Технічні науки

УДК 004.63

Бондаренко Нікіта Володимирович

студент

Національного технічного університету України

«Київський національний інститут імені Ігоря Сікорського»

Бондаренко Никита Владимирович

студент

Национального технического университета Украины

«Киевский политехнический институт имени Игоря Сикорского»

Bondarenko Nikita

Student of the

National Technical University of Ukraine

«Igor Sikorsky Kyiv Polytechnic Institute»

Іванішин Іван Володимирович

студент

Національного технічного університету України

«Київський національний інститут імені Ігоря Сікорського»

Иванишин Иван Владимирович

студент

Национального технического университета Украины

«Киевский политехнический институт имени Игоря Сикорского»

Ivanishyn Ivan

Student of the

National Technical University of Ukraine

«Igor Sikorsky Kyiv Polytechnic Institute»

МЕТОДИ КОНВЕРТАЦІЇ ДОКУМЕНТІВ ФОРМАТІВ PDF ТА DOC
МЕТОДЫ КОНВЕРТАЦИИ ДОКУМЕНТОВ ФОРМАТОВ PDF И DOC
CONVERSION METHODS OF PDF AND DOC DOCUMENTS

Анотація. *У роботі розглянуто та проаналізовано найпопулярніші формати документів DOC, DOCX та PDF. А також способи їх конвертації для подальшої роботи з ними.*

Ключові слова: *DOC, DOCX, PDF, конвертація документів.*

Аннотация. *В работе рассмотрены и проанализированы самые популярные форматы документов DOC, DOCX и PDF. А также способы их конвертации для дальнейшей работы с ними.*

Ключевые слова: *DOC, DOCX, PDF, конвертация документов.*

Summary. *This paper discusses and analyzes the most popular document formats DOC, DOCX and PDF. As well as methods of their conversion for further work on them.*

Key words: *DOC, DOCX, PDF, file conversion.*

Вступ. В сучасному світі документи грають важливу роль як спосіб обміну інформацією. За значенням, документ – це об'єкт, що містить певні дані або інформацію, що можна зберігати або поширювати. Фактично, документ є записом для будь-яких видів транзакцій або комунікацій між двома або більше організаціями або особами. Для будь-якої сучасної компанії у світі, створення і зберігання електронної документації є основою її функціонування.

Кожна комп'ютерне обладнання розробляється за певними стандартами, що мають свої вимоги до даних, якими система оперує. Так само кожен стандарт документу сприяє формуванню особливої структури, що має свої

переваги та недоліки щодо опрацювання, збереження та перегляду інформації. Звісно ж, єдиного золотого стандарту не існує: різні дії над даними потребують різних форматів документів. Гнучкість у роботі з інформацією, у такому разі, досягається завдяки конвертації документу у інші формати.

Стандарт DOC. Формат файлу DOC - це сукупність записів та структур, які визначають текст, таблиці, поля, зображення, вбудовану розмітку XML та інший вміст документа. Вміст можна друкувати на сторінках різного розміру або відображати на різних пристроях [1].

Формат DOC почав використовуватись компанією Microsoft Word ще 35 років тому у першій версії Word для операційної системи MS-DOS. Очевидно, що це було власне розширення для управління документами Microsoft. Єдиною програмою, яка офіційно підтримувала файли DOC, була саме Word.

У 1990-х - на початку 2000-х років з файлами DOC могли працювати вже декілька файлів-конкурентів, хоча не всі формати та налаштування Word були повністю сумісні з іншими текстовими процесорами. Починаючи з 2008 року чимало постачальників інтегрували формат DOC у безкоштовні та платні програми для обробки текстових документів.

