

# Automated Data Extraction from PDF Documents: Application to Large Sets of Educational Tests

Karina Wiechork<sup>1,2</sup> <sup>a</sup> and Andrea Schwertner Charão<sup>1</sup> <sup>b</sup>

<sup>1</sup>*Department of Languages and Computer Systems, Federal University of Santa Maria, Santa Maria, Brazil*

<sup>2</sup>*Information Technology Coordination, Federal Institute of Education Science and Technology Farroupilha, Frederico Westphalen, Brazil*

**Keywords:** Dataset Collection, Ground Truth, Performance Evaluation, PDF Extraction Tools.

**Abstract:** The massive production of documents in portable document format (PDF) format has motivated research on automated extraction of data contained in these files. This work is mainly focused on extractions of natively digital PDF documents, made available in large repositories of educational exams. For this, the educational tests applied at Enade were used and collected automatically using scripts developed with Scrapy. The files used for the evaluation comprise 343 tests, with 11.196 objective and discursive questions, 396 answers, with 14.475 alternatives extracted from the objective questions. For the construction of ground truth in the tests, the Aletheia tool was used. For the extractions, existing tools were used that perform data extractions in PDF files: tabular data extractions, with Excalibur and Tabula for answer extractions, textual content extractions, with CyberPDF and PDFMiner to extract the questions, and extractions of regions of interest, with Aletheia and ExamClipper for the cutouts of the questions. The results of the extractions point out some limitations in relation to the diversity of layout in each year of application. The extracted data provide useful information in a wide variety of fields, including academic research and support for students and teachers.

## 1 INTRODUCTION

With the development of information technology and the wide spread of the Internet, a large amount of electronic documents are stored in PDF files (Fang Yuan and Bo Lu, 2005).

PDF is one of the most widely used document formats for storing text based data. This file format was designed by Adobe in 1993 with the purpose of representing a document, regardless of the platform used, and preserving the layout on the screen and in printing.

While this is an efficient way to store the visual representation of a document, the resulting structure is difficult to work with if the aim is to extract specific parts of the text in a structured manner (Budhiraja, 2018).


One of the great advances in the digital era has been to enable us to store vast amounts of documents electronically (Øyvind Raddum Berg, 2011). The substitution of physical document storage for elec-


tronic storage, provides advantages such as: cost reduction, easy storage and sharing, optimization in searches and queries, documents are not damaged and digital documents have a structural standardization, for example in paragraphs, sections, titles, figures, which can be useful to detect regions and extract information in high demand, in an automated or semi-automated manner.

This research aims to carry out an exploratory analysis on tools used in data extraction in documents born in PDF, whose objective is to discover their effectiveness and limitations. Retrieving relevant information in the questions of these tests is a difficult task, since the layout is not geometrically simple.

Extracting information from PDF files is an important job, since these extracted questions are very valuable knowledge assets for research, providing useful and timely information for several users who may benefit, for example, research material to students who intend to study for tests, courses or public contests, using as an object of study interesting for learn to retain new knowledge.

This information can also help course coordinators to analyze the effectiveness of Pedagogical

<sup>a</sup>  <https://orcid.org/0000-0003-2427-1385>

<sup>b</sup>  <https://orcid.org/0000-0003-3695-8547>

Course Projects, mapping students' knowledge and discovering gaps from the results of the questions in reports. In addition to becoming a set of interesting material to be used in the classroom by teachers, in order to facilitate understanding, as well as use these questions for exercises. In the case of a teacher, he can have a database of questions and answers and from there generate new tests.

For this work, the extractions of the educational tests were carried out through the ENADE tests (National Student Performance Test) applied in the years 2004 to 2019. The set of downloaded files, consists of 386 tests and 396 answers, totaling 782 files, however, only tests with more than two applications were used. The number of tests, pages for extractions and questions are 343, 6.834, 11.196 respectively, while the total of alternatives in the 396 objective responses is approximately 14.475. Our dataset for PDF extraction totals 739 files, accounting for 343 tests and 396 answers.

