الجمهورية العربية السورية

وزارة التعليم العالـــــي

الجامعة الافتراضية السورية

دراسة أعدت لنيل درجة الماجستير في علوم الويب

تطوير آلية جديدة لترجيح المصطلحات

من خلال تحليل الوثائق النصية ومعالجة اللغة الطبيعية

Developing a New Terms Weighting Schema

Through Text Analysis and Natural Language Processing

إعداد الطالبة : منال شيخ أوغلي

بإشراف الدكتور محمد مازن المصطفى

الملخص:

احتلت أنظمة استرجاع المعلومات IR خلال العقود الخمسة الأخيرة مركزاً متقدماً في مجال البحث العلمي حيث شهد المحتوى الرقمي العالمي تزايداً ملحوظاً (خاصة المحتوى النصي) نظراً للمعلومات المتدفقة إليه عبر عالم الويب والنمو المضطرد للشبكة العنكبوتية، الأمر الذي جعل من الضرورة إيجاد أنظمة استرجاع معلومات تساعد المستخدم في الحصول على المعلومة التي يرغب بها في ذلك الكم الهائل من المعلومات الرقمية، فأنظمة استرجاع المعلومات الفعّالة هي التي توفر طريقة قادرة على استعادة عدد من المستندات ذات الصلة العالية بمتطلبات المستخدم.

في هذه الورقة سنلقي الضوء على النماذج الأهم لأنظمة استرجاع المعلومات مع التطرّق لمحاولات الباحثين تحسين هذه النظم من خلال استعراض الأبحاث التي تم من خلالها دراسة تأثير أهمية المصطلح والعمل على ابتكار أسلوب جديد للتوزين ومقارنة هذه الأساليب الجديدة وأثرها في تصنيف الوثائق.

حيث بيّنت الدراسات أن الوزن المناسب لأهمية المصطلح يؤثر بشكل فعال على نتائج الاسترجاع فتوزع المصطلح وموقعه ودلالاته وتزامن حدوثه مع المصطلحات الأخرى في الوثيقة تعتبر عوامل لابد أن تؤخذ في الاعتبار عند قياس التشابه بين الوثائق أو بين الاستعلام والوثيقة.

كما سنقوم بتطوير آلية جديدة لترجيح المصطلحات وذلك اعتماداً على تحليل النصوص وإعادة ترجيح المصطلحات الواردة فيها اعتماداً على معايير تقييم للوثائق مستنتجة من خلال المعلومات المستقاة من الأبحاث السابقة ومناقشة طرق معالجة الوثائق وإجراء الاختبارات الرياضية المناسبة للوصول إلى طريقة ترجيح جديدة للمصطلحات تعزز قدرة النظام المقترح على استرجاع المعلومات الأكثر ملاءمة لطلب المستخدم وبالتالي الحصول على تصنيف أكثر دقة الأمر الذي يساهم في تطور هذا المجال المعرفي.

فمشكلة البحث الرئيسية تتركز في مناقشة نقاط الضعف في خوارزمية الترجيح TF-IDF التقليدية في النموذج الشعاعي لاسترجاع المعلومات والتي تتجاهل العديد من التفاصيل والميزات المعرّفة للوثائق ومن ثم استعراض بعض الخوارزميات التي تمت لتحسين أدائها والمقارنة بين هذه الخوارزميات بهدف ايجاد أسلوب جديد لترجيح المصطلحات يساهم في الحصول على تصنيف للوثائق أكثر دقة وفعالية وبالتالي استرجاع الوثائق الأكثر ارتباطاً باستعلام المستخدم أو الوثائق الأكثر تشابهاً من خلال التحليل الدقيق لكل من الوثائق والاستعلامات وبالتالي الحصول على نتائج أكثر صلة برغبة المستخدم وهو الهدف الأهم الذي تسعى إليه جميع أنظمة استرجاع المعلومات.

الفصل الأول: دراسة نظرية لنماذج استرجاع المعلومات

مقدمة:

إن أنظمة البحث والتصنيف والفهرسة الحديثة التي تدعمها قوة الحوسبة التي وصلنا إليها اليوم وسرعات الشبكة السريعة وسعة تخزين البيانات غير المحدودة تقريباً تعني أن لدينا سهولة الوصول إلى جميع المعلومات التي نحتاجها عندما نحتاج إليها [1].

وإن عملية استرجاع المعلومات (Information Retrieval) تعتمد إلى حد كبير على عملية المطابقة بين رغبة المستخدم والتي يعبر عنها من خلال الاستعلام وبين مخازن المعلومات لإعادة النتائج ذات الصلة برغبة المستخدم.

ولقد شهد مجال استرجاع المعلومات تطوراً ملحوظاً في العقود الأخيرة وذلك نتيجة حاجة المستخدمين الماسة للبحث في ذلك الكم الهائل من المعلومات الرقمية.

في هذا الفصل نستعرض لمحة تاريخية عن تطور هذا المجال المعرفي وأهم نماذج استرجاع المعلومات والتي أثبتت فعاليتها في عمليات البحث المختلفة موضحين نقاط القوة والضعف في كل نموذج.

لمحة تاريخية:

إن أرشفة المعلومات المكتوبة قد تعود إلى الألف الثالث قبل الميلاد عندما خصص السومريون مناطق خاصة لتخزين أقراص الطين مع النقوش المسمارية ليدرك السومريون منذ ذلك الوقت أهمية التنظيم السليم والوصول إلى الأرشيف كعاملين أساسيين للاستخدام الفعال للمعلومات [2].

وأصبحت الحاجة إلى تخزين واسترجاع المعلومات المكتوبة ذات أهمية متزايدة وخاصة بعد اختراع الأجهزة الحاسوبية، وبدأ التفكير في استخدامها لتخزين واسترجاع كميات كبيرة من المعلومات بشكل آلي.

وفي عام ١٩٤٥ نشر Vannevar Bush مقالة رائدة بعنوان "As We May Think" والتي ولدت فكرة الوصول التلقائي إلى كميات كبيرة من المعرفة المخزنة [2] [1] [3]. ليُستخدم مصطلح استرجاع المعلومات Information Retrieval لأول مرة أثناء تقديم Calvin Mooers لورقة بحثية في مؤتمر عام ١٩٥٠، حيث كتب " المشكلة قيد المناقشة هنا هي البحث الآلي واسترجاع المعلومات من التخزين وفقاً للمواصفات وحسب الموضوع ..." [4] [5].

قام Mooers باستخدام هذا المصطلح لوصف العملية التي يستطيع المستخدم من خلالها تحويل حاجته للمعلومات إلى قائمة فعلية منها ضمن مجموعة من المراجع المفيدة، ووضح أن استرجاع المعلومات هو اسم آخر أكثر عمومية لإنتاج ببليوغرافيا[1] الطلب وقال بأن استرجاع المعلومات يتضمن الجوانب الفكرية لوصف المعلومات ومواصفاتها الضرورية لعملية البحث، فاسترجاع المعلومات أمر حاسم لتوثيق وتنظيم المعرفة [5].

وفي منتصف خمسينيات القرن الماضي تحولت تلك الفكرة إلى وصف أكثر واقعية لكيفية البحث التلقائي في النصوص المخزنة [2].

وبدأ استرجاع المعلومات IR في الظهور كعلم في هذا المجال مع تطورين مهمين هما: كيفية فهرسة الوثائق وكيفية استرجاعها [4].

وبدأت العديد من الأعمال تظهر في تلك الفترة والتي تناولت الفكرة الأساسية المتمثلة في البحث عن نص باستخدام جهاز كمبيوتر ليتم وصف واحدة من أكثر الطرق فاعلية بواسطة H.P. Luhn في عام ١٩٥٧، حيث اقترح استخدام المصطلحات Terms كوحدات فهرسة للمستندات وقياس تداخل الكلمات كمعيار للاسترجاع [2] [6].

لتشهد بداية الستينات مجموعة واسعة من الأنشطة والتي حددت طرق ووسائل تحسين أنظمة استرجاع المعلومات وكان من الشخصيات البارزة التي ظهرت في تلك الفترة Gerard Salton والذي أنشأ وقاد مجموعة IR كبيرة بدأت العمل في جامعة Harvard ثم في جامعة Cornell حيث وضعت هذه المجموعة العديد من الأفكار والمفاهيم والتي لا تزال مجالاً للبحث حتى يومنا هذا [2] [4].

وبدأت العديد من الأفكار الخلّاقة تظهر تباعاً ومن هذه الأفكار كان طريقة تمثيل المستندات والاستعلامات على أنها متجهات ذات N بعد، حيث N هو عدد المصطلحات الفريدة في المجموعة التي يتم البحث فيها. وكان Paul Switzer أول من اقترح في عام ١٩٦٣ طريقة الاعتماد على المتجهات (الأشعة) لتمثيل المستندات والاستعلامات [4].

ليقترح Salton فيما بعد قياس التشابه بين الوثائق والاستعلامات من خلال معامل الجيب Cosine.

[1] **ببليوغرافيا**: جاءت هذه الكلمة أصلا من اللغة اليونانية وهي مركبة من كلمتين هما Biblion :كتيب وهي صورة التصغير للمصطلح Biblios بمعني كتابة، وكلمة Graphia وهي اسم الفعل المأخوذ من Graphein بمعنى توصيف فأبسط تعريف لهذه الكلمة هي توصيف الكتب.

كما كان هناك تحسين آخر ذو أهمية كبيرة أُدخل على أنظمة استرجاع المعلومات في تلك الفترة وهو التغذية الراجعة ذات الصلة Relevance Feedback هذه العملية التي تستخدم اليوم بشكل واسع في محركات البحث الحديثة لدعم عمليات البحث وتمييز المستندات التي تم استردادها سابقاً على أنها ذات صلة في نظام استرجاع المعلومات [4].

وفي منتصف الستينيات ظهرت شركات البحث التجارية والتي قامت بتطوير أنظمة مخصصة لشركات كبيرة أو مؤسسات حكومية وكانت Dialog واحدة من أوائل هذه الشركات والتي تم تشكيلها في عام ١٩٦٦ وقامت بإنشاء نظام استرجاع معلومات للإدارة الوطنية للملاحة الجوية والفضاء في الولايات المتحدة الأمريكية NASA [4].

لتشهد سبعينيات وثمانينيات القرن الماضي العديد من التطورات التي بنيت على التقدم الذي تحقق في الستينيات، فتم تطوير نماذج مختلفة للقيام باسترجاع المستندات وتم إحراز تقدم مهم في هذا المجال. وقد أثبتت هذه النماذج والتقنيات الجديدة بشكل تجريبي أنها فعالة في مجموعات النصوص الصغيرة (عدة آلاف من المقالات) والتي كانت متاحة للباحثين في ذلك الوقت [2].

وكان أحد هذه التطورات الرئيسية في تلك الفترة هو ظهور مفهوم تردد المصطلح Term Frequency (TF) أي تواتر الكلمة في وثيقة ما [4].

والذي استكمل بعمل Spark Jones والتي قدّمت ورقتها حول تردد المستند العكسي Inverse Document Frequency (IDF) ووضّحت من خلال الورقة التي قدمتها أن تواتر حدوث كلمة في مجموعة مستندات يتناسب تناسباً عكسياً مع أهميتها في الاسترجاع [7]، ليتم اعتماد فكرة الجمع بين الأسلوبين (TF، IDF) بسرعة كأسلوب جديد لتوزين المصطلحات [8].

لتتوالى فيما بعد الأبحاث والدراسات المختلفة لإنتاج أشكال مختلفة من مخططات الترجيح بهدف تطوير أنظمة استرجاع المعلومات.

في أواخر عام ١٩٩٠ أنشأ Berners-Lee شبكة الويب العالمية وكان عدد المواقع على شبكة الإنترنت وكمية الصفحات صغيرة نسبياً حتى عام ١٩٩٣. وبالتالي فإن الفهرسة اليدوية التقليدية للمحتوى كانت كافية في تلك السنوات الأولى لظهور شبكة الانترنت العالمية [4].

وفي منتصف عام ١٩٩٣ ، كما جاء في دراسة لـــ Matthew Gray [٢]، كان هناك حوالي ١٠٠ موقع على شـبكة الإنترنت إلا أن العدد بدأ يتضـاعف بسـرعة لتبدأ محركات البحث على شـبكة الإنترنت في الظهور في أواخر عام ١٩٩٣ للتعامل مع هذا النمو المتسارع ولتزداد شعبية محركات البحث يوماً بعد يوم كوسيلة متطورة للوصول إلى المعلومات وكواحدة من أهم تطبيقات أنظمة استرجاع المعلومات [4].

تعريف أنظمة استرجاع المعلومات:

في عام ١٩٨٣ قدّم كل من Salton و McGill هذا التعريف: "نظام اســترجاع المعلومات هو نظام يسـتخدم لتخزين عناصـر المعلومات التي تحتاج إلى معالجتها والبحث عنها واسـترجاعها ونشـرها على مجموعة من المستخدمين" [9] [10].

وأما Calvin Mooers فقد عرّف اســترجاع المعلومات بأنها الطرق التي تمكّن المســتخدم من تحويل حاجته للمعلومات إلى قائمة مسـتندات فعلية في وحدات التخزين وبأنها عملية البحث أو الاكتشـاف فيما يتعلق بالمعلومات المخزنة [5] [9].

كما تم تعريفه أيضاً "بـأنه العلم الذي يبحث عن مواد (عادةً ما تكون المستندات) ذات طبيعة غير منظمة (عادةً ما تكون نصـية) تلبي حاجة المعلومات من داخل مجموعات كبيرة (عادةً ما يتم تخزينها على أجهزة الكمبيوتر)" [11].

ويُعرّف نظام استرجاع المعلومات من قبل الموسوعة العربية لمصطلحات المكتبات والمعلومات والحاسبات على أنه "مجموعة من الإجراءات المميكنة التي تســتخدم في الرجوع إلى البيانات التي تحويها الوثائق Documents وتكشيف تلك البيانات واختزانها بطريقة يمكن استعادتها عند الطلب" [12].

ووفقاً لقاموس العلوم والتكنولوجيا Science and Technology Dictionary، فإن استرجاع المعلومات هو تقنية وعملية البحث عن المعلومات واستعادتها وتفسيرها من كميات كبيرة من البيانات المخزنة [9].

في حين تم تعريف استرجاع المعلومات في موسوعة Britannica Concise Encyclopedia على أنه استعادة المعلومات لا سيما في قاعدة البيانات المخزنة في جهاز الكمبيوتر [9].

[٢] تم إنشاء زاحف الويب الأول في منتصف عام ١٩٩٣ وتمت تسميته Wanderer. وقام Matthew Gray بإنشاء فهرس يسمى Wandex وكان الهدف منه قياس حجم الويب ولم يستخدم أبدًا لأغراض استرجاع المعلومات.

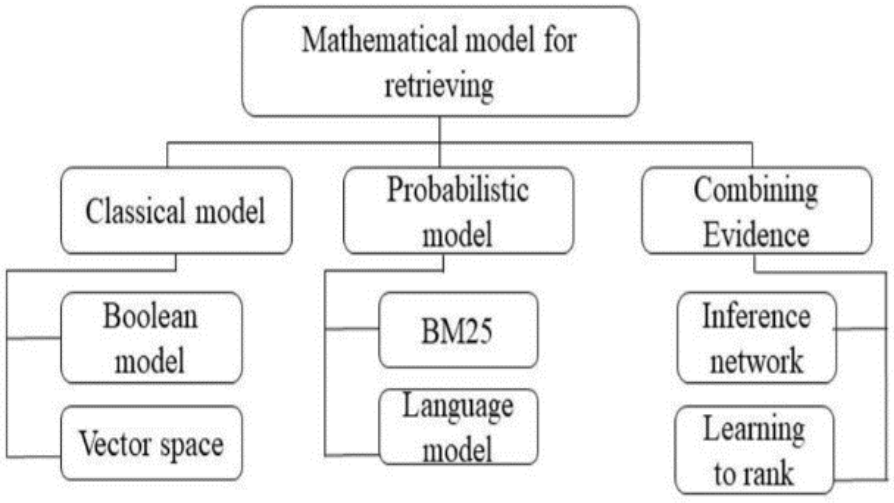
نماذج استرجاع أنظمة المعلومات:

تعتبر نماذج استرجاع المعلومات بمثابة مخطط لتنفيذ نظام استرجاع فعلي حيث أن نظام الاسترجاع يتنبأ ويشرح ما الذي يريده المستخدم من خلال تحليل الاستعلام المحدد من قبل المستخدم [13]. حيث توفر النماذج تقنيات وأساليب مختلفة للمطابقة بين الوثائق المخزنة وبين الاستعلام. وقد كان النموذج الأكثر بساطة يقوم بمطابقة المصطلحات الواردة في الاستعلام بمصطلحات فهرس الوثائق لتصبح عملية المطابقة فيما بعد أكثر تطوراً من خلال تطبيق أساليب إحصائية ورياضية لتحديد الوثائق الأكثر ملاءمة للاستعلام من خلال تحديد مدى الارتباط بين الاستعلام والوثيقة.

فالهدف الرئيسي لنماذج استرجاع المعلومات إيجاد الوثائق ذات الصلة باحتياجات المعلومات من مجموعة كبيرة من الوثائق [9].

وتختلف نماذج أنظمة استرجاع المعلومات فيما بينها بشكل عام بطريقة تمثيل الوثائق والاستعلامات، وتابعي المطابقة والترتيب ويمكن تصنيف هذه النماذج وفقاً لبعدين: الأساس الرياضي وخصائص النموذج.

وتُصنف نماذج الاسترجاع وفقاً لبعد الأساس الرياضي كنماذج كلاسيكية (نماذج الفضاء المنطقي والشعاعي)، والنماذج الاحتمالية (BM–25 ونماذج اللغة) والجمع بين نماذج الأدلة (شبكة الاستدلال ونماذج الترتيب) [14].



الشكل ١ - التصنيف الرياضي لنماذج استرجاع المعلومات [14]

وفيما يلي نستعرض أهم ما يميز كل نموذج من هذه النماذج موضحين أوجه القصور في كل منها:

النماذج الكلاسيكية Classical Models:

١. النموذج المنطقي Boolean Model:

يعتبر النموذج البولياني النموذج الأول لاسترجاع المعلومات [13]. وهو أحد أقدم وأبسط النماذج في هذا المجال حيث يعتمد على الجبر المنطقي [9]، ومبدأ التطابق التام Exact Match [13].

في هذا النموذج يتم تمثيل الوثائق من خلال مجموعة من المصطلحات (تعرَّف أيضاً باسم مصطلحات الفهرس) [9] [15]. ويُعرض كل مصطلح على أنه متغير منطقي فتكون قيمة المصطلح في مستند ما صحيحة إذا كان المصطلح موجوداً في المستند وخاطئة في خلاف ذلك [15]. وبالتالي يتم تصنيف الوثائق في هذا النموذج إلى فئة وردت فيها مصطلحات الاستعلام، وفئة لم ترد فيها المصطلحات وهذا التصنيف يعني أنه ليس هناك أي نوع من الترتيب في تقييم مدى صلة الوثائق بالاستعلام [16].

ويتم تحديد احتياجات المستخدم من المعلومات من خلال مزيج من المعاملات المنطقية حيث عرّف George Bool ثلاث معاملات أساسية وهي (AND , OR , NOT) وتعني التقاطع و الجمع والفرق، والتي تستخدم أثناء صياغة الاستعلام [9] [13] [14]. وتكون النتائج في حال استخدام هذه المعاملات كالتالي:

١. التقاطع And: ويعني أنه يجب اقتران المصطلحات الواردة في الاستفسار ووجودها مجتمعة في الوثائق.
٢. الجمع OR: يقصد به عدم اشتراط اقتران المصطلحات مع بعضها في الوثائق، حيث يكفي ورود أحد المصطحات في الوثيقة ليتم استرجاعها.
٣. الفرق Not: يشير إلى استبعاد المصطلحات التي تلي هذا المعامل بحيث يتم استرجاع الوثائق التي لا تحوي تلك المصطلحات [16].

وببساطة يمكن القول بأن نموذج الاسترجاع المنطقي هو نموذج التطابق التام حيث يطرح الاستعلام على شكل تعبير منطقي من خلال الجمع بين المصطلحات باستخدام المعاملات المنطقية الثلاث ليعيد النتائج التي تطابق الاستعلام من غير ترتيب للأهمية فيما بينها فالمستند إما أن يحوي هذا المصطلح أو أنه لا يحويه وبالتالي لا إمكانية للمطابقة الجزئية.

وعلى الرغم من أوجه القصـــور التي تعاني منها هذه النماذج إلا أنها ما زالت مســـتخدمة في العديد من أنظمة الاسترجاع [16]. كما أنه يعطي المستخدمين الخبراء شعوراً بأنهم قادرين على السيطرة على النظام بقدر أكبر مقارنة مع غيره [2].

ولقد جرى تطوير هذا النماذج فظهر على سبيل المثال النموذج المنطقي الموسع Extended Boolean Model والذي تم وصـــفه في مقال نشـــر في مجلة ACM عام ١٩٨٣ من قبل كل من (Gerard Salton, Edward A. Fox, and Harry Wu).

من خلال هذا النموذج نستطيع إجراء المطابقة الجزئية ووزن المصطلح. فهو يجمع بين خصائص نموذج Vector Space Model (سـيتم شـرحه في الفقرة التالية) مع خصـائص النموذج المنطقي من خلال صياغة استعلام منطقي Boolean query واستخدام أسلوب قياسي لحساب التشابه بين الاستعلام والوثيقة [15]. وذلك في محاولة لترتيب النتائج حسب صلتها بالاســتعلام وذلك عن طريق تحديد الوثائق التي وردت فيها كل مصطلحات البحث الواردة في الاستعلام ووضعها في مرتبة أفضل من تلك التي لم يرد فيها أحد المصطلحات. ولكن بقي الحكم في هذا النموذج على مدى أهمية الوثائق بالنسبة للاستعلام معتمداً على ورود المصطلح فيها أو عدم وروده دون الوضع في الاعتبار مدى تكراره أو وروده في الوثيقة الواحدة والتي تراعيه النماذج الأخرى [16].

كما ظهر النموذج الضبابي Fuzzy Set والذي يعتمد فكرة درجة عضوية المصطلح في الوثيقة Degree Of Membership [15]، وتأخذ درجة عضوية المصطلح قيمة بين الـ (٠، ١) [15] [16].

ويتم تمثيل الوثائق والاستفسارات وفقاً لهذا النموذج بواصفات متصلة جزئياً بدلالات المحتوى فالمطابقة فيه تتم بشكل تقريبي وليس كلي كما أن عملية المضاهاة لا تقتصر على المصطلح المطابق تماماً لمصطلح البحث ولكن يتم تجاوزه إلى مصـــطلحات أخرى مطابقة له جزئياً بالاعتماد على العلاقة الدلالية Semantically التي تربط بين المصطلحات [16].

وبالتالي فإن هذا النموذج ينظر إلى درجة ارتباط المصـــطلحات فيما بينها وليس تواتر حدوث المصـــطلح ضمن الوثيقة كما أن هذا النموذج لا يعتبر من النماذج المستخدمة على نطاق واسع [15].

كما تم تطوير نموذج المنطقة Region Model كتطوير للنموذج المنطقي Boolean Model أيضـاً حيث يقوم بنمذجة مجموعات الوثائق كسـلسـلة خطية من الكلمات وإن أي تسـلسـل من الكلمات المتتالية يسمى منطقة والمناطق تعرف من خلال موضع البداية وموضع النهاية [13].

ويعتبر الاختلاف الأهم بين النموذج المنطقي ونموذج المنطقة أن الأخير مصـــمم للبيانات الشـــبه مهيكلة semi-structured data.

٢. نموذج الفضاء الشعاعي Vector Space Model:

تم اقتراح هذا النموذج من قبل Gerard Salton وزملاؤه في عام ١٩٨٣ [10]. وكان هذا النموذج يعتمد على معيار التشـــابه الذي اقترحه Hans Peter Luhn عام ١٩٥٧ والذي كان أول من اقترح النموذج الاحصـــائي للبحث عن المعلومات معتمداً على معيار التماثل بين الاســتفســارات والوثائق، حيث صـــاغ H.P Luhn معيار التشـابه كالتالي: "كلما اتفق تمثيلان في العناصـــر المحددة وتوزيعها كلما زاد احتمال تمثيلها لمعلومات مماثلة" [6].

وبالاعتمـــاد علـــى هـــذا المعيـــار اعتبـــر Salton وزمـــلاؤه أنـــه يمكـــن تمثيـــل كـــل مـــن الوثـــائق والاسـتعلامات كأشـعة فـي الفضـاء الاقليـدي بحيـث يـتم تعيـين بعـد مسـتقل لكـل مصـطلح ومـن ثـم قـاموا بحسـاب التشـابه Similarity بـين الأشـعة باسـتخدام جيـب الزاويـة Cosine بـين المتجهـات الممثِّلة لكل من الوثيقة والاستعلام [10].

كـان هـذا النمـوذج مـن أكثـر النمـاذج اسـتخداماً حتـى نهايـة القـرن الماضـي [15]. وهـو نمـوذج مـن النمـاذج الجبريـة يـتم تمثيـل الـنص فيـه مـن خـلال شـعاع مـن المصـطلحات والمصـطلحات هـي عبارة عن كلمات وعبارات تمثل مصطلحات الفهرسة [1] [2] [14].

ووفقاً للنموذج الاحصـــائي ينظر إلى محتوى الوثيقة على أنه حقيبة من الكلمات *Bag-of-Words* [15] [17]. وذلك يعني أن محتوى الوثيقة يشتمل على مصطلحات غير مرتبة وذات ترددية غير منتظمة داخل محتوى الوثيقة [17].

فـإذا مـا تـم اختيـار الكلمـات كمصـطلحات للفهرسـة فـإن كـل كلمـة سـيكون لهـا بعـد مسـتقل فـي فضـاء الأشعة.

ويـتم تعيـين أوزان لمصـطلحات الفهـرس فـي كـل مـن الوثـائق والاسـتعلامات. ومـن ثـم قيـاس التشـابه Similarity أو المسـافة Distance بـين الوثيقـة والاسـتعلام وذلـك مـن خـلال عـدة طـرق نـذكر منها على سبيل المثال:(Dot Product, Euclidean, Manhattan, Cosine…). [14]

فإذا كان الشعاع $\vec{D}$ يمثل الوثيقة، والشعاع $\vec{Q}$ يمثل الاستعلام فإن التشابه $\text{Sim}(\vec{D}, \vec{Q})$ بين الاستعلام والوثيقة يمكن حسابه بعدة طرق من خلال المعادلات الرياضية المبينة في الجدول التالي:

Dot Product	$\sum Wt_{iQ}.Wt_{iD}$
Cosine	$\frac{\sum Wt_{iQ}.Wt_{iD}}{\sqrt{\sum(Wt_{iQ})^2}.\sqrt{\sum(Wt_{iD})^2}}$
Jaccard's Coefficient	$\frac{\sum Wt_{iQ}.Wt_{iD}}{\sum(Wt_{iQ})^2.\sum(Wt_{iD})^2-(\sum Wt_{iQ}.Wt_{iD})}$
Dice's Coefficient	$\frac{2\sum Wt_{iQ}.Wt_{iD}}{\sum(Wt_{iQ})^2+\sum(Wt_{iD})^2}$

الجدول ١ – طرق قياس التشابه بين الاستعلام والوثائق في نموذج الفضاء الشعاعي

ويعتبر التحدي الأكبر الذي يواجه هذا النموذج هو تعيين القيمة المناسبة لمكونات الشعاع وهذه المشكلة تعرف باسم ترجيح المصطلحات Term Weighting [13].

فالفكرة الرئيسية التي بُني عليها نموذج الفضاء الشعاعي هي أن استخدام الوزن الثنائي الذي اعتمد عليه النموذج البولياني (ذو صلة أو ليس ذو صلة) والذي يعبر عنه بالثنائية (0,1) يحد جداً من عملية الاسترجاع والترتيب الطبقي للنتائج [17].

وبناءً عليه قدّم هذا النموذج إطار عمل جديد يعتمد على ما يعرف بالمطابقة الجزئية (أي أن درجة اتصال أو عدم اتصال الوثيقة بالاستعلام يحدد من خلال أوزان متفاوتة بين قيمتين أدناها يرمز له بالرقم 0 وأعلاها يرمز له بالرقم 1، ويتم ذلك من خلال مجموعة من المعادلات الخاصة بوزن المصطلح بحيث تستخدم هذه الأوزان في نهاية المطاف لحساب درجة التشابه والتماثل بين كل من الوثيقة المختزنة في النظام وبين استفسار المستخدم) [17].

النماذج الاحتمالية Probabilistic Models:

تم اقتراح الفكرة الأولية للنماذج الاحتمالية في أنظمة استرجاع المعلومات من قبل Maron, Kuhns في ورقة بعنوان Probabilistic Indexing and Information Retrieval نشرت عام ١٩٦٠ [2]

[18]. حيث اعتبر أول عمل علمي يتطرق إلى استخدام المنهج الاحتمالي في استرجاع المعلومات وعليه ظهر ما يُعرف بالتكشـيف الاحتمالي Probabilistic Indexing ومنذ ذلك الحين تم تطوير العديد من النماذج التي تعتمد على تقنيات مختلفة لتقدير الاحتمالات.

فهذا النموذج هو عبارة عن عائلة من النماذج والتي تعتمد على مبدأ عام وهو أن الوثائق في المجموعة يجب أن ترتب حسـب احتمال ارتباطها بالاسـتعلام [2] [9] [15]. وهذا المبدأ يدعى بمبدأ الترتيب الاحتمالي Probabilistic Ranking Principle PRP [2] [15] والذي يقول:

"إذا كانت اسـتجابة نظام الاسـترجاع المرجعي لكل طلب هي ترتيب الوثائق في المجموعة من خلال الترتيب التنازلي لاحتمالية أهميتها للمسـتخدم الذي أرسـل الطلب، حيث يتم تقدير الاحتمالات بأكبر قدر ممكن من الدقة على أسـاس بيانات يتم تقييمها وإتاحتها للنظام لهذا الغرض، فإن الفعالية العامة للنظام بالنسبة للمستخدم ستكون أفضل ما يمكن الحصول عليه على أساس تلك البيانات" [19].

ففي النموذج الاحتمالي بدلاً من مطابقة نفس المصطلحات الواردة في الاستعلام مع المصطلحات الواردة في الوثائق يتم تقدير الاحتمالات التي يمكن أن تكون فيها الوثيقة ذات صلة باستعلام معين ومن ثم ترتب الوثائق المسترجعة ترتيباً تنازلياً وفقاً لاحتمالات صلتها بالاستعلام وفائدتها بالنسبة للمستخدم. وتلعب التغذية الراجعة دوراً مهماً في مثل هذه النماذج حيث تقوم بتوظيف المعلومات التاريخية لاسـتخدام تلك الوثيقة في احتساب احتمالات صلتها بالاستعلام [20].

وتختلف النماذج الاحتمالية بناءً على الافتراضات التي يعتمد عليها [15]. فالنموذج الاحتمالي الكلاسيكي والذي قدمه كلٌّ من S.E. Robertson ، K.Sparck Jones في عام ١٩٧٦ يفترض اسـتقلال المصـطلح وعرّف هذا النموذج باسـم نموذج الاسـترجاع القائم على الثنائيات المسـتقلة The binary independence retrieval model (BIR)، واعتمد هذا النموذج على مفهوم أن مجموعة الوثائق المخزنة في نظام اسـترجاع المعلومات تنقسـم إلى مجموعتين ثنائيتين مسـتقلتين عن بعضـهما البعض، المجموعة الأولى تعرف بمجموعة الصـلة والتي يتسـم محتواها بالصـلة بالاسـتعلام، أما المجموعة الثانية فتعرف بمجموعة اللاصلة والتي يتسم محتواها بعدم الصلة بالاستعلام [17] [21].

حيث تسـمى مجموعة جميع النتائج المحتملة فضـاء العينة وبالتالي فإن احتمال P(R) في فضـاء العينة يمكن أن يحمل إحدى قيمتين {ملائمة Relevant ، غير ملائمة Irrelevant} حيث يكون R متحول عشوائي يأخذ إحدى قيمتين {0,1} حيث Irrelevant=0 ، Relevant=1 [13].

ولابد أن نذكر أن القيمة P(…) تتغير بتغير المتحول العشـــوائي R فعندما نعين قيم مختلفة للمتحول العشــوائي R أو قيم مختلفة لفضـــاء العينة فإننا بالتالي حكماً نتحدث عن قيم مختلفة للاحتمال P [13]. فهذا النموذج يعتمد على طريقة استخدام نظرية الاحتمالات كأساس لعملية المعالجة حيث يتم وفقاً للنموذج الاحتمالي المتبّع إحصـــاء أو تقدير الاحتمالات التي تكون فيها الوثيقة ذات صـــلة باســـتعلام المســـتخدم. فنموذج الاحتمالات يعمل على تحليل المصطلحات في كل من الاستفسارات والوثائق من النواحي النحوية Syntactic أو الدلالية Semantic أو الواقعية Pragmatic، ومن ثم ترتب الوثائق المســـترجعة ترتيباً تنازلياً وفقاً لاحتمالات صلتها بالاستعلام [16].

فالفكرة الأساسية لهذا النموذج تتمثل في فرضية احتمال أن نظام استرجاع المعلومات يشتمل على وثائق تتصل باستفسار المستفيد تمام الصلة وهناك مجموعة أخرى بمنأى عن هذه الصلة، فوفقاً لهذا النموذج تسمى مجموعة الوثائق ذات الصـــلة بمجموعة الجواب المثالي ideal answer set، وبتوفير توصـــيف كامل لهذة المجموعة من الوثائق (مجموعة الجواب المثالي) تتضـــاءل مشـــاكل اســـترجاع محتوى الوثائق، ورغم ذلك تظهر عقبة أخرى في صعوبة معرفة ماهية هذة الخصائص والسمات بشكل قاطع [17].

فالجهد الأســـاســـي لهذا النموذج يتمثل في التخمين الأولي لتحديد خصـــائص الكلمات المفتاحية الواردة في الوثائق والتي تحظى بدلالات لغوية واصطلاحية تساهم في وسم هذه الخصائص مما يسمح بإنشاء وصف أولي احتمالي لمجموعة الجواب المثالية على الاستفسار من الوثائق [17].

وعلى الرغم من النتائج التي حققها هذا النموذج إلا أن البعض رأى بأنه ليس أفضـــل من النتائج التي حققتها النماذج الكلاســـيكية، الأمر الذي أدى من وجهة نظرهم إلى عدم اقتناع مطوري النظم بالتحول إلى هذا النموذج بدرجة كبيرة [16].

ومن النماذج الاحتمالية نذكر:

١. النموذج Best Match(BM25) :

تم تطوير هذا النموذج في ســـبعينيات وثمانينيات القرن العشـــرين بواســـطة كلٍّ من Stephen E. Robertson, Karen Spärck Jones وآخرين.

وهو من النماذج الاحتمالية المســتخدمة على نطاق واســع حالياً، حيث يتم تصــنيف المســتندات بناءً على الاحتمال المقدر لملاءمة المستندات في الاستعلام [15].

ولقد تم تطوير هذا النموذج من نموذج الاستقلال الثنائي (BIM) وهو نموذج احتمالي كلاسيكي يمثل كلاً من المستندات والاستعلامات كمتجهات ثنائية من خلال دمج تردد المدى والتطبيع بالنسبة إلى طول المستند Document Length Normalization [15].

ويشار إلى هذا النموذج عادة باسم Okabi BM25 لأن نظام الاسترجاع Okabi والذي تم تنفيذه في مدينة لندن في الثمانينات كان أول نظام قام بتطبيق هذا النموذج.

إن نموذج BM25 هو نموذج احتمالي كما ذكرنا إلا أن لديه الكثير من العوامل المشتركة مع خوارزمية ترجيح المصطلحات TF-IDF فكلتا الخوارزميتين تستخدم تردد المصطلح وتردد الوثيقة العكسي إلا أن تعريف العوامل Factors يختلف قليلاً بين النموذجين [22].

فكلا النموذجين يعرّف الوزن الذي يعطى لكل مصطلح كنتيجة لجمع التردد العكسي للوثيقة و تردد المصطلح ومن ثم حساب وزن المصطلح لكامل المستند بالنسبة للاستعلام المحدد ويعتبر الاختلاف الأبرز بين النموذجين هو أن BM25 يأخذ بعين الاعتبار طول الوثيقة في حين أنه ليس له أي تأثير في الطريقة التقليدية TF-IDF المستخدمة في النموذج الكلاسيكي [22].

