АЛГОРИТМ ФОРМУВАННЯ РЕФЕРАТУ ТЕКСТОВОЇ РОБОТИ ЗАСОБАМИ APACHE POI

Якушко А.П., Двухглавов Д.Е.
Національний технічний університет «Харківський політехнічний інститут», м. Харків

Автоматизація аналізу текстових документів у сфері навчального процесу є важливим питанням. Бібліотека Apache POI надає потужні інструменти для програмного аналізу документів Microsoft Office, зокрема формату DOCX.

Одним із практичних застосувань цієї бібліотеки є формування реферату, що містить кількісні характеристики пояснювальної записки до курсової або дипломної роботи. Реферат включає такі параметри: кількість сторінок, рисунків, таблиць, джерел та додатків. Для кожного з цих параметрів потрібен окремий алгоритм обчислення. Визначення кількості рисунків використовує функціональність Apache POI для отримання як вбудованих зображень (getAllPictures()), так і діаграм (getCharts()). Загальна кількість рисунків визначається як їх сума. При підрахунку таблиць виникає специфічна проблема: метод getTables() класу XWPFDocument повертає всі об'єкти-таблиці, але не враховує, що одна логічна таблиця може бути представлена декількома фізичними об'єктами через розриви сторінок. Тому спочатку пропонується отримати таблиці методом getTables(), потім знайти перелік абзців, відформатованих стилем "Tablenumber", що містят текст "Кінець таблиці" або "Продовження таблиці". Їх кількість треба відняти від загальної кількості таблиць. Кількість джерел інформації визначається як кількість абзаців між абзацем з текстом "Список джерел інформації" типу Heading 1. Також цей заголовок є «орієнтиром» для визначення кількості додатків – це будуть всі наступні абзаци типу Heading 1 до кінця тексту.

Оскільки формат DOCX не містить прямої інформації про кількість сторінок, цей метод використовує складний алгоритм. Спочатку документ конвертується у формат PDF за допомогою компонентів бібліотеки JODConverter - класів LocalOfficeManager та LocalConverter для запуску LibreOffice та подальшої конвертації. Після успішної конвертації застосовується метод getNumberOfPages() класу PDDocument з бібліотеки PDFBox для підрахунку сторінок у PDF-документі. Сформований реферат представляється у вигляді рядка, що містить фразу "Реферат: K стор., L рис., M табл., N джерел, P додатків", де K, L, M, N та P - відповідні кількісні значення елементів. Автоматичне формування реферату значно прискорює процес оформлення текстових робіт і зменшує ймовірність помилок при підрахунку елементів документа.

Варто зауважити, що застосування такого алгоритму можливе лише за умови отримання документу, в якому має бути використаний визначений перелік стилів, структура документа та оформлення таблиць, ілюстрації та заголовків має відповідати вимогам університету. Тому студенти мають отримувати шаблон для підготовки документів, що не є проблемою. А тому проблем із впровадженням цього алгоритму в практику підготовки документів немає.