Extraction de contextes de citations dans un corpus de publications scientifiques

Valérie Verdenet

Extraction de contextes de citations dans un corpus de publications scientifiques

VERDENET
Valérie

Sous la direction de :

Haralambous Yannis

Ringeval Fabien

Laboratoires : LabSTICC, LIG

UFR LLASIC
Département I3L
Section Industrie de la Langue

Mémoire de Master 2 mention Sciences du Langage - 20 crédits

Parcours : Industries de la Langue

Année universitaire 2016-2017

Extraction de contextes de citations dans un corpus de publications scientifiques

VERDENET
Valérie

Sous la direction de :

Haralambous Yannis

Ringeval Fabien

Laboratoires : LABSTICC, LIG

UFR LLASIC
Département I3L

Mémoire de Master 2 mention Sciences du Langage - 20 crédits

Parcours : Industries de la Langue, orientation professionnelle

Année universitaire 2016-2017

Remerciements

Je remercie les différentes personnes qui m'ont permis de réaliser ce travail :

- L'équipe pédagogique du master, dirigée par le Pr Georges Antoniadis, pour ses enseignements et son soutien.
- Pr. Yannis Haralambous, le maître de stage qui m'a donné ma chance.
- Pr Fabien Ringeval, pour avoir accepté d'encadrer mon stage.
- Damian Marek Janickowski, élève en 2e année d'école d'ingénieur, pour son aide en programmation au cours des 2 dernier mois.
- Les étudiants du master pour leur aide ponctuelle.

UNIVERSITÉ Grenoble Alpes

DÉCLARATION

1. Ce travail est le fruit d'un travail personnel et constitue un document original.
2. Je sais que prétendre être l'auteur d'un travail écrit par une autre personne est une pratique sévèrement sanctionnée par la loi.
3. Personne d'autre que moi n'a le droit de faire valoir ce travail, en totalité ou en partie, comme le sien.
4. Les propos repris mot à mot à d'autres auteurs figurent entre guillemets (citations).
5. Les écrits sur lesquels je m'appuie dans ce mémoire sont systématiquement référencés selon un système de renvoi bibliographique clair et précis.

NOM : VERDET

PRENOM : Valérie

DATE : 04-09-2017 SIGNATURE : [redacted]

Sommaire

Introduction

Le Laboratoire des Sciences et Techniques de l'Information, de la Communication et de la Connaissance (LabSTICC) existe depuis 2008, en partie grâce à la volonté du CNRS de regrouper diverses unités(« Présentation - Lab-STICC »). Il est composé de trois pôles scientifiques multi-sites et multi-établissements. Le fonctionnement repose sur la coopération entre ces différents pôles et regroupe au totale 562 personnes. Nous pouvons noter des partenariats internationaux. De ce fait, les chercheurs ont besoin de connaître la portée de leurs recherches. Pour diffuser leurs résultats, ils passent par la rédaction et la diffusion d'articles, souvent écrits en anglais pour augmenter leur visibilité. Un bon indice de popularité est l'ensemble de citations que l'on peut faire de ces écrits.

La discipline qui s'en charge est la bibliométrie. Ses origines datent de 1850 indépendamment de l'informatique. Elle a été modélisée en 1934 par Samuel C. Bradford, et est une ressource pour les bibliothécaires. Il s'agit d'une méthode statistique appartenant aux sciences de l'information pour évaluer la recherche à différentes échelles (chercheurs, laboratoires, établissements…). Des raisons variées conduisent à utiliser ce système d'évaluation (choix d'un sujet de thèse, état de l'art…).

Plusieurs types d'outils ont été créés parmi lesquels nous pouvons compter :

- Ceux en accès libres :
    - Google Scholar, CiteseerX, Citebase, Publish Or Perish, SCImago Journal & Country Ran
- Parmi les plus utilisés (d'après le site de l'Université de Bretagne Loire, Bertignac), en accès restreint :
    - Web of Science (Thomson Reuters), Journal Citation Report (Thomson Reuters), Scopus (Elsevier), SIGAPS.

Les principaux indicateurs sont :

- Le facteur d'impact (impact factor) (Deboin, Fovet-Rabot, & Lambert, 2017) : c'est un indicateur de notoriété des revues que l'on calcule sur deux ans. Il s'agit du

nombre moyen de citations d'un article de la revue rapporté au nombre d'articles que publie la revue.

- L'indice *H* (Gallets) : cela signifie qu'un auteur a au moins *H* publications citées *H*fois.

Il en existe d'autres comme :

- Scimago journal Rank : c'est une variante du facteur d'impact d'une revue. Son calcul est plus complexe parce qu'il pondère chaque citation par le « prestige » de la revue citante et ne prend pas en compte les autocitations de la revue.

- L'Eigenfarter : il est basé sur le nombre de citations des articles d'une revue sur 5 ans. Il ajoute au facteur d'impact une pondération sur les revues citées, et ne tient pas compte des autocitations de la revue.

- ou l'indice *G* : c'est une variante de l'indice *H*. « Il lui est toujours supérieur. C'est le rang le plus élevé au niveau duquel les publications cumulent un nombre de citation g^2. »…

Cependant, nous pouvons noter des limites à ces indicateurs de popularité. Il peut conduire à des dérives, comme le précisent les auteurs de l'article *Les dérives de l'évaluation de la recherche. Du bon usage de la bibliométrie* (Gringas & Caraco, 2014), comme augmenter l'autocitation pour permettre une meilleure visibilité dans les moteurs de recherches. Ce fait est nécessaire mais fausse les résultats en termes de popularité.

De plus, même si cette discipline regroupe de nombreux critères de classement, elle semble se situer uniquement au niveau quantitatif. Or nous ne pouvons négliger l'aspect qualitatif d'un article. La manière dont ce dernier sera cité pourrait influencer notre approche vis-à-vis des théories défendues et de sa crédibilité. En outre, ce type d'informations pourrait permettre au chercheur d'approfondir sa réflexion.

Le projet du LabSTICC est de proposer à son équipe un outil qui lui permettra d'avoir une perception plus précise de l'impact de ses écrits sur le monde scientifique. Dans ce but, nous parserons un corpus d'écrits scientifiques afin d'extraire des citations pour observer comment celles-ci sont intégrées. Pour cette observation, nous analyserons

syntaxiquement et sémantiquement les phrases extraites, comprenant des citations. À partir de là, nous établirons une grille de scores permettant de quantifier l'impact du travail réalisé sur la communauté.

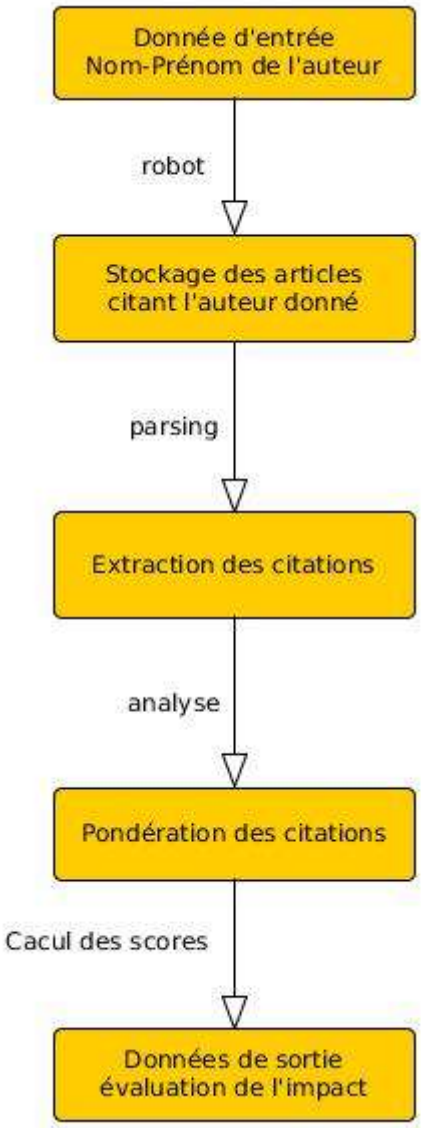


Figure 1. Schéma du modèle à concevoir

Chapitre 1. Analyse documentaire

Notre travail nous a conduit à nous intéresser à deux points essentiels : la structure des documents et la notion de contexte linguistique. Ce sont ces deux aspects que nous allons développer dans ce chapitre.