٢. نموذج اللغة Language Models:

تم تطبيق نماذج اللغة على استرجاع المعلومات من قبل عدد من الباحثين في أواخر التسعينيات نذكر منهم:(Ponte and Croft 1998, Hiemstra and Kraaij 1998, Miller etal. 1999) [13].

وهي تعتبر من النماذج الاحتمالية والتي تم تطويرها لأنظمة التعرف التلقائي على الكلام في أوائل الثمانينيات، فهو يدرس التوزيع الاحتمالي على جميع تسلسلات الكلمات في اللغة [13] [15].

حيث يُقدر نموذج اللغة احتمالية تسلسلات الكلمات ويكون هناك نموذج لغوي مرتبط بكل وثيقة وقد يحتوي هذا النموذج على الاستعلامات الأكثر صلة به. وتعتبر الأساليب القائمة على نموذج اللغة من النماذج المستخدمة على نطاق واسع [15].

فنموذج اللغة هو توزيع احتمالي على جميع سلاسل اللغة. بعبارة أخرى، يعين نموذج اللغة احتمالية لكل كلمة/مصطلح في اللغة. لذلك، فإن المهمة الأساسية في هذا النموذج هي بناء دالة توزيع احتمالية. فبمجرد معرفة احتمالات الكلمات الفردية فإنه يمكننا حساب الاحتمالات لأي عبارة أو جملة في اللغة. وكلما زاد احتمال الجملة زادت احتمالية أن تكون الجملة صحيحة في اللغة. على سبيل المثال، افترض أن

الاحتمالات المرتبطة بالكلمات (المعلومات ، الاســـترجاع ، النماذج) هي (٠,١- ٠,١٥ - ٠,٠٥) على التوالي. عندها يكون احتمال عبارة: نماذج استرجاع المعلومات هو ٠,٣ [15].

وقد تحدد بعض النماذج الاحتمالية الأخرى احتمالاً كبيراً للغاية لكلمة "استرجاع"، مما يشير إلى أن احتمال هذه الرسالة التي نقوم بكتابتها على سبيل المثال ستكون مرشحة بقوة للاسترجاع إذا كان الاستعلام يحتوي على هذه الكلمة [14] [13].

وتأخذ نماذج اللغة نفس نقطة البداية التي يتبعها نموذج الفهرسة الاحتمالية والذي وضع من قِبل (Bill Maron and Larry Kuhns) وهو أن يتم تعيين قيمة احتمال لمصطلحات الفهرسة المختلفة والتي يحتويها المستند فيكون لكل مستند مجموعة من مصطلحات الفهرسة ولكل مصطلح قيمة احتمالية تحدد مدى أهميته بالنسبة للاستعلام بما يحتويه من مصطلحات. فيتم تصميم نموذج اللغة لكل وثيقة باتباع هذا النهج [14] [13].

ومن النماذج المتطورة والتي ظهرت بالاعتماد على هذا التقنية هو نموذج معالجة اللغة الطبيعية Natural Language Processing Model.

حيث لا يتم الاعتماد على مصطلحات الاستفسار والوثيقة فقط ولكنه يعالج الجمل والصيغ ويعمل على مضاهاتها ويتطلب بناء النظم التي تعمل على معالجة نصوص اللغة الطبيعية ثلاثة مستويات من المعالجة هي:

- **التحليل النحوي (Syntactic analysis)**: يتطلب فهم بناء الجمل ويعتمد على الكلمات التي تضمنتها القواميس (Lexicon) والمعلومات المقترنة بها مثل القواعد والعلامات النحوية.
- **التحليل الدلالي (Semantic analysis)**: يتعامل مع معاني الكلمات في الجمل وفقاً لما هو مخزن في قاعدة المعرفة.
- **التحليل الواقعي أو العملي (Pragmatic analysis)**: يأخذ في الاعتبار السياق الذي جاءت فيه المصطلحات[20] .

نماذج الجمع بين الأدلة Combining Evidence:

في هذه النماذج يتم استخدام تقنية " فهم المحتوى" لكل من المستندات والاستعلامات ومن ثم استخدامها لاستنتاج العلاقات المحتملة بين المستندات والاستعلامات، فعملية استرجاع المعلومات هي عملية

الاستدلال أو التفكير المنطقي التي نقدر فيها احتمال مدى ملاءمة المستند للاستعلام الذي يحدد حاجة المستخدم.

من هذه النماذج:

١. شبكة الاستدلال Inference Network:

في هذا النموذج تتم نمذجة استرجاع الوثائق كعملية استدلال منطقية، تقدر من خلالها احتمالية تلبية حاجة المستخدم من المعلومات والتي يتم التعبير عنها من خلال واحد أو اكثر من الاستعلامات وذلك من خلال تحليل المستند حيث أن شبكة الاستدلال ستكون هي الآلية لاستنتاج هذه الأنواع من العلاقات [23].

ويمكن تطبيق معظم التقنيات المستخدمة بواسطة أنظمة استرجاع المعلومات ضمن هذا النموذج [2].

وفي أبسط تطبيق لهذا النموذج تقوم الوثيقة بإعطاء المصطلح قوة معينة ومن ثم يتم جمع القيم للمصطلحات الواردة ضمن الاستعلام لحساب النتيجة الرقمية للاستعلام بالنسبة للوثائق.

وبوصف آخر يمكن اعتبار القوة التي تعطى للمصطلح على أنها وزن المصطلح في الوثيقة. وبالتالي يصبح تصنيف المستندات في هذا النموذج مشابهاً للترتيب في نموذج الفضاء الشعاعي أو النماذج الاحتمالية. وإن قوة المصطلح غير محددة وبالتالي يمكن استخدام أي خوارزمية أو شكل لقوة المصطلح ضمن الوثيقة أو الاستعلام [2].

وإن معظم الأبحاث الاخيرة في مجال استرجاع المعلومات تشير إلى أن تحسين عملية الاسترجاع تتطلب تقنيات تساعد في فهم محتوى كل من الوثائق والاستعلامات والتي يمكن استخدامها لاستنتاج العلاقات المحتملة بين الوثائق والاستعلام [23].

ولقد تم استخدام تمثيلات الشبكة في استرجاع المعلومات في أوائل الستينيات كما أن الفكرة القائلة بأن عملية الاسترجاع هي عملية استدلال منطقية ليست جديدة فالاسترجاع المنطقي لـــ Cooper يعتمد على استنتاج العلاقات بين تمثيل المستندات والحاجة للمعلومات [23].

ويعتمد هذا النموذج على ثلاث أمور أساسية:

١. دعم استخدام مخططات تمثيل المستندات المتعددة حيث أظهرت الأبحاث أن استعلاماً معيناً سيسترد مستندات مختلفة عند تطبيقه على تمثيلات مختلفة.

٢. السماح بدمج النتائج من أنواع استعلامات مختلفة. فيمكن استخدام وصف المعلومات المطلوبة لإنشاء العديد من تمثيلات الاستعلام (على سبيل المثال: الاحتمالية، المنطقية)، كل منها يستخدم استراتيجية استعلام مختلفة وكل منها يلتقط جوانب مختلفة من المعلومات المطلوبة. وبالتالي استرداد وثائق مختلفة لنفس الحاجة المحددة للمعلومات.

٣. المطابقة المرنة بين المصطلحات أو المفاهيم المذكورة في الاستعلامات وتلك المعينة للمستندات. يبدو أن المطابقة الضعيفة بين المفردات المستخدمة للتعبير عن الاستعلامات والمفردات المستخدمة لتمثيل المستندات هي سبب رئيسي لبيانات مسترجعة قليلة وبالتالي يمكن تحسين الاستدعاء باستخدام المطابقة المعرفية لمفاهيم الاستعلام والوثائق وتمثيلاتها دون أن يؤدي ذلك إلى تدهور الدقة بشكل كبير [23].

٢. تعلم الترتيب Learning To Rank:

تعد خوارزمية تعلم الترتيب الخوارزمية جزءاً من استرجاع المعلومات الخاص بالمستندات الكبيرة. تتكون البيانات من الاستعلامات والوثائق التي يتم تمثيلها كأشعة [15].

وهي مقسمة إلى ثلاثة نماذج (نقطية Pointwise، زوجية Pairwise، قائمة Listwise). في النموذج الأول يتم الترتيب كعملية تصنيف تقليدية فتكون النتيجة صنف Class وبالتالي يكون الهدف هو تقليل التصنيف الخاطئ للاستعلامات والوثائق. وفي النموذج الثاني Pairwise وهي عملية تحويل الترتيب إلى عملية تصنيف نقطية والهدف من هذه العملية هو زيادة عدد الازواج والتي صنفت خارج الترتيب وأما الثالث فهو مشابه تماماً للنموذج الزوجي إلا أنه يتعامل مع قوائم من الصفوف والفئات [15] [24].

حيث يتم تطبيق هذا النموذج على مجموعة اختبار Training Set وفرز الوثائق وفقاً لدرجة ارتباطها وأهميتها [15] [24].

ويمكن استخدام العديد من النماذج في عمليات الترتيب المستنتجة والتي تختلف باختلاف النموذج المستخدم في عملية استرجاع المعلومات.

مقارنة بين النماذج الرئيسية لأنظمة استرجاع المعلومات:

من خلال استعراض النماذج السابقة نجد أن النماذج الأساسية التي تعتمد عليها بقية النماذج الأخرى هي ثلاثة: النموذج المنطقي والنموذج الشـــعاعي والنموذج الاحتمالي وجميع النماذج الأخرى ما هي إلى محاولات لتطوير هذه النماذج الأساسية الثلاث.

ولقد وجدنا أن نموذج الاسترجاع المنطقي يتيح للمستخدمين صياغة عبارات منطقية معقدة. ومع ذلك، قد يكون إنشاء استعلامات منطقية صعباً بالنسبة للمستخدم العادي وتعتبر جميع المصطلحات التي تم إدخالها في الاســـتعلام ذات أهمية مماثلة نظراً للطبيعة الثنائية للنتائج. ولا يوفر هذا النموذج ترتيباً للمســـتندات التي تم اســـترادها. وســـتكون مجموعة المســـتندات التي تم إرجاعها إما فارغة تقريباً (وهو استدعاء منخفض لأن العديد من المستندات ذات الصلة لن يتم استردادها) أو ستتضمن العديد من المستندات (وبالتالي ستكون الدقة منخفضة حيث سيكون المستند غير ذي صلة أيضاً في المجموعة) بســـبب اســـتخدام معيار المطابقة التامة. إن هذا النموذج أكثر فائدة لاســـتعادة البيانات من اســـترجاع المعلومات لأن جميع المصطلحات متساوية في الوزن [25].

بينما وجدنا في نموذج الاسترجاع الشعاعي VSM أنه يمكن تطبيق مجموعة من القيم على كل مصطلح سواء في تمثيل المستندات أو في استعلام المستخدم. بالإضافة إلى ذلك، لا يُفضل تطبيع المستندات الطويلة بسبب استخدام قيمة تردد المستند المعكوس في ترجيح المصطلحات في هذا النموذج فإنه لا تعتبر المصطلحات الشائعة مهمة بينما يتم إعطاء الأهمية للمصطلحات النادرة.

ففي نموذج VSM إن الوثيقة الطويلة التي يمكن أن تحتوي على نفس المصـــطلحات التي وردت في الاستعلام فقط في العنوان والملخص فقد تكون ذات صلة كبيرة جداً بالاستعلام إلا أنها في هذا النموذج ستكون ذات أهمية منخفضة مقارنة بمستند قصير يحتوي على نفس المصطلحات في التذييل. ويمكن للتطبيق الأكثر تقدماً حساب أهمية المصطلحات بشكل مختلف، على سبيل المثال عن طريق تفضيل المصـــطلحات التي تظهر في بداية المســـتند وهذا ما يعتبر أحد عيوب هذا النموذج. كما أن هناك عيب آخر في تمثيل مستندات VSM هو أن ترتيب المصطلحات مفقود ولا يمكن تفضيل المستندات التي تحتوي على مصطلحات استعلام قريبة من بعضها البعض على المستندات التي تحتوي على مصطلحات منفصلة في أجزاء مختلفة من المستند [25].

وأما نموذج الاسترجاع الاحتمالي فيعتمد على الافتراضات التي تم إجراؤها بشكل صريح، مثل افتراض أن ٥٠٪ من المستند الذي يحتوي على مصطلح وثيق الصلة بهذا المصطلح. ولكن ليس بالضرورة أن تتناسب كل الافتراضات مع الواقع. لذلك يجب تخمين العدد الإجمالي للوثائق ذات الصلة كما أن حساب قيمة الاحتمال P(..) والذي هو ثابت لا يكون دائماً صحيحاً. لذلك يتطلب نموذج الاسترجاع الاحتمالي لتحقيق نتائج دقيقة أن تكون المصطلحات مستقلة فهو يتجاهل حساب الوزن لتكرار المصطلح والموضع داخل المستندات، وبالتالي فهو مناسب للوثائق الطويلة أكثر منه للمستندات القصيرة [25].

وفيما يلي جدول يلخص أهم الفروق بين نماذج الاسترجاع الرئيسية الثلاثة:

	المنطقي	الشعاعي	الاحتمالي
الإيجابيات	- نموذج بسيط وغير معقد وبالتالي فهو سهل التطبيق والتحقيق. - يمكن التنبؤ به وسهل الشرح. - يشعر الخبراء بقدرة أكبر على التحكم بالنظام.	- أبسط نموذج يعتمد على الجبر الخطي. - يعتمد على توزين المصطلحات - يعتمد على حساب درجة التشابه بين الاستعلامات والوثائق - المطابقة الجزئية ممكنة	- نموذج فعال. - نموذج رياضي ونظري. - مناسب للوثائق الطويلة.
السلبيات	- يجب أن تكون مفردات الفهارس هي نفسها مفردات الاستعلام فهو يستخدم المطابقة التامة وليس هناك إمكانية لتطبيق المطابقة الجزئية. - الوثائق المستعادة غير مرتبة أو مصنفة. - لا يوجد ترجيح لمصطلحات الفهرس أو الاستعلام - صعوبة بناء استعلامات منطقية إذا كانت طويلة.	- يفترض أن المصطلحات مستقلة إحصائياً من الناحية النظرية. - يتم تمثيل الوثائق الطويلة بشكل سيء وبالتالي تعتبر القدرة التعبيرية له محدودة. - يجب أن تكون الكلمات المفتاحية مطابقة تماماً لمصطلحات الوثيقة. - يفتقد إلى البنية اللغوية لتمثيل الميزات اللغوية الهامة.	- الاحتمالات صعبة التقدير - افتراضات غير واقعية بسبب استقلالية المصطلح - العلاقات المنطقية مهملة - وجود العديد من النماذج وبالتالي صعوبة تحديد أفضلها لأنه بحاجة إلى معرفة مسبقة.

الجدول ٢ – مقارنة بين نماذج استرجاع المعلومات الأساسية

التقييم في أنظمة استرجاع المعلومات:

إن نماذج استرجاع المعلومات بشكل عام تتكون من أربعة عناصر هي:

- الوثائق Documents والتي تمثل مجموعة تتكون من تمثيلات الوثائق في المجموعة.
- الاستعلامات Queries وتمثل مجموعة تتكون من تمثيلات احتياجات المستفيدين من المعلومات.
- الإطار Framework وهو إطار لنمذجة تمثيلات الوثائق والاستعلامات والعلاقة بينهما.
- الترتيب Ranking وهي وظيفة تشترك فيها تمثيلات الوثائق والاستعلامات حيث يتم تحديد ترتيب الوثائق في النتيجة حسب ما جاء في الاستعلام [16].

وفعالية الاسترجاع في أنظمة استرجاع المعلومات هي مقياس لمدى تلبية المستندات التي يتم استردادها بواسطة النظام لاحتياجات المستخدمين ويشار إلى عملية تحديد فعالية الاسترجاع لاستعلام معين باسم تقييم الفعالية [15].

إن قياسات فعالية أنظمة استرجاع المعلومات هي (الدقة Precision وهي النسبة المئوية للوثائق ذات الصلة بالنسبة للمجموعة المستردة، الاستدعاء Recall النسبة المئوية للوثائق ذات الصلة في المجموعة المستردة بالنسبة لجميع المستندات) [15] [26].

إن هذه القياسات مستخدمة على نطاق واسع في تقييم أداء أنظمة استرجاع المعلومات في تصنيف النصوص [26].

ومع ذلك فإن التقييم من خلال هاذان العاملان بمعزل عن بعضهما البعض لا معنى له الدقة والاستدعاء يعارضان بعضهما البعض ويترابطان عكسياً لأن الحصول على مستويات أعلى من الدقة يكون من خلال قيم استدعاء منخفضة [15] [26].

حيث أن أنظمة استرجاع المعلومات الفعالة يجب أن تسترد أكبر عدد ممكن من المستندات ذات الصلة (أي لديها استدعاء عالي)، ويجب أن تسترد عدداً قليلاً جداً من المستندات غير ذات الصلة (أي ذات دقة عالية). ولسوء الحظ فقد أثبت هذان الهدفان أنهما متناقضان تماماً على مر السنين [2].

لذلك استخدم في بعض أنظمة التقييم مقياس F1 كمقياس للجمع بين الدقة والاستدعاء [26].

وفي هذا المجال يتم استخدام التعابير التالية:

- True Positive TP: عدد المستندات التي تنتمي إلى الفئة C والتي تم تصنيفها بشكل صحيح من قبل النظام لتكون في الفئة C.

- False Positive FP: هو عدد المستندات التي لا تنتمي إلى الفئة C والتي تم تصنيفها بشكل غير صحيح من قبل المصنف.
- False Negative FN: عدد المستندات التي تنتمي إلى الفئة C والتي تم تحديدها بشكل غير صحيح من قبل المصنف.
- True Negative TN: عدد المستندات التي لا تنتمي إلى الفئة C والتي تم تصنيفها بشكل صحيح من قبل المصنف بحيث لا تكون ضمن الفئة C [26].

وهذه القيم يمكن تمثيلها ضمن الجدول التالي:

	المستندات ذات الصلة Relevant	**المستندات غير ذات صلة** Non_relevant
المستندات التي تم استدعاؤها Retrieved	TP	FP
المستندات التي لم يتم استدعاؤها Not Retrieved	FN	TN

الجدول ٣ – معنى التعابير المستخدمة في أنظمة تقييم استرجاع المعلومات

وفيما يلي نوضح طرق حساب كل من الـ Recall، Precision، F1:

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

$$= \frac{2 * \mathrm{TP}}{2 * \mathrm{TP} + \mathrm{FN} + \mathrm{FP}}$$

ملخص الفصل الأول

بعد استعراض النماذج المختلفة لأنظمة استرجاع المعلومات وجدنا أن النموذج الاكثر ملاءمة والذي سيتم دراسته في بحثنا هو النموذج الشعاعي أو ما يعرف باسم نموذج حيز المتجهات باعتبار أنه يمثل واحداً من النماذج الأكثر انتشاراً حتى اليوم والتي تعتمد نتائجه بشكل كبير على عملية توزين المصطلحات حيث تعتبر المشكلتين الرئيسيتين في النموذج الشعاعي:

١. استقلالية المصطلحات.

٢. ترجيح المصطلحات [15].

وبالتالي سنحاول من خلال هذا البحث تجاوز هذه النقاط دون أن ندخل في تعقيدات العمليات الحسابية للنموذج الاحتمالي.

وذلك من خلال تحديد واصفات للمصطلحات في الوثائق بحيث تعطي تلك الواصفات مؤشرات كمية أو نوعية تحدد قيمة المعلومات فيها ومدى أهميتها للوثيقة ومن ثم يتم تحديد قيمة تمثل درجة عضوية المصطلح في الوثيقة بناء على هذه الواصفات وهو ما يسمى بعملية ترجيح المصطلحات.

حيث سنقوم بداية بإجراء مجموعة من عمليات المعالجة الأولية لنصوص المستندات ومن ثم تطبيق وإجراء الاختبارات على خوارزميات ترجيح المصطلحات المختلفة والمقارنة بين النتائج للوصول إلى خوارزمية جديدة لترجيح المصطلحات تأخذ بعين الاعتبار النقاط الهامة التي أغفلتها طريقة ترجيح المصطلحات التقليدية حيث سنعمل جاهدين على أن تأخذ الطريقة الجديدة في الاعتبار العلاقة بين المصطلحات وورودها ضمن الوثائق وموقعها النحوي.[٣]

[٣] تم نشر معلومات الفصل الأول كورقة علمية بعنوان مقارنة بين نماذج استرجاع المعلومات الأساسية بتاريخ ٢٠٢١/٩/٢٣ في مجلة: International Journal Of Engineering Research & Technology (IJERT) والورقة مرفقة في نهاية البحث.

الفصل الثاني: تحليل النصوص ومعالجة اللغة الطبيعية

مقـــدمة:

تعتبر الحاجة إلى تحليل النصـوص قبل اسـترجاعها واحدة من أكبر العقبات التي تواجهها نظم اسـترجاع المعلومات لأن عملية استرجاع المعلومات تعتمد بشكل كبير على فهم محتوى النصـوص المراد استرجاع المعلومات منها وتحليل الكلمات التي يتم اســتخدامها لبناء اســتعلامات البحث Queries ومن ثم إجراء عملية ربط معنوي بين الكلمات المفتاحية وبين قاعدة البيانات النصية للوصول إلى المستند الصحيح بوقت قياسي ومن هنا ظهرت الحاجة إلى عملية التحليل النصي.

في النموذج الشــعاعي يتم تمثيل المصــطلحات على أنها مجموعة من المفاهيم التي يتم تمثيلها كشــعاع حيث يمثل الشعاع الكلمات المفتاحية والمصطلحات التي تم استخراجها من الوثائق.

ويمكن القول بأن العمليات في النموذج الشــعاعي تنقســم إلى مراحل ثلاث: المرحلة الأولى هي الفهرســة وفيها يتم استخلاص المصطلحات من المستند النصي، المرحلة الثانية هي ترجيح المصطلحات المفهرسة والمرحلة الأخيرة هي تصنيف الوثيقة بالنسبة للاستعلام ووفقاً للتشابه [27].

حيث لا تعمل أنظمة البحث الفعالة بشكل مباشر مع المستندات أو الاستعلامات وإنما يتم استخدام تقنيات واستراتيجيات مختلفة لتمثيل المعنى الرئيسي على هيئة أجزاء من المستندات أو الاستفسارات وهذه العملية تسمى الفهرسة [28].

فبعد جمع النص الذي نريد البحث عنه، فإن الخطوة التالية هي تحديد ما إذا كان ســـيتم تعديلها أو إعادة هيكلتها بطريقة ما لتبســـيط البحث. وتســـمى التغييرات التي يتم إجراؤها في هذه المرحلة عملية تحويل النص Text transformation أو معالجة النصوص Text processing [29].

فعملية الفهرســة هي عملية تحديد الكلمات الرئيســية لتمثل وثيقة بناءً على محتوياتها. وتعتبر مرحلة مهمة جداً من نظام استرجاع المعلومات.

وتعرّف الفهرسة بأنها عملية تحديد الكلمات الرئيسية أو المصطلحات الوصفية أو ما يطلق عليها كلمات مفتاحية Index Term والتي تمثل الوثيقة بناء على محتوياتها بهدف الوصول الفعال للوثائق [30].

ومن هنا نجد أن عملية معالجة وتحليل النصوص تعتبر الخطوة الأولى لعملية الفهرسة في أنظمة استرجاع المعلومات من أجل فعالية عملية الاسـترجاع ولابد لتحقيق ذلك من وجود بنية مناسـبة للفهرس وإن هيكل البيـانـات الأكثر اســـتخـدامـاً هو الفهرس المقلوب Inverted Index وهو عبـارة عن آليـة موجهـة للمصطلحات.

إن بنية الفهرس المقلوب يحتوي على مكونين أساسيين: المفردات وقائمة الوثائق، والمفردات هي عبارة عن مجموعة المصـــطلحات المختلفة المســـتخرجة من الوثائق. حيث يتم تمثيل كل وثيقة بقائمة من بعض الكلمات المرجعية ويتم تخزينها أبجدياً في ملف [29] [31]. وتعتبر الفهارس المقلوبة من أكثر هياكل الفهارس كفاءة ومرونة [29].

كما يمكن تخزين معلومات إحصـــائية عن كل مصـــطلح في كل وثيقة مثل تكرار المصـــطلح وموضـــع المصطلح وغيرها من الميزات التي تفيد في عملية الاسترجاع.

أي أنه يمكننا القول بأن الفهرسـة تتضـمن تعريف كل مسـتند بالكلمات الرئيسـية والمصـطلحات التي تمثله وتخزينها بهدف إنشاء الفهرس المناسب [30].

ولإنشاء الفهرس يمر النص بمراحل تحليل ومعالجة ضمن أنظمة استرجاع المعلومات نذكر منها:

التحليل اللغوي (استخراج الرموز) Token Extraction:

في هذه المرحلة يتم تحليل النص واختيار المصـــطلحات المميزة وإزالة علامات الترقيم والرموز غير الضرورية وغالباً ما يتم الإشارة إلى هذه العملية باسم Tokenization.

حيث يتم تقسيم المستند إلى وحدات تسمى الرموز Tokens وينتج عن هذه العملية مجموعة من الكلمات ذات المعنى الدلالي [32].

ولابد لنا من التمييز بين مصطلحين هما: Term ، Token.

Token: الرمز المميز وهو عبارة عن تمثيل لسلسلة من الأحرف في مستند معين تم تجميعها معاً كوحدة دلالية مفيدة للمعالجة [11] [29].

أما الـ Term: فهو عبارة عن Token تم معالجته ليكون في قاموس نظام IR.

والسؤال الأهم الذي نطرحه في هذه مرحلة هو: ما هي الرموز الصحيحة التي ينبغي على النظام معالجتها وتخزينها؟ [11].

التجذيع والتجذير Stemming & Lemmatization:

يطلق على هذه العملية أيضاً اسم التجريد عامة، وتشير عادة مصطلحات التجريد والتجذيع والتجذير إلى عملية تغيير بنية الكلمة واختزال شكل المصطلحات إلى شكل مشترك.

فالتجذيع Stemming تشير إلى اقتطاع جزء من نهاية الكلمة وإزالة اللواحق المشتقة منها أي إزالة أي إضافات من الكلمة في حين أن التجذير Lemmatization يعتمد على التحليل الصرفي للكلمات لإزالة النهايات التصريفية فقط وإعادتها إلى الشكل الأساسي للكلمة كما ورد في القاموس اللغوي وهو ما يعرف باسم الجذر Lemma [11] [32].

الهدف من التجريد أو التجذيع هو تقليل الأشكال المختلفة للكلمة التي تتولد بسبب التصريف وأحياناً الأشكال الاشتقاقية ذات الصلة بالكلمة إلى صيغة مشتركة [11] [29].

على سبيل المثال:

am, are, is, was ⇒ be
car, cars, car's, cars' ⇒ car

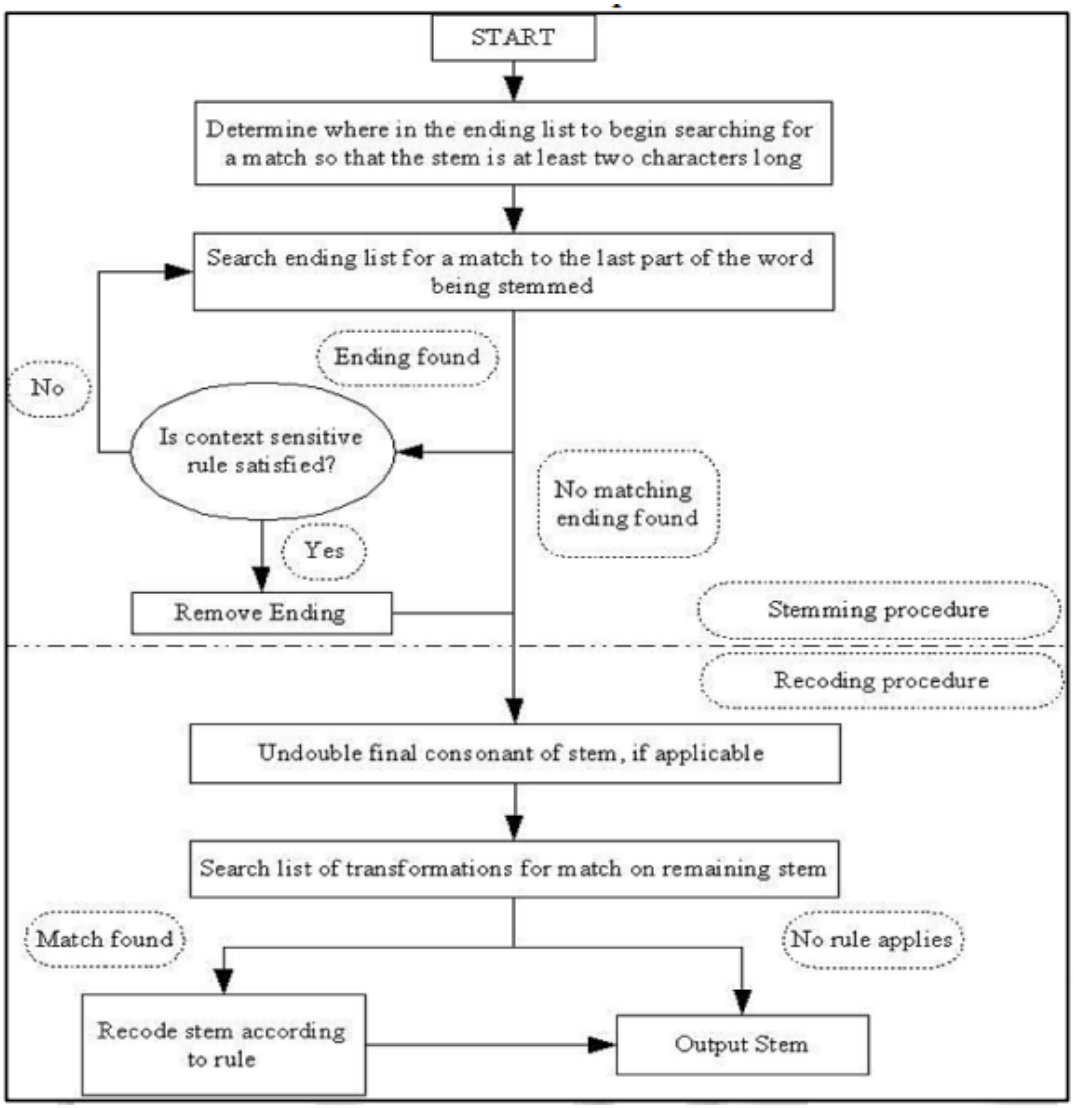
فعندما يحدد المستخدم مصطلحاً للبحث فإنه من الضروري استرجاع الوثائق التي تحتوي على المتغيرات النحوية لهذا المصطلح الأمر الذي يمنع التطابق التام بين مصطلح الاستعلام والمستند الذي يحوي على هذا المصطلح.

فمن خلال هذه العملية يتم تمثيل الأشكال النحوية للمصطلحات في صيغة أساسية مشتركة كما يساعد على تقليل حجم المستندات وزيادة سرعة عملية البحث من خلال البحث عن المصطلح المجرّد عوضاً عن البحث عن كامل المصطلح. وعادةً ما يكون للمصطلحات ذات الجذع المشترك معاني متشابهة، على سبيل المثال:

Connect, Connected, Connecting, Connection, Connections

وهي ذات معنى واحد ومرتبطة مع بعضها فإذا ماتم دمج مثل هذه الكلمات في مصطلح واحد فإن هذا سيؤدي إلى تحسين أداء نظام IR وذلك من خلال إزالة اللواحق -ING, -ED, -IONS, -ION.

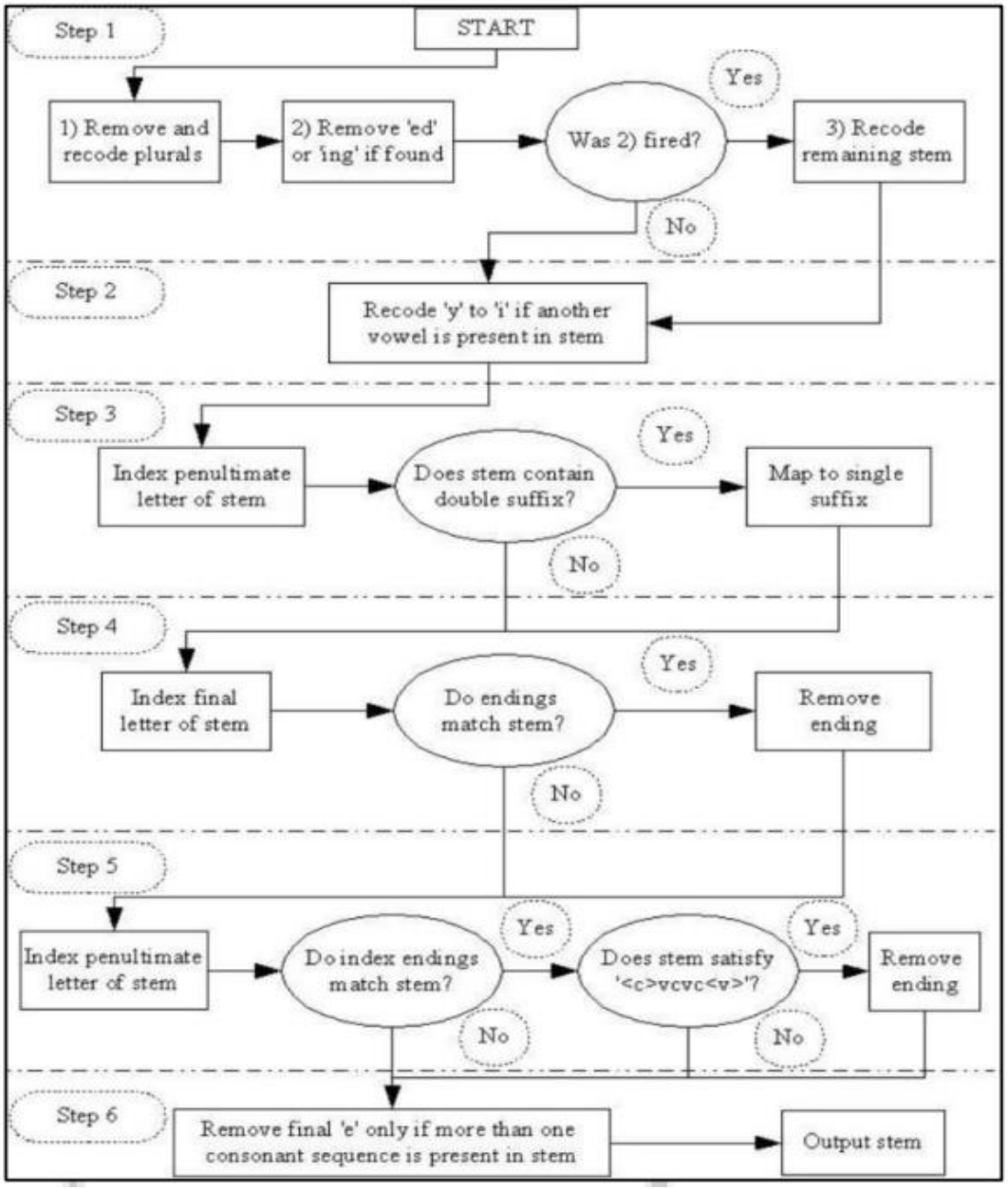
وكان أول مجرّد للغة الانكليزية قد طُوِّر عام ١٩٦٨ من قبل Julie Beth Lovins والذي قدّم فكرة التجريد بالاعتماد على قاموس اللواحق Suffix الشـــائعة. حيث اعتمدت هذه الخوارزمية على مبدأ إزالة اللواحق بناء على التطابق الأطول. ولقد أدت هذه الخوارزمية إلى نتائج معقولة في مجال اســـترجاع المعلومات [33].



