دراسة مقارنة لاستخراج جذور الكلمات وتصنيفها بمناهج مختلفة باللغة العربية

إعداد

ربى اسماعيل العمري

إشراف

أ.د غسان كنعان

الملخص

يدرس هذا البحث التصـــنيف التلقائي للنص العربي باســـتخدام تقنيات ومناهج مختلفة من المعالجة المسـبقة للنصـوص. يمكن تحليل المقالات الإخبارية إلى البيانات التي تمت معالجتها عن طريق تطبيق معالجة اللغة العربية الطبيعية (NLP) بطرق مختلفة.

لتحليل المســتندات، اســتخدمنا أربعة خوارزميات تصــنيف شــعبية مختلفة (KNN وSVM وNB وRF). تتضمن تقنيات المعالجة المسبقة للنصوص دراسة تأثير إزالة كلمات التوقف، وتطبيع النص معًـا، ثم تطبيق stemming. اســـتخـدمنـا ثلاثـة أنواع مختلفـة من الجـذوع (P–Stemmer، Snowball Stemmer، وخوجا Stemmer).

كما تبحث الدراســة في دقة اســتخدام خطوات المعالجة المســبقة بتقنيات تصــنيف مختلفة بالنســبة للفئات.

عينة الدراسـة تتألف من 3750 مقال اخباري مقسم على خمس فئات وقد تم تجميع البيانات من اثنا عشر جريدة اخبارية.

أظهرت النتائج أنSVM متفوق على المصـــنفات الأخرى، كما اعطى p-stemmer نتائج جيده على كل من (SVM,RF and NB) في حين أظهر خوجا تقدما على KNN .

تتمتع فئة الرياضــة بأفضــل نتيجة على جميعا لفئات الأخرى وأعطتsnowball أعلى النتائج على فئة الرياضة بدقة98.3%

A Comparative Study for Arabic Language Stemming and Classification

Prepared by:

Ruba I. Aalomari

Supervised by:

Prof. Ghassan Kanaan

Abstract

This research studied the automatic Arabic text classification using pre-processing text techniques. Converting the news articles to processed data, can be analyzed by applying Arabic natural language processing (NLP) in different ways.

To analyze the documents, we used four different popular classification algorithms (KNN, SVM, NB, and RF). Texts pre-processing techniques involve studying the effect of tokenization, removing stop words, and text normalization together, then apply stemming. We used three different types of stemming (P-Stemmer, Snowball Stemmer, and Khoja Stemmer).

The experiments applied on corpus consist of 3750 documents collected from twelve News paper.

The study also investigates the accuracy of classification using the pre-processing steps with different classification techniques on the categories.

The results showed that SVM superior over all the other classifiers. It also the P stemmer gave good results with each of SVM, NB and RF where P-Stemmer with SVM gave an average of (93.1 %) accuracy results while Khoja stemmer progressed in KNN. Finally the Sport category has the best result all over the other categories. The snow ball stemmer gave the highest results on Sport category with 98.3 % accuracy.