UNIVERSIDADE FEDERAL DE SANTA MARIA
CENTRO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Karina Wiechork

EXTRAÇÃO AUTOMATIZADA DE DADOS DE DOCUMENTOS EM FORMATO PDF: APLICAÇÃO A GRANDES CONJUNTOS DE EXAMES EDUCACIONAIS

Santa Maria, RS
2021

Karina Wiechork

EXTRAÇÃO AUTOMATIZADA DE DADOS DE DOCUMENTOS EM FORMATO PDF: APLICAÇÃO A GRANDES CONJUNTOS DE EXAMES EDUCACIONAIS

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação (PPGCC), da Universidade Federal de Santa Maria (UFSM, RS), como requisito parcial para obtenção do grau de **Mestra em Ciência da Computação**.

ORIENTADORA: Prof.ª Andrea Schwertner Charão

Santa Maria, RS
2021

Wiechork, Karina
Extração Automatizada de Dados de Documentos em Formato PDF: Aplicação a Grandes Conjuntos de Exames Educacionais / Karina Wiechork.- 2021.
73 p.; 30 cm

Orientadora: Andrea Schwertner Charão
Dissertação (mestrado) - Universidade Federal de Santa Maria, Centro de Tecnologia, Programa de Pós-Graduação em Ciência da Computação , RS, 2021

1. PDF 2. Extração Automatizada 3. Avaliação 4. Exames Educacionais 5. Ground Truth I. Schwertner Charão, Andrea II. Título.

Sistema de geração automática de ficha catalográfica da UFSM. Dados fornecidos pelo autor(a). Sob supervisão da Direção da Divisão de Processos Técnicos da Biblioteca Central. Bibliotecária responsável Paula Schoenfeldt Patta CRB 10/1728.

Declaro, KARINA WIECHORK, para os devidos fins e sob as penas da lei, que a pesquisa constante neste trabalho de conclusão de curso (Dissertação) foi por mim elaborada e que as informações necessárias objeto de consulta em literatura e outras fontes estão devidamente referenciadas. Declaro, ainda, que este trabalho ou parte dele não foi apresentado anteriormente para obtenção de qualquer outro grau acadêmico, estando ciente de que a inveracidade da presente declaração poderá resultar na anulação da titulação pela Universidade, entre outras consequências legais.

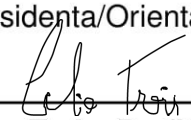
Karina Wiechork

EXTRAÇÃO AUTOMATIZADA DE DADOS DE DOCUMENTOS EM FORMATO PDF: APLICAÇÃO A GRANDES CONJUNTOS DE EXAMES EDUCACIONAIS

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação (PPGCC), da Universidade Federal de Santa Maria (UFSM, RS), como requisito parcial para obtenção do grau de **Mestra em Ciência da Computação**.

Aprovado em 16 de abril de 2021:

Andrea Schwertner Charão, Dra. (UFSM)
(Presidenta/Orientadora)



Celio Trois, Dr. (UFSM)

Marcos Didonet Del Fabro, Dr. (UFPR)

Santa Maria, RS
2021

DEDICATÓRIA

À minha família, amigos, professores, colegas e para quem este trabalho possa vir a ser proveitoso.

AGRADECIMENTOS

A conclusão deste trabalho encerra mais uma etapa muito importante na minha carreira profissional. Durante essa caminhada, diversas pessoas me apoiaram e me incentivaram.

Agradeço aos meus pais, irmãos e familiares por todo suporte e incentivo. Ao meu companheiro, pela parceria e por me motivar nos momentos difíceis.

Aos meus amigos que estiveram sempre incentivando e acreditando nesse sonho. A todos e a todas que de forma direta ou indireta auxiliaram e incentivaram nessa etapa da minha vida.

Ao IFFar pela liberação e incentivo à minha capacitação e aos colegas que seguraram as pontas enquanto estive ausente.

À UFSM e aos professores do PPGCC, pelo ensino de qualidade, aos servidores sempre solícitos. Aos meus colegas do mestrado pelas trocas de experiências.

Gostaria de agradecer principalmente à minha orientadora pela confiança, pelos ensinamentos, conselhos e por me incentivar em todos os momentos. Nunca esquecerei que sempre me ajudou e me apoiou. Muito obrigada.

"Dreams do come true, but not without the help of others, a good education, a strong work ethic, and the courage to lean in".

(Ursula Burns)

RESUMO

EXTRAÇÃO AUTOMATIZADA DE DADOS DE DOCUMENTOS EM FORMATO PDF: APLICAÇÃO A GRANDES CONJUNTOS DE EXAMES EDUCACIONAIS

AUTORA: Karina Wiechork
ORIENTADORA: Andrea Schwertner Charão

A produção massiva de documentos em formato PDF tem motivado pesquisas sobre extração automatizada de dados contidos nesses arquivos. Muitos exames educacionais utilizam provas disponibilizadas em formato PDF, que servem como material de estudo e pesquisa. Segmentar, identificar e extrair automaticamente o conteúdo de uma prova em PDF representa um desafio, pois o *layout* deste tipo de documento pode apresentar muitas variações. Pesquisas nas áreas de análise e reconhecimento de documentos, visão computacional e recuperação de informação têm produzido algoritmos e ferramentas que podem ser aplicados a esta tarefa, mas determinar sua eficácia para um dado conjunto de documentos não é uma tarefa trivial. Este trabalho propõe uma abordagem em avaliar ferramentas de extrações de dados em PDF nativamente digitais, disponibilizados em repositórios de exames educacionais. Para isso, foram utilizados os exames educacionais aplicados no Enade, entre os anos de 2004 até 2019. Os arquivos utilizados para a avaliação compreendem 343 provas, com 11.196 questões objetivas e discursivas, além de todos os 396 gabaritos, com 14.475 alternativas extraídas das questões objetivas. Para a construção de *ground truth* nas provas utilizou-se a ferramenta Aletheia, cuja finalidade é definir as regiões de interesse em cada questão. Para as extrações, utilizou-se ferramentas existentes que realizam extrações de dados em arquivos PDF, definidas para três categorias: extrações de dados tabulares, extrações de conteúdo textual e extrações de regiões de interesse. Os resultados das extrações apontam algumas limitações em relação a diversidade de *layout* em cada ano de aplicação da prova do Enade, a dificuldade em identificar e extrair questões quando dispostas em duas colunas na mesma página ou em colunas múltiplas. Os dados extraídos fornecem informações úteis, podendo auxiliar estudantes que pretendem estudar para outras provas, professores no intuito de utilizar essas questões para exercícios em sala de aula, além de coordenadores de cursos auxiliando a mapear dificuldades dos alunos a partir de questões em relatórios.

Palavras-chave: PDF. Extração Automatizada. Avaliação. Exames Educacionais. *Ground Truth*.

ABSTRACT

AUTOMATED DATA EXTRACTION FROM PDF DOCUMENTS: APPLICATION TO LARGE SETS OF EDUCATIONAL TESTS

AUTHOR: Karina Wiechork
ADVISOR: Andrea Schwertner Charão

The massive production of documents in PDF has motivated research on automated extraction of data contained in these files. Many educational tests use tests available in PDF format, which serve as study and research material. Segmenting, identifying and automatically extracting the content of a test in PDF represents a challenge, as the layout of this type of document can have many variations. Research in the areas of document analysis and recognition, computer vision and information retrieval have produced algorithms and tools that can be applied to this task, but determining their effectiveness for a given set of documents is not a trivial task. This work proposes an approach to evaluate native digital PDF data extraction tools, available in large educational test repositories. For this, the educational tests applied at Enade were used, between the years 2004 to 2019. The files used for the evaluation comprise 343 tests, with 11.196 objective and discursive questions, in addition to all 396 answers, with 14.475 alternatives extracted from the questions objectives. For the construction of ground truth in the tests, the Aletheia tool was used, whose purpose is to define the regions of interest in each question. For the extractions, existing tools were used that perform data extractions in PDF files, defined for three categories: extractions of tabular data, extractions of textual content and extractions of regions of interest. The results of the extractions point out some limitations in relation to the diversity of layout in each year of application of the Enade test, the difficulty in identifying and extracting questions when arranged in two columns on the same page or in multiple columns. The extracted data provide useful information, which can assist students who intend to study for other tests, teachers in order to use these questions for classroom exercises, as well as course coordinators helping to map students' difficulties from questions in reports.

Keywords: PDF. Automated Extraction. Evaluation. Educational Tests. Ground Truth.

LISTA DE FIGURAS

LISTA DE GRÁFICOS

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

$BPMN$	Business Process Model and Notation
$CBIES$	Sistema de Extração de Informação Baseado em Coordenadas
CID	Caractere Identificador
DAS	International Workshop on Document Analysis Systems
$Enade$	Exame Nacional de Desempenho dos Estudantes
$ICDAR$	International Conference on Document Analysis and Recognition
$IJDLS$	International Journal of Digital Library Systems
IE	Extração de Informações
$INEP$	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
$JCDL$	Joint Conference on Digital Libraries
JPG	Joint Photographic Experts Group
OCR	Optical Character Recognition
PDF	Portable Document Format
PPC	Projetos Pedagógico do Curso
ROI	Região de Interesse
$Sinaes$	Sistema Nacional de Avaliação da Educação Superior
XML	Extensible Markup Language

SUMÁRIO

1 INTRODUÇÃO

Com o desenvolvimento da tecnologia da informação e o amplo uso da Internet, um grande número de documentos eletrônicos é armazenado em arquivos Portable Document Format (PDF) (YUAN; LIU; YU, 2005). Desse modo, a demanda por análises baseadas em texto e a extração de informações significativas e específicas de documentos digitais está aumentando. Isso porque de acordo com (Clausner; Antonacopoulos; Pletschacher, 2017), tem havido um interesse crescente na análise de documentos nos últimos anos.

O formato PDF é comumente usado para a troca de documentos na web e existe uma necessidade progressiva de entender e extrair ou redefinir os dados contidos nesse tipo de arquivo (HASSAN, 2009). Extrair informações de um documento PDF é uma tarefa importante. As vantagens desses documentos, em comparação aos físicos (armazenados em papel) são numerosas: eles não sofrem rasuras; há redução de custos; o armazenamento é fácil; e há uma otimização em pesquisa, consulta e compartilhamento.

Analisar o *layout* de um documento é uma parte essencial da extração de texto, pois garante a consistência e o fluxo textual (BUDHIRAJA, 2018). Embora a maior parte da pesquisa atual de análise de *layout* tenha sido aplicada a páginas de documentos baseados em imagens, existe uma atenção cada vez maior em documentos de *layout* fixo digital nativo, normalmente em PDF (Tao et al., 2014).

Em particular, analisar uma lista de arquivos PDF para extrair dados específicos e migrá-los para outros destinos para uso posterior é tedioso e frustrante de se fazer manualmente (Parizi et al., 2018). Porém, atualmente, no mercado, há diversas ferramentas, métodos e algoritmos que extraem automaticamente textos específicos de documentos digitais. Essas ferramentas podem economizar esforço humano e também otimizar o tempo na exploração dos arquivos. Sendo assim, à medida que a quantidade de informações armazenadas usando documentos PDF aumenta, a análise baseada em texto requer processos cada vez mais automatizados, pois o processamento de documentos é demorado e exige trabalho repetitivo para quem irá fazê-lo (BUDHIRAJA, 2018).

Um exemplo de utilização desse tipo de ferramenta consiste nas provas aplicadas em exames educacionais, que normalmente são publicadas online em formato de PDF, sendo que cada site fornece arquivos com seu próprio *layout*. Um exame educacional em formato digital desempenha um papel importante, fornece informações úteis e oportunas para diferentes atores: estudantes, professores, administradores e pesquisadores. No entanto, essas provas podem abranger uma grande quantidade de questões com diferentes categorias, como tabelas, gráficos ou imagens. Dessa forma, a extração de informações de uma prova em PDF trata-se de uma tarefa difícil já que o *layout* não é geometricamente simples.

Em um trabalho anterior (CHARãO et al., 2020), foram explorados os dados ex-

traídos de arquivos PDFs do Exame Nacional de Desempenho dos Estudantes (Enade), visando reunir contribuições para revisão do Projeto Pedagógico de um Curso (PPC) da Universidade Federal de Santa Maria (UFSM). Para isso, processou-se a totalidade das provas aplicadas ao curso de Bacharelado em Ciência de Computação (cinco provas aplicadas trienalmente, de 2005 a 2017). Para extração dos dados, utilizaram-se as ferramentas Tabula, que extrai dados de tabelas em PDFs, e também o ExamClipper que é um software em desenvolvimento na UFSM, o qual identifica regiões no arquivo PDF e extrai recortes de questões. O estudo colocou em evidência a utilidade dos dados, mas também a dificuldade de realizar um trabalho semelhante em maior escala.

A presente dissertação de mestrado se preocupa em tentar localizar questões de provas contidas em documentos PDFs natos digitais, para extrair essas informações. Uma das principais motivações desta pesquisa é justamente a extração das informações, que têm potencial de serem muito úteis em uma ampla variedade de campos, desde estudos acadêmicos ao suporte para uma gama de estudantes e professores. Todavia, o desafio da investigação está relacionado à dificuldade de extração de dados em arquivos PDFs. Os *layouts* da página das provas no conjunto de dados proposto são diversos e essas provas podem abranger um grande número de questões com diferentes categorias, como tabelas, gráficos ou imagens, além de possuir questões em coluna única e outras em colunas múltiplas. A partir dessas constatações, apresenta-se, na sequência, os objetivos e justificativas desta pesquisa.

1.1 OBJETIVOS E JUSTIFICATIVAS

Esta dissertação tem como objetivo principal realizar uma análise exploratória sobre métodos e ferramentas utilizadas em extrações de dados em documentos PDF nativamente digitais, com a finalidade de descobrir sua eficácia e limitações presentes, aplicado em um grande conjunto de exames educacionais. Esta pesquisa não tem o propósito de trabalhar com abordagens em extrações de PDFs digitalizados.

As informações extraídas são ativos de conhecimento valiosos para a pesquisa, concedendo informações úteis para diversos usuários que poderão se beneficiar. O principal objetivo é extrair de maneira automatizada, questões de exames educacionais de arquivos PDFs nato digitais. Este trabalho é motivado pelo objetivo de identificar e extrair questões, podendo servir como material de pesquisa a estudantes, por exemplo, que pretendem estudar para exames/provas, como o Exame Nacional do Ensino Médio (ENEM), cursos preparatórios para ingresso em universidades ou concursos públicos, no intuito de aprender e reter novos conhecimentos.

As próprias questões fornecem recursos que os alunos podem usar para treinar e praticar. As informações também poderão auxiliar coordenadores de cursos a analisarem

a eficácia dos PPC, mapeando o conhecimento dos alunos e descobrindo lacunas a partir dos resultados das questões em relatórios. Além disso, pode se tornar um conjunto de material interessante para ser utilizado em sala de aula pelos professores, com vistas a facilitar o entendimento dos estudantes, bem como disponibilizar essas questões para exercícios. Os professores poderão ainda, ter um banco de dados de perguntas e respostas, a partir das quais eles possam gerar novas provas.

Como exemplo, cita-se o trabalho de (DENNY et al., 2008), em que os autores implementaram um sistema, PeerWise[1], o qual os alunos criam perguntas de múltipla escolha e respondem as questões de seus colegas. No artigo, os pesquisadores relatam alguns resultados quantitativos a respeito dos estudantes que usam o PeerWise, comprovando que existe um desempenho melhor nos exames finais em comparação aos estudantes que não o utilizam. Mencionam, ainda, uma correlação significativa entre o desempenho em perguntas escritas (e não apenas de múltipla escolha) e a atividade do PeerWise, sugerindo que o uso ativo do sistema pode contribuir para um aprendizado profundo.

O escopo deste trabalho não visa disponibilizar as questões extraídas em base de dados para estudos ou sistemas, como o PeerWise, mas essa pode ser uma sugestão de trabalho futuro. A referida pesquisa foi citada no intuito de trazer à tona as questões nesse sistema, sendo que possivelmente algumas tenham sido extraídas de exames educacionais em PDF, uma das vantagens em realizar extrações de informações desses arquivos.

1.2 QUESTÃO DE PESQUISA

O presente trabalho concentra-se em responder a seguinte questão de pesquisa:

1. Como analisar o desempenho de ferramentas para a extração de dados de provas em PDF, aplicando em um conjunto de provas?

Para responder essa questão, foram pesquisados trabalhos em anais de conferências e periódicos na área de análise e reconhecimento de documentos.

