Analyse des données de microblogs

Houssem Eddine Dridi

Partie orale de l'examen pré-doctoral

Juillet, 2012

Introduction

- L'analyse de ces informations est une tâche indispensable.

Contexte

- Collaboration avec `Rali` et `MediaBadger` (Halifax).
- `MediaBadger` voudrait développer un système de détection de préoccupations des internautes et d'analyse d'opinions à partir de données textuelles dans le web.
- En 2010, `MediaBadger` a développé un système qui fonctionnait sur des données en anglais.

Objectifs ?

- Extraire automatiquement des données à partir des microblogs.
- Détecter les événements majeurs et les préoccupations des utilisateurs.
- Déterminer l'opinion publique sur chaque événement.
- Traiter des messages écrits dans plus d'une langue. Principalement en arabe et en français.

Microblogs : présentation

- Plates-formes permettant aux utilisateurs de publier des messages courts (entre 140 et 200 caractères au maximum) sans titre.
- Un message peut contenir des données multimédias (vidéos, images. . .)
- Au départ : *qu'est-ce qu'on est en train de faire?*
- Actuellement : les internautes profitent de ce service pour exprimer leurs opinions sur différents sujets, diffuser des informations et faire des discussions.
- Les microblogs sont devenus d'excellents outils pour des entreprises pour faire des publicités sur leurs produits et services ou pour les célébrités pour communiquer avec leurs fans.

Microblogs : écritures

- Messages sont généralement très courts.
- Parfois incompréhensibles par les non-initiés.
- Les utilisateurs commettent fréquemment des fautes d'orthographe et de grammaire.
- Utilisation des abréviations (*pcq* = *parce que*) et des mots étirés (*merciiii* = *merci*).
- Utilisation d'onomatopées (*ha ha* = *rire*).

Twitter : présentation

- Plate-forme de microbloggage la plus populaire.
- Plus de 475 millions utilisateurs inscrits, 175 millions de messages (*tweets*) envoyés chaque jour.
- Slogan actuel : *Découvrez ce qui se passe, en ce moment, avec les personnes et les organisations qui vous tiennent à cœur*.

Twitter : conventions d'écriture

- Nombre de caractères autorisés dans un tweet est 140.
- Le nom d'utilisateur doit être précédé toujours par *@* (*@username*).
- On peut étiqueter les sujets d'un *tweet* par un hashtag *#*. Exemple : *Les manifestants se sont dispersés. #manifencours #GGI*
- Retweeter : renvoyer un tweet sans changer son contenu. Un retweet commence par *RT @Y* où *Y* est le titulaire du tweet.
- Utilisation de service de réduction d'URL (*t.co*) : rend la page accessible par l'intermédiaire d'une très courte URL. *http://www.iro.umontreal.ca/rubrique.php3?id_rubrique=13* $\Leftrightarrow$ *http://t.co/NrUGjAtx*

Twitter : recherches

- Nombre important d'utilisateurs inscrits.
- Toutes catégories d'âge.
- Enorme quantité de données.
- Les données sont par défaut publiques.

État de l'art

- **Classification de sentiments**

  (Go et al. 2009), (Barbosa et Feng 2010) techniques TALN.
  (Jiang et al. 2011) techniques TALN + relation entre les tweets.
  (Tan et al. 2011) relation entre les utilisateurs.

- **Événements vs. tweets**

  (Doan et al. 2011) tremblements de terre au Japon en 2011.
  (Lampos et Cristianini 2010) prévalence de la maladie H1N1 au Grande-Bretagne.
  (Metzler et al. 2012) résumé d'un événement particulier.
  (OConnor et al. 2010) corrélation entre les tweets et les sondages.

Tweets des tunisiens

- Tweet peut être écrit avec plus d'une langue.
- Les utilisateurs peuvent écrire un mot arabe avec des alphabets français et des chiffres qui donne la même prononciation du mot arabe.

Pourquoi ce genre de tweets ?

- Nos compétences qui nous permettent de comprendre le français et l'arabe.
- La façon d'écriture employée par les tunisiens.
- Nous connaissons les événements qui se déroulent en Tunisie.

