العنوان:	تصميم خوارزمية حديثة لزيادة كفاءة تصنيف النص العربي آليا باستخدام الانظمة الذكية
المؤلف الرئيسي:	سليمان، عبدالعزيز حسن خرساني
مؤلفين آخرين:	أحمد، عوض حاج علي(مشرف)
التاريخ الميلادي:	2013
موقع:	الخرطوم
الصفحات:	1 - 158
رقم MD:	855316
نوع المحتوى:	رسائل جامعية
اللغة:	Arabic
الدرجة العلمية:	رسالة دكتوراه
الجامعة:	جامعة النيلين
الكلية:	كلية علوم الحاسوب وتقانة المعلومات
الدولة:	السودان
قواعد المعلومات:	Dissertations
مواضيع:	تصميم الخوارزميات، تصنيف النص العربي، النظم الذكية، علوم الحاسبات
رابط:	http://search.mandumah.com/Record/855316

جامعة النيلين
كلية الدراسات العليا
كلية علوم الحاسوب وتقانة المعلومات

بحث مقدم لنيل درجة الدكتوراه في علوم الحاسوب بعنوان:

تصميم خوارزمية حديثة لزيادة كفاءة تصنيف النص العربي آليا باستخدام الانظمة الذكية

إعداد الطالب:
عبد العزيز حسن خرساني سليمان

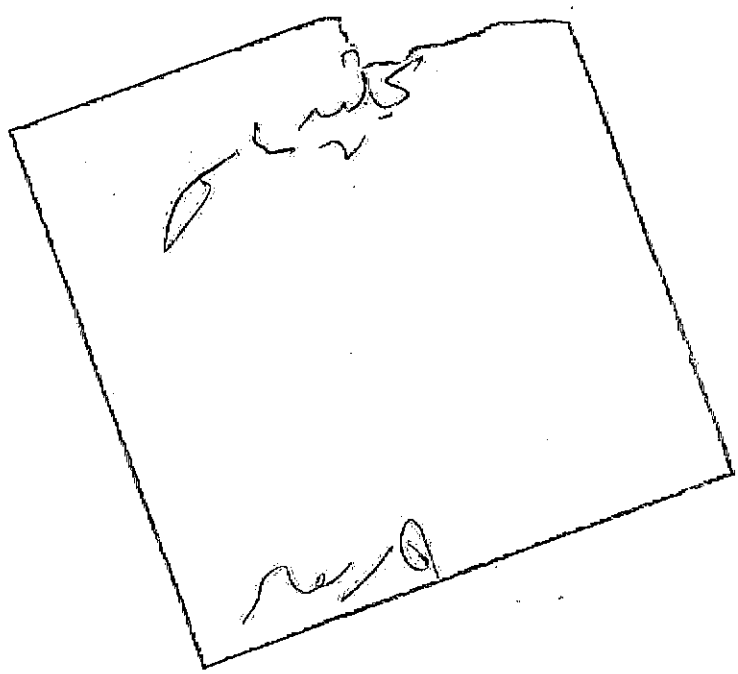
إشراف الأستاذ الدكتور:
عوض حاج علي احمد

(2013)

الاستهلال

(وَلَقَدْ نَعْلَمُ أَنَّهُمْ يَقُولُونَ إِنَّمَا يُعَلِّمُهُ بَشَرٌ لِسَانُ الَّذِي يُلْحِدُونَ إِلَيْهِ أَعْجَمِيٌّ وَهَذَا لِسَانٌ عَرَبِيٌّ مُبِينٌ)

القرآن الكريم : سورة النحل : الآية: (103)



الإهـــداء

إلى...

والدىّ...

اللذان سقياني من كأس الحب والحنان ، وأطعماني العزيمة والصبر والأمل وألبساني الشموخ والتواضع

إلى...

زوجتي وأبنائي وإخوتي الأعزاء وجميع أفراد العائلة الكريمة....

حفظهم الله ورعاهم

إلى...