1.3 ESTRUTURA DO TRABALHO

Os capítulos a seguir abordarão os seguintes aspectos da pesquisa mencionados nesta introdução:

- O Capítulo 2 é destinado a fundamentação teórica de alguns conceitos básicos importantes, bem como os principais elementos utilizados nesta pesquisa. Esse capí-

[1] Disponível em: https://peerwise.cs.auckland.ac.nz/

tulo apresenta a base teórica dos temas abordados na dissertação e que serviram para subsidiar a pesquisa.

- O Capítulo 3 descreve os trabalhos relacionados a esta pesquisa, apresentando abordagens de extrações em arquivos PDFs nato digitais e em documentos PDF digitalizados.
- O Capítulo 4 explica o conjunto de dados e o processo proposto para esta pesquisa. Esse capítulo também apresenta informações de como foi construído o *ground truth* em cada página.
- O Capítulo 5 apresenta as ferramentas avaliadas e uma descrição concisa abordando suas funcionalidades, os métodos utilizados para essas avaliações, juntamente com os experimentos realizados. O capítulo também descreve o relato de erros encontrados em algumas extrações e apresenta os resultados experimentais.
- Por fim, o Capítulo 6 descreve as considerações finais com base em avaliações de desempenho, indica possibilidades de trabalhos futuros e apresenta as publicações obtidas.

2 FUNDAMENTAÇÃO

Apresenta-se brevemente, neste capítulo, a base teórica dos temas abordados nesta dissertação, bem como alguns conceitos considerados importantes para subsidiar este trabalho e que serão úteis para os leitores menos familiarizados com essa área de pesquisa.

2.1 EXAME NACIONAL DE DESEMPENHO DE ESTUDANTES - ENADE

O Enade avalia o rendimento dos concluintes dos cursos de graduação em relação aos conteúdos programáticos previstos nas diretrizes curriculares dos cursos, o desenvolvimento de competências e habilidades necessárias ao aprofundamento da formação geral e profissional, e o nível de atualização dos estudantes com relação à realidade brasileira e mundial (INEP, 2020).

Aplicado pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) desde 2004, o Enade integra o Sistema Nacional de Avaliação da Educação Superior (Sinaes), composto também pela Avaliação de cursos de graduação e pela Avaliação Institucional. Juntos, eles formam o tripé avaliativo que permite conhecer a qualidade dos cursos e instituições de educação superior brasileiras. Os resultados do Enade, aliados às respostas do Questionário do Estudante, são insumos para o cálculo dos Indicadores de Qualidade da Educação Superior (gov.br, 2021b).

Os resultados das provas do Enade compõem o cálculo de diversos indicadores, entre eles o conceito do curso, que varia de zero a cinco. Com base nas análises dos dados obtidos pela aplicação do Enade, é possível analisar o desempenho tanto das instituições quanto dos estudantes, e então calcular indicadores de qualidade que poderão oportunizar decisões de melhorias no processo de ensino.

A prova do Enade é composta de quarenta questões, sendo dez questões da parte de formação geral e trinta da parte de formação específica da área. Ambas as partes contém questões discursivas e de múltipla escolha (gov.br, 2021a), conforme distribuição apresentada na Tabela 2.1.

Tabela 2.1 – Valores e pesos para cada parte da prova.

	Formação Geral	Formação Específica
Discursiva	2	3
Múltipla Escolha	8	27
Peso	25%	75%

Fonte: Autoria própria.

A Tabela 2.2 mostra o quantitativo de todos os documentos PDFs de provas por ano, aplicadas entre 2004 a 2019. Algumas dessas provas possuem especificidades. As provas de engenharias, por exemplo, disponibilizam mais de 40 questões específicas relativas ao componente específico profissionalizante dos cursos. No caso de Engenharia Elétrica os componentes são: Controle e Automação, Telecomunicações, Eletrônica e Sistemas de Energia Elétrica. Em outras provas identificou-se que as questões dos cursos de licenciatura e bacharelado estão contidas no mesmo arquivo de prova, contabilizando em mais de 40 questões. Normalmente as provas com mais questões no mesmo arquivo são as de Engenharias, excepcionalmente a prova de Comunicação Social aplicada no ano de 2009 possui 105 questões, unindo cursos de Cinema, Editoração, Jornalismo, Publicidade e Propaganda, Radialismo e Relações Públicas.

Tabela 2.2 – Número total de provas do Enade separadas por ano.

Ano	Quantidade de Provas
2004	13
2005	20
2006	15
2007	15
2008	30
2009	22
2010	19
2011	33
2012	17
2013	17
2014	41
2015	26
2016	18
2017	44
2018	27
2019	29
Total	386

Fonte: Autoria própria.

Já a aplicação do Enade de 2020 foi adiada pelo Inep, devido as restrições impostas pela pandemia de Covid-19 e o impacto no cronograma de aulas das instituições de ensino superior em todo o país. A seguinte seção descreve, de maneira sucinta, o formato de documento PDF, o qual é utilizado nas extrações de questões das provas do Enade para esta pesquisa.

2.2 *PORTABLE DOCUMENT FORMAT* - PDF

O PDF é um dos formatos de documento mais utilizados para armazenar dados baseados em texto. Esse formato de arquivo foi projetado pela empresa Adobe em 1993 com a finalidade de representar um documento, independente da plataforma subjacente, e preservar os *layouts* na tela e na impressão. Apesar de ser uma maneira eficiente para salvar a representação visual de um documento, o trabalho com a extração de partes específicas do texto, de maneira estruturada, torna-se um quesito complicador (BUDHIRAJA, 2018).

O PDF foi desenvolvido pela Adobe até a versão 1.7. Atualmente, é um padrão aberto e oficial reconhecido pela Organização Internacional de Padronização (ISO), como ISO 32000-2: 2917 (ISO32000-2:2017, 2017).

O sucesso do PDF pode ser atribuído ao fato de ser um formato orientado para impressão, garantindo consistência de apresentação entre diferentes plataformas de computação, telas e dispositivos de impressão. No entanto, essa também é a maior desvantagem do PDF, já que a maioria dos arquivos dessa natureza contém pouca ou nenhuma informação estrutural, tornando a extração de informações uma tarefa desafiadora (Hassan; Baumgartner, 2007). Para complicar ainda mais, não há regras que regem a ordem em que o texto é codificado no documento. Por exemplo, para produzir uma página com um layout de duas colunas, a página pode ser desenhada iniciando pela primeira linha da coluna esquerda, depois a primeira à direita, posteriormente a segunda à esquerda e assim sucessivamente.

De acordo com (BERG, 2011), parte do problema enfrentado ao extrair informações de arquivos PDF é a variedade de ferramentas usadas para produzir os arquivos, além de todas as maneiras diferentes em que eles os codificam. Para ilustrar a variação em termos de ferramentas usadas para produzir documentos, o autor desenvolveu um *script* que coletava algumas estatísticas em cerca de 27 mil documentos com variações de *layout* nos arquivos, da coleção do *Norwegian Open Research Archives (NORA)* [1] que é uma coleção de pesquisa de literatura. A distribuição das ferramentas pode ser vista na Figura 2.1.

[1] Disponível em: http://www.ub.uio.no/nora/search.html

Figura 2.1 – Visão geral das ferramentas usadas para criar arquivos PDF.

Number of documents	Name of tool
1063	214 other programs with less than 50 documents
62	Windows NT
64	AFPL Ghostscript PDF Writer
64	ESP Ghostscript
65	easyPDF SDK
75	GPL Ghostscript PDF Writer
89	HP Digital Sending Device
106	AFPL Ghostscript
122	PDF PT
123	FrameMaker
150	OmniPage Pro
160	Pscript.dll
163	Acrobat Distiller
180	Writer
211	PrimoPDF
221	PDFCreator
225	Aladdin Ghostscript
228	Canon iR EUR
232	Acrobat Distiller for Windows
392	Adobe InDesign
641	GNU Ghostscript
874	dvips
2199	TeX
2538	Adobe Acrobat
3641	PScript5.dll
6750	Word
248	Broken documents
6222	Documents without metadata

Fonte: (BERG, 2011)

Destaca-se que as diferentes ferramentas utilizadas afetam a aparência tipográfica do conteúdo, a maneira e a ordem em que o conteúdo é escrito, a quantidade e o tamanho dos espaços em branco, a representação do conteúdo gráfico, a quantidade, o formato e a codificação das fontes. Tudo isso afetará e complicará frequentemente a tarefa em questão, especialmente devido à necessidade emergente de contornar informações ausentes e evitar suposições que algumas vezes se revelarão falsas (BERG, 2011).

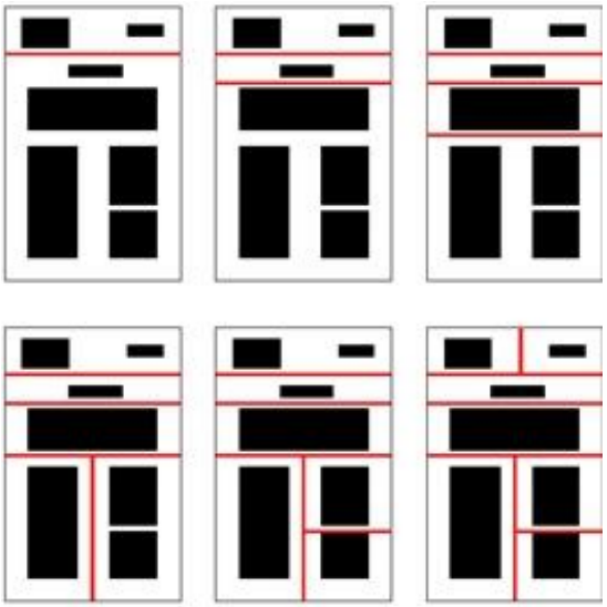
2.3 SEGMENTAÇÃO DE DOCUMENTOS

A análise do *layout* do documento ou segmentação de página é a tarefa de decompor as unidades do documento em regiões diferentes, como textos, imagens, separadores e tabelas. Ainda é um problema desafiador devido à variedade de *layouts* de documentos (TRAN; NA; KIM, 2016).

A segmentação de página identifica as regiões de interesse (ROI) no documento. A

Figura 2.2 demonstra exemplos de segmentação em regiões, ou seja, divisão de páginas do arquivo em colunas e blocos.

Figura 2.2 – Um exemplo de cortes sequenciais que segmentam uma página em diferentes regiões. A cor preta indica áreas onde o conteúdo está presente.



Fonte: (SASIREKHA; CHANDRA, 2013).

Para avaliação dos resultados de qualquer sistema de segmentação/reconhecimento, as informações de um *ground truth* desempenham um papel significativo. O *ground truth* é abordado detalhadamente na Seção 2.4.

2.4 *GROUND TRUTH* E AVALIAÇÃO DE SEGMENTAÇÃO

O termo *ground truth* pode ser traduzido como: “verdade fundamental”, “verdade básica”, “verdade do solo” e, também, “verdade do terreno”. Nesta dissertação será utilizado o termo *ground truth*, sem tradução, da mesma maneira que outros pesquisadores brasileiros também utilizam.

Ground truth deriva originalmente da área de geoprocessamento, onde as informações extraídas das imagens de satélite são confirmadas pelas pessoas que visitam o local para serem estudadas no terreno (KONDERMANN, 2013). Também pode ser utilizado para outros propósitos, tendo sempre como principal característica instruir o que é dito como verdade.

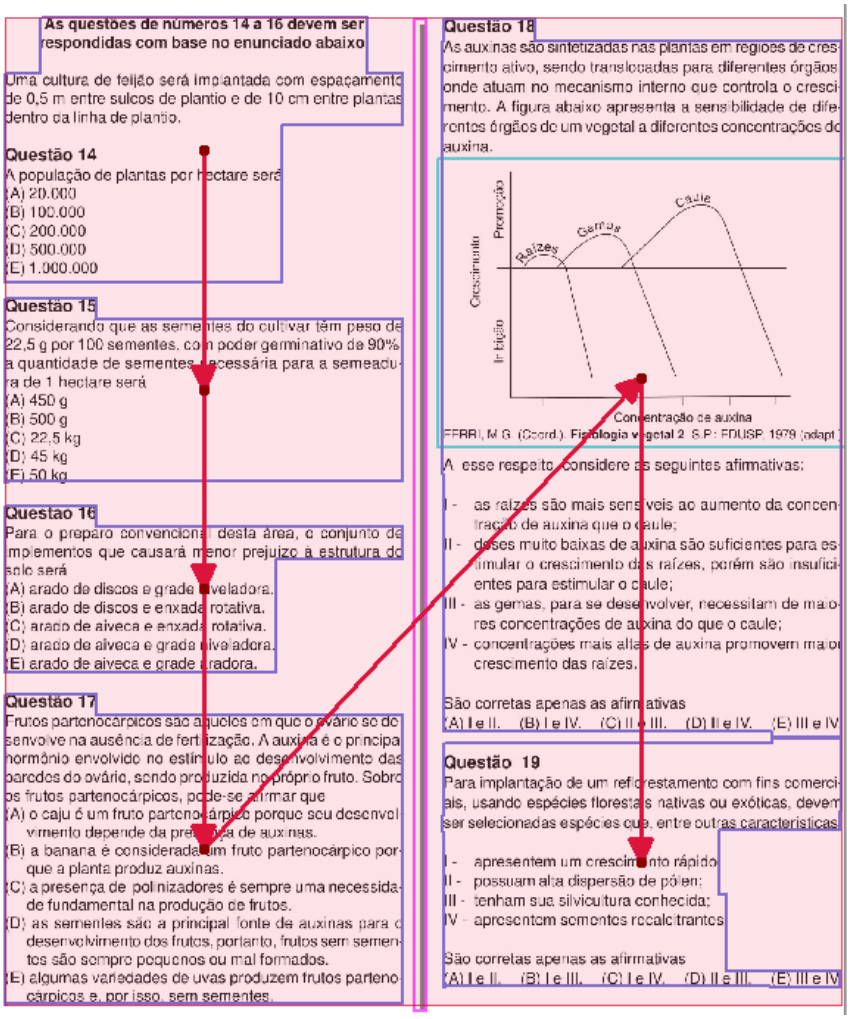
No contexto da visão computacional, os dados de *ground truth* incluem um conjunto de imagens e um conjunto de marcação nas imagens e a definição de um modelo para reconhecimento de objeto, incluindo a contagem, localização e relação dos principais recur-

sos. As marcações são adicionadas por um ser humano ou automaticamente por análise de imagem, dependendo da complexidade do problema. A coleção de marcações, como pontos de interesse, regiões, formas e histogramas, formam um modelo (KRIG, 2014).

A construção de um *ground truth* é uma tarefa demorada e sujeita a erros (Alaei; Nagabhushan; Pal, 2011). Frequentemente, o *ground truth* necessário deve ser criado manualmente, limitando o tamanho do conjunto de dados. A complexidade do *ground truth* também é um fator inibidor: quanto mais detalhe que ele deve fornecer, mais esforço deve ser feito para gerá-lo. Esse é o principal motivo pelo qual existem tão poucos conjuntos de *ground truth* para análise de *layout* (Strecker et al., 2009).

A estrutura armazenada em um arquivo de *ground truth* também contém a ordem de leitura (TKACZYK; SZOSTEK; BOLIKOWSKI, 2014). Um exemplo de ordem de leitura, pode ser visto na Figura 2.3. A ordenação entre as regiões é mostrada com setas.

Figura 2.3 – Ordem de leitura na prova do Enade.



Fonte: Autoria Própria.

No entendimento da imagem do documento, conjuntos de dados públicos com *ground truth* são uma parte importante do trabalho científico. Eles não são apenas úteis

para o desenvolvimento de novos métodos, mas também fornecem uma maneira de comparar o desempenho (Strecker et al., 2009). Os dados de referência com base no *ground truth* consistem em um conjunto de informações que incluem os resultados desejados. O *ground truth* é como uma referência para comparar e avaliar os resultados dos experimentos, pois possui dados detalhados em vários aspectos.

3 TRABALHOS RELACIONADOS

O problema da extração de metadados é amplamente estudado na literatura. Sendo assim, neste capítulo são analisados e descritos trabalhos relacionados a esta dissertação, desenvolvidos por outros pesquisadores da área. Dentre algumas opções de busca utilizadas, foram pesquisados trabalhos em anais de conferências e periódicos relevantes para a área de Análise e Reconhecimento de Documentos, como: *International Conference on Document Analysis and Recognition (ICDAR)* [1], *Joint Conference on Digital Libraries (JCDL)* [2], *International Workshop on Document Analysis Systems (DAS)* [3] e *International Journal of Digital Library Systems (IJDLS)* [4].