Knowing that it is possible to extract data in this type of test, it will also be possible to extract from other tests, just by reusing the tools used in this work.

The remainder of this paper is organized as follows. Section 2 discusses some basic topics and related work that included experiments involving extracting data from PDF files. In section 3, we present the dataset used and extracted, together with the methodology used to obtain these data. Section 4 details the experiments and the results we obtained. Finally, section 5 concludes this article with a brief summary and suggestions for future research.

## 2 BACKGROUND AND RELATED WORK

### 2.1 Enade

In Brazil, the National Institute of Educational Studies and Research Anísio Teixeira (INEP) is responsible for applying Enade. The results of the tests present several indicators, among them the concept of the course that varies from 0 to 5 for the courses. Based on the analysis of the data obtained by the application of Enade, it is possible to analyze the performance of both institutions and students, and then calculate quality indicators that may provide opportunities for improvement decisions in the teaching process.

Enade assesses the performance of graduates of undergraduate courses in relation to the syllabus foreseen in the curricular guidelines of the courses, the

development of competencies and skills necessary for the deepening of general and professional formation, and the level of updating of students in relation to the reality Brazilian and world (INEP, 2020).

Applied by Inep since 2004, Enade is part of the National Higher Education Assessment System (Sinaes), which also comprises undergraduate courses and institutional assessment. Together, they form the evaluative tripod that allows to know the quality of Brazilian higher education courses and institutions. The results of Enade are inputs for calculating the Higher Education Quality Indicators (gov.br, 2021).

The test consists of 40 questions, where 10 questions make up the general formation and 30 specific formation in the area, both parts contain discursive and multiple choice questions. The general formation part has 25% of the test and 75% is for specific formation, showed at Table 1.

Table 1: Values for each part of the test.

|                 | General Formation | Specific Formation |
|-----------------|-------------------|--------------------|
| Discursive      | 2                 | 3                  |
| Multiple Choice | 8                 | 27                 |
| Peso            | 25%               | 75%                |

### 2.2 Ground Truth

For automatic evaluation of results of any segmentation/recognition system, the ground truth information plays a significant role. However, it is an error prone and time consuming task (Alaei et al., 2011).

In document image understanding, public datasets with ground truth are an important part of scientific work. They are not only helpful for developing new methods, but also provide a way of comparing performance. Generating these datasets, however, is time consuming and cost-intensive work, requiring a lot of manual effort (Strecker et al., 2009).

To assist in this research in the task of creating the ground truth on each page with question in the Enade tests, the software was used Aletheia, belonging to one of the PRIMA research group (Pattern Recognition & Image Analysis Research Lab) at the University of Salford Manchester. The fact that it has been widely adopted in similar studies, is maintained by a research group, updated and presents several options for working, are contributing factors for use in this work.

The figure 1 shows a page in Aletheia with the regions of interest marked. The workflow of Aletheia consists of the steps of input, which includes the page,

and output, where the segments are classified and saved in eXtensible Markup Language (XML).

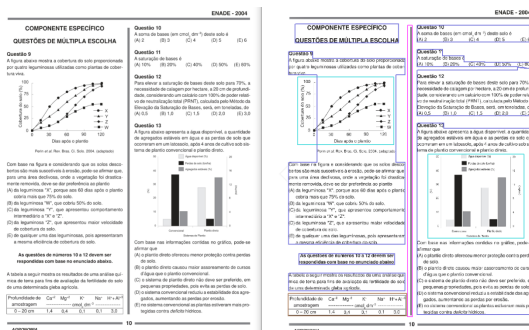


Figure 1: Example of input (left) and output (right) with 12 marked regions belonging to 4 different types: image (light blue), table (brown), text (dark blue) and separator (pink).

About 6.834 pages of the PDF tests were collected. For each page of the Enade tests that contains questions, a ground truth file is available and the corresponding XML file built with the Aletheia tool. The files are in XML format and contain the coordinates of the regions in a hierarchical structure.