الشكل ٢ – مخطط عمل المجرّد Lovins [33]

ثم جاء بعده Martin Porter حيث نشـــر ورقة بحثية عام ١٩٨٠ في مجلة Program ليُوصّـــف خوارزمية بســـيطة جداً من حيث المفهوم تتحكم بهذه الخوارزمية قواعد لتحديد فيما إذا كانت اللاحقة يجب أن تحذف أو لا معتمداً على الحد الأدنى المتبقي بعد إزالة اللاحقة. ولقد أثبتت هذه الخوارزمية بشـــكل متكرر أنها فعالة جداً من الناحية التجريبية [11].

وفيما يلي مخطط يوضح طريقة عمل خوارزمية Porter:

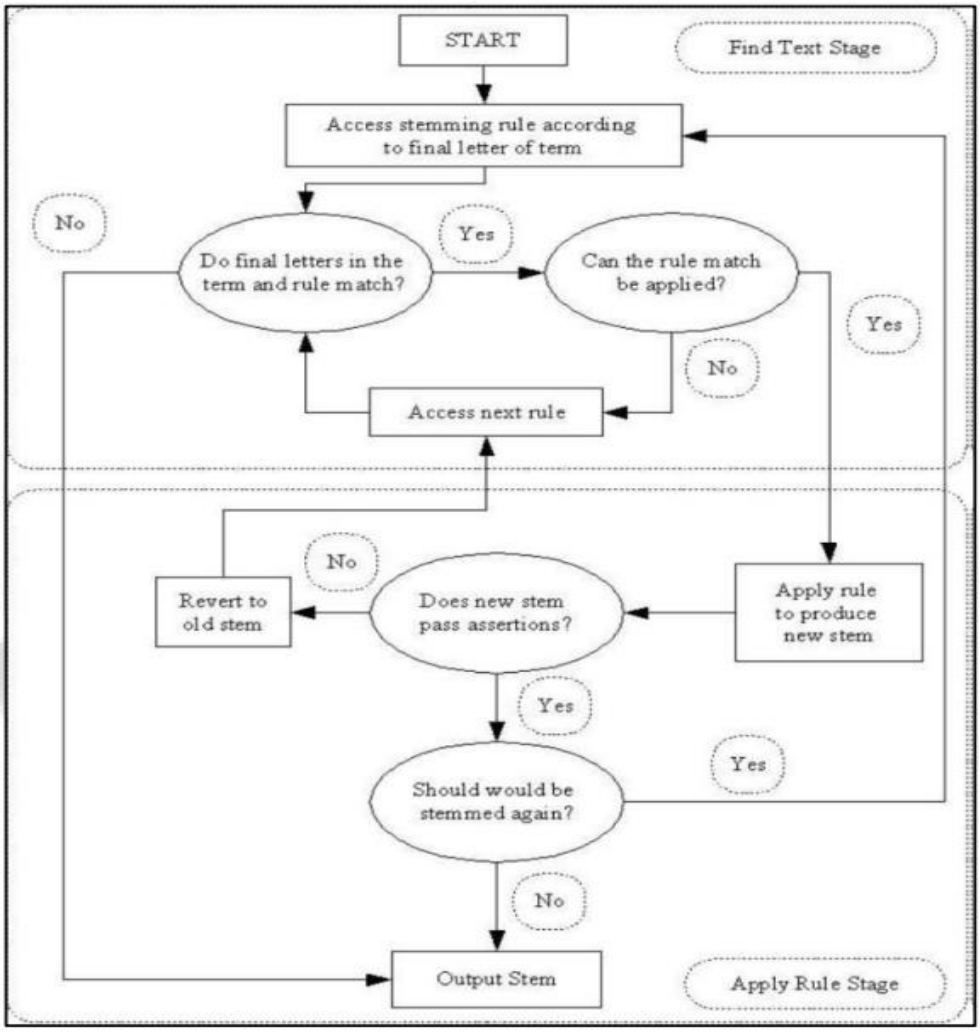


الشكل ٣ – مخطط عمل المجرّد Porter [33]

وأصبح المجرد الذي طوّره Porter واسع الانتشار ومستخدماً في الكثير من الخوارزميات التي تعالج اللغة الإنكليزية حيث لم يعتمد في خوارزميته على قاموس تجريد وإنما اســـتخدم قوائم Suffix وتم ربط كل Suffix بمعيار لحذف هذا الـ Suffix من الكلمة لكي نحصل عند تطبيق هذا المعيار على كلمة مجرّدة صـــحيحة. وتعتمد هذه الخوارزمية على عدة مراحل تحتوي كل واحدة منها على مجموعة من القواعد لإزالة اللواحق Suffix وهي متوفرة لعدة لغات إلا أنها لا تتضمن اللغة العربية [29].

وفي عام ١٩٩٠ نُشـــر لأول مرة المجرِّد Paice/Husk stemmer والذي تم تطويره من قبل Chris Paice بمسـاعدة Gareth Husk حيث يسـتخدم جدولاً من القواعد المفهرسـة والتي تحدد فيما إذا سـيتم حذف أو استبدال اللواحق [33].

والمخطط التالي يبين خطوات تسلسل العمل في هذا المُجرّد:



الشكل ٤– مخطط عمل المجرّد Paice [33]

فإذا كان لدينا النص التالي:

> Such an analysis can reveal features that are not easily visible from the variations in the individual genes and can lead to a picture of expression that is more biologically transparent and accessible to interpretation.

سيتم تجريده على النحو التالي:

Original Text	*Lovins Stemmer*	*Porter Stemmer*	*Paice Stemmer*
such	such	such	such
an	an	an	an
analysis	analys	analysi	analys
can	can	can	can
reveal	reve	reveal	rev
features	feature	feature	feat
that	that	that	that
are	ar	ar	are
not	not	not	not
easily	eas	easily	easy
visible	vis	visibl	vis
from	from	from	from
the	th	the	the
variations	vari	variat	vary
in	in	in	in
the	th	the	the
individual	individu	individu	individ
genes	gen	gene	gen
and	and	and	and
can	can	can	can
lead	lead	lead	lead
to	to	to	to
a	a	a	a
picture	picture	picture	pict
of	of	of	of
expression	expres	express	express
that	that	that	that
is	is	is	is
more	mor	more	mor
biologically	biology	biolog	biology
transparent	transpar	transpar	transp
and	and	and	and
accessible	acces	access	access
to	to	to	to
interpretation	interpres	interpret	interpret

الجدول ٤ - مثال لمقارنة عملية التجريد بين (Lovins, Porter , Paice)

ونجد أن عملية التجذيع Stemming تزيد من عدد الوثائق المســـترجعة وتؤثر إلى حد ما على الدقة Precision [29].

إزالة الكلمات الشائعة Stop Words:

هي قائمة من الكلمات الشائعة من الناحية اللغوية والتي لديها تأثير محدود على عملية تصنيف الوثائق واختيار الوثائق التي تتناسب مع احتياجات المستخدم.

وهي تعتبر كلمات وظيفية ليس لها معنى كما تعتبر جزءاً من كيفية وصف الأسماء في النص والتعبير عنها [29]. مثل الضمائر، حروف العطف، حروف الجر والتي تظهر في جميع المستندات النصية [32].

ونادراً ما تشير مثل هذه الكلمات إلى أي شيء يتعلق بموضوع الوثيقة وبالتالي فمثل هذه الكلمات الوظيفية لن تساعدنا في عمليات البحث [29].

إن الاستراتيجية العامة لتحديد هذه القائمة من الكلمات هي فرز المصطلحات حسب تكرار التجميع (العدد الإجمالي لعدد المرات التي يظهر فيها كل مصطلح في مجموعة الوثائق) ثم تؤخذ المصطلحات الأكثر شيوعاً والتي غالباً ما تتم تصفيتها يدوياً لمحتواها الدلالي القليل بالنسبة للوثائق التي يتم فهرستها [29].

فهناك العديد من المصطلحات المتكررة للغاية والتي لا تحقق قيمة كبيرة في عملية تمثيل المستندات كما أن إزالة هذه الكلمات يساهم بتقليل حجم المستند الذي تم إنشاؤه.

ويميز البعض بين نوعين من هذه الكلمات:

- الكلمات العلائقية (Inside ، outside ، Below ، above ...)
- الكلمات الغير علائقية (is ، am ، are ، a ، an .....)

فالكلمات العلائقية قد تشير إلى أهمية دلالية وقد تكون ذات أهمية لرفع كفاءة عملية الاسترجاع أما الكلمات غير العلائقية فتؤدي إلى تقليل طول المستند مما يؤدي إلى زيادة سرعة البحث.

ترجيح المصطلحات Term Weighting:

ذكرنا أن المصطلحات هي واصفات المحتوى التي يتم تعيينها للمستندات والوثائق والتي تستخدم في عملية الفهرسة ومن خلال هذه المصطلحات يتم تقييم مدى ارتباط المستندات بالاستعلامات.

وغالباً ما تُصنّف المصطلحات إلى مصطلحات موضوعية وغير موضوعية (objective and nonobjective) حيث تتم عملية الترجيح على المصطلحات غير الموضوعية والتي تعكس محتوى المعلومات في الوثيقة ومن ثم يتم تعيين وزن لهذه المصطلحات يشير إلى درجة أهمية هذا المصطلح بالنسبة لمحتوى معلومات المستند [15].

ومن هنا نستطيع القول أن ترجيح المصطلح عبارة عن عملية لحساب وتعيين قيمة رقمية لكل مصطلح من أجل تقدير مساهمته في تمييز وثيقة معينة عن غيرها من الوثائق.

معالجة اللغة الطبيعية Natural Language Processing NLP:

يستخدم تعبير "اللغة الطبيعية" للإشارة إلى اللغات الإنسانية كاللغة العربية واللغة الإنكليزية والفرنسية وغيرها لتمييزها عن اللغات غير الطبيعية كلغات برمجة الحاسب.

ولكل لغة مفرداتها ونحوها وقواعدها اللغوية التي تحدد كيفية بناء الجمل كما تحدد المعنى الدلالي لهذه الجمل.

واللغة هي وسيلة الإنسان للتفكير والتعبير عن الأفكار والمعلومات ووسيلة للتواصل ومن هنا تاتي أهمية موضوع معالجة اللغة الطبيعية.

تعتبر معالجة اللغات الطبيعية طريقة لتحليل النصوص عن طريق الحاسوب وتتضمن معالجة اللغات الطبيعية NLP جمع المعارف حول كيفية فهم البشر واستخدامهم للغة، ويتم ذلك من أجل تطوير الأدوات والتقنيات المناسبة التي تجعل أنظمة الكمبيوتر قادرة على فهم اللغات الطبيعية ومعالجتها لأداء مختلف المهام المطلوبة" [34].

ويُعرِّف الدكتور Michael J.Garbade معالجة اللغات الطبيعية NLP بأنها فرع من الذكاء الصناعي Artificial Intelligence الذي يعالج موضوع التفاعل بين أجهزة الكمبيوتر والبشر باستخدام اللغة الطبيعية ويبين أن الهدف النهائي من معالجة اللغات الطبيعية هو قراءة اللغات البشرية وفهمها وإدراكها بطريقة قيّمة واستخلاص المعنى المطلوب منها [35].

مجموعة أدوات معالجة اللغة الطبيعية Stanford CoreNLP:

تعتبر Stanford CoreNLP من أكثر الأدوات المستخدمة على نطاق واسع حيث توفر معظم خطوات معالجة اللغة الطبيعية الأساسية الشائعة NLP من الترميز Tokenization وحتى تحليل التبعية Coreference resolution من خلال الجمع بين عدة مكونات لتحليل اللغة الطبيعية [36].

كان من أهداف الإصدار الأولي الذي تم تطويره في عام ٢٠٠٦ الحصول على التعليقات التوضيحية الخطية Annotators Pipeline بسرعة وتوفير إطار خفيف الوزن من خلال استخدام كائنات الجافا Java Object وتطبيقها على أي نص بدلاً من تطبيقه على جملة واحدة فقط.

في عام ٢٠٠٩ تم تطوير النظام ليكون أكثر سهولة وليستخدم من قبل نطاق أوسع من المستخدمين حيث وفر النظام واجهة سطر الأوامر والقدرة على الكتابة خارج التعليقات التوضيحية وبتنسيقات مختلفة بما في ذلك XML.

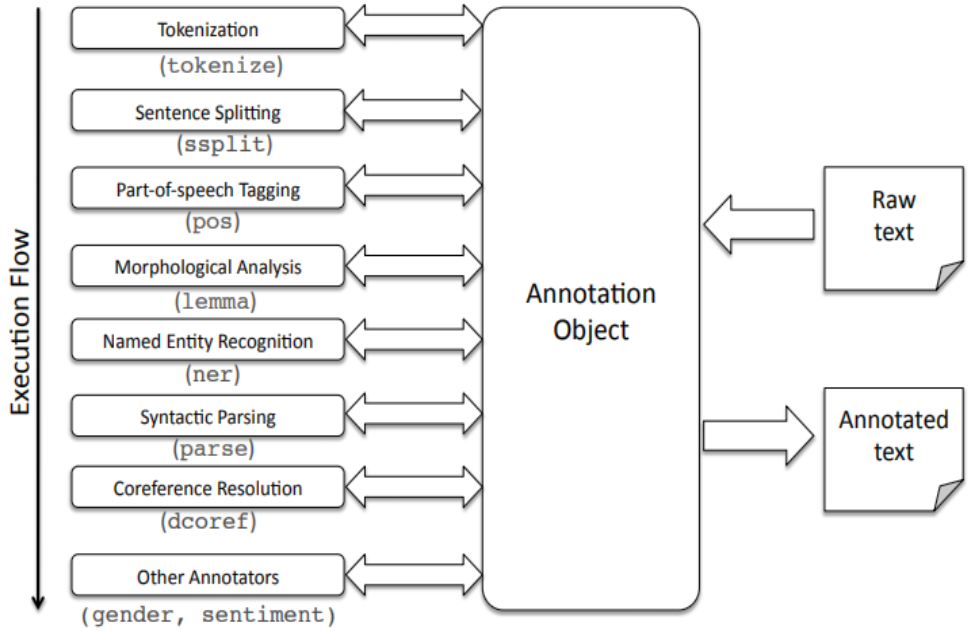
لقد وفرت Stanford CoreNLP مجموعة من مكونات التحليل اللغوي المستقرة والقوية وعالية الجودة والتي يمكن استدعاؤها بسهولة وهي تعتبر واحدة من أكثر مجموعات معالجة اللغة الطبيعية استخداماً [36].

يتم التحكم بالتعليقات التوضيحية في الخصائص الخاصة بالكائن Properties object، كما يتم تغليف مجموعة الأدوات Stanford CoreNLP بحيث يمكن الوصول إليها بسهولة من العديد من اللغات كـ Python، Ruby، Perl، Scala، JavaScript، Net. بما في ذلك C# [36].

و يمكن تنفيذ الاستخراج الآلي للمعلومات الدلالية من الأوصاف النصية من خلال معالجة نتائج تطبيق Stanford CoreNLP [37].

ويتضمن الإصدار الحالي مجموعة من أدوات المعالجة المصمّمة لأخذ مدخلات نصية أولية وإخراج تحليل نصي كامل وتعليقات توضيحية لغوية مناسبة لتحليل نصي فعال [37].

وفيما يلي نستعرض بنية النظام التي من خلالها تقوم Stanford Core NLP بتحليل النص ومعالجته:



الشكل ٥- بنية نظام مجموعة أدوات معالجة اللغة الطبيعية **Stanford CoreNLP** [36]

حيث تعطي المعالجة النتائج التالية: الترميز، علامات جزء الكلام (POS)، والتعرف على كيان الاسم (NER) بالإضافة إلى التحليل الدلالي للأسماء والأفعال [37].

وفيما يلي نستعرض بعض أهم مهام معالجة اللغة الطبيعية والتي تقوم بها مجموعة أدوات CoreNLP:

- الترميز Tokenization: والذي يقوم بتحويل النص إلى سلسة من الرموز المميزة والتي تتوافق إلى حد كبير مع الكلمات (Words).
- التجذيع Lemma: وهي عملية توليد الشكل الأساسي (Lemma) لجميع الرموز المميزة.
- التعرف على كيان الاسم Named Entity Recognition: حيث يتم التعرف على الأسماء كواحد من الأشكال (PERSIN، LOCATION، ORGANIZATION..) والأرقام (SET، DURATION، TIME، DATE، NUMBER، MONEY).
- علامات جزء الكلام Part-Of-Speech Tagging: تصنيف الرموز حسب قسم الكلام Part-of-Speech التي تنتمي له والتي تم ترميزها من خلال مجموعة من الـ Tags.

نذكر على سبيل المثال: NN اسم مفرد، NNS اسم جمع، NP عبارة اسمية، PRB ضمير شخصي ... [37].

ونورد جميع الـ Tags مع شرح يوضح كل منها في الجدول التالي:

Tag	Meaning	Explanation
CC	Coordinating conjunction	' & n and both but either et for less minus neither nor or plus so therefore times v. versus vs. whether yet
CD	Cardinal number	mid-1890 nine-thirty forty-two one-tenth ten million 0.5 one forty -seven 1987 twenty '79 zero two 78-degrees eighty-four IX '60s .025 fifteen 271,124 dozen quintillion DM2,000...
DT	Determiner	all an another any both del each either every half la many much nary neither no some such that the them these this those
EX	Existential there	There
FW	Foreign word	gemeinschaft hund ich jeux habeas Haementeria Herr K'ang-si vous lutihaw alai je jour objets salutaris fille quibusdam pas trop Monte terram fiche oui corporis...
IN	Preposition or subordinating conjunction	stride among upon whether out inside pro despite on by throughout below within for towards near behind atop around if like until below next into if beside...
NN	Noun, singular or mass	common-carrier cabbage knuckle-duster Casino afghan shed thermostat investment slide humour falloff slick wind hyena override subhumanity machinist...
NNS	Noun, plural	undergraduates scotches bric-a-brac products bodyguards facets coasts divestitures storehouses designs clubs fragrances averages subjectivists apprehensions muses factory-jobs...
NNP	Proper noun, singular	Motown Venneboerger Czestochwa Ranzer Conchita Trumplane Christos Oceanside Escobar Kreisler Sawyer Cougar Yvette Ervin ODI Darryl CTCA Shannon A.K.C. Meltex Liverpool...
NNPS	Proper noun, plural	Americans Americas Amharas Amityvilles Amusements Anarcho-Syndicalists Andalusians Andes Andruses Angels Animals Anthony Antilles Antiques Apache Apaches Apocrypha...
RB	Adverb	occasionally unabatingly maddeningly adventurously professedly stirringly prominently technologically magisterially predominately swiftly fiscally pitilessly...
RBR	Adverb, comparative	further gloomier grander graver greater grimmer harder harsher healthier heavier higher however larger later leaner lengthier less-perfectly lesser lonelier longer louder lower more

RBS	Adverb, superlative	best biggest bluntest earliest farthest first furthest hardest heartiest highest largest least less most nearest second tightest worst
RB	Particle	aboard about across along apart around aside at away back before behind by crop down ever fast for forth from go high i.e. in into just later low more off on open out over per pie raising start teeth that through under unto up up-pp upon whole with you
SYM	Symbol	% & ' " ".)). * + ,. < = > @ A[fj] U.S U.S.S.R * ** ***
UH	Interjection	Goodbye Goody Gosh Wow Jeepers Jee-sus Hubba Hey Kee-reist Oops amen huh howdy uh dammit whammo shucks heck anyways whodunnit honey golly man baby diddle hush sonuvabitch ...
WDT	Wh-determiner	that what whatever which whichever
JJ	Adjective	third ill-mannered pre-war regrettable oiled calamitous first separable ectoplasmic battery-powered participatory fourth still-to-be-named multilingual multi-disciplinary...
JJR	Adjective, comparative	bleaker braver breezier briefer brighter brisker broader bumper busier calmer cheaper choosier cleaner clearer closer colder commoner costlier cozier creamier crunchier cuter
JJS	Adjective, superlative	calmest cheapest choicest classiest cleanest clearest closest commonest corniest costliest crassest creepiest crudest cutest darkest deadliest dearest deepest densest dinkiest...
LS	List item marker	A A. B B. C C. D E F First G H I J K One SP-44001 SP-44002 SP-440 SP-44007 Second Third Three Two * a b c d first five four one six three two
MD	Modal	an cannot could couldn't dare may might must need ought shall should shouldn't will would
PDT	Predeterminer	all both half many quite such sure this
POS	Possessive ending	's
PRP	Personal pronoun	hers herself him himself hisself it itself me myself one oneself ours ourselves ownself self she thee theirs them themselves they thou thy us
PRP$	Possessive pronoun	her his mine my our ours their thy your

VB	Verb, base form	ask assemble assess assign assume atone attention avoid bake balkanize bank begin behold believe bend benefit bevel beware bless boil bomb boost brace break bring broil brush build ...
VBD	Verb, past tense	dipped pleaded swiped regummed soaked tidied convened halted registered cushioned exacted snubbed strode aimed adopted belied figgered speculated wore appreciated contemplated ...
VBG	Verb, gerund or present participle	telegraphing stirring focusing angering judging stalling lactating hankerin' alleging veering capping approaching traveling besieging encrypting interrupting erasing wincing ...
VBN	Verb, past participle	multihulled dilapidated aerosolized chaired languished panelized used experimented flourished imitated reunifed factored condensed sheared unsettled primed dubbed desired ...
VBP	Verb, non-3rd person singular present	predominate wrap resort sue twist spill cure lengthen brush terminate appear tend stray glisten obtain comprise detest tease attract emphasize mold postpone sever return wag ...
VBZ	Verb, present tense, 3rd person singular	bases reconstructs marks mixes displeases seals carps weaves snatches slumps stretches authorizes smolders pictures emerges stockpiles seduces fizzes uses bolsters slaps speaks pleads ...
TO	to" as preposition or " infinitive marker	To
WP	WH-pronoun	that what whatever whatsoever which who whom whosoever
WP$	WH-pronoun, possessive	Whose
WRB	Wh-adverb	how however whence whenever where whereby wherever wherein whereof why

الجدول ٥ – توصيف الـ Tags في مجموعة أدوات معالجة اللغة الطبيعية Stanford CoreNLP

وعلى الرغم من بعض الملاحظات لبعض المهام التحليلية لهذه الأدوات إلا أنه يمكن القول بأنها مجموعة من الأدوات سهلة الفهم والتي يمكن أن تستخدم كمكون داخل نظام أكبر بكثير قابل للتطوير [36].

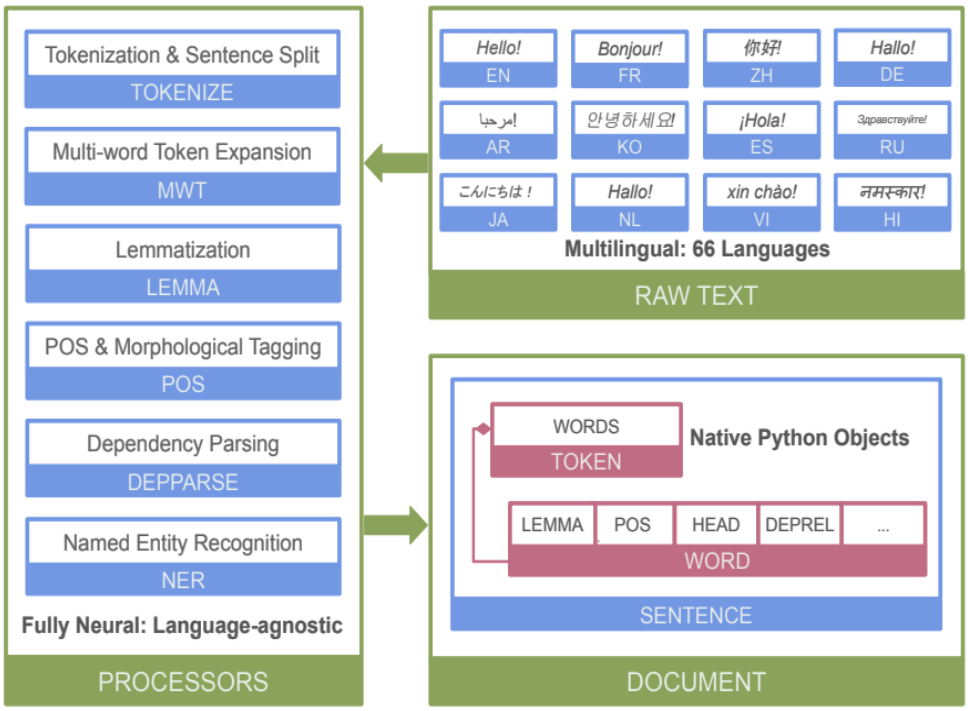
مجموعة أدوات معالجة اللغة الطبيعية Stanza:

Stanza هي تطوير لمجموعة أدوات معالجة اللغة الطبيعية Stanford CoreNLP بلغة Paython وهي تدعم ما يقارب ٧٠ لغة بشرية في حين أن CoreNLP لم تكن تدعم أكثر من ٦ لغات.

ولقد توفرت العديد من هذه الأدوات بدءاً من CoreNLP عام ٢٠١٤ ليأتي بعدها FLAIR ومن ثم SpaCy وكذلك UDPipe والتي كانت بلغة C++ [38].

إلا أن جميعها كانت تعاني من محددات فغالباً ما كانت تدعم عدداً قليلاً من اللغات بالإضـــافة إلى محدودية بعض الأدوات فيما يتعلق بدقة النتائج وذلك بسبب التركيز على موضوع الكفاءة.

وفيما يلي بنية النظام في مجموعة أدوات Stanza:



الشكل ٦ - بنية نظام مجموعة أدوات معالجة اللغة الطبيعية Stanza [38]

ورغم دعم Stanford coreNLP للغة العربية إلا أنها كانت محدودة الامكانات مقارنة باللغة الإنكليزية والتي تبين عدم إمكانية دعم اللغة العربية في العديد من أدواتها:

Annotator	Ara-bic	Chi-nese	Eng-lish	Fre-nch	Ger-man
Tokenize	✓	✓	✓	✓	✓
Sent. split	✓	✓	✓	✓	✓
Truecase			✓		
POS	✓	✓	✓	✓	✓
Lemma			✓		
Gender			✓		
NER		✓	✓		✓
RegexNER	✓	✓	✓	✓	✓
Parse	✓	✓	✓	✓	✓
Dep. Parse		✓	✓		
Sentiment			✓		
Coref.			✓		

الشكل ٧- دعم مكونات تحليل CoreNLP للغات المختلفة [36]

فاللغة العربية من اللغات ذات التصريف الشديد وتعتبر عملية الـــ Stemming جزءاً أساسياً وفاعلاً في عملية البحث [29].

فجاءت مجموعة الأدوات Stanza لتدعم اللغة العربية بشـــكل أكبر وفيما يلي مقارنة بين دعم أدوات Stanza لكل من اللغتين العربية والإنكليزية:

Treebank	System	Tokens	Sents	Words	POS	Lemmas
Arabic	**Stanza**	99.98	80.43	97.88	94.89	93.27
English	**Stanza**	99.01	81.13	99.01	95.40	97.21

ولقد أظهرت النتائج أن Stanza استطاعت بالإضافة إلى إمكانية معالجة العديد من اللغات تحقيق نتائج أكثر دقة من مجموعات الأدوات الأخرى [38].

المنهجية المقترحة لعملية معالجة النص والفهرسة:

في هذه الدراسة نستخدم مجموعة Stanford Core NLP كأدوات لتنفيذ المكونات الأساسية لمعالجة اللغة الطبيعية وتحليل النص. حيث سنقوم بتحليل لغوي ونحوي بسيط للمصطلحات الواردة في الوثائق من خلال هذه المكتبة اعتماداً على القيود والقواعد العامة للغة والاستفادة من هذا التحليل في عملية فهرسة المصطلحات وتحديد ميزات المصطلحات الواردة في الوثائق لرفع كفاءة عملية الاسترجاع.

تمتلك كل لغة قواعد عامة لبنية الجملة فيها وإن فهم البنية العامة للجمل يمكننا من فهم أكبر لمتطلبات المستخدم وحاجته من المعلومات.

وتعدُّ اللغة الإنكليزية من اللغات ســـهلة التعلُّم فقواعدها واضـــحة وقليلة على عكس قواعد اللغة العربية المتشـابكة وتحتوي اللغة الإنكليزية على قواعد رئيسـية وبنية عامة للجمل شـأنها شـأن بقية اللغات والتي يمكن الاســـتفادة منها في عملية التحليل اللغوي للمصـــطلحات لذلك تم اختيار اللغة الإنكليزية كنموذج لتطبيق المنهجية الجديدة ليتم الاستفادة منها كأساس لبناء منهجية خاصة باللغة العربية.

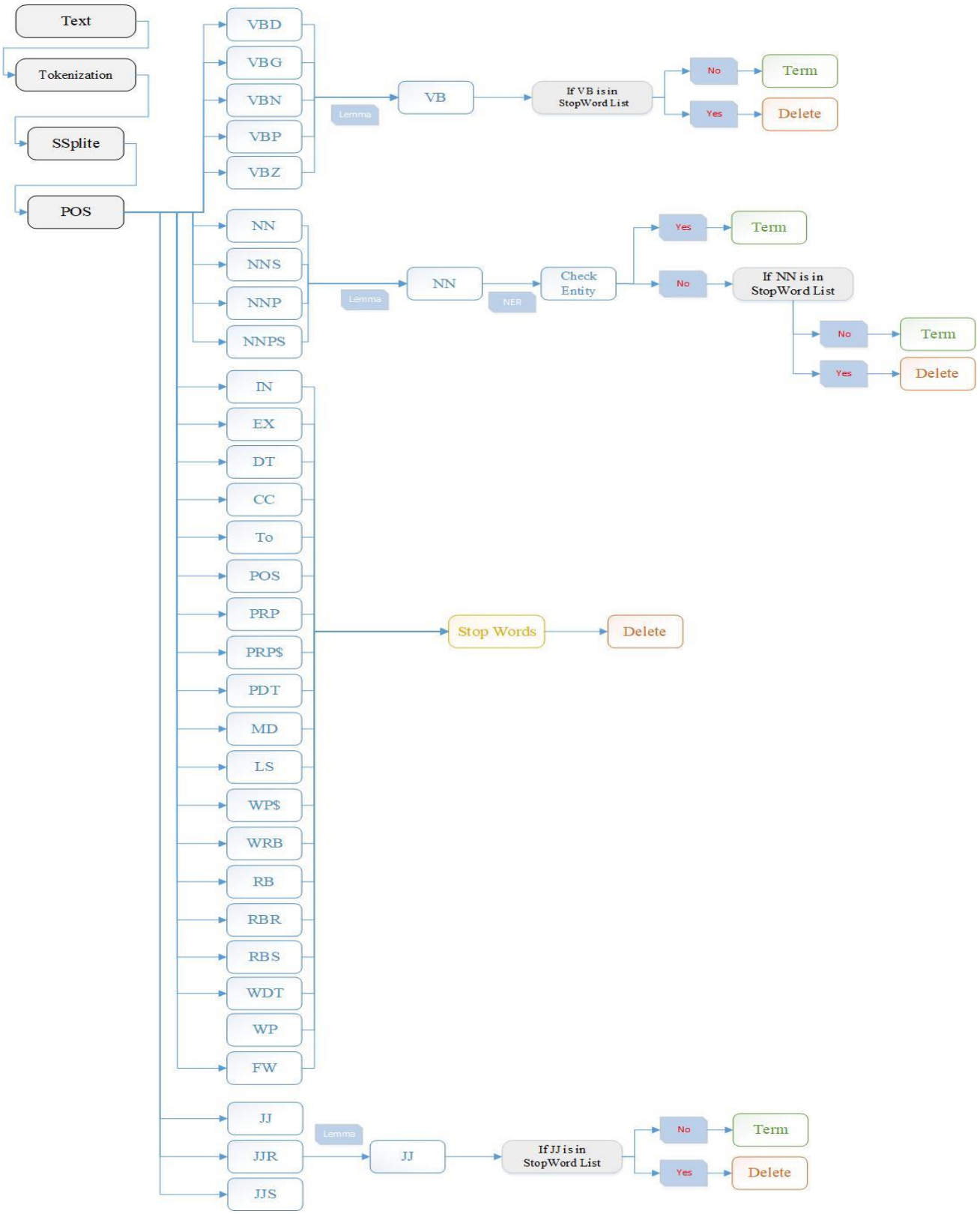
فطبيعية اللغة الإنكليزية تســـاعد في عملية التحليل النصـــي ســـواء في عملية التجريد حيث أن جميع الإضافات على جذور الكلمات تتم في نهاية الكلمة باستثناء بعض الحالات الخاصة.

كما تتميز بسهولة التحليل وإعراب الجملة لتحديد طريقة ورود الكلمات في الجمل Part-Of-Speech.

تم الاســـتفادة من إمكانية هذه الأدوات في عملية تحليل النص واســـتخراج المميزات لكل وثيقة Feature لأهم المصـــطحات في كل منها وذلك من خلال منهجية في عملية تحليل النصـــوص اعتمدت على الخطوات التالية:

١. تحليل النص من خلال مجموعة أدوات Stanford CoreNLP وتقســـيمه إلى مجموعة من الرموز Tokens.
٢. تصنيف الـــ Tokens من خلال تحليل POS إلى مجموعات ترتبط بموقع المصطلح النحوي حيث يتم فرز الرموز إلى أربعة مجموعات رئيسية: (الرموز الوظيفية، الأفعال، الأسماء، الصفات).
٣. حذف الرموز الوظيفية باعتبار أنها تؤدي مهمة وظيفية في النص ولا تلعب دوراً في تحديد موضوع الوثيقة.
٤. تطبيق التجذيع Lemma على كل من (الأفعال، الصفات، الأسماء).
٥. التصـــفية التلقائية لبعض الرموز غير الوظيفية حيث تمّ تحديد مجموعة من الرموز في النظام المقترح والتي تعتبر StopWords وهي عبارة عن بعض الأفعال والأســـماء والصـــفات شـــائعة الاستخدام ولا تساهم في تحديد موضوع الوثيقة حيث بلغ عدد الرموز ضمن هذه القائمة حوالي ٢٠٠ رمز نذكر منها على سبيل المثال: (do، like، good، great، ....).
٦. فهرسة الرموز الناتجة عن العمليات السابقة كمصطلحات Terms ضمن قاعدة بيانات النظام المقترح مع بعض الخصائص مثل (POS، NER، Order) والتي سيتم بحث إمكانية الاستفادة منها في عملية ترجيح المصطلحات في الفصل القادم.

وفيما يلي مخطط يبين خطوات منهجية الفهرسة المقترحة:



الشكل ٨- مخطط معالجة النص والفهرسة في النظام المقترح

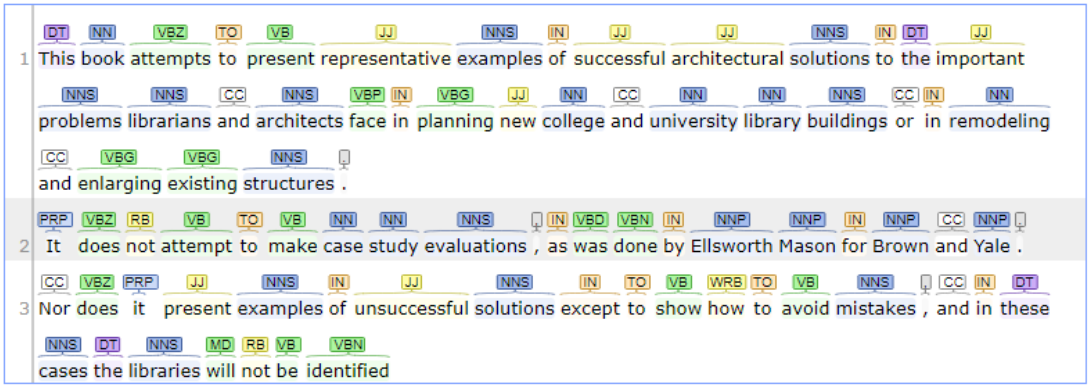
فيما يلي نستعرض مثالاً عملياً لعملية المعالجة التي تمت على أحد نصوص قاعدة البيانات:

لدينا النص التالي:

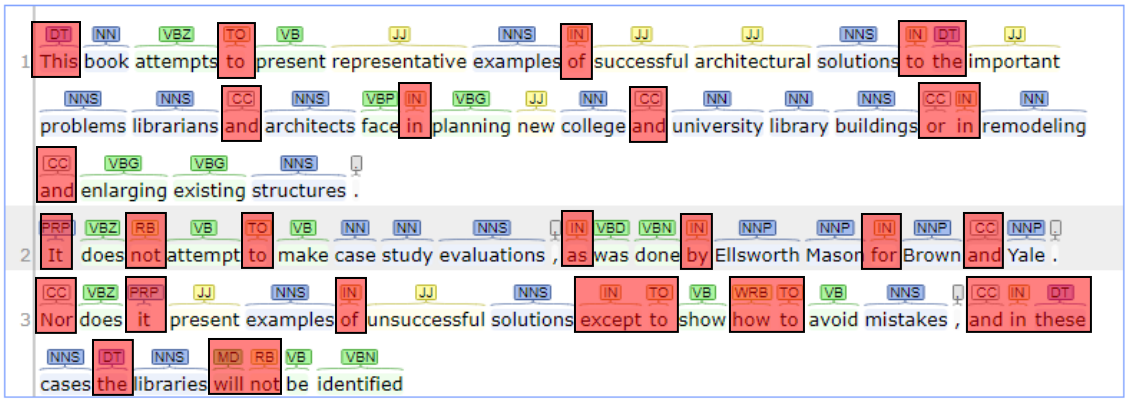
> This book attempts to present representative examples of successful architectural solutions to the important problems librarians and architects face in planning new college and university library buildings or in remodeling and enlarging existing structures. It does not attempt to make case study evaluations, as was done by Ellsworth Mason for Brown and Yale. Nor does it present examples of unsuccessful solutions except to show how to avoid mistakes, and in these cases the libraries will not be identified.

