التعلم الآلي واستخراج البيانات الببليوجرافية من المواد النصية: نموذج مقترح للمواد النصية باللغة العربية[1]

أ. محمد حسين أحمد حسين

مسؤول الأنظمة الآلية بمكتبة الفجيرة الرقمية
Mohamedhussein12397@gmail.com

تاريخ الاستلام: 9 إبريل 2022 **تاريخ القبول: 14 مايو 2022**

المستخلص.

في الآونة الأخيرة ذاع وانتشر مصطلح الذكاء الاصطناعي وتطبيقاته المختلفة، كالتعلم الآلي والتعلم العميق ومعالجة اللغة الطبيعية ورؤية الحاسب الآلي، واستخدم في العديد من القطاعات، ونتج عنه تطوير في الأعمال من حيث الأداء والسرعة والجودة، هذا التطوير امتد أيضًا إلى المكتبات ومراكز المعلومات باعتبارها مؤسسات تقدم خدمات معرفية، ويلقي البحث الضوء على عملية استخراج البيانات الببليوجرافية من مصادر المعلومات وبخاصة المواد النصية التي تتضمن (الكتب والمقالات العلمية)، و يَعِد البحث تشجيعًا لمؤسسات المعلومات ومؤسسات صناعة المعرفة وبالتحديد الناشرين والمكتبات ومراكز المعلومات على تبني استخدام أدوات استخراج البيانات الببليوجرافية، ويوفر النموذج المقترح إطارا عاما لاستخراج البيانات الببليوجرافية من مصادر المعلومات – النصية – العربية، وتسهيل عمل المفهرسين وليس إلغاء دورهم كاملًا، وإن كان من الممكن أن تحجم التقنية من دور المفهرس، وقد تم الاعتماد على المنهج الوصفي التحليلي لتوضيح ماهية الذكاء الاصطناعي وتطبيقاته، والتعريف بماهية الفهرسة الوصفية، وتوضيح دور الناشر في عملية إنشاء التسجيلات الببليوجرافية، والاستفادة من إمكانات التعلم الآلي في استخراج البيانات الببليوجرافية من مصادر المعلومات النصية، وتضمنت الدراسة عرضا لبنية ومكونات النموذج المقترح لاستخراج البيانات الببليوجرافية خروجًا من ذلك بعدد من النتائج والتوصيات.

الكلمات المفتاحية: التعلم الآلي (1)؛ الذكاء الاصطناعي (2)؛ التعلم العميق (3)؛ الفهرسة الوصفية (4)؛ خوارزميات التعلم الآلي (5).

(1) **تم عرضه في المؤتمر العلمي الثاني عشر لقسم المكتبات والوثائق وتقنية المعلومات " ثورة البيانات وتأثيرها على مؤسسات المعلومات العربية: بين الواقع وطموحات المستقبل" في الفترة من 30-31 مارس 2022م.**

1/ **المقدمة**:

في الآونة الأخيرة ذاع وانتشر مصطلح الذكاء الاصطناعي وتطبيقاته المختلفة كالتعلم الآلي والتعلم العميق ومعالجة اللغة الطبيعية وروية الحاسب الآلي، واستخدم في العديد من المجالات ونتج عنه تطوير في الأعمال من حيث الأداء والسرعة والجودة أيضًا، وباعتبار المكتبات أحد أكبر المجالات المعرفية والخدمية كان يجب أيضًا عليها أن تذهب إلى الاستفادة من تطبيقات الذكاء الاصطناعي في وظائف المكتبات المختلفة فظهرت الأنظمة الخبيرة في الخدمات المرجعية، واستخدام التعلم الآلي والتعلم العميق في عملية استخراج البيانات الوصفية والببليوجرافية والتكشيف الآلي والاستخلاص الآلي، وسيركز البحث على استخدام تطبيقات الذكاء الاصطناعي في استخراج البيانات الببليوجرافية أو الوصفية، في محاولة لوضع نموذج مقترح لاستخراج البيانات الببليوجرافية لمصادر المعلومات – النصية– المكتوبة باللغة العربية.

2/1 أهمية البحث :

1. إلقاء الضوء على عملية استخراج البيانات الببليوجرافية من مصادر المعلومات وبخاصة المواد النصية التي تتضمن (الكتب والمقالات العلمية).

2. يَعِد البحث تشجيعًا لمؤسسات المعلومات ومؤسسات صناعة المعرفة وبالتحديد دور النشر والمكتبات ومراكز المعلومات على تبني استخدام أدوات استخراج البيانات الببليوجرافية.

3. يوفر النموذج إطارا عاما لاستخراج البيانات الببليوجرافية من مصادر المعلومات – النصية – العربية.

4. تسهيل عمل المفهرسين وليس إلغاء دورهم كاملًا، وإن كان من الممكن أن تحجم التقنية من دور المفهرس.

3/1 أهداف البحث:

- توضيح دور الناشر في عملية إنشاء التسجيلات الببليوجرافية.
- الاستفادة من إمكانات التعلم الآلي في استخراج البيانات الببليوجرافية من مصادر المعلومات النصية.
- توضيح بنية ومكونات النموذج المقترح لاستخراج البيانات الببليوجرافية.

4/1 مشكلات البحث:

تتمثل مشكلة البحث الرئيسية في تلك الصعوبات التي تواجه المفهرسين في عملية الفهرسة الوصفية:

- عملية الفهرسة تعتبر عملية آلية – بنسبة تصل إلى 90٪ - باستثناء عملية استخراج الموضوعات الرئيسية للكتاب.

- شغلت الفهرسة أذهان المؤسسات الكبرى منذ عقود طويلة، فأصدروا العديد من المعايير، محاولةً منهم للتوصل إلى نوع من التقنين الدولي، إلى أن ظهرت تقنيات الويب وتعقدت الأمور أكثر، فبنية تسجيلة الفهرسة في شكلها الحالي – شكل مارك 21 -لا تواكب محركات البحث، فأصدروا معايير مثل معيار وصف المصادر وإتاحتها "وام RDA " كمعيار وصف للمحتوىContent، ومعيار BibFrame كمعيار تكويد / بنية Structure

- يصدر عن الإنسان – المفهرس- أخطاء جسيمة لا تدركها محركات البحث.

- الوقت المستغرق – من قبل المفهرس- لإنشاء التسجيلات يعتبر كثيرا جدا.

– تكلفة تسجيلة الفهرسة مرتفعة من حيث راتب المفهرس ومنافذ الاستخدام الخاصة بالمفهرسين terminals.

وستتم الإشارة إلى عدة مشكلات أخرى أثناء عرض مكونات النموذج.

5/1 تساؤلات البحث:

– هل الاعتماد على تطبيقات التعلم الآلي يساهم في عملية الفهرسة – تحديدا استخراج البيانات الببليوجرافية– ويطورها؟

6/1 مجال البحث وحدوده:

– الحدود الموضوعية: إمكانية الاستفادة من التعلم الآلي والتعلم العميق في استخراج البيانات الببليوجرافية من مصادر المعلومات النصية المكتوبة باللغة العربية.

– الحدود النوعية: كتب، وبحوث، ومقالات علمية، وأطروحات.

– الحدود اللغوية: المصادر النصية المكتوبة باللغة العربية.

– الحدود المكانية: المصادر النصية المكتوبة باللغة العربية أيًّا كان مكان النشر بالعالم.

7/1 منهج البحث وأدواته:

تم الاعتماد على المنهج الوصفي التحليلي في تحليل ماهية الذكاء الاصطناعي وبالأخص تطبيقات التعلم الآلي والتعلم العميق ودورها في عملية استخراج البيانات الببليوجرافية، ومن ثم وضع نموذج مقترح لاستخراج البيانات الببليوجرافية من مصادر المعلومات النصية المكتوبة باللغة العربية.

تم عمل استبيان لقياس مدى تكرارية عمل الفهرسة الوصفية بواسطة المهنيين. (انظر الصفحات 20–22)

8/1 المراجعة العلمية:

تم الاستناد إلى المصادر التي نشرت منذ عام 2016 حتى 2021 التي تتضمن موضوع استخراج المعلومات الببليوجرافية، وأيضا الدراسات التي تناولت تطبيقات الذكاء الاصطناعي في خدمات المكتبات، وذلك من عدد من قواعد البيانات منها: Sprninger, google scholar, HaL، وتم البحث في القواعد العربية مثل دار المنظومة ولم يكن فيه ذكر للموضوع.

قدمت دراسة قومار وشيشادري (Kumar & Sheshadri, 2019) عرضا لتطبيقات الذكاء الاصطناعي منها: النظم الخبيرة ومعالجة اللغة الطبيعية والتعرف على الأنماط Pattern Recognition والتعلم الآلي والروبوتات، ونظام هاملت HAMLET2، ثم تناولت تطبيقات الذكاء الاصطناعي في خدمات المكتبات الأكاديمية حيث أشارت إلى كيف يمكن الاستفادة من تطبيقات الاصطناعي في خدمات المكتبات الأكاديمية بالاعتماد على نظام خبير، وكذلك في الفهرسة والتصنيف والتكشيف والتزويد، ثم أوضحت كيف يمكن الاستفادة من معالجة اللغة الطبيعية في خدمات المكتبة وكذلك التعلم الآلي والروبوتات، والواجهات الذكية لقواعد البيانات المتاحة على الخط المباشر، فبذلك تشير الدراسة فقط إلى إمكانية الاستفادة من تطبيقات الذكاء الاصطناعي في خدمات المكتبة.

---

(2) وهو نظام تم تطويره من قبل مركز برمينغهام للإنترنت والمجتمع في هارفارد ويستخدم خوارزمية تدعى doc2vec تقوم بإنشاء محاكة تفرق بين مختلف المستندات.

قدمت دراسة (Tkaczyk & Collins & Sheridan & Beel, 2018) تقييما وعرضا لعشر أدوات مفتوحة المصدر تستخدم في استخراج البيانات الببليوجرافية من المقالات والدراسات العلمية (Anystyle-Parser, Biblio, CERMINE, Citation-Citation Parser, GROBID, ParsCit, PDFSSA4MET, ReferenceTagger and Science Parse) وقدمت المقارنة تفوق GROBID ثم CERMINE ثم ParsCit. وتؤكد الدراسة أيضًا أن ضبط نماذج بيانات خاصة بمهمة محددة يؤدي إلى زيادة في جودة عملية الاستخراج.

تناولت الورقة البحثية (Khemakhem & Foppiano & Romary, 2017) تجربة تكويد المعاجم واستخراجها في قالب رقمي يعرف بمبادرة تشفير النصوص (Text Encoding Initiative) TEI ، بالاعتماد على نظام GROPID (GeneRation Of Bibliographic Data) وهو نظام مفتوح المصدر لتعلم الآلة يقوم باستخراج البيانات الببليوجرافية من المقالات العلمية وبخاصة تلك المستندات النصية ذات تنسيق PDF وتطبيقه على المعاجم، واتبعت في ذلك خوارزمية CRF (Conditional Random Fields) عن طريق الاعتماد على عينتين مختلفتين من القاموس، وعرضت نقاط القوة والقيود التي برزت في التجربة.

وتتناول دراسة (Velden, et al, 2017) الإطار لكيفية وصف وتمييز المناهج أو العمليات التي تعمل على استخراج الموضوع بواسطة استخراج البيانات الببليوجرافية من المنشورات العلمية، ومقارنة الحلول التي توفرها المناهج لاستخراج الموضوعات، وتتم هذه المقارنة دون رجوع إلى حقيقة أساسية، حيث تفترض الدراسة وجهات نظر متعددة ومتساوية الأهمية لتجنب التحيز، وقدمت الدراسة هذه المقارنة من خلال تطبيقه على موضوع الفيزياء الفلكية Astrophysics، وعرضت الدراسة لمجموعة البيانات التي تتبناها في هذه الدراسة The Astro Data Set، وعرض لآلية الاستناد في استخراج الموضوعات – عنونة الموضوع Topic labeling – سواء بالاعتماد على التقنيين أو الاشتقاق وذلك باستخدام مكنز الفلك الموحد (UAT) أو من خلال اللغة الطبيعية من خلال النص ذاته.