In terms of presentation, each page with the content of the test is composed of 3 parts: 1) one with identified ground truth; 2) the original page in jpg format; 3) an XML description of the attributes contained and built with Aletheia, according to the selected regions. This files contains detailed data that can meet other extraction requirements and be used in other research. In this case, it was used to compare with the output of the textual extraction tools, PDFMiner and CyberPDF, in addition to being used for the metrics count script, which will be explained in the section 3.2.

### 2.3 PDF Extraction Tools

In this section, works by other authors related to this research are analyzed and described. The authors (Constantin et al., 2013) present a system designed to reconstruct the logical structure of academic PDF articles, the PDFX tool. The output is an XML or HyperText Markup Language (HTML) document that describes the logical structure of the input article in terms of title, sections, tables, references, etc. When using HTML output, the figures are also extracted and are available at the end of the file, but not in the order of reading.

In this direction, the work of (Ramakrishnan et al., 2012), the authors develop a tool for extracting text from PDF with layout recognition (LA-PDFText) whose objective is accurately extract text from scientific articles. The scope of the system is only in

extractions in the textual content of the research articles. In (Hadjar et al., 2004), attackers describe an approach in which they extract all objects from a PDF document, including text, images and graphics, entitled Xed (eXtracting electronic documents). The output of the extracted objects is in SVG (Scalable Vector Graphics) format.

A number of works in the field of extracting tables from PDF files, such as (Hassan and Baumgartner, 2007), are available. The work (Liu et al., 2007) it presents a system capable of extracting tables and table metadata from PDF documents, for this purpose the PDFBox is used to extract raw text, which is later processed to identify tables. The Tabula (Manuel Aristaarán, Mike Tigas, Jeremy B. Merrill, Jason Das, David Frackman and Travis Swicegood, 2018) tool allows users to select tables for extracting tabular data from PDF documents. Excalibur is a web tool for extracting tabular data from text-based PDFs and not from scanned documents (Excalibur, 2018). The extraction of tables has become useful for this work in performing the extraction of the answers of objective questions.

PDFMiner is a tool for extracting information from PDF documents. Unlike other PDF-related tools, it focuses entirely on obtaining and analyzing text data (Yusuke Shinyama, 2014). In (Parizi et al., 2018) the authors propose a technique that allows users to consult a representative PDF document and extract the same data from a series of files in the form of batch analysis quickly, CyberPDF is an automatic PDF batch extraction tool based on coordinates.

Other approaches focus on evaluating PDF extraction tools. In the research by (Bast and Korzen, 2017), the authors provide an assessment of 14 PDF extraction tools to determine the quality and scope of its functionality, based on a benchmark that they built from parallel TeX and PDF data. They used 12.098 scientific articles and for each article, the benchmark contains a ground truth file in addition to the related PDF file. In (Lipinski et al., 2013) the authors evaluate the performance of tools for the extraction of metadata from scientific articles. The comparative study is a guide for developers who want to integrate the most appropriate and effective metadata extraction tool into their software.

Specific approaches to extract figures and captions from PDF are being proposed. This is the case of the work by (Choudhury et al., 2013), where the authors were concerned with extracting figures and associated captions from PDF documents. In the work of (Li et al., 2018), they present a system for extracting figures and associated legends from scientific publications, PDFFigCapX.

Although several sophisticated and even complex approaches have been proposed, they are still limited in many ways (Strecker et al., 2009).

In the research of (Lima and Cruz, 2019), the authors propose an approach to detect and extract data from unstructured data sources available online and spread across multiple web pages, to store the data in a Data Warehouse properly designed for this. Almost all files are published in PDF and there are files with different layout. For this process, the authors use pre-existing tools.

Through these works, it is possible to identify some tools used in the extraction of data files PDF. This study in related works helped in the exploratory research of some approaches used in this research.

### 3 DATASET AND METHODOLOGY

In this section, we present the dataset collection and the methodology used to conduct our experiments with Excalibur and Tabula, for extracting answers, CyberPDF and PDFMiner, for extracting textual content, Aletheia and ExamClipper, for extracting regions of interest. There are generic tools that process images and allow you to make cutouts. With that we use Aletheia, to do an experiment with some tests, which contemplates this category and allows you to make cuts in regions of interest in PDF files.

