Découverte automatique des textes littéraires qui présentent les caractéristiques statistiques d'un texte de qualité

Hamza Maaouia

▶ **To cite this version:**

Hamza Maaouia. Découverte automatique des textes littéraires qui présentent les caractéristiques statistiques d'un texte de qualité. Sciences de l'Homme et Société. 2014. dumas-01066867

HAL Id: dumas-01066867

https://dumas.ccsd.cnrs.fr/dumas-01066867

Submitted on 23 Sep 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Découverte automatique des textes littéraires qui présentent les caractéristiques statistiques d'un texte de qualité

Nom : Maaouia
Prenom : Hamza

UFR Langage, lettres et arts du spectacle, information et communication

---

Mémoire de master 2 recherche - 30 **crédits** - **Mention** Sciences du Langage
Spécialité : Industrie de la langue
Parcours : Traitement Automatique du Langage Naturel
Sous la direction de Eric Gaussier, Quentin Pleplé, Gilles Bisson, Garbay Catherine et Claude Ponton

Année universitaire 2013-2014

Découverte automatique des textes littéraires qui présentent les caractéristiques statistiques d'un texte de qualité

Nom : Maaouia
Prenom : Hamza

UFR Langage, lettres et arts du spectacle, information et communication

---

Mémoire de master 2 recherche - 30 **crédits** - **Mention** Sciences du Langage
Spécialité : Industrie de la langue
Parcours : Traitement Automatique du Langage Naturel
Sous la direction de Eric Gaussier, Quentin Pleplé, Gilles Bisson, Garbay Catherine et Claude Ponton

Année universitaire 2013-2014

Remerciements

C'est une habitude saine que de remercier au début d'un tel travail tous ceux qui, plus ou moins directement, ont contribué à le réaliser. C'est avec mon enthousiasme le plus vif et le plus sincère que je voudrais rendre mérite à tous ceux qui à leur manière m'ont aidé à mener à bien ce travail.

Je tiens d'abord à témoigner de ma plus profonde gratitude à mes encadreurs de recherche, monsieur Eric Gaussier, monsieur Quentin Pleplé, monsieur Gilles Bisson, madame Garbay Catherine et monsieur Claude Ponton . Ils ont su, par leur extrême dévouement, leur disponibilité et leur gentillesse débordante rendre mon travail fort agréable, voire même amusant. De conseils judicieux en mots d'encouragements, ils étaient toujours d'une aide précieuse et je leurs en suis très reconnaissant.

Mes remerciements s'adressent également à tous mes enseignants à l'université Stendhal Grenoble 3, pour ces deux années d'apprentissage.

Je remercie tous les membres de Short Édition qui m'ont aidé surtout Eudes F. Hasselbaink, Manon Landeau et Kauffeisen Pascale à m'intégrer dans l'équipe et à me donner ses précieux conseils. Ainsi que tous les autres membres, j'ai passé des bons moments avec eux.

J'exprime ma gratitude à mon infermière et à tous mes amis qui m'ont toujours encouragé et supporté tout au long de la réalisation de ce mémoire, merci infiniment à vous.

Finalement, une attention toute particulière est dirigée vers ma famille, et en particulier mon père, ma mère, ma sœur, et surtout ma fiancée Manel, qui n'ont pas négligé les sacrifices tout au long des mes études. Merci infiniment d'avoir toujours été si attentionnés et dévoués.

UNIVERSITÉ Stendhal GRENOBLE 3

Déclaration anti-plagiat
Document **à scanner** après signature
et **à intégrer** au mémoire électronique

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 : Maaouia **Prenom :** Hamza **Date :** 10/07/2014

Signature :

Contents

1 État de l'art **9**
- 1.1 Introduction . . . 9
- 1.2 Problématique . . . 10
- 1.3 Critère d'évaluation . . . 11
  - 1.3.1 Quantitatif . . . 11
  - 1.3.2 Richesse . . . 11
  - 1.3.3 Cohésion-Cohérence . . . 12
  - 1.3.4 Lisibilité . . . 12
- 1.4 Méthodes . . . 14
  - 1.4.1 Méthodes fondées sur les automates finis ou les expressions régulières 14
  - 1.4.2 Méthodes basées sur des règles . . . 16
  - 1.4.3 Méthodes stochastiques . . . 17
  - 1.4.4 Analyse sémantique latente (LSA) . . . 18
- 1.5 Outils de traitement automatique des langues . . . 22
  - 1.5.1 TreeTagger . . . 22
  - 1.5.2 NLTK . . . 22
  - 1.5.3 Cordial Analyseur . . . 23
  - 1.5.4 Langage Tool . . . 24
- 1.6 Logiciels . . . 25
  - 1.6.1 Essay Grade . . . 25
  - 1.6.2 SATO . . . 25
  - 1.6.3 Intelligent Essay Assessor . . . 25
  - 1.6.4 E-Rater . . . 26
- 1.7 Piste de recherche et l'hypothèse . . . 26

2 Méthodologie **28**
- 2.1 Introduction . . . 28
- 2.2 Démarche . . . 30
- 2.3 Pré-traitement . . . 30
  - 2.3.1 Les erreurs de prétraitement . . . 31
  - 2.3.2 Nos solutions . . . 31

3 Représentation des indicateurs textuels **33**
- 3.1 Bag of words . . . 33
- 3.2 Indicateurs généraux . . . 37
  - 3.2.1 Longueur des mots . . . 37
  - 3.2.2 Longueur des phrases . . . 37
  - 3.2.3 Longueur des paragraphes . . . 38

3.3 Indicateurs lexicaux . . . 39
3.3.1 Richesse Vocabulaire . . . 39
3.3.2 Densité lexicale . . . 40
3.3.3 Entropie . . . 41
3.4 Indicateurs grammaticaux . . . 42
3.4.1 Fréquence relative des adjectifs . . . 42
3.4.2 Fréquence relative des noms . . . 42
3.4.3 Fréquence relative des pronoms . . . 43
3.4.4 Fréquence relative des verbes auxiliaires . . . 43
3.4.5 Fréquence relative des prépositions . . . 43
3.4.6 Fréquence relative des déterminants . . . 43
3.4.7 Fréquence relative des verbes . . . 44
3.4.8 Fréquence relative des conjonctions . . . 44
3.4.9 Distribution des adverbes . . . 44
3.5 Indicateurs de complexité . . . 45
3.5.1 Nombre moyen de phrases complexes . . . 46
3.5.2 Cohésion du texte . . . 46
3.5.3 Lisibilité . . . 46
3.6 Fautes d'écriture . . . 48

4 Évaluation **50**
4.1 Qualité ? . . . 50
4.2 Évaluation visuelle des indicateurs . . . 50
4.2.1 Indicateurs généraux . . . 51
4.2.2 Indicateurs lexicaux . . . 52
4.2.3 Indicateurs grammaticaux . . . 54
4.2.4 Indicateurs de complexité . . . 58
4.2.5 Fautes d'écriture . . . 58
4.3 Évaluation de la pertinence statistique des indicateur (Corrélation) . . . 59

5 Matrice de classification **61**
5.1 Représentations textuelle . . . 62
5.1.1 Représentation binaire . . . 62
5.1.2 Représentation fréquentielle . . . 63
5.1.3 Représentation tf-idf . . . 64
5.2 Construire la matrice . . . 65

6 Classification **67**
6.1 Apprentissage . . . 67
6.1.1 Apprentissage supervisé . . . 67
6.1.2 Apprentissage non supervisé . . . 68
6.2 Algorithme de classification . . . 68
6.2.1 Support Vector Machine (SVM) . . . 68
6.2.2 Descente de gradient stochastique (SGD) . . . 69

6.2.3 Arbres de Décision 70
6.2.4 Random Forest 71
6.3 Résultats de classification 71
6.3.1 Scikit Learn 71
6.3.2 Orange Canvas Learn 72
6.3.3 Indices d'évaluation 72
6.3.4 Résultat 73

7 Conclusion **77**

8 Annexe **82**

Résumé

Le domaine du traitement automatique des langues naturelles a connu des évolutions très rapides ces dernières années, et spécialement les méthodes de statistique textuelle. Elles ont été mises en lumière par plusieurs disciplines : l'étude des textes, la linguistique, l'analyse du discours, la statistique, l'informatique, le traitement des enquêtes.

Ce projet de recherche s'inscrit dans le cadre du problématique de Short Édition qui concerne l'éditeur communautaire de littérature courte. L'objectif est d'assister le travail du comité de lecture en effectuant une première catégorisation des textes. Notre travail implique la conception et la mise en œuvre d'un prototype permettant de repérer les textes qui présentent les caractéristiques d'un texte de qualité et de trouver une méthode de classification en nous fondant sur les principes de la fouille de données permettant de bien classer nos textes.

MOTS-CLÉS : Traitement Automatique de la Langue - Évaluation automatique de textes - Classification - Apprentissage.

Abstract

The field of natural language processing has witnessed very rapid developments in recent years, particularly with respect to methods used for statistical text analysis. These methods have been brought into focus by several disciplines in particular : the study of texts, linguistics, discours analysis, statistics, computer sciences, and survey processing.

This research project develops within the framework of an issue that concerns the publishing company Short Editions. It relies on contributions in a field that employs a vast variety of designations (lexical statistics, statistical linguistics, quantitative linguistics, etcetera). Our work involves the creation of a prototype that allows for the identification of texts that present the characteristics of a quality text and to find appropriate methods of classification of these texts based on data and text mining principles.

KEYWORDS : Automatic Language Processing - Automatic evaluation of texts - Classification - Learning.

1 État de l'art

1.1 Introduction

Une histoire est un récit court, elle est plus courte que le romane, elle est centrée sur un seul événement. Les personnages sont peu nombreux et sont moins développés que dans le roman.

L'attribution de facteurs de qualité à des histoires courtes a toujours été un centre d'intérêt pour les éditeur. Plusieurs techniques ont été développées dans ce sens en se basant sur l'évaluation automatique de texte et que nous prendrons comme point de départ lors de notre recherche. Le lien linguistique entre ces méthodes est par ailleurs important à noter. Pour élaborer une histoire courte qui soit efficace, il faut apprendre à poser la forme qui servira de squelette de base en utilisant le scénario, le thème, les personnages, le dénouement et la conclusion.

D'un point de vue général, la rédaction d'un texte est basée sur deux composantes essentielles telles que le contenu et l'apparence. Les mots utilisés dans le texte doivent apparaître d'une manière compréhensible afin d'atteindre l'intérêt du lecteur. L'apparence est un indice sur l'attention du lecteur, la manière de placer les composants d'un texte, tels que les mots, les phrases et les paragraphes.

Dans les années soixante, ce domaine a commencé à voir la lumière, il est né avec E.B. Page. Cela a préparé le terrain à l'émergence de plusieurs logiciels tels que E-rater qui peut détecter automatiquement les erreurs d'orthographe ou de grammaire, la richesse lexicale et la longueur de l'essai(Chodorow et Leacock, 2004 Burstein, 2009). Aujourd'hui, les États-Unis sont l'un des principaux pays dans le développement des logiciels pour l'évaluation automatique des textes. Nous constatons un nombre considérable de logiciels d'aide à évaluer le travail d'étudiants, précisément, ils permettent d'évaluer l'écriture des étudiants et donner à la suite des notes basées sur des critères bien définis selon les objectifs des matières évaluées.

Notre projet traite une problématique de Short-Édition[1] qui concerne l'éditeur communautaire de littérature courte. C'est une maison d'édition de nouvelle génération qui essaie de concevoir un modèle qui soit à mi-chemin entre le numérique et le papier. C'est dans ce but qu'elle anime avec passion la communauté des lecteurs intéressés dans la littérature courte, les nouvelles, BD courtes, poèmes et très très courts (nouvelles de moins de 4 min de lecture). L'enjeu de la maison d'édition est de savoir repérer les œuvres de qualité pour les mettre en avant, qu'elles soient plus lues, voire publiées en livre papier par Short Édition. Depuis 3 ans, ce repérage repose sur un comité éditorial composé d'une trentaine de personnes qui lisent et évaluent quasiment toutes les œuvres entrantes.

L'objectif de notre projet est de créer un algorithme qui va apprendre à partir des évaluations humaines des membres du comité, une fonction permettant de repérer les textes qui présentent les caractéristiques d'un texte de qualité. Ces textes seront, par la suite, mis en avant sur Short Édition et l'activité de la communauté sera mesurée et interprétée pour confirmer ou infirmer la prédiction de l'algorithme.

Notre travail sera organisé comme suit : Nous commençons par une présentation de la problématique du sujet et qui sera suivie par trois sections. Dans la première, nous décrirons les différents niveaux d'analyse utilisés pour évaluer les textes. Dans la deuxième, nous présentons les méthodes mises en œuvre dans ce domaine. Enfin dans la troisième section nous complétons par les logiciels existants.

1.2 Problématique

Comme nous l'avons mentionné ci-dessus, l'objet d'étude principal de ce mémoire est l'évaluation automatique de la qualité des histoires courtes. La recherche dans ce domaine est toujours d'actualité car les résultats obtenus aujourd'hui sont encore sujet d'amélioration. Il y a potentiellement de nombreux facteurs qui agissent sur la qualité des textes tels que la distribution des catégories de mots, des sentiments et des connotations. Au premier abord, le problème que nous traitons est composé de deux parties.

- D'un côté, nous sommes en présence d'une grande base de données comprenant des textes de types différents (les nouvelles, BD courtes, poèmes et textes très courts).
- De l'autre côté, nous sommes en présence de méthodes et indicateurs qu'il faut sélectionner et appliquer à notre corpus. L'objectif est donc de concevoir une application informatique capable de percevoir la qualité des textes de Short-édition.

Bien que la définition soit relativement simple, la mise en place d'une solution est loin d'être immédiate vu le nombre de facteurs à prendre en considération. La figure ci-dessous donne le cadre général de nos tâches et nous amène à nous poser des questions telles que : quelles sont les méthodes, les indicateurs et les outils efficaces qui pourraient nous donner une évaluation de qualité ? Selon quels facteurs et critères ?

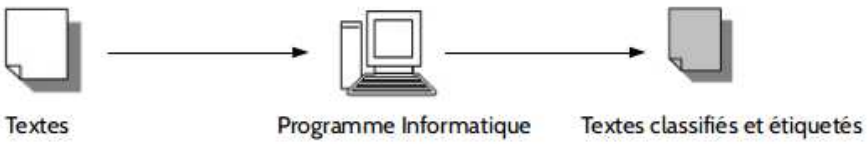


Figure 1: Cadre général du projet

1.3 Critère d'évaluation

Notre objectif s'inscrit dans le cadre du traitement automatique de la langue (TAL). Dans un premier temps nous définissons un ensemble de critères d'analyse selon les différents niveaux lexical, syntaxique, sémantique, pragmatique ou structurel comme indiqué dans la figure ci-dessous, afin de les employer dans notre algorithme.

1.3.1 Quantitatif

Dans cette section, nous allons décrire une suite de critères permettant de déterminer le poids (longueur) d'un histoire. Ce type de critère basé sur l'analyse lexicale permet de transformer une suite de caractères en une suite de mots, de phrases, ou même de paragraphes (appelée tokens en anglais). La première caractéristique d'une histoire courte est sa taille, généralement, elle varie de 500 mots à 5 000 mots. Il nous semblait nécessaire de qualifier les histoires en fonction de leur longueur. Nous pouvons la déterminer à différents niveaux : nombre de caractères, de mots, de phrases, etc.

Parmi les premières approches dans ce domaine, nous trouvons celle de (Serraf Guy, 1964). Il a fait une étude comparative portant sur deux échantillons de textes de la langue française, il a distingué les textes faciles et difficiles. Pour le texte facile, il a sélectionné tous les textes de très grande diffusion destiné à un public dont le niveau d'instruction est le certificat d'études primaires : quotidien populaire, hebdomadaire de grande information, texte publicitaire, roman d'auteur connu et jouissant de tirages importants, etc. Et pour les textes difficiles, il a retenu les textes destinés à un public restreint et traitant des problèmes d'une discipline particulière, impliquant une écriture (à l'exception des chiffres et symboles) et un vocabulaire spécialisés. Pour distinguer les textes, il s'est appuyé sur des critères quantitatifs tels que la longueur de mots, de phrase, nombre des syllabes, etc.

1.3.2 Richesse

Une histoire courte, est une succession de caractères limités et organisés dans le but de transmettre un contenu et/ou des émotions. C'est un ensemble d'idées organisé sous la distribution des catégories grammaticales : nom, déterminant, verbes, etc. Parmi les caractéristiques des histoires courtes, nous trouvons égalment la distribution des personnages. La plupart des histoires sont écrites à la première, à la deuxième ou à la troisième personne. À partir de ce contexte, pour déterminer la richesse du histoire, nous nous pouvons nous intéresser d'un coté à la proportion d'adjectifs, de verbes, d'adverbes, de phrases coordonnées, etc. et, de l'autre coté à la distribution des idées.

La distribution des catégories grammaticale peut être un critère d'évaluation des histoires. Prés de cette approche, nous trouvons le travail de (Thierry Trubert-Ouvra, en 2002), qui a utilisé un corpus de grande dimension (L'Encyclopædia Universalis com-

prend environ vingt-huit millions de mots) pour déterminer la place des adjectifs en position épithète dans le groupe nominal.

Ainsi, nous pouvons considérer la distribution des idées dans une histoire parmi les critères utilisés pour évaluer la qualité d'écriture. La densité des idées, qui correspond au ratio entre le nombre de propositions sémantiques et le nombre de mots dans une histoire reflète la qualité informative des propositions langagières. Ici, nous trouvons l'approche de (Hyeran Lee, Philippe Gambette, Elsa Maillé, Constance Thuillier, en 2010). Ils ont proposé une méthode basée sur un étiquetage morphosyntaxique et des règles d'ajustement, inspirée du logiciel CPIDR, dans le but de calculer automatiquement la densité des idées dans un corpus. Ils ont utilisé un corpus de quarante entretiens oraux transcrits.

1.3.3 Cohésion-Cohérence

D'un point de vue sémantique, la cohérence peut être un indice d'évaluation de la qualité d'écriture. Il est important de tenir compte de la continuité du histoire qui est exprimée par sa structure. C'est un critère difficile à définir, nous avons tous lu dans des corrections, des remarques telles que "manque de cohérence", il se manifeste au niveau global du texte (champ lexical, progression des idées, relation entre passages, etc). Plus l'histoire cohérente plus elle est compréhensible et plus elle est facile à lire.

La cohérence s'inscrit dans la signification générale d'histoire, mais pour la cohésion, elle est un peu plus spécifique. Elle s'intéresse particulièrement aux relations locales du texte telles que les règles morphologiques et syntaxiques, les connecteurs argumentatifs et les organisateurs, etc. Parfois, nous pouvons tomber dans des cas où un texte peut être cohérent, mais sans cohésion : erreurs de temps, de grammaire, etc. De même qu'un texte peut avoir de la cohésion, mais sans être cohérent : erreurs sémantique (par exemple, des mots ne fonctionnant pas ensemble).

Dans cette approche, nous trouvons l'étude de (Yves Bestgen, 2012), il s'est appuyé sur le calcul de cohésion lexicale pour évaluer un corpus de 223 textes à travers la méthode d'analyse sémantique latente (section 1.4.4). La cohésion considérée parmi les critères fréquemment utilisés dans l'évaluation de qualité littéraire.

1.3.4 Lisibilité

Mesure de lisibilité est une analyse statistique, généralement réalisée en comptant les syllabes, les mots, les phrases et par la comparaison de la fréquence des mots par rapport à d'autres textes. Selon (Robert Gunning, 1968) la lisibilité repose essentiellement sur la longueur des phrases et des mots. Et selon (Flesch, 1946) la lisibilité repose sur le nombre de phrases et de syllabes. Nous aborderons dans cette section les deux formules les plus connues dans ce domaine "Flesh" et "Gunning".

Flesch : Le plus connu est l'indice de lisibilité de Flesch(1948) intégrant le nombre de mots dans les phrases et de syllabes dans les mots. Il est basé sur la double hypothèse suivante : statistiquement, plus une phrase est longue, plus elle est complexe et plus un mot est long, plus il est rare. Les résultats varient entre 0 et 100. La moyenne se situe entre 60 et 70. Plus le nombre est élevé, plus le texte est lisible.

$$Flesch(T) = 206.84 - 0.85W - 1.02S$$

où le nombre 206.84 est un constante ; W : le nombre moyen de syllabes par 100 mots ; S : longueur moyenne des phrases par mots.

Afin d'appliquer cette équation, il faut faire la correspondance du résultat obtenu via le tableau de référence suivant pour avoir le pourcentage de lisibilité.

–	Flesch(T)	Pourcentage correspondant %
1	90 à 100	90
2	80 à 90	86
3	70 à 80	80
4	60 à 70	75
5	50 à 60	40
6	30 à 50	24
7	0 à 30	4.5

Table 1: Table de référence de Flesch

Avant de passer à la deuxième approche, nous présentons ici un exemple simple pour éclairer le mécanisme de cette équation. Supposons un petit texte T contient 95 mots(compter 120 syllabes) et qui comprend par exemple 7 phrases,donc :

$$Flesch(T) = 206.84 - 0.85(120) - 1.02(\frac{95}{7}) = 91.27$$

D'après le tableau 1 de référence 91.27 équivalant à 90 % taux de lisibilité.

R. Gunning : D'autre part, R. Gunning, l'un des premiers consultants en lisibilité, ne retient que le nombre de mots moyen par phrase et le pourcentage de mots de plus de 3 syllabes. On additionne les deux chiffres et on multiplie par $0,4$. Le résultat est ce que Gunning appelle "Fog index", donc :

$$Fogx(T) = 0.4(LM(T) + LS(T))$$

où, LM : nombre moyen de mots par phrase du texte T ; LS : nombre moyen de mots de plus 3 syllabes du texte T.

La multiplication par $0,4$ permet d'avoir un indice correspond au niveau nécessaire d'années de scolarité pour lire et comprendre le texte comme il est indiqué dans le tableau 8. Un indice élevé correspondra donc à un texte difficile pour la compréhension.

8-9	Littérature junior et ado
10-11	Libération
14-15	Le Monde Diplomatique, L'express
16-17	Rapports parlementaires
17-18	Article universitaire
22 et plus	Directives européennes

Table 2: Gunning Fog Type d'écrit

1.4 Méthodes

Dans cette section, nous mettrons en avant un ensemble de méthodes utilisées dans le domaine TAL. En général, nous pouvons les classer en trois types, les méthodes stochastiques, les méthodes basées sur des règles et d'autres fondées sur des automates finis ou des expressions régulières. Tout d'abord, nous nous intéresserons au trois types de méthodes. Ensuite nous présenterons une méthode du nom d'analyse sémantique latente (LSA) basée sur la représentation vectorielle.

1.4.1 Méthodes fondées sur les automates finis ou les expressions régulières

Les expressions régulières[2] **:** est considéré comme une méthode qui permet d'effectuer des recherches de segments dans les textes comme les phrases, les mots, les entités nommées, un mot particulier, les formes fléchies d'un lemme, etc. Autrement dit c'est une formule permettant de caractériser un ensemble de chaînes de caractères. Une expression régulière est inscrit dans un ensemble d'alphabets de symboles Σ et de mots vides ε. Dans notre approche, nous nous appuyons sur le langage de programmation PYTHON pour développer notre prototype. C'est dans ce contexte que nous présentons brièvement dans le tableau ci-dessous un ensemble de symboles permettant de gérer des opérations complexes sur les expressions régulières.

Symboles	Descriptions
.	Remplace n'importe quel symbole sauf le retour chariot $\backslash n$
[ ]	Remplace l'un quelconque des symboles placés entre les crochets
[$\wedge$]	Remplace l'un quelconque des symboles qui ne sont pas entre les crochets
$\backslash w$	Remplace n'importe quel caractère alphanumérique (lettre ou chiffre) plus le caractère
$\backslash W$	Remplace n'importe quel symbole qui n'est ni un caractère alphanumérique, ni le caractère
$\backslash d$	Remplace n'importe quel chiffre
$\backslash D$	Remplace n'importe quel symbole qui n'est pas un chiffre
$\backslash s$	Remplace l'un quelconque des caractères d'espacement, $\backslash t$, $\backslash n$, $\backslash f$, $\backslash r$, $\backslash v$
$\backslash S$	Remplace n'importe quel symbole qui n'est pas un caractère d'espace

Table 3: Opérations complexes sur les expressions régulières.

Les expressions régulières jouent un rôle important, cela est dû au nombre de fonctions, qu'elles peuvent réaliser. Elles permettent de segmenter les textes en détectant les séparateurs de phrases et de mots. Les expressions régulières sont utilisées dans l'analyse et l'étiquetage morphosyntaxiques.

Exemple : Trouver tous les adverbes en "-ment" dans la phrase P.
$P =$ *"Il s'était prudemment déguisé mais fut rapidement capturé par la police."*
Expression : `re.findall(r"\w+ment", PH)`
Résultat : ['prudemment', 'rapidement']

Automates finis : Ils représentent des méthodes fondamentales en mathématiques discrètes et en informatique. Ils permettent de gérer plusieurs tâches dans l'étude des langages formels et en compilation : les protocoles de communication, modélisation de processus, la théorie de la calculabilité, etc.

