

## АВТОМАТИЗОВАНИЙ ПАРСИНГ БІБЛІОГРАФІЧНИХ ПОСИЛАНЬ

Безносик О. Ю., Кулик О. В.

## АВТОМАТИЗИРОВАННЫЙ ПАРСИНГ БИБЛИОГРАФИЧЕСКИХ ССЫЛОК

Безносик А. Ю., Кулик А. В.

## AUTOMATIZED PARSING OF BIBLIOGRAPHIC REFERENCES

Beznosyk O., Kulyk O.

Національний технічний університет України  
«Київський політехнічний інститут імені Ігоря Сікорського»

Київ, Україна

[o.beznosyk@gmail.com](mailto:o.beznosyk@gmail.com)

*Дане дослідження присвячене створенню парсеру для автоматизованого пошуку, декомпозиції та коректного занесення до бази даних бібліографічних посилань з різних джерел інформації. У статті розглядається алгоритм роботи парсеру, що здатний інтерпретувати, обробляти та зберігати бібліографічні посилання, представлені в найбільш розповсюджених стандартах.*

**Ключові слова:** автоматизований парсер, бібліографічне посилання

*Данное исследование посвящено созданию парсера для автоматизированного поиска, декомпозиции и корректного занесения в базу данных библиографических ссылок из разных источников информации. В статье рассматривается алгоритм работы парсера, который способен интерпретировать, обрабатывать и сохранять библиографические ссылки, представленные в наиболее распространенных стандартах.*

**Ключевые слова:** автоматизированный парсер, библиографическая ссылка

*The given investigation is devoted to the development of a parser for automatized search, decomposition and correct input into the database of bibliographic references from different information sources. An algorithm of the parser's work intended to interpret, process and keep bibliographic references presented in the most widespread standards is considered in the paper.*

**Keywords:** automatized parser, bibliographic reference

### ВСТУП

За роки дослідницької діяльності будь-якої кафедри чи іншої наукової установи зазвичай накопичується значна кількість викладацьких та студентських публікацій, що потребують внесення в базу обліку. Створити саму базу даних для обліку публікацій не є складним завданням, але її заповнення за допомогою ручного вводу бібліографічними посиланнями за достатньо великий проміжок часу потребує значних зусиль на обробку джерел посилань:

- пошук бібліографічних посилань, наприклад, у файлах різного роду звітів;
- аналіз структури знайдених посилань;
- ручний перенос даних кожного посилання у відповідні поля бази даних.

Як наслідок, це приводить до значних витрат часу, що також ускладнюється

відсутністю перевірки на дублікати при внесенні в систему, оскільки в такому випадку користувач має сам виконувати відповідну перевірку. Таким чином, виникає задача дослідження можливостей автоматизації процесу заповнення бази даних публікацій.

Актуальність цього дослідження полягає у тому, що постійно зростає кількість різноманітної інформації, яку має обробляти підрозділ, скажімо, рівня кафедри для подальшого формування відповідної звітності, яка базується на інформації про публікації викладачів та студентів. Збір такої інформації вручну потребує багато часу. Створення та своєчасне наповнення бази даних відповідною інформацією дозволить швидко отримувати всі необхідні звіти за різними параметрами (наприклад, скільки сторінок займають публікації викладачів кафедри, що представлені у *Scopus*, за певний рік). Складність роботи полягає у тому, що існуючі стандарти (ДСТУ та *IEEE*) допускають багато варіантів коректного представлення бібліографічного опису, тобто в рамках такого опису може бути присутня або відсутня певна необов'язкова інформація. Також часто бібліографічні описи бувають представлені взагалі не у відповідності до стандартів, що також суттєво ускладнює їх автоматичну інтерпретацію.

На даний момент у відкритому доступі відсутні засоби пошуку та декомпозиції бібліографічних посилань з використанням семантичного аналізу, що означає фактичну відсутність аналогів розробленої системи та обумовлює наукову новизну даного дослідження.

### АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ

Процес внесення нових публікацій в розроблену базу обліку публікацій є досить простим. Користувач вносить дані у необхідні поля опису публікації і, за умови повноти та відповідності типу наданих даних, може створити новий запис з автоматично згенерованим бібліографічним посиланням. Проте, у випадку внесення публікацій за минулі роки, найчастіше в наявності є деякі файли з бібліографічними посиланнями або безпосередньо самі файли публікацій, які й слугують джерелом інформації для користувача системи. В такому випадку процес внесення існуючого масиву публікацій стає повторюваним копіюванням частин файлу у відповідні поля, на що витрачається значна кількість часу та не гарантується відсутність помилок. Виникає питання, яким чином можливо автоматизувати цей процес.

Проблема автоматизованого внесення даних відома людству вже давно та є добре дослідженою [1]. На даний момент для її вирішення використовують засоби розпізнавання символів для отримання тексту з його фізичних носіїв, з подальшим використанням інтелектуальних парсерів [2] – програм, що здатні встановити семантичну відповідність між шуканим полем об'єкту предметної області та фрагментом отриманого тексту, перетворюючи «сирі» текстові дані на блоки, що можна зберігати та оброблювати [3].

Таким чином, програми-парсери, що допустимо використовувати у нашому випадку, мають шукати поля посилань (автори, видавництво тощо) у файлі чи веб-сторінці, що значно звужує можливості пошуку серед наявних рішень. Ускладнює завдання пошуку задовільних рішень і необхідність підтримки ними української та російської мови, оскільки значна частина наявних бібліографічних посилань та самих файлів публікацій представлена саме ними.

Існує значна кількість комерційних рішень автоматизованого вводу даних з різних джерел [4], проте у відкритому доступі присутні лише декілька безкоштовних

рішень, що задовольняють потреби предметної області дослідження. Серед них проект з відкритим кодом для парсингу публікацій у *pdf*-форматі *Science Parse* версії 2 [5], засіб аналізу веб-статей *Newspaper3k* [6] та відкриті класифікатори іменованих сутностей *StanfordNLP/NamedEntityRecognition* [7] і *MIT Information Extraction* [8].

*Science Parse* версії 2 є парсером *pdf*-файлів публікацій. Він використовує багатошарову двонаправлену нейронну мережу на архітектурі довгої короткочасної пам'яті [9]. На даний момент парсер дозволяє визначати заголовок, авторів публікації, місце проведення заходу чи видавництва та рік публікації, а також формує бібліографічні посилання у форматі *APA* (Американська психологічна асоціація) [10].

Робота з засобом відбувається за допомогою командного інтерфейсу. Присутня можливість надання класифікатору парсера власного набору підготовлених тестових даних та ручного його навчання.

До переваг даного рішення можна віднести представлення результату пошуку у зручному форматі *JSON* (*JavaScript Object Notation*), наявність розгорнутої документації та можливість розгорнути готовий образ проекту у *Docker*, використовуючи готовий репозиторій *allenai/spv2:2.10* [5], що дозволяє просто інтегрувати засіб у існуючу систему обліку.

Незважаючи на невідповідність стандарту *APA* до ДСТУ (Державний стандарт України), інформація, наведена у цих посиланнях, може бути використана при побудові посилань бажаного формату, а отже сама невідповідність може не вважатися вагомим недоліком.

Значним же недоліком є відсутність натренованого класифікатору для статей українською та російською, що викликає необхідність збору та попередньої підготовки великої кількості тестових даних перед початком використання засобу. Також варто звернути увагу на те, що *pdf*-документ не є простим текстовим файлом-контейнером – він містить атрибути, що можуть допомогти у визначенні полів посилання на публікацію: розміщення тексту, гарнітуру, кегль, стиль та колір. Це теоретично може обмежувати ефективність використання даного засобу для необроблених текстових файлів, конвертованих у формат *pdf*.

*Newspaper3k* – це бібліотека для мови *Python*, що включає в себе натренований класифікатор полів інтернет-публікацій та дозволяє завантажити статтю для попередньої обробки з наданого гіперпосилання. *Newspaper3k* дозволяє провести семантичний аналіз та визначити назву, авторів, дату публікації та ключові слова.

