MIDI
— INNOVANT

Détection d'événements à partir d'un flux de tweets

MidInnovant de l'UMRsu

Sophie Baillargeon

25 janvier 2018

THALES

Partenariat **THALES** -

UNIVERSITÉ
LAVAL

Simon Hallé : Spécialiste et Manager du Portfolio Big Data,
Thales Canada Recherche et Technologie (TRT).

Sophie Baillargeon : Moi, étudiante au doctorat en statistique

- direction : **Thierry Duchesne**, professeur au
Département de mathématiques et de statistique
- codirection :

Christian Gagné : Professeur au Département de génie électrique
et de génie informatique.

Via un stage **Mitacs** Accélération au sein de l'**UMRSU**.

Projet de Thales Canada - Crowdstack

Ajouter des capacités d'analyse à Crowdstack

Faciliter le travail d'un analyste humain qui souhaite :

- **extraire de l'information** à partir d'une banque de tweets passés ;
- faire de la **surveillance en temps réel** de flux de tweets, dans le but de détecter des événements.

Description des données

Exemple de flux de tweets

Carrie Phillips @C_f_p

@British_Airways bravo! Yoghurt covered honeycomb is back in the EDI lounge!
Happy New Year indeed! But ssshhh! Don't tell everyone...! #onehappycustomer

[Full document...](#)

Arjun Sampath @Arjunbr84

the @VirginAmerica Elevate program was great, but now ceases to exist as all pts converted to @AlaskaAir miles. moral of the story is domestic air travel stinks. virgin provided a temporary respite, but its gone like the wind.

[Full document...](#)

Jessica Plascencia @jessplas

Looking at plane tickets now @AmericanAir help me <https://t.co/XHqjNAziG0>

[Full document...](#)

حسن مجراتين @Hsajwanzaton

#Emirates cancels its flights #EK701 & #EK702 #Dubai - #Mauritius until tomorrow
18th Jan 2018 as the islands embrace for Cat3 storm #cyclone #Berguita
<https://t.co/BaIC2vgY3m>

[Full document...](#)

Ben Jones @Begelbeard

Hey @SouthwestAir - 100+ min wait times right now. Any chance I can get help with a travel credit / new booking?

[Full document...](#)

Tiorde @godsgood00210

@Delta This is just a perfect example of the sorry state of affairs of customer service for large companies. No employee can actually do anything. Just cite policies and blame others.

[Full document...](#)

Georgia Singh @softmenbys1975

RT @noelle_morgan11: My husband Graham,tucked in,praying hands,thanking the Lord for a comfortable night in his @Qantas PJs.#QFINnovation h...

[Full document...](#)

Exemple de tweet

Home Moments Search Twitter Have an account? Log in

Thales Group @thalesgroup [Follow](#)

Professor Tremblay, can you tell us what are the benefits of the collaboration between Thales and Urban Science of Quebec?
[@SmartCityexpo](#)

Joint Research Unit in Urban Science

- Consortium of Universities, Industries and Governments
- Synergy between developers and researchers, and
- Promoting innovation of technology in

372 views 0:13 / 1:20

9:14 AM - 16 Nov 2016

8 Retweets 13 Likes

© 2017 Twitter About Help Center Terms Privacy policy Cookies Ads info

Thales Group @thalesgroup
Leading international electronics and systems group, addressing global #defence, #space, #aerospace, #security and #transportation markets.
Paris, France
thalesgroup.com
Joined March 2009

Exemple de retweet

The image shows a screenshot of a Twitter interface. At the top, there are navigation options: 'Home', 'Moments', 'Search Twitter', and 'Have an account? Log In'. The profile of 'UMRsu @UMRsuQc' is visible in the top left. The main focus is a tweet by 'Christian Gagné @chgagne' from November 17, 2017, at 11:17 AM. The tweet text is: 'Laval Turns AI Research into Real-World Solutions' followed by a link 'researchinfosource.com/pdf/CIL2017-La ...' and mentions '@universitelaval @CRDM_UL @UMRsuQc'. It has 6 retweets and 11 likes. Below the tweet, a retweet by 'UMRsu @UMRsuQc' from November 21 is shown. The retweet text is: '#Flashinnovation de la semaine : Une tuque signée @thenorthface dont la laine provient d'une ferme dont le mode de fonctionnement permet d'emmagasiner annuellement l'équivalent des émissions polluantes de 850 véhicules. bit.ly/2VaYGu'. The retweet includes a photo of a dark grey knit beanie against a landscape background.

Exemple de tweet cité

The image shows a screenshot of a Twitter interface. At the top, there are navigation options: Home, Moments, Search Twitter, and a link to 'Have an account? Log in'. The main content is a tweet from UMRsu (@UMRsuQc) with a 'Follow' button. The tweet text reads: 'Profitez de cette conférence pour en apprendre un peu plus sur ce phénomène en émergence. #VoituresAutonomes'. Below the text is a quote tweet from ITIS (ULaval) (@itis_ul) with a small image of a car. The quote tweet text says: 'Comment fonctionnent les véhicules autonomes? Conférence de Philippe Giguère @BibliQc jeudi 16 nov 19h. Entrée libre! bit.ly/2AdM241'. The tweet is dated '6:58 AM - 7 Nov 2017' and has '1 Like'. At the bottom left, there is a circular logo for UMRsu. At the bottom right, there is a footer with copyright information: '© 2017 Twitter About Help Center Terms Privacy policy Cookies Ads info'.