كل من يحمل جزءاً من حياتنا وأعاننا على ترتيب أفكارنا

إليكم جميعاً أهدي ثمرة جهدي المتواضع

الشكـر والعرفان

الشكر لله وحده من قبل ومن بعد، الذي يسر لي إكمال هذا البحث وأنعـم علي بإتمامه في هذه الصورة

الشكر إلى من أنعم علينا بنعمة العلم وجعله لنا سراجا منيرا نستنير بـه ونقهـر الجهل

الشكر إلى من جاد علي بنصح وتوجيه وإرشاد ومتابعة أستاذي الجليل الدكتور:

أ.د/عوض حاج علي

الذي كان نعم المعين في إخراج هذا البحث بهذه الصورة، لما قدمه لي من نصائح ثرة جزاه الله عني كل الخير وجعل ذلك في ميزان حسناته.

وكذلك الشكر أ.د/ السماني عبد المطلب احمد الذي لم يبخل عني بالتوجيه والمعلومات

كما أتقدم بالشكر لكل من جامعة النيلين وجامعة سلمان بن عبدالعزيز، ولكل من شارك وجاد علي بنصح وتوجيه وإرشاد، أو قدم إلي أي مساعدة .

ولكم الشكر وجزأكم الله خيرا . .

مستخلص الدراسة

تصنيف النص Text categorization هو عملية تصنيف المستندات والوثائق إلى مجموعات محددة مسبقا من الفئات استنادا علي مضمون ومحتوي المستندات والوثائق . وهدفت الدراسة الي تصميم خوارزمية لتحسين دقة تجريد النص العربي وتصنيفه وخاصة الافعال المضعفة والأسماء الثنائية والحرف المكرر وذلك باستخدام برامج ذكية تساهم في تصنيف النص العربي . تم استخدام حزمة برنامج visual studio2010 وحزمة #c لأنها تناسب برامج النظم الذكية المستخدمة في تصنيف النصوص . وخلصت الدراسة الي مجموعة نتائج اهمها زيادة نسبة التصنيف للنصوص العربية الي 96% . تم تحسين خوارزمية التجريد وذلك بزيادة الإفعال المضعفة والاسماء الثنائية . تم تحسين خوارزمية التجريد وذلك بازلة الحرف المكرر(ـــــــــــــــــــــــ) في النص العربي . تم عمل واجهة يمكن من خلالها زيادة وزن الكلمة يدويا في النظام وذلك اذا راي المستخدم اهمية الكلمة لتحديد فئة ما . وهنالك حاجة ماسة لاستخدام اللغة العربية لعلم التصنيف الإلي وذلك لمواكبة اللغات الأخرى وتقنيات العصر اوصت الدراسة باستخدام معجم عربي حديث يستفيد من المعالجة الآلية، نجد مثلا الكثيرين ممن لديهم خلفية في الحاسب ليس لديهم معرفة باللغة العربية وعلماء اللغة العربية ليس لديهم معرفة بالحاسب إلا قليلا، وضرورة ربط الترجمة الالية بالتصنيف الالي

للنصوص العربية المجمعة، وان يصبح تصنيف النص العربي الياً ضمن مقررات لمناهج علوم الحاسوب، ويمكن استخدام المصنف Support Vector Machines (SVM) لتصنيف النص العربي.

Abstract

Text categorization is the process of classifying documents into a predefined set of categories based on the content of the documents. The study aimed to design an algorithm to improve the accuracy of the Arabic text stemming and classification especially verbs multiple reward and bilateral names and duplicate character. And using intelligent software contribute to the Arabic text classification . Software package was used visual studio2010 and c # package to fit it smart software systems used to classify texts . And The study concluded that the most important result set to increase the percentage of classification of Arabic texts to 96% . Abstraction algorithm has been improved by increasing verbs multiple reward and bilateral names. abstraction algorithm has been improved and duplicate by deleting the character in the Arabic text(————————) . been working interface from which you can increase Weight word manually in the system and the user so if Ray importance of the floor to determine the category. There is an urgent need for the use of the Arabic language for automated taxonomy in order to keep up with other languages, techniques, and study

recommended. There is no modern Arabic dictionary benefit from processing machine, We find, for example, many people who have a background in the computer have no knowledge of the Arabic language and Arab linguists have no knowledge of the computer only slightly, the need to link machine translation automatic with classification of Arabic texts collected, the Arabic text classification becoming within the decisions of the computer science approaches, can be used Support Vector Machines (SVM) to the Arabic text classification.

