REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE MOHAMED BOUDIAF - M'SILA
FACULTE DES MATHEMATIQUES ET DE L'INFORMATIQUE

DEPARTEMENT D'INFORMATIQUE

MEMOIRE de fin d'étude

Présenté pour l'obtention du diplôme de MASTER

Domaine : Mathématiques et Informatique

Filière : Informatique

Spécialité : Technologie de l'Information et de Communication

Par : LAHRACHE Fatma

SUJET

Classification des textes prophétiques

Soutenu publiquement le : / /2016 devant le jury composé de :

Nom et prénom Enseignant	**Université de M'sila**	**Président**
Mr : Brahimi Belkacem	**Université de M'sila**	**Rapporteur**
.....	**Université de M'sila**	**Examinateur**
.....	**Université de M'sila**	**Examinateur**

Promotion : 2015 /2016

Table des matières :

CHAPITRE 3 : LA CLASSIFICATION DU TEXTE

CHAPITRE 4 : LES TACHES DE PRETRAITEMENT D'UN TEXTE

CHAPITRE 5 : RESULTATS ET ANALYSES

Introduction générale :

La disponibilité croissantes de la quantité énorme de l'information et la cause de l'inflation du volume des données, lorsqu'on parle des données massives nous sources, du volume qui arrive à des centaines de téraoctets ou béta octets.

A partir de là le Data Mining aperçus comme une technique conçu parlons des quantités que nous ne peuvent pas imaginer, des données des différents types et des différentes pour extraire des connaissances et à partir de la quantité énorme des informations, cette technique basée sur des algorithmes sont basés sur l'exploration de données, il est dérivé de nombreuses des sciences telles que les statistiques, la logique, l'intelligence artificielle, systèmes experts, etc.

La fouille de donnée (DM) et la fouille de donnée textuelle (TM) sont des technologies modernes qui sont utilisées dans le système d'information, le Text mining (TM) à savoir de l'extraction de l'information utile à partir de gros volumes de contenus textes.

Les différentes recherches précédentes de la classification des textes conçus beaucoup plus sur des textes française et des textes anglaise, la classification des textes arabes sont moins nombreux que les autres langues.

Le Coran et la Sunna sont les deux principales sources de la théologie islamique, El hadith est l'ensemble des paroles et des actes de l'envoyer de Dieu prophète MOHAMMED.

Lorsqu'on a fait des recherches sur les classifications des textes prophétiques on a remarqué que peu de recherches sur la classification des textes prophétiques.

Notre objectif de travail est de mener une étude sur la classification des textes prophétiques, faire une comparaison entre les algorithmes de classification supervisée, entre les modèles de représentations de texte (stem arabic, stemlight et n-gram).

Pour réaliser cet objectif, nous avons créé une collection des textes prophétique de Sahih Elboukhari qui représente deux corpus ayant un nombre de classe différent pour étudier l'effet du nombre de classes sur les performances des classifieurs.

1. On a appliqué des méthodes de classification supervisé (Naive bayes, KNN et SVM), le but est de faire une étude comparative pour répondre à la question suivante :
   - Quel est le meilleur algorithme adapté à la classification des textes prophétique ?
2. pour chaque méthode de classification on a étudié l'effet des différents étapes du prétraitement (tokenization,stopwords,stem arabic,stemlight et n-gram) sur la classification des textes prophétiques.

3. L'objectif et de comparer les différents techniques de représentation du textes prophétiques (stem arabic,stemlight et n-gram) sur la base des critères de mesures de classification (F-mesure et Accurancy).

Ce travail contient 5 chapitres :

Le premier intitulé « Data mining » dans lequel nous avons parlé principalement des notions sur la fouille de donnée et la fouille de donnée textuelle (TM).

Le deuxième intitulé « Le texte prophétique Sahih Elboukhari » dans ce chapitre, nous avons parlé sur Sahih Elboukhari et sa description, nous avons également présenté sur corpus qui nous avons collecté et étudié dans ce travail, aussi parlé sur les recherches faites dans les domaines de classification sur les textes prophétiques et leur résultats.

Le troisième intitulé « La classification de texte » ce chapitre présente les tâches de la classification, on a décrit les algorithmes les plus utilisés dans l'apprentissage automatique et fouille de donnée textuelle.

Le quatrième chapitre « les tâches de prétraitement d'un texte » dans lequel on a présenté les différents techniques et opérations de prétraitement sur le texte afin de le préparé pour la tâche de fouille de donnée textuelle (classification).

Cinquième chapitre intitulé « Résultats et analyse » concerne la partie pratique dans lequel nous avons appliqué les différents algorithmes d'apprentissage supervisé sur notre corpus, et définie les résultats obtenus et les résultats de l'étude comparative.