До переваг даного рішення можна віднести зручність та простоту налаштування, добру підтримку (оскільки проект активно оновлюється розробниками) та підтримку аналізу публікацій англійською, українською та російською мовами, що дозволяє використовувати засіб для всіх наявних публікацій без обмежень за мовою.

До недоліків можна віднести відсутність визначення кількості сторінок, міста видавництва та видавника. Як і в розглянутому раніше засобі, текст має додаткові атрибути, що лімітує ефективність використання засобу попередньо відформатованими сторінками. Таким чином, використання тільки цього засобу без подальших модифікацій не дасть бажаного результату та не дозволить автоматично формувати повні бібліографічні посилання, проте наявний функціонал дозволяє розглядати це рішення як фундамент більш складного методу, що у комбінації з іншими рішеннями може дати задовільний результат.

Для уточнення результатів, отриманих за допомогою *Newspaper3k*, можна використовувати засіб розпізнавання іменованих сутностей (PIC)

### *StanfordNLP/NamedEntityRecognition.*

Розпізнавання іменованих сутностей – це задача встановлення відповідності між словом у тексті та його семантичною групою, з наступною поміткою слова тегом в залежності від налаштувань програми-розпізнавача [11]. Найчастіше зустрічаються наступні сутності: особа, організація, локація, час, гроші, гіперпосилання. Бібліотека *StanfordNLP/NamedEntityRecognition* є частиною програмного комплексу *CoreNLP* та є одним з найточніших засобів серед наявних у відкритому доступі. Ця бібліотека дозволяє виявити організації (видавники), числа та числові діапазони (сторінки) та локації (міста видавництва). Тобто, провівши попередній аналіз за допомогою *Newspaper3k* та відсіявши отримані результати, можна класифікувати залишок та надати користувачеві знайдені результати.

На жаль, всі засоби РС мають значний недолік при використанні з елементами предметної області роботи. Ключовим фактором правильності розпізнавання є наявність сформованих моделей мови, що містять значну кількість слів з різних джерел, представлених у різних відмінках та родах. Для отримання таких моделей необхідний доступ до великої кількості джерел інформації та значна обчислювальна потужність, що змушує можливість вибору бібліотек РС до тих, що мають готові моделі української та російської мови. Враховуючи їх відсутність у *StanfordNLP/NamedEntityRecognition*, використання засобу є можливим лише для бібліографічних посилань, записаних англійською мовою. Додатково, факт відсутності підтримки нашого мовного простору імен та організацій ставить під сумнів точність визначення транслітерованих сутностей.

Проект *MIT Information Extraction* включає в себе набір безкоштовних сучасних засобів для отримання даних [12]. На даний момент включає в себе як самі засоби РС та визначення бінарних відносин між словами, так і скрипти для тренування власних моделей екстракторів даних. Це значно спрощує процес отримання моделі мови.

Бібліотека може бути встановлена як утиліта для командного рядка чи програма для *C*, *Python* чи *Java*, що дозволяє використовувати її в межах стеку технологій системи обліку публікацій.

Проблема відсутності моделі РС для української мови була успішно вирішена у 2017 році спільнотою *lang-uk*, що змогли акумулювати в моделі близько 217 тисяч токенів розпізнавання та 6700 іменованих сутностей в категоріях «особи», «локації» та «організації» [13]. Також спільнотою була переорганізована під стандарт моделі *MIT Information Extraction* модель російської мови. Тестування продемонструвало, що модель потребує додаткового тренування на прикладах предметної області публікацій, проте навіть в поточному стані здатна визначати прості сутності.

Таким чином, проаналізовані існуючі рішення для парсингу файлів та веб-сторінок, що можуть бути використані для парсингу публікацій, не дозволяють повністю задовольнити наші потреби і можуть бути використані лише як частина процедури парсингу відповідних джерел. Враховуючи встановлені недоліки описаних рішень, потребу у розробці власних алгоритмів парсингу посилань можна вважати обґрунтованою.

