مكنز عربي للمساعدة في الاستنباط الآلي للعبارات المفتاحية

محمد اللقماني١ وحسني المحتسب٢

مقدمة

تُظهِر العبارات المفتاحية في مستند ما المواضيع الأساسية المناقشة في ذلك المستند. ونظراً لعدم توفر العبارات المفتاحية في كثير من مراكز المحتوى الرقمي، فقد أصبحت الحاجة ملحة إلى خوارزميات عالية الكفاءة لاستخراج العبارات المفتاحية في النصوص.

وتهدف خوارزميات استخراج العبارات المفتاحية آليا إلى الاستفادة من التقدم في الحوسبة من حيث السرعة والكفاءة لحساب حل مشاكل استكشاف واستخدام العبارات المفتاحية مع تقليل التكاليف (في الجهد والوقت) المرتبطة بعمل البشر في تصنيف المستندات.

ندرس في هذا العمل البحثي بعضا من السمات التي يمكن استخدامها لتحسين جودة استخراج العبارات المفتاحية وتطبيقها على خوارزمية تدعى "خوارزمية استخراج العبارات المفتاحية". ونجري أيضاً دراسة تحليلية للخوارزمية المحسنة مقارنة مع بعض الخوارزميات المستخدمة في نفس المجال.

ونستخدم في هذه الدراسة التحليلية مكنزين من مكانز البيانات. يحوي المكنز الأول مستندات تمثل أبحاثا علمية باللغة الإنجليزية. بينما يحوي المكنز الثاني مستندات أعددناها كجزء من هذا العمل تمثل وثائق باللغة العربية.

يحوي المكنز العربي الذي نعرضه هنا مستندات عربية تحوي ٤٠٠ مستندا موزعة على ١٨ موضوعاً. وقد جمعنا هذه المستندات من مصدرين رئيسين هما: الويكيبيديا العربية ومبادرة الملك عبد الله لإثراء المحتوى العربي.

وقد قمنا بدارسة المستندات واستخراج الكلمات المفتاحية يدويا حتى تكون مرجعا لبحوث الاستنباط الآلي للكلمات المفتاحية العربية. وتغطي مجموعتنا العديد من فروع المعرفة وذلك لإضفاء بعض الشمولية على محتوياتها. كما تختلف مستندات المكنز من حيث الحجم وعدد الصفحات.

تعريف خوارزميات استخراج العبارات المفتاحية

تهدف خوارزميات استخراج العبارات المفتاحية آليا إلى استنباط عبارات ذات جودة عالية لتصف محتوى المستند كاملاً. ويتم استخراج هذه العبارات آلياً دون تدخل بشري، وفي بعض الأحيان مع قليل من التدخل البشري. وتنقب خوارزميات الاستخراج الآلي على العبارات المفتاحية في مكانز المستندات وتقوم بوسم المستندات بما يصفها من العبارات المفتاحية. ويمكننا تصنيف الخوارزميات حسب المجال المستخدم، فهناك خوارزميات تهدف لاستخلاص العبارات من مستندات ذات تخصص معين، وهناك من لا يشترط أن تكون المستندات حول مجال محدد بل عامة ومتنوعة. وتميل الخوارزميات التي تستهدف المستندات العامة إلى استخدام سمات بحث غير مرتبطة بنوع المستند، ونذكر من الأمثلة على هذا النوع من الخوارزميات (KP-Miner) وهو ما قام به البلتاجي ورافع في [١] وكذلك العمل المذكور في [٢].

يمثل مجال البحوث العلمية أحد المجالات التي تركز عليها خوارزميات استخراج الكلمات المفتاحية. وهذا المجال متخصص من حيث ترتيب المستندات ونوعيتها. ومن أمثلة البحوث في هذا المجال أعمال الباحثين في [٣] و[٤] و[٥]. وقد يكون من أسباب كثرة الخوارزميات في هذا المجال هو توفر العديد من المكانز المدققة والتي يمكن استخدامها للتحقق من جودة هذه الخوارزميات. ومع ذلك، لا تقتصر المجالات المتخصصة على المستندات العلمية، فهناك خوارزميات تركز على استخراج العبارات المهمة من توتير (Twitter) [٦]. كما أن هناك خوارزميات تستنبط كلمات مفتاحية من محاضرات الفيديو مثل ما ورد في [٧].

من جهة أخرى فأنه يمكننا تصنيف خوارزميات استخراج العبارات المفتاحية آلياً حسب الطريقة المتبعة للاستخراج إلى صنفين رئيسين وهما الفهرسة الحرة والفهرسة المتحكم بها (مقيدة).

الفهرسة الحرة

تعتمد خوارزميات استخراج العبارات المفتاحية آلياً المبنية على الفهرسة الحرة على أساليب الذكاء الاصطناعي. ويمكننا تقسيم هذا المجال إلى مجالين فرعيين هما:

١. الخوارزميات المعتمدة على التعلم: وهي خوارزميات تنتهج التعلم الآلي تحت الإشراف (supervised) أو دون إشراف (unsupervised) بحيث يتم استخراج العبارات المفتاحية بواسطة أمثلة أو مكانز تدريب. ومن الأمثلة على ذلك ما ورد في [٨]، [٩] و[١٠].