تقترح دراسة (Myanak & et al, 2016) إطار عمل مفتوح المصدر ++OCR مصمم لمجموعة متنوعة من مهام استخراج المعلومات من المقالات العلمية بما في ذلك البيانات الوصفية: (العنوان، وأسماء المؤلفين، والانتساب، والبريد الإلكتروني)، وذلك على مجموعة متنوعة من المقالات العلمية المكتوبة باللغة الإنجليزية لفهم أنماط الكتابة العامة وصياغة القواعد لتطوير هذا الإطار المختلط، وتوضح عمليات التقييمات الشاملة أن الإطار المقترح يتفوق على الأدوات الحديثة الموجودة بهامش كبير في استخراج المعلومات الهيكلية إلى جانب تحسين الأداء في مهام استخراج البيانات الوصفية والمراجع، سواء من حيث الدقة (تحسين حوالي 50 ٪)، ووقت المعالجة (حوالي 52 ٪ تحسن)، وأجريت الدراسة تجربة المستخدم بمساعدة 30 باحثًا حيث وجدوا أن هذا النظام مفيد جدًا، كهدف إضافي، وأوضحت الدراسة بنية النظام فهو مكتوب بلغة برمجة PYTHON واستخدام نموذج (CRF(conditional random field

تناولت دراسة (Lajeunesse, 2015) عرضا لحزمة Metagear المضافة إلى R3، حيث عرضت وظائف الحزمة التي تتضمن فرزا وغربلة واستخراج المعلومات الببليوجرافية من أعداد كبيرة من الدراسات العلمية، وتتضمن الحزمة أدوات تقييم جهد الفحص التي تتم عبر العديد من المتعاونين / المراجعين، ويتم تقييم وثوقية هؤلاء المراجعين باستخدام إحصائيات kappa. وتتضمن الحزمة أيضًا إمكانية تنزيل ملفات بامتداد PDF لأتمة استرجاع مقالات المجلات من قواعد البيانات على الإنترنت، وتتم علمية الاستخراج الآلي للبيانات من خلال scatter-plots, box-plots and bar-plots. وأيضًا تدعم الحزمة مخططات تدفق PRISMA.

(3) **هي لغة برمجية تستخدم في الحوسبة الإحصائية والرسومات الجرافيك، وهي معروفة بين الإحصائيين ومنقبي البيانات** Data miners **لتطوير البرمجيات الإحصائية، وتستخدم أيضًا في تحليل البيانات** Data Analysis .(R -project)

وتعرض دراسة (Tkaczyk & et al, 2015) نظام CERMINE وهو نظام مفتوح المصدر يستخدم في استخراج البيانات الوصفية meta data أو الببليوجرافية من المقالات العلمية، وتعتمد عمليات تنفيذ معظم الخطوات على أساليب تعلم الآلة سواء التي تخضع لإشراف وغير الخاضعة للإشراف، وتقدم الورقة البحثية تخطيط بنية سير العمل الإجمالية وتفاصيل حول تنفيذ الخطوات الفردية، ومقارنة CREMINE مع حلول مماثلة منها GROBID, PDFX, Pdf-Extract, Parscit، وأشارت الدراسة تفوق CRIMINE.

نجد في هذه الدراسات منها ما يشير إلى تطبيقات الذكاء الاصطناعي واستخدامها في خدمات المكتبات، وباقي الدراسات جاءت تطبيقية وتقييمية لعدد من نظم مفتوحة المصدر التي تستخدم في استخراج البيانات الببليوجرافية من مصادر المعلومات النصية المكتوبة باللغة الإنجليزية، ودراسة تناولت مناهج استخراج الموضوعات بالاعتماد على البيانات الببليوجرافية من المصادر الإنجليزية، ومن ثم جاءت الدراسة لتضع نموذجا مقترحا يستند على فكرة استخراج البيانات الببليوجرافية من مصادر المعلومات النصية وبخاصة المكتوبة باللغة العربية.

2/ الذكاء الاصطناعي والبيانات الوصفية / الببليوجرافية Attificial intellegance and Bibligraphic data

إلى أي مدى يمكن أن تصل إمكانات الآلة إلى قدرات البشر؟ استكشف مؤتمر في جامعة دارتموث عام 1956 هذا السؤال والذي أدى إلى صياغة مصطلح الذكاء الاصطناعي (AI)

في الستينيات من القرن العشرين، اهتمت وزارة الدفاع الأمريكية بهذا النوع من العمل وزادت من التركيز على تدريب أجهزة الكمبيوتر على محاكاة التفكير البشري.

ثم توالت الأعوام وتمت إضافة مصطلح التعلم العميق في معجم الذكاء الاصطناعي لتعكس القدرة على تسخير قوة حوسبة Computing جديدة تضيف إلى الذكاء الاصطناعي، وأدرك العامة مصطلح الذكاء الاصطناعي من خلال ألعاب الشطرنج مثل Deep mind. (Thrall and etl, 2018)

1/2 تطبيقات الذكاء الاصطناعي:

سيتم التعرف على أحد أهم تطبيقات الذكاء الاصطناعي وسنركز على التعلم الآلي والتعلم العميق لتعلقهما بشكل أساسي ومباشر بموضوع البحث.

1/1/2 التعلم الآلي والتعلم العميق:

يُعرّف التعلم بأنه: اكتساب المعرفة أو المهارة، في مجال معين.(Saloky & Šeminský, 2005)

هذا التعريف مرتبط بالبشر في علم النفس، تم اقتراح العديد من التعريفات المعممة للتعلم، والكثير منها يفسر التعلم بأنه تغيير في سلوك كائن ما.

1/1/1/2 تعلم الألة / الآلي Machine learning:

الهدف من التعلم الآلي هو تنفيذ مهام جديدة دون تعليمات واضحة من المطورين Developers، حيث يتضمن استخدام التجارب السابقة لعمل التنبؤات وصياغة حلول جديدة للمشاكل ذات الحد الأدنى من التدخل، الإدخال البشري. (Wells III, 2019)

وكان أول تعريف للتعلم الآلي من قبل آرثر صموئيل Arthur Samuel's عام 1959 حيث عرفه على بأنه: مجال الدراسة الذي يمنح أجهزة الحاسب الآلي القدرة على التعلم دون أن تكون مبرمجة بشكل صريح.، ثم في عام 1998 اقترح ميتشل – باحث في مجال التعليم الآلي – تعريفًا أكثر دقة من تعريف آرثر، حيث اقترح أحد برامج الحاسب الذي يقوم بالتعلم بالاعتماد على التجربة مشيرًا إليها بحرف E والمهام المطلوب القيام بها T وقياس أداء تلك المهام P، إذا تحسن أداءه في T وقياسه بواسطة P تتحسن التجربة E

خوارزميات التعلم الآلي:

لأجل تحقيق هدف التعلم الآلي يجب وصف الخوارزميات التي سيعتمد عليها نموذج التعلم، وتنقسم هنا آليات عمل الخوارزميات إلى: (Puget, 2016)

– التعلم تحت الإشراف Supervised Learning: حيث يتم فيه إعطاء الخوارزمية بيانات التدريب التي تحتوي على الإجابة الصحيحة لكل مثال، ويندرج تحتها نوعان من الخوارزميات:

* الانحدار / التراجع Regression: حيث يجب فيها أن تكون الإجابة التي يجب تعلمها قيمة مستمرة، على سبيل المثال: يمكن تغذية الخوارزمية بسجل مبيعات المنازل بسعرها، وتتعلم كيفية تحديد أسعار المنازل.

* التصنيف Classification: حيث يجب فيها أن تكون الإجابة التي يجب تعلمها واحدة من العديد من القيم المحتملة، على سبيل المثال بطاقة الائتمان، يجب أن تتعلم الخوارزمية كيفية العثور على الإجابة الصحيحة بين "الاحتيال" و "صادق"، عندما يكون هناك اثنان فقط من القيمة المحتملة نقول إنها مشكلة تصنيف ثنائي.

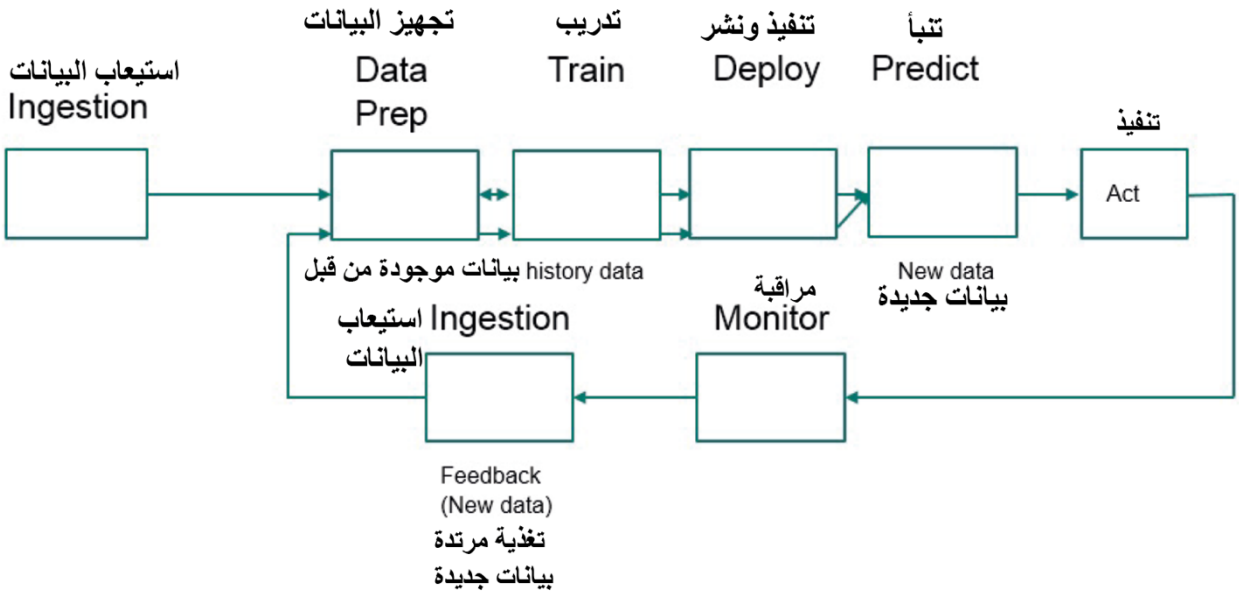
– التعليم غير الخاضع للإشراف Unsupervised Learning: تبحث فيها الخوارزمية عن بنية بيانات التدريب، مثل: البحث عن الأمثلة المتشابهة مع بعضها البعض وتجميعها في مجموعات، يندرج تحتها نوعان من الخوارزميات:

* التجزئة Segmentation: حيث تكون البنية المراد تعلمها عبارة عن تكوين مجموعات clusters متشابهة على سبيل المثال: يهدف تجزئة السوق إلى تجميع العملاء في مجموعات من الأشخاص الذين لديهم سلوك شراء مماثل.

* تحليل الشبكات Network analysis: البنية المراد تعلمها هي معلومات حول أهمية العُقّد nodes ودورها في الشبكة على سبيل المثال: تقوم خوارزمية ترتيب الصفحات بتحليل الشبكة المصنوعة من صفحات الويب وارتباطاتها التشعبية. (Puget, 2016)

سير عمل التعلم الآلي:

تبدأ خطوات التعلم الآلي بتحديد البيانات والحصول عليها Ingestion، ثم عمل تجهيزها وإعدادها للتدريب Data Prep، ثم تأتي مرحلة التدريب train وفقًا لأحد الأساليب سواء أكانت خاضعة للإشراف أو غير خاضعة للإشراف، بعد أن يتم التدريب تنتشر البيانات Deploy ثم يتم التحقق منها Predict وعمل تقييم لها وفقا لقواعد Act ثم مراقبتها في سياقها الجديد Monitor لينتج عنها تغذية مرتدة بإحدى المشاكل أو العيوب التي تحتاج إلى إعادة العملية من جديد.



