

TRAVAIL PERSONNEL ENCADRE



RAPPORT FINAL : PARTIE THÉORIQUE

THÈME

L'information géolocalisée des réseaux sociaux pour la Ville Intelligente

RÉDIGÉ PAR

MILORME Pierre Rubens

ENCADRE PAR

- CAMP Olivier Collaboration externe (ESEO, Angers, France)
- Dr. Ho Tuong Vinh (IFI)

**Année Académique 2016-2017
Promotion 21**

SOMMAIRE

Introduction et analyse du sujet.....	5
Présentation de Twitter	6
Exploitation des données de Twitter	6
Technologies à mettre en œuvre.....	6
Etat de l'art.....	8
Partie I.....	9
Twitter et la géolocalisation.....	9
Analyse des Sentiments (sentiment analysis).....	9
Twitter et Analyse de sentiment.....	10
PARTIE II.....	11
Méthodes et techniques utilisées.....	11
Acquisition des données depuis Twitter.....	11
Stockage des données.....	11
Méthodes de traitement et de classification des messages Twitter.....	12
TF-IDF.....	12
BOW.....	12
LAD.....	12
SVM, Naive Bayes, N-gram.....	12
Méthodes de visualisation des messages de Twitter sur une carte.....	12
Outils de visualisation spatiale.....	14
Solutions propriétaires.....	14
Solutions Opensource.....	15
Outils de navigation et de visualisation de la chronologie.....	16
Outils de visualisation de diagramme.....	17
PARTIE III.....	18
Exemples d'applications.....	18
Solution proposée.....	22
Solution proposée.....	23
Outils à utiliser.....	26
Expérimentations, évaluation et validation.....	28
Risques encourus, difficultés et éventuelles contraintes.....	28
Planification des tâches.....	29
Conclusion.....	30
References.....	31

Cette page est laissée vide intentionnellement.

Mots clés : *TPE (travail personnel encadre), information, géolocalisation, information géolocalisée, réseaux sociaux, Ville Intelligente, Twitter, sentiment analysis, topic classification, TF-IDF, NLP, UGC, POI (Places of Interest), topic extraction.*

INTRODUCTION

Analyse du sujet

Les applications pour la Ville Intelligente s'appuient sur de la collecte de données et l'utilisation de ces données pour fournir des services pertinents aux utilisateurs (*OpenData*). Ces données peuvent provenir de plusieurs sources : caméra, capteurs météo, ... et d'autres technologies telles que les téléphones intelligents (*smartphones*), le *Cloud Computing*, l'*Internet des Objets (IoT)* ou encore les réseaux sociaux. Dans notre cadre d'étude, nous pensons que ces derniers peuvent constituer une source d'informations riche. L'idée de ce travail est d'étudier comment un réseau social tel que Twitter peut être utilisé pour produire des informations pertinentes pouvant servir à la Ville Intelligente. En d'autres termes, notre objectif est d'étudier la façon dont Twitter peut être utilisé comme capteur dans le cadre de la Ville Intelligente.

Présentation de Twitter

Twitter [7] est une plateforme de réseau social de microblogage permettant à ses utilisateurs d'envoyer des messages courts (pas plus de 140 caractères). Son principe de fonctionnement s'énonce ainsi : des utilisateurs diffusent des *tweets* accessibles par tous, certains utilisateurs peuvent en suivre d'autres (*followers*), certains *tweets* peuvent être '*retweetés*'. Du point de vue du format des messages les *tweets* contiennent en plus du contenu textuel des métadonnées dont des mots clefs (*hashtag*) permettant de catégoriser les informations qu'ils véhiculent. Certains *tweets* peuvent contenir également une information permettant de les géolocaliser. En raison du caractère spontané et immédiat des messages publiés et la facilité d'accès offerte par ses API, Twitter constitue une plateforme idéale d'expérimentation dans le cadre du travail que nous aurons à effectuer dans ce projet.

Exploitation des données de Twitter

Plusieurs études se sont déjà penchées sur l'analyse des réseaux sociaux tels que *Twitter* (ou *FourSquare*) dans le cadre d'une ville soit pour étudier la dynamique d'une ville [2], soit pour analyser la géolocalisation des *tweets* (pour trouver des lieux depuis lesquels sont produits beaucoup de *tweets* non géo-étiquetés) [1], soit pour analyser le contenu des *tweets* afin déterminer "l'humeur" générale des habitants d'une ville par exemple (sentiment analysis) [3]. Les systèmes de recommandation de services contextuels personnalisés[5,6] ou le suivi de déplacement et des centres d'intérêt des populations et communautés[4] sont autant d'exemples potentiels d'application qui peuvent découler de l'analyse des données provenant de ces plateformes.

Dans notre cadre d'étude, nous commencerons tout d'abord par une étude approfondie des travaux qui constituent l'état de l'art en nous appuyant sur les études précédemment citées notamment. Il s'agira essentiellement d'étudier les travaux effectués sur l'utilisation des réseaux sociaux dans l'extraction d'informations sur une ville et sur ses habitants. Nous dresserons la liste des informations pouvant être déduites, recenserons les techniques mises en œuvre pour la déduction de ces informations, étudierons et repertorions également les types d'applications existants utilisant ces informations.

Technologies à mettre en œuvre

Nous avons comme objectif pratique dans le contexte de ce travail personnel encadré de proposer un outil qui permettra de déduire des informations sur une ville ou ses habitants à partir des données récupérées depuis Twitter. Des exemples d'informations pouvant être déduites sont : les lieux fréquentés en fonction du jour, de l'heure ou de la saison, les utilisateurs ou communautés

influent, les communautés d'utilisateurs partageant des intérêts communs, les centres d'intérêt des utilisateurs.

Dans la partie pratique de notre travail nous serons amenés à manipuler plusieurs plateformes et technologies afin d'atteindre cet objectif. Twitter propose plusieurs API publiques permettant aux développeurs d'accéder aux données et de les manipuler. Une bonne compréhension du fonctionnement de ces API s'avère donc nécessaire pour la phase d'acquisition des données sur lesquelles reposent les applications pour la Ville Intelligente. La phase de recherche bibliographique et de l'état de l'art qui suivra cette mise en contexte sera l'occasion d'approfondir sur ce sujet. Afin de garder les relations telles que l'abonnement à un compte (*followers*) et les connexions qui existent entre les données, il est possible qu'on mette à contribution une base de données de type graphe telle que *Neo4j* [8]. La visualisation des informations obtenues se faisant sur une carte, *OpenStreetMap (OSM)* [9] semble être un bon compromis pour accomplir une telle tâche. En exploitant les API proposés par cette plateforme, nous espérons être capables d'assurer l'exploration des informations déduites dans le temps et dans l'espace.

Toutefois, il faudra patienter et attendre d'avoir fait une étude de l'état de l'art du domaine pour justifier nos choix des plateformes et technologies que nous envisageons d'utiliser à l'issue de ce projet : les technologies citées précédemment sont donc sujet au changement et ne sont en aucun cas définitives.

ÉTAT DE L'ART

De nombreuses applications exploitant les données générées par les réseaux sociaux dans le cadre de la Ville Intelligente s'appuient sur les informations de localisation qui accompagnent les messages transmis [12]. Connaître le lieu exact depuis lequel une information a été émise s'avère donc essentiel et parfois crucial et nécessaire dans le contexte de ces applications [11][12]. Nous avons assisté récemment à un accroissement et une multiplication fulgurante de ces réseaux sociaux. La majorité d'entre eux offrent la possibilité d'incorporer des informations de localisation dans les messages (*Twitter*, *Foursquare*, *Facebook*, ...), certains pour leur part se basent essentiellement sur ces informations pour fonctionner (*Foursquare*), tandis que d'autres proposent le choix à l'utilisateur de partager ou pas ces informations. Dans notre cadre d'étude nous nous intéressons à l'information géolocalisée des réseaux sociaux dans le cadre de la Ville Intelligente. Twitter étant le réseau social que nous allons considérer dans ce travail.

