Analyse lexicale d'une page web

David Reymond, Kouamvi Couao-Zotti, Alaric Tabariès, Amandine Lebourgeois, Lauren Campos

Revue française des sciences de l'information et de la communication

24 | 2022
Data Paper : émergence d'une nouvelle donne scientifique

Analyse lexicale d'une page web

Extractions du message hypertextuel pour comparaisons

Lexical analysis of a web page : verification of hypertextual communicative coherence

David Reymond, Kouamvi Couao-Zotti, Alaric Tabariès, Amandine Lebourgeois et Lauren Campos

OpenEdition Journals

Édition électronique
URL : https://journals.openedition.org/rfsic/12365
DOI : 10.4000/rfsic.12365
ISSN : 2263-0856

Éditeur
Société Française de Sciences de l'Information et de la Communication

Référence électronique
David Reymond, Kouamvi Couao-Zotti, Alaric Tabariès, Amandine Lebourgeois et Lauren Campos, « Analyse lexicale d'une page web », *Revue française des sciences de l'information et de la communication* [En ligne], 24 | 2022, mis en ligne le 01 janvier 2022, consulté le 30 août 2023. URL : http://journals.openedition.org/rfsic/12365 ; DOI : https://doi.org/10.4000/rfsic.12365

Analyse lexicale d'une page web

Extractions du message hypertextuel pour comparaisons

Lexical analysis of a web page : verification of hypertextual communicative coherence

David Reymond, Kouamvi Couao-Zotti, Alaric Tabariès, Amandine Lebourgeois et Lauren Campos

Région Sud, programme ASERVENA

Pôle Information, Numérique, Prévention, Santé de l'Université de Toulon

Société Web Notoriété (Hyères)

Nos plus sincères remerciements aux relecteurs de la revue qui ont grandement contribué à la clarification de l'article.

Contexte

1 Pour une organisation, la communication sur le web passe par l'élaboration d'un site web. Nous avons proposé (Pinède *et al.*, 2012 ; Pinède & Reymond, 2013) l'expression "Site web organisationnel" (SWO) pour caractériser ces sites web associés à une instance énonciatrice de type organisationnel. Le SWO se définit comme un produit hypertextuel d'informations et de services placé sous une seule et même responsabilité. Associé à une image (Rouquette, 2009), le SWO est censé véhiculer un ou des messages en phase avec les objectifs communicationnels de l'organisation. Pinède (2018) propose un modèle global de cette projection numérique de l'image et de l'identité d'une organisation, en la considérant au carrefour de logiques complexes. Le site web organisationnel en tant qu'unité cohérente de texture (Brügger, 2012) permet de discuter de l'identité numérique d'une organisation et couvre, en communication, un objectif d'image : un message en phase avec la stratégie est inscrit dans une cohérence formelle (design, architecture, charte, etc.) et une cohérence sémantique de contenus textuels (Zacklad, 2019). Sous l'angle de l'indexation, la visibilité du site web se contrôle via les techniques de référencement web. Métadonnées et réseaux de liens internes ou externes (*netlinking*) participent à performer sa place sur les index (Domenget, 2014 ;

Chartier & Martin, 2021) qui enregistrent une représentation sémantique composée et dynamique du site web. Celle-ci est une fonction directe du vocabulaire véhiculé par l'ensemble des pages qui constituent le réseau autour du site web. Si le contenu textuel d'un site web est contrôlable en général par son éditeur, il est cependant difficile d'apprécier la cohérence avec des contenus textuels des pages qui le citent. Dans ce qui suit, nous proposons un outil qui donne à voir au plan lexical les contenus d'une page, une extraction ciblée pour synthétiser les contenus textuels par les termes les plus représentatifs de la page qui participent de fait à cette indexation. Au-delà de cet outil décomposé en trois scripts réutilisables, nous présentons le jeu de données produites sur notre terrain d'étude pour montrer la pertinence de cette extraction et le lexique qui en résulte. La sauvegarde de ces données sera utile à des fins d'historicisation pour pouvoir à terme analyser les évolutions.

