UNIVERSITÉ DE LORRAINE

Les mots de la campagne présidentielle française 2012

Projet Tutoré
MASTER 1 Sciences de la Cognition et Applications – 2011 / 2012

Etudiants

Bruno ANDRIAMIARINA-M.
Cécile DESHAYES

Tuteurs

Maxime AMBLARD
Bruno GUILLAUME

Remerciements,
A nos tuteurs de projet, M. Amblard et M. Guillaume,
Aux différents professeurs qui ont répondu positivement à nos sollicitations, M. Barrandon et M. Rapezynski,
Aux personnes qui ont participé à nos tests.

Table des Matières

Un événement politique important a lieu cette année en France. Il s'agit de l'élection présidentielle qui ne se produit que tous les 5 ans.
Cet événement va donner lieu à toutes sortes de publications, commentaires, articles, communiqués, plus ou moins vertueux, sur l'ensemble des réseaux de communications, Internet, télévision, radio, meeting, bouche à oreille, pendant tout le temps de la campagne.

Un des flux les plus ouverts et accessibles à tous est aujourd'hui le Web. On y chuchote, on y parle, on y crie, sur tout et aussi sur l'élection présidentielle française 2012.
L'enjeu y est important car la communication peut y être à la fois maitrisée et spontanée. Une information ou un événement y est relaté, raconté, déformé, démenti, pressenti, inventé, confirmé …

Le sujet de notre projet tutoré nous amène à infiltrer cette toile pour y récupérer les informations circulant sur l'élection présidentielle française 2012 de manière à pouvoir les trier et les analyser.
Il nous invite à identifier, au cœur de cette nébuleuse médiatique, quelques principaux thèmes de la campagne, personnes à l'origine de courant de pensée, et autres débats entre sites interposés.

Dans un premier temps, nous établirons un périmètre de travail ; trouver les sites importants, pertinents, représentatifs et homogènes en termes de couleur politique.

Dans un second temps, nous mettrons en place les passerelles qui permettront la récupération de l'information. Nous parlerons alors d'API et de base de données.

Puis, dans un troisième temps, nous verrons comment ces données doivent être normalisées grâce à des outils comme un Lemmatiseur ou le WOLF avant d'être étudiées.

Enfin, ces analyses seront mises en ligne sur Internet, venant alors elles-mêmes alimenter la production de documentation sur la toile concernant la campagne présidentielle française 2012.

Ainsi va le flux.

A. Périmètre de l'étude

1. Sujet de l'étude

Cette étude concerne la campagne présidentielle française 2012 et tous les documents relatifs à ce sujet circulant sur Internet.
En effet, l'utilisation de l'information circulant sur le Web est d'un enjeu capital pour les candidats car elle peut être très positive comme très négative, contrôlée ou libre.
Certains sites ont une maîtrise totale de cette information, tels que les sites de campagne des candidats, où on l'on trouve de l'information traitée souvent comme de la communication.
D'autres sites sont au contraire capables du meilleur comme du pire, comme dans le cas par exemple, des réseaux sociaux où l'on trouve de l'information comme de la désinformation.
C'est dans le cadre de cette diversité ambivalente que le sujet de notre projet tutoré prend toute son importance et démontre sa pertinence.

2. Objet de l'étude

Il s'agit d'identifier, récupérer, trier et analyser la production de documentations relatives à la campagne présidentielle sur Internet.
Dans un premier temps, nous avons différencié plusieurs sources d'informations permettant de garantir une notion d'objectivité dans le cadre de cette campagne. Cette notion est basée sur un équilibre entre la représentativité des forces politiques (gauche, centre, droite) et un équilibre entre la représentativité des types d'interlocuteurs (politique, média, public).
Dans un second temps, plusieurs types de traitements ont été mis en place pour récupérer ces informations, le maximum de traitements devant être fait de manière automatique. C'est l'activité des sites qui a permis de définir la fréquence du traitement associé (quotidien, ou hebdomadaire).
Dans un troisième temps, la normalisation de ces données grâce à des outils tels que le lemmatiseur a permis d'obtenir une base de données de mots utilisés dans le cadre de cette campagne avec pour chaque référence, le site sur lequel il a été trouvé, la date où il a été écrit, l'article dans lequel il est présent et combien de fois on l'a répertorié ce jour là et sur ce site, dans cet article.
Enfin, l'exploitation de ces données à travers différents traitements a permis de mettre en évidence certains courants de pensée, temps de présence médiatique et autres flux observables durant cette campagne présidentielle.

3. But de l'étude

Il s'est agit d'essayer, grâce à différentes analyses, de dégager les thématiques abordées, de repérer l'émergence de leader et d'identifier les flux d'informations.
A travers les représentations graphiques liées à l'activité représentée par ces différentes analyses, nous allons étudier l'utilisation de la langue naturelle dans le contexte de la campagne présidentielle française.
Nous avons défini une base de données normalisée et des traitements automatisés en nous basant sur des arguments scientifiques.
Un site Internet a été créé pour servir de support à la diffusion de ces résultats et les rendre ainsi accessibles aux plus grand nombre.

4. Cycle de fonctionnement général

Le cycle de l'application se décrit en quelques étapes résumées par le schéma en Figure 1.

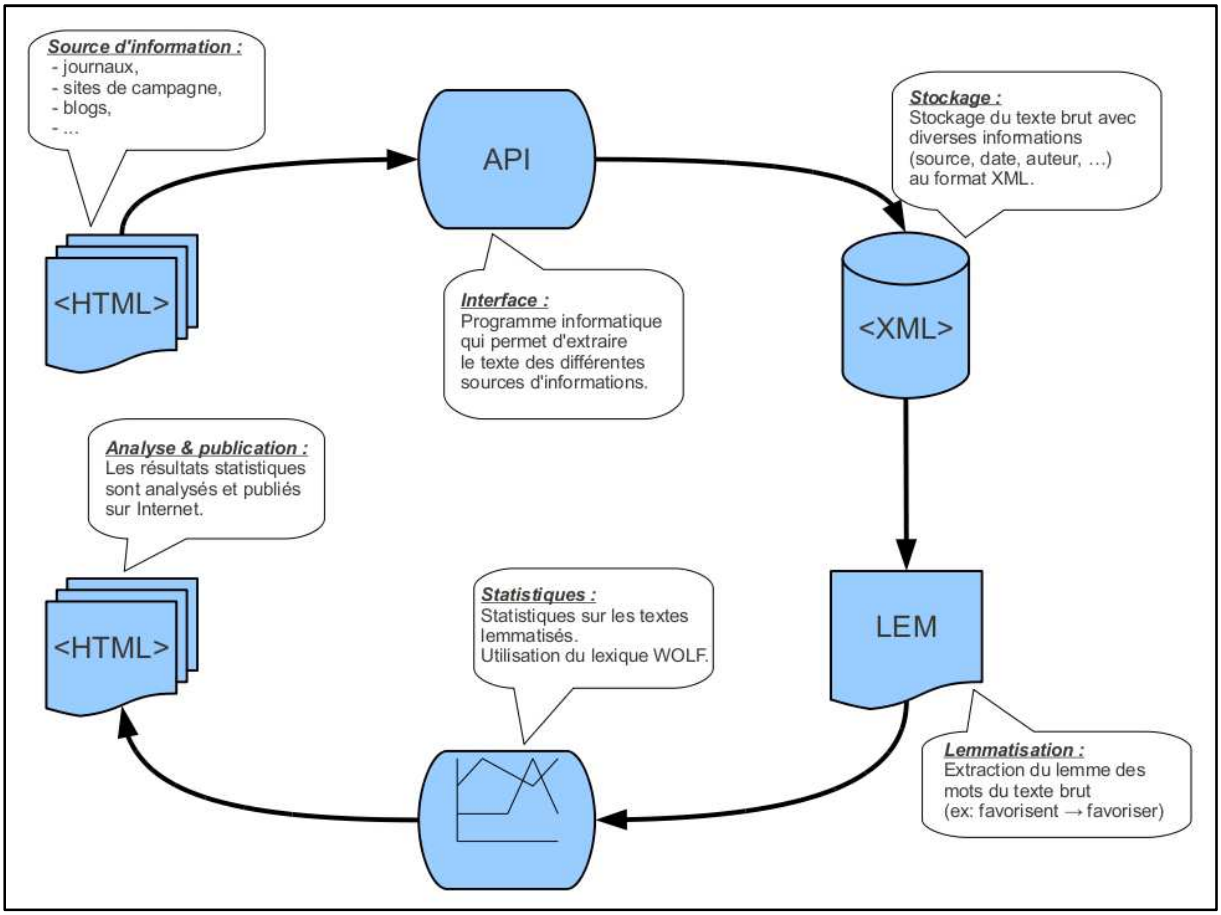


Figure 1 : Cycle de fonctionnement général

Dans un premier temps, une interface extrait le texte contenu dans les différentes sources d'informations disponibles sur Internet (journaux en lignes, sites de campagne des différents candidats, réseaux sociaux, …). Ces informations sont stockées dans des documents au format XML avec diverses informations telles que la date, l'auteur, … Pour pouvoir ensuite traiter ces informations, on lemmatise les textes (on dégage les lemmes des mots) et on leur applique des traitements statistiques. Les résultats ainsi obtenus sont analysés et mis en ligne.

5. Etat de l'art

Le sujet de notre projet tutoré est loin d'être anodin puisque plusieurs médias s'y sont également intéressés, en proposant eux aussi des analyses basées sur l'utilisation de la documentation liée à la campagne présidentielle 2012 circulant sur Internet.
Déjà en 2007, Jean Véronis, linguiste, avait créé un blog autour de cette thématique à l'occasion de l'élection présidentielle 2007[1]. On le retrouve encore cette année, sur son blog, réorganisé autour de l'élection présidentielle 2012 mais également sur le site du Monde, où, en association avec ce journal, il propose une analyse des discours de la campagne[2].

[1] http://blog.veronis.fr/

[2] http://www.lemonde.fr/election-presidentielle-2012/article/2012/ 03/06/explorez-et-analysez-tous-les-discours-de-la-campagne_1652544_1471069.html

C'est sur une autre thématique, celle de la visibilité médiatique, que Libération, avec l'institut TrendyBuzz propose une évaluation en continu[3]. La notion ici utilisée est différente en ce qu'elle se focalise sur la dynamique d'intérêt et non spécifiquement sur les mots.
On peut compléter le tableau avec le baromètre des opinions sur Internet proposé par Le Figaro[4].

A ce panel de sites proposant leur lecture de la campagne à travers différentes analyses, nous proposons d'ajouter le notre[5].
Il se fait fort de deux propositions pour les internautes :

- D'une part, il met à disposition des outils très dynamiques de recherche linguistique, offrant comme source de données tout ce qui a été récupéré sur les sites de campagne ou les sites médiatiques ; chaque requête donnant immédiatement lieu à une représentation graphique ; l'interprétation étant laissée à l'utilisateur.
- D'autre part, il met à disposition les études réalisées par nos soins, de leur structuration à leur interprétation, en passant par leur représentation graphique.

Il nous sert également de support à la diffusion de nos autres réalisations dédiées au projet à savoir, le présent rapport, la fiche de synthèse, le poster et la vidéo.

B. Contexte de l'étude

1. Détermination du champ d'application

a. Format des informations

Au début de notre analyse, plusieurs types de documents ont été identifiés :

- Tout d'abord des documents textuels prenant différentes formes tels que, les programmes de campagne, des discours, des communiqués de presse, des commentaires, des articles de presse,
- Puis des documents plus éclectiques tels que des caricatures ou des vidéos.

Afin de pouvoir automatiser au maximum les traitements de récupération des informations, nous avons décidé de limiter le champ de l'étude aux documents textuels récupérés sur Internet.
Ce choix exclu de fait les documents non textuels (caricature ou vidéo) nécessitant un traitement différent même s'il aurait été intéressant de les prendre en compte.

b. Sélection des sites : Enjeu fonctionnel

Il était important de pouvoir faire une sélection de sites donnant une bonne représentativité de la blogosphère politique. Cette représentativité, nous avons tenté de la mesurer en prenant en compte différents paramètres comme la fonction du site (média, blog, site de campagne), sa couleur politique ou encore sa popularité.
Ces sites devaient être des acteurs pertinents de la campagne présidentielle sur la toile.

[3] http://politivox.liberation.fr
[4] http://elections.lefigaro.fr/le-scan
[5] http://quine.loria.fr/lmdlc/index.html

(1) Fonction du site

Pour garantir une bonne représentativité, il s'est agit là surtout d'identifier la plateforme auquel un site se rattache ; c'est une manière de le catégoriser en tenant compte des gens qui le font vivre.
Nous avons pu ainsi identifier plusieurs sortes de plateforme qui de part leur diversité, pouvait apporter un panel représentatif de sites.

- Plateforme politique

La plateforme politique est une plateforme dans laquelle on va trouver les sites de campagne des candidats ou encore les sites de parti politique.
Ce sont des sites qui ont à la fois des parties statiques et des parties dynamiques. Dans les parties statiques, on trouve la présentation des équipes de campagne, le programme du candidat ou ses dates de meeting ; dans les parties dynamiques, on trouve soit les discours de campagne, soit des billets sur des thèmes de campagne ou des sujets d'actualité. Cette partie dynamique est mise à jour plusieurs fois par jour par des professionnels. Souvent, l'information y est traitée comme de la communication. Elle y est très contrôlée puisqu'elle a une fonction de représentation du candidat.
Cette plateforme va nous apporter la parole officielle des candidats.

- Plateforme professionnelle

La plateforme professionnelle est une plateforme dans laquelle on va trouver les sites des journaux, quotidiens ou hebdomadaires, des sites de télévisions ou d'autres médias.
A ce niveau, s'est posée à nous la question de savoir si nous souhaitions rester positionnés sur des sites en accès gratuits ou aller également sur des sites payants.
L'intérêt d'aller sur des sites payants ne nous est pas apparu essentiel en sachant que souvent, si ces sites peuvent être à l'origine de scoop, ils n'en restent pas longtemps l'unique diffuseur. Concernant les autres sites gratuits, ils proposent aussi souvent une partie payante qui ne concerne pas la diffusion de l'information mais la possibilité de mettre des commentaires.
Nous avons donc décidé de rester dans tous les cas, sur des sites proposant un accès gratuit car cela nous garantissait quand même la pluralité des informations.
Ce sont des sites qui sont essentiellement dynamiques dans le sens où ils sont alimentés régulièrement par les informations à traiter. Quelques uns disposent de parties moins dépendantes de l'actualité où l'on trouve des analyses ou des outils sur la campagne (tels que ceux que nous proposons sur notre site Web).
La plus part d'entre eux ont mis en place des parties spécifiques dédiées à la campagne présidentielle française, présentant ainsi l'avantage de donner une classification de facto de leurs articles ce qui nous a permis de localiser très facilement les articles qui nous intéressaient.
Ces sites présentent l'intérêt de traiter de tous les sujets relatifs à la campagne. Ils nous ont permis de récupérer de l'information sur tous les candidats. C'est aussi par ces sites que nous avons toutes les dépêches officielles des agences de presses telle que l'AFP ou Reuters.
Leur contenu est normalement réputé objectif de part la caution journalistique que leur apportent leurs auteurs.

- Plateforme publique

La plateforme publique est une plateforme dans laquelle on va trouver les sites des réseaux sociaux, les blogs et les forums.
Ce sont des sites très dynamiques sur lesquels on trouve essentiellement des commentaires, des billets d'humeurs qui agissent telles des réponses « réflexes » à des stimuli extérieurs et qui ne prennent tout leur sens que dans l'espace collectif où ils s'inscrivent.
Le principal problème des réseaux sociaux est de pouvoir attribuer de manière fiable un écrit à la campagne politique ou pas. Tout peut être écrit, de l'information comme de la désinformation. Ces sources sont donc à manipuler avec précaution.
Paradoxalement, on retrouve aussi sur ces réseaux de l'information très contrôlée puisque chaque candidat a une page Facebook ou un compte Twitter. Cela lui permet d'être en temps réel sur les mêmes réseaux que son électorat ; démarche inverse de celle des sites de campagne où ce sont les électeurs qui viennent à la rencontre des candidats.
Sur ces sites, nous avons donc trouvé de tout mais nous avons essentiellement pu y regarder de grande tendance de masse et avoir une vision globale des échanges.

