التعلم الآلي واستخراج البيانات الببليوجرافية من مصادر المعلومات النصية: نموذج مقترح لمصادر المعلومات النصية المكتوبة باللغة العربية

محمد حسين أحمد

إخصائي وثائق وأرشيف، ومطور نظام الوثائق بمركز جمعة الماجد للثقافة والتراث

2020

المستخلص

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، ويلقي البحث الضوء على عملية استخراج البيانات الببليوجرافية من مصادر المعلومات خاصة المواد النصية التي تتضمن (الكتب والمقالات العلمية)، حيث يحدّد البحث تشجيعًا لمؤسسات المعلومات ومؤسسات صناعة المعرفة وبالتحديد دوور النشر والمكتبات ومراكز المعلومات على تبني استخدام أدوات استخراج البيانات الببليوجرافية، ويوفر النموذج أطار عام لاستخراج البيانات الببليوجرافية من مصادر المعلومات – النصية – العربية، وتسهيل عمل المفهرسين وليس إلغاء دورهم كاملًا، وأن كان من الممكن إن تحجم التقنية من دور المفهرس، **ويقدم البحث** توضيح لماهية الذكاء الاصطناعي وتطبيقاته. والتعريف بماهية الفهرسة الوصفية، وتوضيح دور الناشر في عملية إنشاء التسجيلات الببليوجرافية، والاستفادة من إمكانيات التعلم الآلي في استخراج البيانات الببليوجرافية من مصادر المعلومات النصية، بالإضافة إلى عرض لبنية ومكونات النموذج المقترح لاستخراج البيانات الببليوجرافية خروجًا من ذلك بعدد من النتائج والتوصيات

المؤتمر العلمي الثاني عشر لقسم المكتبات والوثائق وتقنية المعلومات
"ثورة البيانات وتأثيرها على مؤسسات المعلومات العربية : بين الواقع وطموحات المستقبل"
بالمكتبة المركزية الجديدة جامعة القاهرة 30-31 مارس 2022م

الكلمات المفتاحية

التعلم الآلي، الذكاء الاصطناعي، الفهرسة الوصفية، TEI، CRF، خوارزميات التعلم الآلي، GROBID، التعلم العميق، Datasets

قائمة المحتويات

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 الباحث على استخدام تطبيقات الذكاء الاصطناعي في استخراج البيانات الببليوجرافية أو الوصفية ليخرج منه بوضع نموذج مقترح لاستخراج البيانات الببليوجرافية أو الوصفية لمصادر المعلومات -النصية- المكتوبة باللغة العربية.

2/1 أهمية البحث

1. يلقي البحث الضوء على عملية استخراج البيانات الببليوجرافية من مصادر المعلومات خاصة المواد النصية التي تتضمن (الكتب والمقالات العلمية).
2. يؤكد البحث تشجيعًا لمؤسسات المعلومات ومؤسسات صناعة المعرفة وبالتحديد دوور النشر والمكتبات ومراكز المعلومات على تبني استخدام أدوات استخراج البيانات الببليوجرافية.
3. يوفر النموذج أطار عام لاستخراج البيانات الببليوجرافية من مصادر المعلومات – النصية – العربية.
4. تسهيل عمل المفهرسين وليس إلغاء دورهم كاملًا، وأن كان من الممكن إن تحجم التقنية من دور المفهرس

3/1 أهداف البحث

يحاول الباحث من خلال البحث أن يتوصل إلى عدة نقاط

- توضيح ماهية الذكاء الاصطناعي وتطبيقاته.
- التعريف بماهية الفهرسة الوصفية

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

4/1 تساؤلات البحث

- ما هو الذكاء الاصطناعي وما هي تطبيقاته؟
- ما هو التعلم الآلي والتعلم العميق واللغة الطبيعية ؟
- ما هو الوصف الببليوغرافي ؟
- هل الاعتماد على تطبيقات التعلم الآلي يساهم في عملية الفهرسة – تحديدا استخراج البيانات الببليوجرافية- ويطورها؟

5/1 مجال البحث وحدوده

الحدود النوعية: أمكانية الاستفادة من التعلم الآلي والتعلم العميق في استخراج البيانات الببليوجرافية من مصادر المعلومات النصية المكتوبة باللغة العربية.

6/1 مشكلات البحث

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

- عملية الفهرسة تتعبر عملية آلية – بنسبة تصل إلى 90% - باستثناء عملية استخراج الموضوعات الرئيسية للكتاب.
- شغلت الفهرسة أذهان المؤسسات الكبرى منذ عقود طويلة، فأصدروا العديد من المعايير، محاولةُ منهم للتوصل إلى نوع من التقنين الدولي، إلى أن ظهرت تقنيات الويب وتعقدت الأمور أكثر فبنية تسجيله 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 كثيراً جدا.
- تكلفة تسجيله الفهرسة مرتفعة من حيث راتب المفهرس ومنافذ الاستخدام الخاصة بالمفهرسين terminals.

سيشير الباحث إلى عدة مشكلات أخرى أثناء عرض مكونات النموذج.

7/1 منهج البحث وأدواته

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

قام الباحث بعمل استبيان لقياس مدى تكرارية عمل الفهرسة الوصفية بواسطة المهنيين.

8/1 المراجعة العلمية

تم الاستناد إلى المصادر التي نشرت منذ عام 2016 التي تتضمن موضوع استخراج المعلومات الببليوجرافية، وأيضا الدراسات التي تناولت تطبيقات الذكاء الاصطناعي في خدمات المكتبات، من عدد من قواعد البيانات منها Sprninger, google scholar, HaL، وقام الباحث بالبحث في القواعد العربية ولم يجد ذكًرا للموضع.

قدمت دراسة قومار وشيشادري (Kumar & Sheshadri, 2019) عرض لتطبيقات الذكاء الاصطناعي منها النظم الخبيرة ومعالجة اللغة الطبيعية والتعرف على الأنماط Pattern Recognition والتعلم الآلي والربوتات، ونظام هاملت HAMLET[1]، 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 لمعالجة اللغة الطبيعية في خدمات المكتبة وكذلك التعلم الآلي والربوتات، والوجهات الذكية لقواعد البيانات المتاحة على الخط المباشر. فبذلك تشير الدراسة فقط إلى إمكانية الاستفادة من تطبيقات الذكاء الاصطناعي في خدمات المكتبة.

قدمت دراسة (Tkaczyk & Collins & Sheridan & Beel, 2018) تقييم وعرض لعشرة أدوات مفتوحة المصدر تستخدم في استخراج البيانات الببليوجرافية من المقالات والدراسات العلمية (Anystyle-Parser, Citation-Parser, GROBID, ParsCit, ,Biblio, CERMINE, Citation 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.

[1] وهو نظام تم تطويره من قبل مركز برمينغهام للإنترنت والمجتمع في هارفارد ويستخدم خوزرامية تدعى doc2vec تقوم بإنشاء محاكة تفرق بين مختلف المستندات.

تناولت الورقة البحثية (Khemakhem & Foppiano & Romary, 2017) تجربة تكويد المعاجم واستخراجها في قالب رقمي يعرف بمبادرة تشفير النصوص TEI (Text Encoding Initiative) ، بالاعتماد على نظام GROPID (GeneRation Of Bibliographic Data) وهو نظام مفتوح المصدر لتعلم الآلة يقوم باستخراج البيانات الببليوجرافية من المقالات العلمية خاصة 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.

