UMONS
Université de Mons

Extraire du contenu textuel à partir de sites Internet. Un outil pour les analystes du discours : le Détricoteur

Papotages – 9 novembre 2023

Romuald Dalodiere
romuald.dalodiere@umons.ac.be

Faculté
de Traduction
et d'Interprétation
Ecole d'Interprètes
Internationaux

Utilisation de corpus en linguistique

- En linguistique de corpus
- En traitement automatique du langage (TAL)
- … mais aussi en analyse du discours (AD)
  - Une épistémologie différente dans les trois disciplines
  - Une herméneutique de l'AD que n'ont pas le TAL ou la linguistique de corpus (Sitri et Barats, 2017:11)
  - Épistémologie ≠ outils : travaux en AD avec les outils de la linguistique de corpus (p.ex. Alexander, 2009 ; Fernández-Vázquez et Sancho-Rodríguez, 2020)

Extraction de contenu textuel à partir du Web pour la constitution de corpus

- Deux approches selon De Schryver (2002) :
  - *Web as corpus*
    - « Le web est un corpus »
  - *Web for corpus*
    - « Le web est une ressource (pour constituer des corpus) »
- Le Détricoteur : *Web for corpus*
  - Développé par Jordan (2020)
  - Pour les besoins d'une thèse (Dalodiere, 2023)
  - Problématique : comment extraire facilement le contenu textuel de sites Internet rapidement et de façon uniforme ?

Une problématique ancienne

- Une multitude de programmes depuis au moins 20 ans (p.ex. Gupta, 2003)
- Sujet récurrent du TAL
  - Deux problèmes majeurs :
    - Gestion du bruit
    - Gestion des doublons
  - Une thèse traitant de ces deux questions : Pomikálek (2011)
    - Développe l'algorithme JusText

Les doublons

- Plusieurs causes :
  - Miroirs de sites Web
  - Styles d'affichage d'un même site
  - Citations (sur forums)
  - Reprises à l'identique d'actualités sur différentes plateformes
  - Etc.
- « Contenu dupliqué interne » (dans un même site)
  - Conséquences statistiques
  - Gêne pour l'analyste (concordancier…)
  - Doublons et quasi-doublons

Le bruit

- Éléments indésirables qui « obscurcissent » les résultats
  - Au-delà des seuls sites Internet : Paratexte (Genette, 1982)
    - préfaces, postfaces, avant-propos, notes, titres et sous-titres…

