# Classification des utilisateurs de Twitter en fonction de leurs comportements à l'aide d'un algorithme des K-means

Vincent Brault, Jean-Marc Francony, Adeline Samson, Matthieu Meynet

# Classification des utilisateurs de Twitter en fonction de leurs comportements à l'aide d'un algorithme des $K$-$means$

Vincent Brault*, Jean-Marc Francony**,
Adeline Leclercq-Samson*, Matthieu Meynet***

*Univ. Grenoble Alpes, LJK, F-38000 Grenoble, France
Univ. Grenoble Alpes, CNRS, F-38000 Grenoble, France
vincent.brault@univ-grenoble-alpes.fr et adeline.leclercq-samson@univ-grenoble-alpes.fr,
https://www-ljk.imag.fr/membres/Vincent.Brault/ et http://adeline.e-samson.org/fr+en/
**Univ. Grenoble Alpes, UMR PACTE, F-38000 Grenoble, France
jean-marc.francony@univ-grenoble-alpes.fr
https://www.pacte-grenoble.fr/membres/jean-marc-francony
***Univ. Grenoble Alpes, GRESEC, F-38000 Grenoble, France
matthieu.meynet@univ-grenoble-alpes.fr

**Résumé.** Dans cette présentation, nous proposons une analyse des comportements des utilisateurs de Twitter durant la campagne des élections européennes de 2014 dans le but de former des groupes d'utilisateurs qui avaient tendance à faire des actions aux mêmes moments durant cette campagne. Pour ce faire, nous effectuons une classification des jours de la campagne électorale puis, à l'aide de la classification obtenue, nous proposons une classification des utilisateurs dans le but d'étudier leurs comportements. Cette classification imbriquée nous oblige à transformer le modèle pour permettre une exécution plus rapide de l'algorithme des $K$-$means$. Nous concluons cette présentation par une analyse des résultats obtenus.

## 1 Introduction

Twitter sert de plateforme de réseau social (ou bien encore site de microblogage) qui a progressivement été investie par les agents du champs politique (hommes politiques, journalistes...) depuis 2008 et l'élection de Barack Obama (voir par exemple Heinderyckx (2011)). Dans ce travail, nous étudions les *retweets* (c'est-à-dire les *tweets* qui citent d'autres *tweets*) et les *mentions* dans un tweet (les *tweets* dans lesquels sont mentionnés d'autres utilisateurs de Twitter) durant la campagne des élections européennes de 2014 dans le but de voir s'il y a des consignes au sein des partis et/ou des stratégies de communication militantes (voir par exemple Bruns (2012)).

Pour ce faire, nous avons proposé une stratégie de classification des utilisateurs faite en deux temps : nous avons commencé par faire une classification des jours afin que les événements exceptionnels (comme le jour de l'élection) ne perturbent pas l'analyse des comporte-

ments puis une classification des utilisateurs afin d'étudier les comportements communs dans chaque groupe.

# 2 Données, modélisation et notations

Dans cette étude, nous disposons des *retweets* et *mentions* de $n = 7\,193$ utilisateurs comportant certains hashtags particuliers (comme **#EE2014** ou **#Europeennes2014**) durant la campagne de l'élection européenne 2014. Plus précisément, la période étudiée s'étale du 27 avril 2014 au 2 juin 2014 ; soit un peu avant le début officiel de la campagne et jusqu'à une semaine après le vote du dimanche 25 mai 2014 (soit un total de $J = 37$ jours). Notons qu'un même tweet peut contenir plusieurs *mentions* et donc faire référence à plusieurs utilisateurs, dans ce cas, le tweet en lui même sera compté pour chaque *mention*. Au total, nous avons $119\,673$ *retweets* et $203\,587$ *mentions*.

Afin de contourner le caractère instantané d'un tweet, nous avons choisi de regrouper les informations par plages horaires d'une heure et nous noterons $N^R_{i,j,h}$ (resp. $N^M_{i,j,h}$) le nombre de *retweets* (resp. *mentions*) faits par l'utilisateur $i \in \{1, \dots, n\}$ durant l'heure $h \in \{1, \dots, 24\}$ du jour $j \in \{1, \dots, J\}$.

Ainsi, le vecteur $N^R_{i,\cdot,h}$ représente tous les *retweets* faits par l'utilisateur $i$ à l'heure $h$ de chaque jour de la période étudiée. S'il a un comportement journalier régulier sur Twitter, toutes les valeurs de ce vecteur devraient être proches.