Home Moments Search Twitter Have an account? Log in

UMRsu @UMRsuQc Follow

Profitez de cette conférence pour en apprendre un peu plus sur ce phénomène en émergence. #VoituresAutonomes

ITIS (ULaval) @itis_ul
Comment fonctionnent les véhicules autonomes? Conférence de Philippe Giguère @BibliQc jeudi 16 nov 19h. Entrée libre! bit.ly/2AdM241

6:58 AM - 7 Nov 2017

1 Like

© 2017 Twitter About Help Center Terms Privacy policy Cookies Ads info

UMRsu @UMRsuQc
L'Unité Mixte de Recherche en sciences urbaines est un réseau de recherche et d'innovation - UMRsu is a research and innovation network
Québec, Canada
umr-su.com
Joined November 2016

© 2017 Twitter About Help Center Terms Privacy policy Cookies Ads info

Exemple de réponse à un tweet

The screenshot shows a Twitter interface with a tweet and its reply. The background features a dark grey header with navigation options like 'Home', 'Moments', and 'Search Twitter'. On the left, there is a profile card for 'Two Hat Security' with a logo and bio. On the right, there are large teal and red circular graphics, one containing a smartphone icon and the word 'COMMUNITIES'.

Original Tweet:
User: **FSG ULAVAL** @FSGulaval · Oct 31
Entente de 1.7 M\$ entre le CRDM_UL @TwoHatHQ et @MitacsCanada pour 118 stages sur quatre ans à Kelowna et Québec pour les étudiants UL.
Media: **ulavalmédiás** @ulavalmedias
Contenus toxiques en ligne: Centre de rech. en données massives @universitelaval signe entente avec Two Hat Security ulaval.ca/notre-universi...

Reply:
User: **Two Hat Security** @TwoHatHQ
Replied to @FSGulaval @MitacsCanada
C'est un honneur pour nous de travailler avec l'Université Laval, Mitacs et l'Université Simon Fraser sur ce projet!
4:59 PM - 2 Nov 2017
2 Likes

S'agit-il de données massives ?

Sommes-nous en présence de :

- **V**olume ?
- **V**élocité ?
- **V**ariété ?
- **V**éracité potentiellement douteuse ?

Volume et Vitesse

- Volume croissant, influencé par les filtres
- Vitesse parfois grande, variable dans le temps et entre les flux

Variété : Un tweet, ce n'est pas que du texte !

Véracité : bruit et incertitude

- Bruit : plusieurs tweets non pertinents à filtrer
- Incertitude associée à des variables dérivées

- localisation du tweet inférée à partir du profil de l'auteur

- certaines métadonnées du CNRC proviennent de modèles
 - sentiments : probabilités d'être négatif, neutre ou positif.
 - émotions : scores de tristesse, joie, peur, aversion, colère, surprise.

Autres défis propres aux tweets

- redondance (retweets officiels, faux retweets)
- particularités du texte, par exemple :

RT @Flarouche: Tjs impressionné par @RTCQuebec à chaque visite de #Qc. Écrans dans bus, marketing, propreté. @stminfo devrait s'en inspirer...

Nothing new lol <https://t.co/goV2vnbbZV>

Les midibus (800000\$ chaque) du @RTCQuebec ont des prises USB !! Wow, on va pouvoir recharger nos iPhones 10-15min de charge 🙌🏻🙌🏻 !! #villequebec

Approche proposée

Fonctionnement de l'approche proposée

1. Regrouper les tweets similaires afin de réduire la quantité d'information à traiter.
2. Mettre en évidence les grappes de tweets intéressantes pour l'analyste.

Données arrivant en temps réel

État des grappes à priori

Traitement des données par petits-lots

Classement des groupes déjà vus

Lot de tweets à traiter

*comparaison des identifiants
de groupe de retweets*

Grappes actives

Classement par contrainte

Lot de tweets à traiter

Classement non supervisé

Lot de tweets à traiter

*calcul de plusieurs distances,
basées sur différents attributs*

Grappes actives
mises à jour

Formation de nouvelles grappes (clustering hiérarchique)

Lot de tweets à traiter

Grappes actives
mises à jour

Qualification des grappes

Grappes actives

Des grappes peuvent être déclarées inactives.

Mise à jour de :

- score d'intérêt
- modèle d'activité
- statistiques agrégées (si la grappe a accueilli de nouveaux tweets)

État des grappes à postériori

Morceaux statistiques de l'approche

Clustering hiérarchique

Pour chaque distance considérée, basée sur une des nombreuses variables :

Group-group distance matrix

	g ₁	g ₂	g ₃	g ₄	g ₅	g ₆
g ₁	0.0	0.1	0.7	0.9	1.0	0.8
g ₂	0.1	0.0	0.4	0.6	0.9	1.0
g ₃	0.7	0.4	0.0	0.8	1.0	0.7
g ₄	0.9	0.6	0.8	0.0	1.0	0.9
g ₅	1.0	0.9	1.0	1.0	0.0	0.6
g ₆	0.8	1.0	0.7	0.9	0.6	0.0

Dendrogram from hierarchical clustering

Partition after cutting the dendrogram

Clustering par ensemble

Distance matrix 1

(feature 1, distance 1)

	g_1	g_2	g_3	g_4
g_1	0	1	0	1
g_2	1	0	1	1
g_3	0	1	0	1
g_4	1	1	1	0

Clustering 1

Distance matrix D

(feature P, distance 3)

	g_1	g_2	g_3	g_4
g_1	0	5	--	6
g_2	5	0	--	1
g_3	--	--	0	--
g_4	6	1	--	0

Clustering D

g_i : group of retweets
 c_j : cluster of groups
-- : missing value

Consensus clustering

Modèle d'activité par grappe

Modélisation de l'arrivée des tweets dans la grappe :

Travaux futurs

Travaux futurs

- Poursuivre l'implémentation de l'approche

- Tester l'approche sur différents flux de tweets

- Publier l'approche

Merci !