Visées du travail de recherche

2 Apprécier la cohérence de communication sur le plan des messages véhiculés entre deux pages web reliées par un lien hypertexte est un problème difficile qui se pose dans le domaine de la communication web, de l'indexation et du référencement en général. Dans le cadre des SWO, nos travaux visent à instrumenter la construction d'une vue partielle du message véhiculé par une page. Cette vue est focalisée sur le lexique et simule de cette façon le fonctionnement des index. Ainsi, en proposant de reconstituer le fonctionnement des systèmes de signification extraits des sites web, nous adoptons le principe limitatif volontaire d'extraire les occurrences terminologiques ciblées sur les verbes, noms et adjectifs du message qu'une page délivre. La capture de la terminologie dominante (par ses occurrences), filtrée sur les termes triviaux (ou idiolecte du média web pour des sites SWO – « contact, accueil, lire la suite » par ex.) permet de transcrire le sens global véhiculé et d'opérer des comparaisons et des analyses. Ce procédé adapte la méthode proposée par Barthes (1964) pour analyser les **grandes unités signifiantes** du discours par l'utilisation de la linguistique (et du TAL). Nous posons les prémices d'un cadre d'analyse sémiotique plus large appliqué au modèle discret du discours à reconstituer comme somme de deux pages web reliées par un lien hypertexte.

Terrain d'étude

3 Le Parc National de Port-Cros (PNPC[2]) construit sur son site web une image en phase avec ses objectifs historiques : protéger la nature et permettre à chacun de la visiter. La notoriété du parc à l'international fait de lui l'une des destinations touristiques les plus fréquentées du littoral varois avec plus d'un million de visiteurs chaque année. Dans ce contexte, le parc cherche depuis plusieurs années à concilier les activités humaines (économique, tourisme, trafic maritime, etc.) et sa mission de préservation de la nature quelquefois antagonique (Deldreve & Michel, 2019).

Mode opératoire

4 Pour analyser le lien entre le territoire physique du parc et le territoire virtuel qui sera à terme reconstitué par le contenu lexical de sa communication web et l'inscription

topologique de son site (hors réseaux sociaux), nous nous focalisons sur les acteurs du tourisme autour du PNPC. Notre point de départ a été de cibler les organisations qui avaient un site web et identifiées dans le domaine via le site web officiel http://visitvar.fr. Le nombre d'acteurs à considérer étant très important nous avons constitué un ensemble de catégories pour les classer.

5 La typologie de sites est la suivante (réalisée *a priori*) :

1. Offices de tourisme (OT)
2. Mairies
3. Parcs nationaux (parcs)
4. Hébergement
5. Organisations
6. Sites web gouvernementaux
7. Gouvernement (Gouv)
8. Restauration
9. Locations
10. Services

6 Le classement des 379 sites web issus du site web officiel dans la typologie ci-dessus est visible dans le fichier Ressources/Sites.json. Ce fichier est utilisé comme point de départ pour l'outil de collecte pour construire une synthèse par catégorie comme somme des contenus textuels extraits de chaque site web.

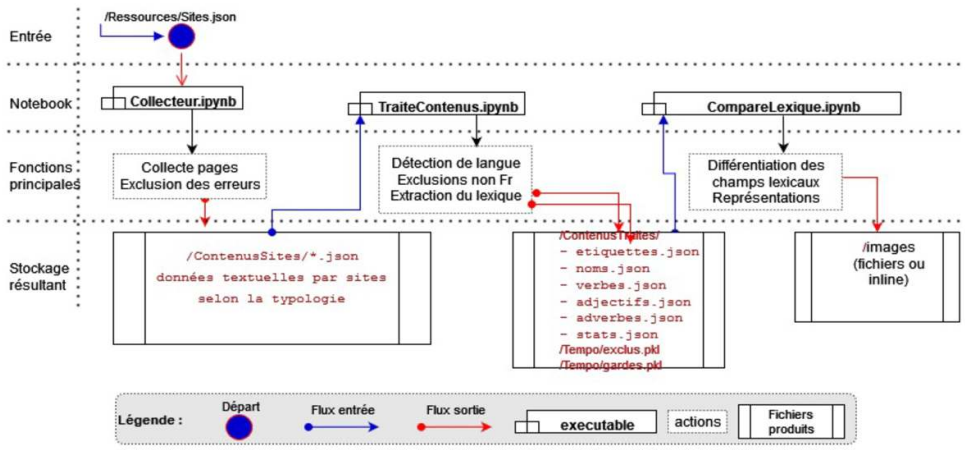
Instrumentation et données produites

7 Données et exécutables de traitement sont disponibles sur Git (REYMOND *et al.*, 2022) pour à la fois reproduire les traitements à partir de nos données, reproduire la collecte sur nos données d'entrée pour disposer de données récentes, soit encore adapter les données d'entrée pour traiter soi-même d'autres sites..

