Proposition pour l'acquisition d'un corpus de Tweets V5

Julien Longhi

Proposition pour l'acquisition d'un corpus de Tweets

V5

Version du 2 mai 2014

Pour citer ce document :

Longhi J. (2014). Proposition pour l'acquisition d'un corpus de Tweets *(cmr-polititweets-tei-v1-manuel.pdf)*. In Longhi, J., Marinica, C., Borzic, B., Alkhouli, A. *Polititweets : corpus de tweets provenant de comptes politiques influents*. Banque de corpus CoMeRe. Ortolang.fr : Nancy. [cmr-polititweets-tei-v1]

Coordinateur de la tâche : Julien Longhi

Participants : Julien Longhi, Claudia Marinica, Boris Borzic, Abdulhafiz Alkhouli.

- **Objectif de ce document :**

Ce document explique les conditions dans lequel la collecte des données de ce qui allait devenir le corpus

> *Longhi, J., Marinica, C., Borzic, B., Alkhouli, A. Polititweets : corpus de tweets provenant de comptes politiques influents. Banque de corpus CoMeRe. Ortolang.fr : Nancy. [cmr-polititweets-tei-v1]*

a été faite. Il cite le projet de recherche pour lequel la collecte a été effectuée, les méthodes de sélection de comptes Twitter, les descriptions des personnalités tenant les principaux comptes de départ. Il évoque le contexte juridique à partir duquel il a été décidé de mettre le corpus en accès libre. Enfin, il indique la structure XML intermédiaire dans laquelle les messages ont été décrits, avant passage en TEI dans le corpus proprement dit.

Contenu

1. Objectifs et contraintes de départ

La constitution d'un corpus de tweets correspond à un double objectif, pensé dès le départ comme des éléments solidaire d'une même recherche :

- participer dans le cadre du projet CoMeRe à la constitution d'un ensemble de corpus de communications médiatisées par les réseaux ;
- se doter d'un corpus pour réaliser une recherche centrée sur le lexique politique, à partir d'analyses d'observables issus des nouveaux moyens de communication.

Le cadre institutionnel est donc à la fois :

- le projet « Humanité numériques et data journalisme : le cas du lexique politique » http://www.u-cergy.fr/ ;
- le projet CoMeRe http://comere.org

Il s'agit donc d'un travail collaboratif, à plusieurs niveaux (au sein de l'Université de Cergy-Pontoise, et dans le cadre plus large de CoMeRe).

Dans ce cadre également, un sujet « Projet d'Initiation à la recherche » (PIR) a été proposé au laboratoire Etis, ce qui permet le travail sur l'état de l'art. Un stage intitulé « Analyse de réseaux sociaux à l'aide d'indicateurs linguistiques » aura également lieu du 3 avril au 2 octobre 2014 (6 mois).

Participants :

- Julien Longhi, MCF, Université de Cergy-Pontoise, CRTF EA1392 (Centre de recherche textes et francophonies) : coordination du projet de la Fondation, mise en place des échanges scientifiques et collaborations, participant à Comere
- Claudia Maricina ETIS UMR 8051 (Equipe MIDI) : travail dans la base de données et mise en forme des tweets au formal xml
- Boris Borzic, Abdulhafiz Alkhouli : extraction des tweets et constitution d'une base de données les regroupant

2. Etat de l'art sur Twitter et ses analyses

Projet étude Master Pierre Thomazot (PIR, soutenu le 26 mars 2014) :

Il présente les travaux actuels sur Twitter, chez les linguistes et les informaticiens.

Sur les corpus de tweets, visiblement en France rien encore d'existant, mais cela semble bouger.

D'autres projets européens équivalents à CoMeRe semblent avoir utilisé des tweets.

Sur les méthodes d'analyse, elles relèvent souvent de l'informatique, comme la LSA. La bibliographie référence certains de ces travaux.

Les informations principales sur les tweets :

Twitter est un médium de microbloggage qui permet aux utilisateurs de faire de courtes déclarations publiques, pour partager leur sentiment du moment.

Techniquement, un tweet se compose de peu d'éléments, à savoir un utilisateur, un message de 140 caractères ou moins, et un espace de commentaire à la suite pour re-tweeter.

Une des particularités de twitter est le "techno-langage" impliqué dans chaque tweet.

On y retrouve le hastag (#), permettant de 'tagger' le tweet, et dont le nom est laissé libre de choix à l'utilisateur (ex : #France, #democratie), l'arobase (@) pour adresser son message à un utilisateur particulier, et des URLs réduites.

3. Acquisition dans le cadre du projet Humanités Numériques et Data journalisme

Travail en interaction avec une recherche du laboratoire ETIS UMR 8051 (Equipe MIDI : Claudia Marinica, Boris Borzic, Abdulhafiz Alkhouli), dont un des objectifs est de trouver une méthode automatique pour détecter les comptes (twittos) de personnalités politiques sans prendre les journalistes, blogeurs, influenceurs, militants ....

3.1. Boris Borzic, Abdulhafiz Alkhouli : extraction de tweets

1) on est parti de 7 personnalités de 6 groupes politiques différents : JLMelenchon, Bayrou, Copé, Fillon, Lepen, Ayraut, Cohn-Bendit (on avait commencé par Copé-Fillon ce qui explique le doublon, et ce premier résultat nous a donné l'idée de partir des 6 groupes politiques, d'ou le chiffre 7)
2) on a récupéré toutes les listes ou ils étaient cités => 7087 listes
3) on a sélectionné parmi ces listes, celles qui avaient au moins 6 twittos et qui contenait la chaine de caractère *politic* dans le nom ou descriptif de la liste = 120 listes (11K lignes)
4) Sur ces 120 listes on a récupéré 2934 twittos
5) pour être sûr de sélectionner les twittos politiciens (et non journalistes ...), nous travaillons par seuil. En ne retenant que les comptes présents dans plus de 12 listes nous avons 205 twittos politiciens.