(2) Couleur politique

Pour garantir une bonne représentativité, il a également fallu tenir compte de la tendance politique des sites de manière à pouvoir être sûr de rester impartial.
Le problème a donc été de connaitre la tendance politique des sites, ce qui n'est pas forcément une donnée qui s'affiche ouvertement ; surtout dans la mesure où l'on parle bien ici de parti pris (clairement affiché) mais aussi de tendance (souvent cachée).
Cette problématique n'a donc pas été la même suivant les plateformes ; certaines plateformes affichant ouvertement leurs couleurs politique, là ou d'autres essayaient de s'en défendre.

- Plateforme politique

Concernant les sites de la plateforme politique, le problème ne s'est absolument pas posé. La couleur politique des sites de campagne dépendait évidement du candidat référent et de son adhésion à tel ou tel parti. Nous nous sommes basé sur le référencement des partis politiques pour les cataloguer.

- Plateforme professionnelle

Concernant les sites de la plateforme professionnelle, le problème a été plus délicat car par principe les médias se défendent souvent d'avoir une étiquette politique. Ils revendiquent leur impartialité. Nous avons été obligés de nous en remettre à la notion relative de « notoriété publique » pour déterminer des tendances.

- Plateforme publique

Concernant la plateforme publique, le problème a été propre à chaque site. Sur les réseaux sociaux, dans la mesure où chaque internaute peut s'y exprimer, il n'y avait pas de problème dans un choix à faire. Par contre, pour sélectionner les blogs, il a fallu utiliser des moyens extérieurs pour garantir notre impartialité dans nos choix. Sur le site du journal Le Monde, un outil de recherche dans la blogosphère politique française[6] permet de rechercher des

[6] http://www.lemonde.fr/politique/visuel/2012/02/02/cartographie-de-la-blogosphere-politique-en-2012_1635269_823448.html

sites et de connaître leur couleur politique. C'est cet outil qui nous a permis de sélectionner des blogs en respectant un équilibre gauche / centre / droite.

(3) La popularité

Pour garantir une bonne représentativité, il a également fallu tenir compte de la popularité des sites. Cette donnée n'a pas été utilisée dans la sélection des sites de la plateforme politique. Par contre, nous en avons eu besoin pour les deux autres plateformes. Nous sommes partis du principe qu'un site populaire était plus représentatif qu'un site moins populaire.
Pour mesurer cette popularité, nous avons utilisé un site mesurant les buzz sur Internet[7] ainsi qu'un site qui permet de relever la fréquentation des sites[8].
Lorsqu'un choix était à faire entre deux sites de même plateforme et de même tendance politique, celui préféré a été celui ayant le plus de fréquentation. Si cette donnée n'existait pas, celui préféré a été celui ayant fait le plus de buzz. Les sites pressentis avaient forcément une information disponible sur leur popularité (fréquentation ou buzz), sinon, ils n'étaient pas du tout retenus pour faire partis de notre sélection.

c. Schéma de départ

Après avoir détaillé nos principes de sélection de site, voici la représentation graphique symbolisant la taxonomie de notre première classification.
On y distingue les 3 plateformes, politique, publique et professionnelle ainsi que les sites sélectionnés sur leur aspect fonctionnel.

Dans la plateforme politique, on retrouve les sites de campagne ainsi que les pages Facebook ou Twitter des candidats. On a tous les sites sur lesquels le candidat communique officiellement.
Dans la plateforme professionnelle, on retrouve les sites de journaux, hebdomadaires ou quotidiens, nationaux ou régionaux, des sites aussi de télévision.
Dans la plateforme publique, on retrouve les réseaux sociaux, Facebook et Twitter mais également des blogs ou des forums.

Le schéma en Figure 2 représente notre blogosphère politique ciblée.

[7] http://labs.ebuzzing.fr/top-blogs

[8] http://www.ojd-internet.com/chiffres-internet

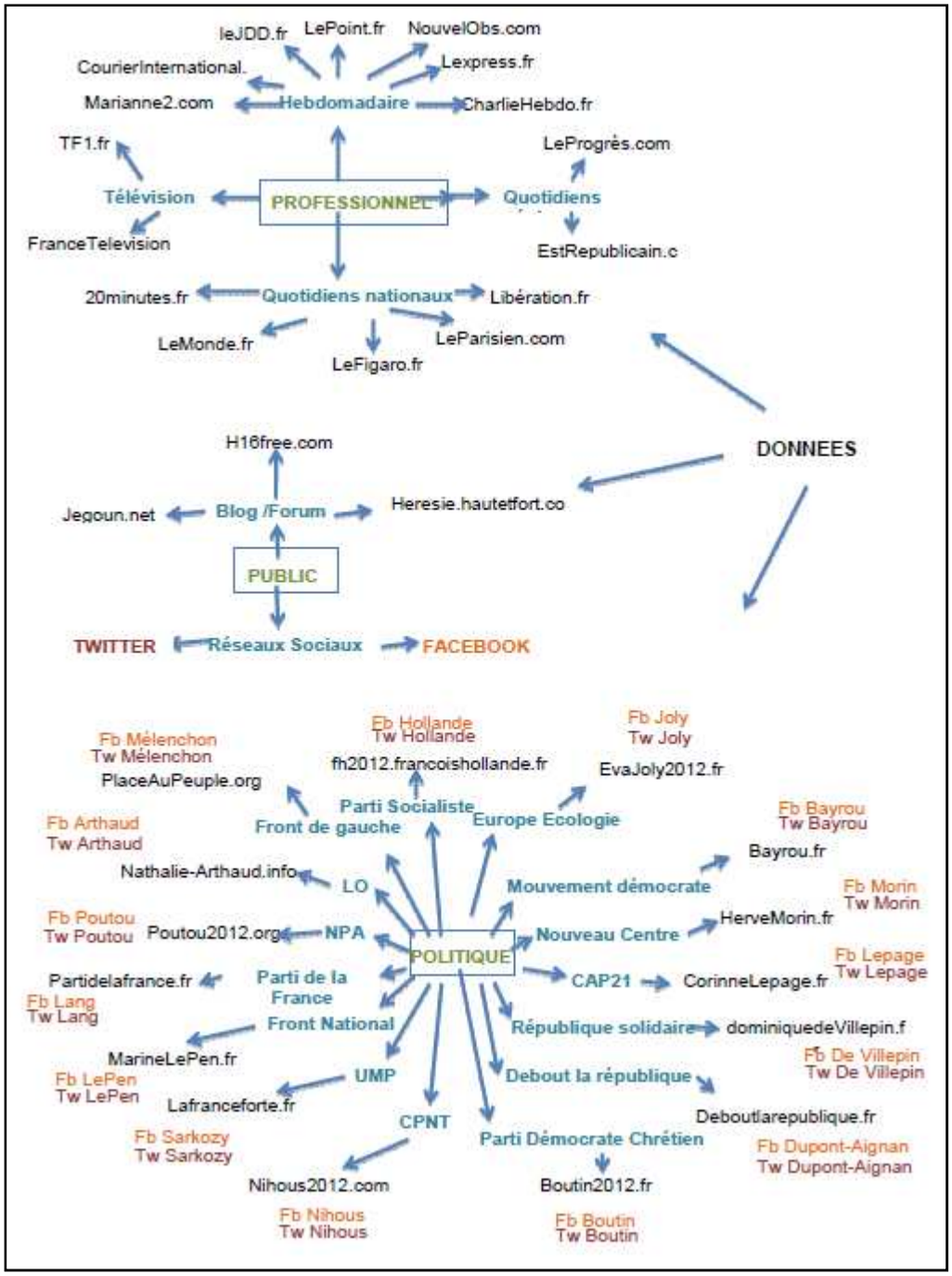


Figure 2 : Blogosphère politique ciblée

d. Sélection des sites : Enjeu technique

Une des principales problématiques de ce projet réside dans la récupération des informations depuis les diverses sources retenues. C'est dans cette optique de récupération des textes sur Internet qu'il a fallut utiliser des APIs.

(1) API (Application Programming Interface)

Une API (Application Programming Interface) est « une interface fournie par un programme informatique. Elle permet l'interaction de programmes entre eux, de manière analogue à une Interface Homme-Machine, qui rend possible l'interaction entre un homme et une machine. Du point de vue technique, une API est un ensemble de fonctions, procédures ou

classes mises à disposition par une bibliothèque logicielle, un système d'exploitation ou un service. » (Wikipedia)
On peut donc décrire une API comme étant un programme informatique qui permettra de faire le lien entre notre application et les différentes sources d'informations issues d'Internet.

(2) Les APIs déjà disponibles

Dans un premier temps, le travail a consisté à faire l'état des lieux des APIs existantes et essayer d'en sélectionner celles qui pourront nous être utiles. Plusieurs problèmes se sont posés : d'un côté, les réseaux sociaux (Twitter, Facebook) mettent à disposition des APIs qui en théorie devraient permettre de récupérer aisément le texte issu de cette source d'information. Dans la pratique, il en est autrement : le langage de programmation choisi (Python), ne possède pas d'APIs utilisables. Pour la plupart, elles correspondent toutes à des projets abandonnés ; ou dans le cas contraire, elles étaient tellement peu documentées qu'elles en étaient inutilisables. La seule API de réseaux sociaux que l'on a décidé de retenir est celle de Twitter, ou plus précisément une API qui permet d'accéder à la partie publique du réseau social Twitter (celle à laquelle tout internaute lambda peut accéder). La suite du projet a malheureusement eu raison de cette source, les informations y étant difficilement maîtrisables du fait de leur volume imposant et de leur contenu instable.

De l'autre côté, les sources d'informations plus conventionnelles, c'est-à-dire les sites Web classiques telles que les sites de campagne, la presse sur Internet ou encore les divers blogs ne possèdent tout simplement pas d'API. La question ne s'est donc pas posée : pour chacune de ces sources d'informations, il faut créer nos propres interfaces afin d'en extraire les textes.

(3) Les APIs mises en place

Afin de récupérer et de manipuler les textes des différents sites Internet, il a fallu mettre en place des API. Voici les spécifications de ces dernières :

- Le langage de programmation retenu est Python dans sa version 2.7 car elle possède de nombreuses bibliothèques et en particulier NLTK[9]. Cette dernière est spécialisée pour le TAL[10] mais ne fonctionne malheureusement que dans cette version de Python.

- Elles doivent récupérer des zones ciblées des différents sites Web : nous avons décidé de ne garder que les informations pertinentes telles que la date des différents articles, leur lien Internet, leur titre mais aussi et surtout leur contenu. Ces informations ainsi récupérées sont stockées dans des fichiers au format XML[11]. Ce choix de stockage est motivé par une spécificité du format XML : les informations peuvent être stockées en arborescence. Nous avons choisi de représenter une source d'information comme étant une racine de cette arborescence. Chaque racine possède des nœuds filles qui correspondent ni plus ni moins à une information : une

[9] Natural Language ToolKit : http://nltk.org
[10] Traitement Automatique des Langues
[11] Extensible Markup Language : http://www.w3schools.com/xml

information correspond à un article, son lien, sa date, son auteur, son titre, son contenu ainsi que son auteur.

Ainsi, une source d'informations peut être illustrée par le schéma en Figure 3.

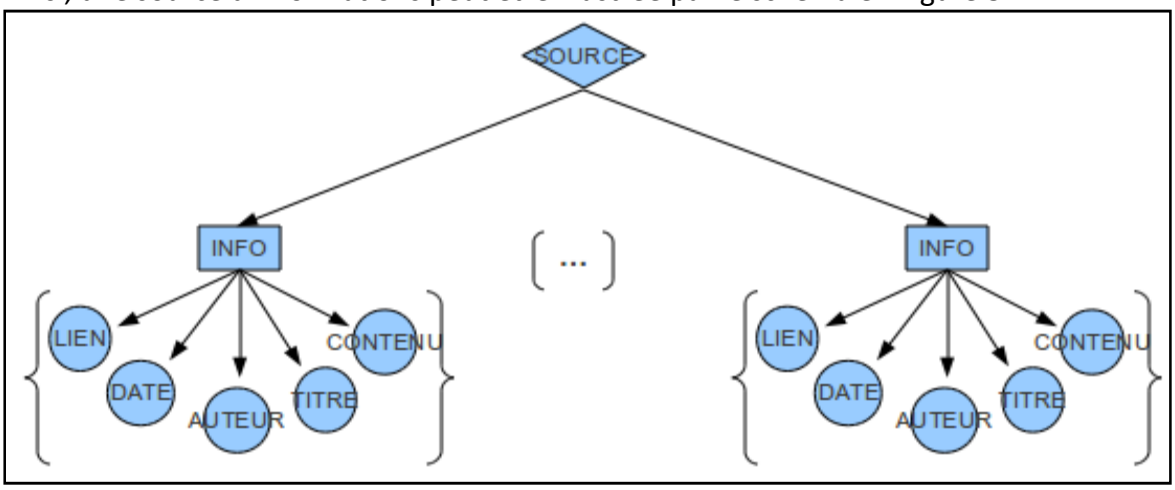


Figure 3 : Source d'information

À partir des fichiers au format XML, les APIs doivent pouvoir récupérer les informations selon des critères bien définis : il devra être possible de récupérer les informations par source, date, période. Il devra également être possible de récupérer uniquement les titres, ou encore les contenus, ou finalement les titres et les contenus à la fois, chacune de ces formes de la donnée pouvant être porteuse d'informations différentes et donc de faire l'objet de traitements différents.
Pour pouvoir satisfaire ces spécifications, plusieurs étapes bien distinctes ont été réalisées.

Première étape : étude de la source d'informations

Dans un premier temps, il est important de connaître la source d'informations. Pour le projet nous concernant, cette source d'informations correspond à un site Web (sites de campagne, blogs, presses sur Internet).
Ainsi, la vraie problématique à cette étape du processus de réflexion est de comprendre ce qu'est un site Web et comment une page Web est structurée. La description basique d'un site Web serait : un site Web est un ensemble de pages Web liées les une aux autres par des liens hypertextes. Ces pages Web sont au format HTML.

Deuxième étape : comment cibler les informations à récupérer ?

La première étape de la réflexion a permis de dégager une information importante : les pages web sont écrites dans le langage de balises HTML[12]. Un fichier HTML est organisé sous forme d'une arborescence.
Il devient donc assez aisé de repérer les informations qui nous intéressent : pour un site Web donné, pour peu qu'il ait été créé à l'aide de logiciels de création de sites Web, le même type d'information (par exemple un titre, une date, un certain contenu) correspond à un « pattern » bien défini, c'est-à-dire à une structure récurrente qui permettra d'identifier le type d'information. Une fois l'information ciblée, il suffit alors de l'extraire.

[12] Les informations sont contenues dans des balises où chacune correspond à un rendu visuel et/ou hiérarchique.

Troisième étape : choix des outils

Maintenant que l'on sait ce qu'on souhaite récupérer, et comment, il nous appartient choisir les outils pour réaliser la tâche.

Dans un premier temps, nous avons décidé d'utiliser une librairie[13] incluse dans Python 2.7 : la librairie HTMLParser. Un problème lié à cette librairie s'est alors posé : elle s'avère très limitée pour réaliser la tâche et difficile à mettre en œuvre.

Il a donc fallu trouver une autre librairie pour pouvoir manipuler les pages HTML et donc d'en extraire l'information. Sous les conseils de nos encadrants, nous avons opté pour la librairie LXML. Cette librairie Python, utilisable gratuitement sous licence BSD[14], présente de multiples avantages :

- Elle est très bien documentée ;
- Elle est facile à utiliser ;
- Elle est performante et est peu gourmande en ressources ;
- Elle permet de manipuler à la fois les documents HTML et XML ;
- Elle permet de manipuler à la fois en lecture et en écriture.

Les deux derniers points ont également été décisifs dans le choix de la méthode de stockage des données : la même librairie serait donc utilisée pour l'extraction, le stockage et la restitution de l'information. Nous avons fait ce choix pratique.

Dernière étape : création des APIs

Une fois les connaissances et les outils décidés, nous avons créé les APIs pour chaque source d'informations.

