REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE MINISTERE DE L'ENSEIGNEMENT
SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE
UNIVERSITE SAÂD DAHLAB BLIDA
FACULTE DÉS 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

Examinateur M .....UNIVERSITE DE BLIDA

Examinateur 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 touts mes collègues du Cerist

A toutes les personnes qui m'aiment

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

Remerciements

J e 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 du 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 الاولية و نهج خاضع للرقابة (خوارزمية بايز الساذج) لتصنيف الوثائق الجديدة.

الهدف الرئيسي من عملنا هو توفير تصنيف موثوق للوثائق الطبية.

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

Liste des figures

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

Table des matières

Chapitre2 : Codage des textes

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

Chapitre 4 : Etude et conception

Chapitre 5: Implémentation et expérimentations

Conclusion générale

Annexe

Bibliographie