وتعرض دراسة (Velden, et al, 2017) الإطار لكيفية وصف وتمييز المناهج أو العمليات التي تعمل على استخراج الموضوع من 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 التقينين 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، أسماء المؤلفين، الانتساب، البريد الإلكتروني)، وذلك على مجموعة متنوعة من المقالات العلمية المكتوبة باللغة الإنجليزية لفهم أنماط الكتابة العامة وصياغة القواعد لتطوير هذا الإطار المختلط. توضح 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 30باحثًا حيث وجدوا أن هذا النظام مفيد جدًا. كهدف إضافي، وأوضحت الدارسة بنية النظام فهو مكتوب بلغة برمجة PYTHON واستخدام نموذج (CRF(conditional random field

تناولت دراسة (Lajeunesse, 2015) عرض لحزمة Metagear المضافة إلى R^2، حيث عرضت وظائف الحزمة التي تتضمن فرز 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 موثوقية هؤلاء المراجعين باستخدام إحصائيات kappa. وتتضمن الحزمة أيضًا أمكانية تنزيل ملفات بامتداد PDF لأتمتة استرجاع مقالات المجلات من قواعد البيانات

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

على الإنترنت. وتتم علمية الاستخراج الآلي للبيانات من خلال scatter-plots, box-plots and bar-plots. وأيضًا تدعم الحزمة مخططات تدفق PRISMA.

تعرض دراسة (Tkaczyk & et al, 2015) نظام CERMINE وهو نظام مفتوح المصدر يستخدم في استخراج البيانات الوصفية meta dataأو الببليوجرافية من المقالات العلمية، وتعتمد عمليات تنفيذ معظم الخطوات على أساليب تعليم الآلية التي تخضع لإشراف وغير الخاضعة للإشراف، تقدم 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 في خدمات المكتبات، وباقي الدارسات 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، ومن ثم جاءت الدراسة لتضع نموذج مقترح يستند على فكرة استخراج البيانات الببليوجرافية من مصادر المعلومات النصية وخاصة المكتوبة باللغة العربية.

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

إلى أي مدى يمكن أن تصل إمكانيات الآلة إلى قدرات البشر؟ استكشف مؤتمر في جامعة دارتموث عام 1956 هذه 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 وتم إضافة مصطلح التعلم العميق في معجم الذكاء الاصطناعي لتعكس القدرة على تسخير قوة حوسبية 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 لتعلقهم 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 على أنه 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 أنه 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 أقترلحد برامج الحاسب يقوم بالتعلم بالاعتماد على التجربة مشير ً ا إليها بحرف E والمهام المطلوب القيام بها T وقياس أداء تلك المهام P، أذا تحسن أدائه في T وقياسه بواسطة P تتحسن التجربة E

خوارزميات التعلم الآلي

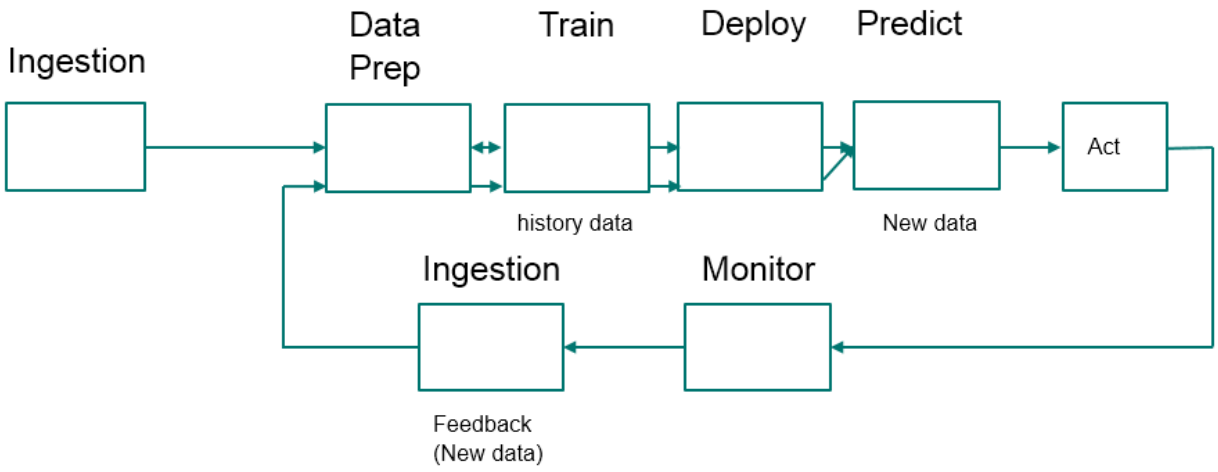
لأجل تحقيق هدف التعلم الآلي يجب وصف الخوارزميات التي سيعتمد عليها نموذج التعلم، وتنقسم هنا آليات عمل الخوارزميات إلى: (Puget, 2016)

- التعلم تحت الإشراف Supervised Learning: حيث يتم فيه إعطاء الخوارزمية بيانات التدريب التي تحتوي على الإجابة الصحيحة لكل مثال، يندرج تحتها نوعين 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 اثنين فقط من القيمة المحتملة نقول إنها مشكلة تصنيف ثنائي.
- التعليم الغير خاضع للإشراف Unsupervised Learning: حيث تبحث فيها الخوارزمية عن بنية بيانات التدريب، مثل البحث عن الأمثلة المتشابهة مع بعضها البعض وتجميعها في مجموعات، يندرج تحتها نوعين من الخوارزميات:

- التجزئة Segmentation: حيث تكون البنية المراد تعلمها عبارة 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 حيث 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:

تبدء خطوات التعلم الآلي تحديد البيانات والحصول عليها Ingestion، ثم عمل تجهيزها وإعدادها للتدريب Data Prep، ثم تأتي مرحلة التدريب train وفقًا لأحدى الأساليب سواء كانت خاضعة للأشراف أو غير خاضعة للأشراف، 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 بأحدي 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. تتضمن التطبيقات الشائعة التعرف على الصور والكلام. حيث كانت بداية التعلم العميق في عام 2006، حينما ركز مشكلة تصنيف الصور المعروفة بمشكلة MNIST.

تقنيًا يُعد التعلم العميق هو التعلم الآلي ووظائفه مماثلة له، ولكن الاختلاف بينهم في القدرات، إذا كانت خوارزمية الذكاء

الاصطناعي تُرجع تنبؤًا 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, 2020)

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. تتطلب 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)

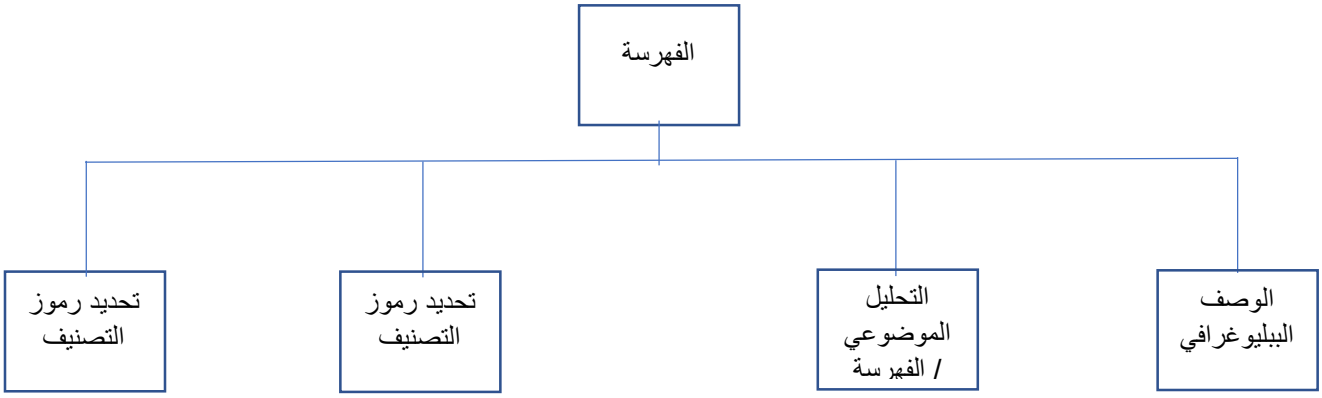
معالجة اللغة الطبيعية هي قدرة أجهزة الكمبيوتر على تحليل وفهم وتوليد اللغة البشرية، بما في ذلك الكلام.