Sur ces 205 comptes nous avons récupéré les 200 derniers tweets de chacun au 27 mars 2014, soit 34273 tweets. Cela permet d'avoir un corpus centré sur l'entre-deux tours des élections municipales 2014, ou pour les comptes moins actifs une prise en compte de cette élection voire des précédentes (car selon la densité de publication de tweets, la temporalité différera d'un compte à l'autre : le plus ancien date du 2009-03-04 11:59:49).

3.2. Aspects techniques

On a développé une application sur mesure en 3 étapes :

1) qui fait appel à l'API de Twitter, nous appelons une dizaine de fonction de l'API selon nos besoins, ensuite nous récupérons toutes les informations sous format JSON que nous convertissons et
2) qui nous permet d'enrichir une base de données Etis avec un design de base qui nous est propre (dizaine de table cinquantaine de champs). Ensuite nous avons des programmes qui calculent des indices pour enrichir des champs supplémentaires ...
3) Ensuite nous pouvons faire un export sur mesure, avec une sous partie des informations stockées dans n'importe quel format de données...

- **Mise en forme des tweets : Claudia Marinica, voir point 6**

4. Description des 7 comptes/personnalités de départ

Synthèse des pages wikipedia

4.1. JL Melenchon

Né le 19 août 1951 à Tanger au Maroc, est un homme politique français.

Militant socialiste à partir de 1977, il est successivement élu conseiller municipal de Massy (1983), conseiller général de l'Essonne (1985) puis sénateur du même département en 1986, 1995 et 2004, enfin député européen en 2009 dans la circonscription Sud-Ouest. Il est ministre de l'Enseignement professionnel de 2000 à 2002, dans le gouvernement de cohabitation de Lionel Jospin.

Il fait partie de l'aile gauche du Parti socialiste jusqu'au congrès de Reims, en novembre 2008, date à laquelle il quitte ce parti pour fonder le Parti de gauche (PG). D'abord président du bureau national, il en est actuellement coprésident aux côtés de Martine Billard. Il est le candidat du Front de gauche à l'élection présidentielle de 2012, où il se positionne quatrième au premier tour, avec 11,10 % des voix.

Si Jean-Luc Mélenchon se qualifie lui-même de socialiste républicain, son ambition ultime est de parvenir à « être le rassembleur de toute la gauche » sur une ligne politique antilibérale voire anticapitaliste issue de la « révolution par les urnes » pour gouverner et transformer profondément (révolution socialiste) la France dans le sens du progrès massif de l'intérêt général (refondation républicaine) à l'instar des expériences sud-américaines boliviennes et vénézuéliennes respectivement pilotées par Evo Morales et Hugo Chávez77 et l'Autre Gauche.

4.2. F Bayrou

François Bayrou est un homme politique français né le 25 mai 1951 à Bordères dans les Pyrénées-Atlantiques.

Ministre de l'Éducation nationale sous trois gouvernements différents de droite, ancien député des Pyrénées-Atlantiques de 1986 à 2012, ancien député européen et conseiller du président du Parlement européen, deux fois président du conseil général des Pyrénées-Atlantiques, conseiller municipal de Pau, il a été aussi président des partis Centre des démocrates sociaux (CDS), Force démocrate, l'Union pour la démocratie française (UDF) et du Mouvement démocrate (MoDem) qu'il a fondé.

Après avoir obtenu 6,84 % des voix à l'élection présidentielle de 2002 (4e du premier tour), il réunit 18,57 % des suffrages au premier tour de l'élection présidentielle de 2007 (3e du premier tour), et fonde peu après un nouveau parti qui se veut au centre de l'échiquier politique : le MoDem, successeur de l'UDF, dont il est le président. Lors de l'élection présidentielle de 2012, il recueille 9,13 % des suffrages (5e du premier tour) ; il n'est pas réélu député aux élections législatives qui suivent.

François Bayrou a souvent mis en cause l'objectivité des médias français appartenant à de grands groupes industriels, arguant de leur forte tendance à la bipolarisation de la vie politique française, autour de l'UMP et du PS. Il accuse ces médias d'une surexposition de

ces partis et de leurs candidats voire de connivence avec certains de ces candidats ; il affirme que cette inclination s'exprime notamment dans le contenu des questions posées et dans celles qui justement ne le sont pas. Il propose à cet effet de rendre impossible la détention des groupes de médias par des groupes industriels et financiers dépendant des commandes de l'État.

4.3. Jean-François Copé

Jean-François Copé est un homme politique français, né le 5 mai 1964 à Boulogne-Billancourt (Hauts-de-Seine).

Maire de Meaux et député de la sixième circonscription de Seine-et-Marne, il a occupé plusieurs fonctions ministérielles dans les gouvernements Raffarin et Villepin. Nommé secrétaire général de l'UMP en 2010, il est élu président du parti en novembre 2012.

Après la défaite de Nicolas Sarkozy à l'élection présidentielle de 2012, il autorise la création de courants au sein de l'UMP. Le 26 août 2012, à Châteaurenard, il se déclare candidat à la présidence du parti, vacante pendant le mandat présidentiel de Nicolas Sarkozy, et compose un « ticket » avec Luc Chatel pour la vice-présidence du parti et Michèle Tabarot pour le secrétariat général.

Le soir du scrutin, le 18 novembre 2012, les deux candidats revendiquent chacun la victoire. Jean-François Copé est proclamé vainqueur le lendemain, par 50,03 % des voix, par la Commission d'organisation et de contrôle des opérations électorales (COCOE). Ce résultat est contesté par les partisans de François Fillon. Après une semaine d'atermoiements politiques et médiatiques, la Commission nationale des recours (CONARE) proclame à nouveau vainqueur, cette fois avec 50,28 % des voix. François Fillon et ses partisans refusent toujours de reconnaître l'élection, et fin décembre 2012, un accord est finalement trouvé entre Jean-François Copé et François Fillon. Le 30 juin 2013, à l'issue d'un congrès extraordinaire, les militants de l'UMP votent pour le maintien de Jean-François Copé à la tête du parti jusqu'en 2015.

