

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE MINISTERE DE L'ENSEIGNEMENT  
SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE  
UNIVERSITE SAÂD DAHLAB BLIDA  
FACULTE DES SCIENCES  
DEPARTEMENT D'INFORMATIQUE



Mémoire présenté pour l'obtention du diplôme de

Master

en

INFORMATIQUE

Option : Ingénierie des logicielles

**Classification des documents médicaux  
basée sur le Text Mining**

Présenté par

**Dahmani Houria**

Soutenu en 2012 devant la commission du jury composée de :

Président (e) M.....UNIVERSITE DE BLIDA

Examineur M .....UNIVERSITE DE BLIDA

Examineur M .....UNIVERSITE DE BLIDA

Promoteur Mme A.ELMAOUHEB, Chef du département réseaux, CERIST

A mon mari Ali

A mes enfants Samy, Maria et Ilyas

A mes parents et beaux parents

A mes sœurs et mes frères

A Guilen

A tous mes collègues du Cerist

A toutes les personnes qui m'aiment

**Qu'ils trouvent ici l'expression de ma sincère gratitude**

## Remerciements

Je remercie ma structure le CERIST, qui m'a permis de rejoindre les bancs de l'université et d'aborder des perspectives nouvelles.

Je remercie Mme ELMAOUHAB Aouaouche d'avoir accepté la lourde charge de promoteur, et d'accorder de son temps pour suivre mon travail.

Je souhaite également adresser mes remerciements à l'ensemble des membres du jury, d'avoir accepté de lire, de juger et de discuter mon travail.

Je remercie également Mr BALA, Mr Ben Nouar et Melle Salhi, pour leur précieuse assistance, pendant le cursus du Master.

Un grand merci pour Insaf, Bdiaa, pour leurs précieuses orientations, qui m'ont permis d'affiner mon travail.

Mes remerciements vont aussi à toute l'administration de la faculté informatique, pour leur gentillesse.

Mes remerciements les plus distingués pour mon mari, Ali, pour son assistance morale et ses encouragements, pour son aide précieuse pendant toute ma formation.

Je remercie également mes chères collègues du Cerist, pour leur soutien moral, spécialement : Amal, Ikram, Hassina, Kahina, Rabab, Nassima et Rym.

Que toutes les personnes qui ont attribué de près ou de loin à l'élaboration de ce travail trouvent ici l'expression de ma haute gratitude.

## Résumé/ Abstract/ ملخص

Avec l'avènement de l'informatique et l'explosion de nombre de documents stockés sur les supports électroniques et sur le web, qui sont à plus de 80% de type texte, l'utilisation de technologie facilitant leur traitement et leur analyse est devenu indispensable, pour aider les utilisateurs de ces masses de données à les explorer puis à les organiser.

Ainsi, le Text Mining et précisément la classification automatique de textes, qui consiste à assigner un document à une ou plusieurs catégories, s'impose de plus en plus comme une technologie clé, les résultats obtenus sont utiles aussi bien pour la recherche d'information que pour l'extraction de connaissance aussi bien sur internet (moteurs de recherche), qu'au sein des entreprises (classement de documents internes, dépêches d'agences, etc.).

A l'égard des différentes approches de classification automatique de textes, décrites dans l'état de l'art, nous avons utilisé l'approche non supervisée (algorithme Kmeans) pour étiqueter nos documents et l'approche supervisée (algorithme Naive Bayes) pour classer les nouveaux documents.

L'objectif principal de notre travail, est d'offrir un modèle fiable de classification de documents médicaux.

Nous utilisons MEDLINE comme corpus de textes, sur lequel nous menons nos expérimentations.

**Mots Clés** : Catégorisation, clustering, Classification, Texte, Apprentissage, Text Mining, Evaluation, Kmeans, Naïve Bayes, MEDLINE.

With the advent of computers and the explosion of the number of documents stored on electronic media and on the web, which are more than 80% of type text, the use of technology to facilitate their processing and analysis has become essential to help users to explore the masses data and to organize them.

Thus, Text Mining and precisely automatic text classification, which consists in assigning a document to one or more categories, is becoming increasingly recognized as a key technology, the results are also useful for finding information in knowledge extraction or on the Internet (search engines) and within companies (ranking internal documents, news agencies, etc.).

With regard to the different approaches of automatic text classification, described in the prior art, we used an unsupervised approach (algorithm Kmeans) to label our documents and supervised approach (Naive Bayes algorithm) for classifying new documents.

The main objective of our work is to provide a reliable classification of medical documents.

We use MEDLINE as text corpus, in which we conduct our experiments.

**Keywords:** categorization, clustering, classification, Text, Learning, Text Mining, Evaluation, Kmeans, Naïve Bayes, MEDLINE.

مع ظهور الحواسيب وانفجار عدد الوثائق المخزنة على وسائط الإعلام الإلكترونية وعلى شبكة الإنترنت، والتي هي أكثر من 80% على شكل نص، استخدام التكنولوجيا لتسهيل المعالجة والتحليل أصبح ضروري لمساعدة المستخدمين على استكشاف البيانات وتنظيمها.