شكل رقم (1) سير عمل التعلم الآلي Machine Learning Workflow (Puget, 2016)

2/ 1/ 1/ 2 التعلم العميق Deep learning:

يستخدم التعلم العميق شبكات عصبية ضخمة بها العديد من طبقات وحدات المعالجة، مستفيدة من التقدم في قوة الحوسبة وتقنيات التدريب المحسنة لتعلم أنماط معقدة بكميات كبيرة من البيانات، وتتضمن التطبيقات الشائعة التعرف على الصور والكلام. وكانت بداية التعلم العميق في عام 2006، حينما ركزت مشكلة تصنيف الصور المعروفة بمشكلة MNIST. تِقْنِيًّا يُعد التعلم العميق هو التعلم الآلي ووظائفه مماثلة له، ولكن الاختلاف بينهما في القدرات، إذا كانت خوارزمية الذكاء الاصطناعي تُرجع تنبؤً غير دقيق، فيجب على المبرمج التدخل وإجراء التعديلات، ولكن باستخدام نموذج التعلم العميق، يمكن للخوارزمية تحديد ما إذا كان التنبؤ دقيقًا أم لا من خلال شبكتها العصبية الخاصة. (Grossfeld, 2021)

2/1/2 الشبكة العصبية: A neural network (Thompson & Bolen)

الشبكة العصبية هي نوع من التعلم الآلي المستوحى من أعمال الدماغ البشري، إنه نظام حوسبة يتكون من وحدات مترابطة (مثل الخلايا العصبية) يقوم بمعالجة المعلومات من خلال الاستجابة للمدخلات الخارجية، ونقل المعلومات بين كل وحدة، وتتطلب العملية تمريرات متعددة على البيانات للعثور على الاتصالات واستنباط المعنى من البيانات غير المحددة.

3/1/2 رؤية الحاسب: Computer vision (Thompson & Bolen)

تعتمد رؤية الكمبيوتر على التعرف على الأنماط والتعلم العميق للتعرف على ما يوجد في صورة أو مقطع فيديو، عندما تتمكن الآلات من معالجة الصور وتحليلها وفهمها، فيمكنها التقاط الصور أو مقاطع الفيديو في الوقت الفعلي وتفسير محيطها.

4/1/2 معالجة اللغة الطبيعية: NLP (Natural language processing): (Thompson & Bolen)

معالجة اللغة الطبيعية هي قدرة أجهزة الكمبيوتر على تحليل وفهم وتوليد اللغة البشرية، بما في ذلك الكلام.

2/2 الفهرسة / البيانات الوصفية (الببليوغرافي):

تعرف الفهرسة بأنها: العملية التي يتم بمقتضاها توفير الوصول إلى المصادر عن طريق إنشاء مداخل استرجاع وإتاحتها في فهارس، حيث تتضمن عملية الفهرسة عدة عمليات منها: الوصف الببليوغرافي، والتحليل الموضوع وتحديد رموز التصنيف، فهي بذلك أداة للضبط الببليوغرافي وهي أداة للوصول إلى مصادر المعلومات، (الشامي، 2018)، (Reitz, 2014)

يقوم الوصف الببليوغرافي بتحليل وتنظيم العناصر الأساسية – المؤلف والعنوان والتواريخ وغيرها - لمصادر المعلومات، (الشامي، 2018) ، (Reitz, 2014)

ويقابل مصطلح الوصف الببليوغرافي مصطلح الفهارس الوصفية بوصف الكيان المادي أو الملامح المادية لأوعية المعلومات عن طريق تقديم مجموعة من البيانات الببليوجرافية: (اسم المؤلف، العنوان، وبيانات النشر، وغيرها من البيانات التي تهتم بوصف الملامح الخارجية للمصدر المعلومات). (الشامي، 2018)

وتضم عملية الفهرسة عدة معايير منها:

1/2/2 معايير المحتوى: Content standard(حسام الدين، 2011)

تقدم هذه المعايير مجموعة الإرشادات والتعليمات في كيفية فهرسة مصادر المعلومات، حيث تقوم بتحديد نوعيات مصادر المعلومات، وتحديد عناصر البيانات، وتحديد مصادر مصادر الحصول على هذه البيانات، وكيفية صياغتها وترتيبها، وتحديد علامات الترقيم، واللغة التي تكتب بها البيانات، وتحديد مستويات الوصف والمستوى الببليوجرافي، وكيفية تسجيل البيانات التي تتعلق بالعلاقات بين هذه المصادر.

وبدأت هذه المعايير بمعيار Panizzi عام 1841، ثم cutter عام 1876، ومنذ عام 1902 حتى عام 1949 بدء ظهور قواعد الفهرسة الأنجلو أمريكية، ثم نشرت الطبعة الأولى لقواعد الفهرسة الأنجلو أمريكية (1) AACR عام 1967، ثم الطبعة الثانية عام 1978 AACR (2).

إلى أن ظهر معيار وصف المصادر وإتاحتها "وام" RDA، من قبل اللجنة التوجيهية RDA كجزء من خطتها الإستراتيجية (2005-2009) لتحل محل الطبعة الثانية من قواعد الفهرسة الأنجلو أمريكية (2) AACR.

وهي حزمة من عناصر البيانات والإرشادات والتعليمات الخاصة بإنشاء البيانات الوصفية لموارد المكتبات والتراث الثقافي والتي تم تشكيلها بشكل جيد وفقًا للنماذج الدولية لتطبيقات البيانات المرتبطة التي تركز على المستخدم. (rdatoolkit, 2016)

وتعتبر الركيزة الأساسية التي تم الاعتماد عليها في بناء المعيار هي الوثيقة الصادرة عام 2009 بعنوان: " بيان المبادئ العالمية للفهرسة"، ويشتمل البيان على الأساس النظري المنطقي في عائلة 2010 – 1998 4FRBR (المتطلبات الوظيفية للتسجيلات

(4) **"مجموعة من النماذج المفاهيمية التي صيغت في شكل نماذج علاقات بين الكيانات** "ERD" Entity Relationship Diagram

الببليوجرافية FRBR 1998، والمتطلبات الوظيفية للبيانات الاستنادية FRAD 2007، والمتطلبات الوظيفية للبيانات الاستنادية FRSAD 2010.

2/2/2 معايير التكويد / الشكل: Structure standards

هو قالب Form تسكن فيه البيانات، لتتمكن نظم إدارة قواعد البيانات ومحركات البحث التعرف عليها وإجراء عمليات الاختزان والمعالجة والاسترجاع والعرض، (حسام الدين، 2011)

أمثلة عليها:

- الفهرسة المقروءة آلِيًّا (MARC 21 (Machine Readable Cataloging

بدأ مفهوم الفهرسة الآلية مع ظهور مارك - فهو معيار تنسيق رقمي Digital format لوصف العناصر الببليوجرافية التي طورتها مكتبة الكونجرس - في ستينيات القرن الماضي على يد المطور هنرييت أفرام Henriette Avram، في غضون عام 1971 إلى أن أصبح مارك هو المعيار الوطني الأمريكي لنشر البيانات الببليوجرافية، وبعدها بعامين أصبح المعيار الدولي، (Reitz, 2014)

وظهر MARC 21 عام 1999 نتيجة لدمج معايير MARC الأمريكية والكندية .UNIMARC.

- Dubline Core

تم إطلاقه في عام 1995 من خلال ورشة عمل مشتركة بين NCSA و OCLC في مدينة دبلن بأوهايو " OCLC/NCSA Metadata Workshop"، حيث ناقش أكثر من 50 شخصا مجموعة أساسية من الدلالات semantics للمصادر المتاحة في بيئة الويب، وفائدة المعيار تتمثل في عمل تصنيف لمحتويات الويب، ومن ثم تسهل من البحث والاسترجاع. (dublincore.org) ومن ثم يخدم المعيار مفهوم الويب الدلالي5

- مبادرة تشفير النصوص(TEI (The Text Encoding Initiative

هي تكتل consortium غير هادف للربح، مكون من مؤسسات أكاديمية ومشاريع بحثية وعلماء فرديين من جميع أنحاء العالم، يعمل على تطوير وصيانة معيار لتمثيل النصوص في شكل رقمي، مهمتها الرئيسية هي وضع مجموعة من الإرشادات التي تحدد طرق الترميز للنصوص المقروءة آلِيًّا، وبخاصة في العلوم الإنسانية والعلوم الاجتماعية واللغويات، وبدأت فكرتها منذ عام 1987 حتى تم إصدار النسخة الأولى من الإرشادات في عام 1994، وتم استخدام إرشادات TEI على نطاق واسع من قبل المكتبات والمتاحف والناشرين والباحثين الفرديين لتقديم النصوص للبحث عبر الإنترنت والتدريس والحفظ، بالإضافة إلى الإرشادات نفسها، (TEI: Text Encoding Initiative)

---

(5) وهو عبارة عن إضافة المزيد من واصفات البيانات إلى المحتويات والبيانات الموجودة على الويب ومن ثم تحديد هوية كل كيان object / resources **موجود على الويب، يجعل الويب الدلالي أجهزة الحاسب قادرة على تقديم تفسيرات ذات معنى مماثل للطريقة التي يعالج بها البشر المعلومات لتحقيق أهدافهم.**

3/ التعلم الآلي واستخراج البيانات الببليوجرافية:

ظهرت العديد من النظم مفتوحة المصدر التي تقوم باستخراج البيانات الببليوجرافية من المصادر الإلكترونية النصية – وبخاصة الأبحاث، المقالات العلمية – ومنها التالي:

- parsCit
- CERMINE
- Science Parse
- science Parse v2
- Metatagger
- BILBO

– ويعرض البحث نوعين من هذه النظم وهما: GROBID و CERMINE، وبالتحديد هذان النظامان لما لوحظ من خلال الدراسات السابقة ارتقائهما في الترتيب على بقية النظم.

3/ 1 Grobid (GeneRation Of Bibliographic Data)

هو نظام مفتوح المصدر يقوم باستخراج البيانات الببليوجرافية من المقالات والأبحاث / الأوراق العلمية، التي تكون بامتداد من نوع PDF

تستغل الأداة "الحقول العشوائية الشرطية" (CRF)6 ، وهي تقنية للتعلم الآلي لاستخراج المحتوى وإعادة هيكلته تِلْقَائِيًّا من مصادر خام وغير متجانسة إلى مستندات قياسية TEI (مبادرة تشفير النصوص).

GROBID هي مكتبة تعلم آلي Machine learning تستخدم في استخراج وتحليل المستندات الأولية مثل PDF وإعادة هيكلتها في هيكل مبني على XML / TEI بطريقة منظمة، بالتركيز وبشكل خاص على المنشورات التقنية والعلمية، التطورات الأولى بدأت في عام 2008، ثم أصبح الشفرة المصدرية Source Code متاحة منذ عام 2011، كان العمل على GROBID ثابتًا كمشروع جانبي منذ البداية ويتوقع أن يستمر على هذا النحو.

يقدم الوظائف التالية:

– استخراج الرأس Header وعمل Parse7 للمقالات ذات امتداد PDF. يغطي الاستخراج هنا المعلومات الببليوغرافية (مثل: العنوان والمستخلص والمؤلفين والانتساب Affiliation8 والكلمات المفتاحية وغيرها).