Le début du mandat de Jean-François Copé est notamment marqué par l'engagement de l'UMP dans l'opposition au projet de loi ouvrant le mariage aux personnes de même sexe.

4.4. François Fillon

François Fillon est un homme politique français, né le 4 mars 1954 au Mans.

Assistant parlementaire de profession, et membre du RPR puis de l'UMP, il est nommé successivement ministre de l'Enseignement supérieur et de la Recherche dans le gouvernement d'Édouard Balladur (1993-1995), puis ministre des Technologies de l'Information et de la Poste (1995), et ministre délégué chargé de la Poste, des Télécommunications et de l'Espace au sein des deux gouvernements d'Alain Juppé (1995-1997). Après la réélection de Jacques Chirac à l'Élysée, François Fillon est nommé ministre des Affaires sociales, du Travail et de la Solidarité (2002-2004) ; il mène des réformes structurelles sur la durée du travail et sur les retraites. Nommé ministre de l'Éducation nationale, de l'Enseignement supérieur et de la Recherche (2004-2005), il fait voter par le Parlement, la loi Fillon pour l'éducation.

À la suite de la victoire de Nicolas Sarkozy à l'élection présidentielle de 2007, François Fillon est nommé Premier ministre le 17 mai 2007 et forme son premier gouvernement. Il est reconduit le 18 juin suivant, à la suite de la victoire de la droite aux élections législatives : son deuxième gouvernement est, en durée, le deuxième gouvernement le plus long de la Ve République après celui de Lionel Jospin. Il forme son troisième gouvernement le 14 novembre 2010. Le 10 mai 2012, il remet la démission de son gouvernement, à la suite de l'élection à la présidence de la République de François Hollande. Il est le deuxième Premier ministre, après Georges Pompidou, dans l'ordre de durée de présence en continu à Matignon.

Il est élu député de Paris en juin 2012. Candidat à la présidence de l'UMP, il conteste les résultats annoncés par les instances du parti qui, à la suite du vote du 18 novembre 2012, placent Jean-François Copé à sa tête ; il décide alors de la création d'un groupe parlementaire distinct du groupe UMP, le Rassemblement-UMP, qui est dissous après la conclusion d'un accord avec Jean-François Copé.

4.5. Marine Le Pen

Marine Le Pen, de son vrai nom Marion Anne Perrine Le Pen, est une femme politique française, née le 5 août 1968 à Neuilly-sur-Seine (Hauts-de-Seine).

Avocate de profession, elle occupe plusieurs mandats locaux à partir de 1998 (conseillère régionale du Nord-Pas-de-Calais, conseillère régionale d'Île-de-France, conseillère municipale d'Hénin-Beaumont) et siège au Parlement européen depuis juillet 2004.

Un temps vice-présidente exécutive du Front national, Marine Le Pen annonce à plusieurs reprises son intention de briguer la succession de Jean-Marie Le Pen à la présidence du Front national. À la suite d'une réunion du bureau politique du FN, le 12 avril 2010, son père annonce qu'il quittera ses fonctions au prochain congrès. Marine Le Pen confirme son intention de se porter candidate, contre Bruno Gollnisch. Les membres du Front national sont alors appelés à voter pour leur nouveau président et les cent membres du comité central. Le congrès du parti, organisé à Tours les 15 et 16 janvier 2011, voit l'élection de Marine Le Pen à la présidence du parti avec 67,65 % des voix des militants.

En avril 2011, Marine Le Pen est classée parmi les 100 personnes les plus influentes au monde par le magazine américain Time.

Candidate à l'élection présidentielle française de 2012, elle obtient 17,90 % des suffrages exprimés, soit le meilleur résultat jamais obtenu par un candidat du FN au premier tour d'une élection présidentielle française.

4.6. Jean-Marc Ayraut

Jean-Marc Ayrault, né le 25 janvier 1950 à Maulévrier en Maine-et-Loire, est un homme politique français. Il est Premier ministre depuis le 15 mai 2012.

Conseiller général de la Loire-Atlantique et maire de Saint-Herblain dans les années 1970-1980, il est maire socialiste de Nantes de 1989 à 2012, député de 1986 à 2012 et président du groupe socialiste à l'Assemblée nationale de 1997 à sa nomination à la tête du premier gouvernement de la présidence de François Hollande le 15 mai 2012, puis du second le 18 juin suivant.

Considéré pendant la campagne présidentielle comme un des favoris pour le poste de Premier ministre, il est nommé à cette fonction par le nouveau président de la République François Hollande le 15 mai 201234. Il soumet la composition de son gouvernement au chef de l'État, le 16 mai. Ce gouvernement est le 35e gouvernement de la Ve République française. Il renonce alors à son mandat de maire de Nantes, tout en restant conseiller municipal

4.7. Daniel Cohn-Bendit

Daniel Cohn-Bendit, né le 4 avril 1945 à Montauban (Tarn-et-Garonne), est un homme politique de nationalité allemande, présent dans la vie politique allemande, française et européenne. Né en France de parents allemands, il opte pour la nationalité allemande en 1959. Il fait ses études supérieures en France, à l'université de Nanterre ; durant le mouvement de mai 1968 dont il est l'un des leaders, le gouvernement utilise le fait qu'il n'est pas français pour l'expulser en Allemagne. Durant les années 1970, il s'insère dans la vie politique allemande comme élu du parti écologiste Die Grünen à Francfort. À partir des années 1990, il participe aussi à la vie politique en France avec les partis écologistes qui s'y sont succédé. Favorable à la mise en place d'une Europe fédérale, il est député européen depuis 1994 et coprésident du groupe Verts/ALE au Parlement européen depuis 2002. En septembre 2010, il cofonde le Groupe Spinelli, initiative visant à renforcer la tendance fédéraliste au Parlement européen. Daniel Cohn-Bendit dit aujourd'hui être favorable au capitalisme et à « une écologie qui prenne acte de l'économie de marché pour mieux la réguler »3. Membre des Verts Allemands depuis 1984, il a déclaré dans L'Humanité à l'occasion de la campagne pour les élections européennes de 1999 prôner un « réformisme écologico-social lié à une tradition libertaire qui est effectivement non étatique ». Il a revendiqué à cette même occasion l'étiquette de « libéral-libertaire »

5. Question juridique

5.1. Wikipedia

Droits d'auteur

Le problème des droits d'auteur s'appliquant à un message sur Twitter est loin d'être évident. Par exemple, si on recopie un tweet d'autrui, on ne peut invoquer le droit de courte citation, car le caractère « court » de la citation se rapporte à la longueur de l'œuvre dont elle est extraite. Les retweets, pour leur part, peuvent même être accusés de violer les droits moraux de l'auteur quand le message est modifié.