#### 3.1 Dataset Collection

The dataset for this research consists of tests and evaluation answers from Enade, composed from the years 2004 to 2019. In an automated way to download these tests and answers on the INEP website, the Scrapy tool was applied to assist in the download these files in an automated way. All automation scripts are available at: <https://github.com/karinawie/scrapy>. The use of these scripts was essential, due to the fact that there is a large set of documents to be collected.

As a result, 386 tests and 396 answers from objective alternatives were collected. This difference in the amount of tests for the answers, although in some years the same test was applied for similar courses, but with different answers.

However, the data set totals 343 tests and 396 answers, totaling 739 files. Not all tests were used due to the large volume and the delay in carrying out the experiments, so it was decided to remove the tests with less than two applications. The dataset is available at: <https://github.com/karinawie/PDFExtraction/tree/master/dataset> with 739 files downloaded.

From the quantitative data contained in the dataset, the number of tests, pages for extractions and questions are 343, 6.834, 11.196 respectively, while the total extraction of responses from the 396 files, totaled 14.475 approximately. In this count, the blank pages, the covers of the tests and the pages of the questionnaire of perception of the test were not included. In all tests the part of General Formation questions was counted only once, in cases where the layout pattern was the same for all tests in the year evaluated. In a few years, more than one layout pattern was identified, with which more than one ground truth was generated for General Formation. Answers are not accounted for with essay questions, only with objective questions.

Table 2 presents an overview with quantitative data from pages and questions that were used in the extractions of the tests.

Table 2: Overview of the data used in the tests.

| Year         | Number of pages | Number questions |
|--------------|-----------------|------------------|
| 2004         | 151             | 430              |
| 2005         | 391             | 884              |
| 2006         | 171             | 340              |
| 2007         | 188             | 440              |
| 2008         | 453             | 939              |
| 2009         | 328             | 595              |
| 2010         | 236             | 550              |
| 2011         | 546             | 956              |
| 2012         | 292             | 520              |
| 2013         | 342             | 520              |
| 2014         | 884             | 1222             |
| 2015         | 439             | 580              |
| 2016         | 351             | 520              |
| 2017         | 905             | 1270             |
| 2018         | 537             | 610              |
| 2019         | 620             | 820              |
| <b>Total</b> | <b>6.834</b>    | <b>11.196</b>    |

#### 3.2 Methodology

This section presents the methodology used to perform the data extractions, in addition to the set of metrics and criteria established for the evaluations applied in the experiments.

After obtaining our dataset, was performed a comparative evaluation of 6 PDF file extraction tools. According to the need for this research, we performed data extractions for 3 categories: data tables, text content and regions of interest for image format. In each category there are 2 tools that extract the same content. After this extraction, a comparison is made with the tool output belonging to the same category. For

this, a set of criteria was established that allow an assessment of the extraction tools comparing the results of the tools with the ground truth.

Performance evaluation is necessary to compare and select the most suitable methods for a given application. Different algorithms have different deficiencies considering all the metrics of evaluation. Ground truth contains sufficient and detailed data in several aspects and it is necessary to use it as a reference to evaluate the results of the experiments (Fang et al., 2012).

To quantify the accuracy when analyzing the performance of the tools, the metrics listed in the Table 3 are used. When evaluating a tool, each of its output files is compared with the equivalent ground truth file, then with the competing tool. The following evaluation criteria are measured.

To compare the metrics of the question extractions with the ground truth, a metric count script was created and is available at [https://github.com/karinawie/XML\\_aletheia](https://github.com/karinawie/XML_aletheia). The ground truth was built with Aletheia, the XML files were created and saved. These XML files were used to create our script. The script counts the metrics for each question according to the year and the test area and exports it to a text file with this information. The script analyzes the XML file, detecting the beginning of the question using a regular expression: "QUESTION or Question or DISCURSIVE QUESTION or Discursive Question", followed by numbers between 0 to 9 with two digits. With that, the quantity of each region in each test was informed in a spreadsheet in an automated way with the help of this script.