1. Structure des documents

Pour mener à bien ce travail, il a fallu constituer un corpus d'articles enregistrés sous format PDF et issus de différentes revues. Avant d'engager concrètement l'extraction de citations pour en analyser leurs contextes, il paraît important de comprendre comment se structure ce type de document d'une part, et les articles scientifiques d'autre part.

1.1. LE PDF

Le document PDF n'est pas un fichier texte linéaire ordinaire (Zoonekynd, 2001), (« Structure et construction d'un fichier PDF »). Il est composé d'objets. Il a huit types de données de bases (booléens, numériques, chaînes de caractères, noms, tableaux, dictionnaires, flux et objet nul).

Le fichier est structuré par :

- Un en-tête qui spécifie la version du PDF.

- Le corps du document qui est composé des objets.

- Xref qui est une table de références croisées.

- Trailer qui indique le 1er objet à lire.

Les objets sont des dictionnaires qui s'appellent les uns les autres. Chacun donne une information sur la structure, permettant de fournir un tout cohérent et organisé pour le résultat final.

Il existe différentes manières de créer un PDF. Le langage LaTex permet de décrire l'aspect final du document qui sera produit. La plupart des supports de diffusion (revues, livres…) donnent un modèle type à suivre avec des polices de caractères spécifiques, des liens et des agencements particuliers, pour avoir une certaine uniformité dans leurs articles. Nous avons également la possibilité de faire une conversion d'un document Word ou OpenOffice directement en PDF, ce qui peut entraîner pour nous des difficultés de

traitement. En effet, nous avons noté que l'encodage automatique était moins robuste que celui réalisé directement en LaTex. Nous avions plus d'erreurs de conversion avec des fichiers issus de Word par exemple.

1.2. Structure des articles

Pour cette étude, nous nous sommes concentrés sur les textes en anglais, plus présents dans la littérature. Nous pouvons noter d'un article scientifique à l'autre des correspondances, dans leur structure. Il convient d'en délimiter les grandes lignes comme suit : un résumé, le corps du texte puis les références bibliographiques. Ces trois grandes parties vont permettre de délimiter l'importance de la citation, et récupérer les références dont il est question.

1.2.1. Le résumé

Le titre est, en général, de la même police que le texte, éventuellement en gras. Cependant, dans certaines revues, comme Elsevier, il peut y avoir une disposition particulière comme des espaces entre les lettres des mots, un tiret entre le titre et le texte.

1.2.2. Le corps du texte

Le corps du texte débute de plusieurs manières :

- Le titre de la section, « Introduction », qui est facilement repérable et est mis en place dans les mêmes dispositions que le résumé, en ce qui concerne la typographie et dont nous devrons tenir compte dans le programme pour l'extraction des données.
- L'utilisation d'une numérotation (chiffres arabes ou romains) suivie du terme « Introduction ».

Le deuxième cas peut-être problématique, dans la mesure où nous pouvons trouver ce type de graphie à l'intérieur même du texte du résumé. Il convient de déterminer un ordre de priorité dans l'exécution du programme pour limiter au maximum ce risque.

1.2.3. Références

Il s'agit de la bibliographie et sa mise en forme répond à un ordre précis :

1) Un éventuel numéro de référence,

2) le(s) auteur(s),

3) le titre de l'ouvrage/article,

4) des renseignements complémentaires tels que la revue, l'année, l'édition, les pages…

Pour repérer l'élément introduisant la bibliographie, nous devons tenir compte de différentes formulations qui peuvent être employées dans l'ordre de priorité suivant :

1) « Reference », en veillant à avoir en priorité le pluriel, le singulier étant souvent employé dans le corps du texte pour faire appel à un point particulier de l'argumentation. Nous évitons ainsi des ambiguïtés. Toutefois, nous devons tout de même le proposer en dernier recours dans le programme qui sera créé.
2) « Bibliography », qui est moins utilisé et évite toute ambiguïté.
3) « Literature », rarement utilisée mais dont nous devons tenir compte.

Comme pour le résumé et le corps du texte, il faudra prévoir les variations typographiques telles que la casse, les espacements…

Il faut également faire la différence entre citation, référence et appel de références. Ces termes sont définis dans le glossaire.

2. Le contexte linguistique

Pour déterminer le contexte linguistique de notre corpus, nous avons dû d'abord nous interroger sur sa définition. Ensuite, pour pouvoir proposer un programme automatique d'analyse d'autres citations possibles, nous avons d'abord effectué ce travail manuellement. Cela nous a permis de décider des outils existants à utiliser et des améliorations à leur apporter.

2.1. Définition

Il existe toute une littérature sur le contexte linguistique. Georges Kleber, Strasbourg II, l'évoque dans le 4e Colloque de Pragmatique de Genève (Kleiber, 1989, p. 241-258) et propose une approche sur le processus interprétatif. Selon lui, « les données linguistiques ne suffisent pas à « donner » le référent visé ». Si une approche pragmatique est importante, il ne faut pas négliger la sémantique. Il plaide pour plus de sémantique dans l'analyse, sans occulter, cependant, les marqueurs référentiels.

De même, O. Ducrot et J-M. Schaffer nous dressent un panorama de la situation du discours dans le *Nouveau dictionnaire encyclopédique des sciences du langage* (Ducrot &

Schaeffer, 1995), en se référant à quelques grands noms de la linguistique tels que Tesnière et Martinet. Ils nous exposent ainsi ce qu'est le contexte linguistique, « l'entourage linguistique d'un élément selon la terminologie traditionnelle ». Il concerne les relations sémantiques entre les phrases, l'anaphore et les relations syntagmatiques des éléments.

Tous s'accordent pour évoquer les principes de référence et d'anaphore, tout en admettant que le rapport entre les deux est délicat à cerner. En effet, il s'agit de déterminer précisément les frontières entre les mots et les liens de dépendance entre chacun, tout en tenant compte des problèmes d'ambiguïté. Nous avons été confrontés à cette difficulté dans l'élaboration de notre programme. En ce qui nous concerne, nous nous sommes fixés sur le fait qu'il s'agit de « l'entourage linguistique d'un élément », l'appel de référence, noté :

- [10]

- [10, 12]

- (nom de l'auteur, année)…

2.2. Analyse pour notre corpus

Pour avoir une bonne connaissance des liens de dépendance, nous avons procédé à une analyse manuelle de notre corpus. Nous avons utilisé à la fois des étiquettes existantes et nous en avons créées spécifiquement pour notre usage. Par exemple, nous désignons l'appel de référence par l'appellation « REF ». Ce type d'étiquette n'existe pas au niveau grammatical, mais il est nécessaire pour reconnaître automatiquement les références.

Nous avons ainsi déterminé les liens de dépendance au niveau syntaxique (nominal, verbal...) et les caractéristiques positives et négatives à prendre en compte au niveau sémantique. Par exemple, « low cost » sera considéré positif alors que « high cost » sera plutôt négatif.

Nous avons séparé les citations dont l'appel de références était intégré dans les phrases des autres citations. Les premières ont plus de valeur qualitative, car l'auteur montre qu'il s'appuie concrètement sur sa source pour évoquer sa théorie.

Cela nous a permis d'élaborer notre programme et de déterminer nos besoins. Nous avons choisi d'utiliser des outils déjà existants (NLTK, parseur de Stanford) pour obtenir la correspondance jeton-étiquette et les types de relations syntaxiques. Nous avons adapté ces outils à notre objectif qui est de déterminer l'impact d'une citation dans un article.