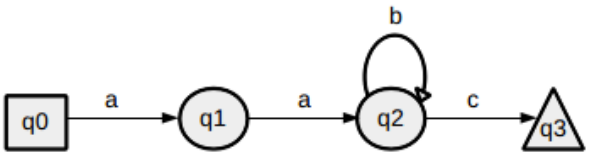
Les automates finis servent à caractériser des langages composées de mots acceptés. Nous trouvons plusieurs extensions d'automates finis, parmi les plus connus les automates de Büchi, les automates de Rabin et les automates de Muller. Dans le traitement automatique de la langue, ces méthodes utilisées dans le but de recherche des motifs dans le texte. Autrement dit, il vise à répondre à la question est-ce que le mot W appartient au langage L ? Il existe deux types d'automates d'états finis : déterministe et non-déterministe. Un automate à état fini est un quintuplé $A = (S, E, T, S_0, F)$ avec :

S : un ensemble fini d'états	$S_0 \subseteq S$: l'ensemble des états initiaux
E : un alphabet	$F \subseteq S$:l'ensemble des états terminaux
T : fonction de transition $\texttt{T : S} \times \texttt{E} \rightarrow \texttt{S}$	

Il est possible de représenter un automate d'état fini sous forme de symboles indiquant des transitions. Ce type de représentation est appelé réseau de transitions simples. C'est une machine abstraite destinée à reconnaître un langage régulier, c'est-à-dire un langage défini par une expression régulière.

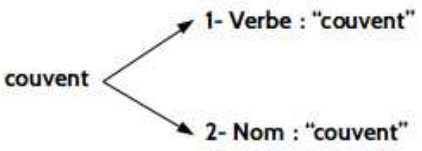
Initial : □ | Intermédiaire : ◯ | Terminal : △

Exemple : Prenons la chaîne (**aab*c**) comme exemple.



1.4.2 Méthodes basées sur des règles

Ces méthodes sont fondées sur un ensemble de règles dans le but d'affecter une catégorie à chaque mots dans un texte, et de résoudre de potentiels problèmes d'ambiguïté. Illustrons par un exemple : le mot "couvent" peut être étiqueté comme un verbe ou un nom, dans le célèbre exemple : "les poules de couvent couvent"



Le but est de délimiter clairement les catégories afin d'éviter toute ambiguïté. Elles sont appelées des règles contextuelles, car elles s'appuient sur les mots précédent ou suivant le mot analysé. La plupart des règles sont établies de manière classique (à la main). La mise au point des règles se fait à travers des tests qui nous permettent de savoir lesquelles parmi les règles choisies répondent à nos besoins. Chomsky, parmi les premiers chercheurs dans ce domaine, a proposé un ensemble de règles servant de base pour de nombreux travaux.

```
Ph  ⇒ (SP) SN SV (SP)     | SN ⇒ (SN) (SP) : complément du nom
SP  ⇒ Prép SN             | SAdv ⇒ (SAdv) Adv
SN ⇒ (Dét) N (SP) (SA)    | SV  ⇒(AUX) V (SN) (SP) (SA) (SAdv)
SA ⇒ (SAdv) A (SP)        | SAdv ⇒(SAdv) Adv
```

Les éléments entre parenthèses () sont facultatifs.

Ces règles[2] seront augmentées par Z. Harris dans le but de faciliter la transaction avec les phrases complexes et les mettre en équivalent avec les phrases simples. Graphiquement nous pouvons transformer une suite de règles d'une phrase dans un arbre (figure 2).

Exemple P : *"Le petit chat est gris"*.

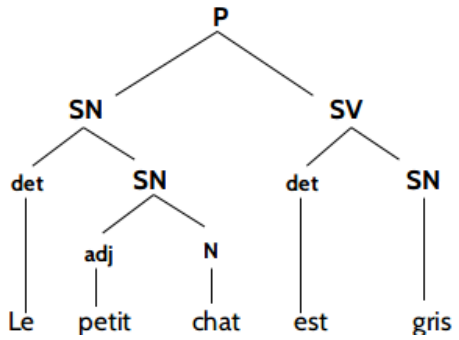


Figure 2: Représentation graphique du P

1.4.3 Méthodes stochastiques

Les méthodes stochastiques[3] font partie de la famille des méthodes probabilistes, sont basées sur des calculs statistiques sur des textes. Nous trouvons parmi ces méthodes une qui était employée pour la correction orthographique par (Kernighan et al. 1990 et Jurafsky et Martin, 2000). Elle repose sur deux hypothèses fortes :

- Les erreurs orthographiques détectées par l'absence des mots d'une liste fermée (lexique, dictionnaire), si un mot n'appartient pas à cette liste, il est considéré comme mal orthographié.
- L'erreur peut peut provenir d'un de ces quatre indices :
  - Insertion d'une lettre parasite;
  - Manquante lettre;
  - Remplacement d'une lettre par une autre;
  - Inversion de deux lettres adjacentes.

Plusieurs approches étudient ce type de méthodes comme le travail de (Naber, en 2003). Il a proposé un correcteur grammatical basé sur un étiquetage probabiliste Qtag crée par (Tufis et Mason, 1998). Ceci a donné le correcteur libre An Gramadóir s'appuyant sur des règles limitant l'ambiguïté, pouvant être soit écrites à la main, soit construites automatiquement par apprentissage à l'aide de l'algorithme de (Brill, en 1995). Nous pouvons aussi citer qui (Carlberger et al, 2002; Knutsson et al., 2002, 2003b,c, 2007),ont proposé le vérificateur grammatical Granska basé sur les Modèles de Markov Cachés (HMM) et sur des règles d'erreurs.

La méthode “n-gram”, fait partie des méthode stochastiques. Les techniques basées sur les n-grammes présentent plusieurs avantages :

- Comparativement à d'autres techniques, les “n-gram” capturent automatiquement les racines des mots les plus fréquents (Grefenstette, 1995).
- Elles opèrent indépendamment des langues (Dunning, 1994), contrairement aux systèmes basés sur les mots dans lesquels il faut utiliser des dictionnaires spécifiques (féminin-masculin, singulier-pluriel, conjugaisons, etc.) pour chaque langue.
- Elles sont tolérantes aux fautes d'orthographes et aux déformations causées lors de l'utilisation des lecteurs optiques. Lorsqu'un document est scanné, la reconnaissance optique est souvent imparfaite.
- Enfin, ces techniques n'ont pas besoin d'éliminer les mots outils, ni de procéder à la lemmatisation. Ces traitements augmentent la performance des systèmes basés sur les mots. Par contre, pour les systèmes “n-gram”, de nombreuses études (Sahami, en 1999) ont montré que la performance ne s'améliore pas après l'élimination des ”Stop Words” et de ”Stemming”.

1.4.4 Analyse sémantique latente (LSA)

LSA est une approche qui fournit une représentation statistique de la connaissance du monde basée sur l'analyse de corpus pour calculer la similarité sémantique entre les mots, les phrases et les paragraphes (Landauer, McNamara, Dennis et Kintsch, 2007). Elle est utilisée dans plusieurs domaines de recherches tels que le TAL, la psychologie, et les sciences de l'éducation (Landauer et al. , 2007).

Elle permet de créer un espace sémantique de petite dimension à partir de l'analyse statistique des occurrences dans un corpus de textes pour estimer la similarité sémantique entre des mots, des phrases, des paragraphes et même des textes. Elle est basée sur un modèle algébrique, et elle est utilisée dans l'indexation et le classement par pertinence. Elle permet donc de découvrir via une approche statistique et algébrique. La sémantique cachée et sous-jacente (latente) de mots dans un corpus de documents.

Ici, nous expliquons les étapes de cette méthode pour calculer la cohésion au niveau des phrases :

- Construction de la matrice des occurrences X
  La première étape consiste donc à construire une matrice des occurrences des termes dans les textes de taille $W \times D$. Chaque ligne de la matrice représente un texte et chaque colonne représente un mot. L'intersection de chaque ligne et de chaque colonne représente le nombre d'occurrence du mot représenté par la colonne, dans le texte représenté par la ligne. Dans cette étape nous ne prenons pas en compte les mots outils, nous construirons le matrice X sur la base de mots pleins.

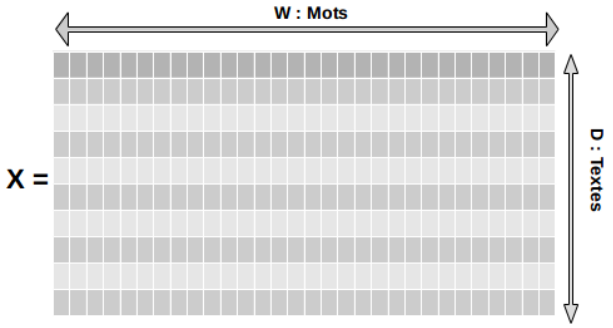


Figure 3: Matrice des occurrences X

- Décomposition de la matrice en valeurs singulières :
  La deuxième étape consiste à déterminer les valeurs singulières de la matrice X afin de le décomposer en trois matrices U, Σ, V^T, dont la multiplication donne la matrice X.

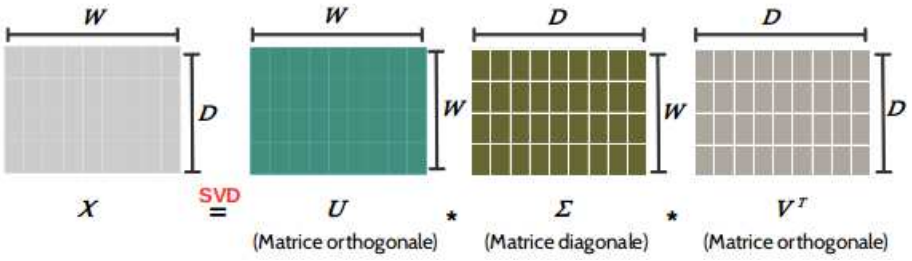


Figure 4: Maquette de décomposition de la matrice par des valeurs singulières

où Σ est une matrice diagonale de taille $W \times D$ de valeurs singulières et U, V^T, deux matrices orthogonales de taille $W \times W$ et $D \times D$.

- Réduction des matrices au rang k :
  Dans le but de réduire la dimension d'espace, nous employons une technique de Data-Minnig. Cette technique nous permet de nous concentrer sur les plus fortes valeurs singulières, autrement dit les informations les plus importantes. La première étape consiste à annuler la diagonale de Σ au-delà d'un certain indice k, et à recalculer la matrice de départ. Le résultat sera des données filtrées, représentant l'information dominante de l'ensemble de départ. Dans notre cas, commençant à réduire la matrice diagonale Σ de taille $D \times W$ à une nouvelle matrice de taille $k \times k$, tel que $K \leq min(W, D)$(cf. figure 6). Rappelons que les valeurs singulières dans Σ sont triées par ordre décroissante.

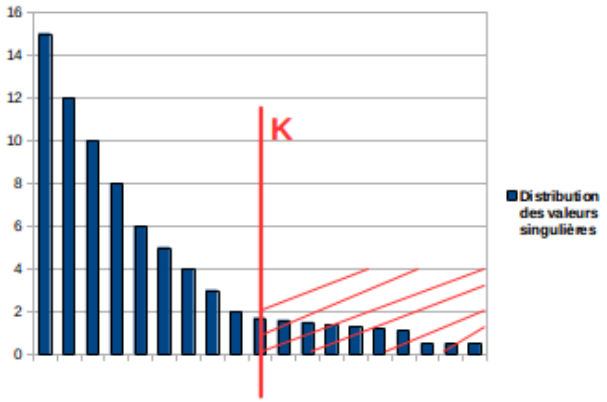


Figure 5: Fixer la valeur de K

La figure précédente (figure 5) devient comme suit avec la nouvelle matrice, qui est une approximation de notre matrice précédente X, donc $X \simeq X_k$:

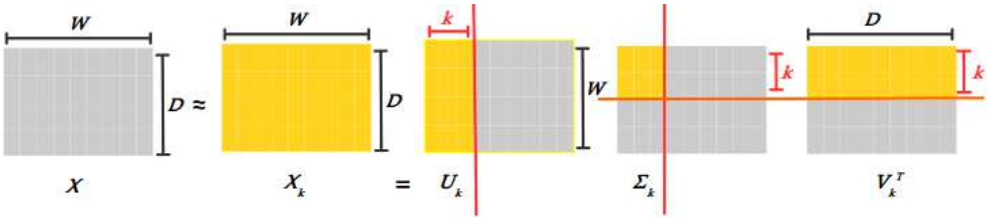


Figure 6: Réduire la matrice diagonale

- Projeter chaque phrase de texte sur l'espace sémantique latente :
  La matrice X_k obtenu est une matrice approchée de la matrice de départ X ne contenant que la sémantique des mots les plus importants d'un point de vue statistique pour le corpus.
  Maintenant, nous passons à l'étape suivante consistant à transformer chaque phrase de chaque texte sous la forme de vecteur P, et projeter chaque vecteur dans l'espace sémantique latent, autrement dit, multiplier chaque vecteur de phrases P par la matrice V_k^T ($P \in \Re^W$). Cet vecteur P contient le nombre d'occurrences de chaque mot de la matrice X dans la phrase p, donc il est de dimension $W \times 1$(cf.figure 8).

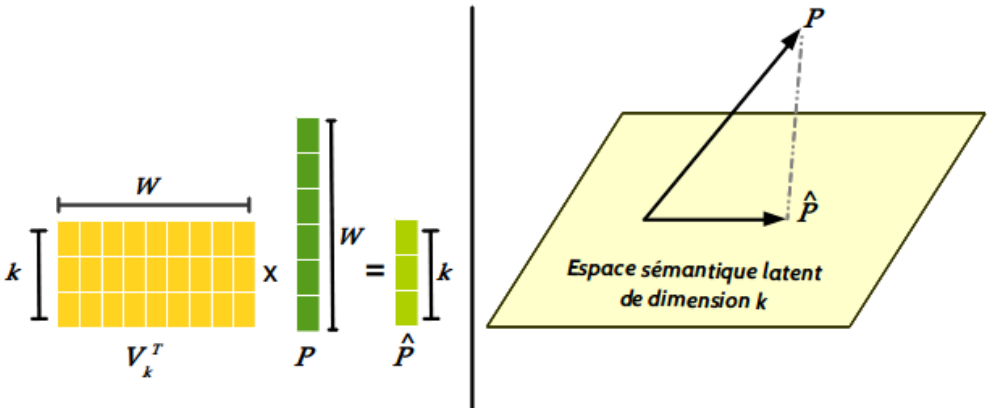


Figure 7: Projection P sur le sous-espace sémantique Latent

- Calculer la distance entre les vecteurs :
  Ensuite, nous passons à estimer l'indice de cohésion entre les phrases adjacentes, pour atteindre cet objective, nous calculons la distance entre les vecteurs de deux phrases adjacentes, $\cos(\hat{P}_i, \hat{P}_{i+1})$, donc pour estimer l'indice moyen de tout le texte (T), nous calculons la moyenne de tous les indices de cohésion de chaque paire phrases adjacentes (cf. figure9), le calcul se fait à la base de l'équation suivante :

$$Indice_Cohesion(T) = \frac{1}{n-1}\sum_{i=1}^{n-1}\cos(\hat{P}_i, \hat{P}_{i+1})$$

où,n ; Nombre de phrases dans texte T, P_i; Phrase d'indice i dans le texte T

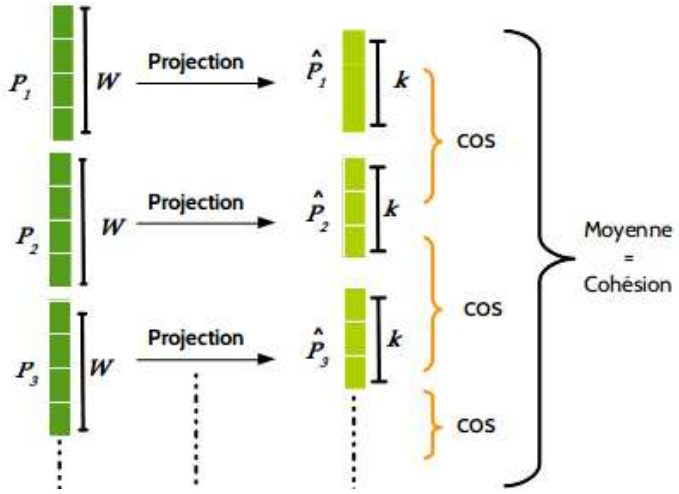


Figure 8: Dernière étape de cohésion

1.5 Outils de traitement automatique des langues

Pour rappel, l'analyse morphosyntaxique faite partie de l'analyse syntaxique. Elle consiste à identifier les mots par ses catégories. En d'autres termes, elle permet d'associer chaque mot à sa catégorie (noms, verbes, proposition, etc). Notons que cette étape est précédée par deux tâches : une qui consiste à segmenter le texte en mots ou phrases et l'autre se concentre sur la lemmatisation qui associe un lemme morphosyntaxique à chaque mot. Dans cette section, nous présentons les quatre types d'outils. Les plus courants.

1.5.1 TreeTagger

TreeTagger[4], est un outil d'annotation de texte proposant des informations sur les parties du discours et des informations de lemmatisation. Il a été développé par (Helmut Schmid, 1994) dans le cadre du projet (TC) dans le ICLUS (Institute for Computational Linguistics of the University of Stuttgart). Il permet l'étiquetage de l'allemand, l'anglais, le français, l'Italien, l'espagnol, le bulgare, Le russe, le grec, le portuguais, le chinois et les textes en Français ancien. Il est adaptable à d'autres langues si des lexiques et des corpus étiquetés manuellement sont disponibles.

TreeTagger peut égalment être utilisé comme un (chunker) pour l'anglais, l'allemand et le français (étiquetage des parties du discours, délimitation des groupes syntaxiques, étiquetage des groupes). Nous pouvons le considérer comme un pont permettant de passer des formats bruts des données (texte, article, données web etc) à une nouvelle forme structurée facilitant le traitement de données.

Nous présentons dans le tableau ci-dessous un exemple de résultat de treetagger effectué sur la phrase suivante : *"Il permet d'annoter plusieurs langues."*.

il	PRO:PER	il
permet	VER:pres	permettre
d'	PRP	de
annoter	VER:infi	annoter
plusieurs	PRO:IND	plusieurs
langues	NOM	langue
.	SENT	.

Table 4: Exemple de résultat de Treetageer

1.5.2 NLTK

NLTK[5] [Bird, et al. 2009] est un outil pour l'élaboration de programmes PYTHON permettant de travailler sur des données de langage humain et des données textuelles. Il fournit des interfaces faciles à utiliser sur plus de 50 corpus et ressources lexicales telles que WordNet, avec un ensemble de bibliothèques de traitement de texte pour la

classification, la création de jetons, le marquage, l'analyse et le raisonnement sémantique. NLTK est disponible pour Windows, Mac OS X et Linux. C'un outil open source : projet libre offert par la communauté pour la communauté.

Dans notre approche, nous nous intéressons à cet outil, car il nous permet d'effectuer différents traitements automatiques des langues à des niveaux différents tels que lexicale, syntaxique, et sémantique. Parmi les tâches traitées par NLTK nous trouvons la segmentation de texte (en phrases ou mots), l'étiquetage, la Chunk parsing, les grammaires hors-contexte, grammaires de dépendance etc. En plus, il fournit des démonstrations graphiques, des échantillons de données, des tutoriels, ainsi que la documentation de l'interface de programmation (API). Cependant le problème avec ce logiciel est qui il est pleinement opérationnel seulement pour l'anglais, les fonctions avancées sont limitées pour les autres langues comme le français (figure 9).

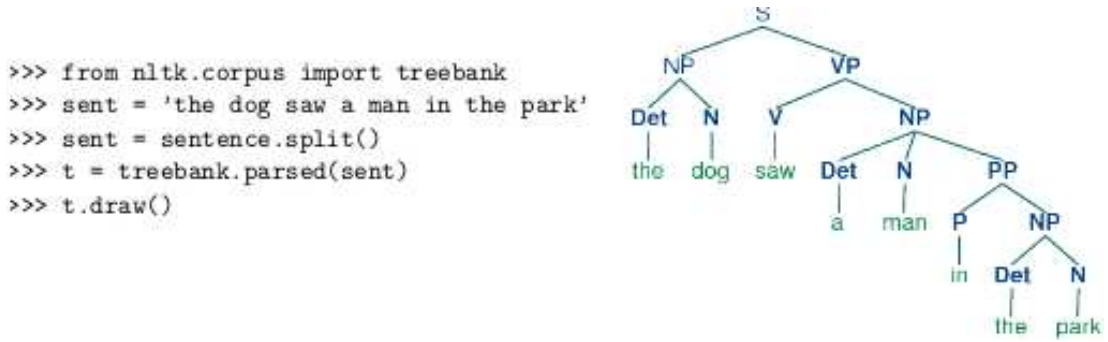


Figure 9: Arbre d'analyse avec NLTK

1.5.3 Cordial Analyseur

Cordial Analyseur[6] est un outil payant permettant d'effectuer des tâches dans le TAL et ressemble un peu à TreeTagger. Il a été crée essentiellement pour la langue française. Parmi les taches que nous pouvons effecteur à l'aide de Cordial Analyseur :

- Etiquetage morpho-syntaxique des textes en français.
- Étiquetage au format EASY fournissant l'ensemble des composants et des relations.
- Analyse statistique des caractéristiques stylistiques des textes.
- Aide à l'analyse terminologique et sémantique de corpus.

En plus de ces fonctions, Cordial Analyseur permet de calculer le nombre d'occurrence de mots (lemmatisé ou non), d'extraire des informations à statistiques à partir les textes, par exemple le pourcentage de verbes infinitif, etc, de chercher les mots-clés, les phrases-clés, etc.

Afin de démontrer ces capacités, nous prenons la figure ci-dessous comparant Cordial Analyseur et TreeTagger, faite par un groupe d'étudiants[7].

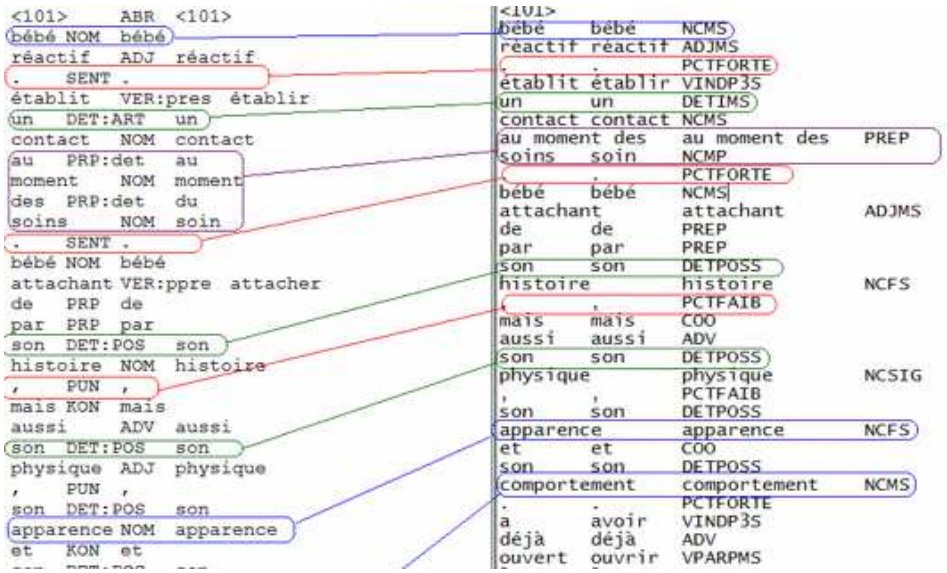


Figure 10: Comparaison TreeTagger et Cordial Analyseur

Cette comparaison montre que l'étiquetage avec Cordial Analyseur se fait de manière plus détaillée que TreeTagger. Cordial Analyseur nous donne plus d'informations sur les verbes tels que mode, genre, temps, etc, alors que Treetagger ne donne que le temps et l'infinitif verbal. Ainsi, TreeTagger a des problèmes avec les termes complexes par exemple "au moment des soins", il les traite par unités, c'est-à-dire, il décompose en mots. Par contre Cordial le traite comme une seule unité lexicale. Malgré ces inconvénients, TreeTagger est très utilisé dans le TAL, du fait de sa gratuité. En plus son résultat est acceptable en général.

Pour résumer, le principal avantage de TreeTagger sur Cordial Analyseur est qu'il permet de traiter un plus grand nombre de langues, rappelons que cordial ne fonctionne qu'avec la langue française.

1.5.4 Langage Tool

Langage Tool[8] est un correcteur grammatical libre plurilingue pour le français, l'anglais, l'allemand, le polonais, le breton, l'espéranto et plus de 20 autres langues ce qui est déjà énorme mais là où il se différencie des autres outils ou services du même type, c'est au niveau de son cœur de recherche et de correction. En effet, il trouve de nombreuses erreurs qui ne peuvent pas être signalées par un simple correcteur orthographique comme les confusions d'homonyme (des, dès, dés...), les erreurs de grammaire telles que les accords en genre ou en nombre, les conjugaisons incorrectes, etc.

Langage Tool se présente de différentes façons, que soit un service web, une extension Firefox ou bien encore une extension pour LibreOffice ou OpenOffice. Nous nous intéressons avec cet outil d'un côté du fait de sa gratuité et d'autre côté il se présente sous une extension gratuite en python (language_check.LanguageTool).

1.6 Logiciels

Dans cette partie nous verrons une liste de quatre logiciels d'évaluation automatique que nous allons utiliser plus tard comme référence dans le développement de notre modèle (Wajdi Zaghouani, 2002).

