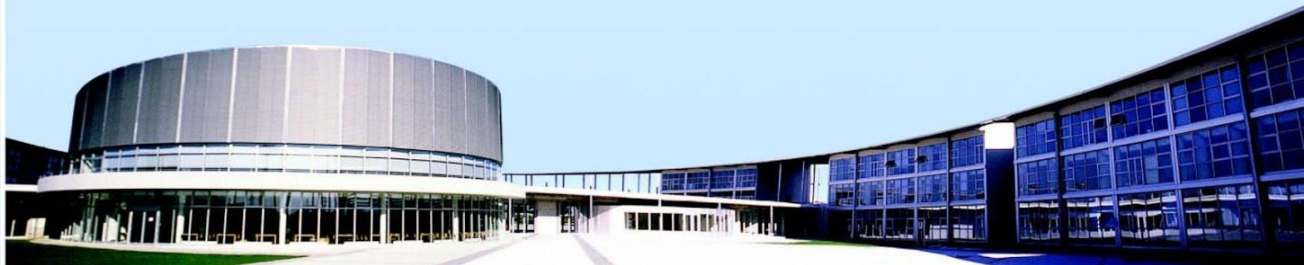
11/11/2020

Cahier des charges : Extraction de données Twitter

Version 1.0

Andromeda

Evann Bonnaventure
Gabriel Duciel
Clément Le Galeze
Nao Nicolas



www.utt.fr
Université de technologie de Troyes • 12 rue Marie Curie • BP 2060 • 10010 Troyes cedex

Sommaire

Préface

Ce cahier de charge a été rédigé par les équipes d'Andromeda, cabinet de conseil en informatique. Il est destiné aux équipes de Synevent, entreprise d'organisation en événement qui nous ont demandé un outil d'analyse concernant leur communication à travers le réseau social Twitter.

L'entièreté de ce cahier des charges est à l'usage unique des équipes de Synevent. Il ne peut en aucun cas être partagé à des tiers partis sans l'accord des organisations ici présentées.

Le contenu de ce document est sujet à modification selon les exigences de Synevent. La version ici présentée est la : **1.0.**

Introduction

La société d'événementiel Synevent organise des salons, forums et conférences pour le compte de ses clients. Celle-ci utilise majoritairement Twitter dans le cadre de sa communication : le client et la société choisissent conjointement un hashtag à utiliser pour communiquer sur l'événement.

Dans ce cadre, Synevent souhaiterait disposer d'un outil en ligne de commande permettant de juger de l'efficacité de sa communication. Cet outil devra permettre une fine analyse des tweets concernant un événement précis, de leur efficacité et de leur portée, à travers une série de spécifications fonctionnelles bien précises, qui seront spécifiées dans le présent cahier des charges.

La finalité de cette analyse sera la rédaction d'un rapport d'impact social media, qui sera assurée par la société Synevent. En outre, la société fournira elle-même la liste des tweets concernant un hashtag précis via une API qui fournira une liste de tweets au format CSV.

L'outil produit devra donc être capable entre autres d'extraire de ces données différents graphiques exploitables afin de faciliter le travail des rédacteurs du rapport. Il devra aussi, dans cet objectif, fournir aux rédacteurs des tweets à inclure dans leur rapport sous une forme exploitable, c'est-à-dire un format textuel raccourci, structuré et facilement lisible, comme par exemple [ID, URL, Auteur, Présentation de l'auteur, Date, Texte du tweet, nombre de retweet, hashtags associés].

Spécification générale

Les chefs de projet de Synevent ont eux-même précisé des éléments qu'ils attendaient de cette application. Nous avons par ailleurs ajouté quelques exigences qui offriraient à Synevent des fonctionnalités étendues, en lien avec leur utilisation actuelle du produit.

1. Exigences fonctionnelles

Au total, 10 exigences fonctionnelles ont été mises en place :

SPEC 1
Extraire le nombre de tweets sur un hashtag pour une période donnée par journée.

SPEC 2
Extraire le top 10 des tweets comportant un hashtag ayant été le plus retweeté.

SPEC 3
Extraire le Top 10 des auteurs de tweets avec le plus d'informations à leur sujet.