- من خلال استخدام Stanford CoreNLP سيتم تقسيم النص إلى رموز وربط كل Token بنوع من الـ Tags فتكون نتيجة عملية تحليل النص POS:

Part-of-Speech:



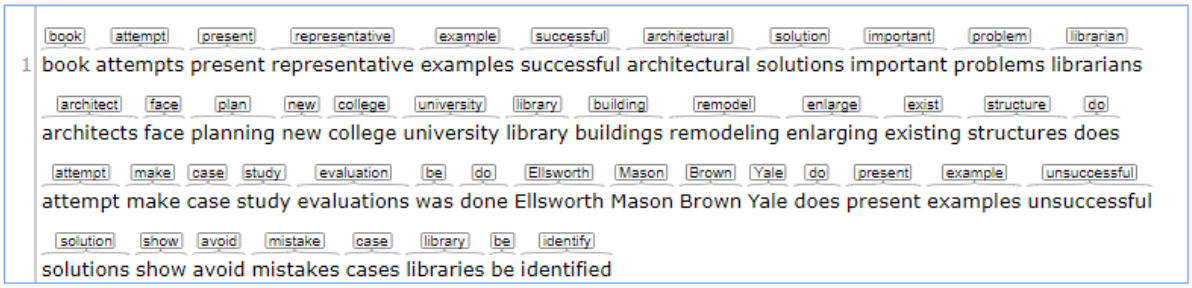
- حسب المخطط الذي تم اعتماده في النظام المقترح سيقوم النظام المقترح بحذف الـــ Tokens التي تمتلك خصائص وظيفية ولا ترتبط بموضوع النص ولن تفيد في عملية الاسترجاع:



نتيجة الخطوة السابقة سيصبح بعد حذف الـــ StopWords ٤٧ رمزاً Token في حين كان قبل حذف الرموز الوظيفية ٧٨ رمزاً.

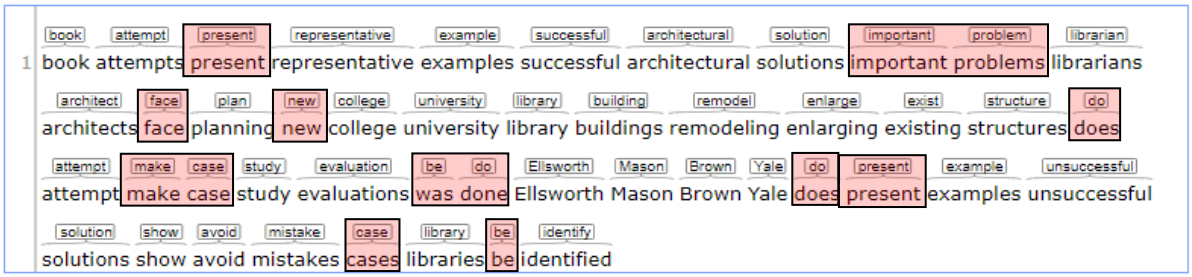
- و بتطبيق الـ Lemma نحصل على النتيجة التالية:

Lemmas:



- ومن ثم سيقوم النظام بتصفية مجموعة الرموز الواردة في قائمة الكلمات الشائعة في النظام المقترح:

Lemmas:



- يصبح النص بعد خطوات المعالجة السابقة باستخدام Stanford CoreNLP على الشكل التالي:

book attempt representative example successful architectural solution librarian architect plan college university library building remodel exist enlarge structure attempt study evaluation **Ellsworth Mason Brown Yale** example unsuccessful solution avoid mistake library identify

ويصبح عدد الرموز التي تم اختيارها ٣٢ رمزاً مصنفين إلى (اسم، صفة، فعل).

- فهرســـة الأفعال **Verbs** كما هي (attempt، face، plan، remodel، avoid، ...) وكذلك فيما يتعلق بالرموز التي تدل على كيانات **Entity** (Yale، Brown، ...) بالإضافة إلى كل من الأسماء والصفات.

لتصبح فهرسة النص بالشكل النهائي التالي مع بعض الخصائص المميزة للمصطلح:

Original Word	Term	Freq	POS	Order	NER
book	book	1	NN	2	-
attempts, attempt	attempt	2	VBZ, VB	3,39	-
representative	representative	1	JJ	6	TITLE
examples,examples	example	2	NNS,NNs	7,61	-
successful	successful	1	JJ	9	-
architectural	architectural	1	JJ	10	-
solutions, solutions	solution	2	NNS,NNS	11,64	-
librarians	librarian	1	NNS	16	-
architects	architect	1	NNS	18	-
planning	plan	1	VBG	21	-
college	college	1	NN	23	-
university	university	1	NN	25	-
library, libraries	library	2	NN,NNS	26,78	-
buildings	building	1	NNS	27	-
remodeling	remodeling	1	NN	30	-
enlarging	enlarge	1	VBG	32	-
existing	exist	1	VBG	33	-
structures	structure	1	NNS	34	-
study	study	1	NN	43	-
evaluations	evaluation	1	NNS	44	-
Ellsworth	Ellsworth	1	NNP	50	PERSON
Mason	Mason	1	NNP	51	PERSON
Brown	Brown	1	NNP	53	PERSON
Yale	Yale	1	NNP	55	ORGANIZATION
unsuccessful	unsuccessful	1	JJ	63	-
avoid	avoid	1	VB	70	-
mistakes	mistake	1	NNS	71	-
identified	identify	1	VBN	82	-

الجدول ٦ – فهرسة أحد النصوص في النظام المقترح

وبالتالي تم فهرسة هذا النص بعدد من المصطلحات Terms بلغ ٢٨ مصطلحاً فقط.

وفيما يلي مقارنة نتيجة التجريد بعد حذف الـــ StopWord بين المنهجية المقترحة (NLP Index) وبين مجرد Porter:

Original Word	NLP Index	Porter
book	book	book
attempts, attempt	attempt	attempt
representative	representative	repres
examples, examples	example	example
successful	successful	success
architectural	architectural	architectur
solutions , solutions	solution	solut
librarians	librarian	librarian
architects	architect	architect
planning	plan	plan
college	college	colleg
university	university	univers
library, libraries	library	librari
buildings	building	build
remodeling	remodeling	remodel
enlarging	enlarge	enlarge
existing	exist	exist
structures	structure	structur
study	study	studi
evaluations	evaluation	evalu
Ellsworth	Ellsworth	Ellsworth
Mason	Mason	Mason
Brown	Brown	Brown
Yale	Yale	Yale
nor	-	nor
unsuccessful	unsuccessful	unsuccess
except	-	except
avoid	avoid	avoid
mistakes	mistake	mistak
identified	identify	identifi

الجدول ٧ – مقارنة نتيجة الفهرسة بالاعتماد على مجرّد Porter والفهرسة في النظام المقترح

ملخص الفصل الثاني:

من خلال مقارنة النتائج النهائية والتي حصـــلنا عليها باســـتخدام أدوات معالجة اللغة الطبيعية في عملية التحليل والتجريد نلاحظ إمكانية الاستفادة من هذه الأدوات في عملية الفهرسة وتصفية الكلمات الشائعة الأمر الذي يساهم في توسيع الوثائق المسترجعة من خلال عدم الاعتماد على المطابقة التامة في عمليات البحث.

ويمكن تلخيص الميزات المستفادة من مجموعة أدوات معالجة اللغة الطبيعية في منهجية الفهرسة المطبقة بما يلي:

- تصنيف الرموز إلى مجموعات الأمر الذي قد يلعب دوراً في تحديد أهميتها في النصوص.
- التصفية الآلية للرموز التي تلعب دور وظيفي في الوثائق ولا ترتبط بموضوع الوثيقة.
- تخفيض قائمة الكلمات الشائعة من قائمة تزيد عن ٨٠٠ كلمة إلى قائمة لا تتجاوز الـ ٢٠٠ كلمة.
- التجريد الدقيق للأفعال والذي يعطي الصيغة المجردة للفعل.

بالإضـــافة إلى ما ذكر فإنه يمكن الاســـتفادة من هذه الأدوات في عملية اســـتخلاص المزيد من الميزات Features الخاصة بالمصطلحات والتي يمكن استخدامها في تطوير عمليات ترجيح المصطلحات بحيث لا يؤثر زيادة عدد الوثائق المسترجعة على دقة النتائج.

وبما أننا نعالج هنا الوثائق النصـية ومع تطور التكنولوجيا والإمكانيات الضـخمة التي تمتلكها اليوم فيمكن المحافظة على كفاءة أنظمة الاسترجاع وفعاليتها رغم عمليات المعالجة التي تحتاج لها مثل هذه العمليات.

ورغم الميزات التي ذكرت إلا أن هذه الأدوات لم تعالج اللواحق في الأســـماء والصـــفات الأمر الذي يعد قصوراً في هذه الأدوات واقتصرت عملية التجريد فيها على بعض العمليات البسيطة.

وبالتالي لا بد من دراســـة عملية معمقة تأخذ بعين الاعتبار مزايا اللغة وقواعدها بالإضـــافة إلى الحالات الشاذة فيها للوصول إلى عملية تجريد دقيقة وخاصة فيما يتعلق بالأسماء والصفات.

الفصل الثالث: تطوير مخطط الترجيح TF-IDF

مقدمة:

إن ترجيح المصطلحات يعتبر فرعاً في مجال استرجاع المعلومات يدرس مسألة مدى أهمية كلمة أو عبارة في نص معين حيث يعتبر موضوع تحديد أهمية الكلمات الأساسية موضوعاً رئيسياً ومؤثراً في فاعلية أنظمة استرجاع المعلومات الحديثة.

ما هو مدى أهمية المصطلح؟ سؤال هام تبحث عن إجابته العديد من أنظمة استرجاع المعلومات [39].

ويعتبر وجود المصطلح في الاستعلام والوثيقة من أهم العوامل التي تؤثر في أهمية المصطلح في الوثيقة بالنسبة للاستعلام، كما أن العديد من الأنظمة تعتمد في عملية الترجيح على عدد تكرارات المصطلح في المستند والذي يشار إليه عادة باسم تردد المصطلح Term Frequency [11] [40].

في حين تشير بعض الدراسات إلى أهمية عوامل ثلاث في تحديد مدى أهمية المصطلح في النص هي: (طول النص Document Length، تردد المصطلح Term Frequency، تردد الوثيقة المعكوس Inverse Document Frequency) [39].

ورغم أهمية تكرار المصطلح والذي يعتبر طريقة شائعة في تحديد أهمية مصطلح في المستند إلا أن تكرار المصطلح يتجاهل كيفية تفاعل المصطلح مع سياق النص [40].

لذا تقترح هذه الدراسة إيجاد استراتيجية جديدة داعمة لخوارزمية ترجيح المصطلحات التقليدية تعتمد في جزءها الأساسي على تكرار المصطلح إلا أنها تضيف معاملات تأخذ بعين الاعتبار أهم ميزات المصطلح في النص والتي تم استخلاصها من أدوات معالجة اللغة الطبيعية بهدف تحسين دقة استرجاع المعلومات.

مخطط الترجيح التقليدي TF-IDF:

إن TF-IDF تعتبر خوارزمية بسيطة وفعالة لمطابقة الكلمات في الاستعلامات مع المستندات ذات الصلة بالاستعلام كما أنها تعتبر الأساس للعديد من الخوارزميات الأخرى في أنظمة استرجاع المعلومات [41].

TF-IDF هو اختصار لـــ Term Frequency–Inverse Document Frequency، حيث عرّف العالم Gerard Salton مصطلح TF لأول مرة بينما تم تعريف مصطلح IDF في عام ١٩٧٢ من قبل العالم Karen Sparck-Jones.

وتم تعريفه بأنه إحصـــاء رقمي يهدف إلى التعبير عن مدى أهمية كلمة في وثيقة في مجموعة وغالباً ما يستخدم كعامل مرجح في عمليات البحث ضمن أنظمة استرجاع المعلومات [42].

يطلق على تكرار المصطلح TF أيضاً مصطلح Local Term Weight ويتم تعريفه على أنه عدد مرات تكرار المصطلح قيد البحث في وثيقة.

$$TF(t) = \frac{\text{Number of times term t appears in a document}}{\text{Total number of terms in the document}}$$

$$= (\text{العدد الكلي للمصطلحات في الوثيقة / عدد مرات ظهور المصطلح في الوثيقة})$$

بينما يطلق على الـ IDF مصطلح Global Term Weight ويعبر عن تواتر المصطلح ضمن مجموعة من الوثائق.

$$IDF(t) = \log_e \frac{\text{Total number of documents}}{\text{Number of documents with term t in it}}$$

$$= \log_e (\text{عدد الوثائق التي تحوي المصطلح / عدد الوثائق الكلي}).$$

فالتردد المعكوس للوثيقة هو لوغاريتم نسبة عدد المستندات في المجموعة إلى عدد الوثائق التي تحتوي على كلمة معينة وهذا يعني أن الكلمات النادرة ستكون قيمة IDF لها عالية في حين أن الكلمات الشائعة ستكون ذو قيمة IDF منخفضة، حيث ستكون قيمة IDF تساوي /٠/ إذا ظهر المصطلح في جميع الوثائق [43].

ويتم احتساب ترجيح المصطلحات باستخدام TF، IDF من خلال المعادلة:

$$\boldsymbol{TF - IDF_{t,d} = TF_{t,d} * IDF_t}$$

إذاً TF يخبرنا فيما إذا كان المصطلح t مكرر كثيراً في الوثيقة d وبالتالي يكون المصطلح t مهم جداً بالنسبة لها وبشكل متشابه فإن التردد العكسي للوثائق IDF يقيس ندرة مصطلح t فيما يتعلق بالمجموعة الكاملة أي أن المصطلحات التي تظهر في العديد من الوثائق هي أقل فائدة.

فعلى سبيل المثال: الكلمات الشائعة Stop Words مثل a، an، the ... الخ هي أقل فائدة لأنها مصطلحات غالباً ما ترد في جميع الوثائق بالمجموعة، بينما المصطلح النادر جداً في المجموعة الكاملة والمتكرر بالوثيقة يعتبر هاماً جداً [44].

وبالتالي فإن مخطط ترجيح المصطلح TF-IDF يعتبر مؤشراً لأهمية المصطلح في تمثيل الوثائق، وهو من أكثر طرق ترجيح المصطلحات استخداماً والأكثر شيوعاً في استرجاع المعلومات [42].

وبشكل رئيسي فإن طرق ترجيح المصطلحات هي نموذجين: المشرف عليها Supervised، والغير مشرف عليها Unsupervised وتسمى أيضاً بالطرق التقليدية.

إن تصنيف طرق الترجيح إلى هاتين الفئتين يعتمد على المعلومات الاحصائية المستخدمة في حساب وزن المصطلح فيما إذا كان على مستوى مجموعة أو على مستوى فئة.

فالتمييز بين الطرق المشرف عليها Supervised وغير المشرف عليها Unsupervised هو ان الأول يستخدم معلومات مصنّفة بينما الأخير يحسب على مستوى المجموعة [26].

إن العديد من المعلومات الاحصائية يتم استخدامها من قبل طرق الترجيح غير المشرف عليها المختلفة، هذه المعلومات تتضمن على سبيل المثال: عدد الوثائق، طول الوثيقة، تكرار المصطلح على مستوى المجموعة. وبالتالي فإن العديد من مخططات الترجيح يمكن وضعها في فئة الأصناف الغير مشرف عليها مثل: Term Frequency-Inverse Document Frequency (TF-IDF)، Document Frequency، Glasgow weight وكذلك Entropy.

ومن جهة أخرى فإن تقنيات الترجيح المشرف عليها تستخدم معلومات حول الفئة Category(Class) والتي تنتمي الوثيقة لها تعرف بعضوية الوثيقة document's membership وتدخل بحساب وزن المصطلح [26].

قيود مخطط الترجيح التقليدي TF-IDF:

إن تقنية TF-IDF التقليدية مستخدمة بشكل شائع في عمليات التنقيب في النصوص واسترجاع المعلومات.

وبالرغم من المزايا العديدة للطريقة التقليدية TF-IDF فإنه يوجد العديد من العيوب والتي لا يمكن تجاهلها وسنلقي الضوء على بعض منها فيما يلي:

- تفترض الطريقة التقليدية بأن حساب تردد المصطلحات يقدم دليلاً مستقلاً على التشابه الأمر الذي لا يعتبر دائماً صحيح [44].
- هذه الطريقة تحسب ترجيح المصطلحات على أساس تردد المصطلح ولا تُعنى بموضع المصطلح في النص [41] [45].

- إن TF-IDF التقليدية هي تقنية لاختيار ميزة Feature غير مشـــرف عليها فهي مقيدة فقط بالوثيقة [44].

ورغم ذلك فهي لا تناقش دلالته وورود المشـترك مع مصـطلحات أخرى في الوثيقة. كما لا تهتم بموضوع العلاقة بين الكلمات وأهمية المصطلح بالنسبة للنص نفسه [41] [45].

لذلك كان وما زال هذا المجال حافزاً للكثير من الباحثين لدراســة إمكانية زيادة فعالية أنظمة اســترجاع المعلومات من خلال تطوير ترجيح المصطلحات TF-IDF التقليدية وقد ظهرت العديد من الدراسات في هذا المجال نورد بعضاً منها فيما يلي.

الدراسات السابقة:

- في عـام ٢٠١٧ قـام مجموعـة من البـاحثين وهم Thabit Sabbah, Ali Selamat, Md Hafiz Selamat, Fawaz S.AlAnzi, Enrique Herrera Viedma, Ondrej Krejcar, Hamido Fujita باقتراح أربعة مخططات لترجيح المصـطلحات بالاعتماد على مخطط ترجيح المصـطلحات TF-IDF وهي mTF، mTFIDF، TFmIDF وmTFmIDF [26].

اقترح الباحثون في هذه الدراسـة أن تأخذ مخططات ترجيح المصـطلح المقترحة في الاعتبار إحصـاء المصـطلحات الغائبة مع حسـاب أوزان المصـطلحات الحالية لتحسـين أداء تصـنيف النصـوص TC Text Categorization. فقاموا بتعديل مخطط تردد المصطلح TF ومخطط تردد الوثيقة المعكوس IDF كما يلي:

- **مخطط تردد المصطلح المعدل mTF:**

كان المقترح الأول يدرس نسـبة عدد المصـطلحات المفقودة في مسـتند إلى إجمالي عدد المصـطلحات في مجموعة ما ليقدم ترجيح جديد باسم mTF.

في هذا المخطط أخذ الباحثون بالاعتبار نسبة العدد الاجمالي لكافة الوثائق المعرّفة بـــ Tt إلى العدد الاجمالي للرموز Token في المجموعة والمعروفة بـــــــ Tc ولكن وحيث أن Tc أكبر من Tt يتم تطبيق النسبة المقلوبة بمساعدة الجذر التربيعي لقياس تأثير الفرق بين Tt و Tc.

إضافة إلى ذلك اقترح الباحثون أن يتم أخذ طول الوثيقة بعين الاعتبار Length d لـــ Tc في عملية التطبيع Normalization عند حساب وزن المصطلح t في الوثيقة d.

وكانت المعادلة المقترحة من قبل الباحثون في هذه الدراسة على الشكل التالي:

$$mTF_{t,d} = \frac{\mathrm{tf}_{t,d} \times \log\frac{\sqrt{Tc}}{Tt}}{\log|\sum_{t=1}^{n}(\mathrm{tf}_{t,d}^2) \times (\frac{length_d^2}{\sqrt{Tc}})|}$$

حيث : tf يمثل التردد الأولي للمصطلح في الوثيقة d والذي هو ببساطة عبارة عن عدد ورود الرمز. بينما Tt هو العدد الاجمالي للتردد الأولي للمصطلح t في كافة الوثائق.

- **مخطط تردد الوثيقة المعكوس المعدل mIDF:**

أما المقترح الثاني فيدرس نسبة عدد المستندات التي لا يوجد فيها مصطلح إلى إجمالي عدد المستندات في المجموعة وهو ما أطلق عليه الباحثون اسم mIDF.

ذكر الباحثون في هذه الورقة أن تردد الوثيقة القياسي هو عبارة عن نسبة عدد كافة الوثائق والتي يرمز لها بـ N لعدد الوثائق في المجموعة والتي يظهر فيها المصطلح t.

وبالتالي فعدد الوثائق التي لا يظهر فيها المصطلح t يمكن ملاحظتها من خلال $N - DF_t$.

وبالتالي تكون الصيغة المعدلة لتردد الوثيقة المعكوس كما يلي:

$$mIDF_t = log[\frac{N}{1/((N - DF_t) + 1)}]$$

وبناء على المقترحين السابقين قدم الباحثون أشكالاً مختلفة من مخطط الترجيح القياسي لمصطلح TF−IDF بناءً على مخططات mTF وmIDF المقترحة [26]:

$$\boldsymbol{mTF - IDF_{t,d} = mTF_{t,d} * IDF_t}$$
$$\boldsymbol{TF - mIDF_{t,d} = TF_{t,d} * mIDF_t}$$
$$\boldsymbol{mTF - mIDF_{t,d} = mTF_{t,d} * mIDF_t}$$

ونلاحظ أن الباحثون في هذه الدراسة اعتمدوا على فكرة أنه لابد أن تختفي بعض المصطلحات عند وجود مصطلحات أخرى في المستند والعكس بالعكس وبالتالي لابد أن تتأثر أوزارن المصطلحات بذلك.

ونجد أن الباحثين في هذه الدراسة أغفلوا العديد من الميزات الأخرى الهامة كدلالة المصطلح وتوزعه وطول الوثيقة على سبيل المثال، الأمر الذي يؤثر بشكل كبير في عملية ترجيح المصطلح فأهمية اختيار الخصائص Feature Selection لوثيقة ما والاعتماد على أسلوب ترجيح للمصطلحات يأخذ بعين الاعتبار ميزات مختلفة للوثيقة يؤثر وبشكل كبير في عملية تحسين أداء تصنيف النصوص.

- في عام ٢٠١٨ قام كل من Rajendra Kumar Roul ، Jajati Keshari Sahoo و Kushagr Arora من جامعة Zuarinagar الهندية بدراسة بيّنت عيوب طريقة TF-IDF ومن ثم اقترحوا أربع تقنيات مختلفة لترجيح المصطلحات بهدف تجاوز هذه العيوب من خلال تعديل طريقة TF-IDF التقليدية حيث وضحت الدراسة أن عملية تمثيل النصوص المبني على أساس اللغة Language Base مثل النموذج الشعاعي Vector Space Model VSM يؤثر بشكل كبير وخاصة في المجالات التي يتم فيها معالجة اللغة طبيعياً كمجال استرجاع المعلومات وإن عملية تحويل المستندات إلى أشعة يعطي إمكانية إجراء أي عملية حسابية على المستندات الممثلة شعاعياً وبالتالي فإن ترجيح المصطلح يلعب دوراً كبيراً ليكون تمثيل الوثائق أكثر دقة [44].

وتضمنت الدراسة أربع تقنيات مقترحة لتعديل خوارزمية الترجيح التقليدية TF−IDF حيث ناقشت الدراسة النقاط التالية:

١. TF-IDF يعتمد على التشتت بين الصفوف inter-class dispersion ويُعرّف التشتت بين الصفوف بأنه مساهمة المصطلح في الفئة والذي يساعد المصنّف باتخاذ القرار الصحيح أثناء تصنيف الوثيقة.

٢. إن الأسلوب التقليدي TF-IDF يعتمد على تكرار الوثيقة المعكوس المعدل: وبالتالي فإن الأخذ بعين الاعتبار الكلمات التي تم تكرارها بشكل كبير بالوثائق وبداخل المجموعة بشكل كامل (باستثناء StopWords) يساعد على زيادة القوة التمييزية لتكرار الوثيقة المعكوس التقليدي.

٣. يعتمد الأسلوب التقليدي على تكرار الفئة Class frequency: تردد الفئة يساعد على تحديد فيما إذا كان المصطلح وثيق الصلة بفئة خاصة أم لا. الأمر الذي يعطي وزناً للمصطلحات

في خصـائص الفئة وبالتالي فإن تصـنيف النص مضـافاً إليه وزن المصـطلحات بالاعتماد على أساس تكرار فئتهم سوف يُحسّن صحة التصنيف.

٤. يعتمد الأسـلوب التقليدي على الطول الطبيعي Normalized Length: من خلال إضـافة عامل الطول الطبيعي إلى الأسـلوب التقليدي سـوف يزيد أهمية العامل TF تكرار المصـطلح ويقلل من وزن المصـطلحات الأقل تكراراً ولكنها نسـبياً ذات وزن TF weighting أعلى في الوثيقة [44].

وفيما يلي نستعرض التقنيات الأربع المقترحة بالتفصيل:

- **ترجيح المصطلح المعدل بالاعتماد على التشتت داخل الفئة:**

TF-IDF Based on inter-class dispersion (W1)

تبين المعادلة التالية طريقة حساب الترجيح بالاعتماد على التشتت داخل الفئة:

$$Weight\ (i,j) = \boldsymbol{W_{1ij} = TF(i,j) * IDF(i) * D(i)}$$

حيث : D(i) يمثل معامل التشتت داخل الفئات للمصطلح i ويتم حسابها كما يلي:

$$D\ (i) = \frac{1}{n} \sum_{i=0}^{n} (F(t,i) - avg(F(t,i))^2$$

هنا يكون n عدد الفئات ، F(t,i) عدد الوثائق التي تحتوي المصطلح t ومرتبط بالفئة التي يرتبط بها المصطلح t أيضاً.

$$avg\big(F(t,i)\big) = \frac{1}{n} \sum_{i=0}^{n} F(t,i)$$

إلى جانب مزايا TF-IDF التقليدية فإن هذه الطريقة تقوم بالتحقق من مدى مساهمة المصطلح الجيدة في عملية التصنيف.

فإذا كان المصـطلح موزعاً بشـكل موحد بين الفئات فإن التشـتت داخل الفئة D سـيكون منخفضـاً وبالتالي سيكون الوزن الذي يساهم فيه المصطلح منخفضاً.

ولكن إذا كان لدى المصـطلح اختلاف عالٍ فذلك يعني أنه يعتبر ميزة جيدة لاسـتخدامه في عملية التصنيف ووفقاً لذلك فإن قيمة D ستكون عالية بالنسبة لذلك المصطلح.

- **ترجيح المصطلح المعدل بالاعتماد تردد الوثيقة المعدّل:**

TF-IDF Based on modified inverse document frequency (W2)

سيتم تعديل تقنية ترجيح المصطلحات التقليدية من خلال المعادلة التالية:

$$Weight\ (i, j) = \boldsymbol{W_{2ij} = TF(i,j) * Modified_{IDF(i)}}$$

حيث:

$$\boldsymbol{Modified_{IDF(i)} = log_{10}\left(\frac{no.of\ document\ in\ P+1}{document\ frequency\ of\ term\ i}\right)}$$

إذا جاء المصطلح في كل وثيقة فإن IDF التقليدي سوف يعطي وزن (0) لجميع حالات الورود في المصفوفة الثنائية (مصطلح – وثيقة)، فتم إضافة القيمة (1) إلى العدد الإجمالي للوثائق لزيادة تردد الوثيقة العكسي للمصطلحات الفريدة unique والغير فريدة non-unique، وبالتالي لن يقلل القوة التمييزية لتردد الوثيقة المعكوس.

فإذا كان لدى المصطلح تردد عال في الوثيقة وهو في الواقع ليس كلمة فريدة فعندئذ يكون الوزن الذي سيعطى من خلال TF-IDF التقليدية سيكون (0) بينما من خلال التقنية المقترحة سيتم تعيين وزن يتناسب مع تردد المصطلح.

وبهذا لن يتم تجاهل الكلمة الهامة لكافة الوثائق والتي لديها تردد عالي High TF في الوثيقة وبالتالي ستسمح لتردد المصطلح TF بالتأثير على الوزن النهائي.

وهذا يتطلب من أجل ذلك معالجة خاصة للكلمات الشائعة Stop Words حيث أنها متكررة جداً في كافة الوثائق.

- **ترجيح المصطلح المعدل بالاعتماد على تردد الفئة:**

TF-IDF Based on class frequency (W3)

المعادلة التالية تستخدم لتعديل تقنية ترجيح المصطلحات التقليدية بالاعتماد على تردد الفئة:

$$Weight\ (i, j) = \boldsymbol{W_{3ij} = TF(i,j) * Modified_{IDF(i)} * Class\ frequency}$$

هنا $\boldsymbol{Modified_{IDF(i)}}$ لها نفس المعنى الذي ذكر في الطريقة السابقة ويتم تعريف تردد الفئة Class Frequency كما يلي:

$$Class\ frequency = \frac{\boldsymbol{n(C_{ij})}}{\boldsymbol{N(C_i)}}$$

حيث تدل $n(C_{ij})$ على العدد الإجمالي للوثائق التي تحتوي المصـــطلح j ومصـــنفة في الفئة C_i والتي تنتمي لها الوثيقة i.

و $N(C_i)$ تمثل العدد الإجمالي لوثائق الفئة C_i.

ومن أجل تطبيع تردد الفئة Normalization فإنه من الضروري تقسيمها على عدد الوثائق في الفئة C_i وذلك للأسباب التالية:

الحالة الأولى: الفئة C_i لديها عشر وثائق والمصطلح i يظهر في الفئة C_i عشر مرات.

الحالة الثانية: الفئة C_i لديها مئة وثيقة والمصطلح i يظهر في الفئة C_i عشر مرات.

فالمصطلح يمثل الفئة في الحالة الأولى بشكل أفضل من الحالة الثانية لذلك فإن التقسيم على عدد الوثائق سيجعل التمثيل أفضل، حيث أنه بعد التقسيم فإن معامل تردد الفئة بالنسبة للمصطلح يصبح /١/ في الحالة الاولى و /٠,١/ في الحالة الثانية. وبالتالي فإن عدد الوثائق في الفئة ســـتطبِّع تردد المصطلحات في الفئة وتعطي أهمية للمصطلحات اعتماداً على تردد فئتها.

- **ترجيح المصطلح المعدل بالاعتماد على الطول الطبيعي:**

TF-IDF Based on normalized length (W4)

عدَّلت تقنية TF-IDF بالاعتماد على الطول الطبيعي من خلال المعادلة التالية:

$$Weight\ (i, j) = W_{4ij} = TF(i,j) * Modified_{IDF(i)} * Normalized\ length$$

حيث:

$$Normalized\ length = log_{10}\left(\frac{L}{L-TF(i,j)+1}\right)$$

وطول الوثيقة L هو مجموع ترددات المصطلحات الفريدة.

في المعادلة الســـابقة يتم إضـــافة معامل الطول الطبيعي إلى TF-IDF من أجل تطبيع المصـــطلح بالنسبة للطول، وذلك لإعطاء فرص متساوية لكافة الوثائق القصيرة والطويلة وسيتم مناقشة مثال فيما يلي:

لنفترض لدينا ٢٠ وثيقة في المجموعة، ثلاث وثائق منها (d1,d2,d3) تحتوي على المصـــطلح t وليكن:

TF (t, di)/ (document length of i), i € {1, 2, 3}

$\frac{15}{1000}$ ، $\frac{15}{200}$ ، $\frac{2}{10}$ على التوالي وباستخدامنا صيغة TF-IDF التقليدية فإن أوزان الوثائق الثلاثة هي: 28,456 ، 28,456 ، 3,794 على التوالي وبإجراء التعديل فإن قيم الـــ TF-IDF ستكون 0,412 ، 2,118 ، 0,410 على التوالي.

ووفقاً لذلك فإن الوثيقة بطول ٢٠٠ كلمة هي الوثيقة ذات المرتبة الأعلى في الترتيب والتي من المحتمل أن تكون الوثيقة المطلوبة (مرتبطة relevant، مفيدة useful).

وهذا منطقي لأن الوثيقة القصيرة ستتضمن معلومات قليلة جداً والوثائق الكبيرة جداً ولديها نفس تردد المصـــطلحات كما في الوثيقة متوســـطة الحجم والتي ربما تحتوي على معلومات متعلقة بجزء من الوثيقة ولن يكون الباقي مفيداً، لذلك تكون الأفضـلية للوثيقة متوسـطة الحجم والتي تكون مرغوبة في هذه الحالة [44].

وبالتالي نلاحظ أن الباحثين في هذه الدراسة اقترحوا أربع أساليب مختلفة أخذت كل طريقة جانب من الجوانب المميِّزة للوثائق وهذا بدوره يميز كل أسلوب بجانب دون آخر.

- في عام ٢٠١٩ اقترح كل من Ting ZHANG، Shuzhi Sam Ge من جامعة العلوم الالكترونية والتكنولوجيا الصينية طريقة جديدة لترجيح المصطلحات أطلقوا عليها اسم TF-IDF-ρ.

قام الباحثون في هذه الورقة من خلال دراســـتهم باقتراح فكرة جديدة تعتمد على القوة التمييزية للفئة والتي يؤثر بها المصطلح Class Discriminative Strength والاستفادة منها لتحسين خوارزمية ترجيح المصطلحات TF-IDF التقليدية [45].

حيث بيّنت الدراسة أن اختيار وحساب الوزن للعناصر المميزة للمستندات يحدد وبشكل كبير فيما إذا كان المستند قد تم تصنيفه بشكل صحيح أو لا.

وأظهرت الدراسة أن TF-IDF لا يظهر توزع الميزات Feature في كل فئات النصوص المصنّفة فعلى سبيل المثال إذا كان لدينا مصطلحين مميزين للوثائق في n وثيقة ولديهم نفس التردد في الوثائق المحددة إلا أن أحد المصـــطلحين يظهر في فئة محددة في حين أن الآخر يظهر في n من الفئات وبحســب خوارزمية TF-IDF التقليدية فإن وزن كلا المصــطلحين ســيكون واحداً وهذا غير معقول وبالتالي هذا سيؤثر بشكل سلبي على دقة تصنيف النصوص ولذلك تم وزن المصطلح من خلال إضـــافة عامل يتم من خلاله مفاضـــلة الوثائق وهو القوة التمييزية للفئة Class Discriminative

Strength والتي تمثل قوة تمييزية للمصـطلح والمسـاوي للعدد الإجمالي للفئة في المجموعة مقسـمة على عدد الفئات التي ورد فيها المصطلح كما يلي:

$$P = \frac{C}{C_i}$$

عندما تظهر الميزة في كافة الفئات، فإن C هي مساوية لـــ C_i وتكون قيمة P المقترحة هي /١/ ولا شيء سيميز الفئة التي يتعلق بها النص.

فالمصطلح سيكون له وزناً ذو قيمة منخفضة ولا يساوي الصفر عندما يظهر t_i في فئات أقل، وكلما كانت قيمة P أكبر فهذا ســـيزيد قدرة المصـــطلح على التمايز ولذلك فإن الطريقة المتبعة في هذه الدراسة هي أن وزن الميزة (المصطلح) متعلقاً ليس فقط بتكراره ضمن النص وبتردد الوثيقة المعكوسة في المجموعة وإنما أيضاً يتعلق بعدد الفئات التي تحتويه.

