أتمتة التكشيف الموضوعي في مكتبة (ZBW): كيف تتحول نتائج البحث إلى ممارسة دائمة

آنا كاسبرجيك



الملخص:

يتناول هذا المقال تجربة مركز لايبنتس للمعلومات الاقتصادية (ZBW) في نقل حلول أتمتة التكشيف الموضوعي من البحث التطبيقي إلى خدمة إنتاجية دائمة. تستعرض المؤلفة مسار الجهود منذ عام 2000، مرورًا بقرار إجراء البحث داخل المكتبة واعتماد البرمجيات مفتوحة المصدر، وصولًا إلى خدمة (AutoSE) المبنية على إطار (Annif)، التي تُثري سجلات قاعدة بيانات (EconBiz) بواصفات مكنز (STW) آليًّا، وتقدم مقترحات للمكشفين البشريين. ويعرض المقال تدفقات البيانات، ومنهج ضمان الجودة القائم على المقاييس والتنبؤ بالجودة ومبدأ «الإنسان في الحلقة»، والمراحل التي لم تكتمل بعد. ويخلص إلى أنه لا توجد بعد أنظمة مفتوحة المصدر جاهزة للاستخدام في أتمتة التكشيف الموضوعي، وأن تجاوز صيغة «المشروع» إلى مهمة دائمة مدعومة بالموارد الكافية أمر جوهري، مع ضرورة استمرار التواصل بين الباحثين ومهندسي البرمجيات وأصحاب المصلحة، وإيجاد حوافز تكافئ نقل نتائج البحث إلى الممارسة.

الكلمات المفتاحية: التكشيف الموضوعي، الأتمتة، تعلم الآلة، الذكاء الاصطناعي، البيانات الوصفية

1. مقدمة - السياق:

يُعد التكشيف الموضوعي، أي الإثراء الدلالي لسجلات البيانات الوصفية بالواصفات، أحد الأنشطة الجوهرية للمكتبات. ومن البديهي أنه نظرًا إلى تكاثر الوثائق الرقمية لم يعد من الممكن وصف كل وثيقة على حدة وصفًا بشريًّا، ولهذا نحتاج إلى استكشاف إمكانات الأتمتة على كل المستويات. ويمكن أن تبدأ الأتمتة بتدابير صغيرة، مثل استخدام نصوص برمجية وإجراءات بسيطة لمعالجة البيانات الوصفية، وأن تمتد حتى استخدام أساليب الذكاء الاصطناعي (Artificial Intelligence)، ولا سيما أساليب مجال تعلم الآلة.

بدأت الجهود الرامية إلى أتمتة عملية التكشيف الموضوعي في (ZBW) منذ عام 2000. وقد أسفر مشروعان مع شركاء خارجيين و/أو باستخدام برمجيات تجارية عن بعض الأفكار حول ما وصلت إليه التقنية في ذلك الوقت، لكنهما أظهرا في الغالب أن الحلول التي قُيّمت لن تكفي لتلبية متطلبات المكتبة، وأنه ما زالت هناك عقبات عدة ينبغي تجاوزها، سواء من حيث جودة المخرجات أو من حيث التنفيذ التقني. ومع ذلك لم يكن التخلي عن هذا المسعى خيارًا مطروحًا، لأن الحاجة إلى الأتمتة صارت مع مرور الوقت أوضح وأكثر إلحاحًا. وأفضت مرحلة إعادة توجيه نحو عام 2014 إلى قرار يقضي بأن يُجرى البحث التطبيقي اللازم منذ ذلك الحين داخل المؤسسة، وأن تُستخدم البرمجيات مفتوحة المصدر وحدها وتُنشأ برمجيات من هذا النوع. ولهذا الغرض أُنشئت داخل المكتبة وظيفة بدوام كامل لمهندس بحثي، مع إمكانية الحصول على درجة الدكتوراه في علوم الحاسوب. وسُميت المرحلة الأولى من الأنشطة بعد إعادة التوجيه هذه «مشروع (AutoIndex)»، واستمرت حتى عام 2018. وبعد تغيير في العاملين عام 2018، رُقّي دور تنسيق أتمتة التكشيف الموضوعي إلى وظيفة دائمة بدوام كامل، وشغلتها متخصصة في علوم الحاسوب حاصلة على تأهيل إضافي في المكتبات.

غير أن الحلول النموذجية الأولية القائمة على تعلم الآلة التي طُوّرت في مشروع (AutoIndex) لم تكن جاهزة بعد للدمج في العمليات الإنتاجية للمكتبة. ولكي يتسنى التصدي لهذا التحدي على النحو الصحيح، أُجريت تعديلات إضافية عدة على المستوى الاستراتيجي، أهمها أن أتمتة التكشيف الموضوعي في (ZBW) لم تعد تُعد مشروعًا، بل أُعلنت مهمة دائمة (أُطلق عليها اسم «(AutoSE)»). وقد دفع ذلك بدوره إلى إطلاق مرحلة تجريبية (بدأت عام 2020، ويُخطط لأن تستمر حتى عام 2024) هدفها نقل نتائج البحث التطبيقي في سياق (AutoSE) إلى خدمة إنتاجية، عبر بناء بنية برمجية مناسبة تتيح التكشيف الموضوعي في الوقت الفعلي باستخدام نماذج (AutoSE) المدرّبة، ودمجها في مسارات العمل الأخرى الخاصة بالبيانات الوصفية في (ZBW). ولتلبية هذه المتطلبات مُنح فريق (AutoSE) وظيفة أخرى بدوام كامل؛ فمنذ بداية المرحلة التجريبية يتألف الفريق من ثلاثة موظفين يغطون أدوار القيادة/التنسيق، والبحث التطبيقي، وتطوير البرمجيات/البنية.