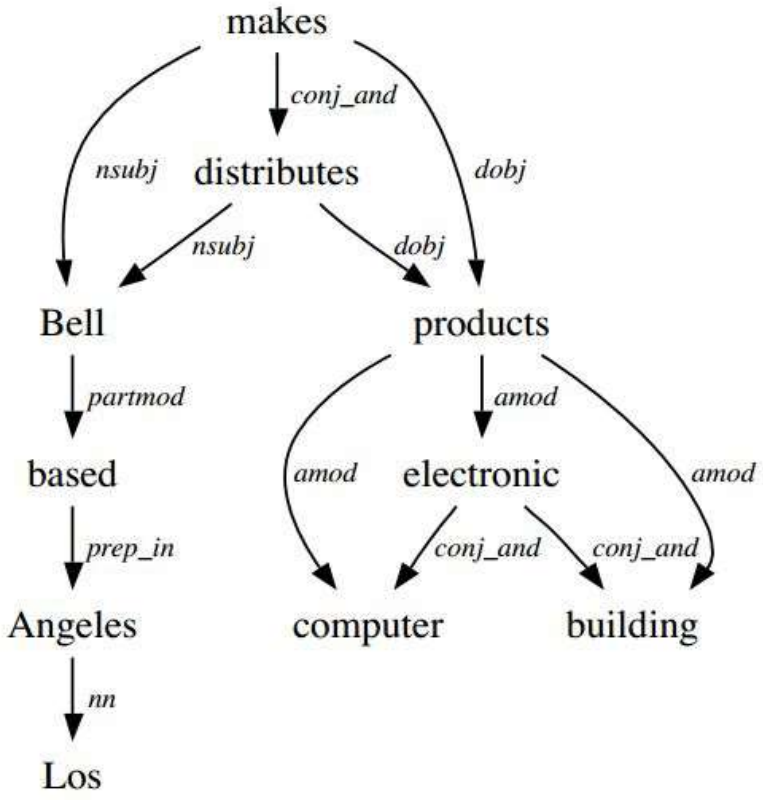


Figure 2. Liens de dépendance: source Stanford

Chapitre 2. Choix des outils

Notre travail, de l'extraction des citations à l'analyse quantitative de ces citations, a été réalisé sous Unix avec le langage Python. Nous avons également utilisé des outils pré-existants (Google Scholar, PDFMiner, NLTK, parseur de Stanford).

1. Extraction des citations

1.1. Modules python sous Unix

Le langage de programmation Python a été choisi pour la diversité de modules qu'il propose en ce qui concerne la réalisation de programmes. Pour l'extraction des citations, il était nécessaire de convertir les documents PDF vers un format texte. Après plusieurs tests, expliqués dans la section 1.2, nous avons choisi une conversion au format HTML.

Pour la conversion d'un document PDF vers un autre format nous avions le choix entre deux outils :

- Pdftotext : un module spécialement conçu pour le langage Unix uniquement.
- PDFMiner : un module Python qui permet la conversion des PDF aussi bien sous Unix que sous DOS, à condition de créer un script permettant d'utiliser les différentes parties du module.

PDF Miner a été retenu car la conversion vers le HTML était la plus efficace, malgré quelques ajustements à faire, tels que l'organisation des différents segments.

1.2. Tests de conversions

Pour choisir un format de conversion des textes, nous avons effectué plusieurs essais. Le XML a été écarté car il produisait un caractère par balise.

PDF vers le format Txt : Cela semble être la solution la plus simple. Cependant, la mise en forme initiale (police et taille de caractères) est supprimée alors qu'elle pourrait être utile pour situer les citations dans le texte.

PDF vers le format HTML : l'avantage de cette solution est que nous conservons une trace de la mise en forme initiale via les balises. Ainsi, nous avons pu détecter les problèmes de conversion et les corriger :

- Mauvaise détection des colonnes entraînant des déplacements de bloc des textes dans le document.

Nous avons ainsi choisi la conversion vers le HTML pour deux raisons :

- Résoudre ce problème d'organisation graphique de blocs de texte.
- Pouvoir envisager de détecter les différents types de textes (titres, corps, références…) grâce aux données CSS incluses dans les balises HTML en utilisant des moyens statistiques.

2. Analyse qualitative des citations

Pour mettre en place notre programme, comme précisé plus haut (introduction du chapitre 2), nous avons utilisé des outils d'analyse linguistique: NLTK (Natural Language ToolKit) et l'analyseur de Stanford. Le but était de les adapter à notre projet.

2.1. NLTK (Bird, Klein, & Loper, 2009)

Cet outil, conçu en langage python, nous a permis de parser les éléments de citations, jeton par jeton, pour déterminer les étiquettes grammaticales de chaque terme. Cela nous a permis de faire un premier tri et d'obtenir la base de notre lexique pondéré.

Voici un extrait de ce lexique pondéré :

Lexème	Tag	Pondération
A/D	*NNP*	*0*
adopted	*VBN*	*0.5*
adopts	*VBZ*	*0.5*
adopt	*VBP*	*0.5*
a	*DT*	*0*
a	*DT*	*0*
A	*DT*	*0*
advanced	*VBD*	*0.5*
advancements	*NNS*	*0.5*
advances	*NNS*	*0.5*
advantage	*NN*	*0.5*
advantageous	*JJ*	*0.5*
advantages	*NNS*	*0.5*
AFE	*NNP*	*0*
affected	*VBN*	*-0.5*

Nous exprimons la valence d'un sentiment exprimé. Chaque jeton a par défaut un score de 0 (neutre). Le but est que les utilisateurs puissent, par la suite, incrémenter cette base de données, pour compléter le programme et l'entraîner sur des données plus précises.

2.2. L'analyseur de Stanford (« The Stanford Natural Language Processing Group »)

Pour obtenir automatiquement les dépendances syntaxiques entre les différents éléments de chaque citation, nous avons utilisé le modèle de Stanford, qui est repris dans de nombreux parseur comme Maltparser et programmé en langage Java. Il nous a permis de constituer un premier tableau d'analyse, en attribuant à chaque jeton un identifiant, une étiquette et un lien de dépendance. Nous avons, par la suite, modifié ce premier résultat en fonction de nos objectifs. Nous pouvons noter que l'analyse sur les liens de dépendance se fait en plusieurs étapes.

```
nsubj(makes-11, Bell-1)
det(company-4, a-3)
appos(Bell-1, company-4)
nsubjpass(based-7, which-5)
auxpass(based-7, is-6)
rcmod(company-4, based-7)
prep(based-7, in-8)
pobj(in-8, LA-9)
root(ROOT-0, makes-11)
cc(makes-11, and-12)
conj(makes-11, distributes-13)
nn(products-15, computer-14)
dobj(makes-11, products-15)
```

Figure 3. Exemple de données produites par l'analyseur de Stanford

Dans un premier temps, l'outil décompose la phrase en plusieurs groupes grammaticaux pour ensuite déterminer les liens entre les lexèmes, auxquels il attribue des identifiants.

1	Bell	_	NNP	NNP	_	11	nsubj	_	_
2	,	_	,	,	_	1	punct	_	_
3	a	_	DT	DT	_	4	det	_	_
4	company	_	NN	NN	_	7	nsubjpass	_	_
5	which	_	WDT	WDT	_	0	erased	_	_
6	is	_	VBZ	VBZ	_	7	auxpass	_	_
7	based	_	VBN	VBN	_	4	rcmod	_	_
8	in	_	IN	IN	_	0	erased	_	_
9	LA	_	NNP	NNP	_	7	prep_in	_	_
10	,	_	,	,	_	1	punct	_	_
11	makes	_	VBZ	VBZ	_	0	root	_	_
12	and	_	CC	CC	_	0	erased	_	_
13	distributes	_	VBZ	VBZ	_	11	conj_and	_	_
14	computer	_	NN	NN	_	15	nn	_	_
15	products	_	NNS	NNS	_	11	dobj	_	_

Figure 4. Modèle de Stanford attendu

Dans un second et dernier temps, nous réunissons dans un tableau, comme ci-dessus, les différents éléments : identifiant, forme, tag, identifiant de tête de dépendance et type de dépendance. Nous avons également pu, à partir de ce module, proposer notre pondération des liens de dépendance.

Dans notre programme, nous avons un fichier composé, comme dans l'exemple suivant, du type de dépendance, de sa pondération, et de sa définition, selon la documentation de Stanford.

Type de dépendence	*Pondération*	*definition de la dépendance*
xcomp	*0*	*open clausal complement: An open clausal complement (xcomp) of a verb or an adjective is a predicative or clausal complement without its own subject. The reference of the subject is necessarily determined by an argument external to the xcomp (normally by the object of the next higher clause, if there is one, or else by the subject of the next higher clause. These complements are always non-finite, and they are complements (arguments of the higher verb or adjective) rather than adjuncts/modifiers, such as a purpose clause. The name xcomp is borrowed from Lexical-Functional Grammar.*
xsubj	*0*	*controlling subject: A controlling subject is the relation between the head of a open clausal complement (xcomp) and the external subject of that clause. This is an additional dependency, not a basic depedency.*
lowcost	*0.5*	*when "low" is followed by "cost"*
longterm	*0.5*	*when "long" is followed by "term"*

Figure 5. Pondération des liens de dépendance

Chapitre 3. Création d'un module spécifique

À partir des outils déjà existants et des recherches documentaires réalisées sur le sujet, nous avons pu élaborer un programme pour atteindre nos objectifs. (cf. annexe 1 p 39)

1. Création de scripts pour l'extraction de citations à partir d'un corpus

Les modules python déjà présents ne sont pas suffisants pour extraire correctement les citations. Dans un premier temps, nous avons effectué un traitement de normalisation des fichiers html, afin de corriger les erreurs de conversion ; dans un second temps, nous avons segmenté le document pour séparer abstract (s'il existe), corps du texte et références ; et enfin nous procéderons à l'extraction des citations.

to the sensitivity to the level of random switching activity.