2/2 الفهرسة / البيانات الوصفية (الببليوغرافي)

تعرف الفهرسة على أنها العملية التي يتم بمقتضاها توفير الوصول إلى المواد 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 الفهرس الوصفية بوصف الكيان المادي أو الملامح المادية لأوعية المعلومات عن طريق تقديم مجموعة من البيانات الببليوجرافية (اسم المؤلف، العنوان، وبيانات النشر، إلى أخره 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 المصادر 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 الببليوجرافية، وكيفية تسجيل البيانات التي تتعلق بالعلاقات بين هذه المصادر

وبدئت 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 (2) AACR.

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)

ويعتبر الركيزة الأساسية التي تم الاعتماد عليها في بناء المعيار هي الوثيقة الصادرة عام 2009 بعنوان " بيان المبادئ العالمية للفهرسة"، حيث يشتمل البيان على الأساس النظري المنطقي في عائلة FRBR 1998 – 2010[3] (المتطلبات الوظيفية للتسجيلات

[3] "مجموعة من النماذج المفاهيمية التي صيغت في شكل نماذج علاقات بين الكيانات Entity RelationshipDiagram "ERD" "

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

2/2/2 معايير التكويد / الشكل Structure standards

فهو قالب Form تضع 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 آلي ًا (MARC 21 (Machine Readable Cataloging

بدء مفهوم الفهرسة الآلية مع ظهور مارك - فهو معيار تنسيق رقمي Digital format لوصف العناصر الببليوجرافية التي طورتها مكتبة الكونجرس - في ستينات 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 شخص مجموعة أساسية من الدلات 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 الدلالي[4]

- 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 آلي ًا، وخاصة في العلوم الإنسانية والعلوم الاجتماعية واللغويات. وبدئت فكرتها منذ عام 1987 حتى تم إصدار النسخة الأولى من الإرشادات في عام 1994، وتم استخدام إرشادات TEI على نطاق

[4] 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.

واسع من قبل المكتبات والمتاحف والناشرين والباحثين الفرديين لتقديم النصوص للبحث عبر الإنترنت والتدريس والحفظ. بالإضافة إلى الإرشادات نفسها. (TEI)

3 التعلم الآلي واستخراج البيانات الببليوجرافية

ظهرت العديد من النظم مفتوحة المصدر التي تقوم باستخراج البيانات الببليوجرافية ومنها التالي:

- parsCit
- CERMINE
- Science Parse
- science Parse v2
- Metatagger
- BILBO
- ويعرض الباحث نوعين من هذه النظم وهما GROBID و CERMINE، وبالتحديد هذين النظامين لما وجده الباحث من خلال الدراسات السابقة في ارتقائهم في الترتيب على بقية النظم

Grobid (GeneRation Of BIbliographic Data) 1/3

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)[5] وهي تقنية للتعلم الآلي لاستخراج المحتوى وإعادة هيكلته تلقائيًا 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 بشكل خاص على المنشورات التقنية والعلمية. التطورات الأولى بدأت في عام 2008. ثم أصبح الشفرة المصدرية Source Code متاحة منذ عام 2011 . كان العمل على GROBIDثابتًا كمشروع جانبي منذ البداية ويتوقع أن يستمر على هذا النحو.

يقدم الوظائف التالية

[5] تعد إطار عمل احتمالي لتصنيف البيانات وتنظيم وهيكلة البيانات (hanna, 2005)

- استخراج الرأس Header وعمل Parse[6] للمقالات ذات امتداد PDF. يغطي الاستخراج هنا المعلومات الببليوغرافية (مثل العنوان والمستخلص والمؤلفين والانتساب Affiliation[7] والكلمات المفتاحية وغيرها).
- 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 ، وخاصة 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 مجزء لهيكلة النص.

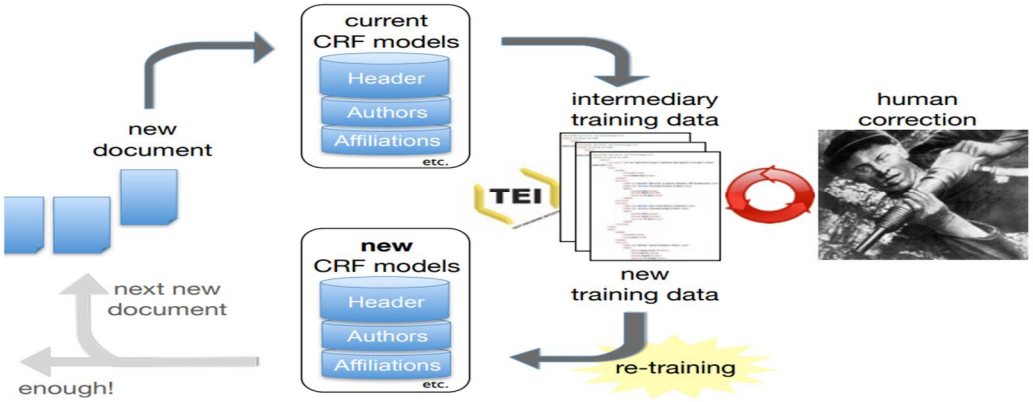
يتضمن GROBID معالجة الُدفعات ، وواجهة برمجة تطبيقات RESTful API شاملة ، واجهة برمجة تطبيقات JAVA API ، وعدم لاستخدام docker إطار تقييم عام نسبيً ا (الدقة ، الاسترجاع ، إلخ) وإنشاء بيانات التدريب شبه التلقائي.

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

[6] وهي قراءة وتحليل سلسلة المحارف

[7] عرفها الشامي:" في الببليومتريقا تعرف بانها المؤسسة أو العمل الذي ينتسب إليه الشخص، وتحليل الانتساب يمكن استخدامه لتقويم أو مقارنة المؤسسات أو مجلات الأبحاث".

