УДК 004.91

АНАЛІЗ СУЧАСНИХ СПЕЦІАЛІЗОВАНИХ ПРОГРАМНИХ РОЗШИРЕНЬ ДЛЯ АВТОМАТИЗАЦІЇ РОБОТИ З ЦИФРОВИМИ ДОКУМЕНТАМИ НАВЧАЛЬНИХ МАТЕРІАЛІВ

Слободзян В.О., Мазурець О.В.
Україна, Хмельницький національний університет
E-mail: vitaliy.slobodzian@gmail.com

Вступ. Онтологія навчального матеріалу є методом виявлення сенсу навчального матеріалу та засобом вирішення ряду практичних задач. Модель онтології навчального матеріалу може складатися з ключових слів, ключових термінів, структури навчального матеріалу, атрибутів ключових слів та ключових термінів, що забезпечують їх прив'язку до елементів структури навчального матеріалу [1].

Основними з етапів побудови онтології навчального матеріалу є пошук ключових термінів у контенті навчального матеріалу та побудова його логічної структури. Вхідними даними є електронний документ навчального матеріалу. Для автоматизації виконання обох наведених етапів потрібна програмна обробка відповідних цифрових файлів (зазвичай формату .DOCX). Для ефективної реалізації програмної обробки цифрових документів є доцільним використання спеціалізованих програмних комплексів, що надають об'єктно-орієнтований інструментарій для програмної роботи з контентом відповідних файлів.

Метою роботи є аналіз сучасних існуючих спеціалізованих програмних розширень для автоматизації обробки цифрових документів і визначення найбільш ефективного програмного розширення.

Основна частина. На сучасному етапі для зберігання електронних документів навчальних матеріалів використовується файл з розширенням .DOCX (або створених на його основі .HTML, .PDF тощо). На відміну від файлів .DOC, які зберігають дані документа в одному бінарному файлі, файли .DOCX створюються за допомогою відкритого формату XML, в якому зберігаються документи як колекції окремих файлів і папок в стиснутому пакеті. .DOCX-файли містять .XML-файли і три папки, docProps, Word, і _rels (Рис. 1), які містять властивості документа, його зміст і відношення між файлами, тему та включені файли. .DOCX-файли розроблені, щоб зробити вміст документів доступним і відкритим – так, текстовий документ чи

зображення зберігаються як окремі прості файли в такій колекції у складі одного файлу .DOCX.

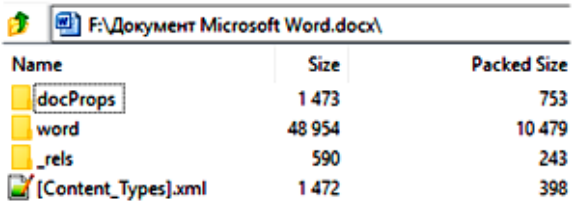

Рис. 1 – Структура .DOCX-файлу

Office OpenXML є форматом електронних офісних документів, таблиць, презентацій та діаграм, розроблений компанією Microsoft. Даний формат є міжнародним стандартом, який затверджений в Міжнародній організації зі стандартизації (ISO). Він включає в себе такі первинні мови розмітки:

- WordprocessingML – для обробки текстових документів;
- SpeadsheetXL – електронних таблиць;
- PresentationML – для презентацій;
- DrawingML – для векторної графіки, діаграм.

Хоча Word реалізує свою функціональність через об'єктну модель, збереження даних відбувається у наборі XML-файлів, що ускладнює можливості для автоматизації прямого програмного парсингу. Оскільки файл стилів та текст знаходяться в різних файлах у складі .XML, потрібно зчитувати одразу ряд файлів та на базі їх взаємозв'язків програмно формувати об'єктну модель документа, що є не найкращим рішенням. Тому для реалізації програмної обробки цифрових документів є доцільним використання спеціалізованих програмних комплексів, що надають об'єктно-орієнтований інструментарій для програмної роботи з контентом відповідних файлів. Найбільш відомими з таких спеціалізованих програмних комплексів є розширення Microsoft.Office.Interap.Word.dll [2], DocumentFormat.OpenXml.dll [3] та Spire.Doc.dll [4].

Бібліотека ***Microsoft.Office.Interap.Word.dll*** [2] дозволяє керованому коду взаємодіяти з MS Office та об'єктною моделлю на базі COM.

За допомогою процесів автоматизації Office Automation програми, написані мовами, такими як Visual C# .NET, отримують можливість програмно керувати іншими програмами. Автоматизація Microsoft Word дозволяє виконувати відповідні дії, такі як створення нового документу, додавання даних у документ або створення

таблиць. З програмами Word та іншими програмами Microsoft Office практично всі дії, які користувач може виконувати вручну за допомогою користувацького інтерфейсу, також можуть бути виконані програмним шляхом за допомогою автоматизації Office Automation.