---

(6) **تعد إطار عمل احتمالي لتصنيف البيانات وتنظيم وهيكلة البيانات** (hanna, 2005)

(7) **وهي قراءة وتحليل سلسلة المعارف.**

(8) **عرفها الشامي" في الببليومتريقا تعرف بأنها: المؤسسة أو العمل الذي ينتسب إليه الشخص، وتحليل الانتساب يمكن استخدامه لتقويم أو مقارنة المؤسسات أو مجلات الأبحاث".**

– استخراج المراجع وعمل لها Parse من المقالات ذات امتداد PDF ، وتدعم المراجع الموجودة في الحواشي السفلية، على الرغم من أنها لا تزال قيد التقدم، وهي نادرة في المقالات الفنية والعلمية، ولكنها متكررة في المنشورات في العلوم الإنسانية والاجتماعية.

– عملية Parsing للمراجع تكون معزولة عن باقي العمليات.

– عمل Parsing الأسماء، وبخاصة أسماء المؤلفين في رأس المقال Header، وأسماء المؤلفين في المراجع.

– عمل Parsing للانتساب وكتل العناوين.

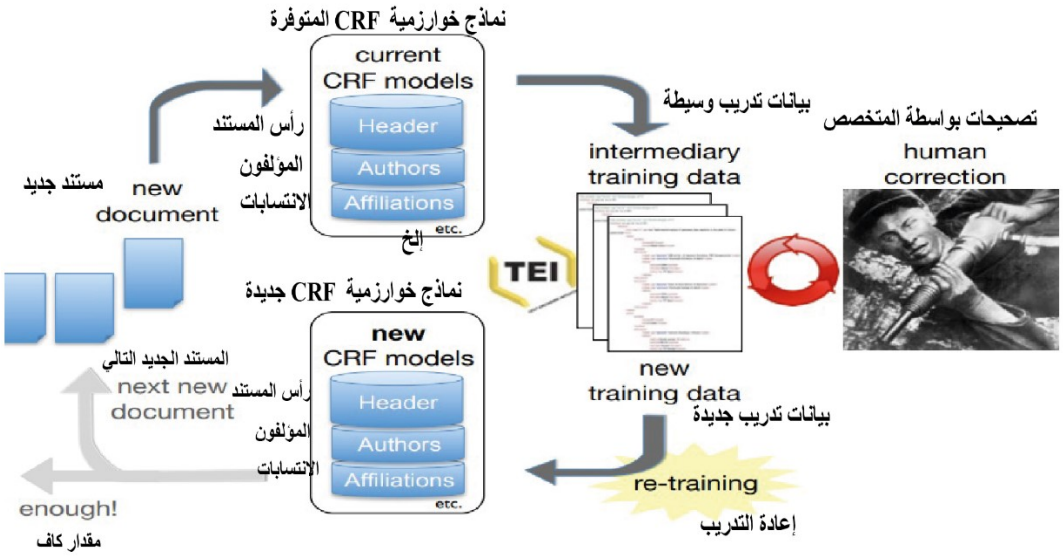
– عمل Parsing التواريخ.

– استخراج النص الكامل من مقالات ذات تنسيق PDF ، في شكل كامل أو مجزأ لهيكلة النص.

يتضمن GROBID معالجة اَلدُّفُعَات، وواجهة برمجة تطبيقات RESTful API شاملة ، وواجهة برمجة تطبيقات JAVA API ، وعدم لاستخدام docker ، وإطار تقييم عام نِسْبِيًّا (الدقة ، الاسترجاع ، إلخ) وإنشاء بيانات التدريب شبه التلقائي.

يمكن اعتبار GROBID مُعدة للإنتاج، لاحتوائها على بيانات مُدربة Datasets من (ResearchGate، HAL Research Archive ، the European Patent Office ، INIST، Mendeley، CERN، Internet Archive ، …).

الشكل التالي يوضح آلية عمل GROBID



الشكل رقم (2) آلية عمل GROBID

CERMINE 2 /3

هي عبارة عن مكتبة برمجية مكتوبة باللغة الجافا Java library مبنية على الويب Web based application تقوم باستخراج البيانات الوصفية Metadata والبيانات الببليوجرافية من المقالات العلمية في شكل رقمي مكود مبني على XML

حيث يقوم النظام بتحليل محتوى ملف بتنسيق PDF ويحاول استخراج معلومات:

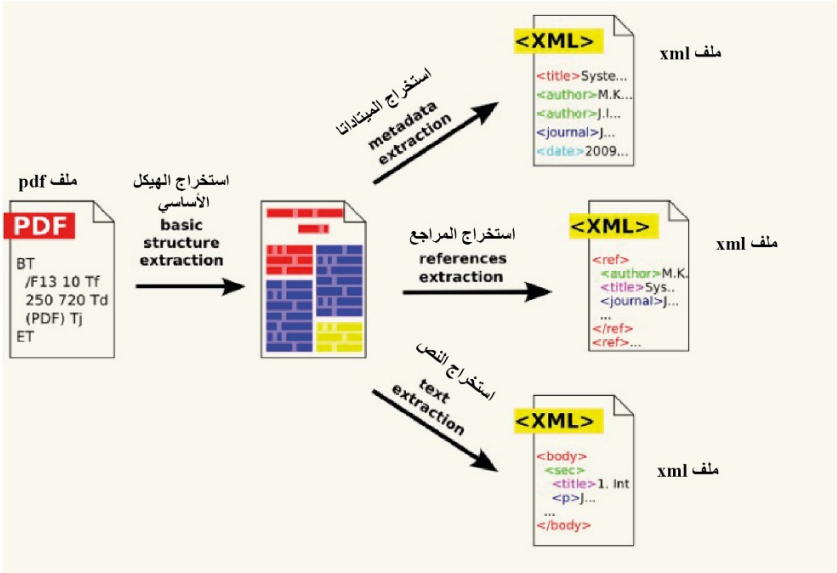
– عنوان المقال.

- معلومات المجلة (العنوان،العدد، وعيره).
- بيانات ببليوجرافية (الطبعة، العدد، عدد الصفحات،).
- المؤلفين والانتسابات affiliations
- الكلمات الدالة / المفتاحية.
- المستخلص.
- المراجع.

يعتمد CERMINE على نمذجة خطوات سير العمل حيث تضمن بنيته architecture إمكانية الحفاظ على خطوات سير العمل الفردية بشكل منفصل، لذلك من السهل إجراء عمليات التقييم والتدريب والتحسين والاستبدال دون تنفيذ خطوة واحدة دون تغيير الأجزاء الأخرى من سير العمل.

تستخدم معظم الخطوات التنفيذية تقنيات التعلم الخاضعة للإشراف supervised وغير الخاضعة للإشراف unsupervised.

الشكل التالي يوضح آلية عمل نظام CERMINE



شكل رقم (3) آلية عمل CERMINE

4/ النموذج المقترح لاستخراج البيانات الببليوجرافية من المصادر العربية النصية:

يقترح هنا نموذجا لاستخراج البيانات الببليوجرافية من المواد النصية العربية، وبخاصة الأبحاث / الأوراق العلمية والكتب والأطروحات.

1/4 أهداف النموذج:

- توفير بيانات وصفية وببليوجرافية للمقالات والأبحاث / الأوراق العلمية ومن ثم توفير وقت ومجهود المكتبات في عمل ذلك، ودعم المكتبات الرقمية وبخاصة التي تحتوي على مواد نصية مكتوبة باللغة العربية.

– وضع منهج Approach لاستخراج البيانات الببليوجرافية من المصادر النصية العربية.

– متضمن في ذلك حث دور النشر والمكتبات وقواعد المعلومات على توفير Datasets وفقًا للمنهج المقترح، لحل مشكلة تنوع في مخططات والأساليب Styles التي تتبعها دور النشر العربية.

2/4 في البداية علينا توضيح عدة أشياء منها الإجابة على سؤال، عملية الفهرسة من أين تبدأ الفهرسة؟

– تتم عملية الفهرسة قبل النشر سواء من قبل الناشر أو المكتبة وتعرف بالفهرسة قبل النشر / أثناء النشر.

في صناعة النشر ظهر مصطلح الفهرسة أثناء النشر CIP (cataloging in publication)ووضع فكرتها winsor jastinعام 1886، ثم بدأت مكتبة الكونجرس عام 1971 باستخدام المصطلح، وعرب المصطلح على يد د. سعد الهجرسي (عبد الهادي، 2010).

وهي أن تقوم دار النشر بالاعتماد على مكتبة أو بالاعتماد على موظفيها في عمل بطاقة فهرسة تضع في الكتاب، وتشمل بيانات الفهرسة أثناء النشر على: (رقم تصنيف ديوي، ورقم تصنيف مكتبة الكونجرس، ورقم تدمك ISBN، والعنوان، والمؤلف، وغيرها).

منذ عام 2007 حل برنامج ECIP (Electronic Cataloging in Publication-ECIP) محل برنامج CIP حيث أصبح الناشرون يقدمون طلبات الفهرسة أثناء النشر إلى مكتبة الكونجرس إِلِكْتِرُونِيًّا.

– بعد النشر من قبل المكتبات / مراكز المعلومات / شركات الفهرسة، وتتم بعد أن تقوم المكتبة بشراء الكتاب من الناشر أو الموزع، من قبل مجموعة أخصائي الفهرسة بالمكتبة.

3/4 آلية عمل المفهرس:

للوقوف على نموذج يضاهي عمل المفهرس وجب أولًا معرفة آلية عمل المفهرس ومعرفة كمْ مقدار تكرار العمل أثناء عمله في مهمة الفهرسة الوصفية.

– استخراج العناصر الموجودة على صفحة العنوان أو الغلاف.

– قراءة موجزة للمقدمة (عند الحاجة).

– تقنين العناصر المستخرجة وفقًا لمعايير الوصف المتبعة سواء ACCR2 أو RDA، واستخدام خطط التصنيف وقوائم رؤوس الموضوعات، ملفات الاستناد(إنشاء ملف استناد، واستخدام ما هو مُعد مسبقًا).

– إدخال البيانات على النظام الفرعي للفهرسة، في أحد قوالب مارك 21 أو Bibframe.

– مراجعة تسجيلة الفهرسة في شكلها النهائي وتدقيقها، والسماح بعرضها في واجهة البحث OPAC.

هل عمل المفهرس وتحديدًا في الفهرسة الوصفية / الببليوجرافية يعتبر عملا روتينيا متكررا؟

للإجابة على هذا السؤال، تم تصميم استبيان بواسطة نماذج جوجل Google forms، معتمدا على الأسلوب المغلق في وضع الأسئلة، وطرحه على المجموعات المتخصصة في المكتبات على منصة فيس بوك، موجه إلى أخصائي الفهرسة، تم طرح الاستبيان لمدة شهر، و قام بالرد تسعة عشر متخصصا فقط.

كم عدد سنوات الخبرة التي عملت فيها في مجال الفهرسة؟

- أقل من خمس سنوات
- أكثر من خمس سنوات
- غير ذلك...

هل العمل كمفهرس مهم أم لا

- مهم
- غير مهم
- غير ذلك...

هل عملك به تكرار وتشعر بالملل؟

- نعم
- لا

في أي مهمة يوجد به تكرار؟ رجاء تحديد في خيار غير ذلك طبيعة التكرار إذا رغبت.

- الفهرسة الوصفية
- التحليل الموضوعي (أختيار رؤوس الموضوعات)
- تحديد أرقام التصنيف
- عمليات أخرى
- غير ذلك...