فعندما يكون المصـطلح المميز t_i ظاهراً في جميع الفئات فإن وزن المصـطلح سـيكون مسـاوٍ للوزن الحاصل عليه بالخوارزمية التقليدية. إن هذا لا يحافظ فقط على ميزات الخوارزمية التقليدية وإنما أيضاً يأخذ بالاعتبار القوة التمييزية لفئة المصطلح.

وبناءً على التحليل أعلاه، فإن معادلة الخوارزمية المحسنة والمقترحة في هذه الورقة هي كالتالي:

$$\boldsymbol{TF - IDF_{t,d} - P = \frac{N_{i,j}}{\sum_k N_{i,k}} * \log_{10}\left(\frac{N}{DF_i} + \frac{C}{C_i} + 1\right)}$$

وبالتالي نجد أنه من خلال هذه الطريقة تم تخصـيص وزناً أكبر للعناصـر المميزة والتي تظهر بنسـب أكبر كقوة لتمييز التصـــنيف أو الفئة وذلك من أجل تســليط الضـــوء على قدرة هذا المصـــطلح على التمييز بين أنواع مختلفة من النصوص [45].

وعلى الرغم من أن هنـاك العديد من المشـــاكل التي تؤثر في عملية تمثيل النصـــوص وترجيح المصـــطلحات إلا أنه لم يتم مناقشـــتها في هذه الورقة حيث اقتصـــرت الدراســـة على مناقشـــة عيوب خوارزمية ترجيح المصـطلح التقليدية TF-IDF ومن ثم تم حسـاب قيمة القوة التمييزية للفئة وإضـافتها للمعادلة.

وفي مواجهة المعلومات النصية الضخمة ما تزال النقطة التي يركز عليها الباحثون في مجال معالجة اللغة الطبيعية وأنظمة استرجاع المعلومات هي كيفية التعبير عن جميع الوثائق وتمثيلها بنجاح ومن ثم إعطاء التوزين المناسب للمصطلحات الأمر الذي يؤدي إلى تصنيفها بطريقة صحيحة وفعّالة.

مجموعة البيانات المعيارية CISI:

تم اختيار CISI كمجموعة بيانات Data Set معيارية وهي مجموعة من المقالات العلمية يبلغ عددها ١٤٦٠ تم نشرها بين عامي ١٩٦٩ و١٩٧٧. وتضم في معلوماتها اسم الكاتب وعنوان المقال بالإضافة إلى الملخص. كما تم تزويد هذه المجموعة بمجموعة استعلامات و نتائج الخبراء لكل استعلام [46].

ولقد تم استخدام مجموعة جزئية من هذه البيانات ضمّت ٣٠٠ نص لإجراء الاختبارات عليها. وتم مقارنة النتائج مع خوارزمية ترجيح المصطلحات التقليدية TF-IDF ولم يتم مقارنة نتائج هذا البحث مع نتائج الدراسات السابقة باعتبار أن الدراسات السابقة كانت في معظمها تطبق على النصوص المشرف عليها Supervised أي المصنّفة والطريقة التقليدية تعتبر من الطرق غير المشرف عليها Unsupervised أي أنها تطبق على مستوى كامل المجموعة.

كما أن البحث يركز على دراسة إمكانية تحليل النص باستخدام أدوات معالجة اللغة الطبيعية ومساهمتها في تطوير آلية داعمة لخوارزمية ترجيح المصطلحات التقليدية وبالتالي رفع كفاءة عمليات الاسترجاع بشكل إيجابي وفعال.

المنهجية المقترحة في ترجيح المصطلحات:

إن النهج الشائع في العديد من مجالات معالجة اللغة الطبيعية هو:

١. البحث عن الميزات Features في الوثيقة.
٢. تحديد الأهمية بالنسبة لتلك الميزات.
٣. إرسال الميزات الموزونة لإجراء القرار المناسب [47].

من خلال مخطط الفهرسة الذي تم توضيحه في الفصل الثاني تم اختيار الـ Terms المحددة لكل وثيقة ومن خلال عملية التوزين المقترحة سيتم تحديد أهمية كل مصطلح في الوثيقة.

وذلك من خلال المنهجية المقترحة التالية:

- **إضـافة معامل POS:** يحدد هذا المعامل مدى التطابق النحوي بين المصـطلح الوارد في الاسـتعلام والمصطلح الوارد في الوثيقة فعلى سبيل المثال:

المصطلح Book يختلف في معناه بين العبارتين التاليتين:

- ***Book*** a study seat in the Syrian Virtual Unversity to develop your scientific level.
- The Syrian Virtual University website has many important digital ***books***.

إن المصطلح book سيأخذ شكلاً موحداً بعد عملية التجريد في كلا النصين وبالتالي لابد من التمييز بين المصطلحين حين وروده كفعل أو حين وروده كاسم ومدى تطابقه مع رغبة المستخدم.

لذا تقترح الدراسة تصنيف الـ POS إلى فئات مطابقة لتصنيفهم ضمن عملية الفهرسة وهي:

- الفئة الأولى: NN، NNS، NNP، NNPS.
- الفئة الثانية: VB، VBG، VBN، VBP، VBZ، VBD.
- الفئة الثالثة: JJ، JJR، JJS.

ومن ثم يتم إعطاء قيم للمصطلحات المشتركة بين الاستعلام والنص حسب الجدول التالي:

POS=POS NER =NER	POS=POS One NER	POS=POS Not NER	POS ≠ POS Same Class	POS ≠ POS Not Same Class
1	**0.8**	**0.5**	**0.3**	0.1

وتكون معادلة حساب الـ POS للنص كما يلي:

$$POS_{\ Value} = \frac{\sum_{t=0}^{n} \text{POS Value}}{Common\ Term\ (n)}$$

حيث: n عدد المصطلحات المشتركة بين الاستعلام والنص.

ولأهمية عدد الـــــ Tokens الإجمالي للوثائق فقد تم أخذ هذه القيمة بعين الاعتبار لمعامل Pos. حيث أظهرت الاختبارات أن هذه القيمة تتأثر بنسبة المصطلحات المشتركة إلى العدد الكلي لمصطلحات النص لذلك تم تعديل المعادلة السابقة كما يلي:

$$POS = \frac{\sum_{t=0}^{n} \text{POS Value}}{Common\ Term\ (n)} \times \frac{Common\ Term\ n}{Total\ Count\ Of\ Term\ N}$$

لتصبح المعادلة بشكلها النهائي:

$$POS = \frac{\sum_{t=0}^{n} \text{POS Value}}{Total\ Count\ Of\ Term\ (N)}$$

- **إضافة معامل الترابط Correlation:**

هذا المعامل يدرس مدى ترابط الكلمات فيما بينها من خلال دراســة موقع الكلمات وحســاب التباعد بينها في النص. ويتم حساب معامل الترابط من خلال المعادلة التالية:

$$Corr = \frac{Common\ Terms\ (n)^2}{[\sum_{k=0}^{n-1} dis(term) + 1] * Count\ Term\ Of\ Query}$$

حيث: *n* *Common Terms* : عدد المصطلحات المشتركة بين الاستعلام والنص.

dis: هي التباعد بين المصطلح i والمصطلح i+1 أي:

$$Dis = Order\ Term_{i+1} - Order\ Term_i$$

وباعتبار أن المصـطحات سـتكون موزعة ضـمن النص مع احتمال وجود التكرار تم حسـاب القيمة الأقل للتباعد بين المصطلحات وحساب الترابط بالاعتماد على هذه القيم.

<u>وفيما يلي نورد مثالاً تطبيقياً لطريقة حساب هذا المعامل:</u>

لدينا الاستعلام التالي: $Information_1\ systems_2$

والنصين التاليين:

The evolution described below of one aspect of the NASA ***system***$_{11}$ and Technical ***Information***$_{14}$ Facility's machine search ***system***$_{18}$ may be of general interest to the documentation profession.

Recently a number of articles, books, and reports dealing with ***information***$_{11}$ ***systems***$_{12}$, i.e., document retrieval systems, have advanced the doctrine that such ***systems***$_{23}$ are to be evaluated in terms of the degree or percentage of relevancy they provide.

فتكون قيمة معامل ترابط مصطلحات الاستعلام مع النص الأول:

$$Corr_{text1} = \frac{2^2}{(3+1)*2} = \frac{4}{8} = 0.5$$

بينما تكون قيمة معامل ترابط مصطلحات الاستعلام مع النص الثاني:

$$Corr_{text2} = \frac{2^2}{(1+1)*2} = \frac{4}{4} = 1$$

وبالتالي نلاحظ أن موقع المصطلحات تعطي قيمة مضافة يمكن من خلالها إجراء عملية تفاضل بين الوثائق بناء على قيمتها.

- **خوارزمية ترجيح المصطلحات المقترحة NLP-TF-IDF:**

إن الطريقة المقترحة هي طريقة داعمة لخوارزمية ترجيح المصطلحات التقليدية وذلك من خلال الميزات التي تم فهرستها باستخدام مجموعة أدوات معالجة اللغة الطبيعية CoreNLP وبالتالي سيتم حساب الترجيح على مراحل هي:

المرحلة الأولى: حساب قيمة TF-IDF التقليدية.

المرحلة الثانية: حساب التشابه باستخدام Cosine Similarity بين الاستعلام والوثائق.

المرحلة الثالثة: تطبيق المعادلة التالية:

$$\boldsymbol{NLP\ Similarity = Cosine\ Sim_{TF-IDF} * \frac{[1 + (10 * POS) + Corr]}{2}}$$

الاختبارات النظرية:

تم اختيار الاستعلامات الـ ٣٠ الأولى لاجراء الاختبارات وباعتبار أن النصوص المختارة كمجموعة تجريبية هي ٣٠٠ نص وجدنا ثلاثة استعلامات نتائجها صفرية ضمن الوثائق المختارة وبالتالي لم يتم عرض نتائجها ضمن نتائج الاختبارات.

ولقد تم إجراء الاختبارات على عدة مراحل:

- **المرحلة التجريبية الأولى:**

في هذه المرحلة تم اختبار ترتيب الوثائق المسترجعة لكل من المعاملين: (POS، Corr) ومقارنتها بنتائج التصنيف باستخدام مقياس التشابه Cosine للوثائق المفهرسة حسب منهجية النظام المقترحة وباستخدام المجرد Porter وباعتبار أن القيم لم يتم إجراء Normalize فيما بينها لذلك تم استرجاع عدد محدد من الوثائق أي تثبيت القيمة TP + FP =25.

ولقد أظهرت الاختبارات النتائج التالية:

QueryID	TP+ FN	Porter	NLP index		POS		Corr	
		TF-IDF	TP	New	TP	New	TP	New
Query1	**17**	**10**	**11**	**3**	**7**	**2**	**5**	**2**
Query2	**5**	**2**	**0**	**0**	**1**	**0**	**0**	**0**
Query3	**10**	**7**	**6**	**0**	**4**	**1**	**7**	**1**
Query5	**10**	**2**	**3**	**1**	**2**	**1**	**0**	**0**
Query8	**3**	**0**	**0**	**0**	**1**	**1**	**0**	**0**
Query9	**5**	**3**	**3**	**0**	**2**	**0**	**1**	**0**
Query10	**6**	**5**	**4**	**0**	**2**	**0**	**2**	**0**
Query11	**31**	**6**	**6**	**1**	**10**	**5**	**6**	**2**
Query12	**3**	**0**	**1**	**1**	**1**	**1**	**2**	**2**
Query13	**32**	**16**	**15**	**3**	**10**	**1**	**7**	**4**
Query14	**1**	**1**	**1**	**0**	**1**	**0**	**1**	**0**
Query15	**22**	**9**	**9**	**1**	**7**	**4**	**3**	**0**
Query16	**4**	**2**	**2**	**0**	**0**	**0**	**1**	**1**
Query17	**4**	**1**	**1**	**0**	**1**	**0**	**1**	**0**
Query18	**3**	**2**	**2**	**0**	**1**	**0**	**1**	**0**
Query19	**12**	**6**	**6**	**0**	**4**	**0**	**3**	**0**
Query20	**16**	**5**	**7**	**3**	**6**	**3**	**5**	**2**
Query21	**7**	**2**	**2**	**0**	**0**	**0**	**0**	**0**
Query22	**22**	**2**	**3**	**1**	**3**	**2**	**3**	**1**
Query23	**23**	**6**	**6**	**1**	**2**	**1**	**2**	**2**
Query24	**13**	**6**	**5**	**1**	**2**	**0**	**1**	**1**
Query25	**7**	**2**	**3**	**1**	**0**	**0**	**0**	**0**
Query26	**12**	**7**	**7**	**0**	**5**	**2**	**4**	**1**
Query27	**31**	**10**	**12**	**4**	**10**	**3**	**5**	**4**
Query28	**8**	**4**	**4**	**0**	**3**	**0**	**4**	**1**
Query29	**18**	**5**	**6**	**3**	**4**	**1**	**3**	**1**
Query30	**23**	**9**	**10**	**1**	**9**	**3**	**8**	**3**
		Sum=130	**Sum=135**	**25**	97	31	75	27

*الجدول ٨ – **نتائج TP لكل من POS,Cor,CosineSimilarity***

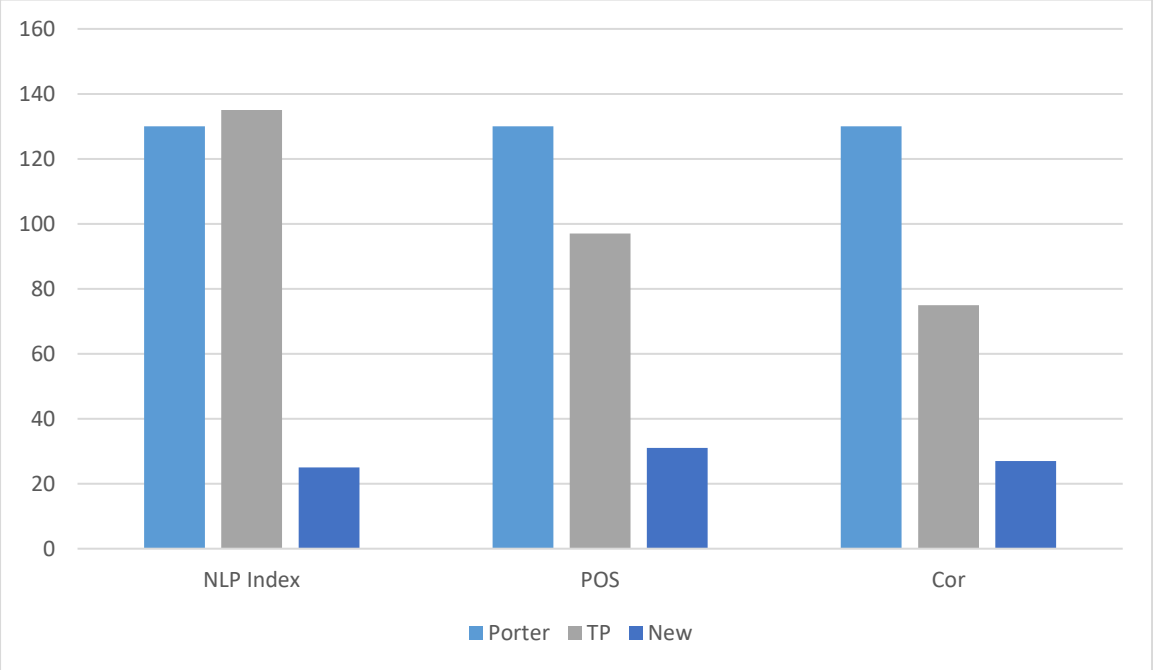
وبدراسة النتائج نجد ما يلي:

– تفوقت الفهرسة المعتمدة على مجموعة أدوات معالجة اللغة الطبيعية على المجرد Porter حيث بلغ مجموع الوثائق المسترجعة ذات الصلة TP باستخدام Porter /١٣٠/ وثيقة/ في حين كان عدد الوثائق ذات الصلة المسترجعة باستخدام الفهرسة المقترحة /١٣٥/ وثيقة/.

– بلغ عدد الوثائق الجديدة باستخدام فهرسة NLP /٢٥/ وثيقة جديدة مختلفة عن الوثائق التي ظهرت في فهرسة Porter الأمر الذي يشير إلى أهمية عملية الفهرسة في أنظمة استرجاع المعلومات.

– أظهر المعاملان POS و Corr عدداً من الوثائق الجديدة لم تظهر ضمن نتائج الـ Cosine similatrity بنسبة زادت عن الـ ٣٠% في نتائجهما ونسبة زادت عن ٢٠% وثائق جديدة بالنسبة لنتائج فهرسة Porter.

وفيما يلي مخطط يظهر ملخص نتائج المرحلة الأولى:



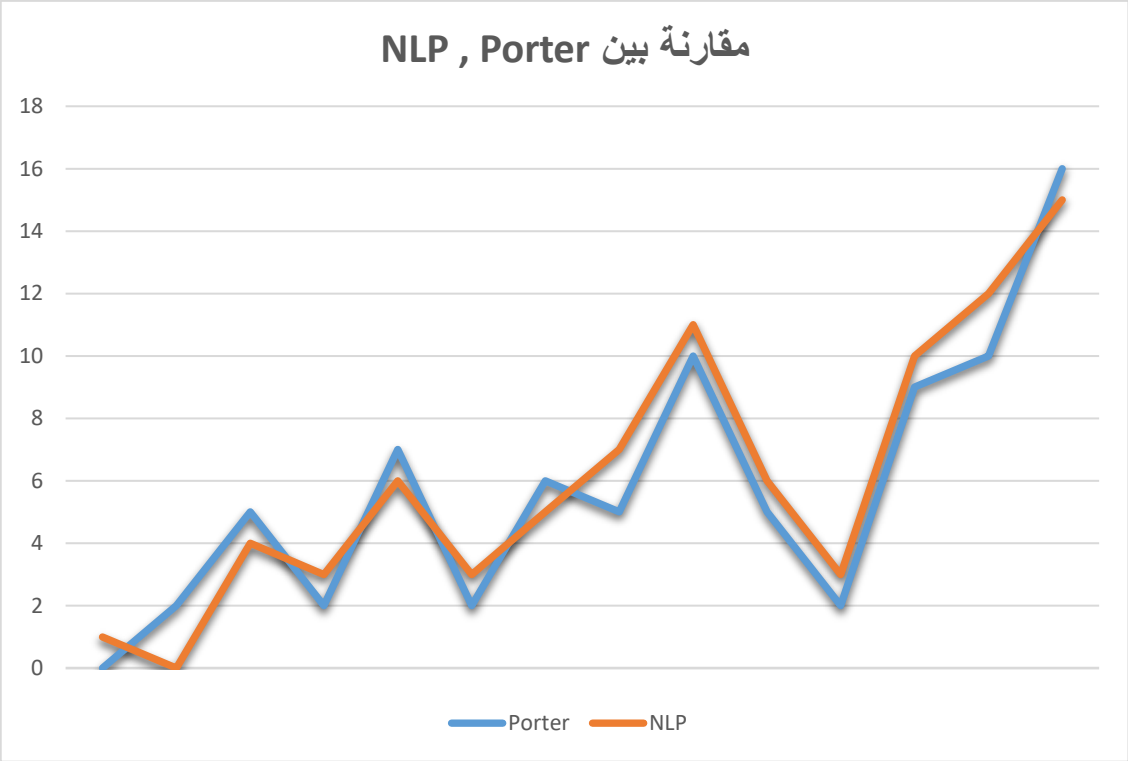
- **المرحلة التجريبية الثانية:**

في هذه المرحلة تم مقارنة عوامل تقييم أنظمة استرجاع المعلومات الثلاثة (Recall, Presicion, F1) لنتائج TF-IDF التقليدية بين فهرسة Porter ومنهجية الفهرسة المقترحة NLP باستخدام Cosine:

Query	Porter Index			NLP Index			$F1_{NLP}$-$F1_{Porter}$
	Recall	Precision	F1	Recall	Precision	F1	
1	**0.58824**	**0.4**	**0.47619**	**0.64706**	**0.44**	**0.52381**	**0.04762**
2	**0.4**	**0.08**	**0.13333**	**0**	**0**	**0**	**-0.1333**
3	**0.7**	**0.28**	**0.4**	**0.6**	**0.24**	**0.34286**	**-0.0571**
5	**0.2**	**0.08**	**0.11429**	**0.3**	**0.12**	**0.17143**	**0.05714**
8	**0**	**0**	**0**	**0**	**0**	**0**	**0**
9	**0.6**	**0.12**	**0.2**	**0.6**	**0.12**	**0.2**	**0**
10	**0.83333**	**0.2**	**0.32258**	**0.66667**	**0.16**	**0.25806**	**-0.0645**
11	**0.19355**	**0.24**	**0.21429**	**0.19355**	**0.24**	**0.21429**	**0**
12	**0**	**0**	**0**	**0.33333**	**0.04**	**0.07143**	**0.07143**
13	**0.5**	**0.64**	**0.5614**	**0.46875**	**0.6**	**0.52632**	**-0.0351**
14	**1**	**0.04**	**0.07692**	**1**	**0.04**	**0.07692**	**0**
15	**0.40909**	**0.36**	**0.38298**	**0.40909**	**0.36**	**0.38298**	**0**
16	**0.5**	**0.08**	**0.13793**	**0.5**	**0.08**	**0.13793**	**0**
17	**0.25**	**0.04**	**0.06897**	**0.25**	**0.04**	**0.06897**	**0**
18	**0.66667**	**0.08**	**0.14286**	**0.66667**	**0.08**	**0.14286**	**0**
19	**0.5**	**0.24**	**0.32432**	**0.5**	**0.24**	**0.32432**	**0**
20	**0.3125**	**0.2**	**0.2439**	**0.4375**	**0.28**	**0.34146**	**0.09756**
21	**0.28571**	**0.08**	**0.125**	**0.28571**	**0.08**	**0.125**	**0**
22	**0.09091**	**0.08**	**0.08511**	**0.13636**	**0.12**	**0.12766**	**0.04255**
23	**0.26087**	**0.24**	**0.25**	**0.26087**	**0.24**	**0.25**	**0**
24	**0.46154**	**0.24**	**0.31579**	**0.38462**	**0.2**	**0.26316**	**-0.0526**
25	**0.28571**	**0.08**	**0.125**	**0.42857**	**0.12**	**0.1875**	**0.0625**
26	**0.58333**	**0.28**	**0.37838**	**0.58333**	**0.28**	**0.37838**	**0**
27	**0.32258**	**0.4**	**0.35714**	**0.3871**	**0.48**	**0.42857**	**0.07143**
28	**0.5**	**0.16**	**0.24242**	**0.5**	**0.16**	**0.24242**	**0**
29	**0.27778**	**0.2**	**0.23256**	**0.33333**	**0.24**	**0.27907**	**0.04651**
30	**0.3913**	**0.36**	**0.375**	**0.43478**	**0.4**	**0.41667**	**0.04167**

الجدول ٩ – مقارنة قيم التقييم بين فهرسة Porter ، NLP

وفيما يلي مخطط يوضح فارق قيم TP بين Porter، NLP مع TP+FN:



وقد بينت نتائج هذه المرحلة مايلي:

- عامل التقييم Recall ارتفع وسطياً بقيمة ٠,٧٢% تقريباً باستخدام الفهرسة في النظام المقترح عن الفهرسة باستخدام المجرِّد Porter.
- عامل التقييم Precision ارتفع وسطياً بقيمة ٠,٧٤% باستخدام الفهرسة في النظام المقترح عن الفهرسة باستخدام المجرّد Porter.
- عامل التقييم F1 ارتفع بقيمة وسطية ٠,٧٢% باستخدام الفهرسة في النظام المقترح عن الفهرسة باستخدام المجرّد Porter.

وبالرغم من تفوّق فهرسة Porter بـ/٥/ استعلامات إلا أن النتائج الإجمالية لمعاملات التقييم الثلاثة كانت تشير جميعها إلى تفوق النتائج التي تم الوصول إليها باستخدام نظام الفهرسة المقترح، حيث تفوقت الفهرسة باستخدام مجموعة أدوات معالجة اللغة الطبيعية بـ/٩/ استعلامات وتشابهت النتائج في /١٣/ استعلام آخر. مشيرين إلى أنه لم يتم تعديل خوارزمية ترجيح المصطلحات التقليدية في هذه المرحلة وإنما تم استخدامها وحساب التشابه باستخدام Cosine Similarity.

- **المرحلة التجريبية الثالثة:** في هذه المرحلة تم تعديل آلية حساب التشابه بين الاستعلام والوثائق ومقارنة نتائج TF-IDF التقليدية وNLP-TF-IDF المعدّلة حيث ظهرت النتائج التالية:

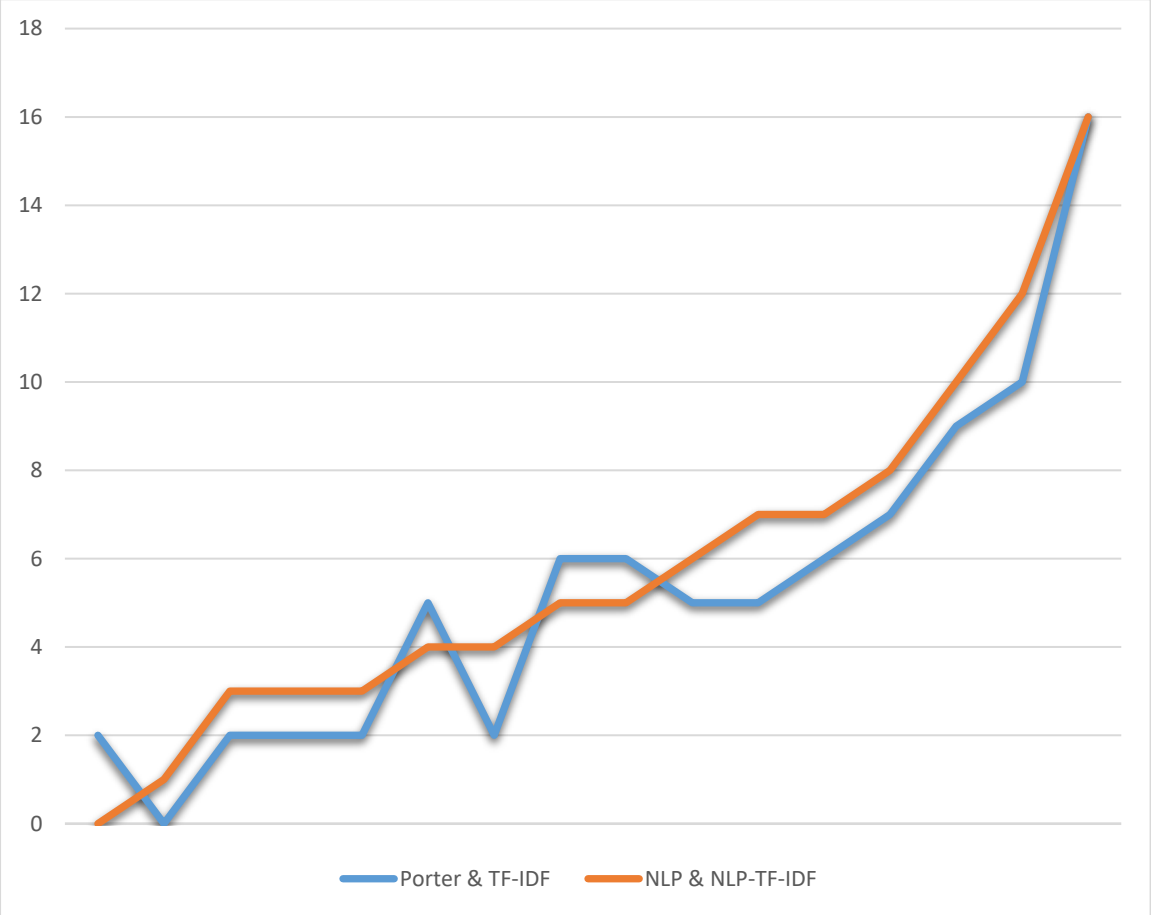
QueryID	TP+ FN	Porter & TF-IDF	NLP & TF-IDF		NLP & NLP-TF-IDF	
		TP	TP	New	TP	New
Query1	**17**	**10**	**11**	**3**	**10**	**2**
Query2	**5**	**2**	**0**	**0**	**0**	**0**
Query3	**10**	**7**	**6**	**0**	**8**	**1**
Query5	**10**	**2**	**3**	**1**	**3**	**1**
Query8	**3**	**0**	**0**	**0**	**0**	**0**
Query9	**5**	**3**	**3**	**0**	**3**	**0**
Query10	**6**	**5**	**4**	**0**	**4**	**0**
Query11	**31**	**6**	**6**	**1**	**7**	**1**
Query12	**3**	**0**	**1**	**1**	**1**	**1**
Query13	**32**	**16**	**15**	**3**	**16**	**2**
Query14	**1**	**1**	**1**	**0**	**1**	**0**
Query15	**22**	**9**	**9**	**1**	**9**	**1**
Query16	**4**	**2**	**2**	**0**	**3**	**1**
Query17	**4**	**1**	**1**	**0**	**1**	**0**
Query18	**3**	**2**	**2**	**0**	**2**	**0**
Query19	**12**	**6**	**6**	**0**	**6**	**0**
Query20	**16**	**5**	**7**	**3**	**7**	**3**
Query21	**7**	**2**	**2**	**0**	**2**	**0**
Query22	**22**	**2**	**3**	**1**	**4**	**2**
Query23	**23**	**6**	**6**	**1**	**5**	**1**
Query24	**13**	**6**	**5**	**1**	**5**	**1**
Query25	**7**	**2**	**3**	**1**	**3**	**1**
Query26	**12**	**7**	**7**	**0**	**7**	**0**
Query27	**31**	**10**	**12**	**4**	**12**	**5**
Query28	**8**	**4**	**4**	**0**	**4**	**0**
Query29	**18**	**5**	**6**	**3**	**6**	**3**
Query30	**23**	**9**	**10**	**1**	**10**	**1**
		Sum=130	**Sum=135**	**25**	**Sum=139**	**27**

الجدول ١٠ – مقارنة قيم المعامل TP بين فهرسة Porter ، NLP وآلية الترجيح المقترحة

ولقد بينت نتائج التجارب ما يلي:

- ارتفع عدد الاســـتعلامات التي تفوّق فيها النظام المقترح على Porter ليصـــبح ١١ اســـتعلام وتشابهت النتائج في ١٢ استعلام.
- انخفض عدد الاستعلامات التي كان قد تفوّقت بها فهرسة Porter بتطبيق آلية الترجيح المقترحة إلى ٤ استعلامات فقط.
- وكنتيجة إجمالية ارتفع عدد الوثائق الصحيحة باستخدام أسلوب الترجيح المقترح إلى /١٣٩/ وثيقة في حين كان قد بلغ باستخدام Porter /١٣٠/ وثيقة فقط.
- بلغ عدد الوثائق الجديدة التي ظهرت /٢٧/ وثيقة مختلفة عن تلك التي ظهرت باســتخدام فهرســة Porter مع خوارزمية ترجيح المصطلحات التقليدية.

وفيما يلي مخطط بياني لاختلافات النتائج بين الطريقة التقليدية(Porter&TF-IDF) والآلية المقترحة:



وفيما يلي نستعرض قيم معاملات التقييم الثلاثة:

Query	Porter Index & TF-IDF			NLP Index & TF-IDF			NLP Index & NLP-TF-IDF		
	Recall	Prec	F1	Recall	Prec	F1	Recall	Prec	F1
1	0.58824	0.4	0.47619	0.64706	0.44	0.52381	0.58824	0.4	0.47619
2	0.4	0.08	0.13333	0	0	0	0	0	0
3	0.7	0.28	0.4	0.6	0.24	0.34286	0.8	0.32	0.45714
5	0.2	0.08	0.11429	0.3	0.12	0.17143	0.3	0.12	0.17143
8	0	0	0	0	0	0	0	0	#DIV/0!
9	0.6	0.12	0.2	0.6	0.12	0.2	0.6	0.12	0.2
10	0.83333	0.2	0.32258	0.66667	0.16	0.25806	0.66667	0.16	0.25806
11	0.19355	0.24	0.21429	0.19355	0.24	0.21429	0.22581	0.28	0.25
12	0	0	0	0.33333	0.04	0.07143	0.33333	0.04	0.07143
13	0.5	0.64	0.5614	0.46875	0.6	0.52632	0.5	0.64	0.5614
14	1	0.04	0.07692	1	0.04	0.07692	1	0.04	0.07692
15	0.40909	0.36	0.38298	0.40909	0.36	0.38298	0.40909	0.36	0.38298
16	0.5	0.08	0.13793	0.5	0.08	0.13793	0.75	0.12	0.2069
17	0.25	0.04	0.06897	0.25	0.04	0.06897	0.25	0.04	0.06897
18	0.66667	0.08	0.14286	0.66667	0.08	0.14286	0.66667	0.08	0.14286
19	0.5	0.24	0.32432	0.5	0.24	0.32432	0.5	0.24	0.32432
20	0.3125	0.2	0.2439	0.4375	0.28	0.34146	0.4375	0.28	0.34146
21	0.28571	0.08	0.125	0.28571	0.08	0.125	0.28571	0.08	0.125
22	0.09091	0.08	0.08511	0.13636	0.12	0.12766	0.18182	0.16	0.17021
23	0.26087	0.24	0.25	0.26087	0.24	0.25	0.21739	0.2	0.20833
24	0.46154	0.24	0.31579	0.38462	0.2	0.26316	0.38462	0.2	0.26316
25	0.28571	0.08	0.125	0.42857	0.12	0.1875	0.42857	0.12	0.1875
26	0.58333	0.28	0.37838	0.58333	0.28	0.37838	0.58333	0.28	0.37838
27	0.32258	0.4	0.35714	0.3871	0.48	0.42857	0.3871	0.48	0.42857
28	0.5	0.16	0.24242	0.5	0.16	0.24242	0.5	0.16	0.24242
29	0.27778	0.2	0.23256	0.33333	0.24	0.27907	0.33333	0.24	0.27907
30	0.3913	0.36	0.375	0.43478	0.4	0.41667	0.43478	0.4	0.41667

الجدول ١١ – مقارنة قيم التقييم بين فهرسة Porter ، NLP وآلية الترجيح المقترحة

وتكون النتائج الإجمالية كما يلي:

Query	Comparesion with Porter			Comparision with NLP		
	Recall	Prec	F1	Recall	Prec	F1
1	**0**	**0**	**0**	**-0.05882**	**-0.04**	**-0.04762**
2	**-0.4**	**-0.08**	**-0.13333**	**0**	**0**	**0**
3	**0.1**	**0.04**	**0.05714**	**0.2**	**0.08**	**0.11428**
5	**0.1**	**0.04**	**0.05714**	**0**	**0**	**0**
8	**0**	**0**	**0**	**0**	**0**	**0**
9	**0**	**0**	**0**	**0**	**0**	**0**
10	**-0.16666**	**-0.04**	**-0.06452**	**0**	**0**	**0**
11	**0.03226**	**0.04**	**0.03571**	**0.03226**	**0.04**	**0.03571**
12	**0.33333**	**0.04**	**0.07143**	**0**	**0**	**0**
13	**0**	**0**	**0**	**0.03125**	**0.04**	**0.03508**
14	**0**	**0**	**0**	**0**	**0**	**0**
15	**0**	**0**	**0**	**0**	**0**	**0**
16	**0.25**	**0.04**	**0.06897**	**0.25**	**0.04**	**0.06897**
17	**0**	**0**	**0**	**0**	**0**	**0**
18	**0**	**0**	**0**	**0**	**0**	**0**
19	**0**	**0**	**0**	**0**	**0**	**0**
20	**0.125**	**0.08**	**0.09756**	**0**	**0**	**0**
21	**0**	**0**	**0**	**0**	**0**	**0**
22	**0.09091**	**0.08**	**0.0851**	**0.04546**	**0.04**	**0.04255**
23	**-0.04348**	**-0.04**	**-0.04167**	**-0.04348**	**-0.04**	**-0.04167**
24	**-0.07692**	**-0.04**	**-0.05263**	**0**	**0**	**0**
25	**0.14286**	**0.04**	**0.0625**	**0**	**0**	**0**
26	**0**	**0**	**0**	**0**	**0**	**0**
27	**0.06452**	**0.08**	**0.07143**	**0**	**0**	**0**
28	**0**	**0**	**0**	**0**	**0**	**0**
29	**0.05555**	**0.04**	**0.04651**	**0**	**0**	**0**
30	**0.04348**	**0.04**	**0.04167**	**0**	**0**	**0**
AVG	**0.024106**	**0.013333**	**0.014926**	**0.016914**	**0.005926**	**0.007678**

الجدول ١٢ – فرق قيم التقييم بين آلية الترجيح المقترحة وكل من فهرسة Porter ، NLP .