Table 3: Metric notations.

| Notations | Signification   |
|-----------|---|
| 1C        | one column on the page  |
| 2C        | two columns on the page   |
| MC        | mixed columns on page   |
| 1QP       | one question per page/column                                    |
| 1QV       | one question that starts on one page/column and ends on another |
| VQP       | multiple questions in one page/column                           |
| QF        | questions with figure/graph                                     |
| QT        | questions with tables   |
| -         | not available in the selected test set                          |
| N         | tool does not recognize   |

Several approaches are currently available for extracting data from PDF. To carry out the comparison of the tools, they must largely have the same general objectives, for this the Excalibur comparison will be performed with Tabula, CyberPDF with PDFMiner,

Aletheia with ExamClipper, since they have resources for similar extractions. ExamClipper is software under development by a research group at the Federal University of Santa Maria, used to extract regions of interest. The Aletheia tool, for providing an option, among the several, to extract similar to ExamClipper, was chosen to use it. For this, the XML files were used, with their respective images, created during the development of the ground truth, that is, the same files were reused to perform the extractions. Obviously this is not a fair comparison, although the output of the two tools and the type of extractions are similar and fit the same extraction category.

In the extraction of the answers, the Excalibur and Tabula tools were used, both work in the same objective: extraction of tabular PDF data. These tools were selected to extract the answers of objective questions.

To calculate the performance of the tools, a simple rule of three was used, where only the complete extractions of the questions are counted. In the following formula, the value of "total questions to extract", is equivalent to the total of questions identified in the ground truth. Then, the simple arithmetic mean in each year was applied, to know the average in each metric. Finally, the average formula was again applied to obtain the average value of the entire dataset for each tool.

$$\frac{\text{questions\_extracted\_by\_tool} * 100}{\text{total\_questions\_to\_extract}}$$

## 4 PERFORMANCE EVALUATION AND RESULTS

### 4.1 Performance Evaluation

To verify the performance of the selected tools, experiments were performed in the dataset. The evaluation criteria introduced in the section 3.2 are easily interpretable, but measuring them is not trivial.

Starting with the category of questions: general formation, for all tests of the same year these questions are always the same. The experiments were carried out only once on these questions, for each year. Then, counting occurred only on specific questions in the area. This approach reduced the amount of computation required and, therefore, reduced the time required to perform the analysis. Initially the work would be applied to 14.386 questions. Finally, we decreased to 11.226 questions objective and discursive used to extract.

Text extraction plays an important role for data processing workflows in digital libraries. Complex

file formats make the extraction process error prone and make it very difficult to verify the accuracy of the extraction components.

Based on digital preservation and information retrieval scenarios, three quality requirements in terms of effectiveness of text extraction tools are identified: 1) is a certain text snippet correctly extracted from a document, 2) does the extracted text appear in the right order relative to other elements and, 3) is the structure of the text preserved (Duretec et al., 2017).

The tools were executed to obtain the final output and then compared with the results of its competitor. Then, both results are compared with ground truth.

Excalibur and Tabula that extract tabular data, both used to extract the answers, the evaluations were only in the objective answers. Discourse answers are not included in the count.

The main objective of the evaluation is to analyze each tool, comparing its output files with the ground truth files, using the set of established metrics and criteria. This was more difficult than expected, especially the part of comparing tool outputs. Then, we will present the results of these experiments.

## 4.2 Results

This section presents the results obtained from experiments carried out using the extraction tools. For each tool, a concise result is provided, according to the criteria addressed. The full results are available at: <https://github.com/karinawie/PDFExtraction>.

The information from the ground truth of each page of the tests, was informed in a spreadsheet to make comparisons with the information extracted from the extraction tools. The analysis of the experimental results demonstrates the effectiveness of the suggested measures and provides valuable information on the performance and characteristics of the evaluated tools.

The Table 4 provides an overview of the evaluation results for each of the PDF extraction tools, in relation to the average time in seconds, required to extract the data from a single PDF file. The value obtained from the average time was calculated in 5 equal tests for all tools, only at the time of extraction without counting the time to attach the tests to the tools. The ExamClipper tool took approximately 4 minutes to complete this task, the Aletheia tool took about 3 minutes. Emphasizing that the results of Aletheia have a bias previously configured in the ground truth, even so it was accounted for.