Conclusion Générale :

Dans le cadre de ce mémoire, ont c'est basé sur l'étude des différents méthodes de classification des connaissances sur les deux corpus des différents nombres de classes qui contiennes des textes prophétique de Sahih Elboukhari.

On a fait une étude comparative entre ces deux corpus dont on a appliqué les trois méthodes de classification supervisé (Naive bayes, KNN et SVM), pour chaque méthode on a appliqué des différents modèles (stem arabic, stemlight, n-gram), après avoir appliqué ces méthodes on à comparer les résultats obtenus avec les mesures de classification (F-mesure moyenne et Accurancy).

Après ces études nous nous sommes intéressés dans ce mémoire à comparer les différents résultats obtenu par l'environnement de l'apprentissage RapidMiner.

Pour les perspectives de ce travail. Nous proposons pour l'amélioration de cette étude les points suivants :

- Collecter un corpus plus large à partir de Sahih Elboukhari et autres (Imam Mouslem).
- Appliquer autres algorithmes de classifications (Réseaux de neurones,).
- Pour enrichir les textes prophétiques, l'utilisation des dictionnaires ontologie parait une solution prometteuse.
- Combiner les techniques de classification non supervisés (K-means, E-M...) avec les techniques de catégorisation pour atteindre les meilleurs résultats.
- Améliorer les performances de classification en utilisant les techniques de sélection d'attributs (Feature Selection).

Bibliographie :

Ouvrage :

[1] CAMPEDEL Marine, HOOGSTOËL Pierre,Marine Campedel,Pierre Hoogstoël , Sémantique et multimodalité en analyse de l'information] LAVOISIER,2011] .

[3] Dong, Guozhu, Pei, Jian, Sequence Data Mining, Springer Edition, 2007].

[14] Juan-Manuel Torres-Moreno ,Résumé automatique de documents, LAVOISIER/,rue Lavoisier 75008 paris /ISBN 978-2-7462-3212-9 /ISSN 1968-8008 .

[17] Liu,Bing ,Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data ,Second edition ;2011 .

[20] Michael J. A. Berry Gordon S. Linoff , Mastering Data Mining: The Art and Science of Customer Relationship Management,1st edition 2000.

[21] Michael J. A. Berry, Gordon S. Linoff, Data Mining Techniques For Marketing, Sales, and Customer Relationship, Management, Second Edition, april 2004.

[29] S. PRABHU, N. VENKATESAN, Data Mining and Warehousing, New Age International (P) Ltd., Publishers, New Delhi, 2007.

[35] Tom M. Mitchell, Machine Learning, (March 1, 1997 .

Article :

[2] Dominik francoeur ,Machines A Vecteurs de support une introduction /CaMUS 1 (2010).

[4] Fatma Karem* , Mounir Dhibi* Arnaud Martin , Combinaison de classification supervisée et non-supervisée par la théorie des fonctions de croyance « ch3 » « article ».

[5] G. DONG, J. PEI, Sequence Data Mining, Springer Edition, 2007.'ch1' « cours ».

[15] Kanaan G., Al-Shalabi R., Ghwanmeh S., "A comparison of text-classification techniques applied to Arabic text", Journal of the American Society for Information Science and Technology, 60(9), pp. 1836 – 1844, 2009.

[16] Laurent denoue, classification suppervisé de document , 2011,pdf « ch3 » « article ».

[18] Luc Lamontagne ,Apprentissage a base d'exemple / Concepts avancés pour systèmes intelligents .

[22] Mierswa I., Wurst M., Klinkenberg R., Scholz M., Euler T., "YALE: Rapid Prototyping for Complex Data Mining Tasks", *in the Proc. of the 12th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining KDD-06*, 2006.

[23] Mohamadally Hasan Fomani ,SVM : Machines à Vecteurs de Support ou Séparateurs à Vastes Marges ,16 janvier 2006 .

[24] Mohammad Arshi Saloot1 · Norisma Idris1 · Rohana Mahmud1 · Salinah Ja'afar1 · Dirk Thorleuchter2 · Abdullah Gani1, Hadith data mining and classification: a comparative analysis, 8 January 2016ch2 ». « article ».

[25] Osmar R. Zaian, Principles of Knowledge Discovery in Databases, CMPUT690, University of Alberta, 1999.

[26] Ph. PREUX , Fouille de données Notes de cours Université de Lille 3 ,26 mai 2011].

[27] Ph. PREUX, Fouille de données : Notes de cours, Université de Lille 3, 9 octobre 2008.

[28] Pio NardielloAffiliated withMercurioWeb SNC, Fabrizio Sebastiani, Alessandro Sperduti, Discretizing Continuous Attributes in AdaBoost for Text Categorization, Volume 2633 of the series Lecture Notes in Computer Science pp 320-334, 15 April 2003,,» « ch1 » « article ».