Different solutions such as Edge Memories (EMs) [7] have

been suggested to solve this latching problem. In [15],

authors introduced an original idea to reduce the number of

clock cycles required to decode one codeword and to replace

the EMs by simple deterministic shufflers. The proposed

architecture, that uses multiple streams in parallel to

represent the same probability, presents the best architecture

 the

 ABSTRACT

Figure 6. Erreur de conversion: du corps de texte avant l'abstract

1.1. Récupération des articles

Avant d'extraire les citations, nous avons dû récupérer les articles dans lesquels elles étaient insérées. Pour cela, nous avons utilisé un script de *Google Scholar* qui permet de parser les bibliothèques et obtenir les résultats recherchés du point de vue quantitatif.

Une fois ces bases de données trouvées, nous avons programmé un robot qui récupère les articles sur les sites n'ayant pas de barrières anti-robots. Nous avons limité le nombre de pages par document à 60 afin d'éviter les ralentissements de téléchargement et de conversion. En effet, le script ne fait pas de distinction entre un article ou une thèse par exemple.

Nous avons ensuite pu, grâce à PDFMiner, convertir les PDF au format HTML et commencer la normalisation de nos documents pour en extraire les citations.

1.2. La normalisation de la structure du document converti

Pour normaliser les documents html, nous avons procédé en plusieurs étapes. Nous nous sommes basés sur les informations de positionnement basés dans les balises « » et les balises « <a page>».

Avant d'arriver à cette étape, nous avons nettoyé le document en supprimant les éléments inutiles tels que les balises d'images, les retours chariots superflus, les balises vides, les tirets de césure. En effet, la lecture des PDF se fait colonne par colonne et la fin d'une ligne est considérée comme un retour à la ligne, symbolisé par la balise «
 ». La finalité était d'avoir sur une seule ligne les mêmes éléments encadrés dans une balise « ».

Une fois cette première manipulation réalisée en plusieurs étapes, nous avons pu procéder à la remise dans l'ordre des différents éléments. Pour cela, nous avons défini une barrière séparant colonne de gauche et colonne de droite à partir de la marge gauche inscrite dans les balises « ». Les différents éléments ont été extraits dans des listes séparées pour nous focaliser sur les marges de gauche, droite et les numéros de pages. Les indices des listes nous ont permis de conserver leur structure. Grâce à une fonction de tri, nous avons pu rétablir le bon ordre des paragraphes.

1.3. La segmentation du document en vue de l'extraction des citations

Pour faciliter l'extraction des citations et leur détection dans les références, nous avons choisi de segmenter les documents en 3 parties : résumé, corps du texte, et références.

Pour cela, nous avons déterminé des frontières en fonction des appellations pour l'introduction et les références sous forme d'expressions régulières.

```
#repérage du début de l'article
introAbst                                                                                    =
re.search(r'((<span(.*?)>)(.*?)((I|i)(N|n)(T|t)(R|r)(O|o)(D|d)(U|u)(C|c)(T|t)(I|i)(O|o)(N|n)))',load)
intro = re.search(r'((<span(.*?)>)((I|i)(N|n)(T|t)(R|r)(O|o)(D|d)(U|u)(C|c)(T|t)(I|i)(O|o)(N|n)))',load)
introRom = re.search(r'((<span(.*?)>)(.*?)((V|X|C|M|L)*I+(V|X|C|M|L)*\s*\.\s*(\w*?)))',load)

introChiffre = re.search(r'((<span(.*?)>)(.*?)([0-9]\.\s))',load)
```

Figure 7- Expressions régulières pour la segmentation du document en fonction du titre « introduction » et de la manière dont il est introduit.

Il nous a paru opportun de séparer le résumé du reste du corps du texte. C'est un endroit du texte qui permet avant tout de résumer les grandes lignes de l'article. La norme est de ne pas citer d'auteur dans cette section.

1.4. Extraction des références et des citations

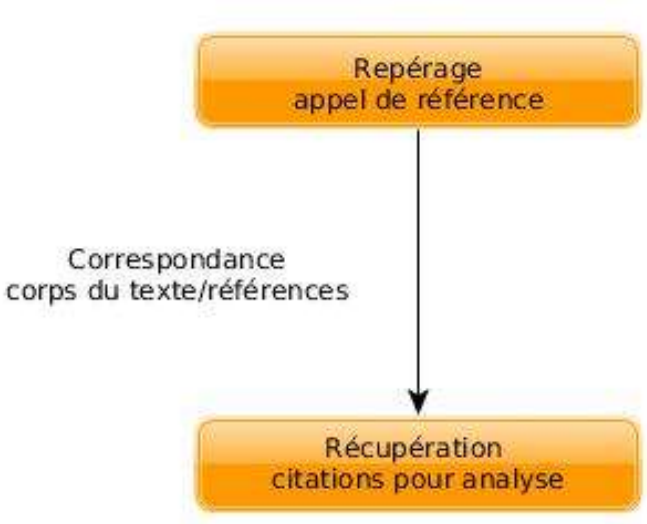


Figure 8. Extraction des références

Dans un premier temps, nous avons isolé les phrases avec des citations présentes dans l'article en les repérant grâce aux appels de références.

Ensuite, nous avons placé les références dans un dictionnaire avec pour clé l'appel de référence.

Pour finir, nous avons opéré une correspondance entre les citations et les références en plaçant dans un tableau HTML la citation avec la référence correspondante.

Pour le moment, il ne s'agit pas du résultat final attendu, mais d'une première base pour isoler des citations et comprendre leur fonctionnement pour les évaluer et pouvoir

reproduire cette évaluation automatiquement. Ce programme sera utile dans la recherche des citations d'un même ouvrage dans divers articles.

2. L'analyse qualitative des citations

2.1. Analyse manuelle

Nous avons effectué une analyse manuelle du corpus de citations, ainsi que des parcours d'arbres syntaxiques. Cela nous a permis d'identifier les besoins et difficultés à venir pour notre programme. Pour cette analyse, nous avons procédé par groupes grammaticaux, comme montré dans l'exemple ci-dessous. La première colonne désigne l'étiquette du groupe principal, la deuxième la composition étiquetée du groupe (soit des sous-groupes, soit des terminaux) définie dans la première et la troisième représente le morceau textuel désigné. Quand nous avons un indicateur nous permettant de déterminer que la référence est intégrée dans la citation, nous l'avons signalé dans une quatrième colonne par l'expression « ref in quot ». Ce travail a été réalisé sur 500 citations.

```
P          SAPPO SN SV    In case [25], 17 use case have been defined
SAPPO      SPREPREF PUNCT      In case [25],
SPREP      IN REF  In case [25]      quot in sentence
SN         CD SN   17 use case
SN         NN NN   use case
SV         VBP SV  have been defined
SV         PASSIVB           been defined
PASSIVB          BE VBN been defined
```

Figure 9. Analyse syntaxique manuelle.

Notre corpus a été mieux maîtrisé. Nous avons également testé la conception d'un outil *ad hoc*. Même s'il n'a pas abouti, il a eu pour utilité d'accéder à une plus grande connaissance de notre sujet d'étude et des contraintes de l'analyse syntaxique automatique.

Nous avons ensuite pu élaborer notre pondération en fonction des critères essentiels qui constituent l'ensemble du contexte linguistique : les liens de dépendance et la sémantique.

2.2. Adaptation du modèle de Stanford

Les citations ont été traitées par le modèle de Stanford dans un premier temps. Nous avons ensuite parser ce résultat pour faire les modifications en fonction:

- du lexique (REF, REFMIN...)

- des liens de dépendance s'il y a lieu (prepref...)