In the Table 5, below, the results of the Excalibur and Tabula tools are shown together with the performance in extracting the tabular data, which in this

Table 4: Overview of the results of the evaluation process of extracting information from PDFs.

| Tools       | Time |
|-------------|------|
| Excalibur   | 20   |
| Tabula      | 16   |
| PDFMiner    | 16   |
| CyberPDF    | 22   |
| ExamClipper | 240  |
| Aletheia    | 180  |

work were the objective answers. For these tools, only the QF metric was calculated, as the answers are in tables.

Table 5: Overview of the results of the tabular data extraction tools.

| Metrics / Tool | Excalibur | Tabula |
|----------------|-----------|--------|
| 1QP            | N         | N      |
| 1QV            | N         | N      |
| VQP            | N         | N      |
| QF             | N         | N      |
| QT             | 99,4      | 97,7   |

In the Figure 2, in yellow, the ground truth compared to Excalibur, in blue, and Tabula, in red. It is observed that between the years 2005 to 2007, the Tabula tool had a slight difficulty in extracting the alternatives from the answers.

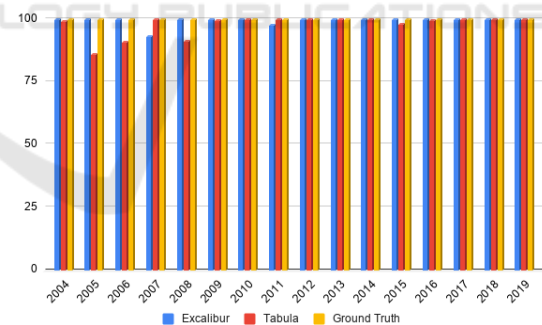


Figure 2: Excalibur and Tabula extraction results detailed by year and compared to ground truth.

The Figure 3 shows that the Excalibur and Tabula tools achieve a result of extraction quantity very close, however the Excalibur tool presents a better performance in data extraction. Regarding the performance of the time to extract the data, Tabula was a little faster.

The metrics 1C, 2C, MC, 1QP, 1QV, VQP, QFG and QT, Table 6, are presented with the simple arithmetic mean of the percentages obtained for the years 2004-2019. According to this table, the metrics for 2 columns (2C) have a relatively low recovery rate,

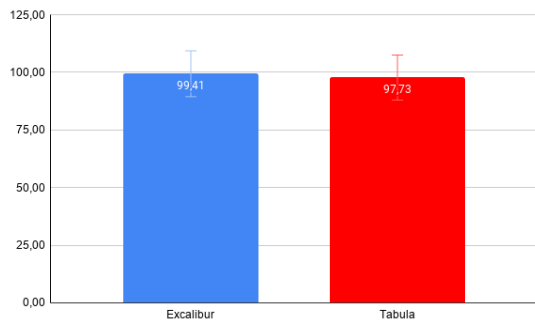


Figure 3: General comparison of Excalibur and Tabula extraction, the higher the result, the more efficient the tool is.

as the tools do not identify that the page contains 2C, so the extraction ends up being performed as a single column (1C). This is also true for mixed column (MC) metrics.

Table 6: Overview of the results of the textual data extraction tools.

|     | CyberPDF |      |      | PDFMiner |      |      |
|-----|----------|------|------|----------|------|------|
|     | 1C       | 2C   | MC   | 1C       | 2C   | MC   |
| 1Q  | 74,1     | 15,4 | -    | 76,7     | 39,5 | -    |
| 1QV | 68,1     | 0    | -    | 88,8     | 62,5 | -    |
| VQP | 80,4     | 18,6 | 41,9 | 74,6     | 55,5 | 33,6 |
| QFG | N        | N    | N    | N        | N    | N    |
| QT  | N        | N    | N    | N        | N    | N    |

The Table 7 show the results of the extractions performed by Aletheia and ExamClipper, which identifies regions and extracts the clippings from the PDFs.

