ДЕРЖАВНИЙ УНІВЕРСИТЕТ
ІНФОРМАЦІЙНО-КОМУНІКАЦІЙНИХ ТЕХНОЛОГІЙ
НАВЧАЛЬНО-НАУКОВИЙ ІНСТИТУТ
ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ
КАФЕДРА ІНЖЕНЕРІЇ ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ

КВАЛІФІКАЦІЙНА РОБОТА

на тему: «Розробка методики формування метаданих про заклади позашкільної освіти на основі технології NLP»

на здобуття освітнього ступеня магістра
зі спеціальності 121 Інженерія програмного забезпечення
(код, найменування спеціальності)
освітньо-професійної програми «Інженерія програмного забезпечення»
(назва)

Кваліфікаційна робота містить результати власних досліджень. Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело

_____ Юрій АБРОСКІН
(підпис)

Виконав: здобувач вищої освіти групи ПДМ-64
Юрій АБРОСКІН

Керівник: Оксана ЗОЛОТУХІНА
к.т.н., доцент

Рецензент: _____
науковий ступінь, вчене звання *Ім'я, ПРІЗВИЩЕ*

Київ 2024

ДЕРЖАВНИЙ УНІВЕРСИТЕТ
ІНФОРМАЦІЙНО-КОМУНІКАЦІЙНИХ ТЕХНОЛОГІЙ

Навчально-науковий інститут інформаційних технологій

Кафедра Інженерії програмного забезпечення

Ступінь вищої освіти Магістр

Спеціальність 121 Інженерія програмного забезпечення

Освітньо-професійна програма «Інженерія програмного забезпечення»

ЗАТВЕРДЖУЮ
Завідувач кафедри
Інженерії програмного забезпечення
_____ Ірина ЗАМРІЙ
«_____» _____ 2023 р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

Аброскіну Юрію Юрійовичу

1. Тема кваліфікаційної роботи: «Розробка методики формування метаданих про заклади позашкільної освіти на основі технології NLP»

керівник кваліфікаційної роботи Оксана ЗОЛОТУХІНА к.т.н., доцент,

затверджені наказом Державного університету інформаційно-комунікаційних технологій від «19» жовтня 2023 р. №145.

2. Строк подання кваліфікаційної роботи «29» грудня 2023 р.

3. Вихідні дані до кваліфікаційної роботи: науково-технічна література, документація бібліотек python, веб-сторінки закладів позашкільної освіти.

4. Зміст розрахунково-пояснювальної записки (перелік питань, які потрібно розробити)

1.Дослідження предметної області веб-ресурсів закладів позашкільної освіти.
2.Аналіз методів Natural Language Processing.
3.Розробка методики формування метаданих на основі обробки природної мови.

5. Перелік графічного матеріалу: *презентація*

1. Формати представлення метаданих.
2. Види метаданих про заклад позашкільної освіти.
3. Структура XML-файлу з метаданими про заклад позашкільної освіти.
4. Методика формування метаданих.
5. Визначення повноти та точності витягнутих метаданих.

6. Дата видачі завдання«19» жовтня 2023 р.

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів кваліфікаційної роботи	Строк виконання етапів роботи	Примітка
1	Аналіз наявної науково-технічної літератури	19.10-05.11.23	
2	Аналіз інформаційних ресурсів про заклади позашкільної освіти	06.11-12.11.23	
3	Аналіз методів Natural Language Processing	13.11-19.11.23	
4	Аналіз особливостей формування метаданих	20.11-26.11.23	
5	Визначення критеріїв відбору метаданих	27.11-29.11.23	
6	Розробка моделі формування метаданих	30.11-04.12.23	
7	Розробка методики формування метаданих на основі NLP	05.12-11.12.23	
8	Оцінка ефективності впровадженого методу	12.12-14.12.23	
9	Оформлення роботи: вступ, висновки, реферат	15.12-20.12.23	
10	Розробка демонстраційних матеріалів	21.12-29.12.23	

Здобувач вищої освіти _____ Юрій АБРОСКІН
(підпис)

Керівник
кваліфікаційної роботи _____ Оксана ЗОЛОТУХІНА
(підпис)

РЕФЕРАТ

Текстова частина кваліфікаційної роботи на здобуття освітнього ступеня магістра: 67 стор., 9 табл., 19 рис., 31 джерел.

Мета роботи – спростити процес формування метаданих про заклади позашкільної освіти на основі методів обробки природної мови..

Об'єкт дослідження – процес формування метаданих про заклади позашкільної освіти.

Предмет дослідження – методи та засоби вилучення інформації з текстових даних.

Короткий зміст роботи: У роботі проведено аналіз науково-технічної літератури за темами Natural Language Processing та метадані. Проаналізовано веб-ресурси закладів позашкільної освіти. Проаналізовані методи NLP, досліджено алгоритм NER. Були сформовані критерії відбору даних про заклади позашкільної освіти. Розроблено формальну модель та на її основі було створено блок-схему методу. Розроблено методику формування метаданих на основі технологій обробки природної мови.

КЛЮЧОВІ СЛОВА: NLP, NATURAL LANGUAGE PROCESSING, ОБРОБКА ПРИРОДНОЇ МОВИ, NER, NAMED ENTITY RECOGNITION, ВИЗНАЧЕННЯ ІМЕНОВАНИХ СУТНОСТЕЙ, МЕТАДАНІ.

ABSTRACT

Text part of the master's qualification work: 67 pages, 19 pictures, 9 table, 31 sources.

The purpose of the work – simplify the process of generating metadata about out-of-school.

Object of research – the process of generating metadata about out-of-school educational institutions.

Subject of research – methods and tools for extracting information from text data.

Summary of the work: In today's digital world, an important task is to automate and simplify the processes of generating metadata about out-of-school educational institutions. The study is aimed at developing a methodology based on natural language processing (NLP) technology for the effective extraction and structuring of information about educational institutions.

The educational process is defined as a key element in the formation of a personality, and the school curriculum aims to provide basic knowledge in various subjects. Out-of-school institutions act as an additional opportunity to gain educational knowledge that can expand the school curriculum or explore specific areas of study in greater depth. According to the official statistics of the Ministry of Education and Science of Ukraine, at the beginning of 2021, there were 1351 out-of-school educational institutions. It is important to note that in the current context, out-of-school education institutions also serve to provide children with the opportunity to escape the influence of the military information space.

The object of the study is the process of forming metadata about out-of-school education institutions, and the subject is methods and tools for extracting information from text data. The methodology is divided into three blocks: pre-processing, extraction of named entities, and metadata generation.

The first block uses an algorithm for cleaning HTML pages, tokenization, stop word extraction, and text lemmatization. These methods help to separate useful information from the HTML code and cut out unnecessary words, which facilitates further operations.

The second block includes an algorithm for recognizing and extracting named entities with an additional subject area dictionary to accurately extract relevant data. The use of a specially prepared dictionary improves the accuracy of identifying named entities, ensuring adequate metadata generation.

The third block implements the process of generating and storing metadata in XML format. This allows structuring information about out-of-school education institutions in a convenient and standardized way.

The developed methodology simplifies the process of generating metadata about out-of-school education institutions and allows processing a large amount of information in a short time. The use of NLP technology ensures effective recognition and analysis of text data, facilitating automation and unification of processes. The developed methodology can be used as a basis for further research in the field of natural language processing and metadata generation.

KEYWORDS: NLP, NATURAL LANGUAGE PROCESSING, NATURAL LANGUAGE PROCESSING, NER, NAMED ENTITY RECOGNITION, DEFINITION OF NAMED ENTITIES, METADATA.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ 10
ВСТУП 11
РОЗДІЛ 1 АНАЛІЗ ПІДХОДІВ ДО ФОРМУВАННЯ МЕТАДАНИХ ПРО ЗАКЛАДИ ПОЗАШКІЛЬНОЇ ОСВІТИ 13
1.1 Аналіз інформаційних ресурсів про заклади позашкільної освіти 13
1.1.1 Особливості доступу до інформації про заклади позашкільної освіти 13
1.1.2 Аналіз структури та вмісту контенту інформаційних ресурсів про заклади позашкільної освіти 19
1.2 Аналіз методів Natural Language Processing для обробки контенту веб-сторінок 23
1.2.1 Методи попередньої обробки текстових даних 23
1.2.2 Методи вилучення інформації про сутності та їх зв’язки 27
1.3 Аналіз особливостей формування метаданих 29
1.3.1 Використання метаданих для опису інформації про об’єкти 29
1.3.2 Формати представлення метаданих 35
1.3.3 Методи та інструменти формування метаданих інтернет-ресурсів 40
РОЗДІЛ 2 РОЗРОБКА МЕТОДИКИ ФОРМУВАННЯ МЕТАДАНИХ ПРО ЗАКЛАДИ ПОЗАШКІЛЬНОЇ ОСВІТИ НА ОСНОВІ ТЕХНОЛОГІЇ NLP 44
2.1 Визначення критеріїв відбору метаданих 44
2.2 Розробка моделі формування метаданих 47
2.3 Розробка методу формування метаданих на основі NLP 50
2.4 Опис інструментарію, методи, які будуть використовуватись у проекті ... 56
РОЗДІЛ 3 ЕКСПЕРИМЕНТАЛЬНЕ ДОСЛІДЖЕННЯ ЕФЕКТИВНОСТІ МЕТОДИКИ ФОРМУВАННЯ МЕТАДАНИХ ПРО ЗАКЛАДИ ПОЗАШКІЛЬНОЇ ОСВІТИ НА ОСНОВІ ТЕХНОЛОГІЇ NLP 58
3.1 Оцінка ефективності впровадженого методу 58
3.2 Екрані форми додатку 61
ВИСНОВКИ 67
ПЕРЕЛІК ПОСИЛАНЬ 68
ДЕМОНСТРАЦІЙНІ МАТЕРІАЛИ (Презентація) 73

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ

NLP	-	Natural Language Processing
NER	-	Named Entity Recognition
HTML	-	HyperText Markup Language
XML	-	eXtensible Markup Language
JSON	-	JavaScript Object Notation
CSV	-	Comma-Separated Values
RDF	-	Resource Description Framework
DCMI	-	Dublin Core Metadata Initiative

ВСТУП

Освітній процес визначається як суттєвий фактор у формуванні особистості, а шкільна програма спрямована на надання основних знань з загальних предметів, позашкільні заклади виступають як доповнення до цього процесу, надаючи можливість здобути додаткові знання, що можуть виходити за рамки шкільної програми або ж детально вивчати конкретні напрямки навчання. Станом на перше січня 2021 з офіційної статистики МОН України нараховується 1351 закладів позашкільної освіти. На даний момент заклади позашкільної освіти також є одним з факторів, який дозволяє дитині відволіктись від воєнного інформаційний простір.

Робота є надзвичайно актуальна у сучасному інформаційному суспільстві, де великий обсяг даних генерується кожною секундою. Використання технологій обробки природної мови (NLP) та розпізнавання іменованих сутностей (NER) має величезний потенціал у різних областях, включаючи обробку текстової інформації, аналіз великих даних, пошукові системи та багато іншого.

Застосування NLP дозволяє розуміти та обробляти природну мову, що відкриває широкі можливості для автоматизації обробки текстової інформації. Розпізнавання іменованих сутностей (NER) у сучасному світі має вагому значущість. У контексті метаданих, які формуються, важливо розуміти, як роль відіграють NLP та NER. Вони можуть покращити якість та повноту зібраних даних, забезпечуючи точні та релевантні результати.

Таким чином завдання на розробку методу формування метаданих на основі обробки природної мови є сучасним та актуальним.

Мета роботи – спростити процес формування метаданих про заклади позашкільної освіти на основі методів обробки природної мови.

Об’єкт дослідження – процес формування метаданих про заклади позашкільної освіти.

Предмет дослідження – методи та засоби вилучення інформації з текстових даних.

Практична значущість результатів полягає в використанні розробленої методи для спрощення процесу формування метаданих про заклади позашкільної освіти. З можливістю використання методики для інших потреб: аналізу, складання статистики тощо. .

Для досягнення мети вирішувались наступні завдання.

1. Аналіз інформаційних веб-ресурсів про заклади позашкільної освіти
2. Аналіз науково-технічної літератури про NLP, NER та метадані.
3. Дослідження методів обробки природної мови та вилучення іменованих сутностей.
4. Розробка критеріїв відбору метаданих.
5. Розробка формальної моделі методу формування метаданих.
6. Розробка програмного додатку формування метаданих на основі технологій обробки природної мови.
7. Визначення повноти та точності витягнутих метаданих на основі обробки природної мови.

1 АНАЛІЗ ПІДХОДІВ ДО ФОРМУВАННЯ МЕТАДАНИХ ПРО ЗАКЛАДИ ПОЗАШКІЛЬНОЇ ОСВІТИ

1.1 Аналіз інформаційних ресурсів про заклади позашкільної освіти

1.1.1 Особливості доступу до інформації про заклади позашкільної освіти

Заклади позашкільної освіти є ключовими установами, юридичними особами, які забезпечують надання позашкільної освіти.

Відповідно до статті 1 Закону України «Про позашкільну освіту» «заклад позашкільної освіти – складова системи позашкільної освіти, яка надає знання, формуючи вміння та навички за інтересами, забезпечує потреби особистості у творчій самореалізації та інтелектуальний, духовний і фізичний розвиток, підготовку до активної професійної та громадської діяльності, створює умови для соціального захисту та організації змістовного дозвілля відповідно до здібностей, обдарувань та стану здоров’я вихованців, учнів і слухачів» [1].

Аналіз інформаційних ресурсів про заклади позашкільної освіти показує, що вони представлені різними даними, що відображають різноманітність та розгалуженість системи позашкільної освіти.

Перш за все, інформаційні ресурси про заклади позашкільної освіти представлені різними відомчими підпорядкуваннями.

Варто відзначити, що заклади позашкільної освіти функціонують у системі освіти, системі культури та системі спорту.

Відповідно дані про заклади позашкільної освіти подано за різним відомчим підпорядкуванням, а саме:

- Міністерство освіти і науки України;
- Міністерство молоді та спорту України;
- Міністерство культури та інформаційної політики України.

Інформація про заклади позашкільної освіти на інформаційному ресурсі Міністерства освіти і науки України (рис. 1.1).

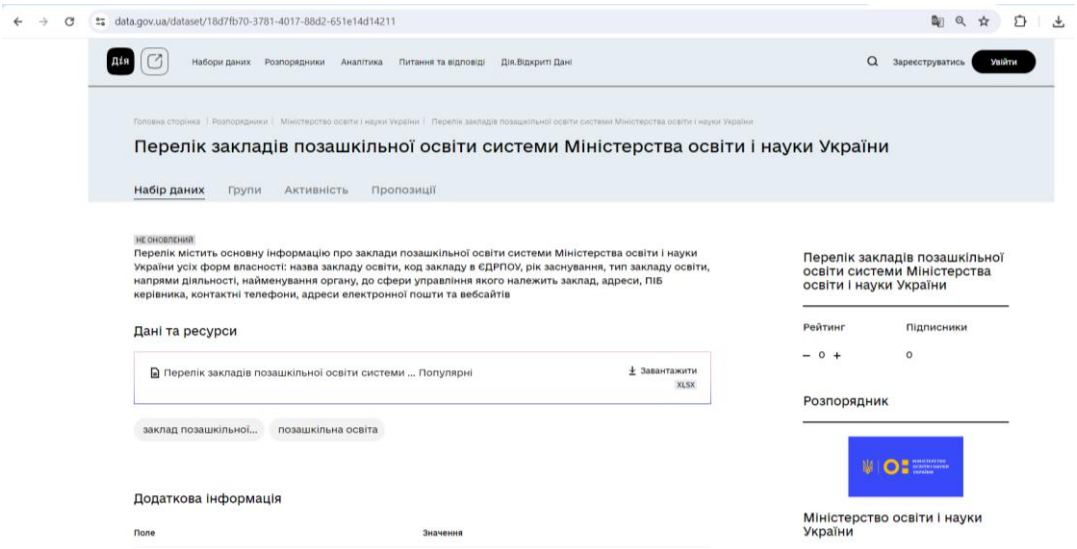


Рис. 1.1. Заклади позашкільної освіти на сайті Міністерства освіти і науки України

Інформація про заклади позашкільної освіти на інформаційному ресурсі Міністерства молоді та спорту України (рис. 1.2).

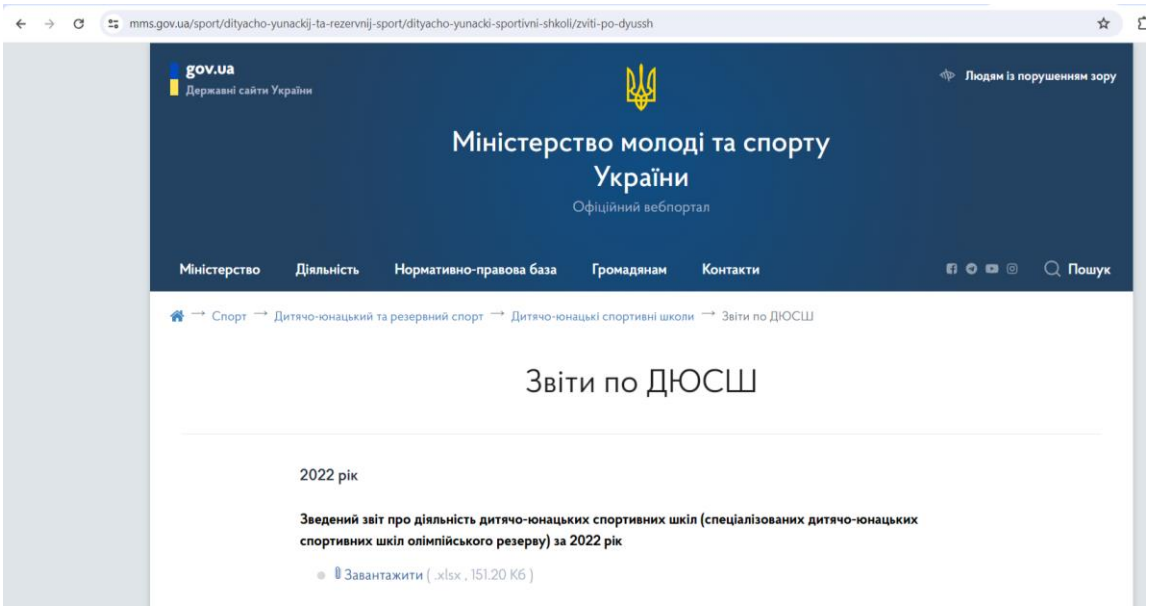


Рис. 1.2. Заклади позашкільної освіти на сайті Міністерства молоді та спорту України

Інформація про заклади позашкільної освіти на інформаційному ресурсі Міністерства культури та інформаційної політики України (рис. 1.3).