Il est cependant important de remarquer que d'une part, les articles de chaque source d'information ne sont pas tous datés ; et que d'autre part, les sites de campagne de certains candidats (notamment Nicolas Sarkozy) ne sont apparus qu'au dépôt officiel de candidature de ces derniers et sont devenus inactifs à des moments différents. Ainsi, nous avons des APIs pour les sources d'informations suivantes, ainsi que les périodes de couverture d'informations de chacune d'elles :

- le site de campagne du candidat François Bayrou : bayrou.fr
  - période de couverture : du 01/01/2012 au 13/05/2012
- le site de campagne du candidat Eva Joly : evajoly2012.fr
  - période de couverture : du 01/01/2012 au 15/04/2012
- le site de campagne du candidat François Hollande : francoishollande.fr
  - période de couverture : du 01/01/2012 au 29/04/2012
- le site de campagne du candidat Jean-Luc Mélenchon : www.placeaupeuple2012.fr
  - période de couverture : du 01/01/2012 au 22/04/2012
- le site de campagne du candidat Nicolas Sarkozy : www.lafranceforte.fr
  - période de couverture : du 04/03/2012 au 29/04/2012
- le site de presse Le Monde : www.lemonde.fr
  - période de couverture : du 01/01/2012 au 13/05/2012

[13] Une librairie informatique est un ensemble de fonctions utilitaires, regroupées et mises à disposition afin de pouvoir être utilisées sans avoir à les réécrire. *Wikipédia*

[14] La licence BSD (Berkeley Software Distribution license) est une licence libre utilisée pour la distribution de logiciels. Elle permet de réutiliser tout ou une partie du logiciel sans restriction, qu'il soit intégré dans un logiciel libre ou propriétaire. *Wikipédia*

- le site de presse Le Figaro : elections.lefigaro.fr
  - période de couverture : du 19/02/2012 au 13/05/2012
- le blog de centre : heresie.hautetfort.com
  - période de couverture : du 01/01/2012 au 13/05/2012
- le blog de gauche : www.jegoun.net
  - période de couverture : du 01/01/2012 au 13/05/2012
- le blog de droite : h16free.com
  - période de couverture : du 01/01/2012 au 13/05/2012

(4) Influence sur le choix des sources

La difficulté liée à la création des APIs a pesé lourd dans le choix des sources : cette difficulté est étroitement liée à la structure HTML du site concerné, mais également à l'accessibilité de ce dernier. En effet, certains sites ont une structure très disparate que la théorie de la récurrence des « patterns » que nous avons formulé dans les précédents paragraphes est faussée. Certains sites également, empêche des programmes automatiques (tels que le notre) d'accéder au code source de leurs pages. C'est par exemple le cas du site de campagne de Marine Le Pen (www.marinelepen2012.fr). Ces raisons font que de nombreux sites restent inexploitables.

e. Schéma d'arrivée

Après avoir détaillé nos contraintes techniques d'accès aux sites, voici la représentation taxonomique de notre classification finale.
On y distingue toujours nos 3 plateformes, politique, publique et professionnelle ainsi que les sites sélectionnés sur leurs aspects fonctionnels et techniques, ne restent que les sites avec lesquels nous avons réellement travaillé.

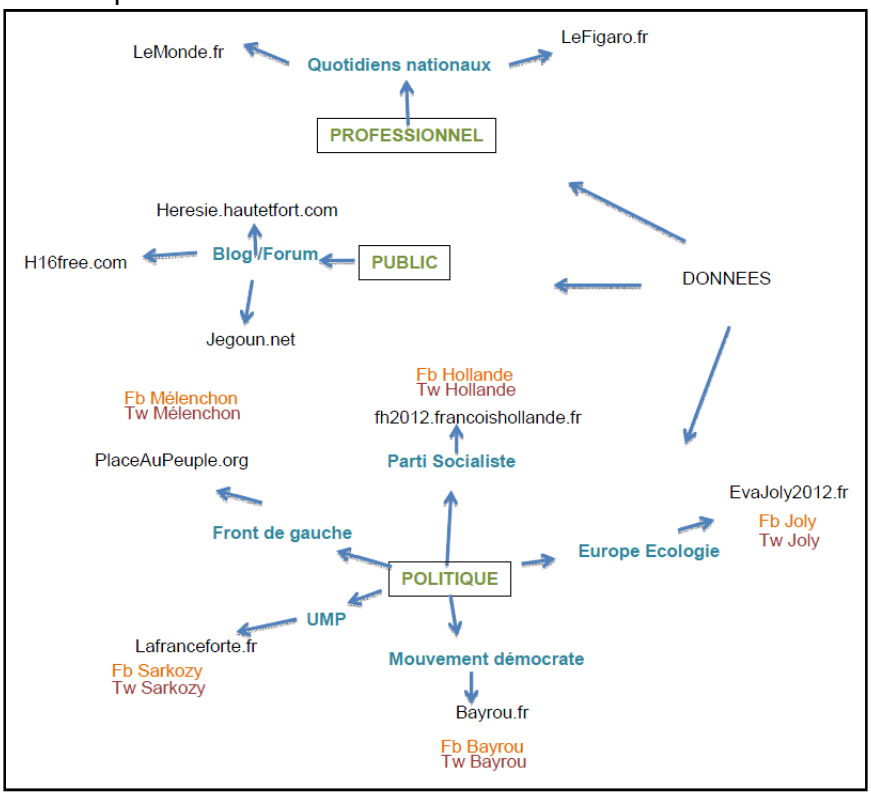


Figure 4 : Blogosphère politique cible

2. Moyens mis à disposition

a. Outils de formalisation des données

(1) Lemmatiseur

Au début du projet, tous les travaux de statistiques étaient effectués sur du texte brut, c'est-à-dire du texte avec les mots pris tels qu'ils se présentent dans le texte. S'est alors posé un problème : les statistiques comptabilisaient les mots « président » et « présidents » comme étant deux mots différents, alors que le second est en fait la forme plurielle du premier. Il fallait donc un outil, qui pour les mots bruts du texte – dits sous forme fléchie –, retrouve chaque lemme afin que les cas similaires à l'exemple précédent ne soient pas source d'erreur.
On a donc mis à notre disposition le lemmatiseur de l'équipe SEMAGRAMME du LORIA de Nancy. Cet outil crée à partir d'un fichier au format texte contenant du texte brut un fichier contenant du texte lemmatisé : ce texte contient alors chaque mot du texte brut initial, enrichi d'informations telles que les lemmes possibles du mot, avec leur catégorie grammaticale. Ainsi, pour le mot brut « polémiquent », on obtient grâce au lemmatiseur son lemme « polémiquer » qui est un verbe à l'infinitif.
Le lemmatiseur permet en outre de reconnaître les noms propres, les noms de lieux. Cela nous a permis d'envisager de nouvelles applications pour nos études.
Cependant, le lemmatiseur ne tient pas compte du contexte dans lequel un mot est dit. Ainsi, pour certains mots, il y a ambigüité : par exemple, tous les lemmes possibles du mot « invité » sont « inviter » en tant que verbe, « invité » en tant qu'adjectif et « invité » en tant que nom. Pour lever cette ambigüité, nous avons décidé de ne retenir que le premier lemme du mot et ignorer les autres, conscients que cela pourrait avoir une influence sur nos résultats.

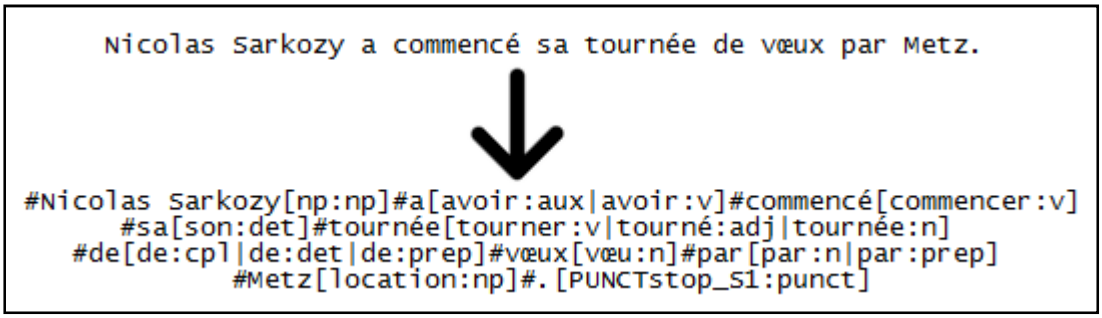


Figure 5 : Lemmatisation

En Figure 5, nous avons le résultat de la lemmatisation de la phrase « Nicolas Sarkozy a commencé sa tournée de vœux par Metz ». On constate que la lemmatisation a reconnu les noms de personnes et de lieux, mais donne également tous les lemmes possibles pour chaque mot.

(2) WOLF

Un des buts de ce projet est de voir l'évolution des thématiques tout au long de la campagne présidentielle. Or, jusque là, nos études statistiques ne portent que sur les mots – les lemmes – et non sur des thèmes.
Il nous fallait alors un outil qui permettrait d'agréger plusieurs mots autour d'un sens commun. Un moyen est de trouver des groupes de mots synonymes, où chaque groupe correspondrait à un sens.

Pour ce faire, nous avons utilisé WOLF[15], un lexique issu des travaux de l'équipe ALPAGE du laboratoire INRIA de Paris 7.
Une surinformation peut alors survenir, liée à la structure interne de WOLF : certains groupes de synonymes ont des sens tellement proches que même les êtres humains ne pourraient les distinguer. Par exemple, les groupes de synonymes [gouvernement, état] et [gouvernement, état, pouvoir] sont très proches sémantiquement, alors que WOLF les distingue. Pour résoudre ce conflit, nous avons pondéré les groupes de synonymes par rapport au nombre de mots qu'ils contiennent : ainsi, un plus grand groupe de synonymes aura tendance à prendre le pas sur un plus petit groupe de synonymes. De ce fait la thématique qui émergera sera celle qui agrège le plus de synonymes.
Cependant, au fil du projet, nous avons décidé de retenir des traitements et d'en exclure certains. Les traitements utilisant WOLF font partie de ceux que nous avons exclu, notamment parce que les résultats sont peu concluants.

b. Serveur

(1) Stockage des données

Une des problématiques de ce projet est la notion de stockage des informations récupérées des différentes sources d'informations. En effet, une quantité assez importante de texte est récupérée d'Internet. Il faudrait pouvoir les conserver pour une utilisation ultérieure.
Pour résoudre ce problème, nos encadrants ont mis à notre disposition une machine distante située au LORIA : cette machine s'appelle Quine. Quine nous sert de serveur afin que nous puissions stocker aussi bien les scripts des traitements que les résultats des extractions.

(2) Périodicité des traitements

La récupération des informations sur Internet est automatisée. Pour mettre en place cette tâche automatique, nous utilisons le serveur mentionné dans le paragraphe précédent. Des traitements périodiques sont également mis en place sur Quine.
En effet, cette dernière offre la possibilité de mettre en place des tâches automatiques dont on peut définir la périodicité.

c. Module d'extraction des données

Les traitements des données de cette étude doivent se faire de façon automatisée. Un module d'extraction de données a donc été créé afin de pouvoir être utilisé dans tous les traitements qui ont été mis en place.
Ce module permet dans un premier temps de récupérer les mots déjà lemmatisés, et dans un deuxième temps d'en faire un comptage brut.

(1) Filtrage

Lors de la récupération des mots des textes lemmatisés, on ne veut en retenir que les mots qui ont un sens et exclure les mots qui représentent sémantiquement peu d'intérêts en général.

[15] Le WOLF (Wordnet Libre du Français) est une ressource lexicale sémantique (wordnet) libre pour le français. *http://alpage.inria.fr/~sagot/wolf.html*

Ainsi, les mots outils (mots appartenant aux classes grammaticales fermées) tels que les pronoms, déterminants, prépositions, conjonctions ou auxiliaires ont été exclus car ne sont pas porteurs de sens dans le cadre de cette étude.
Seuls les mots lexicaux (mots appartenant aux classes grammaticales ouvertes) tels que les noms, les adjectifs, les verbes et les adverbes ont été retenus.

```
'un', 'une', 'des',
'le', 'la', 'les',
'du', 'de', 'des',
'au', 'à', 'aux',
'l', 'd',
'je', 'tu', 'il', 'elle', 'on', 'nous', 'vous', 'ils', 'elles',
'me', 'te', 'lui', 'leur',
'moi', 'toi', 'elle', 'eux',
'se',
'y', 'en',
'm', 'j', 't', 's',
'mon', 'ton', 'son', 'ma', 'ta', 'sa', 'notre', 'votre', 'leur',
'mes', 'tes', 'ses', 'nos', 'vos', 'leurs',
'ce', 'cet', 'cette', 'ces', 'c',
'mais', 'ou', 'et', 'donc', 'or', 'ni', 'car',
'ne', 'n', 'que', 'qui', 'qu', 'comme', 'dans', 'plus', 'donc', 'rien', 'sinon', 'aucun',
'jamais', 'pour', 'ainsi', 'parce', 'depuis', 'puis', 'puisque', 'où', 'là', 'ici', 'pas', 'avec', 'tout',
'si', 'peu', 'quand', 'jusqu', 'autre', 'autres', 'ceci', 'cela', 'celà', 'ça', 'alors',
'tout', 'toute', 'tous', 'toutes', 'aussi', 'entre', 'dont', 'sur', 'très', 'non',
'même', 'à', 'de', 'pour', 'sans', 'par',
'quel', 'quels', 'quelle', 'quelles',
'soi',
'être', 'suis', 'es', 'est', 'sommes', 'êtes', 'sont',
'étais', 'était', 'étions', 'étiez', 'étaient',
'fus', 'fut', 'fûmes', 'fûtes', 'furent',
'serai', 'seras', 'sera', 'serons', 'serez', 'seront',
'serais', 'serait', 'serions', 'seriez', 'seraient',
'sois', 'soit', 'soyons', 'soyez', 'soient',
'fusse', 'fusses', 'fût', 'fussions', 'fussiez', 'fussent',
'étant', 'été', 'étée', 'étées', 'étés',
'avoir', 'ai', 'as', 'a', 'avons', 'avez', 'ont',
'avais', 'avait', 'avions', 'aviez', 'avaient',
'eus', 'eut', 'eûmes', 'eûtes', 'eurent',
'aurai', 'auras', 'aura', 'aurons', 'aurez', 'auront',
'aurais', 'aurait', 'aurions', 'auriez', 'auraient',
'aie', 'aies', 'ait', 'ayons', 'ayez', 'aient',
'eusse', 'eusses', 'eût', 'eussions', 'eussiez', 'eussent',
'ayant', 'eu', 'eue', 'eues'
```

Figure 6 : Les mots exclus

En Figure 6, la liste des mots exclus lors de la récupération des données.

(2) Comptage

Après le filtrage, vient l'étape du comptage. Il s'agit plus précisément d'un comptage brut des mots retenus à l'étape précédente : pour chaque mot trouvé, on compte ses occurrences.
Cette étape a été réalisée à l'aide d'une bibliothèque Python performante pour cet exercice : NLTK[16].

(3) Informations extraites

Deux informations peuvent alors être extraites grâce à ce module :

- D'une part, tous les mots lexicaux avec leurs occurrences dans le corpus source donné,
- D'autre part, la taille du corpus source donné.

Ces informations seront cruciales dans la suite, où des traitements statistiques plus poussés sont mis en œuvre.

Les fonctions qui composent ce module sont présentées en annexe 3.

[16] Natural Language Toolkit (NLTK) est une bibliothèque logicielle en Python permettant un traitement automatique des langues. *Wikipédia, http://nltk.org/*

C. Traitements sur les données

Maintenant que la chaîne des traitements a été décrite, nous nous attachons dans cette partie à montrer l'intérêt de notre projet à travers l'exploitation des données que nous allons réaliser à travers différentes études.

1. Le Top 20

Chacun des 10 candidats à l'élection présidentielle française a préparé un programme pour présenter ses idées. Derrière des thématiques propres à chaque parti, nous avons souhaité savoir s'il était possible de faire ressortir des mots, utilisés par tous qui pourraient par la suite faire émerger des similitudes entre programmes. C'est cette liste de 20 mots que nous appelons le Top 20.

Pour ce faire, nous avons récupéré tous les programmes des candidats sur les sites de campagne et les avons soumis à la lemmatisation pour avoir la possibilité de traitements identiques grâce à leur nouvelle présentation homogène.