It was decided to follow the extraction with the Aletheia tool. The disadvantage of using the same XML file as the one configured in the ground truth for the extractions, is that the comparison with the ExamClipper is a result that does not match the way used in the other extractions, since there were manual adjustments so the XML was organized some corrections in the segmentations, for example, joining lines of questions in the same region. This ended up favoring 100% extractions for all the metrics evaluated. The advantage is that all questions have been extracted and are available on GitHub: <https://github.com/karinawie/PDFExtraction/tree/master/extractions/aletheia>. Obviously this is not a fair comparison, although the output of the two tools and the type of extractions are similar and fit the same extraction category.

The values of the ExamClipper were not obtained with the manual selection on each region, but with the detection that the tool makes available in the cutout interface.

Table 7: Overview of the extractions performed by the ExamClipper.

|     | Aletheia |     |     | ExamClipper |      |      |
|-----|----------|-----|-----|-------------|------|------|
|     | 1C       | 2C  | MC  | 1C          | 2C   | MC   |
| 1Q  | 100      | 100 | -   | 85,8        | 61,1 | -    |
| 1QV | 100      | 100 | -   | 70,2        | 55,0 | -    |
| VQP | 100      | 100 | 100 | 64,7        | 48,2 | 36,9 |
| QFG | 100      | 100 | 100 | 68,3        | 37,1 | 56,6 |
| QT  | 100      | 100 | -   | 70,8        | 48,9 | -    |

## 5 CONCLUSIONS

This article presented a performance of the PDF extraction tools: Excalibur, Tabula, CyberPDF, PDFMiner, Aletheia and ExamClipper. We ran Excalibur and Tabula on the 396 answers, CyberPDF, PDFMiner, Aletheia and ExamClipper on the 343 tests.

According to the settings used in the tools for this work, it was possible to evaluate the extraction tools. Based on the extracted data, the Excalibur tool recognizes more tables compared to the uses of Tabula, however it takes a few more seconds for the extraction. PDFMiner is able to automatically identify multiple questions in all of the stipulated metrics, while CyberPDF cannot automatically identify questions that start on one page/column and end on another and that are in two columns. The PDFMiner tool also extracts more quickly. Although the extractions the Aletheia use a bias and the results are all at 100%, it was possible to obtain all the extractions of the questions used in this research. ExamClipper offers the option to manually adjust regions for cutouts, taking longer. If this had been applied, the extractions would also have been 100%.

These results can change within certain limits, for example, manually adjusting some identifications that the tools select, changing input settings, among others. The results of the extractions were valued in the automatic identifications that the tools allow without manual interference, except with Aletheia. If the tests used a standard layout for all courses in all years, the extractions would be more efficient, at least using the CyberPDF tool where it uses the coordinates as a standard for the other files.

As a suggestion for future work, it is intended to carry out experiments with other extraction tools not covered in the study. This extracted information is very valuable knowledge assets for research, providing useful, informative and timely information for several users who may benefit, and it can serve as research material for students, for example, who intend

to study for other tests, using as an interesting object of study to learn and retain new knowledge. In addition to becoming a set of interesting material to be used in the classroom by teachers, in order to facilitate understanding, as well as use these questions for exercises. In the case of a teacher, he may have a database of questions and answers from which he can generate new tests. Another important aspect is the possibility of creating a database of questions with these questions extracted. The objective of this work was not to make these extracted questions available in databases or systems, but it can be a suggestion for future works.