ملاحظات / تعقيبات / مقترحات لتطوير مهنتك

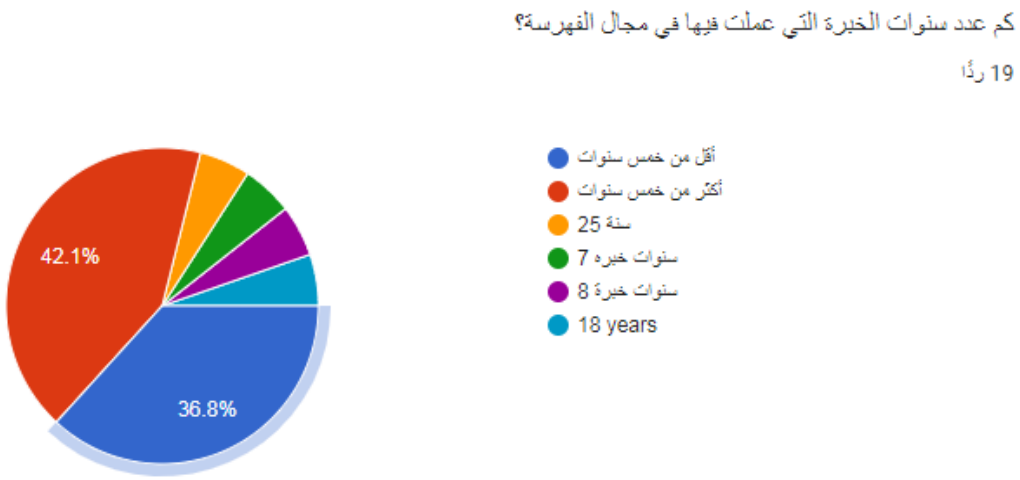
نص الإجابة الطويلة

الهدف من الأسئلة الكشف عن ضرورة البدء في استخدام التعلم الآلي والتعلم العميق في الفهرسة الوصفية.

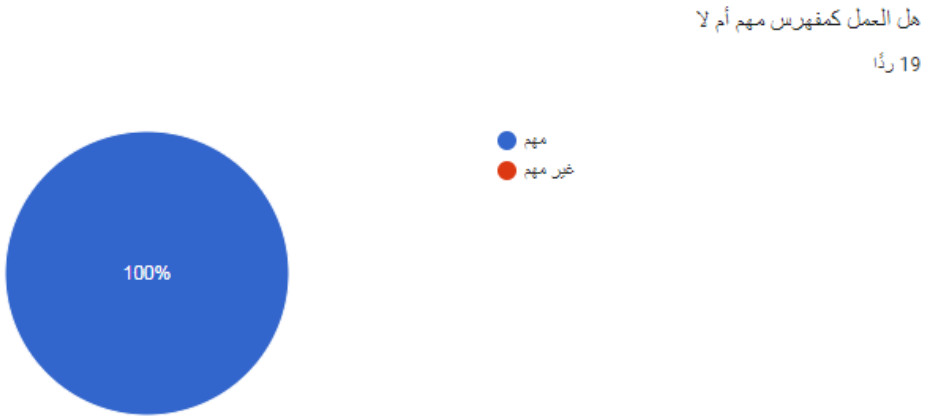
جاءت نتائج الاستبيان كالتالي:

- عدد سنوات الخبرة: (أكثر من خمس سنوات 63.2%، أقل من خمس سنوات 36.8 %).

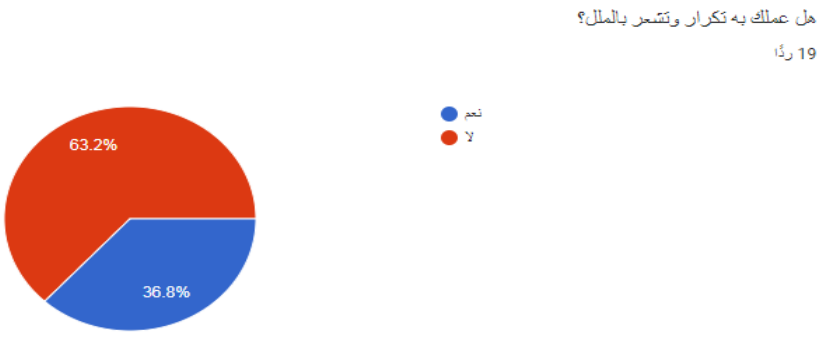
إذاً فعينة الدراسة الأغلبية العظمى لهم باع طويل في العمل في الفهرسة الوصفية.



- هل العمل كفهرس مهم أم لا: مهم % 100، أجمعت عينة الدراسة على أهميتها.

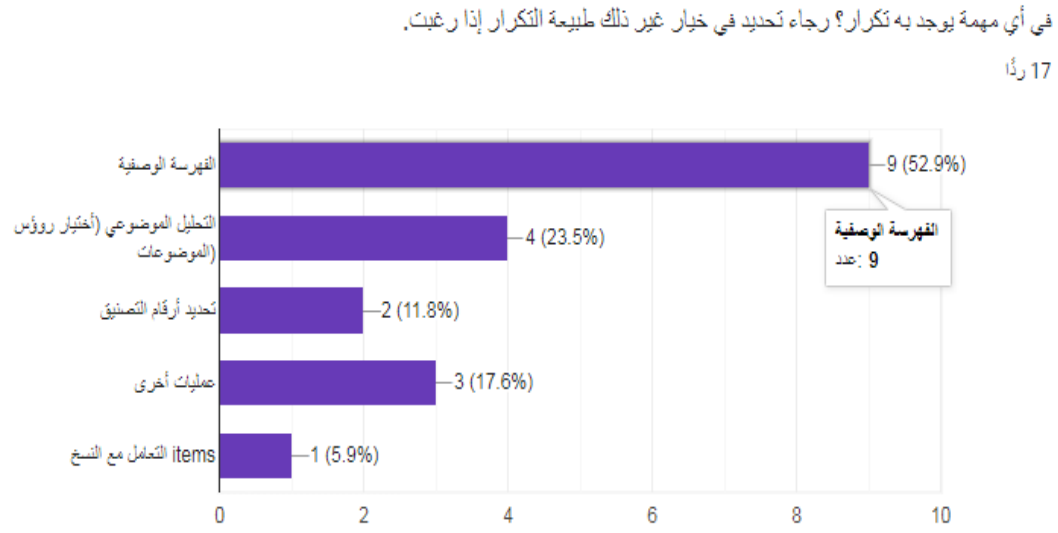


- هل عملك به تكرار وتشعر بالملل: نعم بنسبة % 63.2، لا بنسبة % 36.8.



إذا أغلب المهنيين يشعرون بالملل من تكرار العمل وشعورهم بالملل.

- في أي مهمة يوجد تكرار: قام بالإجابة على هذا السؤال سبعة عشر فقط، وجاءت الفهرسة الوصفية بنسبة % 52.9



أوضحت نتائج الاستبيان أن مهمة الفهرسة الوصفية مهمة تكرارية، تجعل المفهرس يشعر بالملل، ومن ثم حدوث أخطأ في مهام أخرى، التعلم الآلي هدفه الرئيسي القيام بأية مهمة متكررة لتوفير وقت وجهد الإنسان، ليركز على الأعمال الأخرى التي تحتاج إلى جانب إبداعي.

4/4 مكونات النموذج المقترح:

ويُقترح بناء النموذج بالاعتماد على نظام جاهز وعمل تعديل عليه وتدريبه من خلال مجموعات البيانات، ونظام GROPID هو الأفضل لما له من مزايا سبق ذكرها، يتناول البحث مكونات النموذج بالتطبيق على: (كتاب صادر عن قسم الشؤون الفكرية والثقافية في العتبة الحسينية، وبحث علمي منشور بدار المنظومة، وأطروحة دكتوراه صادرة عن قسم المكتبات بكلية الآداب جامعة حلوان.

يتكون النموذج المقترح من عدة خطوات:

4/ 4/ 1 تحديد المخططات: Layouts وأماكن البيانات البيبليوجرافية في الكتب والأبحاث / الأوراق العلمية – المتوفرة بقواعد المعلومات – والأطروحات، حيث يتم تحديد مواقع البيانات بواسطة إحداثيات النقاط Coordinates of a points وهي عبارة عن زوج من الأرقام يحدد موضع نقطة ما على مستوى ثنائي الأبعاد.

– الكتب: توجد البيانات البيبليوجرافية عادة في صحفة العنوان والصفحة التي تليها أو من الممكن أن تتوفر بطاقات فهرسة تم إعدادها قبل / أثناء النشر، ومثال على ذلك كتاب (الإمام موسى بن جعفر الكاظم ورواياته الفقهية، صادر عن قسم الشؤون الفكرية والثقافية في العتبة الحسينية.

(نموذج للبيانات البيبليوجرافية من النص الأصلي)

تحديد Coordinates of a points: مثال تكملة العنوان. (دراسة تحليلية)

<Coords points="174,293 230,293 230,291 238,291 238,294 306,294 306,320 278,320 278,315 209,315 209,321 174,321"/>

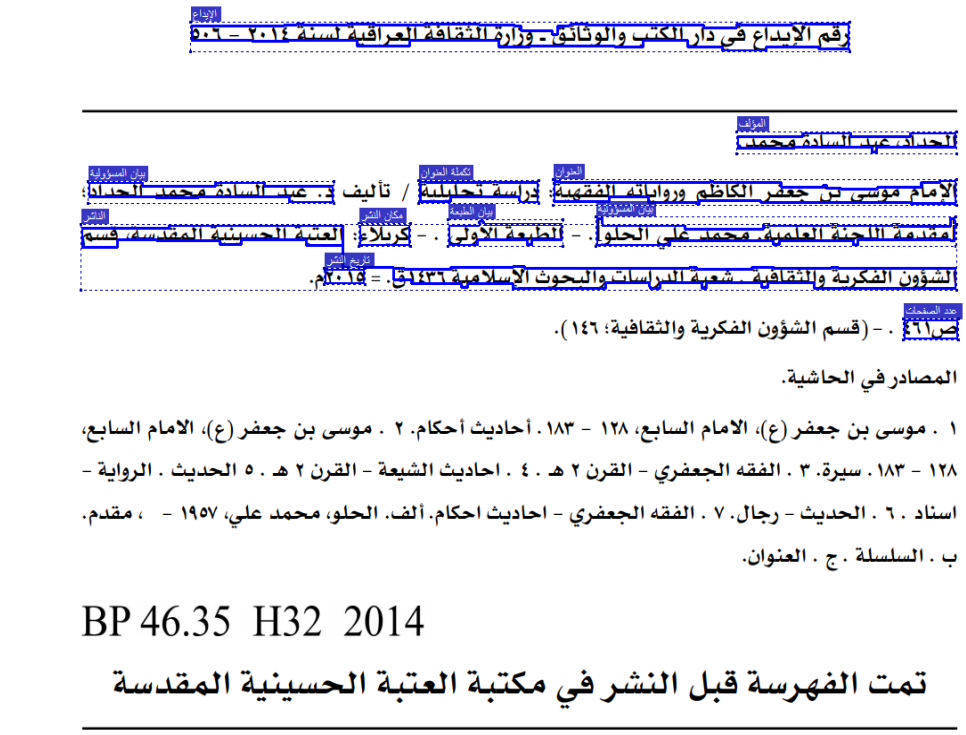
(نموذج للبيانات البيبليوجرافية المُعدة قبل النشر من قبل مكتبة العتبة الحسينية المقدسة)

دراسة تحليلية

عبد الساده محمد المناد

بيان الطبعة
الطبعة الأولى
١٤٣٦هـ ٢٠١٥م

بيان مكان النشر
العراق: كربلاء المقدسة – العتبة الحسينية المقدسة
قسم الشؤون الفكرية والثقافية – هاتف: ٣٢٦٤٩٩
www.imamhussain-lib.com
E-mail: info@imamhussain-lib.com



رقم الإيداع في دار الكتب والوثائق - وزارة الثقافة العراقية لسنة ٢٠١٤ – ٥٠٦

الحداد، عبد السادة محمد

الإمام موسى بن جعفر الكاظم وروايته الفقهية: دراسة تحليلية / تأليف د. عبد السادة محمد الحداد؛ المقدمة اللجنة العلمية محمد علي الحلو. - الطبعة الأولى . - كربلاء: العتبة الحسينية المقدسة، قسم الشؤون الفكرية والثقافية، شعبة الدراسات والبحوث الإسلامية ١٤٣٦هـ. = ٢٠١٥م.

ص٤٦١ . - (قسم الشؤون الفكرية والثقافية؛ ١٤٦).

المصادر في الحاشية.