الشكل رقم التالي يوضح آلية عمل GROBID



الشكل رقم (2) آلية عمل GROBID

CERMINE 2/3

هي عبارة عن مكتبة برمجية مكتوبة بالغة الجافات 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

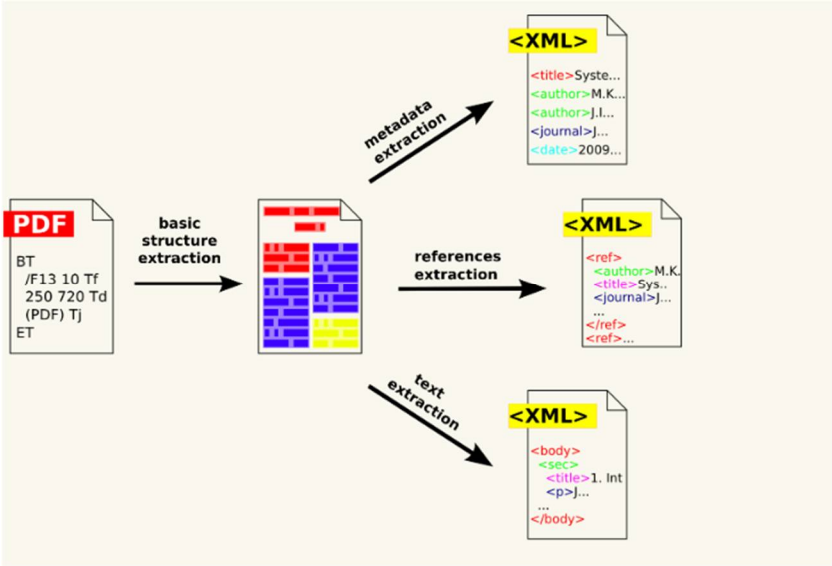
حيث يقوم النظام بتحليل محتوى ملف بتنسيق PDF ويحاول استخراج معلومات:

- عنوان المقال
- معلومات المجلة (العنوان، إلخ)
- بيانات ببليوجرافية (الطبعة، العدد، عدد الصفحات، إلخ)
- 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 والغير خاضعة للإشراف unsupervised.

الشكل التالي يوضح آلية عمل نظام CERMINE



الشكل رقم (3) آلية عمل CERMINE

4 النموذج المقترح لاستخراج البيانات الببليوجرافية من المصادر العربية النصية

يقترح الباحث هنا نموذج لاستخراج البيانات الببليوجرافية من المواد النصية العربية، وخاصة الأبحاث / الأوراق العلمية والكتب والأطروحات

1/4 أهداف النموذج

- استخراج البيانات الببليوجرافية من الأبحاث / الأوراق العليمة والكتب والأطروحات الصادرة عن قواعد البيانات المتاحة على الويب مثل دار المنظومة
- توفير بيانات وصفية وببليوجرافية للمقالات والأبحاث / الأوراق العلمية ومن ثم توفير وقت ومجهود المكتبات في عمل ذلك، ودعم المكتبات الرقمية خاصة 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 و الأساليب Styles التي تتبعها دوور النشر العربية.

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

- تتم عملية الفهرسة قبل النشر سواء من قبل الناشر أو المكتبة وتعرف ب الفهرسة قبل النشر / أثناء النشر

في صناعة النشر ظهر مصطلح الفهرسة أثناء النشر CIP (cataloging in publication)ووضع فكرتها winsor jastinعام 1886، ثم بدئت مكتبة الكونجرس عام 1971 باستخدام المصطلح، وعرب المصطلح على يد د. سعد الهجرسي (عبد الهادي، 2010)[8]

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، العنوان، المؤلف، وغيرها).

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

- بعد النشر من قبل المكتبات، 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 كم 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 ُمعد مسبقًا).
- إدخال البيانات على النظام الفرعي للفهرسة، في أحدى قوالب مارك 21 أو Bibframe.
- مراجعة تسجيله الفهرسة في شكلها النهائي وتدقيقها، والسماح بعرضها في واجهة البحث OPAC.

هل عمل المفهرس وتحديدًا في الفهرسة الوصفية / الببليوجرافية يعتبر عمل روتيني متكرر؟

للإجابة على هذا السؤال، قام الباحث بإجراء استبيان على عينة عشوائية من المهنين في الفهرسة عددهم عشرون بواسطة نماذج جوجل Google forms، حيث حدد عدة عناصر لقياس هل الفهرسة الوصفية عمل تكراري أم لا. جاءت أسئلة الاستبيان كالتالي:

8

شكل رقم (4) أسئلة الاستبيان التي تم طرحها على المهنين في مجال الفهرسة، تم انشائها من خلال google forms

كان لكل سؤال هدف خاص ساعد الباحث على وضع مبرر لضرورة ان يتم البدء في استخدام التعلم الآلي وتطبيقات الأصطناعي في الفهرسة الوصفية.

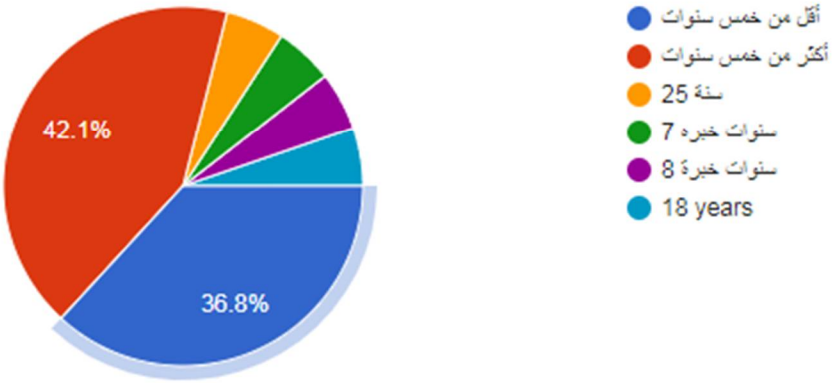
جاءت النتائج كالتي

- 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.

