Faculté des arts et des sciences
Département d'informatique
et de recherche opérationnelle

rali

Détection automatique d'évènements à partir de Twitter

Houssem Eddine DRIDI

Soutenance de thèse, 2 octobre 2014

Houssem dridi
Voir ma page de profil

33	164	20
TWEETS	ABONNEMENTS	ABONNÉS

Écrire un nouveau Tweet...

Suggestions · Actualiser · Tout afficher

Mohamed Ali Chebâane @MedAl...
Suivi par Eya Turki et d'autres
Suivre

Yassine Brahim @Yassine_Brahim
Suivre

Parcourir les catégories · Trouver vos amis

Tendances · Modifier

#galaartis
#PCPointer
#ThingsNotToDoOnAFirstDate
#bcdebate2013
#cdnpoli
Oakville
CBC
Bernabeu
Madrid

Tweets

Oueld Ifriqiya @Khamousss — 1 min
Le FLN, un État dans l'État. #Algérie
Ouvrir

Tunisia News @tunisinews — 1 min
#Tunisie #Tunisia Tunisie - Politique : Soumaya Ghannouchi s'attaque à la France bit.ly/14Q4tKu
Ouvrir

Tunisia News @tunisinews — 1 min
Urgent : Report de la grève des transporteurs de carburants : Tunisie : Actualités : Société :... goo.gl/8wS0F #Tunisie #Tunisia
Ouvrir

SAT-7 Arabic @sat7ar — 1 min
برنامج كرمالك عراق - القداسة: youtu.be/qEzWEEUC07w?a via @YouTube
Afficher le média

koora @KooraTunisie — 2 min
Mi-temps : RM-DOR 0-0 (1-4)
Ouvrir

radioexpressfm @radioexpressfm — 3 min
L'interview de Hedi Ben Abbes dans Tac au Tac fb.me/IHx5mvzb
Ouvrir

@khawlamagrebi **الاخوان عدو الانسان** — 4 min
علماء: فيروس التهاب الكبد الفيروسي عمره أكثر من 82 مليون سنة fb.me/PRHIuygX #أووووووووف
Ouvrir

Houssem dridi
Voir ma page de profil

33	164	20
TWEETS	ABONNEMENTS	ABONNÉS

Écrire un nouveau Tweet...

Suggestions · Actualiser · Tout afficher

Mohamed Ali Chebâane @MedAl...
Suivi par Eya Turki et d'autres
Suivre

Yassine Brahim @Yassine_Brahim
Suivre

Parcourir les catégories · Trouver vos amis

Tendances · Modifier

#galaartis
#PCPointer
#ThingsNotToDoOnAFirstDate
#bcdebate2013
#cdnpoli
Oakville
CBC
Bernabeu
Madrid

Tweets

Oueld Ifriqiya @Khamousss — 1 min
Le FLN, un État dans l'État. #Algérie
Ouvrir

Tunisia News @tunisinews — 1 min
#Tunisie #Tunisia Tunisie - Politique : Soumaya Ghannouchi s'attaque à la France bit.ly/14Q4tKu
Ouvrir

Tunisia News @tunisinews — 1 min
Urgent : Report de la grève des transporteurs de carburants : Tunisie : Actualités : Société :... goo.gl/8wS0F #Tunisie #Tunisia
Ouvrir

SAT-7 Arabic @sat7ar — 1 min
youtu.be/qEzWEEUC07w?a via @YouTube :برنامج كرمالك عراق - القداسة
Afficher le média

koora @KooraTunisie — 2 min
Mi-temps : RM-DOR 0-0 (1-4)
Ouvrir

radioexpressfm @radioexpressfm — 3 min
L'interview de Hedi Ben Abbes dans Tac au Tac fb.me/IHx5mvzb
Ouvrir

@khawlamagrebi الاخوان عدو الانسان — 4 min
علماء: فيروس التهاب الكبد الفيروسي عمره أكثر من 82 مليون سنة fb.me/PRHIuygX #أووووووووف
Ouvrir

Houssem dridi
Voir ma page de profil

33 TWEETS | 164 ABONNEMENTS | 20 ABONNÉS

Écrire un nouveau Tweet...

Suggestions · Actualiser · Tout afficher

Mohamed Ali Chebâane @MedAl...
Suivi par Eya Turki et d'autres
Suivre

Yassine Brahim @Yassine_Brahim
Suivre

Parcourir les catégories · Trouver vos amis

Tendances · Modifier

#galaartis
#PCPointer
#ThingsNotToDoOnAFirstDate
#bcdebate2013
#cdnpoli
Oakville
CBC
Bernabeu
Madrid

Tweets

Oueld Ifriqiya @Khamousss — 1 min
Le FLN, un État dans l'État. #Algérie
Ouvrir

Tunisia News @tunisinews — 1 min
#Tunisie #Tunisia Tunisie - Politique : Soumaya Ghannouchi s'attaque à la France bit.ly/14Q4tKu
Ouvrir

Tunisia News @tunisinews — 1 min
Urgent : Report de la grève des transporteurs de carburants : Tunisie : Actualités : Société :... goo.gl/8wS0F #Tunisie #Tunisia
Ouvrir

SAT-7 Arabic @sat7ar — 1 min
youtu.be/qEzWEEUC07w?a via @YouTube برنامج كرمالك عراق - القداسة:
Afficher le média

koora @KooraTunisie — 2 min
Mi-temps : RM-DOR 0-0 (1-4)
Ouvrir

radioexpressfm @radioexpressfm — 3 min
L'interview de Hedi Ben Abbes dans Tac au Tac fb.me/IHx5mvzb
Ouvrir

@khawlamagrebi الاخوان عدو الانسان — 4 min
علماء: فيروس التهاب الكبد الفيروسي عمره أكثر من 82 مليون سنة fb.me/PRHIuygX #أووووووووف
Ouvrir


Accueil
Connecter
Découvrir
Moi
Rechercher
Houssem dridi
Voir ma page de profil
33
TWEETS
164
ABONNEMENTS
20
ABONNÉS
Écrire un nouveau Tweet...
Suggestions · Actualiser · Tout afficher
Mohamed Ali Chebâane @MedAl...
Suivi par Eya Turki et d'autres
Suivre
Yassine Brahim @Yassine_Brahim
Suivre
Parcourir les catégories · Trouver vos amis
Tendances · Modifier
#galaartis
#PCPointer
#ThingsNotToDoOnAFirstDate
#bcdebate2013
#cdnpoli
Oakville
CBC
Bernabeu
Madrid
Tweets
Oueld Ifriqiya @Khamousss 1 min
Le FLN, un État dans l'État. #Algérie
Ouvrir
Tunisia News @tunisinews 1 min
#Tunisie #Tunisia Tunisie - Politique : Soumaya Ghannouchi s'attaque à la France bit.ly/14Q4tKu
Ouvrir
Tunisia News @tunisinews 1 min
Urgent : Report de la grève des transporteurs de carburants : Tunisie : Actualités : Société :... goo.gl/8wS0F #Tunisie #Tunisia
Ouvrir
SAT-7 Arabic @sat7ar 1 min
youtu.be/qEzWEEUC07w?a via @YouTube :برنامج كرمالك عراق - القداسة
Afficher le média
koora @KooraTunisie 2 min
Mi-temps : RM-DOR 0-0 (1-4)
Ouvrir
radioexpressfm @radioexpressfm 3 min
L'interview de Hedi Ben Abbes dans Tac au Tac fb.me/IHx5mvzb
Ouvrir
@khawlamagrebi الاخوان عدو الانسان 4 min
علماء: فيروس التهاب الكبد الفيروسي عمره أكثر من 82 مليون سنة fb.me/PRHIuygX #أووووووووف
Ouvrir

Faculté des arts et des sciences
Département d'informatique
et de recherche opérationnelle

rali

Notre Objectif

- Détection de l'ensemble des évènements ***EV***.
- Trouver les dates saillantes.
- Traiter des *tweets* qui portent sur la Tunisie.
- Corpus :
  - Nous avons extrait 276 505 tweets, entre le 08/02/2012 et 15/04/2012 (67 jours).