Mais un tweet n'est pas forcément protégé par le droit d'auteur, car celui-ci ne s'applique qu'aux créations originales. Il est rare qu'un message aussi court puisse être considéré comme une telle création, mais pas impossible (c'est le cas des slogans publicitaires).

Twitter lui-même encourage les utilisateurs à placer leurs messages dans le domaine public, ne revendiquant lui-même aucun droit dessus — ce qui lui vaut les félicitations des défenseurs des contenus libres en comparaison de Facebook

5.2. Blog SI Lex

par Lionel Maurel. Juriste & bibliothécaire

http://scinfolex.com/2009/06/14/twitter-et-le-droit-dauteur-vers-un-copyright-2-0/

Excellente nouvelle : Twitter ne revendique aucun droit de propriété intellectuelle sur les contenus produits par les utilisateurs du service. On sait que ce n'est pas forcément le cas de tous les services 2.0, qui peuvent se comporter comme de véritables prédateurs de ce point de vue

5.3. Bibliothèque

Twitter archivé à la Bibliothèque du Congrès : un patrimoine impossible ?

http://scinfolex.com/2010/05/01/twitter-archive-a-la-bibliotheque-du-congres-un-patrimoine-impossible/

5.4. https://twitter.com/tos

- **2. Confidentialité**

Toute information que vous communiquez à Twitter est soumise à notre Politique de Vie Privée, qui régit la collecte et l'utilisation de vos informations. Vous comprenez qu'en utilisant nos Services, vous consentez à la collecte et l'utilisation (ainsi qu'il est énoncé dans la Politique de Vie Privée) de cette information, y compris le transfert de cette information aux États-Unis et / ou dans d'autres pays à des fins de stockage, de traitement et d'utilisation par Twitter

➢ 5. Vos droits

Vous conservez vos droits sur tous les Contenus que vous soumettez, postez ou publiez sur ou par l'intermédiaire des Services. En soumettant, postant ou publiant des Contenus sur ou par le biais des Services, vous nous accordez une licence mondiale, non-exclusive, gratuite, incluant le droit d'accorder une sous-licence, d'utiliser, de copier, de reproduire, de traiter, d'adapter, de modifier, de publier, de transmettre, d'afficher et de distribuer ces Contenus sur tout support par toute méthode de distribution connu ou amené à exister.

Astuce Cette licence signifie que vous nous autorisez à mettre vos Tweets à la disposition du reste du monde et que vous permettez aux autres d'en faire de même.

Vous consentez à ce que cette licence comprenne le droit pour Twitter de fournir, de promouvoir et d'améliorer les Services et de mettre les Contenus publiés ou transmis au travers des Services à disposition d'autres sociétés, organisations ou individus en partenariat avec Twitter pour l'agrégation, la diffusion, la distribution ou la publication de ces Contenus sur d'autres supports, médias et services, dans la limite Des termes de ces Conditions pour l'utilisation de ces Contenus.

Astuce Twitter applique un ensemble évolutif de règles sur la manière dont les partenaires de l'écosystème peuvent interagir avec vos Contenus. Ces règles ont été conçues pour mettre en place un écosystème ouvert, tenant compte de vos droits. Mais ce qui vous appartient vous appartient – vous restez propriétaire de vos Contenus (et vos photos font partie de ces Contenus).

Ces usages supplémentaires par Twitter, ou d'autres sociétés, organisations ou individus en partenariat avec Twitter, peuvent être faits sans compensation à votre égard en ce qui concerne les Contenus que vous soumettez, postez, transmettez ou rendez disponible au travers des Services.

Nous pouvons modifier ou adapter vos Contenus afin de les transmettre, afficher ou distribuer sur des réseaux informatiques et sur différents médias et / ou apporter des changements nécessaires à vos Contenus afin de les rendre conformes aux exigences ou limitations de tous réseaux, équipements, services ou médias.

➢ 9. Politique de Copyright

Twitter respecte les droits de propriété intellectuelle d'autrui et s'attend à ce que les utilisateurs des Services en fassent de même. Nous répondons aux notifications relatives à une violation des droits d'auteur dès lors qu'elles sont conformes à la législation applicable et nous sont adressées correctement. Si vous pensez que vos Contenus ont été reproduits ou diffusés de manière contrefaisante, veuillez nous fournir les informations suivantes : (i) une signature physique ou électronique du titulaire de droits ou d'une personne autorisée à agir en son nom, (ii) l'identification de l'œuvre protégée qui selon vous a fait l'objet d'une utilisation contrefaisante, (iii) l'identification des contenus qui selon vous porte atteinte à ces droits ou fait l'objet d'activités contrefaisantes et qui doit être enlevé ou dont l'accès doit être désactivé, ainsi que des renseignements raisonnablement suffisants pour nous permettre de localiser ces contenus, (iv) vos coordonnées, notamment vos adresse, numéro de téléphone et une adresse e-mail, (v) une déclaration de votre part selon laquelle vous estimez de bonne foi que l'utilisation des contenus en cause n'est pas autorisée par le titulaire de droits, son mandataire ou la loi, et (vi) une déclaration selon laquelle les informations contenues dans la notification sont exactes et, sous peine de parjure, que vous êtes autorisé à agir pour le compte du titulaire de droits.