## REFERENCES

- Alaei, A., Nagabhusan, P., and Pal, U. (2011). A benchmark kannada handwritten document dataset and its segmentation. In *2011 International Conference on Document Analysis and Recognition*, pages 141–145.
- Bast, H. and Korzen, C. (2017). A Benchmark and Evaluation for Text Extraction from PDF. In *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 1–10.
- Budhiraja, S. S. (2018). *Extracting Specific Text From Documents Using Machine Learning Algorithms*. Thesis of computer science, Lakehead University, Canada.
- Choudhury, S. R., Mitra, P., Kirk, A., Szep, S., Pellegrino, D., Jones, S., and Giles, C. L. (2013). Figure metadata extraction from digital documents. In *2013 12th International Conference on Document Analysis and Recognition*, pages 135–139.
- Constantin, A., Pettifer, S., and Voronkov, A. (2013). PDFX: Fully-Automated PDF-to-XML Conversion of Scientific Literature. In *Proceedings of the 2013 ACM Symposium on Document Engineering, DocEng '13*, page 177–180, New York, NY, USA. Association for Computing Machinery.
- Duretec, K., Rauber, A., and Becker, C. (2017). A text extraction software benchmark based on a synthesized dataset. In *Proceedings of the 17th ACM/IEEE Joint Conference on Digital Libraries, JCDL '17*, page 109–118. IEEE Press.
- Excalibur (2018). Excalibur: Pdf table extraction for humans. Accessed: 2020-11-29.
- Fang, J., Tao, X., Tang, Z., Qiu, R., and Liu, Y. (2012). Dataset, ground-truth and performance metrics for table detection evaluation. In *2012 10th IAPR International Workshop on Document Analysis Systems*, pages 445–449.
- Fang Yuan and Bo Lu (2005). A new method of information extraction from PDF files. In *2005 International Conference on Machine Learning and Cybernetics*, volume 3, pages 1738–1742 Vol. 3.
- gov.br (2021). Exame Nacional de Desempenho dos Estudantes (Enade). Accessed: 2020-01-16.
- Hadjar, K., Rigamonti, M., Lalanne, D., and Ingold, R. (2004). Xed: a new tool for extracting hidden structures from electronic documents. In *First International Workshop on Document Image Analysis for Libraries, 2004. Proceedings.*, pages 212–224.
- Hassan, T. and Baumgartner, R. (2007). Table recognition and understanding from pdf files. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, volume 2, pages 1143–1147.
- INEP (2020). Exame Nacional de Desempenho dos Estudantes (Enade). Accessed: 2020-10-07.
- Li, P., Jiang, X., and Shatkay, H. (2018). Extracting figures and captions from scientific publications. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM '18*, page 1595–1598, New York, NY, USA. Association for Computing Machinery.
- Lima, R. and Cruz, E. F. (2019). Extraction and multi-dimensional analysis of data from unstructured data sources: A case study. In *ICEIS*.
- Lipinski, M., Yao, K., Breiting, C., Beel, J., and Gipp, B. (2013). Evaluation of header metadata extraction approaches and tools for scientific pdf documents. *JCDL '13*, page 385–386, New York, NY, USA. Association for Computing Machinery.
- Liu, Y., Bai, K., Mitra, P., and Giles, C. L. (2007). Table-seer: Automatic table metadata extraction and searching in digital libraries. In *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL '07*, page 91–100, New York, NY, USA. Association for Computing Machinery.
- Manuel Aristarán, Mike Tigas, Jeremy B. Merrill, Jason Das, David Frackman and Travis Swicegood (2018). Tabula is a tool for liberating data tables locked inside pdf files. Accessed: 2020-07-20.
- Parizi, R. M., Guo, L., Bian, Y., Azmoodeh, A., Dehghantanha, A., and Choo, K. R. (2018). Cyberpdf: Smart and secure coordinate-based automated health pdf data batch extraction. In *2018 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*, pages 106–111.
- Ramakrishnan, C., Patnia, A., Hovy, E., and Burns, G. A. (2012). Layout-aware text extraction from full-text PDF of scientific articles. *Source Code for Biology and Medicine*, 7(1):7.
- Strecker, T., v. Beusekom, J., Albayrak, S., and Breuel, T. M. (2009). Automated ground truth data generation for newspaper document images. In *2009 10th International Conference on Document Analysis and Recognition*, pages 1275–1279.
- Yusuke Shinyama (2014). Python pdf parser and analyzer. Accessed: 2020-05-21.
- Øyvind Raddum Berg (2011). *High precision text extraction from PDF documents*. Thesis en informatics, University of Oslo.