# 3 Classification

La classification se fait en deux temps : d'abord la classification des jours puis, étant donnée une classification des jours, la classification des utilisateurs.

Dans la suite, nous présenterons les procédures sur les *retweets* sachant que les *mentions* ont été traitées de la même façon.

## 3.1 Classification des jours

Nous avons d'abord cherché à classifier les jours en fonction du comportement global des utilisateurs. Pour cela, nous avons fait la somme sur tous les *retweets* de tous les utilisateurs pour chaque heure de chaque jour, noté $N^R_{+,j,h}$. Nous obtenons ainsi 37 individus dans $\mathbb{N}^{24}$ et nous avons utilisé un algorithme des $K$-$means$ (voir par exemple Hartigan (1975)) afin de regrouper les jours.

Nous pouvons voir sur la figure 1 les répartitions des classes. Les deux pics des *retweets* correspondent respectivement au jour du débat télévisé *Des paroles et des actes* sur France 2 (le 22 mai) et le jour de l'élection (le 25 mai). Nous voyons que l'algorithme a très vite séparé ces jours puis il a regroupé les jours d'avant et d'après la période électorale et a enfin partitionné progressivement les jours de la campagne (par exemple en séparant les jours suivants les deux jours importants de la campagne). À l'opposé, il semblerait que le jour de la semaine n'intervienne pas trop dans le comportement des utilisateurs de Twitter.

Du point de vue de l'analyse, il a été choisi de se concentrer sur $5$ classes de jour incluant trois classes de comportements différents (faiblement actifs hors période, moyennement et

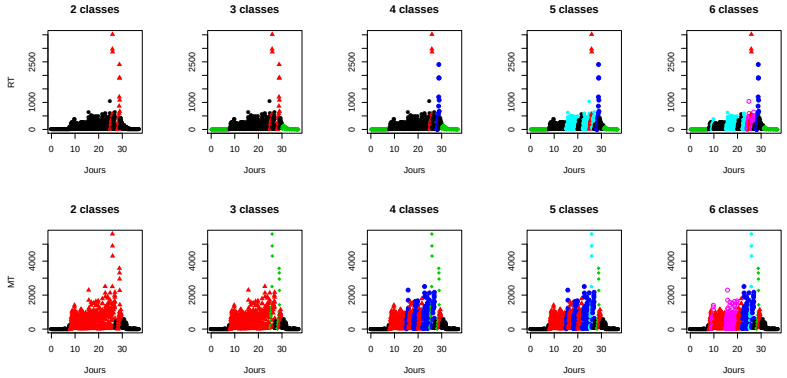


FIG. 1 – *Représentation des classes de jours obtenues pour les retweets (ligne du haut) et les mentions (ligne du bas) pour un nombre de classes allant de 2 à 6 (colonnes) : pour chaque graphique, l'abscisse est mis à partir du premier jour d'étude (jour 0), chaque point représente une heure et les couleurs symbolisent les classes.*

plutôt actifs durant la période électorale) et deux classes ne contenant chacune qu'un jour particulier. Toutefois, nous avons étudié également les cas avec un nombre de classes différent.

## 3.2 Classification des utilisateurs

Étant donnée une classification des jours en $K$ classes, nous avons cherché à classifier les utilisateurs en groupes de telle sorte que :

— les utilisateurs d'une même classe aient des comportements similaires et différents de ceux des utilisateurs d'autres classes,

— pour une classe de jours donnée, chaque utilisateur ait le même comportement chaque jour de cette classe (par exemple, il va avoir l'habitude de *retweeter* vers 10h tous les jours d'une même classe).

Pour ce faire, il est important de coder l'algorithme des $K$-$means$ de façon à limiter le temps de calcul. Nous notons $\mathbf{w} = (w_{j,k}) \in \mathcal{M}_{J\times K}(\{0,1\})$ la matrice binaire de classification des jours telle que $w_{j,k}$ vaut 1 si et seulement si le jour $j$ est dans la classe $k$. Nous notons $N^R_{i,\cdot,\cdot}$ la tranche $i$ de taille $H \times J$ de la matrice en trois dimensions $N^R$. Alors le comportement moyen de l'individu $i$ pour chacune des classes de jour se résume par la matrice $\mathbf{C}^i = \left(N^R_{i,\cdot,\cdot}\right)^T \overline{\mathbf{w}}$ de taille $H \times K$ avec $\overline{\mathbf{w}}$ la matrice moyennée de $\mathbf{w}$ où chaque case est divisée par la somme des cases de la colonne dans laquelle elle appartient et $\mathbf{A}^T$ la transposée de la matrice $\mathbf{A}$ :