Faculté des arts et des sciences
Département d'informatique
et de recherche opérationnelle

rali

Dialecte tunisien

j'ai voté, ta7ya tounes #TnElec #Vote
تصويرة بن علي رجعت في حلق الوادي Retour de Ben Ali à La Goulette http://t.co/RqVXr5Hu #tunisie #tnelec

Faculté des arts et des sciences
Département d'informatique
et de recherche opérationnelle

rali

Détection des évènements : *Twitter*

- *Twitter* constitue un excellent moyen pour diffuser des informations, pour discuter des évènements et pour donner des avis.

- Données accessibles via APIs.

- Plusieurs recherches ont montré que le contenu de *Twitter* reflète étroitement l'intérêt et les préoccupations des utilisateurs en temps réel (**Becker2011, Ozdikis2012**).

Faculté des arts et des sciences
Département d'informatique
et de recherche opérationnelle

rali

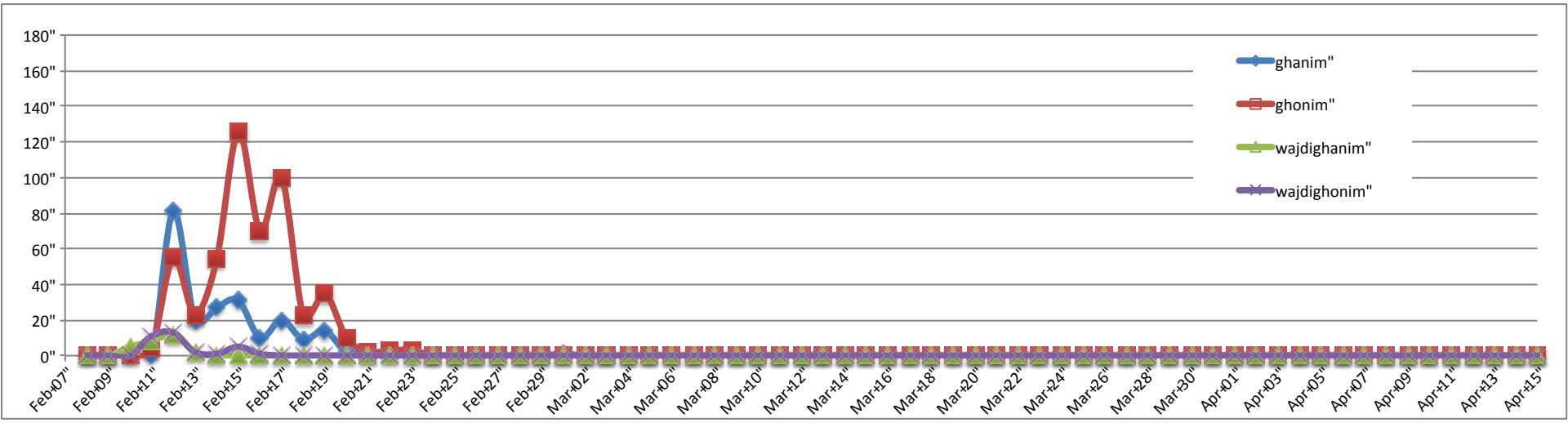
Détection des évènements

- Un évènement est représenté par un ensemble de termes.
  - Disparition de l'avion *Malaysia Airlines*: {*#PrayForMH370*, *#MH370*, *#MH370Flight*, *#MalaysiaAirlines*, etc.}.
- Objectif :
  - Regrouper automatiquement les termes représentant un même sujet.
  - Trouver l'ensemble d'évènements ***EV*** (fréquence des termes, etc.).
- Défi :
  - Supporter les conventions d'écriture, les fautes, la taille réduite d'un message, etc.

Faculté des arts et des sciences
Département d'informatique
et de recherche opérationnelle

rali

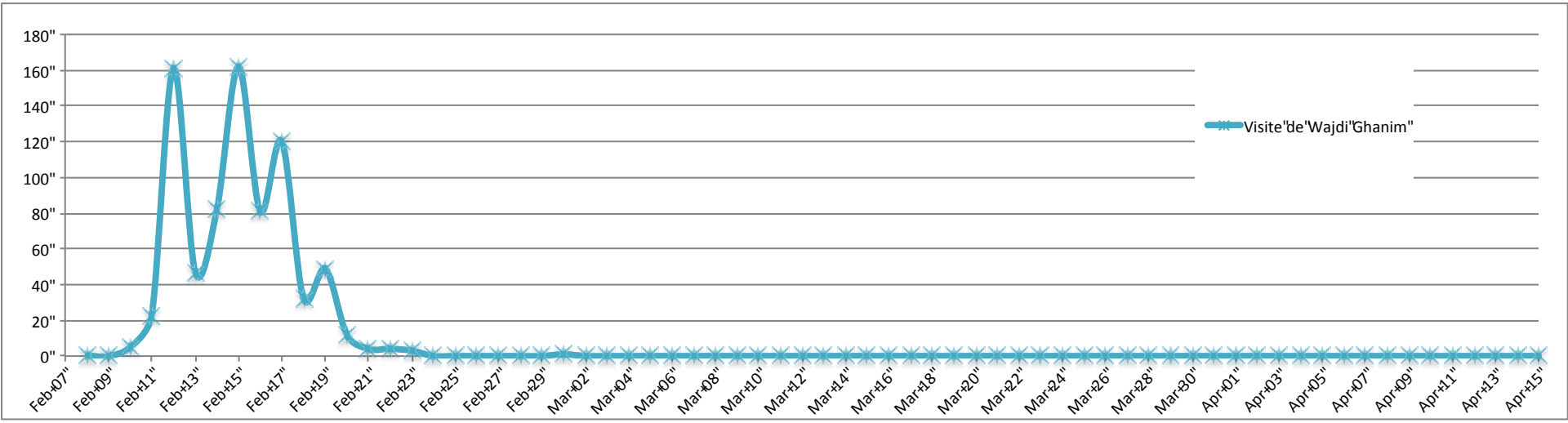
Regroupement : pourquoi ?



Faculté des arts et des sciences
Département d'informatique
et de recherche opérationnelle