كم عدد سنوات الخبرة التي عملت فيها في مجال الفهرسة؟

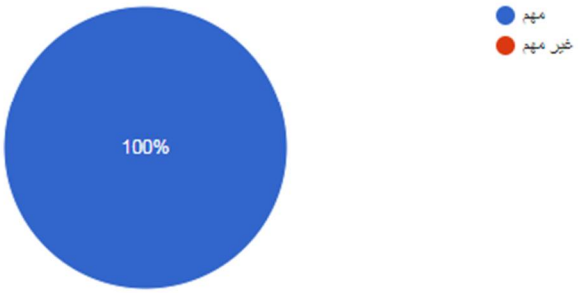
19 ردًا



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

هل العمل كمفهرس مهم أم لا

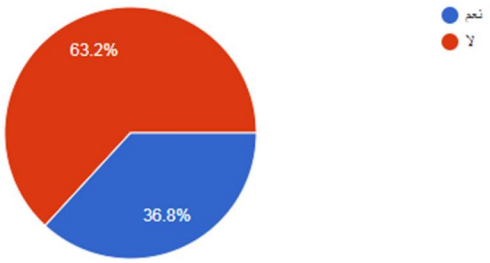
19 ردًا



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

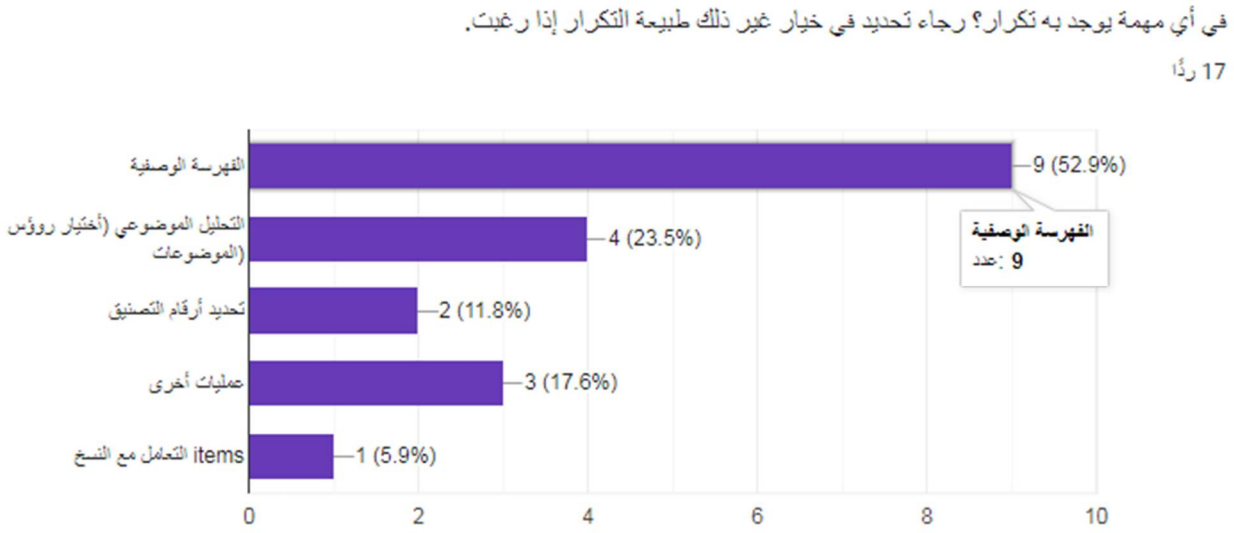
هل عملك به تكرار وتشعر بالملل؟

19 ردًا



إذا اغلب المهنين يشعرون بالملل من تكرار العمل وشعورهم بالملل.

- في أي مهمة يوجد تكرار: قام بالاجابة على هذا السؤال سبعة عشر فقط من العشرون، ووجاءت الفهرسة الوصفية بنسبة 52.9 %



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

4/4 مكونات النموذج المقترح

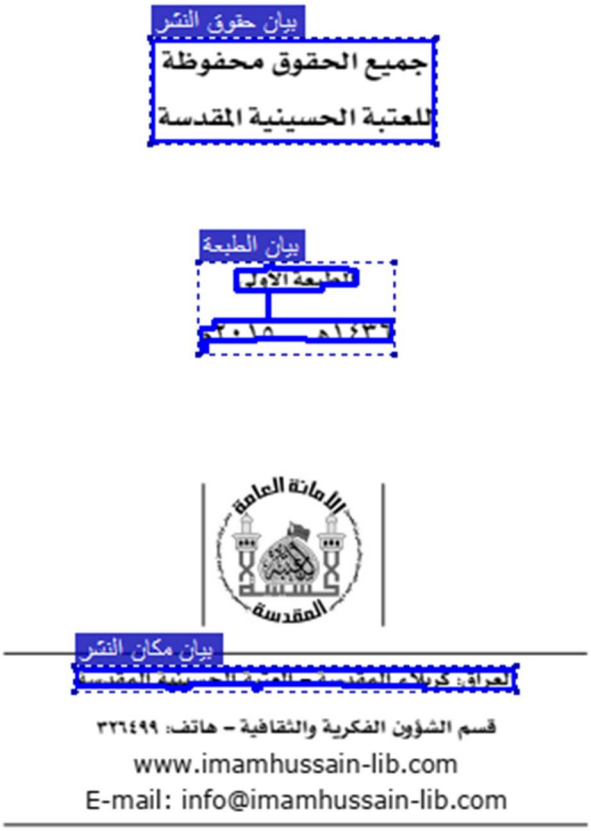
ويقترح الباحث بناء النموذج بالاعتماد على نظام جاهز وعمل تعديل عليه وتدريبه من خلال مجموعات البيانات، ويرى الباحث أن نظام GROPID هو الأفضل لما له من مزايا سبق ذكرها. يتناول الباحث مكونات النموذج بالتطبيق على (كتاب صادر عن قسم الشؤون الفكرية والثقافية في العتبة الحسينية، وبحث علمي منشور بدار المنظومة،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

يتكون النموذج المقترح من عدة خطوات:

1/4/4 تحديد المخططات Layout وأماكن البيانات الببليوجرافية 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 البيليوجرافية عادة في صحفة العنوان والصفحة التي تاليها أو من الممكن أن يتوفر 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.

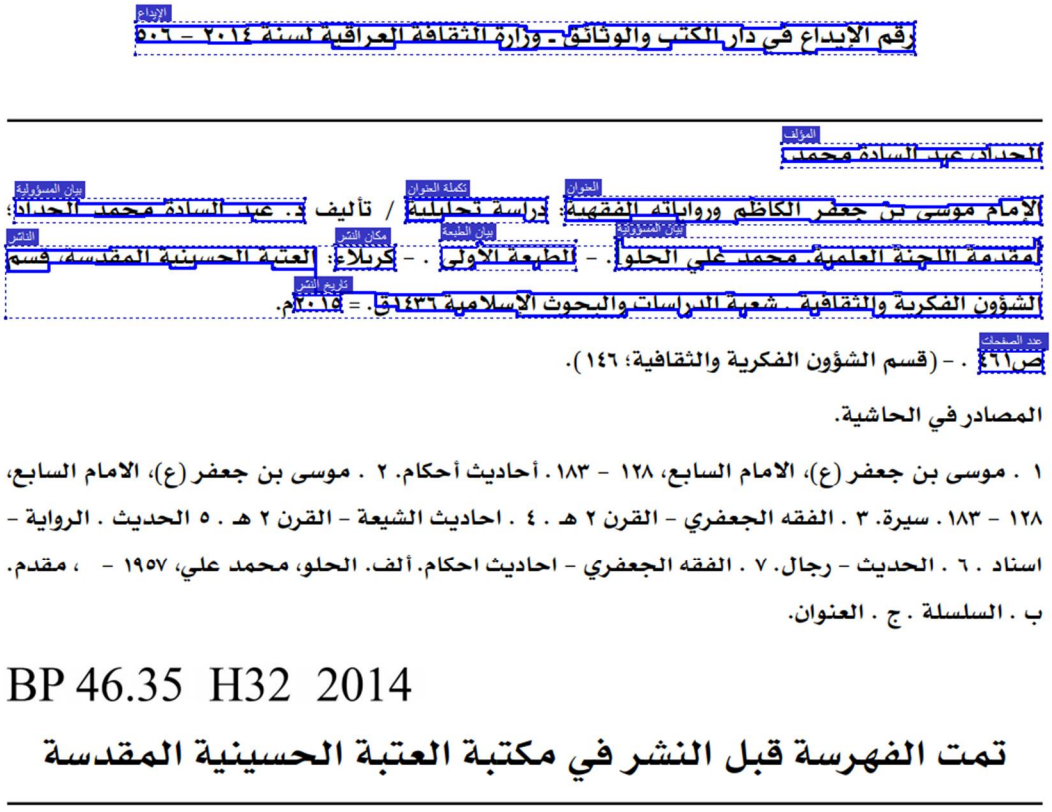


(نموذج للبيانات البيليوجرافية من النص الأصلي)

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 الببليوجرافية المُعدة قبل النشر من قبل مكتبة العتبة الحسينية المقدسة)



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 ؛ مقدمة 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.