[30] Said D., Wanas N., Darwish N., Hegazy N., "A Study of Arabic Text preprocessing methods for Text Categorization", In the 2nd Int. conf. on Arabic Language Resources and Tools, Cairo, Egypt, 2009.

[31] Sebastian Raschka ,Naive Bayes and Text Classification I Introduction and Theory/ October 4, 2014 .

[32] Taghva, K., Elkhoury, R., Coombs, J., "Arabic stemming without a root dictionary", Information Technology: Coding and Computing, ITCC, Vol. 1, pp. 152 – 157, 2005.'

[33] Taïeb Baccouche L'Information Grammaticale Année 1998 Volume 2 Numéro 1 pp. 49-54 Fait partie d'un numéro thématique : Numéro spécial Tunisie .

[34] Tanagra_Naive_Bayes_Classifier_Explained.pdf.

[36] Waad A. Al-Harbi1 and Ahmed Emam Ph.D, EFFECT OF SAUDI DIALECT PREPROCESSING ON ARABIC SENTIMENT ANALYSIS ; ISSN:2319-7900.

Mémoire :

[37] Yasmine Hanane zeggane Mokhtar ,Algorithmesd'apprentissage pour la classification de documents ,Université de Mostaganéme -Algérie- - licence 2009] .

[19] Matallah hocine ; classification automatique de textes approche orientée agent ; UNIVERSITE ABOUBEKR BELKAID-TLEMCEN FACULTE DES SCIENCES DEPARTEMENT D'INFORMATIQUE .

[37] Z Simon,Outils classificatoires par objets pour l'extraction de connaissances dans les bases de donnée.thése de doctorat de l'université Henri Poincaré-Nancy 1,Nancy,2000.

Site web :

[7] http://america.pink/arabic-diacritics_442541.html.

[8] http://scholarpedia.org/article/Text_categorization» .

[9] http://www.iqrashop.com/Sahih-Al-Boukhari-arabe-francais-Al-imam-Zain-oud-Din-Ibn Abdoullatif-Az-Zoubaidi-Livre-livres-Hadiths-p597-.html .

[10] http://www.openml.org/a/estimation-procedures/1 .

[11] http://www.r-bloggers.com/classifieur-naif-bayesien/ .

[12] https://abjadia.wordpress.com/tag/alif-wasla.

[13] https://rapidminer.com/products/studio/.

ملخص:

في هذا البحث يتم اجراء دراسة مقارنة بين مجموعتين من الأحاديث النبوية لصحيح البخاري مختلفة من حيث العدد الكلي للأحاديث، والتي طبقنا عليها أساليب التصنيف المختلفة (Naive bayes,KNN,SVM) بعد تطبيق هذه الأساليب تمت مقارنة النتائج المتحصل عليها على أساس مقاييس التصنيف (F-mesure moyenne, Accurancy).

أظهرت النتائج أن أفضل مصنف مطبق على الأحاديث النبوية هو SVM كما أظهرت أن أفضل نموذج مطبق هو 3gram دون إبقاء الشروط.

للقيام بالتصنيف نستخدم RapidMiner.

الكلمات المفتاحية:

الأحاديث النبوية، التصنيف، اللغة العربية 3gram، SVM، Naive bayes، KNN .

Résumé :

Dans ce mémoire, une étude comparative est faite sur deux corpus de textes prophétiques de Sahih Elboukhari de différent nombre de classe, dont on a appliqué les méthodes de classification (Naive bayes, KNN et SVM), après avoir appliqué ces méthodes on a comparé les résultats obtenus sur la base de mesures de classification (F-mesure moyenne et Accurancy).

Les résultats obtenus montrent que le meilleur algorithme de classification appliquée sur les textes prophétiques est le SVM (Support Vector Machine) et le meilleur modèle est 3gram sans keepterms.

Pour la classification on a utilisé l'environnement de RapidMiner.

Les mots clés :

Classification, Naïve bayes, KNN, SVM, Text prophétique, 3gram, Langue Arabe.

Abstract :

In this thesis, a comparative study is made on two corpuses of prophetic texts of Saheeh Elboukhari of various number of class, the methods of classification of which we applied (Naive Bayes, KNN and SVM), after having applied these methods were compared the results obtained on the basis of measures of classification (average F-measure and Accurancy).

The obtained results show that the best algorithm of classification applied on the prophetic texts is the SVM (Support Vector Machine) and the best model 3gram without keepterms.

For the classification. we used the environment of RapidMiner.

Keywords :

Classification, Naive bayes, KNN, SVM, Prophetic Texts, 3gram, Arabic Language.