1.6.1 Essay Grade

Project Essay Grade (PEG) est le premier système d'évaluation automatique par ordinateur. Il est développé par le précurseur du domaine, E.B. Page, depuis les années soixante [Page 1994]. Ce système d'évaluation se base sur la méthode de régression multiple, qui consiste à calculer une équation à partir des traits linguistiques d'un texte choisi comme modèle par le correcteur humain. Cette équation sert par la suite pour attribuer des points à un travail. Par ailleurs d'autres critères secondaires sont pris en compte dans le calcul de la note tels que la fréquence des fautes orthographiques et syntaxiques.

1.6.2 SATO

Le logiciel SATO[9], conçu et développé par François Daoust, est disponible gratuitement, dans sa version Internet, sur le site du Centre d'analyse de textes par ordinateur de la Faculté des sciences humaines à l'Université du Québec à Montréal.
Il est, depuis longtemps, utilisé dans des disciplines faisant un fort usage de textes, telles que la sociologie, le droit ou la linguistique. Avec l'utilisation de SATO, la notion d'analyse de contenu des textes recouvre une grande variété de significations. Voici quelques exemples:

- Pour le Spécialiste en éducation, SATO est utilisé comme outil pour le calibrage des textes ou documents pédagogiques.
- Pour le Journaliste, SATO permet de retracer une information écrite ou vérifier le style ou la lisibilité de ses textes.
- Pour le Linguiste, SATO est utilisé pour analyser le fonctionnement de la langue dans son aspect lexical ou dans sa dimension syntagmatique.

1.6.3 Intelligent Essay Assessor

Le logiciel Intelligent Essay Assessor (IEA) se distingue des autres logiciels par le fait qu'il utilise une technique d'analyse sémantique latente (Latent Semantic Analysis, LSA) que nous avons déjà présentée. Cette méthode permet, en fait, de concevoir une matrice

qui contient tout le vocabulaire du texte corrélé à un vocabulaire de base fourni par le correcteur humain.

Le but de cette opération est la vérification du contenu sémantique du texte. Ainsi le logiciel peut détecter si l'élève a abordé tel ou tel sujet (Hearst 2000). Bien que le logiciel permette la détection des fautes fonctionnelles de l'anglais, il ne peut pas détecter la plupart des irrégularités syntaxiques et lexicales. IEA est plutôt orienté vers l'évaluation des connaissances dans des domaines particuliers comme la psychologie.

1.6.4 E-Rater

Selon (Burstein, 1998), E-rater se base sur la combinaison d'une méthode statistique et d'une technique de traitement de la langue naturelle. Cette technique permet d'extraire les traits et les mots clés du texte afin de vérifier si le sujet est abordé par l'élève.
Par ailleurs, E-rater dispose d'outils permettant l'analyse syntaxique et morphologique de la phrase afin de détecter les fautes courantes de la langue anglaise.

Dans le but de résumer les critères d'évaluation des logiciels précédents nous présentons dans le tableau ci-dessous un ensemble de critères pour chaque logiciel.

Les critères d'évaluation	E-rater	IEA	PEG	SATO
La stylistique et les fautes de style	X			X
La cohérence textuelle et sémantique	X			X
La variété lexicale	X	X	X	X
Les fautes syntaxiques	X		X	
L'orthographe d'usage	X	X	X	X
La longueur des phrases	X	X	X	X
Le nombre de paragraphes	X			
La fréquence des mots et des phrases	X	X	X	X

Table 5: Critères d'évaluation

1.7 Piste de recherche et l'hypothèse

L'évaluation automatique d'un texte reste toujours un problème d'actualité qui intéresse les chercheurs depuis relativement longtemps. Comme nous l'avons mentionné, notre but est de définir et mettre en œuvre un outil permettant d'évaluer le corpus posté par les auteurs sur le site Short Édition. Rappelons que les textes sont déjà évalués par un comité éditorial composé de lecteurs et auteurs de la communauté de Short Édition, volontaires et bénévoles : ce sont des lecteurs aguerris, férus de littérature, mais ce ne sont pas des "pros".

Par conséquent, le choix de la publication d'une nouvelle repose entièrement sur un groupe de personnes physiques. C'est donc avant la publication que mon outil interviendra afin de permettre un premier tri et faciliter la tâche des relectures humaines, proposant une évaluation de toutes les nouvelles postées. Il laisse à la communauté le choix de valider ou non la publication sur le site. Cette étape apporte une précision supplémentaire au niveau de la relecture puisqu'elle est automatique. Ce contexte nous pousse à nous demander si ces textes sélectionnés sont déjà de bonne qualité. Sont-ils bien structurés ? Les indicateurs sont-ils bien indiqués ? Si non, comment définissons-t-nous ces indicateurs ? Quelles pistes pour mesurer la qualité d'un texte ? Est-ce qu'il y a des niveaux de qualité ? Si oui quels sont ces différents niveaux de qualité ? Évaluer, mais selon quelle conception de l'évaluation ? Quelles méthodologies sont liées à l'évaluation ? Parmi ces méthodologies, lesquelles seraient des entités mesurables, quantifiables, évaluables ? Comment nous y prendre pour réaliser cette évaluation ? Quels sont les aspects d'évaluation qui sont efficaces ? Comment et en quoi sont-ils efficaces ? Sur quoi portent l'évaluation, et éventuellement ses limites ?

La question générale, est : comment évaluer ? Une série de questions tourne dans notre tête. Comme point de départ, nous prendrons en compte les quatre niveaux d'analyse que nous avons exposé au début.

Conclusion : Nous avons présenté un éventail de travaux et d'outils réalisés sur l'évaluation automatique de textes avec des finalités différentes. Cette étude nous a permis d'avoir une idée générale sur un domaine en plein essor et ses défis. Elle nous a permis surtout de relever des points qui doivent être pris en considération dans le cadre de notre projet.
La recherche dans ce domaine est toujours très pertinente, car les résultats obtenus aujourd'hui sont encore sujets à amélioration, donc notre mission principale et d'essayer de fournir le meilleur outil possible, pour l'ensemble des données, ici l'analyse de nouvelles, tout en prenant en compte leurs spécificités.

2 Méthodologie

2.1 Introduction

Pour nos expérimentations, nous nous appuyons sur un corpus qui traite d'un ensemble de textes écrits en français et qui sont issus du domaine littéraire de format court, lus en moins de 20 minutes. Ce corpus est présent dans la base de données de la maison d'édition électronique Short-Édition. Le corpus comporte plus de 23 000 textes, mais ils ne sont pas tous publiés, nous avons 13 000 textes publiés et 10 000 non publiés.

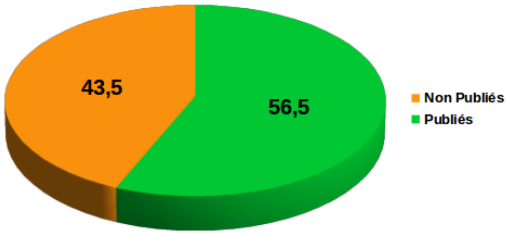


Figure 11: Taux de textes publiés et non publiés

Avant d'entrer dans les détails, nous allons aborder une explication simple sur le mécanisme d'évaluation de Short-Édition. Comme nous l'avons indiqué dans le paragraphe précédent, il y a des textes publiés et d'autres non, cela-ci nous conduit à nous demander :

- Par qui les textes sont écrits ?
- Par qui ils sont évalués ? Sur quels critères ?

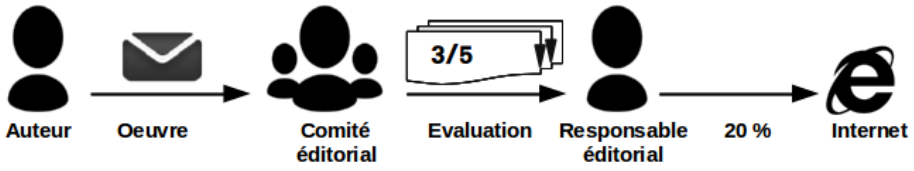


Figure 12: Mécanisme d'évaluation

Tentons de répondre à ces questions. La figure 12 illustre en bref le mécanisme d'évaluation. Actuellement les auteurs envoient leurs œuvres à l'aide d'un formulaire dans le site, transmis au comité éditorial[1]. Les évaluations s'effectuent "à l'aveugle" : Sans connaître le nom de l'auteur ni les évaluations des autres, qui apparaîtront une fois la note et l'évaluation validée. Pour donner son évaluation, le lecteur doit saisir une note entre 1 et 5 (cf.figure 13). La publication est décidée chaque jour par la Direction Éditoriale sur la base des évaluations et des commentaires des membres du Comité.

[1]Le Comité Éditorial est une partie de la communauté de Short Édition : il est composé de plus de 60 internautes, lecteurs ou auteurs de Short Édition, volontaires et bénévoles : ce sont des lecteurs aguerris, férus de littérature.

Exemple :

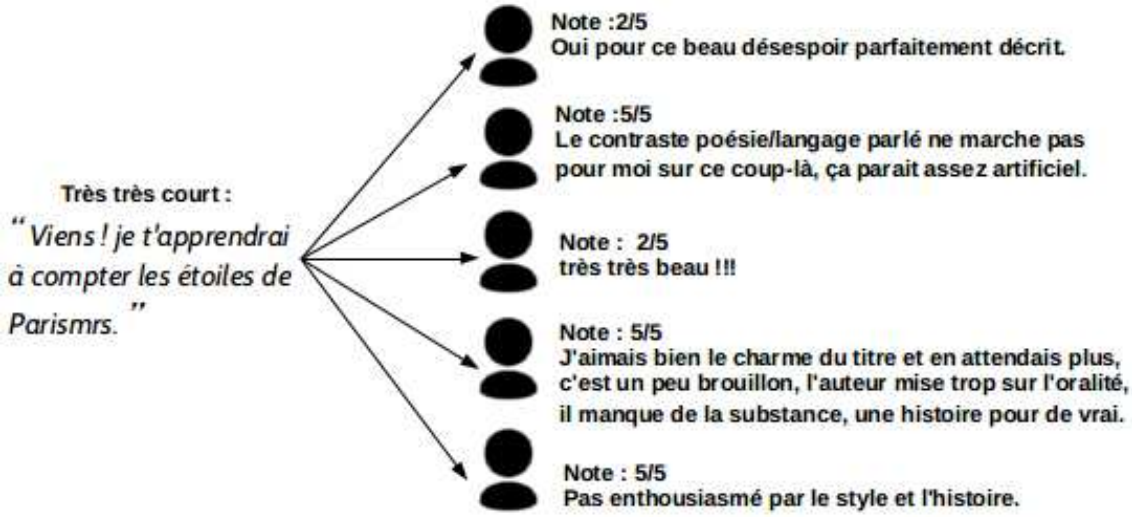


Figure 13: Exemple d'évaluation de comité

Cependant, recevant en ce moment de l'ordre de 2 000 œuvres par mois, il ne sera bientôt plus possible de toutes les lire. Cela donne la naissance de notre objectif de définir et mettre en œuvre un outil permettant d'évaluer et filtrer les textes à l'aide d'un tri rapide.

Types de textes Rappelons que la maison d'édition traite quatre catégories de texte : nouvelles ("Short Story", c'est un texte court variant de 1200 à 5 000 mots), micro-nouvelles (micro roman, variant de 500 à 1200 mots, ça fait moins de 5 minutes de lecture), poèmes (la structure, les strophes et la rime) et bandes dessinées (fondés sur la succession d'images dessinées accompagnées le plus généralement de textes).

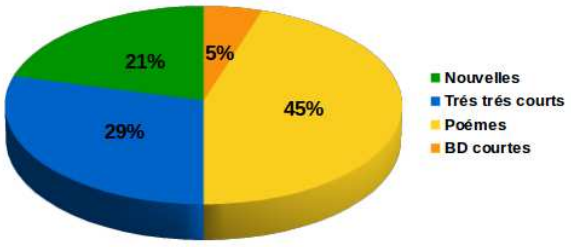


Figure 14: Taux de textes publiés et non publiés

Le figure 14 indique la distribution des catégories dans la base de données de Short Édition. Il est clair les poèmes dominent avec un taux égal à 45%, suivis par les nouvelles avec 40% et les BD avec 5%. Dans notre approche, nous ne prenons pas en compte

les BD et les poèmes, qui constituent un genre littéraire très particulier. Nous nous concentrerons sur les nouvelles et micro-nouvelles.

2.2 Démarche

Commençons de rappeler notre objectif, consistant à prédire la qualité de texte, autrement dit, à classifier les textes par rapport leurs qualités. La classification automatique est une technique utilisée dans plusieurs domaines. Sa capacité prédictive la rend rapide et efficace. Elle doit être obligatoirement précédée par des phases de préparation. Dans notre approche, ces dernières consistent à pré-traiter le corpus, suivi par l'extraction des indicateurs, fini par la préparation de matrice. La figure suivante indique les différents étapes, que nous appliquons pour atteindre notre objectif.

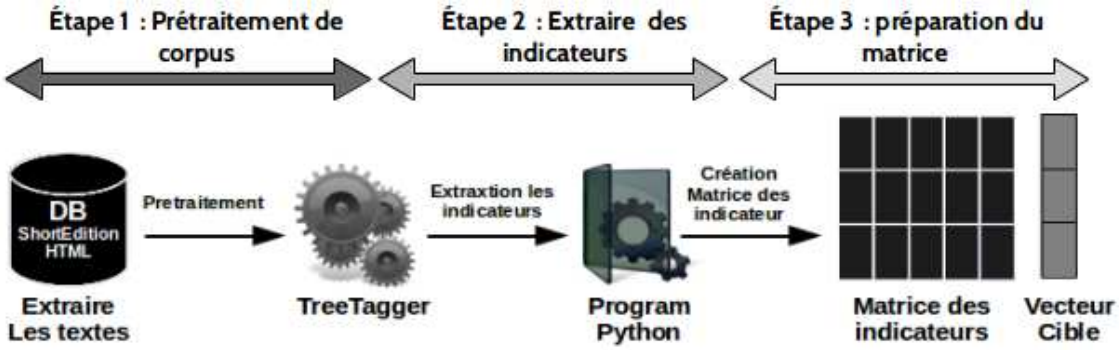


Figure 15: Chaîne de traitement par défaut

2.3 Pré-traitement

La première phase de l'algorithme implémenté, consiste à préparer le corpus dans le but d'extraire la liste des indicateurs (section 3). Nous commençons notre prétraitement par le nettoyage des textes : supprimer tous les balises HTML.

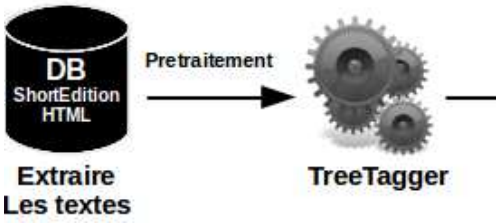


Figure 16: Chaîne de pré-traitement

Dés que les textes sont propres, ils sont passés à TreeTagger (cf. figure 16). Comme nous l'avons mentionné précédemment (section 1.5) il nous permet d'obtenir pour chaque

texte analysé, sa liste de mots après les étapes segmentation et lemmatisation. Chaque mot est associé à sa catégorie morphosyntaxique correspondante et avec son lemme.

2.3.1 Les erreurs de prétraitement

TreeTagger est un analyseur morphosyntaxique pouvant traiter plusieurs langues telles que le français. Avec les textes bien formulés linguistiquement, il est capable de les traiter avec une très bonne qualité. Dans notre cas, la rédaction des textes se fait d'une manière publique ce qui nous oblige de vérifier la qualité d'étiquetage. Une analyse de la qualité de l'étiquetage montre 87% d'étiquetage acceptable. Cette analyse est effectuée à l'aide d'une fonction Python permet de calculer le pourcentage d'erreurs dans l'étiquetage. Nous essayons dans cette partie d'analyser et de présenter les types des erreurs.

L'encodage : Le problème d'encodage, est un problème connu en traitement automatique de langue. Dans la pratique, nous nous appuyons sur l'encodage UTF-8 pour traiter nos textes. Malheureusement, le site de Short Édition étant public, les textes sont rédigés par plusieurs auteurs utilisant des systèmes très différents.

La segmentation Peu d'erreurs sont dues à cette étape, qui sont liées à des erreurs de typographie, telles que l'absence d'espace après une virgule ou un point.

Ponctuations : Les erreurs à ce niveau sont dues au fait que TreeTagger rencontre des difficultés pour reconnaître certains signes de ponctuation et n'arrive pas à les normaliser.

Polylexicales : Les erreurs à ce niveau viennent du fait que Treetagger ne parvient pas à détecter les mots composés comme une unité et les découpe en plusieurs mots. Toutefois, l'impact de ces erreurs sur l'étiquetage est faible et leur traitement ne s'impose pas comme priorité par rapport aux autres types d'erreurs présentées ci-dessus.

2.3.2 Nos solutions

Pour répondre aux problèmes soulevés dans la section précédente, nous introduisons plusieurs étapes de pré-traitement (figure 12) que nous décrivons ci-dessous.

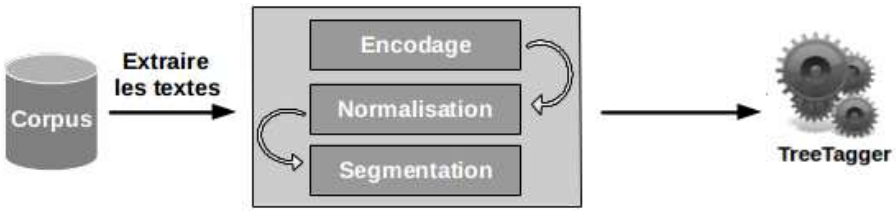


Figure 17: Nouvelle chaîne de traitement

Encodage : Les erreurs de type encodage proviennent essentiellement des caractères accentués qui sont encodés différemment selon les systèmes choisis tels que MS-DOS, iso-latin-1, utf-8, utf-16, ANSI, MIME, etc. Ils sont identifiés par le lemme unknown sous TreeTagger. Nous avons implémenté une fonction qui permet de rendre tous les textes compatibles en utf-8.

Normalisation : Les marques de ponctuations inconnues de Treetagger ont été normalisées comme indiqué dans le tableau 5.

Ponctuations Inconnus	Remplacer par
”	"
“	"
”	"
..	.
’	,
«	"
´´	"
´	'
‘	'
’	'
—	-
–	-

Table 6: Normalisation des marques de ponctuations inconnus par Treetagger

Segmentation : Nous avons implémenté une fonction permettant de corriger les erreurs de typographie selon les règles de la table 6.

Type de ponctuation	Signe	Espace avant	Espace après
Virgule	,	Non	Oui
Point	.	Non	Oui
Point Virgule	;	Oui	Oui
Deux points	:	Oui	Oui
Points de suspension	...	Non	Oui
Apostrophe	’	Non	Non
Exclamation ou d’interrogation	! ?	Oui	Oui
Tiret entre deux mots	-	Non	Non
Signes mathématiques	$- + */ =>< \%$	Oui	Oui

Table 7: Les règles d’espacement des Signes et de la Ponctuation

3 Représentation des indicateurs textuels

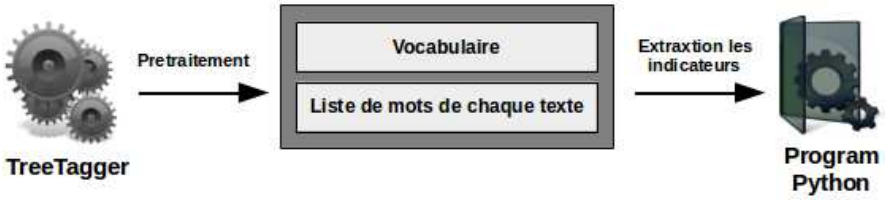


Figure 18: Extraction les valeurs des indicateurs

Dans cette deuxième étape, nous continuons le traitement commencé précédemment. Cette étape est venue après l'étape de préparation des corpus, que nous avons expliquée dans la section précédente (section 2). Maintenant, nous expliquons la première phase de l'algorithme implémenté permettant de préparer les données sous un format que les fonctions d'apprentissage et de classification comprennent.

Ce format se présente comme suit:

```
<Indice>:<Val>   <Indice>:<Val>   <Indice>:<Val>   <Indice>:<Val>
<Indice>:<Val>   <Indice>:<Val>   <Indice>:<Val>   <Indice>:<Val>
<Indice>:<Val>   <Indice>:<Val>   <Indice>:<Val>   etc.
```

`<Indice>` : un entier qui représente l'indice de l'indicateur
`<Val>` : la valeur de l'indicateur.

⇒ `Chaque ligne représente un document.`

Dans notre cas, l'extraction des indicateurs se décompose en deux parties, la première consiste à construire un vocabulaire de mots selon la méthode Bag of Words (section 3.1), et la deuxième s'intéresse à calculer des indicateurs supplémentaires (section 3.2).

3.1 Bag of words

Dans notre cas nous avons construit un dictionnaire de mots de tous les textes. Pour un texte donné les valeurs de `<Indice>` correspondent aux indices de ses mots dans le vocabulaire. Les valeurs `<Val>` correspondent au nombre d'occurrences de chaque mot de vocabulaire dans chaque texte.

Le préprocessing est très important dans le processus de classification car la construction du modèle est basée sur les données préparées. En effet des données mal préparées risquent de donner un modèle non performant. Ainsi nous avons procédé à un nettoyage des textes pour lemmatiser les mots et supprimer les stop words et les symboles de ponctuations.

Élimination de stop words : Cette phase consiste à supprimer tous les mots standards dans un texte, ce sont des mots très communs et utilisés dans pratiquement tous les textes. Leur présence peut dégrader la performance de l'algorithme de classification en terme de coût et en terme de précision de la classification.

Nombre d'occurrence : Après l'étape de préprocessing et l'élimination de stop words, nous calculons la fréquence de chaque mot de vocabulaire dans chaque texte.

Si dans le texte à classer un mot n'est jamais apparu dans le corpus d'apprentissage nous lui attribuons un poids nul (égal 0). Les vecteurs caractéristiques des textes sont composés des indices de leurs lemmes dans le vocabulaire accompagnés des poids correspondants.

Exemple Nous prenons l'exemple de trois textes.

Texte 1: Ce matin vous vous êtes réveillé dans la clarté d'une journée que le printemps vole à l'hiver, en arborant le sourire onctueux de ceux qui, après un instant de panique, se souviennent, qu'on est le week-end.
Texte 2: Depuis samedi soir, je ne dors plus. Je cherche à comprendre. Je sais que je n'y arriverai pas. On ne peut pas comprendre ça. Trois jours qu'il est là-bas.
Texte 3: Depuis combien de temps suis-je là ? Ici, il n'y a plus aucune notion de durée. L'obscurité est si épaisse qu'il est impossible de distinguer quoi que que ce soit. Lentement, le froid s'insinue à travers mes habits.

Première étape : Tokenization et lemmatisation

Texte 1: ['ce', 'matin', 'vous', 'vous','être','réveiller', 'dans', 'le','clarté', 'de', 'un','jour', 'que', 'le', 'printemps', 'voler', 'le', 'hiver', 'en', 'arborer', 'le', 'sourire', 'onctueux', 'de', 'celui', 'qui', 'un', 'instant', 'de', 'panique', 'se souvenir', 'que', 'on', 'être', 'le', 'week-end']
Texte 2: ['depuis', 'samedi', 'soir', 'je', 'ne', 'dormir', 'plus', 'je', 'chercher', 'à' ,'comprendre', 'je', 'savoir', 'que', 'je', 'ne', 'y', 'arriver', 'pas', 'on', 'ne', 'pouvoir', 'pas', 'comprendre','ça', 'trois', 'jour', 'que', 'il', 'être', 'là-bas']
Texte 3: ['depuis', 'combien', 'de', 'temps', 'être', 'je', 'là','ici', 'il', 'ne', 'y', 'avoir', 'plus', 'aucun', 'notion', 'durée','le','obscurité', 'être', 'si', 'épaisseur','être, 'que', 'il', 'impossible', 'de', 'distinguer', 'quoi', 'que', 'que', 'ce', 'être', 'lentement', 'le', 'froid', 's'insinuer', 'à','travers', 'mon', 'habit']

Deuxième étape : Supprimer les mots outils (Stop words)

Texte 1: ['matin', 'être', 'réveiller', 'clarté', 'jour', 'printemps', 'voler', 'hiver', 'arborer', 'sourire', 'onctueux', 'celui', 'instant', 'panique', 'souvenir', 'être', 'week-end']
Texte 2: ['samedi', 'dormir', 'chercher', 'comprendre', 'savoir', 'arriver', 'soir', 'pouvoir', 'soir', 'comprendre', 'ça', 'trois', 'jour', 'être', 'là-bas']
Texte 3: ['combien', 'temps', 'être', 'là', 'avoir', 'aucun', 'notion', 'durée', 'obscurité', 'être', 'épaisseur', 'être', 'impossible', 'distinguer', 'être', 'froid', 'insinuer', 'travers', 'habit']

Troisième étape : Construction du vocabulaire

Rang	Mot	Rang	Mot	Rang	Mot
0	matin	21	comprendre	42	impossible
1	être	22	savoir	43	distinguer
2	réveiller	23	arriver	44	froid
3	clarté	24	soir		
4	jour	25	pouvoir		
5	printemps	26	soir		
6	voler	27	comprendre		
7	hiver	28	ça		
8	arborer	29	trois		
9	sourire	30	jour		
10	onctueux	31	durée		
11	celui	32	là-bas		
12	instant	33	combien		
13	panique	34	temps		
14	souvenir	35	travers		
15	obscurité	36	là		
16	week-end	37	avoir		
17	samedi	38	aucun		
18	dormir	39	notion		
19	chercher	40	habit		
20	insinuer	41	épaisseur		

Quatrième étape : calcul des occurrences
Nous calculons le nombre d'occurrences de chaque mot de vocabulaire dans chaque texte.