٢. الخوارزميات التي لا تعتمد على التعلم: في هذا النوع نعتمد على طرق إحصائية أو حقائق مبنية على قواعد اللغة أو المعلومات المعجمية لاستخراج العبارات المفتاحية من دون الحاجة إلى مستندات تدريبة. على الرغم بأن جودة النتائج في هذه المجال هي أقل من نظيرتها المعتمدة على التعلم لكنها من جهة أخرى فهي أقل تكلفة في التنفيذ لكونها لا تحتاج إلى مكانز تدريب معدة مسبقا. من الأمثلة التي وجدناها تنهج هذا المجال ما قام به الباحثين في [٣] و[٥].

الفهرسة المقيدة

تستخدم خوارزميات استخراج العبارات المفتاحية آلياً المبنية على الفهرسة المقيدة قواميسا ومرادفات معرّفة مسبقاً، وعادة ما تعتمد هذه الخوارزميات على التعلم. ومن أمثلة هذا النوع ما ذُكر في [١٠] و[١١].

وعند البحث عن تطبيق هذه الخوارزميات على اللغة العربية، نجد القليل من البحوث التي تُعنى باستخراج الكلمات المفتاحية من النصوص العربية. ومن أبرز هذه البحوث خوارزمية KP-Miner والمقدمة من البلتاجي [١] وكذلك مستخرج صخر (Sakhr Extractor) [١٢] وكذلك العمل المقدم من الششتاوي [١٣]. ولاحظنا أن المكانز المستخدمة في هذه البحوث كانت صغيرة وتحوي عددا قليل من المستندات. فمثلاً استخدم اختبار KP-Miner ١٠٠ مستندا من ويكيبيديا (Wikipedia) بينما استُخدم ٥٠ مستندا في العمل البحثي [١٣].

خوارزمية استخراج العبارات المفتاحية KEA

استُخدمت خوارزمية KEA في العديد من البحوث المتعلقة باستخراج العبارات المفتاحية مثل [٣] و[١٤] و[١٦] وهي خوارزمية متوفرة للعامة. وقد عُرضت هذه الخوارزمية لأول مرة في عام ١٩٩٩ كخوارزمية تتبع مجال التعلم الآلي تحت الإشراف واختُبرت على مجموعة من البحوث العلمية. وأضافت خوارزمية KEA++ الموضحة في [١٥] إمكانية استخدام القواميس المقيدة بصيغة النظام المبسط لتنظيم المعرفة (Simple Knowledge Organization System) أو ما يعرف اختصاراً بصيغة SKOS [١٧]. تستخدم خوارزمية KEA أربع سمات لاستخراج العبارات المفتاحية وهي:

١. معكوس تكرار الكلمة بالنسبة لتكرار المستند (TF-IDF) وتعكس هذه السمة أهمية كلمة مذكورة في مستند ما بالنسبة لجميع مستندات المكنز. ويطلق على الجزء (TF) بالوزن المحلي للمصطلح بينما (IDF) هو الوزن العام للمصطلح بالنسبة لكل المستندات.

٢. الظهور الأول للعبارة المرشحة (First occurrence).

٣. طول العبارة المفتاحية (Phrase length).

٤. درجة ارتباط العبارة المرشحة بالعبارات الأخرى أو ما يعرف ب (Node degree).

المكانز

تلعب مكانز المستندات دوراً كبيراً في التحقق من كفاءة خوارزميات استخراج الكلمات والعبارات المفتاحية. استخدمنا في بحثنا هذا مكنزين، أحدهما باللغة الإنجليزية والآخر باللغة العربية.

المكنز الإنجليزي

اخترنا المكنز الذي قدم في بحث سُلم لورشة عمل التقييم الدلالي والتي أقيمت في عام ٢٠١٠ [١٤]. ويتوفر هذا المكنز لعامة الباحثين لكي يتم استخدامه في مجال استخراج العبارات المفتاحية آليا. ويمكن معرفة المزيد عن المكنز بالرجوع إلى [١٨]. ويحوي هذا المكنز على ٢٤٤ مستندا جمعت من مكتبة ACM الرقمية. وتحوي المجموعة على أوراق عمل من مؤتمرات وورش عمل يتراوح طول صفحاتها من ٦ إلى ٨ صفحات. وللتأكد من تنوع مواضيع المكنز، اختيرت المواضيع من أربعة فروع مختلفة في البحث حسب تصنيف ACM، وهي: الأنظمة الموزعة (Distributed Systems)، والبحث عن المعلومات واسترجاعها (Information Search and Retrieval)، والذكاء الصناعي الموزع (Distributed Artificial Intelligence)، والعلوم الاجتماعية والسلوكية التابعة للاقتصاد (Social and Behavioral Sciences – Economics).