Après avoir été traité par le module, tous les mots lemmatisés ont été parcourus de manière à les insérer dans une base de données avec leur fréquence totale et le nombre de programme dans lesquels ils sont présents. Une fois trié, ce tableau présente de manière ordonnée, les mots les plus utilisés en commençant par ceux présents dans tous les programmes. Pour coller au plus prêt des textes, il a été fait une différence pour les mots « devoir » et « pouvoir » de manière à pouvoir faire la part des choses entre le verbe et le nom.

Le Tableau 1 présente le résultat brut du traitement :

		Nombre	Programme			Nombre	Programme			Nombre	Programme
1	organization	860	10	11	grand	281	10	21	pouvoir_n	204	10
2	Public	530	10	12	national	259	10	22	loi	188	10
3	devoir_v	496	10	13	état	247	10	23	date	184	10
4	Politique	440	10	14	pays	245	10	24	travail	180	10
5	Social	359	10	15	service	243	10	25	financier	174	10
6	Plaire	359	10	16	contre	231	10	26	tout	173	10
7	Mettre	339	10	17	permettre	228	10	27	place	172	10
8	Droit	316	10	18	français	220	10	28	location	159	10
9	entreprendre	297	10	19	nouveau	215	10	29	aussi	155	10
10	pouvoir_v	295	10	20	emploi	208	10	30	économique	149	10

Tableau 1 : Top 30

De cette liste, on doit retirer quelques termes avant de parler de liste définitive. En effet certains termes ne sont pas significatifs ou sont vides sens. Pour être exploitée, cette liste doit présenter des termes significatifs hors contexte.

Organization : ce terme est donné par le lemmatiseur pour tout nom ou abréviation. On peut donc l'exclure de la liste, il n'est pas significatif.

Plaire : ce terme est donné par le lemmatiseur pour les mots de la famille du verbe plaire et donc aussi pour le terme « plus ». On peut donc l'exclure de la liste car il représente essentiellement la négation en « plus » et non le verbe plaire. Ce n'est pas significatif.

Mettre, grand, contre, nouveau, tout, place, aussi : ces termes ne sont pas révélateur de sens hors contexte, ils ne sont donc pas gardés car non significatifs.

Date : ce terme est donné par le lemmatiseur pour les toutes les dates. On peut donc l'exclure de la liste, il n'est pas significatif.

Location : ce terme est donné par le lemmatiseur pour les tous les lieux. On peut donc l'exclure de la liste, il n'est pas significatif.

Voici, en Tableau 2, le résultat définitif, le top 20 :

		Nombre	Programme			Nombre	Programme
1	Public	530	10	11	service	243	10
2	verbe devoir	496	10	12	permettre	228	10
3	Politique	440	10	13	français	220	10
4	Social	359	10	14	emploi	208	10
5	Droit	316	10	15	nom pouvoir	204	10
6	entreprendre	297	10	16	loi	188	10
7	verbe pouvoir	295	10	17	travail	180	10
8	National	259	10	18	financier	174	10
9	Etat	247	10	19	économique	149	10
10	Pays	245	10	20	crise	144	10

Tableau 2 : Top 20

On retrouve dans ce tableau les grandes questions de société.

2. Proximité thématique dans les programmes

Le traitement 1 a permis de mettre en évidence les 20 mots dont la fréquence est la plus élevée, tous programmes confondus.
Ce top 20 peut-il être révélateur de proximité dans les programmes des candidats ? Est ce que nous pouvons, à travers l'utilisation de ces mots, retrouver les tendances politiques, telles que la gauche, la droite ou le centre ? C'est à ces questions que nous allons tenter de répondre.
La source utilisée est le tableau récapitulatif du comptage des mots dans chaque programme. Il s'agit du top 20 plus détaillé. Il y a une ligne par mot soit 20 lignes et une colonne par programme soit 10 colonnes. Les croisements représentent le comptage du mot dans le programme.

	NPA	LO	F DE GAUCHE	PS	EELV	MODEM	DEBOUT	UMP	FN	PROGRES
crise	43	25	17	1	13	4	1	5	33	2
devoir_v	30	8	34	1	42	30	3	75	232	41
droit	39	13	95	11	41	19	4	18	68	8
économique	6	6	17	4	14	4	4	21	62	11
emploi	15	8	59	15	4	5	7	22	69	4
entreprendre	21	35	68	21	7	8	7	31	94	5
état	12	1	14	2	18	4	6	11	164	15
financier	8	9	56	5	18	1	3	9	30	35
français	9	2	13	6	6	14	4	32	126	8
loi	7	3	62	15	12	5	4	4	62	14
national	10	5	52	5	20	6	4	18	128	11
pays	15	10	30	5	13	28	6	40	84	14
permettre	11	11	48	10	24	7	3	23	78	13
politique	28	35	90	7	38	17	6	30	174	15
pouvoir_n	12	28	33	3	28	17	4	10	57	12
pouvoir_v	21	43	33	6	18	31	4	29	78	32
public	32	6	189	23	26	1	6	43	167	37
service	14	5	74	13	10	2	4	16	100	5
social	33	17	135	18	20	2	6	44	81	3
travail	42	24	44	4	5	7	2	17	31	4

Tableau 3 : Détail Top 20

Pour essayer de trouver des caractéristiques partagées entre les programmes à travers l'utilisation propre de mot commun, il nous est apparu opportun d'effectuer une analyse en composante principale.

Pour ce faire, nous avons travaillé avec le logiciel R[17], qui permet à partir de données telles que notre tableau, d'effectuer ce type d'analyse et de produire des graphiques appropriés. L'utilisation de R pour cette analyse a été possible avec les packages Ellipse, ScoutterPlot3D et FactoMineR.

Après avoir intégré les données dans le logiciel, nous avons effectué successivement les étapes suivantes (cf. annexe 1 : console R avec la liste des commandes exécutées):

- créer les représentations graphiques,
- calculer les valeurs propres entre les facteurs,
- calculer les valeurs pour les individus,
- calculer les corrélations pour les variables.

Plusieurs graphiques sont nécessaires à la compréhension des résultats.

En effet les valeurs propres indiquent que dans notre cas, pour avoir une bonne représentativité des résultats (80% sont généralement demandées en statistique), il faudrait avoir une représentation en 4 dimensions. Or, ceci est très difficilement représentable et

[17] http://www.r-project.org

analysable sur papier. Nous avons donc opté pour conserver 3 dimensions ce qui nous permet d'atteindre presque 70% de représentation.

Cela implique donc d'avoir 4 schémas :

- Deux schémas représentants les individus (les mots)
- Deux schémas représentants les variables (les programmes)

A chaque fois, nous avons un schéma avec l'axe 1 et l'axe 2 qui permet une lecture classique en 2D, et un schéma avec l'axe 1 et l'axe 3 qui permet de donner de la profondeur ou de l'épaisseur aux informations et autorise ainsi une interprétation en 3D.

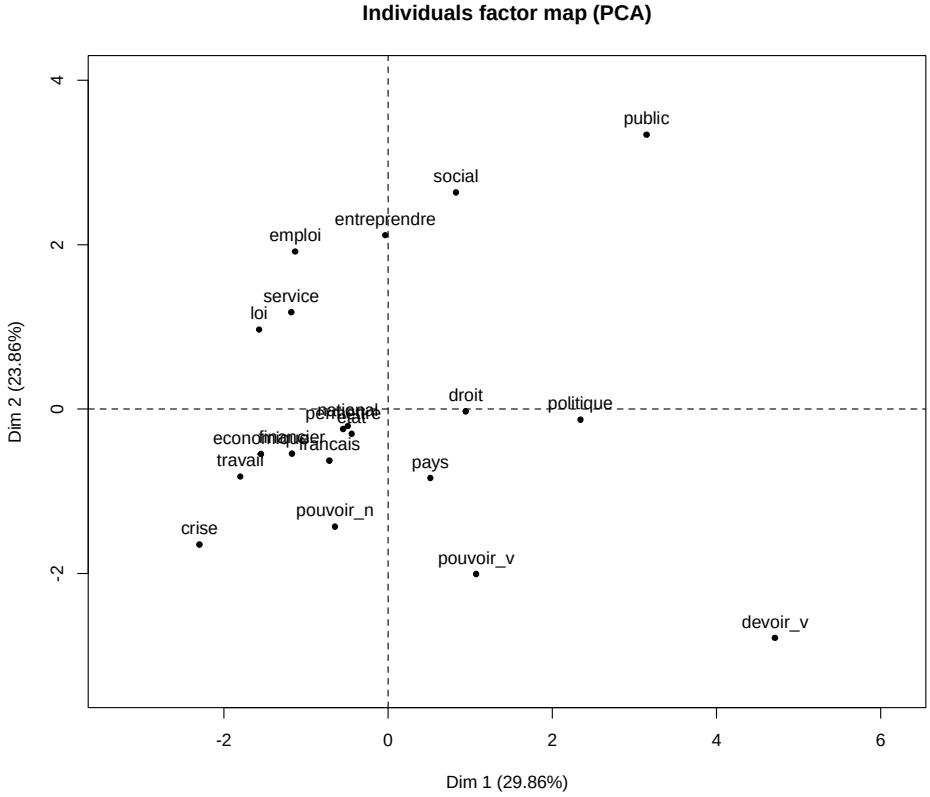
Voici les graphiques obtenus avec nos données.

- Plus les données sont proches du centre, plus elles s'inscrivent dans la moyenne, seule les données en périphérie sont vraiment riche de sens.
- Plus les données sont proches les unes des autres, plus elles sont corrélées.

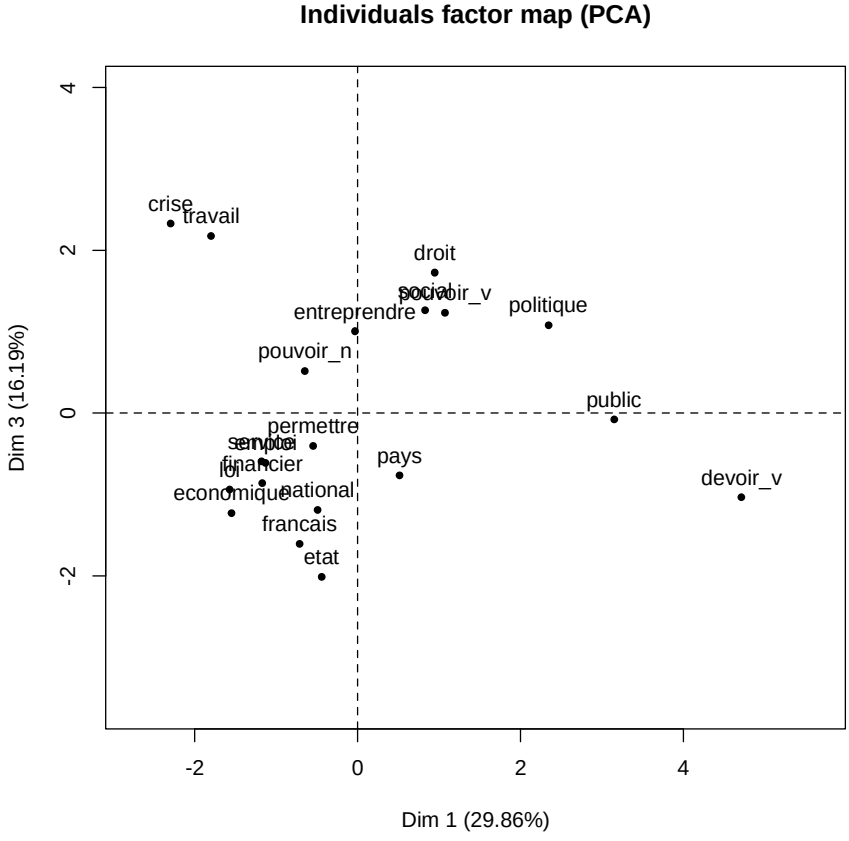
Pour les graphiques Individu1 et Programme1, on peut effectuer une lecture standard « à plat ». Par contre, les graphiques Individu2 et Programme2 apportant le relief, leur lecture ne se fait qu'en étant appliquée au premier graphique.

Exemple avec les individus « économique » et « travail » (situé en bas à gauche sur Individu1):
Ils sont très proches sur la dimension 1 (entre -1,5 et -2) et la dimension 2 (vers -1) sur Individu1 mais en regardant la dimension 3 sur Individu2, on constate qu'ils sont très éloignés (à 2 pour « travail » et -1 pour économique).
On peut alors imaginer sur le graphique 1, un troisième axe qui le traverserait d'avant en arrière, lui donnant ainsi de la profondeur et de l'épaisseur.
Sur ce troisième axe, le mot « économique » se situerait en profondeur -1 ; alors que le mot « travail » se situerait en épaisseur 2.
« Economique » est dans une vallée quand « travail » est sur une colline. On comprend alors que leur proximité reste très relative.

L'enjeu à ce niveau est de déterminer si cette représentation est fiable.



Individu 1



Individu 2

Analyse des contributions :
Les contributions, dont on trouve les valeurs en annexe 1, permettent de voir ce qui a une représentation significative. On détermine l'importance des concepts sur chaque axe.
Pour chaque individu, on dira qu'il a une représentation significative si sa contribution est supérieure à 5% car il y a 20 mots (100/20 = 5).

- Sur la dimension 1 de Individu1, on a une représentation significative pour crise, devoir_v, politique, public, travail,
- Sur la dimension 2 de Individu1, on a une représentation significative pour crise, devoir_v, emploi, entreprendre, pouvoir_v, public, social,
- Sur la dimension 3 de Individu2, on a une représentation significative pour crise, droit, état, emploi, français, travail.

On peut ainsi interpréter que la crise, l'emploi ou le travail sont des mots qui ont été très importants dans la constitution des programmes.

Analyse des cosinus2:
Les cosinus2, dont on trouve les valeurs en annexe 1, permettent de voir ce qui a une bonne représentation.
Pour chaque individu, on dira qu'il est bien représenté si son cosinus2 est proche de 1. Cette valeur peut être étudiée axe par axe ou en prenant conjointement 2 axes (voire en prenant les 3 axes)[18].

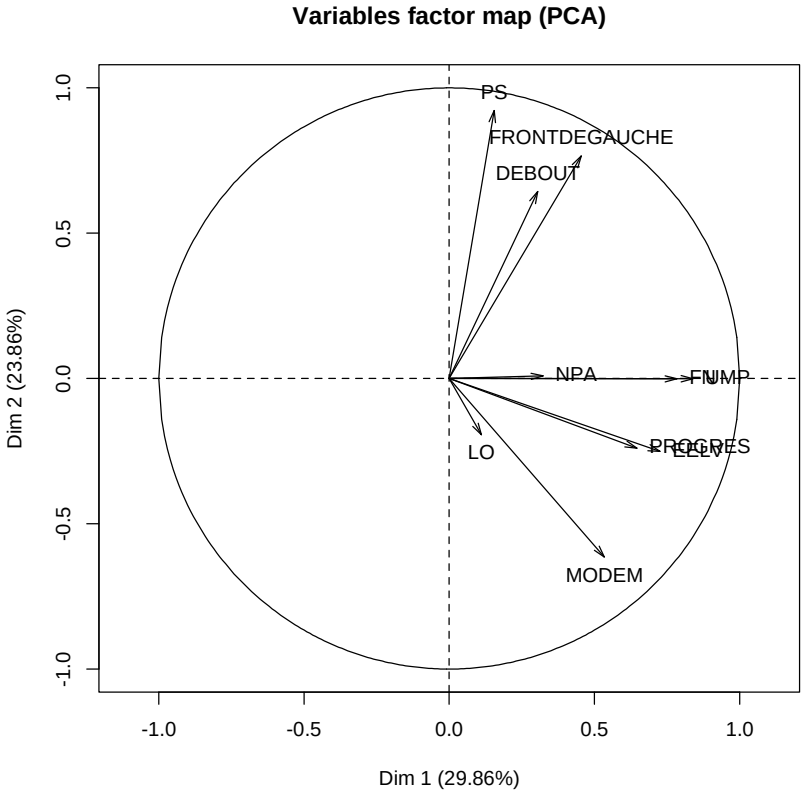
- Sur les axes 1 et 2, on a une très bonne représentation de devoir_v, et public.
- Sur les axes 1 et 2 et 3, on a une très bonne représentation de crise, économique, emploi, loi, service, social et travail. On a également une représentation assez bonne de état, français, national, politique, pouvoir_n et pouvoir_v

Les mots : droit, entreprendre, financier, pays et permettre, ne sont pas bien représentés sur ces 3 axes. Ils auraient gagné à être représentés sur les dimensions non exploitées ici.