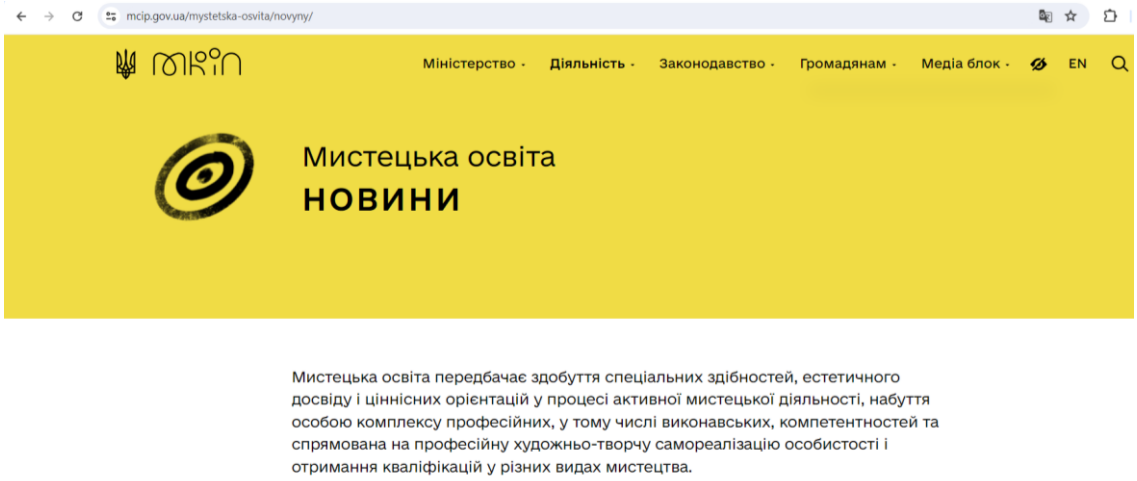


Рис. 1.3. Заклади позашкільної освіти на сайті Міністерства культури та інформаційної політики України

По друге, інформаційні ресурси про заклади позашкільної освіти представлена різними типами закладів позашкільної освіти.

Відповідно до постанови Кабінету Міністрів України від 06.05.2001 р. № 433 «Про затвердження переліку типів позашкільних навчальних закладів і Положення про позашкільний навчальний заклад», заклади позашкільної освіти поділяються на різні типи, а саме:

«1. Дитячо-юнацькі спортивні школи: комплексні дитячо-юнацькі спортивні школи, дитячо-юнацькі спортивні школи з видів спорту, дитячо-юнацькі спортивні школи для осіб з інвалідністю, спеціалізовані дитячо-юнацькі школи олімпійського резерву, спеціалізовані дитячо-юнацькі спортивні школи для осіб з інвалідністю паралімпійського та дефлімпійського резерву.

2. Клуби: військово-патріотичного виховання, дитячо-юнацькі (моряків, річковиків, авіаторів, космонавтів, парашутистів, десантників, прикордонників, радистів, пожежників, автолюбителів, краєзнавців, туристів, етнографів, фольклористів, фізичної підготовки та інших напрямів).

3. Мала академія мистецтв (народних ремесел).

4. Мала академія наук учнівської молоді.

5. Оздоровчі заклади для дітей та молоді: дитячо-юнацькі табори (містечка, комплекси): оздоровчі, заміські, профільні, праці та відпочинку, санаторного типу, з денним перебуванням; туристські бази.

6. Мистецькі школи: музична, художня, хореографічна, хорова, школа мистецтв тощо.

7. Центр, палац, будинок, клуб художньої творчості дітей, юнацтва та молоді, художньо-естетичної творчості учнівської молоді, дитячої та юнацької творчості, естетичного виховання.

8. Центр, будинок, клуб еколого-натуралістичної творчості учнівської молоді, станція юних натуралістів.

9. Центр, будинок, клуб науково-технічної творчості учнівської молоді, станція юних техніків.

10. Центр, будинок, клуб, бюро туризму, краєзнавства, спорту та екскурсій учнівської молоді, туристсько-краєзнавчої творчості учнівської молоді, станція юних туристів.

11. Центри: військово-патріотичного та інших напрямів позашкільної освіти.

12. Дитяча бібліотека, дитяча флотилія моряків і річковиків, дитячий парк, дитячий стадіон, дитячо-юнацька картинна галерея, дитячо-юнацька студія (хорова, театральна, музична, фольклорна тощо), кімната школяра, курси, студії, школи мистецтв, освітньо-культурні центри національних меншин» [2].

Таким чином, кожен з даних типів закладу позашкільної освіти має свої інформаційні ресурси.

По третє, різними є організаційно-правові форми закладів позашкільної освіти – державна, комунальна та приватна й відповідне їх представлення.

Так, інформація про державні заклади позашкільної освіти представлена на інформаційних ресурсах органів державної влади (рис. 1.4).

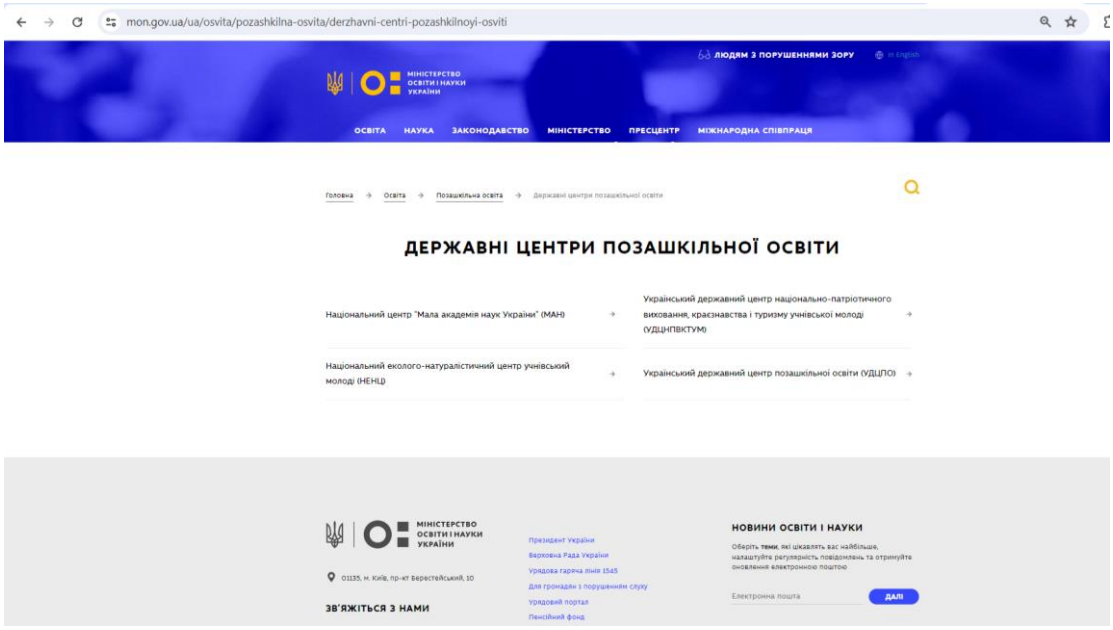


Рис. 1.4. Державні заклади позашкільної освіти на інформаційному ресурсі органу державної влади

Інформація про комунальні заклади позашкільної освіти представлена на інформаційних ресурсах департаментів / управлінь освіти і науки / культури / молоді та спорту (рис. 1.5).

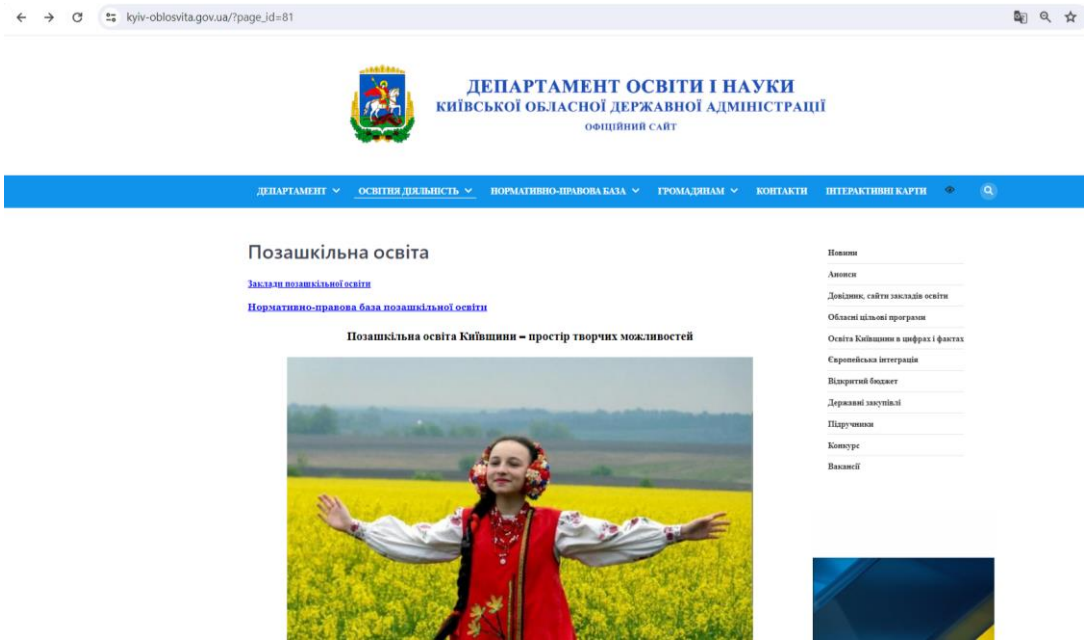


Рис. 1.5. Комунальні заклади позашкільної освіти на інформаційному ресурсі обласного департаменту освіти і науки

Варто відзначити, що інформація про приватні заклади позашкільної освіти представлена на власних інформаційних ресурсах (рис. 1.6).

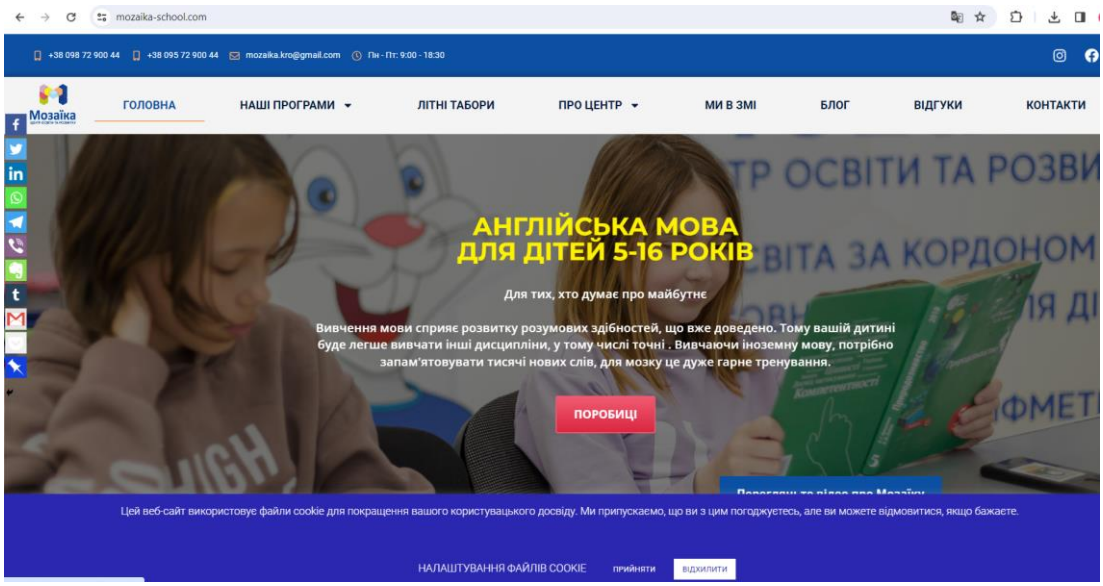


Рис. 1.6. Приватні заклади позашкільної освіти на інформаційному ресурсі на власному інформаційному ресурсі

Таким чином, аналіз інформаційних ресурсів закладів позашкільної освіти показав їх різноманітність, водночас не зовсім структурованість для пошуку широкого загалу.

У наш час дуже важливо мати відкритий доступ до інформації про заклади позашкільної освіти з багатьох причин.

По-перше, це дозволяє батькам та учням знайти та зрозуміти, які саме заклади пропонують позашкільну освіту, які гуртки та активності доступні, і чим вони є унікальними. Це допомагає спрямувати вихованців ,учнів, слухачів до тих гуртків та активностей, які цікаві саме їй і що можуть допомогти розвивати потенціал кожної дитини.

По-друге, відкритий доступ до інформації допомагає порівнювати різні заклади позашкільної освіти та їх пропозиції. Батьки можуть зрозуміти, які можливості та ресурси надаються в конкретному закладі позашкільної освіти. Також зможуть оцінити якість навчання, репутацію педагогів, особливості форм та

методів навчання. Це допоможе батькам зробити освітній вибір, який відповідає потребам і цілям їхньої дитини.

По-третє, наявність відкритого доступу до інформації сприяє розвитку конкуренції між закладами позашкільної освіти. Конкуренція змушує заклади удосконалювати свою роботу, покращувати якість навчання. Це стимулює інновації та покращення якості освіти в цілому.

Отже, у наш час доступ до інформації про заклади позашкільної освіти є дуже важливим. Він допомагає батькам та учням знаходити найкращі можливості для навчання, виховання, розвитку та соціалізації, порівнювати та обирати найбільш зручні заклади позашкільної освіти, а також сприяє розвитку конкуренції та покращенню якості позашкільної освіти.

1.1.2 Аналіз структури та вмісту контенту інформаційних ресурсів про заклади позашкільної освіти

Під час пошуку інформації про заклади позашкільної освіти, було проаналізовано структуру та вміст контенту різних інформаційних ресурсів, а саме:

– Портал «Дія» [3].

– Веб-ресурс Департаменту освіти і науки виконавчого органу Київської міської ради [4].

– Веб-ресурс Центру бенчмаркінгу та веб-менеджменту (ЦБВМ) [5].

– Веб-ресурс «Позашкільна освіта», проект «Карта закладів позашкільної освіти» [6].

Варто відзначити, що на порталі «Дія» представлено перелік закладів позашкільної освіти системи Міністерства освіти і науки України. Особливістю переліку є інформація про назви закладу позашкільної освіти, телефон, електронну адресу, веб-сайт, керівника. Водночас даний перелік подано у вигляді таблиці Excel (рис. 1.7).

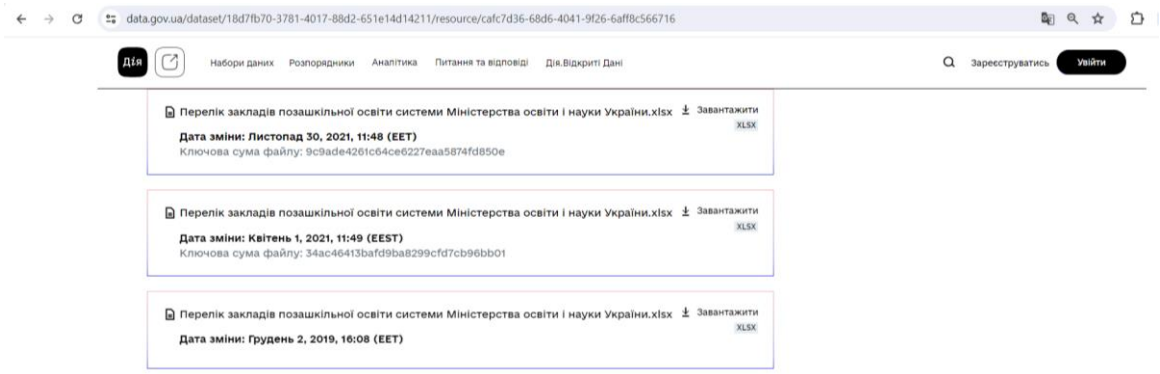


Рис. 1.7. Інформація про заклади позашкільної освіти на порталі «Дія»

Аналіз доступу до інформації про заклади позашкільної освіти на сайті Департаменту освіти і науки виконавчого органу Київської міської ради показав, що зібрано декілька закладів позашкільної освіти м. Києва, поділених за районами (рис. 1.8).