Une fois les changements apportés, nous avons réalisé la correspondance avec les scores attribués pour obtenir la figure suivante:

```
ID      FORM    TAG     ID_POST_TAG     POST_TAG        REL     SCORE
1 Consequently , both , the FET tripler and doubler benefit in terms of bandwidth from a balanced topology[1] .
1       Consequently    RB      10      VB      ADVMOD  0
2       ,       PUNCT   1       RB      APPOS   0
3       both    DT      10      VB      DET     0
4       ,       PUNCT   1       RB      PUNCT   0
5       the     DT      7       NN      DET     0
6       FET     NNP     7       NN      COMPNOUN        0
7       tripler NN      3       DT      CONJ:AND        0
8       and     CC      7       NN      CC      0
9       doubler NN      3       NN      CONJ:AND        0
10      benefit VB      0       ROOT    ROOT    0.5
11      in      IN      12      NN      SPREP   0
12      terms   NNS     10      VB      NMOD    0
13      of      IN      14      NN      SPREP   0
14      bandwidth       NN      12      NN      NMOD:OF 0
15      from    IN      18      NN      SPREP   0
16      a       DT      18      NN      DET     0
17      balanced        JJ      18      NN      AMOD    0
18      topology        NN      10      VB      NMOD    0.5
19      [1]     REF     0       ROOT    APPOS   0
20      .       PUNCT   0       ROOT    PUNCT   0
```

La dernière colonne est le score réalisé selon la pondération du lexique et du lien de dépendance que nous avons pu initialiser dans nos grammaires et lexiques. L'idéal serait de pouvoir présenter plusieurs types de scores:

- Une moyenne pondérée en fonction de la longueur de la phrase, sur un indice de 1. Une phrase de 10 mots peut avoir plus de poids qu'une phrase de 20 mots.

- Une moyenne pondérée de « retours négatifs » en fonction du nombre de jetons par phrase puis sur le nombre total de phrases.

- Une moyenne pondérée de « retours positifs » en fonction du nombre de jetons par phrase puis sur le total.

Figure 10. Schéma de notre modèle d'analyse

Chapitre 4. Évaluation

Le programme créé est une base d'évaluation pour permettre aux chercheurs d'aller plus loin dans l'analyse de leurs écrits et de leurs recherches. La simple extraction des citations peut déjà donner un premier aperçu qualitatif au sujet de l'influence qu'un scientifique a sur ses pairs.

Les fonctions supplémentaires d'analyse ne peuvent être pleinement évaluées dans l'immédiat. Pour cela, il faudrait un panel d'utilisateurs qui pourraient juger manuellement les citations, en parallèle de l'outil. Nous pourrions ainsi noter le taux d'erreurs et de bonnes réponses plus précisément. Cela n'a pas été possible pour une question de temps et d'éventuelle disponibilité d'un public en mesure d'évaluer les productions.

1. Extraction des citations

1.1. Critères d'évaluation

L'évaluation du programme se situe sur 3 axes principaux : la normalisation des documents, la segmentation et l'extraction des citations. Pour cela nous avons créé un corpus de 100 documents. Ils ont été convertis au format choisi.

En ce qui concerne la normalisation, nous avons 11 étapes qui correspondent aux différentes étapes du processus afin de pouvoir isoler les points d'amélioration du programme et déterminer si nous pouvons corriger les erreurs à partir de notre programme ou s'il faudrait agir directement sur la conversion du PDF vers le html. Nous avons établi un score de la manière suivante :

- 1 à chaque étape réalisée correctement par document, avec une mention d'un éventuel correctif à envisager pour plus d'efficacité.
- 0.5 si l'étape été réalisée partiellement. Par exemple, nous avons pu supprimer ce qu'il y avait avant l'abstract (titre du document et auteur) mais il y avait également des informations du corps du texte qui se situaient dans cette partie et ils ont été eux aussi supprimés.
- 0 si l'étape n'a pas pu être réalisée.

Pour la segmentation, nous avons le même système de notation avec 3 parties (abstract, corps et références).

Pour l'extraction de citations, la procédure est différente. Nous basons l'évaluation sur le ratio de correspondances entre le nombre de citations détectées et les références. Par exemple, si 20 citations sont détectées dont 19 sont associées à des références, notre score est de 95%. Cela nous permet d'évaluer de la robustesse du programme.

Tous les points qui doivent attirer notre attention pour une correction ou une amélioration sont surbrillés en rouge, afin d'être traités en priorité.

1.2. Résultats d'évaluation (annexe 2 p 40)

Dans l'ensemble nous obtenons de bons résultats.

Lors de la première passe (annexe 2 p39), nous avons obtenus :

Normalisation	Segmentation			Extraction/correspondance
	Résumé	Corps du texte	Références	
93.66%	92.08%	92.08%	93.56%	86.5%

De manière générale, les résultats sont bons. Cependant, des erreurs restaient à corriger. Pour cela, Damian Janickowski, étudiant de 2^{e} année en école d'ingénieur a repris le programme pour l'améliorer.

La difficulté principale d'extraction des citations réside dans le fait que le format PDF est prévu pour être fixe, comme une image du texte. Nous avons donc dû faire face à des problèmes de réorganisation des différents segments d'écriture et à la structure du PDF en pour ce qui concerne la mise en forme. Ils ont pu être réglés.

2. Analyse qualitative des citations

En ce qui concerne l'analyse qualitative des citations, nous avons dû gérer à la fois les liens de dépendances et la sémantique, deux éléments importants dans l'intentionnalité du discours. Cet ensemble peut conduire à de nombreuses erreurs dans nos résultats. Il est donc nécessaire de bien les comprendre pour les corriger.

2.1. Critère d'évaluation

Pour évaluer notre parseur qualitatif, nous avons utilisé le principe de la mesure BLEU. Nous somme parti du nombre de jetons par phrase, puis nous avons calculé le nombre de bonnes réponses :

- en termes de choix du Tag/étiquette ;

- de lien de dépendance ;

- du type de dépendance ;

- de la pondération attendue.

Cela nous a permis d'obtenir un comparatif entre le réel et ce que nous attendions effectivement d'un tel programme. Nous avons ainsi pu connaître les points d'amélioration à apporter et envisager un avenir à ce programme.

2.2. Résultats (annexe 3 p 41)

Les résultats obtenus restent au-dessus de 68% de bonnes réponses, comme nous pouvons le noter dans l'annexe 2, sur un corpus aléatoire de 14 citations. Cependant, ils sont perfectibles. Notre système a été évalué sur 4 points comme présenté dans le tableau suivant :

Étiquette attendue	Lien de dépendance attendu	Type de dépendance attendu	Pondération attendue
92%	70.37%	68.01%	83.56%

- Les erreurs observées au niveau des étiquettes attendues sont dues à des ambiguïtés dans le lexique :
    - Le mot peut être un nom ou un verbe au gérondif par exemple.
        - Coding = le codage
        - Coding = en codant
    - Le parseur de Stanford a été paramétré pour considérer tout mot commençant par une majuscule comme un nom propre.
- Pour mieux comprendre le résultat sur les liens de dépendance attendus, il nous faudrait étudier à la fois notre analyse manuelle et la manière dont a été

paramétré le modèle de Stanford. Nous n'avons pas accès aux scripts Java du modèle pour agir directement dessus. Cependant, nous avons réussi, grâce à un traitement supplémentaire, à obtenir de bons résultats.

- Pour le type de dépendance attendue, nous avons pu ajouter nos propres types, mais pour progresser, il faudrait, comme pour les liens de dépendance, avoir accès aux scripts du modèle de Stanford.
- Les bons résultats de la pondération semblent être liés essentiellement à la sémantique, puisque les liens de dépendance et les tags ne sont pas toujours les bons.

Une citation affiche des résultats particulièrement bas. Il se trouve qu'elle contient un calcul. Les liens de dépendances sont ainsi faussés. Cela reste rare, mais il faudra envisager une solution à terme. De même, le lien de dépendance concernant les conjonctions de coordination restent erronées.

3. Analyse des difficultés rencontrées et solutions

3.1. Les bloqueurs de robots

La plupart des systèmes de sécurité des sites des revues qui proposent les articles à parser sont équipés de systèmes de sécurité refusant l'accès aux robots ou exigent un abonnement pour limiter la diffusion massive de leurs productions payantes. Des erreurs étaient provoquées dans nos programmes. Nous avons choisi de placer une règle d'exception pour ne sélectionner que les sites accessibles.