5.5. **Réponse de Twitter**

```
Mise à jour par : YLee, 17 janv. 08:39 AM:
Hello,
Please see this help page
https://support.twitter.com/articles/114233
Thank you,
YLee
```

- **Extraits de cette page :**

Directives relatives à la diffusion des Tweets

Nous acceptons et encourageons l'utilisation de Twitter dans les émissions. Nous faisons le nécessaire pour que leur contenu soit bien attribué aux utilisateurs de Twitter et pour assurer à votre audience la meilleure expérience qui soit. Vous trouverez ci-dessous des suggestions basées sur ce qui a le mieux fonctionné pour les intégrations de Tweets déjà réalisées, ainsi que les coordonnées auxquelles vous pouvez nous joindre si vous avez des questions.

[…]

Utilisation du contenu

Nous voulons garantir la meilleure expérience possible aux personnes qui voient vos émissions et simplifier pour vous la procédure d'autorisation et d'octroi de licence. Si vous suivez ces directives, vous n'avez pas à contacter Twitter pour d'autres autorisations d'affichage ou en relation avec les marques déposées. Dans certains cas, l'autorisation du créateur du contenu peut toutefois être nécessaire, les utilisateurs de Twitter conservant les droits sur le contenu qu'ils postent.

Quoi qu'il en soit, un contenu Twitter ne peut pas être utilisé sans autorisation explicite de son créateur dans les cas suivants :

- dans des publicités
- pour impliquer le soutien d'un produit ou service quel qu'il soit

Par ailleurs :

- N'impliquez aucun soutien ni sponsorisation de votre production par Twitter, ni aucune fausse association de celle-ci avec Twitter.
- N'utilisez pas les marques déposées Twitter dans le titre de votre production sans d'abord valider leur utilisation avec nous.
- Assurez-vous de respecter toutes les lois, ordonnances et réglementations applicables à votre fourniture de contenu à votre public, notamment toutes les normes et exigences en matière de diffusion.

Affichage graphique de Tweets, @pseudonymes et hashtags

Consignes :

Placez l'oiseau Twitter www.twitter.com/logo à proximité immédiate des Tweets pour la durée de l'affichage de ceux-ci dans votre émission.

Indiquez le nom d'utilisateur et le pseudonyme Twitter (@pseudonyme) avec chaque Tweet.

Reprenez le texte complet du Tweet. Vous ne pouvez éditer ou modifier le texte d'un Tweet que si nécessaire en raison de limitations techniques ou liées au moyen utilisé (il est possible de supprimer les hyperliens par exemple).

Vérifiez que la taille du logo Twitter ou de l'icône représentant l'oiseau est raisonnable par rapport au contenu. La bonne taille est légèrement plus grande qu'une ligne de texte.

À ne pas faire :

Vous ne devez pas supprimer, masquer ni modifier l'identification de l'utilisateur. Vous pouvez exceptionnellement afficher des Tweets sous forme anonyme, par exemple en cas d'atteinte potentielle à la vie privée de l'utilisateur.

Montrer des données non attribuées sous forme agrégée ou graphique est autorisé, mais vous devez tout de même inclure l'oiseau Twitter officiel.

6. Mise en forme du corpus, Standards et mise aux normes

Travail avec Claudia Maricica (MCF UCP, UMR ETIS)

Le corpus est parti d'une recherche à partir de 7 twittos, qui par extensions en a délimité 205 au total. Nous avons fait 7 fichiers. Chaque fichier aura une sous-organisation et contiendra tous les twittos (et tous les tweets correspondant) associés à une personne politique, dite principale. Il n'y a pas de critère spécifique de regroupement des twittos dans chaque fichier.