Texte	**1:**	0:1	1:1	2:1	3:1	4:1	5:1	6:1	7:1	8:1	9:1
10:1	11:1	12:1	13:1	14:1	15:0	16:1	17:0	18:0	19:0		
20:0	21:0	22:0	23:0	24:0	25:0	26:0	27:0	28:0	29:0		
30:0	31:0	32:0	33:0	34:0	36:0	37:0	38:0	39:0	40:0		
41:0	42:0	43:0	44:0	45:0							

Texte	**2:**	0:0	1:1	2:0	3:0	4:0	5:0	6:0	7:0	8:0	9:0
10:0	11:0	12:0	13:0	14:0	15:0	16:0	17:1	18:1	19:1		
20:0	21:1	22:1	23:1	24:1	25:1	26:1	27:1	28:1	29:1		
30:1	31:0	32:1	33:0	34:0	36:0	37:0	38:0	39:0	40:0		
41:0	42:0	43:0	44:0								

Texte	**3:**	0:0	1:4	2:0	3:0	4:0	5:0	6:0	7:0	8:0	9:0
10:0	11:0	12:0	13:0	14:0	15:0	16:0	17:0	18:0	19:0		
20:1	21:0	22:0	23:0	24:0	25:0	26:0	27:0	28:0	29:0		
30:0	31:1	32:0	33:1	34:1	36:1	37:1	38:1	39:1	40:1		
41:1	42:1	43:1	44:1								

Comme nous l'avons mentionné, nous ne renseignons que les mots de poids non nul donc le résultat devient comme suit :

Texte	**1:**	0:1	1:1	2:1	3:1	4:1	6:1	7:1	8:1	9:1	10:1
11:1	12:1	13:1	14:1	15:1	16:1						

Texte	**2:**	1:1	17:1	18:1	19:1	20:1	21:1	22:1	23:1	24:1
25:1	26:1	27:1	28:1	29:1	30:1	32:1				

Texte	**3:**	1:4	15:1	20:1	31:1	33:1	34:1	36:1	37:1	38:1
39:1	40:1	41:1	42:1	43:1	44:1					

Les travaux en qualité textuelle ont toujours visé à paramétrer les textes sous la forme de variables qui constituent de bons indices pour prédire la qualité d'un texte. Le mécanisme pour prédire la qualité textuelle se base sur d'autres dimensions du texte que le Bag of words, telles que le nombre de caractères, le nombre de mots, lisibilité, la complexité et la richesse du texte notamment dans la diversité du vocabulaire.

Une fois le corpus pré-traité, le poids (nombre d'occurrences) de chaque mot de texte est calculé. L'étape suivante consiste à identifier un ensemble de caractéristiques linguistiques aussi appelées indicateurs. Celles-ci doivent entretenir une relation de corrélation significative avec la qualité de texte.

Toutefois, un bon indicateur doit égalment répondre à d'autres conditions. En particulier, être le moins corrélé possible avec les autres indicateurs afin d'éviter des redondances d'informations. Dans cette optique, nous proposons de classer nos indicateurs

en cinq familles: indicateurs généraux, lexicaux, grammaticaux, complexité et fautes d'écriture.

3.2 Indicateurs généraux

3.2.1 Longueur des mots

La plupart des lecteurs d'histoires courtes préfèrent les histoires les plus lisibles et compréhensibles. La longueur des mots peut être vue comme un indicateur de lisibilité d'un texte. Plus le mot court et contient moins de syllabes plus la lecture sera facile. Prenons cet indicateur parmi les indicateurs qui peuvent influer sur la qualité de texte.

$$AvgLengthword(T) = \frac{1}{n}\sum_{i=1}^{n} Cart_i$$

où, n: nombre total de mots du texte T ; $Cart_i$: nombre de caractères dans le i-ème le mot.

3.2.2 Longueur des phrases

La longueur des phrases peut être vue comme un indicateur de lisibilité d'un texte. Nous trouvons parfois des phrases qui pourraient facilement être scindées en deux. Plus elle est courte, plus elle est simple à lire, plus facile à retenir. Mais d'un point de vue littéraire, une phrase trop courte ne permet pas de donner assez de détails. Il faut donc viser à former des phrases d'une longueur raisonnable, ni trop longues ni trop courtes.

Nombre moyen de caractères par phrase

$$AvgLengthSentChar(T) = \frac{1}{n}\sum_{i=1}^{n} CPhar_i$$

où,n: nombre de phrase du texte T ; $CPhar_i$: nombre de caractères dans la i-ème phrase

Nombre moyen de mots par phrase

$$AvgLengthSentWord(T) = \frac{1}{n}\sum_{i=1}^{n} MPhar_i$$

où, n: nombre de phrase du texte T ; $MPhar_i$: nombre de mots dans la i-ème phrase.

La phrase idéale selon les travaux de (Flesch, 1948; Gunning, 1952) comprendrait en moyenne 14.5 mots pour un texte littéraire.

3.2.3 Longueur des paragraphes

Un paragraphe est une section de texte en prose développant une idée précise et comptant normalement plusieurs phrases. Un texte sans paragraphes peut contenir exactement les mêmes mots, mais pouvant être à comprendre étant donné qu'il n'existe pas de délimitation entre les concepts ou les arguments présentés dans un texte.

Un paragraphe est composé de phrases elle-mêmes composées de mots et de caractères. Il y a donc 3 façons de calculer la longueur d'un paragraphe.

Nombre moyen de caractères par paragraphe

$$AvgLengthParagCart(T) = \frac{1}{n}\sum_{i=1}^{n} CParag_i$$

où, $CParag_i$: Nombre des caractères dans le i-ème paragraphe ; n : nombre total des paragraphes dans le texte T

Nombre moyen de mots par paragraphe

$$AvgLengthParagWord(T) = \frac{1}{n}\sum_{i=1}^{n} MParag_i$$

où, $MParag_i$: Nombre des mots dans le i-ème paragraphe ; n : nombre total des paragraphes dans le texte T

Nombre moyen de phrases par paragraphe

$$AvgLengthParagSent(T) = \frac{1}{n}\sum_{i=1}^{n} PhParag_i$$

où, $PhParag_i$: Nombre des phrases dans le i-ème paragraphe ; n : nombre total des paragraphes dans le texte T.

Exemple Dans cet exemple, nous sélectionnons trois textes de types différents : le premier de la base de Short Édition, le deuxième extrait des Rêveries du promeneur solitaire - 1782 - structuré en une seule phrase longue qui comprend quelques mots peu usuels (JJ Rousseau) et le troisième, c'est un extrait de (la Disparition, 1989) structuré en phrases très courtes (ne contient pas le caractère "e"). Nous passons les trois textes sur notre prototype que nous avons créé.

Texte 1 : Short Édition (Publié, Note : 1), Amours-amies
Des amours se déguisant en amitiés. Des amitiés rougissant d'amour avant de se laisser glisser dans les douceurs de l'intimité...Un ami, jamais vraiment oublié qui reparaît aux mémoires fidèles d'un jour, une nuit, toujours enfoui et vivant. Un souvenir en partage, une promesse autrefois gagnée un matin et que l'on tient dans sa main comme un trésor enfui.

Texte 2 : JJ Rousseau, Rêveries du promeneur solitaire
Quand le lac agité ne me permettait pas la navigation, je passais mon après-midi à parcourir l'île en herborisant à droite et à gauche, m'asseyant tantôt dans les réduits les plus riants et les plus solitaires pour y rêver à mon aise, tantôt sur les terrasses et les tertres, pour parcourir des yeux le superbe et ravissant coup d'œil du lac et de ses rivages couronnés d'un côté par des montagnes prochaines et de l'autre élargis en riches et fertiles plaines, dans lesquelles la vue s'étendait jusqu'aux montagnes bleuâtres plus éloignées qui la bornaient.

Texte 3 : Georges Perec, La Disparition
Anton Voyl n'arrivait pas à dormir. Il alluma. Son Jaz marquait minuit vingt. Il poussa un profond soupir, s'assit dans son lit, s'appuyant sur son polochon. Il prit un roman, il l'ouvrit, il lut; mais il n'y saisissait qu'un imbroglio confus, il butait à tout instant sur un mot dont il ignorait la signification. Il abandonna son roman sur son lit. Il alla à son lavabo; il mouilla un gant qu'il passa sur son front, sur son cou.

Indicateur	Texte 1	Texte 2	Texte 3
Longueur du texte en caractères	277	465	354
Longueur du texte en mots	63	102	85
Longueur du texte en phrases	4	1	7
Longueur du texte en paragraphes	1	1	1
Nombre moyen de caractères par mot	4.39	4.55	4.16
Nombre moyen de caractères par phrase	69.25	465	51
Nombre moyen de mots par phrase	15.75	102	12.14
Nombre moyen de caractère par paragraphe	277	465	354
Nombre moyen de mots par paragraphe	63	102	85
Nombre moyen de phrase par paragraphe	4	1	7

Table 8: Exemple de résultat des indicateurs généraux

3.3 Indicateurs lexicaux

3.3.1 Richesse Vocabulaire

Il est difficile de définir la notion de richesse vocabulaire. En effet, il n'existe pas de définition unique de la richesse vocabulaire, pour déterminer cet indicateurs, nous appuyons sur la méthode la plus populaire : la fraction des mots absents, de la liste de

mots outils (Stop words). Cette dernière contient 271 mots les plus utilisés en français,[2] donc moins il y a de mots outils plus le vocabulaire est riche. Le calcul se fait de la manière suivante :

$$RichesseVocab(T) = \frac{N_{NonOutils}}{N_{total}}$$

où, $N_{NonOutils}$: Nombre de mots qui n'appartiennent pas à la liste de mots outils; N_{total} : nombre de mots total du texte T.

3.3.2 Densité lexicale

La densité lexicale permet de mesurer le degré d'information d'un texte. Pour déterminer la valeur de la densité lexicale, nous la calculons sur deux niveaux, premièrement, la distribution de mots rares et distincts, deuxièmement, la distribution de mots différents dans le texte.

Fréquence relative de mots distincts et rares : Nous entendons par rare, un mot de signification très caractérisée dont l'extension est très petite, sa fréquence d'apparition dans le texte est extrêmement faible.

Dans notre cas, nous définissons comme mots rares les mots qui n'appartiennent pas à la liste française de Gougenheim. Cette liste, créée sur la base du corpus "Élaboration du français fondamental" (163 textes, 312.135 mots et 7.995 lemmes différents), présente pour 8774 mots de fréquence supérieure à 20 leur répartition. Notons, que le calcul se fait sans prendre en compte les répétitions.

$$MotDistRar(T) = \frac{N_{NonGoug}}{N_{total}}$$

où, $N_{NonGoug}$: nombre de mots qui n'appartiennent pas à la liste de Gougenheim T ; N_{total} : nombre de mots total du texte T.

Distribution de mots différents : Cette approche n'est pas trop loin de l'approche précédente. Nous mesurons la distribution de mots différents, et qui s'appellent mots uniques. Cette étape nous permet de reconnaître l'effet de répétition des mots (Bowers, 2000). Par extension, moins le champ lexical d'un texte est varié, plus celui-ci est supposé facile.

$$MotDiff(T) = \frac{N_{SansRepetition}}{N_{total}}$$

où, $N_{SansRepetition}$: nombre de mots différents dans T ; N_{total}: nombre de mots total du texte T.

[2]Mots outils : le, la, l', un, une, ma, ta, sa, mon, ton, son, ce, les, des, mes, tes, ses, ces ; du, au ; quel, quelle, je, tu, il, elle, nous, vous, ils, elles, en, y, tout, on, etc.

3.3.3 Entropie

Nous partons de l'approche de (Howes et Solomon, 1951) qui ont considéré que la fréquence de mots d'un texte peut être un indicateur d'évaluation de sa richesse lexicale. Cette méthode nous permet de mesurer la quantité moyenne d'information attribuable à un texte constitué par un ensemble de mots, représentant le degré d'incertitude où l'on est de l'apparition de chaque mot (Piéron 1963, Ling. 1972).

$$Entropie(T) = -\sum_{i=1}^{n} P(w_i) \log_2 P(w_i)$$

où, n : Nombre total de mots du texte T ; w_i : Mots d'indice i dans T.

Exemple Pour repère, nous calculons l'entropie de deux textes, le premier de 7 mots tous différents et le deuxième de 7 fois le même mot.

Texte 1: Depuis combien de temps suis-je là ?

Texte 2: Depuis depuis depuis depuis depuis depuis depuis.

$$Entropie(T_1) = 0.8$$

$$Entropie(T_2) = 0$$

D'un point de vue mathématique, l'entropie comprise entre $[0, \log(n)]$ avec n nombre total de mots du texte T.

Exemple Reprenons les mêmes exemples de la section précédente.

Texte 1 : Short Édition (Publié, Note : 1), Amours-amies Des amours se déguisant en amitiés. Des amitiés rougissant d'amour avant de se laisser glisser dans les douceurs de l'intimité...Un ami, jamais vraiment oublié qui reparaît aux mémoires fidèles d'un jour, une nuit, toujours enfoui et vivant. Un souvenir en partage, une promesse autrefois gagnée un matin et que l'on tient dans sa main comme un trésor enfui.
Texte 2 : JJ Rousseau, Rêveries du promeneur solitaire Quand le lac agité ne me permettait pas la navigation, je passais mon après-midi à parcourir l'île en herborisant à droite et à gauche, m'asseyant tantôt dans les réduits les plus riants et les plus solitaires pour y rêver à mon aise, tantôt sur les terrasses et les tertres, pour parcourir des yeux le superbe et ravissant coup d'œil du lac et de ses rivages couronnés d'un côté par des montagnes prochaines et de l'autre élargis en riches et fertiles plaines, dans lesquelles la vue s'étendait jusqu'aux montagnes bleuâtres plus éloignées qui la bornaient.

Texte 3 : Georges Perec, La Disparition
Anton Voyl n'arrivait pas à dormir. Il alluma. Son Jaz marquait minuit vingt. Il poussa un profond soupir, s'assit dans son lit, s'appuyant sur son polochon. Il prit un roman, il l'ouvrit, il lut; mais il n'y saisissait qu'un imbroglio confus, il butait à tout instant sur un mot dont il ignorait la signification. Il abandonna son roman sur son lit. Il alla à son lavabo; il mouilla un gant qu'il passa sur son front, sur son cou.

Indicateur	Texte 1	Texte 2	Texte 3
Richesse vocabulaire	0.43	0.41	0.43
Fréquence relative de mots distincts et rares	0.16	0.18	0.17
Distribution des mots différents	0.74	0.65	0.64
Entropie	1.6	1.7	1.5

Table 9: Exemple de résultat des indicateurs lexicaux

3.4 Indicateurs grammaticaux

Dans cette famille d'indicateurs, nous nous intéressons à calculer la distribution des catégories grammaticales des mots tels que, nom, déterminant, adjectif, pronom, verbe, adverbe, préposition, et conjonction.

3.4.1 Fréquence relative des adjectifs

L'utilisation des adjectifs dans la rédaction peut rajouter de la puissance à l'écriture parce qu'ils jouent un rôle sémantique très important. Mais l'utilisation excessive des adjectifs peut conduire à des résultats négatifs. Dans notre évaluation nous décidons de prendre en compte le nombre moyen d'adjectifs comme indicateur de qualité :

$$FreqRelativADJ(T) = \frac{N_{Adj}}{N_{total}}$$

où, N_{Adj}: nombre de adjectifs dans T ; N_{total}: nombre de mots total du texte T.

3.4.2 Fréquence relative des noms

D'une point de vue sémantique, le nom joue un rôle important dans le texte. Il utilisé pour désigner une catégorie de personne, d'animal ou de chose. Ici, nous calculons la fréquence relative de nom de chaque texte de notre corpus :

$$FreqRelativNom(T) = \frac{N_{nom}}{N_{total}}$$

où, N_{nom} : nombre de noms dans T ; N_{total} : nombre de mots total du texte T.

3.4.3 Fréquence relative des pronoms

D'un côté, le but d'employer les pronoms à la place d'un nom dans le texte est d'éviter la répétition. Et d'un autre côté le pronom sert presque toujours à préciser de qui nous parlons. Ainsi, les pronoms rappellent l'information et assurent la cohésion du texte, ils sont considérés comme un lien permettant de maintenir le sens de texte dans l'esprit de lecteur. Ici, nous nous intéressons à calculer la fréquence relative des pronoms :

$$FreqRelativPronPer(T) = \frac{N_{Pron}}{N_{total}}$$

où, N_{Pron} : nombre de pronoms dans T ; N_{total}: nombre de mots total du texte T.

3.4.4 Fréquence relative des verbes auxiliaires

Les verbes auxiliaires jouent deux rôles différents. Ils peuvent être utilisés comme verbe à part entière permettant d'introduire un attribut. Sinon, ils se combinent à un verbe principal pour constituer un temps composé. Dans la rédaction, ils sont fréquemment utilisés. Ce qui nous a incités à calculer leur fréquence relative dans chaque texte.

$$AvgNombDistrubVrA(T) = \frac{N_{VAux}}{N_{total}}$$

où, N_{VAux} : nombre de verbes auxiliaires dans T ; N_{total} : nombre de mots total du texte T.

3.4.5 Fréquence relative des prépositions

Le rôle de préposition est d'introduire un nom, un pronom ou une proposition relative. C'est un mot invariable, qui fait partie des mots-outils. Bien qu'elle qu'elle n'influe pas trop sur la rédaction, elle fréquemment utilisée. Nous calculons ici la fréquence relative des prépositions dans chaque texte.

$$FreqRelativProp(T) = \frac{N_{Prop}}{N_{total}}$$

où, N_{Prop} : nombre de prépositions dans T ; N_{total} : nombre de mots total du texte T.

3.4.6 Fréquence relative des déterminants

D'un point de vue syntaxique, le rôle du déterminant est d'introduire un nom, il reçoit le genre et le nombre du nom. Et d'un point de vue sémantique, certains déterminants établissent un lien avec un groupe nominal déjà désigné dans le texte. Les déterminants font partie des mots outils, ils sont fréquemment utilisés dans les textes. Ici, nous calculons la distribution des déterminants dans chaque texte du corpus (fréquence relative de déterminants).

$$FreqRelativDet(T) = \frac{N_{Det}}{N_{total}}$$

où, N_{Det} : nombre de déterminants dans T ; N_{total} : nombre de mots total du texte T.

3.4.7 Fréquence relative des verbes

Le verbe est le noyau de la rédaction, il permet d'indiquer, ce que nous faisons, ce qui se passe, ce que nous ressentons, etc. Nous ne pouvons pas décrire une phrase ou un paragraphe ou un texte sans employer de verbes. Ici, nous calculons la distribution des verbes dans notre corpus (dans chaque texte).

$$FreqRelativVerb(T) = \frac{N_{Verb}}{N_{total}}$$

où, N_{Verb} : nombre de verbes dans T ; N_{total} : nombre de mots total du texte T.

3.4.8 Fréquence relative des conjonctions

Les conjonctions sont fréquemment utilisées dans la langue française, ils relient deux mots, deux groupes de mots, deux propositions de même nature, etc. Par ailleurs, ils jouent le rôle de connecteur, entre les phrases. Le but d'employer les conjonctions dans les textes est d'améliorer la compréhension du texte.

$$FreqRelativConj(T) = \frac{N_{Conj}}{N_{total}}$$

où, N_{Conj} : nombre de conjonctions dans T ; N_{total} : nombre de mots total du texte T

3.4.9 Distribution des adverbes

Un adverbe est un mot invariable qui permet de modifier un verbe, un adjectif, ou un autre adverbe. L'une de ses utilités est d'indiquer le degré d'une qualité ou d'un défaut et de donner des informations sur ce que pense celui qui parle.

$$FreqRelativAdv(T) = \frac{N_{Adv}}{N_{total}}$$

où, $NAdv$: nombre d'adverbes dans T ; N_{total}: nombre de mots total de texte T

Exemple Reprenons les mêmes exemples avec des indicateurs grammaticaux.

Texte 1 : Short Édition (Publié, Note : 1), Amours-amies
Des amours se déguisant en amitiés. Des amitiés rougissant d'amour avant de se laisser glisser dans les douceurs de l'intimité...Un ami, jamais vraiment oublié qui reparaît aux mémoires fidèles d'un jour, une nuit, toujours enfoui et vivant. Un souvenir en partage, une promesse autrefois gagnée un matin et que l'on tient dans sa main comme un trésor enfui.

Texte 2 : JJ Rousseau, Rêveries du promeneur solitaire
Quand le lac agité ne me permettait pas la navigation, je passais mon après-midi à parcourir l'île en herborisant à droite et à gauche, m'asseyant tantôt dans les réduits les plus riants et les plus solitaires pour y rêver à mon aise, tantôt sur les terrasses et les tertres, pour parcourir des yeux le superbe et ravissant coup d'œil du lac et de ses rivages couronnés d'un côté par des montagnes prochaines et de l'autre élargis en riches et fertiles plaines, dans lesquelles la vue s'étendait jusqu'aux montagnes bleuâtres plus éloignées qui la bornaient.

Texte 3 : Georges Perec, La Disparition
Anton Voyl n'arrivait pas à dormir. Il alluma. Son Jaz marquait minuit vingt. Il poussa un profond soupir, s'assit dans son lit, s'appuyant sur son polochon. Il prit un roman, il l'ouvrit, il lut; mais il n'y saisissait qu'un imbroglio confus, il butait à tout instant sur un mot dont il ignorait la signification. Il abandonna son roman sur son lit. Il alla à son lavabo; il mouilla un gant qu'il passa sur son front, sur son cou.

Indicateur	Texte 1	Texte 2	Texte 3
Fréquence relative des adjectifs	0.032	0.8	0.02
Fréquence relative des noms	0.25	0.18	0.21
Fréquence relative des pronoms	0.08	0.07	0.22
Fréquence relative des verbes auxiliaires	0	0	0
Fréquence relative des verbes	0.16	0.13	0.2
Fréquence relative des prépositions	0.19	0.2	0.1
Fréquence relative des conjonctions	0.06	0.07	0.02
Fréquence relative des déterminants	0.14	0.14	0.16
Fréquence relative des adverbes	0.06	0.06	0.03

Table 10: Exemple de résultat des indicateurs grammaticaux

3.5 Indicateurs de complexité

L'analyse de la complexité du texte amène à distinguer les dimensions du texte, correspondant aux différents niveaux textuels tels que, lexical, syntaxique, sémantique, etc.

Dans cette famille, nous intéressons à calculer la complexité à travers la mesure du nombre moyen de phrases complexes, le pourcentage de cohésion, et la lisibilité d'un texte.

3.5.1 Nombre moyen de phrases complexes

Une phrase complexe est constituée de plusieurs propositions liées entre elles. Les propositions peuvent s'enchaîner de différentes manières, par juxtaposition, coordination ou subordination. Ici, nous calculons le nombre moyen des phrases complexes dans chaque texte de corpus.

$$AvgNombSentComplx(T) = \frac{N_{Prop}}{N_{total}}$$

où, N_{Prop}: nombre de propositions dans T ; N_{total}: nombre total de phrases.

3.5.2 Cohésion du texte

La mesure de cohésion textuelle est considérée parmi les indicateurs importants d'évaluation de la qualité textuelle. Elle s'intéresse particulièrement aux relations locales du texte telles que les règles morphologiques et syntaxiques, ou les connecteurs argumentatifs. Elle s'attache aussi à vérifier si les actions présentées sont bien situées dans le temps, grâce aux verbes et aux points de repère, et que les relations entre les phrases sont bien marquées grâce à des mots-liens.

Différentes approches ont été proposées et restes jusqu'à aujourd'hui un sujet de débat. Nous avons décidé d'utiliser une technique de mesure de la cohésion inter phrastique introduite par (Foltz, Kintsch et Landauer, 1998 ; Spooren, 2006 ; Foltz, 2007). Elle consiste à calculer le cosinus moyen de toutes les paires de phrases adjacentes.

L'approche la plus fréquemment employée pour atteindre cet objectif est l'Analyse Sémantique Latente (Section 1.4). C'est une technique mathématique qui vise à extraire un espace sémantique de petite dimensions à partir de l'analyse statistique. Cette technique nous permet d'estimer les similarités sémantiques entre des mots, des phrases ou des paragraphes, donc plus deux phrases sont sémantiquement proches, plus leur cosinus est élevé.

Dans notre approche nous nous intéressons au calcul de la cohésion au niveau des phrases. Pour réaliser notre objectif nous suivons les étapes de la méthode LSA que nous avons expliqué auparavant (section 1.4.4).

3.5.3 Lisibilité

Nous avons mentionné au début de ce mémoire les deux méthodes les plus connues dans la mesure de lisibilité (Robert Gunning, 1968 ; Flesch, 1946). Dans notre cas, nous calculons le taux de lisibilité sur la base de longueur de mots. Si la longueur moyenne

des mots en français écrit est de 4 à 8 lettres, il serait faux de croire que la plupart des mots comptent 4 à 8 lettres.