١ . موسى بن جعفر (ع)، الامام السابع، ١٢٨ - ١٨٣ . أحاديث أحكام. ٢ . موسى بن جعفر (ع)، الامام السابع، ١٢٨ - ١٨٣ . سيرة. ٣ . الفقه الجعفري - القرن ٢ هـ . ٤ . احاديث الشيعة - القرن ٢ هـ . ٥ الحديث . الرواية - اسناد . ٦ . الحديث - رجال. ٧ . الفقه الجعفري - احاديث احكام. ألف. الحلو، محمد علي، ١٩٥٧ - ، مقدم. ب . السلسلة . ج . العنوان.

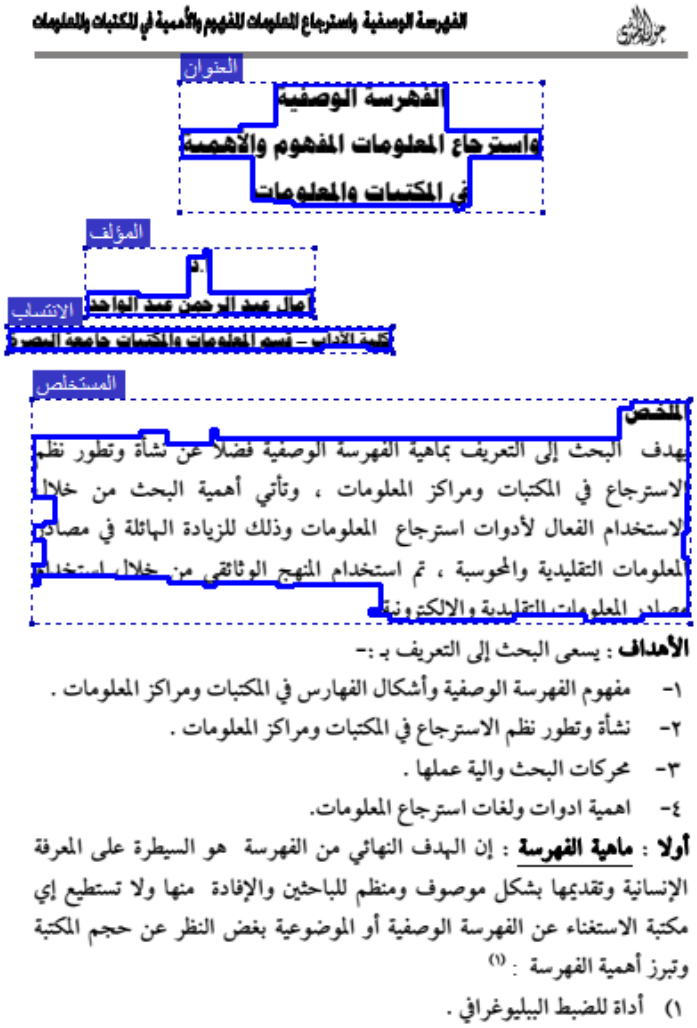
BP 46.35 H32 2014

تمت الفهرسة قبل النشر في مكتبة العتبة الحسينية المقدسة

- الأبحاث / الأوراق العلمية:

توجد البيانات الببليوجرافية عادة في الأبحاث / الأوراق العلمية المنشورة بقواعد المعلومات في رأس الصفحة الأولى من البحث أو من الممكن أن توفر قواعد المعلومات بيانات ببيلوجرافية، مثال على ذلك بحث: (الفهرسة الوصفية واسترجاع المعلومات المفهوم والأهمية في المكتبات والمعلومات) متاح على قواعد معلومات دار المنظومة.

(نموذج للبيانات البيليوجرافية من النص الأصلي)


الفهرسة الوصفية واسترجاع المعلومات المفهوم والأهمية في المكتبات والمعلومات

العنوان

الفهرسة الوصفية
واسترجاع المعلومات المفهوم والأهمية
في المكتبات والمعلومات

المؤلف

أ.د
أمال عبد الرحمن عبد الواحد

الانتساب

كلية الآداب – قسم المعلومات والمكتبات جامعة المنصورة

المستخلص

الملخص

يهدف البحث إلى التعريف بماهية الفهرسة الوصفية فضلاً عن نشأة وتطور نظم الاسترجاع في المكتبات ومراكز المعلومات ، وتأتي أهمية البحث من خلال الاستخدام الفعال لأدوات استرجاع المعلومات وذلك للزيادة الهائلة في مصادر المعلومات التقليدية والمحوسبة ، تم استخدام المنهج الوثائقي من خلال استخدام مصادر المعلومات التقليدية والالكترونية

الأهداف : يسعى البحث إلى التعريف بـ :-

١- مفهوم الفهرسة الوصفية وأشكال الفهارس في المكتبات ومراكز المعلومات .
٢- نشأة وتطور نظم الاسترجاع في المكتبات ومراكز المعلومات .
٣- محركات البحث والية عملها .
٤- اهمية ادوات ولغات استرجاع المعلومات.

أولا : ماهية الفهرسة : إن الهدف النهائي من الفهرسة هو السيطرة على المعرفة الإنسانية وتقديمها بشكل موصوف ومنظم للباحثين والإفادة منها ولا تستطيع إي مكتبة الاستغناء عن الفهرسة الوصفية أو الموضوعية بغض النظر عن حجم المكتبة وتبرز أهمية الفهرسة : [1]

١) أداة للضبط البيليوغرافي .

- ١٢٣ -

تحديد Coordinates of a points: على سبيل المثال المؤلف (أ.د أمال عبد الرحمن عبد الواحد)

<Coords points="86,187 115,187 115,189 143,189 143,167 152,167 152,163 155,163 155,187 209,187 209,186 213,186 213,199 162,199 162,200 86,200"/>

(نموذج للبيانات البيليوجرافية المُعدة قبل النشر من قبل قواعد المعلومات "دار المنظومة")

الأطروحات:

توجد البيانات البيليوجرافية في الأطروحات في الصفحة الأولى والصفحة الثانية، مثال على ذلك: أطروحة دكتوراه بعنوان: (بناء نظام مفتوح المصدر لتحويل ونقل بيانات المكتبات بين النظم الآلية المتكاملة لإدارة المكتبات: دراسة تجريبية) صادرة عن قسم المكتبات والمعلومات بكلية الآداب جامعة حلوان.

تحديد Coordinates of a point: على سبيل المثال مكان النشر (القاهرة)

<Coords points="284,706 340,706 340,729 284,729"/>

2/4/4 تحديد آلية الاستخراج وفقًا لخوارزميات التعلم الآلي:

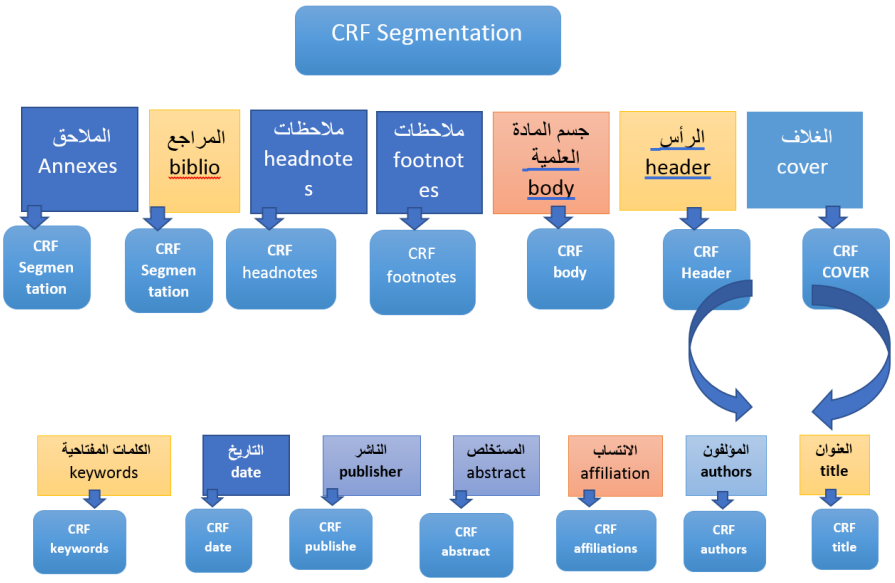
يستخدم GROBID خوارزميات استخراج البيانات بالاعتماد على استخدام مكتبات مثل: Wapiti CRF، أو مكتبة DeLFT للتعلم العميق.

يعتمد منهج التعلم الآلي Machine learning approach في نظام GROBID على 11 نموذجا من CRF، يستخدم كل نموذج نفس النموذج العام المبني على إطار يغطي عملية التدريب والتقييم وفك التشفير وتعيين الرموز المميزة tokenization. كل نموذج له مجموعة من المميزات ومجموعة من البيانات المدربة والمُطبعة normalization.

نموذج التجزئة Segmentation model المعتمد على منهج التعلم الآلي CRF، يقوم بتجزئة كل عنصر وتجزئته إلى عناصر أصغر، قبل البدء في الآلية من الضروري الإشارة إلى عناصر الوصف –سواء من النص الأصلي أو الواردة في بطاقات الفهرسة المُعدة قبل النشر – لكل من الكتب والأبحاث / الأوراق العلمية والأطروحات التي تم استخراجها من المخططات السابقة.

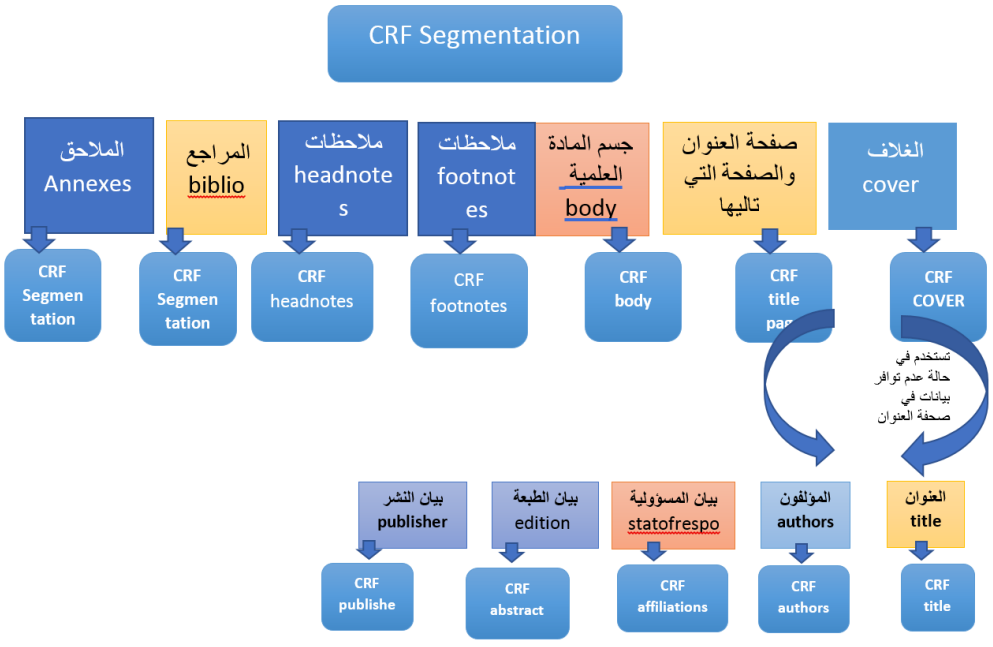
عناصر الوصف (الكتب)	عناصر الوصف (الأبحاث / الأوراق العملية بقواعد البيانات (دار المنظومة))	عناصر الوصف (الأطروحات)
البيانات الببليوجرافية من النص الأصلي		
العنوان	العنوان	العنوان
تكملة العنوان	تكملة العنوان	تكملة العنوان
المؤلف	المؤلف / الانتساب	المؤلف / الانتساب
الناشر	المستخلص	الإشراف / الانتساب
مكان النشر		الناشر
تاريخ النشر		مكان النشر
بيان الطبعة		تاريخ النشر
حقوق النشر		بيان الإطروحة
بيان مكان النشر		المستخلص
		الكلمات المفتاحية
البيانات الببليوجرافية من نموذج قبل النشر		
العنوان	العنوان	
تكملة العنوان	تكملة العنوان	
المؤلف	المؤلف	
بيان المسؤولية	الناشر	
الناشر	العدد	
مكان النشر	تاريخ النشر	
تاريخ النشر	حقوق النشر	
بيان الطبعة	عدد الصفحات	
حقوق النشر	الرابط	
بيان مكان النشر	المصدر (مصدر الحصول على المقالة)	
الإيداع	الموضوع	
عدد الصفحات	بيان التحكيم	
	نوع المحتوى	
	قواعد المعلومات	

الآلية المتبعة في عملية استخراج البيانات الببليوجرافية من الأبحاث / الأوراق العملية بقواعد البيانات والأطروحات، حيث تتشابه إلى حد كبير عناصر البيانات الببليوجرافية ومواقعها في الأبحاث / الأوراق العملية بقواعد البيانات والأطروحات، لذلك من المهم دمج الآلية لتشمل الاثنين معًا.