Tweets des tunisiens : exemples

Tweet	**Explication**
j'ai voté, ta7ya tounes #TnElec #Vote	L'auteur a informé qu'il a déjà voté. Les mots *ta7ya* et *tounes*, dans le deuxième tweet, sont des néographes des mots arabes qui correspondent respectivement à *vive* et *la Tunisie*.
تصويرة بن علي رجعت في حلق الوادي Retour de Ben Ali à La Goulette http://t.co/RqVXr5Hu #tunisie #tnelec	Le texte arabe signifie que la photo Le texte arabe signifie que la photo de *Ben Ali* (président tunisien déchu) a été republiée dans *La Goulette* (ville tunisienne)

Extraction des données

- Nous avons implémenté un programme Java.
- Nous avons utilisé la bibliothèque `Twitter4j`[1] pour accéder à `Twitter API`.

Deux types d'API

- `Search API`[2].
- `Streaming API`[3].

[1] `http://twitter4j.org/en/index.html`
[2] `https://dev.twitter.com/docs/api/1/get/search`
[3] `https://dev.twitter.com/docs/streaming-api`

Search API

- Retourne les tweets qui répondent à une requête.
- On peut filtrer les résultats selon plusieurs critères :

  Langue des tweets spécifier la langue avec laquelle le tweet est écrit.

  La période trouver les tweets écrits entre deux dates.

  Type de résultats spécifier le type de tweets retournés les plus populaires (les plus retweetés), les plus récents ou mixtes (mélange entre les plus populaires et les plus récents).

