Constitution d'un jeu de données social à partir des données twitter

Violaine Guichet, Mathilde Plard

CONSTITUTION D'UN JEU DE DONNÉES SOCIAL PARTIR DES DONNÉES TWITTER

AUTEURES

Violaine GUICHET (IE), Mathilde PLARD (CNRS, UMR 6590 ESO)

RESUME

Le présent article présente les étapes de construction d'un jeu de données sociales sur un événement de course à pied. Les informations sources sont extraites du réseau social Twitter.

MOTS-CLES

Graphe du web, réseau social, Twitter, scraping, extraction, logiciel de visualisation et d'analyse réseaux

https://running-datalab.com/

CONTEXTE ET BESOIN

CONTEXTE INSTITUTIONNEL

Le programme de recherche CHALLENGE vise à documenter l'actuel engouement pour les événements sportifs associés à la course à pied à l'échelle internationale. Le Running DataLab (RDL) s'occupe d'une part de recenser et de qualifier ces événements pour alimenter la réflexion de CHALLENGE. D'autre part, le projet RDL analyse ces événements de courses à pied pour mieux comprendre leurs impacts sociaux et spatiaux.

BESOIN

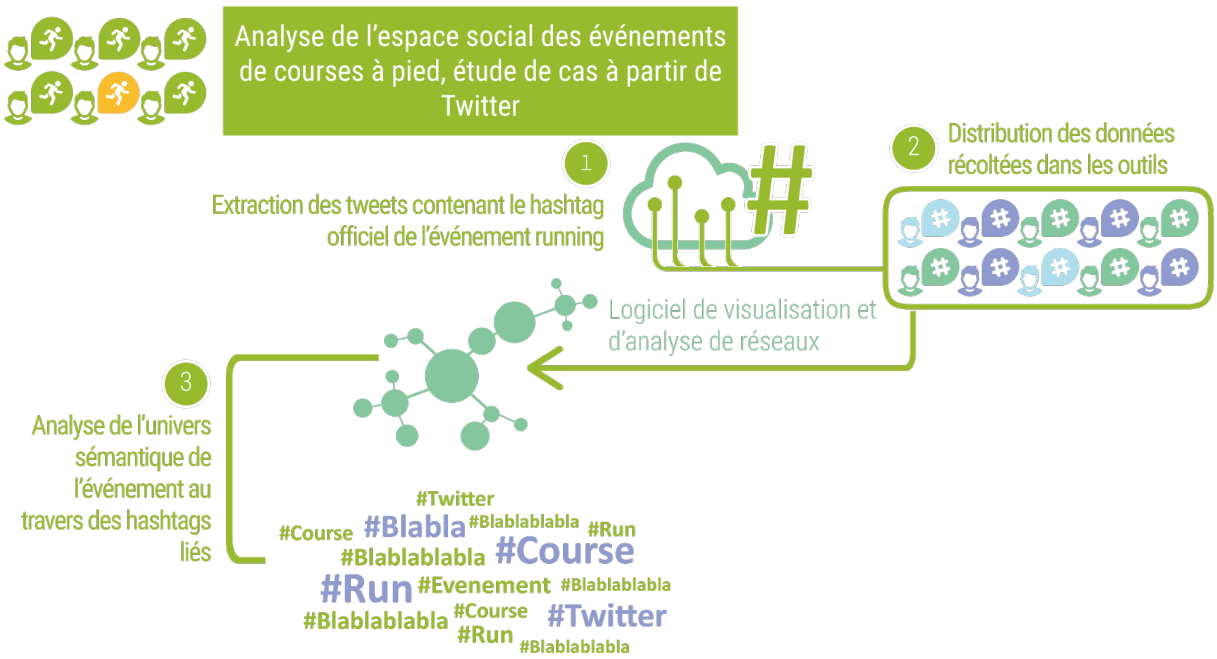
L'objectif est d'étudier l'espace social des événements de courses à pied. Plusieurs interrogations sont portées sur la résonnance sociale ou territoriale d'un événement. Ici, le travail présenté est un travail exploratoire. Il s'agit à partir des données Twitter d'étudier les possibilités d'analyse d'un événement de courses à pied en fonction des variables extraites et de leurs potentielles exploitations.

MÉTHODOLOGIE GLOBALE

METHODE

Dans un premier temps, un événement de course à pied est sélectionné. Une fois sélectionnés, tous les tweets contenant l'hashtag officiel de l'événement sont extraits. Les données extraites sont ensuite insérées dans un logiciel de visualisation et d'analyse réseaux afin de réaliser les graphes.

