

Raconte moi un match...

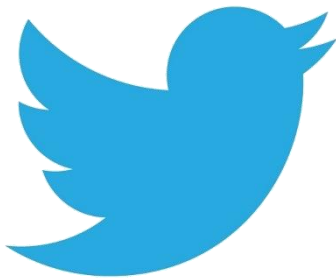
HackaTAL 2016



ROCKET
= labs =

Contexte

- Des tweets
 - Sur l'Euro 2016
 - En français, en arabe, en anglais
- Peut-on raconter le match à partir de ces tweets ?



Tâche

- A partir des tweets, **générer un fichier d'événements** :
 - Timestamp, type d'évènement, infos complémentaires
- Exemple fichier *Albanie Suisse* :

TEMPS ABSOLU (HH:MIN)	EVENEMENT	ANNOTATION COMPLEMENTAIRE
15:01:00	D1P	
15:06:00	BUT	Schär

- Cerise sur le gâteau : un joli résumé en langue naturelle...



Principe de notre approche (I)

- Trop de bruit dans le texte
- Seuls quelques twittos sont nécessaires
- Identifier les influenceurs !



Principe de notre approche (2)

- Les influenceurs s'expriment :
 - De manière **compréhensible** pour notre moteur
 - Sur **tous** les matchs
 - En **cohérence** avec les événements réels
- Les humains font pareil : personne ne suit 30 000 flux en 3 langues pour savoir comment se passe le match 😊

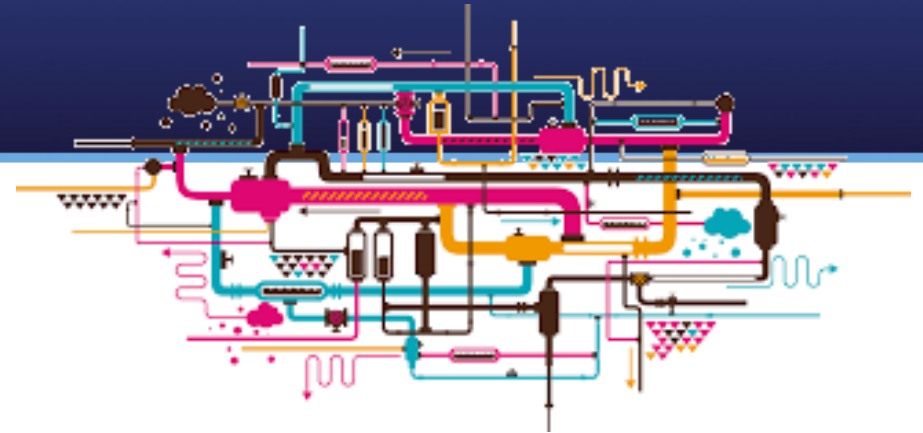


Intérêt de l'approche

- Parmi tous les twittos, au moins 2 ou 3 twittent de manière claire avec une bonne orthographe
- Ceux là sont suffisants !
- Le bruit est géré simplement
- Devrait identifier les influenceurs institutionnels, i.e. qui twittent correctement sur tous les matchs (vs. le supporter d'une équipe qui ne twitte que pour les matchs de son équipe...)



Pipeline



1. Training
 1. Identifie les influenceurs
2. Test
 1. Analyse les tweets des influenceurs uniquement
 2. Extrait les évènements
3. Output
 1. Convertit le fichier évènements en HTML

Entraînement

- Pour chaque match
 - Pour chaque événement
 - Extraire les tweets dans les 2 minutes suivant l'événement
 - Pour chaque tweet
 - Extraire les entités parmi Score, Type évt, Joueur, Pays
 - Si les entités détectées sont conformes avec l'événement
 - +1 pour le tweetos
 - Stocke l'écart temporel entre l'événement et le tweet
- Extrait les tweetos ayant le score le plus élevé -> ce sont les influenceurs
- Le temps moyen d'émission par tweetos est également calculé



Test

real life.