rali

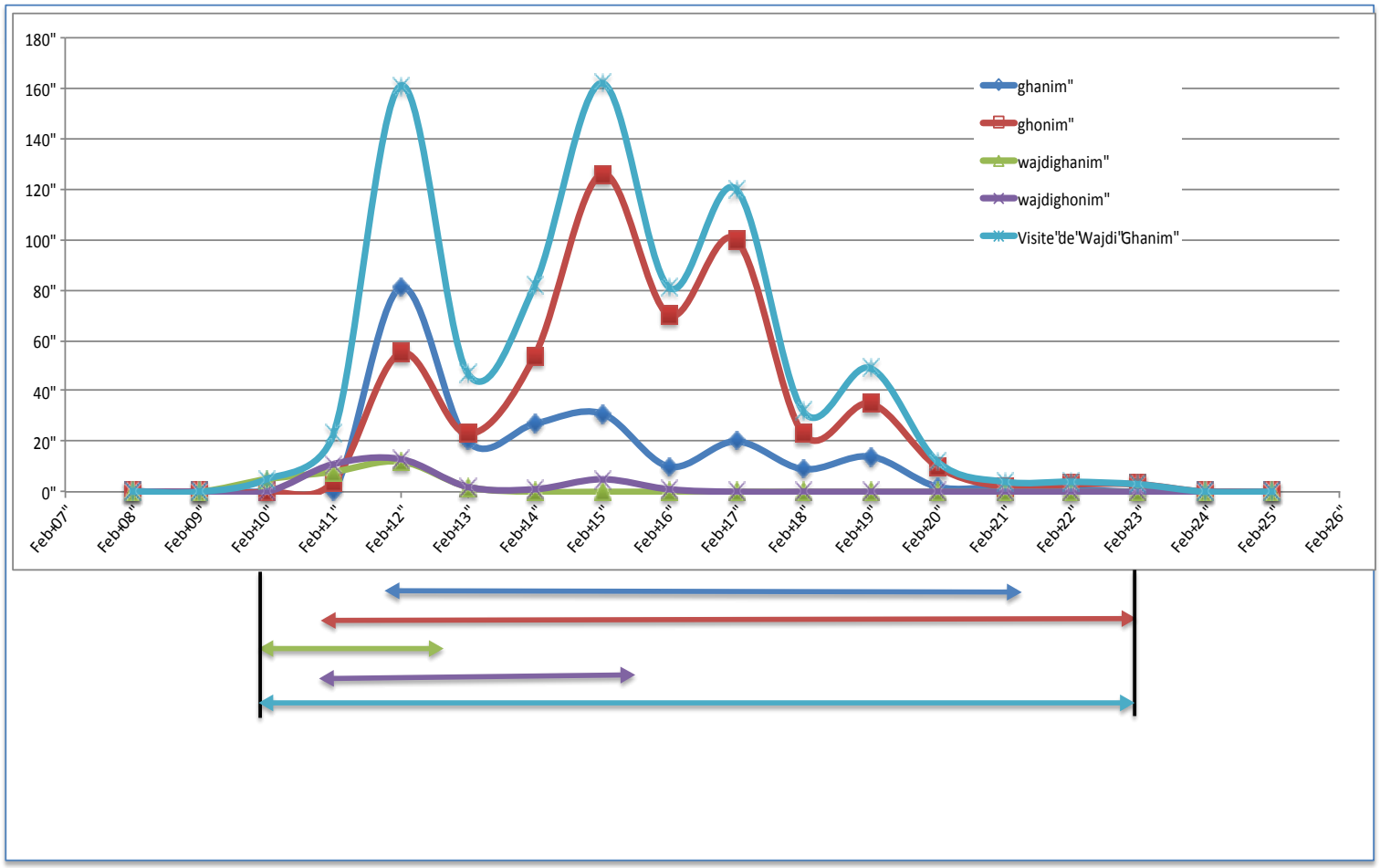
Regroupement : pourquoi ?



Faculté des arts et des sciences
Département d'informatique
et de recherche opérationnelle

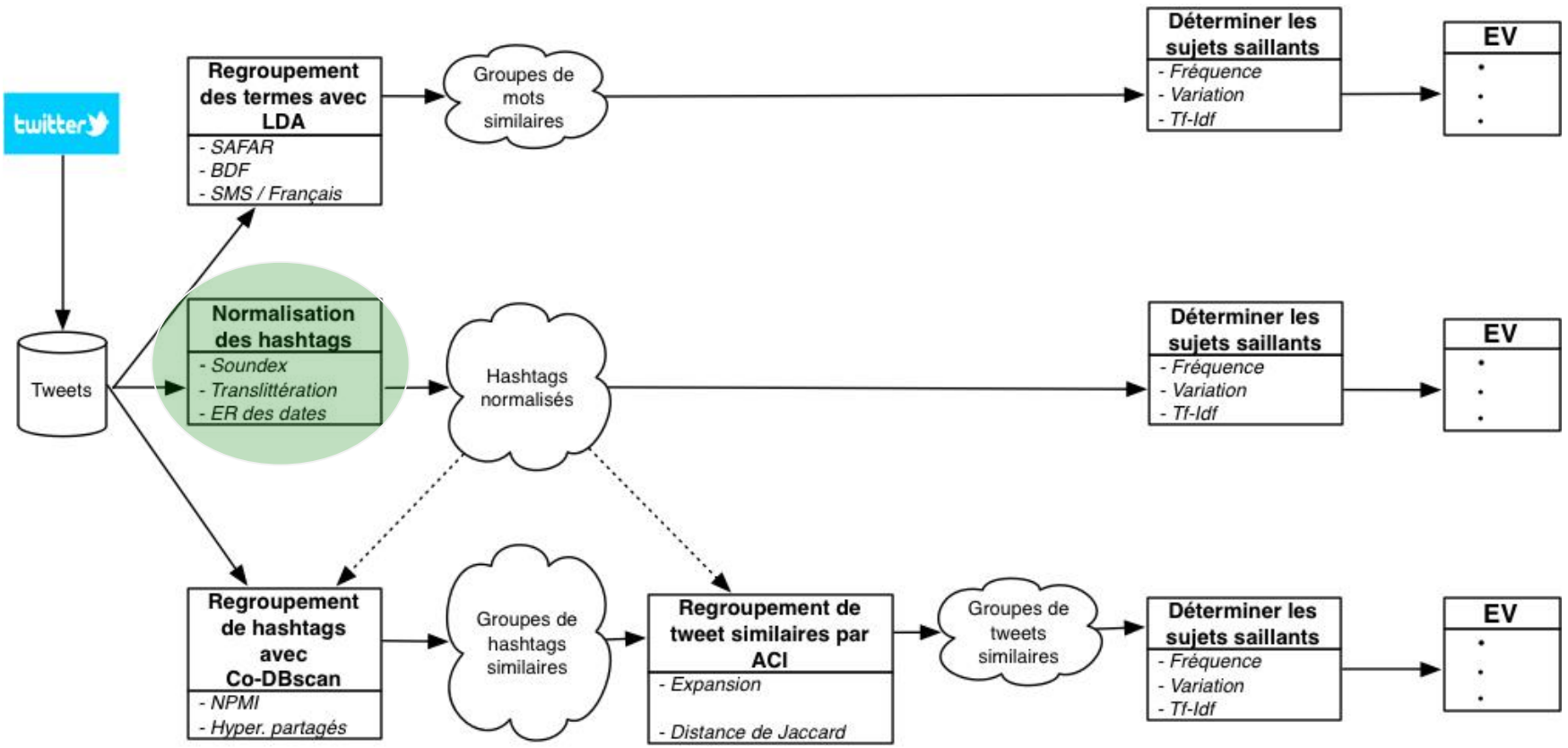
rali

Regroupement : pourquoi ?



rali

Démarche : vue globale



Faculté des arts et des sciences
Département d'informatique
et de recherche opérationnelle

rali

Normalisation des hashtags

- **Soundex** *[Russel et Odell, 1922]***:**
  - Normaliser les termes qui ont une même prononciation.
  - Code les mots qui ont la même prononciation par la même chaine de caractères.

  GHNM0 = {*#ghanim, #ghenim, #ghnaim, #ghnim, #ghoneim, #ghonem, …*}.