وهكذا، تحليل النص و التصنيف الدقيق له، والذي يعود الى اسناده الى صنف واحد ، أصبحت من التكنولوجيات الرئيسية. كذلك فإن النتائج هي أيضا مفيدة للعثور على معلومات جيدة لاستخراج المعرفة خاصة على شبكة الإنترنت (محركات البحث) وداخل الشركات (ترتيب الوثائق الداخلية ، وكالات الأنباء، وما إلى ذلك) .

وفيما يتعلق بالأساليب التلقائية لتصنيف النص فقد استخدمنا نهج غير خاضع للرقابة (خوارزمية ك مينز) لتسمية وثائقنا الاولى و نهج خاضع للرقابة (خوارزمية بايز الساذج) لتصنيف الوثائق الجديدة. الهدف الرئيسي من عملنا هو توفير تصنيف موثوق للوثائق الطيبة.

استخدمنا نصوص ميدلاين لإجراء تجاربنا.

الكلمات الرئيسية: التصنيف، التجميع، النص، التعلم، تحليل النص ، التقييم، ك مينز، بايز الساذج، ميدلاين

## Tableaux et Figures

### Liste des tableaux

- p.9-Table: *Les phases du CRISP-DM.*
- p.41-Tableau 2 : *Matrice de contingence de la classe Ci*
- p.64-Tableau 3 : *Représentation des classes d'objets.*
- p.77-Tableau 4: *Labellisation des clusters.*
- p.79-Table 5 : *Les probabilités à priori.*
- p.78-Table 6 : *Matrice de contingence globale de tout le corpus.*

### Liste des figures

- p.8- Figure 1 : *Les phases de Crisp DM*
- p.11-Figure 2 : *La chaîne de traitement pour le processus de fouille de textes*
- p.17- Figure 3 : *Regroupement d'objets similaires*
- p.19- Figure 4 : *La tâche de classification*
- p.19-Figure 5 : *Démarche de la catégorisation de textes*
- p.21-Figure 6 : *Entraînement d'un système de classification automatique de textes*
- p.22-Figure 7 : *Classification d'un nouveau document*
- p.37-Figure 8 : *Hyperplan avec distance maximal (marge) aux exemples de classes Positives et négatives*
- p.42- Figure 9 : *Agglomération*
- p.49-Figure 10 : *Représentation des trois axes de description d'un système.*
- p.52-Figure 11 : *Diagramme des cas d'utilisation du système.*
- p.53-Figure 12 : *Processus de classification initial.*
- p.54-Figure 13 : *Processus de classification d'une nouvelle notice.*
- p.55-Figure 14 : *Diagramme d'activités pour le cas d'utilisation Préparer les données.*
- p.57-Figure 15 : *Diagramme d'activités pour le cas d'utilisation Regrouper les notices.*
- p.59-Figure 16 : *Diagramme d'activités pour le cas d'utilisation Créer le modèle de classification.*
- p.61-Figure 17 : *Diagramme d'activités pour le cas d'utilisation Classifier une nouvelle notice.*
- p.62-Figure 18 : *Diagramme de classes du système.*
- p.71-Figure 19 : *Exemple d'une notice bibliographique extraite de MEDLINE.*
- p.72-Figure 20 : *Exemple d'une notice simple MEDLINE.*
- p.73-Figure 21 : *Matrice du vocabulaire.*
- p.74-Figure 22 : *L'ACP.*
- p.75-Figure 23 : *Dendrogramme avec CAH.*
- p.76-Figure 24 : *Test de coude.*

*p.77-Figure 25 : La silhouette de Kmeans.*

# Table des matières

## Introduction Générale

<b>1-Contexte</b> .....	<b>2</b>
<b>2-Problématique</b> .....	<b>2</b>
<b>3-Objectifs</b> .....	<b>2</b>
<b>4-Organisation du mémoire</b> .....	<b>3</b>

## Chapitre 1 : Text mining et la classification automatique de textes

### Partie 1 : Text Mining

<b>1-Introduction</b> .....	<b>6</b>
<b>2-Fouille de données (Data Mining)</b> .....	<b>6</b>
2.1-Définitions.....	6
2.2- Processus du Data mining.....	7
<b>3-Fouille de textes (Text Mining)</b> .....	<b>9</b>
3.1 Text Mining versus Data Mining.....	9
3.2- Définitions.....	10
3.3- Approches du Text Mining .....	10
3.3.1-Approche statistique.....	10
3.3.2-Approche Sémantique.....	10
<b>3.4- Chaîne de traitement pour le processus de fouille de données textuelle</b> .....	<b>11</b>
<b>3.5- Text Mining et la classification de textes</b> .....	<b>12</b>
<b>4-Conclusion</b> .....	<b>13</b>

### Partie 2 : Classification automatique de textes

<b>1-Introduction</b> .....	<b>14</b>
<b>2-Pourquoi automatiser la classification ?</b> .....	<b>14</b>
<b>3-Vocabulaire utilisé dans les systèmes de classification</b> .....	<b>16</b>
3.1-Catégorisation (classification Supervisé) .....	16
3.2-Clustering (Regroupement, classification automatique).....	17