Cela a pour conséquence de réduire le nombre de citations à analyser et fausser en partie nos résultats. Actuellement, nous n'avons pas d'autre solution. Il faudrait pouvoir établir une forme de partenariat avec toutes les plateformes pour autoriser l'accès à notre programme, ce qui est impossible pour le moment. L'utilisateur a, toutefois, la possibilité d'ajouter des documents en sa possession manuellement pour les faire analyser.

3.2. Le format PDF

Nous avons déjà évoqué le format PDF. Dans ce cas, il est très difficile d'agir. En effet, le format PDF est prévu pour être non modifiable. Nous avons sélectionné le convertisseur qui nous paraissait le plus adapté à notre projet présent et aux évolutions futures.

L'erreur produite se situe dans la récupération de données au niveau de la mise en forme. Grace au format HTML, nous avons pu récupérer l'organisation initiale du texte

dans la plupart des cas. Toutefois, il reste des erreurs dues à la conception du PDF en soi. Par exemple, nous n'avons pas la transcription des tableaux ou des notes de bas de pages; ceci crée une confusion dans la mise en forme avec des insertions inopinées dans le corps du texte.

3.3. Les ambiguïtés

Malgré toute la rigueur des analyses syntaxiques et sémantiques, nous devons tenir compte des problèmes d'ambiguïté. En effet, certains termes, selon leur emploi, peuvent être perçus négativement ou positivement. Tout dépend du contexte linguistique. Par exemple, le nom « cost », s'il accompagne:

- l'adjectif « lost » (qui, isolé, peut-être perçu négativement), sera considéré comme positif.

  - This method has a low cost of energy contrary to the first. Il nous est indiqué que le coût en énergie est bas ce qui est positif car il y a une économie.

  - We achieve low performance. Dans ce cas les performances sont basses, ce qui est négatif, car on espère, en général, des performances élevées.

-l'adjectif « high » (qui, isolé, peut-être perçu positivement), sera considéré comme négatif.

  - This method has a high cost of energy contrary to the first. Il nous est indiqué que le coût en énergie est haut, ce qui est négatif, car aucune économie par rapport à une autre méthode.

  - We achieve high performance. Dans ce cas les performances sont élevées, ce qui est positif, car on espère, en général, des performances élevées.

Nous avons créé des classes de dépendance particulières pour ce type de cas. Cependant, nous ne pouvons pas connaître toutes les possibilités. L'idéal serait de permettre aux utilisateurs de proposer une correction ou un ajout selon le cas trouvé.

3.4. Le modèle de Stanford

Le modèle de Stanford, comme vu dans l'évaluation présente des failles que nous avons palliées, en ajoutant nos règles après sa première analyse.

1	According	VBG	0	ROOT	erased	0.5
2	to	TO	9	VBP	prepc_according_to	0.5
3	Wasserman	NNP	9	VBP	pobj	0.0
4	and	CC	0	ROOT	erased	0
5	Faust	NNP	9	VBP	pobj	0.0
6	[1]	REF	8	NNS	refApp	0.5
7	social	JJ	8	NNS	amod	0.0
8	networks	NNS	9	VBP	nsubj	0.0
9	contain	VBP	0	ROOT	root	0.0
10	two	CD	13	NNS	num	0.5
11	different	JJ	13	NNS	amod	0.5
12	information	NN	13	NNS	nn	0.5
13	dimensions	NNS	16	VBN	nsubjpass	0.5
14	that	WDT	0	ROOT	erased	0.5
15	are	VBP	16	VBN	auxpass	0.0
16	represented	VBN	13	NNS	rcmod	0.5
17	by	IN	0	ROOT	erased	0
18	two	CD	19	NNS	num	0.5
19	variables	NNS	16	VBN	agent	0.5
20	:	:	13	NNS	punct	0.0
21	a	DT	23	NN	det	0.0
22	structural	JJ	23	NN	amod	0.0
23	one	NN	13	NNS	dep	0.5
24	and	CC	0	ROOT	erased	0
25	a	DT	27	NN	det	0.0
26	compositional	JJ	27	NN	amod	0.0
27	one	NN	23	NN	conj_and	0
28	.	.	9	VBP	punct	0.0

Figure 11. Erreurs sur le modèle de Stanford

Description de l'exemple 11 :

- Colonnes de gauche à droite : identifiant du jeton, jeton, étiquette du jeton, identifiant de dépendance, étiquette de dépendance, type de dépendance, pondération.
- Phrase : « According to Wasserman and Faust [1] social network contain two different information dimensions that are represented by two variables : a structural one and a compositional one. »

Sur ce tableau, nous remarquons qu'il y a plusieurs ROOT et qu'ils ne sont pas adaptés à l'analyse. Les liens de dépendance sont ainsi erronés. Il convient de pouvoir modifier ce fait.

1	*Consider*	*VB*	*4*	*VBP*	*dep*	*1.0*
2	*the*	*DT*	*3*	*NN*	*det*	*0.0*
3	*polynomial*		*NN*	*1*	*VB*	*dobj*
0.0						
4	*function*	*VBP*	*0*	*ROOT*	*root*	*0.0*
5	*-LRB-*	*-LRB-*	*6*	*CD*	*punct*	*0.0*
6	*1*	*CD*	*11*	*CD*	*nsubj*	*0.0*
7	*,*	*,*	*6*	*CD*	*punct*	*0.0*
8	*2*	*CD*	*6*	*CD*	*dep*	*0.5*
9	*-RRB-*	*-RRB-*	*6*	*CD*	*punct*	*0.0*
10	*=*	*SYM*	*11*	*CD*	*dep*	*0.5*
11	*0.4375*	*CD*	*4*	*VBP*	*ccomp*	*0.0*
12	*0.125*	*CD*	*4*	*VBP*	*nsubj*	*0.0*
13	*-LRB-*	*-LRB-*	*15*	*NNS*	*punct*	*0.0*
14	*1*	*CD*	*15*	*NNS*	*num*	*0.0*
15	*+*	*NNS*	*12*	*CD*	*dep*	*0.5*
16	*2*	*CD*	*15*	*NNS*	*dep*	*0.5*
17	*-RRB-*	*-RRB-*	*15*	*NNS*	*punct*	*0.0*
18	*1*	*CD*	*19*	*CD*	*number*	*0.0*
19	*2*	*CD*	*12*	*CD*	*dep*	*0.5*
20	*defined*	*VBN*	*19*	*CD*	*vmod*	*0.0*
21	*in*	*IN*	*22*	*REF*	*prepref*	*0.5*
22	*[1]*	*REF*	*20*	*VBN*	*refmod*	*0.5*
23	*.*	*.*	*4*	*VBP*	*punct*	*0.0*

Figure 12. Erreur liée à l'analyseur du modèle de Stanford

Description de l'exemple 11 :

- Colonnes de gauche à droite : identifiant du jeton, jeton, étiquette du jeton, identifiant de dépendance, étiquette de dépendance, type de dépendance, pondération.
- Phrase : « Consider the polynomial function –LRB- 1,2 –RRB- = 0,4375 0,125 –LRB- 1+5-RRB-1 2 defined in [1]. »

Nous observons sur cet exemple des erreurs dans l'analyse automatique. Le nom « function » est analysé comme un verbe, qui serait la racine du la phrase, son moteur. Or, le véritable noyau est le verbe « Consider », dont le sujet est implicite. Le reste des éléments constitue le complément d'objet. Cependant, nous voyons que nos modifications sont bien prises en compte (22 [1] REF 20 VBN refmod).

Nous avons donc proposé des solutions d'algorithmes pour obtenir des résultats plus fiables.

Conclusion

L’utilité de ce projet est d’avoir une représentation de l’impact des écrits scientifiques. Il présente un intérêt linguistique : comprendre comment un auteur peut laisser transparaitre son opinion à travers sa manière de construire ses phrases. Il s’est avéré nécessaire de vérifier nos acquis par un travail manuel avant l’automatisation, et ceci malgré les outils performants déjà existants. Cela est vérifiable avec les exemples dans le modèle de Stanford (cf. Chapitre 4.2).

Ce module Python, bien qu'avancé, pourrait bénéficier d'évolutions dans l'avenir pour le rendre plus efficace. Pour cela, il faudrait travailler sur plusieurs points.

Il serait intéressant de mener une évaluation plus poussée du module pour déterminer de manière plus détaillée les améliorations à apporter en ce qui concerne la pondération et l'impact effectif des références. Plusieurs possibilités s'offrent à nous :

- Le faire utiliser par des personnes autres que nous et établir un questionnaire sur des points tels que la facilités d'utilisation, les bugs, les performances...