Le bruit

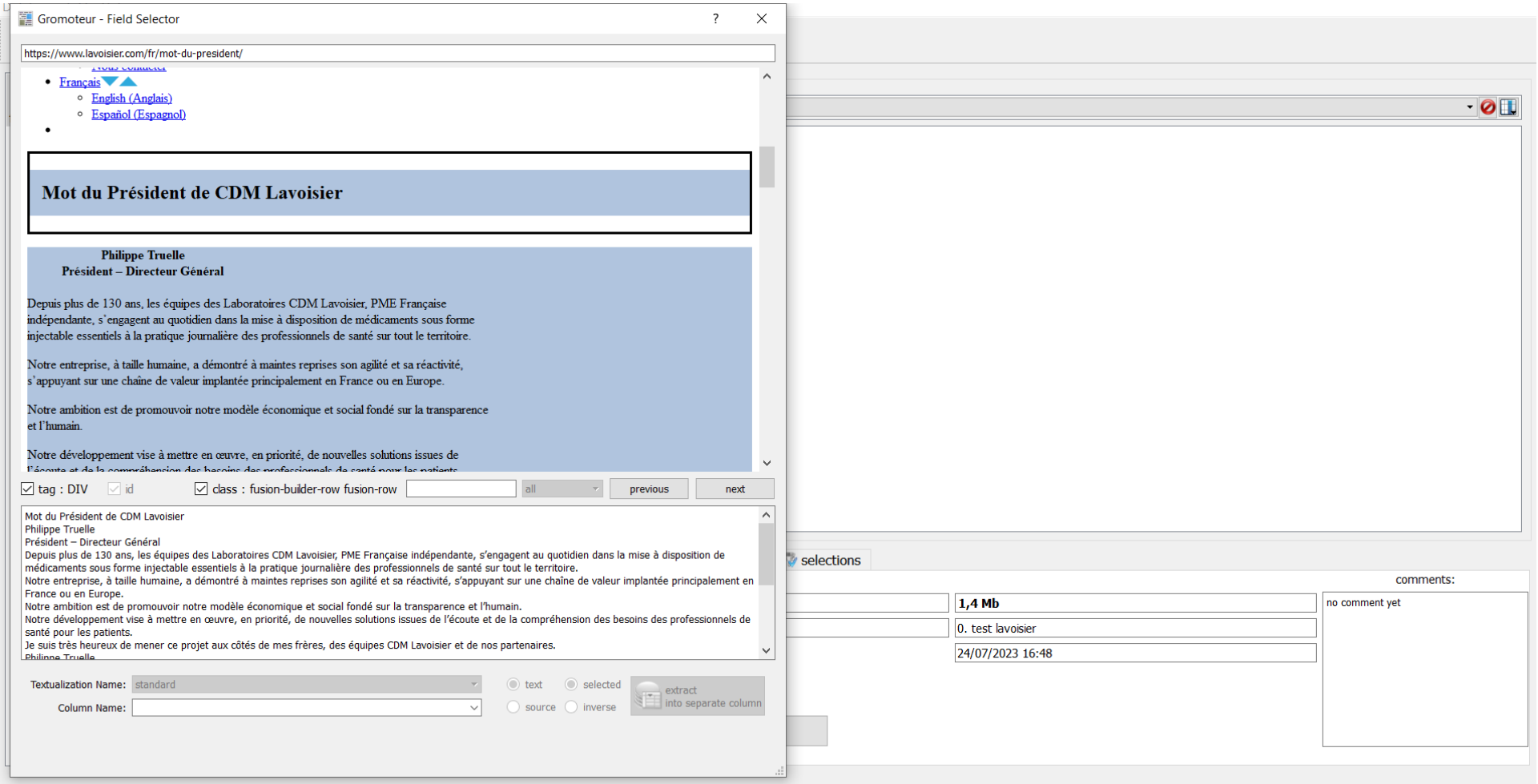
- Un véritable défi en TAL
  - Une multitude d'outils pour l'extraction automatique de contenu prétendant pouvoir discriminer entre *boilerplate* et contenu textuel pertinent
    - P.ex. JusText (Pomikálek, 2011)
  - Des résultats variables ; jamais entièrement satisfaisants
    - « *Le fait qu'il y ait autant d'outils disponibles est en réalité un indicateur de la disparité dans la qualité des résultats obtenus* » (Lejeune et Barbaresi, 2020:47).
    - Comparaison d'outils dans la littérature (p.ex. Barbaresi et Lejeune, 2020)
    - Le Détricoteur n'est pas exempt de défauts !

Un troisième problème (pour les profanes) ?

- Les doublons
- Le *boilerplate*
- … et l'accessibilité des programmes
  - Des programmes généralement sous forme de *packages*
    - Modules de code (souvent Python)
    - Quasi systématiquement disponibles sur des plateformes de partage (type GitHub)…
    - … mais inutilisables par des usagers non-développeurs
  - Très peu de logiciels « clés en main » (*standalones*)

Logiciels *standalones*

- Gromoteur (Gerdes, 2014)