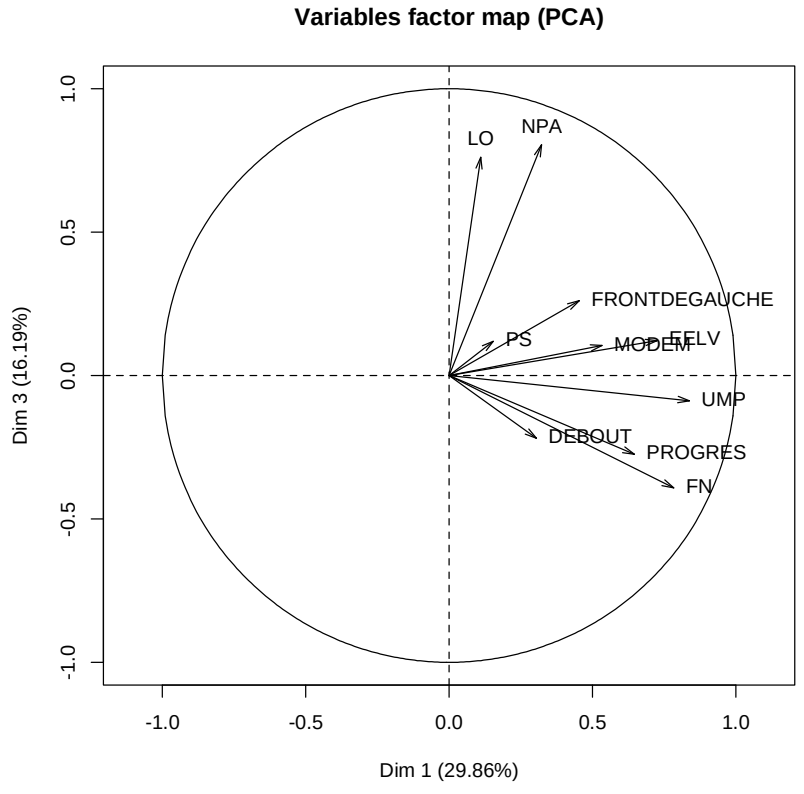
Ce que nous disent en plus les graphiques :
Les mots comme crise, travail, politique, droit, social, pouvoir_v devraient apparaître en relief sur Individu 1.
Les mots comme économique, loi, financier, devoir_v devraient apparaître en retrait sur le Individu 1.

Cela signifie que la proximité graphique entre les mots économique et travail doit être relativisée par le relief du troisième axe.
Cela peut également s'appliquer également aux mots français et pouvoir_n.

[18] Les valeurs s'additionnent quand il y a plusieurs axes



Programme 1



Programme 2

Analyse des corrélations :
Les corrélations, dont on trouve les valeurs en annexe 1, mettent en avant les variables liées. La corrélation est donnée entre 2 variables (plus elle est proche de 1, plus les variables sont liées).

Sur la base du Top 20, on a donc le programme du PS très corrélé avec le programme du front de gauche à 0,78 et le programme de l'UMP corrélé avec le programme du FN à 0,67. On trouve également une corrélation moyenne entre le programme du PS et le programme de Debout la République à 0,58.

Ce que nous disent en plus les graphiques :
Les programmes de LO, NPA, PS, Front de gauche, Modem et EELV devraient apparaître en relief sur le graphique 1.
Les programmes comme UMP, Debout la république, Progrès et Solidarité, FN devraient apparaître en retrait sur Programme1.

Cela signifie que la proximité graphique entre les programmes du PS, Front de gauche et Debout la République doit être relativisée par le relief du troisième axe.
Cela s'applique également aux programmes de Progrès et solidarité et de EELV.

3. Utilisation des mots du Top 20 dans un programme

Le traitement 1 a permis de mettre en évidence les 20 mots dont la fréquence est la plus élevée dans les programmes.
Nous allons ici pouvoir regarder si ces mots correspondaient réellement à une thématique abordée par un candidat ou si ce n'était que des effets d'annonce. Il s'agit de comparer la courbe de vie de plusieurs mots sur un site.

Deux sources ont été nécessaires à la réalisation de ce traitement.
La première source reprend tous les mots que nous avons récupérés via les APIs et qui ont été lemmatisés.
La seconde source correspond à la liste des mots composant le Top 20.

Cette liste est donnée en entrée du module. On récupère alors en sortie une base de données contenant uniquement les mots du Top 20 et leur nombre d'occurrence avec un détail par semaine et par site source.
Il est alors possible de sélectionner un site source au choix et de choisir plusieurs mots du Top 20.
Le résultat est alors présenté sous forme d'une courbe avec en abscisse le temps et en ordonnée la part d'occurrence sur le site sélectionné.

En Figure 7, un exemple : entreprendre et pays sur le site de Hollande.

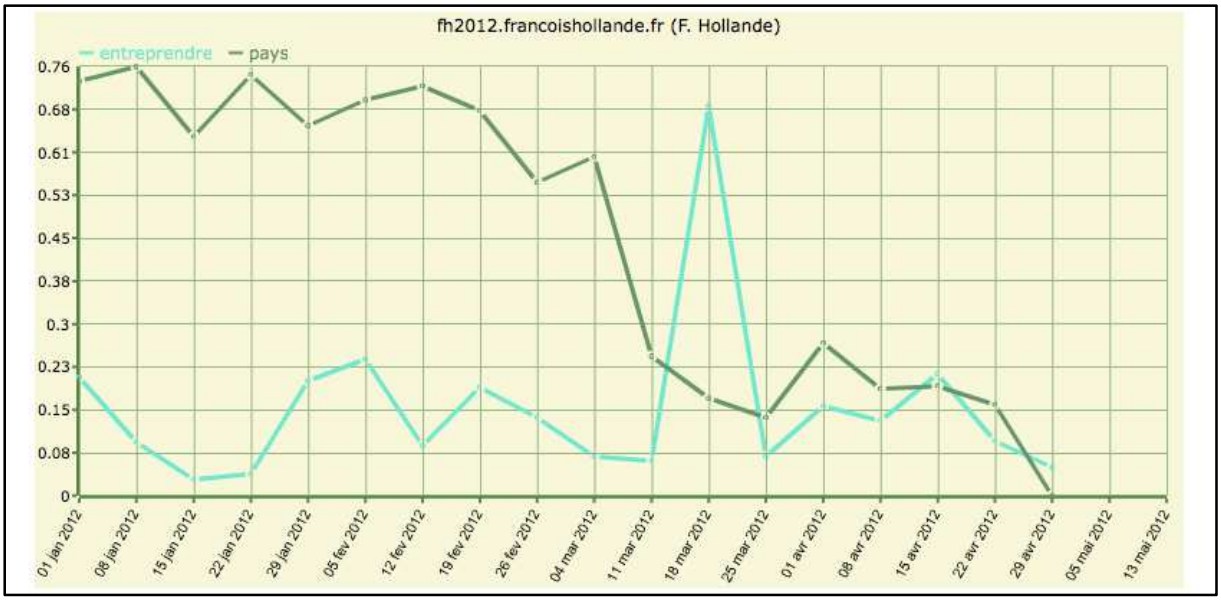


Figure 7 : Les mots « entreprendre » et « pays » dans le site de campagne de F. Hollande

Le mot pays sur le site de Hollande, a été très présent sur la première partie de la campagne avant d'en disparaître presque complètement.
Le mot entreprendre sur le site de Hollande n'a presque jamais existé sauf la semaine du 18 mars.

4. Utilisation d'un mot du Top 20 dans les programmes

Nous allons regarder comment est-ce que les mots du Top 20 ont vécu au fil des semaines de la campagne présidentielle. Il s'agit de comparer la courbe de vie d'un mot sur plusieurs sites.

Pour ce traitement, les mots récupérés via les APIs constituent la première source et les mots composant le Top 20 constitue la seconde source.

Le Top 20 est déjà traité par le module. On récupère alors en sortie une base de données contenant ces mots avec leur nombre d'occurrence pour les semaines, par site.
Il est alors possible de sélectionner un mot du Top 20 au choix et de choisir plusieurs sites sources.
Le résultat est présenté sous forme d'une courbe avec en abscisse le temps et en ordonnée le pourcentage d'utilisation du mot sur chaque site.

En Figure 8 : social sur les sites de Hollande et Mélenchon.

Figure 8 : Le mot « social » dans les sites de campagne de F. Hollande et J-L. Mélenchon

On voit que cette thématique a été traitée en parallèle par les deux candidats pendant plus de 14 semaines.

5. Mesure de l'activité des sites

Quel type d'information peut-on tirer du volume de mot récupéré sur les sites ? Cela peut-il être révélateur de l'activité des équipes de campagne ?

Pour ce traitement, on reprend tous les mots que nous avons récupérés via les APIs et qui ont été lemmatisés.

Après l'action du module sur la source, tous les mots ont été comptés, site par site, semaine par semaine.
Il est alors possible de choisir plusieurs sites sources.
Le résultat est présenté sous forme d'une courbe avec en abscisse le temps et en ordonnée le volume de mots.

Le graphique en Figure 9 présente le volume de mots dans le temps sur les sites de Hollande, Mélenchon, Joly, Bayrou et Sarkozy.

Figure 9 : Volume de mots sur les sites de campagne des candidats

A part sur le site de Mélenchon qui est constant, on s'aperçoit que les volumes de mot sur les sites sont très fluctuants. Le flux saisi sur les sites de campagne n'est pas homogène. Il y a des semaines ou il y a moins de choses à dire que d'autres.
On remarque aussi que certains candidats ont pris un peu de recul avec la campagne avant le premier tour des élections (comme Bayrou ou Hollande) alors que d'autres ont continué à s'exposer (comme Sarkozy).

6. Baromètre de popularité

Est-ce possible de savoir qui à parler de qui sur les sites sources ? Est-ce que les candidats parlent d'eux et est-ce que les journaux parlent de tous les candidats de manière équilibrée ?

Ce traitement utilise tous les mots que nous avons récupérés via les APIs et qui ont été lemmatisés.

Après utilisation du module, seuls les noms des candidats sont retenus, formant ainsi une base de données avec le nombre d'occurrence, site par site, semaine par semaine.
Pour ce traitement, il est possible de choisir plusieurs sites.
Le résultat est alors présenté sous forme d'un histogramme avec en abscisse le nom des candidats et en ordonnée la part des mots.

En Figure 10 : les candidats mentionnés sur les sites de médias Le Monde et Le Figaro.

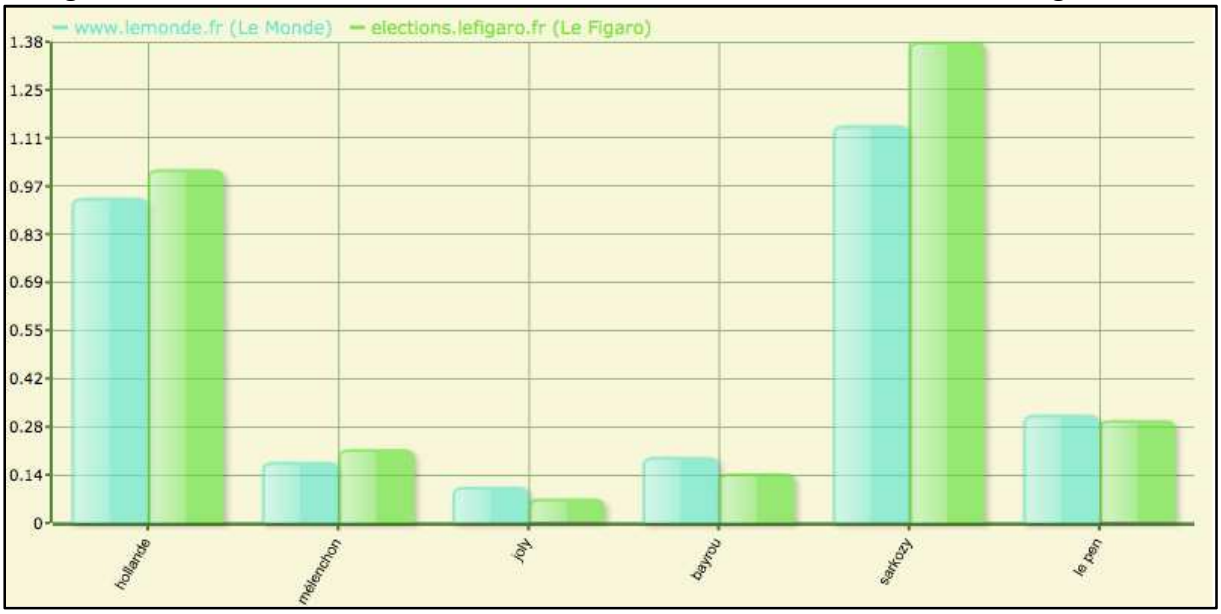


Figure 10 : Les candidats mentionnés dans Le Monde et Le Figaro

On remarque, que les deux sites ont des chiffres à peu prêt similaires. Ils ont beaucoup parlé de Sarkozy, un peu moins de Hollande et quasiment pas des autres candidats.

7. Thématique

Peut- on retrouver une thématique abordée sur une semaine par un candidat ? Voici un outil qui permet de cibler une semaine pour savoir ce qui s'est dit sur un site, à combiner avec l'outil de mesure de l'activité des sites.

Pour ce traitement, on a besoin de tous les mots que nous avons récupérés via les API.
Après l'utilisation du module, tous ces mots sont classés, site par site, semaine par semaine par nombre d'occurrence et on sélectionne alors les 5 mots les plus fréquents.

Dans le traitement, il est possible de choisir un seul site source et une seule semaine.
Le résultat est alors présenté sous forme d'un histogramme avec en abscisse les mots et en ordonnée le nombre d'occurrence.

La Figure 11 montre en exemple le site de Sarkozy entre les deux tours de vote.

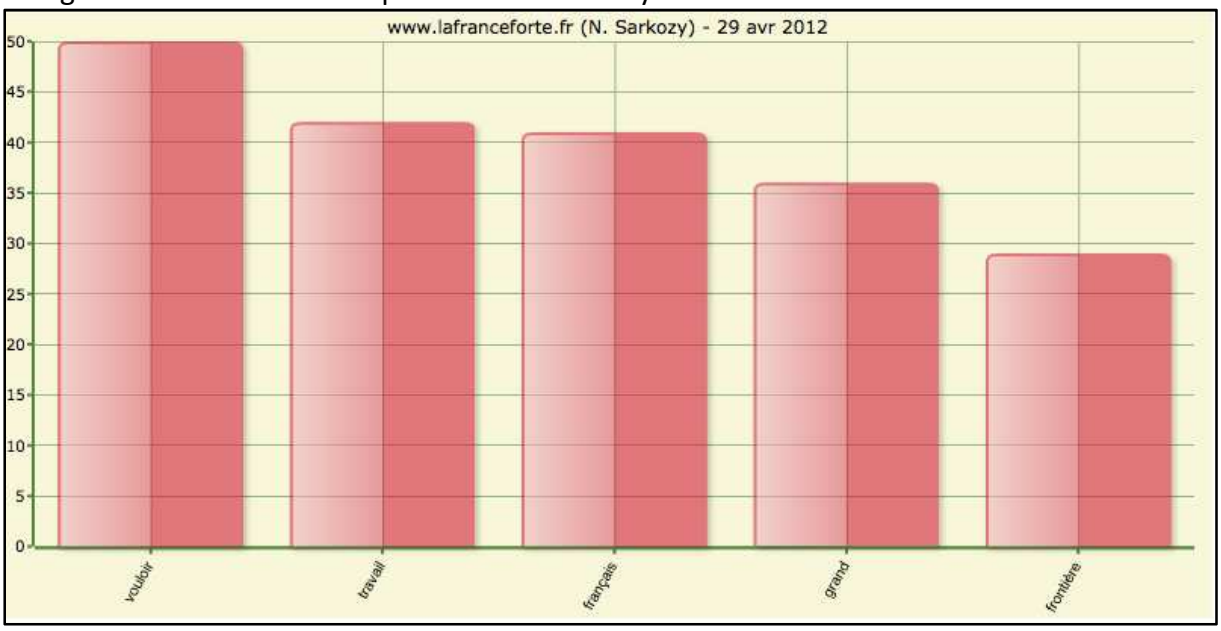


Figure 11 : Top 5 des mots sur le site de campagne de N. Sarkozy la semaine du 29 avril 2012

On observe les principaux mots écrits sur le site de Sarkozy après le premier tour des élections. Il est intéressant de regarder l'outil suivant avant d'aller plus loin dans la réflexion.