Inicialmente, são apresentados os estudos relacionados a abordagens que trabalham com documentos PDFs nativamente digitais. Em seguida, são demonstradas brevemente algumas pesquisas que trabalham com extrações em PDFs baseados em imagens, como saídas de digitalização.

3.1 EXTRAÇÃO EM PDF NATO DIGITAL

O documento digital nativo é aquele que já nasce digital, o qual pode ser produzido, por exemplo, pelo Microsoft Office, LibreOffice ou outros pacotes de software para escritório. Dentre algumas pesquisas que focam em extração nesses documentos, está o trabalho de (BEEL et al., 2013) que extrai títulos de arquivos PDF acadêmicos. Os pesquisadores desenvolveram a ferramenta Docear [5] e criaram uma coleção de testes com 500 arquivos para avaliar a precisão do PDF Inspector da Docear, em comparação com outras ferramentas existentes de extrações dos títulos. Já a ferramenta Cermine [6] (TKACZYK et al., 2015) é um sistema web de código aberto para extrair metadados e conteúdo de artigos científicos em formato digital nativo. O sistema é capaz de processar documentos em formato PDF e extrair metadados, incluindo título, autores, resumo, palavras-chave, nome do periódico, volume e edição, referências bibliográficas analisadas a estrutura das seções, títulos das seções e parágrafos.

Os autores (CONSTANTIN; PETTIFER; VORONKOV, 2013) apresentam a ferramenta PDFX, que é um sistema baseado em regras, projetado para reconstruir a estrutura lógica de artigos acadêmicos em formato PDF. A saída do sistema é um documento eX-

[1] Disponível em: https://icdar2021.org/
[2] Disponível em: https://www.jcdl.org/
[3] Disponível em: https://www.vlrlab.net/das2020/
[4] Disponível em: http://www.igi-global.com/IJDLS
[5] Disponível em: http://www.docear.org.
[6] Disponível em: http://cermine.ceon.pl

tensible Markup Language (XML) ou HyperText Markup Language (HTML) que descreve a estrutura lógica do artigo de entrada em termos de título, seções, tabelas, referências, etc. Também o vincula a marcadores de composição geométrica no PDF original, como quebras de parágrafo e coluna. Quando utilizada a saída para HTML, as figuras também são extraídas e ficam disponíveis no final do arquivo, porém não na ordem de leitura. Em (Parizi et al., 2018), é proposta uma técnica que permite aos usuários consultar um documento PDF representativo e extrair os mesmos dados de uma série de arquivos na forma de análise de lote rapidamente. A técnica é implementada no software CyberPDF, uma ferramenta de extração de lote de dados PDF automática baseada em coordenadas. A ferramenta PDFMiner também extrai informações de documentos PDF, mas, contrário de outras ferramentas relacionadas ao formato, ela se concentra inteiramente na obtenção e análise de dados de texto (Yusuke Shinyama, 2014).

O trabalho de (Fang Yuan; Bo Lu, 2005), apresenta outro método para extrair informações de arquivos PDF. Primeiro ele analisa os arquivos para obter informações de texto e formato, injeta tags em informações de texto para transformá-las em texto semiestruturado e, finalmente, um algoritmo de correspondência de padrão baseado no modelo de árvore é aplicado para obter a solução. Já no trabalho de (BUI et al., 2016), os autores desenvolveram um sistema extrativo de resumo de texto que visa ajudar revisores humanos no desenvolvimento de revisões sistemáticas, com objetivo de aumentar a produtividade e reduzir erros no processo tradicional de extração de dados. Outra abordagem, proposta por (JIANG; YANG, 2009) é realizar a conversão de PDF para HTML usando detecção de texto. A saída é um arquivo HTML que preserva as informações da fonte e o *layout* de todo o documento. O processo funciona detectando fragmentos de texto. A conversão é possível usando as informações de coordenadas coletadas com a biblioteca Apache PDFBox. Dentre as ferramentas utilizadas para extração de texto em PDF, todos esses trabalhos utilizaram a biblioteca PDFBox. O texto extraído, usando essa biblioteca, contém características semelhantes às de copiar e colar manualmente de um leitor de PDF.

A biblioteca Apache PDFBox [7] é uma ferramenta Java de código aberto para trabalhar com documentos PDF. Esta ferramenta permite a criação de novos documentos PDF, manipulação de documentos existentes e a capacidade de extrair conteúdo de documentos. Também inclui vários utilitários de linha de comando e está publicada sob a Licença Apache v2.0 (The Apache Software Foundation, 2020). Outra biblioteca semelhante é Apache Tika [8], que também segue na linha de extrações de dados textuais. Um exemplo de utilização de Apache Tika é Give me Text! [9], um serviço da web de código aberto, gratuito, que extrai o conteúdo textual de PDFs e outros documentos.

No campo específico da extração de tabelas contidas em arquivos PDF, há uma série de trabalhos publicados e ferramentas disponíveis no mercado (Hassan; Baumgart-

[7] Disponível em: https://pdfbox.apache.org/
[8] Disponível em: https://tika.apache.org/
[9] Disponível em: http://givemetext.okfnlabs.org/

ner, 2007). A segmentação de páginas e a detecção de tabelas desempenham um papel importante na compreensão da estrutura dos documentos (He et al., 2017). No trabalho (LIU et al., 2007) é descrito um sistema capaz de extrair tabelas e metadados de tabelas de documentos PDF, para isso é utilizado o PDFBox para extração do texto bruto, que é posteriormente processado para identificar tabelas. Na análise bibliográfica de (SILVA; JORGE; TORGO, 2006), os autores exploram abordagens existentes para extrair informações de tabelas. Nessa direção, o trabalho de (FAN; KIM, 2015) também contribui para detectar regiões das tabelas em documentos PDF, desenvolvendo PDFExtra.

Reconhecer e extrair tabelas é uma tarefa importante porque este é um formato comum de apresentar e estruturar dados com alta densidade de informações. Apesar de ser um formato estruturado, extrair dados de tabelas nem sempre é uma tarefa fácil, visto que estas podem ter formatos variados. Quando há uma variedade de *layouts*, o trabalho é mais custoso e propenso a erros. A ferramenta Tabula (Manuel Aristarán, Mike Tigas, Jeremy B. Merrill, Jason Das, David Frackman and Travis Swicegood, 2018) é um exemplo que permite aos usuários selecionar tabelas para extração de dados tabulares em documentos PDF. Outro exemplo é o trabalho dos autores (YILDIZ; KAISER; MIKSCH, 2005), que desenvolveram heurísticas que identificam e extraem tabelas em arquivos PDF e armazenam os dados extraídos em um formato de dados estruturado (XML), para facilitar a reutilização. Também nesta linha, a ferramenta Excalibur é uma aplicação web para extrair dados tabulares de PDFs baseados em texto e não com documentos digitalizados (EXCALIBUR, 2018). A extração de tabelas tornou-se útil para o presente trabalho, a fim de realizar a extração dos gabaritos das questões objetivas.

No que diz respeito ao desempenho das ferramentas, os autores (Fang et al., 2012) projetaram um conjunto de dados representativo para avaliação de detecção de tabela. Eles abordam um conjunto de métricas de desempenho, que são uma mistura de pontuações de penalidade orientada a aplicativos e cálculo quantitativo baseado em conteúdo. Além disso, avaliam dois projetos de detecção de tabela de código aberto para demonstrar a confiabilidade do conjunto de dados e a eficácia das medidas de desempenho. Essas pesquisas apresentaram ideias que foram utilizadas como modelo para esta dissertação.

Abordagens específicas com o objetivo de extrair figuras e legendas de documentos PDF vêm sendo propostas. É o caso do trabalho de (Choudhury et al., 2013), em que os autores se preocuparam em extrair figuras e legendas associadas de documentos PDF. No trabalho de (LI; JIANG; SHATKAY, 2018), é apresentado um sistema para extração de figuras e legendas associadas de publicações científicas, PDFigCapX[10]. Em (Clark; Divvala, 2016), os autores desenvolveram um algoritmo que extrai figuras, tabelas e legendas de documentos, intitulado PDFFigures [11].

Muitas aplicações de extração de dados podem exigir a combinação de diferentes

[10]Disponível em: https://www.eecis.udel.edu/∼compbio/PDFigCapX
[11]Disponível em: https://github.com/allenai/pdffigures2

estratégias e ferramentas. Por exemplo, no trabalho de (WU et al., 2015), é apresentada uma estrutura de extração de conhecimento de várias entidades para documentos acadêmicos no formato PDF. Essa ferramenta, PDFMEF, é implementada com uma estrutura que encapsula ferramentas de extração de código aberto. Atualmente, ele aproveita PDFBox e TET para extração de texto completo, GROBID para extração de cabeçalho, ParsCit para extração de citação e PDFFigures para extração de figura e tabela. A extração para cada tipo de região fica separada por pastas no computador utilizado pelo usuário.

Diante dos estudos apresentados, entende-se que várias questões desafiadoras permanecem como trabalhos futuros, como a identificação de casos nos quais pares de legenda e figura abrangem mais de uma página. Na presente dissertação, esse problema também é um desafio, visto que algumas questões do Enade estão contidas em mais de uma página. A extração de informações é de suma importância em várias aplicações do mundo real e, embora várias abordagens sofisticadas e até complexas tenham sido propostas, elas ainda são limitadas em muitos aspectos (Oro; Ruffolo, 2008).

No estudo de (LIMA; CRUZ, 2019), é proposta uma abordagem para detectar e extrair dados de fonte de dados não estruturados disponíveis online e espalhados por diversas páginas da Web para armazenar os dados em um Data Warehouse devidamente projetado para isso. Quase todos os arquivos são publicados em PDF e há arquivos com formatações diferentes. Para isso, os autores utilizaram ferramentas pré-existentes.

Já na pesquisa de (Bast; Korzen, 2017), é desenvolvida uma avaliação de 14 ferramentas de extração de PDF para determinar a qualidade e o alcance de suas funcionalidades, com base em um *benchmark* que construíram a partir de dados paralelos de TeX e PDF. Os autores utilizaram 12.098 artigos científicos e, para cada artigo, o *benchmark* contém um arquivo de *ground truth* além do PDF relacionado. Percebe-se, então, que a extração precisa de metadados é uma tarefa importante para automatizar o gerenciamento de bibliotecas digitais. Em (LIPINSKI et al., 2013) é realizada uma avaliação do desempenho de ferramentas para a extração de metadados de artigos científicos. O estudo comparativo é um guia para desenvolvedores que desejam integrar a ferramenta de extração de metadados mais adequada e eficaz em seus softwares.

Os autores (Hadjar et al., 2004) descrevem uma abordagem Xed (*eXtracting electronic documents*), na qual extrai todos os objetos de um documento PDF, incluindo texto, imagens e gráficos. A saída dos objetos extraídos está no formato SVG (*Scalable Vector Graphics*). Os resultados de análises de *layout* complexas são usados como entrada para gerar um algoritmo de classificação topológica para partes de texto. Os dados são extraídos do PDF usando a ferramenta JPedal.

3.2 EXTRAÇÃO EM PDF DIGITALIZADO

Conforme mencionado anteriormente, o documento PDF digitalizado é aquele que representa de maneira digital um documento físico. O arquivo digital pode ser criado por meio da digitalização do documento ou a partir de ferramentas externas, como scanner, câmera fotográfica, entre outros, sendo que o documento gerado é convertido para um formato digital.

O método de Reconhecimento Óptico de Caracteres (OCR) tem sido usado para converter texto impresso em texto editável e trata-se de uma alternativa muito útil e popular em várias aplicações. Sua precisão pode depender do pré-processamento de texto e dos algoritmos de segmentação (PATEL; PATEL; PATEL, 2012).

Para documentos digitalizados, foram encontradas algumas ferramentas disponíveis para extração, como a ferramenta LAREX (Análise de Layout e Extração de Região), cujo objetivo é apoiar os usuários na segmentação e classificação de regiões em imagens (REUL; SPRINGMANN; PUPPE, 2017). O foco da aplicação é em análise de livros antigos impressos, isso é, a ferramenta não foi projetada para segmentar automaticamente qualquer documento fornecido. Já ABBYY [12] é uma ferramenta proprietária que está atualmente disponível no mercado e também foi desenvolvida para digitalizar livros antigos.

No trabalho de (PATEL; PATEL; PATEL, 2012), os autores realizam uma comparação entre as ferramentas Transym [13] e Tesseract [14], sendo que ambas utilizam a detecção do OCR nos arquivos. Em (CHANDARANA, 2014), se apresenta uma visão geral dos métodos de extração de recursos para reconhecimento de caracteres. A seleção do método de extração de recursos é o único fator mais importante para alcançar um alto desempenho em sistemas de reconhecimento de caracteres. Para diferentes representações dos personagens, diferentes métodos de extração de recursos são projetados.

3.3 SÍNTESE GERAL

Os trabalhos discutidos ao longo deste capítulo ilustram que há interesse e avanços que contribuem na extração de dados de documentos em PDF, mas não foram encontrados trabalhos que abordassem a extração de questões em provas e exames disponibilizados em PDFs. Nessa dissertação, não foi desenvolvida uma ferramenta para buscar resolver esse problema, pois entendeu-se que antes disso seria preciso identificar e conhecer como as abordagens existentes se comportam em extrações de dados em arquivos PDFs.

O presente trabalho tem o diferencial de abordar informações com foco em PDF

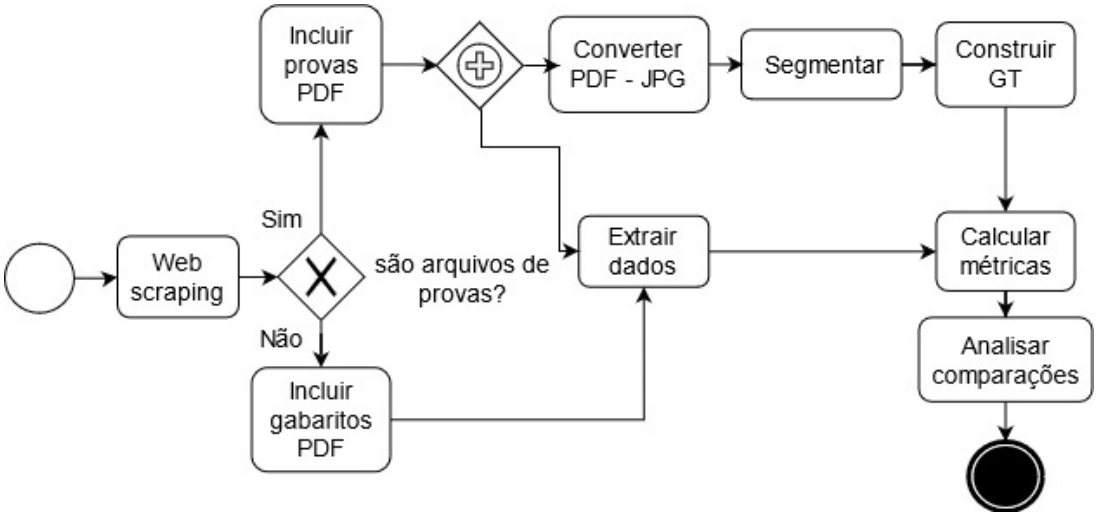
[12] Disponível em: https://www.abbyy.com/
[13] Disponível em: https://transym.com/
[14] Disponível em: https://opensource.google/projects/tesseract

nato digital e, de maneira sucinta, algumas ferramentas que trabalham com o documento PDF digitalizado.

4 PROCESSO PROPOSTO

A fim de fornecer uma medição objetiva de desempenho das ferramentas, um grande conjunto de dados é necessário em todas as tarefas de um processo de segmentação e extração. Neste capítulo será apresentado o processo proposto para a execução do roteiro das tarefas, com a finalidade de reunir todos os dados. A Figura 4.1 fornece uma visão geral do processo proposto, apresentado em linguagem BPMN (*Business Process Model and Notation*), que é amplamente utilizada para representação de processos (CRUZ; MACHADO; SANTOS, 2019). O processo é composto de nove etapas:

Figura 4.1 – Processo Proposto.



Fonte: Autoria Própria.

Web Scraping [1] Inicialmente a etapa do processo consiste em recolher os arquivos do site do INEP. Para essa tarefa, o Scrapy foi utilizado na criação dos *scripts* de recolhimento das provas e dos gabaritos [2]. Esta etapa ocorreu de maneira automatizada, informando no parâmetro do código o ano de interesse para o recolhimento dos arquivos. A Seção 4.1 aborda essa etapa com mais detalhes.