Une étude plus fine, grâce à un programme en python, nous permet de montrer la répartition des mots selon leur nombre de lettres. Nous avons remarqué nettement que les mots de 2, 3, 4 et 5 lettres sont les plus fréquents, avec une nette dominance pour les mots de 2 lettres. (Mesnager, 1979) établit un lien évident entre longueur des mots, rareté et difficulté de lecture : plus un mot est long, plus il est rare et plus il sera difficile à interpréter. Ici, nous calculons le nombre moyen de mots de plus 9 caractères sous la base de l'équation suivante.

$$lisibilit(T) = 1 - \frac{N_{long}}{N_{total}}$$

où, N_{total} : nombre total de mots du texte T ; N_{long} : nombre de mots de plus 9 caractères du texte T.

Exemple Nous appliquons notre proposition sur l'exemple suivant :

Texte(T) *L'**interdépartementalisation** de l'**assainissement francilien** est en marche, tandis que la station d'Achères est mise en service en 1940.* [Julie Védie, Île-de-France]

$$lisibilit(T) = 0.84$$

Exemple Reprenons les mêmes exemples avec des indicateurs de complexité.

Texte 1 : Short Édition (Publié, Note : 1), Amours-amies
Des amours se déguisant en amitiés. Des amitiés rougissant d'amour avant de se laisser glisser dans les douceurs de l'intimité...Un ami, jamais vraiment oublié qui reparaît aux mémoires fidèles d'un jour, une nuit, toujours enfoui et vivant. Un souvenir en partage, une promesse autrefois gagnée un matin et que l'on tient dans sa main comme un trésor enfui.

Texte 2 : JJ Rousseau, Rêveries du promeneur solitaire
Quand le lac agité ne me permettait pas la navigation, je passais mon après-midi à parcourir l'île en herborisant à droite et à gauche, m'asseyant tantôt dans les réduits les plus riants et les plus solitaires pour y rêver à mon aise, tantôt sur les terrasses et les tertres, pour parcourir des yeux le superbe et ravissant coup d'œil du lac et de ses rivages couronnés d'un côté par des montagnes prochaines et de l'autre élargis en riches et fertiles plaines, dans lesquelles la vue s'étendait jusqu'aux montagnes bleuâtres plus éloignées qui la bornaient.

Texte 3 : Georges Perec, La Disparition
Anton Voyl n'arrivait pas à dormir. Il alluma. Son Jaz marquait minuit vingt. Il poussa un profond soupir, s'assit dans son lit, s'appuyant sur son polochon. Il prit un roman, il l'ouvrit, il lut; mais il n'y saisissait qu'un imbroglio confus, il butait à tout instant sur un mot dont il ignorait la signification. Il abandonna son roman sur son lit. Il alla à son lavabo; il mouilla un gant qu'il passa sur son front, sur son cou.

Indicateur	Texte 1	Texte 2	Texte 3
Nombre moyen de phrases complexes	1.75	13	1
Cohésion	0.25	0.20	0.27
Lisibilité	0.95	0.91	0.96

Table 11: Exemple de résultat des indicateurs de complexité

3.6 Fautes d'écriture

En linguistique, nous parlons d'orthographe pour tout ce qui touche à la manière d'écrire les mots. En général, les fautes varient selon le niveau de langue par exemple :

Orthographe lexicale (d'usage) : correspond à la manière d'écrire les mots tels qu'elle apparaît dans les dictionnaires.

Orthographe grammaticale : erreur relative aux accords (du déterminant, de l'adjectif, du verbe, du part. passé), aux règles grammaticales et aux conjugaisons.

Typographique : erreur portant sur l'usage de signes de ponctuation et d'autres signes graphiques.

En informatique, la distinction entre les types de fautes se fait en fonction d'un outil automatique qui est capable de la détecter. Il existe deux techniques de mesure le nombre de fautes, toutes deux fondées sur le principe du pattern-matching, c'est-à-dire sur la correspondance exacte entre un élément et un modèle.

- Technique 1 : Basée sur des règles de grammaire, qui décrivent des patterns grammaticalement corrects. Si une partie du texte ne correspond à aucun pattern, une erreur est détectée.

- Technique 2 : Basée sur des règles d'erreurs, qui comparent le texte non pas à des modèles corrects, mais à des modèles de fautes.

Dans notre projet nous nous intéressons à un outil qui s'appelle "Langage Tool" pour détecter le nombre de fautes d'écriture. Notons, qu'il n'indique pas le type de faute. Mais grâce à son efficacité et sa gratuité nous l'utilisons pour traiter notre indicateur.

Ici, nous calculons le nombre de fautes d'écriture dans chaque texte, à travers l'équation suivante.

$$FauteMoyen(T) = \frac{N_{Fautes}}{N_{total}}$$

où N_{Fautes} : Nombre de fautes dans un texte T ; N_{total} : Nombre total de mots dans un texte T

Exemple Reprenons les trois textes précédents comme exemple.

Texte 1 : Short Édition (Publié, Note : 1), Amours-amies
Des amours se déguisant en amitiés. Des amitiés rougissant d'amour avant de se laisser glisser dans les douceurs de l'intimité...Un ami, jamais vraiment oublié qui reparaît aux mémoires fidèles d'un jour, une nuit, toujours enfoui et vivant. Un souvenir en partage, une promesse autrefois gagnée un matin et que l'on tient dans sa main comme un trésor enfui.

Texte 2 : JJ Rousseau, Rêveries du promeneur solitaire
Quand le lac agité ne me permettait pas la navigation, je passais mon après-midi à parcourir l'île en herborisant à droite et à gauche, m'asseyant tantôt dans les réduits les plus riants et les plus solitaires pour y rêver à mon aise, tantôt sur les terrasses et les tertres, pour parcourir des yeux le superbe et ravissant coup d'œil du lac et de ses rivages couronnés d'un côté par des montagnes prochaines et de l'autre élargis en riches et fertiles plaines, dans lesquelles la vue s'étendait jusqu'aux montagnes bleuâtres plus éloignées qui la bornaient.

Texte 3 : Georges Perec, La Disparition
Anton Voyl n'arrivait pas à dormir. Il alluma. Son Jaz marquait minuit vingt. Il poussa un profond soupir, s'assit dans son lit, s'appuyant sur son polochon. Il prit un roman, il l'ouvrit, il lut; mais il n'y saisissait qu'un imbroglio confus, il butait à tout instant sur un mot dont il ignorait la signification. Il abandonna son roman sur son lit. Il alla à son lavabo; il mouilla un gant qu'il passa sur son front, sur son cou.

Indicateur	Texte 1	Texte 2	Texte 3
Nombre moyen de fautes	0	0.009	0.02
Type de fautes	*****	Typographie (Manque espace après le point: **...Un ami**)	Typographie (Manque espace après le point-virgule: **lut;** et **lavabo;**)

Table 12: Exemple de résultat Fautes d'écriture

4 Évaluation

L'évaluation humaine de la qualité de texte demande plusieurs participants, chacun évaluant le système en fonction de critères précis, tels que les fautes d'orthographes, cohésion, etc. Ce type d'évaluation donne la mesure la plus exacte des performances de système de la qualité d'un écrit, mais elle sollicite plusieurs experts, ce qui rend la tâche coûteuse.

De plus, ce type d'évaluation pose des problèmes de non-reproductibilité et de variabilité inter-annotateurs. C'est pourquoi plusieurs mesures automatiques et objectives ont été développées, dont l'objectif est d'être corrélées avec les scores que produirait une évaluation humaine, tout en étant beaucoup moins coûteux.

4.1 Qualité ?

Le terme qualité a évolué depuis son apparition, ce qui rend impossible, une définition précise, non contradictoire et sans ambiguïté. Ce terme vient du latin qui signifie "la manière d'être plus ou moins distincte". Selon Larousse : "ce qui rend quelque chose supérieur à la moyenne".

Dans notre cas, nous intéressons à la qualité textuelle, mais la définition de la qualité dans ce domaine semble encore incertaine, c'est ce qui nous a incité à utiliser le terme "bien écrit". Donc un texte est bien écrit s'il respecte les règles d'écriture, qui sont réparties en deux catégories, superficielles et profondes.

Pour la première catégorie, nous trouvons, la longueur de texte qui nous permet de savoir s'il est conforme ce une attente (par exemple c'est une histoire courte), la longueur de mots (plus un texte contient de mots long, plus la probabilité qu'il soit complexe est forte. Cela peut nuire au degré d'écriture), la longueur de phrase (Plus une phrase est longue, plus elle exige d'efforts de compréhension de la part lecteur, elle faut être ni trop longues ni trop courtes).

La deuxième catégorie contient des règles plus profondes aux niveaux orthographique, grammatical et typographique. En plus de ces règles, la richesse du vocabulaire, et la cohérence de texte peuvent influer sur la valeur du texte. De ce fait, nous ne pouvons pas savoir par avance lequel de nos indicateurs est un bon prédicateur de la qualité de nos textes.

4.2 Évaluation visuelle des indicateurs

Ici, nous considérons que les textes publiés sont des textes bien écrits, Les textes non-publiés étant de moins bonne qualité. Rappelons, que le jugement était fait à l'avance par le comité éditorial, chaque membre du comité donne une note entre 1 et 5 accompagnée avec des commentaires plus des arguments justifiant la note. La publication est décidée

chaque jour par la Direction Éditoriale sur la base des évaluations et des commentaires. Notons que les textes sont classés dans la base de données d'une manière binaire (+1) textes publiées et (-1) non publiées. L'objectif de ce calcul est d'utiliser le modèle de classification obtenu de la phase d'apprentissage pour classer les nouveaux textes.

Rappelons que notre base textuelle de short-édition est composée par deux grands types de textes des nouvelles et des poèmes. Comme nous l'avons indiqué dans la [Section 2] le poème est une forme de texte hybride qui n'est ni une nouvelle, ni une histoire brève. C'est pour cette raison que nous avons décidé de séparer les deux types de texte. Dans un premier temps, nous appliquerons nos calculs que sur les nouvelles.

Dans le but d'illustrer et de rendre plus compréhensibles nos évaluations, nous utilisons des graphes de types histogramme et nuage des points. Remarquons que les deux graphes nous fournissent des informations précieuses sur la répartition des valeurs. Nous utilisons les histogrammes pour représenter la répartition des valeurs d'indicateurs par rapport au statut des textes soit publiés (+1) soit non publiés (-1). L'axe des X représente les valeurs d'indicateurs, et l'axe des Y représente la densité de probabilité.

Comme il indiqué plus tôt, nos indicateurs sont nombreux, c'est pour cela nous sélectionnons pour chaque famille, les indicateurs qui semblent pertinents pour discriminer entre les documents publiés et ceux non-publiés. Mais cela ne signifie pas que nous allons nous passer du reste des indicateurs, nous les mettrons dans la partie annexe.

4.2.1 Indicateurs généraux

Pour rappel, nos indicateurs sont classés dans cinq familles différentes, tout d'abord, nous commençons par la famille **indicateurs généraux**, elle permet de calculer le poids du texte, à travers la longueur moyenne de mots, de phrases, et de paragraphes. Après analyse de ce calcul, nous remarquons que la longueur moyenne de mots permet de dégager quelques tendances, comme il est indiqué dans les deux figures ci-dessous.

Longueur moyenne de mots : On peut constater visuellement que la longueur moyenne des mots est assez nettement corrélés avec la qualité des textes : les textes comportant, en moyenne, des mots plus longs ont tendances très plus refusés. Il est surprenant de voir qu'un tel indicateur simple apporte une information utile.

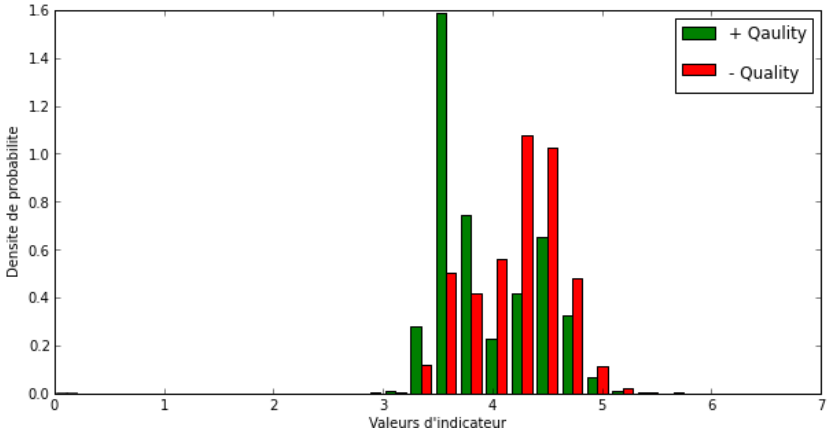


Figure 19: La longueur moyenne de mots

La figure ci-dessus montre la répartition de longueurs moyennes des mots dans les textes. En vert, nous avons les textes de meilleure qualité et en rouge les textes de moins bonne qualité. Il nous est possible de tirer plusieurs enseignements de ce figure. Tout d'abord, concernant les valeurs comprises entre (3...6), on note que pour des valeurs comprise entre $(3...4)$ il y a plus de textes de bonne qualité que de textes de moins bonne qualité. La probabilité d'avoir des textes de meilleure qualité est forte tandis que celle d'obtenir des textes de moins bonne qualité est plus faible. En revanche, pour l'intervalle $(4...6)$ on s'aperçoit qu'il y a plus de textes de moins bonne qualité que de textes de meilleure qualité. La probabilité d'avoir des textes de moins bonne qualité est forte tandis que celle d'obtenir des textes de meilleure qualité est plus faible.

D'après ces observations, nous pouvons retenir l'indicateur de longueur moyenne de mot dans un texte parmi les indicateurs qui permettent d'évaluer les textes de Short Édition.

4.2.2 Indicateurs lexicaux

Densité Vocabulaire Ici, nous nous intéressons à calculer la fréquence relative des mots qui ne sont pas outils. Le calcul se fait de manière suivante : Premièrement nous supprimons la répétition dans le texte. Deuxièmement nous calculons le nombre de mots absents de la liste de mots outils. Après ce calcul, l'indicateur densité vocabulaire nous permet de tirer plusieurs enseignements.

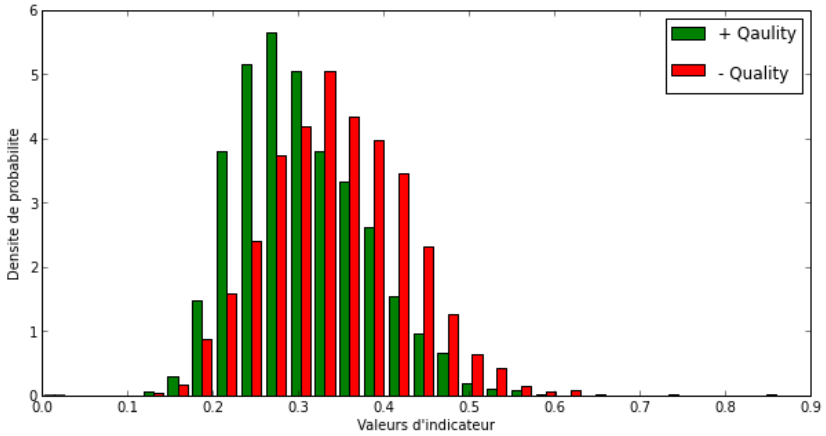


Figure 20: Distribution de densité Vocabulaire

D'après l'histogramme ci-dessus les textes de meilleure qualité ont une distribution spécifique c'est-à-dire si le nombre de mots absents de la liste de mots outils dans un texte T est inférieur ou égal au tiers (0 ... 0.32) du nombre total de mots dans T la probabilité d'avoir un texte de meilleure qualité est forte tandis que celle d'avoir un texte de moins bonne qualité est plus faible (0.32 ... 0.7).

D'après cet indicateur, pour avoir un texte de bonne qualité, il faut respecter l'équilibre entre les mots outils et les mots pleins.

Fréquence relative des mots différents La fréquence de mots différents ou de mots uniques sélectionnée parmi les indicateurs prédictifs. Pour rappel, la fréquence de mots différents égal le nombre de mots différents divisé par le nombre de mots total.

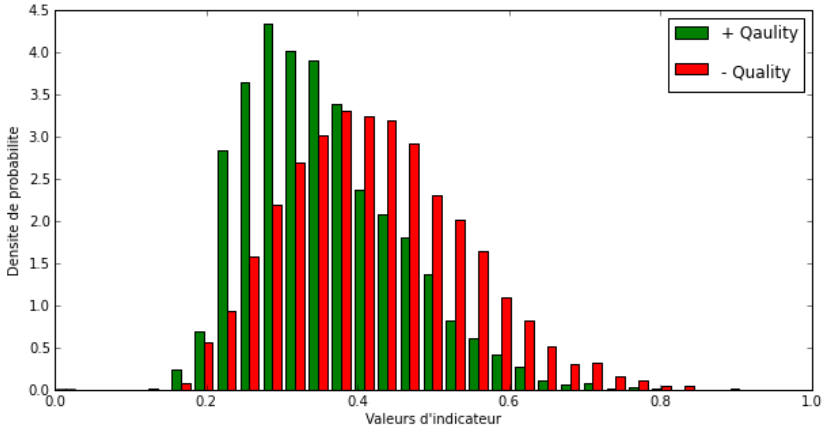


Figure 21: Distribution des mots différents

La plupart des gens pensent que pour être de bonne qualité, un texte doit contenir des mots variés c'est-à-dire des mots différent le plus souvent possible. Toutefois, cela peut l'affecter et de le rendre incompréhensible.

Dans notre corpus, nous remarquons un résultat un peu étonnant. La figure ci-dessus nous montre la distribution des mots différents. Les valeurs varient entre (0 ... 1). Nous notons que pour des valeurs comprises entre (0...0.4) il y a plus de textes de bonne qualité que de textes de moins bonne qualité. La probabilité d'avoir des textes de meilleure qualité est forte tandis que celle d'obtenir des textes de moins bonne qualité est plus faible. En revanche, pour l'intervalle (0.4...0.8) on s'aperçoit qu'il y a plus de textes de moins bonne qualité que de textes de meilleure qualité. La probabilité d'avoir des textes de moins bonne qualité est forte tandis que celle d'obtenir des textes de meilleure qualité est plus faible.

4.2.3 Indicateurs grammaticaux

Dans cette partie nous parlons des indicateurs grammaticaux interagissant avec notre corpus. Nous avons huit indicateurs dans cette catgorie (cf. 3.4) , et 5 d'entre-eux s'avèrent très pertinents pour discriminer les deux classes de document.

Fréquence relative des adjectifs : Le premier indicateur qui attire notre attention dans cette famille, représente la distribution des adjectifs dans les textes. Les deux figures ci-dessous nous montre la répartition des fréquences relatives d'adjectifs dans notre corpus.

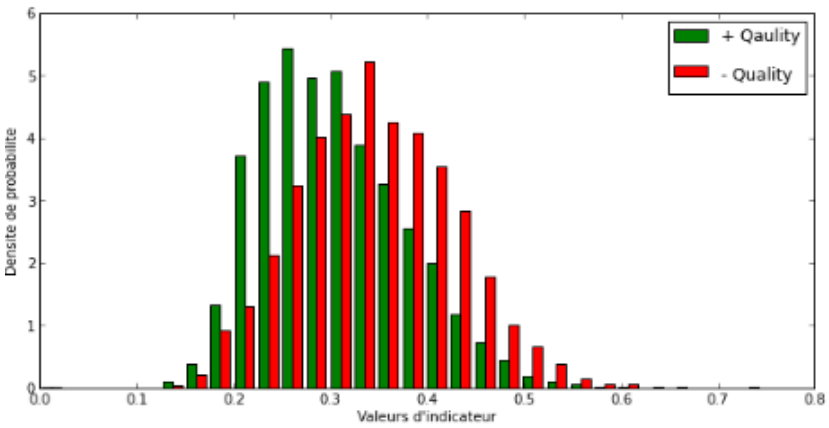


Figure 22: Distribution des adjectifs

D'après l'histogramme ci-dessus les textes de meilleure qualité ont une distribution spécifique c'est-à-dire si le nombre d'adjectifs dans un texte T est inférieur ou égal au tiers (0 ... 0.3) du nombre total de mots dans T la probabilité d'avoir un texte de

meilleure qualité est forte tandis que celle d'avoir un texte de moins bonne qualité est plus faible (0.3 ... 0.8).

Ce résultat nous renvoie à l'approche de [Franck Scandolera, 2013] trop d'adjectifs conduit à tuer le sens d'une phrase voire d'un texte. *Beaucoup d'adjectifs sont un peu comme les enveloppes à bulles, ils masquent et amortissent plus qu'ils ne révèlent.* Nous ne nions pas l'importance des adjectifs. Il faut seulement les utiliser à bon escient.

Fréquence relative des déterminants : Dans la langue française, le déterminant joue un rôle important dans la structure de texte, il permet d'identifier le nom. Ce catégorie varie selon les grammaires, plus qu'il permet d'identifier le nom, il peut accompagner les adjectifs.

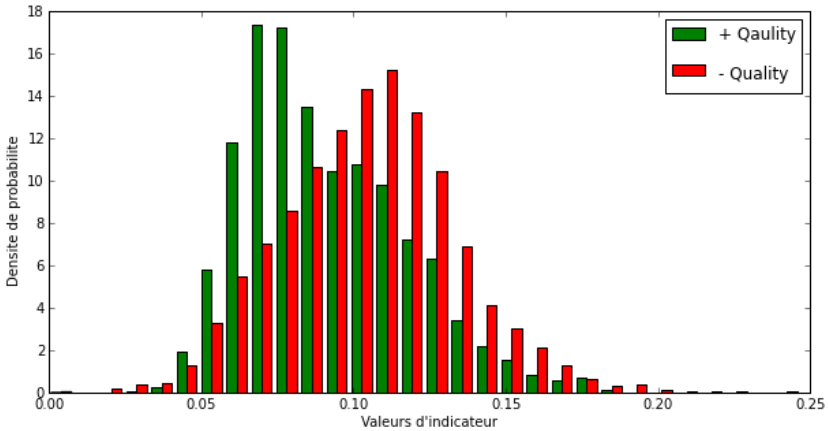


Figure 23: Distribution des déterminants

La figure ci-dessous montre la distribution des déterminants dans notre corpus, il nous est possible de tirer plusieurs enseignements de cette figure. Les fréquences relatives des déterminants se situent dans le quartier des textes entre (0 ... 0.25). Concernant les textes de meilleure qualité en vert les valeurs comprises entre (0.05 ... 0.1), et nous notons que pour des valeurs comprises entre (0.1 ... 0.25) et (0 ... 0.05) il y a plus des textes de moins bonne qualité que de textes de bonne qualité. Dans cet intervalle, la probabilité d'avoir des textes de moindre qualité est forte tandis que celle d'obtenir des textes de meilleure qualité est faible.

Autrement dit, dans ce graphique, nous remarquons que les textes ayant entre 5 et 10 pourcent des déterminants sont ceux de meilleure qualité et les textes ayant (10 et 25) et (0 et 5) pourcent de préposition sont ceux de moindre qualité.

Fréquence relative des noms : Passons maintenant à un nouvel indicateur intéressant. Nous ne distinguons pas le nom commun (nom, sans autre précision) et

le nom propre. Notons qu'un nom commun désigne des êtres, des choses ou des idées, en général (Cheval. Ville. Fille) et qu'un nom propre désigne les mêmes choses, mais en les distinguant par leur appellation (Bucéphale. Toulouse. Martine).

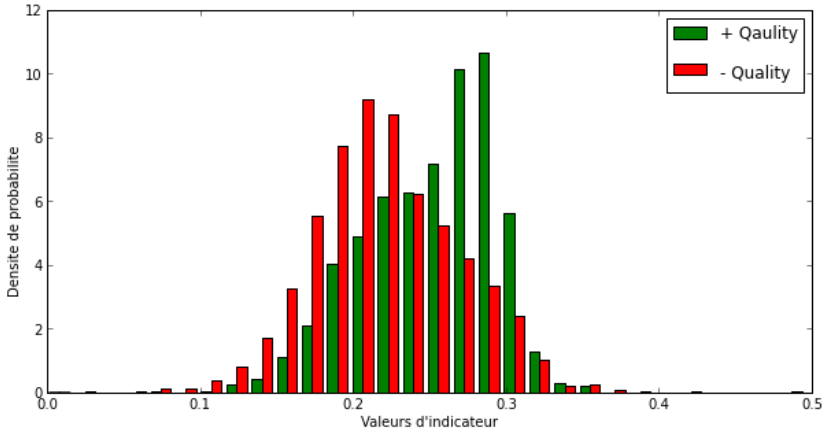


Figure 24: Distribution des noms

Concernant les valeurs comprises entre (0.25 ... 0.35) nous remarquons une forte présence de textes de bonne qualité et inversement pour les valeurs comprises entre (0 ... 0.25). Nous observons que les textes contenant un pourcentage de 30% de noms sont des textes de meilleure qualité. En revanche les textes ayant 20% de noms sont de moins bonne qualité.

Fréquence relative des prépositions : Le rôle de préposition est d'introduire un nom, un pronom ou une proposition relative.