Word реалізує цю програмну функціональність через об'єктну модель. Об'єктна модель є набором класів і методів, які служать аналогами логічних компонентів Word. Наприклад, існує об'єкт Application, об'єкт Document та об'єкт Paragraph, кожен з яких містить функціональність цих частин Word.

Щоб використовувати функції програми MS Office в проекті, можна використовувати первинний взаємозв'язок PIA (Primary Interop Assembly). PIA дозволяє керованому коду взаємодіяти з MS Office та об'єктною моделлю на базі COM. При створенні нового проекту Office, Visual Studio додає посилання на PIA, необхідний для побудови проекту. При цьому, у деяких сценаріях доводиться додавати посилання на додаткові PIA (наприклад, якщо необхідно використовувати функцію MS Office Word в проекті для MS Office Excel).

Бібліотека класів .NET ***DocumentFormat.OpenXml.dll*** [3] дає можливість розробникам програмного забезпечення працювати з пакетом Microsoft Office. Компанія надає дану бібліотеку безкоштовно в повному доступі, й її можна завантажити з офіційного сайту Microsoft або засобами Visual Studio.

Для роботи з файлами формату .DOCX, який має зображену на рисунку 1 структуру, використовується мова розмітки WordprocessingML. На рисунку 2 зображено приклад структури цієї мови розмітки.

```
<?xml version="1.0" encoding="utf-8"?>
<w:document
xmlns:w="http://schemas.openxmlformats.org/wordprocessin
gml/2006/main">
  <w:body>
    <w:p>
      <w:r>
        <w:t> Текст </w:t>
      </w:r>
    </w:p>
  </w:body>
</w:document
```

Рис. 2 – Приклад WordprocessingML-розмітки .DOCX-документу

Теги, наведені на рисунку 2, мають такі властивості:

- *document* – є основою частиною документа
- *body* – контейнер для збору структур рівня блоку, які складають основну історію;
- *p* – абзац (paragraph);
- *r* – контейнер який містить в собі текст з однаковими властивостями (run);
- *t* – власне текстовий елемент (text).

Всі інші теги, які можуть використовуватися, є допоміжними до вищенаведених і додають їм стильових властивостей.

Spire.Doc.dll [4] для .NET є повністю незалежною бібліотекою класів .NET Word, спеціально створеною для розробників, й дозволяє швидко генерувати, відкривати, редагувати та зберігати документи Word різних версій. Конвертація функціональних можливостей дозволяє розробникам здійснювати перетворення Word в інші актуальні формати документів (.PDF, .EPub, .HTML, .RTF, .Image, .XML тощо). Можна створювати та обробляти вже існуючі документи Word динамічно.

Бібліотека Spire.Doc.dll забезпечує роботу з майже всіма елементами документа Word, а саме: сторінки, розділи, заголовки, колонтитули, абзаци, переліки, таблиці, текст, виноски, поля, гіперпосилання, закладки, коментарі, зображення, стилі, фонові параметри, функції друку, налаштування документа та захисту. Крім того, підтримуються графічні об’єкти, включаючи форми, текстові поля, зображення, OLE-об’єкти та елементи керування. Бібліотека Spire.Doc.dll підтримує функцію пошуку та заміни, вирівнювання, перерву сторінки, поле заповнення, об’єднання документів, копіювання документів, друк тощо.

Аналіз. Хоч бібліотека Microsoft.Office.Interap.Word.dll надає доступ до всіх функцій MS Office, так як працює безпосередньо з PIA, для її роботи потрібна ліцензія для MS Office на кожному клієнтському комп’ютері. Крім того, при використанні Automation, MS Office завантажується у фоновому режимі, внаслідок чого займає деякий обсяг оперативної пам’яті та завантажує велику кількість файлів і DLL. Додатки MS Office були розроблені як додатки для користувацького інтерфейсу, і через це Microsoft.Office.Interop.Word.dll працює повільно. Microsoft не рекомендує використовувати Office Automation (або будь-який Office Interop) на сервері.

На відміну від Microsoft.Office.Interap.Word.dll, DocumentFormat.OpenXml.dll та Spire.Doc.dll не вимагають наявності MS Office на машині користувача та його запуску в фоновому режимі.