6.1. Description de chaque champ du fichier XML :

```
<?xml version="1.0" encoding="UTF-8"?>
<corpus>
Description : balise contenant l'ensemble des informations du fichier
   <meta>
Description : balise contenant la description du contenu du fichier
         <name>
Description : Le nom de la personnalité politique choisie pour ce fichier
Format : String – chaine de caractères
Valeur par défaut :  pas de valeur par défaut – ce champ doit être renseigné
</name>
         <name>
Description : Le nom de la personnalité politique suivante
Format : String – chaine de caractères
Valeur par défaut :  pas de valeur par défaut – ce champ doit être renseigné
</name>
         <name>
Description : Le nom de la personnalité politique qui suit
Format : String – chaine de caractères
Valeur par défaut : pas de valeur par défaut – ce champ doit être renseigné
</name>
                  etc.... pour 27 autres
   </meta>
   <tweets>
Description : balise qui contient l'ensemble de tweets relatifs à la première
personnalité présentée dans la balise meta ci-dessus
         <meta>
Description : description de la première personnalité politique
<author>
Description : description de la première personnalité politique
                     <user_id>
Description : L'identifiant de personnalité politique auteur des tweets
Format : String – chaine de caractères
Valeur par défaut :  pas de valeur par défaut – ce champ doit être renseigné
                     </user_id>
                     <name>
Description : Le nom de l'auteur des tweets
Format : String – chaine de caractères
Valeur par défaut :  pas de valeur par défaut – ce champ doit être renseigné
                     </name>
                     <screen_name>
Description : Le nom du compte de l'auteur des tweets
Format : String – chaine de caractères
Valeur par défaut :  pas de valeur par défaut – ce champ doit être renseigné
                     </screen_name>
               </author>
         </meta>
         <tweet id=" ">
Description : La description complète d'un tweet. Elle contient un attribut
« id » qui contient l'identifiant du tweet courant. Il s'agit d'une chaine de
caractères et la valeur de l'attribut est toujours renseignée.
<author>
```

```
Description : description de la personnalité politique auteur du tweet
                        <user_id>
Description : L'identifiant de personnalité politique auteur du tweet
Format : String – chaine de caractères
Valeur par défaut :  pas de valeur par défaut – ce champ doit être renseigné
                        </user_id>
                        <name>
Description : Le nom de l'auteur du tweet
Format : String – chaine de caractères
Valeur par défaut :  pas de valeur par défaut – ce champ doit être renseigné
                        </name>
                        <screen_name>
Description : Le nom du compte de l'auteur du tweet
Format : String – chaine de caractères
Valeur par défaut :  pas de valeur par défaut – ce champ doit être renseigné
                        </screen_name>
                </author>
                < !-- datetime, defaut  NULL-->
                <creation_date>
Description : La date de création du tweet
Format : String – chaine de caractères
Valeur par défaut :  pas de valeur par défaut – ce champ doit être renseigné
                </creation_date>
                <tweet_text>
Description : Le texte qui est contenu dans le tweet
Format : String – chaine de caractères
Valeur par défaut : pas de valeur par défaut – ce champ doit être renseigné
                </tweet_text>
                <entities_hashtags>
Description : Les hashtags (#text1, #text2, etc…) qui sont contenus dans le
texte du tweet
Format : String – chaine de caractères. Ce champs contient que le texte qui suit
le caractère # pour tous les hashtags d'un tweet. Les textes des hashtags sont
séparés par une virgule et sans espace. Par exemple, si dans un tweet nous avons
deux hashtags #ump et #campagne, cette balise contiendra la chaine de caractères
suivante : ump,campagne
Valeur par défaut : la valeur par défaut est la chaine vide ""
                </entities_hashtags>
                <entities_urls>
Description : Les adresses des sites web qui sont contenus dans le texte du
tweet
Format : String – chaine de caractères. Les adresses web sont séparés par une
virgule et sans espace. Par exemple, si dans un tweet nous avons deux adresse
web http://dominiquedevillepin.fr/ et http://fb.me/1GbnVzcTo , cette balise
contiendra la chaine de caractères suivante :
http://dominiquedevillepin.fr/,http://fb.me/1GbnVzcTo
Valeur par défaut : la valeur par défaut est la chaine vide ""
                </entities_urls>
                <entities_user_mentions>
Description : L'identifiant des utilisateurs twitter qui sont référencés dans le
texte du tweet
Format : String – chaine de caractères. Les identifiants des utilisateurs sont
séparés par une virgule et sans espace. Par exemple, si dans un tweet nous avons
deux utilisateurs qui sont référencés (par leur user_id) 39793149 et 148281934 ,
cette balise contiendra la chaine de caractères suivante : 39793149,148281934
Valeur par défaut : la valeur par défaut est la chaine vide ""
                </entities_user_mentions>
                <geo_lat>
Description : La coordonnée latitude de la position géographique où l'auteur du
tweets se trouvait quand il a tweeté
Format : valeur Decimal  – une nombre réel
Valeur par défaut : la valeur par défaut est 0.0
                </geo_lat>
                <geo_long>
Description : La coordonnée longitude de la position géographique où l'auteur du
tweets se trouvait quand il a tweeté
Format : valeur Decimal  – une nombre réel
Valeur par défaut : la valeur par défaut est 0.0
```

```
                </geo_long>
                <source>
Description : La source utilisée par l'auteur pour tweeter le tweet : « web » si
c'est via le site web www.twitter.com, ou d'autres chaines de caractères en
fonction de l'application iOS, Android, etc. utilisée.
Format : String– chaine de caractères
Valeur par défaut : pas de valeur par défaut – ce champ doit être renseigné
                </source>
<isTruncated>
Description : Ce champ indique si le texte de ce tweet a été tronqué par twitter
avant publication (1 – le champ a été tronqué, 0 – le champ n'a pas été
tronqué). Comme ce procédé a été arrêté depuis quelque temps – twitter rejette
directement les tweets avec plus de 140 caractères, de manière générale ce champ
contient 0. Dans notre corpus tous les tweets ont la valeur 0 pour ce champ.
Format : Integer – nombre entier
Valeur par défaut : la valeur par défaut est 0
                </isTruncated>
                <isFavorited>
Description : Ce champ indique si le tweet a été marqué comme favori par
l'utilisateur ayant récupéré les tweets (1 – le tweet a été marqué comme favori,
0 – le tweet n'a pas été marqué comme favori). Dans notre corpus tous les tweets
ont la valeur 0 pour ce champ.
Format : Integer – nombre entier
Valeur par défaut : la valeur par défaut est 0
                </isFavorited>
                <inReplyToStatusId>
Description : Si le tweet est une réponse à un autre tweet, ce champ va contenir
l'identifiant de ce dernier tweet. Si c'est pas le cas, la valeur dans le champ
sera -1.
Format : String – chaine de caractères
Valeur par défaut : la valeur par défaut est -1
                </inReplyToStatusId>
                <inReplyToUserId>
Description : Si le tweet est une réponse à un autre tweet, ce champ va contenir
l'identifiant de l'utilisateur ayant tweeté ce dernier tweet. Si c'est pas le
cas, la valeur dans le champ sera -1.
Format : String – chaine de caractères
Valeur par défaut : la valeur par défaut est -1
                </inReplyToUserId>
                <inReplyToScreenName>
Description : Si le tweet est une réponse à un autre tweet, ce champ va contenir
le nom du compte de l'utilisateur ayant tweeté ce dernier tweet. Si c'est pas le
cas, la valeur dans le champ sera NULL.
Format : String – chaine de caractères
Valeur par défaut : la valeur par défaut est NULL
                </inReplyToScreenName>
                <place>
Description : Contient le pays d'où le tweet a été réalisé. Cette information
peut être attachée au tweet si on le souhaite. Donc, la plus part des tweets ne
contiennent pas cette information, mais ils contiennent NULL.
Format : String – chaine de caractères
Valeur par défaut : la valeur par défaut est NULL
                </place>
                <favoritecount>
Description : Contient le nombre de fois que le tweet a été mis en favori par
les autres utilisateurs.
Format : Integer – un nombre
Valeur par défaut : la valeur par défaut est 0
                </favoritecount>
                <retweetcount>
Description : Le nom de retweets du tweet courant.
Format : Integer – nombre entier
Valeur par défaut : la valeur par défaut est "0"
                </retweetcount>
                <isRetweet>
Description : Si le tweet est un retweet (valeur 1) ou pas (valeur 0)
Format : Integer – un nombre
Valeur par défaut : la valeur par défaut est 0
```

```
                </isRetweet>
                <retweetedstatus_id>
Description : Si le tweet est un retweet , ce champ contient l'identifiant du
tweet de base. Cela marche que pour le premier retweet, mais pas pour les
retweets des retweets. Si c'est pas un retweet la valeur est -1.
Format : String – chaine de caractères
Valeur par défaut : la valeur par défaut est -1
                </retweetedstatus_id>
                <contributors>
Description : Si le tweet a des contributeurs, ce champ contient l'identifiant
de ces contributeurs. Cette option est permise dans une version beta, et pas bcp
de comptes y ont accès. Donc, dans notre cas, ce champ sera vide ""
Format : String – chaine de caractères
Valeur par défaut : la valeur par défaut est ""
                </contributors>
                <getIsoLanguageCode>
Description : La langue détectée par twitter pour le tweet, par exemple :
« fr », « en », etc. Si pas de langue détectée, alors la langue est « und ».
Format : String – chaine de caractères
Valeur par défaut : la valeur par défaut est "und"
                </getIsoLanguageCode>
        </tweet>
   </tweets>
   <tweets>
        ….
   </tweets>
</corpus>
```