  S0000 = {#sousse, #suisse, #ouais, …} (Noisy Soundex)
- **Normalisation des dates :**
  - *9avril = {#9avril, #9april, #9avil, #9avirl,* #9أفريل *}*
- **Translittération** :

غنيم ⟶ GHNM0

Faculté des arts et des sciences
Département d'informatique
et de recherche opérationnelle

rali

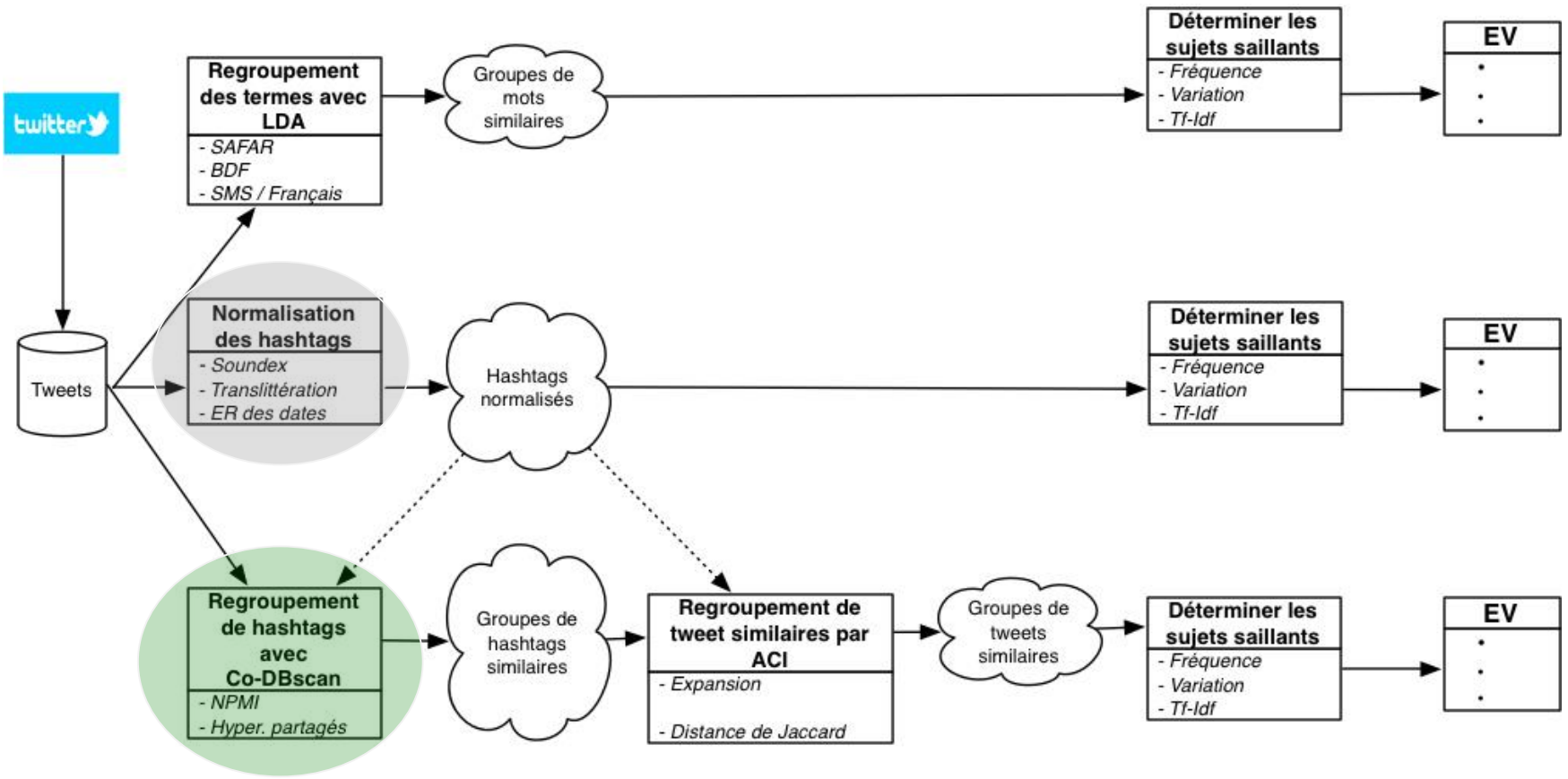
Normalisation des hashtags : Résultats

Nb. hashtags	12 218
Nb. hashtags écrits en latin	11 693
Nb. groupes Soundex initiaux	7 810
Nb. groupes de Soundex après la normalisation des dates	7 781
Nb. groupes de Soundex en disjoignant les *Noisy Soundex*	8 750
Nb. groupes de Soundex après la translittération	9 033

Précision : ≈ 96 %. (15 plus importants clusters)

rali

Démarche : vue globale

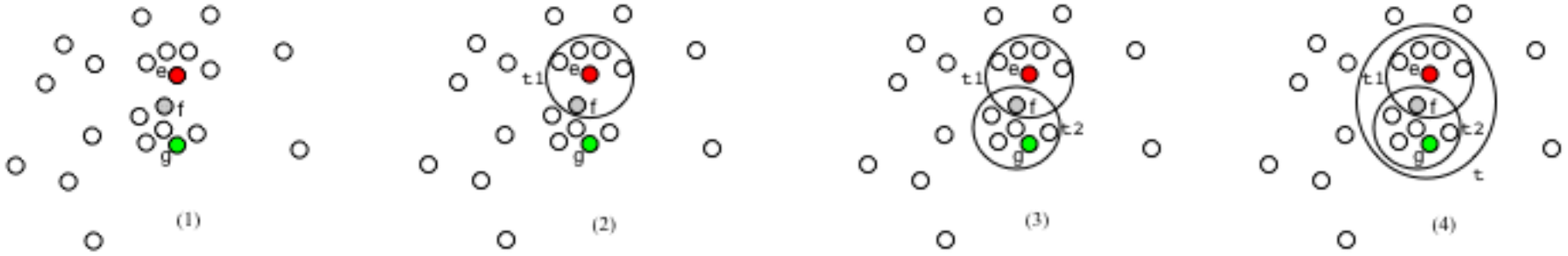


Faculté des arts et des sciences
Département d'informatique
et de recherche opérationnelle

rali

Regroupement hashtags : *DBScan* [Ester, 1996]

- Nombre de *clusters* est déterminé par l'algorithme.
- Paramètres : *MinPts*, *epsilon*.



Faculté des arts et des sciences
Département d'informatique
et de recherche opérationnelle

rali

Regroupement hashtags : *DBScan* [Ester, 1996]

- **Mesures de similarité :**
  1. ***NPMI*** (*Normalized Pointwise Mutual Information (NPMI)*) **:** Les hashtags qui apparaissent ensemble, sont sémantiquement similaires.
  2. ***Hyper*** **:** Deux hashtags apparaissant avec les mêmes hyperliens suggère que ces hashtags portent sur la même sujet.
- **Hashtags considérés :**
  - ***NPMI*** **:** 9 793 /12 218 hashtags.
  - ***Hyper*** **:** 6 008 /12 218 hashtags.