الفهرس العام

فهرس الأشكال

فهرس الجداول

الفصـــل الأول

Chapter one

أساسيات البحث

The essentials of search

أولا :خطة البحث

1.1 تمهيد :

لم يكن من السهل على القطاع الخاص أن ينجز مخططا شاملا ذا رؤية مستقبلية لإعداد المجتمع العربي لعصر المعلومات، واستخدام الحواسيب، نظرا لوضعه الداخلي اجتماعيا واقتصاديا وسياسيا، ولأن هذا المخطط بكل المقاييس يرتبط ارتباطا عضويا بالدول العربية في إطار التسابق الدولي للتمكن من الإعلام والانترنت، الشيء الذي يسمح بالتحكم من جهة في دواليب الأجهزة القومية، ومسايرة التطور العلمي من جهة أخرى للسيطرة آليا على كل التوجهات العلمية على الصعيد الدولي. يمكن القول في هذا الصدد إن وعي الأنظمة العربية كان غائبا بأهمية الطفرة الإعلامية التي تحققت دوليا في مجال التخطيط لعصر المعلومات، أي تصاميم الحواسيب، وأنظمتها، ووضع برامجها، ونظم تشغيلها، وتطويع تقنياتها، حيث تلعب اللغات الوطنية حجر الزاوية، إما أن تكون قوة وذات فعالية في التخطيط القومي من أجل تنمية مستديمة والدفع بعجلة الصناعة الوطنية التي أصبحت خاضعة للتقنيات الحديثة في مضمار ما أضحت تفرضه العولمة من آليات جديدة للتعامل مع التطور الصناعي العالمي، وإما ألا تكون، فتغرق حينذاك في تبعية مطلقة، وتفقد هويتها وقدرتها على اللحاق بالركب الحضاري الذي أصبحت سمته الأساسية قائمة على وسائل الاتصال.

تعد اللغة في هذا المضمار الوجه اللامع والبارز في الخريطة الإعلامية، وأساس كل المشاريع العلمية،لا باعتبارها أداة التخاطب والتواصل والتعليم والثقافة، بل لكونها

أضحت معيارا قائما على التحدي لارتباطها بالتخطيط والتصنيع والتنمية، وأي لغة لم تدخل مجال التقنيات الإلكترونية ولم تستوعب التطور الحاصل في مجال الإعلاميات فإن مصير أهلها سيعرف تدهورا وانحطاطا.

إن أهم ما يميز عصرنا هو تضخم المعلومات، وضرورة الإحاطة بها لمعالجتها وتحليلها وتنظيمها وفهرستها وتصنيفها واسترجاعها، الأمر الذي لم يعد بإمكان الطاقات البشرية القيام به، هذا بالإضافة إلى أن شيوع استخدام الحواسيب في عملية التعليم والتعلم ومعالجة اللغات الطبيعية.

أضحى حجر الزاوية للتزود بالبحث العلمي ومواكبة الفيض الهائل من المعلومات، مما يدعو حتما ضرورة التعامل مع الحاسوب، وتطوير أدواته على قاعدة اللغة الوطنية، أي إدخالها في غمار التقنيات الحديثة، واستخدامها في ضوء الوسائل العلمية المتاحة لكي تصبح أداة عملية في مجال النشر الإلكتروني.

2.1 مشكلة البحث:

اللغة العربية هي اللغة الأم لأكثر من 300 مليون نسمة ونجد انه من الصعب انتشار تقنية حديثة دون أن تستوعب إمكانية التعامل مع اللغة العربية. ومن أهم التقنيات الحديثة واسعة الانتشار التي تستوعب في التعامل معها إمكانيات اللغة العربية، تقنية إرسال الرسائل عبر شبكة الانترنت والتي تسمى البريد الالكتروني، والتي تتيح التواصل بين كثير من فئات المجتمع عبر الرسائل الالكترونية وكذلك تبادل المعلومات .