6.2. Exemples :

- **un exemple classique :**


```
<tweet id="1305615008">
   <author>
        <user_id>
             23341062
        </user_id>
        <name>
             Daniel Cohn-Bendit
        </name>
        <screen_name>
             danycohnbendit
        </screen_name>
   </author>
   <creation_date>
        2009-03-10 15:42:52.0
```

```
   </creation_date>
   <tweet_text>
          Je soutiens l'initiative de "Black-out" de la quadrature du Net contre
l'adoption de la loi Internet et Création par l'Assemblée Nationale
   </tweet_text>
   <entities_hashtags>

   </entities_hashtags>
   <entities_urls>

   </entities_urls>
   <entities_user_mentions>

   </entities_user_mentions>
   <geo_lat>
          0.0
   </geo_lat>
   <geo_long>
          0.0
   </geo_long>
   <source>
          web
   </source>
   <isTruncated>
          0
   </isTruncated>
   <isFavorited>
          0
   </isFavorited>
   <inReplyToStatusId>
          -1
   </inReplyToStatusId>
   <inReplyToUserId>
          -1
   </inReplyToUserId>
   <inReplyToScreenName>
          null
   </inReplyToScreenName>
   <place>

   </place>
   <favoritecount>
          4
   </favoritecount>
   <retweetcount>
          1
   </retweetcount>
   <isRetweet>
          0
   </isRetweet>
   <retweetedstatus_id>
          -1
   </retweetedstatus_id>
   <contributors>

   </contributors>
   <getIsoLanguageCode>
          fr
   </getIsoLanguageCode>
</tweet>
```

- **un exemple avec des éléments dans les balises <entites-hashtags>, <entities_urls> et <entities_user_mentions>**


```
<tweet id="377812110891548672">
<author>
   <user_id>
                23341062
         </user_id>
         <name>
                Daniel Cohn-Bendit
   </name>
   <screen_name>
         danycohnbendit
         </screen_name>
</author>
   <creation_date>
   2013-09-11 17:13:21.0
</creation_date>
<tweet_text>
   Fellow Europeans - Register & RT #EUNOW @thunderclapit #ittakesseconds
http://t.co/YQ1DnHGJP9 cc @demagistris @serracchiani @DavidSassoli
</tweet_text>
   <entities_hashtags>
   EUNOW,ittakesseconds
   </entities_hashtags>
<entities_urls>
   http://thndr.it/1cwP3vg
</entities_urls>
<entities_user_mentions>
   561518048,29416670,35298549,54988878
</entities_user_mentions>
<geo_lat>
   0.0
</geo_lat>
<geo_long>
         0.0
</geo_long>
   <source>
   <a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for
iPhone</a>
</source>
<isTruncated>
   0
</isTruncated>
<isFavorited>
   0
</isFavorited>
<inReplyToStatusId>
   -1
```

```
</inReplyToStatusId>
<inReplyToUserId>
   -1
   </inReplyToUserId>
<inReplyToScreenName>
   null
</inReplyToScreenName>
<place>

</place>
<favoritecount>
   0
</favoritecount>
<retweetcount>
         4
</retweetcount>
   <isRetweet>
   0
</isRetweet>
<retweetedstatus_id>
   -1
</retweetedstatus_id>
<contributors>

</contributors>
<getIsoLanguageCode>
   en
</getIsoLanguageCode>
</tweet>
```

- **- un exemple avec une position géographique (ici Paris) – donc des valeurs pour les balises <geo_lat> et <geo_long>**


```
<tweet id="164987255373561856">
                <author>
                      <user_id>
                            29499020
                      </user_id>
                      <name>
                            DominiquedeVillepin
                      </name>
                      <screen_name>
                            Villepin
                      </screen_name>
                </author>
```

```
                <creation_date>
                        2012-02-02 09:23:21.0
                </creation_date>
                <tweet_text>
                        "Un éventuel assouplissement des systèmes d'embauche doit
être traité en début de quinquennat, avec les partenaires sociaux"
                </tweet_text>
                <entities_hashtags>

                </entities_hashtags>
                <entities_urls>

                </entities_urls>
                <entities_user_mentions>

                </entities_user_mentions>
                <geo_lat>
                        48.87856
                </geo_lat>
                <geo_long>
                        2.28265
                </geo_long>
                <source>
                        <a href="http://twitter.com/download/iphone"
rel="nofollow">Twitter for iPhone</a>
                </source>
                <isTruncated>
                        0
                </isTruncated>
                <isFavorited>
                        0
                </isFavorited>
                <inReplyToStatusId>
                        -1
                </inReplyToStatusId>
                <inReplyToUserId>
                        -1
                </inReplyToUserId>
                <inReplyToScreenName>
                        null
                </inReplyToScreenName>
                <place>
                        France
                </place>
                <favoritecount>
                        0
                </favoritecount>
                <retweetcount>
                        3
                </retweetcount>
                <isRetweet>
                        0
                </isRetweet>
                <retweetedstatus_id>
                        -1
                </retweetedstatus_id>
                <contributors>

                </contributors>
                <getIsoLanguageCode>
                        fr
                </getIsoLanguageCode>
        </tweet>
```

