12º Congresso de Inovação, Ciência e Tecnologia do IFSP - 2021

IMPLEMENTAÇÃO DE MÉTODOS PARA EXTRAÇÃO DE CONTEÚDO TABULAR EM DOCUMENTOS PDF

SOUZA. LUCAS EDUARDO PADUAM DE[1], CORRÊA, ANDREIWID SHEFFER[2]

[1] Graduando em Tecnologia de Análise e Desenvolvimento de Sistemas, Bolsista PIBIFSP, IFSP, Câmpus Campinas, lucas.paduam@aluno.ifsp.edu.br.
[2] Docente, IFSP, Câmpus Campinas, andreiwid@ifsp.edu.br
Área de conhecimento (Tabela CNPq): 1.03.03.04-9 Sistemas de Informação

RESUMO: O acesso à informação e o volume de dados gerados cresceu com o advento da web. Um dos temas de grande interesse da sociedade são os Dados Abertos Governamentais (*Open Government Data*). A quantidade de dados gerados pelos órgãos públicos é enorme e devem ser disponibilizados ao público de forma aberta, estruturada e legível por máquina. Atualmente, esses dados são disponibilizados periodicamente em portais governamentais e de forma aberta, sendo, em grande parte, publicados em formato PDF (*Portable Document Format*). Este tipo de documento eletrônico é muito popular, no entanto, estes documentos estão repletos de dados contidos em tabelas, motivo este de preocupação para comunidade de Dados Abertos, uma vez que, essa prática, torna difícil a extração dos dados e a automatização do processamento por máquina. Este projeto tem como objetivo pesquisar, identificar e implementar melhorias em softwares abertos que se dispõem a solucionar a extração de dados tabulares de arquivos PDF, contribuindo para comunidade, e para tal, identificou como o mais promissor e reconhecido software aberto para extração de dados tabulares o software Tabula, no qual o processo não é realizado apenas por linhas de comando, possuindo também uma interface gráfica, facilitando a utilização por um usuário comum.

PALAVRAS-CHAVE: dados abertos, extração, dados tabulares, tabula.

IMPLEMENTATION OF METHODS FOR EXTRACTION OF TABULAR CONTENT IN PDF DOCUMENTS

ABSTRACT: Access to information and the volume of data generated has grown with the advent of the web. One of the topics of great interest to society is Open Government Data. The amount of data generated by public agencies is enormous and must be made available to the public in an open, structured, and machine-readable way. Currently, these data are periodically made available on government portals and in an open form, and are mostly published in PDF (Portable Document Format). This type of electronic document is very popular, however, these documents are full of data contained in tables, a reason for much concern for the Open Data community, since this practice makes it difficult to extract the data and automate the machine processing. This project aims to research, identify and implement improvements in open source software that are willing to solve the extraction of tabular data from PDF files, contributing to the community, and to this end, identified as the most promising and recognized open source software for tabular data extraction the Tabula software, in which the process is not only carried out by command lines, it also has a graphical interface, facilitating its use by a common user.

KEYWORDS: open data, extraction, tabular data, tabula.

INTRODUÇÃO

Um dos mais populares tipos de documentos em formato digital é o PDF (*Portable Document Format*). Toda essa popularidade se tornou uma preocupação quando a comunidade de Dados Abertos Governamentais identificou que tabelas contendo uma infinidade de dados passaram a ser incluídas em PDF e posteriormente disponibilizadas ao público (RIBEIRO; ALMEIDA, 2011). A preocupação se deve ao fato de que este método de disponibilização, além de não ser uma boa prática, compromete o processamento dos dados tornando mais difíceis tarefas como a fiscalização por órgãos governamentais e pela sociedade (CÔRREA; ZANDER, 2017). Este projeto tem como objetivo estudar os métodos de extração de dados já concebidos e contribuir com melhorias nas ferramentas utilizadas pela comunidade. Dentre eles, tem-se o software aberto mais promissor e reconhecido pela comunidade, atualmente, denominado Tabula (CREATIVE COMMONSa, s.d.). Este software apresenta uma interface de fácil utilização para o usuário comum, extraindo dados tabulares conforme demarcado pelo usuário ou através de métodos de detecção automática de tabelas. No entanto, os resultados da extração nem sempre são apresentados com exatidão exigindo tratamento dos dados após a extração, fato este que dificulta muito a tarefa de extrair e tratar uma enorme quantidade de dados e inviabiliza o processamento por máquina.