تميز اللغة العربية عن بقية اللغات العامة في توجيه كتابتها من اليمين إلي اليسار وان عدد حروفها 28 حرفا، وإنها تحتوي علي صيغة مختلفة لكل من المفرد والمثني والجمع، وأيضا المؤنث والمذكر وأكثر ما يميزها أن معظم كلماتها أصلها فعل وبسبب هذا التعقيد في التركيبة الصرفية للغة العربية فإنها تحتاج إلي مجموعة من الإجراءات والخوارزميات لتجريدها وتصنيفها.

تكمن مشكلة الدراسة في صعوبة تجريد وتصنيف النص العربي آليا بناءً علي التركيبة الصرفية للغة العربية.

3.1 أهداف البحث :

تهدف الدراسة(باستخدام تقنيات الأنظمة الذكية) إلي إيجاد حلول ناجعة لمشكلة تجريد وتصنيف النص العربي وذلك بالاتي :

- بناء خوارزمية ذكية لتصنيف النص العربي.
- بناء خوارزمية لتجريد النص العربي لدراسة الافعال المضعفة والاسماء الثنائية.
- بناء خوارزمية لتجريد النص العربي لدراسة الحرف المكرر.
- تحديد أوزان للكلمات المفتاحية اليا ويدويا اذا لزم الامر .
- بناء قاموس الكتروني للغة العربية.

4.1 أهمية البحث :

هناك حاجة ماسة إلى منظور عربي للمحتوى على الشبكة العنكبوتية العالمية، بوصفه أحد اهتمامات علم المعلومات، وبما يسهم في الوقوف على الوضع الراهن لطبيعة وخصائص هذا المحتوى، والتعامل معه حفظًا وبحثًا وتصنيفا واسترجاعًا وأمنا.

في هذا الصدد يتم تسليط الضوء على مجالات المحتوى العربي على الشبكة العنكبوتية العالمية، تجريده وتصنيفه وفئاته، وأنواعه، والسمات العامة التي تميزه عن غيره من المحتويات الأخرى. ولما كانت اللغة العربية ذات خصائص بنائية ونحوية وإملائية قد تتشابه مع (أو تختلف عن) غيرها من اللغات، فقد أصبح لزامًا التطرق لهذه الخصائص، وتسليط الضوء عليها، بحيث تكون نقطة الانطلاق عند تناول المحتوى العربي على العنكبوتية العالمية.

5.1 حدود وفترة الدراسة:

تتطرق الدراسة لمجموعة نصوص عربية مجمعة، لعدد 1250 مستند تم تجميعها من صحف (قوون، الراي العام، الأهرام) في الفترة (2010م-2012م)

6.1 منهجية البحث:

سوف تتناول الدراسة بالبحث والتطبيق تصميم خوارزمية ومن ثم تنفيذ برنامج ذكي لتجريد وتدريب وتصنيف النص العربي اليا وذلك باستخدام أدوات النظم الذكية المستخدمة في مجال علوم الحاسوب، النموذج العملي فقد اعتمد الباحث علي حزمة visual studio2010 وحزمة c# لتكوين النظام الذكي.

7.1 تقسيمات البحث:

تم تقسيم الدراسة الي سبعة فصول موضحة كالآتي:

الفصل الأول: أساسيات البحث ويحتوي علي تمهيد، مشكلة البحث، أهداف البحث، أهمية البحث، حدود البحث، منهجية البحث، تقسيمات البحث، مصادر جمع البيانات.

الفصل الثاني : الذكاء الاصطناعي Artificial Intelligence

الفصل الثالث: معالجة اللغات الطبيعية والمفاهيم المرتبطة بمعالجة اللغات الطبيعية و اللغة العربية .

الفصل الرابع : الاطار التحليلي:

الفصل الخامس : الاطار التصميمي.

الفصل السادس : الاطار التطبيقي.

الفصل السابع : النتائج والتوصيات والمراجع والمصادر.

8.1 مصادر جمع البيانات:

مصادر رئيسية : الكتب والمراجع والانترنت.

مصادر ثانوية : المجلات، الدوريات، البحوث والدراسات السابقة.