8. Libre choix

La conclusion de la mise en place de l'ensemble de ces traitements nous a montré qu'il était nécessaire de laisser l'utilisateur libre de son objet d'étude.
Voici un outil qui permet une saisie simple d'un mot pour regarder son évolution sur la période sur plusieurs sites.

En entrée du module, on a mis tous les mots que nous avons récupérés via les API et qui ont été lemmatisés puis on les a comptés, site par site, semaine par semaine.

Il est possible de saisir un mot au choix et de sélectionner les sites source.
Le résultat est alors présenté sous forme d'un graphique avec en abscisse les dates et en ordonnée la part des mots.

La Figure 12 présente en exemple les sites de Sarkozy et Hollande pour le mot « frontière ».

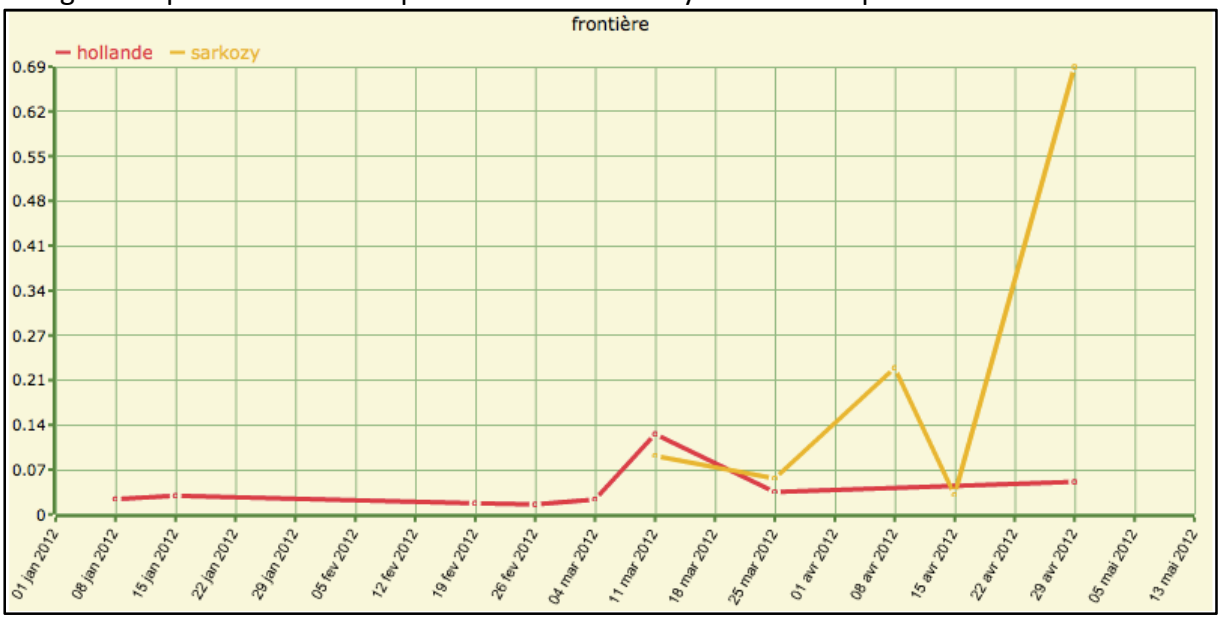


Figure 12 : Le mot frontière sur les sites de campagne de F. Hollande et N. Sarkozy

Ce mot était ressorti sur le traitement précédent. On voit sur ce graphique qu'il est représentatif d'une thématique abordée par Sarkozy à l'entre-deux tour et non repris par Hollande.

D. Diffusion des résultats

Pour diffuser nos résultats, nous avons créé un site Web via Internet qui présente le double avantage d'apporter une large diffusion de nos traitements et d'offrir la dynamique nécessaire à l'optimisation de nos résultats et des représentations graphiques.

1. Caractéristiques techniques

Les pages statiques ont été développées en HTML 5[19]. Les pages dynamiques ont été développées en PHP 5[20]. La feuille de style a été réalisée en CSS 3[21].
La base de données est composée de fichiers dans lesquels on a stocké mots et nombre d'occurrence par semaine et par source. Les graphiques ont été générés grâce à la librairie Open Flash Chart[22].
Ce site est hébergé par le LORIA, il fonctionne sur à un serveur APACHE[23].

2. Présentation du site

Une page d'accueil permet la présentation du projet par l'introduction de phrases clés mettant en avant le mot « MOT » et l'importance de son contenu dans le cadre de la campagne présidentielle (exemple : Si vous aviez votre **mot** à dire pour qualifier cette campagne, lequel choisiriez-vous ?).
Il est également mis en avant le nombre de mots récupérés et les sites sources pour apporter une crédibilité aux traitements proposés dans le menu. 6 liens permettent ainsi d'accéder aux études réalisées.

Le premier lien permet d'accéder aux traitements relatifs au Top 20 des mots les plus utilisés dans les programmes et d'étudier leurs intérêts dans la campagne. C'est notamment sur cette page que va être donnée la constitution du Top20 et les graphiques relatifs à son utilisation, sur un site ou sur plusieurs sites…
Le second lien propose d'accéder aux éléments permettant d'effectuer l'analyse en composante principale sur la proximité thématique.
Le troisième lien donne l'accès au graphique permettant de mesurer l'activité des sites.
Le quatrième lien permet d'accéder à la représentation du baromètre de popularité.
Le cinquième lien ouvre la porte de la thématique au fil des semaines.
Le sixième lien donne l'accès au traitement du libre choix.

[19] http://www.w3schools.com/html5/default.asp
[20] http://www.php.net
[21] http://www.css3.info
[22] http://teethgrinder.co.uk/open-flash-chart
[23] http://httpd.apache.org

1. Exemple de page

Figure 13 : Page d'accueil

Figure 14 : Page de présentation des traitements relatifs au Top 20 des mots les plus utilisés dans les programmes

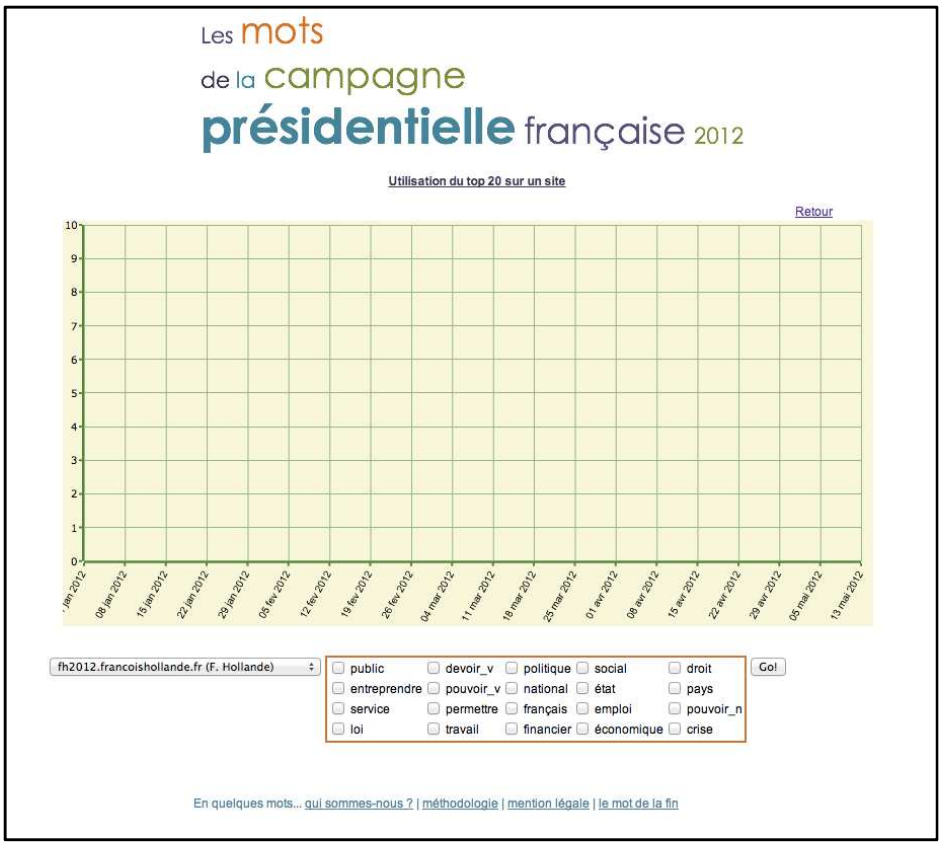


Figure 15 : Page de présentation du graphique d'utilisation du top 20 sur un site

Les autres pages du site ont été mises en annexe 2.

2. Protocole d'expérimentation

Nous avons demandé à trois personnes de naviguer sur le site et de nous faire un rendu de leurs impressions et commentaires. Il n'a été donné en amont ni consignes, ni guide utilisateur. Ces personnes n'avaient qu'une connaissance du contexte du projet sans en savoir les détails.
Le panel était composé d'un professionnel du développement Internet et de 2 internautes.

3. Résultat de l'expérimentation

a. Retours du panel

Le professionnel :
Tout de suite, il a envie de fermer. Il apprécie le graphisme du titre mais n'est pas séduit par le reste. Il y a trop de texte, trop de couleurs, trop de casses différentes.
Le découpage de la feuille est compréhensible.
Les titres avec les liens sont trop longs.
Il trouve les pages avec les graphiques intéressants mais l'accès n'est pas valorisé, le graphisme des tableaux est sympa.
Le nuage de mot n'est pas conventionnel (normalement, il y a moins de couleurs) mais c'est très accrocheur et ça devrait être en page d'accueil.
Il se demande quelle est la cible du site.
Il y a un problème de cohérence entre les couleurs du menu et les couleurs des pages d'arrivée (lorsque le lien est bleu, on s'attend à une ambiance bleue).

Il aurait fallu mettre des choses en avant pour mettre tout de suite en place l'ambiance.
Que fait le menu en bas à gauche ???

Internaute 1 :
Il trouve le texte sympa. Il est impressionné par l'encart avec les références chiffrées. Il trouve une faute d'orthographe sur la page d'accueil.
Il a du mal à comprendre les graphiques qui n'ont pas de légende sur l'axe des ordonnées.
Il trouve la page 3 intéressante grâce au tableau mais ne comprend rien à la proposition d'analyse en composante principale.

Internaute 2:
Il y a trop de couleur, l'écriture est trop petite, mais c'est beau.

Remarques communes :
On ne sait pas où cliquer. Ce qui est souligné n'est pas cliquable et ce qui n'est pas souligné est cliquable.

b. Axes d'amélioration

Il existe des règles d'accessibilité[24] communes sur Internet que le site ne respecte pas. Les liens doivent être soulignés (cela est notamment utile pour les personnes ayant un déficit visuel).
L'idéal aurait été d'avoir une feuille de style reflétant notre charte graphique dans laquelle on aurait pu dire que tous les liens cliquables étaient dans une même couleur.

Les noms des liens doivent être courts quitte à mettre un sous titre explicatif.
Il manque un titre au site.
Le menu doit être en haut ou à droite.

La page d'accueil doit être plus attractive. Elle doit donner envie de continuer. Il doit y avoir des mots clés et des images fortes.
Elle doit être également porteuse du « Look and Feel ».Il s'agit de la sensation que n'importe quel internaute a en arrivant sur une page. Les gens doivent savoir qu'ils sont sur un site qui va parler de mots utilisés pendant la campagne présidentielle et récupérés sur Internet.

c. Améliorations apportées sur le site

Nous avons pu prendre en compte les modifications que nous avons jugées prioritaires, même si toutes les remarques étaient pertinentes.

- Accessibilité : nous avons souligné tous les liens actifs et des-soulignées les phrases qui n'étaient que des titres.
- Compréhension : nous avons retiré des couleurs sur la page d'accueil et mis des légendes aux graphiques.

[24] http://www.w3.org/WAI

CONCLUSION

Ce projet tutoré s'est inscrit dans l'actualité des événements politiques français de ce premier semestre 2012.
On peut même dire que son sujet a été très tendance puisque plusieurs sites Internet ont été dédiés à cette thématique.

Notre problématique était de comprendre l'utilisation de la langue naturelle dans ce contexte particulier et d'en proposer une lecture à travers des représentations graphiques.

Nous avons donc commencé par étudier la blogosphère politique de manière à cibler au mieux les sites qui pouvaient être une source fiable d'informations.

Puis, nous avons mis en place les programmes informatiques permettant de récupérer périodiquement et automatiquement ces données et de normaliser leur format.

Ensuite, nous avons réalisé des traitements statistiques dont les résultats graphiques ont été mis en accès libre sur Internet.

Nous avons dû adapter notre rythme de travail au calendrier électoral de manière à appréhender le mieux l'entrée en campagne des différents candidats.
Nous avons également dû être en veille quasi-permanente sur les sites ciblés, afin de palier aux éventuelles modifications qui auraient rendues obsolètes nos traitements.

Notre travail tout au long de ce projet nous a permis d'aborder, en situation réelle, de nombreuses notions, découvertes depuis notre entrée en Master SCA telles que, la programmation pour le TAL, l'analyse de donnée, l'ergonomie cognitive en analyse ascendante, ou encore la représentation des connaissances.

Ce projet a également été l'occasion pour nous de mettre en pratique des connaissances que nous avions déjà, telles que, les processus de programmation informatique ou la gestion de projet.

Par contre, il nous a manqué de vraies connaissances en ergonomie appliquée au Web, celles-là même qui nous auraient permis de mieux structurer notre site.
Il nous a également manqué de temps, pour approfondir certaines des nos analyses et exploiter au mieux les traitements.

La pluridisciplinarité de ce projet nous a permis d'entrevoir les multiples facettes de la formation et de ses débouchés.

Enfin, il est important de souligner que le travail réalisé dans le cadre de la campagne présidentielle peut être également exploitable dans le cadre de la campagne législative qui débute mais également dans tout autre travail de collecte et d'analyse de données textuelles issues d'Internet.