2. البحث التطبيقي والعمليات الإنتاجية:

2.1. البحث التطبيقي - الأساليب:

من منظور تعلم الآلة، يُعد التكشيف الموضوعي مهمة تُسمى التصنيف متعدد التسميات (multi-label classification)، أي إنه يمكن أن تُسند إلى كل منشور تسميات عدة (تقابل الموضوعات تقريبًا). ومنذ نهاية آخر «شتاء» للذكاء الاصطناعي (نحو عام 2012) ظهرت نماذج متزايدة لتعلم الآلة لهذه المهمة - وهي نماذج قابلة للاستخدام فعلًا! - ويتاح جزء كبير منها برمجياتٍ مفتوحة المصدر. وفي المشروع السابق، (AutoIndex)، طُوّر نموذج أولي للتكشيف الموضوعي الآلي في (ZBW)، يجمع بين أساليب عدة ثم تُرشّح مخرجاتها المجمّعة باستخدام قواعد إضافية (Toepfer & Seifert, 2018a). وفي الوقت نفسه بدأ فريق في المكتبة الوطنية الفنلندية (National Library of Finland, NLF) إنشاء مجموعة الأدوات مفتوحة المصدر (Annif) (Suominen et al., 2023a)، التي تقدم نماذج متنوعة لتعلم الآلة للتكشيف الموضوعي الآلي، وتتيح كذلك دمج نماذج المستخدم الخاصة. وكانت المؤسستان على تواصل، وتبادلتا المعلومات حول ما أنجزته كل منهما.

وفي بداية المرحلة التجريبية اعتمد فريق (AutoSE) إطار (Annif) من أجل الجمع بين عدة نماذج من أحدث ما وصلت إليه التقنية، ويُستخدم منها حاليًّا أربعة: صيغتان ((parabel) و(bonsai)) من (omikuji) (Dong & Suominen, 2022)، وهي خوارزميات لتعلم الآلة قائمة على الأشجار، و(fastText) (Facebook Inc., 2022; Joulin et al., 2016)، الذي

يستخدم تضمينات الكلمات (word embeddings)، و(stwfsa) (ZBW, 2022a)، وهي خوارزمية معجمية قائمة على الآلات ذات الحالات المنتهية (finite-state automata)، طُوّرت في (ZBW) وهي مُحسّنة لمكنز «Standard-Thesaurus Wirtschaft» (STW) (ZBW, 2023)، وهو مكنز مجال الاقتصاد الذي تستضيفه (ZBW) وتستخدمه في التكشيف الموضوعي، لكن يمكن استخدامها مع مفردات أخرى أيضًا. ثم تُدمج مخرجات هذه الأساليب بواسطة أسلوب آخر - هو (nn-ensemble) (Suominen et al., 2023b) - يوازن بينها، فيُنتج في النهاية مجموعة من الموضوعات اجتازت جميعها عتبة ثقة محددة. وتُدرَّب النماذج في (AutoSE) على نصوص قصيرة من سجلات البيانات الوصفية التي تقوم عليها بوابة البحث (EconBiz) التابعة لـ(ZBW) ، وتحديدًا العناوين و(إن توافرت) الكلمات المفتاحية التي يضعها المؤلفون للمنشورات باللغة الإنجليزية. وبالتوازي مع ذلك يستمر البحث التطبيقي لاستكشاف أساليب أخرى لتعلم الآلة تتجاوز الأساليب الكلاسيكية، ومنها مقاربات التعلم العميق (Deep Learning) في صورة نماذج المحوّلات اللغوية الكبيرة (large language transformer models)، ولا سيما المدرّبة مسبقًا منها (ومن أبرز أمثلتها (GPT) (Radford et al., 2018))، وهي واعدة بوجه خاص في التكشيف الموضوعي متعدد اللغات ((Transformer (machine learning model), 2023)).

ويشارك فريق (AutoSE) مشاركة فعالة في التطوير المستمر لإطار (Annif)؛ إذ يتحقق مع المكتبة الوطنية الفنلندية على فترات منتظمة من إمكانية دمج نتائج من سياق (AutoSE) بوصفها وظائفَ جديدة، ويساعد المكتبة الوطنية الفنلندية في تقديم دروس تعليمية، ويساعد مؤسسات أخرى بتقديم المشورة حول كيفية تشغيل (Annif) في الممارسة. وقد استكمل الفريق نسخة (Annif) الخاصة بـ(ZBW) بمكونات من إعداده لإعداد التجارب، وتحسين المعاملات الفائقة (hyperparameter optimisation)، وآليات إضافية متنوعة لضبط الجودة (انظر القسم 3)، وواجهات برمجة تطبيقات (APIs) للتواصل مع مسارات العمل الداخلية والخارجية الخاصة بالبيانات الوصفية.

2.2. العمليات الإنتاجية - تدفقات البيانات:

دخلت نسخة أولى من خدمة (AutoSE) الإنتاجية حيز التشغيل عام 2021. وتعمل البرمجية على عنقود (Kubernetes) مؤلف من خمس آلات افتراضية، وتُستخدم تقنيات مثل (helm) و(GitLab) و(prometheus) و(grafana) لنشر البرمجيات والتكامل المستمر والمراقبة. ومع استمرار البحث التطبيقي وإدماج الفريق مزيدًا من المتطلبات الأصلية والتحسينات التكميلية، تتطور البنية باستمرار، ويحافظ تصميمها القائم على وحدات مستقلة (modular) على قابليتها للتكيف مع التطورات المستقبلية بعد المرحلة التجريبية.