المكنز العربي

يساهم عملنا البحثي في هذه الورقة بتقديم مكنز جديد باللغة العربية أعددناه لاستخدامه في هذا البحث ولدعم الأعمال المستقبلية والمهتمة باللغة العربية. وقد جمعنا مستندات المكنز العربي من مصدرين هما ويكيبيديا العربية [١٩] ومبادرة الملك عبد الله لإثراء المحتوى العربي [٢٠].

أخترنا ويكيبيديا العربية كمصدر أساسي للمكنز كونها تحوي أكثر من ١٩٨،٣٤٩ مقال. واخترنا من هذا المصدر ٣٦٥ مقال، منها ٢٠٠ مقال اختيرت من العمل الذي قام به شعبان في [٢١] بينما جمعنا باقي المقالات عن طريق برنامج BzReader [٢٢]. ويساعد BzReader على تصفح مكانز مقالات ويكيبيديا من غير الحاجة للاتصال بشبكة الأنترنت.

كان مصدرنا الثاني هو مبادرة الملك عبد الله لإثراء المحتوى العربي والتي تهدف إلى إثراء محتوى اللغة العربية المتوفر على شبكة الأنترنت بعد ملاحظة قلتها. فحسب القائمين على هذه المبادرة فإن محتوى اللغة العربية لا يتعدى ٣,٠٪ بين لغات العالم في ذلك الوقت. استخدمنا في مكنزنا ٣٥ مقال من المجال الطبي.

يغطي المكنز العربي مجالات وفئات عدة كالتاريخ والدين والجغرافيا والتقنية والرياضة وغيرها. وقد صنفنا مستندات المكنز حسب مجال المقال إلى ١٨ صنفاً. واخترنا مواضيع مختلفة لإضفاء الشمولية والتنوع وعدم التقيد بمجال معين. ونود الإشارة إلى أن عدد صفحات المستندات المختلفة تتراوح بين ١ إلى ٣٠ صفحة بينما تحوي تقريباً ما بين ١٧٢ إلى ١٧،٥٨٩ كلمة. وقد حفظت جميع مستندات المكنز على شكل ملفات نصية بامتداد (.txt) بترميز (UTF-٨). أكبر فئة من حيث عدد المستندات كان من نصيب مجموعة الأشخاص وتحوي ٥٩ ملفا بينما أصغر فئة فكانت فئة المأكولات واحتوت على ٣ ملفات فقط. ولقد حسبنا معدل الكثافة لكل فئة، ويعبر هذا المقياس عن متوسط كثافة الكلمات في كل ملف من كل فئة. ويهدف هذا المعيار لإظهار غنى الفئات على حسب طول الملفات وليس حسب عدد الملفات تحت تلك الفئة. وقد وجدنا أن فئة الدول كانت الأعلى بتسجيلها متوسط كلمات قدره ٧،٣٦٦ كلمة، تلتها فئة الأديان بمعدل ٥،٩٦٠ كلمة في كل ملف، علماً بأن هذه الفئة حوت على ١٦ ملفا فقط. ومن ناحية أخرى وجدنا بأن فئة الصحة والطب سلجّت معدل كثافة قدره ١،٩٥٠ كلمة لكل ملف على الرغم من أن عدد ملفات هذه الفئة هو ٥١ ملفا. وكان أكبر ملف في مكنزنا من فئة التاريخ ويحكي عن الدولة العثمانية، بينما الأصغر كان من فئة البيئة وتحدث عن التلوث الإشعاعي.

لاحظنا عند تحويل مقالات الويكيبيديا من الصيغة النصية بواسطة برنامج BzReader بلزوم القيام بعمليات تنقية للملفات. وشملت عملية التنقية حذف بعض النصوص التي قد تؤثر سلباً في استخراج العبارات المفتاحية بالنسبة للقراء والخوارزميات.

بعد عملية التنقية استعنا ببعض المتطوعين لقراءة المستندات واستخراج ١٠ عبارات مفتاحية من كل مستند. وحفظت هذه العبارات في ملفات منفصلة تحت نفس اسم الملف ولكن بصيغة (.key)، إذ أن هذه الصيغة تستخدمها خوارزمية KEA وبعض الخوارزميات في نفس المجال. ويحوي كل ملف بصيغة (.key) العبارات المفتاحية مفصلة كل عبارة في سطر على حده مرتبة من أعلى إلى أسفل حسب الأهمية.

وكانت الخطوة الأخيرة في تجهيز المكنز العربي التحقق أو التدقيق. وشملت هذه الخطوة مراجعة العبارات المفتاحية وتصحيح الأخطاء المطبعية وتعديل ما يلزم.