Incluir provas PDF Nessa etapa as provas em PDF são incluídas no conjunto de dados, separadas por pastas conforme o ano da aplicação da prova.

Incluir gabaritos PDF Essa etapa inclui os gabaritos em PDF no conjunto de dados, porém apenas os gabaritos das questões objetivas.

Converter PDF – JPG Na fase de construção dos *ground truths* foi necessário realizar a conversão dos arquivos das provas em PDF para o formato *Joint Photographic*

[1] Traduzido para o português como: raspagem da web. Será utilizado em inglês: *web scraping.*
[2] Disponível em: https://github.com/karinawie/scrapy

Experts Group (JPG) para a utilização com a ferramenta Aletheia, a qual será apresentada na Seção 4.1. Esses arquivos foram convertidos em JPG, de maneira automatizada, para em seguida construir o *ground truth*. O conjunto de dados pode seguir o caminho de segmentação e extração dos dados sem precisar passar na etapa da conversão.

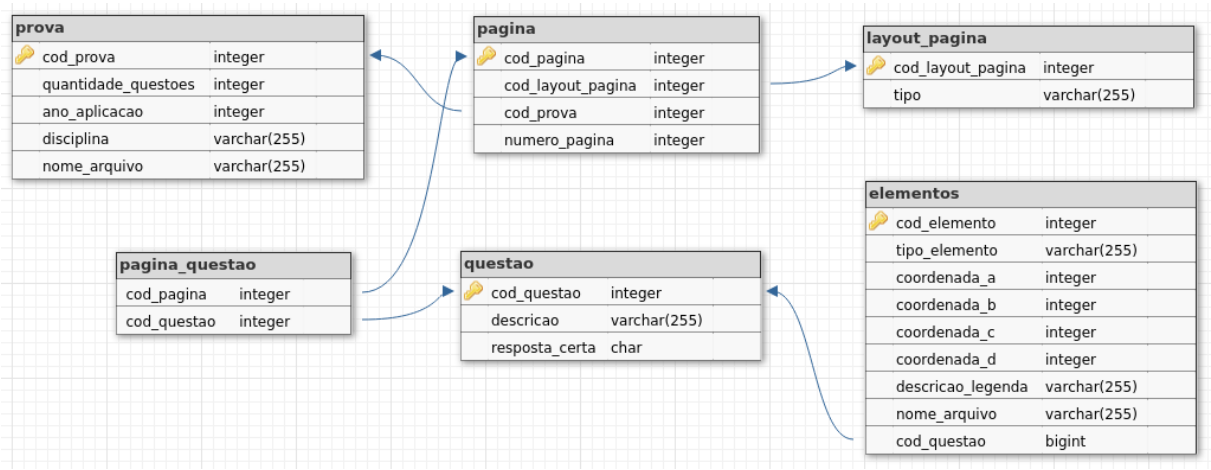
Segmentação A segmentação foi realizada em cada página da prova que continha questões, e utilizou-se a ferramenta Aletheia antes de seguir para o *ground truth*. Essa tarefa consiste em identificar as figuras/gráficos, tabelas, colunas e o conteúdo textual das questões, a segmentação dessas regiões ocorre automaticamente. Não houve necessidade de realizar a segmentação dos gabaritos, pois esses estão identificados em tabelas.

Construção de *ground truth* Nessa etapa foram criados os *ground truths* de maneira automática pelo software Aletheia, no entanto, alguns ajustes manuais precisaram ser feitos por um especialista humano, definindo essa etapa como semi-automatizada. Para cada página com questão, um arquivo de *ground truth* está disponível. A construção do *ground truth* não foi realizada para os gabaritos, visto que os dados tabulares já estão estruturados em tabelas. A Seção 4.2 apresenta de maneira mais detalhada essa tarefa.

Extração A extração dos dados é a etapa mais demorada e mais importante do processo. Antes de começar a extração foi necessário configurar as ferramentas. Nessa etapa é contabilizado o tempo que cada uma demora para realizar a extração. As ferramentas de extrações trabalharam de maneira automatizada e os experimentos são abordados na Seção 5.4.

Calcular métricas As métricas utilizadas neste trabalho foram definidas e criadas com base nas divergências encontradas nos *layouts* das questões das provas. A Figura 4.2 demonstra a modelagem de um banco de dados que, neste trabalho, foi importante pelo fato de auxiliar a identificar possíveis métricas a serem calculadas nas ferramentas de extração de dados. Essa modelagem pode contribuir para trabalhos futuros, auxiliando em desenvolvimento de algum sistema para disponibilização dessas questões extraídas.

Figura 4.2 – Modelagem de dados.



Fonte: Autoria Própria.

Diante disso, todas as extrações do conjunto de dados foram utilizadas para calcular os valores que cada ferramenta obteve em relação a quantidade de métricas que deveriam ter sido extraídas. A quantidade de métricas identificadas em cada ferramenta de extração foi informada de maneira manual em planilhas no Google Sheets[3]. Já a contagem para o *ground truth* dessas métricas, foi feita de maneira semi-automatizada, onde um *script* identifica as regiões de cada página das provas e contabiliza o total dessas métricas. Mais informações relacionadas estão disponíveis na Seção 5.2.

Analisar comparações Uma vez que os dados são recolhidos, juntados com o conjunto de dados e extraídos, eles podem ser comparados com o resultado da ferramenta concorrente de extração e, em seguida, com o *ground truth*. Essa comparação das extrações ocorreram de maneira manual, exceto com os dados textuais que ocorreram de maneira semi-automatizada. Concluindo essa etapa, é possível obter os resultados, que estão disponíveis na Seção 5.6.

A seguir, a Seção 4.1 detalha o conjunto de dados construído a partir do processamento de provas e gabaritos do Enade, para uma avaliação da análise de *layout*.

4.1 CONJUNTO DE DADOS

O conjunto de dados para esta pesquisa é composto de provas e gabaritos de avaliações do Enade, com aplicações dos anos de 2004 a 2019. Para automatizar o download

[3]Links das planilhas disponveís em: https://github.com/karinawie/PDFExtraction

dessas provas e gabaritos no site do INEP, foi aplicada a ferramenta Scrapy. Trata-se de uma biblioteca em Python que auxilia na captação de dados web de forma automatizada.

Todos os arquivos foram coletados no endereço de avaliações do Enade [4]. Desde o dia 31 de julho de 2020, o site do INEP passou a integrar o portal único do Governo Federal e pode ser acessado em gov.br/inep (gov.br, 2021c).

Visando essa possível alteração desses documentos, por parte do INEP, ou modificação dos links dos arquivos, surgiu a preocupação com a garantia da confiabilidade dos dados. Com base nisso, o conjunto obtido está disponível juntamente com os links e a data de obtenção dos mesmos [5]. As informações podem contribuir para uma possível conferência com as extrações. Essa conferência não é objeto desta pesquisa, mas contribui, em parte, para a garantia da confiabilidade da extração.

Foram recolhidos 386 provas e 396 gabaritos. Essa diferença na quantidade de provas para os gabaritos acontece porque, em alguns anos, o mesmo arquivo de prova contém questões separadas por cursos semelhantes, mas com os gabaritos em arquivos diferentes. Este é o caso, por exemplo, das provas de Comunicação Social e suas habilitações (Cinema, Editoração, Jornalismo, Publicidade e Propaganda, Radialismo e Relações Públicas).

O conjunto de dados desta pesquisa, no entanto, é composto de 343 provas e 396 gabaritos, totalizando 739 arquivos. Não foram utilizadas todas as provas devido ao grande volume e a delonga em executar as extrações, os *ground truths* e as comparações, por isso optou-se por remover provas de cursos com menos de duas aplicações. O total de provas removidas foram 43.

A quantidade de páginas para extrações e questões são 6.834 e 11.196 respectivamente, enquanto o total de alternativas objetivas nos 396 gabaritos é de 14.475. Nesta contagem não foram incluídas as páginas em branco, nem as capas das provas e tampouco as páginas de questionário de percepção da prova, que se encontram apenas nas últimas edições.

Em todas as provas, a parte de questões de formação geral foi contabilizada apenas uma vez nos casos em que o padrão de *layout* foi o mesmo para todas as provas do ano avaliado. Nos anos 2004, 2005 e 2007, identificou-se mais de um padrão de *layout*, com isso foi gerado mais de um *ground truth* para formação geral. Os gabaritos não estão contabilizados com as questões dissertativas, apenas com as alternativas objetivas. A Tabela 4.1 apresenta uma visão geral dos dados quantitativos de páginas e questões que foram utilizadas nas extrações das provas.

[4] http://portal.inep.gov.br/web/guest/educacao-superior/Enade/provas-e-gabaritos
[5] Disponível em: https://github.com/karinawie/PDFExtraction/tree/master/dataset

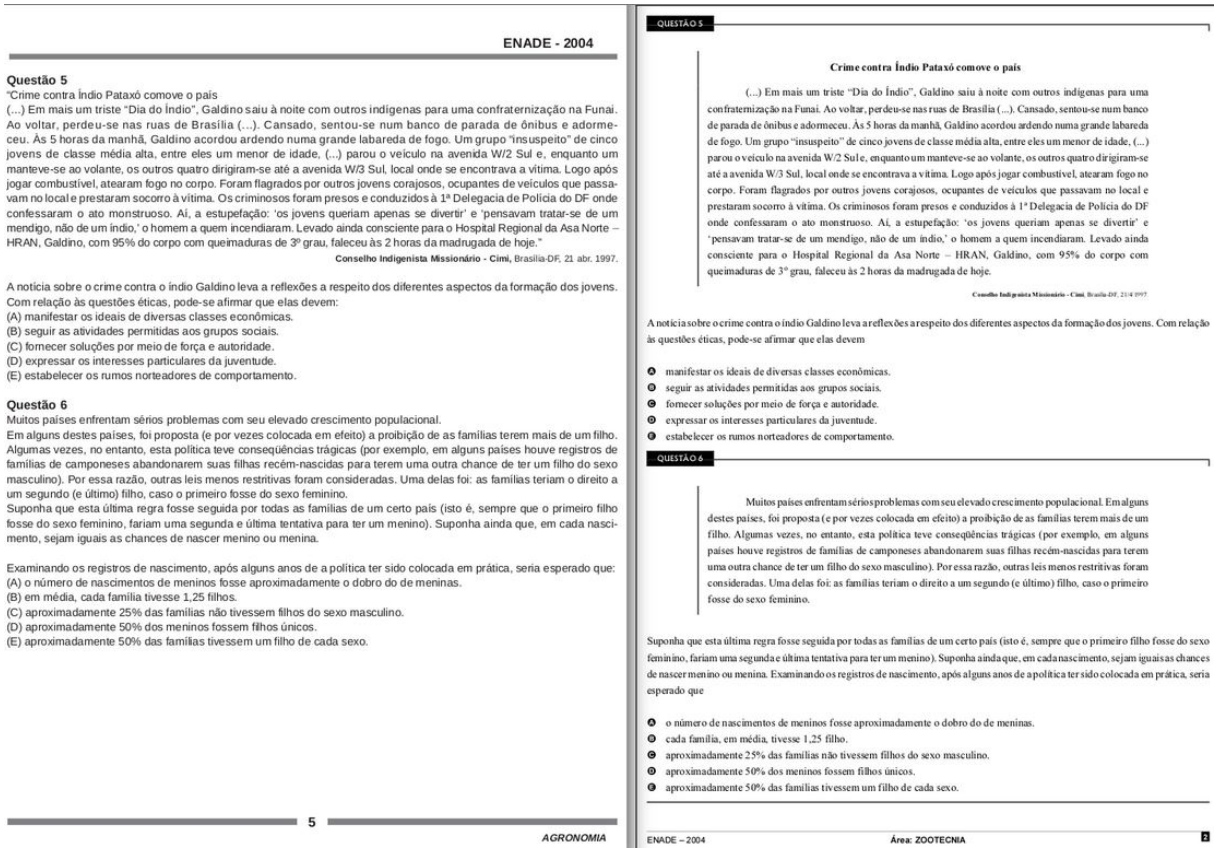
Tabela 4.1 – Visão geral dos dados utilizados nas provas.

Ano	Quant. de provas	Quant. de páginas	Quant. de questões
2004	13	151	430
2005	20	391	884
2006	11	171	340
2007	14	188	440
2008	24	453	939
2009	17	328	595
2010	18	236	550
2011	27	546	956
2012	17	292	520
2013	17	342	520
2014	40	884	1.222
2015	19	439	580
2016	17	351	520
2017	42	905	1.270
2018	20	537	610
2019	27	620	820
Total	343	6.834	11.196

Fonte: Autoria Própria

A Figura 4.3 apresenta a página com as mesmas questões, questão 5 e questão 6, pertencentes ao mesmo ano de aplicação, mas com *layout* diferente. Esse conjunto de dados contém uma ampla variedade de questões e pode ser considerado representativo dos documentos no mundo real, embora todos se tratem de arquivos PDF, de origem digital. Os PDFs analisados representam provas com diversos padrões, como colunas, figuras, tabelas, questões em mais de uma página.

Figura 4.3 – Provas de Agronomia (esquerda) e Zootecnia (direita) de 2004, mesmo ano de aplicação, porém com *layout* divergente.



ENADE - 2004

Questão 5

"Crime contra Índio Pataxó comove o país

(...) Em mais um triste "Dia do Índio", Galdino saiu à noite com outros indígenas para uma confraternização na Funai. Ao voltar, perdeu-se nas ruas de Brasília (...). Cansado, sentou-se num banco de parada de ônibus e adormeceu. Às 5 horas da manhã, Galdino acordou ardendo numa grande labareda de fogo. Um grupo "insuspeito" de cinco jovens de classe média alta, entre eles um menor de idade, (...) parou o veículo na avenida W/2 Sul e, enquanto um manteve-se ao volante, os outros quatro dirigiram-se até a avenida W/3 Sul, local onde se encontrava a vítima. Logo após jogar combustível, atearam fogo no corpo. Foram flagrados por outros jovens corajosos, ocupantes de veículos que passavam no local e prestaram socorro à vítima. Os criminosos foram presos e conduzidos à 1ª Delegacia de Polícia do DF onde confessaram o ato monstruoso. Aí, a estupefação: 'os jovens queriam apenas se divertir' e 'pensavam tratar-se de um mendigo, não de um índio,' o homem a quem incendiaram. Levado ainda consciente para o Hospital Regional da Asa Norte – HRAN, Galdino, com 95% do corpo com queimaduras de 3º grau, faleceu às 2 horas da madrugada de hoje."

Conselho Indigenista Missionário - Cimi, Brasília-DF, 21 abr. 1997.

A notícia sobre o crime contra o índio Galdino leva a reflexões a respeito dos diferentes aspectos da formação dos jovens. Com relação às questões éticas, pode-se afirmar que elas devem:
(A) manifestar os ideais de diversas classes econômicas.
(B) seguir as atividades permitidas aos grupos sociais.
(C) fornecer soluções por meio de força e autoridade.
(D) expressar os interesses particulares da juventude.
(E) estabelecer os rumos norteadores de comportamento.

Questão 6

Muitos países enfrentam sérios problemas com seu elevado crescimento populacional.
Em alguns destes países, foi proposta (e por vezes colocada em efeito) a proibição de as famílias terem mais de um filho. Algumas vezes, no entanto, esta política teve conseqüências trágicas (por exemplo, em alguns países houve registros de famílias de camponeses abandonarem suas filhas recém-nascidas para terem uma outra chance de ter um filho do sexo masculino). Por essa razão, outras leis menos restritivas foram consideradas. Uma delas foi: as famílias teriam o direito a um segundo (e último) filho, caso o primeiro fosse do sexo feminino.
Suponha que esta última regra fosse seguida por todas as famílias de um certo país (isto é, sempre que o primeiro filho fosse do sexo feminino, fariam uma segunda e última tentativa para ter um menino). Suponha ainda que, em cada nascimento, sejam iguais as chances de nascer menino ou menina.

Examinando os registros de nascimento, após alguns anos de a política ter sido colocada em prática, seria esperado que:
(A) o número de nascimentos de meninos fosse aproximadamente o dobro do de meninas.
(B) em média, cada família tivesse 1,25 filhos.
(C) aproximadamente 25% das famílias não tivessem filhos do sexo masculino.
(D) aproximadamente 50% dos meninos fossem filhos únicos.
(E) aproximadamente 50% das famílias tivessem um filho de cada sexo.