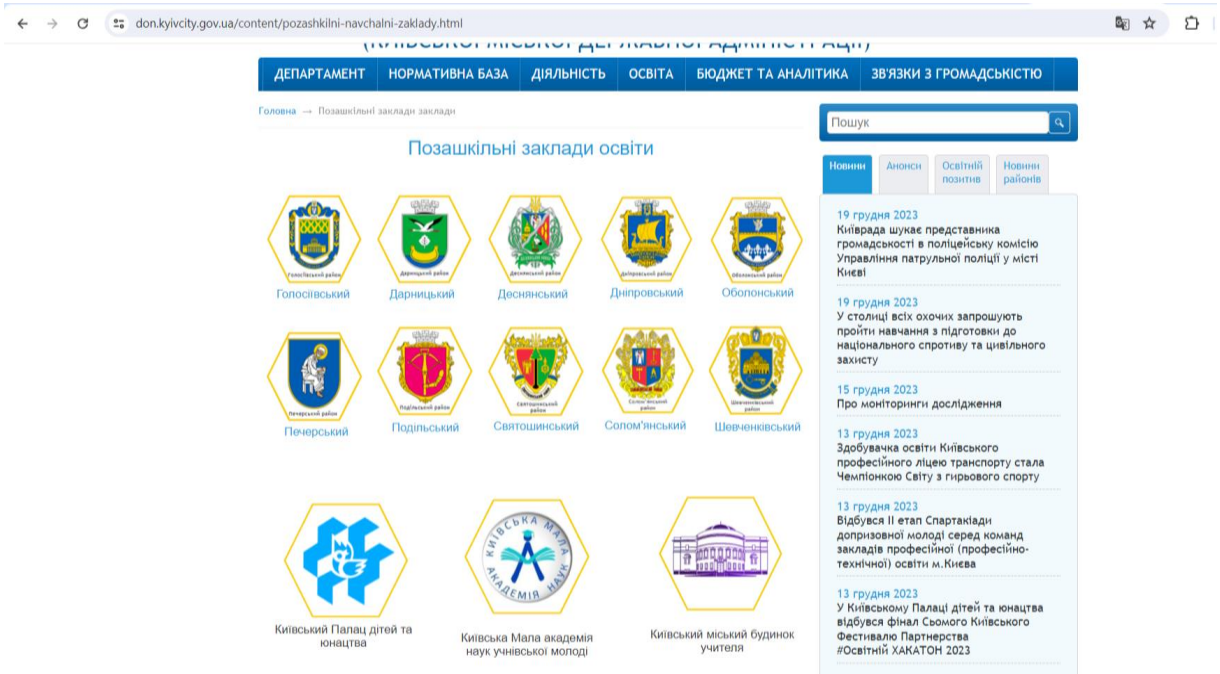
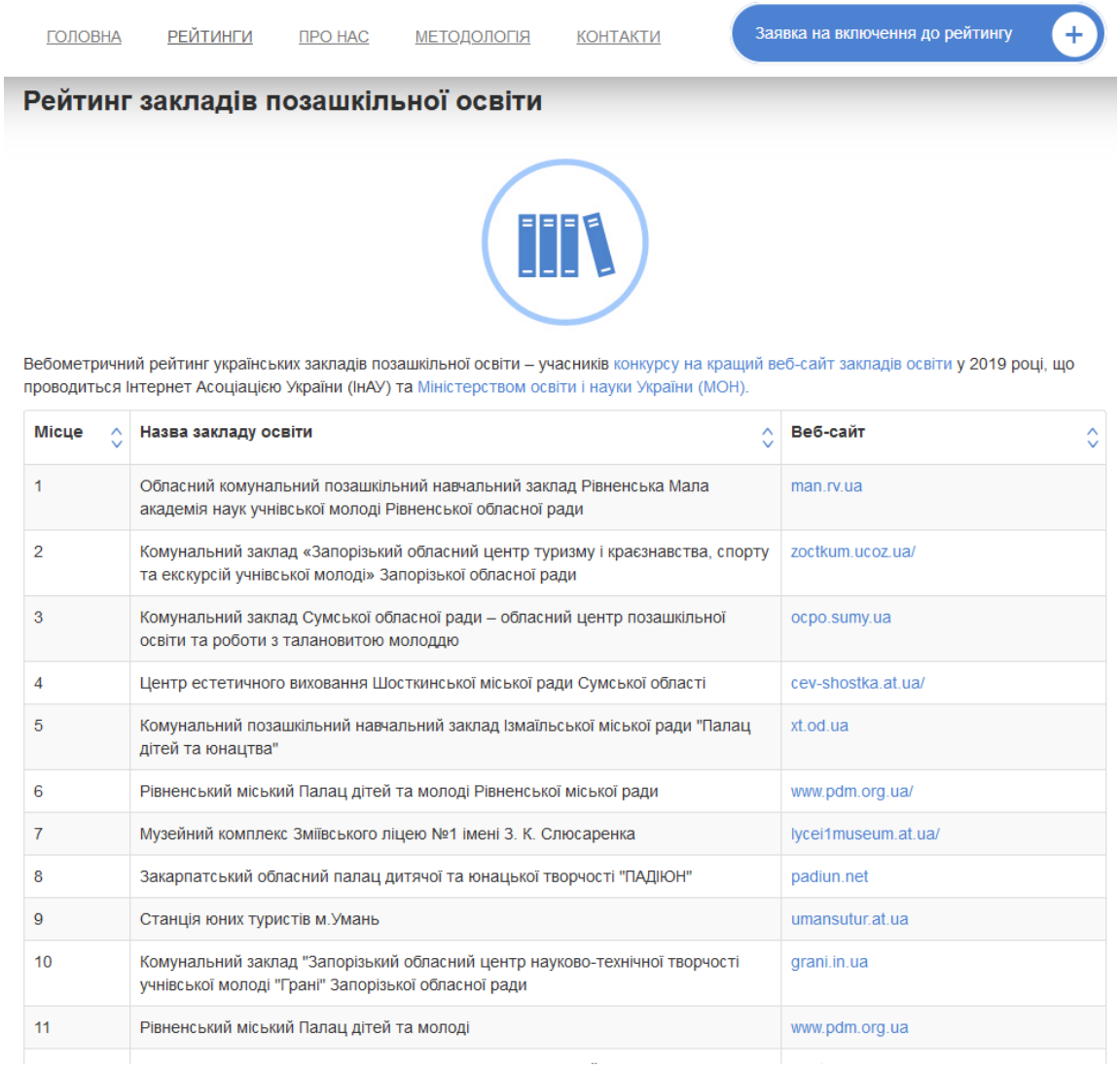


Рис. 1.8. Інформація про заклади позашкільної освіти на веб-ресурсі Департаменту освіти і науки виконавчого органу Київської міської ради

Встановлено, що на сайті Департаменту освіти і науки Київської міської ради зазначені лише назви закладів, у деяких з них є посилання на веб-сторінку або

сторінки у соціальних мережах, деякі зовсім не актуальні і перенаправляють на помилкові веб-сторінки.

Особливістю доступу до інформації про заклади позашкільної освіти на веб-ресурсі Центру бенчмаркінгу та веб-менеджменту є список учасників, які брали участь у конкурсі «X Всеукраїнський конкурс на кращий веб-сайт закладу освіти». На ресурсі зібрано 114 учасника впорядкованих відповідно зайнятому місцю (рис. 1.9).



ГОЛОВНА РЕЙТИНГИ ПРО НАС МЕТОДОЛОГІЯ КОНТАКТИ Заявка на включення до рейтингу

Рейтинг закладів позашкільної освіти

Вебометричний рейтинг українських закладів позашкільної освіти – учасників конкурсу на кращий веб-сайт закладів освіти у 2019 році, що проводиться Інтернет Асоціацією України (ІнАУ) та Міністерством освіти і науки України (МОН).

Місце	Назва закладу освіти	Веб-сайт
1	Обласний комунальний позашкільний навчальний заклад Рівненська Мала академія наук учнівської молоді Рівненської обласної ради	man.rv.ua
2	Комунальний заклад «Запорізький обласний центр туризму і краєзнавства, спорту та екскурсій учнівської молоді» Запорізької обласної ради	zoctkum.ucoz.ua/
3	Комунальний заклад Сумської обласної ради – обласний центр позашкільної освіти та роботи з талановитою молоддю	ocpo.sumy.ua
4	Центр естетичного виховання Шосткинської міської ради Сумської області	cev-shostka.at.ua/
5	Комунальний позашкільний навчальний заклад Ізмаїльської міської ради "Палац дітей та юнацтва"	xt.od.ua
6	Рівненський міський Палац дітей та молоді Рівненської міської ради	www.pdm.org.ua/
7	Музейний комплекс Зміївського ліцею №1 імені З. К. Слюсаренка	lycei1museum.at.ua/
8	Закарпатський обласний палац дитячої та юнацької творчості "ПАДІЮН"	padiun.net
9	Станція юних туристів м.Умань	umansutur.at.ua
10	Комунальний заклад "Запорізький обласний центр науково-технічної творчості учнівської молоді "Грані" Запорізької обласної ради	grani.in.ua
11	Рівненський міський Палац дітей та молоді	www.pdm.org.ua

Рис. 1.9. Інформація про заклади позашкільної освіти на веб-ресурсі Центру бенчмаркінгу та веб-менеджменту

Доцільно відзначити, що сайт з рейтингом надає посилання до усіх закладів учасників, але варто брати до уваги, те що цей ресурс пропонує заклади з найкращим зовнішнім виглядом веб-сторінки закладу. Враховуючи це, не всі заклади можуть відповідати бажаним вимогам споживача.

Відзначимо, що на веб-ресурсі «Позашкільна освіта» представлено проект «Карта закладів позашкільної освіти». Присутня карта України, поділена на області, де на кожну область є свій список закладів, які доступні у відповідній області (рис. 1.10).

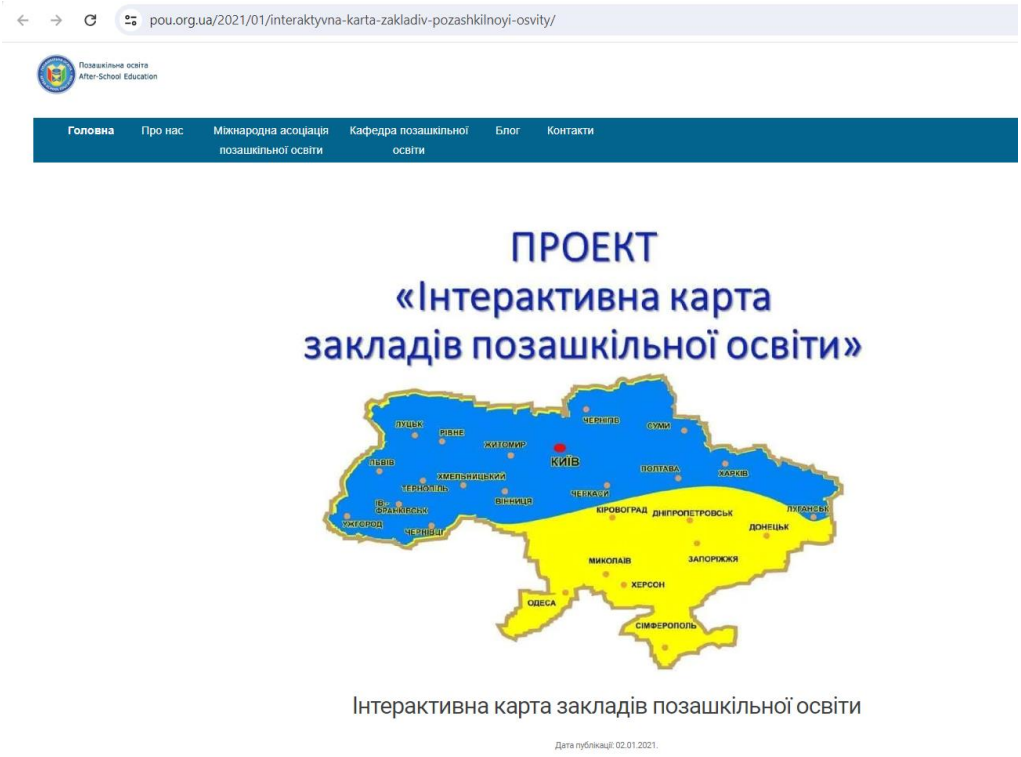


Рис. 1.10. Інформація про заклади позашкільної освіти на веб-ресурсі «Позашкільна освіта», проект «Карта закладів позашкільної освіти»

Встановлено, що даний веб-ресурс може виступати як «каталог», оскільки є список закладів позашкільної освіти, поділених за областями України.

Наведена інформація стосовно обраного закладу також є вичерпною, бо надається місце розташування закладу, почтовий індекс, ПІБ керівника, контактний номер телефону, посилання на соціальну мережу, електрону поштову скриню, веб-сторінку (за наявністю). Водночас, аналіз показав, що так само, як у

випадку з сайтом Київської міської ради, більшість посилань перенаправляють на сторінки, які не працюють або на видалені групи у соціальних мережах.

Таким чином, аналіз інформаційних ресурсів про заклади позашкільної освіти показав, що інформація є недостатньою. Це потребує розробки методики формування метаданих про заклади позашкільної освіти на основі технології NLP.

1.2 Аналіз методів Natural Language Processing для обробки контенту веб-сторінок

1.2.1 Методи попередньої обробки текстових даних

Завдяки штучному інтелекту наше життя значно стало простіше, на даний етап розвитку ми можемо за рахунок лише голосових команд отримати потрібну нам інформації, і це все стає доступним завдяки Natural Language Processing.

Обробка природної мови (Natural Language Processing, NLP)[7] - це галузь комп'ютерних наук, яка розробляє та досліджує методи взаємодії комп'ютера з людською мовою. NLP має широкий спектр застосувань, наприклад: перефразувань тексту, переклад тексту, відповідь на запитання по тексту та підвести короткий підсумок про що йдеться в тексті.

Свій початок обробка природної мови бере з 40-х років 20 століття, коли перед науковцями стало питання в перекладі однієї мови на іншу. Спочатку це була машина, яка за допомогою словників інших мов перекладала слова на відповідну мову та розставляла перекладені слова у відповідному порядку до правил мови. Результатом стала невдача, після чого Хомський впровадив поняття «Синтаксичної структури», яке дало початок ідеї генеративної граматики. Саме після цього стали з’являтися різні напрямки обробки природної мови, одна з них це «Розпізнавання мови». Тоді науковці розділились на два табори, одні вірили у успіхи теоретичної генеративної граматики, а інші у успіх статистичної інформації, це були спільноти які займаються мовною обробкою та мовленням, відповідно.

Як зазначено у роботі Елізабет Д. Лідді, в 1966 році було зроблено висновок, що досягти, на той час, машинного перекладу, який би було неможливо відрізнити

від перекладу людей – не можливо і тому було прийнято рішення не фінансувати цей напрям. Що призвело до припинення розвитку машинного перекладу і інших робіт у сфері обробки природної мови. У наступні 20 років було дослідження у напрямі обробки природної мови, але найбільший вплив та зріст у цьому напрямі відбувся під кінець 80-х та початок 90-х, коли з’явились комп’ютери здатні оброблювати значну більше інформації ніж раніше. Також важливим фактором став доступ до інтернету та великої кількості тексту.

Аналізуючи методи NLP була досліджена робота Костянтина Онищенка, Яни Данієль та Романа Каменєва «Аналіз Методів Обробки Природної Мови»[8]. В даній роботі розглянуті підходи та методи, що використовуються при задачах оброки природної мови. Нижче наведено методи NLP, які розглядались у роботі:

1. Глибинне навчання в NLP – розділ машинного навчання, більша частина технологій NLP функціоную завдяки глибокому навчанню, яке базується на використанні декількох шарів розпізнавання вхідних даних. Завдяки збільшеної кількості даних, які можна надати для тренування та збільшена продуктивності обчислювальних пристроїв.
2. Вкладання слів - методологія NLP, яка відображає слова або словосполучення у вигляді векторів дійсних чисел. В основі вкладання слів лежить ідея того, що слова, які часто зустрічаються поруч або мають подібний контекст у текстах, отримують векторні представлення, які знаходяться близько одне до одного. Ці вектори можуть бути використані для розв'язання різноманітних завдань у сфері обробки природної мови, таких як машинний переклад, аналіз настрою, розпізнавання іменованих сутностей та інші.
3. Машинний переклад – метод перетворення тексту з однієї природної мови в текст зрозумілий носію іншої мови. Тобто машина чи програма замінює у поданому тексті слова або словосполучення з однієї мови на слова та словосполучення іншої мови які будуть відповідати значенню та сенсу цих слів. У такому перекладі використовується аналіз статистики вживання слів у контексті.

4. Голосові помічники – метод що описувалось на початку підрозділу. Завдяки методам розпізнаванню голосу та обробкам природної мови надається можливість використовувати команди голосових помічників за для пошуку інформації, прослуховування та відтворення конкретних ключових слів та фільтрація зовнішнього шуму.
5. Система «Питання-Відповідь» - метод отримує інформацію у вигляді запитів, документів, зображень тощо. В результаті видає відповідь на поставлену йому задачу у вигляді декількох речень. Більшість задач в NLP можна класифікувати таким чином: ми надаємо запит та очікуємо коротку відповідь.
6. Стислий переказ тексту – метод який скорочено передає вхідні дані, відфільтровуючи найважливіше за допомогою відкидання речень в яких слова не так частіше повторюються ніж в інших.

Веб-сторінки містять велику кількість даних переважно в текстовому форматі. Для вилучення великих об’ємів та використання цієї інформації можна використовувати методи NLP, що є більш раціональний підходом ніж використовувати людський ресурс та вилучати інформацію ручними методами.

Для обробки контенту веб-сторінки можна використовувати наступні методи NLP:

1. Метод вилучення інформації – завдяки даному методу будуть витягуватись інформація з веб-сторінки у вигляди неструктурованого тексту
2. Розпізнавання мови – дана методика дізнається на якій мові написаний текст веб-сторінки, це потрібно щоб правильно обробляти вхідний текст: підібрати словники та граматичні правила
3. Розпізнавання частин мови – завдяки даному методу визначається частина мови кожного слова, це дає більш правильне розуміння контексту та значення речення
4. Синтаксичний аналіз – методика визначення структури речення. Це необхідно для визначення відносин між словами, в свою чергу це допоможе краще оброблювати та аналізувати текст

5. Семантичний аналіз
6. Аналіз настроїв
7. Методи попередньої обробки тексту
8. Метод вилучення іменованих сутностей

Для того щоб розпочати оброблювати або аналізувати текстові дані їх потрібно піддати процесу попередньої обробки. Вона потрібна для того щоб позбутися розділових знаків, посилань та стоп-слів, а також оброблення текст для подальшого впливу методів NLP таких як виявлення іменованих сутностей.

Зазвичай процес попередньої обробки тексту поділяють за наступними пунктами:

1. Видалення HTML тегів – якщо сторінка HTML використовується в якості вхідних даних, то для подальшої обробки тексту треба прибрати теги які використовує мова HTML, щоб не було не потрібних та зайвих слів.
2. Переведення у нижній регістр– перетворення всіх слів у нижній регістр за для однорідність та узгодженість, не дозволяючи моделі розглядати слова з різними регістрами як різні сутності.
3. Видалення знаків пунктуації та спеціальних символів
4. Видалення URL посилань.
5. Токенізація – розбивання тексту на речення або одразу на слова. Дозволяє розуміти структуру тексту перебираючи його менш значущими одиницями.
6. Обробка чисел – в залежності від задачі числа або прибираються зовсім, або замінюються на словникові аналоги.
7. Видалення стоп-слів – видалення загальних слів, які не вкладають значного зміст в тексті. Це потрібно щоб позбавитись від слів, які можуть навантажити текст зайвою інформацією.
8. Перевірка орфографії – виправлення потрібно при можливих орфографічних помилках аби текст був точніший, зрозуміліший для подальшого використання.

9. Лематизація та/або стемінг – дані терміни виконують майже однакові процес. Стемінг скорочую слова до їх кореневих форм, тобто відкидає закінчення та суфікс. Лематизація в свою чергу зводить слово до початкової форми. Обираючи між цими двома методами лематизація буде більш точнішою ніж стемінг, через те що визначення леми відбувається за допомогою відповідних словників, коли в свою чергу стемінг якби вкорочує слово, а не зводить його до спільно кореневого, хоча це може бути.