Annexe 1

```
R version 2.14.1 (2011-12-22)
Copyright (C) 2011 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
Platform: i386-apple-darwin9.8.0/i386 (32-bit)

R est un logiciel libre livré sans AUCUNE GARANTIE.
Vous pouvez le redistribuer sous certaines conditions.
Tapez 'license()' ou 'licence()' pour plus de détails.

R est un projet collaboratif avec de nombreux contributeurs.
Tapez 'contributors()' pour plus d'information et
'citation()' pour la façon de le citer dans les publications.

Tapez 'demo()' pour des démonstrations, 'help()' pour l'aide
en ligne ou 'help.start()' pour obtenir l'aide au format HTML.
Tapez 'q()' pour quitter R.

[R.app GUI 1.43 (5989) i386-apple-darwin9.8.0]

[Espace de Travail restauré depuis /Users/ceciledeshayes/.RData]
[Historique recherché depuis /Users/ceciledeshayes/.Rapp.history]
> data=read.table("/Users/ceciledeshayes/Desktop/resultat2.txt",header=TRUE)
> data
            NPA LO FRONTDEGAUCHE PS EELV MODEM DEBOUT UMP  FN PROGRES
crise        43 25            17  1   13     4      1   5  33       2
devoir_v     30  8            34  1   42    30      3  75 232      41
droit        39 13            95 11   41    19      4  18  68       8
economique    6  6            17  4   14     4      4  21  62      11
emploi       15  8            59 15    4     5      7  22  69       4
entreprendre 21 35            68 21    7     8      7  31  94       5
etat         12  1            14  2   18     4      6  11 164      15
financier     8  9            56  5   18     1      3   9  30      35
francais      9  2            13  6    6    14      4  32 126       8
loi           7  3            62 15   12     5      4   4  62      14
national     10  5            52  5   20     6      4  18 128      11
pays         15 10            30  5   13    28      6  40  84      14
permettre    11 11            48 10   24     7      3  23  78      13
politique    28 35            90  7   38    17      6  30 174      15
pouvoir_n    12 28            33  3   28    17      4  10  57      12
pouvoir_v    21 43            33  6   18    31      4  29  78      32
public       32  6           189 23   26     1      6  43 167      37
service      14  5            74 13   10     2      4  16 100       5
social       33 17           135 18   20     2      6  44  81       3
travail      42 24            44  4    5     7      2  17  31       4
> res$eig
        eigenvalue percentage of variance cumulative percentage of variance
comp 1   2.9855317              29.855317                             29.85532
comp 2   2.3856124              23.856124                             53.71144
comp 3   1.6193921              16.193921                             69.90536
comp 4   1.0791290              10.791290                             80.69665
comp 5   0.6901602               6.901602                             87.59825
comp 6   0.5416685               5.416685                             93.01494
comp 7   0.3346800               3.346800                             96.36174
comp 8   0.1795310               1.795310                             98.15705
comp 9   0.1295391               1.295391                             99.45244
comp 10  0.0547559               0.547559                            100.00000
*** pour voir les individus sur le schéma (mot)
> plot(res,axes=c(1,3))
*** pour voir les variables sur le schéma (parti)
> plot(res,axes=c(1,3),choix="var")
> cor(data)
                    NPA           LO FRONTDEGAUCHE          PS        EELV       MODEM
NPA           1.00000000  0.412768646    0.37355540  0.07567583  0.29694529  0.12187007
LO            0.41276865  1.000000000   -0.00424598 -0.02296743  0.10592355  0.37536911
FRONTDEGAUCHE 0.37355540 -0.004245980    1.00000000  0.78675990  0.28942262 -0.29530550
PS            0.07567583 -0.022967431    0.78675990  1.00000000 -0.14922514 -0.37848942
EELV          0.29694529  0.105923553    0.28942262 -0.14922514  1.00000000  0.42666577
MODEM         0.12187007  0.375369107   -0.29530550 -0.37848942  0.42666577  1.00000000
DEBOUT       -0.20201407 -0.004107983    0.40954665  0.58375755 -0.05842993 -0.01897706
UMP           0.25628199  0.001855581    0.26655758  0.13923360  0.37907507  0.49408992
FN            0.02954083 -0.198214629    0.19386386  0.01047375  0.49928838  0.28961311
PROGRES      -0.06645060 -0.047624175    0.16572256 -0.09963198  0.47302872  0.35114022
```

```
                  DEBOUT         UMP           FN     PROGRES
NPA           -0.202014074 0.256281986  0.02954083 -0.06645060
LO            -0.004107983 0.001855581 -0.19821463 -0.04762417
FRONTDEGAUCHE  0.409546646 0.266557583  0.19386386  0.16572256
PS             0.583757551 0.139233601  0.01047375 -0.09963198
EELV          -0.058429931 0.379075070  0.49928838  0.47302872
MODEM         -0.018977058 0.494089921  0.28961311  0.35114022
DEBOUT         1.000000000 0.268771298  0.33519170 -0.06322243
UMP            0.268771298 1.000000000  0.67135578  0.45129711
FN             0.335191700 0.671355779  1.00000000  0.45997380
PROGRES       -0.063222432 0.451297107  0.45997380  1.00000000
> res$ind
$coord
                   Dim.1       Dim.2       Dim.3       Dim.4       Dim.5
crise        -2.29706950 -1.64900061  2.32784049 -1.17627610 -0.86306985
devoir_v      4.71271450 -2.78443952 -1.03403795 -0.70039329 -0.94858841
droit         0.94669819 -0.02733585  1.72433253 -0.97204334 -0.14032846
economique   -1.54810751 -0.54625328 -1.22832711 -0.06062105 -0.01584411
emploi       -1.13132266  1.91750745 -0.61161584  1.19901744 -0.37796989
entreprendre -0.03341084  2.11560711  1.00548471  2.14859427  0.12538896
etat         -0.44298868 -0.30081621 -2.01162404  0.06147042 -0.55252755
financier    -1.16999880 -0.54236638 -0.86055263 -1.44444369  1.92152079
francais     -0.71373571 -0.62690877 -1.60633331  0.62379781 -1.07062472
loi          -1.57098377  0.96762845 -0.93848681 -0.47543060  0.81309370
national     -0.48954286 -0.20572862 -1.18910234 -0.49551608 -0.26192616
pays          0.51507965 -0.83935461 -0.76474792  1.73478616 -0.39858224
permettre    -0.54617312 -0.24186085 -0.40329150 -0.63798332  0.33518736
politique     2.34491356 -0.13000477  1.07970679  0.68466727  0.18403094
pouvoir_n    -0.64742272 -1.43133981  0.51668816  0.54182182  1.05398934
pouvoir_v     1.07277465 -2.00800732  1.23099554  1.58503853  1.64817308
public        3.14998257  3.33941660 -0.07952897 -1.60145525  0.62533665
service      -1.17902415  1.17870104 -0.59319821 -0.43031628 -0.39223318
social        0.82736537  2.63698775  1.26279440 -0.02505581 -0.77369294
travail      -1.79974818 -0.82243180  2.17300401 -0.55965891 -0.91133332

$cos2
                    Dim.1        Dim.2        Dim.3        Dim.4        Dim.5
crise        0.3303108332 1.702222e-01 0.3392196350 8.661500e-02 4.663011e-02
devoir_v     0.6702615143 2.339794e-01 0.0322682179 1.480424e-02 2.715549e-02
droit        0.1015406734 8.466078e-05 0.3368674015 1.070504e-01 2.231046e-03
economique   0.5091670820 6.339371e-02 0.3205428965 7.807373e-04 5.333271e-05
emploi       0.1756002107 5.044594e-01 0.0513227017 1.972437e-01 1.960047e-02
entreprendre 0.0001013167 4.062339e-01 0.0917607836 4.190009e-01 1.427003e-03
etat         0.0251489788 1.159678e-02 0.5185954891 4.842476e-04 3.912397e-02
financier    0.1519797437 3.265881e-02 0.0822185379 2.316414e-01 4.099256e-01
francais     0.0905488383 6.985809e-02 0.4586478567 6.916649e-02 2.037429e-01
loi          0.4156826388 1.577015e-01 0.1483456569 3.807086e-02 1.113524e-01
national     0.0894410177 1.579591e-02 0.5277081034 9.163699e-02 2.560431e-02
pays         0.0401586836 1.066404e-01 0.0885252771 4.555359e-01 2.404731e-02
permettre    0.1440503321 2.824781e-02 0.0785401538 1.965496e-01 5.425364e-02
politique    0.5235165159 1.609145e-03 0.1109911561 4.463092e-02 3.224471e-03
pouvoir_n    0.0774192236 3.784064e-01 0.0493094273 5.422330e-02 2.051849e-01
pouvoir_v    0.0891031033 3.121810e-01 0.1173245198 1.945161e-01 2.103204e-01
public       0.3982133301 4.475491e-01 0.0002538342 1.029270e-01 1.569377e-02
service      0.3535861235 3.533924e-01 0.0895054499 4.710048e-02 3.913258e-02
social       0.0642005463 6.521692e-01 0.1495578945 5.887909e-05 5.614115e-02
travail      0.2951188189 6.162731e-02 0.4302238407 2.853778e-02 7.567073e-02

$contrib
                   Dim.1        Dim.2       Dim.3        Dim.4        Dim.5
crise         8.83683190  5.699171745 16.73109724  6.410843768  5.396497263
devoir_v     37.19551568 16.249712835  3.30134526  2.272901389  6.518920581
droit         1.50096795  0.001566157  9.18036675  4.377920864  0.142663070
economique    4.01375220  0.625400499  4.65849957  0.017027210  0.001818677
emploi        2.14348916  7.706270228  1.15498258  6.661125906  1.034986018
entreprendre  0.00186949  9.380805852  3.12154015 21.389738259  0.113903917
etat          0.32864996  0.189658613 12.49429115  0.017507696  2.211708765
financier     2.29255178  0.616532018  2.28650871  9.667137090 26.749165798
francais      0.85314563  0.823718477  7.96689914  1.802952701  8.304138844
loi           4.13325045  1.962399234  2.71940778  1.047299508  4.789622226
national      0.40135600  0.088707339  4.36572576  1.137659099  0.497024510
pays          0.44432127  1.476593918  1.80573742 13.944037497  1.150948683
permettre     0.49958452  0.122603046  0.50217619  1.885885354  0.813945492
politique     9.20877786  0.035423273  3.59939616  2.171979770  0.245358878
pouvoir_n     0.70197912  4.293936429  0.82428046  1.360221462  8.048084050
pouvoir_v     1.92737104  8.450855896  4.67876195 11.640624859 19.680027450
public       16.61745929 23.372830812  0.01952849 11.883004381  2.833008174
```

```
service       2.32805762  2.911906641  1.08646979  0.857970198  1.114573502
social        1.14641800 14.574254125  4.92360586  0.002908799  4.336679533
travail       5.42465110  1.417652863 14.57937961  1.451254188  6.016924569

$dist
       crise     devoir_v        droit   economique       emploi entreprendre         etat
    3.996803     5.756374     2.970924     2.169556     2.699753     3.319303     2.793396
   financier     francais          loi     national         pays    permettre    politique
    3.001184     2.371898     2.436637     1.636901     2.570305     1.439041     3.240870
   pouvoir_n    pouvoir_v       public      service       social      travail
    2.326824     3.593868     4.991720     1.982783     3.265337     3.312938

>
```

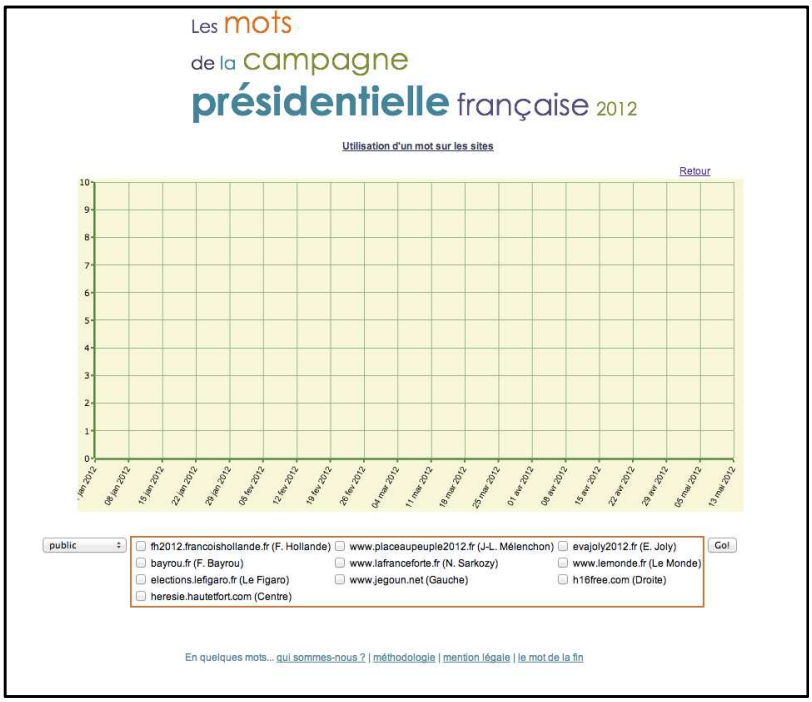
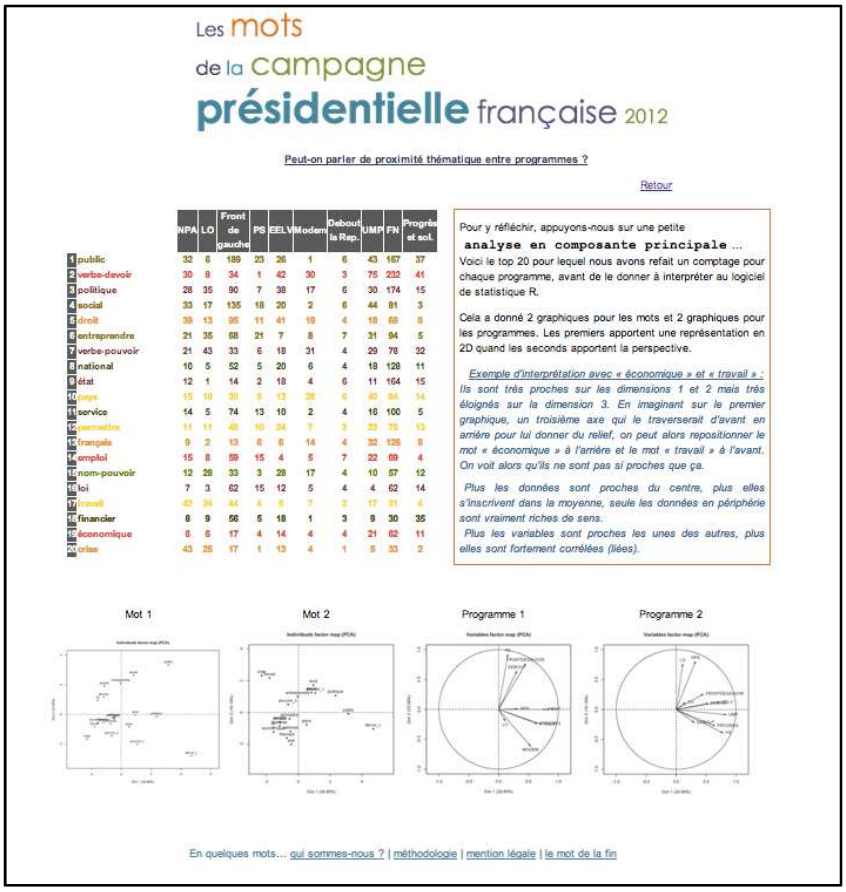


Figure Annexe 1 : Page sur l'utilisation d'un mot sur les sites



	NPA	LO	Front de gauche	PS	EELV	Modem	Debout la Rep.	UMP	FN	Progrès et sol.
public	32	6	189	23	26	1	6	43	167	37
verbe-devoir	30	8	34	1	42	30	3	75	232	41
politique	28	35	90	7	38	17	6	30	174	15
social	33	17	135	18	20	2	6	44	81	3
droit	39	13	95	11	41	10	4	18	68	8
entreprendre	21	35	68	21	7	8	7	31	94	5
verbe-pouvoir	21	43	33	6	18	31	4	29	78	32
national	10	5	52	5	20	6	4	18	128	11
état	12	1	14	2	18	4	6	11	164	15
pays	15	10	30	5	13	28	5	40	84	14
service	14	5	74	13	10	2	4	16	100	5
permettre	11	11	48	10	24	7	3	33	78	13
français	9	2	13	6	6	14	4	32	126	8
emploi	15	8	59	15	4	5	7	22	69	4
nom-pouvoir	12	28	33	3	28	17	4	10	57	12
loi	7	3	62	15	12	5	4	4	62	14
travail	42	34	44	4	8	7	3	17	31	4
financier	8	9	56	5	18	1	3	9	30	35
économique	6	6	17	4	14	4	4	21	62	11
crise	43	25	17	1	13	4	1	6	33	2