تُستخدم مخرجات الخدمة حاليًّا لغرضين: الأول هو التكشيف الموضوعي الآلي بالكامل للمنشورات باللغة الإنجليزية التي لولا ذلك لما وُصفت بأي موضوعات من مكنز (STW) إطلاقًا. إذ تُفحص قاعدة بيانات (EconBiz) كل ساعة بحثًا عن سجلات بيانات وصفية جديدة مؤهلة، ثم تُثريها (AutoSE) بموضوعات (STW)، وتُعاد كتابتها في قاعدة البيانات فورًا. وإذا تصادف أن المنشور ينتمي إلى المجموعة الأساسية من الإنتاج الفكري المخصصة لأن يصفها المتخصصون البشريون في (ZBW)، فإن موضوعات (AutoSE) تُحجب بعد ذلك في كشاف البحث وفي صفحة العرض المفردة لهذا المنشور، بمجرد إتمام التكشيف الموضوعي البشري. وقد فُعّل الربط بين (AutoSE) وقاعدة بيانات (EconBiz) في يوليو 2021، وفي الأشهر الستة الأولى من التشغيل أُدخل أكثر من 100,000 سجل بيانات وصفية موصوف آليًّا في قاعدة البيانات عبر صلاحية الكتابة المباشرة. والعدد الإجمالي للسجلات التي أثرتها أساليب (AutoSE) في قاعدة البيانات أعلى من ذلك، لأن الفريق يعالج أيضًا كميات كبيرة من السجلات بأثر رجعي،

ثم تُعاد كتابتها في قاعدة البيانات عبر عملية دفعية (batch process) . وحتى ديسمبر 2022 كانت قاعدة بيانات (EconBiz) تضم نحو 1.3 مليون سجل مكشّف موضوعيًّا بواسطة (AutoSE) ، وهو ما يعادل نحو ربع مقتنيات (ZBW).

والغرض الثاني لمخرجات الخدمة هو التكشيف الموضوعي بمساعدة الآلة؛ إذ تُتاح الموضوعات التي تولّدها (AutoSE) مقترحاتٍ عبر واجهة برمجة تطبيقات للمنصة المستخدمة في التكشيف الموضوعي البشري في (ZBW) («Digitaler Assistent (DA-3)»؛ (DA-3)). وكان هذا الربط أول ما نُفّذ، وذلك عام 2020. وداخل (DA-3) تُوسم مقترحات (AutoSE) بأنها مولّدة آليًّا لدواعي الشفافية، ويمكن اعتمادها بنقرة واحدة على زر «إضافة» في أثناء وصف المنشور. وتُخزن السجلات الموصوفة حديثًا في الفهرس الموحد، وتُنسخ مرة أخرى إلى قاعدة بيانات (EconBiz)، حيث يجمعها فريق (AutoSE) ويحسب مقياس (F1) (2023 ,«F-score») من الفرق بينهما، لمراقبة أداء النظام الخلفي (backend) الإنتاجي الحالي (انظر أيضًا القسم 3).

ومن المراحل التي لم تكتمل بعد ضمن المرحلة التجريبية:

- الاستعداد لاستخدام المستخلصات، وربما قوائم المحتويات أيضًا، إلى جانب العناوين والكلمات المفتاحية التي يضعها المؤلفون؛ فإلى جانب جمع القدر اللازم من بيانات التدريب لإجراء التجارب من أجل تطوير نماذج مُحسّنة لهذه الأنواع من المواد النصية، كان على الفريق أيضًا أن يستوضح الوضع فيما يتعلق بحقوق التنقيب في النصوص (text mining) للمستخلصات، لأن معظم الرخص لا تذكر استخدام المستخلصات لأغراض إنتاجية غير تجارية مثل خدمة (AutoSE)، حتى لو كان استخدامها لأغراض البحث مسموحًا به صراحة.
- الاستعداد لدمج حلول للغات غير الإنجليزية، ومن المقاربات المطروحة إجراء ترجمة آلية مسبقة قبل التكشيف الموضوعي، واستخدام نماذج محوّلات شبيهة بـ(BERT) ((Transformer (machine learning model), 2023))، وهو الخيار الأكثر وعدًا.
- إتمام واجهة مستخدم على الويب ونشرها، تقدم عرضًا تفاعليًّا للنظام الخلفي الإنتاجي، وإحصاءات عن الأداء الحالي والسابق لخدمة (AutoSE)، ومعلومات إضافية عن الأساليب التي يتضمنها النظام الخلفي.
- أتمتة عمليات متنوعة لتعلم الآلة، مثل تحسين المعاملات الفائقة والتدريب، ليتسنى إعادة تدريب النماذج بسهولة أكبر عند تراكم قدر كافٍ من سجلات البيانات الوصفية الجديدة أو عند إتاحة نسخة جديدة من مكنز (STW).
- توثيق متطلبات العمليات الإنتاجية (المستقبلية) (انظر أيضًا القسم 4).

ومن الخطط لما بعد المرحلة التجريبية توسيع البنية لدمج مسارات عمل للاستخلاص الآلي للبيانات الوصفية من أجل توليد مزيد من المدخلات لـ(AutoSE)، والجمع بين تعلم الآلة والمقاربات الرمزية، أي إدماج مزيد من المعلومات الدلالية من (STW) ومن مصادر خارجية وسيلةً للتحقق من معقولية مخرجات نماذجنا المدرّبة.