<b>4-Avantages et inconvénients.....</b>	<b>18</b>
<b>5-Définition de la catégorisation automatique de textes.....</b>	<b>18</b>
<b>6-Démarche de la catégorisation automatique de textes.....</b>	<b>19</b>
<b>7-Quelques problèmes rencontrés dans la catégorisation automatique de textes.....</b>	<b>22</b>
7.1- Sur-apprentissage.....	22
7.2- L'homographie.....	23
7.3- Polysémie (Ambiguïté).....	23
<b>8-Conclusion.....</b>	<b>23</b>
<b>Chapitre2 : Codage des textes</b>	
<b>1-Introduction.....</b>	<b>25</b>
<b>2-Caractéristique de la donnée textuelle .....</b>	<b>25</b>
<b>3-Prétraitement.....</b>	<b>26</b>
<b>4-Définition des descripteurs.....</b>	<b>27</b>
4.1-Représentation en «sac de mots».....	27
4.2-Représentation des textes par des phrases.....	28
4.3-Représentation des textes par des racines lexicales et des lemmes.....	28
<b>5-Sélection de descripteurs (Réduction).....</b>	<b>29</b>
5.1-Pourquoi réduire?.....	29
5.2-Le nombre de descripteurs conservés.....	29
5.3-Méthodes de sélection de descripteurs.....	30
<b>6-Pondération.....</b>	<b>30</b>
6.1-Formules de pondération.....	31
6.1.1- Term frequency (TF).....	31
6.1.2- Inverse document frequency (IDF).....	31
6.1.3- TF-IDF.....	31
6.2-Modèles de représentation de document.....	32
6.2.1-Le modèle vectoriel.....	32
6.2.1.1-Représentation binaire.....	32
6.2.1.2-Représentation fréquentielle.....	32
6.2.1.3-Vecteur TF-IDF.....	33
<b>7-Conclusion.....</b>	<b>34</b>

## Chapitre 3 : Algorithmes d'apprentissage automatique appliqués à la classification de textes

<b>1-Introduction.....</b>	<b>36</b>
<b>2-Algorithmes d'apprentissage supervisé.....</b>	<b>36</b>
2.1-Machine à vecteur support: SVM.....	37
2.2-Naive Bayes.....	38
2.3-Evaluation.....	40
2.3.1-Matrice de contingence.....	40
2.3.2-Précision et Rappel.....	41
<b>3-Algorithmes d'apprentissage non supervisé.....</b>	<b>42</b>
3.1-Hiérarchique.....	42
3.2-Non-hiérarchique.....	43
3.2.1-Kmeans.....	43
3.3-Evaluation (Validation des classes).....	44
<b>4-Formules pour calcul de distance.....</b>	<b>45</b>
4.1-Calcul de distance.....	45
4.1.1-Définition de la distance.....	45
4.1.2-Variantes de la distance.....	45
4.1.2.1- La distance Euclidienne.....	45
4.1.2.2- La distance Manhattan.....	45
4.1.2.3- La distance Cosinus.....	45
<b>5-Conclusion.....</b>	<b>46</b>

## Chapitre 4 : Etude et conception

<b>1-Introduction.....</b>	<b>48</b>
<b>2-Présentation de la méthode de conception.....</b>	<b>48</b>
2.1-La méthode OMT.....	48
<b>3-Aspect fonctionnel.....</b>	<b>50</b>
3.1-Identification des acteurs.....	50
3.2-Identification des cas d'utilisation.....	50
3.3-Description textuelle des cas d'utilisation.....	51
3.4-Diagramme des cas d'utilisation.....	52

<b>4-Aspect dynamique.....</b>	<b>53</b>
4.1-Elaboration des diagrammes d'activités.....	53
<b>5-Aspect statique.....</b>	<b>62</b>
5.1-Elaboration du modèle objet.....	62
<b>6-Conclusion.....</b>	<b>64</b>

## **Chapitre 5: Implémentation et expérimentations**

<b>1-Implémentation.....</b>	<b>66</b>
1.1-Introduction.....	66
1.2-Configuration matérielle.....	66
1.3-Langages de programmation.....	66
1.4-Quelques algorithmes.....	68
<b>2-Expérimentations.....</b>	<b>70</b>
2.1-Introduction.....	70
2.2- Corpus de textes MEDLINE.....	70
2.3-Descriptions de l'échantillon utilisé.....	73
2.4-Résultats du prétraitement des textes.....	73
2.5- Résultats de Kmeans.....	76
2.6- Résultats de labellisation.....	77
2.7-Validation des résultats de Kmeans.....	77
2.8- Résultats de naïve Bayes.....	78
2.8.1-Apprentissage.....	78
2.8.2-Test.....	78
2.9-Validation des résultats de Naïve Bayes.....	78
2.10-Conclusion.....	79

## **Conclusion générale**

<b>1-Conclusion générale.....</b>	<b>81</b>
<b>2- perspectives.....</b>	<b>81</b>

## **Annexe**

## **Bibliographie**