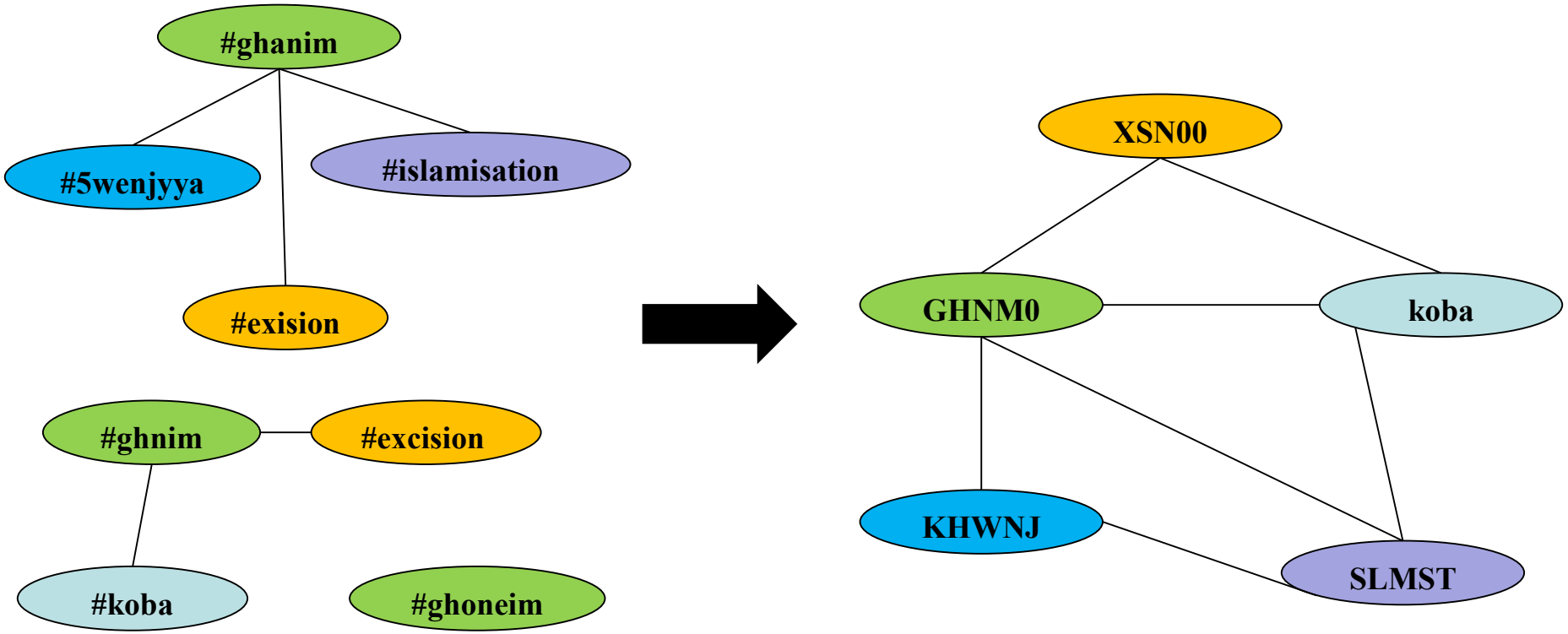
- **Problème :** création d'un immense *cluster*. **P.ex :** $Dbscan_{npmi}$ *:* 6639 hashtags dans un même cluster.

- **Solution :** vérification de similarité avant l'ajout d'un nouvel élément. (***CoDbscan***)

Faculté des arts et des sciences
Département d'informatique
et de recherche opérationnelle

rali

CoDBScan : Amélioration avec *Soundex*



Faculté des arts et des sciences
Département d'informatique
et de recherche opérationnelle

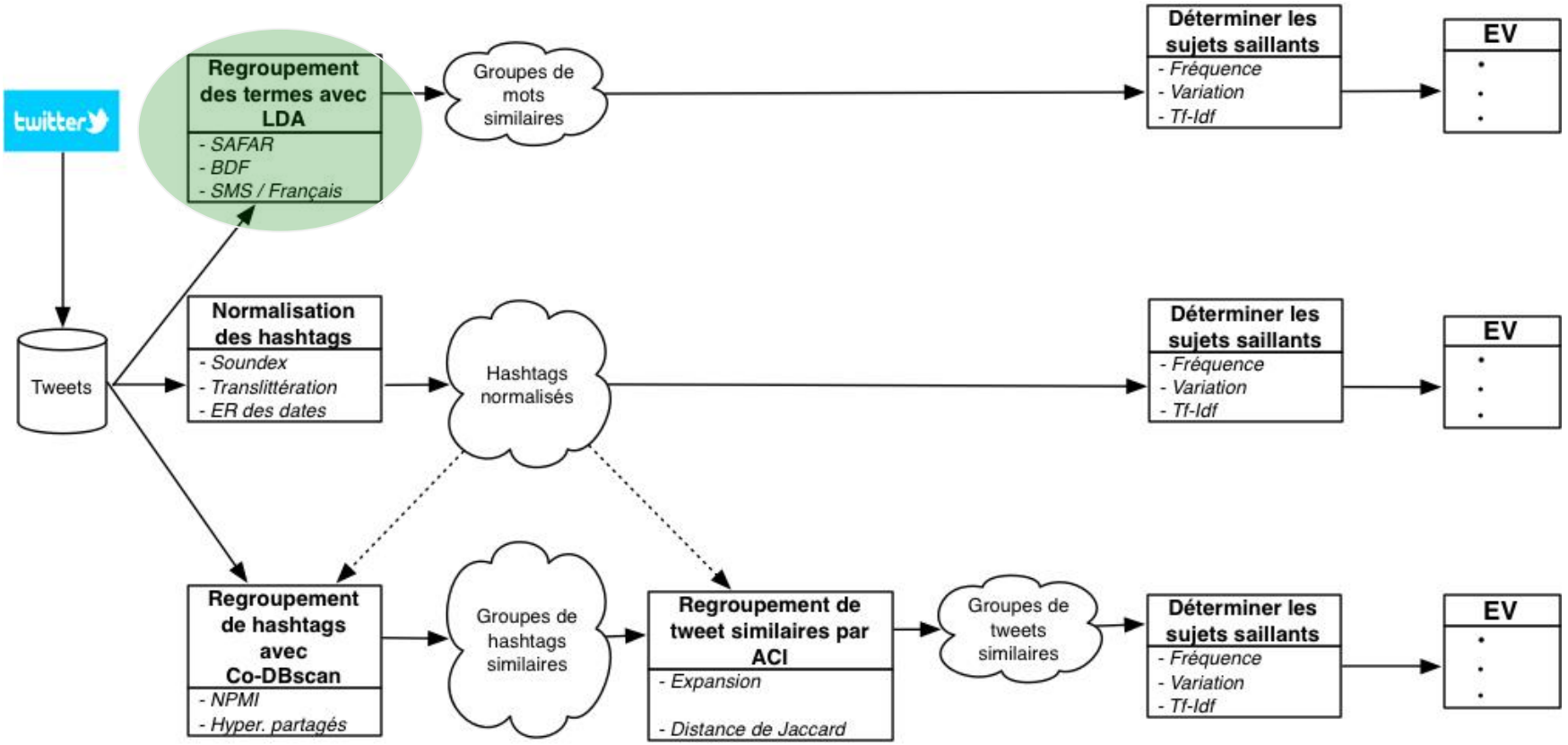
rali

CoDBScan : Validation manuelle

	NH considérés	Nb *clusters*	NH regroupés	NH 20 *clusters*	Précision (20 *clusters*)
CoDBscan $_{npmi}$ **(Ɛ=0,7)**	9 973	928	5 011	485	94 %
CoDBscan $_{npmiWithSndx}$ **(Ɛ=0,6)**	10 929	908	6 633	509	90 %
CoDBscan $_{Hyper}$ **(Ɛ=0,5)**	6 008	1 213	3 556	457	92 %
CoDBscan $_{HyperWithSndx}$ **(Ɛ=0,4)**	7 431	925	4 660	1 048	91 %

rali

Démarche : vue globale



rali

Regroupement : termes

- Utiliser des techniques de *Topic Model* (p.ex. **LDA [Blei et al. , 2003]**) pour regrouper les termes liés à un même sujet.
- Chaque document (tweet) peut être représenté comme un mélange de sujets latents, ou un sujet est lui-même représenté comme une distribution des mots qui ont tendance à co-occurrer.
- Les mots fortement liés à un sujet donné ont les valeurs de probabilité plus grandes.