3. إدارة الجودة:

3.1. مقاربات متنوعة لمفهوم ضمان الجودة:

إن أتمتة التكشيف الموضوعي تغيير دفعت إليه الإمكانات التقنية الجديدة، لكنها تؤثر أيضًا في ممارسات التكشيف الموضوعي على المستوى الثقافي. وفي مسعى للأتمتة كهذا يكون ضبط الجودة أمرًا محوريًّا، سواء بسبب الآثار (الإيجابية أو السلبية) لجودة البيانات الوصفية في الاسترجاع، أو لأن رضا أصحاب المصلحة (ولا سيما خبراء التكشيف الموضوعي) عن مخرجات الخدمة أمر حيوي لضمان قبولها واستخدامها على المدى الطويل.

ويعمل فريق (AutoSE) على مفهوم شامل لضمان الجودة يستخدم مقاربات مختلفة، من أجل التمكن من ضمان أعلى جودة ممكنة للتكشيف الموضوعي بوجه عام (انظر أيضًا (Kasprzik) (2022)). ويشمل ذلك من الناحية التقنية العمل بالمقاييس الشائعة الاستخدام في مجال تعلم الآلة (إذ نهدف حاليًّا إلى تعظيم مقياس (F1)، لكننا نخطط لتقييم توليفات مختلفة الأوزان من الإحكام (precision) والاستدعاء (recall)، أو مقاييس أخرى مثل الكسب التراكمي المخصوم المعياري (Normalized Discounted Cumulative Gain)، أو مقاييس تأخذ تراتبية المكنز في الحسبان أيضًا)، وتحديد عتبات معقولة (مثل الحد الأدنى المطلوب من مستوى الثقة). وبعد عملية التكشيف الموضوعي الآلي ذاتها تُطبّق تلك العتبات على المخرجات (إلى جانب مرشحات أخرى وقواعد شرطية من نوع «إذا – فإن»، مثل القوائم السوداء والمطابقات (mappings) ، انظر 3.2). ومنذ عام 2022 صار ضبط الجودة في (AutoSE) يتضمن أيضًا تطبيق مقاربة قائمة على تعلم الآلة للتنبؤ بالجودة الإجمالية للتكشيف الموضوعي لسجل بيانات وصفية معين. وبعبارة أدق، يتنبأ أسلوب (qualle) بالاستدعاء لذلك السجل اعتمادًا على درجات الثقة للموضوعات المفردة، إضافة إلى أساليب استدلالية (heuristics) أخرى، مثل طول النص، والمحارف الخاصة، ومقارنة العدد المتوقع من التسميات بعدد التسميات المقترحة فعلًا. ويستند (qualle) إلى نموذج أولي موصوف في (Toepfer and Seifert) (2018b)، غير أنه لكي يصبح قابلًا للاستخدام في العمليات الإنتاجية كان لا بد من إعادة تنفيذ شفرته البرمجية من الصفر (ZBW, 2022b). وقبل إطلاق (qualle) طلب الفريق من خبراء التكشيف الموضوعي في (ZBW) مراجعة بشرية لمخرجاته للتأكد من أنه سيتفوق على الأسلوب السابق؛ فحتى ذلك الحين كان يُطبّق مرشح دلالي أقل دقة بكثير لضبط الجودة على مستوى سجل البيانات الوصفية (قاعدة «(min2VB)»)، إذ كان على المخرجات أن تتضمن موضوعين على الأقل من أحد المجالين الاقتصاديين الأساسيين، الممثَّلين في (STW) بوصفهما مكنزين فرعيين. ومن ثم فإن هذه القاعدة الاستدلالية تُهمل مجالات أخرى مرتبطة بالاقتصاد، في حين أن (qualle) يتعلم ببساطة من بيانات التدريب الشكل الذي ينبغي أن يكون عليه التكشيف الموضوعي المناسب، دون تمييز بين المكانز الفرعية. ويُظهر ذلك أن الأسلوب القائم على تعلم الآلة، إذا دُرّب على بيانات مناسبة، يمكن أن يكون أكثر مرونة من قاعدة يضعها الإنسان.

3.2. «الإنسان في الحلقة»:

غير أن أحد أهم مكونات ضمان الجودة هو العنصر البشري، وسيظل كذلك. وقد صاغ مجال تعلم الآلة عبارة «الإنسان في الحلقة» (human in the loop) للدلالة على هذا النموذج، الذي يتناول «الطرق الصحيحة لتفاعل البشر وخوارزميات تعلم الآلة من أجل حل المشكلات» (Monarch & Manning, 2021). ومن التفسيرات والتطبيقات الممكنة لذلك:

- كون بيانات التدريب تُوسم عادة بواسطة البشر (وهو ما ينطبق أيضًا على (AutoSE))
- كون نظم تنظيم المعرفة والمطابقات بينها تُنشأ وتُصان عادة بواسطة البشر (وهو ما ينطبق على (STW) أيضًا)
- التكشيف الموضوعي بمساعدة الآلة (مثل المقترحات المولّدة آليًّا في (DA-3)، انظر أعلاه)

- وطرق متنوعة لجمع التغذية الراجعة البشرية لمقاربات مثل التعلم المتتابع (Online Learning) (حيث تعيد الآلة تدريب نفسها مباشرة، على أساس بيانات التغذية الراجعة البشرية مثلًا) والتعلم النشط (Active Learning) (حيث يمكن للآلة أن تطلب من الإنسان تفاعليًّا، عند نقاط معينة، وسومًا أو تقييمات لبيانات مفردة).