Cette section est structurée en trois grandes parties. Dans la **Partie I** sont exposés le concept de la géolocalisation dans le contexte de Twitter, les problématiques inhérentes aux applications nécessitant ces informations de localisation pour fonctionner et les solutions proposées aux contraintes évoquées. Est développé également le concept de l'analyse de sentiment (*sentiment analysis*). Nous poursuivons dans la **Partie II** en répertoriant les informations pouvant être déduites depuis les données de Twitter, les méthodes et techniques utilisées et les solutions de visualisation des informations sur une carte. Nous concluons dans la **Partie III** en présentant certaines applications réelles ayant rapport avec notre sujet d'étude utilisant les informations et techniques mentionnées notamment.

PARTIE I

Twitter et la géolocalisation

Twitter est un réseau social de microblogage. Les caractéristiques de ce réseau social telles que la longueur très réduite des messages publiés (140 caractères) et l'aspect presque temps réel lui confèrent une place de choix dans le cadre des études menées sur les réseaux sociaux afin d'en extraire des informations[11][12]. Mais la quantité très réduite des messages comportant les informations de géolocalisation [12][37][38][39] (de l'ordre de 1.5% à 3% des tweets seulement) constitue un obstacle majeur au développement et à l'utilité d'applications exploitant ces informations. Certaines études et travaux considèrent seulement les tweets explicitement géolocalisés [13][15][16], d'autres tendent à contourner ce problème en proposant des méthodes pour estimer la localisation des tweets. Parmi ces propositions, certaines considèrent une échelle spatiale assez grande [30][31][39][40][41](l'échelle d'une ville ou de régions plus grandes qu'une ville) alors que d'autres évoluent dans un contexte spatial de granularité plus fine, au niveau d'un quartier d'une ville [12][28][29][32]. La proposition de Pavlos Paraskevopoulos et Themis Palpanas dans [2] est particulièrement intéressante : l'utilisation de la technique TF-IDF (*Term Frequency-Inversed Document Frequency*) afin d'extraire des mots clés dans les tweets suivi du calcul de similarité basé sur le contenu et les caractéristiques d'évolution temporelle entre un tweet non géolocalisé et un ensemble de tweets constituant l'ensemble d'entraînement(dont la localisation est connue) permet d'inférer une localisation et l'attribuer à ce tweet. Il convient de considérer ces différents travaux qui ont été réalisés afin d'avoir des analyses et des résultats avec un niveau de granularité beaucoup plus fin lors de développement d'applications reposant sur les informations géolocalisées de Twitter.

Analyse des Sentiments (sentiment analysis)

L'analyse du sentiment (*Sentiment Analysis*), également appelée extraction d'opinions, est le domaine d'étude qui analyse les opinions, les sentiments, les évaluations, les attitudes, et les émotions envers des entités telles que des produits, des services, des organisations, individus, problèmes, événements, sujets et leurs attributs relatifs [20][22]. C'est un domaine faisant partie d'un domaine beaucoup plus vaste qui est la fouille (ou analyse) de données textuelles et le traitement des langues naturelles[21]. Ci dessous un arbre décrivant les techniques utilisées pour faire l'analyse des sentiments[13]. Deux grandes catégories de méthodes sont identifiées : celles basées sur l'apprentissage automatique et celles qui font appel à l'analyse du lexique :

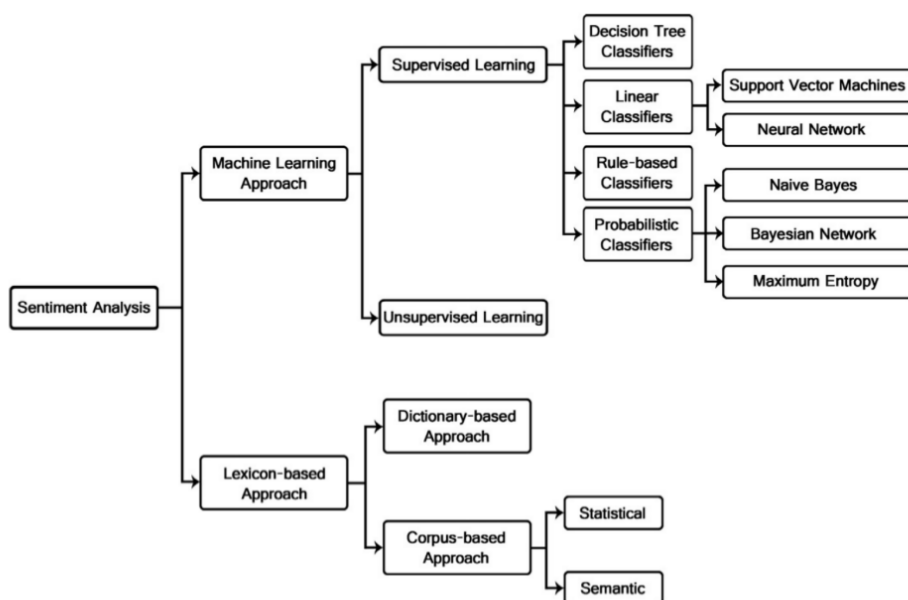


Figure 1 : Techniques utilisées pour l'analyse de sentiment

Twitter et Analyse de sentiment

L'analyse de sentiment dans les messages publiés sur Twitter a été l'objet de beaucoup de travaux[24][25]. Certains de ces travaux s'intéressent à l'analyse de l'humeur des habitants dans le cadre d'une ville en représentant ces informations sur une carte(*sentiment map*)[13][23] afin d'évaluer les fluctuations périodiques des sentiments dans le temps et dans l'espace [16]. Dans [16], les chercheurs utilisent un algorithme de classification supervisée basé sur les *emoticons* ayant comme ensemble d'entraînement des tweets comportant des *emoticons* et déterminent le sentiment général des blocs de recensement en agrégeant les sentiments dans ces blocs. Parmi les autres informations qui peuvent être étudiées, la propagation de maladies : les informations déduites peuvent être également d'une grande utilité au domaine médical et sanitaire (*Infodemiology and Infoveillance*) [26][27].

Beaucoup d'autres informations peuvent être déduites en analysant les données de Twitter dont l'analyse du trafic urbain[33][36]. Les systèmes de recommandations contextuelles[17][18] exploitant les informations spatio-temporelles des données de Twitter constituent également un enjeu majeur parce qu'ils permettent de prendre en compte dans le processus de recommandation le contexte spatio-temporel de l'utilisateur. Dans [14], ce sont les habitudes de déplacement des habitants qui sont analysées où l'on essaie de caractériser et déterminer des communautés d'individus partageant les mêmes centres d'intérêts. Des informations sur les désastres naturels[34]

[35] peuvent en outre être obtenues. Dans [15], on essaie d'analyser la dynamique de la ville de New York en fonction du jour de la semaine et de l'heure à travers le regard de Twitter.

PARTIE II

Méthodes et techniques utilisées

A) Acquisition des données depuis Twitter

Dans le cadre de l'analyse des données provenant du réseau social Twitter, la première étape à entreprendre est la phase d'acquisition des données. Trois APIs (*Application Programming Interface*) sont à notre disposition pour accomplir cette tâche : l'*API Search* [48] permettant de consulter des messages datant au plus des sept derniers jours, l'*API REST 1.1* limité à 450 demandes toutes les 15 minutes permettant à certaines applications de publier des tweets et de s'abonner à d'autres comptes Twitter et de récupérer des données historiques notamment et l'*API Streaming* permettant de récupérer et traiter les tweets en temps réel à partir du moment où la requête a été soumise [42][47] ; ce dernier vise l'ensemble des messages répondant à une requête dont la limite difficilement atteignable est de l'ordre de 1% des messages à un instant t .



Figure 2 : APIs de Twitter

B) Stockage des données

Les données récupérées doivent être stockées de manière adéquate et bien organisée afin de pouvoir être retrouvées et traitées. Plusieurs solutions de stockage ont été proposées dans les études passées et récentes parmi lesquelles l'utilisation d'une base de données relationnelles [47] et de bases de données NoSQL dont le moteur de bases de données MongoDB [45] (qui est une base de données de type document) et récemment les bases de données de graphes [43][27]. Les bases de données de type graphe sont particulièrement intéressantes dans le cadre de l'analyse des réseaux sociaux compte tenu du fait que la structure de stockage en nœuds et relations utilisée facilite la représentation de la structure des réseaux sociaux. Cette représentation facilite l'organisation des données ayant un haut degré de connexion entre elles et simplifient grandement les applications basées sur la recommandation et l'analyse de données textuelles telles que le sentiment analysis ou le traitement du langage naturel (Natural Language Processing -NLP).