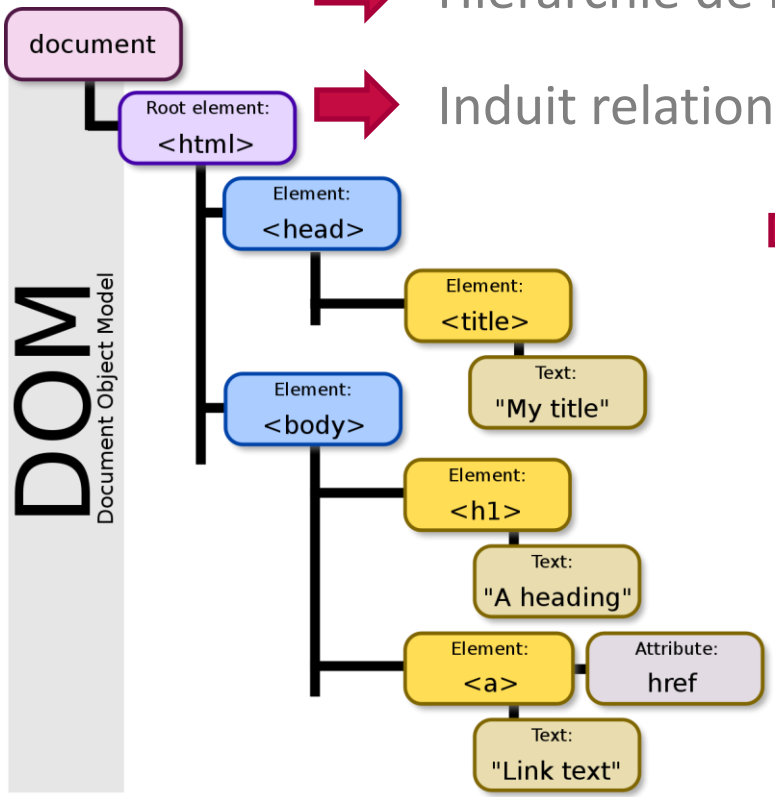


Logiciels *standalones*

- SketchEngine (Kilgarriff *et al.*, 2014)
  - Permet nettoyage de corpus
  - Performances critiquées (Fernández-Vázquez et Sancho-Rodríguez, 2020)
  - Repose sur JusText (Pomikálek, 2011)
    - Pas de contrôle de l'usager sur l'extraction

Le Détricoteur

- Exploitation du DOM (*Document Object Model*) pour l'extraction
  - Hiérarchie de balises HTML
  - Induit relations de subordination
  - Extraction à partir du DOM
    - Similaire à d'autres programmes (p.ex. Gromoteur)
    - La façon la plus efficace d'identifier les balises HTML avant extraction (Tripathy *et al.*, 2012:3)



Source : https://en.wikipedia.org/wiki/Document_Object_Model

Le Détricoteur

- Principe de fonctionnement :
  - Règles d'exclusion
    - On sélectionne ce qu'on veut supprimer
    - Tout ce qui n'est pas spécifiquement indésirable est (potentiellement) intéressant
    - Règles cumulables et ré-importables
  - Règles d'exclusion cumulables = possibilité d'affiner
    - Après exclusion du contenu strictement indésirable : ne reste que du contenu désirable ou au statut incertain
    - Possibilité de recommencer l'extraction en ajoutant de nouvelles règles

Le Détricoteur

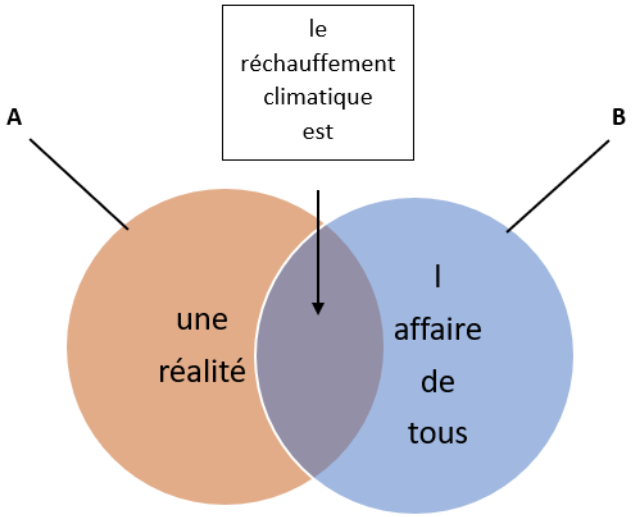
- Répond à deux problématiques sur trois :
  - Logiciel *standalone* (pas besoin de savoir coder)
  - Sélection manuelle du contenu indésirable
    - Import préalable d'une liste d'URLs
    - Les règles d'exclusion sont appliquées sur toutes les URLs de la liste
    - Attention à ne prendre que des URLs d'un même nom de domaine (pas de norme en matière d'architecture de site Web)!
  - Quid des doublons ?

Le Détricoteur

- Doublons :
  - Partir du « tableau lexical entier » (TLE)
    - P.ex. pour :

« Le réchauffement climatique est une réalité » – A
« Le réchauffement climatique est l'affaire de tous » – B

	A	B	TOTAL
affaire	0	1	**1**
climatique	1	1	**2**
de	0	1	**1**
est	1	1	**2**
l	0	1	**1**
le	1	1	**2**
réalité	1	0	**1**
réchauffement	1	1	**2**
tous	0	1	**1**
une	1	0	**1**
TOTAL	**6**	**8**	**14**



  - Utilisation de l'indice de Jaccard
    - Niveau d'identité du stock lexical entre deux textes
    - Formule : $J(A,B) = \frac{|A \cap B|}{|A \cup B|}$
    - Si J(A,B) = 1 : identité lexicale parfaite ; 0 = aucune identité