Дані кроки зроблять попередню обробку тексту, звісно можна долучати будь які інші методи обробки, або навпаки вилучати не потрібні методи. Це все залежіть від конкретних потреб при попередній обробці тексту та подальших маніпуляцій з обробленими даними.

1.2.2 Методи вилучення інформації про сутності та їх зв’язки

Інформація про сутності є невід'ємною складовою у сфері NLP та аналізу текстових даних. Термін "сутність" в цьому контексті вказує на конкретні об'єкти чи поняття, які можуть бути виділені та ідентифіковані в тексті. Інформація про сутності включає в себе розпізнавання іменованих сутностей (Named-entity recognition, NER)[9] та взаємозв'язків між ними, розширюючи розуміння контексту та структури текстових даних.

Метою збирання та аналізу інформації про сутності є отримання глибшого розуміння тексту, виявлення ключових об'єктів, осіб, місць, організацій та подій, що відображені в текстовому матеріалі.

В процесі аналізу текстових даних методами NLP, важливо розуміти, як інформація про сутності може бути використана для покращення обробки та розуміння контенту. Визначення необхідних сутностей та їх взаємодій стає ключовим етапом у розробці ефективних алгоритмів обробки тексту та вилученні цільової інформації.

У контексті методів вилучення інформації про сутності та їх зв'язки, важливо розуміти різні типи сутностей та їх взаємовідносини. Сутності, які можна виділити

завдяки алгоритмам та моделям, можуть бути поділені на кілька категорій, загальні категорії зазначені на таблиці 1.1

Таблиця 1.1

Загальні категорії іменованих сутностей

Абревіатури	Категорія	Опис
PER	Люди	Включає в себе інформацію про конкретних осіб, таку як їх імена, прізвища, титули, посади, а також зв'язані атрибути, такі як дата народження чи професійні досягнення.
LOC	Місця	Відображає географічні об'єкти. Це може бути інформація про країни, міста, регіони, географічні об'єкти або точні локації, зокрема географічні координати.
ORG	Організації	Включає в себе дані про різні організаційні структури, такі як компанії, установи, урядові агентства, неприбуткові організації, та інші групи.
EVENT	Події	Представляє конкретні події, такі як конференції, виставки, відзначення в часі, наукові або культурні заходи. Вони можуть включати дати, час та місце проведення.
PRODUCT	Продукти	Включає в себе дані продуктів, товарів, послуг, брендів, технологій, або інших об’єктів, що можуть бути ідентифіковані в тексті.
MISC	Інше	Включає інші сутності, які не належать до попередніх категорій. Це може бути числа, електронні адреси, або будь-яка інша інформація, яка не відповідає чітко визначеним категоріям.

Три найбільш поширені системи розпізнавання іменованих сутностей[9] включають наступне:

Системи на основі навчання з учителем – використовують моделі машинного навчання, навчені на текстах, які люди попередньо позначили категоріями

іменованих сутностей. Для таких систем використовуються алгоритми, такі як умовні випадкові поля та максимальна ентропія — дві складні статистичні мовні моделі. Цей метод ефективний для розбору семантичних значень та інших складнощів, хоча він вимагає великих обсягів тренувальних даних.

Системи на основі правил – використовують правила для вилучення інформації. Правила можуть включати в себе капіталізацію або титули, такі як "Dr." Цей метод вимагає великої кількості людського втручання для введення, відстеження та виправлення правил, і він може пропускати текстові варіації, які не включені у тренувальні анотації. Вважається, що системи на основі правил не долучаються до обробки складнощів на рівні моделей машинного навчання.

Системи на основі словників – використовують словник із великим словниковим запасом та збіркою синонімів для перевірки та ідентифікації іменованих сутностей. Цей метод може мати проблеми при класифікації іменованих сутностей з варіаціями в написанні.

1.3 Аналіз особливостей формування метаданих

1.3.1 Використання метаданих для опису інформації про об’єкти

Зростаючий обсяг інформації в інтернеті і не тільки вимагає ефективних методів організації та класифікації контенту для забезпечення користувачам точного та швидкого доступу до необхідних даних. Одним із ключових елементів цього процесу є метадані, які визначають основні характеристики об'єктів та їх зв'язки.

Метадані[10] – це поглиблена інформація, що може бути структурованою або не структурованою, яка описую детально об’єкт або інші дані про цей об’єкт. Дану інформацію можна використовувати для пошуку, сортування, аналізу. Як приклад на рис. 1 зазначені метадані звичайного документа Word описують автора файлу, коли був створений, ким відредагований, кількість сторінок, слів, рядків, і тощо. Переглянути наведену інформацію можна за допомогою провідника Windows.

В свою чергу дуже велику роль метадані відіграють у сфері бібліотекознавство. В роботі Н. В. Стрішенець «Метадані у сучасному бібліотекознавстві. Метадані – нове чи старе поняття?»[11], де розглянуто вплив слова «метадані» на американську сферу бібліотекознавства та поширення терміну в українській сфері. Як зазначено в роботі, термін метаданих проник з інформаційних технологій, а саме з управлінь баз даних.

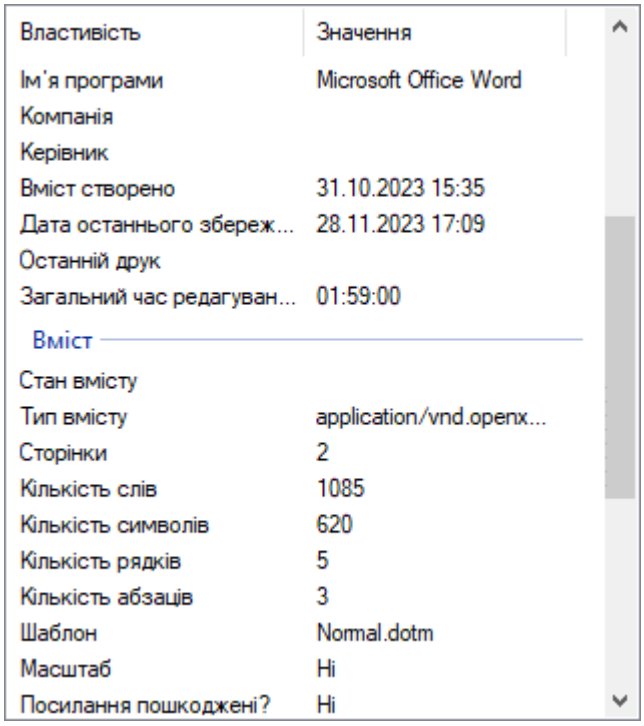


Рис. 1.11. Приклад метаданих Word файлу

Типи метаданих можна класифікувати за різними критеріями. Один із найпоширеніших критеріїв - це функціональний аспект метаданих. За цим критерієм метадані можна поділити на такі типи:

Описові метадані - описують зміст набору даних. Вони використовуються для полегшення пошуку, розуміння та використання даних. Наприклад, описові метадані для книги можуть включати в себе такі поля:

Таблиця 1.2

Можливі атрибути описових метаданих

Назва поля	Короткий опис
Автор	Ім’я автора файлу або документа
Назва	Назва файлу або документа
Рік видання	Рік в якому був виданий, створений файл або документ
Жанр	Жанр в якому написаний даний файл або документ
Ключові слова	Ключові слова за якими можна шукати даний файл або документ

Адміністративні метадані - описують управління набором даних. Вони використовуються для відстеження змін в наборі даних, для визначення прав доступу до даних тощо. Наприклад, адміністративні метадані для книги можуть включати в себе такі поля:

Таблиця 1.3

Можливі атрибути адміністративних метаданих

Назва поля	Короткий опис
Ідентифікатор	Ідентифікаційний номер за яким можна відстежувати файл або документ
Дата створення	Дата створення файлу або документа у форматі дд.мм.рр
Дата останнього оновлення	Дата останнього оновлення файлу або документа, якщо дана дія відбувалась
Власник	Ім’я особи, яка створила файл або документ
Хто оновлював інформацію	Ім’я особи, яка оновлювала інформацію файлу або документа, якщо дана дія відбувалась

Структурні метадані - описують структуру набору даних. Вони використовуються для розуміння взаємозв'язків між даними. Наприклад, структурні метадані для книги можуть включати в себе такі поля:

Таблиця 1.4

Можливі атрибути структурних метаданих

Назва поля	Короткий опис
Розділи	Кількість розділів файлу або документа
Підрозділи	Кількість підрозділів файлу або документа
Сторінки	Кількість сторінок файлу або документа

Мовні метадані - описують мову, на якій написаний набір даних. Вони використовуються для забезпечення можливості перекладу даних на інші мови. Наприклад, мовні метадані для книги можуть включати в себе такі поля:

Таблиця 1.5

Можливі атрибути мовних метаданих

Назва поля	Короткий опис
Мова	Мова якою написаний файл або документ
Переклади	Перелік перекладів доступних для файлу або документу

Технічні метадані - описують технічні аспекти набору даних. Вони використовуються для управління набором даних та забезпечення його сумісності з іншими системами. Наприклад, технічні метадані для книги можуть включати в себе такі поля:

Таблиця 1.6

Можливі атрибути технічних метаданих

Назва поля	Короткий опис
Формат файлу	Назва формату в якому зберігається файл або документ
Розмір файлу	Розмір збереженого файлу або документа у кілобайтах
Кодування	Назва кодування за яким відбувається кодування символів файлу або документа

Юридичні метадані - описують правові аспекти набору даних. Вони використовуються для захисту прав інтелектуальної власності та забезпечення дотримання авторських прав. Наприклад, юридичні метадані для книги можуть включати в себе такі поля:

Таблиця 1.7

Можливі атрибути юридичних метаданих

Назва поля	Короткий опис
Авторські права	Правовий статус файлу або документа, ім’я власника
Ліцензії	Ліцензійні умови за якою поширюється файл або документ
Умови використання	Умови за якими дозволяється використання файлу або документа

Інший критерій класифікації метаданих - це ступінь структурованості метаданих. За цим критерієм метадані можна поділити на такі типи:

- Структуровані метадані мають певну структуру. Вони можуть бути представлені у вигляді таблиць, списків або інших ієрархічних структур. Структуровані метадані легше обробляти та аналізувати, ніж неструктуровані метадані.
- Неструктуровані метадані не мають певної структури. Вони можуть бути представлені у вигляді текстових документів, зображень або інших неструктурованих даних. Неструктуровані метадані складніше обробляти та аналізувати, ніж структуровані метадані.

1.3.2 Формати представлення метаданих

Формат метаданих – це стандартизований спосіб представлення та структурування інформації, призначеної для опису ресурсів або даних. В іншому контексті, це набір правил та конвенцій, які визначають, як конкретна категорія інформації повинна бути представлена з метою ефективного обміну та обробки. Формат метаданих визначає структуру запису, типи даних, правила ідентифікації та взаємодії, що дозволяє послідовно та стандартно передавати ключові характеристики ресурсу чи даних. Кожен формат метаданих має свої унікальні особливості, специфікації та призначення, що робить його важливим елементом в контексті оптимізації обробки та використання інформації.

Найбільш поширеними форматами метаданих є RDF, JSON, XML, CSV, DCMI[12, 13].

Формат RDF (Resource Description Framework) є стандартом для вираження метаданих та взаємозв'язків між ресурсами у веб-середовищі. У RDF дані представлені у вигляді триплетів, що включають суб'єкт, предикат та об'єкт. Цей гнучкий формат дозволяє створювати зв'язки між різнорідними даними, що робить його ідеальним для визначення семантичних взаємозв'язків. Використання RDF сприяє інтеграції та обміну даними між різними джерелами, забезпечуючи стандартизований підхід до опису ресурсів.

JSON є легким та зрозумілим форматом обміну даними, що використовується для структуризації інформації у вигляді ключ-значення. Він дозволяє представляти метадані у зручній формі, забезпечуючи читабельність та легкість обробки програмами.

XML використовується для визначення та структуризації даних у текстовому форматі. Він надає гнучкість у визначенні власних тегів та атрибутів, що робить його ефективним для опису різноманітної інформації, включаючи метадані.

CSV представляє дані у вигляді табличної структури, де значення розділені комою. Цей формат ефективний для простих структур та легко інтегрується з багатьма програмами обробки даних.

DCMI визначає стандартний набір метаданих для опису ресурсів, що використовуються в інтернет-середовищі. Ключові елементи Dublin Core включають назву, автора, дату, що забезпечує консистентність та стандартизацію в описі цифрових ресурсів.

Таблиця 1.8

Недоліки та переваги загальних форматів метаданих

Назва формату	Переваги	Недоліки
RDF	Забезпечує гнучкість у визначенні та розширенні семантичних зв'язків між ресурсами; Стандартизований формат дозволяє ефективну інтеграцію даних з різних джерел; Підтримує семантичний веб та розширює можливості метаданих.	Деяка складність в роботі може виникнути для користувачів без глибокого розуміння семантичних технологій; Зберігання та передача великих обсягів даних може призвести до значної величини файлів.

Продовження таблиці 1.8

Недоліки та переваги загальних форматів метаданих

Назва формату	Переваги	Недоліки
JSON	Легка читабельність для людей і ефективність для машин; Підтримка складених структур даних.	Займає більше місця в порівнянні з іншими форматами; Може виникати проблема з безпекою при використанні ненадійних джерел.
XML	Велика гнучкість у визначенні структури; Підтримка вбудованих метаданих.	Займає більше місця через теги; Менш ефективний у порівнянні з іншими форматами для обробки великих об'ємів даних.
CSV	Простота та легкість інтеграції; Ефективний для представлення табличних даних.	Не підтримує складені структури; Може виникнути непорозуміння при розділенні значень комами.
DCMI	Простий у використанні та розумінні, що полегшує його впровадження; Використовується в різних галузях; Легко інтегрується з іншими метаданими та стандартами	Для деяких конкретних випадків опису ресурсів може бути недостатнім у забезпеченні детальності; Не завжди придатний для вирішення складних метаданих завдань

Також існують інші формати, які спеціалізуються у конкретному напряму файлів[12, 13, 14, 15, 16]:

- BIBFRAME, формати сімейства MARC, MODS
- Теги, метатеги
- GILS , EAD
- vCard, FOAF
- CDWA
- PRISM, ONIX
- CIF
- VICAR
- NewsXML
- EXIF (Exchangeable Image File Format)
- ID3 (методані)
- ISO 19115
- PREMIS (Preservation Metadata Implementation Strategies)
- EPIcore (Event, Place, Individual, Object, Role, Expression)
- METS (Metadata Encoding and Transmission Standard)
- PREMIS Rights Markup Language (RiML)
- DCAT (Data Catalog Vocabulary)
- RDA (Resource Description and Access)
- TEI (Text Encoding Initiative)
- IIIF (International Image Interoperability Framework)

BIBFRAME (Bibliographic Framework), MARC (Machine-Readable Cataloging), та MODS (Metadata Object Description Schema) є форматами для опису бібліографічної інформації. BIBFRAME є сучаснішим форматом, MARC — традиційним, а MODS є більш гнучким та розширеним.

Теги в мовах розмітки (наприклад, HTML, XML) використовуються для визначення елементів документа. Метатеги, зазвичай, вказують метадані про документ, такі як мова, автор чи ключові слова.

GILS (Government Information Locator Service) — стандарт для управління та локалізації інформації урядових ресурсів. EAD (Encoded Archival Description) — формат для опису архівних зібрань, зокрема, використовується для опису фондів архівів.

vCard — стандарт для обміну контактною інформацією, використовується для представлення персональних метаданих.

FOAF (Friend of a Friend) — формат для визначення осіб та їх взаємодій у соціальних мережах.

CDWA (Categories for the Description of Works of Art) — стандарт для опису художніх творів, включаючи художні стилі, матеріали та техніки.

PRISM (Publishing Requirements for Industry Standard Metadata) — стандарт для обміну метаданими у видавничій індустрії.

ONIX (Online Information Exchange) — формат для обміну метаданими про книги та інші видання.

CIF (Crystallographic Information File) — формат для представлення кристалографічних даних.

VICAR (Video Image Communication and Retrieval) — формат для обробки та обміну візуальної інформації, часто використовується у космічних програмах.

NewsXML — формат для представлення новинових статей та інформації.

EXIF — стандарт для вбудованого зберігання метаданих у фотографіях, включаючи інформацію про камеру, налаштування та дату зйомки.

ID3 — формат для вбудованого зберігання метаданих у звукових файлів, таких як MP3. Містить інформацію про виконавця, назву треку, жанр і т. д.

ISO 19115 визначає стандарт для метаданих геопросторової інформації. Включає в себе елементи, які описують просторові дані, їхню точність, тематичний зміст та контекст.

PREMIS (Preservation Metadata Implementation Strategies) визначає стандарт для метаданих зберігання та збереження цифрових об'єктів. Включає інформацію про формати, правила зберігання та стратегії забезпечення довгострокового доступу.

EPIcore (Event, Place, Individual, Object, Role, Expression) це стандарт для метаданих в області ділової документації, який визначає структуру для опису подій, місць, осіб, об'єктів, ролей та виразів.

METS (Metadata Encoding and Transmission Standard) визначає стандарт для представлення метаданих та структури цифрових об'єктів. Використовується для опису зв'язків між різними частинами складних цифрових об'єктів.

PREMIS Rights Markup Language (RiML) — це розширення стандарту PREMIS, яке визначає метадані для управління правами доступу та використання цифрових об'єктів.

DCAT (Data Catalog Vocabulary) визначає формат для метаданих каталогів даних. Використовується для опису наборів даних, їх властивостей та зв'язків між ними.

RDA (Resource Description and Access) визначає стандарт для метаданих, що використовуються в бібліотечній галузі для опису та доступу до ресурсів, включаючи книги, періодичні видання та електронні ресурси.

TEI (Text Encoding Initiative) визначає формат для метаданих, використовуваних для текстового кодування та анотації літературних та історичних текстів.

IIIF (International Image Interoperability Framework) визначає метадані для стандартізації представлення та обміну зображень в цифрових бібліотеках та музеях.

1.3.3 Методи та інструменти формування метаданих інтернет-ресурсів

Існує два основних методи формування метаданих інтернет-ресурсів: ручний і автоматичний:

Ручний метод передбачає введення метаданих людиною. Цей метод може бути утомливим і трудомістким, але він дозволяє забезпечити високу точність і якість метаданих. Ручний метод часто використовується для формування метаданих для специфічних ресурсів, наприклад, для наукових публікацій, навчальних матеріалів або корпоративних ресурсів.

