ايجاد الجذر باستخدام الحاسوب

مشاكل وحلول

فكرة مشروعنا قائمه على تصنيف تغريدات موقع التواصل الاجتماعي (تويتر) الى تغريدات سياسيه واقتصاديه وتقنيه وعرضها على الخارطة الجغرافية حسب الدولة التي ينتمي اليها.

ولمعرفة الخبر الى أي دولة ينتمي وما هو تصنيفه ، ولان من الممكن ان يتحدث المستخدمين عن نفس الخبر لكن بأسلوب لغوي مختلف نحن بحاجه الى خوارزميه لإيجاد جذر الكلمات للتمييز بين الاخبار المتشابهة. .

بالبداية قمنا بتجربة خوارزمية ، لايجاد جذر الكلمات ،ومراقبة النتائج احتوت الخوارزمية على مشاكل ونتائج غير صحيحة مثل :-

الكلمات المشددة (مستمر – سمر)

والكلمات محذوفة اللام ، مثل قاض ورام – روم

نهضت – نهضت

يسألني - يسألني

بلديات – لدي

مسألة – مسألة

تاريخيان – تاريخيان

الخاصة – خاصة

العامة – العامة

أقامت – أقامت

يمرون – رين

اهانة – اهانة

من هنا بدأنا بالعمل على تحسين الخوارزمية لتلافي هذه الأخطاء، قمنا بحل كثير من الأخطاء لكن ما زال هناك بعضها، المشكلة أنه في بعض الأحيان عندما ننتهي من حل مشكلة معينة نكتشف أننا بحلنا هذا خلقنا مشكلة أخرى كانت غير موجودة.

مثال : حللنا مشكلة الكلمات المبدوئة بال التعريف لكن ظهرت عندنا مشكلة الكلمات التي تبدأ بال أصلية. مثل الجأ

مشكلة أخرى هي أنا نعتمد على مستخدمي موقع تويتر في جلب النصوص والأخبار فنحن نتعامل مع مستخدم غير معياري حيث لا يتقيد بقواعد اللغة ولا بقوانينها. كأن يكتب ألعاب هكذا <العاب، بدون همزة على الألف > واستمرَ هكذا استمر بدون شدة > كل هذه الاخطاء تؤثر على عمليه ايجاد الجذر.

بالاضافة الى هذا كله فان هناك مشكلة أرى أن سببها يعود لكوننا نتعامل مع حاسوب لا يفهم الا لغة (1,0)، المشكلة ظهرت عندما تم تمرير الكلمات على الأوزان التي لدينا، كان الحاسوب يجد وزن للكلمة مع انه ليس وزنها الحقيقي السبب في ذلك أنه يجد أن الكلمة قد حققت شروط هذا الوزن فيعتبره هو وزنها. مثال : اهانة على وزن افعلة ، عندما تم تمرير الكلمة على الأوزان مرّ على وزن

فعالة ووجد أن الشروط الموضوعة على هذا الوزن تنطبق على هذه الكلمة، أي أن "اهانة" ثالثها ألف واخرها تاء مربوطة، تحقق الشرط اذن هي على وزن فعالة وهذا ليس بالصحيح. بدأنا عاكفين على حل هذه المشكلة التي كانت برأينا هي الأبرز والأهم وقمنا بوضع شروط منطقية لكي لا يخطأ بالأوزان.

الان نحن بصدد حل المشاكل الأخيرة التي ظهرت لدينا هذه المشاكل هي :

1- الفعل المحذوف لامه مثل: يدعون
2-الأسماء التي تحتوي على تاء منقلبة عن واو مثل: اتساع
3- الأسماء والأفعال التي تحتوي في بدايتها على ال أصلية مثل: الجأ
4- بعض الكلمات المشددة.

بالنهاية أود القول بأن ما يجول في خاطرنا وما نجده مشكلة حقيقية قائمة هو أن المحتوى العربي على الشبكة العنكبوتية ليس بالجودة المطلوبة، أراه على وجهين : محتوى مهم لكن ليس بالكمية ولا الأماكن الموثوقة ، دائما نجد أن المحتوى العربي يتغلغل بالمنتديات التي تجعلنا نادرا ما نثق بالمحتوى. الوجه الاخر هو محتوى تافه ليس له أهمية في الناحية العلمية والتعليمة وغيرها من النواحي، هذا يجعل تحليل النصوص العربية المحوسبة المعتمدة على نصوص مأخوذة من هذه الشبكة أمرا صعبا مما يؤدي الى عزوف المختصين بالحاسوب لعمل تطبيقات وتحليلات للنصوص العربية.

الطالبات :-

مرام نعيم

اسماء ابو نعمه

شذا بني فضل

جامعه النجاح الوطنية كلية تكنولوجيا المعلومات