وكنتيجة نهائية نجد:

أن آلية ترجيح المصطلحات المقترحة قد رفعت القيمة الوسطى للمعامل F1 بقيمة (٠,٧%) عن استخدام خوارزمية ترجيح المصطلحات التقليدية للبيانات المفهرسة باستخدام مجموعة أدوات معالجة اللغة الطبيعية،

وازدادت قيمة المعامل F1 كقيمة وسطية بنسبة (**١,٥%**) تقريباً مقارنة بـــــ Porter والطريقة التقليدية لترجيح المصطلحات بينما بلغت القيمة العظمى لارتفاع قيمة المعامل F1 (**٩,٧%**).

ملخص الفصل الثالث:

إن عملية تحليل النصوص ومعالجة اللغة الطبيعية هي وسيلة لفهم رغبة المستخدم وبالتالي فإن التحليل الصحيح للنصوص يؤدي إلى نتائج استرجاع أكثر دقة وتلبي رغبة المستخدم.

وإن مجموعة أدوات معالجة اللغة الطبيعية استطاعت تحقيق أشواط متقدمة في هذا المجال وبالتالي يمكن الاستفادة من إمكانية هذه الأدوات في رفع كفاءة أنظمة استرجاع المعلومات من خلال تطوير آليات جديدة لترجيح المصطلحات تعتمد في مضمونها على الخصائص المميزة التي يتم استخلاصها من تحليل هذه الأدوات للنصوص.

كما أن ظهور مجموعة أدوات كنسخ مطوَّرة عن مجموعة أدوات Stanford CoreNLP يساعد الباحثين على إيجاد المزيد من الميزات التي تساهم في تحديد الوثائق ذات الصلة برغبة المستخدم بشكل أكثر صحة الأمر الذي يزيد من فعالية ودقة نتائج أنظمة استرجاع المعلومات.

ولابد أن نذكر بأنه وعلى الرغم من قصور هذه الأدوات في معالجة لواحق الأسماء والصفات إلا أن النتائج بينت تفوّق أسلوب النظام المقترح كقيم إجمالية على الأسلوب التقليدي في جميع معاملات تقييم نظم استرجاع المعلومات وبالتالي فإن تطوير عملية الفهرسة ومعالجة هذا القصور قد يؤدي إلى نتائج أكثر فعالية ودقة.[٤]

[٤] تم نشر معلومات الفصل الثاني والثالث كورقة علمية بعنوان تطوير آلية جديدة لترجيح المصطلحات من خلال تحليل النصوص ومعالجة اللغة الطبيعية بتاريخ ٢٠٢١/١٢/٢٩ في مجلة: International Journal Of Engineering Research & Technology (IJERT) والورقة مرفقة في نهاية البحث.

الفصل الرابع: التطبيق العملي

مقدمة:

التطبيق عبارة عن Desktop App يتم من خلاله تحليل نصـــوص الوثائق والاســـتعلامات باســـتخدام مجموعة الأدوات Stanford CoreNLP ومن ثم فهرستها بأسلوب منظم ومعرِّف لبعض ميزات كل وثيقة ومن ثم الاســـتفادة من هذه الميزات المعرِّفة للوثائق في عملية تطوير آلية ترجيح المصـــطلحات التقليدية TF-IDF بهدف الحصـول على النتائج الأقرب لرغبة المسـتخدم. كما سـيتم فهرسـة النصـوص باسـتخدام المجرّد Porter وذلك من أجل عمليات الاختبار ولمقارنة نتائج الآلية الجديدة للترجيح بالطريقة التقليدية.

كما يمكن أن تكون إجراءات هذا التطبيق نواة لإنشـاء تطبيق ويب Web App يحاكي محرك بحث أولي Prototype باستخدام آليه الفهرسة وترجيح المصطلحات المقترحة.

الدراسة التصميمية:

- **تصميم قاعدة البيانات:**

١. قاعدة البيانات "NLP_DB":

هي قاعدة البيانات الخاصــة بفهرســة الوثائق من خلال اســتخدام مجموعة أدوات معالجة اللغة الطبيعية Stanford CoreNLP. وهي تتألف من الجداول التالية:

الرقم	Table Name
١	Documents
٢	DocumentTerms
٣	Tokens
٤	Terms

الجدول ١٣ – جداول قاعدة بيانات NLP Database للنظام المقترح

٢. قاعدة البيانات "Porter_DB":

لإجراء عمليات المقارنة واختبار نتائج النظام المقترح تم إنشـــاء قاعدة بيانات منفصـــلة بحيث يتم فهرســة النصــوص باســتخدام Porter ولقد ضــمت قاعدة البيانات جداولاً مشــابهة لقاعدة بيانات NLP_DB.

- **توصيف جداول قاعدة البيانات:**

نســتعرض فيما يلي الحقول الخاصــة بكل جدول من جداول قاعدة البيانات وخصــائص هذه الحقول وأنماط بياناتها:

بطـاقة جـــدول الوثائق				
الرقـــم		**اسم الجـــدول**		
1		**Documents**		
اسم الحقل باللغة العربية	**اسم الحقل ضمن قاعدة البيانات**	**النمـــط**	**Null**	**القيمـة الافتراضيـة**
معرف الوثيقة	DocumentID	INT	×	Auto
عنوان الوثيقة	Title	NVARCHAR(Max)	✓	-
اسم الكاتب	Author	NVARCHAR(150)	✓	-
مسار ملف	Path	NVARCHAR(200)	✓	-
نص الوثيقة	Text	NVARCHAR(Max)	×	-
طول الوثيقة	Length	INT	×	-
القـيود				
اسـم القيد	**النوع**	**الحقول المشاركة**		**الحقول الخارجية المشاركة**
DocumentID_PK	Primary Key	ID		–
DocumentID_Identity	Identity	ID		–
T- SQL				
إنشاء الجدول				

```
-- Create Documents Table
USE NLP_DB
GO
CREATETABLE Documents
      (
      [DocumentID] [int] IDENTITY(1,1) NOT NULL,
      [Author] [nvarchar](150) NULL,
      [Title] [nvarchar](max) NULL,
      [Path] [nvarchar](200) NULL,
      [Text] [nvarchar](max) NOT NULL,
      [Length] [decimal](18, 3) NOT NULL,
      CONSTRAINT [PK_Documents] PRIMARY KEY CLUSTERED ([DocumentID])
      )
```

الجدول ١٤ – بطاقة جدول الوثائق

بطـاقة جـــدول المصطلحات				
الرقـــم		**اسم الجـــدول**		
2		**Terms**		
اسم الحقل باللغة العربية	**اسم الحقل ضمن قاعدة البيانات**	**النمـــط**	**Null**	**القيمـة الافتراضيـة**
معرف المصطلح	TermID	INT	×	Auto
المصطلح	Term	NVARCHAR(100)	×	-
تردد الوثيقة المعكوس	IDF	DECIMAL(18,3)	✓	-
القـيود				
اسـم القيد	**النوع**	**الحقول المشاركة**	**الحقول الخارجية المشاركة**	
TermID_PK	Primary Key	TermID	–	
TermID_Identity	Identity	TermID	–	
T- SQL				
إنشاء الجدول				

```
-- Create Terms Table
USE NLP_DB
GO
CREATETABLE Terms
      (
      [TermID] [int] IDENTITY(1,1) NOT NULL,
      [Term1] [nvarchar](100) NOT NULL,
      [IDF] [decimal](18, 3) NULL,
      CONSTRAINT [PK_Terms] PRIMARY KEY ([TermID])
      )
```

الجدول ١٥ – بطاقة جدول المصطلحات

بطـاقة جـــدول الرموز				
الرقـــم		**اسم الجـــدول**		
3		**Tokens**		
اسم الحقل باللغة العربية	**اسم الحقل ضمن قاعدة البيانات**	**النمـــط**	**Null**	**القيمة الافتراضية**
معرف الرمز	TokenID	INT	×	Auto
معرف المصطلح	TermID	INT	✓	-
معرف الوثيقة	DocumentID	INT	×	-
الكلمة الأصلية	Word	NVARCHAR(100)	×	-
تجذير الكلمة	Lemma	NVARCHAR(100)	×	-
التحليل النحوي	POS	NVARCHAR(100)	×	-
تحليل الكيانات	NER	NVARCHAR(100)	×	-
ترتيب المصطلح	Order	INT	×	-

القـيود

اسم القيد	النوع	الحقول المشاركة	الحقول الخارجية المشاركة
TermID_PK	Primary Key	TokenID	–
TermID_Identity	Identity	TokenID	–
TokenID_TermID_FK	Foreign Key	TermID	Terms(TermID)
TokentID_DocID_FK	Foreign Key	DocumentID	Documents(DocumentID)

T- SQL

إنشاء الجدول

```
-- Create Tokens Table
USE NLP_DB
GO
CREATE TABLE Tokens (
	[TokenID] [int] IDENTITY(1,1) NOT NULL,
	[TermID] [int] NULL,
	[DocumentID] [int] NOT NULL,
	[Word] [nvarchar](100) NOT NULL,
	[Lemma] [nvarchar](100) NOT NULL,
	[POS] [nvarchar](100) NOT NULL,
	[NER] [nvarchar](100) NOT NULL,
	[Order] [int] NOT NULL,
	CONSTRAINT [PK_Tokens] PRIMARY KEY CLUSTERED ([TokenID])
)
GO
ALTER TABLE [dbo].[Tokens]  WITH CHECK ADD  CONSTRAINT
[FK_Tokens_Documents] FOREIGN KEY([DocumentID])
REFERENCES [dbo].[Documents] ([DocumentID])
GO

ALTER TABLE [dbo].[Tokens] CHECK CONSTRAINT [FK_Tokens_Documents]
GO

ALTER TABLE [dbo].[Tokens]  WITH CHECK ADD  CONSTRAINT
[FK_Tokens_Terms] FOREIGN KEY([TermID])
REFERENCES [dbo].[Terms] ([TermID])
GO

ALTER TABLE [dbo].[Tokens] CHECK CONSTRAINT [FK_Tokens_Terms]
GO
```

الجدول ١٦ – بطاقة جدول الرموز

بطـاقة جـــدول مصطلحات الوثائق

الرقـــم	اسم الجـــدول
4	DocumentTerms

اسم الحقل باللغة العربية	اسم الحقل ضمن قاعدة البيانات	النمـــط	Null	القيمـة الافتراضيـة
معرف الرمز	ID	INT	×	Auto
معرف الوثيقة	DocumentID	INT	×	-
معرف المصطلح	TermID	INT	×	-

تكرار المصطلح	Frequency	INT	×	-
تردد المصطلح	TF	DECIMAL(18,3)	×	-
قيمة الترجيح	TF_IDF	DECIMAL(18,3)	✓	-

القـيود

اسم القيد	النوع	الحقول المشاركة	الحقول الخارجية المشاركة
ID_PK	Primary Key	ID	–
ID_Identity	Identity	ID	–
ID_TermID_FK	Foreign Key	TermID	Terms(TermID)
ID_DocID_FK	Foreign Key	DocumentID	Documents(DocumentID)

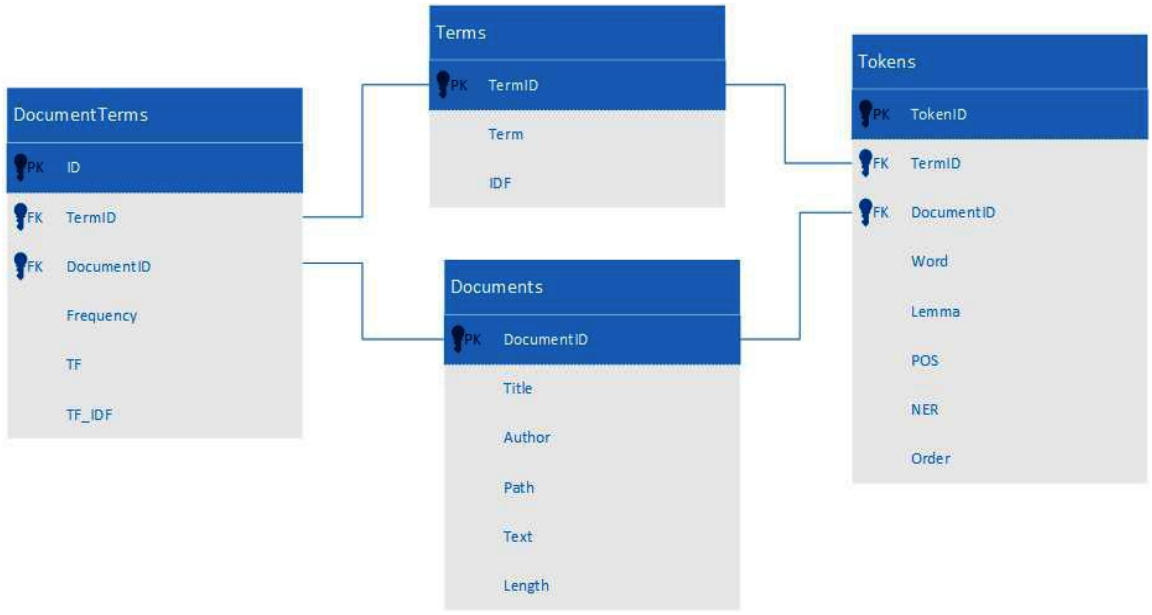
T- SQL

إنشاء الجدول

```
-- Create DocumentTerms Table
USE NLP_DB
GO
CREATE TABLE DocumentTerms(
      [ID] [int] IDENTITY(1,1) NOT NULL,
      [DocumentID] [int] NOT NULL,
      [TermID] [int] NOT NULL,
      [Frequency] [int] NOT NULL,
      [NormalizedFrequency] [decimal](18, 2) NOT NULL,
      [TF_IDF] [decimal](18, 3) NULL,
      CONSTRAINT [PK_DocumentTermFrequencies] PRIMARY KEY CLUSTERED
([ID])
GO

ALTER TABLE [dbo].[DocumentTermFrequencies]  WITH CHECK ADD
CONSTRAINT [FK_DocumentTermFrequencies_Documents] FOREIGN
KEY([DocumentID])
REFERENCES [dbo].[Documents] ([DocumentID])
GO

ALTER TABLE [dbo].[DocumentTermFrequencies] CHECK CONSTRAINT
[FK_DocumentTermFrequencies_Documents]
GO

ALTER TABLE [dbo].[DocumentTermFrequencies]  WITH CHECK ADD
CONSTRAINT [FK_DocumentTermFrequencies_Terms] FOREIGN KEY([TermID])
REFERENCES [dbo].[Terms] ([TermID])
GO

ALTER TABLE [dbo].[DocumentTermFrequencies] CHECK CONSTRAINT
[FK_DocumentTermFrequencies_Terms]
GO
```

الجدول ١٧ – بطاقة جدول مصطلحات الوثائق

- **مخطط UML لقاعدة البيانات NLP_DB:**



الشكل ٩ - مخطط UML لقاعدة بيانات NLP في النظام المقترح

الدراسة التنفيذية:

- **الأدوات البرمجية:**

١. SQL Server

٢. Stanford CoreNLP

٣. Visual Studio

٤. C#

- **لمحة عن التطبيق:**

تم بناء ClassLibrary خاصة بكل من عمليات NLP وأخرى لعمليات Porter:

١. **الـــ ClassLibrary الخاصة بعمليات NLP:** تقوم بالفهرسة من خلال معالجة اللغة الطبيعية وبالإضافة إلى عمليات البحث وفيما يلي نستعرض أهم خطوات العمل البرمجي:

- **عمليات الفهرسة Index:** تتم الفهرسة من خلال مجموعة من الإجرائيات وهي:

١. Annotation(): إجراء يقوم باســـتدعاء مكتبة Stanford CoreNLP لتحليل النص وأجراء عمليات معالجة اللغة الطبيعية حيث يقوم بداية بتقســـيم النص إلى مجموعة من الرموز Tokens ومن ثم فرز هذه الرموز إلى مجموعات حسب موقعها من الكلام POS

٢. ()Processtokens: يقوم بمالعجة الرموز Tokens حسب تصنيفهم (POS) ومن ثم تصفية الرموز المصنفة كـــكلمات شائعة StopWords ومن ثم فهرسة بقية الرموز كمصطلحات Terms ضمن قاعدة بيانات النظام المقترح NLP_DB.

- **عمليات البحث Search:**

تتم عمليات البحث من خلال الإجرائيات:

١. ()QueryAnnotation: تقوم بمعالجة الاستعلام من خلال مكتبة معالجة اللغة الطبيعية Stanford CoreNLP وتقسيم الاستعلام إلى مجموعة من الرموز.

٢. ()ProcessQueryTokens: يقوم بمعالجة الرموز الخاصة بالاستعلام وتصفية الكلمات الشائعة Stopwords.

٣. ()TFIDFSearch: حساب الـــ Consine Similarity بين مصطلحات الاستعلام والنصوص المفهرسة وإعادة النصوص الـ ٢٥ الأكثر تقارباً مع الاستعلام.

٤. ()FormulaSearch: يقوم هذا الإجراء بحساب عدة معاملات وهي:

- Cosine Sililarity: حساب التشابه بين الاستعلام والنصوص المفهرسة.
- المعامل POS:حساب التقارب بين الاستعلام والنص وفقاً لموقع المصطلحات النحوي حسب المعادلة المقترحة.

$$\boldsymbol{POS = \frac{\sum_{t=0}^{n} \textbf{POS Value}}{\textbf{Total Count Of Term } (N)}}$$

- المعامل Corr: حساب التقارب بين الاستعلام والنص وفقاً لموضع المصطلحات وتقاربها حسب المعادلة المقترحة.

$$\boldsymbol{Corr = \frac{\textbf{Common Terms } (n)^2}{[\sum_{k=0}^{n-1} dis(term) + 1] * \textbf{Count Term Of Query}}}$$

- حساب آلية الترجيح الجديدة المقترحة بالاعتماد على المعادلة:

$$\boldsymbol{NLP\ Similarity = Cosine\ Sim_{TF-IDF} * \frac{[1 + (10 * POS) + Corr]}{2}}$$

٢. **الـــ ClassLibrary الخاصة بعمليات Porter:** تقوم بعمليات الفهرسة وإدارة عمليات البحث الخاصة بالمجرّد Porter وآلية ترجيح المصطلحات التقليدية TF-IDF:

- **عمليات الفهرسة Index:**

١. ()Annotation: إجراء يقوم بتحليل النص وتقسيم النص إلى مجموعة من الرموز Tokens.
٢. ()Processtokens: يقوم بمعالجة الرموز Tokens واستدعاء عمليات Porter لمعالجة اللواحق ومن ثم تصفية الرموز المصنفة كـكلمات شائعة StopWords وبعد ذلك يتم فهرسة بقية الرموز كمصطلحات Terms ضمن قاعدة البيانات Porter_DB.

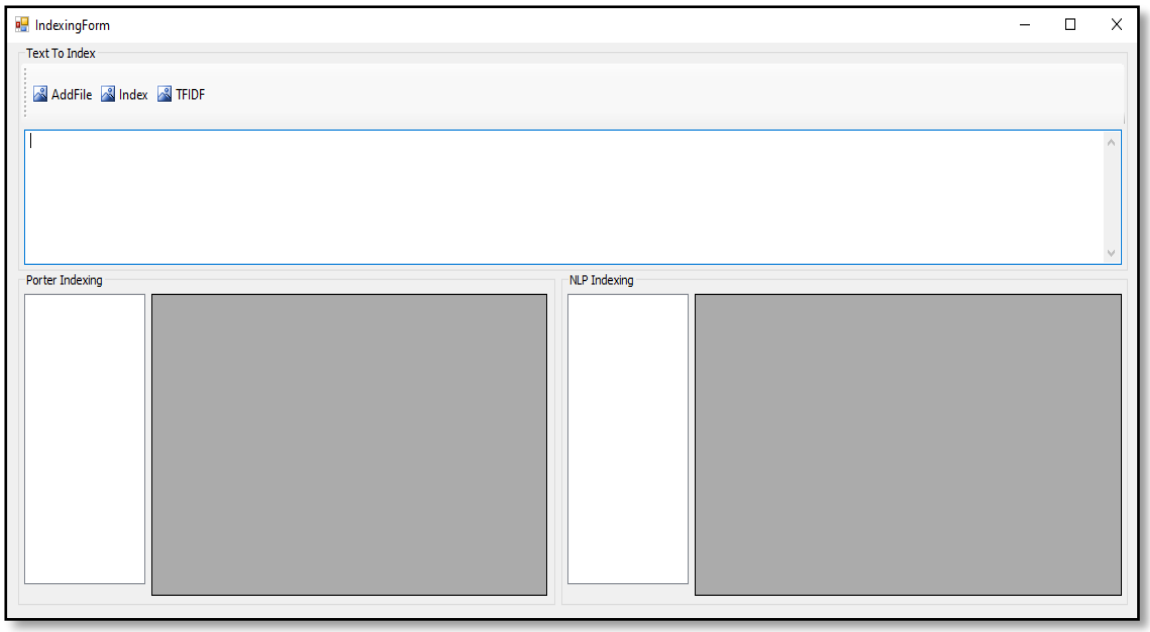
- **عمليات البحث Search:**

١. ()QueryAnnotation: تقوم بمعالجة الاستعلام وتقسيمه إلى مجموعة من الرموز Tokens.
٢. ()ProcessQueryTokens: يقوم بمعالجة الرموز الخاصة بالاستعلام واستدعاء عمليات Porter لمعالجة اللواحق بالإضافة إلى تصفية الكلمات الشائعة Stopwords.
٣. ()Search: حساب ذالـــ Consine Similarity لخوارزمية الترجيح التقليدية TF-IDF بين مصطلحات الاستعلام والنصوص المفهرسة وإعادة النصوص الـ ٢٥ الأكثر تقارباً مع الاستعلام.

واجهات التطبيق:

تم بناء تطبيق وهو عبارة Desktop App وذلك لإجراء الاختبارات اللازمة والبحث حيث تضمن التطبيق أربعة واجهات رئيسية نستعرضها فيما يلي:

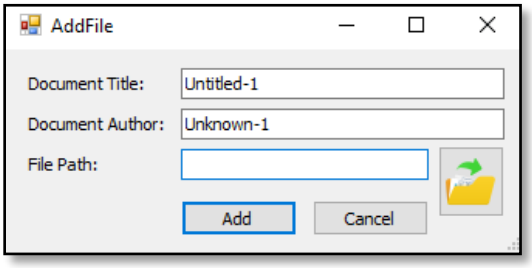
- **واجهة الفهرسة:**



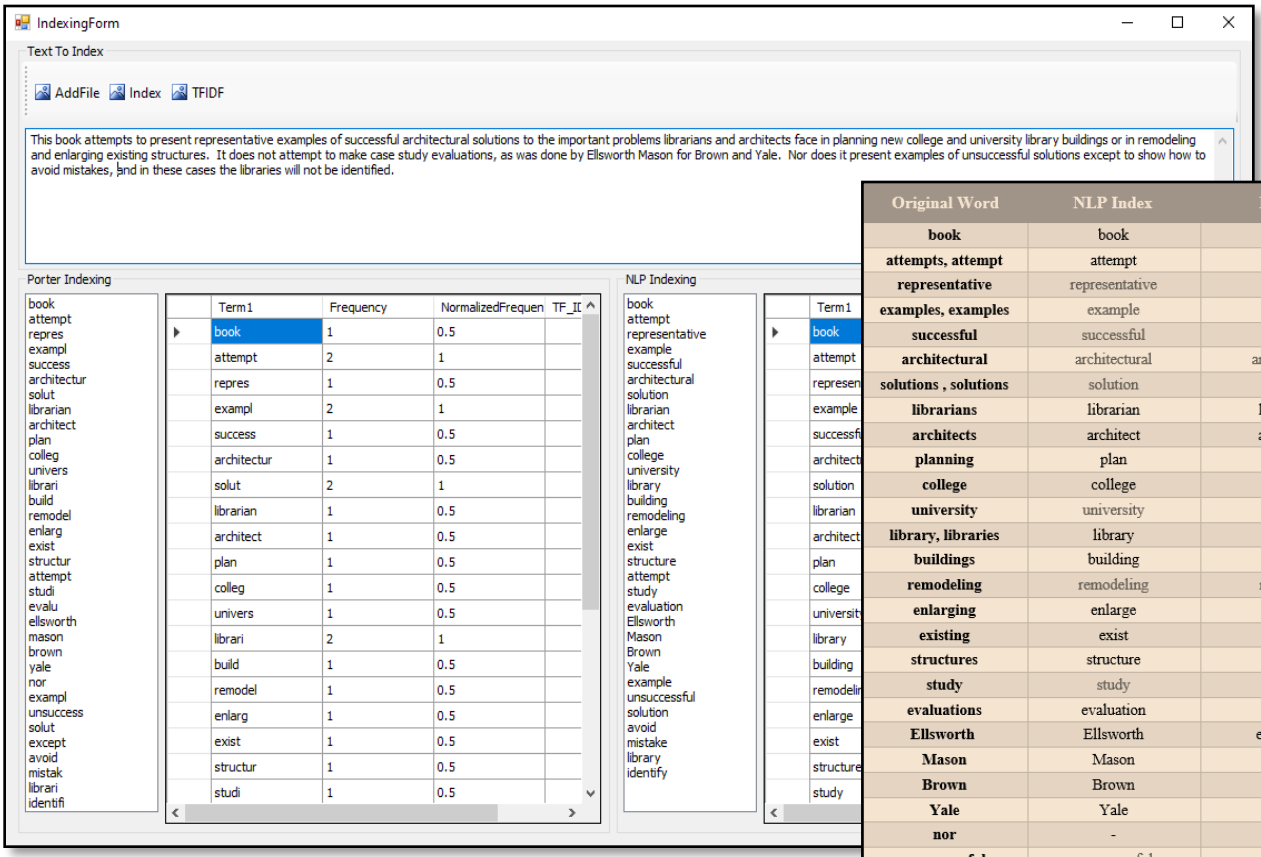
الشكل ١٠ - واجهة الفهرسة ضمن التطبيق

تتيح هذه الواجهة إمكانية فهرسة نص بإحدى طريقتين:

- عملية النسخ واللصق ضمن النافذة المحددة لذلك.
- من خلال الزر Add File ثم تحديد مسار الملف الذي يحوي على نص الوثيقة.



وفيما يلي نستعرض نتيجة فهرسة أحد النصوص ضمن التطبيق:



Original Word	NLP Index	Porter
book	book	book
attempts, attempt	attempt	attempt
representative	representative	repress
examples, examples	example	exampl
successful	successful	success
architectural	architectural	architectur
solutions , solutions	solution	solut
librarians	librarian	librarian
architects	architect	architect
planning	plan	plan
college	college	colleg
university	university	univers
library, libraries	library	librari
buildings	building	build
remodeling	remodeling	remodel
enlarging	enlarge	enlarg
existing	exist	exist
structures	structure	structur
study	study	studi
evaluations	evaluation	evalu
Ellsworth	Ellsworth	ellsworth
Mason	Mason	mason
Brown	Brown	brown
Yale	Yale	yale
nor	-	nor
unsuccessful	unsuccessful	unsuccess
except	-	except
avoid	avoid	avoid
mistakes	mistake	mistak
identified	identify	identifi

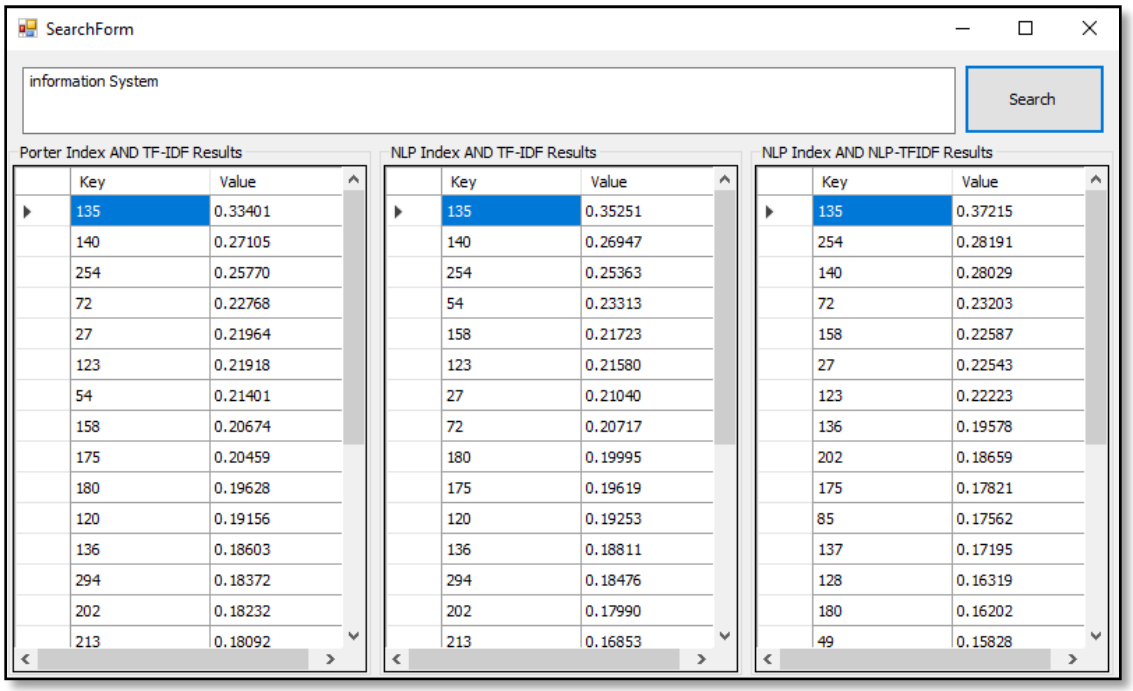
الشكل ١١ـ نتيجة فهرسة أحد النصوص في النظام المقترح ضمن التطبيق

ونذكر هنا أن الزر TFIDF: هو لإعادة حساب الـــ IDF لمصطلحات الوثائق باعتبار أن هذا المعامل يتم حسابه على مستوى المجموعة وإن إضافة أي نص وفهرسته يؤدي إلى تعديل قيمة هذا المعامل.

- **واجهة البحث:**

هذه الواجهة مخصصة لإجراء عمليات البحث التقليدية من خلال إدخال استعلام وإظهار النتائج على ثلاثة مستويات:

(Porter Index & TF-IDF ، NLP Index & TF-IDF ، NLP Index & NLP-TFIDF).



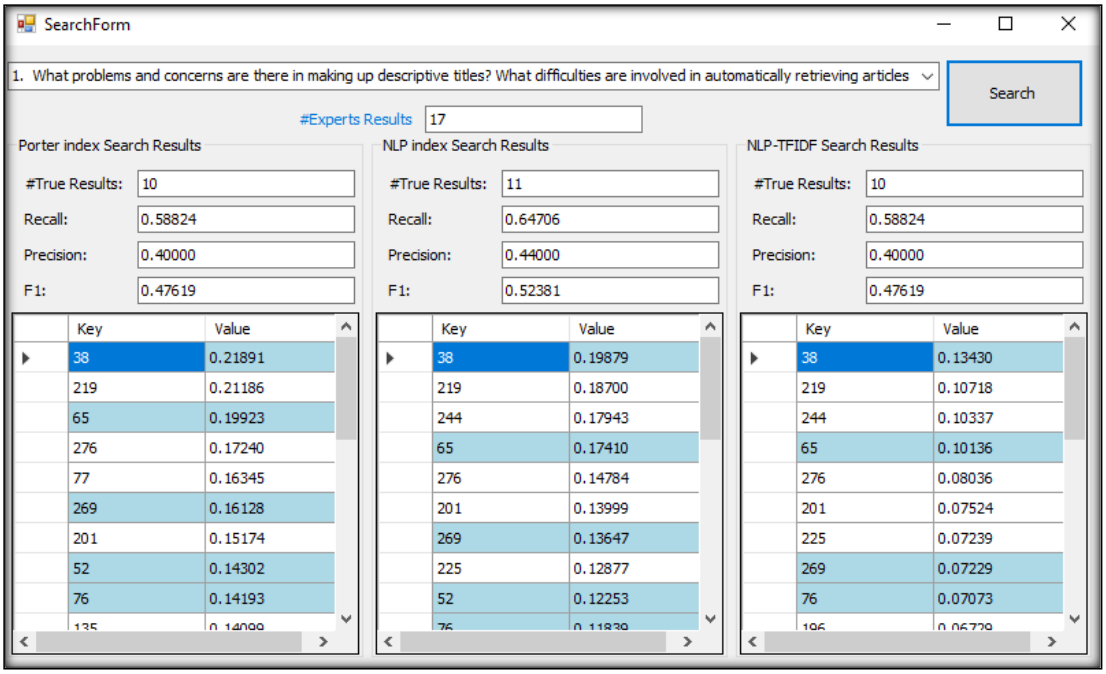
الشكل ١٢ - واجهة البحث ضمن التطبيق

- **واجهة الاختبارات:**

هذه الواجهة مخصصة لاختبار نتائج الاستعلامات الـــ ٣٠ ضمن الـــ ٣٠٠ نص وهي المجموعة الجزئية التي تم اختيارها من قاعدة البيانات المعيارية CISI. حيث يتم المقارنة على ثلاثة مستويات:

(Porter Index & TF-IDF ، NLP Index & TF-IDF ، NLP Index & NLP-TFIDF) بالإضـــافة إلى عرض قيم معاملات التقييم الثلاثة (Recall، Precision، F1) مع بيان الوثائق الصحيحة والتي تتطابق مع نتائج الخبراء.