Search API

- Retourne les tweets qui répondent à une requête.
- On peut filtrer les résultats selon plusieurs critères :

  Langue des tweets spécifier la langue avec laquelle le tweet est écrit.

  La période trouver les tweets écrits entre deux dates.

  Type de résultats spécifier le type de tweets retournés les plus populaires (les plus retweetés), les plus récents ou mixtes (mélange entre les plus populaires et les plus récents).

- Le nombre de tweets retournés par requête ne peut pas dépasser 1500.
- On ne peut pas trouver les tweets qui étaient envoyés il y a plus qu'une semaine.

Expériences 1 et 2

Expérience 1

- Déterminer la polarité des tweets en anglais.

Expérience 2

- Nous avons collecté des tweets qui portent sur les élections tunisiennes d'octobre 2011.
- Trois langues : arabe, anglais, français.
- Statistiques sur le nombre de mots trouvés dans les tweets.
- Statistiques sur le nombre de mots trouvés dans les vocabulaires.

Streaming API

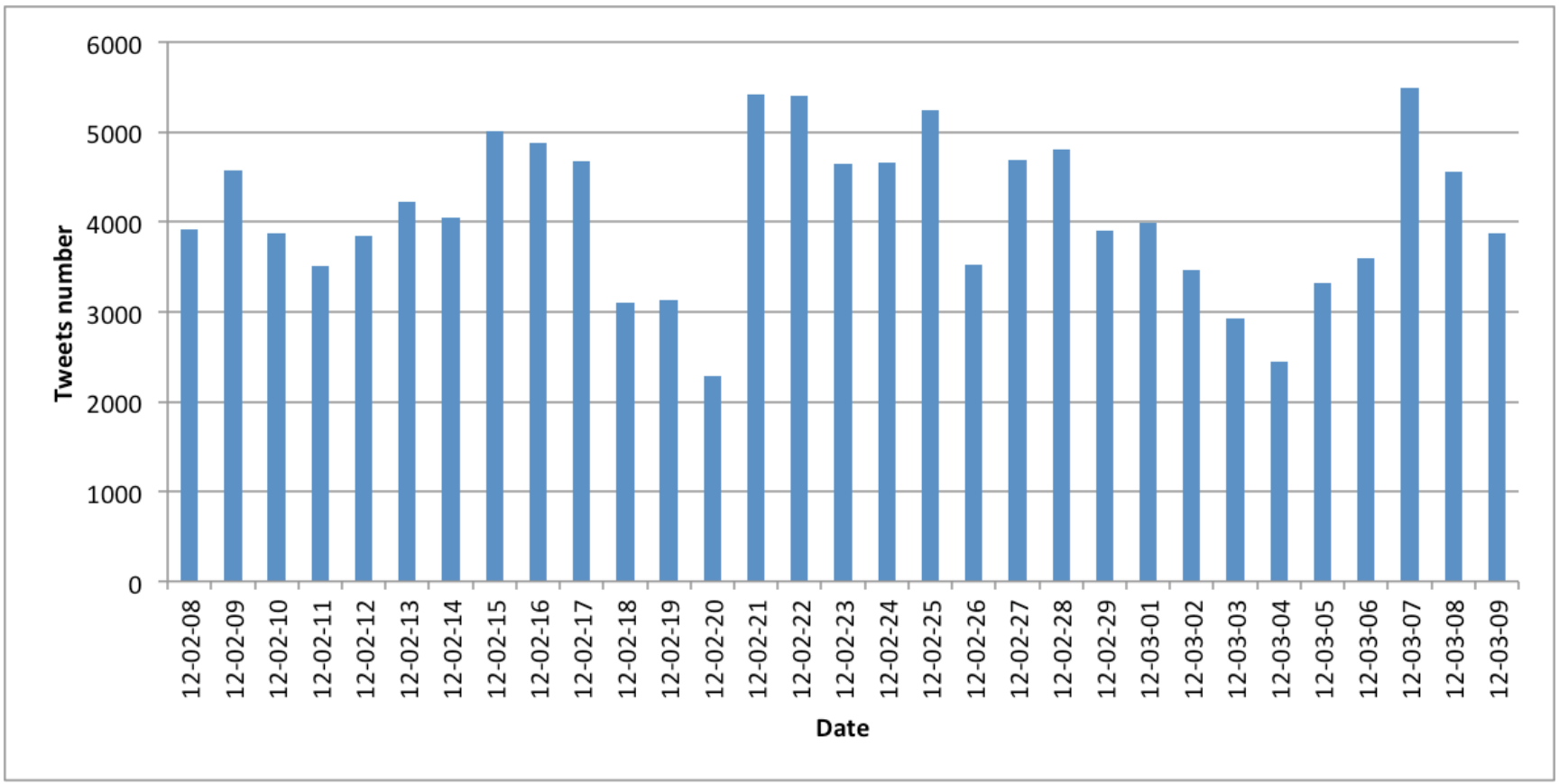
- Obtenir des tweets en temps réel.
- On peut filtrer les tweets avec plusieurs (jusqu'à 400) mots-clés.
- On peut également filtrer les tweets selon leur positionnement géographique.