SPEC 4
Extraire la liste des hashtags associés à un hashtag de référence.

SPEC 5
Corréler la présence de mots clés dans les tweets en fonction du hashtag. L'utilisateur pourra entrer un hashtag, et recevoir en sortir les mots clés les plus importants.

SPEC 6
Visualiser la proportion de tweets par pays/région, selon le choix de l'utilisateur. Il pourra donc choisir un pays, ou le monde entier.

SPEC 7
Corréler le succès d'un tweet avec la présence d'un média en fonction du hashtag.

SPEC 8
Extraire des listes de tweets selon différents critères de recherche. L'utilisateur pourra spécifier un ou plusieurs critères de recherche, portant sur différents éléments du tweet : contenu, auteur, hashtags, nombre de retweets ...

SPEC 9
Afficher un rapport regroupant les résultats des fonctionnalités du logiciel que l'utilisateur aura sélectionnées.

SPEC 10
Représenter les données sous forme de graphiques.

2. Exigences non fonctionnelles

Les exigences non fonctionnelles dans notre cas ne seront que des formalités déjà attendus du logiciel :

SPEC_NF_1
L'application cliente doit pouvoir fonctionner de façon fluide sur du matériel ancien et diversifié

SPEC_NF_2
Le code doit être proprement documenté pour pouvoir être réutilisé par une autre équipe

Spécification détaillée

1. Exigences détaillés

Identifiant	SPEC 1
Titre	Extraire le nombre de tweets sur un hashtag pour une période donnée par journée.
Objectifs	L'utilisateur aimerait connaître le nombre de tweets qui ont été publiés sur un hashtag par rapport à une période choisie sur une journée.
Précondition(s)	Des tweets doivent exister, en format lisible par le logiciel.
Postcondition(s)	Une fonctionnalité a été appelée
Entrées	Le hashtag cherché, la période de recherche, le jour de la recherche.
Données	La date et le hashtag des tweet seront utilisés
Traitements	L'utilisateur rentre sur le logiciel le hashtag cherché, l'heure de début et de fin de la recherche et le jour de la recherche.
Sorties	Le nombre de tweets qui ont été publiés selon la période sélectionnée est affiché.
Gestion des erreurs	Message d'erreur si l'utilisateur rentre un hashtag qui n'existe pas.

Identifiant	SPEC 2
Titre	Extraire le top 10 des tweets comportant un hashtag ayant été le plus retweeté.
Objectifs	L'utilisateur aimerait voir les 10 tweets possédant un hashtag qui ont été le plus retweeté.
Précondition(s)	Des tweets doivent exister en format lisible par le logiciel.
Postcondition(s)	Une fonctionnalité a été appelée
Entrées	/
Données	Le contenu et le hashtag des tweets seront utilisés
Traitements	L'utilisateur sélectionne sur le logiciel l'option pour afficher le top 10 des tweets comportant un hashtag ayant été le plus retweeté.
Sorties	Les 10 tweets sont affichés sur le logiciel et l'utilisateur peut les parcourir.
Gestion des erreurs	/

Identifiant	SPEC 3
Titre	Extraire le Top 10 des auteurs de tweets avec le plus d'informations à leur sujet.
Objectifs	L'application doit permettre à son utilisateur de rapidement identifier la liste des 10 auteurs de tweets les plus influents avec lesquels la marque travaille.
Précondition(s)	Des tweets doivent exister en format lisible par le logiciel. Les informations sur les autours des tweets en question doivent avoir été récupérées également.
Postcondition(s)	Une fonctionnalité qui ne peut pas être représentée par un graphique a été appelée.
Entrées	L'événement ou Hashtag que l'utilisateur souhaite analyser.
Données	Attributs parmi la liste suivante : <hashtag> ; <user_name>; <user_screen_name> ; <user_description> ; <user_favourites_count> ; <user_followers_count> ; <user_friends_count> ; <user_listed_ount> ; <user_statuses_count> ; <user_verified>
Traitements	L'utilisateur sélectionne sur l'interface l'option pour extraire le Top 10 des auteurs de tweets. Il rentre ensuite des options supplémentaires s'il le souhaite : Évènement en question, ou un hashtag. Le logiciel parcourt son jeu de données.
Sorties	Un tableau comportant ces 10 utilisateurs et les informations correspondantes doit être proposé. Exemple d'informations : certification de l'utilisateur, localisation, présence d'une photo de profil, nombre d'amis, taille de la description.
Gestion des erreurs	Si l'utilisateur rentre des options invalides, la fonctionnalité doit le détecter et ne pas lancer la recherche de données.

