25	1796,1600
cluster n°3	c_kmeans_3	1	0,0000

articel	Cluster_KMeans_1
tweet 1	c_kmeans_2
tweet 2	c_kmeans_2
tweet 3	c_kmeans_3
tweet 4	c_kmeans_2
tweet 5	c_kmeans_2
tweet 6	c_kmeans_2
tweet 7	c_kmeans_2
tweet 8	c_kmeans_2
tweet 9	c_kmeans_2
tweet 10	c_kmeans_2
tweet 11	c_kmeans_2
tweet 12	c_kmeans_2
tweet 13	c_kmeans_2
tweet 14	c_kmeans_2
tweet 15	c_kmeans_2
tweet 16	c_kmeans_2
tweet 17	c_kmeans_2
tweet 18	c_kmeans_2
tweet 19	c_kmeans_2
tweet 20	c_kmeans_2
tweet 21	c_kmeans_2
tweet 22	c_kmeans_2
tweet 23	c_kmeans_2
tweet 24	c_kmeans_2
tweet 25	c_kmeans_2
tweet 26	c_kmeans_2

5. Test

5.1 Collections de tests

Une collection des tweets est utilisée pour l'évaluation de notre approche. La Collection comporte 100 tweets des publications sur twitter des chainestélévisées anglo-saxonne : BBC,CNN, ABC, ce choix est justifié par la Syntaxe correcte de tweets et leur contenu bien varié.

5.2 Résultats

Nombre total de tweets	**100**
Nombre de mots distincts	**520**
La moyenne de nombre de caractères par tweet	**60**
La durée moyenne éculée entre 2 tweets	**50 min**

Tableau 3 : Analyse générale des tweets

Apartir de notre approche qui faire le pré-clustering et le résultat d'algorithme de k-means sous Tanagra le tableau suivant représenter la comparaison entre le pré-clustering et k-means.

Partie	Nombre de Clusters (évènement)	Rappel	Precision	F-mesure
Pré-clustering	25	1	0,7	0.82
k-means	3	1	1	1

Tableau 4 : les résultats de l'évaluation.

Pour même collecte des tweets et même traitement des données nous remarquant que les résultats sont toujours bons dans le pré-clustering et l'algorithme de k-means.

Le clustering avec k-means est appliqué sur un fichier Excel c générées par le logiciel Tanagra. De bonnes valeurs des mesures d'évaluation ont été obtenues, ce qui prouve que notre approche pré-clustering des résultats très intéressants.

6. Conclusion

Dans ce chapitre nous avons proposé les différentes phases constituant notre approche de fouille dans les tweets et plus précisément la détection des événements à partir des tweets.

Conclusion

Dans cette étude, nous avons proposé un système de détection des événements à partir des messages tweets en utilisant le calcule de TF-IDF ainsi une classification qui définit bien les événements.

Notre contribution dans les objectifs du travail est que par sa capacité à extraire les sujets émergents d'une collection tweets. Après collecté les tweets sous format XML notre system fait un prétraitement ainsi un calcule de TF-IDF qui définit le poids de chaque mot dans tous les messages tweets, ensuit faire le pré-clustering pour classifier l'ensemble des termes.

Il reste un énorme travail à accomplir dans le domaine du tweet mining, le plus difficile serait de suivre le rythme infernal imposé par les utilisateurs de twitter, qui font introduire chaque jour des nouveaux termes et phrases et de nouvelles manières d'exprimer. Les résultats que nous obtenons devraient être considérés comme des résultats préliminaires et l'étude peut être prolongée dans plusieurs directions. Tout d'abord, nous prévoyons de prolonger l'analyse pour une plus longue collection de tweets périodes, y compris d'autres événements. En outre, l'étude sur des tweets comprenant plusieurs hashtags pour la détection.

Bibliographie

[1] **Kyle W. Prier et al**. Identifying Health-Related Topics on Twitter An Exploration of Tobacco-Related Tweets as a Test Topic. 2011.

[2] **URL:** http://twitter.about.com/od/Twitter-Basics/a/The-Real-History-Of-Twitter-In-Brief.htm

[3] **Mario Cataldi,Luigi Di Caro, Claudio Schifanella**. Emerging Topic Detection on Twitter based on Temporal and Social Terms Evaluation. 2010 .

[4] **Christopher Horn**. Analysis and Classification of Twitter messages. 2010.

[5] **Laurent Mignon, Elodie Buch**. Twitter en action. 2012.

[6] **Dusty Reagan**. Twitter Application Development for Dummies. 2010.

[7] **Graham Williams**. Data Science with R Text Mining. 2014.

[8] **Fabien Chevalier**. La classification. 2012.

[9] **Maurice ROUX.** Algorithmes de classification. 2013.

[10] **Hamou Reda Mohamed , Ahmed Lehireche** . La classification non supervisee (clustring) de document textuels. 2008.

[11] **Sriram et al.** Short text classification in twitter to improve information filtering. 2010.

[12] **Sakaki et al**. Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors. 2011.

[13] **Dridi**. Détection des évènements a partir des tweets. 2012.

[14] **Ozdikis et al.** Semantic Expansion of Hashtags for Enhanced Event Detection in Twitter. Middle East Technical University Ankara, Turkey. 2012.

[15] **Jones, K.** A statistical interpretation of term specificity and its application in retrieval. 1972.

[16] **Christopher D. Manning, Prabhakar Raghavan.** Introduction to Information Retrieval. Cambridge University Press. 2008.

[17] **URL :** http://www.java.com/fr/download/faq/whatis_java.xml.

[18] **Site officiel de My SQL. URL** : http://www.mysql.com/.