الشكل رقم (5) يوضح آلية استخراج البيانات الببليوجرافية من الأبحاث / الأوراق العملية بقواعد البيانات والأطروحات

الآلية المتبعة في استخراج البيانات الببليوجرافية من الكتب تم فصلها عن الأوراق العلمية والأطروحات بسبب اختلاف عناصر الفهرسة بها وتنوعها واختلاف مواقعها.

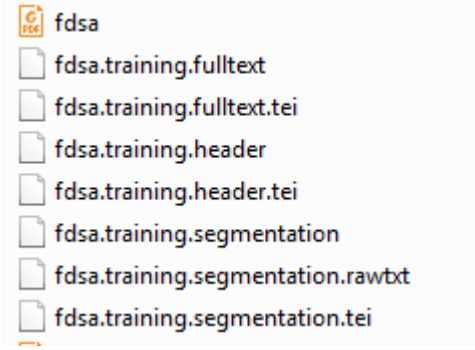


الشكل رقم (6) يوضح آلية استخراج البيانات الببليوجرافية من الكتب

3/4/4 التدريب والتقييم:

بعد أن تم تحديد المخططات وتحديد عناصر البيانات ومواقع البيانات وتم تحديد آلية الاستخراج، يبدأ النموذج في البدء في التدريب، حيث يوفر نظام GROBID إجراء عملية تدريب وتقييم على النموذج العام والنماذج الفرعية .

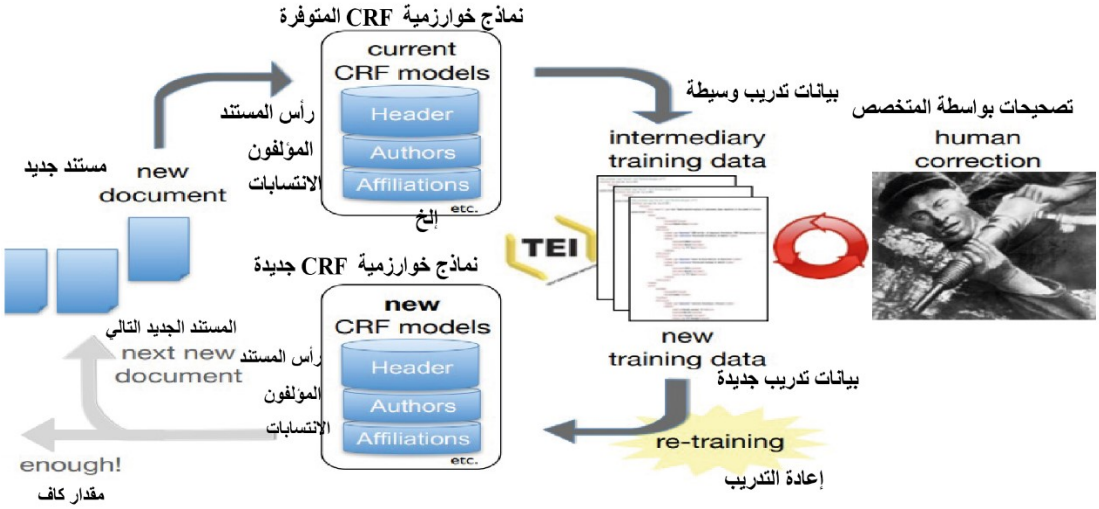
ينتج عن عملية التدريب ملفات متعلقة بكل نموذج ، تم إجراء التدريب ونتج عن ذلك التدريب ملفات، راجع ملحق رقم (3)



4/4/4 إعادة التدريب لما تم استخراجه ومعالجة الأخطاء:

بعد عملية التدريب والتقييم يأتي دور العنصر البشري في مراجعة وتصحيح الأخطاء وتعديلها ومن ثم إدراك الآلة والأخطاء في المرات التالية للتدريب

ويمثل الشكل رقم (7) التالي آلية عمل التدريب.



الشكل رقم (7) التالي آلية عمل التدريب

5/4/4 تصدير البيانات في قالب TEI:

ينتج عن عملية الاستخراج والتدريب بيانات ببليوجرافية في شكل مهيكل بالاعتماد على معيار TEI الذي تم تناول سابقًا، تمت مقابلة للعناصر التي تم استخراجها من النماذج السابقة – للكتب والأوراق / الأبحاث العلمية والأطروحات – مع معيار TEIالمستخدم، راجع الملحق رقم (2).

6/4/4 إدخال البيانات على النظام من خلال مقابلة mappingعناصر TEIمع MARC 21 (RDA) :

تعتبر هذه المرحلة من أهم المراحل التي يتم من خلالها نقل البيانات إلى النظام الآلي المتكامل وبخاصة النظام الفرعي للفهرسة، قبل الشروع لعرض الخطوات سنعرض ماهية نوعية قواعد البيانات المستخدمة في هذه الحالة.

تستخدم أغلب الأنظمة المعتمدة على الذكاء الاصطناعي قاعدة بيانات من نوع NoSQL (Not Only SQL database)، هو approach نهج لتصميم قاعدة بيانات يمكن أن تستوعب مجموعة ضخمة من نماذج البيانات، بما في ذلك تنسيقات key-value المواد النصية document والعمودية columnar والرسم البياني graph. يعد NoSQL، الذي يشير إلى "SQL ليس فقط"، بديلاً لقواعد البيانات الترابطية التقليدية التي يتم فيها وضع البيانات في الجداول، ويتم تصميم مخطط schema البيانات بعناية قبل إنشاء قاعدة البيانات، وقواعد بيانات NoSQL مفيدة بشكل خاص للعمل مع مجموعات كبيرة من البيانات الموزعة. (rouse, 2017)

نموذجنا هنا يعتمد على قاعدة بيانات NOSQL ويستخدم معيار TEI مبنيا على لغة الترميز XML في هيكلة البيانات التي يتم استخراجها ومن ثم حفظها في شكل ملفات.

– الخطوة الأولى فهذه المرحلة هي مقابلة عناصر TEIمع معيار التكويد المستخدم، والمذاع انتشاره حاليًا هو MARC 21 وذلك وفقًا لقواعد وصف محتوى مثل RDA، راجع ملحق رقم (1).

– الخطوة الثانية هي عملية تصدير البيانات:

* من الممكن تحميل الملفات بشكل مباشر على النظام في حالة دعمه لعملية الاستيراد.

* وأيضًا من الممكن الاستفادة من خدامات وجهة برمجة التطبيقات application programming interface (API)، ويتم من خلالها التواصل بين النظام الآلي للمكتبات ونظام التعلم الآلي لاستخراج البيانات الببليوجرافية من مصادر المعلومات النصية المكتوبة باللغة العربية، دون الحاجة إلى نقل ملفات بين النظامين.

يُقترح أن يتبنى نظام GROBID مكتبة Scikit-learn المبنية بلغة بايثون لما يوفره النظام من دعم للغة بايثون بجانب لغة الجافا المستخدمة بشكل أساسي، والتي تعتبر واحدة من مكتبات التعلم الآلي Machine Learning الأكثر شعبية لخوارزميات التعلم الآلي الكلاسيكية، وهي مبنية على مكتبتي NumPy و SciPy. تدعم مكتبة Scikit-Learn معظم خوارزميات التعلم الخاضعة للإشراف supervised وغير الخاضعة للإشراف unsupervised. يمكن أيضًا استخدام Scikit-Learn في استخراج البيانات وتحليل البيانات. (scikit-learn, 2019)

5/ **النتائج والتوصيات**:

خرج البحث بعدد من النتائج والتوصيات.

1/6 النتائج:

– الفهرسة الوصفية عملية تكرارية، أصبح من الضروري الآن في ظل وجود هذه التقنيات أن نفكر في الحد من العمليات التكرارية وإحلال التقنية مكانها.

– عدم توفر مجموعات بيانات Datasets لكافة المواد النصية الصادرة عن دور النشر العربية وبخاصة المصرية.

– تنوع مخططات الكتب يُعد مشكلة في تدريب النموذج.

– لا تنفي هذه الآلية دور المفهرس وإنما تجعله يركز أكثر على المهام الأخرى التي تحتاج إلى إبداع، وأن يساهم بخبرته في بناء مثل هذه النماذج التي تسهل من العمل.

2/6 التوصيات::

– تعاون دور النشر في إمداد النموذج بمجموعات البيانات Datasets (غلاف، صفحة العنوان، النص الكامل "اختياري").

– يساعدنا هذا النموذج – في حالة تبنيه من قبل مؤسسات المعرفة في الوطن العربي – في الوصول إلى ما يعرف بالفهرس المفتوح open catalog خاص بمصادر المعلومات العربية النصية.

– توحيد مخططات layouts الكتب، وإلزام الناشرين بوضع البيانات في الأماكن المتفق عليها.

– أن يتم عمل Hackathon على هوامش مؤتمرات المكتبات والمعلومات المحلية أو الإقليمية أو الدولية لتشجيع المفهرسين والمطورين developers على تطوير مهنة الفهرسة والاستفادة من تطبيقات الذكاء الاصطناعي.