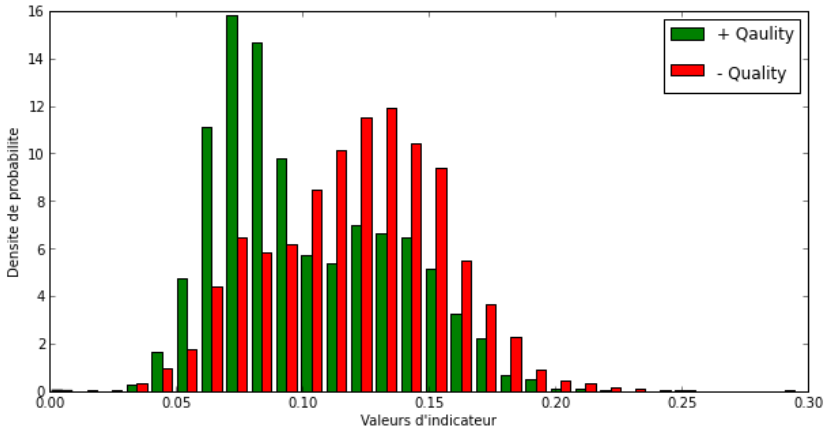


Figure 25: Distribution des prépositions

La figure ci-dessous montre la distribution des prépositions, il nous est possible de tirer plusieurs enseignements de cette figure. Nous remarquons que les fréquences relatives

des prépositions représentent le tiers du nombre total de mots des textes (0 ... 0.3). Concernant les textes de meilleure qualité en vert les valeurs comprises entre (0 ... 0.1), et nous notons que pour des valeurs comprises entre (0.1 ... 0.2) il y a plus des textes de moins bonne qualité que de textes de bonne qualité. Dans cet intervalle, la probabilité d'avoir des textes de meilleure qualité est faible tandis que celle d'obtenir des textes plus de meilleures qualités est forte. Autrement dit, dans ce graphique, nous remarquons que les textes ayant entre 5 et 10 pourcent de préposition sont ceux de meilleure qualité et les textes ayant 10 et 20 pourcent de préposition sont ceux de moins meilleure qualité.

Fréquence relative des adverbes : Beaucoup d'adverbes peuvent avoir des sens différents et appartenir, selon leur utilisation, à plusieurs de ces catégories.

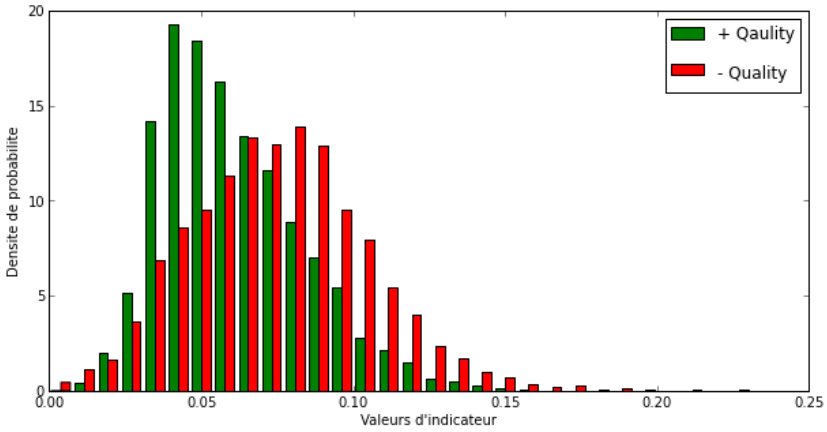


Figure 26: Distribution des adverbes

Il nous est possible de tirer plusieurs enseignements à partir la figure ci-dessus. en général, nous remarquons que la distribution est serrée (courbe très pointue), c'est-à-dire les fréquences relatives des adverbes se situent dans le cinquième des textes entre (0 ... 0.2). Concernant les textes de meilleure qualité en vert les valeurs comprises entre (0.01 ... 0.07), et nous notons que pour des valeurs comprises entre (0.07 ... 0.2) il y a plus des textes de moins bonne qualité que de textes de bonne qualité. Dans cet intervalle, la probabilité d'avoir des textes de meilleure qualité est faible tandis que celle d'obtenir des textes de meilleures qualités est forte.

Autrement dit, les textes ayant entre 1 et 7 pourcent d'adverbes sont ceux de meilleure qualité et les textes ayant 7 et 20 pourcent de préposition sont ceux de moindre qualité.

4.2.4 Indicateurs de complexité

Lisibilité : Nous avons mentionné que la lisibilité faisait partie des indicateurs les plus utilisés dans l'évaluation de textes. Pour rappel, nous souhaitons calculer la fréquence relative des mots de plus de neuf caractères.

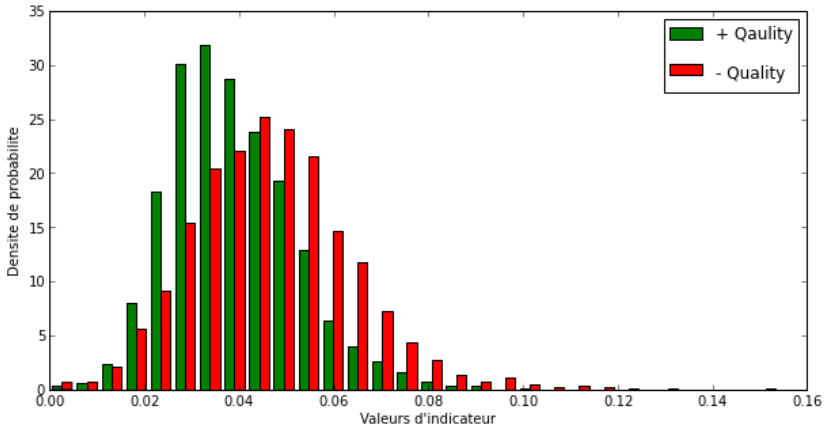


Figure 27: Histogramme de lisibilité

Nous remarquons que la figure ci-dessus est de distribution serrée, elle s'étend entre 0 et 0.12. Concernant les textes contenant entre 1 et 5 pourcent de mots de plus de 9 caractères la probabilité d'avoir un texte de meilleure qualité et forte, c'est-à-dire les valeurs comprises entre (0.01 ... 0.05) il y a plus de textes de meilleure qualité que textes de moins bonne qualité. Et pour les valeurs comprises entre (0.05 ... 0.2) il y a plus de textes de moins bonne qualité ce qui rend la probabilité d'avoir un texte de mauvaise qualité plus forte.

4.2.5 Fautes d'écriture

La logique dit que plus le texte comporte des fautes moins bonne est sa qualité. *Un livre contenant des fautes d'orthographe ou de grammaire se vendra évidemment beaucoup moins bien, voire pas du tout.* Cela nous a amenés de prendre en compte le nombre moyen de fautes comme un indicateur.

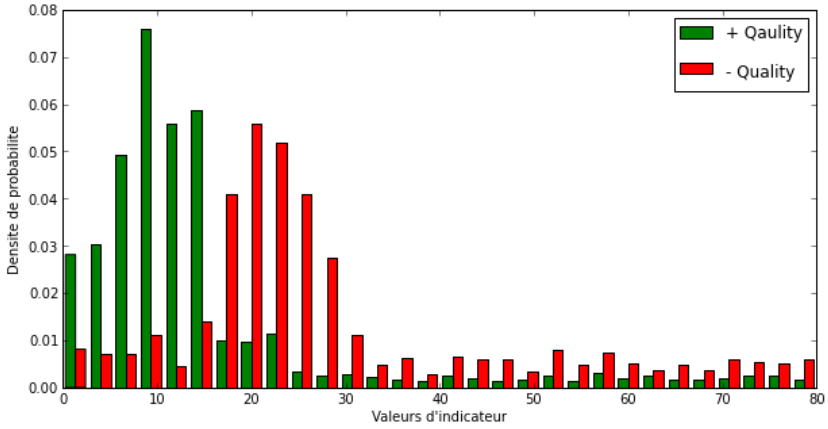


Figure 28: Distribution des fautes d'écriture

La figure ci-dessus montre la fréquence relative de fautes d'écriture dans notre corpus, il nous est possible de tirer plusieurs enseignements de ce tableau. La distribution variante entre (0 ... 80%) du textes. Concernant les textes de meilleure qualité en vert les valeurs comprises entre (0 ... 15), et nous notons que pour des valeurs comprises entre (15 ... 80) il y a plus des textes de moins bonne qualité que de textes de bonne qualité. Dans cet intervalle, la probabilité d'avoir des textes de meilleure qualité est faible tandis que celle d'obtenir des textes de meilleure qualité est forte.

Autrement dit, les textes ayant entre 1 et 15 pourcent des fautes sont ceux de meilleure qualité et les textes ayant 15 et 80 pourcent des fautes sont ceux de moindre qualité. Le nombre de faute semble très un excellent indicateur de la qualité d'un texte.

4.3 Évaluation de la pertinence statistique des indicateur (Corrélation)

Après avoir monter l'aide des histogrammes quels semblaient très les indicateurs les plus discriminer les documents "bien écrits" des autres, nous allons maintenant quantifier de faon plus formelle ce pouvoir discriminant en mesurant la corrélation entre les indicateurs et les classes.

En général, la corrélation nous permet de savoir la relation entre deux valeurs, par exemple : deux caractères quantitatifs X et Y, décrivant le même ensemble d'unités. Nous disons qui il existe une relation entre X et Y si l'attribution des modalités de X et de Y ne se fait pas au hasard, c'est-à-dire si les valeurs de X dépendent des valeurs de Y ou si les valeurs de Y dépendent des valeurs de X. Dire que Y dépend de X signifie que la connaissance des valeurs de X permet de prédire dans une certaine mesure les valeurs de Y. En d'autres termes, si Y dépend de X, nous pouvons trouver une fonction f détermine la corrélation :

$$Y = f(x)$$

Dans notre cas, d'un coté, nous considérons que Y comprend les valeurs de statuts des textes (soit 1 soit -1). La valeur (1) indique que le texte publié et le contraire pour la valeur (-1). Et d'autre coté X comprend les valeurs d'indicateur. Dans notre cas la corrélation est de type binaire (-1,1), c'est pour cela nous appuyons sur le coefficient de corrélation bisériale-point[10], cette dernière est utilisé si l'une de deux variables est binaire. La corrélation est accompagnée avec la valeur-p, permettant de mesurer l'hypothèse de corrélation, c'est un complémentaire de résultat. Les seuils suivants sont généralement pris pour référence :

- $<$ 0,01 : très forte présomption contre l'hypothèse nulle.
- 0,01 -0,05 : forte présomption contre l'hypothèse nulle.
- 0,05 -0,1 : faible présomption contre l'hypothèse nulle.
- $>$ 0,1 : pas de présomption contre l'hypothèse nulle.

Avant de présenter nos résultats nous clarifions un point sur la corrélation, plus le coefficient est proche des valeurs extrêmes -1 et 1, plus la corrélation entre les variables est forte, une corrélation égal à 0 signifie que les variables ne sont pas corrélées. Le tableau ci-dessous, nous indique un résumé sur les valeurs de corrélation.

Corrélation	Positive	Négative
Forte	de 0.7 à 1	de de -1 à -0.7
Moyenne	de 0.3 à 0.7	de -0.7 à -0.3
Faible	de 0 à 0.3	-0.3 à 0

Table 13: Valeurs de corrélation

À partir le tableau 14, nous remarquons qu'il n'y a pas de corrélation forte (de 0.7 à 1 ou de -1 à -0.7). Par ailleurs, il y a une forte présence de corrélations moyennes (de 0.3 à 0.7 ou de -0.7 à -0.3). Il y a des valeurs positives et d'autres négatives.

Une corrélation positive, c'est-à-dire à toute augmentation au niveau d'indicateur X correspond une augmentation au niveau de qualité de texte. L'indicateur et la qualité de texte, varient dans le même sens et avec une intensité similaire. Nous trouvons par exemple, fréquence relative des noms (r = 0.38, p= $2.56e^{-12}$) représente une corrélation positive moyenne, c'est-à-dire plus le texte contient des noms plus la probabilité d'avoir une qualité augmente (égale +1) est forte.

Concernant la corrélation négative, c'est-à-dire à toute augmentation au niveau d'indicateur X correspond une diminution au niveau de qualité de texte. L'indicateur et la qualité de texte varient dans deux sens opposés et avec une intensité similaire, par exemple l'indicateur faute d'écriture présente une corrélation négative proche de forte, si le nombre de fautes augmente dans le texte la qualité diminue (égale -1).

Nous remarquons également qu'il y a une forte présence de corrélations faibles aussi avec des valeur positives et négatives, telles que la longueur moyenne de paragraphe en caractères (-0.101) représente une corrélation faible et négative et la longueur moyenne de texte en caractères (0.12) qui représente une corrélation faible et positive.

Nom d'indicateur	Corrélation	Valeur-p
Nombre moyenne de caractères par texte	0.19	0.024
Nombre moyenne de mots par texte	0.109	$3.01e^{-17}$
Nombre moyenne de phrases par texte	0.2	$3.01e^{-19}$
Nombre moyenne de paragraphes par texte	-0.23	$3.37e^{-7}$
Longueur moyenne de mots	**-0.45**	0.008
Nombre moyenne de caractères par phrase	0.2	$5.63e^{-3}$
Nombre moyenne de mots par phrase	0.24	0.005
Nombre moyenne de caractères par paragraphe	-0.201	$6.55e^{-17}$
Nombre moyenne de mots par paragraphe	-0.31	$1.75e^{-17}$
Nombre moyenne de phrases par paragraphe	-0.27	0.0
Richesse Vocabulaire	**-0.361**	$4.61e^{-13}$
Fréquence relative de mots distincts et rares	**-0.29**	$1.41e^{-9}$
Distribution de mots différents	**0.403**	$3.67e^{-7}$
Entropie	0.21	$1.52e^{-46}$
Cohésion	0.28	$1.21e^{-8}$
Lisibilité	**-0.38**	$5.86e^{-95}$
Nombre de phrases complexes	0.1	0.25
Fréquence relative des noms	**0.38**	$2.56e^{-12}$
Fréquence relative des conjonctions	0.37	$2.07e^{-11}$
Fréquence relative des adverbes	0.35	0.017
Fréquence relative des verbes auxiliaires	0.18	$4.66e^{-14}$
Fréquence relative des verbes	0.28	$3.48e^{-24}$
Fréquence relative des déterminants	**0.41**	0.012
Fréquence relative des pronoms	0.247	$1.21e^{-9}$
Fréquence relative des adjectives	**0.36**	$5.16e^{-10}$
Fréquence relative des prépositions	**-0.39**	$3.37e^{-17}$
Nombre moyen de fautes d'écriture	**-0.66**	0.018

Table 14: Valeurs de corrélation entre la qualité et chaque indicateur

5 Matrice de classification

Notons, que l'étape de classification automatique de textes doit être obligatoirement précédée par une phase de prétraitement de données. Dans les sections précédentes, nous avons fait un passage sur les différentes étapes de notre chaîne de traitement. Nous passons maintenant à l'avant-dernière étape. Cette étape consiste à utiliser les techniques

du TAL pour transformer nos textes en matrice de caractéristiques nécessaires pour un classificateur. Mais avant d'entrer sur les détails de la matrice, nous expliquons les différents types de représentation textuelle, afin d'en utiliser une pour construire notre matrice.

5.1 Représentations textuelle

Dans un premier temps, les textes sont considérés comme un suite des caractères codés au niveau informatique sous la forme ASCII qui fournit 256 caractères différents ou, plus récemment Unicode (65 536 caractères) qui permet de traiter d'autres langues non-alphabétiques comme le chinois. Rappelons que nous nous intéressons dans notre projet au français. Dans les trois sous-sections, nous plaçons les trois représentations vectorielles, les plus connues.

5.1.1 Représentation binaire

Cette représentation est considérée comme la plus ancienne et la plus simple. Malgré cela, elle est encore utilisée grâce à sa souplesse entre complexité et performance des systèmes. Comme son nom indique, ce type de représentation, basée sur deux valeurs binaires (1 et 0) représente les mots du texte T dans une espace vocabulaire V. Ces dernières valeurs indiquent la présence ou l'absence du mot. Nous utiliserons comme descripteurs le lemme des mots, et supprimons les mots outils. Dans le but de réduire cette approche dans une formule, supposons que $TR_{Binaire}$ est la représentation binaire du terme TR du texte T dans vocabulaire V donc :

$$\forall \in [1..V] \begin{cases} TR_{Binaire} & = & 1 & si\ le\ terme\ de\ texte\ apparait\ dans\ V \\ & = & 0 & sinon \end{cases}$$

Exemple Pour clarifier les choses nous donnons l'exemple suivant avec trois petits textes dont chaque texte contient une seule phrase, la représentation binaire est indiqué dans le tableau cidessous avec un échantillon de vocabulaire :

Texte 1:	J'allume mon ordinateur chaque matin et chaque soir.
Texte 2:	Nous regardions les infos chaque matin.
Texte 3:	Chaque jour, je regarde la météo.

Vocabulaire V	Texte 1	Texte 2	Texte 3
allumer	1	0	0
regarder	0	1	1
ordinateur	1	0	0
chaque	1	1	1
matin	1	1	0
soir	1	0	0
infos	0	1	0
jour	0	0	1
météo	0	0	1

Table 15: Exemple de représentation binaire

5.1.2 Représentation fréquentielle

Dans cette section, nous passons à un autre type de représentation textuelle, connus sous le nom vecteur fréquentiel, peu éloignée du vecteur précédent. Au contraire, nous pouvons dire que c'est un complémentaire de la représentation binaire, qui prend en compte le nombre d'occurrences de chaque mot w_i de vocabulaire V dans le texte T_j. En général, la représentation dans l'espace V se fait à travers le nombre d'occurrence de chaque terme de V dans T. Remarquons que le terme peut être un mot ou une phrase. Comme pour l'approche précédente, nous réduisons cette approche dans la formule suivante, supposons que T_{Rfreq} est la représentation fréquentielle du terme TR de vocabulaire V dans le texte T donc :

$$\forall \in [1..V] \left\{ \; TR_{Rfreq} \quad = nombre\ d'apparition\ de\ TR\ dans\ V \right.$$

Exemple En prenant le même exemple précédent et en employant la technique de la représentation fréquentielle, avec la suppression des mots outils, le résultat indiqué dans le tableau ci-dessous :

Vocabulaire V	Texte 1	Texte 2	Texte 3
allumer	1	0	0
regarder	0	1	1
ordinateur	1	0	0
chaque	2	1	1
matin	1	1	0
soir	1	0	0
infos	0	1	0
jour	0	0	1
météo	0	0	1

Table 16: Exemple de représentation fréquentielle

Cet exemple, nous montre l'intérêt de cette approche avec le mots lemmatisés. Ce dernier n'a en effet plus le même poids pour chaque document, contrairement à sa représentation binaire.

5.1.3 Représentation tf-idf

C'est une représentation vectorielle plus informative que les deux représentations précédentes. Elle repose sur la loi de Zipf qui décrit la loi de répartition des mots de vocabulaire V dans un texte T. L'approche la plus utilisée dans la littérature est sans doute le tf-idf (Salton & Yang, 1973), (Salton et al., 1975).

Loi de Zipf Dans l'article (G.K. Zipf en 1930), les auteurs ont montré qu'en classant les mots d'un texte par fréquence décroissante, nous observons que la fréquence d'utilisation d'un mot est inversement proportionnelle à son rang. La loi de Zipf stipule que la fréquence du second mot le plus fréquent est la moitié de celle du premier, la fréquence du troisième mot le plus fréquent, son tiers, etc. Cette loi peut s'exprimer de la manière suivante :

Fréquence d'un mot de rang N = (Fréquence du mot de rang 1)/N

La loi de Zipf dit ainsi que les mots les plus informatifs d'un texte ne sont pas les mots qui apparaissent le plus dans le texte car ceux-ci sont pour la plupart des mots outils, ni les mots les moins fréquents du texte, car ces derniers peuvent en effet être des fautes d'orthographe ou encore des termes trop spécifiques.

TF-IDF : en anglais Term Frequency-Inverse Document Frequency[11]. C'est une méthode de pondération souvent utilisée en recherche d'information et en particulier dans la fouille de textes. Cette mesure statistique permet d'évaluer l'importance d'un terme contenu dans un document, relativement à une collection ou un corpus. Le poids augmente proportionnellement au nombre d'occurrences du mot dans le document. Il varie égalment en fonction de la fréquence du mot dans le corpus. Des variantes de la formule originale sont souvent utilisées dans des moteurs de recherche pour apprécier la pertinence d'un document en fonction des critères de recherche de l'utilisateur. Comme il indique dans le nom de cette méthode, elle est composée de deux parties :

- **Fréquence du terme, TF :**
  La fréquence d'un terme (term frequency) est simplement le nombre d'occurrences de ce terme dans le document considéré, précisément le nombre moyen d'occurrences de mot de vocabulaire dans un texte. Le calcul se fait de la manière suivante :

$$TF_{w_{i,j}} = \frac{Nombre\ occurrences\ w_i}{Nombre\ total\ demots\ T_j}$$

où $Nombre\ occurrences\ w_i$, nombre d'occurrences de w_i de vocabulaire V dans le texte T, $Nombre\ total\ demots\ T_j$ de texte T.

- **Fréquence inverse de document, IDF :**
  La fréquence inverse de document (inverse document frequency) est une mesure de l'importance du terme dans l'ensemble du corpus. Dans le schéma TF-IDF, elle vise à donner un poids plus important aux termes les moins fréquents, considérés comme plus discriminants. Elle consiste à calculer le logarithme de l'inverse de la proportion de documents du corpus qui contiennent le terme :

$$IDF_{w_{i,j}} = \log\left(\frac{Nombre\ occurrence\ w_i\ \ dans\ tous\ les\ textes}{Nombre\ total\ de\ textes\ contient\ w_i}\right)$$

  d'où $Nombre\ total\ de\ textes\ contient\ w_i$, nombre d'occurrence de w_i de vocabulaire V dans tous les texte T, $Nombre\ total\ de\ textes\ contient\ w_i$.

Finalement, le poids s'obtient en multipliant les deux mesures :

$$TF_IDF = IDF_{w_{i,j}}.TF_{w_{i,j}}$$

Exemple Prenons le même exemple précédent, le résultat est indique sur le tableau ci-dessous :

Vocabulaire V	Texte 1	Texte 2	Texte 3
allumer	0	0	0
regarder	0	0	0
ordinateur	0	0	0
chaque	0,02	0,003	0,003
matin	0	0	0
soir	0	0	0
infos	0	0	0
jour	0	0	0
météo	0	0	0

Table 17: Exemple de représentation TF-IDF

5.2 Construire la matrice

L'idée d'une représentation des textes introduit, la notion de matrice numérique. Cette matrice va être constituée de $D \times W$ cellules, où D représente le nombre des lignes de la matrice. Ces lignes représentent le nombre de textes de notre corpus, et W représente le nombre d'occurrences des mots du vocabulaire et les résultats des indicateurs. Plus précisément, chaque ligne de la matrice correspond à un texte T_i, contenant les résultats

des indicateurs et la représentation vectoriel de mots de vocabulaire dans le texte T_i (section 3). Auparavant il faut savoir que nous avons construit un fichier appelé Vocabulaire V, comprenant deux parties, la première représente les noms des indicateurs, et la deuxième représente les mots lemmatisés de notre corpus. Le tableau ci-dessous présente un échantillon de notre fichier V:

Indice	Indicateur	Indice	mots
0	Longueur moyenne de texte en caractère	26	comprendre
1	Longueur moyenne de texte en mots	27	savoir
2	Longueur moyenne de texte en phrases	28	arriver
3	Longueur moyenne de texte en paragraphes	29	soir
4	Longueur moyenne de mots en caractère	23	pouvoir
5	Longueur moyenne de phrase en caractère	30	soir
6	Longueur moyenne de phrase en mots	31	comprendre
7	Longueur moyenne de paragraphes en caractère	32	ça
8	Longueur moyenne de paragraphes en mots	33	trois
9	Longueur moyenne de paragraphes en phrases	34	jour
10	Fréquence relative de mots distincts et rares	35	durée
11	Fréquence relative de mots différents	36	là-bas
12	Longueur moyenne de paragraphes en caractère	37	combien
13	Densité vocabulaire	38	temps
14	Entropie	39	travers
15	Fréquence relative des adjectifs	40	là
16	Fréquence relative des noms	41	avoir
17	Fréquence relative des pronoms personnels	42	aucun
18	Fréquence relative des verbes	43	notion
19	Fréquence relative des propositions	44	habit
20	Fréquence relative des déterminants	45	épaisseur
21	Fréquence relative des conjonctions	46	avoir
22	Fréquence relative des pronoms personnels	47	aucun
23	Fréquence relative des verbes	48	notion
24	Fréquence relative des propositions	49	habit
25	Fréquence relative des déterminants	50	épaisseur
..	....	..	.... etc

Table 18: Échantillon de fichier vocabulaire

En plus, chaque texte est défini par son statut, soit publié (+1), soit non publié (-1), c'est pour cela nous prenons en compte le vecteur qui contient les valeurs de statuts des textes (cf.figure 29).