Автоматичний метод передбачає використання програмного забезпечення для формування метаданих. Цей метод може бути більш ефективним, ніж ручний, але він може призвести до зниження якості метаданих, якщо програмне забезпечення не буде добре налаштовано. Автоматичний метод часто використовується для формування метаданих для великих наборів даних, наприклад, для веб-сторінок або файлів зображень.

У роботі «Використання метаданих у сучасному світі» за авторством Ю. О. Мазура і Н. А. Потапова [13] розглядається цей аспект та наводяться важливі засоби для отримання різноманітної інформації про медіа-файли.

Автори вказують, що звичайний смартфон може служити ефективним засобом отримання різноманітних метаданих, таких як тип, розмір, час створення, параметри камери, географічні координати та інші. Зазначається, що навіть користувач, який робить фотографії, може бути непримітно залучений до збору значущої інформації.

У роботі приводяться приклади онлайн-сервісів для визначення метаданих, зокрема Jeffrey’s Exif Viewer та Pic2Map. Перший з них дозволяє отримати різні дані про зображення, такі як EXIF дані, MakerNotes, JFIF тощо. Другий сервіс використовує дані EXIF для локації та відображення фотографій на мапі.

Додатково, в роботі наводяться інформація про застосунки, які дозволяють редагувати метадані. MetaCleaner та MetaClean є прикладами таких інструментів, які надають можливість видалення метаданих з файлів та забезпечують зручний інтерфейс для користувача.

Висвітлюючи практичний аспект, автори роботи також надають інформацію про вигоди використання метаданих у сучасному світі, які можуть використовуватися для різноманітних цілей, включаючи оцінку фотографій та їхній подальший аналіз.

Існує також ряд інструментів, які можуть використовуватися для формування метаданих інтернет-ресурсів. До таких інструментів відносяться:

- Редактори метаданих - це програмні продукти, призначені для ручного введення метаданих.
- Автоматичні генератори метаданих - це програмні продукти, призначені для автоматичного формування метаданих.
- Метадонні каталоги - це веб-ресурси, які містять набори метаданих для різних типів ресурсів.

Редактори метаданих дозволяють користувачам вводити метадані для конкретних ресурсів. Ці редактори можуть бути простими або складними, залежно від того, які типи метаданих вони підтримують. Деякі популярні редактори метаданих включають:

- MODS Editor - це редактор метаданих для формату MODS (Metadata Object Description Schema).
- Dublin Core Metadata Editor - це редактор метаданих для формату Dublin Core.
- EAD Editor - це редактор метаданих для формату EAD (Encoded Archival Description).

Автоматичні генератори метаданих використовують алгоритми для автоматичного формування метаданих для ресурсів. Ці алгоритми можуть використовуватися для різних типів ресурсів, наприклад, для веб-сторінок, файлів зображень або файлів PDF. Деякі популярні автоматичні генератори метаданих включають:

- Apache Tika - це інструмент для автоматичного розпізнавання і структурування документів.
- ExifTool - це інструмент для читання і запису метаданих в файлах зображень.
- PDFMetaExtractor - це інструмент для читання і запису метаданих в файлах PDF.

Метадонні каталоги містять набори метаданих для різних типів ресурсів. Метадонні каталоги можуть використовуватися для пошуку метаданих для конкретних ресурсів. Деякі популярні метадонні каталоги включають:

- Europeana - це метадоний каталог, який містить метадані для культурних ресурсів з усієї Європи.
- Linked Data Platform - це метадоний каталог, який містить метадані для ресурсів, які пов'язані між собою за допомогою технології Linked Data.
- DataCite - це метадоний каталог, який містить метадані для наукових публікацій.

Вибір методу і інструменту формування метаданих залежить від конкретних потреб і вимог. Для формування метаданих для специфічних ресурсів, які вимагають високої точності і якості, може бути доцільно використовувати ручний метод. Для формування метаданих для великих наборів даних може бути доцільно використовувати автоматичний метод.

2 РОЗРОБКА МЕТОДИКИ ФОРМУВАННЯ МЕТАДАНИХ ПРО ЗАКЛАДИ ПОЗАШКІЛЬНОЇ ОСВІТИ НА ОСНОВІ ТЕХНОЛОГІЇ NLP

2.1 Визначення критеріїв відбору метаданих

В даній роботі дуже важливо визначитись з тією інформацію з якою буде проводитись робота. Так як метою роботи є покращити процес формування метаданих про заклади позашкільної освіти, обробка інформацію відбувається в відповідній галузі освіти. Для пошуку інформації про заклади позашкільної освіти зараз використовують декілька способів:

- Пошук в інтернеті
- Рекомендації знайомих
- Оголошення у загально освітніх закладах
- Оголошення на вулицях
- Оголошення у соціальних мережах
- Сюжетні новини на телебаченні

В контексті даної роботи, будуть розглянуті способи, які пов’язані з доступом до мережі інтернету. Бо саме завдяки метаданим надається можливість спростити та поліпшити процес пошуку та аналізу в інтернеті. Для швидкого ознайомлення з інформацію про заклад можна виділити наступні критерії, котрі будуть важливими для осіб шукаючих та бажаючих ознайомитись з закладом позашкільної освіти:

1. Назва закладу
2. Місце розташування
3. Спеціалізація та напрямки діяльності
4. Кількість учнів та педагогічний склад
5. Інфраструктура
6. Фінансовий стан
7. Інноваційні технології та методи навчання
8. Історія та досягнення

9. Взаємодія з громадою
10. Активності та заходи

Визначення критеріїв для відбору метаданих, пов'язаних із закладами позашкільної освіти, представляє собою невід'ємну частину стратегічного планування та систематизації інформації про ці установи. Цей процес, який має на меті узагальнення та структурування ключових аспектів, виходячи за межі базового визначення метаданих та їхніх типів, є дієвим інструментом для створення докладного та вичерпного опису закладів освіти.

Зазначений вище критерій – назва закладу – виявляється основним елементом для встановлення унікальності та легкої ідентифікації різних освітніх установ. Ця інформація, як своєрідний "цифровий підпис" закладу, слугує не лише засобом внутрішнього класифікування, але й робить його легко розпізнаваним серед широкої громадськості та зацікавлених стейкхолдерів.

Важливим аспектом є місце розташування закладу, яке, в свою чергу, розкриває географічні особливості та можливості. Адреса та географічні координати надають можливість не лише визначити місцезнаходження, а й використовувати геопросторовий аналіз для отримання додаткових інсайтів щодо розподілу освітніх установ у конкретному регіоні.

Інший важливий критерій – це спеціалізація та напрямки діяльності закладу. Детальний опис освітніх послуг та програм дозволяє з'ясувати профіль закладу, що має важливе значення для визначення його ексклюзивних характеристик та акцентів.

Подальше розглядання таких ключових аспектів, як кількість учнів та педагогічний склад, є необхідним для отримання докладної карти розмірів та динаміки освітньої спільноти. Інформація про фінансовий стан закладу, включаючи бюджет та джерела фінансування, розширює обсяг аналізу та дозволяє здійснювати порівняльні оцінки між різними установами.

Надавання інформації про інфраструктуру закладу – це ще один крок у напрямку більш повного відображення освітнього середовища. Опис будівель, класів, спортивних майданчиків та інших приміщень надає не лише технічні дані,

але і створює віртуальний образ життя та умов, в яких здійснюється навчання та розвиток учнів.

Додавання розгорнутої інформації про інноваційні технології та методи навчання є ще одним важливим аспектом у визначенні метаданих. Вказівка на використання сучасних технологій та передових педагогічних підходів надає огляд професійного розвитку закладу та його відповідність сучасним освітнім стандартам. Це стає ключовим показником для батьків та учнів, які враховують такі фактори при виборі освітнього закладу.

Іншим напрямком розширення є опис історії та досягнень закладу. Зазначення попередніх успіхів, рейтингів та репутації додає додатковий вимір до зображення закладу і може служити одним із факторів, що впливають на вибір майбутніх учнів та їхніх батьків.

Необхідно також враховати активності та заходи, що відбуваються в закладі. Опис додаткових освітніх та культурних заходів, конкурсів, семінарів, створює образ динамічного та живого освітнього середовища, що дозволяє учням брати активну участь у різноманітних заходах поза навчанням.

У цілому, розширюючи інформацію про заклад позашкільної освіти за допомогою деталізованих метаданих, ми створюємо повний, багатогранний портрет освітнього закладу. Цей підхід дозволяє учням, батькам та зацікавленим особамам здійснювати обдумані та обґрунтовані вибори, враховуючи всі аспекти життя та діяльності установи.

Для створення власної методики було обрано наступні критерії формування метаданих зазначені на таблиці 2.1. Ці критерії були проаналізовані та обрані, як найбільш потрібна інформація відвідувачу сайту про заклад позашкільної освіти.

Таблиця 2.1

Види метаданих про заклад позашкільної освіти

Назва	Опис
Name	Назва закладу позашкільної освіти
Webpage	Посилання на веб-сторінку закладу
Address	Адреса знаходження закладу в місті
Director	ПІБ директора закладу
Group	Назва гуртка
Group leader	ПІБ керівник гуртка
Schedule	Розклад гуртка
Phone	Контактний телефон
Social_media	Посилання на сторінку закладу у соц. мережах

2.2 Розробка моделі формування метаданих

Завдяки сформованим критеріям метаданих надається можливість сформувати структуру метаданих. Для того щоб сформувати структуру та ієрархію майбутніх метаданих. Аби це зробити було обрано формат метаданих XML.

Обрання формату XML для формування метаданих у даній методиці обґрунтовано кількома важливими перевагами. XML (eXtensible Markup Language) є текстовим форматом, що легко розуміється як людьми, так і програмами, що сприяє його широкому використанню у сфері обміну даними.

Однією з ключових переваг XML є його структурованість. Формат дозволяє організовувати дані у вигляді деревоподібної структури, де інформація представлена в тегах і атрибутах, що спрощує її читання та обробку. Ця

структурованість особливо важлива у метаданих, де потрібно зберігати ієрархічні відносини між іменованими сутностями.

Ще однією перевагою XML є його розширюваність. Завдяки чому можна легко додавати нові елементи або атрибути до XML-документа без необхідності змінювати структуру вже існуючих даних. Це надає гнучкість у розробці та адаптації системи метаданих під змінні умови та вимоги.

Ще однією важливою характеристикою XML є його підтримка веб-технологій, зокрема можливість інтеграції безпосередньо у код сторінок HTML. Це означає, що його можна вбудовувати XML-структури безпосередньо у веб-сторінки, що полегшує взаємодію між метаданими та самим веб-контентом. Використання XML у веб-середовищі забезпечує велику зручність для обробки даних за допомогою JavaScript та інших веб-технологій.

Окрім цього, XML є стандартизованим форматом, що дозволяє легко обмінювати даними між різними системами та забезпечує їхню сумісність. Це особливо важливо, оскільки можливість обміну даними визначається високою сумісністю, що є ключовим аспектом у методиці для забезпечення спільної роботи та обробки метаданих у різних системах із різними конфігураціями та обмеженнями.

Обрана структура має наступний вигляд:

```
<Заклад_позашкільної_освіти>
    <Назва_закладу>Назва закладу позашкільної
освіти</Назва_закладу>
    <Гуртки>
      <Гурток>
        <Назва_гуртка>Назва гуртка</Назва_гуртка>
        <Керівник_гуртка>Ім'я керівника гуртка</Керівник_гуртка>
        <Телефон_керівника>Телефон керівника
гуртка</Телефон_керівника>
```

<Графік_проведення_гуртка>Графік проведення гуртка
</Графік_проведення_гуртка>
</Гурток>
</Гуртки>
</Заклад_позашкільної_освіти>

Враховуючи те що XML формат використовує в своєму описі теги та сам по собі є мовою розмітки. Тому в даній структурі присутній кореневий елементом яким виступає «Заклад позашкільної освіти», в нього присутні два атрибути «Назва закладу» та «Гуртки». У тегу «Гуртки» відповідно знаходяться елементи, які мають наступні атрибути: «Назва гуртка», «Ім'я керівника гуртка», «Телефон керівника гуртка», «Графік проведення гуртка». Як було вказано вище, дану структуру можна змінити та додати інші критерії, які також були зазначені вище у розділі 2.1, за необхідністю.

Також було створено формальну математичну модель (2.3) для подальшого представлення роботи метода. Нехай T - це вхідний текст (HTML-сторінка), тоді T' - очищений текст, та NE – вилучені іменовані сутності, розраховуються за наступними формулами (2.1) та (2.2) відповідно:

$$T' = Lemmatize\left(RemoveStopWords\left(Tokenize\big(RemoveTags(T)\big)\right)\right) \quad (2.1)$$

де
Lemmatize – алгоритм лематизації слів
RemoveStopWords – алгоритм прибирання шумових слів
Tokenize – алгоритм розбивання тексту на слова
RemoveTags – алгоритм прибирання тегів HTML, залишаючи лише текст

$$NE = NER(T', D) \quad (2.2)$$

де

NER - алгоритм розпізнавання іменованих сутностей, який використовує словник предметної галузі D.

D - словник предметної галузі.

Виходячи з цього, отримуємо загальну формулу:

$$XML\ File = SaveToXML\big(CreateMetadata(NE)\big) \tag{2.3}$$

2.3 Розробка методу формування метаданих на основі NLP

Розглянувши попередній розділ, ми зосередимося на ключових аспектах та інноваціях, що входять у створення ефективного та точного механізму вилучення та структурування інформації про заклади позашкільної освіти. Використання NLP в контексті формування метаданих стає важливим етапом для автоматизації процесів аналізу та обробки великої кількості текстової інформації. Завдяки розумінню природної мови та використанню високоточних алгоритмів, ми можемо досягти значного полегшення збору та систематизації даних про заклади позашкільної освіти.

У цьому розділі ми розглянемо кожен етап нашого методу докладно, висвітлимо призначення та деталі кожної етапної операції. Розкриємо стратегії використання NLP для виокремлення іменованих сутностей та аналізу текстових даних, щоб надати чітке розуміння того, як система працює та які переваги вона приносить у контексті формування метаданих.

Важливим аспектом розробки цього методу є його адаптивність та гнучкість. Ми розглянемо можливості розширення та модифікації алгоритмів з урахуванням специфічних потреб дослідника чи вимог конкретного дослідження.

Метадані для закладів позашкільної освіти відіграють важливу роль у поліпшенні управління та розкритті потенціалу освітніх інституцій. Вони надають цінну інформацію, яка може бути використана для різних цілей, сприяючи

оптимізації процесів та підвищенню якості навчання. Ось кілька способів, які метадані можуть допомогти закладам позашкільної освіти:

Організація та Систематизація Інформації - метадані дозволяють організувати та структурувати інформацію про заклади позашкільної освіти. Це включає дані про навчальні програми, графіки проведення гуртків, кадровий склад, інфраструктуру тощо. Систематизована інформація полегшує управління закладом та надає зручний доступ до важливих даних.

- Підвищення Видимості та Реклами - використання метаданих допомагає в підвищенні видимості закладу позашкільної освіти. Наприклад, інформація про різноманітні гуртки та навчальні програми може бути використана для створення ефективних рекламних кампаній та привертання уваги батьків та учнів.
- Аналіз та Оптимізація Процесів - метадані дозволяють проводити аналіз ефективності різних аспектів діяльності закладу. На основі цієї інформації можна впроваджувати оптимізації в роботі, забезпечуючи ефективніше використання ресурсів та підвищення якості навчання.
- Підтримка Процесу Прийняття Рішень - метадані служать як основа для прийняття обґрунтованих рішень. Наприклад, наявність даних про попит на конкретні гуртки чи програми дозволяє адміністрації закладу приймати рішення щодо розвитку та адаптації до потреб учнів.
- Забезпечення Транспарентності - публічні метадані про заклади позашкільної освіти сприяють забезпеченню транспарентності та взаємодії з громадою. Це важливо для батьків, які можуть отримати доступ до інформації та зробити інформований вибір щодо освітнього шляху своєї дитини.
- Взаємодія з Державними Системами - метадані можуть бути використані для взаємодії з державними системами освіти, що полегшує обмін інформацією та виконання адміністративних процедур.

Для покращення процесу формування метаданих про заклади позашкільної освіти, методика повинна збирати інформації з веб-ресурсу та отримувати ключові дані, які були вже визначені у таблиці 2.1. Далі ці дані повинні структуруватись у метадані. Як результат було створено блок-схему (Рис. 2.1.) формування метаданих про заклад позашкільної освіти на основі технологій обробки природної мови.

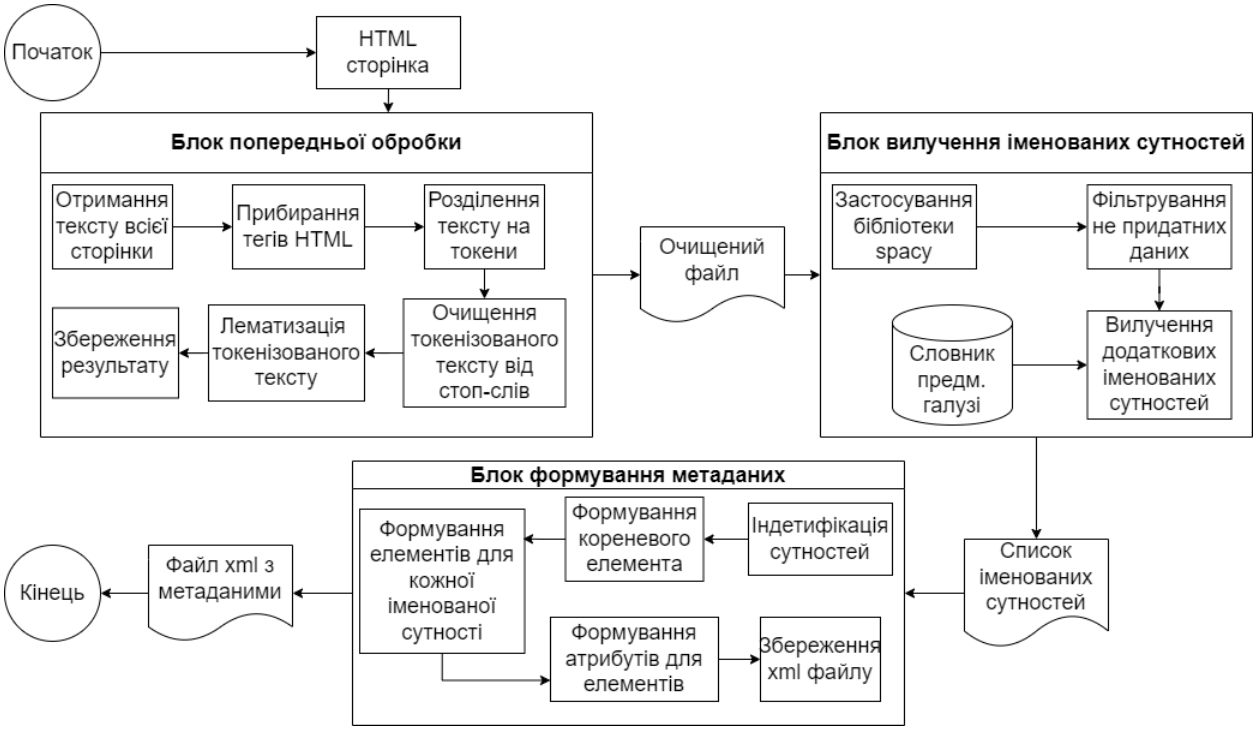


Рис. 2.1. Запропонована методика формування метаданих на основі технологій обробки природної мови

Запропоновану методику можна поділити на 3 блоки. Кожний блок є ключовим етапом, який повинен проходити послідовно.