وفيما يلي نستعرض مثالاً لتطبيق الاختبارات على الاستعلامات:



الشكل ١٣ - واجهة اختبار استعلامات قاعدة البيانات المعيارية CISI

Query	Porter Index & TF-IDF			NLP Index & TF-IDF			NLP Index & NLP-TF-IDF		
	Recall	Prec	F1	Recall	Prec	F1	Recall	Prec	F1
1	0.58824	0.4	0.47619	0.64706	0.44	0.52381	0.58824	0.4	0.47619
2	0.4	0.08	0.13333	0	0	0	0	0	0
3	0.7	0.28							
5	0.2	0.08							
8	0	0							
9	0.6	0.12							
10	0.83333	0.2							
11	0.19355	0.24	0.21429	0.19355	0.24	0.21429	0.22581	0.28	0.25
12	0	0	0	0.33333	0.04	0.07143	0.33333	0.04	0.07143
13	0.5	0.64	0.5614	0.46875	0.6	0.52632	0.5	0.64	0.5614
14	1	0.04	0.07692	1	0.04	0.07692	1	0.04	0.07692
15	0.40909	0.36	0.38298	0.40909	0.36	0.38298	0.40909	0.36	0.38298
16	0.5	0.08	0.13793	0.5	0.08	0.13793	0.75	0.12	0.2069
17	0.25	0.04	0.06897	0.25	0.04	0.06897	0.25	0.04	0.06897
18	0.66667	0.08	0.14286	0.66667	0.08	0.14286	0.66667	0.08	0.14286
19	0.5	0.24	0.32432	0.5	0.24	0.32432	0.5	0.24	0.32432
20	0.3125	0.2	0.2439	0.4375	0.28	0.34146	0.4375	0.28	0.34146
21	0.28571	0.08	0.125	0.28571	0.08	0.125	0.28571	0.08	0.125
22	0.09091	0.08	0.08511	0.13636	0.12	0.12766	0.18182	0.16	0.17021
23	0.26087	0.24	0.25	0.26087	0.24	0.25	0.21739	0.2	0.20833
24	0.46154	0.24	0.31579	0.38462	0.2	0.26316	0.38462	0.2	0.26316
25	0.28571	0.08	0.125	0.42857	0.12	0.1875	0.42857	0.12	0.1875
26	0.58333	0.28	0.37838	0.58333	0.28	0.37838	0.58333	0.28	0.37838
27	0.32258	0.4	0.35714	0.3871	0.48	0.42857	0.3871	0.48	0.42857
28	0.5	0.16	0.24242	0.5	0.16	0.24242	0.5	0.16	0.24242
29	0.27778	0.2	0.23256	0.33333	0.24	0.27907	0.33333	0.24	0.27907
30	0.3913	0.36	0.375	0.43478	0.4	0.41667	0.43478	0.4	0.41667

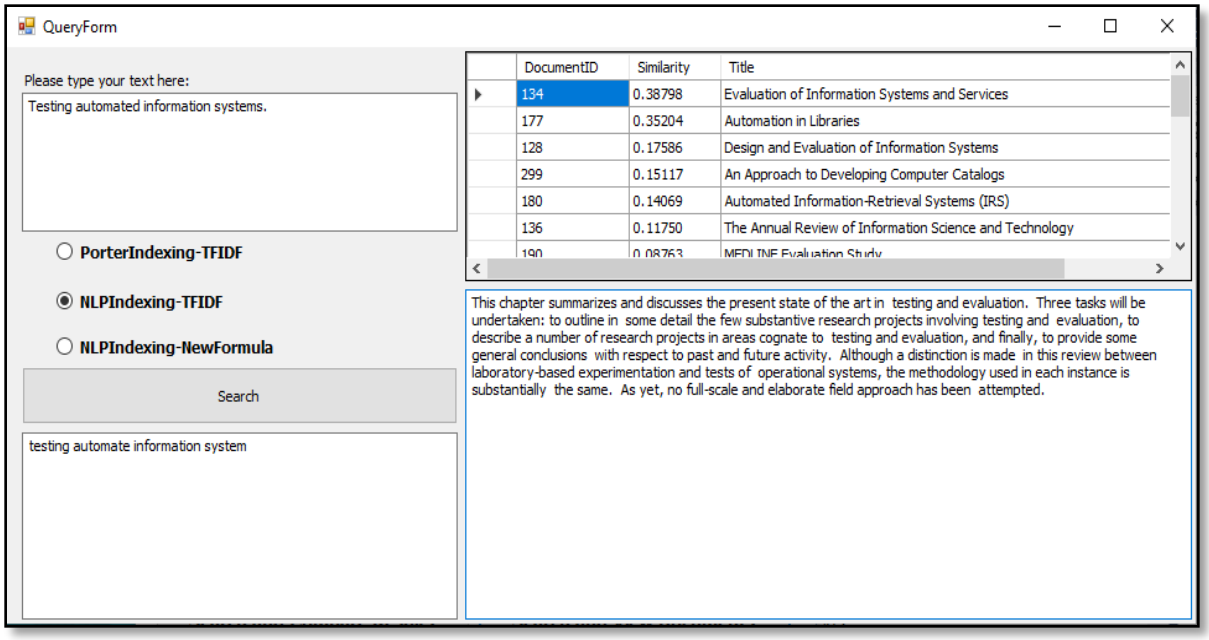
Query	Porter Index & TF-IDF			NLP Index & TF-IDF			NLP Index & NLP-TF-IDF		
	Recall	Prec	F1	Recall	Prec	F1	Recall	Prec	F1
1	0.58824	0.4	0.47619	0.64706	0.44	0.52381	0.58824	0.4	0.47619

- **واجهة استعراض الوثائق:**

هذه الواجهة مخصصة لإجراء عمليات البحث وإظهار الوثائق ذات القيم الـــ ٢٥ الأعلى من حيث التشابه من خلال إدخال استعلام وإظهار النتائج على ثلاثة مستويات:

(NLP Index & NLP-TFIDF ، NLP Index & TF-IDF ، Porter Index & TF-IDF)

وتختلف عن واجهة البحث السابقة بإمكانية استعراض نصوص الوثائق



الشكل ١٤ - واجهة استعراض نصوص نتائج البحث ضمن التطبيق

حيث تختلف نتائج البحث باختلاف أســـلوب الفهرســـة والتوزين كما تتيح هذه الواجهة إمكانية استعراض الاستعلام بعد عملية التجذير.

نتيجة البحث:

إن تحليل النصـوص والاسـتعلامات وفهم رغبة المسـتخدم يؤدي إلى نتائج اسـترجاع أكثر دقة وتزيد من فعالية أنظمة استرجاع المعلومات. وإن مجموعة أدوات معالجة اللغة الطبيعية هي أسلوب مفيد في عملية تحليل النصوص والاستعلامات للوصول إلى النتائج الأكثر ملاءمة لمتطلبات المستخدم.

فقد أكدت لنا نتائج الاختبارات أهمية عملية الفهرسة البالغة وتأثيرها على نتائج الاسترجاع رغم تطبيق آلية ترجيح موحدة وهي خوارزمية ترجيح المصطلحات التقليدية TF-IDF. بالإضافة إلى إمكانية الاستفادة من ميزات أدوات معالجة اللغة الطبيعية في تطوير آلية ترجيح جديدة من خلال إضـافة معاملات تسـاهم في تحديد الوثائق ذات الصـلة برغبة المسـتخدم. ولابد أن نذكر بأنه وعلى الرغم من قصـور هذه الأدوات في

معالجة لواحق الأســـماء والصـــفات إلا أن النتائج بينت تفوّق أســـلوب النظام المقترح كقيم إجمالية على الأســـلوب التقليدي في جميع معاملات تقييم نظم اســـترجاع المعلومات وبالتالي فإن إعادة هيكلة عملية الفهرسة بالاعتماد على مجموعة أدوات معالجة اللغة الطبيعية ودراسة موضوع معالجة اللواحق قد يؤدي إلى نتائج أكثر فعالية ودقة.

توصيات البحث:

- تطبيق مجموعة أدوات معالجة اللغة الطبيعية Stanza واختبار إمكانياتها على اللغة العربية.
- إعادة هيكلة عملية الفهرسة بالاعتماد على مجموعة أدوات معالجة اللغة الطبيعية ودراسة موضوع معالجة اللواحق ضمن فهرسة NLP.
- دراسة إمكانية إضافة ميزات أخرى ودراسة تأثيرها على نتائج عمليات الاسترجاع وفاعلية البحث وخاصة علاقات التبعية والتحليل الدلالي والارتباطات بين المصطلحات.
- إن عملية توسيع الميزات Features المستخلصة من الوثائق يجب أن تتم بدقة وبدراسة موضوعية تدرس مدى قدرة النظام على الاستفادة من هذه الميزات مقارنة بتأثيرها على كفاءة النظام لأن زيادة الميزات والاعتماد عليها بشكل عشوائي قد يؤدي إلى أثر سلبي فيما يتعلق بكفاءة نظم استرجاع المعلومات.

الورقة العلمية الأولى التي نشرت في مجلة IJERT:

International Journal of
Engineering Research & Technology
ISSN : 2278 - 0181, www.ijert.org
(Published by : ESRSA Publications)

CERTIFICATE
OF PUBLICATION

This is to certify that

Manal Sheikh Oghli

Has published a research paper entitled

Comparison of basic Information Retrieval Models

In IJERT, Volume 10, Issue 9 , September - 2021

Registration No: IJERTV10IS090092 **Date:** 23-09-2021 **Chief Editor,IJERT**

International Journal of
Engineering Research & Technology
ISSN : 2278 - 0181, www.ijert.org
(Published by : ESRSA Publications)

CERTIFICATE
OF PUBLICATION

This is to certify that

Muhammad Mazen Almustafa

Has published a research paper entitled

Comparison of basic Information Retrieval Models

In IJERT, Volume 10, Issue 9 , September - 2021

Registration No: IJERTV10IS090092 **Date:** 23-09-2021 **Chief Editor,IJERT**

Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 09, September-2021

Comparison of basic Information Retrieval Models

Manal Sheikh Oghli
Web Science program, Syrian Virtual University
Damascus, Syria

Muhammad Mazen Almustafa
Web Science program, Syrian Virtual University
Damascus, Syria

***Abstract*— it was necessary to provide an information retrieval model capable of meeting user requirements in an effective manner, in light of the increasing growth and the huge amount of digital information in recent decades.**

The Information Retrieval process depends on the matching process largely between representations of the user's desire, which is expressed through the query, and stores of information to return results related to the user's desire.

As the challenge today is no longer, providing information stores, therefore, the biggest challenge that facing researchers is the ability to retrieve an appropriate information related to the needs of the user.

In this paper, we review basic information retrieval models, that are classified according to the mathematical dimension to arrive to a description of the most effective model in retrieval operations, and to demonstrate to the most widespread among other models, which is, Vector Space Model and its strength to excel in the event that the weaknesses points, that it suffers from are addressed, which is represented by not setting fixed standards, for terms' weighting ,in addition to this model that assumes independence of terms from each other.

Keywords*— *Information Retrieval Models IRM, Boolean Model, Vector Space Model VSM, Probabilistic Models, Term Weighting)

I. INTRODUCTION

The term Information Retrieval was first used during Calvin Mooers' presentation of a research paper at a 1950 conference, as he wrote, "The problem under discussion here is machine searching and retrieval of information from storage according to a specification by subject… ". [1] [2]

Mooers used this term to describe the process by which a user could convert their need for information into an actual list containing a set of useful references, and explained that information retrieval is another, more general name for producing a demand bibliography. [2]

Information retrieval models considered a blueprint for implementing an actual retrieval system as the retrieval system predicts and explains what the user wants by analysing the user-defined query. [3]

Models provide different techniques and methods for matching stored documents to a query. The main goal of information retrieval models is to find documents relevant to the information needs of a large group of documents. [4]

II. THE AIM OF THE RESEARCH

A brief historical overview of the emergence of the concept and development of Information Retrieval, and its multiple models, with a brief description of the working method and processing algorithms in each of them, with a focus on the aspects that distinguish each model from the other, and then compare these models with the aim of describing the best model can meet with the requirements of User in terms of performance and related results.

Despite the emergence of many information retrieval models and the development that occurred in this area of knowledge, most models still suffer from their limitations in meeting the user's desire to obtain the required information. In addition, the many models that appeared in information retrieval systems depend in their way of working on the basic models represented by the Boolean model, the probabilistic model and the vector space model.

The aim of this paper is to compare these models to guide workers in this field to the simpler and more efficient model by reviewing the points of strengths and deficiencies in it to overcome them in future researches in order to reach an effective and efficient retrieval model.

III. CLASSIFICATION OF INFORMATION SYSTEMS RETRIEVAL MODELS:

Information retrieval systems models differ among themselves in general in the way of representing documents and queries, and the methods of matching and arranging. These models could be classified according to two dimensions: the mathematical base and the characteristics of the model.

Retrieval models were classified according to the mathematical base dimension into:

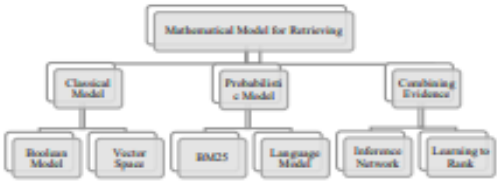


Fig. 1: Classification of Mathematical Information Retrieval Models [5]

The following is a review of the most important features that distinguish each of these models, explaining the Positives and limitations of each.

A. Classical Models:

1) Boolean Model:

The Boolean model is the first form of information retrieval [3]. One of the oldest and simplest models in this field, as it based on logical algebra [4], and the principle of Exact Match [3]. There is no room for partial matching in this form.

Where documents are represented by a set of terms (also known as index terms) [4] [6] .Then it is classification into a class in which the terms of the query mentioned, and a class in which the terms not mentioned. This classification means that there is no sort of arrangement in evaluating the relevance of the documents to the query.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 09, September-2021

The user's information needs are identified by a combination of basic logical transactions defined by George Bool, which are three (AND, OR, NOT) meaning intersection, addition and difference, and are used during formulation of the query [3] [4] [5].

Despite their limitations, these models still used in many retrieval systems. It also gives expert users the feeling that they are able to control the system largely more than others [7].

The development of these models appeared, for example, the Extended Boolean Model, which was described in an article published in ACM in 1983 by (Gerard Salton, Edward A. Fox, and Harry Wu).

Through this model, it became possible to perform partial matching and term weighting. It combines the characteristics of a Vector Space Model (it will be explained in the next paragraph) with the characteristics of a Boolean model [6]. However, the judgment in this model on the importance of the documents to the query, depending on whether or not the term mentioned in them, without taking into account the repetition or its mentioning in the one document that is taken into account by the other models.

2) Vector Space Model:

Gerard Salton and his colleagues suggested this model in 1983 [8]. It was based on the similarity criterion proposed by Hans Peter Luhn in 1957, who was the first one suggested the statistical model for searching for information based on the similarity criterion between inquiries and documents. HP Luhn formulated the similarity criterion as follows:

"The more two representations agreed in given elements and their distribution, the higher would be the probability of their representing similar information…" [9]

Based on this criterion, Salton and his colleagues considered that both documents and queries could be represented as vectors in Euclidean space, so that each term is assigned an independent dimension, and then they calculated the similarity between vectors using the cosine between the vectors representing both the document and the query. [8]

This model was considered one of the algebraic models and the most widespread. The text is represented in it by a vector of terms independent of each other. The terms are words and phrases that represent indexing terms. [7]

According to the statistical model, the content of the document is viewed as a Bag-of-Words [6].

This means that the document content includes unordered and irregular frequency terms within the document content.

Index terms set weights for both documents and queries. Then measuring the similarity or distance between the document and the query through several methods, we mention as example (Dot Product, Euclidean, Manhattan, Cosine…). [5]

The biggest challenge facing this model is to set the appropriate value for the vector components, and this problem is known as Term Weighting in addition to terms independence as this model does not take into account the terms link between them. [3]

B. Probabilistic Models:

Maron, Kuhns suggested the initial idea of probabilistic models in information retrieval systems in a paper titled Probabilistic Indexing and Information Retrieval published in 1960 [7] [10].

It was considered the first scientific work to deal with the use of the probabilistic approach in retrieving information, and on this basis, what is known as Probabilistic Indexing appeared, and since then many models have been developed that rely on different techniques to estimate probabilities.

Probabilistic models are based on the Probabilistic Ranking Principle (PRP) [6] [7]:

"If a reference retrieval system's response to each request is a ranking of the documents in the collection in order of decreasing probability of relevance to the user who submitted the request, where the probabilities are estimated as accurately as possible on the basis of whatever data have been made available to the system for this purpose, the overall effectiveness of the system to its user will be the best that is obtainable on the basis of those data." [11].

Feedback plays an important role in such probability models, as it uses the historical information of the document to calculate its probabilities of relevance to the query.

Probabilistic models differ depending on the assumptions on which they are based [6].

The classic probabilistic model introduced by S.E. Robertson and K. Parck Jones in 1976 assumes the independence of the term as this model was defined as the binary independence retrieval model (BIR). This model was depended on the concept that the set of documents stored in an information retrieval system was divided into two independent binary groups from each other; the first group is known as the join group whose content is related to the query, and the second group is known as not join group whose content is irrelevant to the query. [12]

As the set of all possible outcomes is called the sample space and thus the probability of P(R) in the sample space can carry one of two values {0,1} where Irrelevant = 0, Relevant = 1. [3]

We must mention that the value of P (...) changes with the change of the random variable converter R, so when we assign different values to the random variable converter R or different values of the sample space. Therefore, we are talking about different values of probability P. [3]

This model depends on the method of using probability theory as the basis for the treatment process. Where, according to the followed probability model, the probabilities in which the document is relevant to the user's query are calculated or estimated.

The basic idea of this model is the hypothesis, that the information retrieval system includes documents related to the user's query, that are completely relevant and there is another group that is far from this relevance, according to this model, the related group of documents is called the ideal answer set, and by providing a complete description of this set of documents (ideal answer set). The problems of retrieving the content of documents diminish, and yet another obstacle appears in the difficulty of definitively knowing what these characteristics and features are.

The primary effort of a probabilistic model is the initial guessing to determine the characteristics of the keywords contained in the documents, which have linguistic and idiomatic connotations, which contribute to the marking of these characteristics, allowing the creation of a probabilistic

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 09, September-2021

initial description of the ideal response set to the documents query.

In spite of the results achieved by this model, some saw that it isn't better than the results achieved by the classical models, which led, from their viewpoint, to the system developers not convinced of switching to this model largely.

Among the probabilistic models, we mention:

1) Best Match BM25:

This model was developed in the 1970s and 1980s by Stephen E. Robertson, Karen Spärck Jones, and others. It is a currently widely used probability model, in which documents are classified based on the estimated probability of the documents fit in the query. [6]

This model was developed from the Binary Independence Model (BIM), which is a classic probabilistic model that represents both documents and queries as binary vectors by combining range frequency and Document Length Normalization. [6]

This model is usually referred to as the Okabi BM25 because the Okabi retrieval system implemented in London in the 1980s was the first to implement this model.

The BM25 model has a lot in common with the TF-IDF terms weighting algorithm. Both algorithms use the term frequency and the inverse document frequency, but the definition of factors differs slightly between the two models. [13]

Both models define the weight that is given to each term as a result of combining the inverse document frequency and the term's frequency and then calculating the term's weight for the entire document for the specified query. The most prominent difference between the two models is that the BM25 takes into account the length of the document while it has no effect in the traditional TF- IDF used in the classic form. [13]

2) Language Models:

Language models applied to information retrieval by a number of researchers in the late 1990s, among them we mention: (Ponte and Croft 1998, Hiemstra and Kraaij 1998, Miller etal. 1999). [3]

Language models developed in the early 1980s for automatic speech recognition systems. It studies the probability distribution over all words sequence in a language. [3] [6]

The language model estimates the probability of words sequences. There is a language model associated with each document and this model may contain the queries most relevant to it. Language model-based methods are a widely used model. [6]

Once we know the probabilities of individual words, i.e. assign a probability value to each term, we can calculate the probabilities for any phrase or sentence in the language. The higher the probability of the sentence, the more likely that the sentence will be of interest. For example: Assume that the probabilities associated with words (information, retrieval, forms) are (0.1, 0.15, 0.05) respectively, then the probability of the phrase: Information Retrieval Models is 0.3. [6]

Some other probability models may specify a very high probability of the word "retrieval", indicating that the probability of this message that we are writing, for example, will be a strong candidate for retrieval if the query contains this word. [5] [3]

The language models take the same starting point, as the probability indexing model, that was set by (Bill Maron and Larry Kuhns), is a probability value that is assigned to the different indexing terms, which the document contains, so that each document has a set of indexing terms and each term has a probability value, that determines its importance for the query Including the terms it contains. Then, the language model for each document will be designed to follow this approach. [5] [3]

One of the advanced models that emerged using this technology is the Natural Language Processing Model.

As it does not rely on terms of query and document only, but it processes sentences and formulas, and this model works on matching them. Therefore, it requires building systems that work on natural language texts on three levels of processing: syntactic analysis, semantic analysis, Pragmatic analysis.

C. Combining Evidence:

In these models the technology for understanding the content is used for the documents and queries, and then it used also for concluding the probable relations between documents and queries. Therefore, the information retrieval process is an inference process or logic thinking in which we can evaluate the probability of the extent of the documents' fitness with inquiries which determines the user' need.

One of these models is:

1) Inference Network:

In this model, the documents retrieval is modeled as a process of logical inference, and its probability was evaluated to meet with the user' need for information in which we can express it by one or more of queries by analyzing the document as inference network will be the mechanism of concluding these relations kinds [14].

Most of the techniques used by information retrieval systems could be applied within this model. [7]

In the simplest application of this model, the document gives the term a certain power, and then the values for the terms contained within the query are combined to calculate the numerical result of the query in relation to the documents.

In other words, the power given to a term may be considered as the weight of the term in the document. Thus, the classification of documents in this model becomes similar to the arrangement in the vector space model or the probabilistic models. The strength of the term is indefinite and therefore any algorithm or form of term strength can be used within the document or query. [7]

This model relies on three basic things:

- Support the use of multiple document representation schemes.
- Allow the merging of results from different types of queries, which retrieve different documents for the same specific need for information.
- Flexible matching between terms or concepts mentioned in the queries and those specified for documents. This is done by improving recall by using cognitive matching of query concepts and documents and their representations without significantly degrading accuracy. [14]

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 09, September-2021

	Relevant	Non Relevant
Retrieved	*True Positive TP*	*False Positive FP*
Not Retrieved	*False Negative FN*	*True Negative TN*

2) Learning To Rank:

The learning to rank algorithm is part of the information retrieval for large documents. Data consist of queries and documents which are represented as vectors. [6]

It is divided into three models (Pointwise, Pairwise, and List wise). In the first model, pointwise, the arrangement is done as a traditional classification process, so the result is Class, so the goal is to reduce misclassification of queries and documents. In the second model, Pairwise, which is the process of converting the arrangement to a pointwise classification process, the goal of this process is to increase the number of pairs that were classified out of order, and the third, list wise, is very similar to the pair's wise model except that it deals with lists of rows and classes. [6] [15]

This form is applied to the Training Set test, and the documents are sorted according to their relevance and importance. [6] [15]

IV. Evaluating in IR Systems:

Retrieval effectiveness is a measure of how well the documents retrieved by a system meet users' needs. The process of determining the retrieval effectiveness for a given query is referred to as effectiveness evaluation [6].

One approach to measuring effectiveness of an IR system which is widely used is precision and recall. [6] [16].

However, precision and recall are inversely related. That mean, obtaining higher levels of precision can be obtained through lower recall values [6] [16].

Effective information retrieval systems must retrieve the largest possible number of relevant documents (i.e. have a high recall), and must retrieve a very small number of irrelevant (i.e. high-precision) documents. Unfortunately, experience has shown that these two aims are completely opposed. [7] Therefore, in some evaluating systems, the F1 factor is used as a measure of combining precision and recall. [16]

Below we explain the calculation methods for: Recall, Precision, and F1.

$$Recall = \frac{TP}{TP + FN} \quad , \quad Precision = \frac{TP}{TP + FP}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FN + FP}$$

V. Comparison of Main Models:

By reviewing the previous models, we find that the basic models, which the rest of the other models depend on, are three: Boolean model, the vector space model, the probabilistic model and all other models, which are attempts to develop these three basic models or combine them.

We have found that the Boolean retrieval model allows users to formulate complicated logical expressions and queries that may be difficult for an ordinary user, and this model does not provide an arrangement for the retrieved documents. This set of documents will either be nearly empty (low recall) or contain too many documents (low precision) due to the use of an exact match criterion. Consequently, this model is more useful for data recovery than information retrieval because all terms in it are of equal weight. [17]

While we found in the VSM the possibility of applying a set of values to each term, either in the representation of documents or in the user's query. Common terms are not important in this model due to the application of inverse document frequency as importance is given to rare terms.

In the VSM model, a long document that can contain the same query terms only in the title and the abstract may be very relevant to the query, but in this model it will be of low importance compared to a short document that contains the same terms in the appendix, which is considered one of the drawbacks of this model. Another disadvantage of representing VSM documents is that the terms arrangement is missing and documents containing query terms close to each other cannot be preferred over documents that contain separate terms in different parts of the document. [17]

The probabilistic retrieval model relies on assumptions that have explicitly made, such as assuming that 50% of the document containing the term is closely related to that term. However, not all assumptions correspond to reality. Therefore, the total number of relevant documents must be estimated, and the calculation of the probability value P (…) which is a constant is not always correct. Therefore, the probabilistic retrieval model to achieve accurate results requires that the terms will be independent. It ignores the calculation of weight to repeat the term and position within documents, and thus it is more suitable for long documents than for short documents. [17]

The following table summarizes the most important differences between the three main retrieval models:

Tab.1:

	Positives	Negatives
Boolean Model	– *Simple and uncomplicated form and thus easy to apply and investigate.* – *Predictable and easy to explain.* – *Experts feel more in control of the system.*	– *The vocabularies of indexes are the same, as the vocabularies of inquiries, it uses the complete matching.* – *There is not possible to apply the partial matching.* – *The retrieved documents are not arranged or classified.* – *There is no weighing of the inquiries and index ' terms* – *Difficulty in constructing logic inquiries if they are long.*
Vector Space Model	– *The simplest model based on linear algebra* – *It depends on the weighting of terms.* – *It is based on the calculation of the degree of similarity between inquiries and documents.* – *Partial matching is possible.*	– *The terms were assumed statistically independent in theory.* – *Long documents are poorly represented and thus have limited expressive ability.* – *The keywords must be completely identical to the document terms.* – *It lacks the linguistic structure to represent important linguistic features.*

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 09, September-2021

Probabilistic Model	– *Effective model.* – *Mathematical & theoretical model.* – *Suitable for long documents.*	– *Probabilities are difficult to estimate.* – *Unrealistic assumptions due to independence of the term.* – *Logical relationships are neglected.* – *There are many models, and thus it is difficult to determine the best one, because it requires prior knowledge.*

VI. CONCLUSION:

After reviewing the different models of information retrieval systems, we found that the vector space model considered a flexible and clear at the same time, as it represents one of the most widespread models to date, and whose results depend largely on the process of term weighing, but it has the following two main problems: independence of terms and weighting of terms. [6]Consequently, working to overcome these points enables us to find a sophisticated information retrieval mechanism capable of obtaining better results than those achieved by Boolean models without entering into the complexities of the calculations of the probabilistic model.

We suggest working to increase the effectiveness of terms weighting process by defining descriptors of terms in documents in order to overcome the weaknesses of the vector space model. So that those descriptors give quantitative or qualitative indicators that determine the value of the information in them, and its importance to the document in an objective manner through the analysis of the linguistic structure of the terms and then, a value representing the degree of membership of the term in the document based on these descriptors, which leads to more accurate results and thus obtaining an effective information retrieval system. That is not restricted by exact match and simple

as the case of Boolean retrieval systems and depends in its operations on a thoughtful representation of the terms of texts, and not assumptions that may not correspond to reality as the case of the Probabilistic model, which ignores some important descriptors of the terms.

REFERENCES

[1] M. Sanderson and B. Croft, "The History of Information Retrieval Research," *IEEE*, vol. 100, no. Special Centennial Issue, pp. 1444 - 1451, 2012.

[2] C. Mooers, "Zatocoding applied to mechanical organization of knowledge," *Journal of the Association for Information Science and Technology*, vol. 2, no. 1, pp. 20-32, 1951.

[3] A. Göker and J. Davies, "Information Retrieval Models," *Wiley*, pp. 1-19, 2009.

[4] T. Gondaliya and H. Joshi, "Journey of Information Retrieval to Information Retrieval Tools-IR&IRT A Review," in *CALIBER-2017*, CHENNAI, 2017.

[5] D. Mabrouk, S. Rady , N. Badr and . M. , "Modeling using Term Dependencies and Term Weighting in Information Retrieval Systems," *Eighth International Conference on Intelligent Computing and Information Systems (ICICIS)*, vol. 42, pp. 321-328, 2018.

[6] V. Gudivada, D. L.Rao and A. R.Gudivada, "Information Retrieval: Concepts, Models, and Systems," in *Handbook of Statistics*, vol. 38, United States, 2018, pp. 331-401.

[7] A. Singhal, "Modern Information Retrieval: ABrief Overview," *IEEE*, 2001.

[8] G. Salton and M. J. McGill, Introduction to Modern Information Retrieval, 1983: McGraw-Hill Book Company, New York.

[9] H. Luhn, "A statistical approach to mechanized encoding and searching of literary information," *IBM Journal of Research and Development*, vol. 1, pp. 309-3017, 1957.

[10] M. and K. , "Probabilistic Indexing and Information Retrieval," *Journal of the ACM (JACM)*, vol. 7, pp. 216-244, 1960.

[11] E. Gaussier and F. Yvon, Textual Information Access: Statistical Models, WILEY, 2012.

[12] K. Jones and S. Robertson, "Relevance weighting of search terms," *Journal of the American Society for Information Sciences*, vol. 27, no. 3, pp. 129-146, 1976.

[13] Mastering Elasticsearch 5.x, Third ed., UK: Packt Publishing Ltd, 2017.

[14] H. Turtle and W. B. Croft, "Inference networks for document retrieval," *ACM SIGIR Forum*, vol. 51, no. 2, 2017.

[15] T.-Y. Liu, "Learning to Rank for Information Retrieval," *Foundations and Trends in Information Retrieval*, vol. 3, no. 3, p. 225–331, 2009.

[16] T. Sabbah, A. Selamat, M. H. Selamat, F. S. Al-Anzi, E. H. Viedma and O. Krejcar, "Modified Frequency-Based Term Weighting Schemes for Text Classification," *Applied Soft Computin*, vol. 58, pp. 193-206, 2017.

[17] M. Pannu, A. James and R. Bird , "Comparison of Information Retrieval Models," in *The 19th Western Canadian Conference on Computing Education - In-Cooperation with ACM SIGCSE*, 2014.

ABOUT AUTHORS

Manal SHEIKH OGHLI is an Information Technology IT graduated, she is a student in Web Science master's program at Syrian Virtual University (SVU). Her main research topics are the web science.

Muhammad Mazen ALMUSTAFA is an assistance professor at the international university for science & Technology (IUST). He works also as a part-time at Syrian Virtual University (SVU). He has two masters and two PhD degrees. The first PhD degree in Computer Information Systems (CIS). The second PhD Degree in Computers and their networks. His main research topics are the web science. He has published a number of papers on this topic in conferences and journals.

الورقة العلمية الثانية التي نشرت في مجلة IJERT:

International Journal of
Engineering Research & Technology
ISSN : 2278 - 0181, www.ijert.org
(Published by : ESRSA Publications)

CERTIFICATE
OF PUBLICATION

This is to certify that

Manal Sheikh Oghli

Has published a research paper entitled

Developing A New Term Weighting Schema Through Text-Document Analysis and Natural Language Processing NLP

In IJERT, Volume 10, Issue 12 , December - 2021

Registration No: IJERTV10IS120185 **Date:** 29-12-2021 **Chief Editor,IJERT**

International Journal of
Engineering Research & Technology
ISSN : 2278 - 0181, www.ijert.org
(Published by : ESRSA Publications)

CERTIFICATE
OF PUBLICATION

This is to certify that

Muhammad Mazen Almustafa

Has published a research paper entitled

Developing A New Term Weighting Schema Through Text-Document Analysis and Natural Language Processing NLP

In IJERT, Volume 10, Issue 12 , December - 2021

Registration No: IJERTV10IS120185 **Date:** 29-12-2021 **Chief Editor,IJERT**

Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 12, December-2021

Developing A New Term Weighting Schema Through Text-Document Analysis and Natural Language Processing NLP

Manal Sheikh Oghli
Web Science program, Syrian Virtual University
Damascus, Syria

Muhammad Mazen Almustafa
Web Science program, Syrian Virtual University
Damascus, Syria

***Abstract*—Term weighting is one of the branches concerned with information retrieval IR. It studies the importance of "word" or "phrase" in a certain text, as the issue of determining the importance of keywords is essential and effective in the modern retrieval systems.**

Studies showed that the appropriate weight of the term importance affects the retrieval results. Thus, the distribution, location, indication, and synchronization of the term with other terms in the document are factors that should be taken into consideration upon measuring the similarity between documents or between query and document.

The paper sheds light on some weak points in the traditional term weighting (TF-IDF) of the vector space model. It also reviews some algorithms to improve TF-IDF performance, then develop a new mechanism for term weighting depending on text analysis and natural language processing through features of the terms deduced from the information derived from text analysis and processing, and conducting the appropriate mathematical tests to reach a new way for term weighting. This new way enhances the ability of the suggested system to retrieve the most appropriate information requested by the user; which is the most essential goal all retrieval systems are seeking to achieve.

Keywords*— *Information Retrieval IR, Vector Space Model VSM, Indexing, Term Weighting, TF-IDF, Natural Language Processing NLP)

I. INTRODUCTION

The effective research systems do not work directly with documents or queries as different techniques and strategies are used to represent the essential meaning in the form of parts of a document or inquiries; a process that is called indexing [1].

Terms in the vector space model are represented as a vector getting out of a set of concepts, where the vector represents the keywords and terms extracted from documents. The biggest challenge facing this model; however, is determining the suitable value of the vector constituents or what is known as Term Weighting, in addition to terms independence [2] [3].

According to the VSM model, the long document that could contain the same terms appearing in the query - only in the title and abstract - can be of a great relevance to query, but in this model, it will have less significance in comparison with a short document having the same terms in the footing. A flaw in the VSM document representation appears in that the term arrangement is missing. Documents that do not have close query term cannot be preferred to documents having separate terms in various parts of the document [3].

Processes in the vector space model are three stages: The first stage is "Indexing", where terms are deducted from the text. The second stage is "Term Weighting", and the third is "Classification" as regards query and similarity [4].

This paper suggests a new mechanism for term weighting with the aim of overcoming the shortcomings of the vector space model through identifying the term features of document terms where features give quantitative or qualitative indicators that determine the value of information, significance to document, and the relation between terms and their occurrence within the document and their grammatical position that increases the effectiveness of this model.

II. INDEXING IN INFORMATION RETRIEVAL SYSTEMS

This process implies determining the keywords that represent a document on the basis of its contents. It is a significant stage in the retrieval system. Indexing is defined as "a process that determines keywords or descriptive terms, what is called "Index Terms" that represent the document on basis of its contents to reach an effective access of documents." [5].

Text processing and analysis is the first step of indexing in retrieval systems to get a more effective retrieval. To achieve this, there should be an appropriate structure for indexing. The most used data structure is the Inverted index which is a term-oriented mechanism which is the most competent and flexible index structures [6].

The structure of the Inverted Index has two components; vocabulary and document list. Vocabulary is a set of various terms concluded from documents. Each document is represented by a list of some referential words stored alphabetically [6] [7]. It should be noted that some statistical information could be stored about each term in each document, like term frequency, term position, and other useful features for the retrieval process.

To establish the "Index", text should undergo analysis and processing within information retrieval systems; such as:

A. Linguistic Analysis (Tokens Extraction):

The most important question that should be raised at this stage is the following: What are the right tokens that should the system process and store? [8].

At this stage, a text is analyzed, distinctive terms chosen, and unnecessary symbols and punctuation marks removed. This process is usually referred to as "Tokenization", where the document is divided into units called "Tokens". This results in a set of words of semantic significance [9].

Here, you have to differentiate between "Token" and "Term". "Token" means a distinctive symbol. It is a representation of a series of letters in a certain document

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 12, December-2021

combined to form a good-to-process semantic unit, whereas "Term" is a token processed to be inserted in the IR Index [6] [8].

B. Stemming & Lemmatization:

This process is also called in general "Abstraction". Usually, the terms of abstraction, stemming and lemmatization refer to the change of the word structure and reduction of term form to a common form.

"Stemming" refers to extracting part of the end of the word and removing any suffixes; i.e., removing any additions to the word, whereas "Lemmatization" depends on the morphological analysis of words to remove the inflectional suffixes only and restore the basic form of the word, as stated in the "Linguistic Dictionary", which is known as "root", or "Lemma" [8] [9].

The aim of abstraction or stemming is reducing the different forms of the word generated due to inflection, and sometimes the derivative forms that are related to the word to a common form [6] [8]. Hence, when a user specifies a term to search for, it is necessary to retrieve all documents that have grammatical variables of the term, which prevents any exact match between the query term and the document containing this term.

Through this process, all grammatical forms of terms are represented in a basic common form. This also helps reduce the size of documents and makes search faster through searching for the abstract term instead of searching for the whole term.

The first stemmer of the English language was developed in 1968 by Julie Beth Lovins who introduced the concept of "abstraction" based on the "Dictionary of Common Suffixes". This logarithm was based on the principle of removing suffixing with view to the longest match. This logarithm led to reasonable results in the field of information retrieval [10].

Then came Martin Porter who published a paper in 1980 in the Programme Journal to describe a very simple logarithm, in concept. This logarithm is controlled by certain rules that determine whether the suffix could be omitted or not depending on the minimum left after omission. The logarithm; however, repeatedly proved to be empirically effective [8]. Porter did not depend on an abstraction dictionary; rather, he used lists of suffixes, then linked each suffix with a special criterion to delete the suffix from the word to get a true abstract work when applying the criterion. The logarithm consists of several stages; each of which contains a number of rules to remove suffixes and it is available in different languages except Arabic [6].

The Paice/Husk stemmer was published for the first time in 1990. It was developed by Chris Paice with the help of Gareth Husk, who used a table of indexed rules which determine whether suffixes would be omitted or replaced [10].

C. Removing Stop Words

Stop words is a list of linguistically common words and have a limited effect on the categorization and selection of documents that goes appropriately with user needs.