- **un exemple d'un tweet qui est une réponse à un autre tweet (à regarder les balises <inReplyToStatusId>, <inReplyToUserId> et <inReplyToScreenName>)**


```
<tweet id="448791820211085312">
	<author>
		<user_id>
			499856023
		</user_id>
		<name>
			Roselyne Bachelot
		</name>
		<screen_name>
			R_Bachelot
		</screen_name>
	</author>
	<creation_date>
		2014-03-26 13:01:22.0
	</creation_date>
	<tweet_text>
		@pititecoraya @LeGrand8D8 félicitations !
	</tweet_text>
	<entities_hashtags>

	</entities_hashtags>
	<entities_urls>

	</entities_urls>
	<entities_user_mentions>
		256586875,833259704
	</entities_user_mentions>
	<geo_lat>
		0.0
	</geo_lat>
	<geo_long>
		0.0
	</geo_long>
	<source>
```

```
                    <a href="http://twitter.com/download/iphone"
rel="nofollow">Twitter for iPhone</a>
                </source>
                <isTruncated>
                        0
                </isTruncated>
                <isFavorited>
                        0
                </isFavorited>
                <inReplyToStatusId>
                        448770697893265409
                </inReplyToStatusId>
                <inReplyToUserId>
                        256586875
                </inReplyToUserId>
                <inReplyToScreenName>
                        pititecoraya
                </inReplyToScreenName>
                <place>

                </place>
                <favoritecount>
                        0
                </favoritecount>
                <retweetcount>
                        0
                </retweetcount>
                <isRetweet>
                        0
                </isRetweet>
                <retweetedstatus_id>
                        -1
                </retweetedstatus_id>
                <contributors>

                </contributors>
                <getIsoLanguageCode>
                        fr
                </getIsoLanguageCode>
        </tweet>
```

- **un tweet qui est un retweet (à regarder les balises <isRetweet> et <retweetstatus_id>)**


```
<tweet id="448417240673964032">
   <author>
   <user_id>
```

```
                31177357
            </user_id>
            <name>
                Corinne Lepage
            </name>
            <screen_name>
                corinnelepage
            </screen_name>
    </author>
    <creation_date>
            2014-03-25 12:12:56.0
    </creation_date>
    <tweet_text>
            RT @foeeurope: How could #TTIP affect resistance to #fracking in your
country? ENG,GERM,SP,BG,FR of NoFrackingWay now available http://t.co…
    </tweet_text>
    <entities_hashtags>
            TTIP,fracking
    </entities_hashtags>
    <entities_urls>
            http://bit.ly/1rtcVbr
    </entities_urls>
    <entities_user_mentions>
            238096056
    </entities_user_mentions>
    <geo_lat>
            0.0
    </geo_lat>
    <geo_long>
            0.0
    </geo_long>
    <source>
            <a href="http://twitter.com/download/iphone" rel="nofollow">Twitter
for iPhone</a>
    </source>
    <isTruncated>
            0
    </isTruncated>
    <isFavorited>
            0
    </isFavorited>
    <inReplyToStatusId>
            -1
    </inReplyToStatusId>
    <inReplyToUserId>
            -1
    </inReplyToUserId>
    <inReplyToScreenName>
            null
    </inReplyToScreenName>
    <place>

    </place>
    <favoritecount>
            0
    </favoritecount>
    <retweetcount>
            5
    </retweetcount>
    <isRetweet>
            1
    </isRetweet>
    <retweetedstatus_id>
            448415385466183681
    </retweetedstatus_id>
    <contributors>

    </contributors>
```

```
   <getIsoLanguageCode>
         en
   </getIsoLanguageCode>
</tweet>
```

7. Références

[1] Paveau M-A, "Technodiscursivités natives sur Twitter. Une écologie du discours numérique" Dans Liénard, F. (2013) Culture, identity and digital writing, Epistémè 9, Revue internationale de sciences humaines et sociales appliqués, Séoul : University Korea, Center for applied Cutural Studies, p. 139-176

[2] Morchid M., Dufour R., Linarès G."lia@inex2012 : combinaison de thèmes latents pour la contextualisation de tweets" (2012)

[3] Genc Y., Sakamoto Y., Nickerson J.V., "Discovering Context : Classifying tweets through a semantic transform based on Wikipedia" (2011)

[4] Longhi J., "Essai de caractérisation du tweet politique", L'Information grammaticale, n°136, p.25-32 (2013)

[5] Conover M.D., Gonçalves B., Ratkiewicz J., Flammini A., Menczer F., "Predicting the political alignment of Twitter users" (2010), Center for Complex Networks and Systems Research School of Informatics and Computing Indiana University, Bloomington

[6] Trouvilliez B., "Représentation vectorielle de textes courts d'opinions : Analyse de traitements sémantiques pour la fouille d'opinions par clustering" (2010), Centre de recherche en Informatique de Lens, Université d'Artois, France.

[7] Kohonen T., Kaski S., Laggus K., Salojärvi J., Honkela J., Paatero V., Saarela A. "Self organization of a massive Document Collection" (2000), Neural Network Research Center, Helsinki University of Technology, Espoo, Finland.