Organisation générale des notebooks et des données

8 Trois notebooks Jupyter[3] permettent respectivement de collecter les contenus textuels d'une page web, puis de traiter ces contenus textuels pour extraire les lemmes[4] dominant par leur occurrence, mais aussi les séparer selon les catégories grammaticales. Le troisième notebook permet de visualiser le résultat par des représentations en nuage de mots. Les notebooks sont paramétrés pour pouvoir être adaptés à d'autres sites web. Les données d'entrée et de sortie des traitements sont au format JSON. La figure 1 présente l'organisation générale des dossiers, des flux, des données d'entrée, et des données produites par chaque instrument.

Figure 1. Organisations des notebooks, des dossiers et des données produites



Le collecteur

9 **Collecteur**.ipynb : Le collecteur initial, importe la liste des sites web en entrée et leur typologie (*Ressources/Sites.json*) pour collecter les données et stocker le résultat dans / ContenusSites/. Par ex. *hebergement.json* contient pour chaque URL de la classe hébergement le contenu textuel collecté à cette adresse. Le fichier *BadUrl.json* est la liste des URL en erreur au moment de la collecte donnée par classe de la typologie. Voici le contenu de ce fichier au moment de nos traitements (format json) qui permet d'apprécier s'il en est, pour chaque classe de site, les URL ayant retourné des contenus non utilisables :

```
BadUrls = {"gouv" : [], "parcs" : [], "orga" : ["http://onf.fr"],
"oTourism" : [], "Services" : ["http://spinout.fr"],
"Hebergements" : ["http://parc.fr", "http://
campingsaintpierredeshorts.com"], "restauration" : ["http://
lemediterranee-hyeres.fr"], "locations" : ["http://
levasionbleue.com"], "mairies" : []}
```

Traitement des contenus

10 TraiteContenus.ipynb : réalise les extractions terminologiques.

- Prétraitements :
- Détection de la langue ;
- Suppression des extrêmes (par la taille des données collectées) ;
- Extraction du lexique lemmatisé. Utilisation de la bibliothèque Spacy (Explosion, 2017)
- Sauvegarde des dictionnaires (verbes, adjectifs, etc.) au format JSON. Dans le dossier / *ContenusTraites/* six fichiers sont produits :
- l es fichiers fonction de la catégorie morphologique des termes identifiés ***[verbes, adjectifs, adverbes, noms]Sites.json*** sont structurés par élément de la typologie. Chaque classe de sites est associée à une liste de termes aux occurrences multiples. Ces données sont utilisées pour les représentations (cf. Figure 2). Par ex. verbesSites ["gouv"] = ["aller", "profiter", "agir", "menacer", ...]
- ***Etiquettes.json*** est la liste, par élément de la typologie puis par url de toutes les étiquettes grammaticales identifiées et les formes lexicales associées. C'est le « détail » précis des

fichiers précédents, redondant, mais incluant les autres étiquettes non retenues (entités nommées[5] pour l'essentiel).

- ***CatGrammSites.json*** : comme précédemment, mais sous forme de liste de lemmes associés à leurs occurrences dans la classe. Par ex. "Gouv" : {"MISC" : {"Office" : 1, "Liste rouge des espèces menacées" : 1, ...}}

Instrument de comparaisons de lexiques

11 CompareLexique.ipynb : procédures d'extraction des lexiques pour réaliser des comparaisons et représenter par des visualisations.