ويسرنا، كجزء من هذا العمل، أن نوفر هذه المستندات للباحثين المهتمين في هذا المجال لاستخدامها في أعمال مستقبلية في مجال الاستنباط الآلي. ويمكن الوصول للمكمن عن طريق هذا الرابط: https://github.com/logmani/ArabicDataset

الإعداد للمقارنات التحليلية

أجرينا في هذا البحث دراسة تجريبية احتوت على أربع مقارنات تحليلية اعتمدت على تحسينات على خوارزمية KEA. وقمنا بإجراء كل مقارنة تحليلية مرتين: واحدة بواسطة المكنز الإنجليزي والأخرى عن طريق مكنز اللغة العربية. واخترنا في كل تحليل ١٠٠ مستند. وفيما يلي نعرض مقارناتنا التحليلية.

تجذير الكلمات (Stemming Vs. No-Stemming)

قمنا بتجربة استخدام خوارزمية KEA واعتماد تجذير المستندات إلى أصول الكلمات وقمنا بمقارنة كفاءة الخوارزمية في حال استخدام وعدم استخدام التجذير. بالنسبة لمقارنتنا التحليلية على اللغة العربية فقد استعنّا بتجذير خوجة [٢٣] بعد أن قمنا ببعض التعديلات عليه لحل بعض مشاكل الذاكرة والتي واجهناها عند قيامنا بتجذير بعض الملفات ذات الحجم الكبير. من جهة أخرى فقد اخترنا لمقارنتنا التحليلية على المكنز الإنجليزي تجذير بورتر (Porter) [٢٤] وتجذير لوفنز (Lovins) [٢٥] نظير شعبيتهما في البحوث المعقودة على اللغة الإنجليزية.

صيغة وزن المصطلحات (Top Term Weighting Formula)

قمنا في هذه المقارنة بالتركيز على تعديل سمة (TF-IDF) واستبدالها ببعض السمات التي تعطي أوزاناً مختلفة للمصطلحات. وقد عدلنا خوارزمية KEA لتستخدم أحد ثماني صيغ مختلفة لأوزان المصطلحات. اخترنا خمسة من الثماني صيغ من بحوث علمية سابقة في مجال استرجاع المعلومات وهي: أفضل تطابق أو ما يعرف بـ (Best Match ٢٥) واختصاراً (BM٢٥) [٢٦]، والمعكوس الاحتمالي (Probabilistic Inverse) واختصاراً (IDFP) [٢٧]، وصيغة الوزن اللوغاريثمية (Logarithmic) واختصاراً (LOGA) [٢٧]، واللوغاريثم المضاف (Augmented log) ويعرف اختصاراً بـ (LOGG) [٢٨]، وأخيراً صيغة الجذر التربيعي (Square Root) واختصارها (SQRT) [٢٨]. أما بالنسبة للثلاث الصيغ المتبقية فلقد اقترحناها كجزء من هذا العمل وفيما يلي وصفها.

- LOGA-BM٢٥: في هذه الصيغة قمنا بدمج صيغتي أفضل التطابق (BM٢٥) مع الصيغة اللوغاريثمية (LOGA) وذلك أملاً بأن تعطي نتيجة أفضل. قمنا باستخدام الصيغة اللوغاريثمية كمعامل محلي للمصطلح بينما استخدمنا صيغة أفضل تطابق (BM٢٥) كصيغة للمعامل العام بدلاً من (IDF).
- LOGA-IDFP: في هذه المشاركة قمنا باستخدام الصيغة اللوغاريثمية (LOGA) كمعامل محلي للمصطلح بينما استخدمنا صيغة المعكوس الاحتمالي (IDFP) كصيغة للمعامل العام.
- LOGG-BM٢٥: هي آخر صيغة مقترحة لاستبدال السمة (TF-IDF) المستخدمة في الخوارزمية. قمنا هنا باستخدام الصيغة اللوغاريثمية المضافة كوزن محلي وصيغة أفضل تطابق كمعامل عام.

الظهور الأخير للعبارة المرشحة والظهور الأول معاً (Combined First and Last Occurrence)

تستخدم خوارزمية KEA سمة الظهور الأول لإعطاء وزن أكبر للعبارات المرشحة والتي تذكر في مقدمة المستندات. وفي مقارنتنا التحليلية هذه اقترحنا تعديلين على هذه السمة: الأول يعطي وزنا أكبر للعبارات المرشحة والتي تذكر في آخر المستندات. فمثلاً هناك العديد من المستندات التي تحوي خاتمة للمقال تلخص فيه أهم ما ذكر فيه. واقتراحنا الثاني يهدف إلى إعطاء وزن أكبر لتلك العبارات المرشحة والتي تُذكر في مقدمة وخاتمة المقال. وفي هذه المقارنة التحليلية قمنا بمقارنة الوضع الافتراضي لخوارزمية KEA باستخدام (TF-IDF) وكذلك أفضل التعديلات من المقارنة التحليلية السابقة.