PostgreSQL
the world's most advanced open source database

Figure 3 et 4 : Bases de données relationnelles

Figure 5 et 6 : Bases de données NoSQL



C) Méthodes de traitement et de classification des messages Twitter

Faire une liste afin de répertorier toutes les recherches qui ont été menées à propos de l'extraction et l'analyse d'informations provenant des réseaux sociaux, les méthodes et techniques utilisées est quasiment impossible tant le domaine est vaste. Cependant dans les lignes qui suivent nous avons tenté d'exposer certaines méthodes et techniques que nous avons eu l'occasion d'analyser au cours des recherches entreprises :

a) TF-IDF (Term Frequency - Inverse Document Frequency) est une méthode d'affectation de poids à un mot en se basant sur la fréquence d'apparition de ce mot dans un document mais aussi dans tous les documents du corpus. Si le mot apparaît plusieurs fois dans un document mais pas souvent dans tous les documents alors il aura un poids fort, par contre, si le mot apparaît souvent dans les documents son importance diminue. Cette technique est très utilisée dans l'extraction de mots clés [12].

b) BOW (Bag of Words) est une technique souvent utilisée en combinaison avec la technique TF-IDF pour effectuer l'analyse fréquentielle. Elle permet également de faire la classification de texte.

c) LAD (Latent Dirichlet Allocation) est un algorithme probabiliste de modélisation de sujet afin d'aider à la modélisation des documents et à la classification du texte [46].

d) La plupart des méthodes de classification de sentiment basée sur l'apprentissage automatique sont des méthodes supervisées parmi lesquelles **SVM (Séparateurs à vastes marges ou Support Vector Machine)**, **Naive Bayes**, ou la **méthode N-gram** [13].

D) Méthodes de visualisation des messages de Twitter sur une carte

La représentation des tweets ou des informations dérivées de ces tweets sur une carte permet une analyse plus rapide et une interprétation plus aisée et intuitive, lorsque l'utilisateur auquel le résultat final est destiné n'est pas un expert du domaine surtout. Cette tâche peut être accomplie de plusieurs façons différentes et le choix d'une méthode de représentation est surtout fonction de l'objectif final et de l'information que l'on cherche à extraire et représenter visuellement :

Représentation sous forme de	Descriptions
Markers¹ ou point map (voir figure 7)	Cette méthode de représentation est la plus commune. Ici, il s'agit de représenter les tweets sur une carte par des markers avec des formes et/ou des couleurs différentes.

1. <http://mapd.com>

<p>Heatmap (voir figure 8)</p>	<p>La densité des tweets peut être représentée par une <i>heatmap</i>. Une <i>heat map</i> [43](carte thermique, carte de chaleur) est une représentation graphique de données statistiques qui fait correspondre à l'intensité d'une grandeur variable une gamme de tons ou un nuancier de couleurs sur une matrice à deux dimensions (qui peut elle-même représenter une zone géographique) [48]. Ce procédé permet de donner à des données un aspect visuel plus facile à saisir qu'un tableau de chiffres [49].</p>
<p>Nuages de fumée (voir figure 9)</p>	<p>Une autre méthode de représentation de données sur une carte est la méthode <i>smoky clouds</i> s'inspirant des nuages de fumée [47].</p>
<p>Clusters²</p>	<p>Certaines application, représentent aussi les informations sous forme de clusters qui sont souvent graphiquement des cercles dessinés sur la carte ayant des surfaces, formes et couleurs différentes représentant des regroupements de données ayant une similarité commune.</p>

Figure 7 : Markers ou point map

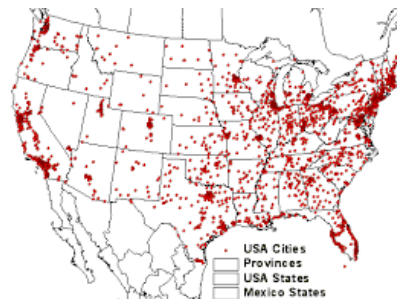


Figure 8 : Heatmap



Figure 9 : Nuages de fumée

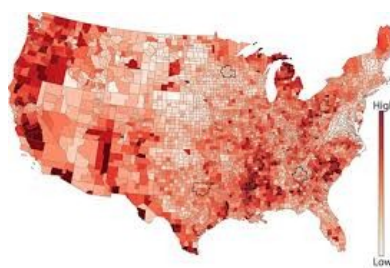


2. <http://onemilliontweetmap.com/>

Figure 10 : Clusters



Figure 11 : Choroplèthe



Outils de visualisation spatiale (bibliothèques et services de cartographie)

Solutions propriétaires

Outils	Descriptions	Logos
Google Maps API	De la branche propriétaire, Google Maps API est la plus populaire, étant utilisée par plus de 350000 sites web. Pour une faible utilisation du trafic, leurs services sont gratuits, mais au-dessus de 25 000 vues par jour, une taxe de 0,50 \$ est facturée pour chaque chargement de 1000 cartes supplémentaires.	
Bing Maps API³	Bing Maps est un service de cartographie web propriétaire. Il est édité par Microsoft.	
HERE Maps API⁴	Récemment HereWeGo est un service de cartographie développé par Nokia utilisé par Yahoo notamment.	

3. <https://www.bing.com/maps>

4. <https://wego.here.com/?map=20.94483,104.9093,10,normal>

Solutions Opensource

Outils	Description
OpenstreetMap ⁵	OpenStreetMap (OSM) est un projet qui a pour but de constituer une base de données géographiques libre du monde (permettant par exemple de créer des cartes sous licence libre), en utilisant le système GPS et d'autres données libres.
OpenLayers ⁶	OpenLayers est une bibliothèque de cartographie web open source pionnière. Contrairement à Google Maps API, OpenLayers est totalement gratuit et il est développé par la communauté open source.
GeoEXT ⁷	Combine OpenLayers et ExtJS (une autre bibliothèque pour créer des applications de cartographie web enrichie) afin d'offrir une interface utilisateur plus agréable.
GeoChart ⁸	Geochart est une carte avec deux modes : un mode région, qui crée une carte choroplète basée sur les valeurs de chaque région et un mode marqueur où chaque marqueur est représenté par des bulles mises à l'échelle selon la valeur du marqueur. Le rendu est basé sur SVG ou VML.
Kartograph ⁹	Un autre framework de mappage Javascript qui crée des cartes SVG sans utiliser de service de cartographie. Le mode projection des données affichées peut être choisie et le format des données utilisées dans la librairie peut être en <i>.shp</i> , <i>.kml</i> ou <i>.geojson</i> . Avec cette bibliothèque, les cartes peuvent être construites de manière très artistique.
Polymaps ¹⁰	Polymaps est un framework de cartographie léger qui utilise SVG pour des données vectorielles avec un rendu très attrayant. Il peut également utiliser des images des cartes web d'OpenStreetMap, CloudMade ou Bing.
Leaflet ¹¹	Leaflet est une autre API de cartographie open source. C'est une bibliothèque Javascript très légère qui peut être combinée avec des panneaux

5. <http://www.openstreetmap.org>

6. <https://openlayers.org/>

7. <http://www.geoext.org/>

8. <https://developers.google.com/chart/interactive/docs/gallery/geochart>

9. <http://kartograph.org/>

10. <http://polymaps.org/>

11. <http://leafletjs.com/>

	de cartes personnalisés créés avec CloudMade, ou MapBox pour produire des applications de carte très attrayantes. Les caractéristiques que cette bibliothèque offre sont assez limitées par rapport à OpenLayers, mais ils peuvent être étendus avec d'autres plug-ins, par exemple comme le réglage des étiquettes ou la création de cartes de chaleur ou des visualisations avancées. Leaflet prend également en charge les données au format <i>.geojson</i> .
--	-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Outils de navigation et de visualisation de la chronologie

L'objectif final de ce travail est de proposer un outil permettant de visualiser les informations dérivées des données de Twitter sur une carte dans l'espace et dans le temps. Les outils de visualisation d'informations sur une carte ayant déjà été vus, voyons à présent dans le tableau ci-dessous les solutions existantes permettant de prendre en compte l'aspect temporel.