Identifiant	SPEC 4
Titre	Extraire la liste des hashtags associés à un hashtag de référence
Objectifs	On souhaite connaître les sujets fréquemment associés à un sujet de référence précis, via les hashtags présents dans les tweets visés.
Précondition(s)	Avoir une liste de tweets existants contenant le hashtag concerné
Postcondition(s)	Une fonctionnalité a été appelée.
Entrées	Une liste de tweets fournie, et le hashtag de référence choisi.
Données	Les hashtags des tweets seront utilisés.
Traitements	La liste de tweet fournie est parcourue. On y recherche tous les tweets contenant le hashtag de référence saisi. Dans ces tweets, on recherche la présence d'autres hashtags et, le cas échéant, on ajoute les hashtags trouvés dans une liste s'ils n'y sont pas déjà.
Sorties	La liste des hashtag fréquemment associés au premier est fournie à l'utilisateur
Gestion des erreurs	Le hashtag saisi n'est présent nulle part/n'existe pas, on propose alors à l'utilisateur d'en saisir un autre

Identifiant	SPEC 5
Titre	Corrélation de mots clés avec un hashtag
Objectifs	L'utilisateur aimerait connaître un certain nombre de mots clés associés à un hashtag
Précondition(s)	Des tweets doivent exister, en format lisible par le logiciel.
Postcondition(s)	Une fonctionnalité a été appelée.
Entrées	Le hashtag cherché , le nombre de mots maximum (optionnel)
Données	Uniquement le texte du tweet sera utilisé
Traitements	L'utilisateur rentre le hashtag cherché, et optionnellement le nombre max de mots retournés. Si le nombre de l'ensemble des tweets comportant le hashtag recherché sont parcourus un à un.
Sorties	Une liste des mots les plus importants, avec leur fréquence d'apparition dans les tweets et le nombre d'apparition total, est affichée à l'utilisateur sous format CSV.
Gestion des erreurs	L'utilisateur doit être informé en l'absence de l'hashtag mis en entrée. Un manque de mots clés (pas de mots qui se démarquent) doit retourner que le hashtag est neutre.

Identifiant	SPEC 6
Titre	Visualiser la proportion de tweets par pays/région
Objectifs	L'utilisateur aimerait connaître la proportion des tweets envoyés selon le lieu d'envoi (pays ou région).
Précondition(s)	Des tweets doivent exister, en format lisible par le logiciel.
Postcondition(s)	Une fonctionnalité a été appelée.
Entrées	Le choix d'échelle entre pays et région, le pays souhaité dans le cas des régions
Données	Le lieu d'envoi et le contenu seront utilisés
Traitements	L'utilisateur fournit un choix d'options correspondant aux entrées précédentes. L'ensemble des tweets sont ensuite parcourus un à un.
Sorties	Un tableau des pays et du nombre de tweets.
Gestion des erreurs	Les pays n'ayant aucun tweet ne seront pas mentionnés (dans la carte, on choisira une couleur neutre).

Identifiant	SPEC 7
Titre	Corréler le succès d'un tweet avec la présence d'un média en fonction du hashtag.
Objectifs	On souhaite connaître la corrélation entre succès d'un groupe de tweets et présence ou non d'un média dans ceux-ci, en fonction de leur hashtag, ceci afin de savoir si certains sujets suscitent plus de réactions s'ils sont accompagnés d'un média.
Précondition(s)	Avoir des tweets existants contenant le hashtag concerné.
Postcondition(s)	Une fonctionnalité a été appelée.
Entrées	Une liste de tweets fournie, et le hashtag de référence choisi.
Données	Le hashtag, le(s) média(s) et le nombre de réactions (likes et retweet) du tweet sont nécessaires.
Traitements	La liste de tweets fournie est parcourue. On en ressort tous les tweets concernant un hashtag, et pour chacun de ces tweets, on regarde si oui ou non ils contiennent un média, et le nombre de réactions au tweet.
Sorties	Un tableau indiquant le succès d'un tweet en fonction du hashtag et de la présence de média.
Gestion des erreurs	Le hashtag saisi n'est présent nulle part/n'existe pas, on propose alors à l'utilisateur d'en saisir un autre.