خوارزمية KEA المحسنة مع خوارزميات أخرى

نقوم في هذه المقارنة التحليلية بمقارنة تعديلات KEA المحسنة والتي تفوقت على نظيراتها من المقارنة السابقة مع خوارزميات معروفة سابقاً. والخوارزميات المستهدفة هي مستخرج العبارات المفتاحية (Keyphrase Extractor) المقدم من كومار [٣] وخوارزمية KP-Miner المقدمة من البلتاجي ورافع [١].

المقاييس المستخدمة للمقارنة

لتقييم أداء التعديلات التي ذكرناها، قمنا باستخدام مقياس التطابق التام (exact matching) [٩]. إذ تعتبر العبارة المستردة آلياً صحيحة في مقياس التطابق التام فقط في حال تطابقها التام مع العبارة المذكورة في قائمة العبارات المعدة يدوياً. وقمنا بحساب متوسط كفاءة الاسترجاع (Recall) ودقة الاسترجاع (precision) كمؤشرات أساسية لتقييم كفاءة الخوارزميات المختبرة.

نتائج المقارنات التحليلية

تنتج الخوارزمية الخاضعة للاختبار، في جميع مقارناتنا التحليلية، ١٠٠ ملف بصيغة 'key.' تحوي العبارات المفتاحية لمستخرجة آلياً. نقوم بعد ذلك بتحليل محتوى هذه الملفات ومقارنتها مع العبارات المفتاحية المستخرجة يدوياً المعدة مسبقا في المكنزين العربي والإنجليزي.

نتائج تجذير الكلمات عدم تجذيرها

في أول مقارناتنا التحليلية لدارسة الأداء مع وبدون خاصية التجذير بواسطة المكنز الإنجليزي، لاحظنا بشكل عام أن أداء الوضع الافتراضي باستخدام تجذير Porter وبدون التجذير حققا نتائج أفضل من تجذير Lovins. وتدعم هذه النتائج الشعبية التي يحظى بها تجذير Porter وأنه التجذير الافتراضي لخوارزمية KEA.

عند قيامنا بإجراء نفس المقارنة على المحتوى العربي وجدنا أن خوارزمية KEA مدعمة بتجذير خوجة حققت ١٥٠ عبارة متطابقة بينما استطعنا الحصول على ١٨٩ عبارة متطابقة عندما لم نستخدم أي تجذير. نعتقد أننا حصلنا على هذه النتيجة عطفاً على صعوبة التجذير الآلي في اللغة العربية. ومن جهة أخرى وجدنا تطابق المقاييس الأخرى (دقة الاسترجاع وكفاءة الاسترجاع) مع مقياس التطابق التام على المكنزين العربي والإنجليزي.

نتائج صيغة وزن المصطلحات

وجدنا في المقارنة التحليلية الثانية (صيغة وزن المصطلحات) على المكنز الإنجليزي أن الوضع الافتراضي لخوارزمية KEA باستخدام سمة معكوس تكرار الكلمة بالنسبة لتكرار المستند (TF-IDF) قد سجل أعلى درجة في مقياسي التطابق التام ودقة الاسترجاع، بينما سجل تعديل الخوارزمية باستخدام سمة أفضل تطابق (BM٢٥) في مؤشر كفاءة الاسترجاع أعلى درجة. نستنتج من هذه النتيجة أن

تعديل الخوارزمية باستخدام سمة أفضل تطابق هو أفضل من الوضع الافتراضي مع المكانز التي تحوي عددا أقل من العبارات المفتاحية المستخرجة يدوياً. يعرض جدول رقم ١ نتائج هذه المقارنة، بينما يعرض جدول رقم ٢ نتائج المقارنة التحليلية الثانية التي أُجريت على مكنز اللغة العربية. نلاحظ هنا أن تحسين الخوارزمية باستخدام سمة أفضل تطابق (BM٢٥) حققت أعلى النتائج في كل مؤشراتنا.

جدول رقم ١: نتائج صيغة وزن المصطلحات على المكنز الإنجليزي

المقياس	الوزن الافتراضي	BM٢٥	IDFP	LOGA	LOGG	SQRT	LOGG+BM٢٥	LOGA+IDFP	LOGA+BM٢٥
مقياس التطابق التام	٦٠	٥٩	٣٣	٥٣	٤٧	٣٤	٥٦	٣٤	٥٠
دقة الاسترجاع	٠,١٥٠	٠,١٤٨	٠,٠٨٣	٠,١٣٣	٠,١١٨	٠,٠٨٥	٠,١٤٠	٠,٠٨٥	٠,١٢٥
كفاءة الاسترجاع	٠,١٥٦	٠,١٥٩	٠,٠٩١	٠,١٣٩	٠,١٣٢	٠,٠٩٢	٠,١٥٥	٠,٠٩٢	٠,١٣١