Expérience 3

- Nous avons extrait des tweets qui portent sur la `Tunisie` (`Streaming API`).
- Nous avons utilisé un ensemble de mots-clés fortement liés à la Tunisie e.g. *Tunisie*, *Tunisia*, *tounes*, *Marzouki* (président actuel), *Ennahdha* (parti au pouvoir)...
- 126 991 tweets entre le 08 février 2012 et le 09 mars 2012.

Expérience 3 : statistiques (1)

Figure: Nombre de tweets par jour sur la `Tunisie` (126 991 tweets entre le 08 février 2012 et le 09 mars 2012). Moyenne = 4096/jour; écart type = 854.



Expérience 3 : statistiques (2)

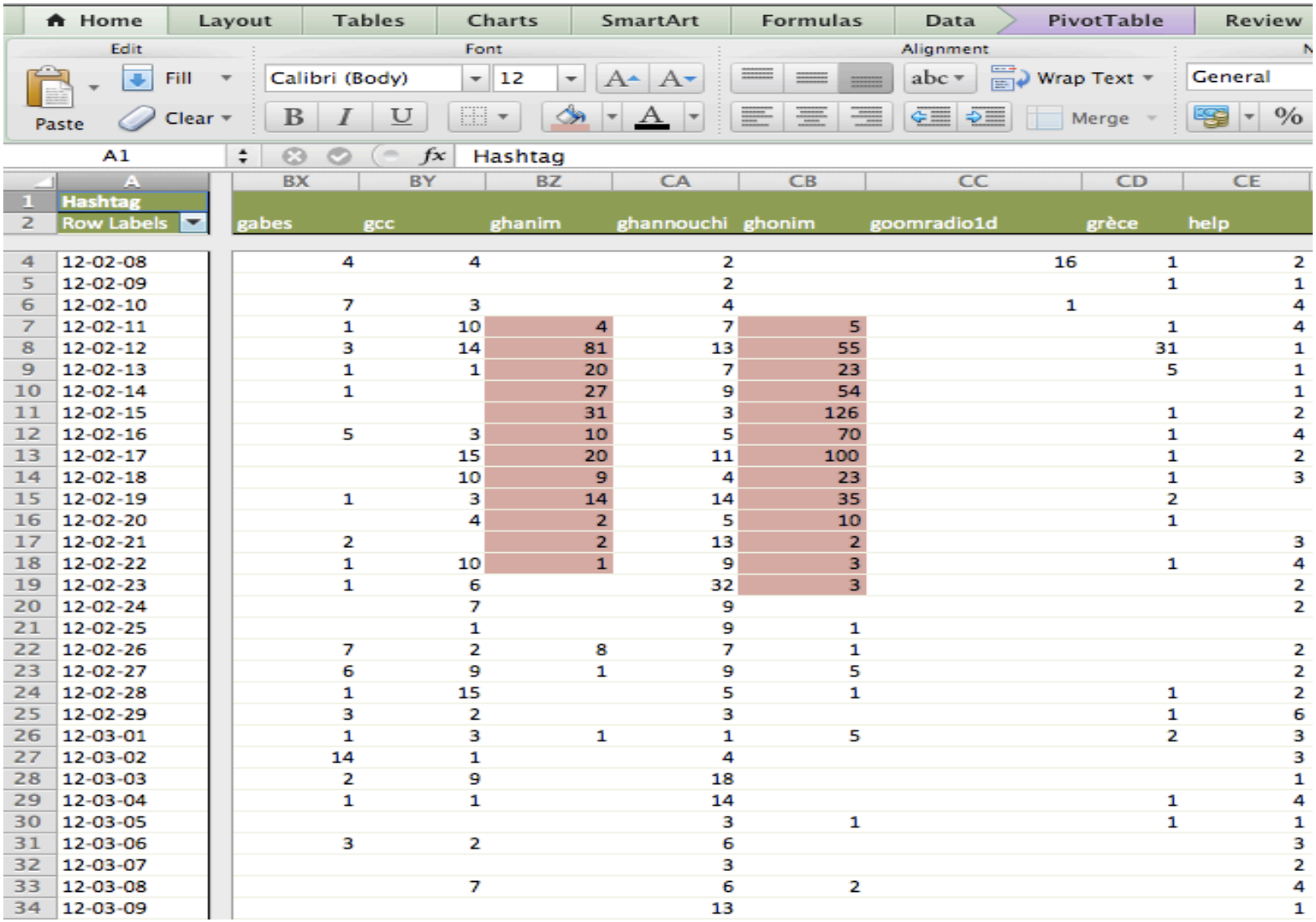
Table: Statistiques sur les tweets qui portent sur la Tunisie publiés entre le 08 février 2012 et le 09 mars 2012).

Numbre de tweets	126 991
Nombre des retweets	23 283
Nombre d'utilisateurs distincts	16 071
Nombre de tweets qui contiennent au moins un hashtag	72 564
Nombre de tweets qui contiennent au moins un utilisateur mentionné	48 750
Nombre de tweets qui contiennent au moins un hyperlien	79 250
Nombre de hashtags distincts	7 879

Hashtags vs. Événements (1)

- Nous avons visualisé les hashtags présents dans le corpus dans d'un *tableau croisé dynamique* d'*Excel*.
- Certains hashtags apparaissent soudainement dans une période déterminée.