Figure 1. Méthodologie globale



Conception/Réalisation : Violaine Guichet, Ingénieure d'études (2018)
Projet Running DataLab, sous la direction de Mathilde Plard, chercheuse CNRS

CIBLAGE DE L'EVENEMENT

ÉVENEMENT DE COURSES A PIED RETENU

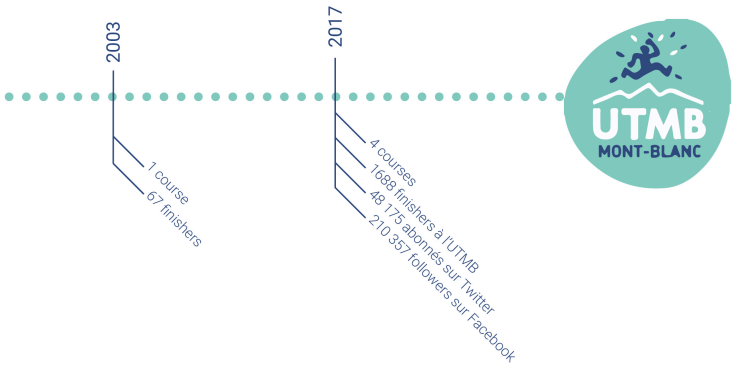
L'événement de courses à pied retenu pour cette manipulation est l'UltraTrail du Mont-Blanc (UTMB).

PRESENTATION DE L'UTMB

L'UTMB a été créé en 2003. Au début, l'événement était constitué d'une course. En quinze ans, l'UTMB s'est développé en proposant plusieurs courses (4 en 2017 et 5 en 2018) qui permettent de cibler un plus large public. Le nombre de finishers (coureu.r.se.s finissant la course) est passé de 67 à 1 688.

L'événement rassemble plus de 250 000 abonnés sur les différents réseaux sociaux.

Figure 2. Frise chronologique simplifiée de l'UTMB



Conception/Réalisation : Violaine Guichet, Ingénieure d'études (2018)
Projet Running DataLab, sous la direction de Mathilde Plard, chercheuse CNRS

EXTRACTION DES DONNEES

RENSEIGNEMENT SUR L'EVENEMENT

Lorsqu'un événement est ciblé, l'objectif est d'extraire les tweets dont son hashtag officiel est mentionné. L'hashtag officiel d'un événement est renseigné sur sa page Twitter.

Figure 3. Page Twitter de l'UTMB

Source : Twitter

RECOLTE DES DONNEES

EXTRACTION

Pour récolter les données, la librairie GetOldTweets est utilisée. Cette librairie écrite en python, permet de récolter les tweets selon un certain nombre de critères (nom de l'utilisateur, date, hashtag, etc.).

Concernant notre étude, les tweets mentionnant l'hashtag officiel de l'événement sont extraits.

Figure 4. **Script d'extraction des tweets**

```
Python Exporter.py -- querysearch « #UTMB »
```

RESULTAT DE L'EXTRACTION

Le script extrait les informations recherchées dans un csv. Concernant l'UTMB, il y a 49 593 tweets extraits qui ont été publiés entre le 14/02/2009 au 3/12/2017 (jour de l'extraction des données).

Pour chaque tweet est renseigné :

— username : l'utilisateur ayant envoyé le tweet

— date : date d'envoi

— retweets : nombre de retweets (= RT) du tweet

— favorites : nombre de favoris pour l'utilisateur

— text : message du tweet

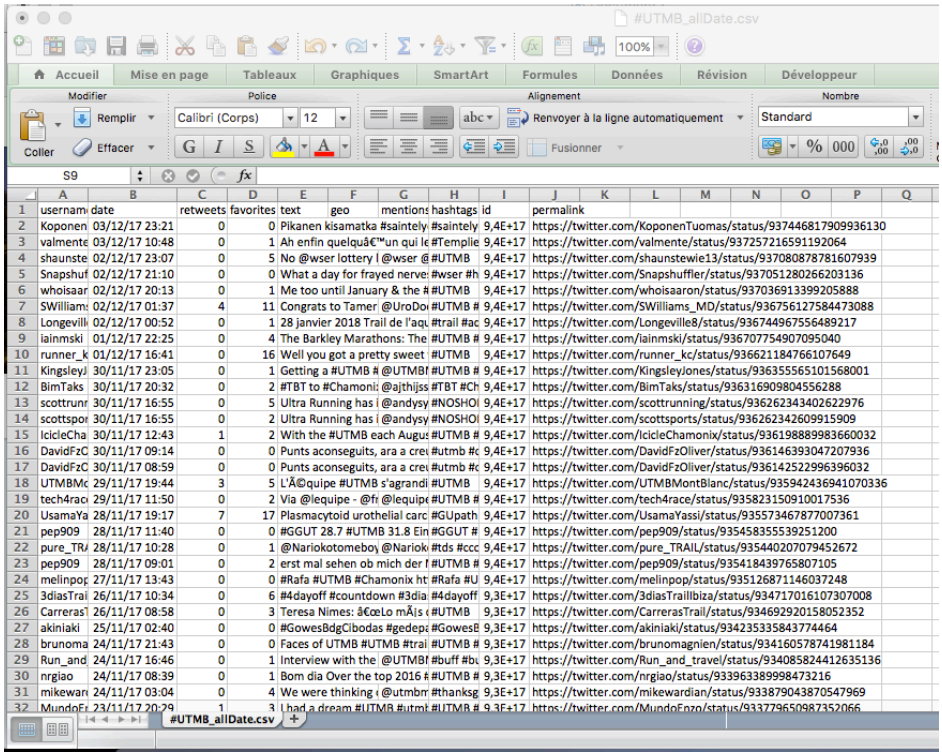
— geo : localisation lors de l'envoi du tweet (— mention : autre(s) utilisateur(s) mentionné(s)