وفيما يتعلق بطرق جمع التغذية الراجعة البشرية، استُخدمت استراتيجيات عدة في سياق (AutoSE). ومن هذه الاستراتيجيات إجراء مراجعة بشرية مرة في السنة تقريبًا، يقيّم فيها فريق من خبراء التكشيف الموضوعي في (ZBW) جودة الموضوعات المولّدة آليًّا لعينة من نحو 1,000 منشور، بإسناد أحد أربعة مستويات للجودة إلى كل موضوع على حدة وإلى مجموع الموضوعات للمنشور المعني. وإذا وجد الخبراء أن موضوعًا ما ناقص، أمكنهم إدخاله في الاستمارة كذلك. واستخدم الفريق في هذا النوع من المراجعة واجهة طُوّرت في مشروع (AutoIndex)، تتيح للخبراء عرض البيانات الوصفية ذات الصلة، والوصول إلى النص الكامل عبر رابط، والتنقل بين السجلات المسندة إليهم (ZBW, 2022c). وبعد كل مراجعة عقد الفريق جلسة موسعة لاستخلاص الملاحظات، أمكن فيها للخبراء أيضًا الإبلاغ عن ملاحظات فردية وعن تحيزات لمسوها في مخرجات (AutoSE). وعلى مدى المراجعات الأخيرة ساعد ذلك في تحديد انحرافات منهجية عن النتيجة المرغوبة ومعالجتها؛ فمثلًا، بسبب التمثيل المفرط في بيانات التدريب، كان موضوعا «النظرية» و«الولايات المتحدة الأمريكية» يظهران في المخرجات خطأً أكثر من الموضوعات الأخرى. وكحل مؤقت، صار موضوع «الولايات المتحدة الأمريكية» يُحجب بعد ذلك إذا لم يرد صراحة في العنوان أو في الكلمات المفتاحية التي وضعها المؤلف، في حين يُحجب موضوع «النظرية» إذا ورد في المخرجات (المرشحة) أيضًا موضوع من قائمة موضوعات أكثر تحديدًا تتعلق بالنظريات الاقتصادية، أعدها خبراء المجال. غير أن هذه المقاربة مرهقة وعرضة للخطأ، وينبغي أن تحل محل مثل هذه الحلول قصيرة ومتوسطة المدى تحسيناتٌ في أساليب تعلم الآلة على المدى الطويل.

ومن طرق جمع التغذية الراجعة الأخرى مقارنة مخرجات (AutoSE) بالتكشيف الموضوعي البشري حيثما توافر؛ فكلما أضاف أحد خبراء التكشيف الموضوعي في (ZBW) موضوعات من (STW) إلى سجل بيانات وصفية سبق أن أثرته (AutoSE)، يُخطَر نظام (AutoSE)، ويُحسب مقياس (F1) من الفرق بين مجموعتي الموضوعات. ويمكّن ذلك الفريق، على سبيل المثال، من جمع أدلة على أن نظامًا خلفيًّا جديدًا يؤدي أداءً أفضل من سابقه قبل إطلاقه في العمليات الإنتاجية وذلك بإجراء اختبار (A/B) يُشغّل فيه نظامان خلفيان بالتوازي مدة معينة لمقارنتهما وفق هذا المقياس.

غير أنه مع أن مقياس (F1) مؤشر أداء مقبول في مهام التصنيف، تجدر الإشارة إلى أنه ليس سوى تغذية راجعة «ثنائية» من نوع ما، بمعنى أنه لا يتيح إلا تحديد ما إذا كان موضوع مولّد آليًّا موجودًا أيضًا في المجموعة المولّدة بشريًّا، وما إذا كان موضوع مولّد بشريًّا غائبًا عن المجموعة المولّدة آليًّا، لكنه لا يتيح مثلًا تحديد ما إذا كان موضوع مولّد آليًّا لم يختره المكشّف البشري عامًّا أكثر مما ينبغي فحسب أم خاطئًا تمامًا. ولذلك، ولأن المراجعات السنوية لا تُنتج إلا قدرًا ضئيلًا من بيانات التغذية الراجعة بسبب نقص الموارد البشرية وما يترتب عليه من صغر العينات، تعاون الفريق في مطلع عام 2022 مع مزوّد منصة (DA-3) لدمج حل في (DA-3) يتيح لأخصائيي المكتبات الموضوعيين تقديم تغذية راجعة متدرجة. ونتيجة لذلك صار خبراء التكشيف الموضوعي قادرين الآن على تقديم تقييمات الجودة عبر (DA-3) باستمرار في أثناء عملهم اليومي، دون الحاجة إلى تغيير البرنامج المستخدم، ويُشجَّعون بقوة على ذلك. وكما في المراجعات السابقة، يمكنهم تقييم الموضوعات منفردة ومجموعها على مستوى سجل البيانات الوصفية. أما الموضوعات الناقصة فتُحسب من الفروق بين مقترحات (AutoSE) والتكشيف الموضوعي البشري الذي يدخله الخبراء في السجل. ويمنح القدر الأكبر من بيانات التقييم الذي يُجمع بهذه الطريقة الفريقَ رؤية أفضل

بكثير لأداء (AutoSE) (كما يدركه خبراء التكشيف الموضوعي)، ويمكّنه من تحسين مجموعة أساليبه على نحو أكثر استهدافًا. كما ستُعرض في المستقبل تمثيلات بصرية مولّدة ديناميكيًّا لهذه البيانات عبر واجهة مستخدم على الويب لزيادة الشفافية.

