ОСОБЛИВОСТІ АВТОМАТИЗІЦІЇ АНАЛІЗУ ФОРМАТУВАННЯ ДОКУМЕНТІВ

Іорданов І.О., Будянський О.О.
Науковий керівник – д.т.н., проф. Ревенчук І.А.
Харківський національний університет радіоелектроніки
(61166, Харків, просп. Науки,14, каф. ПІ, тел. (057) 702-00-00)

The given work is devoted to the investigation of the document formatting validation automation strategies. It describes the main obstacles of formatting verification as well as the analysis of different document formats and tools to work with them. The approach of handling the most complicated issues related to the processing of the Microsoft Word documents is presented. In addition, the work contains the list of advantages and disadvantages of the .doc and .docx formats in the scope of formatting validation. The result of the investigations shows the main problem of current state in the automation.

У наш час існує багато установ, що працюють з постійним потоком документів, що мають відповідати певним строго визначеним вимогам форматування та оформлення. Більшість з таких організацій стикаються з проблемою перевірки оформлення документів, що є рутинною роботою та займає значну частину робочого часу спеціалістів.

Прикладом такої роботи є перевірка наукових робіт у вищих навчальних закладах. Існують певні документовані стандарти, яким мають відповідати наукові роботи, а саме - ДСТУ 3008-2015 «Звіти у сфері науки і техніки» [1].

Провівши аналіз документообігу у вищих наукових закладах було зроблено висновок, що майже усі документи, що потребують перевірки форматування розповсюджуються у форматі Microsoft Word Document (.doc, .docx). Особливістю цього формату є опис документу за допомогою структури xml, що також задокументована у стандартах ECMA (ECMA-376) [2]. Саме ці стандарти дозволяють автоматизувати процес валідації форматування та оформлення документів у форматах .doc та .docx. Саме це і стало нашою основною задачею. Даний формат надає великий набір інструментів для автоматизації аналізу форматування, але, у свою чергу, також має ряд недоліків, що ускладнюють перевірку деяких елементів.

Заданий формат дозволяє автоматизувати перевірку наступних основних груп стандартів форматування документу:

- стиль тексту (шрифт, розмір тексту, формат тощо);
- поля, інтервали та відступи (абзацний відступ, поля документу, інтервал між рядками тощо);
- зміст контенту (семантичний аналіз тексту);
- формули, таблиці, діаграми тощо.

Незважаючи на те, що даний формат надає досить велику кількість можливостей для валідації, він має істотні недоліки:

- відсутність будь-якої інформації про розміщення тексту по сторінках;
- неможливість проаналізувати кількість змісту на певній сторінці;
- неможливість проаналізувати правильність розміщення елементів, що займають декілька сторінок.

Найскладнішою частиною роботи з автоматизації була боротьба з вказаними недоліками. Було проаналізовано декілька варіантів вирішення цих труднощів і в результаті було розроблено механізм конвертації документа у декілька форматів та їх паралельна обробка. Наприклад, для вирішення проблеми з відсутністю інформації про сторінки, документ конвертується у формат PDF, що включає у себе відсутню інформацію та дозволяє визначити розміщення тексту на сторінках документу, поєднавши дані з обох форматів. За допомогою функції пошуку тексту по конкретній позиції у PDF документі, можна зрозуміти чи коректний номер сторінки вказаний, та правильність його розташування.

Важливу роль у автоматизації також відіграють засоби роботи з різними форматами документів. Після порівняння найбільш популярних та доступних засобів було прийнято рішення використати наступні:

- Spire.Doc – бібліотека для роботи зі XML структурою .docx документів
- Spire.Pdf – бібліотека для роботи з PDF документами
- PdfShark – засіб, що дозволяє маніпулювати репрезентацією PDF;
- OpenXML – засіб для низкорівневої роботи з окремими елементами структури документів Microsoft Word.

За результатами проведених досліджень можна побачити проблему, через яку, автоматизація валідації форматування документів наразі слаборозвинена. Причиною цього є факт того, що існуючі формати для написання документів не в повній мірі надають засоби для верифікації форматування. Найкращим рішенням цієї проблеми було б комбінування існуючих форматів для отримання найкращих результатів.

Список використаних джерел

1. ДСТУ 3008-2015 [Електронний ресурс] – режим доступу: http://www.knmu.kharkov.ua/attachments/3659_3008-2015.PDF.

2. ECMA (ECMA-376) [Електронний ресурс] – режим доступу: http://www.ecma-international.org/publications/standards/Ecma-376.htm.

3. Spire.Doc Documentation [Електронний ресурс] – режим доступу: https://www.e-iceblue.com/Tutorials/Spire.Doc.html.