1. Séparation des termes (lemmes communs à tous les sites web, lemmes communs à au moins deux sites web, lemmes exclusifs à chaque catégorie de site web) par catégorie grammaticale ;
2. Cohérence des erreurs de catégorisation (nettoyage des unités lexicales morphologiquement ambiguës, c.-à-d. mots pouvant être catégorisés en deux catégories grammaticales différentes) ;
3. Visualisations : nuages de mots clés, palette de couleurs paramétrable.

12 **Les visualisations sont stockées (pour certaines) dans le dossier IMAGES** : contient les nuages de mots réalisés selon la typologie de sites web et les catégories grammaticales. Cet instrument est fait pour explorer les données. L'analyse nécessite une étude approfondie qui sort du cadre de cet article et peut servir des applications pédagogiques.

Données adaptables

13 Typologie de sites associés (Format JSON) ; Le fichier RESSOURCES/Sites.json donne la possibilité d'adapter à ses besoins la configuration de la typologie des sites web étudiés. Il suffit d'associer autant de classes de sites une liste d'URL selon le format suivant : CatSites [ClasseSite] = [url1, url2, ...]. Il est impératif que les classes soient disjointes : les URL ne peuvent appartenir qu'à une classe et se méfier des sites qui utilisent plusieurs DNS (différentes URL pour désigner le même contenu).

Forces et limites

14 La disparité des formes de communication sur le web nécessite une adaptation contextuelle des collecteurs et traitements.

1. Nous focalisons notre approche sur le contenu HTML : un premier analyseur lexical extrait le texte décrit par les balises du langage. Cet analyseur est générique, donc grossier : l'objectif est de fonctionner au mieux sur un maximum de pages. La réalité du web est qu'un analyseur parfait dépend du site web à analyser. Ici, nous nous satisfaisons d'un générique qui commet des erreurs sur certains contenus HTML, en général ceux qui sont mal balisés .
2. En plus de nous abstraire des balises, nous faisons le choix de représenter le contenu lexical d'une page par un document qui se construit comme la somme des éléments textuels qui auront pu être récupérés lors de la phase précédente. Ce texte est alors analysé par un modèle statistique pré entrainé de spaCy en premier lieu pour identifier les entités nommées (Lample *et al.*, 2016). Un second modèle, plus efficace sur nos données, est alors

utilisé pour extraire les lemmes des fonctions grammaticales identifiées comme pertinentes (verbes, adjectifs, noms et adverbes).

3. Une procédure de normalisation paramétrable exclut les sites web dont la taille de lexique et dont les occurrences s' étendent au-delà d'un espace raisonnable construit autour de la moyenne de la taille des sites web (en nombre de caractères). Ceci est réalisé pour obtenir des données « comparables ».
4. Une exclusion des sites web non rédigés en français a été mise en place. Ils sont identifiés par un utilitaire de détection de langue. Bien que la liste initiale soit constituée de sites web en français, certaines erreurs au moment de la collecte peuvent renvoyer du texte en anglais.
5. Les outils de TAL sont construits pour fonctionner génériquement sur des textes « bien construits » . Dans notre cadre les passages textuels sont extraits d'une forme éditoriale généralement composée (des menus, des cadres, des extraits, et même des métadonnées) qui, alors qu'ils sont isolés visuellement sur la page, sont dans notre cadre concaténés par le traitement d'extraction du contenu textuel du site web. Les modèles linguistiques sont entraînés plutôt sur des textes de type « discours » et, en conséquence, cette procédure « d'aplatissement » de la page reconstitue des textes qui ne sont pas forcément des phrases à l'origine. Alors que l'outil Spacy réalise les tâches d'extraction d'entités nommées, de lemmatisation et d'identification de catégories grammaticales avec des scores de précision très élevés (> 98 %) cette reconstruction artificielle de textes diminue mécaniquement ce score sur nos données.
6. La suppression des formes lexicales communes aux différents sites web n'est pas utile dans les représentations vectorielles , c'est en revanche un moyen d'identifier les signaux faibles plus précisément, d'effectuer des représentations de comparaison très opérationnelles.

Méthodes d'exploitation pour les interpréter et les reproduire

15 La procédure de visualisation permet la représentation du lexique récupéré. En jonglant sur les catégories grammaticales et les différents types de sites web, les représentations s'adaptent pour explorer les lexiques collectés. Des variables d'ajustement sont positionnées pour filtrer les représentations ou servir de seuil selon le nombre d'occurrences.