- Comparer la partie extraction à celle de portails bibliographiques tels que IEEExpore.

- Proposer à des utilisateurs potentiels d'évaluer des citations, en tant que positives, négatives ou neutres. Puis comparer leurs résultats avec ceux trouvés par le programme.

Aucune de ces évaluations n’a été réalisée par manque de temps.

L'un des importants problèmes rencontrés étant le format PDF, il pourrait être intéressant de concevoir un analyseur de PDF plus performant et qui corresponde plus aux exigences du système. Cela induirait une étude approfondie de l'encodage en PDF, de sa structure et de son fonctionnement.

Par ailleurs, lors de l'analyse des articles scientifiques, nous avons pu noter une certaine disparité dans les différentes mises en forme des textes, que ce soit au niveau des titres, des polices de caractères, de la casse ou de la création en PDF. Nous pouvons nous poser la question de la possibilité de proposer un schéma de métadonnées universelles. Ce fait éviterait en partie certains écueils, tout en sachant que chaque revue est en mesure de déterminer ses propres normes.

Comme nous avons constitué un corpus de citations, nous pourrions pousser l'analyse syntaxico-sémantique plus loin, afin de pallier les problèmes d'ambiguïté, générateurs d'erreurs. Cela nous conduirait à analyser plus en détail les liens de dépendance, dans les termes d'une phrase, et ce qui détermine la nature de chacun de ses éléments.

Cette analyse constituerait une avancée vers la création d'un tokeniseur, puis d'un parseur plus précis et plus robuste pour notre projet. Les appels de références, noté « [x] », pourraient être pris en compte comme de véritables jetons. Il deviendrait possible de corriger les erreurs du modèle de Stanford dans la création de son tableau.

Il serait également intéressant de créer une interface homme-machine pour faciliter l'utilisation du module. De cette manière, l'utilisateur ne serait pas obligé de passer par l'invite de commande sous Unix et pourrait choisir les options proposées plus aisément, selon le type d'analyse désirée, ou les informations à implémenter ou à modifier.

De plus, les différents éléments composant le module et concernant directement l'utilisateur seraient plus facilement accessibles, notamment en ce qui concerne la consultation des dictionnaires, grammaires, exemples et documentations.

Travailler sur la conception de cet outil m'a beaucoup apporté autant en terme de nouvelles connaissances qu'en terme d'application de mes savoirs déjà existants. J'ai pu vérifier par moi-même la mise en pratique des notions théoriques reçues en formation. Mes capacités d'analyses linguistiques et de conception d'algorithmes ont été développées grâce à ce stage, particulièrement enrichissant. J'ai ainsi pu progresser de façon notable en programmation python et prendre confiance en moi pour envisager de créer, d’avancer sur d'autres projets et/ou avec d'autres langages de programmation.

Bibliographie

Bertignac, C. (2010). La bibliométrie. Consulté le 20/08/2017 à l'adresse http://guides-formadoct.ueb.eu/bibliometrie

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python* (ed. O'Reilly). Julie Steele.

Deboin, M., Fovet-Rabot, C., & Lambert, M. (2017). Le facteur d'impact et ses indicateurs associés pour évaluer la notoriété d'une revue. Consulté le 20/08/2017 à l'adresse https://coop-ist.cirad.fr/aide-a-la-publication/evaluer-les-publications/revue/le-facteur-d-impact-et-ses-indicateurs-associes/1-familiarisez-vous-avec-le-facteur-d-impact-fi-ou-impact-factor-if

Ducrot, O., & Schaeffer, J.-M. (1995). Situation de discours. In *Le nouveau Dictionnaire encyclopédique des sciences du langage* (éd. Point Seuil, p. 764-775).

Gallet, M. (2011). La bibliométrie [Site institutionnel]. Consulté le 20/08/2017 à l'adresse http://www.bib.uvsq.fr/bibliom%C3%A9trie

Gringas, Y., & Caraco, B. (2014). Les Dérives de l'évaluation de la recherche. Du bon usage de la bibliométrie. *JSTOR*, *69*(2), 296-300. Consulté le 20/08/2017 à l'adresse http://www.jstor.org/stable/43855794

Kleiber, G. (1989). Marqueurs référentiels et processus interprétatifs: pour une approche « plus sémantique » (Vol. 11). Présenté à Marquage linguistique inférence et interprétation dans le discours, Genève: Cahier de linguistique française.

Présentation - Lab-STICC. (2016, 2017). Consulté le 20/08/2017 à l'adresse http://www.lab-sticc.fr/fr/francais/

Structure et construction d'un fichier PDF. (s. d.). [FORUM]. Consulté le 20/08/2017 à l'adresse http://forums.mediabox.fr/wiki/tutoriaux/pao/construction_fichier_pdf

Tesnière, L. (1988). *Éléments de syntaxe structurale* (éd. Klincksieck). Paris.

The Stanford Natural Language Processing Group. (s. d.). Consulté le 20/08/2017 à l'adresse https://nlp.stanford.edu/software/lex-parser.shtml

TLF. Consulté le 29 août 2017, à l'adresse http://atilf.atilf.fr/dendien/scripts/tlfiv4/showps.exe?p=combi.htm;java=no

Zoonekynd, V. (2001, mai 7). PDF (Portable Document Format). Consulté le 20/08/2017 à l'adresse http://zoonek2.free.fr/UNIX/30_UNIX_2000/PDF.html

Glossaire

Anaphore : Procédé consistant à rappeler un mot ou un groupe de mots précédemment énoncé par un terme grammatical (« TLF »)

Appel de référence : Il est inclus dans la citation. Il s'agit d'un élément généralement placé entre crochets qui permet de retrouver la référence complète dans la bibliographie.

Citation : C'est le fait de rapporter les écrits/parole d'un autre auteur (dictionnaire Larousse http://www.larousse.fr/dictionnaires/francais/citation/16228)

Grammaire de dépendance : C'est une approche de la syntaxe fondée par Lucien Tesnière (Tesnière, 1988) selon une hiérarchie. On crée ainsi des arbres de dépendances qui représentent les liens syntaxiques entre les mots.

Lien de dépendance : cf. grammaire de dépendances

Référence : Elle se situe dans la bibliographie et est composée de l'appel de référence correspondant ainsi que des éléments permettant d'identifier la citation et les informations la concernant.

Relation syntagmatique : C'est la relation grammaticale entre différents éléments d'une phrase.

Pondération : C'est l'attribution d'une valeur numérique représentant une certaine qualité d'un objet.

Tag : C'est une étiquette grammaticale du type « nom », « adjectif ». Nous avons utilisé l'ensemble des tags de Stanford et créé ceux qui nous paraissaient nécessaires pour notre système.