Identifiant	SPEC 8
Titre	Extraire des listes de tweets selon différents critères de recherche.
Objectifs	L'utilisateur aimerait recevoir une liste de tweets selon certains critères, en choisissant quelles informations retenir sur le tweet.
Précondition(s)	Des tweets doivent exister, en format lisible par le logiciel.
Postcondition(s)	Une fonctionnalité qui ne peut pas être représentée par un graphique a été appelée.
Entrées	Liste de critères de l'utilisateur, données des critères variables, informations retenus des tweets.
Données	Les données vont dépendre des critères choisis par l'utilisateur.
Traitements	L'utilisateur fournit une liste de critères possibles. Certains critères peuvent être variables, c'est-à-dire exigeant une entrée par l'utilisateur (nombre de retweets, nombre de mots, …). Il choisit ensuite quelles informations il décide de garder sur ces tweets (seulement l'auteur et le contenu, seulement la date, …). L'ensemble des tweets sont enfin parcourus un à un.
Sorties	L'utilisateur reçoit une liste de tweets sous forme d'un tableau Excel.
Gestion des erreurs	L'utilisateur est informé de l'impossibilité de sa requête si aucun tweet ne répond à ses critères. L'utilisateur est alerté pendant l'exécution si le nombre de tweets récupéré dépasse un certain seuil (500 par exemple). Il peut alors choisir d'abandonner la requête, uniquement garder ces premiers tweets trouvés, ou terminer la requête.

Identifiant	SPEC 9
Titre	Afficher un rapport regroupant les résultats des fonctionnalités du logiciel que l'utilisateur aura sélectionné.
Objectifs	L'utilisateur aimerait avoir une vue d'ensemble sur le résultat de plusieurs fonctionnalités existantes.
Précondition(s)	Des tweets doivent exister, en format lisible par le logiciel.
Postcondition(s)	Un rapport est affiché
Entrées	Les fonctionnalités que l'utilisateur veut afficher, le choix du format graphique ou non pour chaque fonctionnalité choisie.
Données	Les données vont être celles des fonctionnalités choisies
Traitements	L'utilisateur peut choisir d'afficher ensemble les différentes fonctionnalités. Lorsque l'utilisateur choisi une fonctionnalité, il doit remplir les entrées de celle-ci. Il décide si il veut l'afficher en graphique ou non.
Sorties	Un rapport avec l'ensemble des fonctionnalités que l'utilisateur aura sélectionnées.
Gestion des erreurs	Les gestions d'erreurs des fonctionnalités que l'utilisateur aura choisies vont s'additionner.

Identifiant	SPEC 10
Titre	Représenter les données sous forme de graphique
Objectifs	Pour chacune des fonctionnalités ci-dessus présentées, il doit être proposé à l'auteur d'en extraire des graphiques.
Précondition(s)	Une fonctionnalité a été appelée.
Postcondition(s)	Un graphique a été généré.
Entrées	/
Données	Les données dépendent trop de la fonctionnalité en question pour être décrite ici.
Traitements	Lorsqu'une fonctionnalité permet la visualisation à travers un graphique, elle doit proposer une à l'utilisateur d'en profiter. Le logiciel charge alors les données sous forme de graphiques, et les affiche à l'utilisateur.
Sorties	L'utilisateur reçoit une visualisation graphique, qu'il peut choisir d'exporter au non sous le format JPEG
Gestion des erreurs	La fonctionnalité ne doit pas proposer cette option pour une fonctionnalité qui ne le permet pas.