5

AGRONOMIA

QUESTÃO 5

Crime contra Índio Pataxó comove o país

(...) Em mais um triste "Dia do Índio", Galdino saiu à noite com outros indígenas para uma confraternização na Funai. Ao voltar, perdeu-se nas ruas de Brasília (...). Cansado, sentou-se num banco de parada de ônibus e adormeceu. Às 5 horas da manhã, Galdino acordou ardendo numa grande labareda de fogo. Um grupo "insuspeito" de cinco jovens de classe média alta, entre eles um menor de idade, (...) parou o veículo na avenida W/2 Sul e, enquanto um manteve-se ao volante, os outros quatro dirigiram-se até a avenida W/3 Sul, local onde se encontrava a vítima. Logo após jogar combustível, atearam fogo no corpo. Foram flagrados por outros jovens corajosos, ocupantes de veículos que passavam no local e prestaram socorro à vítima. Os criminosos foram presos e conduzidos à 1ª Delegacia de Polícia do DF onde confessaram o ato monstruoso. Aí, a estupefação: 'os jovens queriam apenas se divertir' e 'pensavam tratar-se de um mendigo, não de um índio,' o homem a quem incendiaram. Levado ainda consciente para o Hospital Regional da Asa Norte – HRAN, Galdino, com 95% do corpo com queimaduras de 3° grau, faleceu às 2 horas da madrugada de hoje.

Conselho Indigenista Missionário - Cimi, Brasília-DF, 21/4/1997.

A notícia sobre o crime contra o índio Galdino leva a reflexões a respeito dos diferentes aspectos da formação dos jovens. Com relação às questões éticas, pode-se afirmar que elas devem

A manifestar os ideais de diversas classes econômicas.
B seguir as atividades permitidas aos grupos sociais.
C fornecer soluções por meio de força e autoridade.
D expressar os interesses particulares da juventude.
E estabelecer os rumos norteadores de comportamento.

QUESTÃO 6

Muitos países enfrentam sérios problemas com seu elevado crescimento populacional. Em alguns destes países, foi proposta (e por vezes colocada em efeito) a proibição de as famílias terem mais de um filho. Algumas vezes, no entanto, esta política teve conseqüências trágicas (por exemplo, em alguns países houve registros de famílias de camponeses abandonarem suas filhas recém-nascidas para terem uma outra chance de ter um filho do sexo masculino). Por essa razão, outras leis menos restritivas foram consideradas. Uma delas foi: as famílias teriam o direito a um segundo (e último) filho, caso o primeiro fosse do sexo feminino.

Suponha que esta última regra fosse seguida por todas as famílias de um certo país (isto é, sempre que o primeiro filho fosse do sexo feminino, fariam uma segunda e última tentativa para ter um menino). Suponha ainda que, em cada nascimento, sejam iguais as chances de nascer menino ou menina. Examinando os registros de nascimento, após alguns anos de a política ter sido colocada em prática, seria esperado que

A o número de nascimentos de meninos fosse aproximadamente o dobro do de meninas.
B cada família, em média, tivesse 1,25 filho.
C aproximadamente 25% das famílias não tivessem filhos do sexo masculino.
D aproximadamente 50% dos meninos fossem filhos únicos.
E aproximadamente 50% das famílias tivessem um filho de cada sexo.

ENADE – 2004 Área: ZOOTECNIA 2

Fonte: Autoria Própria.

4.2 CONSTRUÇÃO DO *GROUND TRUTH*

Para a avaliação de desempenho da análise de *layout*, uma região diz respeito a um parágrafo em termos de texto (corpo do texto, cabeçalho, nota de rodapé, legenda) ou uma região gráfica (imagens, decisão horizontal/vertical). A representação da região é de fundamental importância em qualquer sistema de avaliação de desempenho. Os elementos compostos de um documento, como tabelas ou figuras com texto incorporado, são considerados cada um como uma única região (composta). A escolha de um esquema de representação da região é crucial para eficiência e precisão (ANTONACOPOULOS; MENG, 2002).

No decorrer desse trabalho, pesquisou-se na literatura por abordagens que pudessem auxiliar na construção dos *ground truths* e, assim, foram encontradas ferramentas específicas para essa demanda. Entretanto, algumas não estão mais disponíveis em seus

endereços informados nos artigos, tais como: WebGT[6] e TrueViz[7]. Já a ferramenta GEDI - *Groundtruthing Environment for Document Images* (DOERMANN; ZOTKINA; LI, 2010) está disponível para baixar, configurar e utilizar. Para isso, são necessárias configurações manuais como: cadastrar as regiões que serão utilizadas e marcar nas páginas todas as ROI. A última versão lançada do GEDI foi em 2013, com isso, para conseguir executá-lo é preciso utilizar a versão do Java 1.6 que está defasada, visto que a versão atual é a 16[8].

No que tange aos sistemas para análise e reconhecimentos de documentos e produção de arquivos de *ground truth*, o software Aletheia[9] é de longe o mais utilizado, visto que essa ferramenta foi mencionada em vários artigos com trabalhos semelhantes e também é utilizada para avaliação de desempenho de análise de *layout* de documentos submetidos em competições do ICDAR (Clausner; Antonacopoulos; Pletschacher, 2017), que é o principal evento para cientistas e pesquisadores envolvidos com análise e reconhecimento de documentos. O software Aletheia pertence a um grupo de pesquisa PRImA (Pattern Recognition & Image Analysis Research Lab), da Universidade de Salford Manchester, e está disponível em duas edições principais: Lite, com a maioria das funcionalidades de visualização e edição habilitadas; e Pro, com todos os recursos de análise e reconhecimento de documentos.

As seções seguintes relatam mais detalhes da construção dos *ground truths* realizados pelo Aletheia.

4.2.1 Construção do *Ground Truth* para Análise Quantitativa

Para este trabalho, a obtenção do *ground truth* foi feita por meio do software Aletheia. O fato de ter sido amplamente adotado em estudos semelhantes, ser mantido por um grupo de pesquisa, estar atualizado e apresentar diversas opções para trabalhar, contribuiu para a sua escolha e utilização. Apesar de ser um software proprietário e atualmente ser distribuído sob licença comercial, a ferramenta foi originada de pesquisas na Universidade de Salford. Acadêmicos podem consultar e obter licenças acadêmicas do Aletheia Pro válida por 183 dias, usada na presente pesquisa.

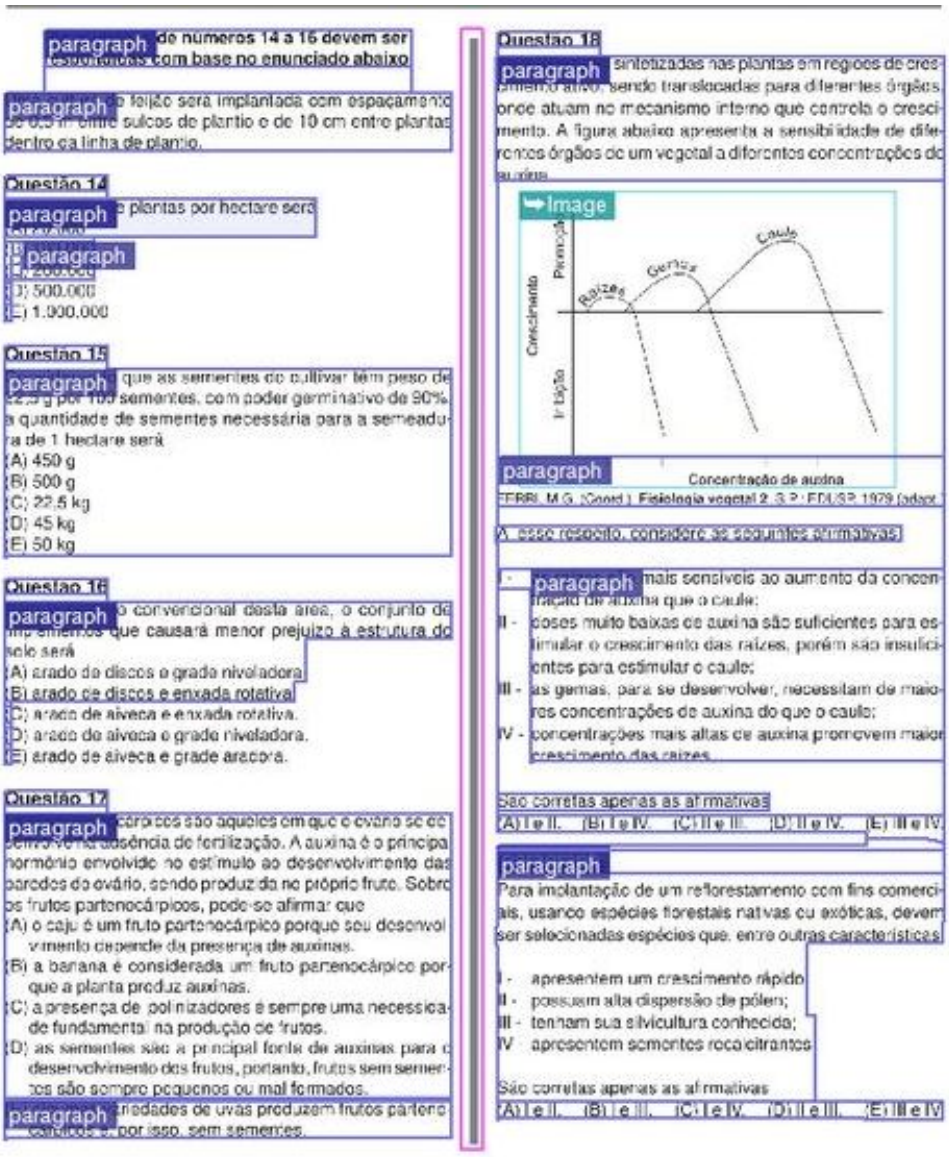
Cada arquivo de *ground truth* deste trabalho, foi gerado por meio do Aletheia de maneira automatizada, mas também com algumas intervenções manuais, que serão melhor relatadas no decorrer deste capítulo. A entrada para o Aletheia é uma imagem de página e, para isso, foi preciso converter todas as páginas das provas para JPG.

O Aletheia fornece uma interface necessária para rotular as regiões identificadas, como mostra uma captura de tela do software na Figura 4.4. O sistema permite que o

[6] http://win-web.cs.bgu.ac.il/ acessado em: 08/08/2020.
[7] www.cfar.umd.edu/∼kanungo/software/software.html acessado em: 08/08/2020.
[8] https://www.oracle.com/java/technologies/javase-downloads.html verificado em: 28 de março de 2021.
[9] Disponível em: https://www.primaresearch.org/tools/Aletheia

usuário altere o retângulo detectado automaticamente e, para resolver casos em que são necessárias formas de região complicadas, o software oferece a opção de usar um método de desenho a mão livre para selecionar componentes.

Figura 4.4 – Aletheia gerando automaticamente *ground truth*.



Fonte: Autoria própria.

Disponibiliza também funções de edição de região, por exemplo, para combinar regiões ou combinar regiões existentes com componentes individuais, enquanto as regiões podem ser separadas em seus componentes constituintes. O software visualiza as informações básicas atribuindo cores diferentes às regiões, dependendo do tipo, isso facilita o processo de rotular as ROI. Contém, ainda, várias opções, como: tipos de regiões definidas (tabela, imagem, texto); função para cadastrar outras regiões, caso necessário; e detecção automática das regiões de um conjunto de imagens sendo que, quando utilizada essa opção, os arquivos XML correspondentes às páginas são salvos automaticamente.

Inicialmente essa opção não era conhecida e o trabalho manual de selecionar página por página, de cada prova, levava em média 6 minutos em uma prova com 19 páginas. Após a utilização da opção *Analyse page*, a etapa reduziu o tempo em uma média de 4 minutos em provas com 19 páginas e, além disso, a análise ocorreu de maneira automática quando selecionada as opções necessárias que, neste caso, foram: *Language: Portuguese, Analysis Depth: Region with text* e *Run for all of collection: choose folder* (selecionando a pasta da prova) e *Run*.

Ademais, o Aletheia oferece duas opções para armazenar a descrição final do *ground truth*. A primeira é exportá-la como um arquivo XML (uma série de regiões individuais, juntamente com seus limites e atributos detalhados) que está em total conformidade com a especificação do *ground truth*. A segunda opção é salvar a representação do *ground truth* no próprio formato do software facilitando a edição posteriormente (ANTONACOPOULOS; KARATZAS; BRIDSON, 2006). Para este trabalho foram salvas as duas opções, o XML para ser utilizado na criação de um *script* e as *screenshot*/representação para facilitar na comparação com os resultados das ferramentas de extrações.

Cerca de 7 mil páginas das provas em PDF foram coletadas. Para cada página das provas do Enade que contém questões, um arquivo de *ground truth* está disponível com o arquivo XML correspondente, construído com a ferramenta Aletheia. Os arquivos estão no formato em XML e contém as coordenadas das regiões em uma estrutura hierárquica. Em termos de apresentação, cada página é composta de três partes: 1) *ground truth* identificado; 2) página original em formato jpg; 3) descrição em XML dos atributos contidos e construídos com o Aletheia, de acordo com as regiões selecionadas. Os arquivos XML possuem dados detalhados e que podem atender outros requisitos. Neste caso, foi utilizado para comparar com a saída das ferramentas de extração textual, PDFMiner e CyberPDF, além de ser utilizado para o *script* de contagem das métricas.

Conforme já mencionado, o objetivo da ferramenta Aletheia é apoiar os usuários finais na segmentação e classificação de regiões em imagens. Mesmo sendo projetado para analisar e marcar regiões automaticamente, em alguns casos, nesta pesquisa, um especialista precisou corrigir, de forma manual, algum *layout* detectado. Na experiência desta pesquisa, a parte mais demorada para construção do *ground truth* foi agrupar regiões que estão separadas quando devem estar em uma só.

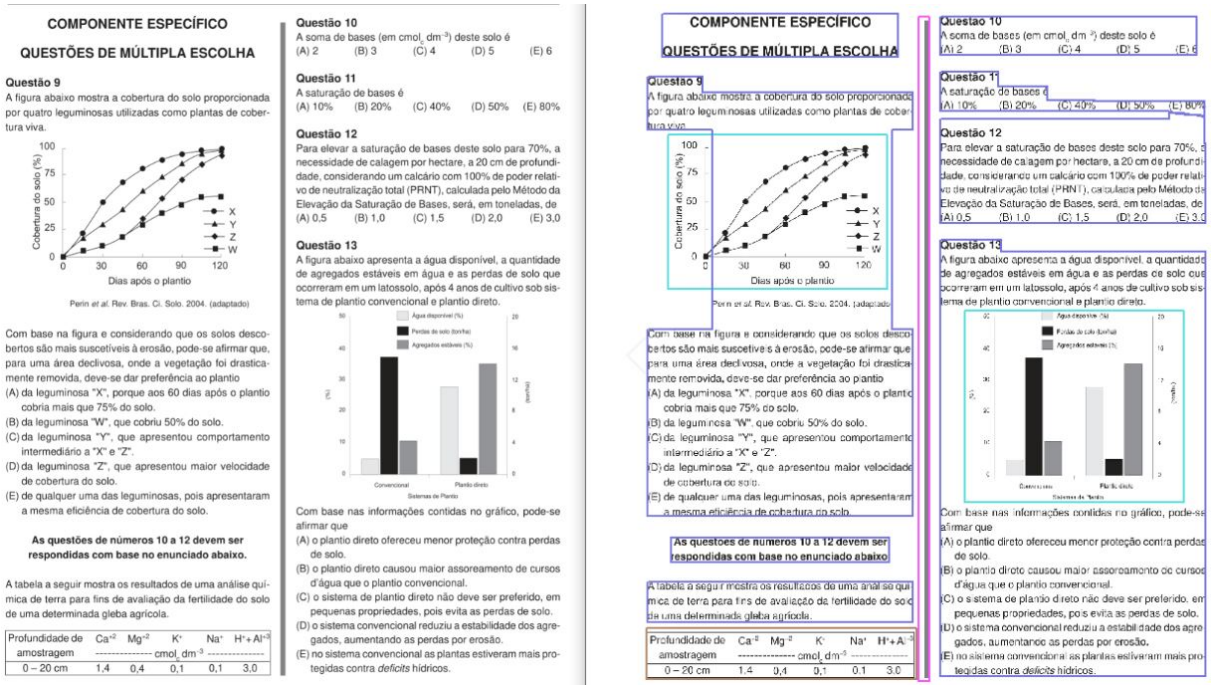
A maioria das intervenções do usuário consistiu na fusão de duas ou mais regiões. Isso ocorreu pelo fato de algumas questões possuírem muitos espaços entre o enunciado e as alternativas, sendo assim, a detecção automatizada do Aletheia reconheceu, por exemplo, como sendo mais de uma região. Constatou-se, ainda, que essa ferramenta facilita a edição (correção) de contornos de região. A construção dos *ground truths* para este trabalho estão em pastas separadas por ano [10]. Dentro dessas pastas contém as *screenshot* dos *ground truths* e os arquivos XML com as imagens originais.