Посилення конкуренції з боку вільного відкритого коду та його конкурентного формату відкритого документа (Open Document Format), сприяло створенню компанією Microsoft ще більш відкритого стандарту наприкінці 2000-х рр. Це призвело до розвитку нового формату файлів - DOCX. Оскільки формат базувався не на менш ефективному старому двійковому форматі, як це було раніше, а на розширюваній мові розмітки, новий стандарт отримав назву "Office Open XML". Ця мова принесла з собою низьку переваг, зокрема зменшились розміри файлів, зменшився ризик пошкодження та кращу якість стислих зображень. Завдяки відкритості стандарту Office Open XML, практично кожний текстовий процесор може читати файли формату DOCX.

Стандарт PDF. PDF використовується для відображення електронних документів незалежно від платформи, обладнання або операційної системи, на якій він переглядається. PDF формат на сьогодні – є найбільш вживаним форматом документу у мережі Інтернет. Його перша версія, 1.0, була опублікована у 1993 році компанією Adobe Systems Incorporated. З тих пір PDF став широко використовуваним стандартом для збору та обміну відформатованими документами в електронному вигляді через Інтернет, електронною поштою та майже будь-яким засобом обміну документами. У 2008 році PDF 1.7 набув вигляд ISO стандарту (ISO 32000-1:2008) [2, с. 3].

PDF може містити не тільки текст, а й зображення, елементи мультимедії, гіперпосилання тощо. Базова структура PDF зображена на рис. 1. [3, с. 39].

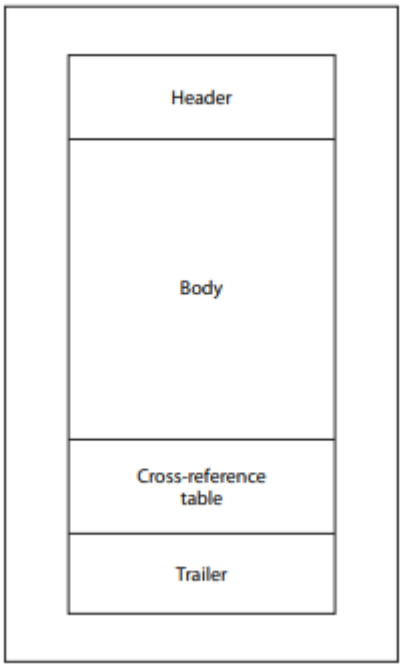

Рис. 1. Базова структура PDF

Елемент «header» містить в собі метадані файлу: версію стандарту PDF та іншу інформацію для програми, що буде відображати цей файл.

Елемент «body» - основна частина документу, що вміщає у себе текст, зображення, гіперпосилання та інші елементи мультимедії. Цей елемент використовується для зберігання усієї інформації, що буде відображена для користувача.

Елемент «xref table» зберігає посилання на усі об'єкти у документі. Її основна ціль – надання довільного доступу до об'єктів у файлі, завдяки цьому програмі не потрібно прочитувати увесь файл, щоб знайти окремий об'єкт.

Елемент «trailer» надає програмі, що читає документ, інструкції по тому, як знайти таблицю посилань та деякі особливі об'єкти у документі. Усі програми PDF-reader повинні починати читання файлу саме з цього елементу [3, с. 3, 9-43].

Мультиплатформеність документу формату PDF досягається завдяки абсолютному позиціонуванню кожного об'єкта (тексту, зображення тощо) на сторінці. Кожна сторінка таким чином є абсолютно незалежною. На відміну від DOC, де вміст абзаців та таблиць автоматично змінює форму, аби заповнювати вільний простір або обтікати навколо різних форм. Позиція кожного елементу є не абсолютною, а відносною щодо інших елементів, полів або абзаців.

Слід зазначити, що у PDF існує дуже обмежений функціонал редагування, адже цей формат був створений саме для передачі інформації, що є платформо-незалежним.