Faculté des arts et des sciences
Département d'informatique
et de recherche opérationnelle

rali

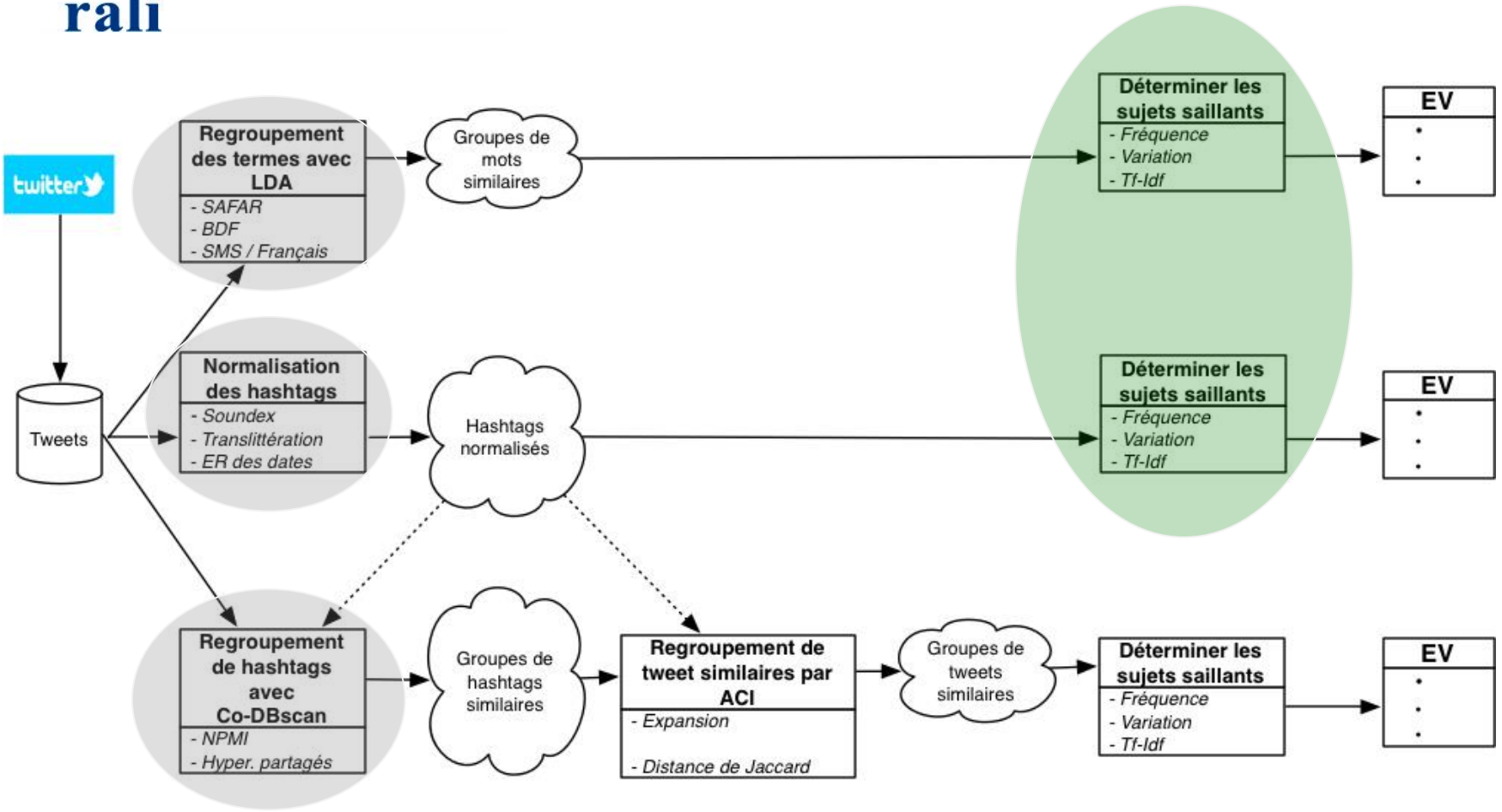
Amélioration

- Regrouper dans un même document les tweets partageant le (s) même (s) hashtag (s).

- Normaliser certains termes :
  - Lemmatiseur **BDF (Français) :** '*répond*' et '*répondu*' sont remplacés par '*répondre*'
  - Corpus parallèle **Français / SMS** : 'parce que' -> 'pcq'.
  - Librairie **SAFAR (arabe)**.

rali

Démarche : vue globale



Faculté des arts et des sciences
Département d'informatique
et de recherche opérationnelle

rali

Fréquence de sujets

- Nous avons supposé que chaque groupe réfère à un sujet.
- Un tweet porte sur **S** (représenté par un groupe **g**), s'il contient au moins un terme de **g**.
- Calculer la fréquence quotidienne de chaque sujet.

Faculté des arts et des sciences
Département d'informatique
et de recherche opérationnelle

rali

Exemple de résultats

Code Soundex	Hashtags	Dates saillantes	DF	Fréq.	Écart-Type	TF-IDF
9avril	#9avril #9april #9avil #9avirl	09/04	14	6070	516.73	4127,27
MPL00	#empl #emplo #emploi	16/03	67	3025	22.12	0
RCRTM	#recrutement	16/03	67	2854	22.63	0
WTHR0	#weather	15/02, 16/02	65	2043	10.25	68,81

Faculté des arts et des sciences
Département d'informatique
et de recherche opérationnelle

rali

Exemple de résultats

Code Soundex	Hashtags	Dates saillantes	DF	Fréq.	Écart-Type	TF-IDF
9avril	#9avril #9april #9avil #9avirl	09/04	14	6070	516.73	4127,27
20mars	#20mars, #20مارس	20/03	15	1422	124,59	924,27
ugtt	#ugtt	25/02	50	1797	97,76	182,93
NHD00	#enahda #enanhda #enhada	21/02	66	1872	66,41	12,23

Faculté des arts et des sciences
Département d'informatique
et de recherche opérationnelle