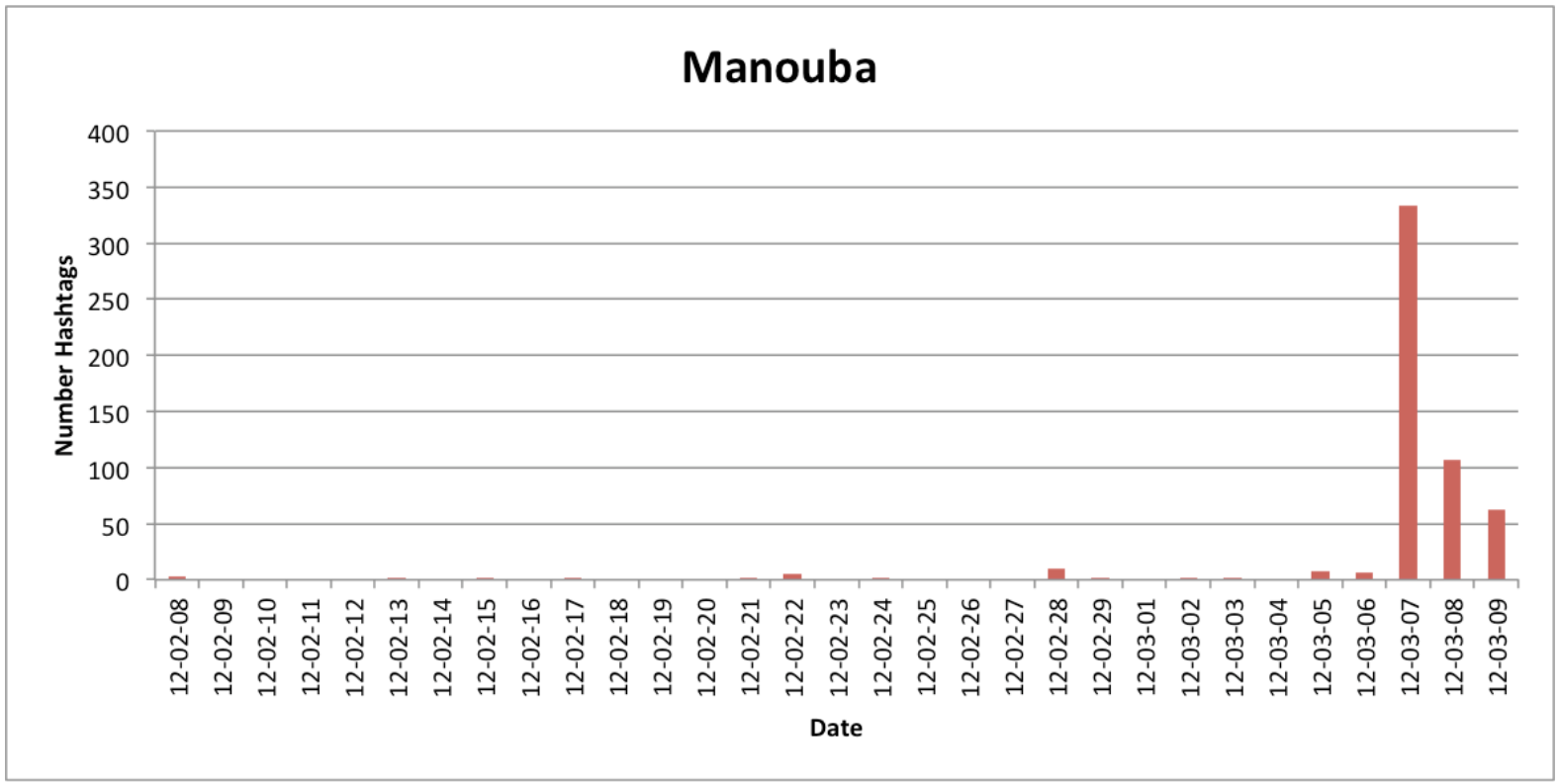
Hashtags vs. Événements (2)



	A	BX	BY	BZ	CA	CB	CC	CD	CE
1	Hashtag								
2	Row Labels	gabes	gcc	ghanim	ghannouchi	ghonim	goomradio1d	grèce	help
4	12-02-08	4	4		2			16	1	2
5	12-02-09				2				1	1
6	12-02-10	7	3		4			1		4
7	12-02-11	1	10	4	7	5			1	4
8	12-02-12	3	14	81	13	55			31	1
9	12-02-13	1	1	20	7	23			5	1
10	12-02-14	1		27	9	54				1
11	12-02-15			31	3	126			1	2
12	12-02-16	5	3	10	5	70			1	4
13	12-02-17		15	20	11	100			1	2
14	12-02-18		10	9	4	23			1	3
15	12-02-19	1	3	14	14	35			2	
16	12-02-20		4	2	5	10			1	
17	12-02-21	2		2	13	2				3
18	12-02-22	1	10	1	9	3			1	4
19	12-02-23	1	6		32	3				2
20	12-02-24		7		9					2
21	12-02-25		1		9	1				
22	12-02-26	7	2	8	7	1				2
23	12-02-27	6	9	1	9	5				2
24	12-02-28	1	15		5	1			1	2
25	12-02-29	3	2		3				1	6
26	12-03-01	1	3	1	1	5			2	3
27	12-03-02	14	1		4					3
28	12-03-03	2	9		18					1
29	12-03-04	1	1		14				1	4
30	12-03-05				3	1			1	1
31	12-03-06	3	2		6					3
32	12-03-07				3					2
33	12-03-08		7		6	2				4
34	12-03-09				13					1

Hashtags : événement *Manouba*

Figure: Distribution par jour de 2 hashtags (*manouba*, *mannouba*).**07 mars** : La mise en berne du drapeau tunisien par un étudiant salafiste.



Événement : détection de pics

- Méthode de [Palshikar G.K, 2009].

02-08	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
0	0	1	33	190	74	102	176	118	155	40	55	14	6	4	4	1

$$S(k,i,x_i,T) = \frac{max(x_i - x_{i-1}; x_i - x_{i-2} \ldots; x_i - x_{i-k}) + max(x_i - x_{i+1}; x_i - x_{i+2} \ldots; x_i - x_{i+k})}{2}$$

k=2

02-08	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
-	-	-15.5	-4	152.5	6.5	6	80	47	76	-26	32	-8	-3	0.5	-	-

x_i est un pic si :

$S_i >$ **moyenne** (valeurs positives de S) et **(S_i -moyenne)** $>$ **h * écart type** (valeurs positives de S)

Événement : détection de pics

- Méthode de [Palshikar G.K, 2009].