1. Блок попередньої обробки тексту
2. Блок вилучення іменованих сутностей
3. Блок формування метаданих

У першому блоці розглядається один із ключових етапів методу формування метаданих – "попередньої обробки". Цей етап визначає основні кроки підготовки вхідного тексту для подальшого аналізу за допомогою технології обробки природної мови.

Попередня обробка тексту визначає невід'ємну та ключову частину процесу розробки методу формування метаданих на основі технології обробки природної мови (NLP). Цей етап є важливою передумовою для ефективного та точного аналізу текстової інформації. Нижче наведено головні аспекти цього процесу.

1. Видалення HTML-тегів:

Починаючи з видалення HTML-тегів, ми очищаємо текст від зайвих елементів та забезпечуємо однорідність джерела даних. Це дозволяє уникнути збоїв та спотворень під час подальшого аналізу та вилучення інформації.

2. Розподіл на токени:

Процес розподілу тексту на токени є першим кроком у створенні лінгвістичної одиниці для подальшого аналізу. Це необхідно для правильного визначення смислових одиниць та подальшої лематизації.

3. Видалення стоп-слів:

Видалення стоп-слів є важливим етапом для ефективного виділення суттєвих термінів. Це дозволяє уникнути врахування часто вживаних слів, які не несуть значущої семантичної ваги.

4. Лематизація:

Лематизація визначає базову форму слова та забезпечує уніфікацію термінів. Це спрощує подальший аналіз та вилучення іменованих сутностей, оскільки дозволяє розглядати слова як єдині лексичні одиниці.

5. Підготовка до використання методів NLP:

Всі ці етапи підготовки тексту мають на меті створити чистий та структурований корпус для подальшого використання алгоритмів NLP. Це є критичним для успішного вилучення іменованих сутностей та формування метаданих.

В наступному блоці відбувається невід'ємний етап методу формування метаданих - "вилучення іменованих сутностей". Цей етап має визначальне значення, оскільки саме тут відбувається розпізнавання та виділення ключових сутностей, які є фундаментальними для формування метаданих. Цей блок спрямований на визначення та виділення у тексті конкретних об'єктів, які мають

ім'я, локальну приналежність, події, тощо. Далі наведено структуру метода вилучення іменованих сутностей:

1. Означення іменованих сутностей:

На першому етапі NER, ми визначаємо, які частини тексту можуть бути іменованими сутностями. Це включає імена осіб, місця, організації, дати, числа тощо. Означення цих сутностей важливо для подальшого уточнення та аналізу інформації.

2. Вилучення іменованих сутностей:

Після означення, ми переходимо до етапу вилучення іменованих сутностей. Тут використовуються різні методи, такі як статистичні моделі, підходи основані на правилах, чи навчання з учителем. Мета - точно та повністю виділити імена та локальні приналежності з тексту.

3. Локалізація та класифікація:

Однією з ключових функцій NER є визначення меж та класифікація вилучених сутностей. Наприклад, визначення, чи є слово іменем людини чи географічною локацією. Це дозволяє вірно ідентифікувати та розмежовувати різні типи іменованих сутностей.

4. Узагальнення та структурування:

Після вилучення іменованих сутностей, важливо провести процес узагальнення та структурування. Це може включати утворення списків, дерев або інших структур для подальшого використання в формуванні метаданих.

5. Адаптація до предметної галузі:

Оскільки мета методу полягає у формуванні метаданих про заклади позашкільної освіти, важливо адаптувати процес вилучення іменованих сутностей до специфіки цієї предметної галузі.

Взагалі, блок "Вилучення Іменованих Сутностей" визначається як етап, який не лише ізолює та виділяє ключові об'єкти у тексті, а й структурує цю інформацію для подальшого використання у формуванні метаданих, надаючи контекст та семантичну цінність отриманої інформації.

У останньому блоці ми зосереджуємося на важливому етапі - "формуванні метаданих". Цей блок визначає кінцевий результат методології, де вся зібрана та оброблена інформація консолідується у зручній та структурованій формі.

Процес формування метаданих на основі списку витягнутих іменованих сутностей є важливим етапом у розробці методу формування метаданих на основі обробки природної мови. Формується структуру та зміст отриманих даних, перетворюючи їх у зрозумілу та легкозрозумілу форму для подальшого використання. Розглянемо кроки цього процесу докладніше:

1. Очищення та підготовка даних:

Спочатку дані, які містять іменовані сутності, проходять через процес очищення. Це включає в себе перевірку на наявність некоректних чи непотрібних елементів, а також перевірку на унікальність та консистентність. Дані готуються для подальшого аналізу.

2. Ідентифікація типів іменованих сутностей:

Кожна іменована сутність позначається своїм типом (наприклад, "назва закладу", "гурток", "керівник гуртка" тощо). Алгоритми визначення типів іменованих сутностей враховують контекст та залежності між словами.

3. Створення структури метаданих:

На основі визначених типів іменованих сутностей формується структура метаданих. Кожен тип може мати свої властивості та атрибути. Наприклад, для "назви закладу" може бути визначена властивість "рівень освіти", а для "гуртка" - атрибут "графік проведення".

4. Встановлення зв'язків між сутностями:

Важливо враховувати взаємозв'язки між іменованими сутностями. Наприклад, "керівник гуртка" може бути пов'язаний з конкретним "гуртком". Встановлення цих зв'язків допомагає створити повнішу та зв'язану інформаційну структуру.

5. Збереження у зручному форматі:

Остаточна структура метаданих зберігається у зручному форматі, часто у вигляді файлу XML чи JSON. Це дозволяє легко обмінюватися та використовувати отримані дані в інших системах.

6. Валідація та Оптимізація:

На завершальному етапі проводиться валідація створеної структури для впевненості в її коректності. Також може проводитися оптимізація для покращення швидкодії та ефективності використання метаданих.

2.4 Опис інструментарію, методи, які будуть використовуватись у проекті

Для практичної реалізації наведеної методики використовується мова програмування Python [14], через її широкі можливості бібліотек для машинного навчання. Python має багато бібліотек з відкритим доступом, спрямованих на обробку природної мови, включаючи NLTK [15], SpaCy[16] та Beautiful Soup[17]. Ці бібліотеки надають готові інструменти для обробки текстової інформації, вилучення ключової інформації та аналізу мови.

Beautiful Soup - це потужна бібліотека для парсингу та обробки документів HTML та XML у мові програмування Python. Використання Beautiful Soup має кілька переваг, які зроблюють її важливим інструментом для роботи з веб-скрапінгом та обробкою HTML-документів.

Одна з ключових переваг полягає в простоті використання. Beautiful Soup надає зручний інтерфейс для навігації та взаємодії з HTML-структурою, дозволяючи легко знаходити та отримувати доступ до елементів, атрибутів та текстового вмісту. Це особливо корисно для швидкого витягування необхідних даних зі складних веб-сторінок.

Ще однією суттєвою перевагою Beautiful Soup є його гнучкість при роботі з різними версіями HTML та вміння адаптуватися до різних структур документів.

Вона може легко вирішувати завдання парсингу навіть у випадках, коли HTML-код має неточності чи нестабільну структуру.

Beautiful Soup дозволяє також ефективно взаємодіяти з різними типами об'єктів, такими як списки, рядки та об'єкти тегів, що полегшує маніпулювання та аналіз даних. Ця гнучкість робить її інструментом вибору для розпаковування та обробки інформації під час веб-скрапінгу та аналізу сторінок.

NLTK (Natural Language Toolkit) використовується для попередньої обробки текстових даних з кількох ключових причин. По-перше, NLTK надає широкий спектр функцій для роботи з текстовою інформацією, таких як токенізація, лематизація, стемінг, та частотний аналіз, що є необхідними для вилучення інформації та аналізу мови у контексті методики формування метаданих.

Додатково, NLTK має підтримку та документацію, що полегшує використання та вивчення бібліотеки. Його інструменти дозволяють проводити детальний аналіз текстової інформації, що є ключовим у розробці методики формування метаданих, оскільки потрібно ефективно вирішувати завдання з обробки природної мови для точного виділення та систематизації інформації.

Вибір NLTK також обґрунтований тим, що бібліотека спеціалізується на лінгвістичних завданнях та має інструменти для роботи з різноманітними мовними особливостями.

SpaCy використовується для витягування іменованих сутностей через його високу ефективність, швидкість та точність у розпізнаванні сутностей у тексті. SpaCy визначає іменовані сутності, такі як імена, локації, організації, числа та інші, з високою точністю, що робить його потужним інструментом для вилучення ключової інформації з текстових даних.

Однією з переваг SpaCy є його оптимізований парсер, який дозволяє швидко та ефективно обробляти текстові дані. Це особливо важливо у великих обсягах даних, які можуть виникати у вивченні закладів позашкільної освіти.

3 ЕКСПЕРИМЕНТАЛЬНЕ ДОСЛІДЖЕННЯ ЕФЕКТИВНОСТІ МЕТОДИКИ ФОРМУВАННЯ МЕТАДАНИХ ПРО ЗАКЛАДИ ПОЗАШКІЛЬНОЇ ОСВІТИ НА ОСНОВІ ТЕХНОЛОГІЇ NLP

3.1 Оцінка ефективності впровадженого методу

Впровадження даної методики спрощує процес формування метаданих про заклад позашкільної освіти на основі технологій обробки природної мови. Для порівняння – ручний процес вилучення метаданих може виконуватись однією або декількома людьми, в залежності від обсягу інформації про заклади, і термін виконання даного процесу може затягнутись. На відміну від ручного процесу, запропонована методика може оброблювати 1 сайт за короткий проміжок часу в залежності від технічних характеристик персонального комп’ютера людини, яка використовуватиме цей метод. Єдиний недолік який може статися при використання методики – неповнота витягнутої інформації про заклади, це може статися через відсутність або некоректне розпізнавання деяких текстових елементів на веб-сторінці. Однак цей недолік можна вирішити шляхом вдосконалення алгоритмів обробки природної мови та регулювання параметрів, що використовуються для вилучення іменованих сутностей.

Порівнюючи з ручним процесом, де велика частина роботи виконується вручну, використання запропонованої методики впрощує та автоматизує процес формування метаданих. Це дозволяє ефективно взаємодіяти з великим обсягом даних, збільшуючи швидкість та точність витягування інформації. Крім того, вона зменшує ризик людських помилок та виключає вплив суб'єктивних факторів на результати.

Підсумовуючи, впровадження запропонованої методики створює ефективний та швидкий спосіб формування метаданих про заклади позашкільної освіти, зменшуючи трудомісткість та часові витрати порівняно із традиційним ручним підходом.

Для того щоб оцінити повноту витягнутих іменованих сутностей буде використовуватись метрика F1. Використання метрики F1 є важливим етапом для оцінки ефективності впровадженої методики в процес обробки текстової інформації.

По-перше, після застосування методики до текстових даних, система буде виявляти іменовані сутності, такі як назви закладів, ключові слова чи описові елементи. Метрика F_1 дозволить оцінити, наскільки ефективно система впоралася із завданням виявлення іменованих сутностей порівняно із заздалегідь відомими правильними результатами (анотаціями).

Другий аспект полягає у збалансованості між точністю та повнотою. Високий показник F_1 свідчить про те, що методика досягла гарного компромісу між тим, щоб виявляти якнайбільше іменованих сутностей (висока повнота) та виявляти їх точно (висока точність).

Крім того, метрика F_1 є особливо корисною в випадках, коли у деякому контексті важливо уникати помилкових виявлених чи невиявлених іменованих сутностей. Наприклад, при обробці інформації про заклади позашкільної освіти, невірне виявлення чи пропуск іменованих сутностей може вплинути на якість збору метаданих.

Отже, використання метрики F_1 дозволить здійснити об'єктивну оцінку та порівняння результатів впровадженої методики, визначити її точність та повноту виявлення іменованих сутностей.

Метрика F_1 набуває значення між 0 та 1, де 1 вказує на ідеальну точність та повноту. Обчислюється дана метрика за формулою (3.1):

$$F_1 = 2 * \frac{precision * recall}{precision + recall} \tag{3.1}$$

де:

Precision (точність) вимірює, яку частину вилучених іменованих сутностей є дійсно іменованими сутностями.

Recall (повнота) визначає, яку частину всіх іменованих сутностей було вилучено алгоритмом.

В свою чергу Precision та Recall розраховується за формулами (3.2) та (3.3) відповідно.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positive} \tag{3.2}$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \tag{3.3}$$

де:

True Positives - Кількість правильно вилучених іменованих сутностей.

False Positives - Кількість неправильно вилучених іменованих сутностей.

False Negatives - Кількість пропущених іменованих сутностей.

Протягом проведення дослідження було взято на обробку 30 веб-сайтів, присвячених гурткам позашкільної освіти. В результаті аналізу цих сайтів були отримані наступні показники: кількість правильно вилучених іменованих сутностей склала 1051, кількість неправильно вилучених – 287, та кількість пропущених становила 124. Отримані значення використовувались для розрахунку метрик Precision і Recall, що вказують, відповідно, на точність та повноту процесу вилучення іменованих сутностей.

Були отримані значення Precision рівне 0,785 та Recall рівне 0,894. Обчислене значення метрики F_1, яке відображає гармонічний баланс між Precision та Recall, склало приблизно 0,836. Отриманий результат свідчить про досягнення задовільного рівня ефективності у витяганні іменованих сутностей з веб-сайтів, що спрощує процес формування метаданих про заклади позашкільної освіти.

3.2 Екрані форми додатку

З метою поліпшення ефективності використання розробленого методу, було вирішено реалізувати програмний інтерфейс, який взаємодіє з користувачем та надає зручний механізм введення веб-посилань для обробки. Користувач подає посилання на веб-сторінку до програмного інтерфейсу, інтерфейс, в свою чергу, автоматично оброблює зазначену веб-сторінку за використанням розробленого методу аналізу та вилучення іменованих сутностей.

У результаті завершення обробки, користувач отримує вихідний файл у форматі XML, що містить зібрані метадані про заклад позашкільної освіти, що був представлений посиланням. Використання програмного інтерфейсу дозволяє здійснити автоматизований та зручний доступ до функціоналу розробленого методу, спрощуючи введення та отримання результатів обробки для користувача.

Такий підхід допомагає зробити процес вилучення метаданих більш доступним і використовуваним, а програмний інтерфейс виступає посередником між користувачем і технічною реалізацією методу, сприяючи його широкому впровадженню та застосуванню.

На рисунку 3.1 представлена перша форма інтерфейсу додатку для формування метаданих. На зображенні відображено загальне вікно програмного інтерфейсу, яке взаємодіє з користувачем під час введення необхідних параметрів та запуску процесу обробки.

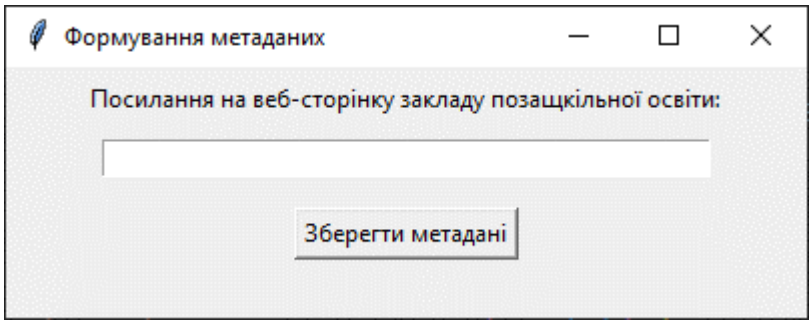


Рис. 3.1. Головна форма додатку формування метаданих

У цій формі користувач має можливість ввести відомості шляхом введення посилання на веб-сторінку відповідним полем. Враховуючи специфіку завдання, обмеження поля забезпечують лише введення посилань для обробки.

Додатково, на формі присутня кнопка, призначена для запуску процесу обробки введених даних та формування метаданих. Ця кнопка ініціює виклик необхідних функцій та алгоритмів, які відповідають за аналіз та вилучення інформації з вказаної веб-сторінки. Крім того, через відповідний інтерфейс користувачеві надається можливість зберегти отримані метадані для подальшого використання.

Застосування цієї форми допомагає забезпечити зручність та доступність введення параметрів для користувача, узагальнюючи основні функціональність додатку та надаючи засоби для взаємодії з його основними функціями.

На рисунку 3.2 представлена інформаційна форма, спрямована на повідомлення користувача про успішне збереження файлу метаданих та надання інформації щодо його назви. Додатково до основної функціональності, форма включає запис ідентифікатора, що складається з назви файлу та дати його створення.

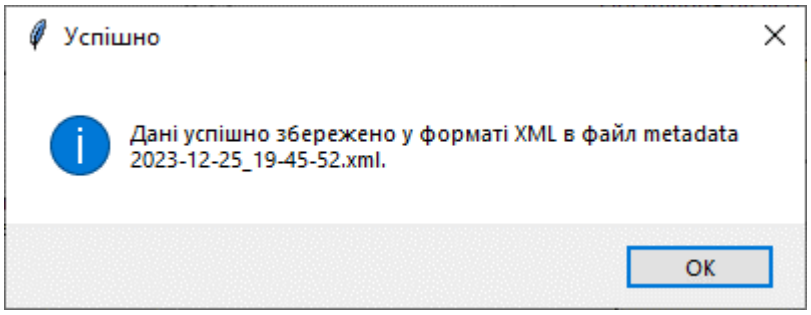


Рис. 3.2 Форма повідомлення про успішне збереження даних

Це ідентифікаційне позначення має на меті спростити процес розпізнавання та управління збереженими файлами метаданих. Завдяки включенню дати створення, користувач може легко визначити послідовність збережених файлів та забезпечити відповідність між конкретним файлом та датою його створення.