При роботі з docx-файлом за допомогою DocumentFormat.OpenXml.dll потрібно дотримуватися тієї ж ієрархії, що і сама структура розмітки тобто: *document-paragraph-run-text*. Також відповідно до розглянутих вище особливостей взаємозв'язків всередині документу, при присвоєнні параграфу деякого стилю, у властивостях параграфа надається лише ідентифікатор (ID) на даний стиль, натомість сам стиль описується окремо у файлі style.xml. Внаслідок необхідності регулярного співставлення ID стиля з контейнером style.xml для одержання характеристик стилів абзаців, зростає складність програмного використання бібліотеки.

Spire.Doc.dll не вимагає встановлення в систему кожного користувача MS Office, тобто є можливість повністю незалежної від нього роботи. Повна версія Spire.Doc.dll є платною, а безкоштовна версія FreeSpire.Doc має ряд обмежень (наприклад обробка не більше 500 абзаців і 25 таблиць). Перевагою Spire.Doc.dll визначено відсутність необхідності співставлення ID стиля з контейнером style.xml, оскільки дана функція реалізована на рівні бібліотеки.

Таким чином, в результаті аналізу сучасних існуючих спеціалізованих програмних розширень для автоматизації обробки цифрових документів, було визначено Spire.Doc.dll оптимальним варіантом для використання в автоматизації обробки цифрових документів. Основними перевагами Spire.Doc.dll встановлено відсутність необхідності наявності MS Office на машині користувача та запуску в фоновому режимі, а також реалізацію функцій автоматичного співставлення стилів текстових блоків їх властивостям на рівні розширення.

Перенесення функцій автоматичного співставлення стилів текстових блоків їх властивостям з рівня функціоналу програмного коду застосунка на рівень функціоналу бібліотеки дозволяє спростити як роботу застосунка з цифровим документом, так і процес програмування.

Практичне застосування. При зчитуванні файлу Spire.Doc.dll перетворює документ в об'єктну модель. Ця модель має структуру, яка починається документом і закінчується об'єктом *TextRange*.

Відповідно до об'єктної моделі документу, MS Office використовує розділи (*Section*), щоб вказати частини документа, що мають різну орієнтацію сторінки, стовпці або заголовки та нижні колонтитули. Розбиття на Section дозволяє користувачеві вказати

місце, де розпочинається і закінчується інше форматування. Тому Section використовуються коли потрібне використання заголовків, колонтитулів, схем, нумерації сторінок, розмірів аркушу, поля чи різні рівні захисту. Об'єкти Section містяться в об'єкті Document, в колекції Selections. А розділи, в свою чергу, містять в собі наступний елемент структури – абзаци (*Paragraph*).

Paragraph в MS Word визначає будь-який текст, який закінчується жорстким поверненням, яке формується, наприклад, при натисканні клавіші Enter. Формат абзацу дозволяє контролювати зовнішній вигляд фрагменту, якщо є окремі абзаци. Наприклад, можна змінити відстань між рядками або вирівнювання тексту по лівому краю на вирівнювання по центру. Можна додати індивідуальні стилі, відступи для абзаців, або визначити заголовки і списки. В об'єктній моделі *Paragraph* знаходиться в *Document -> Sections -> Section -> Paragraphs* (Рис. 3). Одержати відомості про стилі *Paragraph* можна за допомогою методу *GetStyle()*.

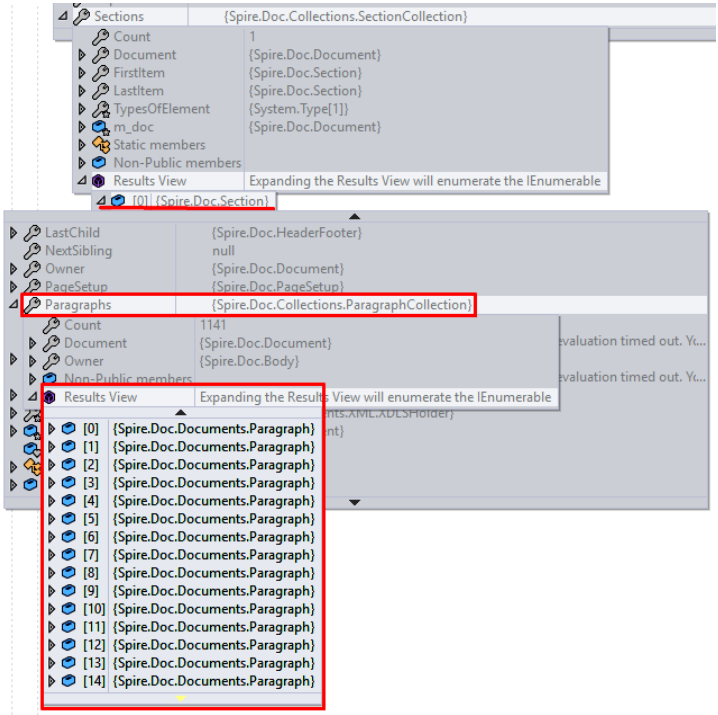

Рис. 3 – Позиція визначеного об'єкта Paragraph в об'єктній моделі документа

TextRange є найнижчим рівнем структури документу, що визначає фрагмент тексту однакового стилю. Тому для того, щоб при розв'язку прикладних задач отримати окремо кожне слово з власними властивостями стилів, потрібна розробка додаткового алгоритму для розбиття цих фрагментів *TextRange* на слова. *TextRange* знаходиться в *Paragraph -> Items*. Через роботу з відповідними властивостями, можна одержати текст та стилі (рис. 4), що відносяться до визначеного *TextRange*.