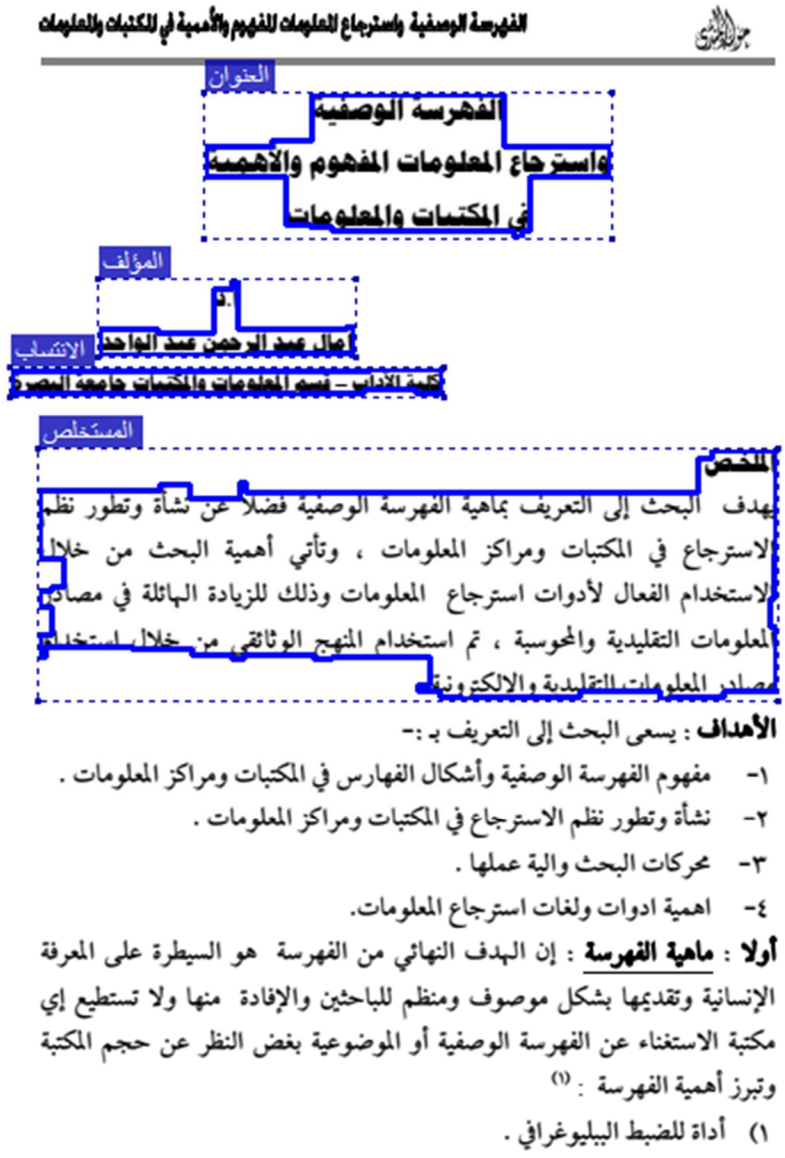
BP 46.35 H32 2014

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 الصحفة 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 الببليوجرافية من النص الأصلي)



الفهرسة الوصفية واسترجاع المعلومات للفهوم 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 الببليوغرافي .