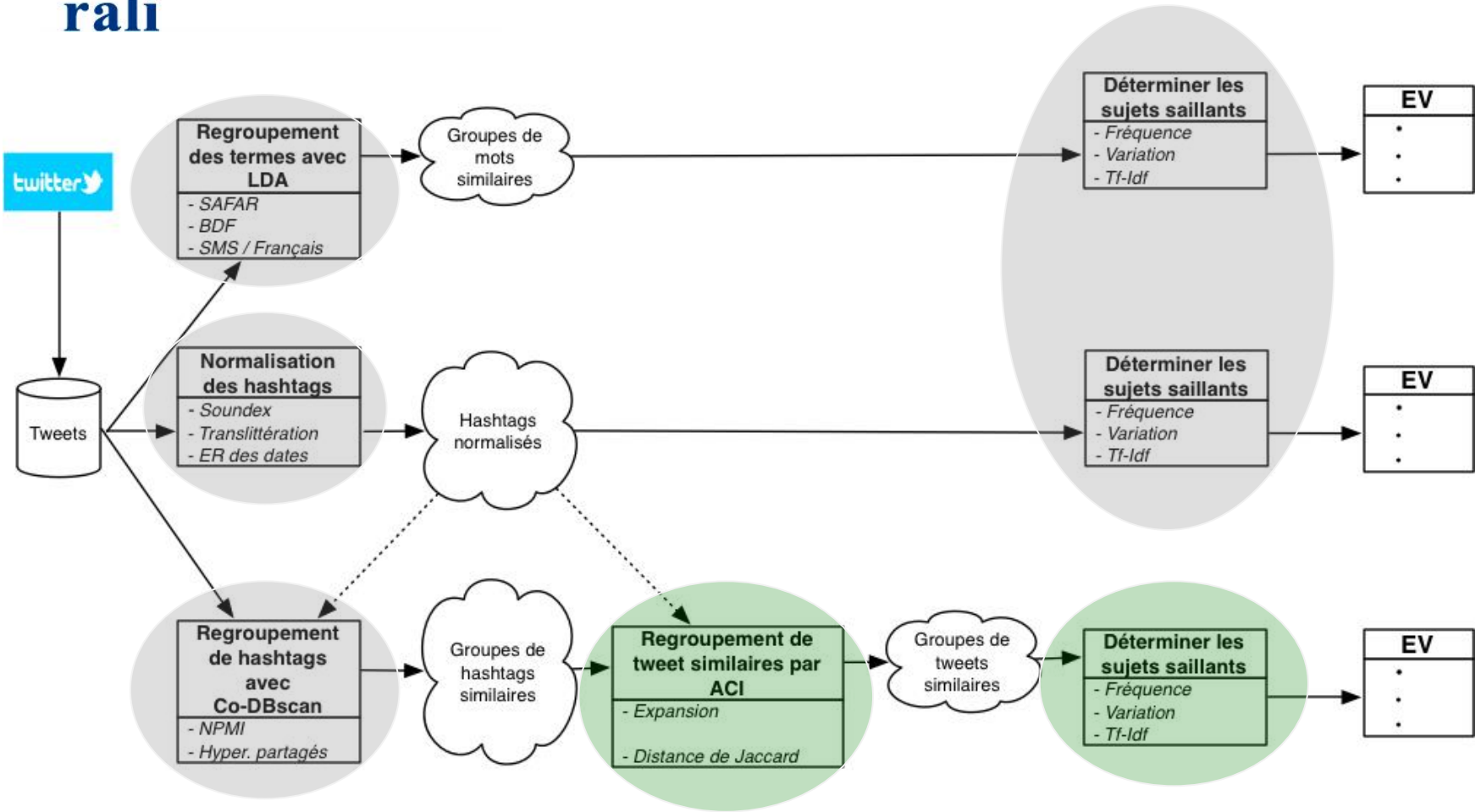
rali

Exemple de résultats

Code Soundex	Hashtags	Dates saillantes	DF	Fréq.	Écart-Type	TF-IDF
9avril	#9avril #9april #9avil #9avirl	09/04	14	6070	516.73	4127,27
TNPHR	#tunpharma	04/04 04/05	15	1710	52,79	1111,47
20mars	#20mars, #20مارس	20/03	15	1422	124,59	924,27
3PLKT	#application	04/04 04/05	15	1355	41,89	880,72

rali

Démarche : vue globale



Faculté des arts et des sciences
Département d'informatique
et de recherche opérationnelle

rali

Regroupement de tweets : *expansion*

- Étendre les tweets par d'autres hashtags reliés afin d'améliorer le regroupement.
- Impact :

	Nb. Tweets non regroupés
Sans expansion	4 296
Expansion $\boldsymbol{CoDBscan}_{npmiWithSoundex}$	3 077
Expansion $\boldsymbol{CoDBscan}_{npmiWithSoundex}$	3 305

Faculté des arts et des sciences
Département d'informatique
et de recherche opérationnelle

rali

Déterminer les évènements : *dates saillantes*

- Méthode de **Palshikar (2009)** :

08-02	09-02	10-02	11-02	12-02	13-02	14-02	15-02	16-02	17-02	18-02	19-02	20-02	21-02	22-02	23-02	24-02
0	0	1	33	190	74	102	176	118	155	40	55	14	6	4	4	1

- k=2

$$S_i = \frac{\max_{1 \le j \le k}(x_i - x_{j-1}) + \max_{1 \le j \le k}(x_i - x_{i+j})}{2}$$

08-02	09-02	10-02	11-02	12-02	13-02	14-02	15-02	16-02	17-02	18-02	19-02	20-02	21-02	22-02	23-02	24-02
-	-	-15,5	-4	152,5	6,5	6	80	47	76	-26	32	-8	-3	0,5	-	-

$\mathbf{x_i}$ est un pic si : $\mathbf{S_i}$ > *moyenne* (*S*>0) **ET** ($\boldsymbol{S_i}$ - *moyenne*) > *écart-type* (*S*>0)

Faculté des arts et des sciences
Département d'informatique
et de recherche opérationnelle

rali

Déterminer les évènements : *EV*

- Sélectionner que les sujets saillants : ***EV*** !
- Adapter la méthode **Palshikar (2009).**
  - $\boldsymbol{k}$ = nombre de sujets,
  - $\boldsymbol{x_i}$: valeur du critère (*fréquence*, *écart-type*, *Tf-Idf*) utilisé.

rali

Validation

- L'évaluation de l'exactitude de la méthode est une tâche difficile : **pas de données de référence**.
- Deux méthodes utilisées :
    1. Vérification de l'importance et de la date de chaque évènement auprès d'un ensemble des medias traditionnels (journaux numériques) fiables.
    2. Recours à des experts, dans notre cas des Tunisiens au courant des évènements qui se sont déroulés en Tunisie, afin de distinguer les évènements importants parmi ceux détectés par nos méthodes.

Faculté des arts et des sciences
Département d'informatique
et de recherche opérationnelle

rali

Validation

http://rali.iro.umontreal.ca:8080/dridihou/

analyse des tweets

home | expériences | annotation | contact

Annotation des événements

Num	Termes	Date (2012)	Jugement		
1	9avril (9avril 9april 9avil 9avirl 9avr)	(04/09)	Évènement	pas-Évènement	pas clair
2	20mars (20mars 20mars1956 20mars 20mars1956 20mars2012)	(03/20)	Évènement	pas-Évènement	pas clair
3	ugtt (ugtt)	(02/25)	Évènement	pas-Évènement	pas clair
4	NHD00 (enahda enanhda enhada ennahada ennahd)	(02/21)	Évènement	pas-Évènement	pas clair
5	3KB00 (acab)	(04/09)	Évènement	pas-Évènement	pas clair
6	MNB00 (mannoub mannouba manouba mnaouba منوبة)	(03/07)	Évènement	pas-Évènement	pas clair
7	TNPHR (tunpharma)	(04/04 04/05)	Évènement	pas-Évènement	pas clair
8	PHRMC (pharmacie pharmaciecentrale pharmacies)	(04/04 04/05)	Évènement	pas-Évènement	pas clair
9	3LGRY (algeria algerian algria)	(02/16 02/21)	Évènement	pas-Évènement	pas clair
10	GYPT0 (egyipt egypet egypt egypte égypte)	(03/02 04/09)	Évènement	pas-Évènement	pas clair
11	BHRN0 (baharin bahrain bahraïn bahrein bahreïn)	(02/14 02/15 03/02)	Évènement	pas-Évènement	pas clair

Faculté des arts et des sciences
Département d'informatique
et de recherche opérationnelle