They are functional words that have no meaning, and are part of how nouns in a text are described and expressed [6]. Like pronouns, connectors and prepositions, that appear in all text documents [9].

Rarely do these words refer to anything related to the subject of the document; hence, such functional words will not be of help in the search processes [6].

The general strategy used to define this list of words depends on sorting terms by aggregation frequency (total number of the times each term appears in a certain set of documents). Then, the most common terms are taken, and filtered, often manually, due to their poor semantic content in relation to the documents being indexed [6].

D. Term Weighting:

Term Weighting is defined as a digital computing aiming to express the importance of a word within a group. It is usually used as a weighing factor in the search processes within information retrieval systems [11].

It is a calculation process and determining a digital value for each term to consider its contribution in distinguishing a certain document. Terms are descriptors of content in the documents used in indexing, and through which the relatedness of documents to queries is evaluated. These terms are classified as objective and non-objective, where the weighting process is applied to the non-objective terms which reflect the contents of a document. Then, these terms are weighted and their significance in relation to the information included in the document is demonstrated [12].

III. TF-IDF ALGORITHM

TF-IDF is an abbreviation of Term Frequency–Inverse Document Frequency, which is a technique commonly used in text mining and information retrieval. The Term Frequency, which was one of the most important developments in the field of information retrieval, was defined, for the first time, by Scientist Gerard Salton in the nineteen seventies [13].

This was supplemented by the work accomplished by Spark Jones who presented her paper about the Inverse Document Frequency (IDF) [14], and resulted in the quick adoption of the two-method combination; i.e., (TF) and (IDF), as two new methods for term weighting [3] [15].

The term frequency (TF) is also called "Local Term Weigh" and is defined as the number of recurrences of the term being searched within a document.

$$TF(t) = \frac{\text{Number of times term t appears in a document}}{\text{Total number of terms in the document}}$$

The IDF, on the other hand, is called "Global Term Weight" and reflects the term frequency within a number of documents.

$$IDF(t) = \log_e \frac{\text{Total number of documents}}{\text{Number of documents with term t in it}}$$

The term weighting could be calculated by use of TF and IDF through the following equation:

$$\boldsymbol{TF - IDF_{t,d} = TF_{t,d} * IDF_t}$$

IV. LIMITATIONS OF TF-IDF:

Despite the numerous features of the traditional way TF-IDF, there are a lot of shortcomings which cannot be ignored. Those shortcomings are like the following:

1) The traditional method assumes that calculating the term frequency gives an independent proof of similarity, which is not always correct [16].
2) This method calculates term weighting on basis of the term frequency, and is not concerned with the term position in the text [17] [18].
3) The traditional TF-IDF is a technique used to select an unsupervised feature as it is only limited only to the document [16]. Nonetheless, it does not discuss the

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 12, December-2021

common connotation or occurrence with other terms in the document. It is not, either, concerned with the relation between words and the importance of the term in relation to the text itself [17] [18].

Therefore, this domain was and is still a motive for a lot of researchers to study the possibility of producing different forms of weighting plans and improve TF-IDF algorithm with the aim of developing information retrieval systems.

Here are three examples:

- In 2017, a number of scholars; namely, Thabit Sabbah, Ali Selamat, Md Hafiz Selamat, Fawaz S.AlAnzi, Enrique Herrera Viedma, Ondrej Krejcar, Hamido Fujita, proposed four schemes for term weighting depending on the TF-IDF scheme. They were: mTF, mTFIDF, TFmIDF, and mTFmIDF. The schemes take into consideration the missing term counting with calculating the weights of current terms to improve the performance of Text Classification (TC).
  The first proposal examines the number of missing terms in a text in comparison to the total number of terms in a certain group to present a new weighting called "mTF". The second proposal; however, examines the percentage of the number of missing term documents to the number of texts in a group, which researchers called "mIDF".
  Based on the previous two proposals, scholars presented different standard weighting schemes of the TF-IDF, on the basis of the proposed mIDF and mTF schemes [19].
  In this study, scholars depended on the idea that some terms should disappear when other terms already exist in the text, and vice versa. Therefore, the weights of terms would, definitely, be affected.
- In 2018, Rajendra Kumar Roul, Jajati Keshari Sahoo, and Kushagr Arora from Zuarinagar University in India, conducted a study that demonstrated the shortcomings of the TF-IDF, then proposed four different techniques for term weighting with view to overcoming those shortcomings by modifying the traditional TF-IDF. The study showed that the text representation based on language, like the BSM Vector Space Model, greatly affect, especially in the fields where language is processed naturally (NLP), the information retrieval (IR), and that the transfer of texts to vector renders the possibility of conducting any mathematical operation on vector-represented texts. Therefore, term weighting plays a big role in representing documents more accurately.
  The four techniques proposed by the researchers in this paper entail a modification of the known TF-IDF algorithm by the addition of some mathematical operations to documents such as Inter-class dispersion, where the term is distributed in a unified way among the different classes, which means term dispersion will be low., therefore, the weight the term will contribute in will also be low. If; however, the term has a big disparity, this means it is good and the weight it is contributing in will be high. The second proposal was about modification of the traditional IDF through giving the value 0 or 1 depending on the frequency of the term in all documents or non-occurrence in any document. The third proposal took into consideration the number of documents containing that term and belong to a definite category, and the total number of documents in this category. The last proposal examined the importance of the document length in the term weighting [16].
- In 2019, Shuzhi Sam Ge and Ting ZHANG from the University of Electronic Sciences and Technology in China suggested a new term weighting way. They called it "TF-IDF-ρ"
  Their study showed that the traditional algorithm TF-IDF is one of the term weighting algorithms and the most common text representation way [18]. Nevertheless, it does not function properly, and has a lot of shortcomings.
  Throughout the study, they worked on improving this algorithm by suggesting a new idea that depends on the "Class Discriminative Strength", which the term affects. They suggesting benefiting from this to improve the traditional TF-IDF algorithm.
  The study also showed that the selection and weight calculation of the distinctive terms of texts defines to a great extent whether the text has been properly classified or not.
  Researchers suggested that the class discriminative strength ρ represents the discriminatory power of a feature item, equal to a total number of the category in corpus divided by the number of classes of feature item occurrence in.
  This way suggested by researchers represents in assigning a greater weight for distinctive terms which appear in greater proportions as a strength that distinguishes classification or category, with the aim of shedding light on the ability of this term to distinguish different texts [18].

V. NATURAL LANGUAGE PROCESSING (NLP):

The need to analyze texts before retrieval is one of the biggest obstacles facing Information Retrieval Systems (IRS) as the latter mainly depend on understanding the content of texts to be retrieved, and analyzing the words used to build queries, then making a link between keywords and the text database, and conducting the appropriate weighting process to reach the proper text. This led to the arising need of a textual analysis process.

Natural Language Processing (NLP) is a way to analyze texts in a computer. It includes collecting knowledge about how humans understand and use language, to develop the appropriate tools and techniques that make computer systems capable of understanding and processing natural languages to perform the various required tasks [20].

Dr. Michael J.Garbade defines Natural Language Processing (NLP) as a branch of Artificial Intelligence (AI) that handles the interaction between computers and humans through the use of natural language. It shows that the ultimate goal of Natural Language Processing is reading, understanding and realizing the human languages in a valuable way and inferring the required meaning from them [21].

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 12, December-2021

VI. STANFORD CORENLP

The Stanford CoreNLP is one of the widely used tools as most of the common essential natural language processing are available, such as tokenization and even coreference resolution through combining several constituents of natural language analysis [22].

One of the initial release goals that was developed in 2006 was obtaining Annotators Pipeline swiftly, and providing a light framework through the use of Java objects and applying them on any text instead of applying them on a single sentence only. In 2009, the system was developed to be used more easily and by a wider range of users, as the system provided the interface of command line and the ability to write outside annotations with different formats, including XML.

Control of annotations could be through the Object Properties, and the Stanford CoreNLP set could be packaged in a way that makes them easy to access by various languages, such as: Python, Ruby, Perl, Scala, JavaScript, Net and even C# [22].

The current release entails a set of processing tools designed to take initial textual inputs, giving whole textual analysis outputs, and linguistic annotations appropriate for the effective textual analysis [23].

The most important tasks for natural language processing which the CoreNLP set of tools carries out are: Tokenization, Lemma, Named Entity Recognition, where names are recognized as one of the shapes (person, location, organization..), numbers (set ، duration ،time ،date ،number ،money), and classification of symbols according to the Part of Speech they belong to and have been symbolized through a set of Tags [23].

In spite of some notes on the analytical tasks of these tools, it could be said that it is a set of easy-to-understand tools that could be used as a constituent within a bigger and scalable system [22].

Here is a review of the system structure through which Stanford CoreNLP analyzes and process texts:

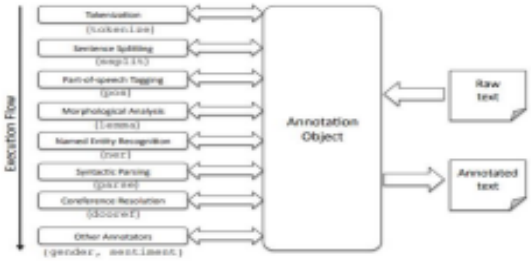


Figure 1: System structure of Stanford CoreNLP [22]

VII. A PROPOSED METHODOLOGY FOR TEXT PROCESSING AND INDEXING:

The CoreNLP set of tools was benefited from in the text analysis and extracting features of each document and the most significant terms through a specific text analysis methodology that used the following:

1) Analyzing the text through the Stanford CoreNLP set of tools and dividing it into a group of tokens.
2) Classifying tokens through POS analysis into groups correlated to the grammatical position of token, then sorting the tokens into four main groups: (Functional tokens, verbs, nouns, and adjectives)
3) Omission of functional tokens as they perform a functional task in the text and do not play a role in defining the subject of the document.
4) Applying Lemma on (verbs, adjectives and nouns).
5) Automatic filtering of some non-functional tokens. Some tokens, considered Stop Words, have been defined in a proposed system. Those are some verbs, nouns, and adjectives commonly used but do not participate in defining the subject of the document. It should be noted that the number of tokens in this list are about 200, such as: (do, like, good, great ...).
6) Indexing the tokens resulting from previous processes like terms within a proposed database system, and some features like (POS, NER, and Order).

VIII. PROPOSED METHODOLOGY IN TERM WEIGHTING:

The common method used in the fields of natural language processing (NLP) is:

1) Looking for features in the document.
2) Defining the significance of these features.
3) Sending the weighted features for the sake of taking the right decision [24].

As long as the term frequency is not considered the only dimension that information retrieval systems (IRS) rely on in determining the relatedness of queries and documents. Therefore, we suggest the following two parameters:

A. Addition of the POS Parameter:

This coefficient defines the lexical matching between the term in the query and the term in the document. For example: The term (book) has two different meanings in the following statements:

- Book a study seat in the Syrian Virtual University to develop your scientific level.
- The Syrian Virtual University website has many important digital books.

The term (book) will take a unified shape after the abstraction process in both texts. Therefore, it is necessary to distinguish between both terms, when it appears as a verb and as a noun, and determine how much it conforms to the user desire.

Therefore, the study suggests classify the POS results into a classes (nouns, verbs, and adjectives) which are identical to their classification within the indexing process.

Then, values are given to terms common to query and text according to the following table:

POS=POS NER =NER	POS=POS One NER	POS=POS Not NER	POS ≠ POS Same Class	POS ≠ POS Not Same Class
1	0.8	0.5	0.3	0.1

Thus, the proposed equation for the POS calculation of the text will be like this:

$$POS = \frac{\sum_{i=0}^{n} \textbf{POS Value}}{\textbf{Total Count Of Term } (N)}$$

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 12, December-2021

B. Addition of the Correlation Parameter:

This parameter examines how much words are correlated through studying the position of words and the calculating distances between them in the text. The correlation parameter could be calculated through the following equation:

$$Corr = \frac{Common\ Terms\ (n)^2}{[\sum_{k=0}^{n-1} dis(term) + 1] * Count\ Term\ Of\ Query}$$

Where (n) Common Terms is the number of common terms between query and text, and (dis) is the distance between term (i) and term (i+1); i.e.:

$$Dis = Order\ Term_{i+1} - Order\ Term_i$$

as long as terms will be distributed within the text with a possibility of frequency, then calculating the less value of distance between terms, and calculating correlation depending on these values.

Thus, we notice that the position of terms gives an added value through which we could make a differentiation process between documents based on their value.

C. Algorithm of the Proposed Term Weighting NLP-TF-IDF:

The proposed way is a supportive way of the traditional term weighting algorithm, through some features indexed by the use of CoreNLP tools, therefore, the weighting will be calculated through the following equation:

$$NLP\ Similarity = Cosine\ Sim_{TF-IDF} * \frac{[1 + (10 * POS) + Corr]}{2}$$

IX. Evaluation In the Information Retrieval Systems:

The effectiveness in information retrieval systems is a measure of retrieved documents by the system to meet the needs of users. The process of identifying the effectiveness of retrieval of a certain inquiry is referred to as "Effectiveness Evaluation" [12].

The measures of effectiveness of retrieval systems are precision (the percentage of relevant documents as for the retrieved group), and recall (the percentage of relevant documents in the retrieved group as for all documents) [12] [19].

Some evaluation systems used the F1 measure as a measure combining precision and recall [3] [19].

Here is how the F1, precision and recall are calculated:

$$Recall = \frac{\text{TP}\ (Retrieved\ \&\ Relevant)}{\text{TP}\ +\ \text{FN}\ (Not\ Retrieved\ \&\ Relevant)}$$

$$Precision = \frac{\text{TP}\ (Retrieved\ \&\ Relevant)}{\text{TP} + \text{FP}\ (Retrieved\ \&\ Non\_Relevant)}$$

$$F1\ = 2\ \times \frac{Precision \times Recall}{Precision + Recall} = \frac{2 * \text{TP}}{2 * \text{TP} + \text{FN} + \text{FP}}$$

X. Experimental Results:

The CISI was chosen to be a normative data set. It is a number of scientific articles; 1460 articles published between 1969 and 1977. They included the author's name, title of article and abstract. The group was provided with a query group and expert results for each query [25].

A partial set of this data was used. This set had 300 texts to be tested. The first 30 queries were chosen to be tested. As long as the selected texts to be a pilot group were 300 texts, we found out that 3 queries had zero results within the selected documents. Therefore, the results were not shown within the test results. Our results were compared to the traditional TF-IDF algorithm and the Porter Index.

A. First Experimental Stage:

At this stage, the order of retrieved documents is tested by the use of Cosine similarity of the traditional Term Weighting algorithm according to the proposed system methodology by use of the Porter Stemmer, where a definite number of documents is retrieved; i.e., fixing the value of TP + FP = 25. The traditional algorithm of Term Weighting TF-IDF is not modified at this stage, but it is used and the similarity is calculated by the use of the Cosine Similarity.

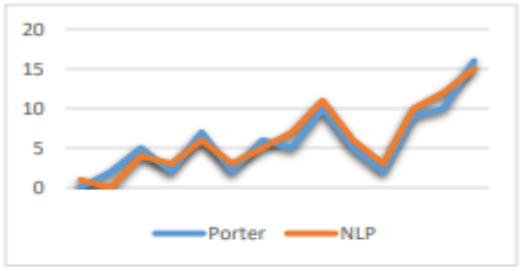
Here is a comparison of the evaluation factors of the three information retrieval systems (Recall, Precision and F1):

Query	Porter Index			NLP Index			$F1_{NLP}$ - $F1_{Porter}$
	Recall	Prec	F1	Recall	Prec	F1	
1	0.5882	0.4	0.4761	0.6470	0.44	0.5238	0.0476
2	0.4	0.08	0.1333	0	0	0	-0.1333
3	0.7	0.28	0.4	0.6	0.24	0.3428	-0.0571
5	0.2	0.08	0.1142	0.3	0.12	0.1714	0.0571
8	0	0	0	0	0	0	0
9	0.6	0.12	0.2	0.6	0.12	0.2	0
10	0.8333	0.2	0.3225	0.6666	0.16	0.2580	-0.0645
11	0.1935	0.24	0.2142	0.1935	0.24	0.2142	0
12	0	0	0	0.3333	0.04	0.0714	0.0714
13	0.5	0.64	0.5614	0.4687	0.6	0.5263	-0.0351
14	1	0.04	0.0769	1	0.04	0.0769	0
15	0.4090	0.36	0.3829	0.4090	0.36	0.3829	0
16	0.5	0.08	0.1379	0.5	0.08	0.1379	0
17	0.25	0.04	0.0689	0.25	0.04	0.0689	0
18	0.6666	0.08	0.1428	0.6666	0.08	0.1428	0
19	0.5	0.24	0.3243	0.5	0.24	0.3243	0
20	0.3125	0.2	0.2439	0.4375	0.28	0.3414	0.0975
21	0.2857	0.08	0.125	0.2857	0.08	0.125	0
22	0.0909	0.08	0.0851	0.1363	0.12	0.1276	0.0425
23	0.2608	0.24	0.25	0.2608	0.24	0.25	0
24	0.4615	0.24	0.3157	0.3846	0.2	0.2631	-0.0526
25	0.2857	0.08	0.125	0.4285	0.12	0.1875	0.0625
26	0.5833	0.28	0.3783	0.5833	0.28	0.3783	0
27	0.3225	0.4	0.3571	0.3871	0.48	0.4285	0.0714
28	0.5	0.16	0.2424	0.5	0.16	0.2424	0
29	0.2777	0.2	0.2325	0.3333	0.24	0.2790	0.0465
30	0.3913	0.36	0.375	0.4347	0.4	0.4166	0.0416

Table-1: Comparison of Evaluation Values between NLP and Porter Index

Here is a chart demonstrating the difference of the TP values between NLP, Porter Index:

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 12, December-2021



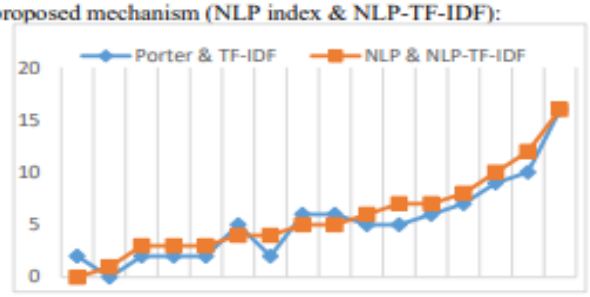
The results of this stage showed the following:

1) The values of the three evaluation parameters (Recall ، Precision ،F1) with the values of (0.72, 0.74, 0.75) increased by the use of the traditional weighting measure of the indexed documents in the proposed indexing system by use of Porter Stemmer.
2) The new documents by use of NLP Index were 25 different from the documents that appeared in the Porter Index, which indicates the indexing significance in the information retrieval systems and their huge impact on results.
3) Although the Porter Indexing outperformed 5 queries, the total results of the three-evaluation parameter indicated the high results achieved by the proposed indexing system. Thus, the indexing with the use of the NLP tool set outperformed 9 queries and the results were similar in 13 other queries.

B. Second Experimental Stage:

At this stage, the mechanism of calculating similarity between query and documents was modified, then the results of the traditional TF-IDF and the modified NLP-TF-IDF was compared. The following results were concluded:

Here is a chart of the differences of the TP values between the traditional method (Porter Index & TF-IDF) and the proposed mechanism (NLP index & NLP-TF-IDF):



Here are the values of the three-evaluation parameters:

Query	Porter Index & TF-IDF			NLP Index & NLP-TF-IDF		
	Recall	Prec	F1	Recall	Prec	F1
1	0.5882	0.4	0.4761	1	0.04	0.0769
2	0.4	0.08	0.1333	0	0	0
3	0.7	0.28	0.4	0.3333	0.04	0.0714
5	0.2	0.08	0.1142	0.6666	0.08	0.1428
8	0	0	0	0.25	0.04	0.0689
9	0.6	0.12	0.2	0.75	0.12	0.2068
10	0.8333	0.2	0.3225	0	0	0
11	0.1935	0.24	0.2142	0.6	0.12	0.2
12	0	0	0	0.6666	0.16	0.2580
13	0.5	0.64	0.5614	0.2857	0.08	0.125
14	1	0.04	0.0769	0.4285	0.12	0.1875
15	0.4090	0.36	0.3829	0.5	0.16	0.2424
16	0.5	0.08	0.1379	0.3	0.12	0.1714
17	0.25	0.04	0.0689	0.8	0.32	0.4571
18	0.6666	0.08	0.1428	0.5	0.24	0.3243
19	0.5	0.24	0.3243	0.5833	0.28	0.3783
20	0.3125	0.2	0.2439	0.3846	0.2	0.2631
21	0.2857	0.08	0.125	0.4375	0.28	0.3414
22	0.0909	0.08	0.0851	0.5882	0.4	0.4761
23	0.2608	0.24	0.25	0.3333	0.24	0.2790
24	0.4615	0.24	0.3157	0.1818	0.16	0.1702
25	0.2857	0.08	0.125	0.4090	0.36	0.3829
26	0.5833	0.28	0.3783	0.2173	0.2	0.2083
27	0.3225	0.4	0.3571	0.4347	0.4	0.4166
28	0.5	0.16	0.2424	0.2258	0.28	0.25
29	0.2777	0.2	0.2325	0.3870	0.48	0.4285
30	0.3913	0.36	0.375	0.5	0.64	0.5614

Table-2: Comparison of Evaluation Values Between Suggested System and Porter indexing & TF-IDF

The results of this stage experiments showed the following:

1) The number of queries in which the proposed system outperformed the Porter increased to be 11 queries, and the results were similar in 12 queries. However,
2) The number of queries in which the Porter Indexing excelled by applying the proposed weighting mechanism decreased to be 4 queries.
3) As a result, the number of true documents that used the proposed weighting method rose to be /139/ documents, whereas it was only /130/ documents with the use of Porter Indexing.
4) The new /27/ documents that appeared were different from those that appeared using the Porter Indexing with the traditional Term Weighting algorithm.
5) The test results of the second stage showed that the proposed mechanism of term weighting raised the medium value of the F1 parameter to be (0.7%) higher than the use of the traditional term weighting algorithm TF-IDF of indexed data using the NLP tool set. The value of F1 as a medium value raised at (1.4%) compared to similarity values of the TF-IDF of the indexed data using the Porter Stemmer. The maximum value of the increase of the F1 parameter value reached (9.7%).

XI. CONCLUSION:

Text analysis and natural language processing is a way to understand the user desire, therefore, the right text analysis leads to more accurate retrieval results that meet user needs.

The NLP tool set could achieve great results in this field; hence, those tools could be benefited from in raising the efficacy of the information retrieval systems through developing new mechanisms for term weighting that depend, in its content, on the distinctive features that could be elicited from the analysis of such tools to texts.

The test results also affirmed the importance of indexing and its effect on the retrieval results, on the one hand, and the possibility of benefiting from the features of the NLP tools in developing the traditional term weighting mechanism TF-IDF through adding parameters that contribute in defining the relevant documents to the user's desire, on the other hand.

It should be noted that despite the shortcomings of these tools in processing the suffixes of nouns and adjectives, the results demonstrated the superiority of the proposed system method as aesthetical values to the traditional method in all parameters of information retrieval parameters evaluation. Therefore, a restructuring of indexing depending on NLP tool set and studying the suffix processing issue will lead to more effective and more accurate results.

REFERENCES

[1] J. Savoy and E. Gaussier, "Information Retrieval," in *Handbook of Natural Language Processing*, Chapman and Hall/CRC, 2010, pp. 455-484.

[2] A. Göker and J. Davies, "Information Retrieval Models," *Wiley*, pp. 1-19, 2009.

[3] M. Sheikh Oghli and M. M. Almustafa, "Comparison of basic Information Retrieval Models," *International Journal Of Enginering Research & technology (IJERT)*, vol. 10, no. 09, pp. 299-303, 2020.

[4] S. JABRI, A. DAHBI, T. GADI and A. BASSIR, "Ranking of Text Documents using TF-IDF Weighting," in *2018 4th International Conference on Optimization and Applications (ICOA)*, Mohammedia, Morocco, 2018.

[5] E. Chauhan and D. Asthana, "Review of Indexing Techniques in Information Retrieval," *International Journal of Engineering Science and Computing IJESC* , vol. 7, no. 7, pp. 13940-13942, 2017.

[6] W. B. Crof, D. Metzler and T. Strohman, Search Engines, Pearson Education, 2015.

[7] B. Saini, V. Singh and S. Kumar, "Information Retrieval Models and Searching Methodologies: Survey," *International Journal of Advance Foundation and Research in Science & Engineering (IJAFRSE)*, vol. 1, no. 2, pp. 57-62, 2014.

[8] C. D.Manning, P. Raghavan and H. Schütze, An Introduction to Information Retrieval, New York: Cambridge University Press, 2008.

[9] V. Murthy, D. B. V. Vardhan, K. Sarangam and P. V. p. Reddy, "A Comparative Study on Term Weighting Methods For Automated Telugu Text Categorization with Effective Classifiers," *International Journal of Data Mining & Knowledge Management Process (IJDKP)*, vol. 3, no. 6, pp. 95-105, 2013.

[10] V. Gurusamy, S. Kannan and K. Nandhini, "Performance Analysis: Stemming Algorithm for the English Language," *IJSRD - International Journal for Scientific Research & Development*|, vol. 5, no. 05, pp. 1933-1938, 2017.

[11] T. Xia and Y. Chai, "An Improvement to TF-IDF: Term Distribution based Term Weight Algorithm," *JOURNAL OF SOFTWARE*, vol. 6, 2011.

[12] V. Gudivada, D. L.Rao and A. R.Gudivada, "Information Retrieval: Concepts, Models, and Systems," in *Handbook of Statistics*, United States, 2018, pp. 331-401.

[13] M. Sanderson and B. Croft, "The History of Information Retrieval Research," *IEEE*, vol. 100, no. Special Centennial Issue, 2012.

[14] S. Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of documentation*, vol. 28, no. 1, pp. 11-21, 1972.

[15] G. Salton and C. Yang, "on the Soecification of Term Value in Autoatic Indexing," *Cornell University*, pp. 73-173, 1973.

[16] R. K. Roul, J. K. Sahoo and K. Arora, "Modified TF-IDF Term Weighting Strategies for Text Categorization," in *14th IEEE India Council International Conference (INDICON)*, 2018.

[17] J. Ramos, "Using TF-IDF to Determine Word Relevance in Document Queries," January 2003.

[18] T. ZHANG and S. G. Sam , "An Improved TF-IDF Algorithm Based on Class Discriminative Strength for Text Categorization on Desensitized Data," in *ICIAI 2019: Proceedings of the 2019 3rd International Conference on Innovation in Artificial Intelligence*, Suzhou, China, 2019.

[19] T. Sabbah, A. Selamat, M. H. Selamat, F. S. Al-Anzi, E. H. Viedma and O. Krejcar, "Modified Frequency-Based Term Weighting Schemes for Text Classification," *Applied Soft Computin*, vol. 58, pp. 193-206, 2017.

[20] S. Joseph, H. Hlomani, K. Letsholo, F. Kaniwa and K. Sedimo, "Natural Language Processing: A Review," *International Journal of Research in Engineering and Applied Sciences*, vol. 6, no. 3, 2016.

[21] D. M. J.Garbade, "A Simple Introduction to Natural Language Processing," *Becoming Human: Artificial Intelligence Magazine*, 2018.

[22] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard and D. McClosky, "The Stanford CoreNLP Natural Language Processing Toolkit," in *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics*, Maryland USA, 2014.

[23] E. Nazaruka, J. Osis and V. Griberman, "Using Stanford CoreNLP Capabilities for Semantic Information Extraction from Textual Descriptions," in *Evaluation of Novel Approaches to Software Engineering*, Riga, Latvia, Riga Technical University, 2020.

[24] C. Buckley, "The importance of proper weighting methods," *Proceedings of the workshop on Human Language Technology*, pp. 349-352, 1993.

[25] H. R. Turtle, Inference networks for document retrieval, University of Massachusetts Amherst, 1991.

المـــــراجع

[1] D. Harman, Information Retrieval: The Early Years, IEEE, 2019, p. 166.

[2] A. Singhal, "Modern Information Retrieval: ABrief Overview," *IEEE,* 2001.

[3] V. Bush, "As We May Think," *Atlantic Monthly,* vol. 176, pp. 101-108, 1945.

[4] M. Sanderson and B. Croft, "The History of Information Retrieval Research," *IEEE,* vol. 100, no. Special Centennial Issue, 2012.

[5] C. Mooers, "Zatocoding applied to mechanical organization of knowledge," *Journal of the Association for Information Science and Technology,* vol. 2, no. 1, pp. 20-32, 1951.

[6] H. Luhn, "A statistical approach to mechanized encoding and searching of literary information," *IBM Journal of Research and Development,* 1957.

[7] S. Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of documentation,* vol. 28, no. 1, pp. 11-21, 1972.

[8] G. Salton and C. Yang, "on the Soecification of Term Value in Autoatic Indexing," *Cornell University,* pp. 73-173, 1973.

[9] T. Gondaliya and H. Joshi, "Journey of Information Retrieval to Information Retrieval Tools-IR&IRT A Review," in *CALIBER-2017*, CHENNAI, 2017.

[10] G. Salton and M. J. McGill, Introduction to Modern Information Retrieval, 1983: McGraw-Hill Book Company, New York.

[11] C. D.Manning, P. Raghavan and H. Schütze, An Introduction to Information Retrieval, New York: Cambridge University Press, 2008.

[12] A. M.Elshami and S. Hassaballah, Arabic Encyclopedia Of Library, Information, and Computer Terms, Cairo: ACADEMIC BOOKSHOP, 2001.

[13] A. Göker and J. Davies, "Information Retrieval Models," *Wiley,* pp. 1-19, 2009.

[14] D. Mabrouk, S. Rady , N. Badr and . M. , "Modeling using Term Dependencies and Term Weighting in Information Retrieval Systems," *Eighth International Conference on Intelligent Computing and Information Systems (ICICIS),* vol. 42, pp. 321-328, 2018.

[15] V. Gudivada, D. L.Rao and A. R.Gudivada, "Information Retrieval: Concepts, Models, and Systems," in *Handbook of Statistics*, United States, 2018, pp. 331-401.

[١٦] ف. س. بامفلح، أساسيات نظم استرجاع المعلومات الإلكترونية، السعودية: مكتبة الملك فهد الوطنية، ٢٠٠٦، p. 329.

[١٧] م. س. النشرتي، "التحديات التي تواجه خوارزميات محركات البحث في استرجاع المحتوى العربي على الشبكة العنكبوتية العالمية،" *Cybrarians Journal*، المجلد ٣٠، ٢٠١٢.

[18] M. and K. , "Probabilistic Indexing and Information Retrieval," *Journal of the ACM (JACM),* vol. 7, pp. 216-244, 1960.

[19] E. Gaussier and F. Yvon, Textual Information Access: Statistical Models, WILEY, 2012.

[٢٠] د. ا. غ. دبور، "نظم استرجاع المعلومات العربية واتجاهات البحوث المعاصرة،" *الاتحاد العربي للمكتبات والمعلومات*، ٢٠١٥.

[21] K. Jones and S. Robertson, "Relevance weighting of search terms," *Journal of the American Society for Information Sciences,* vol. 27, no. 3, pp. 129-146, 1976.

[22] Mastering Elasticsearch 5.x, Third ed., UK: Packt Publishing Ltd, 2017.

[23] H. Turtle and W. B. Croft, "Inference networks for document retrieval," *ACM SIGIR Forum,* vol. 51, 2017.

[24] T.-Y. Liu, "Learning to Rank for Information Retrieval," *Foundations and Trends in Information Retrieval,* vol. 3, no. 3, p. 225–331, 2009.

[25] M. Pannu, A. James and R. Bird , "Comparison of Information Retrieval Models," in *The 19th Western Canadian Conference on Computing Education - In-Cooperation with ACM SIGCSE*, 2014.

[26] T. Sabbah, A. Selamat, M. H. Selamat, F. S. Al-Anzi, E. H. Viedma and O. Krejcar, "Modified Frequency-Based Term Weighting Schemes for Text Classification," *Applied Soft Computin,* vol. 58, pp. 193-206, 2017.

[27] S. JABRI, A. DAHBI, T. GADI and A. BASSIR, "Ranking of Text Documents using TF-IDF Weighting," in *2018 4th International Conference on Optimization and Applications (ICOA)*, Mohammedia, Morocco, 2018.

[28] J. Savoy and E. Gaussier, "Information Retrieval," in *Handbook of Natural Language Processing*, Chapman and Hall/CRC, 2010, pp. 455-484.

[29] W. B. Crof, D. Metzler and T. Strohman, Search Engines, Pearson Education, 2015.

[30] E. Chauhan and D. Asthana, "Review of Indexing Techniques in Information Retrieval," *International Journal of Engineering Science and Computing IJESC ,* vol. 7, no. 7, pp. 13940-13942, 2017.

[31] B. Saini, V. Singh and S. Kumar, "Information Retrieval Models and Searching Methodologies: Survey," *International Journal of Advance Foundation and Research in Science & Engineering (IJAFRSE),* vol. 1, no. 2, pp. 57-62, 2014.

[32] V. Murthy, D. B. V. Vardhan, K. Sarangam and P. V. p. Reddy, "A COMPARATIVE STUDY ON TERM WEIGHTING METHODS FOR AUTOMATED TELUGU TEXT CATEGORIZATION WITH EFFECTIVE CLASSIFIERS," *International Journal of Data Mining & Knowledge Management Process (IJDKP),* vol. 3, no. 6, pp. 95-105, 2013.

[33] V. Gurusamy, S. Kannan and K. Nandhini, "Performance Analysis: Stemming Algorithm for the English Language," *IJSRD - International Journal for Scientific Research & Development|,* vol. 5, no. 05, pp. 1933-1938, 2017.

[34] S. Joseph, H. Hlomani, K. Letsholo, F. Kaniwa and K. Sedimo, "Natural Language Processing: A Review," *International Journal of Research in Engineering and Applied Sciences,* vol. 6, no. 3, 2016.

[35] D. M. J.Garbade, "A Simple Introduction to Natural Language Processing," *Becoming Human: Artificial Intelligence Magazine,* 2018.

[36] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard and D. McClosky, "The Stanford CoreNLP Natural Language Processing Toolkit," in *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics*, Maryland USA, 2014.

[37] E. Nazaruka, J. Osis and V. Griberman, "Using Stanford CoreNLP Capabilities for Semantic Information Extraction from Textual Descriptions," in *Evaluation of Novel Approaches to Software Engineering*, Riga, Latvia, Riga Technical University, 2020.

[38] P. Qi, Y. Zhang, Y. Zhang, J. Bolton and C. D. Manning, "Stanza : A Python Natural Language Processing Toolkit," *Stanford, CA 94305,* 23 Apr 2020.

[39] A. K. Singhal, Term Weighting Revisited, Cornell University. ProQuest Dissertations, 1997.

[40] Z. Dai and J. Callan, "Context-Aware Term Weighting For First Stage Passage Retrieval," in *SIGER*, China, 2020.

[41] J. Ramos, "Using TF-IDF to Determine Word Relevance in Document Queries," January 2003.

[42] T. Xia and Y. Chai, "An Improvement to TF-IDF: Term Distribution based Term Weight Algorithm," *JOURNAL OF SOFTWARE,* vol. 6, 2011.

[43] N. Polettini, "The Vector Space Model in Information Retrieval- Term Weighting Problem," in *Sommarive 14, 38050 Povo (TN)*, Italy, 2004.

[44] R. K. Roul, J. K. Sahoo and K. Arora, "Modified TF-IDF Term Weighting Strategies for Text Categorization," in *14th IEEE India Council International Conference (INDICON)*, 2018.

[45] T. ZHANG and S. G. Sam , "An Improved TF-IDF Algorithm Based on Class Discriminative Strength for Text Categorization on Desensitized Data," in *ICIAI 2019: Proceedings of the 2019 3rd International Conference on Innovation in Artificial Intelligence*, Suzhou, China, 2019.

[46] H. R. Turtle, Inference networks for document retrieval, University of Massachusetts Amherst, 1991.

[47] C. Buckley, "The importance of proper weighting methods," *Proceedings of the workshop on Human Language Technology,* pp. 349-352, 1993.

فهرس الجداول

فهرس الأشكال

فهرس المحتـــويات