02-08	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
0	0	1	33	190	74	102	176	118	155	40	55	14	6	4	4	1

$$S(k,i,x_i,T) = \frac{max(x_i - x_{i-1}; x_i - x_{i-2} \ldots; x_i - x_{i-k}) + max(x_i - x_{i+1}; x_i - x_{i+2} \ldots; x_i - x_{i+k})}{2}$$

k=2

02-08	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
-	-	-15.5	-4	152.5	6.5	6	80	47	76	-26	32	-8	-3	0.5	-	-

x_i est un pic si :

$S_i >$ **moyenne** (valeurs positives de S) et **(S_i -moyenne)** $>$ **h * écart type** (valeurs positives de S)

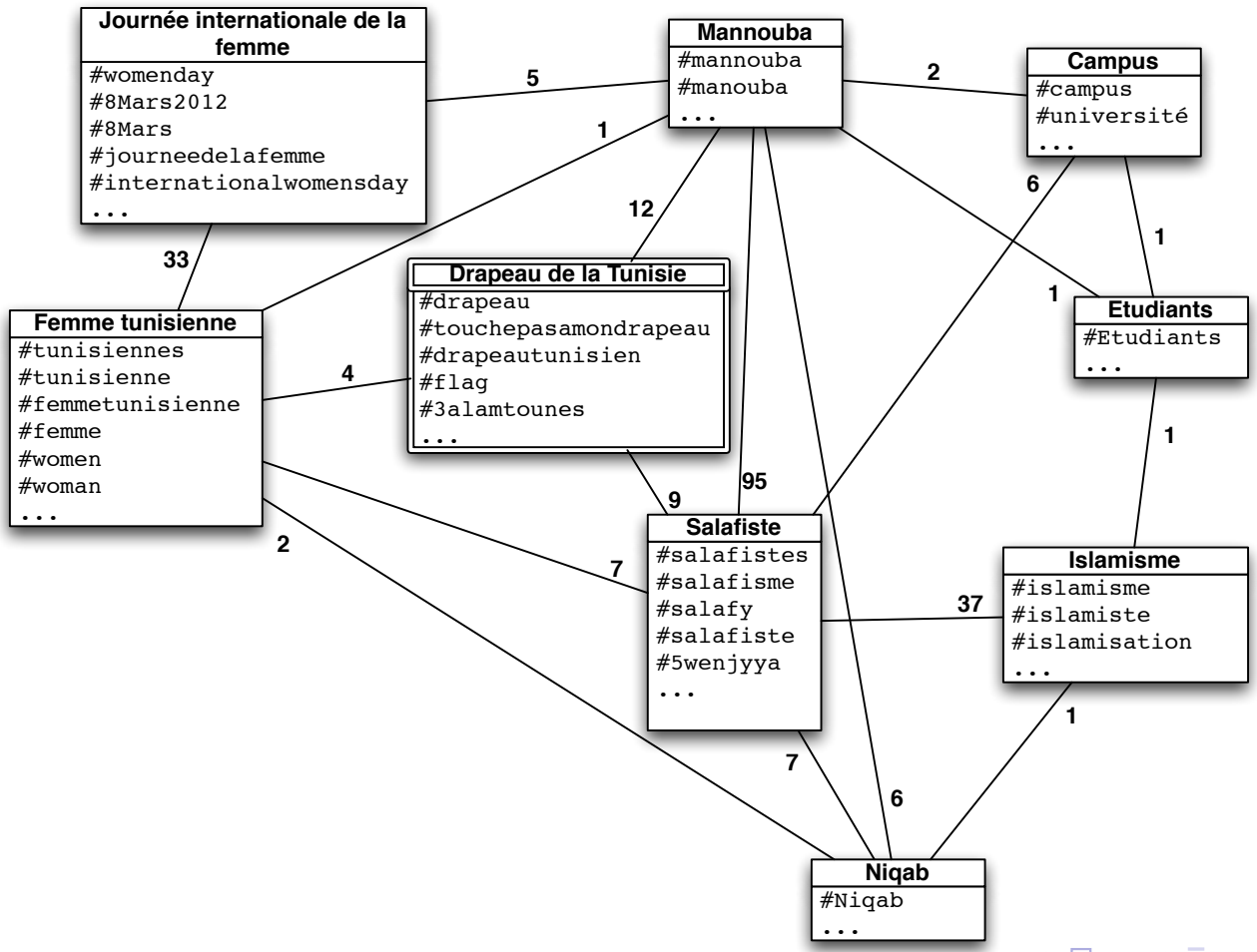
Cooccurrence des hashtags : événement *Manouba*