— hashtags : hashtag(s) contenu(s) dans le tweet

— id : id du tweet

— permalink : page sur laquelle se trouve le tweet

Figure 5. **Aperçu du fichier csv en extraction**



Source : projet Running DataLab

Analyse des croisements possibles

Au vu de ces variables, différentes analyses apparaissent exploitables dans le cadre de l'étude des événements de courses à pied.

Le jeu d'acteurs

Le jeu d'acteurs peut être étudié de différente manière grâce à trois variables que sont username, retweets et favorites, nous permettent d'étudier la résonnance sociale de l'événement de courses à pied : qui publie ? De qui s'agit-il ? Quelle portée ont ces publications ?

La resonnance territoriale

La résonnance territoriale, soit la manière dont un événement fait écho sur d'autres territoires, pourrait-être détaillée à l'aide de trois variables : geo, hashtag et text. Dans ces trois éléments peuvent être mentionnés des territoires. Seulement, la variable geo est souvent erronée, elle renvoie à des éléments qui ne font pas référence à un lieu.

Pour cette étude il faudra alors réaliser des tests à partir des hashtags et du text.

Univers semantique

Enfin, l'univers sémantique ou ce à quoi est associé l'événement de courses à pied peut être exploré à travers une analyse plus fine des hashtags.

Choix pour realiser un premier essai

Ici le choix est fait d'étudier dans un premier temps le jeu d'acteurs et sa résonance.

Traitement dans Gephy

Retranscription d'un tweet en graphe

Un graphe du web est structuré en nœud et en lien. En fonction de leur relation, les nœuds sont connectés les uns aux autres.

Un tweet est constitué des différents éléments, il y a : l'utilisateur qui envoie le tweet, le tweet (soit le message écrit), la langue du tweet, la date et l'heure d'envoi. Dans certains cas, l'utilisateur mentionne un ou plusieurs utilisateur(s) (@user2, @user3, etc.), hashtags (#hashtags1, #hashtags2, etc.), médias (photographie, vidéo, etc.), liens ou la localisation de l'envoi du tweet.

Pour exemple, le tweet ci-dessous contient : un tweet, l'utilisateur qui a envoyé le tweet, la date et l'heure d'envoi, la langue du tweet, un médiat (photographie) et 3 hashtags.

Source : twitter

Figure 1. Retranscription d'un tweet en graphe

Conception/Réalisation : Violaine Guichet, Ingénieure d'études (2018)
Projet Running DataLab, sous la direction de Mathilde Plard, chercheuse CNRS

ADAPTATION DES DONNEES TABLEUR AU LOGICIEL GEPHY

Afin d'insérer les données dans le logiciel de visualisation et d'analyse réseau, il faut structurer les données d'une certaine manière. Pour que le logiciel puisse traiter la donnée, il faut stipuler une source et une cible (= »target »). Chaque source et chaque cible sont un nœud, la cible est reliée à sa source. Plus une source ou cible a de liens avec d'autres nœuds, plus elle est importante.

Concernant les acteurs, les champs mobilisés sont :

— users (= target)

— mentions (= source)

— favorites (= attributs)