[10] Disponível em: https://github.com/karinawie/PDFExtraction/tree/master/ground_truth

Após construídos os *ground truths* das provas, a quantidade de cada região em cada prova foi informada em uma planilha de maneira semiautomatizada com auxílio de um *script*. Para isso, foi realizada uma análise manual no *script* para conferir se os valores estavam de acordo com o total de métricas pré-identificadas em um pequeno grupo de provas. Essa análise aconteceu com a contabilização das métricas que determinada questão possui e considerando se o resultado correspondia com a saída do *script*.

A Figura 4.5 ilustra uma página no Aletheia com as ROI marcadas. O fluxo de trabalho do Aletheia consiste nas etapas de entrada, incluindo a página e a saída, sendo que os segmentos são classificados e salvos em XML.

Figura 4.5 – Exemplo de entrada (esquerda) e saída (direita) com doze regiões marcadas pertencentes a quatro tipos diferentes: imagem (azul claro), tabela (marrom), texto (azul escuro) e separador (rosa).



Fonte: Autoria própria.

4.2.2 Formato de Arquivo de *Ground Truth*

O *ground truth* prevê a representação de vários tipos de regiões, que podem estar sujeitas a diferentes processamentos em sistemas de reconhecimento. Para cada ROI, há uma descrição de seu contorno na forma de um polígono isotético, ou seja, um polígono tendo apenas bordas horizontais e verticais. Além disso, o formato oferece meios para demonstrar a ordem de leitura e relações mais complexas entre as regiões (Antonacopoulos et al., 2009).

Para armazenar os dados dos arquivos de *ground truth* foi utilizado o XML, que contém a estrutura do documento, incluindo regiões, linhas, páginas, palavras, conteúdo com caixas delimitadoras e a ordem dos elementos. A Figura 4.6 mostra um exemplo de arquivo de *ground truth* de uma prova do Enade.

Figura 4.6 – *Ground truth* de uma prova do Enade em formato XML.

```
<?xml version="1.0" encoding="UTF-8"?>
- <PcGts xsi:schemaLocation="http://schema.primaresearch.org/PAGE/gts/pagecontent/2019-07-15
http://schema.primaresearch.org/PAGE/gts/pagecontent/2019-07-15/pagecontent.xsd"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns="http://schema.primaresearch.org/PAGE/gts/pagecontent/2019-07-15">
  - <Metadata>
      <Creator/>
      <Created>2020-09-15T21:51:36</Created>
      <LastChange>2020-11-18T16:04:35</LastChange>
    </Metadata>
  - <Page imageHeight="2170" imageWidth="1619" imageFilename="administracao-11.jpg">
    - <ImageRegion id="r0">
        <Coords points="392,247 392,1205 1234,1205 1234,247"/>
      </ImageRegion>
    - <TextRegion id="r10" type="paragraph">
        <Coords points="106,181 277,181 277,1173 1512,1173 1512,1410 1513,1410 1513,1736
          597,1736 597,1791 238,1791 238,2043 106,2043"/>
      - <TextEquiv conf="0.95318">
          <Unicode>QUESTÃO 07 Promover a integração | SE || Promover a operação e a
            conectividade dos compartilhada de Á modos de transporte FF veículos autônomos
            Disponível em: <https://www.thinglink.com/scene/980079663516745730?
            buttonSource=viewLimits>. Acesso em: 26 jul. 2018 (adaptado). Considerando as
            informações do infográfico, avalie as afirmações a seguir. |. No planejamento das
            cidades, deve-se priorizar o transporte coletivo, situação que está em consonância
            com o que ocorre nas cidades mais populosas do Brasil. Il. O engajamento dos
            cidadãos nos debates e no planejamento das cidades é essencial para o
            desenvolvimento de projetos urbanos viáveis, acessíveis e sustentáveis. Ill. É
            necessário que o planejamento de uma cidade sustentável esteja focado na fluidez
            dos veículos automotores autônomos, na diversidade de opções de mobilidade e nas
            modalidades compartilhadas de transporte. IV. A utilização de painéis solares para
            abastecer veículos e a diminuição da emissão de gases poluentes em uma cidade
            sustentável são metas ainda distantes de serem atingidas no Brasil, devido à primazia
            dos meios de transportes movidos a combustíveis fósseis. É correto apenas o que se
            afirma em A I. B II. C I e III. D II e IV. E III e IV</Unicode>
        </TextEquiv>
      </TextRegion>
    </Page>
</PcGts>
```

Fonte: Autoria própria.

O formato XML é ideal para representar o *ground truth*, pois é o padrão atual do setor. Ele permite que os pesquisadores o entendam e o usem com facilidade para avaliar o algoritmo ou experimentar novas métricas (Fang et al., 2012). O uso do XML facilita a extensão do sistema para gerar outros tipos de documentos.

5 AVALIAÇÃO DE FERRAMENTAS PARA EXTRAÇÃO DE PDF

Neste capítulo, é apresentada a avaliação do comportamento de cinco ferramentas que realizam extrações de informações em arquivos PDFs, para demonstrar a confiabilidade do conjunto de dados e a eficácia das métricas de desempenho. Dessas, três são para extrações das 343 provas e duas para as extrações dos 396 gabaritos. As avaliações das ferramentas com o *ground truth* foram previstas conforme o processo proposto anteriormente, no Capítulo 4. No decorrer do presente Capítulo, apresentam-se também as métricas utilizadas e os resultados das avaliações.

5.1 FERRAMENTAS UTILIZADAS PARA EXTRAÇÃO DE PDF

A escolha de uma ferramenta que produza um formato fácil de analisar é importante, pois reduzirá o tempo de processamento necessário para extrair as informações de formatação necessárias (BUDHIRAJA, 2018). Atualmente, estão disponíveis várias abordagens para extração de dados em arquivos PDFs. A Tabela 5.1 lista as ferramentas que foram utilizadas para esta pesquisa. A fim de realizar a comparação entre elas, é necessário que tenha, em grande parte, os mesmos objetivos gerais e extrações semelhantes.

Tabela 5.1 – Visão geral dos recursos de cinco ferramentas de extração de PDF. A última coluna, fornece os formatos de saída disponíveis.

Ferramenta	Conteúdo Extraído	Formato
Excalibur	Tabelas	CSV, Excel, JSON, HTML
Tabula	Tabelas	CSV, TSV, JSON, Zip, Script
CyberPDF	Texto	XLS
PDFMiner	Texto	TXT, XML
ExamClipper	Regiões de Interesse	TIF

Fonte: Autoria própria.

Inicialmente foi decidido trabalhar com extrações das questões para dados textuais. No decorrer da pesquisa notou-se a problemática com algumas provas, que será relatado na Seção 5.5, e foi decidido utilizar uma extração diferente para as questões, optando pela ROI. Por fim, identificou-se várias ferramentas que trabalham com extrações de dados tabulares e a importância de extrair os gabaritos, concluindo as três categorias para as extrações.

A comparação do Excalibur foi realizada com o Tabula, para extração de respostas; CyberPDF e PDFMiner, para extração de conteúdo textual; e ExamClipper, para extração

de ROI. Não foi encontrada uma ferramenta concorrente ao ExamClipper. No entanto, existem ferramentas que processam imagens e permitem fazer recortes. O Aletheia foi utilizado para fazer os experimentos com recortes em ROI em arquivos PDF. Isso será melhor detalhado no decorrer da Seção 5.6.3.

Excalibur [1] É uma interface web para extrair dados tabulares de arquivos PDFs, detectando automaticamente essas tabelas. Pode ser configurado para cargas de trabalho paralelas e distribuídas, quando houver muitos PDFs. Por padrão, os PDFs são processados sequencialmente. Com essa ferramenta é possível extrair tabelas de vários arquivos de uma só vez, usando uma regra de extração ao iniciar os trabalhos e aplicando em PDFs com estruturas semelhantes. O Excalibur permite salvar as configurações de extração de tabela para um PDF e aplicá-las em novos PDFs ao extrair tabelas com estruturas semelhantes.

Tabula [2] A ferramenta extrai dados tabulares de arquivos PDFs. O Tabula permite importar vários arquivos PDF de uma só vez para realizar a extração, apresenta uma prévia para as extrações dos dados e caso não esteja de acordo com o resultado esperado há a opção de revisar as tabelas selecionadas.

Na extração dos gabaritos foram utilizadas as ferramentas Excalibur e Tabula para formato CSV, ambas trabalham no mesmo objetivo: extração de dados tabulares em arquivos de PDFs. Essas ferramentas foram selecionadas para realizar a extração dos gabaritos das questões objetivas.

CyberPDF [3] É um Sistema de Extração de Informação Baseado em Coordenadas (CBIES). A técnica proposta permite aos usuários consultar um documento PDF representativo e extrair os mesmos dados de uma série de arquivos na forma de análise de lote rapidamente (Parizi et al., 2018). Por ser uma ferramenta que trabalha com um padrão para vários arquivos PDF e os arquivos das provas não possuem um *layout* padrão, para esta pesquisa as coordenadas foram criadas com base na prova que mais possui páginas com questões para cada ano, com intuito de contemplar as demais provas com menos páginas. Essas coordenadas atendem a página inteira e não foram selecionadas coluna por coluna. O conteúdo extraído fica em uma célula em formato Excel.

PDFMiner [4] É uma ferramenta que se concentra em extrair conteúdo textual de arquivos PDFs. Primeiro analisa a estrutura do arquivo PDF e posteriormente o extrai para texto sem formatação. Pode acontecer em alguns cenários que as linhas extraídas

[1] Disponível em: https://github.com/camelot-dev/excalibur
[2] Disponível em: https://tabula.technology/
[3] Disponível em: https://github.com/LeonKwok0/CyberPDF
[4] Disponível em: https://github.com/euske/pdfminer

não são preservadas na ordem de leitura do texto, principalmente ao manusear documentos com mais de uma coluna ou na presença de *layout* complexo.

ExamClipper [5] É um software em desenvolvimento na UFSM que identifica regiões em arquivos PDFs e extrai recortes de questões. O ExamClipper permite detectar as regiões de acordo com as coordenadas informadas na etapa da extração ou possibilita ao usuário selecionar manualmente as ROI em cada página.

Na literatura, as extrações em arquivos PDFs mais abordadas são para conteúdo textual. Algumas das ferramentas apresentadas no Capítulo 3 foram testadas para utilização deste trabalho. É o caso da ferramenta PDFX que é de fácil utilização, por se tratar de uma ferramenta web que não tem a necessidade de executar no computador e disponibilizar várias saídas, no entanto, optou-se por não utilizá-la, pois a ferramenta suporta arquivos com no máximo 5MB e algumas das provas do Enade extrapolam esse tamanho.

O projeto PDFAct[6], que inicialmente era intitulado como Icecite[7], mistura os caracteres das questões no momento da extração, fazendo com que não fiquem na ordem de leitura correta. A ferramenta PDFigCapX pertence a um grupo de pesquisa da Universidade de Delaware[8] e diz respeito a um sistema web que suporta arquivos com tamanho de até 10MB, porém o link para extração não está disponível[9].

A ferramenta PDFMEF também foi testada, no entanto, optou-se por não a utilizar dada a complexidade em executá-la para todas as saídas (texto, imagem, tabela) e considerando que essas saídas ficam separadas umas das outras. Na Seção 5.2, em seguida, serão relatados os métodos de avaliação aplicados sobre as ferramentas utilizadas neste trabalho.

5.2 MÉTRICAS E CRITÉRIOS DE AVALIAÇÃO

Esta seção apresenta o conjunto de métricas e critérios estabelecidos para as avaliações aplicadas nos experimentos. Foi realizada uma avaliação comparativa de cinco ferramentas de extração de arquivos PDFs. De acordo com a necessidade deste trabalho, três categorias de extrações em PDFs estão disponíveis: tabelas, texto e ROI para formato de imagens.

Em cada categoria há duas ferramentas que extraem o mesmo conteúdo, exceto para ROI. Após essa extração, foi realizada a comparação com a saída da ferramenta pertencente a mesma categoria. Para isso, estabeleceu-se um conjunto de critérios que

[5]Disponível em: https://github.com/examclipper/examclipper
[6]Disponível em: https://github.com/ad-freiburg/pdfact
[7]Disponível em: https://github.com/ckorzen/icecite
[8]Disponível em: https://staff.eecis.udel.edu/
[9]Testado em 13 de janeiro de 2021.

permitem uma avaliação das ferramentas de extração comparando a saída das ferramentas com o *ground truth*.

5.2.1 Métricas de Avaliação

De acordo com (KRIG, 2014), métricas e dados de *ground truth* devem andar juntos. Para quantificar a exatidão ao analisar o desempenho das ferramentas, foram criadas as métricas listadas no Quadro 5.1. Métricas adequadas devem ser propostas para, em primeiro lugar, descobrir correspondências entre os resultados reconhecidos e o *ground truth* e, em seguida, avaliar e comparar os algoritmos (Fang et al., 2012). Essas métricas definidas, podem ser utilizadas em qualquer outros tipos de arquivos PDFs.

Quadro 5.1 – Notações das métricas.

Notação	Significado
1C	uma coluna na página
2C	duas colunas na página
MC	misto de colunas na página
1Q	uma questão por página/coluna
1QV	uma questão que começa em uma página/coluna e termina em outra
VQP	várias questões um uma página/coluna
QFG	questões com figuras/gráficos
QT	questões com tabelas
-	não disponível no conjunto de prova selecionado
N	ferramenta não reconhece

Fonte: Autoria própria.

A avaliação de desempenho é necessária para comparar e selecionar os métodos mais adequados para uma determinada aplicação. Algoritmos diferentes têm deficiências distintas considerando todas as métricas de avaliação. *Ground truth* contém dados suficientes e detalhados em vários aspectos e é necessária a utilização como uma referência para avaliar os resultados dos experimentos (Fang et al., 2012). Na próxima seção os critérios de avaliação são descritos em detalhes.