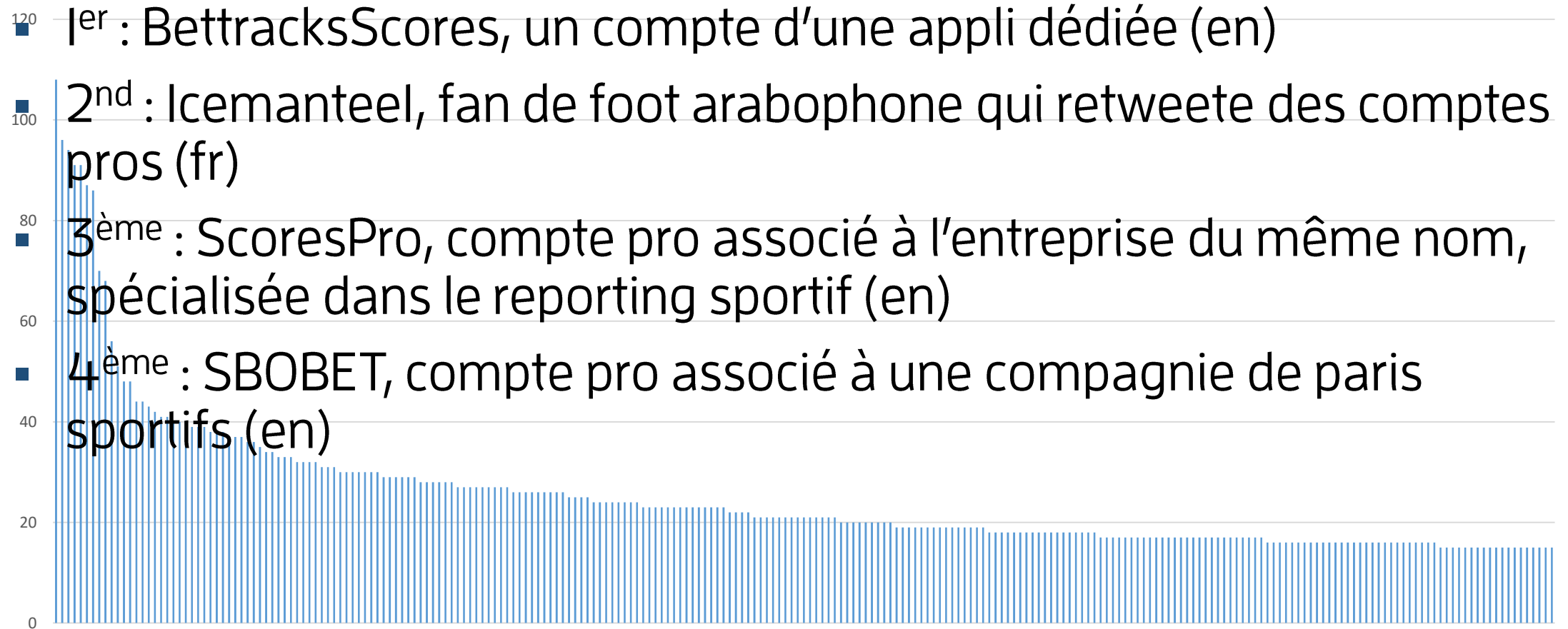
- Pour chaque tweet
 - Extraire le tweet, le timestamp, les entités
- Pour chaque ligne
 - Fusionner les lignes faisant référence au même évènement
 - Type d'évènement égal
 - Timestamp dans un intervalle de 2 minutes
- Pour chaque ligne du fichier fusionné
 - Convertir le format interne en format attendu

L'équipe !

- Corentin Ribeyre : supervision du code Python et dév global
- Xin He : extraction des entités nommées (Python)
- Yu Qian : identification des influenceurs
- Naruemon Saeng : réflexion réseaux de neurones (TensorFlow)
- Aïssa El Ouafi : génération linguistique, rendu Web (Python, HTML)
- Gilles Moyse : conception, présentation, support, outils, animation ☺



Résultats : les influenceurs



Résultats : les reports de match

- Démo !



Perspectives

- Utiliser plus de données sociales
 - Nb followers, nb retweets
- Automatiser la méthode
 - Apprendre les influenceurs
- Approche Bayésienne
 - 3 processus stochastiques émettent des séquences de caractères
 - Comment estimer les paramètres d'un processus latent qui émet des événements ?