16 Les résultats seront identiques si le collecteur n'est pas lancé : la dynamique du web faisant que, d'une part, les contenus textuels peuvent changer tant dans le texte que la forme éditoriale et, d'autre part, les sites web non accessibles au moment de nos collectes peuvent l'être à un autre moment et inversement. Les visualisations produites peuvent aussi varier dans le positionnement des termes, mais ces derniers resteront les mêmes.

17 Ces représentations constituent une synthèse des lexiques véhiculés par les classes de sites web. Ces représentations permettent, dans le cadre de sites web d'organisations, d'extraire le lexique qui sera indexé à partir de ces pages et constituent, en ce sens, une représentation sémiotique de la page.

Notes et cadres d'usages

18 La collecte par le collecteur prend une dizaine de minutes. Le traitement des données par spaCy est un peu plus long (il est préférable d'activer le GPU dans les paramètres du notebook pour ne pas dépasser les vingt minutes de traitement). Les résultats montrent une sensibilité évidente liée au nombre de sites web par catégorie : plus ce nombre est élevé, plus la gamme lexicale sera importante. En même temps, un lexique issu de catégories de sites web bien composées (des sites web choisis comme relevant de la catégorie) est particulièrement pertinent : le traitement laisse apparaître les termes les plus fréquents de cette classe de site permettant de déterminer le registre lexical de la classe de site et d'opérer à des comparaisons pour identifier des points communs ou des singularités lexicales . L'instrument sera efficace aussi sur les pages suffisamment fournies en contenu textuel et respectueuses du langage HTML.

19 Les trois diagrammes de la figure 2 comparent les lemmes (noms, verbes et adjectifs) de la classe restauration à l'ensemble des noms dans cette classe (à droite) en passant par les noms exclusifs de cette classe (au milieu). Les couleurs autres que bleues (diagramme du milieu) sont celles provenant des autres classes de la typologie de sites web ou communes à toutes (en gris) ou à certaines seulement (autres couleurs). Ainsi, les termes communs participent au lexique web de chaque type d'organisations et l'extraction des termes spécifiques permet de cibler « la spécialité » dominante du type de site web et probablement de l'éditeur sous-jacent (taverne, cantine et pizzéria sont majoritaires dans nos données, c'est-à-dire correctement représentées sur le web). Nous avons aussi constaté que, par construction, la classe restauration n'est pas disjointe de celle de l'hébergement et, de fait, la lisibilité par le lexique différencié (propre à une classe) est une dépendante de la séparation correcte de cette classe avec les autres.

Figure 2. Lexiques extraits de la classe « restauration » de notre typologie.

De gauche à droite : 2.a : Ensemble des lemmes (noms, adjectifs, verbes) de cette classe. 2.b : ensemble des noms exclusifs à cette classe (les mots n'apparaissent dans aucune des autres classes de site). 2.c : Ensemble des noms de la classe restauration. En gris les mots communs (identifiés dans d'autres classes de site), les autres couleurs colorent les termes exclusifs des autres classes dans l'ensemble des schémas.

20 Cette instrumentation construite comme un oligoptique – *oligopticon* (Latour, 2006 ; Venturini & Latour, 2010), reproduit approximativement un robot web et améliore la finesse de l'extraction par la lemmatisation. Ces données de traitement sont reprises dans un article complémentaire pour étendre l'analyse sur les liens entrants particuliers partenaires de la communication d'un site web originel, le site web du Parc National de Port-Cros[6]. Cette étape nous permet d'envisager de développer les prémisses d'une herméneutique de la communication hypertexte par l'apport sémiotique de cette instrumentation. Les données serviront de référence dans le cadre

d'une analyse spécifique de l'environnement hypertexte du site web du PNPC afin de mesurer les résultats de recherche-action sortant du cadre de ce document.