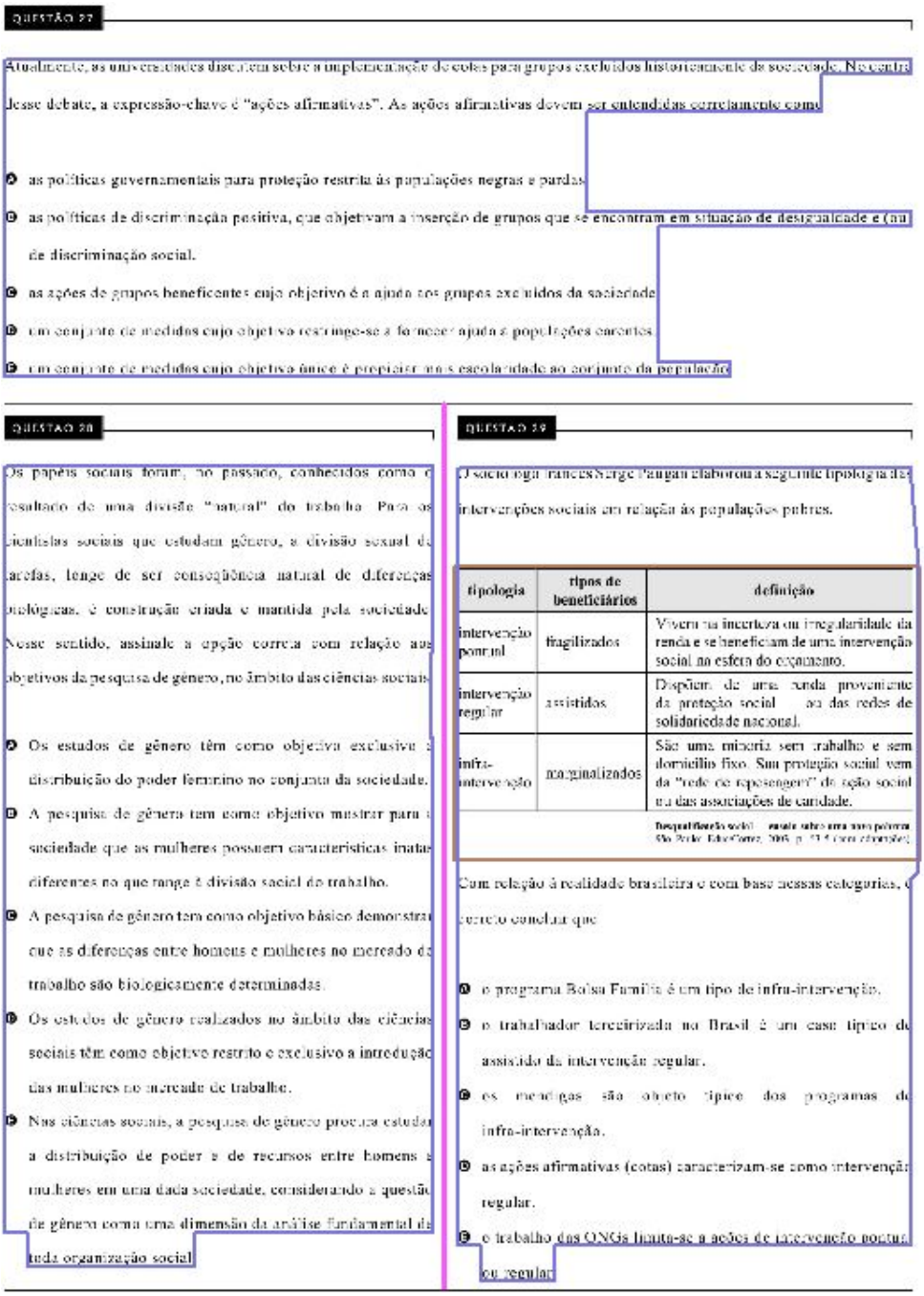
5.2.2 Critérios de Avaliação

Nesta pesquisa ao avaliar uma ferramenta, cada um de seus arquivos de saída é comparado com o arquivo de *ground truth* equivalente e, na sequência, com a ferramenta

concorrente. Os seguintes critérios de avaliação são medidos:

Questões com gráficos ou figuras são contabilizadas na categoria QFG (Questões com Figuras ou Gráficos). Questões que começam em uma coluna de uma página e terminam na página seguinte em duas colunas, devem ser contabilizadas na categoria que a questão iniciou. A Figura 5.1 demonstra um exemplo da notação considerada Coluna Mista (MC) na página. Neste exemplo, a página pertence à MC e foram contabilizadas três questões para notação Várias Questões em uma Página (VQP).

Figura 5.1 – Exemplo de coluna mista na página - MC.



QUESTÃO 37

Atualmente, as universidades discutem sobre a implementação de cotas para grupos excluídos historicamente da sociedade. No centro desse debate, a expressão-chave é "ações afirmativas". As ações afirmativas devem ser entendidas corretamente como

A as políticas governamentais para proteção restrita às populações negras e pardas.
B as políticas de discriminação positiva, que objetivam a inserção de grupos que se encontram em situação de desigualdade e (ou) de discriminação social.
C as ações de grupos beneficentes cujo objetivo é a ajuda aos grupos excluídos da sociedade.
D um conjunto de medidas cujo objetivo restringe-se a fornecer ajuda à populações carentes.
E um conjunto de medidas cujo objetivo único é propiciar uma escolaridade ao conjunto da população.

QUESTÃO 38

Os papéis sociais foram, no passado, conhecidos como o resultado de uma divisão "natural" do trabalho. Para os cientistas sociais que estudam gênero, a divisão sexual das tarefas, longe de ser consequência natural de diferenças biológicas, é construção criada e mantida pela sociedade. Nesse sentido, assinale a opção correta com relação aos objetivos da pesquisa de gênero, no âmbito das ciências sociais.

A Os estudos de gênero têm como objetivo exclusivo a distribuição do poder feminino no conjunto da sociedade.
B A pesquisa de gênero tem como objetivo mostrar para a sociedade que as mulheres possuem características inatas diferentes no que tange à divisão social do trabalho.
C A pesquisa de gênero tem como objetivo básico demonstrar que as diferenças entre homens e mulheres no mercado de trabalho são biologicamente determinadas.
D Os estudos de gênero realizados no âmbito das ciências sociais têm como objetivo restrito e exclusivo a introdução das mulheres no mercado de trabalho.
E Nas ciências sociais, a pesquisa de gênero procura estudar a distribuição de poder e de recursos entre homens e mulheres em uma dada sociedade, considerando a questão de gênero como uma dimensão da análise fundamental de toda organização social.

QUESTÃO 39

O sociólogo francês Serge Paugam elaborou a seguinte tipologia das intervenções sociais em relação às populações pobres.

tipologia	tipos de beneficiários	definição
intervenção pontual	fragilizados	Vivem na incerteza ou irregularidade da renda e se beneficiam de uma intervenção social no esteio do cotidiano.
intervenção regular	assistidos	Dispõem de uma renda proveniente da proteção social ou das redes de solidariedade nacional.
infra-intervenção	marginalizados	São uma miséria sem trabalho e sem domicílio fixo. Sua proteção social vem da "rede de responsáveis" do ação social ou das associações de caridade.

[illegible] social: ensaio sobre uma nova pobreza. São Paulo: Educ/Cortez, 2003, p. 53-4 (com adaptações).

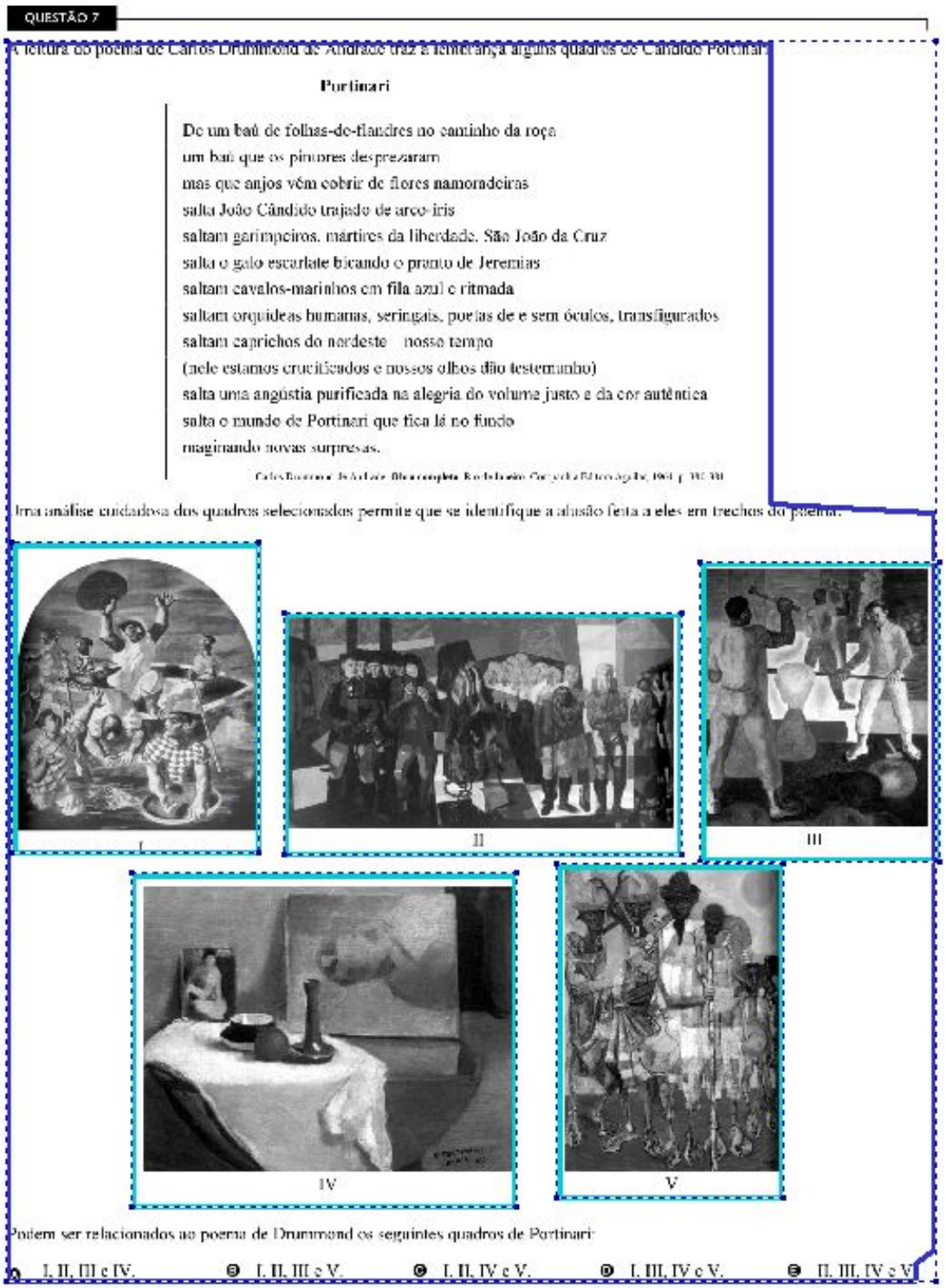
Com relação à realidade brasileira e com base nessas categorias, é correto concluir que

A o programa Bolsa Família é um tipo de infra-intervenção.
B o trabalhador terceirizado no Brasil é um caso típico de assistido da intervenção regular.
C os mendigos são objeto típico dos programas de infra-intervenção.
D as ações afirmativas (cotas) caracterizam-se como intervenção regular.
E o trabalho das ONGs limita-se a ações de intervenção pontual ou regular.

Fonte: Autoria Própria

A MC é considerada quando se tem um texto misturado em diferentes colunas, desconsiderando a posição das imagens, como mostra o exemplo da Figura 5.2. A métrica equivalente para esse exemplo é uma questão na página (1Q) em uma coluna (1C) com cinco imagens (QFG).

Figura 5.2 – Exemplo de extração com métrica equivalente a 1Q-1C.


QUESTÃO 7

A leitura do poema de Carlos Drummond de Andrade traz à lembrança alguns quadros de Cândido Portinari.

Portinari

De um baú de folhas-de-flandres no caminho da roça
um baú que os pintores desprezaram
mas que anjos vêm cobrir de flores namoradeiras
salta João Cândido trajado de arco-íris
saltam garimpeiros, mártires da liberdade. São João da Cruz
salta o galo escarlate bicando o pranto de Jeremias
saltam cavalos-marinhos em fila azul e ritmada
saltam orquídeas humanas, seringais, poetas de e sem óculos, transfigurados
saltam caprichos do nordeste nosso tempo
(nele estamos crucificados e nossos olhos dão testemunho)
salta uma angústia purificada na alegria do volume justo e da cor autêntica
salta o mundo de Portinari que fica lá no fundo
imaginando novas surpresas.

Uma análise cuidadosa dos quadros selecionados permite que se identifique a alusão feita a eles em trechos do poema.

I

II

III

IV

V

Podem ser relacionados ao poema de Drummond os seguintes quadros de Portinari:

A) I, II, III e IV. B) I, II, III e V. C) I, II, IV e V. D) I, III, IV e V. E) II, III, IV e V.

Fonte: Autoria Própria

A validação para que uma questão seja extraída corretamente ocorre por meio de uma mesma sequência de linha da questão original, desconsiderando o enunciado “Questão 00”. A ideia é que a extração da questão esteja pronta para que o usuário copie e cole em outro lugar, por exemplo. As extrações em que as questões estão lado a lado não foram consideradas, visto que a extração para esse caso acontece linha a linha não identificando as colunas que as separam.

Para realizar a comparação das extrações das informações com o *ground truth*, um *script* de contagem de métricas foi desenvolvido e está disponível no GitHub [10]. O *script* que realiza a contagem das métricas para cada questão, de acordo com o ano e área da prova, salva a saída para um arquivo com essas informações pertencentes.

O *script* utiliza o arquivo XML, detectando o início da questão por meio de uma expressão regular: “QUESTÃO|Questão|QUESTÃO DISCURSIVA|Questão Discursiva”, seguindo de números entre 0 a 9, com dois dígitos, ou antecedendo de espaços em brancos com dois dígitos de números entre 0 a 9, seguindo de enter. Esse segundo caso ocorre em questões das provas do ano de 2005, as quais não iniciam com “Questão”, como nas outras provas, mas sim com o número da questão. Isso só foi identificado após conferir que a contabilização para esse ano fechava em zero.

A técnica usada para calcular o desempenho das ferramentas foi a regra de três simples, onde são contabilizadas apenas as questões extraídas por completo para cada prova. Na fórmula a seguir, o valor de “total_de_questoes_para_extrair” equivale ao total de questões identificadas no *ground truth*. Foi realizada a soma total de cada métrica para cada ano e aplicou-se a regra de três simples. Por fim, calculou-se a porcentagem para obter a porcentagem de identificação em todo o conjunto de dados para cada ferramenta.

$$\frac{questoes\_extraidas\_pela\_ferramenta * 100}{total\_de\_questoes\_para\_extrair}$$

Para acessar os valores do *ground truth* em cada prova, as informações estão disponíveis no GitHub [11]. Os resultados das extrações sobre cada ferramenta estão disponíveis para links externos no Google Sheets, sendo que os valores de *ground truth* estão identificados nas células verdes das planilhas.

[10] https://github.com/karinawie/XML_aletheia

[11] Disponível em: https://github.com/karinawie/PDFExtraction

5.3 AMBIENTE EXPERIMENTAL

Nesta seção será apresentada de maneira sucinta as configurações dos computadores utilizados para os experimentos. A importância em apresentar essas configurações está associada com o resultado do desempenho dos experimentos. O trabalho foi realizado em dois computadores distintos.

O fato de utilizar um segundo computador foi uma conveniência pelo ambiente de trabalho utilizado, no caso o Ubuntu. Foi necessário utilizar o Windows para execução do software Aletheia. Identificou-se essa restrição após todas as extrações dos dados já realizadas no Ubuntu.

O Quadro 5.2 apresenta as tarefas executadas em cada computador. Com base nessas execuções, foram obtidos os resultados das avaliações que serão apresentados na Seção 5.6.

Quadro 5.2 – Configurações dos computadores utilizados nos experimentos.

	PC1	**PC2**
Processador	Intel Core i5-8250U	Intel Core i5-4590 (haswell)
Speed	1.60 GHz	3.30GHz
RAM	8GB	8GB
SO	Ubuntu 18.04	Windows 10
Excalibur (exec. e comp.)	X	
Tabula (exec. e comp.)	X	
CyberPDF (exec.)	X	
CyberPDF (comp.)		X
PDFMiner (exec.)	X	
PDFMiner (comp.)		X
ExamClipper (exec. e comp.)	X	
Aletheia (exec. e comp.)		X
Construção *ground truth*		X

Fonte: Autoria própria.

5.4 EXPERIMENTOS REALIZADOS

Para verificar o desempenho das ferramentas selecionadas, foram realizados experimentos extensivos no conjunto de dados. Os critérios de avaliação introduzidos na Seção 5.2 são facilmente interpretáveis, mas medi-los não é trivial.

Inicia-se com as questões da parte de formação geral. Para todas as provas de um mesmo ano essas questões sempre são iguais, excepcionalmente em alguns anos o *layout* dessas questões teve alterações, ou seja, foram reconhecidos mais de um padrão nas mesmas questões de formação geral. A seguir, são listados esses anos com a identificação da formação geral com as provas que foram identificadas com o mesmo padrão:

- 2004 - **Formação Geral 1**: Agronomia, Educação Física, Enfermagem, Farmácia, Medicina, Medicina Veterinária, Nutrição, Odontologia e Serviço Social. **Formação Geral 2**: Fisioterapia, Fonoaudiologia, Terapia Ocupacional e Zootecnia.

- 2005 - **Formação Geral 1**: Arquitetura e Urbanismo, Engenharia Grupo I, Engenharia Grupo II, Engenharia Grupo III, Engenharia Grupo VIII, Pedagogia e Química. **Formação Geral 2**: Biologia, Física, Geografia, História, Letras. **Formação Geral 3**: Ciências Sociais, Computação, Engenharia Grupo VI, Engenharia Grupo V, Engenharia Grupo VI, Engenharia Grupo VII, Filosofia e Matemática.

- 2007 - **Formação Geral 1**: Agronomia, Biomedicina, Enfermagem, Farmácia, Fisioterapia, Medicina, Medicina Veterinária, Odontologia, Serviço Social e Tecnologia em Radiologia. **Formação Geral 2**: Educação Física, Nutrição, Tecnologia em Agroindústria, Terapia Ocupacional e Zootecnia.