Tableau 1. **Tableau de base lors de l'extraction des données**

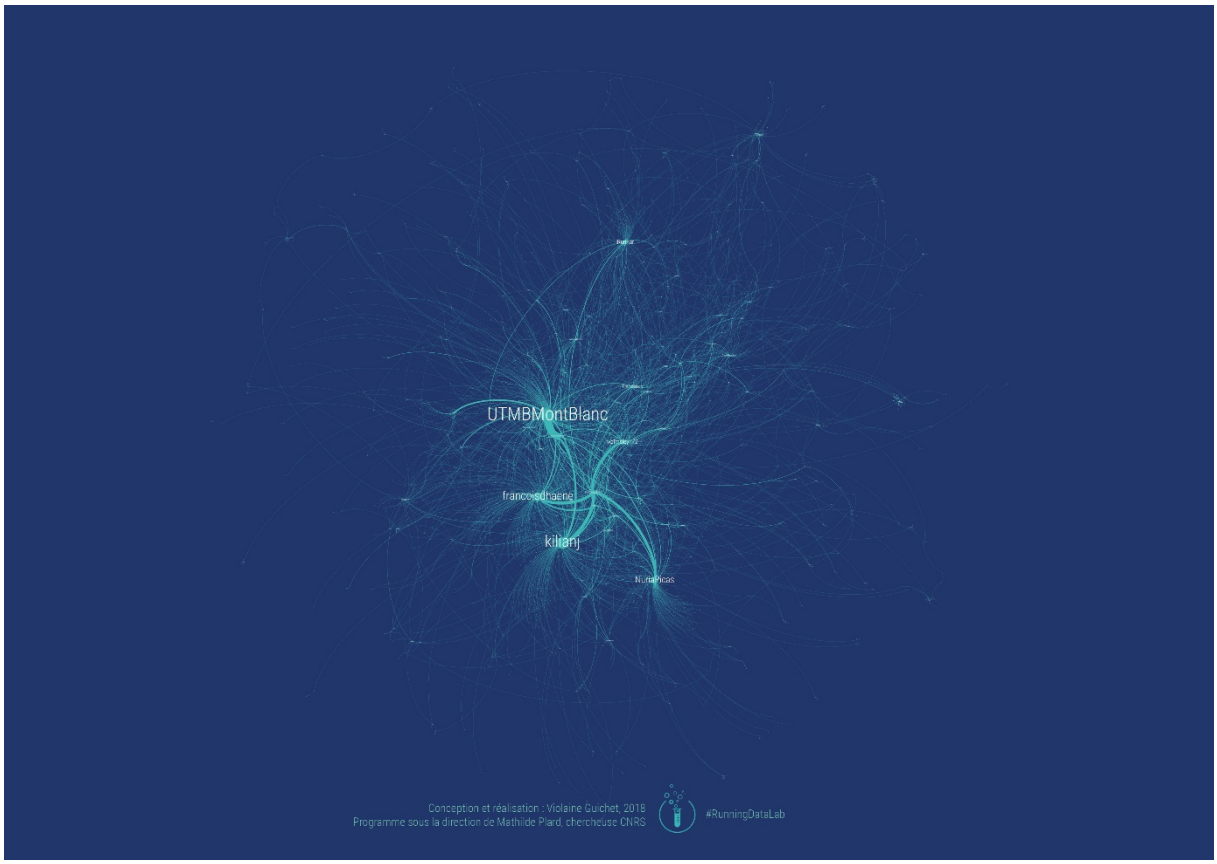
Tableau de base			
username	mentions	retweets	favorite
User1		0	0
User2	@User3, @User5	23	3
User4	@User1	44	132
User3	@User2	3	1
User1	@User4	4	0
User1		11	0
User5	@User3, @User2	6	4

Tableau 2. **Restructuration des données pour import dans le logiciel de visualisation et d'analyse réseau**

Tableau pour Gephi			
source	target	Retweets_total	favorite
User1		15	0
User2	User1	40	3
User2	User5	40	3
User4	User1	44	132
User3	User2	3	1
User1	User4	15	0
User1		15	0
User5	User3	6	4
User5	User2	6	4

RESULTAT

Le graphe



Première catégorisation des acteurs-Twitter

Un utilisateur peut tweeter souvent par rapport aux autres (ex : User1), être actif sur twitter, mais pour autant ne pas avoir de résonnance, puisque non retweeté.

En revanche, d'autres utilisateurs sont moins actifs, mais lorsqu'ils tweetent sont beaucoup repris, il s'agit des influenceurs.

Puis il y a les groupes d'amis ou communautés qui échangent entre eux et sont plus imperméables aux autres utilisateurs.

Piste d'explorations

Ce type de constitution de jeu de données permet d'analyser :

— les acteurs qui tweetent sur l'événement

— le poids de ces acteurs en fonction des RT

— les groupes d'acteurs en fonction des conversations

LIMITES, AMELIORATIONS & PERSPECTIVES

QUALIFIER LES ACTEURS

Pour avoir une analyse plus fine des acteurs, il faudrait télécharger la liste des comptes certifiés Twitter (https://twitter.com/verified/following) qui permettrait de différencier les particuliers, des institutionnels ou personnalités.

QUALIFIER LES HASHTAGS

La catégorisation des hashtags dans différents domaines (économie, acteur, territoire, etc.) permettrait une analyse plus fine des événements de courses à pied. De plus, cela permettrait d'évaluer en fonction de chaque événement le poids de chacun des domaines, et ainsi avoir une « photographie » de l'univers sémantique de chacune des courses. D'autre part, cela permettrait de mettre en lumière la résonance territoriale.