6/ **الملاحق** :

الملحق رقم (1) جدول المقابلة Crosswalk بين عناصر TEI وعناصر MARC 21 (rda)

TEI		MARC 21 (RDA)
<teiHeader xml:lang="__">		040 **$b**
<fileDesc> ├		**n /a**
<titleStmt> ├ |		
<title type="__"> ├ | |	• main • sub • alt • short • desc • translated	• 130 • 210 • 240 • 242 • 245 **ab** • 246 • 247

<author> ├ | |	<author><persName> <author><orgName>	• 100 • 110 • 111 • 700 • 710 • 711
<editor> ├ | |	<persName> <orgName>	• 700 • 710 • 711
<respStmt> ├ | |	<persName> <orgName>	• 700 • 710
<editionStmt> <p> ├ |		250
<publicationStmt> ├ |		**n/a**
<publisher> ├ | |		264_1 **$b**
<distributor> ├ | |		264_2 **$b**
<idno> ├ | |		0285_
<availability> ├ | |	<license>	540
<date when="___" /> └ | |		264_1 **$c**
<seriesStmt> ├ |		**n/a**
<title level="s" type="___"> └ | |	• main • sub • alt • short • desc • translated • filing	• 490 • 8**xx** (**optional**)
<notesStmt> <note> └ |		5**xx**
<sourceDesc> └ |		**n/a**
<biblStruct> ├ |		**n/a**
<analytic> ├ |		**n/a**
<author> ├ | |		**n/a**

<title level="a" type="___"> ├ | |	• main • sub • alt • short • desc • translated • filing	**n /a**
<ptr target="___"> └ | |		**n /a**
<monogr> ├ |		**n /a**
<author> ├ | |	<persName> <orgName>	534 **$a** = 1**st author**
<title level="___" type="___"> ├ | |		534 **$t**
<respStmt> ├ | |		500
<edition> ├ | |		534 **$b**
<imprint> ├ | |		**n /a**
<pubPlace> ├ | | |		534 **$c**
<publisher> ├ | | |		534 **$c**
<date when="___"> *or* <date notBefore="___" notAfter="___"> *or* <date from="___"> └ | | | *or* <date to="___"> *or* <date from="___" to="___">		534 **$c**
<extent> ├ | |		534 **$e**
<note> ├ | |		534 **$n**
<idno> ├ | |		534 **$z for ISBN**
<ptr target="___"> └ | |		856 **$u when** 2**nd indicator** = 2 **and** $3 = **"Source"**
<series> ├ |		534 **$f**
<title level="s"> ├ | |		534 **$f**
<biblScope unit="volume"> ├ | |		534 **$f**
<idno type="ISSN"> ├ | |		534 **$f**
<ptr target="___"> └ | |		534 **$f**

<relatedItem> <biblStruct> └		• 700 **$t** • 710 **$t** • 711 **$t** • 730 • 740
<listRelation> └		**n/a**
<encodingDesc> ├		**n/a**
<projectDesc> <p> ├		
<schemaRef url="___"> ├		856 **$z which should include boilerplate text describing how the TEI document is presented to the user (as page images, text, or both)**
<editorialDecl> ├		**n/a**
<correction status="___" method="___"> ├		**n/a**
<hyphenation eol="___"> ├		**n/a**
<normalization method="___"> ├		**n/a**
<punctuation marks="___" placement="___"> ├		**n/a**
<quotation> ├		**n/a**
<p> ├		• 008/18 • 040 **$e** • 500
<tagsDecl ├		**n/a**
<rendition selector="___" scheme="css"> ├		**n/a**
<namespace name="http://www.tei-c.org/ns/1.0"> <tagUsage> └		**n/a**
<classDecl> <taxonomy xml:id="___"> <bibl> └		500
<samplingDecl> <p> ├		
<appInfo> <app> ├		500

<listPrefixDef> <prefixDef ident="bptl" matchPattern="L([1-5])-v(\d+\.\d+\.\d+[aαβb]?)" replacementPattern="http://www.tei-c.org/SIG/Libraries/teiinlibraries/$2/">		500
<profileDesc>		**n/a**
<langUsage>		**n/a**
<language ident="___">		• 008/35-37 • 041 • 546
<textClass>		**n/a**
<classCode scheme="___">		050-099
<keywords scheme="___"> <term>		6**xx** 2**nd indicator or** 6**xx** $2 **when** 2**nd indicator** = 7
<xenoData>		**n/a**
<revisionDesc> <change when="YYYY-MM-DD" who="[URI]">		**n/a**

الملحق رقم (2)

العنوان	<title> <title type="main">
تكملة العنوان	<title type="sub">
المؤلف (الاسم، الدور، الانتساب) بيان المسؤولية	<author> <persName> <forename> <surname> <roleName> <affiliation> <orgName
بيانات النشر الناشر مكان النشر تاريخ النشر حقوق النشر	**<publicationStmt>** **<publisher>** **<pubPlace>** **<date>** **<availability>** <licencece>
بيان الطبعة الطبعة التاريخ	**<editionStmt>** **<edition>** **<date>**
بيان العدد، المجلد	**<biblScope unit="issue">** **<biblScope unit="volume">**
المحلي / الإيداع الدولي **DOI**معرّف الكيانات الرقمية **URL** محدد موقع الموارد المُوحّد **ISSN**الرقم المعياري الدولي للدوريات	**<idno type="ISBN">** **<idno type="ISSN">** **<idno type="DOI">** **<idno type="URI">**
الوصف المادي عدد الصفحات الابعاد (الطول * العرض) الوصف المادي يمكن عمله تلقائي أثناء عملية الاستخراج دون الحاجة أن يكون مذكور ذلك ضمن النص بشكل مباشر وصريح	**<measure unit="pages" quantity="">** **<dimensions unit="">** **<height>** **<width>**

الملحق رقم (3)

اسم الملف	النموذج
***.training.segmentation.tei.xml**	Segmentation
***.training.header.tei.xml**	Header
***.training.header.affiliation.tei.xml**	affiliation-address
***.training.header.authors.tei.xml**	header
***.training.header.date.xml**	Date
***.training.header-references.xml**	Header
***.training.fulltext.tei.xml**	Fulltext
***.training.figure.tei.xml and *.training.table.tei.xml**	figure, table
***.training.references.referenceSegmenter.tei.xml**	reference-segmenter
***.training.references.tei.xml**	Fulltext
***.training.references.authors.tei.xml**	citation

7/ **المراجع**:

1/7 المراجع العربية:

حسام الدين، مصطفى، (2011)، وصف المصادر وإتاحتها (و م إ : **RDA**) الملامح والبناء والتطبيق في بيئة عربية - **Cybrarians Journal**.- ع 26. تاريخ الاطلاع 15 سبتمبر **2021**. متاح في:
http://www.journal.cybrarians.org/index.php?option=com_content&view=article&id=552:rda&catid=243:2011-08-22-11-46-36&Itemid=79

عبد الهادي، محمد فتحي و جمعة، نبيلة خليفة، (2010)، الفهرسة في البيئة الإلكترونية)،(ط 8)، القاهرة، مصر: الدار المصرية اللبنانية.

2/7 المراجع الأجنبية:

About. (n.d.). Retrieved September 16, **2021, from https://www.rdatoolkit.org/about**

DCMI History. (n.d.). Retrieved September 10, 2021, from https://www.dublincore.org/about/history/

Elshami, A. (2018). Default. Retrieved September 8, **2021, from https://www.elshami.com/**

Grossfeld, B. (2021, January 24). A simple way to understand machine learning vs deep learning. Retrieved September 4, **2021, from https://www.zendesk.com/blog/machine-learning-and-deep-learning/**

III, R. E. W. (2019, December 17). Learn What Machine Learning Is and How It's Changing the World of AI. Retrieved September 15, 2021, from https://www.lifewire.com/what-is-machine-learning-4773696

Khemakhem , M., Foppiano, L., & Romary, L. (2017). Automatic Extraction of TEI Structures in Digitized Lexical Resources using Conditional Random Fields. Hal Archives Ouvertes. Retrieved from https://hal.archives-ouvertes.fr/hal-01508868v2

Lajeunesse, M. J. (2015). Facilitating systematic reviews, data extraction and meta-analysis with themetagearpackage forr. Methods in Ecology and Evolution, 7**(3), 323–330. doi: 10.1111/2041-210x.12472**

Puget, J. F. (2016, May 18). What Is Machine Learning? Retrieved June 4, 2019, from https://web.archive.org/web/20190604140740/https://www.ibm.com/developerworks/community/blogs/jfp/entry/What_Is_Machine_Learning?lang=en

Reitz, J. M. (n.d.). ODLIS. Retrieved September 7, **2021, from https://www.abc-**

clio.com/ODLIS/odlis_b.aspx#bibdescrip

Rouse, M. (2017, March 13). What is NoSQL (Not Only SQL database)? - Definition from WhatIs.com. Retrieved September 29, 2021, from https://searchdatamanagement.techtarget.com/definition/NoSQL-Not-Only-SQL

Saloky, T., & Šeminský, J. (2005). Artificial Intelligence and Machine Learning . Retrieved from http://uni-obuda.hu/conferences/SAMI2005/SALOKY.pdf

scikit-learn. (n.d.). Retrieved September 29, 2021, from https://scikit-learn.org/stable/index.html

Singh, Mayank & Barua, Barnopriyo & Palod, Priyank & Garg, Manvi & Satapathy, Sidhartha & Bushi, Samuel & Ayush, Kumar & Rohith, Krishna & Gamidi, Tulasi & Goyal, Pawan & Mukherjee, Animesh. (2016). OCR++: A Robust Framework For Information Extraction from Scholarly Articles

TEI: Text Encoding Initiative. (n.d.). Retrieved September 28, 2021, from https://tei-c.org/

The Dartmouth Artificial Intelligence Conference: The Next Fifty Years was held at the College. (n.d.). Retrieved September 6, 2021, from http://www.dartmouth.edu/~ai50/homepage.html.

Thompson, W., Li, H., & Bolen, A. (n.d.). Artificial intelligence, machine learning, deep learning and more. Retrieved January 6, 2021, from https://www.sas.com/en_us/insights/articles/big-data/artificial-intelligence-machine-learning-deep-learning-and-beyond.html.

Thrall, J. H., Li, X., Li, Q., Cruz, C., Do, S., Dreyer, K., & Brink, J. (2018). Artificial Intelligence and Machine Learning in Radiology: Opportunities, Challenges, Pitfalls, and Criteria for Success. Journal of the American College of Radiology, 15(3), 504–508. doi: 10.1016/j.jacr.2017.12.026

Tkaczyk, D., Collins, A., Sheridan, P., & Beel, J. (2018, April 19). Machine Learning vs. Rules and Out-of-the-Box vs. Retrained: An Evaluation of Open-Source Bibliographic Reference and Citation Parsers. Retrieved September 13, 2021, from https://arxiv.org/abs/1802.01168

Tkaczyk, D., Szostek, P., Fedoryszak, M., Dendek, P. J., & Bolikowski, Ł. (2015). CERMINE: automatic extraction of structured metadata from scientific literature. International Journal on Document Analysis and Recognition (IJDAR), 18(4), 317–335. doi: 10.1007/s10032-015-0249-8

Velden, T., Boyack, K. W., Gläser, J., Koopman, R., Scharnhorst, A., & Wang, S. (2017). Comparison of topic extraction approaches and their results. Scientometrics, 111(2), 1169–1221. doi: 10.1007/s11192-017-2306-1

Vijayakumar, S., & Sheshadri, K. N. (2019). Applications of Artificial Intelligence in Academic Libraries . International Journal of Computer Sciences and Engineering, 7(16), 136–140. doi: 10.26438/ijcse/v7si16.136140

wallach, hanna m. (2005, May 12). introductionconditional random fields. Retrieved from http://www.inference.org.uk/hmw26/crf/

What is R? (n.d.). Retrieved September 15, 2021, from https://www.r-project.org/about.html

●●●

Machine Learning and Bibliographic Data
Extraction from Textual Information Sources:

a Proposed Model for Text Materials in Arabic[(1)]

Mohamed Hussein Ahmed Hussein

Systems librarian at Fujairah digital library
Mohamedhussein12397@gamil.com

In recent times, the term artificial intelligence and its various applications have spread, such as machine learning, deep learning, natural language processing, and computer vision, and it has been used in many sectors and has resulted in business development in terms of performance, speed and quality. This development has also extended to libraries and information centers as one of the largest institutions that provide knowledge services, and the research sheds light on the process of extracting bibliographic data from information sources, especially text materials that include (books and scientific articles), Where the search is to encourage enterprises and institutions of information industry knowledge, namely publishers, libraries and information centers to adopt the use of bibliographic data extraction tools, The proposed model provides a general framework for extracting bibliographic data from information sources - text - Arabic, And to facilitate catalogers works and not cancel their role fully, Although the technology could limit the role of the catalogers, The researcher relied on the descriptive and analytical approach to explain of what artificial intelligence is and its applications, And the definition of what is descriptive cataloging, clarifying the role of the publisher in the process of creating bibliographic records, and making use of the capabilities of machine learning in extracting bibliographic data from textual information sources, The study included a presentation of the structure and components of the proposed model for extracting bibliographic data out of several results and recommendations.

***Keywords*: *TEI* (1); *CRF* (2); *GROBID* (3); *Datasets* (4); *Artificial Intelligence* (5); *Machine Learning* (6); *Deep Learning* (7); *Descriptive Cataloging* (8); *Machine Learning Algorithms* (9); *Python* (10).**

(1) The research was participated in the twelfth scientific conference of the Department of Libraries Documents and Information Technology, held on March 30-31, 2022.