ومن الخطط المستقبلية المتعلقة بتطبيق علاقة أكثر تقدمًا من نوع «الإنسان في الحلقة» استكشاف إمكانية استخدام بيانات التغذية الراجعة هذه في التعلم التدريجي (التعلم المتتابع). ومن المفاهيم المثيرة الأخرى التي تستحق المتابعة مفهوم التعلم النشط (انظر أعلاه). وحتى الآن يمثل التكشيف الموضوعي الآلي والبشري مسارين منفصلين تقريبًا؛ إذ تُستبعد الموضوعات المولّدة آليًّا بمجرد توافر الموضوعات المولّدة بشريًّا (حتى لو كانت الأخيرة مستوحاة من الأولى). ونود أن نستكشف إمكانات نمط أكثر تفاعلية تحل فيه الآلات والبشر مهمة التكشيف الموضوعي معًا، ويستثمر كذلك نقاط قوة كل منهما (على نحو أفضل)؛ فالحلول الآلية ما زالت تُصمم حاليًّا لمحاكاة الحلول البشرية بأكبر قدر ممكن من الدقة، مع أن الآلات قد تكون قادرة على تحديد أنماط وفروق دقيقة حيث تكون القواعد التقليدية للتكشيف الموضوعي البشري أقل تفصيلًا مما ينبغي. وفضلًا عن ذلك، فإنه متى أصبحت العملية مؤتمتة جزئيًّا على الأقل، فقد يمهد ذلك أيضًا الطريق للتحول نحو ممارسات للفهرسة والتكشيف الموضوعي تقوم على الكيانات وعلى العلاقات المصاغة صياغة رسمية بينها، لا على مداخل (قائمة على السلاسل النصية) في قاعدة بيانات، وهو ما سييسر بدوره تطبيق تقنيات دلالية أكثر تقدمًا. وبطبيعة الحال، يجب إخضاع كل هذه المقاربات الممكنة لدراسات مصممة بعناية من حيث جدواها وفائدتها الفعلية، للتأكد من أن التغييرات المقترحة في ممارسات معالجة المعلومات في المكتبة مستدامة ومصممة على نحو بنّاء وفق احتياجات مختلف المستخدمين وأصحاب المصلحة.

4. الدروس المستفادة:

4.1. الالتزام والتواصل:

ثمة عقبات كثيرة ينبغي تجاوزها في نقل نتائج البرمجيات النموذجية الأولية إلى الممارسة - انظر (Battistella et al. (2015) للاطلاع على مراجعة شاملة للإنتاج الفكري حول تحديات نقل التقنية بوجه عام، وكثير منها ينطبق على حالتنا أيضًا.

سمعت كاتبة هذه الورقة مرارًا وتكرارًا مكتبيين (في هذه المؤسسة وغيرها) يعبّرون عن إحباطهم من انطباع مفاده أنه رغم وجود عدد من حلول الأتمتة الواعدة الناتجة عن أنشطة البحث التطبيقي أو غيرها من المشروعات المؤسسية، فإن هذه الحلول نادرًا ما تتحول إلى خدمات إنتاجية يمكنهم استخدامها فعليًّا في عملهم اليومي. وقد يكون أحد أسباب هذا النقل غير المكتمل إلى البنية التحتية للمؤسسة نقصَ الموارد، ولا سيما العاملين ذوي الخبرة المناسبة، سواء للبحث التطبيقي أو لتحويل نتائج البحث إلى برمجيات قابلة للاستخدام. ويبدو أن صناع القرار في المكتبات كثيرًا ما يترددون في الالتزام بمحاولة كبيرة للتحول الرقمي على المدى الطويل، ويفضلون بدلًا من ذلك وصف أنشطة الأتمتة الداخلية التجريبية بأنها «مشروعات»، تجنبًا لتقييد قدر كبير من الموارد (بما فيها وظائف دائمة لموظفين ذوي تأهيل عالٍ) سنوات عديدة. ولهذا لم يكن التحول الرسمي لـ(AutoSE) من وضع المشروع إلى مهمة دائمة مجرد خطوة رمزية، بل كان أساسيًّا لكسر الحاجز أمام إطلاق نسخة أولى من الخدمة فعليًّا، لأنه أعطى إشارة الالتزام اللازم من جميع المعنيين وإليهم. وفي (ZBW) كان اقتناء البرمجيات والقدرة الحاسوبية اللازمة، وإنشاء وظيفة إضافية لمهندس بنية برمجيات، نتيجتين مباشرتين لذلك الالتزام الأولي. ومع ذلك، غني عن القول إن العملية تطلبت، وما زالت تتطلب، إعادة تفاوض مستمرة على الموارد البشرية والمادية، والتعامل مع الظروف الاستراتيجية

والمالية المتغيرة داخل المؤسسة وفي العالم المحيط بها. وفيما يتعلق بالتوظيف، أظهرت تجارب المرحلة التجريبية لـ(AutoSE) أن الأدوار التي يجب شغلها، في الحد الأدنى اللازم، لضمان عمليات إنتاجية دائمة وتطوير مستمر للخدمة هي أدوار التنسيق، والبحث التطبيقي، وتطوير البرمجيات، وإدارة تقنية المعلومات. ويغطي فريق (AutoSE) هذه الأدوار بثلاثة أشخاص في الوقت الحاضر. وتُظهر الممارسات الفضلى أيضًا أنه ينبغي، على أقل تقدير، توزيع النوعين الأخيرين من الخبرة على أكثر من شخص واحد لكي تكون العمليات الإنتاجية مأمونة من الأعطال. وزيادة عدد العاملين في الفريق خيار جذاب بطبيعة الحال (وترف في معظم المؤسسات)، لكن مما يساعد أيضًا أن تكون المنظومة التقنية (technology stack) للمؤسسة ككل متجانسة قدر الإمكان، لأن ذلك ييسر تبادل الخبرات مع الإدارات الأخرى، ولهذا يكتسب التواصل والتنسيق بين الإدارات هذه الأهمية الكبيرة لأنشطة (AutoSE).