- Nous avons constaté qu'il y a d'autres hashtags qui cooccurrent avec les hashtags principaux.
- Ces hashtags peuvent reflèter la nature des événements et éclaircir le contexte dans lequel ils ont eu lieu.

	Hashtags cooccurrent avec le sujet		**nombre de fois**
	Sujet	**Hashtags**	
Manouba	Tunisie	Tunisie	702
		Tunisia	
		tunisi	
		tn	
	Le gouvernement Tunisien	tngov	126
	L'assemblée constituante	tnac	108
	Salafistes	salafistes	106
		salafiste	
		salafisme	
		salafis	
		salafist	
	Ennahdha	ennahdha	88
		nahdha	
		ennahda	
	Drapeau Tunisien	drapeau	11
		touchepasamondrapeau	
	Niqab	niqab	6

Cooccurrence des hashtags : graphe

Figure: Graphe représentant des hashtags liés à l'événement de la mise en berne de drapeau Tunisien à l'université de manouba. *3alamtounes* : *3alam* = drapeau et *tounes* = Tunisie.



Utilisateurs : comportement

les 5 utilisateurs qui ont le plus grands nombre de tweets dans le corpus.

Pseudonyme	Nb. tweets	Nb. abonnés	Nb. abonnement	NFER[1]	NFAR[2]
tunisinews	1 911	557	1	96	0
tunisieup	1 831	1266	631	16	0
tunisienouvelle	1 520	161	268	21	0
actutunisie	1 407	3818	18	76	0
journaltunisie	1 403	1497	255	88	1

- Nous avons remarqué que ces utilisateurs sont des journaux ou des magazines
- Les tweets sont généralement objectifs (contiennent des nouvelles) et standards (ne contiennent pas des fautes).

les tops 5 utilisateurs retweetés dans le corpus.

Pseudonyme	Nb. tweets	Nb. abonnés	Nb. abonnement	NFER	NFAR
ooouups	803	4 666	631	545	71
toomaa_6	22	2 630	1 063	489	1
nawaat	202	40 434	399	481	8
tn_revo	535	199	286	387	59
arabeman2012	1 081	308	219	324	362

- Nous avons remarqué que ces utilisateurs partagent souvent des opinions.

[1] nombre de fois que l'utilisateur était retweeté

[2] nombre de fois que l'utilisateur a retweeté

Utilisateurs : comportement

les 5 utilisateurs qui ont le plus grands nombre de tweets dans le corpus.

Pseudonyme	Nb. tweets	Nb. abonnés	Nb. abonnement	NFER[1]	NFAR[2]
tunisinews	1 911	557	1	96	0
tunisieup	1 831	1266	631	16	0
tunisienouvelle	1 520	161	268	21	0
actutunisie	1 407	3818	18	76	0
journaltunisie	1 403	1497	255	88	1

- Nous avons remarqué que ces utilisateurs sont des journaux ou des magazines
- Les tweets sont généralement objectifs (contiennent des nouvelles) et standards (ne contiennent pas des fautes).

les tops 5 utilisateurs retweetés dans le corpus.

Pseudonyme	Nb. tweets	Nb. abonnés	Nb. abonnement	NFER	NFAR
ooouups	803	4 666	631	545	71
toomaa_6	22	2 630	1 063	489	1
nawaat	202	40 434	399	481	8
tn_revo	535	199	286	387	59
arabeman2012	1 081	308	219	324	362

- Nous avons remarqué que ces utilisateurs partagent souvent des opinions.

[1] nombre de fois que l'utilisateur était retweeté

[2] nombre de fois que l'utilisateur a retweeté

Utilisateurs : comportement

les 5 utilisateurs qui ont le plus grands nombre de tweets dans le corpus.