### **РОЗРОБКА АЛГОРИТМІВ АНАЛІЗУ БІБЛІОГРАФІЧНИХ ПОСИЛАНЬ**

Загальна структура бібліографічного посилання, описана в стандартах ДСТУ 7.1:2006 [14] та ДСТУ 8302:2015 [15], допускає значну кількість додаткових полів опису, кожне з яких вимагає розміщення в певній послідовності відносно інших інформаційних блоків, а також використання відповідних розділових знаків.

Відповідно, навіть в умовах правильно оформленого посилання, задача його декомпозиції на основні смислових компонентів є нетривіальною.

Відомим рішенням для шаблонного пошуку символів у тексті є регулярні вирази – формальна мова опису підрядків тексту [16], що використовує певну множину символів як умовні позначення відповідних категорій символів [17].

Таким чином, для встановлення відповідності між елементом бібліографічного посилання та відповідним полем бази даних необхідно проаналізувати внутрішню структуру наданого парсеру тексту, взаємне розташування інформаційних блоків посилання та їх контекст.

Для точного визначення приналежності тексту до певного поля посилання, текст повинен мати певні унікальні ознаки. Розглянемо список полів та встановимо наявність (чи відсутність) у них унікальних ознак (табл. 1).

Таблиця 1. Поля та їх зовнішній вигляд для регулярних виразів

Поле	Ознака
Автор(и)	Слова, що починаються літерою верхнього регістру, перед чи після яких наявні ініціали
Назва частини публікації	Довільний набір нечислових символів, що має починатися з літери верхнього регістру
Назва повної публікації	
Відповідальна за публікацію установа	
Країна проведення конференції	
Повна дата проведення конференції	Формат «число слово число»
Номер видання чи журналу	Число, попереду чи після якого наявна позначка номеру чи видання
Номер тому	Число, попереду чи після якого наявна позначка тому
Місто видавника	Довільний набір нечислових символів, що має починатися з літери верхнього регістру
Країна видавництва	
Видавництво	
Рік	Число від 1900 до 2099
Сторінки	Число або числовий діапазон, якому або передую позначення сторінок, або слідує за ним
<i>ISBN, DOI, ISSN</i>	Наявність відповідного ключового слова ( <i>ISBN, DOI, ISSN</i> )
Режим доступу	Має структуру веб-посилання

Отже, частина полів має унікальні ознаки, що дозволяє використовувати для їх пошуку регулярні вирази. Знайшовши таким чином частину відповідностей між текстом та полем бібліографічного посилання, ми значно спростимо подальший їх пошук. На рис. 1 наведено результат розпізнавання одного з видів бібліографічних посилань за допомогою регулярних виразів, більша частина полів визначається вже на цьому етапі.

Виключивши успішно знайдені поля, отримаємо схожі з точки зору зовнішнього вигляду фрагменти тексту. Наступним кроком їх обробки стане використання засобу розпізнавання іменованих сутностей *MITIE*. Цей засіб було обрано через його мовну універсальність. Оскільки у словниках української, російської та англійської мов наявні сутності «Організація» та «Локація» [18], використання цього рішення дозволить встановити наступну частину полів (табл. 2). Проблема встановлення різниці між містом та країною вирішується за допомогою наявної у відкритому доступі бази локалізованих назв населених пунктів [19], що дозволяє визначити приналежність локації до типу «Місто». На рис. 2 наведено