21 Les données et les notebooks sont disponibles sous licence MIT sur le dépôt git : https://github.com/Patent2net/LexiComWeb Du point de vue pédagogique, au-delà de l'analyse de ces données de terrain, chacun des notebooks peut être « dérivé » pour produire des extractions sur d'autres sites web et/ou mettre en œuvre des comparaisons de lexiques. Ces applications sont variées dans le domaine du référencement et de l'indexation, en intelligence économique ou encore en analyse de communication. Par exemple, il est possible de modifier le collecteur pour récursivement collecter autant de pages que souhaité d'un site web afin de construire une représentation lexicale, synthèse de la communication véhiculée par ledit site web, à l'aide de l'extracteur de lemmes. Il faut cependant prendre garde à construire des analyses sur des SWO qui utilisent en général une communication positive. En modifiant l'analyseur syntaxique pour intégrer les formes négatives, il est envisageable de viser d'autres formes de communication comme les réseaux d'hyperliens référencés en tant que preuves de désinformation afin de tenter de faire émerger des incohérences par le lexique véhiculé.

BIBLIOGRAPHIE

Brügger, Niels. « L'historiographie de sites Web : quelques enjeux fondamentaux ». Le Temps des médias, vol. 18, no 1, 2012, p. 159. DOI.org (Crossref), https://doi.org/10.3917/tdm.018.0159.

Chartier, Mathieu, et Alexandra Martin. *Techniques de référencement web : Audit et suivi SEO*. Éditions Eyrolles, 2021.

Deldrève Valérie. et Michel Charlotte., « La démarche de capacité de charge sur Porquerolles (Provence, Parc national de Port-Cros, France) : de la prospective au plan d'actions », *Science Report Port-Cros national Park*, n° 33, 2019, pp. 63-100.

Domenget, Jean-Claude. « Référencer des contenus web ». *E-marketing et e-Commerce*, édité par Thomas Stenger et Stéphane Bourliataux-Lajoinie, Dunod, 2014, p. 233-65, https://hal.archives-ouvertes.fr/hal-01514325.

Explosion, AI. « spaCy-Industrial-strength natural language processing in python ». URL : https://spacy.io, 2017.

Lample, Guillaume, *et al.* « Neural architectures for named entity recognition ». arXiv preprint arXiv :1603.01360, 2016.

Latour, Bruno. *Changer la Société-Refaire de la sociologie*. Traduit par Nicolas Guilhot, La Découverte Poche, 2006.

Pinède, N., et D. Reymond. « Classer les sites Web organisationnels. Une approche taxonomique des liens hypertextes ». *Hermès*, vol. 1, no 66, 2013, p. 187-94.

Pinède, Nathalie, *et al*. « Approche diachronique de marqueurs lexicaux hypertextuels. Application aux sites web d'universités ». Le document à l'ère de la différenciation numérique, *Europia*, 2012, p. 135-52.

Pinède, Nathalie, *et al*. « « Du site web aux identités numériques organisationnelles. Proposition d'un modèle d'analyse ». *Questions de communication*, vol. 34, no 2, 2018, p. 75-94. Cairn.info.

Reymond, David, Amandine Lebourgeois, *et al*. LexiComWeb : donnees et executables : Analyse lexicale d'une page web : extractions du message hypertextuel pour comparaisons. 1.0.0, Zenodo, 2022. DOI.org (Datacite), https://doi.org/10.5281/ZENODO.5919879.

Rouquette, Sébastien. « Sociologie des sites web d'entreprise ». *Esprit Critique : Revue Internationale de Sociologie et de Sciences sociales*, vol. 12, n° 1, 2009, p. 15.

Venturini, Tommaso, et Bruno Latour. « The social fabric : digital traces and quali-quantitative methods ». Proceedings of Future En Seine, 2010, http://www.medialab.sciences-po.fr/publications/Venturini_Latour-The_Social_Fabric.pdf.

Zacklad, Manuel. « Le design de l'information : textualisation, documentarisation, auctorialisation ». Communication langages, no 1, 2019, p. 37-64. Brügger, Niels. « L'historiographie de sites Web : quelques enjeux fondamentaux ». *Le Temps des médias*, vol. 18, n° 1, 2012, p. 159. *DOI.org (Crossref)*, https://doi.org/10.3917/tdm.018.0159.