Конвертація. Оскільки формат DOC став основним до появи його відкритих специфікацій, було багато спроб переробити його, щоб принаймні прочитати дані в документі. Але підхід до проблеми такий, що необхідно отримувати не лише текст, а й графіку та інші дані. Аналіз доступного програмного забезпечення показав, що найпростіший спосіб отримати всю необхідну інформацію з файлів DOCX і DOC - це використовувати бібліотеки взаємодії Microsoft Office Interop або LibreOffice Writer. Розробка власного

рішення займає багато часу, готові реалізації або надто дорогі, або не відповідають необхідним умовам. Отже, раціональним рішенням є придбання пакету MS Office або використання LibreOffice Writer, але у цьому випадку виникнуть проблеми з перетворенням графічних зображень.

Що стосується формату PDF, ситуація дещо інша. Є відкриті безкоштовні бібліотеки, але існує два варіанти вирішення проблеми - встановити програму-конвертер, яка забезпечить перетворення PDF-документів в формат DOC, або скористатися спеціальним онлайн сервісом. Тому, використовуючи безкоштовне програмне забезпечення, можна перетворювати описані вище формати, не витрачаючи багато часу на написання програмного коду [4].

За своєю суттю конвертація з документу, де кожна сторінка і елемент є фіксованим (PDF) у документ з відносним позиціонуванням елементів має свої складнощі, пов'язані з навчанням штучного інтелекту. Символи складаються у рядки, рядки - в абзаци, і у кінці, вимірюються відступи та відстані до полів. У більш складних випадках присутні також таблиці та зображення. В теорії, зміст файлу PDF може бути настільки комплексним, що його неможливо буде представити у DOC файлі. Кожен документ має свою структуру: рекламний памфлет має повністю відмінну від наукової статті структуру, що в свою чергу відрізняєтсья від посадової інструкції тощо. Постає проблема у навчанні нейронної мережі для класифікації подібних макетів і розумінні контексту, для успішної конвертації.

Висновок. Іноді, навіть людині важко розпізнати, де закінчується реклама та починається текст статті; що є частиною графіку, а що є її підписом, де починається чи закінчується комірка таблиці. В залежності від автора, документ може мати різний контекст, що також може значно відрізняти його серед інших: наприклад, документ, написаний арабською мовою, скоріш за все

буде читатися справа наліво, а наукова стаття з математики може мати формули, що неможливо буде описати людині, що її не вивчала. Тому задача конвертації одного документа у формат іншого стає особливо складною.

Таким чином, існує два варіанти: навчити конвертер, що в майбутньому зможе виконувати дуже точну конвертацію, але це потребує великого об'єму даних та ручного позначення кожного елементу документа: тексту, таблиць або малюнків; або створити список певних правил, що систематично описують загальний вигляд таблиці чи формули, що не є доволі точним. Вибір методу конвертації у цілому залежить від задачі, поставленої перед конвертором. Але у кінці кінців успішна конвертація вимагає знання побудови вихідного та цільового форматів, у разі відсутності стандарту формату необхідно використовувати зворотню розробку, і хоча результат буде наближено відповідати стандартам, будуть присутні також помилки, або навіть відсутній деякий функціонал.

Література

1. Microsoft. Documentation. [MS-DOC]: Word (.doc) Binary File Format / Microsoft // Microsoft. 2021. URL: https://docs.microsoft.com/en-us/openspecs/office_file_formats/ms-doc/ccd7b486-7881-484c-a137-51170af7cc22
2. M. Hardy. & Masinter, Larry. The application/pdf Media Type. Internet Engineering Task Force // Internet Engineering Task Force. 2017. URL: https://tools.ietf.org/html/rfc8118
3. Document management — Portable document format — Part 1: PDF 1.7. ISO 32000-1:2008. 2008. URL: https://www.adobe.com/content/dam/acom/en/devnet/pdf/pdfs/PDF32000_2008.pdf

4. Білощицький А. О. Перетворення файлів різних типів до єдиного формату / А. О. Білощицький, О. В. Діхтяренко, Т. О. Лященко // Управління розвитком складних систем. 2014. Вип. 18. С. 140-144. URL: http://nbuv.gov.ua/UJRN/Urss_2014_18_25