Figure Annexe 2 : Page sur la proximité thématique

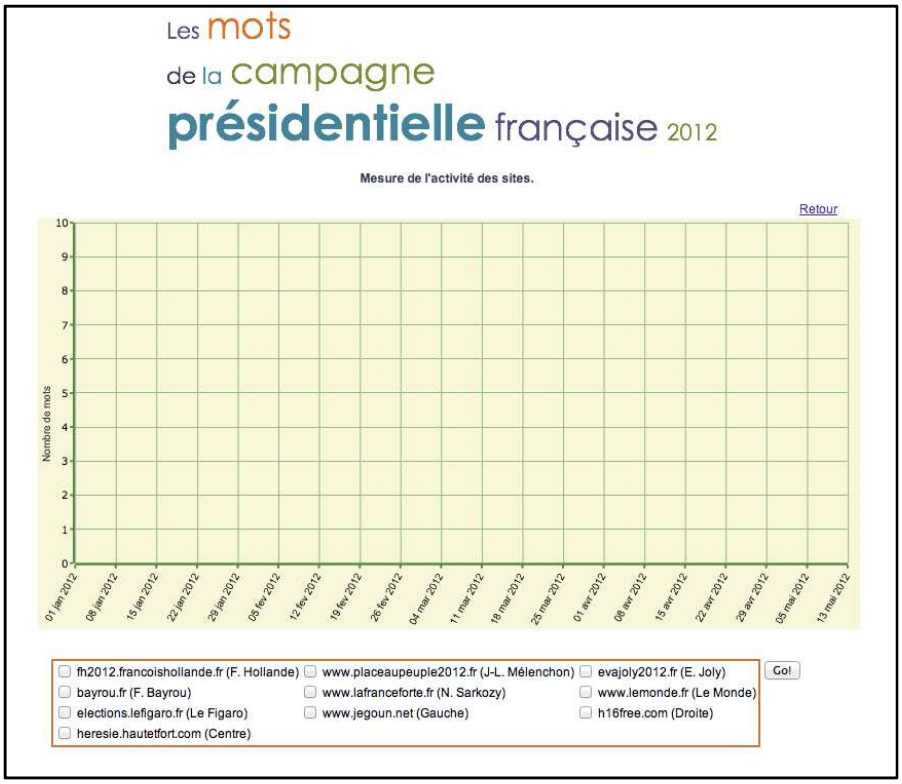


Figure Annexe 3 : page sur la mesure de l'activité des sites

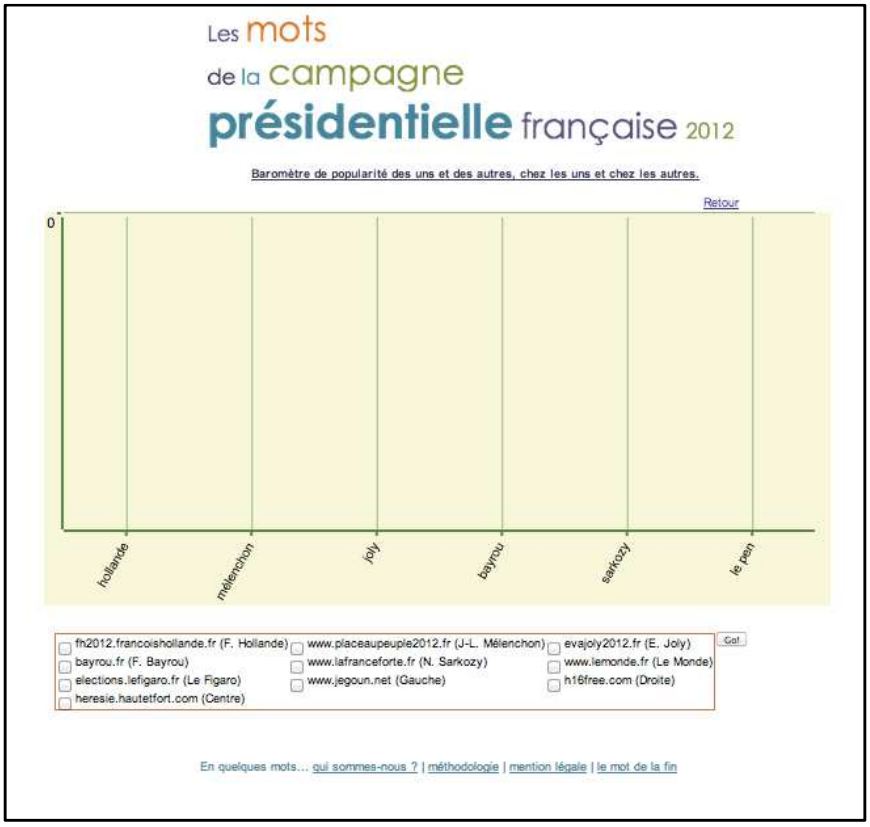


Figure Annexe 4 : Page sur le baromètre de popularité

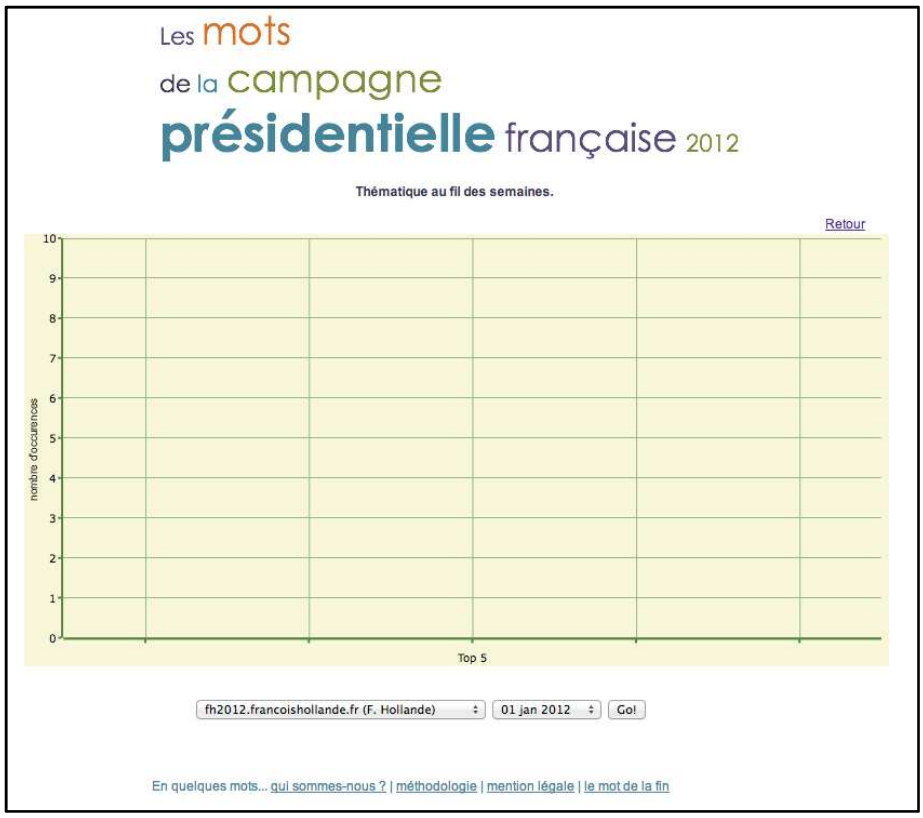


Figure Annexe 5 : Page sur la thématique au fil des semaines

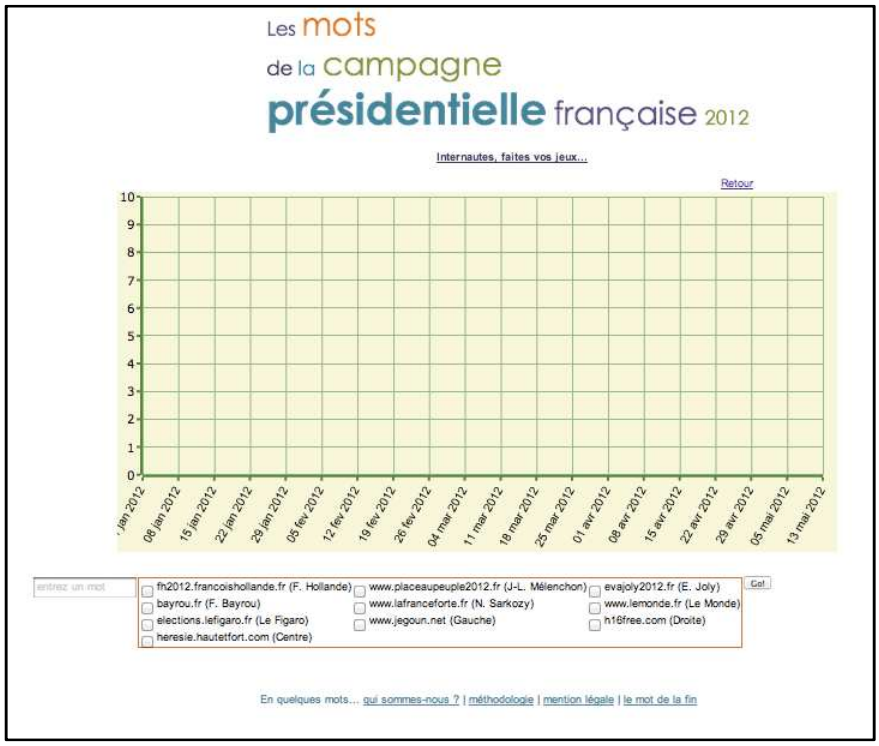
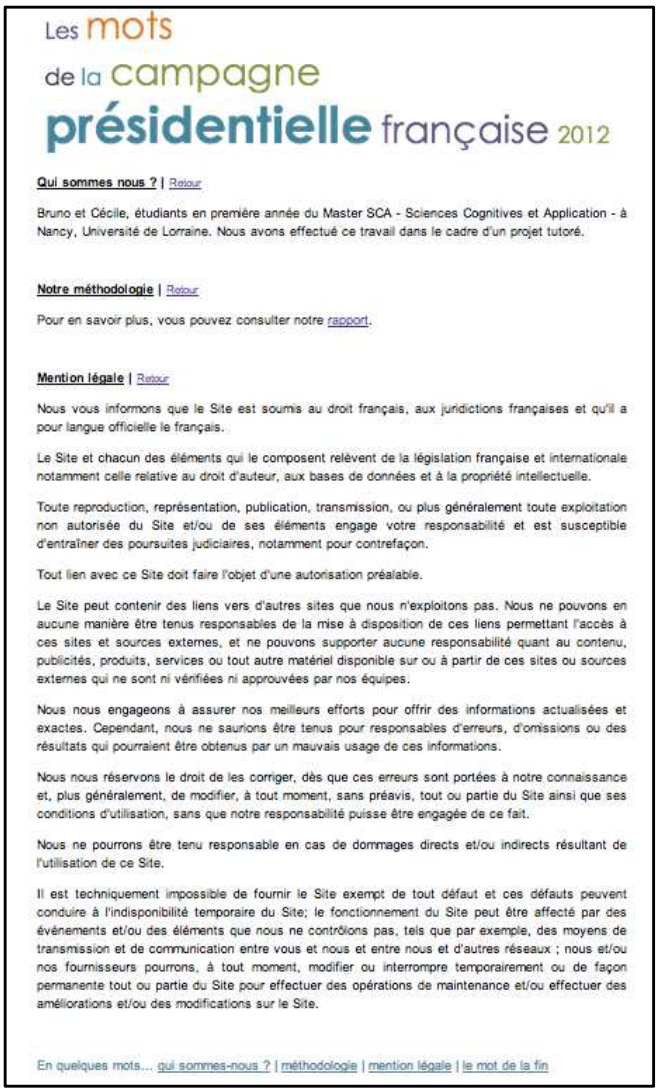


Figure Annexe 6 : Page sur le libre choix



Les mots
de la campagne
présidentielle française 2012

Qui sommes nous ? | Retour

Bruno et Cécile, étudiants en première année du Master SCA - Sciences Cognitives et Application - à Nancy, Université de Lorraine. Nous avons effectué ce travail dans le cadre d'un projet tutoré.

Notre méthodologie | Retour

Pour en savoir plus, vous pouvez consulter notre rapport.

Mention légale | Retour

Nous vous informons que le Site est soumis au droit français, aux juridictions françaises et qu'il a pour langue officielle le français.

Le Site et chacun des éléments qui le composent relèvent de la législation française et internationale notamment celle relative au droit d'auteur, aux bases de données et à la propriété intellectuelle.

Toute reproduction, représentation, publication, transmission, ou plus généralement toute exploitation non autorisée du Site et/ou de ses éléments engage votre responsabilité et est susceptible d'entraîner des poursuites judiciaires, notamment pour contrefaçon.

Tout lien avec ce Site doit faire l'objet d'une autorisation préalable.

Le Site peut contenir des liens vers d'autres sites que nous n'exploitons pas. Nous ne pouvons en aucune manière être tenus responsables de la mise à disposition de ces liens permettant l'accès à ces sites et sources externes, et ne pouvons supporter aucune responsabilité quant au contenu, publicités, produits, services ou tout autre matériel disponible sur ou à partir de ces sites ou sources externes qui ne sont ni vérifiées ni approuvées par nos équipes.

Nous nous engageons à assurer nos meilleurs efforts pour offrir des informations actualisées et exactes. Cependant, nous ne saurions être tenus pour responsables d'erreurs, d'omissions ou des résultats qui pourraient être obtenus par un mauvais usage de ces informations.

Nous nous réservons le droit de les corriger, dès que ces erreurs sont portées à notre connaissance et, plus généralement, de modifier, à tout moment, sans préavis, tout ou partie du Site ainsi que ses conditions d'utilisation, sans que notre responsabilité puisse être engagée de ce fait.

Nous ne pourrons être tenu responsable en cas de dommages directs et/ou indirects résultant de l'utilisation de ce Site.

Il est techniquement impossible de fournir le Site exempt de tout défaut et ces défauts peuvent conduire à l'indisponibilité temporaire du Site; le fonctionnement du Site peut être affecté par des événements et/ou des éléments que nous ne contrôlons pas, tels que par exemple, des moyens de transmission et de communication entre vous et nous et entre nous et d'autres réseaux ; nous et/ou nos fournisseurs pourrons, à tout moment, modifier ou interrompre temporairement ou de façon permanente tout ou partie du Site pour effectuer des opérations de maintenance et/ou effectuer des améliorations et/ou des modifications sur le Site.

En quelques mots... qui sommes-nous ? | méthodologie | mention légale | le mot de la fin

Figure Annexe 7 : Page sur les caractéristiques

Annexe 3

```
def module_creation_stat(fichier_entree, fichier_sortie, lg_mot, nb_mot,
typ_trt = None, liste_mot = None):
    # type trt
    # 1 = catégorie np
    # 2 = liste de mot
    # autre = tableau de tous les mots
    resultat = filtrage_mot(fichier_entree, typ_trt, liste_mot)
    tableau = resultat[0]
    longueur_tot = resultat[1]
    comptage_mot(tableau, longueur_tot, fichier_sortie, lg_mot, nb_mot)

def filtrage_mot(fichier_entree, typ_trt, liste_mot):
    liste = re.split(r'[#]|\n', fichier_entree)
    tab_mot = []
    tab_selection = []
    for mot in liste:
        debut = mot.find('[')
        fin = mot.find(':')
        terminaison1 = mot.find('|')
        terminaison2 = mot.find(']')
        # si les crochets ne sont pas vides
        if debut != -1:
            if fin != -1:
                contenu = mot[debut + 1:fin]
                if terminaison1 != -1:
                    genre_mot = mot[fin + 1:terminaison1]
                else:
                    genre_mot = mot[fin + 1:terminaison2]
                if contenu not in motsGramm:
                    if genre_mot in liste_categorie:
                        if contenu in liste_disso:
                            contenu = contenu + "_" + genre_mot
                        tab_mot.append(contenu)
                        if typ_trt == 1:
                            if genre_mot == 'np':
                                tab_selection.append(mot[:debut])
                        elif typ_trt == 2:
                            if contenu in liste_mot:
                                tab_selection.append(contenu)
    if typ_trt in range(1, 3):
        return (tab_selection, len(tab_mot))
    else:
        return (tab_mot, len(tab_mot))

def comptage_mot(tableau, longueur_tot, fichier_sortie, lg_mot, nb_mot):
    # Ecriture des résultats
    f = open(fichier_sortie, 'w')
    freq = freq_mot(tableau, lg_mot, nb_mot)
    f.write(str(longueur_tot) + '\n')
    for mot in freq:
        res = mot[0] + ':' + str(mot[1])
        f.write(res + '\n')
    f.close()
```

```
def module_lecture_stat(fichier_stat):
    dict_mot = {}
    f = open(fichier_stat, "r")
    contenu = f.read()
    f.close()
    lignes = contenu.split('\n')
    nb_total = lignes[0]
    for ligne in lignes[1:]:
        mot = re.split(r':', ligne)
        if len(mot) == 2:
            dict_mot[mot[0]] = int(mot[1])
    return (nb_total, dict_mot)
```