La grande taille du corpus influe sur la partie pré-traitement, elle a été très longue, notamment à la phase calcul du cohésion et du nombre moyen de fautes d'écriture qui

a duré environ 5 heures pour les 7212 textes. La figure suivante nous donne une idée générale sur le format de matrice :

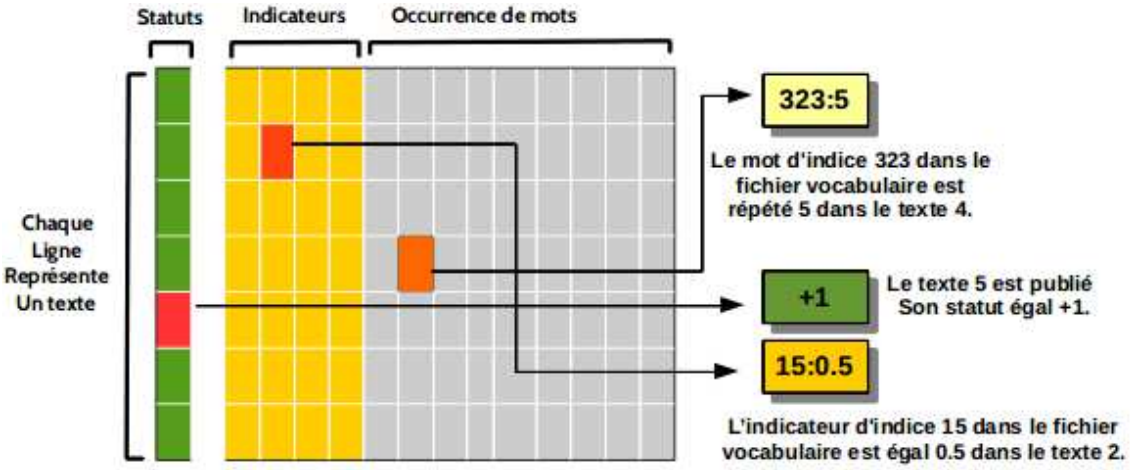


Figure 29: Représentation vectorielle des textes.

6 Classification

Avant d'entrer dans les détails, nous rappelons brièvement notre démarche. Notre parcours consiste d'abord à extraire les textes de la base de données de Short Édition. Cette étape est suivie par l'étape de prétraitement, qui nous permet de calculer nos indicateurs (section 3). Dès que le calcul est effectué, nous passons à l'étape représentant le lien entre la préparation et la classification. Elle consiste à transformer nos données en une matrice utilisable par les fonctions d'apprentissage et de classification.

Dans cette dernière étape nous allons effectuer une classification avec un outil d'apprentissage. Le but est d'apprendre, à partir d'un ensemble d'exemples, une "fonction" capable de prédire le statut d'un texte (publié, non publié) à partir de la liste des indicateurs et des mots du texte (cf. 5.2) qui caractérise ce texte.

6.1 Apprentissage

Pour arriver à l'étape classification, il faut d'abord passer par le processus d'apprentissage, en se fondant sur l'analyse de nos données qui sont présentés dans la matrice (section 5). Il existe deux types d'algorithme d'apprentissage :

6.1.1 Apprentissage supervisé

Nous nous intéressons particulièrement à ce type d'apprentissage, cependant, nous nous attacherons à expliquer l'autre type d'apprentissage. Tout d'abord le mot supervisé signifie *contrôler la réalisation d'un travail accompli par d'autres.* L'analyse

d'apprentissage supervisé se base donc sur un ensemble de classes connues et définies à l'avance. Autrement dit, nous pouvons dire que l'apprentissage est supervisé si les classes sont prédéterminées et les exemples connus. Le processus se passe en deux phases :

- **La première phase**, consiste à apprendre un modèle de données étiquetées.
- **La seconde phase**, consiste à prédire l'étiquette d'une nouvelle donnée.

6.1.2 Apprentissage non supervisé

Contrairement à l'apprentissage supervisé, cette approche se base sur un ensemble de classes inconnues et non définies à l'avance. Elle consiste donc à apprendre et classer sans supervision. Il vise à diviser l'ensemble des données en différents(paquets) Chaque sous-ensemble partage des caractéristiques communes. Nous pouvons le trouver dans plusieurs domaines tels que :

- Médecine: la découverte de variétés de patients qui souffrent de caractéristiques physiologiques communes.
- la transformation de la construction d'un système de reconnaissance vocale de la voix humaine.
- Traitement de l'image

6.2 Algorithme de classification

Dans notre cas, nous décidons d'employer quatre type des classificateurs afin de trouver celui qui sera le plus adapté à nos données. Nous choisissons comme classificateurs les quatre[12] plus connus, indiqués dans le tableau ci-dessous :

Classificateur	**Types d'apprentissage**
SVM linear, Non linear	supervisé
Random Forest Classifier	supervisé
Descente de gradient stochastique	supervisé
Arbres de Décision ADD	supervisé

Table 19: Types classificateurs

6.2.1 Support Vector Machine (SVM)

Support vector machine (SVM) est une méthode de classification basée sur l'apprentissage supervisé, d'une manière statistique et automatique. Créé par [Vapnik, en 1995]. Il permet de traiter un problème de classification bi classe comme notre cas. Il existe deux types de séparation entre les données, un linéaire et l'autre non-linéaire. Le deuxième type non-linéaire utilise un ensemble de fonctions mathématiques, appelées noyaux pour séparer les objets. La figure ci-dessous permet d'expliquer en général le principe de cette méthode :

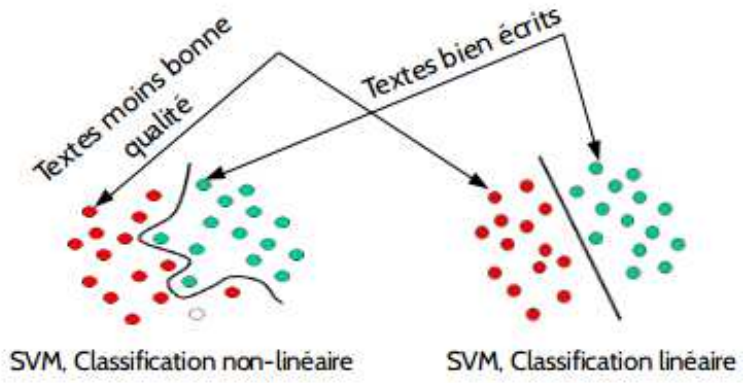


Figure 30: Principe de SVM.

Avantages SVM Les avantages de support vector machines sont:

- Efficacité dans les espaces de grande dimension.
- Utilise un sous-ensemble de points de formation dans la fonction de décision(appelée vecteurs de support), il est égalment efficace au niveau de la mémoire.
- Polyvalent: différentes fonctions du noyau peuvent être spécifiées pour la fonction de décision. Des noyaux communs sont prévus, mais il est égalment possible de spécifier des noyaux personnalisés.

Inconvénients SVM Les inconvénients de support vector machines sont:

- Si le nombre d'éléments est beaucoup plus important que le nombre d'échantillons, le procédé est susceptible de donner des performances médiocres.

6.2.2 Descente de gradient stochastique (SGD)

C'est une approche simple mais très efficace pour l'apprentissage discriminant des classificateurs linéaires sous des fonctions linéaires tels que Support Vecteur Machines et régression logistique. Cette approche d'apprentissage est apparue il y a longtemps. Elle a récemment été mis au devant de la scène dans le cadre de l'apprentissage à grande échelle. SGD a été appliquée avec succès à des problèmes de grande envergure concernant l'apprentissage machine récurrent dans la classification de texte et le traitement du langage naturel.

Les avantages Les avantages de la descente de gradient stochastique sont :

- Efficacité.
- Facilité de mise en œuvre (beaucoup de possibilités pour le code tuning).

Les inconvénients Les inconvénients de la descente de gradient stochastique :

- SGD nécessite un certain nombre des paramètres tels que le paramètre de régularisation, et le nombre d'itérations.
- SGD est sensible à la fonction mise à l'échelle.

6.2.3 Arbres de Décision

Les arbres de décision (DTS) sont une méthode d'apprentissage supervisé non-paramétrique utilisée pour la classification et la régression. Les arbres de décision présentent l'avantage de construire une classification facilement compréhensible pour l'utilisateur car elle se présente sous la forme d'un arbre de règles qui sont assimilables un ensemble de conditions "Si ... Alors ... Sinon" imbrique". Le but est de créer un modèle qui prédit la valeur d'une variable cible par l'apprentissage des règles de décision simples inférées à partir des caractéristiques de données.

Avantages Quelques avantages des arbres de décision :

- Simple à comprendre et à interpréter. Les arbres peuvent être visualisés.
- Nécessite peu de préparation de données. D'autres techniques ont souvent besoin de normalisation de données, de variables indicatrices doivent être créées et des valeurs vides doivent être enlevées. Notons cependant que ce module ne supporte pas les valeurs manquantes.
- Capable de gérer à la fois des données quantitatives et qualitatives. D'autres techniques sont généralement spécialisées dans l'analyse d'ensembles de données qui ont un seul type de variable. Voir algorithmes pour plus d'informations.
- Capable de traiter des problèmes multi-sorties.
- Utilise un modèle de boîte blanche. Si une situation donnée est observable dans un modèle, l'explication de la condition s'explique facilement par la logique booléenne. En revanche, dans un modèle de boîte noire (par exemple, dans un réseau neuronal artificiel), les résultats peuvent être plus difficiles à interpréter.

Inconvénients Les inconvénients des arbres de décision sont:

- Les arbres de décision apprenants peuvent créer des arbres plus complexes que ceux généralisés avec les données. C'est ce qu'on appelle le sur-apprentissage. Des mécanismes tels que la taille (pas pris en charge actuellement), fixant le nombre minimum d'échantillons requis à un nœud feuille ou réglage de la profondeur maximale de l'arbre sont nécessaires pour éviter ce problème.
- Les arbres de décision peuvent être instables en raison de petites variations dans les données ce qui pourrait conduire à un arbre complètement différent. Ce problème est atténué par l'utilisation des arbres de décision au sein d'un ensemble.

- Le problème de l'apprentissage d'un arbre de décision optimale est connu pour être NP-complet sous plusieurs aspects d'optimalité et même des concepts simples. Par conséquent, les algorithmes pratiques d'apprentissage par arbres de décision sont basés sur des algorithmes heuristiques tels que l'algorithme glouton où les décisions optimales sont effectués localement à chaque nœud. Ces algorithmes ne peuvent pas garantir d'adapter l'arbre de décision optimale à l'échelle mondiale. Ceci peut être atténué par la formation de plusieurs arbres dans un apprenant ensemble, où les caractéristiques et les échantillons sont prélevés au hasard avec remise.

- Il y a des concepts qui sont difficiles à apprendre parce que les arbres de décision ne les expriment pas facilement, comme XOR, parité ou multiplexeur problèmes. Les arbres de décision créent des arbres biaisées si certaines classes dominent. Il est donc recommandé d'équilibrer l'ensemble de données avant le montage de l'arbre de décision.

6.2.4 Random Forest

Avantages Les avantages de la forêt aléatoire sont :

- Il est l'un des algorithmes d'apprentissage les plus précis, il produit un classificateur de haute précision.

- Il fonctionne de manière efficace sur de grandes bases de données.

- Il peut gérer des milliers de variables d'entrée sans supprimer des variable.

- Il donne des estimations sur l'importance des variables en les classant.

- Il donne une estimation pertinente et précise même lorsqu'une partie des données sont manquantes.

- Il existe des méthodes pour corriger une asymétrie de la classe de population due à une ou des erreurs.

Inconvénients Les inconvénients de la forêt aléatoire sont :

- Stockage des arbres en mémoire.

6.3 Résultats de classification

Nous avons choisit deux outils "Scikit Learn"[12] et "Orange canvas"[13] pour implémenter les méthodes de classification.

6.3.1 Scikit Learn

C'est développé par David Cournapeau. C'est un open source d'apprentissage basé sur des bibliothèques créées par le langage de programmation Python. Il dispose de diverses

classifications, régression et regroupement des algorithmes, y compris les machines à vecteurs de support , régression logistique, naïf de Bayes, forêts aléatoires et K-means. Il est conçu pour interagir avec les bibliothèques numériques et scientifiques Python Numpy et SciPy. Scikit définie une bibliothèque qui fournit une variété de techniques d'apprentissage à la fois supervisées ou non.

6.3.2 Orange Canvas Learn

C'est une base de composants de data mining et une suite apprentissage automatique, avec une programmation visuelle pour effectuer l'analyse exploratoire des donnes et de visualisation, liaisons Python et des bibliothèques pour les scripts. Il comprend un ensemble de composants pour le prétraitement des données, fonctionnalité notation et de filtrage, la modélisation, l'évaluation des modèles et des techniques d'exploration. Il est mis en œuvre en C++ et Python. Son interface utilisateur graphique s'appuie sur la multi-plateforme Qt. Orange est distribué gratuitement sous la licence GPL . Il est maintenu et développé au Laboratoire de bio-informatique de la Faculté d'informatique et de sciences de l'information à l'université de Ljubljana, Slovénie. Au niveau de l'environnement informatique, orange est prise en charge sur les différentes versions de Linux, Apple Mac OS X, et Microsoft Windows.

6.3.3 Indices d'évaluation

Dans cette section, nous allons présenter trois types de mesure qui nous permettent d'évaluer chaque classificateur de la liste, nous allons découvrir lequel parmi nos classificateurs fonctionne le mieux avec les types de nos données. Nous calculons l'indice (rappel), ensuite l'indice (précision) et enfin (F-mesure).

- **Rappel** Cet indice calcule la moyenne de classification, c'est-à-dire calcule le nombre total de textes bien classés dans sa catégorie divisé par le nombre total de textes de cette catégorie, donc le calcul se fait de la manière suivante :

$$Rappel = \frac{Nombre\ total\ de\ textes\ bien\ classes\ dans\ sa\ categorie}{Nombre\ total\ de\ textes\ appartient\ dans\ cette\ categorie}$$

- **Précision** Cet indice présentant des similitude avec l'indice précédent, calcule la moyenne de classification, entre le nombre total de textes bien classés dans sa catégorie et le nombre total de textes classé dans cette catégorie, donc le calcul se fait de la manière suivante :

$$Precision = \frac{Nombre\ total\ de\ textes\ bien\ classes\ dans\ sa\ categorie}{Nombre\ total\ de\ textes\ classes\ dans\ cette\ categorie}$$

- **F-mesure** Afin de calculer la combinaison entre les deux indices, nous obtenons l'indice F-mesure :

$$F - mesure = \frac{2 \times Rappel \times Precision}{Rappel + Precision}$$

6.3.4 Résultat

Avant de présenter le résultat, nous rappelons qu'Orange Canvas nous offre une interface graphique (cf.figure 31). Cette interface permet de faciliter le travail, en offrant l'utilisateur la possibilité de construire de manière interactive les chaînes de traitement de donnes.

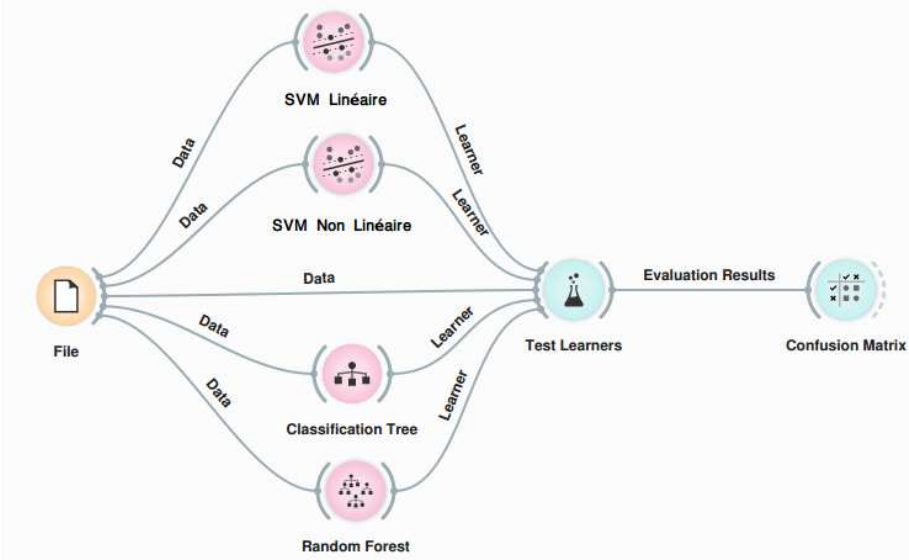


Figure 31: Représentation des classificateurs sous Orange Canvas

Pour évaluer la pertinence de nos différentes méthodes, nous avons choisi d'utiliser les quatre classificateurs ci-dessus pour la classification supervisée, le choix s'est porté sur ces quatre types de classificateur dans le but de trouver l'algorithme d'apprentissage supervisé qui donne le plus souvent les meilleurs résultats pour la classification des textes.

Les résultats de la classification sont présentés en trois termes: Rappel, précision, et F-messure. Notons, que la précision est égal 100% nous pouvons dire dans ce cas que tous les textes sont classées dans la bonne catégorie. Pour atteindre cet objectif, notre mesure est calculée entre deux sous-ensembles : 70% du corpus (du matrice) est utilisé pour l'apprentissage et 30% pour la prédiction (cf.figure 32).

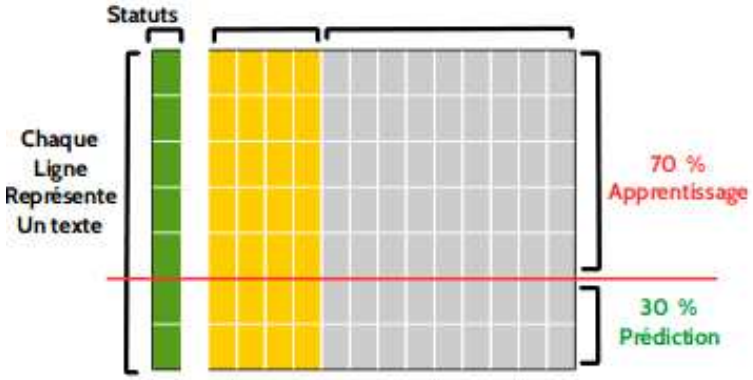


Figure 32: Matrice de classification

Pour aller plus loin, nous avons décidé de faire notre test de classification sur deux niveaux. Dans le premier test, que nous appellerons “avec Bag of Words”, l’apprentissage se fait en prenant en compte les indicateurs ainsi que les mots composant les documents. Dans le second test, que nous appellerons “sans Bag of Words”, nous n’utilisons que les indicateurs pour décrire les documents. Les résultats de ces tests sont respectivement présents dans les tableaux 20 et 21 de ce rapport.

	Orange Canvas Learn			Scikit Learn		
	Rappel	Précision	F-mesure	Rappel	Précision	F-mesure
SVM Non Linéaire	0.71	0.73	0.72	0.71	0.73	0.72
SVM Linéaire	0.72	0.73	0.73	0.73	0.73	0.74
Arbre de décisions	0.71	0.70	0.70	0.7	0.72	0.71
Random Forest	0.88	0.66	0.76	0.86	0.67	0.75

Table 20: Classification avec la présence de Bag Of Words

Pour le test effectué avec la représentation des mots, nous pouvons constater, comme indiqué dans le tableau 16, que l’outil Scikit Learn affiche de meilleurs résultats sur notre corpus avec la méthode Random Forest, affichant un taux de précision de 67% , un taux de rappel de 86% et un taux de F-mesure de 0.75% et mêmes résultats avec Orange canvas, la méthode Random Forest a des meilleurs résultats : précision 66% , rappel 88% et F-mesure 0.76%.

Les deux outils affichent des résultats similaires voire identiques pour les quatre méthodes. Nous pouvons donc dire que la méthode Random Forest est la plus efficace car présente de meilleurs taux sur les deux outils.

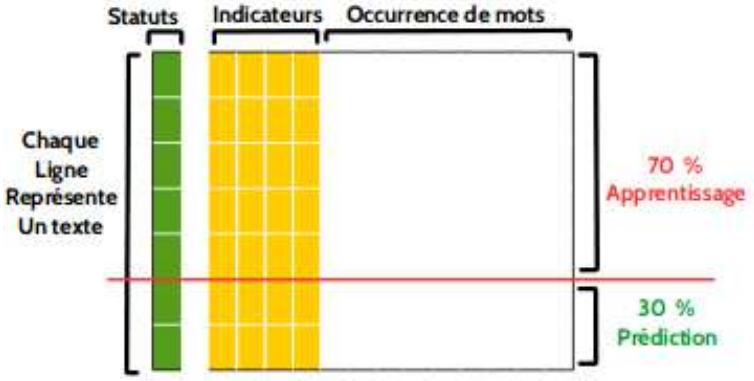


Figure 33: Classification avec l'absence de représentation de vocabulaire

	Orange Canvas Learn			Scikit Learn		
	Rappel	Précision	F-mesure	Rappel	Précision	F-mesure
SVM Non Linéaire	0.58	0.66	0.62	0.59	0.65	0.61
SVM Linéaire	0.67	0.66	0.61	0.60	0.63	0.61
Arbre de décisions	0.62	0.61	0.61	0.61	0.64	0.62
Random Forest	0.86	0.63	0.73	0.79	0.64	0.71

Table 21: Classification avec l'absence de Bag Of Words

Nous passons au deuxième test, en supprimant la représentation de vocabulaire et en ne prenant en compte que les valeurs des indicateurs (cf. figure 33). En s'appuyant sur le tableau 17, nous remarquons, d'un côté, une baisse des valeurs et d'un autre côté, Random Forest reste la plus efficace. La présence de la représentation du vocabulaire agit donc sur la qualité de classification avec toutes les méthodes de classification. Cela dit, les deux outils, que ce soit Orange Canvas Learn ou Scikit Learn montrent que la méthode Random Forest devance les autres classificateurs avec une faible marge. Il présente un taux de précision de 63%, un taux de rappel de 86% et un taux de F-mesure 73%.

Nous pouvons donc dire que la méthode la mieux adaptée pour la classification avec et sans représentation de vocabulaire est Random Forest. Cela est dû à son efficacité avec les grandes bases de données, il peut gérer des milliers des variables d'entrée sans suppression.

Dans le but d'améliorer nos résultats de classification, nous avons décidé d'implémenter la méthode SGD (section 6.2.2) à l'aide de Scikit Learn sous la fonction linéaire SVM. L'implémentation se fait sur les deux niveaux (avec et sans la représentation de vocabulaire).

	SGD : SVM Non Linéaire		
	Rappel	Précision	F-mesure
Avec représentation de vocabulaire	0.87	0.85	0.84
Sans représentation de vocabulaire	0.85	0.83	0.83

Table 22: Implémenter la méthode SGD

Avec ou sans la représentation des mots nous remarquons une augmentation des valeurs. Avec la représentation des mots la méthode DSG, affichant un taux de précision de 85% , un taux de rappel de 87% et un taux de F-mesure de 84%. Sans la représentation de vocabulaire, nous remarquons une stabilité avec DSG. La présence de la représentation du vocabulaire n'agit donc pas trop sur la qualité de classification avec ce type de classification.

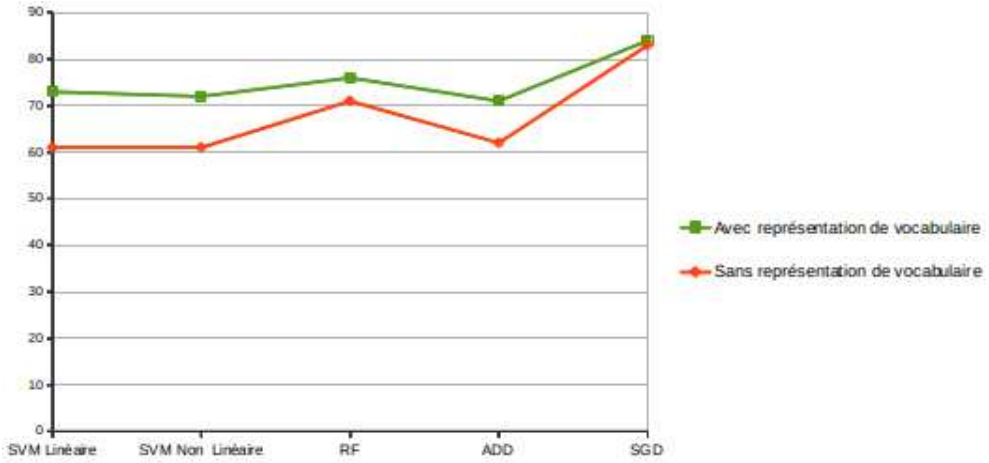


Figure 34: Comparant la classification avec et sans représentation vocabulaire

Comme le montre la figure 34 les meilleurs résultats sont obtenus en appliquant l'algorithme d'optimisation DSG sur un SVM. Le point intéressant est que les résultats deviennent très similaires que l'on utilise ou non les mots des documents. D'une part cela montre que les indicateurs que nous avons définis sont plutôt pertinents pour juger de la qualité d'un texte. D'autre part, s'affranchir du vocabulaire nous permet de réduire très sensiblement la taille des exemples d'apprentissage (et donc le temps de calcul) ; en outre, cette indépendance vis à vis du vocabulaire permet sans doute de traiter une plus large diversité de textes.

7 Conclusion

Bilan d'étude

Repérer les textes de bonne qualité constitue un enjeu majeur pour les maisons d'édition. Dans notre travail, nous avons procédé à la mise en place d'un système qui traite une problématique de Short-Édition visant à la mise en avant de textes intéressants pour la lecture, voire publication. Au premier abord, nous sommes partis d'une grande base de données comprenant 20k des nouvelles. La publication des textes (+1 : publié, -1 : non publié) se fait sur la base des évaluations de comité éditorial (notées entre 1 et 5 avec commentaire).

Comme préparatif, nous avons commencé par l'exploration des différents étapes de prétraitement utilisés en TAL dans le cadre d'extraction d'informations à partir de textes. Ensuite, nous avons commencé notre travail en fixant les indicateurs de qualité à chercher dans les textes en se basant sur des mesures lexicales (vocabulaire, entropie, ...), grammaticales ainsi que des indicateurs de complexité et de fautes d'écriture.

Pour évaluer nos choix d'indicateurs afin de connaître lesquels pouvaient être des indices de classification, d'un côté, nous sommes passés par une représentation graphique de chaque indicateur. Cette représentation se fait sur la base des valeurs d'indicateurs par rapport au statut des textes (soit +1, soit -1). Et d'autre coté, nous sommes appuyés sur le coefficient de corrélation bisériale-point pour déterminer le taux de corrélation entre chaque indicateur et la qualité des textes. Nous avons utilisé cette méthode parce que nous sommes en présence des valeur binaires (+1, -1).