Outils	Description
Envision.js ¹² (voir figure 12)	Envision.js est une bibliothèque qui peut être utilisée pour créer des visualisations de chronologie rapides et interactives en HTML5. Les données peuvent également être visualisées en temps réel afin que le graphique soit mis à jour chaque fois que de nouvelles données sont disponibles. Un avantage à propos d'Envision est que cela permet à l'utilisateur de sélectionner une période de temps spécifique qu'il souhaite examiner plus précisément.
Cubism.js ¹³ (voir figure 13)	Cubism.js est un plug-in basé sur D3 (https://d3js.org/) pour visualiser des séries chronologiques. Avec cet outil plusieurs séries chronologiques peuvent être visualisées, ce qui est intéressant pour la visualisation de l'évolution de sujet dans le temps par exemple. L'utilisateur ne peut pas interagir avec la frise chronologique pour choisir une période de temps spécifique.
TimelineJS ¹⁴ (voir figure 14)	TimelineJS est un outil open source qui peut être utilisé pour créer des frises chronologiques visuellement riches et interactifs. Il est orienté davantage vers la visualisation des événements plutôt que visualisation de données statistiques.

12. <http://www.humblesoftware.com/envision>

13. <https://square.github.io/cubism/>

14. <https://timeline.knightlab.com/>

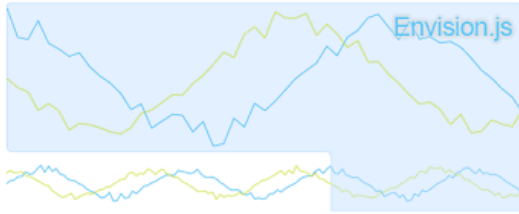


Figure 12 : EnvisionJS

Figure 13 : CubismJS

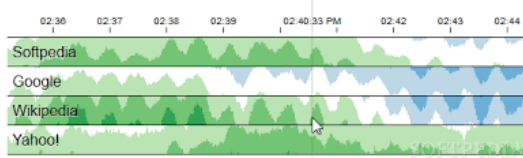




Figure 14 : TimelineJS





Outils de visualisation de diagramme

L'analyse par le biais de la visualisation des informations sur une carte, bien que plus rapide et plus intuitive à l'interprétation peut cependant être complétée par des diagrammes donnant des informations quantitatives plus complexes et plus approfondies. Le tableau ci-dessous présente quelques outils qui peuvent être utilisés pour construire de tels diagrammes.

Outils	Description	
Google Charts ¹⁵	Bibliothèque Javascript qui fournit des tableaux prêts à être utilisés et facilement intégrables dans les pages Web. Les données peuvent être chargées dans les tableaux sous forme de format <i>.json</i> .	
Highcharts ¹⁶	Une autre bibliothèque en HTML5/Javascript pour la visualisation. Le graphique et le design sont très interactifs et attrayants.	

15. <https://developers.google.com/chart/>

16. <https://www.highcharts.com/>

Graphaël ¹⁷	Bibliothèque fournissant entre autres diagrammes en secteurs, en bâtons, etc...	
D3JS ¹⁸	D3 est une bibliothèque de visualisation. D3 permet une grande flexibilité dans la construction de plusieurs types de visualisation comme dans la construction de délais, mais aussi comme une bibliothèque de cartographie car il peut interpréter les fichiers .geojson.	

PARTIE III

Exemples d'applications

Il existe de nombreuses applications exploitant les données de Twitter et représentant ces informations sur une carte. Le tableau ci-dessous énumère quelques unes d'entre elles :

Applications	Description
TwitterViz ¹⁹	TwitterViz une solution complète pour la visualisation et l'analyse des données spatio-temporelles de Twitter en combinaison avec l'analyse du graphique Twitter
TrendsMap ²⁰	TrendsMap est une application web qui permet de visualiser les <i>tweets</i> sur une carte par des zones de texte en les regroupant par sujets d'actualité. La taille de la zone de texte et sa couleur représentent l'importance du sujet qui sont calculées à partir du nombre de personnes qui ont émis des <i>tweets</i> à propos de ce sujet. En cliquant sur une des zones de texte, une liste des derniers messages postés à propos du sujet et les éventuels URL et images contenues dans les <i>tweets</i> sont affichées. Les <i>tweets</i> sont mis à jour

17. <http://jster.net/library/graphael>

18. <https://d3js.org/>

19. <https://www.dropbox.com/s/0vsrsod25m26ikm/demo.mp4?dl=0>

20. <http://trendsmap.com>

	<p>en temps réel et les zones de texte peuvent croître ou décroître selon que le sujet devienne plus ou moins populaire. De nouveaux sujets d'actualité peuvent apparaître. L'utilisateur a la possibilité de choisir une ville et l'application permet d'afficher les sujets d'actualité relatifs à cette zone.</p>
Tweetping ²¹	<p>Une autre application qui se concentre principalement sur la visualisation est Tweetping. L'application Web affiche des tweets en temps réel en tant que pixels lumineux sur une carte à fond sombre, ce qui fait qu'il ressemble à une image par satellite de la Terre. Il n'y a pas de possibilité de zoom avant, après avoir laissé fonctionner l'application pendant un certain temps, les tweets ont tendance à s'accumuler en clusters de zones éclairées et il n'est pas possible d'accéder individuellement aux tweets. Il fournit également un tableau de bord avec des statistiques sur le nombre de tweets repartis par continents, hashtags et mentions. L'application fonctionne en temps réel et ne permet pas de choisir une période de temps antérieur. La visualisation des tweets est l'objectif premier.</p>
Onemilliontweetmap ²²	<p>Onemilliontweetmap est une autre application de visualisation spatiale des données de Twitter. Cette carte affiche en temps réel les 1 000 000 derniers tweets. Il y a deux méthodes de visualisation, une première méthode permet d'afficher les tweets en les regroupant en clusters et une autre comme une carte de chaleur. Outre la visualisation, l'utilisateur peut interagir avec la carte et peut spécifier un mot-clé ou un hashtag pour filtrer les tweets. Les tweets affichés peuvent également être filtrés par les hashtags les plus populaires, qui sont dans ce cas récupérés par l'application. Une composante temporelle est également disponible afin que l'utilisateur puisse filtrer les tweets correspondant à différentes intervalles de temps jusqu'aux 6 dernières heures. La carte peut également être réinitialisée pour une meilleure visualisation de chaque tweet.</p>
Tweereal ²³	<p>Tweereal est une carte d'activités Twitter en</p>

21. <http://tweetping.net/>

22. <http://onemilliontweetmap.com/>

23. <http://tweereal.com/>

	temps réel qui affiche les tweets comme cercles colorés apparaissant / disparaissant. Les tweets apparaissent sur la carte à l'instant où ils sont créés, en temps réel. L'utilisateur peut contrôler la taille, l'opacité des cercles et la vitesse à laquelle ils disparaissent de la carte. Cependant, aucune information sur le contenu du tweet n'est affichée.
Tweetmap ²⁴	Tweetmap est un outil qui utilise une barre de temps (timeline) pour suivre les changements. Les tweets peuvent être filtrés par un mot-clé et une période de temps d'intérêt. Les tweets filtrés sont ensuite affichés comme une carte thermique ou à l'aide de points et leur évolution dans le temps est également affichée sur une échelle de temps en pourcentage du nombre total de tweets.
Twittinfo ²⁵	<i>Twittinfo</i> est une autre application qui utilise la perspective temporelle. C'est un système qui identifie et étiquette les pics d'événements, permet la visualisation du focus et du contexte des événements de longue durée et fournit une vue agrégée du sentiment de l'utilisateur. C'est une application semblable à <i>Tweetmap</i> . La technique de visualisation de la partie cartographique est simple, une carte Google est utilisée et en plus, les broches avec les tweets correspondants sont affichées. Cette application utilise une timeline pour montrer l'évolution des événements avec différentes granularités de temps : l'utilisateur peut choisir de voir le sujet d'intérêt dans les derniers jours ou dans les cinq dernières années.

Autres applications et plateformes

Il est impossible de dresser une liste exhaustive de toutes les applications existantes tant elles sont nombreuses. À titre informatif, en plus des applications déjà mentionnées, il existe également **MapD**²⁶, **Tweepsmap**²⁷, **socialbearing**²⁸, **Tweetpath**²⁹, **Tweereal**³⁰, qui sont soit des applications, soit des plateformes tout aussi bien intéressantes.