2. Format de données

Cette section du cahier des charges a pour objectif de détailler le format de données des tweets utilisés. Cette description, au format ABNF, permettra de s'assurer la parfaite maîtrise du dit format, ce dernier formant une section critique de notre solution.

1. Tweet Brut

<Tweet> = [<coordinates>] ",," [<created_at>] ",," [<hashtag>] ", [<media>] ",," [<urls>] ",," [<favorite_count>] ",," [<id>] ",," [<in_reply_to_screen_name>] ",," [<in_reply_to_status_id>] ",," [<in_reply_to_user_id>] ",," [<lang>] ",," [<place>] , [<possibly_sensitive>] ",," <retweet_count> ",," <reweet_id> ",," [<retweet_screen_name>] ",,"[<source>]",,"<text>",,"<tweet_url>",," <user_created_at>",," [<user_screen_name>] ",," <user_default_profile_image> ",," <user_description> ",," <user_favourites_count> ",," <user_followers_count> ",," <user_friends_count> ",," [<user_listed_count>] ",," [<user_location>] ",," <user_name> ",," <user_screen_name> ",," <user_statuses_count> ",," [<user_time_zone>] ",," <user_urls> ",," <user_verified>

- <coordinates> = ***VARCHAR**
- <created_at> = <Date>
  - <Date> = "Mon" / "Tue" / "Wed" / "Thu" / "Fri" / "Sat" / "Sun" **WSP** "Jan" / "Feb" / "Mar" / "Apr" / "Mai" / "Jui" / "Jui" / Aug" / "Sep" / "Oct" / "Nov" / "Dec" **WSP** <jour_mois> **WSP** <heure> **WSP** "+0000" **WSP** 20 2**Digit**
  - <Heure> = 2**DIGIT**":"2**DIGIT**":"2**DIGIT**
  - <jour_mois> = "1" / "2" / "3" / "4" / "5" / "6" / "7" / "8" / "9" / "10" / "11" / "12" / "13" / "14" / "15" / "16" / "17" / "18" / "19" / "20" / "21" / "22" / "23" / "24" / "25" / "26" / "27" / "28" / "29" / "30" / "31"
- <hashtag> = *(**Alpha** / **Digit** / **WSP**)
- <media> = <inner_link>
  - <inner_link> = "https://twitter.com/" *(**Alpha** / **Digit** / '/')

    - <link> = “https://” *(**Alpha** / **Digit** / “.”) “.com” “ *(**Alpha** / **Digit** / ‘/’)
- <favorite_count> = ***Digit**
- <id> = 1***Digit**
- <in_reply_to_screen_name> = *(**Alpha** / **Digit** / ‘_’)
- <in_reply_to_status_id> = 1***Digit**
- <in_reply_to_user_id> = 1***Digit**
- <lang> = ‘fr’ / ‘en’ / ‘es’ / ‘al’
- <place> = ***varchar** *;Aucun exemple n’a été trouvé dans le data set*
- <possibly_sensitive> = ‘true’ / ‘false’
- <retweet_count> = ***Digit**
- <reweet_id> = ***Digit**
- <retweet_screen_name> = *(**Alpha** / **Digit** / ‘_’)
- <source> = ‘<a href=” ’ <link> “‘ rel=” ’ ***Alpha**/**Digit** “‘”>’ *(**Digit** / **Alpha** / . / ‘_’) ‘’
- <text> = “‘” ***VARCHAR** “‘”
- <tweet_url> = <inner_link>
- <user_created_at> = <date>
- <user_screen_name> = *(**Alpha** / **Digit** / ‘_’)
- <user_default_profile_image> = ‘True’ / ‘false’
- <user_description> = “‘” ***CHAR** “‘”
- <user_favourites_count> = ***Digit**
- <user_followers_count> = ***Digit**
- <user_friends_count> = ***Digit**
- <user_listed_count> = ***Digit**
- <user_location> = ***varchar** *;Aucun exemple n’a été trouvé dans le data set*
- <user_name> = “@” 1*(**Digit** / **Alpha** / **WSP**)
- <user_statuses_count> = <inner_link>
- <user_time_zone> = <date>
- <user_urls> = <link>
- <user_verified> = “false” / “true”

3. Spécification algébrique

Titre : ATTRIBUTE
Sort : AT
Reference : STRING, INT, IMG, DATE, BOOLEAN
Description : Un attribut compose un tweet.