جدول رقم ٢: نتائج صيغة وزن المصطلحات على المكنز العربي

المقياس	الوزن الافتراضي	BM٢٥	IDFP	LOGA	LOGG	SQRT	LOGG+BM٢٥	LOGA+BM٢٥
مقياس التطابق التام	١٨٩	١٩١	١٥٤	١٨٧	١٧٣	١٥٤	١٧٧	١٩٠
دقة الاسترجاع	٠,١٨٩	٠,١٩١	٠,١٥٤	٠,١٨٧	٠,١٧٣	٠,١٥٤	٠,١٧٧	٠,١٩٠
كفاءة الاسترجاع	٠,١٩١	٠,١٩٣	٠,١٥٧	٠,١٨٩	٠,١٧٦	٠,١٥٧	٠,١٨٠	٠,١٩٣

نتائج الظهور الأخير والظهور الأول معا

ركزنا في هذه المقارنة التحليلية على تعديل سمة الظهور الأول وإبدالها بسمتي الظهور الأخير والظهور الأول معا للعبارات في أول وآخر المقال. أخذنا في تجربتنا هذه من التجربة السابقة سمة التطابق الأفضل (BM٢٥) إضافة إلى سمة الوزن الافتراضية لخوارزمية KEA باستخدام سمة معكوس تكرار الكلمة بالنسبة لتكرار المستند (TF-IDF). قمنا بتعديل الخوارزمية لإعطاء وزن أكبر لمكان ظهور العبارات المرشحة. وجدنا في جزئي المقارنة على اللغة العربية والإنجليزية أن الوضع الافتراضي باستخدام الظهور الأول مع TF-IDF حقق نتائج أفضل من باقي التحسينات. وتبين لنا من خلال هذه النتائج أن العبارات المؤثرة تُذكر في أول المقال في المقالات العامة. كما أن المقال قد لا يحتوي على خلاصة لتعيد ذكر العبارات المهمة. يعرض جدول رقم ٣ وجدول رقم ٤ ملخص نتائج هذه المقارنة التحليلية.

جدول رقم ٣: نتائج الظهور الأخير والظهور الأول معا على المكنز الإنجليزي

المقياس	الوضع الافتراضي	أول وآخر ظهور مع TF-DF	أول وآخر ظهور مع BM٢٥	آخر ظهور مع TF-IDF	آخر ظهور مع BM٢٥
مقياس التطابق التام	٦٠	٥٨	٥٧	٤٨	٤٧
دقة الاسترجاع	٠,١٥٠	٠,١٤٥	٠,١٤٣	٠,١٢٠	٠,١١٨
كفاءة الاسترجاع	٠,١٥٦	٠,١٤٣	٠,١٤٩	٠,١١٧	٠,١١٢

جدول رقم ٤: نتائج الظهور الأخير والظهور الأول معا على المكنز العربي

المقياس	الوضع الافتراضي	أول وآخر ظهور مع TF-IDF	أول وآخر ظهور مع BM٢٥	آخر ظهور مع TF-IDF	آخر ظهور مع BM٢٥
مقياس التطابق التام	١٨٩	١٨٥	١٨٧	١٥٣	١٥٨
دقة الاسترجاع	٠,١٨٩	٠,١٨٥	٠,١٨٧	٠,١٥٣	٠,١٥٨
كفاءة الاسترجاع	٠,١٩١	٠,١٨٨	٠,١٩٠	٠,١٥٦	٠,١٦١

خوارزمية KEA المحسنة مع خوارزميات أخرى

قارنا الوضع الافتراضي لخوارزمية KEA واقتراحنا للظهور الأول والأخير معا للعبارات المفتاحية مع مستخرج العبارات المفتاحية وخوارزمية KP-Miner. كان التفوق هنا لخوارزمية KEA في جميع المؤشرات. ورغم سرعة مستخرج العبارات المفتاحية في إظهار النتائج إلا أن النتائج كان ضعيفة وهذا يظهر تفوق الخوارزميات المعتمدة على التعلم. يعرض جدول رقم ٥ تفاصيل التجربة.

جدول رقم ٥: نتائج خوارزمية KEA المحسنة مع خوارزميات أخرى على المكنز الإنجليزي

المقياس	الوضع الافتراضي	أول وآخر ظهور مع TF-IDF	KP-Miner	مستخرج العبارات المفتاحية
مقياس التطابق التام	٦٠	٥٨	٥٢	٤
دقة الاسترجاع	٠,١٥٠	٠,١٤٥	٠,١٣٠	٠,٠١٠
كفاءة الاسترجاع	٠,١٥٦	٠,١٤٣	٠,١٣٦	٠,٠٠٩

في الجزء الثاني من المقارنة والمطبق على المكنز العربي، قارنا اثنين من اقتراحاتنا لتحسين خوارزمية KEA حسب نتائج التجارب السابقة وهذه الاقتراحات هي صيغة وزن المصطلحات للتطابق الأفضل (BM٢٥) وكذلك الظهور الأول والأخير معا للعبارات المفتاحية. وكما في الجزء الأول من هذه المقارنة، أظهرت مقترحاتنا تفوقاً على الخوارزميتين الأخريين في جميع المؤشرات. سجل مقياس دقة الاسترجاع أعلى تحسن في هذه الاختبارات ٠,١٩ ٪ وهي نسبة أقل بقليل من نسبة الاسترجاع اليدوي بواسطة محترفين، ولكن عند الأخذ بعين الاعتبار الفائدة التي حصلنا عليها من الاستخراج الآلي فإن النسبة تعتبر مقبولة. يعرض جدول رقم ٦ تفاصيل التجربة.