Цей підхід також забезпечує можливість зберігання та управління метаданими для різних закладів позашкільної освіти. Кожен файл, створений додатком, буде відзначений унікальною назвою, що містить ім'я файлу та дату створення. Це сприяє систематизації та організації збережених даних, роблячи процес взаємодії з отриманими результатами більш зручним для користувача.

На рисунках 3.4 та 3.5 представлені приклади інформаційних форм, які виводяться в разі виникнення помилок під час введення параметрів користувачем. Вони ілюструють дві можливі ситуації помилок: залишення поля для введення посилання порожнім та введення невірного посилання перед натисканням кнопки збереження.

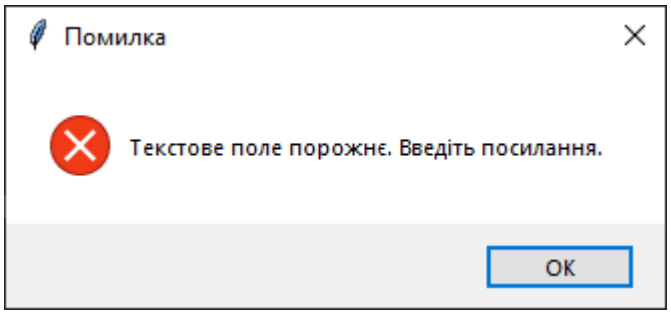


Рис. 3.4 Форма повідомлення про пусте поле посилання

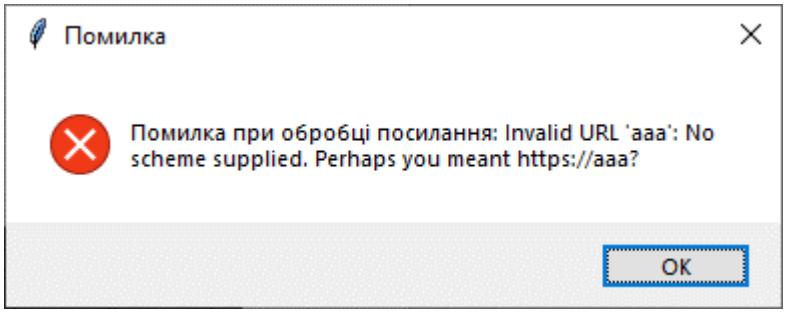


Рис. 3.5 Форма повідомлення про хібне посилання

У випадку залишення поля порожнім користувач отримає відповідне повідомлення про помилку, що нагадує про необхідність введення коректного посилання перед збереженням метаданих.

У ситуації, коли користувач введе невірне посилання, інформаційна форма повідомить його про помилку та виводить відповідне сповіщення про необхідність перевірити та виправити введені дані.

Ці інформаційні форми допомагають користувачеві виявити та виправити помилки, зроблені при введенні параметрів. Такий підхід спрямований на покращення користувацького досвіду та забезпечення надійності введення вхідних даних, що є важливим елементом успішної роботи програмного забезпечення.

На рисунку 3.6 відображено результати успішного збереження метаданих про заклад позашкільної освіти за обраною структурою. Результати представлені у вигляді сформованого файлу, який містить інформацію, витягнуту з вибраної веб-сторінки.

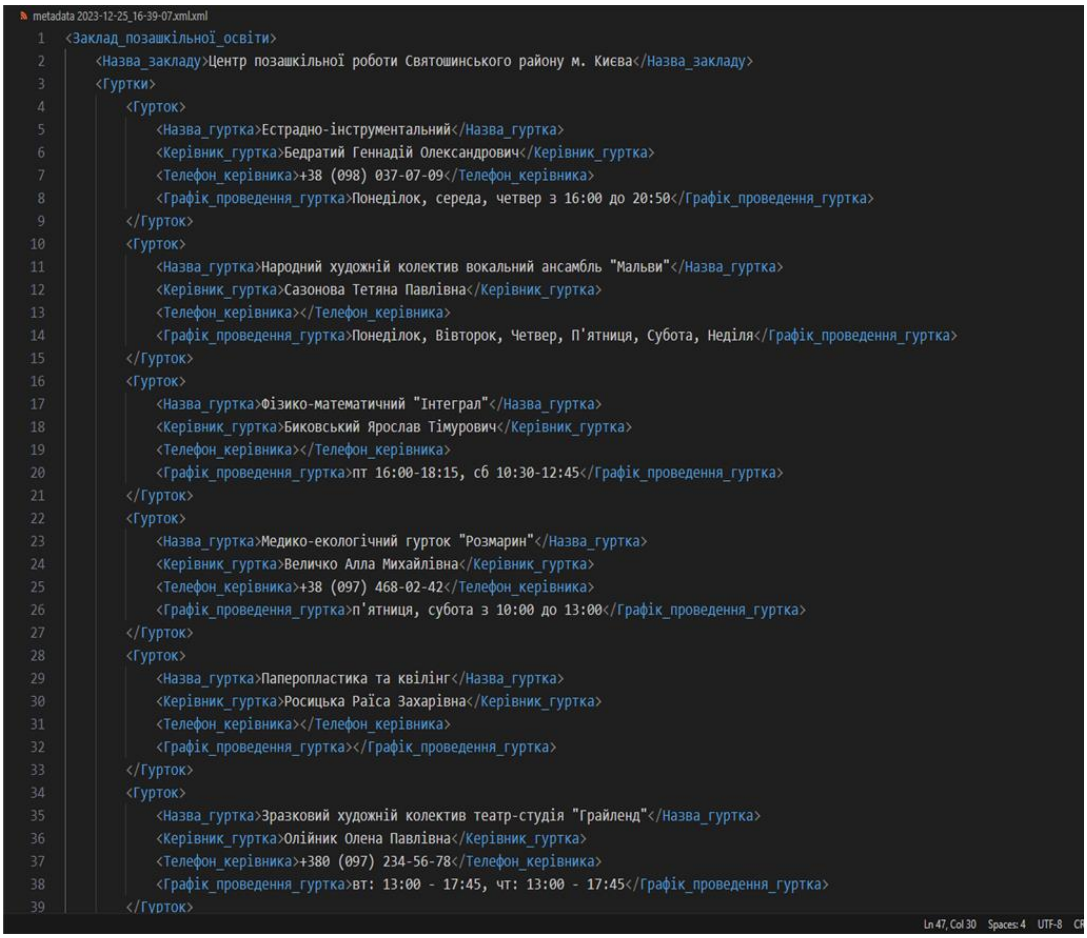


Рис. 3.6 Результат збережених метаданих про заклад

У цьому контексті, структура відображає спосіб організації та представлення вилучених метаданих. Кожен елемент інформації про заклад позашкільної освіти відображений відповідно до визначеної структури, яка може включати назву закладу, назви гуртків, ім'я керівника гуртка, телефон керівника гуртка, графік проведення гуртка та інші важливі деталі.

Цей файл, сформований за вказаною структурою, є результатом успішного використання розробленого методу та програмного інтерфейсу. Він може бути подальше використаний для подальшого аналізу, обробки чи інтеграції в інші інформаційні системи, що робить його цінним елементом в контексті автоматизованого формування метаданих про заклади позашкільної освіти.

Під час розробки програмного інтерфейсу був створений файл, що відображає роботу методу визначення іменованих сутностей (NER) для тонкого налаштування загального методу формування метаданих. У цьому файлі детально описано процес роботи алгоритму NER, який відповідає за визначення та вилучення іменованих сутностей з текстової інформації.

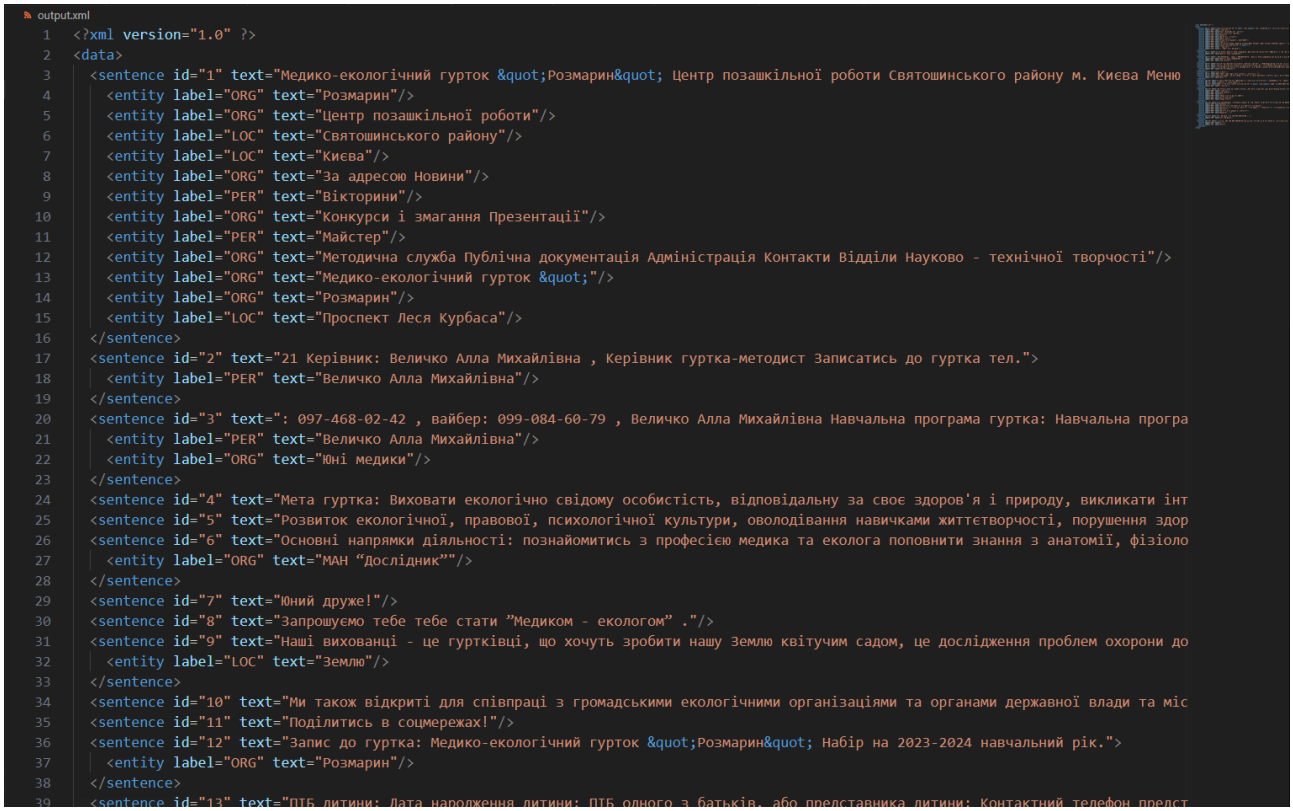


Рис. 3.7 Приклад роботи методу вилучення іменованих сутностей

Аналіз цього файлу дозволяє налаштовувати параметри та входження алгоритму NER відповідно до специфіки завдання з формування метаданих для закладів позашкільної освіти. Розглядання роботи NER є ключовим етапом в оптимізації та вдосконаленні алгоритмічних компонентів, спрямованих на точне визначення та вилучення релевантних іменованих сутностей.

Цей підхід допомагає досягти високої точності та ефективності в роботі із текстовою інформацією, що є критичним у контексті формування метаданих, де важливо точно ідентифікувати та структурувати ключові елементи інформації про заклади позашкільної освіти.

ВИСНОВКИ

1. Проведений докладний аналіз сучасного стану закладів позашкільної освіти виявив потребу у спрощенні процесу формування метаданих про ці заклади. Розглянуті різноманітні аспекти їхньої діяльності, особливості надання послуг та освітні програми.
2. Проведений огляд різних форматів метаданих дозволив визначити найбільш ефективні для використання в контексті закладів позашкільної освіти. В результаті визначено конкретні види метаданих, які будуть створені та використовуватись для опису цих закладів.
3. Розроблено формальну модель, що визначає структуру та основні етапи методики обробки веб-сторінки закладу позашкільної освіти. Побудовано блок-схему методу, яка ілюструє послідовність застосування методів та алгоритмів обробки природної мови, включаючи NER (визначення іменованих сутностей).
4. Розроблено інтерфейс, який спрощує витягування інформації про заклад позашкільної освіти. Ця екранна форма дозволяє користувачеві отримати необхідні дані, використовуючи лише посилання за допомогою запропонованої методики. Отримана інформація може використовуватись для подальшого вдосконалення веб-сторінок закладів позашкільної освіти та для забезпечення можливостей пошуку, аналізу та статистики.
5. Проведено аналіз ефективності впровадженої методики. Система в змозі вилучити до 84% метаданих із різних джерел. Основною перевагою методики є її висока швидкість та ефективність при обробці значної кількості неструктурованої інформації.

ПЕРЕЛІК ПОСИЛАНЬ

1. Закон України «Про позашкільну освіту» [Електронний ресурс] – Режим доступу: https://zakon.rada.gov.ua/laws/show/1841-14 (дата звернення: 18.12.2023). – Назва с екрану.
2. Постанова Кабінету Міністрів України від 06.05.2001 р. № 433 «Про затвердження переліку типів позашкільних навчальних закладів [Електронний ресурс] – Режим доступу: https://zakon.rada.gov.ua/laws/show/433-2001-%D0%BF#Text (дата звернення 18.12.2023). – Назва с екрану.
3. Перелік закладів позашкільної освіти системи Міністерства освіти і науки України - Набори даних - Портал відкритих даних [Електронний ресурс] // Головна сторінка - Data.gov.ua. – Режим доступу: https://data.gov.ua/dataset/18d7fb70-3781-4017-88d2-651e14d14211/resource/cafc7d36-68d6-4041-9f26-6aff8c566716 (дата звернення: 18.12.2023). – Назва з екрана.
4. Позашкільні заклади освіти - Департамент освіти і науки Києва [Електронний ресурс] – Режим доступу: https://data.gov.ua/dataset/18d7fb70-3781-4017-88d2-651e14d14211/resource/cafc7d36-68d6-4041-9f26-6aff8c566716 (дата звернення: 18.12.2023). – Назва з екрана.
5. Рейтинг закладів позашкільної освіти [Електронний ресурс] // Перший Український рейтинговий портал. – Режим доступу: https://ranking.sumdu.edu.ua/ranking/16-rejting-zakladiv-pozashk-osv.html (дата звернення: 18.12.2023). – Назва з екрана.
6. Інтерактивна карта закладів позашкільної освіти – Позашкільна освіта України [Електронний ресурс] – Режим доступу: https://pou.org.ua/2021/01/interaktyvna-karta-zakladiv-pozashkilnoyi-osvity (дата звернення: 18.12.2023). – Назва з екрана.

7. Liddy E. Natural Language Processing [Електронний ресурс] / Elizabeth Liddy // SURFACE at Syracuse University. – Режим доступу: https://surface.syr.edu/istpub/63/ (дата звернення: 18.12.2023). – Назва з екрана.
8. Онищенко К. Г. Аналіз методів обробки природної мови / К. Г. Онищенко, Я. Данієль, Р. Каменєв // Інформаційні системи та технології : матеріали 9-ї Міжнар. наук.-техн. конф., 17-20 листопада 2020 р. – Харків : Друкарня Мадрид, 2020. – С. 186–190.
9. Barney N. What is named entity recognition (NER)? | definition from techtarget [Електронний ресурс] / Nick Barney // WhatIs. – Режим доступу: https://www.techtarget.com/whatis/definition/named-entity-recognition-NER (дата звернення: 18.12.2023). – Назва з екрана.
10. Kranz G. What is metadata and how does it work? [Електронний ресурс] / Garry Kranz // WhatIs. – Режим доступу: https://www.techtarget.com/whatis/definition/metadata (дата звернення: 18.12.2023). – Назва з екрана.
11. Стрішенець Н. В. Метадані у сучасному бібліотекознавстві. Метадані – нове чи старе поняття? / Н. В. Стрішенець // Бібліотекознавство. Документознавство. Інформологія. - 2010. - № 2. - С. 4-11. - Режим доступу: http://nbuv.gov.ua/UJRN/bdi_2010_2_1
12. Qin J. Metadata / Jian Qin, Marcia Lei Zeng. – [Б. м.] : Taylor & Francis Group, 2022.
13. de Keyser P. Metadata formats and indexing [Електронний ресурс] / Pierre de Keyser // Indexing. – [Б. м.], 2012. – С. 143–166. – Режим доступу: https://doi.org/10.1016/b978-1-84334-292-2.50008-8 (дата звернення: 26.12.2023). – Назва з екрана.
14. Guides: metadata & discovery @ pitt: metadata standards [Електронний ресурс] // Home - Guides at University of Pittsburgh. – Режим доступу: https://pitt.libguides.com/metadatadiscovery/metadata-standards (дата звернення: 26.12.2023). – Назва з екрана.

15. Metadata standards: definition, examples, types & more! [Електронний ресурс] // Atlan | Third-Gen Data Catalog. – Режим доступу: https://atlan.com/metadata-standards/ (дата звернення: 26.12.2023). – Назва з екрана.
16. Steinacker A. Metadata standards for Web-based resources [Електронний ресурс] / A. Steinacker, A. Ghavam, R. Steinmetz // IEEE multimedia. – 2001. – Т. 8, № 1. – С. 70–76. – Режим доступу: https://doi.org/10.1109/93.923956 (дата звернення: 26.12.2023). – Назва з екрана.
17. Zakharova O. V. Big data metadata classification [Електронний ресурс] / O. V. Zakharova // Problems in programming. – 2019. – № 4. – Режим доступу: https://doi.org/10.15407/pp2019.04.053 (дата звернення: 18.12.2023). – Назва з екрана.
18. Мазур Ю. О. Використання метаданих у сучасному світі [Електронний ресурс] / Ю. О. Мазур // Вісник студентського наукового товариства ДонНУ імені Василя Стуса. – Режим доступу: https://jvestnik-sss.donnu.edu.ua/article/view/12104 (дата звернення: 18.12.2023). – Назва з екрана.
19. Welcome to python.org [Електронний ресурс] // Python.org. – Режим доступу: https://www.python.org (дата звернення: 18.12.2023). – Назва з екрана.
20. NLTK :: natural language toolkit [Електронний ресурс] // NLTK :: Natural Language Toolkit. – Режим доступу: https://www.nltk.org (дата звернення: 18.12.2023). – Назва з екрана.
21. SpaCy · industrial-strength natural language processing in python [Електронний ресурс] // spaCy · Industrial-strength Natural Language Processing in Python. – Режим доступу: https://spacy.io (дата звернення: 18.12.2023). – Назва з екрана.
22. Beautiful soup: we called him tortoise because he taught us. [Електронний ресурс] // Swear not by the wiki, the fickle wiki, the inconstant wiki. – Режим доступу: https://www.crummy.com/software/BeautifulSoup/ (дата звернення: 18.12.2023). – Назва з екрана.
23. Tkinter – інтерфейс python до tcl/tk [Електронний ресурс] // Python documentation. – Режим доступу:

https://docs.python.org/uk/3/library/tkinter.html#module-tkinter (дата звернення: 26.12.2023). – Назва з екрана.