Pseudonyme	Nb. tweets	Nb. abonnés	Nb. abonnement	NFER[1]	NFAR[2]
tunisinews	1 911	557	1	96	0
tunisieup	1 831	1266	631	16	0
tunisienouvelle	1 520	161	268	21	0
actutunisie	1 407	3818	18	76	0
journaltunisie	1 403	1497	255	88	1

- Nous avons remarqué que ces utilisateurs sont des journaux ou des magazines
- Les tweets sont généralement objectifs (contiennent des nouvelles) et standards (ne contiennent pas des fautes).

les tops 5 utilisateurs retweetés dans le corpus.

Pseudonyme	Nb. tweets	Nb. abonnés	Nb. abonnement	NFER	NFAR
ooouups	803	4 666	631	545	71
toomaa_6	22	2 630	1 063	489	1
nawaat	202	40 434	399	481	8
tn_revo	535	199	286	387	59
arabeman2012	1 081	308	219	324	362

- Nous avons remarqué que ces utilisateurs partagent souvent des opinions.

[1] nombre de fois que l'utilisateur était retweeté

[2] nombre de fois que l'utilisateur a retweeté

Application des techniques de TALN?

- Absence de ressources linguistiques annotées.
- Difficile d'appliquer des méthodes classiques de traitement de la langue sur tel type de texte (courts, plusieurs langues. . .).

Solution :

- Création de corpus d'apprentissage.
- Annotation par des experts (initiés à la façon d'écriture des tunisiens).
- Annotation via un site web[4].

[4] `http://rali.iro.umontreal.ca:8080/dridihou`

Construction de corpus d'apprentissage

Veuillez lire ces instructions avant de commencer

Vous n'êtes pas obligé d'annoter tous les tweets.

Vous pouvez vous connecter une autre fois avec les mêmes login et mot de passe et vous allez trouver un autre ensemble à annoter.

Num	Tweet	Langue	Sentiment			
1	En Tunisie, il est difficile d'assumer ses missions d'universitaires. Aidons-les en diffusant leurs messages! https://t.co/y0tkxPmO		positive ○	négative ○	neutre ○	Information inssufisante ○
2	حاجب العيون : أسعار العلف من نار ... والسماسرة على الخط http://t.co/NeKaTxGj #Tunisie #Tunisia		positive ○	négative ○	neutre ○	Information inssufisante ○
3	العلمانين لا يتحملون ديمقراطيتهم العفنة .. شكوى ضد الداعية وجدي غنيم في تونس http://t.co/rPC2xNvp #tunisie		positive ○	négative ○	neutre ○	Information inssufisante ○
4	Souvenez-vous des guignols qui venaient nous réciter leurs programmes à la #TTN. C'était ça leur campagne électorale! #LesInconnus #tnElec		positive ○	négative ○	neutre ○	Information inssufisante ○
5	Tunisie: L'INRIC qualifie "d'injustes" les accusations contre les journalistes http://t.co/YvEissHM		positive ○	négative ○	neutre ○	Information inssufisante ○
6	Tunisie: Le reflet de l'ombre de Naceur Belhaj Bettaïeb: [La Presse] S'appuyant... http://t.co/7RCdzTZn #artisanat		positive ○	négative ○	neutre ○	Information inssufisante ○
7	Titre alternatif: 'La blague de l'année' http://t.co/SPvWEnCr		positive ○	négative ○	neutre ○	Information inssufisante ○
8	#immobilier TUNISIE LEASING : Les caractéristiques de l'opération d'absorption par la société de sa filiale la... http://t.co/gWANAGFQ		positive ○	négative ○	neutre ○	Information inssufisante ○
9	#TUNISIE #HUMOUR Chômeur avant et après la révolution par @benyaglane https://t.co/6VeZ6Rud		positive ○	négative ○	neutre ○	Information inssufisante ○
10	TV5MONDE : actualites : L'Arabie saoudite n'extradera jamais Ben ... - TV 5 http://t.co/rYX2KDCF #marzouki #MBJ		positive ○	négative ○	neutre ○	Information inssufisante ○

Conclusion

- Les tweets sont généralement courts et non-standards (fautes, abréviations. . .).
- Difficile d'appliquer des méthodes classiques de traitement de la langue sur ce type de texte (courts, plusieurs langues,. . .).
- Les hashtags semblent être de bons points de départ pour détecter les événements et comprendre le contexte.
- Nous avons effectué les statistiques manuellement.

Prochaines étapes

- Éliminer les données qui peuvent fausser notre analyse.
- Implémenter un système de détection des événements.
- Appliquer des techniques de TALN sur le corpus qu'on va obtenir.
- Traiter un autre dialecte.
- Résultats : système de surveillance qui affiche les préoccupation actuelle des internautes et leurs opinions.


twitter