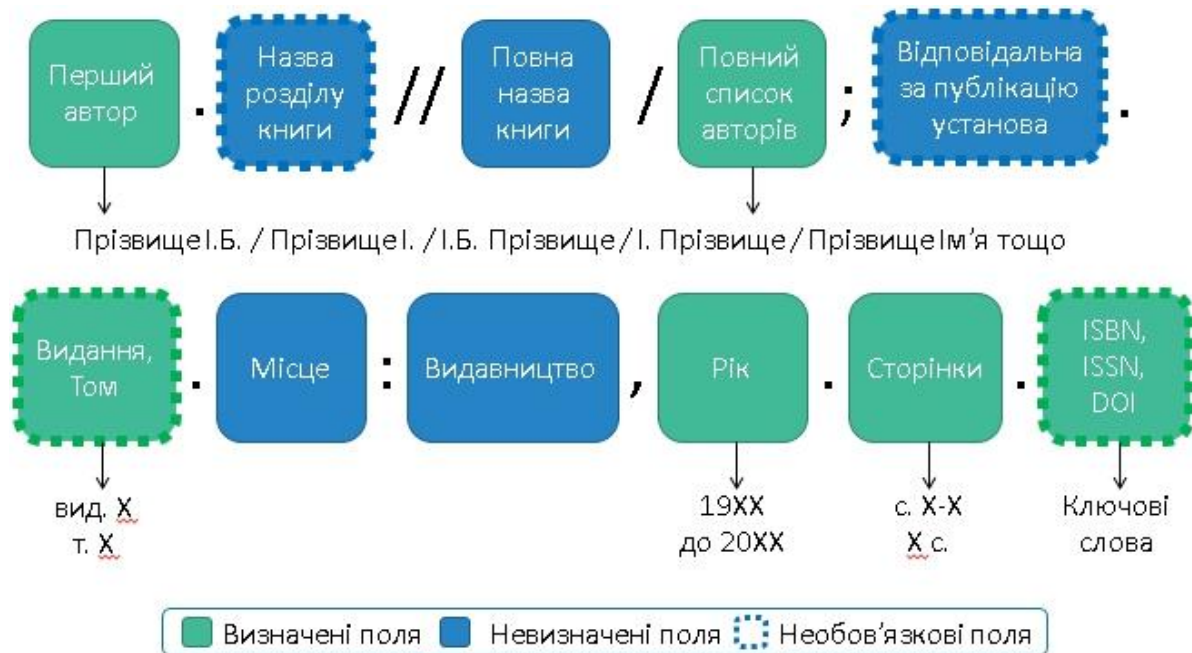


Рис. 1. Поля, що були визначені за допомогою регулярних виразів

Таблиця 2. Поля та їх відповідність знайденим сутностям

Поле	Сутність
Назва частини публікації	<i>Не належить до сутностей</i>
Назва повної публікації	<i>Не належить до сутностей</i>
Відповідальна за публікацію установа	Організація
Країна проведення конференції	Локація
Місто видавника	
Країна видавництва	
Видавництво	Організація