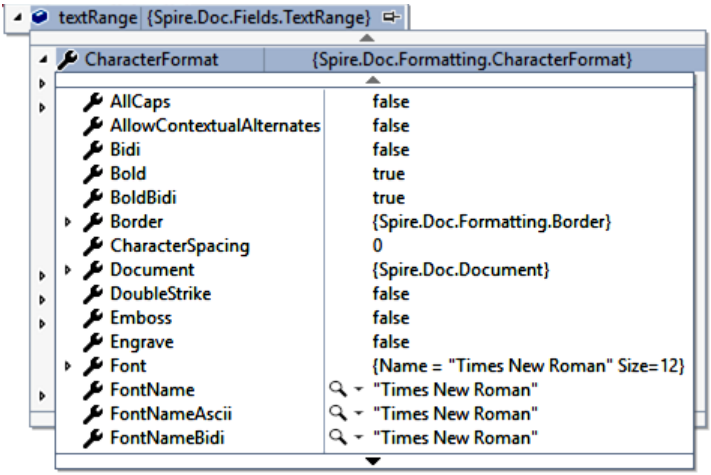

Рис. 4 – Деякі стилі визначеного TextRange

Маючи стилі форматування Paragraph та TextRange, можна визначити заголовки, рівні заголовків, текст для аналізу чи ігнорування, прив'язку певного фрагменту тексту до заголовку.

Paragraph може містити лише один з кількох визначених форматів: Level1, Level2, Level3, Level4, Level5, Level6, Level7, Level8, Level9, Body. Відповідно до цього значення, можна визначити заголовки та їх рівні. Кожен похідний елемент має посилання на батьківський, і таким чином існує можливість визначити прив'язку певного елементу до певного модулю, заголовку чи підзаголовку при визначенні структури змістовних блоків у електронному документі навчального матеріалу.

З метою визначення можливості застосування розширення Spire.Doc.dll в задачах роботи з електронними документами, було проведено його практичну апробацію [5] в рамках розробки програмної системи для вирішення прикладної задачі побудови структури електронного документу та пошуку ключових слів у його

контенті. При цьому, визначення властивостей Paragraph надало можливість для формування структури документу, а обробка властивостей елементів TextRange визначила необхідний для програмної обробки контент, що дозволило визначити множини ключових слів за допомогою методу дисперсійного оцінювання.

Висновки. За результатами використання розширення Spire.Doc.dll в розробленому програмному комплексі встановлено, що дана бібліотека надає достатній інструментарій для роботи з цифровими документами навчальних матеріалів у рамках розглянутих актуальних задач, й може бути ефективно використана в реалізації інформаційних технологій для автоматизації роботи з навчальними матеріалами.

Подальші дослідження спрямовані на впровадження розширення Spire.Doc.dll в розробку автоматизованих систем роботи з цифровими документами навчальних матеріалів для вирішення прикладних задач.

Література

1. Мазурець О. В. Онтологічний підхід до побудови семантичної моделі навчальних матеріалів / О. В. Мазурець // Науковий журнал «Вісник Хмельницького національного університету» серія: Технічні науки. Хмельницький, 2017, №6. – С.223-229.

2. Considerations for server-side Automation of Office [Електронний ресурс]. – Режим доступу: https://support.microsoft.com/en-us/help/257757/considerations-for-server-side-automation-of-office/

3. International Organization for Standardization - Open XML file formatsb [Електронний ресурс]. – Режим доступу: https://www.iso.org/search/x/query/Open%2520XML

4. Spire.Doc for .NET [Електронний ресурс]. – Режим доступу: https://www.nuget.org/packages/Spire.Doc/

5. Мазурець О. В., Ковальчук О. В., Слободзян В. О. Використання спеціалізованих програмних розширень для автоматизації роботи з цифровими документами навчальних матеріалів / О. В. Мазурець, О. В. Ковальчук, В. О. Слободзян // Науковий журнал «Вісник Хмельницького національного університету» серія: Технічні науки. Хмельницький, 2018, №1. – С.61-69.