NOTES

1. La personnalisation est adaptée pour être utilisée via Google Colab (mais adaptable à tout autre dispositif d'hébergement de notebook). Elle permet de travailler avec ces « *executables papers* » directement sur ses propres données en les adaptant au format d'entrée (json).
2. Cf. http://www.portcros-parcnational.fr/fr
3. cf. https://jupyter.org/
4. Lemme : Forme graphique choisie conventionnellement comme adresse dans un lexique. (CNRTL). Ce procédé permet d'utiliser une forme normalisée pour les mots variables dans le genre, la pluralité ou la conjugaison par ex.
5. Expression linguistique qui désigne un nom de lieu, un nom de personne ou un nom d'organisation.
6. Par l'insertion de ce lien, à partir de quelques temps post-publication, les données initiales des liens entrants sur le site du PNPC seront modifiées. La publication sur le web du contenu de cet article entraîne en conséquent une modification du réseau topologique de citation du site web du PNPC. En conséquent il s'agit d'un data paper qui modifie les données... Les exécutables permettent de re-collecter et de retraiter...

RÉSUMÉS

Dans ce data paper, nous décrivons les données produites par une série d'outils adaptables[1] construits pour retrouver les mots employés dans une page d'un site web. Pour dépasser à terme les techniques de référencement actuelles, l'objectif est de disposer d'un instrument capable de

réduire les contenus textuels d'une page web, expurgée des balises HTML et codes informatiques, en un lexique afin de pouvoir saisir le sens global porté par la page. Pour apporter une finesse sémantique, le lexique est lemmatisé et séparé selon les catégories grammaticales (verbe : actions, nom : champ nominal, adjectifs et adverbes : intensité, temporalité, etc.). Les ensembles de données obtenues sont alors combinés pour être représentés en nuages de mots paramétrables afin d'accompagner une lecture distante. Les données collectées dans l'environnement web du Parc National de Port-Cros, sont agrégées selon une typologie de sites. Les traitements et représentations montrent l'intérêt et la pertinence de cette instrumentation pour comparer les lexiques véhiculés par des pages. La sauvegarde de ces extractions ainsi que toute la chaîne de production est d'intérêt autant pour des travaux en continuité que pour les reproduire dans un cadre pédagogique. Les forces et limites sont discutées pour cadrer l'extension de ce procédé à d'autres domaines et applications à la communication web en général.

In this data paper we describe the data produced by a set of adaptable tools built to reconstruct words found on a webpage. To go beyond current SEO techniques, the objective is to use an instrument capable of reducing the textual contents of a web page, stripped of HTML and computer tags, into a lexicon. The lexicon is lemmatized and separated according to parts of speech (verbs : actions, nouns : nominal field, adjectives, and adverbs : intensity, temporality, etc.) to provide semantic sophistication. All the acquired data are then combined and used by a configurable word cloud representation tool to allow for distant reading. Data collected in the territorial context of Port-Cros National Park, aggregated according to a classification of websites, demonstrate the interest and functionality of these representations in comparing their respective resulting lexicons. Saving these data representations as well as the entire process for obtaining them is of interest both for continuing this work and for reproduction purposes in an educational context. The strengths and limitations of the process are discussed to set up a framework for its expansion into other domains and web communication applications in general.

INDEX

Mots-clés : data paper, communication web, site web organisationnel, TAL, scraping, Jupyter notebook
Keywords : data paper, web communication, NLP, Jupyter notebook, scraping

AUTEURS

DAVID REYMOND

David Reymond est MCF-HDR en sciences de l'information et de la communication à l'université de Toulon.

KOUAMVI COUAO-ZOTTI

Kouamvi Couao-Zotti est doctorant en sciences de l'information et de la communication à l'université de Toulon.

ALARIC TABARIÈS

Alaric Tabariès est doctorant en sciences de l'information et de la communication à l'université de Toulon.

AMANDINE LEBOURGEOIS

Amandine Lebourgeois est étudiante en master Langues et Sociétés Parcours Traitement de l'information, Linguistique, Traduction (TILT) à l'université de Toulon.

LAUREN CAMPOS

Lauren Campos est étudiante en master Langues et Sociétés Parcours Traitement de l'information, Linguistique, Traduction (TILT) à l'université de Toulon.