24. <http://bop.worldmap.harvard.edu/bop/#1/search?time=%5B2017-02-09T00:00:00%20TO%202017-05-09T00:00:00%5D&geo=%5B18.299877342757625,-73.70988906249995%20TO%2019.601888723271188,-71.06767714843747%5D>

25. <http://twittinfo.csail.mit.edu/>

26. <https://www.mapd.com/demos/>

27. <https://tweepsmap.com/>

28. <https://socialbearing.com/>

29. <http://www.tweetpaths.com/>

30. <http://tweereal.com>

Figure 15 et 16 : Mapd

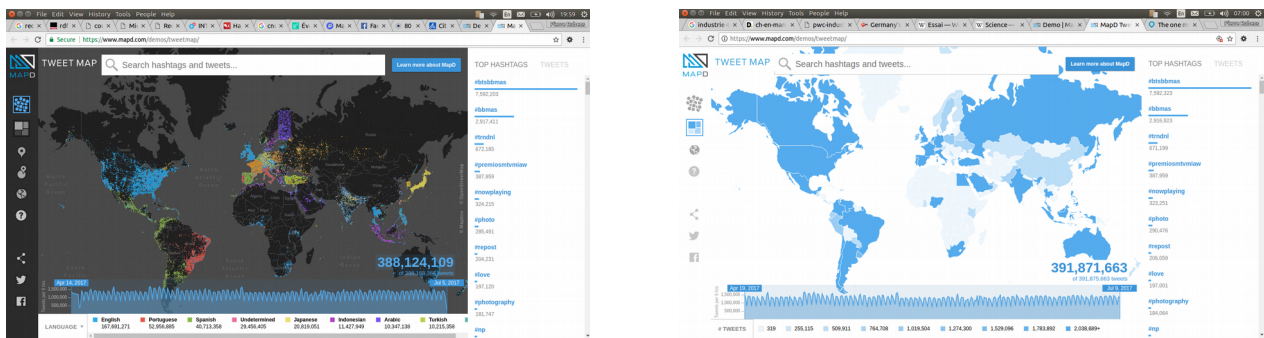
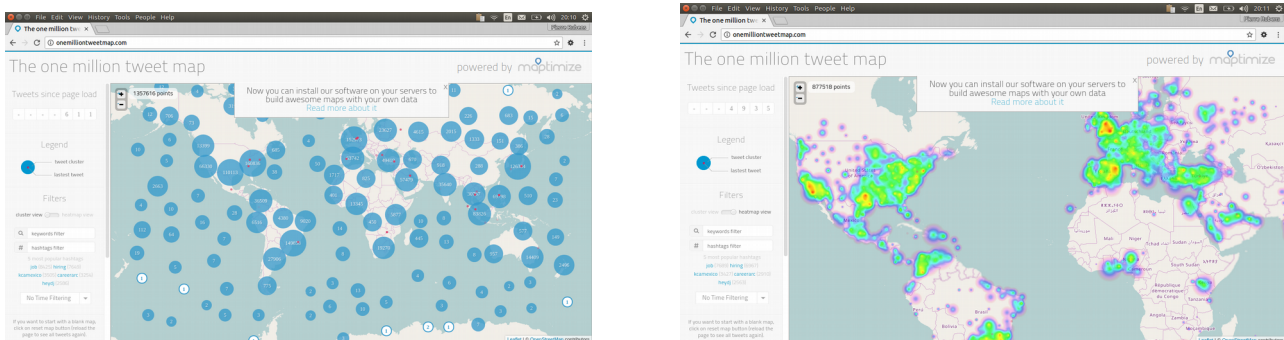


Figure 17 et 18 : Onemilliontweetmap



Pour ce travail de recherche bibliographique, nous avons dans la partie I du rapport présenté le concept de géolocalisation dans le cadre de Twitter et l'analyse de sentiment. Nous avons poursuivi ensuite dans la partie II en étudiant les informations pouvant être déduites des données de Twitter sur une ville et les méthodes, outils et techniques utilisées. Enfin la partie III a été l'occasion de répertorier quelques applications traitant de sujets similaires au nôtre.

Cette étape de recherche bibliographique et d'étude de l'état de l'art du domaine nous aura permis de prendre connaissance et d'avoir les notions nécessaires pour l'étape de proposition de solution qui s'ensuivra. Dans la prochaine section qui va suivre qui constitue justement cette étape de proposition de solution nous exposerons et justifierons nos choix concernant la solution proposée et la démarche d'implémentation envisagée.

SOLUTION PROPOSÉE

Dans la phase d'analyse de ce travail, nous avons tenté de comprendre le sujet qui nous a été attribué et de cerner les enjeux qui y sont associés. Ainsi nous avons vu que dans le contexte de la *Ville Intelligente*, les réseaux sociaux peuvent être utilisés comme des capteurs de données pour des applications. L'étude de l'état de l'art du domaine que nous avons effectuée par la suite, a permis de prendre connaissance des informations pouvant être déduites des données géolocalisées provenant de ces réseaux sociaux, de répertorier certaines méthodes, techniques et outils utilisés et d'évoquer certains exemples d'applications réalisées dans le domaine. Nous nous sommes attardés particulièrement sur l'analyse des sentiments (*sentiment analysis*) [4].

Après ces deux premières étapes de la partie théorique, nous exposons dans cette troisième partie la solution proposée. Nous y développons en détail notre approche, les outils que nous utiliserons pour l'implémentation et les raisons de nos choix. Nous présentons aussi les scénarios de test pour l'évaluation et la validation des résultats obtenus, les contraintes et difficultés auxquelles nous ferons face et la planification des tâches à réaliser.

Solution proposée

Après avoir fait une étude sur l'état de l'art des travaux du domaine, il nous incombe de choisir dans l'étape de solution proposée la ville (ou le quartier de cette ville) et d'identifier les informations pertinentes que nous voulons analyser. Suite à des expérimentations permettant de récupérer des tweets via l'API Streaming de Twitter sur des villes dont Port-au-Prince, Paris, New York et Hanoï, notre choix s'est porté vers cette dernière. Dans le cadre de ce projet, nous proposons d'analyser et visualiser sur une carte l'activité touristique de la ville de Hanoï, c'est-à-dire d'étudier et d'analyser le tourisme et de voir comment cela se reflète-t-il sur Twitter. En effet, la raison de ce choix est un constat sur la très grande affluence de touristes étrangers présents dans cette ville et la très grande intensité qui caractérise le phénomène. Analyser cet aspect nous paraît donc d'une très grande importance dans le contexte de la ville intelligente.

Notre analyse se fera selon deux aspects :

a) Détermination des lieux les plus fréquentés par la communauté touristique de Hanoï (POI), c'est-à-dire les endroits très prisés, très populaires de Hanoï où la communauté touristique aime se rendre.

b) Analyse de l'humeur des touristes (*sentiment analysis*) : il s'agit ici de faire l'analyse des sentiments des touristes en faisant le lien avec les lieux les plus fréquentes. Ceci permettra par exemple de connaître l'humeur des touristes en visitant ces lieux.

Processus

a) Acquisition des données

Récupération et filtrage des tweets provenant de la ville de Hanoï via l'API Streaming de Twitter. Les tweets récupérés sont donc explicitement géolocalisés.

b) Traitements (nettoyage, préparation et *sentiment analysis*)

Après la phase de récupération, les tweets dont la langue de rédaction ou la langue du profil de l'utilisateur est en vietnamien seront écartés. Nous supposons ici que les touristes étrangers ne

parlent pas vietnamiens et ne possèdent pas la nationalité vietnamienne également. Nous procéderons ensuite à la phase de classification afin de déduire l'humeur exprimée dans chaque tweet : les classes de sentiments (polarités) attendues étant *positive*(positif), *negative*(negatif) et *neutral*(neutre). Cependant l'analyse du sentiment donnant de meilleurs résultats avec des modèles élaborés avec des corpus en anglais il se pose un problème de langue avec le corpus de tweets récupérés : ceux-ci étant rédigés dans plusieurs langues différentes (suite à des tests que nous avons menés nous avons constaté que beaucoup de ces tweets sont rédigés dans les langues des pays asiatiques environnants). Le module que nous allons utiliser et que nous évoquons plus loin dans ce document permet en plus de faire la classification de sentiment, la traduction d'une langue vers une autre. Nous utiliserons donc cette fonctionnalité dans le cas où le nombre de tweets récupérés en anglais ne serait pas suffisant afin de faire la transposition d'un tweet rédigé dans une langue autre que l'anglais (par exemple le japonais ou l'allemand) dans la langue anglaise.