$$\mathbf{C}^i = \begin{pmatrix} N^R_{i,1,1} & N^R_{i,2,1} & \cdots & N^R_{i,J,1} \\ N^R_{i,1,2} & N^R_{i,2,2} & & \vdots \\ \vdots & & \ddots & \vdots \\ N^R_{i,1,H} & N^R_{i,2,H} & \cdots & N^R_{i,J,H} \end{pmatrix} \begin{pmatrix} \frac{w_{1,1}}{w_{+,1}} & \frac{w_{1,2}}{w_{+,2}} & \cdots & \frac{w_{1,K}}{w_{+,K}} \\ \frac{w_{2,1}}{w_{+,1}} & \frac{w_{2,2}}{w_{+,2}} & & \vdots \\ \vdots & & \ddots & \vdots \\ \frac{w_{J,1}}{w_{+,1}} & \frac{w_{J,2}}{w_{+,2}} & \cdots & \frac{w_{J,K}}{w_{+,K}} \end{pmatrix}.$$

Par exemple, $C^i_{\cdot,k}$ est le comportement journalier moyen de l'individu $i$ durant les jours de la classe $k$. En agglomérant toutes les matrices $\mathbf{C}^i$, nous obtenons donc une matrice en trois dimensions de taille $n \times H \times K$. Comme la manipulation de ces matrices est complexe, nous faisons le choix de vectoriser [1] chaque matrice $\mathbf{C}^i$ de telle sorte qu'en concaténant les colonnes, nous obtenons le profil moyen des 24 heures de la première classe, puis de la seconde... Par les formules classiques, nous obtenons $\text{Vec}\left(\mathbf{C}^i\right) = \left(Id_H \otimes \overline{\mathbf{w}}^T\right) \text{Vec}\left({N^R_{i,\cdot,\cdot}}^T\right)$ où $\otimes$ est le produit de Kronecker [2]. La concaténation des ces vecteurs $\text{Vec}\left(\mathbf{C}^i\right)$ permet d'obtenir une matrice de taille $(KH) \times n$ qui rend le stockage et l'implémentation de l'algorithme $K\text{-}means$ plus aisés.

L'intérêt de cette représentation est également que nous pouvons inclure facilement d'autres contraintes comme la concaténation des heures de nuits en modifiant légèrement la formule.

# 4 Présentations

Dans cet exposé, nous présenterons la méthode et expliquerons les différentes visions du problème permettant d'améliorer la vitesse d'exécution de l'algorithme $K\text{-}means$. Nous terminerons par une présentation des résultats obtenus à l'aide d'une application shiny.

# Références

Bouveyron, C., P. Latouche, et R. Zreik (2016). The stochastic topic block model for the clustering of vertices in networks with textual edges. *Statistics and Computing*, 1–21.

Bruns, A. (2012). Journalists and twitter : How australian news organisations adapt to a new medium. *Media International Australia 144*(1), 97–107.

Hartigan, J. A. (1975). *Clustering algorithms* (99th ed.). New York, NY, USA : John Wiley & Sons, Inc.

Heinderyckx, F. (2011). Obama 2008 : l'inflexion numérique. *Hermès, La Revue* (1), 135–136.

# Summary

In this talk, we provide an analysis of the behavior of Twitter users during the 2014 European elections campaign. Our goal is forming user groups that tended to do the same action at the same time during this campaign. We propose a classification of the days of the electoral campaign and, using the classification obtained, we propose a classification of the users in order to study their behaviors. This nested classification forces us to transform the model to allow faster execution of the $K\text{-}means$ algorithm. We conclude this talk with an analysis of the results obtained.

1. La vectorisation d'une matrice $\mathbf{C}$ de taille $H \times K$ consiste à la transformer en un vecteur de longueur $HK$ en concaténant les colonnes les unes au-dessus des autres.

2. Le produite de Kronecker d'une matrice $\mathbf{A}$ de taille $n \times p$ et d'une matrice $\mathbf{B}$ de taille $m \times q$ est une matrice de taille $(nm) \times (pq)$ de telle sorte que la matrice $\mathbf{A} \otimes \mathbf{B}$ est composée de blocs $a_{i,j}\mathbf{B}$.