غير أن الالتزام ليس مطلبًا لصناع القرار وحدهم. فإذا أُريد للنقل من البحث وصولًا إلى العمليات الإنتاجية أن ينجح، فيجب أن تظل جميع الأطراف المعنية منخرطة حتى تصبح النتائج مرضية. وينطبق ذلك أيضًا على الباحثين، وهو أمر كثيرًا ما يكون إشكاليًّا، لأن الحاجز الفاصل بين العالم الأكاديمي وعالم حالات الاستخدام الواقعية ما زال يبدو عائقًا أمام التعاون المثمر. ويمكن تصنيف عمليات البحث في علوم الحاسوب، تصنيفًا تقريبيًّا جدًّا، إلى نوعين: البحث الأساسي الذي يُنتج نتائج نظرية تظل صحيحة أيًّا كانت طريقة تطبيقها، والبحث التطبيقي الذي يُنتج برمجيات نموذجية أولية يجب بعد ذلك تحويلها إلى خدمات إنتاجية. ويعتمد تنفيذ هذه الخطوة الأخيرة اعتمادًا كبيرًا على سياق التطبيق، وهو ليس طريقًا ذا اتجاه واحد، بل يتطلب دورات كثيرة لتكييف النتيجة مع حالة الاستخدام. ولا تقتصر هذه المراحل اللاحقة على اختبارات علمية إضافية، بل تشمل أيضًا تطوير البرمجيات واختبارها للعمليات الإنتاجية، وإجراء اختبارات قابلية الاستخدام مع المستخدمين المستهدفين لاكتشاف مزيد من المشكلات وضمان القبول، وما إلى ذلك. وفي العادة تكون هذه هي المراحل التي تتعرض فيها عملية النقل للتعثر، ويصاب فيها المعنيون بالإحباط.

وقد يكون أحد أسباب ذلك أن صناع القرار والباحثين على حد سواء قد استهانوا في الماضي بهذه المراحل الأخيرة، ونتيجة لذلك لا يُكافأ النجاح في هذا المجال بالمكانة نفسها التي يُكافأ بها النجاح الأكاديمي. ومن المؤسف أنه لا يوجد حاليًّا حافز حقيقي للباحثين للبقاء منخرطين بعد مرحلة النموذج الأولي؛ بل على العكس من ذلك، فإن الضغط الناجم عن مؤشرات الأداء الرئيسية الأكاديمية («عدد المنشورات»، و«حجم التمويل المكتسب من أطراف ثالثة»، وما إلى ذلك) شديد إلى حد يمنعهم فعليًّا من ذلك. وقد حظي (AutoSE) بتخصيص وظيفة دكتوراه كاملة داخل فريقه، بحيث استطاع ذلك الموظف التركيز على بحثه وعلى سياق التطبيق معًا، ومن ثم تحقيق أكبر قدر ممكن من التآزر، غير أنه عند التعاون مع باحثين آخرين من خارج الفريق ما زالت هذه الأولويات المتعارضة تمثل تحديًا قد يعترض الطريق أحيانًا. ومن الحلول (الجزئية) أن يضع صناع السياسات مؤشرات أداء رئيسية تقيس أنشطة نقل البحث الناجحة وتكافئها (سواء للمؤسسات أو للأفراد المنخرطين في هذه الأنشطة)، وأن يمنحوها الأهمية نفسها التي تحظى بها المؤشرات الأخرى.

ومن المأمول أن يؤدي توجيه مزيد من الاهتمام إلى أهمية هذه المراحل الأخيرة وتعقيداتها إلى ما يلي: أ) أن يوفر صناع القرار موارد بشرية كافية لشغل جميع الأدوار الأساسية (انظر أعلاه)، وب) أن يزداد تدريس المهارات اللازمة لتلك المراحل الأخيرة في أثناء تعليم الباحثين التطبيقيين المستقبليين. ويشمل ذلك المهارات التقنية اللازمة للمشاركة في مشروعات تطوير البرمجيات واسعة النطاق (البرمجة التعاونية، والاختبار، والنشر، وما إلى ذلك)، والمهارات الشخصية اللازمة لإدارة المشروعات والتواصل مع أصحاب المصلحة، بحيث يُزوّد الباحثون بالأدوات اللازمة

لاكتشاف مواضع التحديات العملية والمساعدة في حلها. وخلاصة القول إن جميع الأطراف، من باحثين تطبيقيين ومهندسي برمجيات وأصحاب مصلحة، يجب أن تتبادل باستمرار المتطلبات والمشكلات من أجل إحداث نقل ناجح لنتائج البحث إلى خدمة إنتاجية ملائمة ومجهزة للتشغيل الدائم.

4.2. الخاتمة:

أظهرت تجارب المرحلة التجريبية حتى الآن ما يلي: لا توجد حتى الآن أنظمة أتمتة مفتوحة المصدر جاهزة للاستخدام في التكشيف الموضوعي؛ إذ يجب تكييف البرمجيات الموجودة وصيانتها باستمرار، وهو ما يتطلب أشكالًا متعددة من الخبرة. والانتقال من صيغة المشروع إلى مهمة دائمة يستحق الجهد المبذول؛ فالبحث عن حلول لأتمتة التكشيف الموضوعي وما يتصل به من عمليات أخرى مهمة دائمة ستلازم المكتبات سنوات عديدة قادمة. وبناءً على ذلك، يجب أن تقوم العمليات الإنتاجية المتسقة مع هذه المهمة على مفهوم طويل المدى مُرسّخ بعناية، وأن تصاحبها موارد كافية (العاملون، والبرمجيات، والقدرة الحاسوبية).