جدول رقم ٦: نتائج خوارزمية KEA المحسنة مع خوارزميات أخرى على المكنز العربي

المقياس	BM٢٥	أول وآخر ظهور مع TF-IDF	KP-Miner	مستخرج العبارات المفتاحية
مقياس التطابق التام	١٩١	١٨٧	١٦٧	٩٥
دقة الاسترجاع	٠,١٩١	٠,١٨٧	٠,١٦٧	٠,٠٩٦
كفاءة الاسترجاع	٠,١٩٣	٠,١٩٠	٠,١٦٩	٠,٠٩٧

الخاتمة

قمنا في بحثنا المقدم بدراسة مجال استخراج العبارات المفتاحية آلياً وذكرنا بعض تصنيفاتها. وعرضنا مكنزا باللغة العربية أعددناه كجزء من هذا العمل. ويحتوي المكنز على ٤٠٠ مستند تحوي مقالات في ١٨ مجالاً ويقابلها ٤٠٠ ملف تحوي العبارات المفتاحية المستخرجة يدوياً. ويسرنا أن نوفر هذا المكنز العربي للباحثين المهتمين في هذا المجال لاستخدامها في أعمال مستقبلية في مجال الاستنباط الآلي.

قمنا أيضاً في هذا البحث بإجراء العديد من التحسينات على خوارزمية KEA وقمنا بعقد بعض المقارنات التحليلية على مكنزين

أحدهما باللغة الإنجليزية والآخر باللغة العربية.

أظهرت بعض تحسيناتنا تفوق بسيطا في كفاءة خوارزمية KEA على وضع الخوارزمية الافتراضي ولكنه ليس بالتفوق المعتبر. ولكن تحسيناتنا أظهرت تفوقاً ملحوظاً على خوارزميتين من البحث العلمي وهما مستخرج العبارات المفتاحية وKP-Miner.

ونتوقع كعمل مستقبلي متابعة هذا البحث بالقيام بإنشاء قاموس مقيد يتبع صيغة النظام المبسط لتنظيم المعرفة (SKOS). وأخيراً نقترح أن يتم زيادة العبارات المفتاحية الخاصة بالمكنز الإنجليزي ليتم إثراء ملفات العبارات المفتاحية اليدوية إلى ١٠ عبارات بدلاً من ٤ كما هو الحال الآن.

المراجع

[١] S. R. El-Beltagy and A. Rafea، "KP-Miner: A keyphrase extraction system for English and Arabic documents،" Inf. Syst.، vol. ٣٤، no. ١، pp. ١٤٤–١٣٢، Mar. ٢٠٠٩

[٢] O. Medelyan and I. H. Witten، "Thesaurus based automatic keyphrase indexing،" Proc. ٦th ACM/IEEE-CS Jt. Conf. Digit. Libr. - JCDL '٠٦، p. ٢٠٠٦ ،٢٩٦.

[٣] N. Kumar and K. Srinathan، "Automatic keyphrase extraction from scientific documents using N-gram filtration technique،" Proceeding eighth ACM Symp. Doc. Eng. - DocEng '٠٨، p. ٢٠٠٨ ،١٩٩

[٤] Nhon Do; LongVan Ho، "Domain-specific keyphrase extraction and near-duplicate article detection based on ontology،" in Computing & Communication Technologies - Research، Innovation، and Vision for the Future (RIVF)، ٢٠١٥ IEEE RIVF International Conference on ، vol.، no.، pp.٢٨-٢٥ ،١٢٦-١٢٣ Jan. ٢٠١٥ - doi: ١٠ ,١١٠٩/RIVF.٢٠١٥ ,٧٠٤٩٨٨٦

[٥] T. Nguyen and M. Kan، "Keyphrase extraction in scientific publications،" Asian Digit. Libr. Look. Back ١٠ Years Forg. New Front.، pp. ٢٠٠٧ ،٣٢٦–٣١٧.

[٦] A. Bellaachia and M. Al-Dhelaan، "Learning from Twitter Hashtags : Leveraging Proximate Tags to Enhance Graph-based Keyphrase Extraction،" pp. ٢٠١٢ ،٣٥٧–٣٤٨.

[٧] A. Balagopalan، L. L. Balasubramanian، V. Balasubramanian، N. Chandrasekharan، and A. Damodar، "Automatic keyphrase extraction and segmentation of video lectures،" ٢٠١٢ IEEE Int. Conf. Technol. Enhanc. Educ.، pp. ١٠–١، Jan. ٢٠١٢.