Pour faire une distinction entre les textes bien écrits et les textes de moins bonne qualité, nous avons exploré dans notre projet de recherche plusieurs techniques de classification binaire telles que les Support Vector Machines (SVM) en présentant ses capacités de séparation linéaire et non-linéaire entre classes en grandes dimensions. Dans un second temps, nous sommes passés par les arbres de décision qui sont plus adaptés à la visualisation tout en gardant une puissance de classification Bayesienne. Ainsi, nous avons implémenté sur nos données l'algorithme de classification Random forest qui nous a donné des meilleurs résultats, et nous avons fini par l'algorithme Descente de gradient stochastique (SGD) implémenté sous la fonction linéaire SVM. Pour aller, un petit peu plus loin, nous avons joué avec la représentation vectorielle de données, cet enjeu nous permet d'avoir des résultats différents.

Ce travail pourrait être encore amélioré sur plusieurs niveaux. Tout d'abord, au niveau de prétraitement des données, l'étape de segmentation peut être améliorée afin de détecter les mots composés (Polylexicaux). Au niveau des indicateurs, la liste des mots outils et Gougenheim peuvent être enrichir. Ainsi, à ce niveau nous pouvons mesurer les autres indicateurs qui peuvent influer sur la qualité de classification tels que

le nombre moyen de personnes singuliers, l'occurrence être/avoir (au présent, au passé, au future, etc) nombre moyen de termes de comparaison métaphores (comme, tel ...) etc.

Bilan personnel

Ce travail de recherche s'inscrit à la problématique de Short-Édition, il s'est déroulé au sein du laboratoire de recherche (LIG) et de la société Short-Édition. Il intègre une partie professionnelle qui m'a permis d'apprendre de nouvelles techniques et connaissances d'un coté, dans le domaine Traitement du Langage Naturel : statistiques textuelles, prétraitement des données et de l'autre cote, dans le domaine apprentissage de machine : les représentations des données textuelles, les types d'algorithme d'apprentissage, des méthodes et des techniques de classification. Pendant ce mémoire de recherche, j'ai pu avoir un contact direct tant avec le monde professionnel qu'avec des chercheurs de haut niveau et des ingénieurs spécialisés dans Traitement du Langage Naturel. Finalement, cette expérience très enrichissante, elle me pousse de poursuivre ma carrière dans ce domaine.

References

[1] Guy Serraf, *Textes faciles et difficiles essai d'une formulation de critères quantitatifs*,1964.

[2] Thierry Trubert-Ouvrard, *la place de l'adjectif en français, méthodologie micrométrique*, 2002.

[3] Hyeran Lee, Philippe Gambette, Elsa Maillé, Constance Thuillier, *Densidées, calcul automatique de la densité des idées dans un corpus oral*, 2010.

[4] Yves Bestgen, *Évaluation automatique de textes et cohésion lexicale*, 2012.

[5] Robert Gunning, *The Technique of clear Writing (New ' York, Me Graw-Hill*, 1968.

[6] Rudolf Flesch, *he Art of Plain Talk (New York, Harper and Row)*, 1946.

[7] Rudolf Flesch, *How to test Readibility (New York, Harper and Row)*,1949.

[8] JURAFSKY D., MARTIN J. H, *Speech and Language Processing. Upper Saddle River : Prentice Hall*, 2000.

[9] KERNIGHAN M., CHURCH K., GALE W, *A Spelling Correction Program Based on a Noisy Channel Model. Proceedings of COLING*, 1990.

[10] Naber, Daniel, *A Rule-Based Style and Grammar Checker. Diplo- marbeit, Technische Fakultät, Universität Bielefeld, Bielefeld*, 2003.

[11] Tufis, Dan et Mason, Oliver, *Tagging Romanian Texts: a Case Study for QTAG, a Language Independent Probabilistic Tagger. In Proceedings of the First International Conference on Language Resources and Evaluation(LREC), pp.589-596,Granada spain*, 1998.

[12] Carlberger, Johan, Domeij, Rickard, Kann, Viggo et Knutsson, Ola, *A Swedish grammar checker. Submitted for Association for Com- putational Linguistics*, 2002.

[13] Knutsson, Ola, Bigert, Johnny et Kann, Viggo, *A robust shallow parser for Swedish. In Proceedings of the 14 th Nordic Conference on Computational Linguistics (NoDaLiDa), Reykjavik, Iceland.*, 2003a.

[14] Knutsson, Ola, Cerrato Pargman, Teresa et Severinson Eklundh, Kerstin, *Transforming Grammar Checking Technology into a Learning Environment for Second Language Writing. In Burstein, Jill et Leacock, Claudia (Eds.). Proceedings of the HLT-NAACL 03 Workshop on Building Educational Applications Using Natural Language Processing, pp. 38–45.*, 2003b.

[15] Knutsson, Ola, Cerratto Pargman, Teresa, Severinson Eklundh, Kerstin et Westlund, Stefan, *Designing and developing a language environment for second language writers. Computers and Education, 49(4):1122–1146*, 2007.

[16] Grefenstette G, *Comparing Two Language Identification Schemes. In Proceedings of the 3rd International Conference on the Statistical Analysis of Textual Data (JADT'95),Rome, Italy*, 1995.

[17] Dunning T *Statistical Identification of Languages. Technical Report MCCS 94-273, Computing Research Laboratory*, 1994.

[18] Sahami M , *Using Machine Learning to Improve Information Access . PhD thesis, Computer Science Department, Stanford University*, 1999.

[19] D. S. McNamara, S. Dennis, W. Kintsc *Handbook of Latent Semantic Analysis. Mahwah,Lawrence Erlbaum Associates*, 2007.

[20] Wajdi Zaghouani, *AUTO-ÉVAL : vers un modèle d'évaluation automatique d'un texte In proceedings of the CESLA colloquim, Montréal, UQAM*, 2002.

[21] Page, Ellis Batten, *Computer grading of student prose, using modern concepts and software. Journal of experimental education 62:127-142*, 1994.

[22] Hearst, A. Marti, *The debate on automated essay grading. IEEE Intelligent systems 15(5):22-37*, 1994.

[23] Burstein, Jill, Karen Kukich, Susanne Wolff, Chi Lu et Martin Chodorow *Computer analysis of Essays. Texte d'une conférence présentée au NCME Symposium on Automated Scoring, Princeton University*, 1998.

[24] Landauer , T.K. et al, *Handbook of Latent Semantic Analysis. Mahwah (N.J.) : L. Erlbaum*, 2007.

[25] Landauer , T.K., Foltz , P.W. et Laham , D, *An Introduction to Latent Semantic Analysis. Discourse Processes 25 (2-3) : 259-284*, 1998.

[26] Mesnager, Jean, *Lisibilité des textes pour enfants : un nouvel outil?, Communication et langages, 79, p.1838*, 1989.

[27] Gerard Salton et C.S. Yang, *On the Specification of Term Values in Automatic Indexing*, 1973.

[28] G. Salton, C. S. Yang, et C. T. Yu, *LA theory of term importance in automatic text analysis. Journal of the American Society for Information Science, pp 33–44*, 1975.

[29] G. K. Zipf, *National unity and disunity. The nation as a bio-social organism*, 1941.

[30] C . Cortes and V. Vapnik, *k, Support vector networks. Machine Learning*,1995.

[31] Corman Julien, *Extraction d'expressions polylexicales sur corpus arboré*,2012.

[32] Radwan Jalam , Jean-Hugues Chauchat, *Pourquoi les n-grammes permettent de classer des textes ? Recherche de mots-clefs pertinents à l'aide des n-grammes caractéristiques*,2002.

[33] Nicolas *Extraction et regroupement de descripteurs morphosyntaxiques pour des processus de Fouille de Textes*,2009.

[34] Nicolas *Amélioration de l'étiquetage morphosyntaxique et de la détection des entités nommées sur des tweets dans le domaine politique*,2012.

Webographie

[1] http://short-edition.com

[2] http://blog.onyme.com/quelques-notions-en-analyse-syntaxique.htm

[3] http://www.cairn.info/zen.php?ID_ARTICLE=LANG_171_0095

[4] http://taln09.blogspot.fr/2009/02/etiquetage-morpho-syntaxique-et.html

[5] http://www.nltk.org

[6] http://www.synapse-fr.com/Cordial Analyseur.html

[7] http://www.tal.univ-paris3.fr/travaux-etudiants.htm

[8] http://www.zinfosweb.fr/2014/01/langage-tool-correcteur-grammatical.html

[9] htm http://www.ling.uqam.ca/sato/outils/sato.htm

[10] http://en.wikipedia.org/wiki/Point-biserial_correlation_coefficient

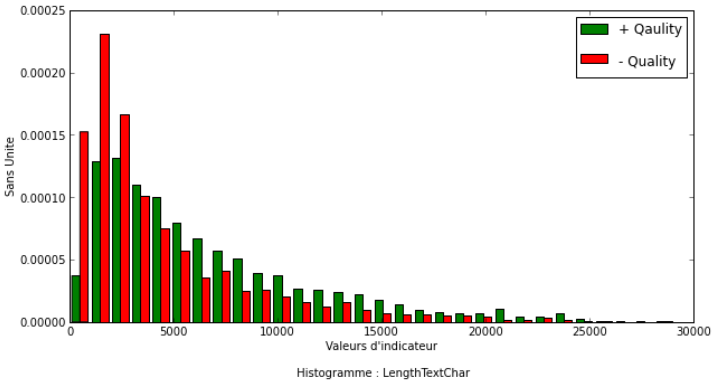
[11] fr.wikipedia.org/wiki/TF-IDF

[12] http://scikit-learn.org/stable/supervised_learning.html supervised-learning

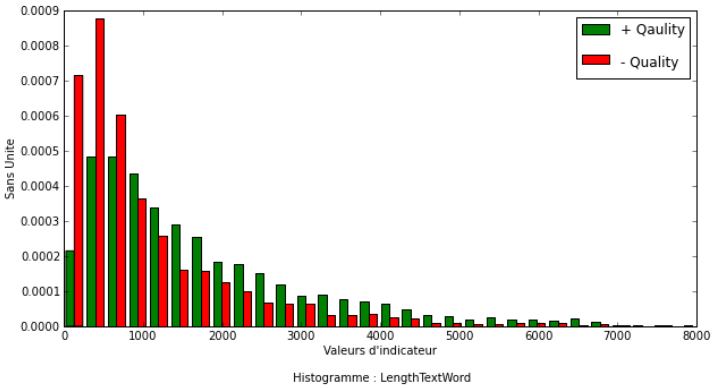
[13] http://orange.biolab.si/

8 Annexe

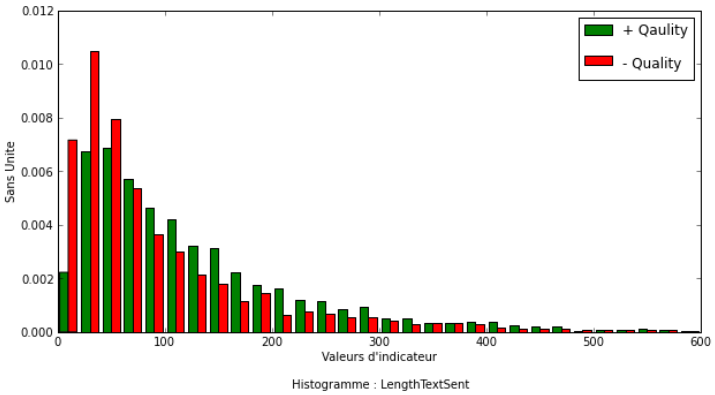
Nombre de caractères par texte



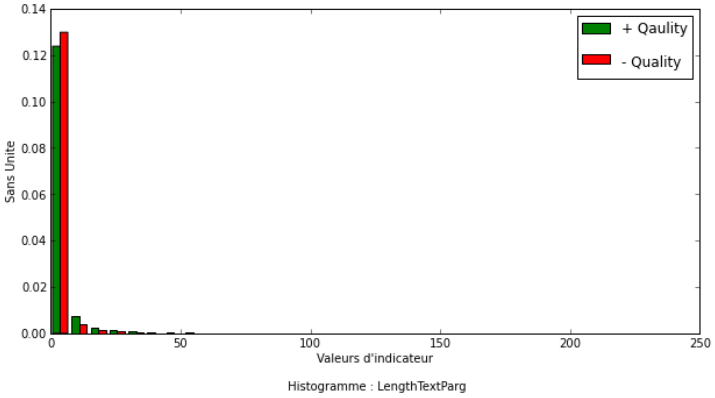
Nombre de mots par texte



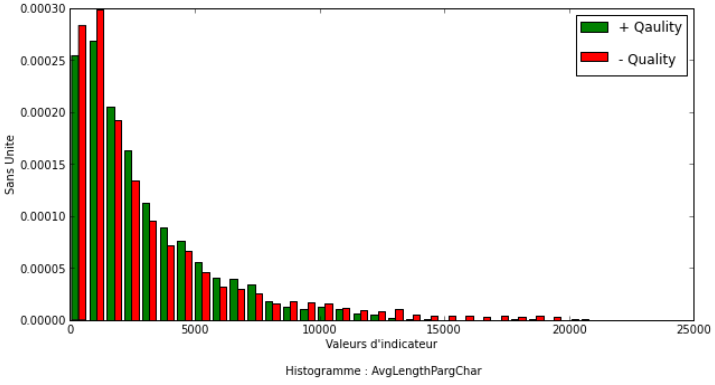
Nombre de phrases par texte



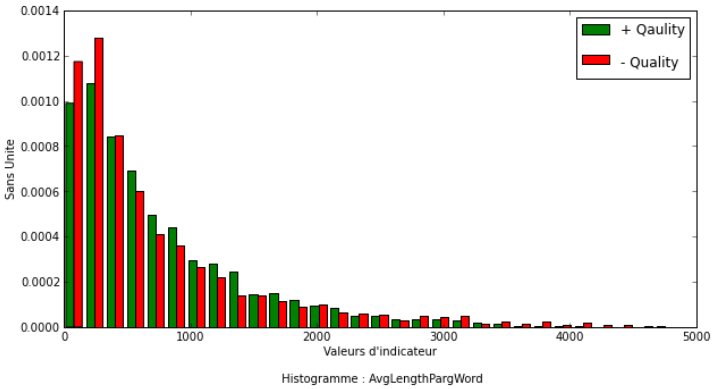
Nombre de paragraphes par texte



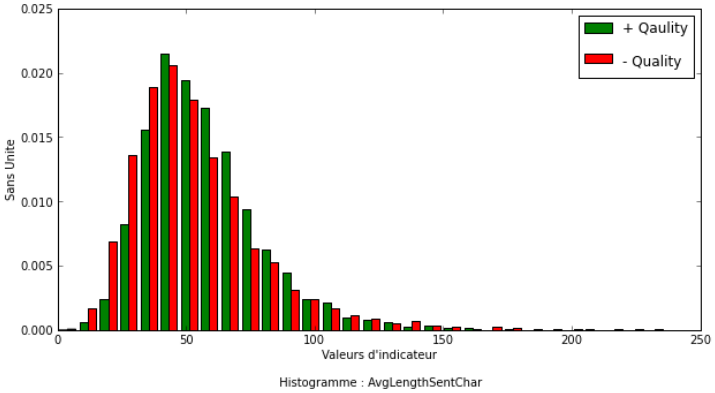
Nombre de caractères par paragraphes



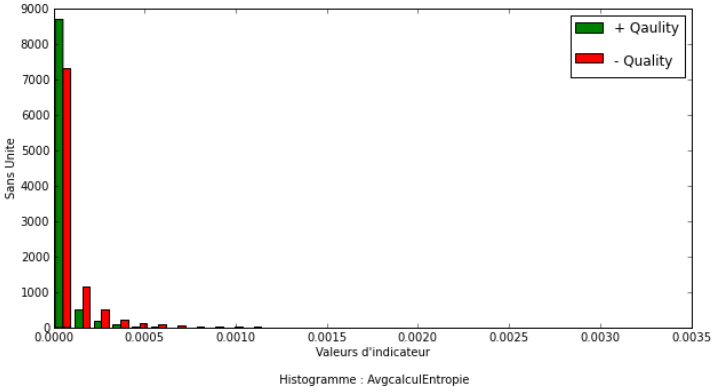
Nombre de mots par paragraphes



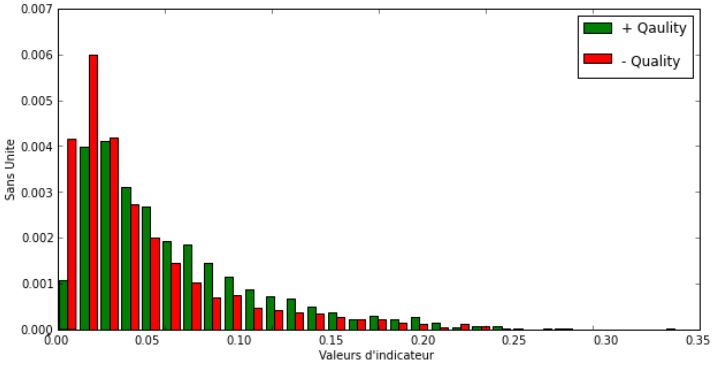
Nombre de caractères par phrase



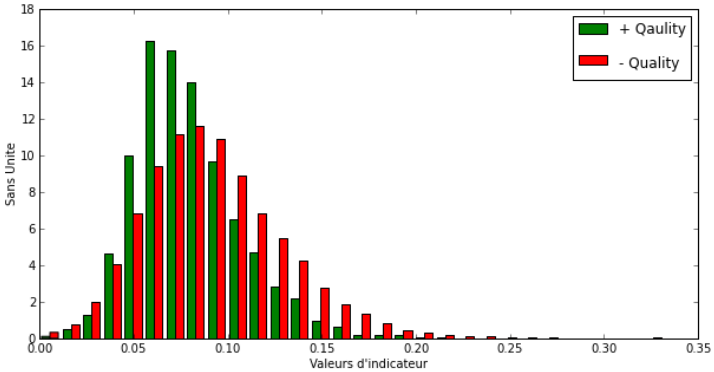
Entropie



Cohésion

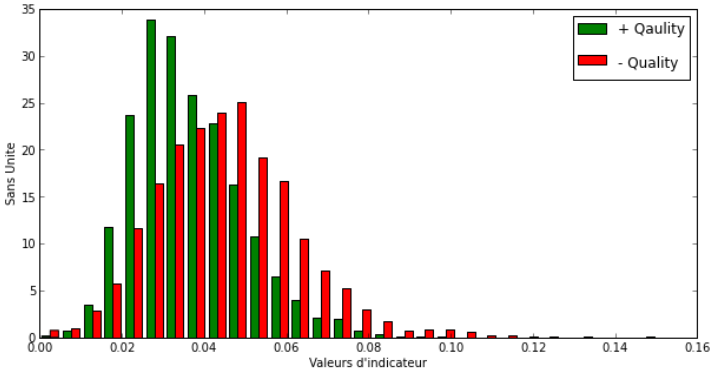


Fréquence relative des Pronoms



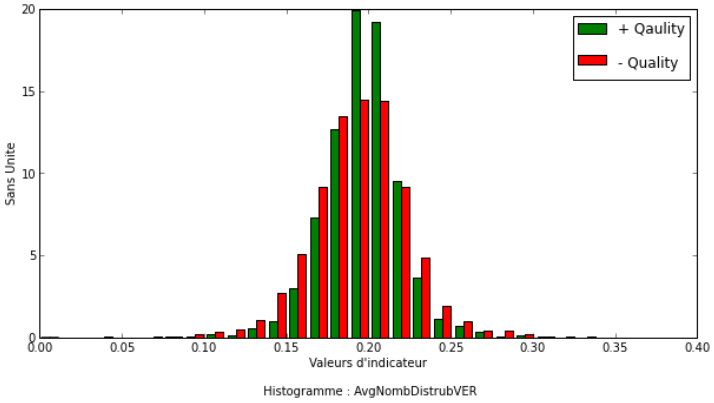
Histogramme : AvgNombDistPrPers

Fréquence relative de conjonctions

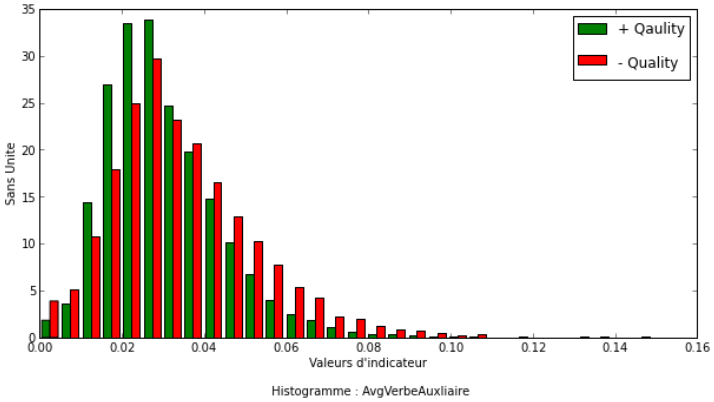


Histogramme : AvgNombDistrubCON

Fréquence relative de verbes



Fréquence relative de verbes auxiliaires



List of Figures

1 Cadre général du projet 10
2 Représentation graphique du P 17
3 Matrice des occurrences X 19
4 Maquette de décomposition de la matrice par des valeurs singulières . . . 19
5 Fixer la valeur de K 20
6 Réduire la matrice diagonale 20
7 Projection P sur le sous-espace sémantique Latent 21
8 Dernière étape de cohésion 21
9 Arbre d'analyse avec NLTK 23
10 Comparaison TreeTagger et Cordial Analyseur 24
11 Taux de textes publiés et non publiés 28
12 Mécanisme d'évaluation 28
13 Exemple d'évaluation de comité 29
14 Taux de textes publiés et non publiés 29
15 Chaîne de traitement par défaut 30
16 Chaîne de pré-traitement 30
17 Nouvelle chaîne de traitement 31
18 Extraction les valeurs des indicateurs 33
19 La longueur moyenne de mots 52
20 Distribution de densité Vocabulaire 53
21 Distribution des mots différents 53
22 Distribution des adjectifs 54
23 Distribution des déterminants 55
24 Distribution des noms 56
25 Distribution des prépositions 56
26 Distribution des adverbes 57
27 Histogramme de lisibilité 58
28 Distribution des fautes d'écriture 59
29 Représentation vectorielle des textes. 67
30 Principe de SVM. 69
31 Représentation des classificateurs sous Orange Canvas 73
32 Matrice de classification 74
33 Classification avec l'absence de représentation de vocabulaire 75
34 Comparant la classification avec et sans représentation vocabulaire 76

List of Tables

1 Table de référence de Flesch 13
2 Gunning Fog Type d'écrit 14
3 Opérations complexes sur les expressions régulières. 15
4 Exemple de résultat de Treetageer 22
5 Critères d'évaluation 26
6 Normalisation des marques de ponctuations inconnus par Treetagger . . . 32
7 Les règles d'espacement des Signes et de la Ponctuation 32
8 Exemple de résultat des indicateurs généraux 39
9 Exemple de résultat des indicateurs lexicaux 42
10 Exemple de résultat des indicateurs grammaticaux 45
11 Exemple de résultat des indicateurs de complexité 48
12 Exemple de résultat Fautes d'écriture 49
13 Valeurs de corrélation 60
14 Valeurs de corrélation entre la qualité et chaque indicateur 61
15 Exemple de représentation binaire 63
16 Exemple de représentation fréquentielle 63
17 Exemple de représentation TF-IDF 65
18 Échantillon de fichier vocabulaire 66
19 Types classificateurs 68
20 Classification avec la présence de Bag Of Words 74
21 Classification avec l'absence de Bag Of Words 75
22 Implémenter la méthode SGD 76

Résumé

Le domaine du traitement automatique des langues naturelles a connu des évolutions très rapides ces dernières années, et spécialement les méthodes de statistique textuelle. Elles ont été mises en lumière par plusieurs disciplines : l'étude des textes, la linguistique, l'analyse du discours, la statistique, l'informatique, le traitement des enquêtes.

Ce projet de recherche s'inscrit dans le cadre du problématique de Short Édition qui concerne l'éditeur communautaire de littérature courte. L'objectif est d'assister le travail du comité de lecture en effectuant une première catégorisation des textes. Notre travail implique la conception et la mise en œuvre d'un prototype permettant de repérer les textes qui présentent les caractéristiques d'un texte de qualité et de trouver une méthode de classification en nous fondant sur les principes de la fouille de données permettant de bien classer nos textes.

MOTS-CLÉS : Traitement Automatique de la Langue - Évaluation automatique de textes - Classification - Apprentissage.

Abstract

The field of natural language processing has witnessed very rapid developments in recent years, particularly with respect to methods used for statistical text analysis. These methods have been brought into focus by several disciplines in particular : the study of texts, linguistics, discours analysis, statistics, computer sciences, and survey processing.

This research project develops within the framework of an issue that concerns the publishing company Short Editions. It relies on contributions in a field that employs a vast variety of designations (lexical statistics, statistical linguistics, quantitative linguistics, etcetera). Our work involves the creation of a prototype that allows for the identification of texts that present the characteristics of a quality text and to find appropriate methods of classification of these texts based on data and text mining principles.

KEYWORDS : Automatic Language Processing - Automatic evaluation of texts - Classification - Learning.