MATERIAL E MÉTODOS

Inicialmente, para desenvolver a pesquisa relacionada aos métodos já empregados na extração de dados tabulares por meio do software Tabula, foi necessário realizar a extração dos dados a fim de identificar funções já estabelecidas e limitações do software. Para isso, foram testados 15 documentos do tipo PDF, contendo tabelas simples e complexas (p. ex. tabelas que se utilizam do recurso de mesclagem de células). Os testes foram realizados utilizando as funções de seleção manual e seleção automática de tabelas. Desta forma, foram identificados pontos onde os dados extraídos perderam a formatação original e foram alocados em células ou colunas incorretas e também comprovado que todos os dados foram extraídos.

Em seguida, se fez necessário entender o código-fonte do software, de modo a compreender como o algoritmo realiza a identificação dos caracteres, linhas e colunas de uma tabela, qual o método utilizado para extração dos dados e quais as configurações para apresentação do resultado, e para tal, foi realizado o download do código-fonte do Tabula (versão 1.2.1) do repositório Github (CREATIVE COMMONSb, s.d.).

Em seguida, foi necessário efetuar o download e instalação do ambiente de desenvolvimento Netbeans (IDE versão 12.3) , uma vez que o algoritmo do Tabula foi escrito e utiliza bibliotecas em linguagem de programação Java. Para isso, também foi preciso realizar a instalação do kit de desenvolvimento JDK JAVA SE (versão 17).

Com o código-fonte em máquina e as ferramentas de desenvolvimento instaladas, buscou-se pelas dependências (bibliotecas) do código (STACKOVERFLOW, 2020). Como o Tabula se utiliza de outros softwares abertos em sua composição, por exemplo, o Fontbox e o Pdfbox, estas dependências devem ser atribuídas ao código para que funcione corretamente. Desta forma, pode ser instanciado e sua utilização iniciada no ambiente Netbeans.

O Tabula, assim como vários softwares abertos, possui diversos colaboradores e se utiliza de outros softwares abertos em sua composição, apresentando diferentes padrões na escrita do código e documentação reduzida, dificultando a compreensão dos métodos utilizados. Também há muitas classes atribuídas ao código que não são utilizadas, mas precisam ser analisadas e testadas. As funções de cada uma das classes estão relacionadas aos nomes dados a estas classes e deste modo buscam ser intuitivas para o leitor, como *ExtractionAlgorithm* e *RetangularTextContainer,* as quais são exemplos de como são denominadas, sendo as classes principais compostas de sub-classes (ROSÉN, 2019). Em geral, o processo de extração pode ser definido em quatro partes: Detecção, Pré-extração, Extração, Escrita.

A princípio é realizada a Detecção da tabela, que pode ser automática ou pré-definida, manipulando suas coordenadas em relação à página, por meio de um retângulo na área desejada. A Pré-extração é responsável pela interpretação do arquivo PDF bruto, uma vez que, quando um arquivo PDF é gerado, perde todas as características do arquivo que o originou, assumindo um código próprio para descrição dos elementos contidos na página. Na parte de Extração, são armazenados os dados em

listas ou índices de forma a gerar uma tabela abstrata, restando a função de escrita a compilação desses dados em forma de linhas e colunas.

RESULTADOS E DISCUSSÃO

Nesta pesquisa, implementou-se o software Tabula no ambiente de desenvolvimento Netbeans e foram realizados testes de extração em 15 documentos em formato PDF contendo tabelas, imprimindo os resultados em tela. Para melhor entendimento dos tipos de tabelas extraídas e dos tipos de problemas encontrados, foram definidos graus de complexidade para as tabelas, bem como os tipos de problemas. A Tabela 1 mostra os resultados obtidos nas extrações realizadas por meio da versão interativa do software.

Tipos de tabela por grau de complexidade

- Simples - Tabela composta por linhas e colunas bem definidas não utilizando recurso de mesclagem de células;
- Média complexidade - Tabelas com linhas e colunas bem definidas que utilizam recursos de mesclagem de células em cabeçalhos;
- Alta complexidade - Tabelas que utilizam recursos de mesclagem nos cabeçalhos e nas linhas e colunas internas da tabela.

Tipos de problemas apresentados

1. Dados apresentados em colunas incorretas;
2. Dados de colunas diferentes atribuídos a uma única coluna;
3. Células mescladas lidas como linhas e colunas diferentes causando confusão na apresentação dos dados;
4. Dados dispersos em diferentes células que impossibilitam a identificação das linhas e colunas a qual pertencem.