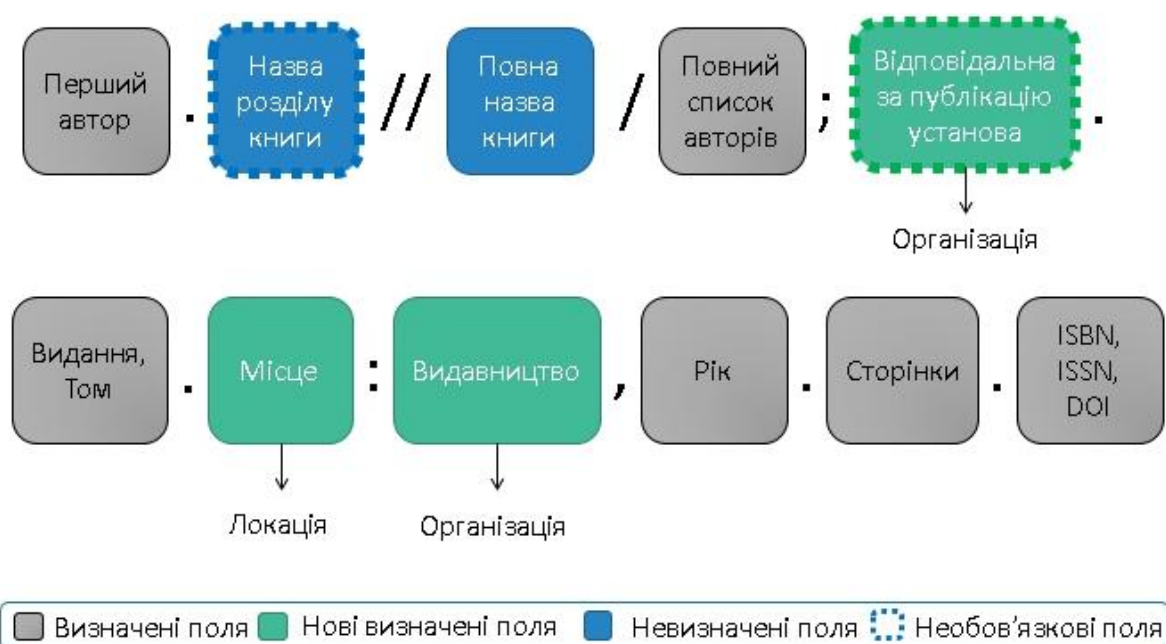


Рис. 2. Поля, що були визначені за допомогою розпізнавання іменованих сутностей

Остаточню ж розподілити наданий парсеру набір текстових даних дозволяє розташування блоків один відносно одного. Досвід обробки великої кількості бібліографічних посилань доводить відсутність прецедентів розміщення повної назви публікації попереду частини, що дозволяє робити припущення про можливість використання цього факту для визначення типу назви. На рис. 3 наведено результат аналізу посилання з точки зору взаємного розташування невизначених на попередніх етапах полів відносно вже ідентифікованих блоків та розділових знаків, з урахування відповідності цих полів стандарту оформлення.

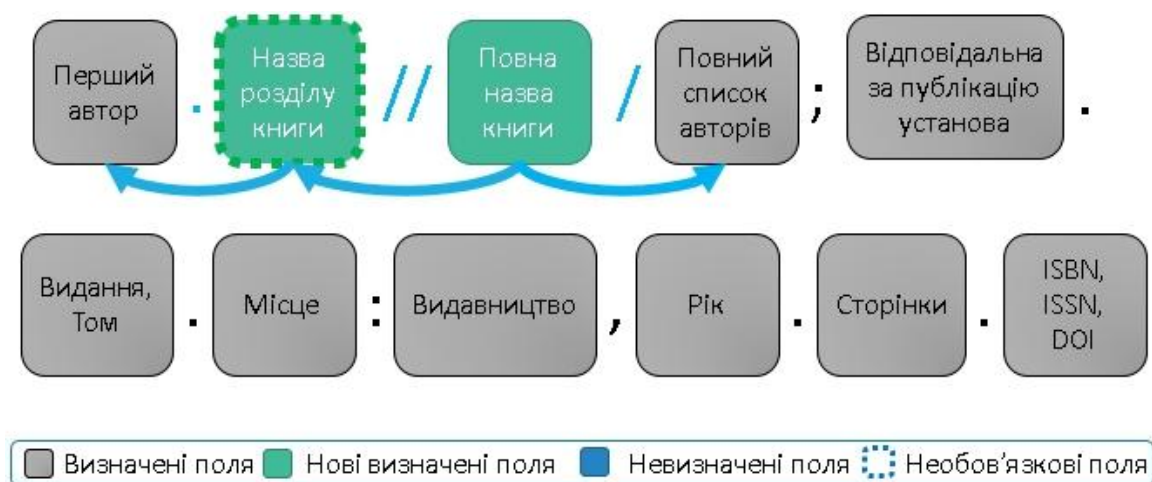


Рис. 3. Поля, що були визначені за допомогою аналізу взаємного розташування блоків

Крім того, для роботи в подальшому механізмів автоматичної генерації бібліографічних посилань на основі отриманих даних необхідно встановити відповідність між наданим бібліографічним посиланням та його типом, а також встановити мову посилання.




Враховуючи специфіку бібліографічного посилання, для вирішення цього завдання можливо використання регулярних виразів для встановлення типу посилання за наявністю коренів ключових слів та символів: семінар, конференція, симпозіум – для матеріалів конференції, номер журналу – для періодичного видання, та методом виключення визначати книгу. Аналогічним шляхом можлива реалізація уточнення типу посилання (навчальний посібник, конспект лекцій, тези доповіді тощо).