CREATE : → AT
EQUAL : AT x AT → Boolean

EQUAL(at1, at2) = TRUE si at1 == at2 sinon FALSE&

Titre : TWEET
Sort : T<AT>
Reference : ATTRIBUTE, BOOLEAN
Description : Un tweet est un court message de 280 caractères max, associés à de nombreux attributs détaillant sa création, éventuellement associé à un ou plusieurs (4 max) médias.

CREATE : → T<AT>
EQUAL : T<AT> x T<AT> → Boolean

EQUAL(ens1, ens2) = TRUE si ens1 == ens2 sinon FALSE

Titre : ENSEMBLE
Sort : E<T>
Reference : TWEET, BOOLEAN, INTEGER
Description : Un ensemble de tweets. On peut en ajouter, retirer, vérifier qu'un tweet y existe déjà. On peut aussi obtenir la cardinalité d'une collection.

CREATE : → E<T>
ADD : E<T> x T → E<T>
REMOVE : E<T> x T → E<T>
UNION : E<T> x E<T> → E<T>
INTERSECTION : E<T> x E<T> → E<T>
CARD : E<T> → Integer
BELONG : E<T> x E<T> → Boolean

ADD(t, ens) = {ens,t}
ADD(t, Create) = {t}

UNION(ADD(ens, t), CREATE) = ADD(ens, t)
UNION(ens, t) = ADD(ens, t)

INTERSECTION(ADD(ens, t), CREATE) = ADD(ens, t)
INTERSECTION(ens, t) = {t} si BELONG(ens, t) sinon FALSE

REMOVE(ens, t) = INTERSECTION(ens, ¬t)
REMOVE (t, t) = CREATE()

CARD(CREATE) = 0
CARD(ADD(ens, t)) = CARD(ens) + 1
CARD(REMOVE(ens, t)) = CARD(ens) - 1
CARD(UNION(ens1, ens2)) = CARD(ens1) + CARD(ens2) -
CARD(INTERSECTION(ens1, ens2)
CARD(INTERSECTION(ens1, ens2)) = CARD(ens1) + CARD(ens2) -
CARD(UNION(ens1, ens2))

BELONG(ens1, ens2) = VRAI si CARD(INTERSECTION(ens1, ens2)) =
CARD(ens1) V CARD(INTERSECTION(ens1, ens2)) = CARD(ens2)
BELONG(CREATE, t) = FAUX
BELONG (ADD(ens, t1), t2) = Vrai ssi t1 = t2 V BELONG(ens, t2)
BELONG(REMOVE(ens, t1), t2) = VRAI ssi (BELONG(ens,t2) Λ t1 != t2)
BELONG(UNION(ens1,ens2),t) = VRAI ssi (BELONG(ens1,t) V
BELONG(ens2,t))
BELONG(INTERSECTION(ens1, ens2), t) = VRAI ssi BELONG(ens1, t) Λ
BELONG(ens2, t)

Conclusion

Cette version **1.0** du cahier des charges vise à intégrer 10 spécifications fonctionnelles et 2 spécifications non fonctionnelles, telles que décrites dans le cahier des charges. Le projet initial a été amélioré par nos équipes pour proposer de nouvelles fonctionnalités, en lien avec les besoins initiaux de Synevent. Nous espérons que ces nouvelles fonctionnalités auront permis d'anticiper d'éventuels besoins de la part de l'utilisateur, tout en restant cohérent avec les exigences initiales.

Afin de gagner en précision et de mettre un travail en accord avec nos exigences qualités, un travail supplémentaire de spécification algébrique et d'analyse des formats de données a été mis en place pour offrir tous les outils essentiels à nos équipes de développement.

Le projet devrait être fini pour le 16/12/2020, il passera ensuite par une phase d'analyse qualité par nos équipes avant la livraison finale du livrable aux équipes de Synevent.

Andromeda reste à l'entière disposition de ses équipes et de son client pour toute révision et modification du cahier des charges.