24. Аброскін Ю.Ю., Золотухіна О.А. “Розробка методики формування метаданих про заклади позашкільної освіти на основі технології NLP” // Науково-практична конференція «Проблеми комп’ютерної інженерії», 10 грудня 2023 року, Державний університет інформаційно-комунікаційних технологій, Київ, Україна (с.129).
25. Honcharenko T. Cluster method of forming metadata of multidimensional information systems for solving general planning problems [Електронний ресурс] / Tetyana Honcharenko // Management of development of complex systems. – 2020. – № 42. – С. 93–101. – Режим доступу: https://doi.org/10.32347/2412-9933.2020.42.93-101 (дата звернення: 26.12.2023). – Назва з екрана.
26. Natural language processing (NLP) in management research: a literature review [Електронний ресурс] / Yue Kang [та ін.] // Journal of management analytics. – 2020. – Т. 7, № 2. – С. 139–172. – Режим доступу: https://doi.org/10.1080/23270012.2020.1756939 (дата звернення: 26.12.2023). – Назва з екрана.
27. Automated question generator using NLP [Електронний ресурс] / Asst Prof Mrs M. M. Phadatare [та ін.] // International journal for research in applied science and engineering technology. – 2023. – Т. 11, № 7. – С. 894–898. – Режим доступу: https://doi.org/10.22214/ijraset.2023.54763 (дата звернення: 26.12.2023). – Назва з екрана.
28. Potential of natural language processing for metadata extraction from environmental scientific publications [Електронний ресурс] / Guillaume Blanchy [та ін.] // Soil. – 2023. – Т. 9, № 1. – С. 155–168. – Режим доступу: https://doi.org/10.5194/soil-9-155-2023 (дата звернення: 26.12.2023). – Назва з екрана.
29. An automated framework for the extraction of semantic legal metadata from legal texts [Електронний ресурс] / Amin Sleimi [та ін.] // Empirical software engineering. – 2021. – Т. 26, № 3. – Режим доступу:

https://doi.org/10.1007/s10664-020-09933-5 (дата звернення: 26.12.2023). – Назва з екрана.

30. Deep learning-based extraction of algorithmic metadata in full-text scholarly documents [Електронний ресурс] / Iqra Safder [та ін.] // Information processing & management. – 2020. – Т. 57, № 6. – С. 102269. – Режим доступу: https://doi.org/10.1016/j.ipm.2020.102269 (дата звернення: 26.12.2023). – Назва з екрана.
31. Sarkisian H. S. Formation of gis «automobile road» metadata bases for monitoring and automobile road cadastre [Електронний ресурс] / H. S. Sarkisian, V. M. Riapukhin // Bulletin of kharkov national automobile and highway university. – 2019. – № 84. – С. 55. – Режим доступу: https://doi.org/10.30977/bul.2219-5548.2019.84.0.55 (дата звернення: 26.12.2023). – Назва з екрана.

ДЕМОНСТРАЦІЙНІ МАТЕРІАЛИ
(Презентація)

ДЕРЖАВНИЙ УНІВЕРСИТЕТ ІНФОРМАЦІЙНО - КОМУНІКАЦІЙНИХ ТЕХНОЛОГІЙ

НАВЧАЛЬНО–НАУКОВИЙ ІНСТИТУТ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ

КАФЕДРА ІНЖЕНЕРІЇ ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ

Магістерська робота

«РОЗРОБКА МЕТОДИКИ ФОРМУВАННЯ МЕТАДАНИХ ПРО ЗАКЛАДИ ПОЗАШКІЛЬНОЇ ОСВІТИ НА ОСНОВІ ТЕХНОЛОГІЇ NLP»

Виконав: студент групи ПДМ-64 Аброскін Юрій Юрійович

Керівник: к.т.н., доц., доцент кафедри ІПЗ Золотухіна Оксана Анатоліївна

Київ - 2024

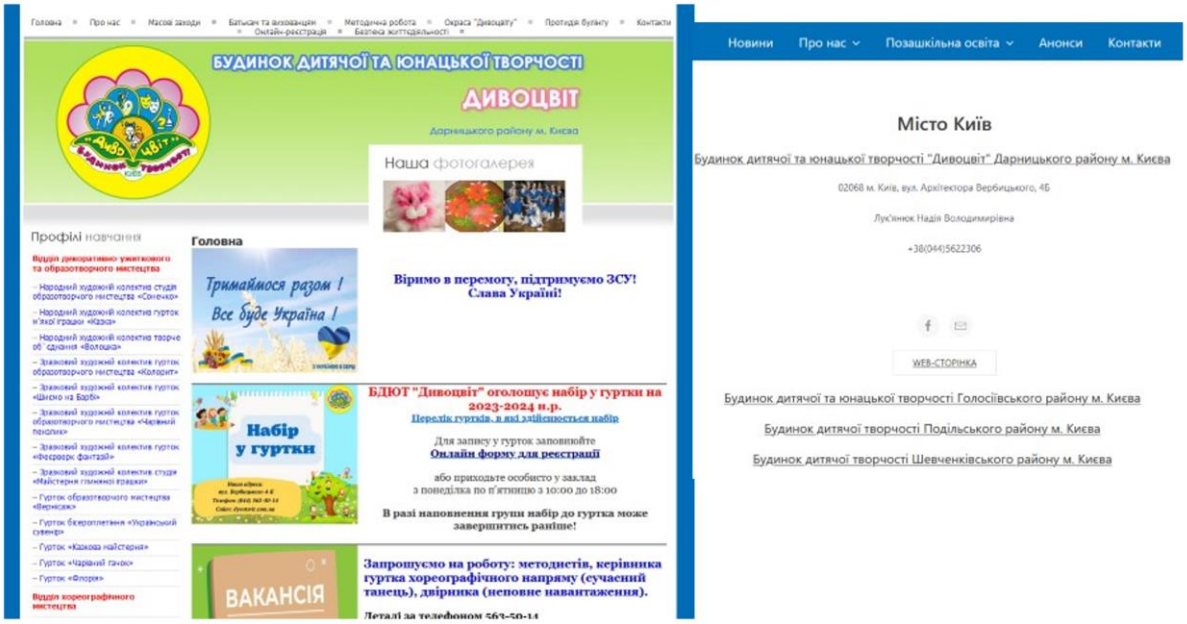
МЕТА, ОБ’ЄКТ, ПРЕДМЕТ ДОСЛІДЖЕННЯ

Мета роботи: спростити процес формування метаданих про заклади позашкільної освіти на основі методівобробки природної мови.

Об’єкт дослідження процес формування метаданих про заклади позашкільної освіти.

Предмет дослідження методи та засоби вилучення інформації з текстових даних.

ПРИКЛАД ВЕБ -СТОРІНКИ З ІНФОРМАЦІЄЮ ПРО ЗАКЛАД ПОЗАШКІЛЬНОЇ ОСВІТИ



3

ФОРМАТИ ПРЕДСТАВЛЕННЯ МЕТАДАНИХ

Формат	Особливості
CSV	Простий текстовий формат, що складається з заголовків стовпців та даних у вигляді рядків. Має обмежені можливості для опису метаданих .
JSON	Структурований текстовий формат, подібний до XML, але має більш компактний формат. Має схожі переваги та недоліки, як і XML.
RDF	Графовий формат метаданих, що використовує трикутники для опису зв'язків між метаданими. Має хорошу масштабованість та придатність для обміну метаданими між різними системами.
DCMI	Стандартизований формат метаданих для опису документів та інших ресурсів. Має широкий набір елементів метаданих, що робить його зручним для використання в різних областях.
XMP	Розширюваний формат метаданих, що може додавати метадані до інших файлів. Має широкий набір можливостей, але вимагає підтримки з боку програмного забезпечення.
XML	Структурований текстовий формат, що використовує теги для опису метаданих. Має більшу гнучкість, ніж CSV, але вимагає більш складних інструментів для роботи з ним.

4

ВИДИ МЕТАДАНИХ ПРО ЗАКЛАД ПОЗАШКІЛЬНОЇ ОСВІТИ

Назва	Опис
Name	Назва закладу позашкільної освіти
Webpage	Посилання на веб-сторінку закладу
Address	Адреса знаходження закладу в місті
Director	ПІБ директора закладу
Group	Назва гуртка
Group leader	ПІБ керівник гуртка
Schedule	Розклад гуртка
Phone	Контактний телефон
Social_media	Посилання на сторінку закладу у соц. мережах

5

СТРУКТУРА XML-ФАЙЛУ З МЕТАДАНИМИ ПРО ЗАКЛАД ПОЗАШКІЛЬНОЇ ОСВІТИ

```
<Заклад_позашкільної_освіти>
  <Назва_закладу>Назва закладу позашкільної освіти</Назва_закладу>
  <Гуртки>
    <Гурток>
      <Назва_гуртка>Назва гуртка</Назва_гуртка>
      <Керівник_гуртка>Ім'я керівника гуртка</Керівник_гуртка>
      <Телефон_керівника>Телефон керівника гуртка</Телефон_керівника>
      <Графік_проведення_гуртка>Графік проведення гуртка </Графік_проведення_гуртка>
    </Гурток>
  </Гуртки>
</Заклад_позашкільної_освіти>
```

6

ФОРМАЛЬНА МОДЕЛЬ ВИДІЛЕННЯ МЕТАДАНИХ

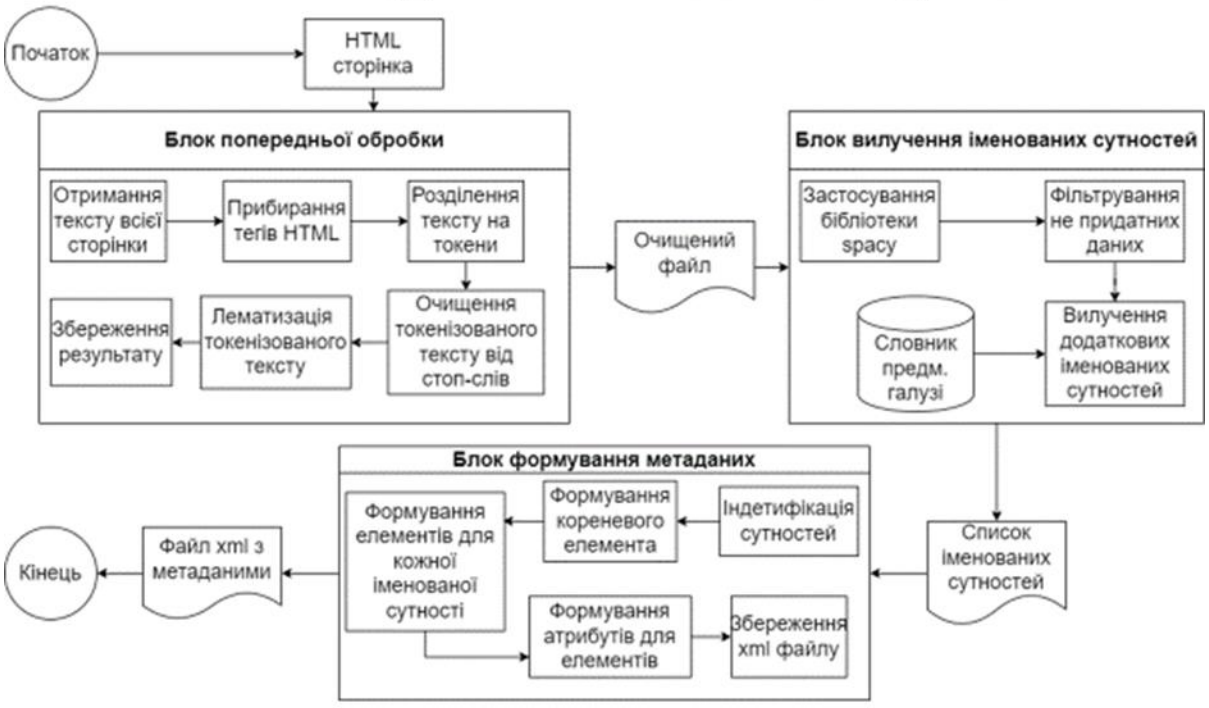
$$XML\ File = SaveToXML\big(CreateMetadata(NE)\big)$$

$$NE = NER(T', D)$$

$$T' = Lemmatize\Big(RemoveStopWords\Big(Tokenize\big(RemoveTags(T)\big)\Big)\Big)$$

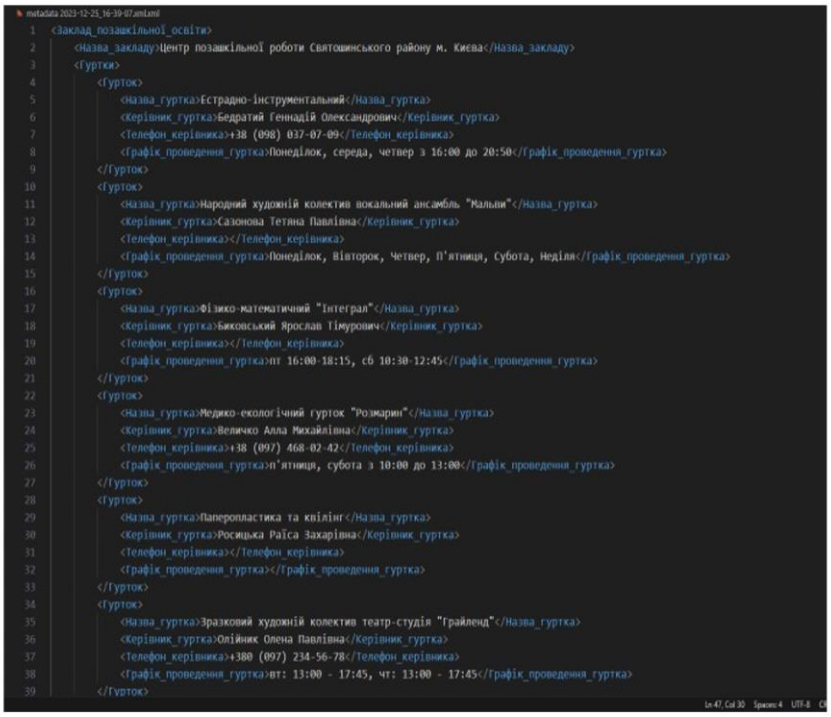
7

ЕТАПИ МЕТОДУ ФОРМУВАННЯ МЕТАДАНИХ



8

ПРИКЛАД ГЕНЕРАЦІЇ МЕТАДАНИХ НА ОСНОВІ ТЕКСТОВОГО КОНТЕНТУ САЙТУ ЗАКЛАДУ ПОЗАШКІЛЬНОЇ ОСВІТИ



9

ВИЗНАЧЕННЯ ПОВНОТИ ТА ТОЧНОСТІ ВИТЯГНУТИХ МЕТАДАНИХ

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positive}$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

10

АНАЛІЗ РЕЗУЛЬТАТІВ ВПРОВАДЖЕННЯ МЕТОДИКИ

Критерій	**До впровадження (людина)**	**Після впровадження (алгоритм)**
Затрачений час	> 20 хв.	< 5 хв.
Кількість оброблених даних за раз	1 сайт	Кількість не обмежується
Можливість обробляти декілька сайтівводночас	Ні	Так
Структурованість даних	Може бути присутня структурованість	По замовчуванню структуроване
Середній відсоток вилучених метаданих	87%	84%

11

ВИСНОВКИ

1. Проведений докладний аналіз сучасного стану закладів позашкільної освіти виявив потребу у спрощенні процесу формування метаданих про ці заклади. Розглянуті різноманітні аспекти їхньої діяльності, особливості надання послуг та освітні програми.
2. Проведений огляд різних форматів метаданих дозволив визначити найбільш ефективні для використання в контексті закладів позашкільної освіти. В результаті визначено конкретні види метаданих, які будуть створені та використовуватись для опису цих закладів.
3. Розроблено формальну модель, що визначає структуру та основні етапи методики обробки веб-сторінки закладу позашкільної освіти. Побудовано блок-схему методу, яка ілюструє послідовність застосування методів та алгоритмів обробки природної мови, включаючи NER (визначення іменованих сутностей).
4. Розроблено інтерфейс, який спрощує витягування інформації про заклад позашкільної освіти. Ця екранна форма дозволяє користувачеві отримати необхідні дані, використовуючи лише посилання за допомогою запропонованої методики. Отримана інформація може використовуватись для подальшого вдосконалення веб-сторінок закладів позашкільної освіти та для забезпечення можливостей пошуку, аналізу та статистики.
5. Проведено аналіз ефективності впровадженої методики. Система в змозі вилучити до 84% метаданих із різних джерел. Основною перевагою методики є її висока швидкість та ефективність при обробці значної кількості неструктурованої інформації.

12

ПУБЛІКАЦІЇ ТА АПРОБАЦІЯ РОБОТИ

Стаття:

1. Золотухіна О.А., Аброскін Ю.Ю. Застосування методів обробки природної мови в задачах підтримки бізнес-процесів в сфері освіти. Наукові записки Державного університету телекомунікацій. № 1. 2024. Прийнято до друку.

Тези доповідей:

1. Аброскін Ю.Ю., Золотухіна О.А. "Розробка методики формування метаданих про заклади позашкільної освіти на основі технології NLP" // Науково-практична конференція «Проблеми комп'ютерної інженерії», 10 грудня 2023 року, Державний університет інформаційно-комунікаційних технологій, Київ, Україна. С. 129 https://duikt.edu.ua/uploads/p_2626_33497568.pdf

13

ДЯКУЮ ЗА УВАГУ!

14