Мову можна встановити за наявністю специфічних символів, характерних лише для певної мови, виконуючи перевірку у такій послідовності: українська, російська, англійська. Таким чином, ми уникаємо розпізнавання тексту посилання, у якому присутні аббревіатури латинською абеткою, як посилання англійською мовою.

Таким чином, було визначено загальну послідовність кроків роботи парсеру для обробки наданих бібліографічних посилань, що відповідають певним стандартам оформлення. Проте, на практиці, надані парсеру посилання не завжди представлені у відповідності до стандартів та можуть мати значну кількість помилок різного типу, що значно ускладнює автоматичну інтерпретацію таких посилань. Але слід відміти, що навіть за таких умов, базуючись на вищенаведених принципах, розроблений парсер в більшості випадків демонструє гарні результати розпізнавання посилань. На рис. 4 наведено приклад штучно зробленого посилання, оформлення якого не відповідає прийнятим стандартам, та результат його успішного розпізнавання.

Різними кольорами показано частини посилання, що були ідентифіковані на різних етапах роботи парсеру.

[Ладогубець В.В., 2011, Методи підвищення ефективності Y-Δ перетворення при побудові схемних макромоделей МЕМС, \*Електроника и связь\* : тематический выпуск «\*Електроника и нанотехнологии\*» / \[Фіногенов О.Д., Безносик О.Ю., № 1.\]\(#\)](#)

-  Аналіз внутрішньої структури полів за допомогою регулярних виразів
-  Аналіз контексту полів за допомогою розпізнавання іменованих сутностей
-  Аналіз розташування полів

Відсутні сторінки – запис потребує уваги користувача

Рис. 4. Аналіз посилання з помилками оформлення

При цьому, за відсутності певної обов'язкової інформації, парсер видає також відповідне інформаційне повідомлення, що має привернути увагу користувача.

Загалом, тестування роботи парсеру було проведено на 122 бібліографічних посиланнях з відкритих джерел, з них після роботи парсеру додаткової уваги користувача потребувало 15. Таким чином, точність роботи на тестовій вибірці склала 88%. Проблеми розпізнавання були пов'язані з відсутністю або суттєвим спотворенням в цих посиланнях інформації щодо авторів (2), назви видання (4), міста видавника (7) та сторінок (2).

## ВИСНОВКИ

Основною метою даної роботи була розробка парсеру бібліографічних посилань, що представлені у визначених стандартах, з метою його подальшого використання для автоматизованого наповнення бази даних публікацій.

Запропонований алгоритм включає в себе три етапи:

- послідовне виключення класифікованих за допомогою регулярних виразів полів;
- семантичний аналіз інших полів та пошук відповідності «сутність – поле»;
- розподіл полів, що залишилися, враховуючи їх взаємне розташування у наданому парсеру посиланні.

Використання регулярних виразів на першому кроці обробки бібліографічних посилань є наслідком загальності структури відповідних елементів. Другий крок покликаний виявити організації (видавництва, відповідальні за публікацію організації) та локації (місто, країна видавника тощо), що можуть бути наявні в тексті та значно спрощують роботу на третьому етапі, залишаючи тільки декілька некласифікованих полів. На останньому кроці, враховуючи початкове положення інформаційного блоку відносно класифікованих полів, робиться висновок щодо приналежності блоків тексту, що залишилися, певним полям бібліографічного посилання. Частина посилання, що не була класифікована за ці три етапи, якщо така



є, записується в спеціальний буфер, що дозволяє користувачеві переглянути її вручну та прийняти рішення щодо її приналежності певним полям бібліографічного опису.