[٨] C. Zhang، H. Wang، Y. Liu، D. Wu، Y. Liao، and B. Wang، "Automatic Keyword Extraction from Documents Using Conditional Random Fields،" vol. ٢٠٠٨ ،٣.

[٩] I. Witten، G. Paynter، and E. Frank، "KEA: Practical automatic keyphrase extraction،" Proc. fourth ACM Conf. Digit. Libr.، pp. ١٩٩٩ ،٢٥٥–٢٥٤.

[١٠] O. Medelyan and I. Witten، "Domain □independent automatic keyphrase indexing with small training sets،" J. Am. Soc. Inf. Sci. Technol.، vol. ٥٩، no. ٧، pp. ٢٠٠٨ ،١٠٤٠–١٠٢٦.

[١١] P. Lopez and L. Romary، "HUMB : Automatic Key Term Extraction from Scientific Articles in GROBID،" no. July، pp. ٢٠١٠ ،٢٥١–٢٤٨.

[١٢] Sakhr، "Sakhr Keywords Extractor." [Online]. Available: http://aramedia.com/keywordextraction.htm. [Accessed: -٠١ Jul٢٠١٢-].

[١٣] T. El-Shishtawy and A. Al-Sammak، "Arabic Keyphrase Extraction using Linguistic knowledge and Machine Learning Techniques،" arXiv Prepr. arXiv١٢٠٣ ,٤٦٠٥، pp. ٢٠١٢ ،٨–١.

[١٤] S. Kim، O. Medelyan، M. Kan، and T. Baldwin، "Semeval٢٠١٠– task ٥: Automatic keyphrase extraction from scientific articles،" Proc. ٥th Int. Work. Semant. Eval.، pp. ٢٠١٠ ،٢٦–٢١.

[١٥] N. Pudota، A. Dattolo، A. Baruzzo، and C. Tasso، "A New Domain Independent Keyphrase Extraction System،" Digit. Libr.، ٢٠١٠.

[١٦] Zefeng Li; Bin He; Yangnan، "Adding Lexical Chain to Keyphrase Extraction،" in Web Information System and Application Conference (WISA)، ١١ ٢٠١٤th ، vol.، no.، pp.١٤–١٢ ،٢٥٧–٢٥٤ Sept. ٢٠١٤ – doi: ١٠ ,١١٠٩/WISA.٢٠١٤ ,٥٣

[١٧] A. Miles and S. Bechhofer، "SKOS simple knowledge organization system reference،" ٢٠٠٩.

[١٨] S. N. Kim، "Dataset for Automatic Keyphrase Extraction from Scientific Articles." [Online]. Available: https://github.com/snkim/AutomaticKeyphraseExtraction. [Accessed: –٠٨Aug٢٠١٢–].

[١٩] Wikipedia، "Arabic Wikipedia." [Online]. Available: http://ar.wikipedia.org/wiki. [Accessed: –٠١Jul٢٠١٢–].

[٢٠] King Abdullah Initiative for Arabic Content، "King Abdullah Initiative for Arabic Content." [Online]. Available: http://www.econtent.org.sa. [Accessed: –٠٧Oct٢٠١٢–].

[٢١] O. Shaaban، "Automatic Diacritics Restoration for Arabic Text،" King Fahd University of Petroleum and Minerals، ٢٠١٣.

[٢٢] V. Tymchenko، "BzReader، an application to browse Wikipedia compressed dumps offline." [Online]. Available: http://code.google.com/p/bzreader/. [Accessed: –٠١Jul٢٠١٢–].

[٢٣] S. Khoja and R. Garside، "Stemming Arabic text،" Lancaster، UK، Comput. Dep. Lancaster Univ.، ١٩٩٩.

[٢٤] M. F. Porter، "An algorithm for suffix stripping، Program،" vol. ١٤، no. ٣، p. ١٩٨٠ ،١٣٧–١٣٠.

[٢٥] J. B. Lovins، "Development of a Stemming Algorithm،" Mech. Transl. Comput. Linguist.، vol. ١١، pp. ١٩٦٨ ،٣١–٢٢.

[٢٦] S. Robertson and K. Jones، "Relevance weighting of search terms،" J. Am. Soc. Inf. Sci.، vol. ٢٧، no. ٣، pp. ١٩٧٦ ،١٤٦–١٢٩.

[٢٧] W. B. CROFT and D. J. HARPER، "Using Probabilistic Models of Document Retrieval Without Relevance Information،" J. Doc.، vol. ٣٥، no. ٤، pp. ٢٩٥–٢٨٥، Dec. ١٩٧٩.

[٢٨] E. Chisholm and T. Kolda، "New term weighting formulas for the vector space method in information retrieval،" Comput. Sci. Math. Div. Oak Ridge Natl. Lab.، ١٩٩٩.