Ci-dessous les étapes nécessaires pour réaliser l'analyse de sentiment :

- **Tokenization** : séparer et récupérer les mots du texte original de chaque tweet.
- **Enlever les stopwords des tokens** : Les stopwords étant les mots utilisés couramment qui ne sont pas pertinents pour l'analyse de données textuelles tels que *I, am, you, are, etc* pour la langue anglaise.
- **Marquage POS(part of speech) des tokens** : sélectionner seulement les tokens significatifs tels que les adjectifs, les adverbes, etc...

Les tokens seront passés ensuite à un classifieur. La bibliothèque que nous utiliserons s'appuie sur le corpus *Movie Reviews* déjà labellisés afin de faire la classification.

c) Stockage des données dans une base de données adéquate

Après avoir récupéré les tweets et effectué les traitements nécessaires nous les sauvegarderons dans Neo4j qui est un système de gestion de base de données orienté graphes. Les tweets originaux seront stockés sous forme de nœuds et de relations entre ces nœuds. Il faut au préalable effectuer des opérations comme l'extraction des hashtags, des utilisateurs et des informations qui les caractérisent, créer les différentes relations entre ces entités, créer la structure adéquate dans la base de données pour recevoir ces différentes entités composant les tweets. Ci-dessous la représentation des tweets sous formes de nœuds et de relations dans la base de données :

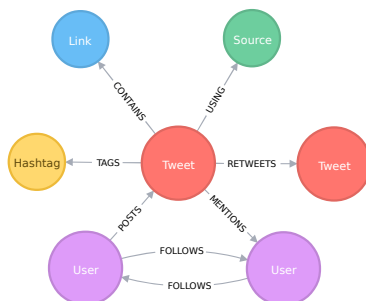


Figure 19 – Représentation des tweets dans la base de données

- Noeuds

User	(nœud représentant l'utilisateur)
Tweet	(nœud représentant le message - tweet)
Link	(les éventuels liens contenus dans le texte)
Source	(source)

- Relations

POSTS	(matérialise le fait qu'un utilisateur a poste un tweet)
MENTIONS	(matérialise le fait qu'un tweet mentionne un utilisateur)
TAGS	(matérialise le fait qu'un tweet comporte un hashtag)
REPLY_TO	(matérialise le fait qu'un tweet soit une réponse a un tweet d'un utilisateur)
RETWEET	(matérialise le fait qu'un utilisateur a retweet un tweet)
FOLLOWS	(matérialise le fait qu'un utilisateur s'est abonne a un autre utilisateur)
CONTAINS	(matérialise le fait qu'un tweet contienne un lien)
USING	(matérialise le fait qu'un tweet s'appuie sur une source)

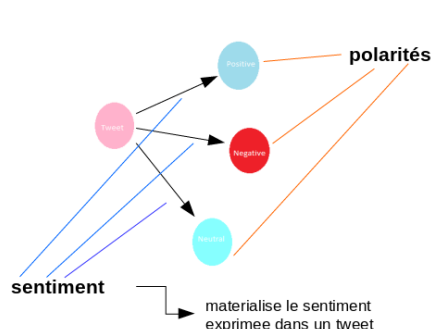


Figure 20 - Matérialisation de la polarité des tweets dans la base de données

d) Affichage dans le temps et dans l'espace

Nous avons choisi d'implémenter un outil basé sur les technologies web. Notre application utilisera un service de cartographie web afin de visualiser les tweets. Dans la phase de recherche bibliographique, nous avons évoqué plusieurs méthodes de représentation des données sur une carte. La méthode de représentation avec des marqueurs ou les tweets sont représentés individuellement sur la carte sera utilisée. L'important ici étant d'analyser la densité ou la concentration des tweets dans les quartiers, régions à différents endroits de la ville. Une représentation sous forme de *heatmap* sera éventuellement donnée également pour représenter la densité des tweets.

La méthode de représentation sous forme de marqueurs a l'avantage d'autoriser l'accès aux tweets individuellement. Mais ici les tweets appartenant à des classes d'une polarité différente seront affichés avec des couleurs différentes selon le sentiment exprimé. Ainsi, avec les deux types de représentation il sera possible de déduire le sentiment général dans la ville et l'humeur prédominante dans les régions bien spécifiques (quartier, places publiques, parcs, lieux touristiques) selon des périodes de temps choisies.

Notons que ces informations seront affichées non seulement dans l'espace mais également dans le temps où il sera possible de naviguer afin de faire des analyses en fonction de l'évolution temporelle.

Cependant, afin de réaliser des analyses plus poussées il est important d'accompagner ces informations visuelles avec des représentations sous forme de diagrammes donnant des informations quantitatives telles que des graphiques, des tableaux, des animations ou d'autres outils de visualisation. Des bibliothèques *javascript* seront mis a profit pour accomplir de telles tâches afin de faire dans les cas les plus simples des statistiques sur nos résultats.

Architecture générale de notre système

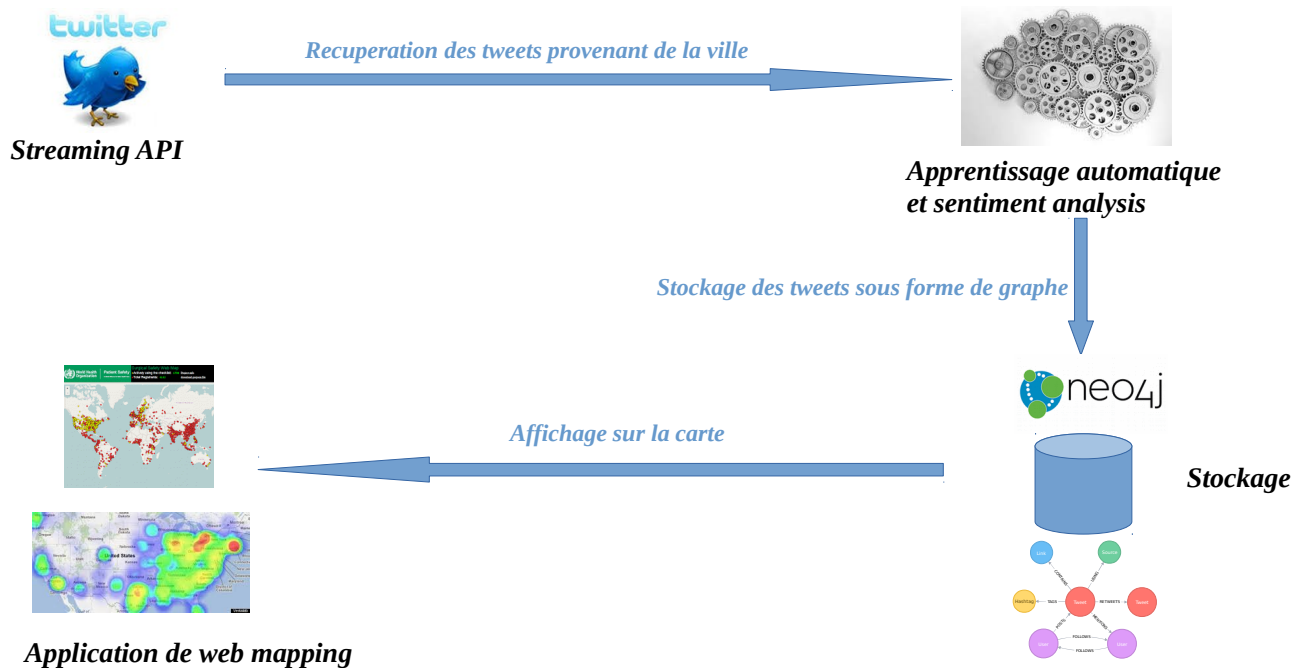


Figure 21 – Architecture du système

Outils à utiliser

✓ logiciels et librairies :

Pour réaliser et proposer l'outil espéré dans le cadre de ce travail, il est nécessaire de procéder aux quatre étapes que sont l'acquisition des données, le stockage, le traitement des données et l'affichage comme nous l'avons vu plus haut dans ce document. Les outils dont nous ferons usage dans notre étude sont donc :

Acquisition des données

- *tweepy*

Module python permettant d'interagir avec les API de Twitter.