### ЛІТЕРАТУРА

1. Automated vs Manual Data Entry – How to Make an Informed Choice. *Floatapp* : веб-сайт. URL: <https://floatapp.com/blog/automated-vs-manual-data-entry-how-to-make-an-informed-choice> (дата звернення: 10.11.2019).
2. Grune D., C. Jacobs. Parsing Techniques: A Practical Guide. New York City : Springer, 2010. 688 с.
3. Multilingual Parsing from Raw Text to Universal Dependencies / D. Zeman, M. Popel et al. *CoNLL 2018*. Brussels, Belgium : Association for Computational Linguistics, 2018. 21 p. – URL: <https://www.aclweb.org/anthology/K18-2001.pdf> (дата звернення: 10.11.2019).
4. Data Entry Software: Review and Pricing. *SoftwareAdvice* : веб-сайт. URL: <https://www.softwareadvice.com/data-entry/> (дата звернення: 12.11.2019).
5. Science-parse version 2. *GitHub* : веб-сайт. URL: <https://github.com/allenai/spv2> (дата звернення: 3.12.2019).
6. Newspaper3k: Article scraping & curation. *Read the Docs* : веб-сайт. URL: <https://newspaper.readthedocs.io/en/latest/> (дата звернення: 10.11.2019).
7. StanfordNLP 0.2.0 - Python NLP Library for Many Human Languages. *GitHub* : веб-сайт. URL: <https://stanfordnlp.github.io/stanfordnlp/> (дата звернення: 3.12.2019).
8. Named Entity Recognition in 140 Characters or Less / Geyer K. et al. Lexington MA, United States : MIT Lincoln Laboratory, 2016. 2 p. URL: [https://www.ll.mit.edu/sites/default/files/publication/doc/2018-05/2016\\_04\\_11\\_GreenfieldK\\_WWW\\_FP2.pdf](https://www.ll.mit.edu/sites/default/files/publication/doc/2018-05/2016_04_11_GreenfieldK_WWW_FP2.pdf) (дата звернення: 3.12.2019).
9. Schuller B., A. Mousa Contextual Bidirectional Long Short-Term Memory Recurrent Neural Network Language Models. *A Generative Approach to Sentiment Analysis* : Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. Valencia, Spain : Association for Computational Linguistics, 2017. P. 1023–1032. URL: <https://www.aclweb.org/anthology/E17-1096/> (дата звернення: 10.11.2019).
10. Publication Manual of the American Psychological Association. Washington, D.C. : American Psychological Association, 2019. – 439 p.
11. Banerjee S. Introduction to Named Entity Recognition. *Medium* : веб-сайт. URL: <https://medium.com/explore-artificial-intelligence/introduction-to-named-entity-recognition-eda8c97c2db1> (дата звернення: 1.12.2019).
12. MITIE: library and tools for information extraction. *GitHub* : веб-сайт. URL: <https://github.com/mit-nlp/MITIE> (дата звернення: 3.11.2019).
13. Дьомкін В. Анонс lang-uk: створюємо умови для повноцінної обробки україномовних текстів. *DOU* : веб-сайт. URL: <https://dou.ua/lenta/columns/lang-uk/> (дата звернення: 3.11.2019).
14. Бібліографічний запис. Бібліографічний опис. Загальні вимоги та правила складання : ДСТУ ГОСТ 7.1:2006. Вид. офіц. - Вперше (зі скасуванням ГОСТ 7.1-84, ГОСТ 7.16-79, ГОСТ 7.18-79, ГОСТ 7.34-81, ГОСТ 7.40-82) ; введ. 2007-07-01. К. : Держспоживстандарт України, 2007. III, 47 с.
15. ДСТУ 8302-2015. Інформація та документація. Бібліографічне посилання. Загальні положення та правила складання – Чинний від 2016-07-01. К. : ДП «УкрНДНЦ», 2016. 16 с.
16. Friedl E. F. Mastering Regular Expressions: Understand Your Data and Be More Productive. Sebastopol, California, United States : O'Reilly Media, 2012. 544 p.
17. Regular expressions 101. *regex101* : веб-сайт. URL: <https://regex101.com> (дата звернення: 3.12.2019).
18. NER annotation corpus. *lang.org.ua* : веб-сайт. URL: <http://lang.org.ua/en/corpora/> (дата звернення: 3.10.2019).
19. Countries States Cities Database. *Github* : веб-сайт. URL: <https://github.com/search?p=2&q=cities+countr&type=Repositories> (дата звернення: 7.11.2019).