ولقد وجدنا أن إجراء البحث التطبيقي وتطوير البرمجيات لـ(AutoSE) داخل الوحدات المكتبية في (ZBW) (لا في إدارة منفصلة للبحث أو لتطوير تقنية المعلومات) مفيد للغاية، لأنه يتيح تعاونًا وتواصلًا وثيقين مع أخصائيي المكتبات الموضوعيين. ومن الضروري إشراك خبراء التكشيف الموضوعي بوصفهم أصحاب مصلحة في العملية، سواء لخبرتهم في مجالات تنظيم المعلومات والمعرفة، أو لزيادة القبول، لأن الشفافية تساعد على تبديد التحفظات وبناء ثقة أساسية في التقنية، ولا سيما في الطرق التي سيستخدمها بها الفريق. ويمكن لتطبيق أساليب الذكاء الاصطناعي أن يساعد المكتبات في رسالتها المستمرة المتمثلة في إعداد مصادر المعلومات وإتاحتها، مع إعادة تشكيل ممارساتها في معالجة المعلومات على نحو جديد. ويقدم مفهوم «الإنسان في الحلقة» مقاربة ممكنة للحفاظ على خبرة التكشيف الموضوعي البشري، مع الجمع بينها وبين الأساليب القائمة على تعلم الآلة، ومن ثم نقلها إلى صورة أكثر تكيفًا مع إمكانات أحدث التقنيات المتاحة اليوم.

المراجع:

- Battistella, C., De Toni, A., & Pillon, R. (2015). Inter-organisational technology/knowledge transfer: a framework from critical literature review. The Journal of Technology Transfer, 41, 1195–1234. https://doi.org/10.1007/s10961-015-9418-7
- Bus factor. (2023, July 5). In Wikipedia. https://en.wikipedia.org/w/index.php?title=Bus_factor&oldid=1154218987
- Dong, T., & Suominen, O. (2022). Omikuji [Computer Software]. https://github.com/tomtung/omikuji
- Eurospider Information Technology AG. (2023, January 17). Subject indexing using the DA. https://www.eurospider.com/en/relevancy-product/digital-assistant-da-3
- F-score. (2023, January 17). In Wikipedia. https://en.wikipedia.org/wiki/F-score
- Facebook Inc. (2022). fastText [Computer Software]. https://fasttext.cc/
- German National Library. (2022, October). Automatic Cataloguing System. https://www.dnb.de/EN/Professionell/ProjekteKooperationen/Projekte/KI/ki_node.html

- Grote, C. (2022, May 11). German National Library launched its “cataloguing machine” [Online forum post]. Annif Users. https://groups.google.com/g/annif-users/c/KVQB-hvLrbA/m/I9RwM9EPBgAJ
- Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). Bag of tricks for efficient text classification. arXiv preprint. https://doi.org/10.48550/arXiv.1607.01759
- Kasprzik, A. (2022, July 26–29). Get everybody on board and get going – the automation of subject indexing at ZBW [Conference presentation] 87th IFLA World Library and Information Congress (WLIC), Satellite Meeting: Information Technology – New Horizons in Artificial Intelligence in Libraries, Dublin, Ireland. https://repository.ifla.org/handle/123456789/2047
- Monarch, R.M., & Manning, C.D. (2021). Human-in-the-loop machine learning – active learning and annotation for human-centered AI. Manning Publications. https://livebook.manning.com/book/human-in-the-loop-machine-learning/
- Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. OpenAI. https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf
- Suominen, O., Inkinen J., Virolainen, T., Fürneisen, M., Kinoshita, B.P., Veldhoen, S., Sjöberg, M., Umstein, P., Neatherway, R., & Lehtinen, M. (2023a). Annif [Computer Software]. https://doi.org/10.5281/zenodo.5654173
- Suominen, O., Inkinen J., Virolainen, T., Fürneisen, M., Kinoshita, B.P., Veldhoen, S., Sjöberg, M., Umstein, P., Neatherway, R., & Lehtinen, M. (2023b). nn-ensemble [Computer Software].
- Toepfer, M., & Seifert, C. (2018a). Fusion architectures for automatic subject indexing under concept drift. International Journal on Digital Libraries, 21, 169–189. https://doi.org/10.1007/s00799-018-0240-3
- Toepfer, M., & Seifert, C. (2018b). Content-based quality estimation for automatic subject indexing of short texts under precision and recall constraints. In E. Méndez, F. Crestani, C. Ribeiron, G. David, & J. Correia Lopes (Eds.), Lecture Notes in Computer Science: Vol. 11057. Digital Libraries for Open Knowledge (pp. 3–15). Springer. https://doi.org/10.1007/978-3-030-00066-0_1
- Transformer (machine learning model). (2023, January 17). In Wikipedia. https://en.wikipedia.org/wiki/Transformer_(machine_learning_model)
- ZBW – Leibniz Information Centre for Economics. (2022a). stwfsapy [Computer Software]. https://github.com/zbw/stwfsapy
- ZBW – Leibniz Information Centre for Economics. (2022b). qualle [Computer Software]. https://github.com/zbw/qualle
- ZBW – Leibniz Information Centre for Economics. (2022c). releasetool [Computer Software]. https://github.com/zbw/releasetool
- ZBW – Leibniz Information Centre for Economics. (2023, January 17). STW Thesaurus for Economics. https://zbw.eu/stw/version/latest/about.en.html

- دمج (Omikuji) في (Annif): https://github.com/NatLibFi/Annif/wiki/Backend%3A-Omikuji
- دمج (fastText) في (Annif): https://github.com/NatLibFi/Annif/wiki/Backend%3A-fastText
- دمج (STWFSA) في (Annif): https://github.com/NatLibFi/Annif/wiki/Backend%3A-STWFSA
- بوابة (EconBiz): https://www.econbiz.de/