Tabela 1. Resultados obtidos por meio da versão interativa do software Tabula considerando a complexidade das tabelas testadas e erros encontrados.

Tabela	Simples	Média Complexidade	Alta Complexidade	Porcentagem
Quantidade analisada	5	5	5	100%
Resultado 100% correto	5	0	0	33,3%
Problema tipo 1	0	4	4	53,3%
Problema tipo 2	0	1	2	20%
Problema tipo 3	0	1	3	26,6%
Problema tipo 4	0	0	2	13,3%

Conforme o teste realizado, foi verificado no processo de extração que existem muitas inconsistências, representando 66,7% dos casos avaliados, sendo que na extração de tabelas de média e alta complexidade 100% dos casos apresentam algum tipo de inconsistência, em sua maioria, 53,3% das tabelas analisadas, dados que são apresentados em colunas que não correspondem às colunas originais. Casos onde as células mescladas são interpretadas como linhas e colunas diferentes causando confusão na apresentação dos dados representam 26,6%.

UNIDADE ORÇAMENTÁRIA		QUANTIDADE						
		AUXÍLIO-ALIMENTAÇÃO	ASSISTÊNCIA PRÉ-ESCOLAR	AUXÍLIO-TRANSPORTE	EXAMES PERIÓDICOS	ASSISTÊNCIA MÉDICA E ODONTOLÓGICA		
CÓDIGO	DESCRIÇÃO					TITULARES	DEPENDENTES	TOTAL
01101	CÂMARA DOS DEPUTADOS	14.597	3.137	111	-	19.606	16.008	35.614
01901	FUNDO ROTATIVO DA CÂMARA DOS DEPUTADOS	-	-	-	-	419	342	761
TOTAL		14.597	3.137	111	-	20.025	16.350	36.375

FIGURA 1. Exemplo de tabela classificada com grau de complexidade alta.
Fonte: Câmara dos Deputados - Departamento de Pessoal e Departamento de Orçamento, Contabilidade e Finanças.

	A	B	C	D	E	F
3	UNIDADE ORÇAMENTÁRIA					
4		AUXÍLIO-	ASSISTÊNCIA	AUXÍLIO-	EXAMES	ASSISTÊNCIA MÉDICA E ODONTOLÓGICA
5	CÓDIGO DESCRIÇÃO	ALIMENTAÇÃO	PRÉ-ESCOLAR	TRANSPORTE	PERIÓDICOS	TITULARES DEPENDENTES TOTAL
6	01101 CÂMARA DOS DEPUTADOS	14597	3137	111	-	19.606 16.008 35.614
7	01901 FUNDO ROTATIVO DA CÂMARA DOS DEPUTADOS	-	-	-	-	419 342 761
8	TOTAL	14597	3137	111	-	20.025 16.350 36.375

FIGURA 2. Ilustra o resultado da extração da tabela referenciada na FIGURA 1, apresentando o tipo de problema 2 (dados de colunas diferentes atribuídos a uma única coluna).

Verificou-se que o método de extração se baseia na identificação de um retângulo principal definido através das coordenadas da página e subsequente identificação de caracteres em linha, calculando os espaços entre eles, para definir células e posição na página em relação às linhas para definir as colunas.

Além da versão final do Tabula, foi instanciado, em paralelo, uma de suas principais bibliotecas, a Pdfbox (JAVAPOINT, s.d.), na qual constatou-se que as principais funções de reconhecimento e extração de dados foram efetuadas pelo Tabula. No intuito de aprofundar-se nos métodos principais de extração, foi localizado um repositório da Apache (VIEWVC, s.d.), responsável pela Pdfbox, onde foram encontradas diversas tentativas de melhorias no algoritmo das classes do software, desde os métodos de identificação dos caracteres até as classes de representação final das tabelas. Neste ponto da pesquisa, foi instanciada a classe PDFStreamStripper (TI-ENXAME, s.d.). Com esta classe, foi possível manipular as coordenadas do retângulo de detecção das tabelas e obter resultados em forma de linhas e colunas contendo as respectivas coordenadas de cada grupo de caracteres. A classe não trouxe melhorias aos resultados apresentados na Tabela 1, mas permitiu obter uma visão mais clara de como o processo de identificação das coordenadas da página estão ligados ao método de extração dos dados.