Token : C'est une entité lexicale dans le cadre d'une analyse

Sigles et abréviations utilisés

CNRS : Centre National de la Recherche Scientifique

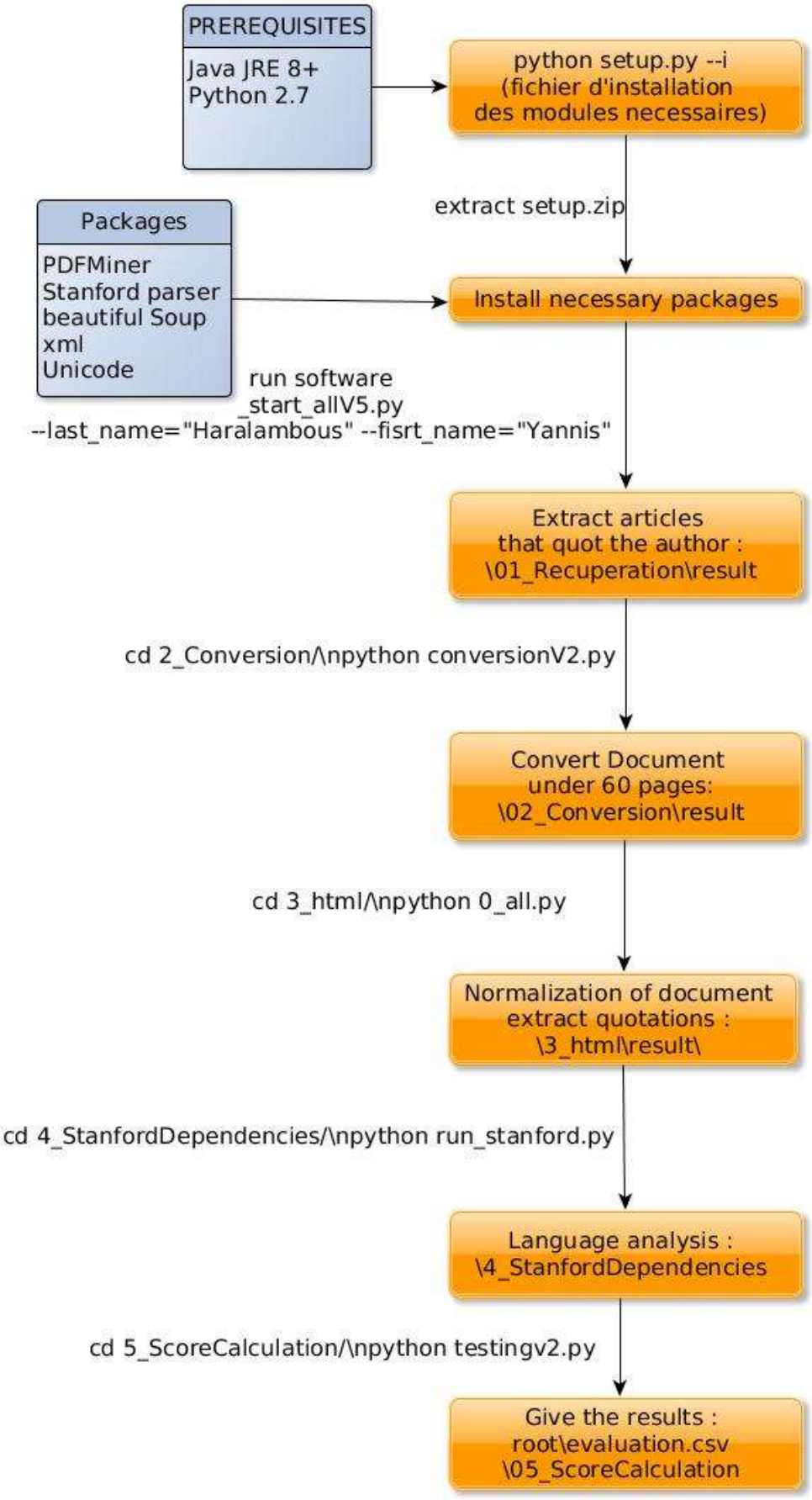
NLTK : Natural Language ToolKit

PDF : Portable Document Format

Table des illustrations

Table des annexes

Annexe 1
Modélisation du programme



Annexe 2
Étapes d'évaluation de la normalisation des documents après la conversion pdf2html

La normalisation des documents s'effectue en 10 étapes sur 101 fichiers HTML. Elle concerne :

- La suppression des tirets de césure, le retour chariot et la balise «
 » les accompagnant. Score de 96.04%
- La suppression des balises «
 » et de leur retour chariot. Score de 95.05%
- La mise en place de chaque paragraphe sur une seule ligne et encadrer par des balises « .. ». Score 94.06
- Suppression des répétitions de balises « .. » vides. Score 97.07%
- Suppression du texte avant le résumé (noms des auteurs du document…). Score 92.57%
- Suppression des balises images qui n'apportent rien à l'analyse. Score 94.06%
- Réorganisation des segments en fonction des indices dans les balises « .. » et récupérer l'ordre initial du PDF. Score 90.59%
- Suppression des balises inutiles « .. ». Score 94.06%
- Suppression de l'ISBN en bas de page. Score 94.06%

- Nous atteignons un score moyen de 93.66%.

La segmentation des documents s'effectue en une seule étape sur les 90% fichiers exploitables :

- Le résumé : 92.08%
- Le corps du texte : 92.08%
- Les références : 93.56%

L'extraction des citations a permis d'extraire 2877 citations. Sur le total nous avons eu 86.50% de correspondance avec les références.

Annexe 3
Évaluations du paramétrage de la pondération et des liens de dépendances

id phrase	nb token	étiquettes attentues	étiquettes erronnée	taux de réussite	taux d'erreur	lien dép. attendues	lien dép. erronées	taux de réussite	taux d'erreur	dép. attendue	dep. erronée	taux de réussite	taux d'erreur	pondération attendue	pondération erronée	taux de réussite	taux d'erreur
1	41	37	4	90,24 %	9,76 %	28	13	68,29 %	31,71 %	27	14	65,85 %	34,15 %	25	16	60,98 %	39,02 %
2	34	31	3	91,18 %	8,82 %	26	8	76,47 %	23,53 %	22	12	64,71 %	35,29 %	31	3	91,18 %	8,82 %
3	10	10	0	100,00 %	0,00 %	7	3	70,00 %	30,00 %	7	3	70,00 %	30,00 %	10	0	100,00 %	0,00 %
4	9	8	1	88,89 %	11,11 %	7	2	77,78 %	22,22 %	7	2	77,78 %	22,22 %	8	1	88,89 %	11,11 %
5	26	25	1	96,15 %	3,85 %	18	8	69,23 %	30,77 %	17	9	65,38 %	34,62 %	23	3	88,46 %	11,54 %
6	12	10	2	83,33 %	16,67 %	9	3	75,00 %	25,00 %	11	1	91,67 %	8,33 %	11	1	91,67 %	8,33 %
7	35	33	2	94,29 %	5,71 %	28	7	80,00 %	20,00 %	28	7	80,00 %	20,00 %	34	1	97,14 %	2,86 %
8	23	19	4	82,61 %	17,39 %	18	5	78,26 %	21,74 %	12	11	52,17 %	47,83 %	18	5	78,26 %	21,74 %
9	30	30	0	100,00 %	0,00 %	23	7	76,67 %	23,33 %	21	9	70,00 %	30,00 %	28	2	93,33 %	6,67 %
10	14	12	2	85,71 %	14,29 %	6	8	42,86 %	57,14 %	6	8	42,86 %	57,14 %	13	1	92,86 %	7,14 %
11	36	34	2	94,44 %	5,56 %	25	11	69,44 %	30,56 %	27	9	75,00 %	25,00 %	34	2	94,44 %	5,56 %
12	27	26	1	96,30 %	3,70 %	14	13	51,85 %	48,15 %	17	10	62,96 %	37,04 %	22	5	81,48 %	18,52 %
total	297	275	22	92,59 %	7,41 %	209	88	70,37 %	29,63 %	202	95	68,01 %	31,99 %	257	40	86,53 %	13,47 %
moyenne	24,75	22,92	1,83	92,59 %	7,41 %	17,42	7,33	70,37 %	29,63 %	16,83	7,92	68,01 %	31,99 %	21,42	3,33	86,53 %	13,47 %

Table des matières

MOTS-CLÉS : contexte linguistique, extraction de citations, parseur, références bibliographiques, bibliométrie

RÉSUMÉ

Pour évaluer leur travail, les chercheurs doivent prouver que leurs travaux ont un réel impact sur la communauté scientifique. La discipline dédiée se nomme bibliométrie. Cependant, la plupart des outils bibliométriques ne comptent que le nombre de citations d'un article selon différents critères. Il ne génère pas d'opinion, positive ou négative, engendrée par ces écrits. Nous avons étudié la possibilité de fournir une application adaptée à une analyse qualitative des contenus pour permettre une vision plus précise de l'influence effective des productions scientifiques. Dans ce but, nous nous sommes basés sur des analyses à la fois sémantiques et syntaxiques. Nous avons proposé une pondération de ces éléments afin de pouvoir réaliser une nouvelle forme d'évaluation qui joindrait à la fois l'analyse quantitative et l'analyse qualitative des textes. Puis, nous avons évalué la robustesse de notre programme et envisagé d'autres évolutions et une meilleure optimisation des données.

KEYWORDS : Linguistic context, quotation extraction, parser, bibliographic references, bibliometrics

ABSTRACT

To evaluate their activities, researchers have to prove that their works have a real impact on the scientific community. The dedicated discipline is called "bibliometrics". However, most bibliometic tools count quotation of an article with different criteria. There do not take into account the opinion, positive or negative, of their readers. We studied the possibility to offer an application providing a qualitative analysis in order to have a better idea of the effective influence of scientific production. We based ourselves on semantic and syntactic analysis. We proposed a weighting of this elements as to obtain an evaluation model, combining quantitative and qualitative analysis of texts. Finally, we have evaluated robustness of our program and considered other developments and better data optimization.