rali

Évaluation

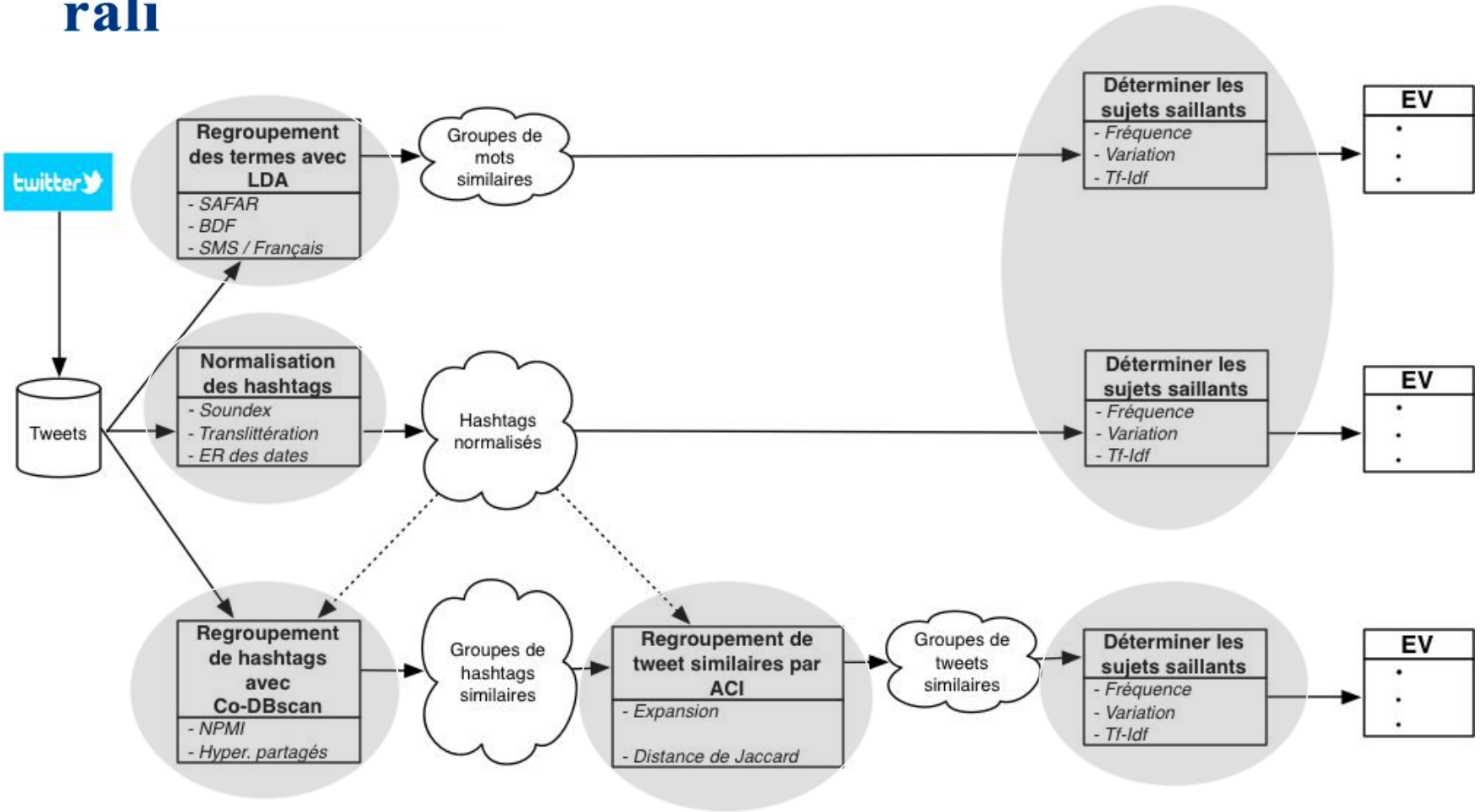
- Annotation par 10 experts.

	Fréquence		***variation***		***Tf-Idf***	
	|EV|	**Pr.**	***|EV|***	**Pr.**	***|EV|***	**Pr.**
Groupe de hashtags (normalisation)	123	0,64	88	0,82	81	**0,95**
Groupe termes (*LDA*)	53	0,73	67	0,87	74	**0,92**
Groupe de tweets + expansion *CoDBscan*$_{npmiWithSoundex}$	98	0,61	52	0,8	46	**0,93**
Groupe de tweets + expansion *CoDBscan*$_{HyperWithSoundex}$	104	0,54	51	0,69	46	**0,94**

Faculté des arts et des sciences
Département d'informatique
et de recherche opérationnelle

rali

Démarche : vue globale



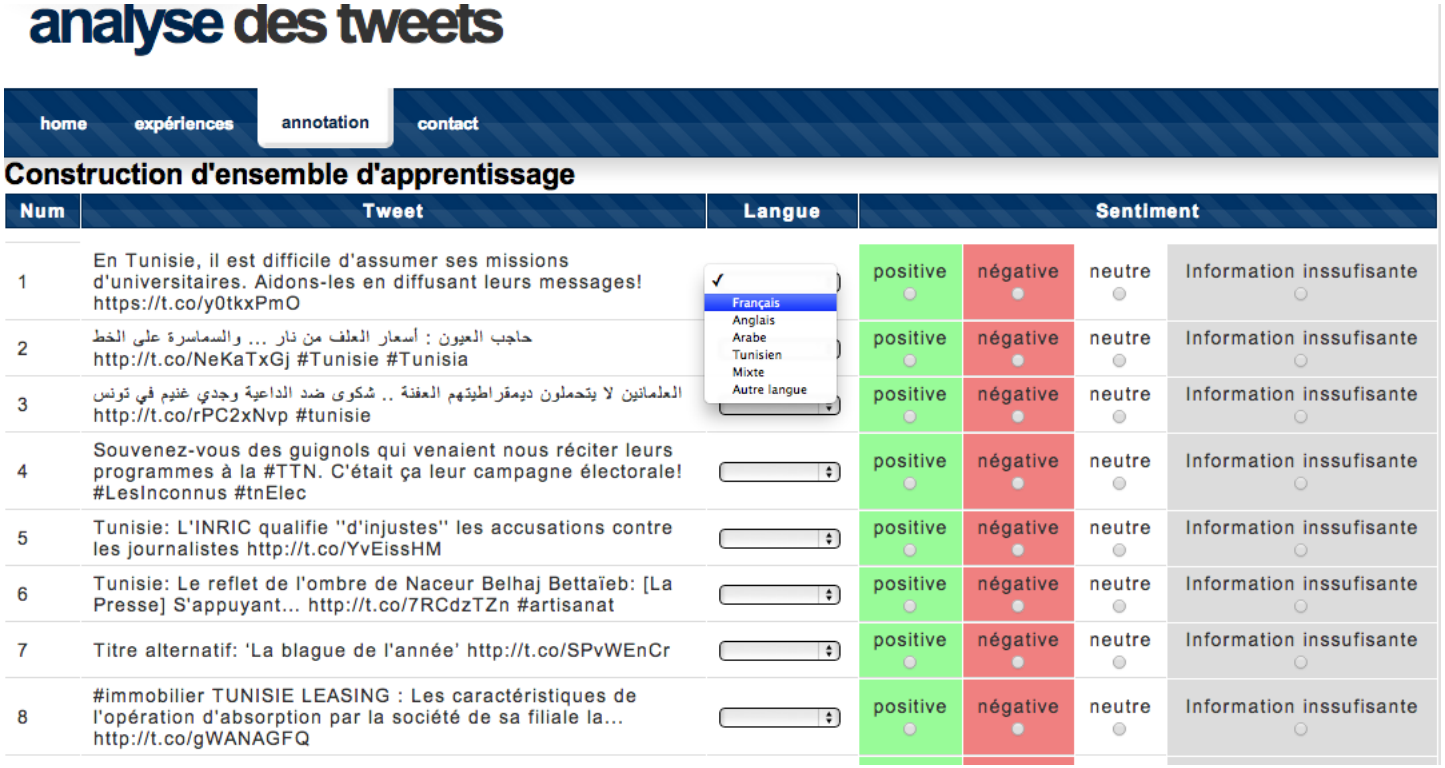
Travaux futurs (1)

- Utiliser d'autres éléments pour identifier ***EV***. P.ex : hyperliens.
- Tester nos méthodes sur d'autres corpus :
  - Environ 600 000 tweets portant sur la Tunisie (octobre 2012 – janvier 2013).
  - Plus que 5 millions de tweets envoyés à partir du Québec. (février 2014 – août 2014).
- Déterminer l'opinion publique pour un évènement :
  - Déterminer la proportion de chaque polarité (positive, négative, neutre) à partir des tweets assignés à l'évènement.
  - Pas de ressources disponibles!

rali

Travaux futurs (2)

- Environ 4 000 tweets annotés.



Faculté des arts et des sciences
Département d'informatique
et de recherche opérationnelle

rali

Conclusion

- Les hashtags sont des indices pour déterminer les évènements.
- La fréquence ne reflète pas l'importance d'un sujet.
- Tâches de validation est coûteuse.
- Notre système : haute précision.

Faculté des arts et des sciences
Département d'informatique
et de recherche opérationnelle

rali

@toutLeMonde: Mrci pr vtr atention. #soutenance #thèse #Houssem