- 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"/>
```

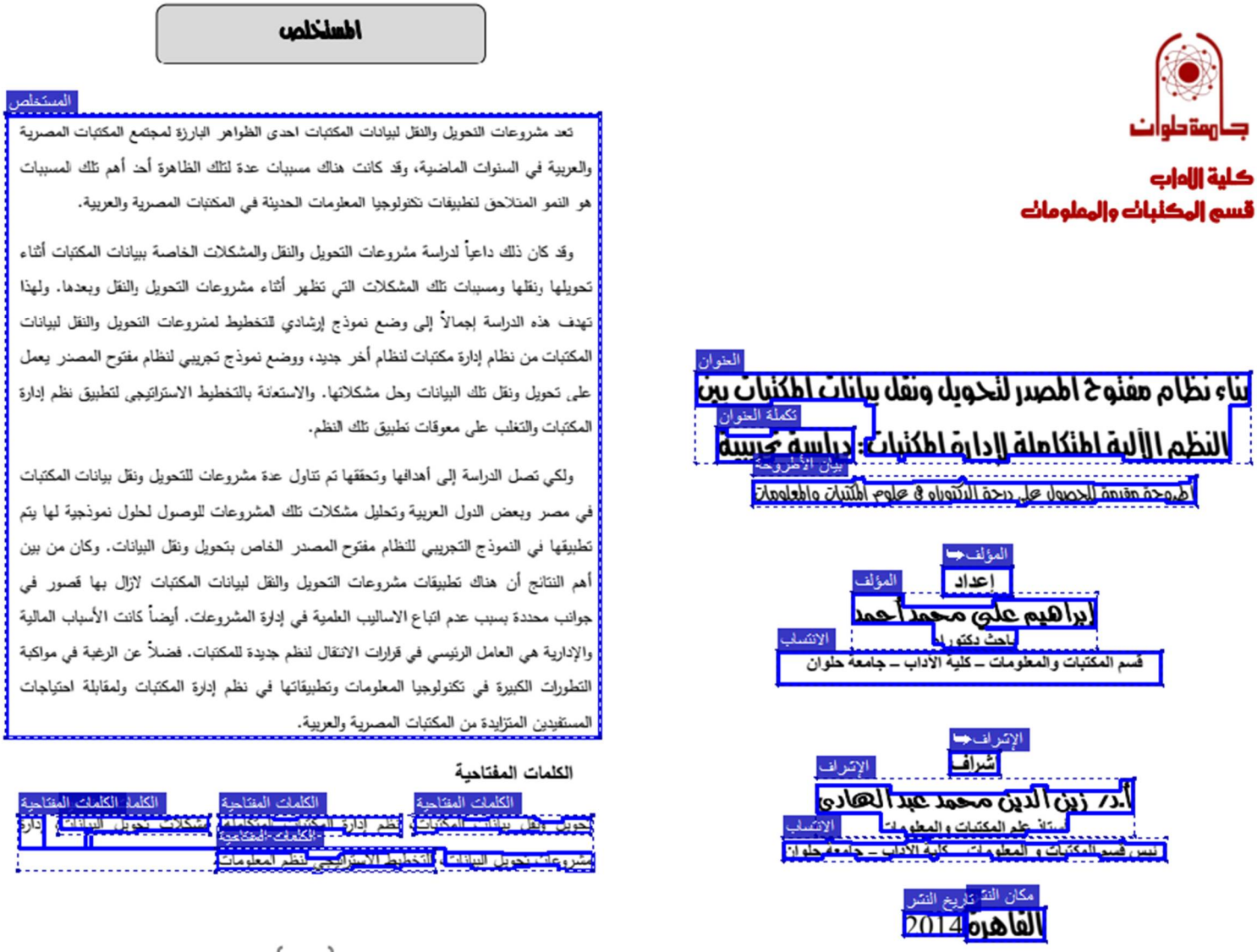
(نموذج للبيانات الببليوجرافية المُعدة قبل النشر من قبل قواعد المعلومات "دار المنظومة")

- الأطروحات

توجد البيانات الببليوجرافية في الأطروحات في الصفحة الأولى والصفحة الثانية، مثال على ذلك أطروحة دكتورة 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.

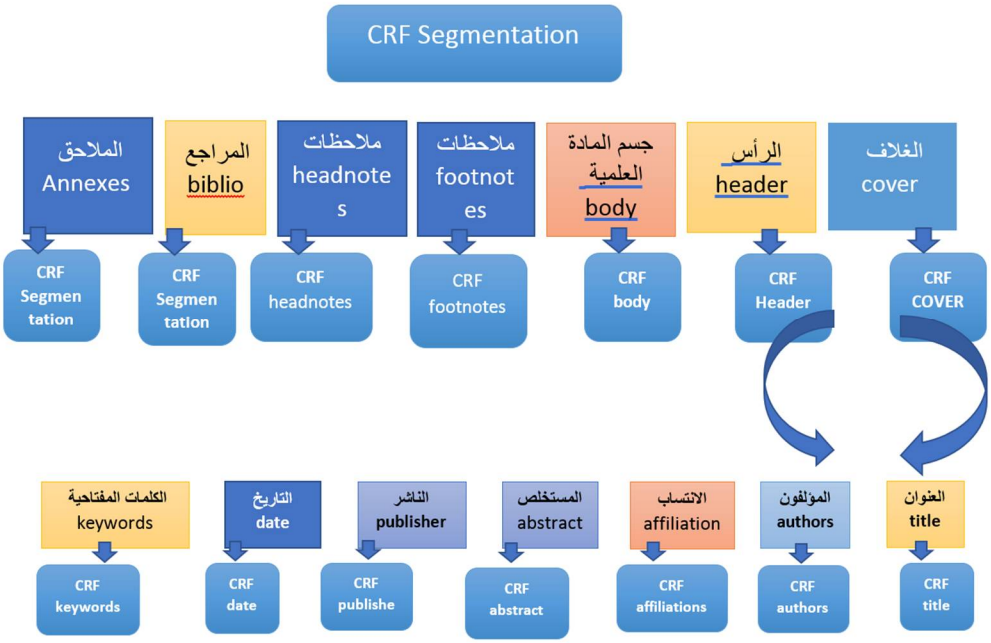
يعتم منهج التعلم الآلي Machine learning approach في نظام GROBID على 11 نموذج من CRF، يستخدم كل نموذج نفس النموذج العام المبني على إطار يغطي عملية التدريب والتقييم وفك التشفير وتعيين الرموز المميزة tokenization لكل نموذج له مجموعة من المميزات ومجموعة من البيانات المدربة والمُطبعة normalization.

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	

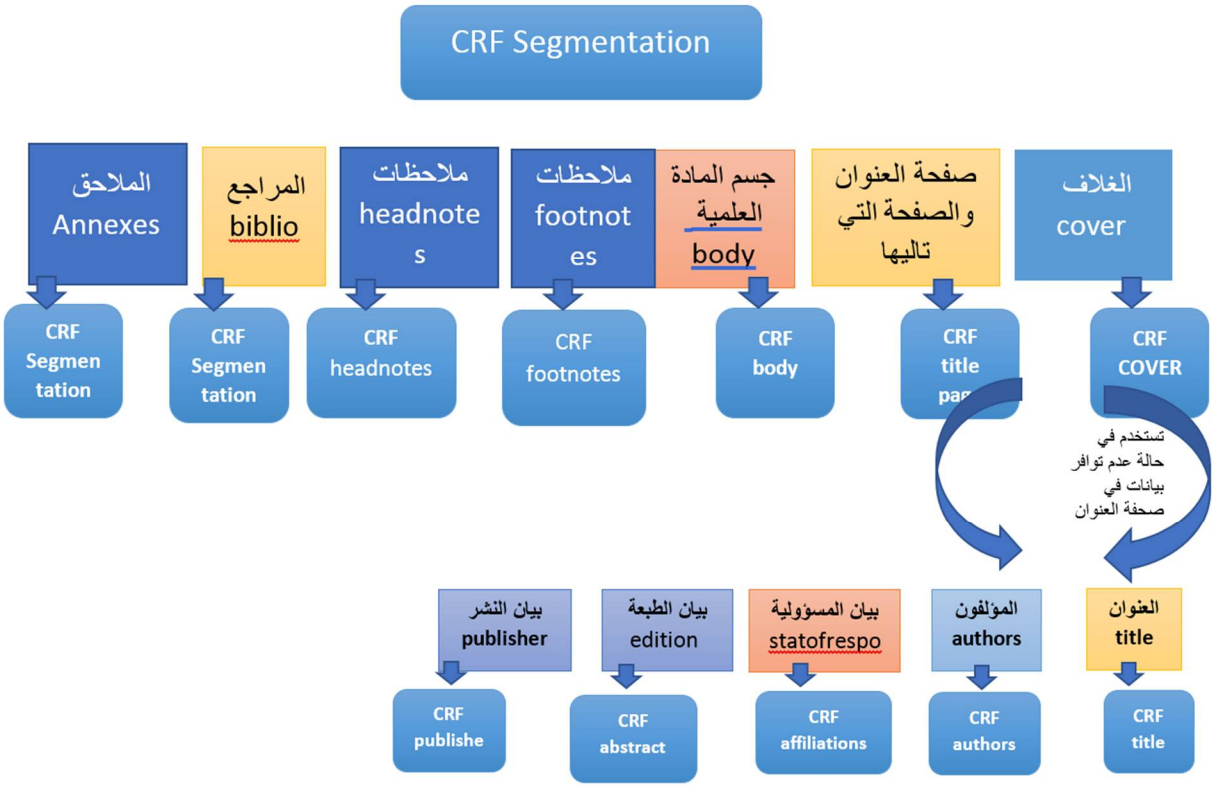
بيان الطبعة	عدد الصفحات	
حقوق النشر	الرابط	
بيان مكان النشر	المصدر (مصدر الحصول على المقالة)	
الإيداع	الموضوع	
عدد الصفحات	بيان التحكيم	
	نوع المحتوى	
	قواعد المعلومات	

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 وجد الباحث أن من المهم دمج الآلية لتشمل الإثنين معاً



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

الآلية المتبعة في استخراج البيانات الببيلوجرافية من الكتب تم فصلها عن الأوراق العلمية والأطروحات بسبب أختلاف عناصر الفهرسة بها وتنوعها وأختلاف مواقعها

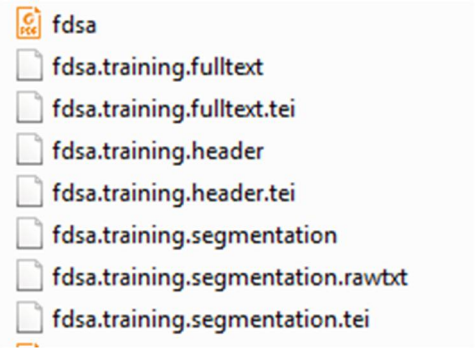


الشكل رقم (6) بوضح آلية استخراج البيانات الببيلوجرافية من الكتب

3/4/4 التدريب والتقييم

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

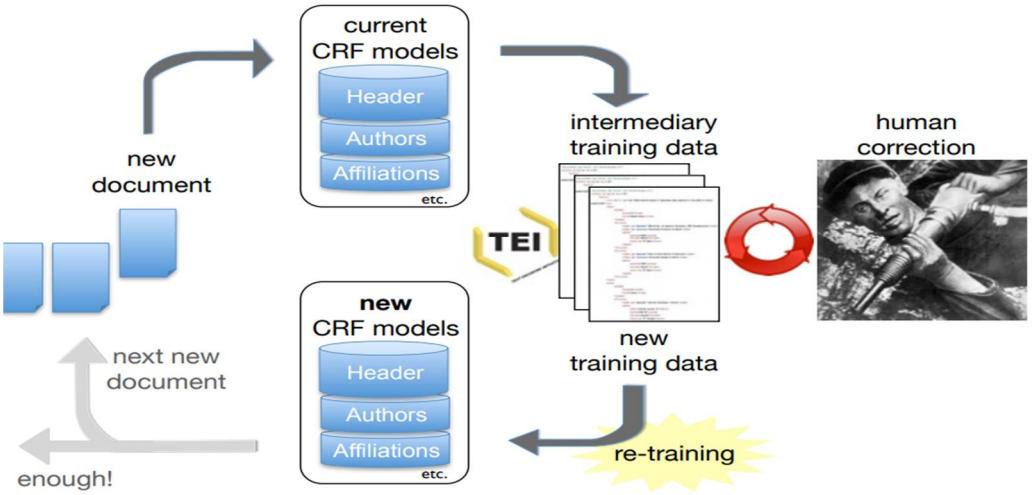
ينتج عن عملية التدريب ملفات متعلقة بكل نموذج ، قام الباحث بإجراء التدريب ونتج عن ذلك التدريب ملفات، راجع ملحق رقم (3)



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 أدراك الآلة الأخطاء في المرات التالية للتدريب

ويمثل 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.

قام الباحث بمقابلة للعناصر التي تم استخراجها من النماذج السابقة - للكتب والأوراق / الأبحاث العلمية والأطروحات – مع معيار TEI المستخدم، راجع الملحق رقم (2).

6/4/4 إدخال البيانات على النظام من خلال مقابلة mapping عناصر TEI مع MARC 21 (RDA)

تعتبر هذه المرحلة من أهم المراحل التي يتم من خلالها نقل البيانات إلى النظام الآلي المتكامل خاصة 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 "SQLليس فقط" ، بديلاً لقواعد البيانات الترابطية التقليدية التي يتم فيها وضع البيانات في

الجداول ويتم تصميم مخطط schema البيانات بعناية قبل إنشاء قاعدة البيانات. قواعد بيانات NoSQL مفيدة بشكل خاص للعمل مع مجموعات كبيرة من البيانات الموزعة. (rouse, 2017)

نموذجنا هنا يعتمد على قاعدة بيانات NOSQL ويستخدم معيار TEI مبني 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 حالياً ما 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 الألي لاستخراج البيانات الببليوجرافية من مصادر المعلومات النصية المكتوبة باللغة العربية، دون الحاجة إلى نقل ملفات بين النظامين

يقترح الباحث أن يتبنى نظام GROBID مكتبة Skikit-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. هي 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 الباحث بعدد من النتائج والتوصيات.

1/6 النتائج :

- الفهرسة الوصفية عملية تكرارية
- عدم توفر مجموعات بيانات Datasets لكافة المواد النصية الصادرة عن دور النشر العربية وبالأخص المصرية.
- تنوع مخططات الكتب يُعد مشكلة في تدريب النموذج.
- لا ينفي هذه الآلية دور المفهرس وأنما 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").
- توحيد مخططات layouts الكتب، وإلزام الناشرين بوضع البيانات في الأماكن المتفق عليها.
- أن يتم عمل Hackathon في المؤتمرات المحلية أو الدولية والإقليمية لتشجيع المفهرسين والمطورين developers على تطوير مهنة الفهرسة والاستفادة من تطبيقات الذكاء الاصطناعي.

6 الملاحق

الملحق رقم (1)

جدول المقابلة Crossoalk 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 $a,$b • 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>		028 5_
├ <availability>	<license>	540
└ <date when="___"/>		264_1 $c
├ <seriesStmt>		n/a
└ <title level="s" type="___">	• main • sub • alt • short • desc • translated • filing	• 490 • 8xx (optional)
└ <notesStmt> <note>		5xx