CONCLUSÕES

Em busca de aprimoramento e maior taxa de acurácia nos softwares livres utilizados para extração de dados tabulares contidos em arquivos PDF, este artigo se propôs ao estudo de um algoritmo ou ferramenta livre neste tipo de atividade. Devido à falta de documentação, tanto do software Tabula quanto de suas dependências (PDFBOX), a identificação das funções realizadas pelas classes se tornou custosa e foi gasto muito tempo em testes para descoberta de funções e parâmetros, reduzindo assim as expectativas de desenvolver um algorítimo para melhoria do processo de extração.

De modo geral, pode-se concluir que a manipulação das coordenadas dos retângulos de seleção em modo manual tende a ter resultados mais exatos que as detecções automáticas (processamento por máquina), inviabilizando os objetivos da comunidade de dados abertos. Este é um problema que ainda parece estar longe de ser resolvido, uma vez que, este software precisa ser analisado completamente, eliminando conteúdo obsoleto, documentando as classes de maneira clara e objetiva. Neste sentido, novas abordagens devem ser implementadas, abordagens estas que partem desde de o reconhecimento de várias linhas em uma única célula até a utilização de novas tecnologias como o Reconhecimento Óptico de Caracteres OCR (*Optical Character Recognition*).

AGRADECIMENTOS

Agradeço ao Programa Institucional de Bolsas de Iniciação Científica e Tecnológica do Instituto Federal de Educação, Ciência e Tecnologia de São Paulo (PIBIFSP) e ao respectivo orientador pela oportunidade de participar deste projeto de pesquisa, abrindo novos horizontes de aprendizado, desenvolvimento e contribuição para a comunidade.

REFERÊNCIAS

CORREA, Andreiwid Sheffer; ZANDER, Par-Ola. Unleashing tabular content to open data: A survey on pdf table extraction methods and tools. 18th International Conference on Digital Government Research, 2017. p. 01–10. Disponível em <http://doi.acm.org/10.1145/3085228. 3085278>. Acesso em: 21 Mar 2021

CREATIVE COMMONSa. Escoladedados.org: Libertando dados com TABULA E ROWS, Tutoriais, s.d. Disponível em: <https://escoladedados.org/tutoriais/libertando-dados-com-tabula-e-rows/> Acesso em: 13 Mar. 2021.

CREATIVE COMMONSb. Escoladedados.org: Usando o TABULA na linha de comando, Tutoriais, s.d. Disponível em: <https://escoladedados.org/tutoriais/usando-o-tabula-na-linha-de-comando/> Acesso em: 29 Mar. 2021.

JAVATPOINT. Javatpoint.com: PDFBox Tutorial, Pdfbox-tutorial, s.d. Disponível em: <https://www.javatpoint.com/pdfbox-tutorial> Acesso em: 12 Mai. 2021.

RIBEIRO, C.J.S.; ALMEIDA, R.F. Dados abertos governamentais (Open Government Data): instrumento para exercício de cidadania pela sociedade. Brasília, 2011. Disponível em: <http://arq.3rengtt.com.br/wp-content/uploads/2015/09/EnancibXII-RibeiroAlmeida.pdf>
Acesso em: 04 Mar. 2021.

ROSÉN, G. Analysis of Tabula: a PDF-Table extraction tool. Suécia, 2019. Disponível em: <http://uu.diva-portal.org/smash/get/diva2:1363917/FULLTEXT01.pdf> Acesso em 07 Mar. 2021.

STACKOVERFLOW. Stackoverflow.com: How can tabula (jar) be called from java?, Questions, 2020. Disponível em: <https://stackoverflow.com/questions/52866639/how-can-tabula-jar-be-called-from-java> Acesso em: 20 Abr. 2021.

SHUTTLEWORTH FUNDED. TABULA: HOW TO USE, s.d. Disponível em: <https://tabula.technology/> Acesso em: 15 Mar. 2021.

TI-ENXAME. Ti-enxame.com: Análise PDF arquivos (especialmente com tabelas) com PDFBOX. s.d. Disponível em : <https://www.ti-enxame.com/pt/java/analise-pdf-arquivos-especialmente-com-tabelas-com-pdfbox/969427550/> Acesso em 15 Mai. 2021.

VIEWVC. Svnapache.org: Contents of. Pdfbox, s.d. Disponível em: <https://svn.apache.org/viewvc/pdfbox/> Acesso em 25 Mai. 2021.