Stockage des données

- *Neo4j database engine*

Système de gestion de base de données de type graphe

- *py2neo*

Module python servant d'interface entre le langage de programmation python et le système de gestion de base de données Neo4j.

Traitement des données et déduction des informations

- *python*

Langage de programmation.

- *time*

Module python permettant de manipuler le temps.

- *json*

Module python permettant de manipuler des fichiers en format *json*.

- *textblob*

Module python permettant de réaliser diverses manipulations sur les données textuelles.

- *re*

Module python permettant de manipuler et d'appliquer les expressions régulières.

Présentation et affichage

- *Django*

Framework web en python.

- *Mapbox [10]*

Service de cartographie et de manipulation de données spatiales.

- *Leaflet*

Service de cartographie web.

- *Bootstrap*

Framework *front-end* web permettant de gérer très facilement l'affichage et le design d'une application web.

- *Javascript*

Langage de programmation de script orienté web.

- *D3js*

Bibliothèque *javascript* permettant de faire la visualisation des données sur le web.

✓ algorithme :

Algorithme Naïves Bayes

L'analyse de l'humeur étant une opération de classification, plusieurs méthodes peuvent être utilisées afin d'accomplir la tâche. La solution et la méthode que nous proposons utilise le modèle

Naives Bayes qui est une méthode probabiliste de classification. Nous utiliserons le module *textblob* implémenté en python. Ce module dispose d'un corpus de tweets et d'un modèle basé sur l'algorithme *Naives Bayes* déjà entraîné. Les classes à prévoir sont :

- ◆ *Positive (positif)*
- ◆ *Negative (negatif)*
- ◆ *Neutral (neutre)*



Figure 22 : polarités attendues

Expérimentations, évaluation et validation

Afin d'évaluer les résultats que nous obtiendrons, nous définissons les critères suivants :

- Analyse de la densité des tweets émis dans les lieux répertoriés dans le temps.
- Analyse de la périodicité des événements, des fluctuations des sentiments dans les espaces répertoriés dans le temps. Il s'agira essentiellement ici de détecter des pics à des périodes et des intervalles de temps réguliers. Cela permettra de déduire la dynamique de la ville de Hanoï du point de vue de l'humeur exprimée par les touristes.
- Enquête faite auprès d'échantillons de touristes dans les POI identifiés afin de s'assurer que les déductions faites par le système automatique sont cohérentes et en adéquation avec le monde réelle.

Ces approches d'analyse permettront de détecter les lieux ou zones d'intérêt pour les touristes et de déduire en même temps l'humeur générale qui prédomine dans ces lieux à travers le temps.

Risques encourus, difficultés et éventuelles contraintes

En faisant choix d'une ville et d'un thème particulier à analyser à propos de cette ville, nous exposons à des difficultés et problèmes inhérents aux données générées. Le cas de Hanoï n'échappe pas à cette règle. Nous avons déjà évoqué les divergences et la diversité linguistiques auxquelles nous devons faire face dans le cas de la ville de Hanoï et auxquelles il convient de trouver une méthode de traitement adéquate pour les données afin d'avoir une bonne représentation de celles-ci et des résultats qui soient le plus en adéquation possible avec la réalité. En ce sens, il serait tout à fait judicieux et utile comme nous l'ont suggéré les membres du jury lors de la présentation de la partie théorique, dans un premier temps d'essayer d'explorer un peu les données

récupérées afin d'effectuer des statistiques sur une période assez longue dans le passé (en utilisant potentiellement l'API REST par exemple) et dans la mesure du possible afin de voir et de dégager les langues prédominantes et les plus importantes dans notre contexte.

Outre ce problème de langues nous pouvons aussi mentionner comme obstacle la possibilité que nous puissions avoir une quantité de données insuffisantes à analyser. En effet, comme nous avons pris le soin de le spécifier au début les données sont nécessaires dans le cadre de développement d'applications pour la Ville Intelligente. Dans le cas où ce scénario se présenterait, nous envisageons aussi de prendre en compte éventuellement d'autres réseaux sociaux tels que Facebook si toutefois les statistiques menées ne seraient pas concluantes et que nous n'aurions pas assez de données à exploiter. Conformément aux directives qui nous ont été données dans le sujet il est possible de prendre en compte éventuellement les autres réseaux sociaux (l'accent ayant été mis cependant sur Twitter qui est obligatoire).

Une autre suggestion faite par les membres du jury concerne le *topic extraction* : faire l'analyse de l'humeur dans les lieux fréquentés identifiés est bien, cependant il serait tout aussi intéressant après avoir effectué le *sentiment analysis* de faire aussi éventuellement un *topic extraction* dans ces lieux afin de déterminer non seulement les raisons pour lesquelles des personnes se regroupent dans de pareils lieux et aussi d'essayer de donner des explications plausibles ou d'émettre des hypothèses sur le motif de la présence de tels sentiments dans ces lieux. Cependant d'ores et déjà nous savons les difficultés existantes dans le processus de classification des tweets afin d'en extraire des sujets (*topic extraction*) telles que la longueur très réduite des tweet (ce processus donnant de meilleurs résultats à partir de corpus comportant des documents de grande taille comme des textes, etc.), le caractère informel et non standard au niveau de l'écriture des messages publiés (mots abrégés, sigles, acronymes, etc ...), etc...

Arriver à contourner les différents risques et problèmes sus-mentionnés est importante dans le travail que nous faisons et prendre en compte les différentes suggestions apportées au projet constituent des alternatives et des perspectives intéressantes auxquelles il convient de penser.

Planification des tâches

Travaux	Descriptions	Durées estimées
Acquisition des données et traitements	<ul style="list-style-type: none"> - Familiarisation avec les outils de développement. - Récupération des tweets et statistiques sur les langues. - Test sur les traitements à effectuer. 	2 semaines
Conception et implémentation	<ul style="list-style-type: none"> - Conception du fonctionnement global de l'outil à réaliser. - Implémentation et intégration des différents composants logiciels. 	7 semaines
Expérimentations et évaluation	<ul style="list-style-type: none"> - Amélioration du code - Évaluation des résultats et interprétation. 	2 semaines

Rédaction du rapport final	<ul style="list-style-type: none"> - Rédaction rapport final - Correction du rapport final et présentation 	3 semaines
----------------------------	------------------------------------------------------------------------------------------------------------------------------------	------------

Conclusion

Dans ce rapport nous avons présenté la totalité de ce que nous avons pu entreprendre pour la partie théorique de notre sujet, à savoir comment nous avons compris le sujet, l'étude de l'état de l'art et finalement la solution que nous proposons et le plan que nous suivrons pour la phase pratique de notre étude, les outils que nous utiliserons, les raisons de nos choix, les critères d'évaluation de nos résultats et la planification des tâches à réaliser. Ainsi la partie théorique se termine ici. La partie pratique de ce travail qui s'ensuivra, sera l'occasion de mettre en œuvre la solution envisagée afin de réaliser l'outil escompté.