Le Détricoteur

- Limites :
  - Pas adapté au TAL
    - Méthode lente (essai-erreur) inadaptée aux (très) gros corpus (jusqu'à plus de 1Mds mots en TAL [p.ex. Pomikálek et al., 2012])
    - P.ex. pour ma thèse (Dalodiere, 2023) : 5 corpus, 20 000 à 25 000 mots environ chacun
    - Différence épistémologique TAL / AD : sacrifier l'exactitude au profit du temps (TAL) ou l'inverse (AD) ?
    - AD : corpus représentatifs / exhaustifs : exactitude essentielle ; mise en place d'une norme de dépouillement

Le Détricoteur

- Deux utilisations envisageables :
  - Collection de pages de différents sites Web pour constituer un corpus thématique (p.ex., uniquement des URLs traitant d'environnement)
  - Analyser la cohérence de la communication d'une seule organisation sur l'ensemble de son site

Le Détricoteur

- Avantages :
  - Accessible à l'utilisateur non-développeur (solution prête à l'emploi)
  - Accès à toutes les pages listées pour un nom de domaine donné (pas de navigation manuelle du site)
  - Affinage progressif de l'extraction par réitération et / ou réimport de règles
  - Critères plus objectifs, partage de règles (copie des pages HTML lors de l'extraction
- Inconvénients :
  - Pas adapté aux gros corpus

Références :

Alexander, R. J. (2009). *Framing Discourse on the Environment: A Critical Discourse Approach*. NewYork/ Abingdon: Routledge.

Barbaresi, A., & Lejeune, G. (2020). *Out-of-the-Box and Into the Ditch? Multilingual Evaluation of Generic Text Extraction Tools.* Proceedings of the 12thWeb as Corpus Workshop, Language Resources and Evaluation Conference (LREC 2020), Marseille, France, 11–16 May 2020, 5-13.

Dalodiere, R. (2023). *Analyse du discours environnemental et sociétal de PME scandinaves et francophones : une approche textométrique*. Thèse de doctorat, Université de Mons, Mons.

De Schryver, G. D. (2002). Web for/as corpus: a perspective for the African languages. *Nordic Journal of African Studies, 11*(2), 266-282.

Fernández-Vázquez, J., & Sancho-Rodríguez, Á. (2020). Critical discourse analysis of climate change in IBEX 35 companies. *Technological Forecasting and Social Change, 157*, article 120063.

Genette, G. (1982). *Palimpsestes.* Paris : Seuil.

Gerdes, K. (2014). *Corpus collection and analysis for the linguistic layman: The Gromoteur.*JADT 2014 : 12e Journées internationales d'Analyse statistique des Données Textuelles, Paris, France, 3-6 juin 2014, 261-269.

Gupta, S., Kaiser, G.E., Neistadt, D., & Grimm, P. (2003). *DOM-based content extraction of HTML documents*. Proceedings of the 12thinternational conference on World Wide Web (WWW '03), Budapest, Hongrie, 20-24 mai 2003, 207-214

Jordan, M. (2020). Le Détricoteur, v.1.6c. Disponible sur demande

Kilgarriff, A., Baisa, V., Busta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography, 1*, 7-36.

Lejeune, G., & Barbaresi, A. (2020). *Bien choisir son outil d'extraction de contenu à partir du Web*. Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22^{e} édition), Nancy, France, 08-19 juin 2020, 46-49.

Pomikálek, J. (2011). *Removing Boilerplate and Duplicate Content from Web Corpora*. Thèse de doctorat, Masaryk University, Brno.

Pomikálek, J., Jakubíček, M., & Rychlý, P. (2012). *Building a 70 billion word corpus of English from ClueWeb*. 8thInternational Conference on Language Resources and Evaluation, Istanbul, Turquie, 21-27 mai 2012, 502-506

Sitri, F., & Barats, C. (2017a). Introduction. In Née, É. (dir.), *Méthodes et outils informatiques pour l'analyse des discours*. Rennes : PUR, 9-16.

Tripathy, A.K., Joshi, N., Thomas, S., Shetty, S., & Thomas, N.H. (2012). *VEDD-a visual wrapper for extraction of data using DOM tree*. 2012 International Conference on Communication, Information & Computing Technology (ICCICT),Mumbai, Inde, 19-20 octobre 2012,1-6.

Merci pour votre attention !

Romuald.dalodiere@umons.ac.be