└ <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 = 1st 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 2nd 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>		6xx 2nd indicator or 6xx $2 when 2nd 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">
الإيداع الدولي / المحلي **معرّف الكيانات الرقمية DOI** **محدد موقع الموارد المُوحّد URL** الرقم المعياري الدولي للدوريات ISSN	<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)

النموذج	اسم الملف
Segmentation	***.training.segmentation.tei.xml**
Header	***.training.header.tei.xml**
affiliation-address	***.training.header.affiliation.tei.xml**
header	***.training.header.authors.tei.xml**

Date	***.training.header.date.xml**
Header	***.training.header-references.xml**
Fulltext	***.training.fulltext.tei.xml**
figure, table	***.training.figure.tei.xml and *.training.table.tei.xml**
reference-segmenter	***.training.references.referenceSegmenter.tei.xml**
Fulltext	***.training.references.tei.xml**
citation	***.training.references.authors.tei.xml**

7 المراجع / المصادر

1/7 العربية

1. حسام الدين، مصطفى. وصف المصادر وإتاحتها (و م إ : RDA) الملامح والبناء والتطبيق في بيئة عربية .- Cybrarians Journal.- ع 26، سبتمبر 2011 . تاريخ الاطلاع 15 فبراير 2020. متاح في: shorturl.at/ilwJ3
2. عبد الهادي، محمد فتحي و جمعة، نبيلة خليفة. (2010). الفهرسة في البيئة الإلكترونية).(ط 8). القاهرة، مصر: الدار المصرية اللبنانية.

2/7 الأجنبية

1. About. (n.d.). Retrieved February 16, 2020,from https://www.rdatoolkit.org/about
2. DCMI History. (n.d.). Retrieved February 10, 2020, from https://www.dublincore.org/about/history/
3. Elshami, A. (2018). Default. Retrieved February 8, 2020, from https://www.elshami.com/
4. Grossfeld, B. (2020, January 24). A simple way to understand machine learning vs deep learning. Retrieved February 4, 2020, from https://www.zendesk.com/blog/machine-learning-and-deep-learning/
5. III, R. E. W. (2019, December 17). Learn What Machine Learning Is and How It's Changing the World of AI. Retrieved February 15, 2020, from https://www.lifewire.com/what-is-machine-learning-4773696
6. 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

7. 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
8. 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
9. Reitz, J. M. (n.d.). ODLIS. Retrieved February 7, 2020, from https://www.abc-clio.com/ODLIS/odlis_b.aspx#bibdescrip
10. Rouse, M. (2017, March 13). What is NoSQL (Not Only SQL database)? - Definition from WhatIs.com. Retrieved January 29, 2020, from https://searchdatamanagement.techtarget.com/definition/NoSQL-Not-Only-SQL
11. Saloky, T., & Šeminský, J. (2005). Artificial Intelligence and Machine Learning . Retrieved from http://uni-obuda.hu/conferences/SAMI2005/SALOKY.pdf
12. scikit-learn. (n.d.). Retrieved January 29, 2020, from https://scikit-learn.org/stable/index.html
13. 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
14. TEI: Text Encoding Initiative. (n.d.). Retrieved January 28, 2020, from https://tei-c.org/
15. the Dartmouth Artificial Intelligence Conference: The Next Fifty Years was held at the College. (n.d.). Retrieved January 6, 2020, from http://www.dartmouth.edu/~ai50/homepage.html.
16. Thompson, W., Li, H., & Bolen, A. (n.d.). Artificial intelligence, machine learning, deep learning and more. Retrieved January 6, 2020, from

https://www.sas.com/en_us/insights/articles/big-data/artificial-intelligence-machine-learning-deep-learning-and-beyond.html.

17. 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
18. 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 January 13, 2020, from https://arxiv.org/abs/1802.01168
19. 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
20. 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
21. 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
22. wallach, hanna m. (2005, May 12). introductionconditional random fields. Retrieved from http://www.inference.org.uk/hmw26/crf/
23. What is R? (n.d.). Retrieved January 15, 2020, from https://www.r-project.org/about.html