FERRAMENTA PARA EXTRAÇÃO DE INFORMAÇÕES DE PORTAIS DE TRANSPARÊNCIA PÚBLICA DISPOSTAS EM FORMATO NÃO ABERTO / DEVELOPMENT OF A TOOL EXTRACTION OF TRANSPARENCY WEB SITES INFORMATION AVALIABLE IN NON-OPEN FORMAT

Evandro Henrique Couto de Paula[1] (aluno autor); Paulo Henrique Pereira Cardoso(aluno co-autor); Andreiwid Sheffer Corrêa[1] (orientador);
1-Instituto Federal de Educação, Ciência e Tecnologia – *Campus* Campinas;

Resumo/Abstract

Este trabalho trata do uso de uma ferramenta desenvolvida em linguagem java que objetiva a extração de dados abertos governamentais dispostos em formatos não abertos, especialmente PDF e HTML.

This work talks about a tool developed in Java language that has the objective the extraction of governamental data disposed in not open fomat specially PDF and HTML.

Introdução/Introduction

Segundo Avritzer (2007) o Brasil vem enfrentando um aumento considerável na participação politica do cidadão na politica brasileira. Isto só é possível graças ao acesso do público a informação.

O trabalho descreve o processo de criação de uma ferramenta para extração de dados públicos dispostos em formato não aberto.

Avritzer (2007) defines that Brasil has facing a rise number of citzens interested in politics, this is just possible through access to information.

This work describe the process of development of a tool to extract data of public documents available in non-open formats.

Materiais e Métodos/ Material and Methods

O aplicativo desenvolvido utiliza como base a linguagem Java, e como suporte a HTML, já que o aplicativo funciona como um Java Applet inserido na página HTML.

O aplicativo utiliza as bibliotecas externas PDFBox e Jsoup e se integrará com a ferramenta Tabula extractor funcionando em um webservice.

A extração do PDF ocorre através da obtenção do documento pdf utilizando sua URL. O documento é então convertido em imagem utilizando PDFBox, utilizando esta imagem a área de seleção de extração é escolhida, então estas informações são enviadas ao webservice para que o Tabula execute a extração.

O HTML é extraído a partir do acesso a página e a biblioteca Jsoup, realiza as buscas tendo como parâmetro as tags de tabela.

Os dados extraídos são retornados para o usuário em formato CSV.

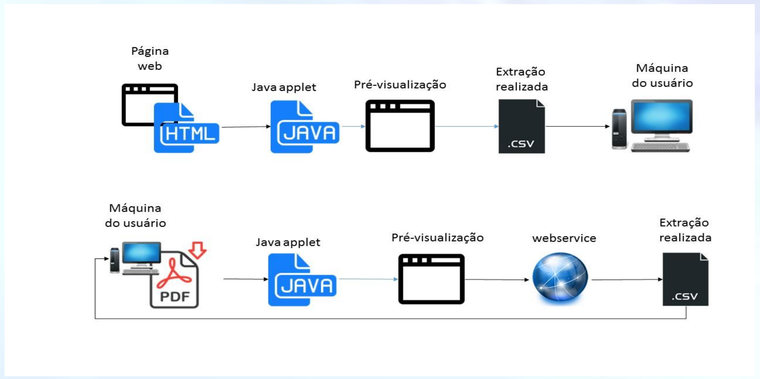

Figura 1: Fases da extração de documento.

The application was developed using Java language and has as support HTML, because it will run as a Java Applet.

The application uses external libraries PDFBox and Jsoup and integrates with Tabula extractor tool woking in a webservice.

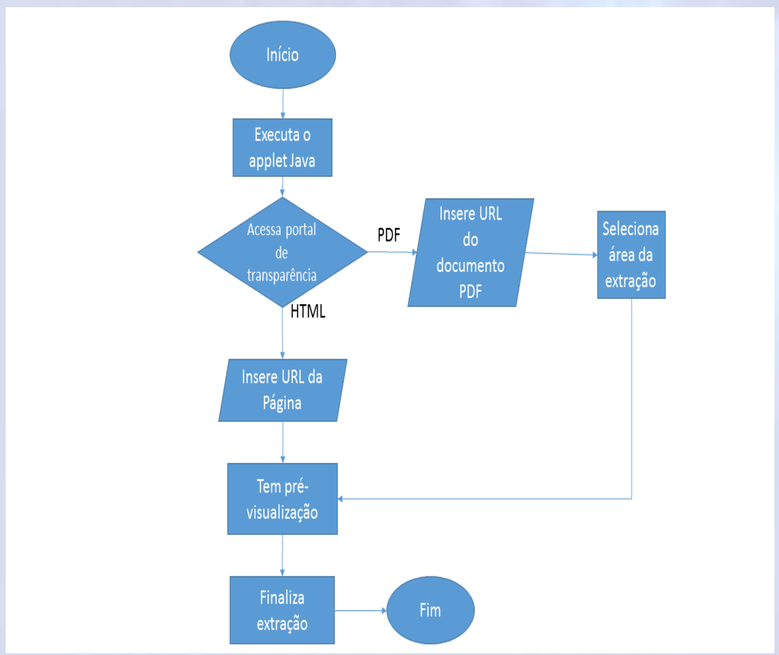

Figura 2: Fluxo de extração do documento.

The PDF extraction occurs saving the PDF using the URL to get information. The document its parsed into a image format and the coordinates for extractions are selected and send to Tabula on the webservice to execute the extraction.

The HTML data are extract from the access of the HTML page and using Jsoup library it performed a search tags that compose a HTML table.

The extracted data its returned to user as a CSV file.

Resultados/Results

As extrações de páginas HTML que possuem informações estáticas geraram resultados positivos e a extração foi possível, em páginas com conteúdo dinâmico ainda não foram obtidos resultados positivos

Os documentos PDF ainda não foram extraídos, pois ainda não foram integrados a aplicação e o webservice.

The HTML extractions were succeed in static page however in HTML pages with dynamic content extraction weren't possible yet.

The PDF documents were not extracted tet bacause the integration with the webservce were not developed yet

Conclusão/Conclusion

O trabalho mostrou que a manipulação de documentos encontrados em formatos não abertos não é tão simples, porém é possível e a extração pode ocorrer.

The work shows that manipulate documents found in non-open format is not ease, however it is possible to extract the data locked inside them.

Referências/ References

AVRITZER, Leonardo. Sociedade civil, instituições participativas e representação: da autorização à legitimidade da ação. **Dados**, v. 50, n. 3, p. 443-464, 2007.

.

Fomento/Promotion