Références bibliographiques

1. Pavlos Paraskevopoulos and Themis Palpanas. 2015. Fine-Grained Geolocalisation of Non-Geotagged Tweets. In Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015 (ASONAM '15), Jian Pei, Fabrizio Silvestri, and Jie Tang (Eds.). ACM, New York, NY, USA, 105G 112. DOI: <https://dx.doi.org/10.1145/2808797.2808869>
2. Justin Cranshaw, Eran Toch, Jason Hong, Aniket Kittur, and Norman Sadeh. 2010. Bridging the gap between physical location and online social networks. In Proceedings of the 12th ACM international conference on Ubiquitous computing (UbiComp '10). ACM, New York, NY, USA, 119G128. DOI=<http://dx.doi.org/10.1145/1864349.1864380>
3. Urbano França, Hiroki Sayama, Colin McSwiggen, Roozbeh Daneshvar, Yaneer BarG Yam. 2016. Visualizing the "heartbeat" of a city with tweets. Complexity 21(6): 280- 287
4. Karla Z. Bertrand, Maya Bialik, Kawandee Virdee, Andreas Gros, Yaneer BarG Yam.2013. Sentiment in New York City: A High Resolution Spatial and Temporal View. CoRR abs/1308.5010
5. Liu, Q., Ma, H., Chen, E., & Xiong, H. (2013). A survey of context-aware mobile recommendations. International Journal of Information Technology & Decision Making, 12(01), 139G172.
6. Uan Yuan, Gao Cong, Kaiqi Zhao, Zongyang Ma, and Aixin Sun. 2015. Who, Where,When, and What: A Nonparametric Bayesian Approach to Context-aware Recommendation and Search for Twitter Users. ACM Trans. Inf. Syst. 33, 1, Article 2 (February 2015), 33 pages. DOI: <http://dx.doi.org/10.1145/2699667>
7. Twitter, <https://twitter.com>
8. Neo4j, <https://neo4j.com>
9. OpenStreetMap(OSM), www.openstreetmap.org
10. <https://www.mapbox.com/>
11. Using Social Media Data to Understand Cities, Dan Tasse, Jason I. Hong
12. Pavlos Paraskevopoulos and Themis Palpanas.2015. Fine-Grained Geolocalisation of Non-Geotagged Tweets. In Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015 (ASONAM'15), Jian Pei, Fabrizio Silvestri, and Jie Tang (Eds.). ACM, New York, NY, USA, 105G 112. DOI: <http://dx.doi.org/10.1145/2808797.2808869>
13. Song, Z and Xia, J. 2016. Spatial and Temporal Sentiment Analysis of Twitter data, chapter 16. In: Capineri, C, Haklay, M, Huang, H, Antoniou, V, Kettunen, J, Ostermann, F and Purves, R. (eds.) European Handbook of Crowdsourced Geographic Information, Pp. 205–221. London: Ubiquity Press. DOI: <http://dx.doi.org/10.5334/bax.p>. License: CC-BY 4.0.
14. Justin Cranshaw, Eran Toch, Jason Hong, Aniket Kittur, and Norman Sadeh. 2010. Bridging the gap between physical location and online social networks. In Proceedings of the 12th ACM international conference on Ubiquitous computing (UbiComp '10). ACM, New York, NY, USA, 119G128. DOI=<http://dx.doi.org/10.1145/1864349.1864380>
15. Urbano França, Hiroki Sayama, Colin McSwiggen, Roozbeh Daneshvar, Yaneer BarG Yam. 2016. Visualizing the "heartbeat" of a city with tweets. Complexity 21(6): 280- 287
16. Karla Z. Bertrand, Maya Bialik, Kawandee Virdee, Andreas Gros, Yaneer Bar Yam.2013. Sentiment in New York City: A High Resolution Spatial and Temporal View. CoRR abs/1308.5010
17. Liu, Q., Ma, H., Chen, E., & Xiong, H. (2013). A survey of context-aware mobile recommendations. International Journal of Information Technology & Decision Making, 12(01), 139G172.

18. Uan Yuan, Gao Cong, Kaiqi Zhao, Zongyang Ma, and Aixin Sun. 2015. Who, Where, When, and What: A Nonparametric Bayesian Approach to Context-aware Recommendation and Search for Twitter Users. *ACM Trans. Inf. Syst.* 33, 1, Article 2 (February 2015), 33 pages. DOI: <http://dx.doi.org/10.1145/2699667>
 19. Exploiting Geographical Neighborhood Characteristics for Location Recommendation.pdf
 20. Bing Liu. *Sentiment Analysis and Opinion Mining*, Morgan & Claypool Publishers, May 2012.
 21. *Natural Language Processing with Python* Steven Bird, Ewan Klein, and Edward Loper
 22. *Fundamentals of Sentiment Analysis: Concepts and Methodology* A.B. Pawar, M.A. Jawale and D.N. Kyatnavar
 23. CH A PT ER 17, Social Networks VGI: Twitter Sentiment Analysis of Social Hotspots Dario Stojanovski*, Ivan Chorbev, Ivica Dimitrovski and Gjorgji Madjarov
 24. Svetlana Kiritchenko, Xiaodan Zhu, Saif M. Mohammad, *Sentiment Analysis of Short Informal Texts*
 25. Eric Baucom, Azade Sanjari, Xiaozhong Liu, Miao Chen, *Mirroring the Real World in Social Media: Twitter, Geolocation, and Sentiment Analysis*
 26. Vincenza Carchiolo, Alessandro Longheu (B) , and Michele Malgeri, *Using Twitter Data and Sentiment Analysis to Study Diseases Dynamics*
 27. Narendran Thangarajan, Nella Green, Amarnath Gupta, Susan Little, Nadir Weibel, *Analyzing Social Media to Characterize Local HIV At-risk Populations*
 28. Fahad Bin Muhaya, Swarup Chandra, Latifur Khan, *Estimating Twitter User Location Using Social Interactions – A Content Based Approach*
 29. Laura Di Rocco, Michela Bertolotto, Giovanna Guerrini, Barbara Catania, Tiziano Cosso, *Microblogs Exploiting Crowdsourced Gazetteers and Social Interactions*
 30. [H.-w. Chang, D. Lee, M. Eltaher, and J. Lee, “@ phillies tweeting from philly? predicting twitter user locations with spatial word usage,” in *ASONAM*, 2012.
 31. S. Kinsella, V. Murdock, and N. O’Hare, “I’m eating a sandwichnin glasgow: modeling locations with tweets,” in *SMUC*, 2011.
 32. *Towards a framework for automatic geographic feature extraction from Twitter* Enrico Steiger, Johannes Lauer, Timothy Ellersiek, Alexander Zipf
 33. Fatma Amin Elsafoury, *Monitoring Urban Traffic Status Using Twitter Messages*
 34. T. Sakaki, M. Okazaki, and Y. Matsuo, “Earthquake shakes twitter users: real-time event detection by social sensors,” in *WWW*, 2010.
 35. A. Crooks, A. Croitoru, A. Stefanidis, and J. Radzikowski, “#earthquake: Twitter as a distributed sensor system,” *Transactions in GIS*, vol. 17, no. 1, pp. 124–147, 2013.
 36. M. Balduini, E. Della Valle, D. Dell’Aglia, M. Tsytsarau, T. Palpanas, and C. Confalonieri, “Social listening of city scale events using the streaming linked data framework,” in *ISWC*, 2013.
 37. K. Leetaru, S. Wang, G. Cao, A. Padmanabhan, and E. Shook, “Mapping the global twitter heartbeat: The geography of twitter,” *First Monday*, vol. 18, no. 5, 2013.
- 11-
38. V. Murdock, “Your mileage may vary: on the limits of social media,” *SIGSPATIAL Special*, 2011.
 39. B. Han, P. Cook, and T. Baldwin, “Text-based twitter usergeolocation prediction,” *JAIR*, 2014.

40. H.-w. Chang, D. Lee, M. Eltaher, and J. Lee, “@ phillies tweeting from philly? predicting twitter user locations with spatial word usage,” in ASONAM, 2012.
41. S. Kinsella, V. Murdock, and N. O’Hare, “I’m eating a sandwich in glasgow: modeling locations with tweets,” in SMUC, 2011.
42. Horn, C., 2010. Analysis and Classification of Twitter messages. Graz: Graz University of Technology.
43. Christodoulos Efstathiades, Helias Antoniou, Dimitrios Skoutas, Yannis Vassiliou. TwitterViz : Visualizing and Exploring the Twittersphere
44. Pulkit Goyal, Sapan Diwakar . Data Mining and Analysis on Twitter
45. Shamanth Kumar, Fred Morstatter, Huan Liu, Twitter Data Analytics
46. Blei, D.M., Ng, A.Y. & Jordan, M.I., 2003. Latent Dirichlet Allocation. Journal of Machine Learning Research , (3), pp.993-1022.
47. Twitter Libraries. <https://dev.twitter.com/docs/twitter-libraries>.
48. Using the Twitter Search API. <https://dev.twitter.com/docs/using-search>.
49. Ian Robinson, Jim Webber & Emil Eifrem. Graph Databases, NEW OPPORTUNITIES FOR CONNECTED DATA.
50. <http://www.linguee.fr/francais-anglais/search?source=auto&query=%22heat+map%22>
51. <http://searchbusinessanalytics.techtarget.com/definition/heat-map>
52. <http://www.cio.com/article/2369317/big-data/125155-8-Cool-Heat-Maps-That-Help-You-Visualize-Big-Data.html>
53. <https://www.uq.edu.au/news/article/2000/10/new-mapping-techniques-see-through-clouds-and-smokescreens>