Após essa identificação para todas as provas utilizadas nesta pesquisa, os experimentos foram efetuados apenas uma vez nas questões pertencentes à formação geral. Nessas provas com mais de um padrão de *layout* a contagem das questões ocorreu uma vez para cada padrão. Em seguida, a contagem se deu apenas nas questões específicas. Essa abordagem reduziu a quantidade de computação necessária.Inicialmente o trabalho seria aplicado em 14.386 questões, mas, por fim, resultou em 11.196 questões objetivas e discursivas utilizadas para extração das questões.

A comparação com a categoria tabelas realizou-se de maneira manual por um especialista. Para a categoria de extração em dados textuais, a comparação foi semiautomatizada utilizado um *script*. A explicação desse *script* é descrita na Seção 5.2.1, juntamente com o notepad++ instalando um plugin específico nomeado de Compare, que permite comparar dois textos conseguindo identificar diferenças entre eles.

Foi adotada uma abordagem tolerante ao julgar se um determinado objeto foi representado corretamente ou não. No caso das fórmulas, muitas vezes são utilizados caracteres que não constam no alfabeto, essas então foram ignoradas. A maioria das alternativas nas questões possuem caracteres romanos, as extrações com esses caracteres também não ocorreram corretamente.

As ferramentas foram executadas para obter a saída final e então comparadas com os resultados de sua concorrente. Em seguida, ambos os resultados foram comparados com o *ground truth*, de acordo com o conjunto de métricas propostas na Seção 5.2.

Na utilização da ferramenta CyberPDF foram criadas as coordenadas com base na prova de cada ano que contém maior quantidade de páginas utilizadas, para que as demais pudessem ser contempladas no momento da extração, conforme descrito anteriormente na Seção 5.1.

Com Excalibur e Tabula, que extraem dados tabulares e foram utilizadas para extrair os gabaritos, as avaliações foram apenas nas respostas objetivas. Não está incluso na contagem as respostas discursivas. Alguns dos gabaritos não estão disponíveis no site do INEP, como ocorre com provas de Arquivologia (2006) e Medicina (2007), em que o gabarito anexado pertence a outra prova.

O principal objetivo da avaliação era analisar cada ferramenta, comparando seus arquivos de saída com os arquivos de *ground truth* usando o conjunto de métricas e critérios estabelecidos. Por fim, ocorreram diversos erros em que as extrações das provas de alguns anos não foram detectadas. Isso será explicado mais detalhadamente na próxima seção.

5.5 ANÁLISE DE ERROS

Durante a etapa de extração do conteúdo textual, mais especificamente utilizando a ferramenta PDFMiner, observou-se que em algumas provas o texto das questões não era extraído ou sua saída apresentava “cid” (*Caractere Identificador*), como mostra a Figura 5.3.

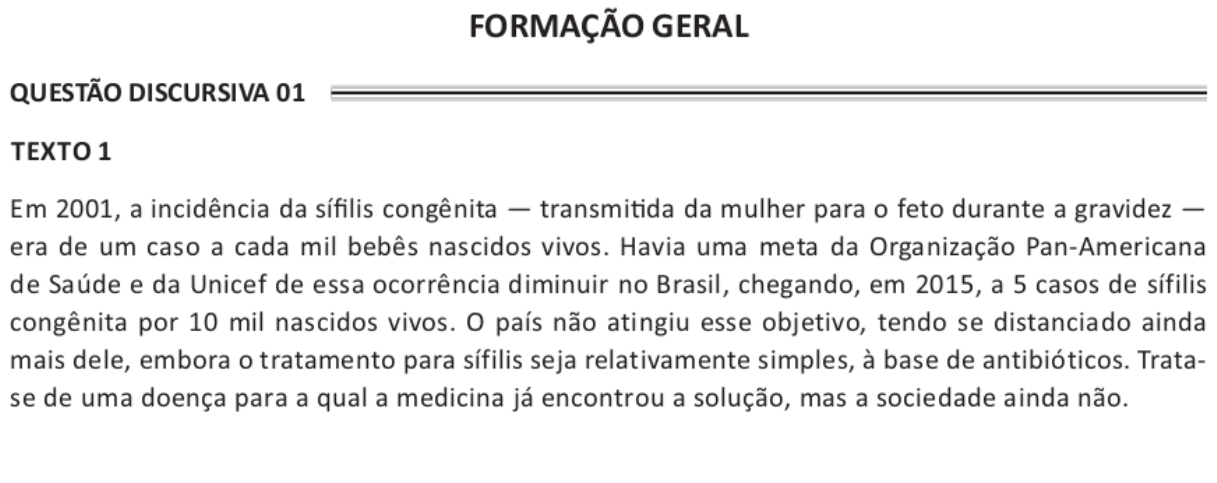
Figura 5.3 – Extração com o PDFMiner realizada na prova de formação geral do ano 2017.

```
QUESTÃO DISCURSIVA 01

TEXTO 1

FORMAÇÃO GERAL

(cid:28)(cid:373)(cid:3)(cid:1006)(cid:1004)(cid:1004)(cid:1005)(cid:853)(cid:3)(cid:258)
(cid:3)(cid:349)(cid:374)(cid:272)(cid:349)(cid:282)(cid:289)(cid:374)(cid:272)(cid:349)
(cid:258)(cid:3)(cid:282)(cid:258)(cid:3)(cid:400)(cid:351)(cid:302)(cid:367)(cid:349)(cid:
400)(cid:3)(cid:272)(cid:381)(cid:374)(cid:336)(cid:289)(cid:374)(cid:349)(cid:410)(cid:258)
(cid:3)(cid:888)(cid:3)(cid:410)(cid:396)(cid:258)(cid:374)(cid:400)(cid:373)(cid:349)(cid:
415)(cid:282)(cid:258)(cid:3)(cid:282)(cid:258)(cid:3)(cid:373)(cid:437)(cid:367)(cid:346)
(cid:286)(cid:396)(cid:3)(cid:393)(cid:258)(cid:396)(cid:258)(cid:3)(cid:381)(cid:3)(cid:
296)(cid:286)(cid:410)(cid:381)(cid:3)(cid:282)(cid:437)(cid:396)(cid:258)(cid:374)(cid:410)
(cid:286)(cid:3)(cid:258)(cid:3)(cid:336)(cid:396)(cid:258)(cid:448)(cid:349)(cid:282)(cid:
286)(cid:460)(cid:3)(cid:888)(cid:3)
```

Fonte: Autoria Própria

A Figura 5.4 apresenta o mesmo trecho com o conteúdo original da prova em PDF, que deveria ter sido extraída. Se a opção copiar e colar for utilizada de maneira manual nessas provas, serão apresentados caracteres estranhos com símbolos, conforme exposto na Figura 5.5.

Figura 5.4 – Questão original a ser extraída da prova de formação geral do ano 2017.



FORMAÇÃO GERAL

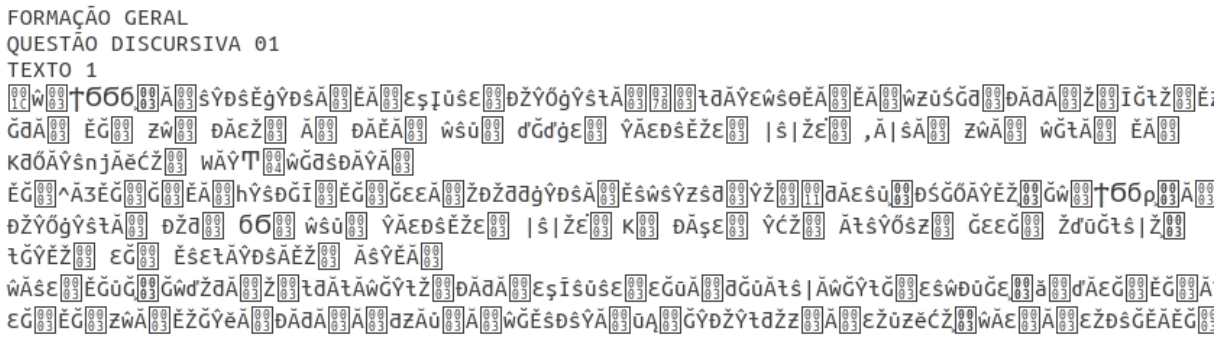
QUESTÃO DISCURSIVA 01

TEXTO 1

Em 2001, a incidência da sífilis congênita — transmitida da mulher para o feto durante a gravidez — era de um caso a cada mil bebês nascidos vivos. Havia uma meta da Organização Pan-Americana de Saúde e da Unicef de essa ocorrência diminuir no Brasil, chegando, em 2015, a 5 casos de sífilis congênita por 10 mil nascidos vivos. O país não atingiu esse objetivo, tendo se distanciado ainda mais dele, embora o tratamento para sífilis seja relativamente simples, à base de antibióticos. Trata-se de uma doença para a qual a medicina já encontrou a solução, mas a sociedade ainda não.

Fonte: Autoria Própria

Figura 5.5 – Caracteres estranhos ao copiar e colar a questão da prova de formação geral do ano 2017.



FORMAÇÃO GERAL
QUESTÃO DISCURSIVA 01
TEXTO 1

Fonte: Autoria Própria

Em busca de resolver esse problema, foi solicitado o parecer de um especialista da Adobe sobre o relato do problema com as extrações nas provas do ano de 2017[12]. O mesmo relata: “Os arquivos em questão eram de 2017 e foram criados com ferramentas que não eram suportadas pela Adobe há muitos e muitos anos. O texto está codificado de uma maneira não padronizada. Não posso falar em nome do seu software não Adobe PDFMiner, mas o software Adobe atual, ou seja, o Acrobat Pro DC, também não consegue entender isso. A única solução/alternativa que posso sugerir seria abrir qualquer arquivo

[12]A autorização de seu nome não foi solicitada/concedida.

PDF do qual você deseja extrair dados no Acrobat Pro DC, salvar as páginas como imagens TIFF de pelo menos 600 dpi, lê-las de volta no Acrobat, convertê-las em PDF, executar OCR (óptico reconhecimento de caracteres) no Acrobat e, em seguida, tente extrair o texto. Isso pode funcionar".

Com base nesses respaldos, não serão apresentados resultados de extração de dados textuais, CyberPDF e PDFMiner, para as provas dos anos de: 2010, 2014, 2015 e 2017, visto que resultaram em nenhuma questão extraída. Os resultados de todas as extrações são melhor apresentados em seguida, na Seção 5.6. Testes adicionais foram realizados com outras ferramentas, Ghostscript[13] e Give me text!, em outro computador e diferente sistema operacional, mas também não realizaram extrações nas mesmas partes das provas com relatos de problemas. Conclui-se, portanto, que o problema não está nas ferramentas e sim nos arquivos das provas.

5.6 RESULTADOS DAS AVALIAÇÕES

Essa seção apresenta os resultados obtidos a partir de experimentos realizados utilizando as ferramentas de extrações de informações em PDF. Para cada ferramenta é fornecida uma descrição concisa, de acordo com os critérios abordados. Os resultados completos estão disponíveis [14]. Esse processo extraiu dados de 343 provas sendo que dessas 130 resultaram em nenhum dado textual extraído, conforme detalhado anteriormente.

As informações do *ground truth* de cada página das provas foram descritas em uma planilha [15], para realizar comparações com as informações extraídas das ferramentas. As métricas para essas comparações foram pré-estabelecidas e já apresentadas na Seção 5.2. A análise dos resultados demonstra a eficácia das métricas propostas e fornece informações sobre o desempenho e as características das ferramentas avaliadas.

A Tabela 5.2 apresenta informações obtidas durante as extrações realizadas em cada ferramenta de extração de PDF. A coluna "Erro" fornece o número de arquivos PDF que não puderam ser processados pela ferramenta. A coluna "Tempo" indica o tempo médio, em segundos, necessário para extrair os dados de um único arquivo PDF. Os melhores valores em cada coluna são apresentados em negrito. O valor obtido do tempo médio foi calculado em cinco provas iguais para todas as ferramentas apenas no momento da extração, sem contabilizar o tempo de anexar as provas nas ferramentas.

[13] Disponível em: https://www.ghostscript.com/download/gsdnld.html
[14] Disponível em: https://github.com/karinawie/PDFExtraction/tree/main/extractions
[15] Disponível em: https://github.com/karinawie/PDFExtraction

Tabela 5.2 – Informações obtidas durante o processo de extração no conjunto de dados.

Ferramenta	Erro	Tempo
Excalibur	0	20
Tabula	0	**16**
CyberPDF	130	22
PDFMiner	130	**16**
ExamClipper	0	240

Fonte: Autoria Própria.

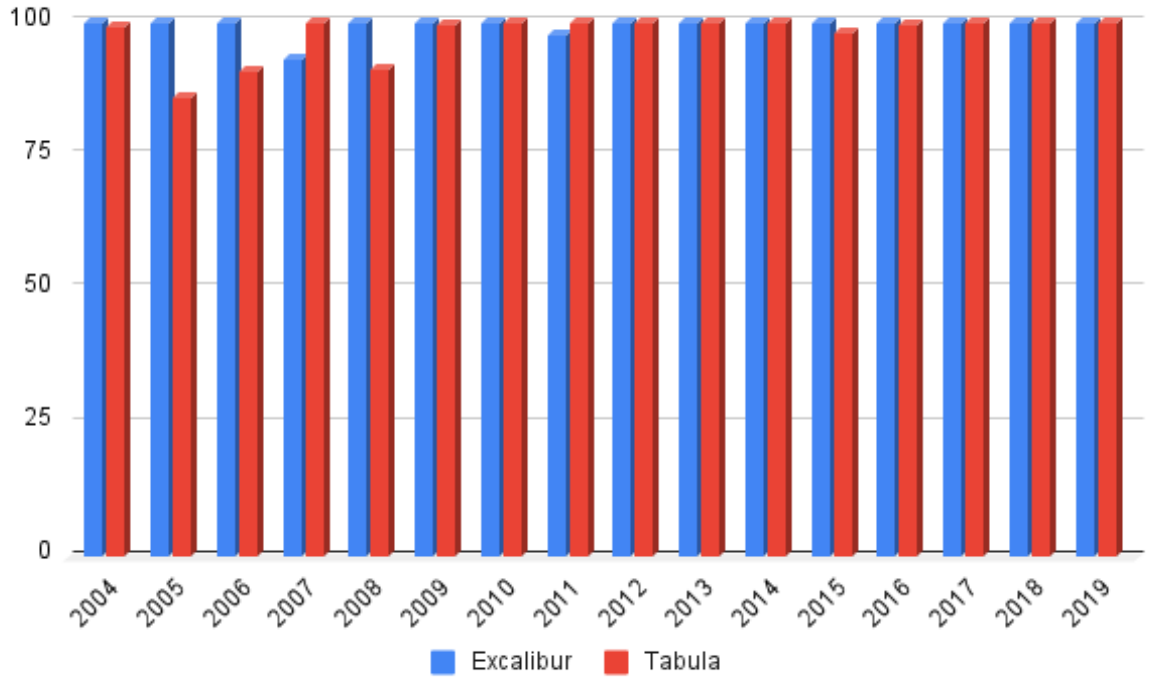
5.6.1 Avaliação de Dados Tabulares

Os resultados experimentais das extrações de dados tabulares podem ser vistos no Gráfico 5.1. Observa-se que entre os anos 2005, 2006 e 2008, a ferramenta Tabula teve dificuldade em extrair todas as alternativas dos gabaritos. A extração com o Excalibur para o ano de 2007 apresentou dificuldade, pois o *layout* dos gabaritos não são os mesmos. Identificou-se que o gabarito para a prova de Educação Física foi configurado na posição paisagem e não centralizada, enquanto todos os demais estão como retratos e com a tabela centralizada.

Os resultados detalhados estão disponíveis na planilha [16]. É possível identificar a quantidade de alternativas que a ferramenta deveria ter identificado dentro das tabelas e a quantidade de alternativas que foram corretamente extraídas.

[16] Disponível em: https://github.com/karinawie/PDFExtraction

Gráfico 5.1 – Resultados da extração do Excalibur e Tabula detalhados por ano.



Fonte: Autoria Própria.

As informações na Tabela 5.3 estão relacionadas apenas para as extrações dos dados tabulares nos gabaritos. Nessa extração, foi contabilizado apenas Questões com Tabelas (QT). Como os gabaritos de cada prova estão dispostos em uma única tabela, a contagem foi aplicada para cada alternativa extraída, na tabela do gabarito.

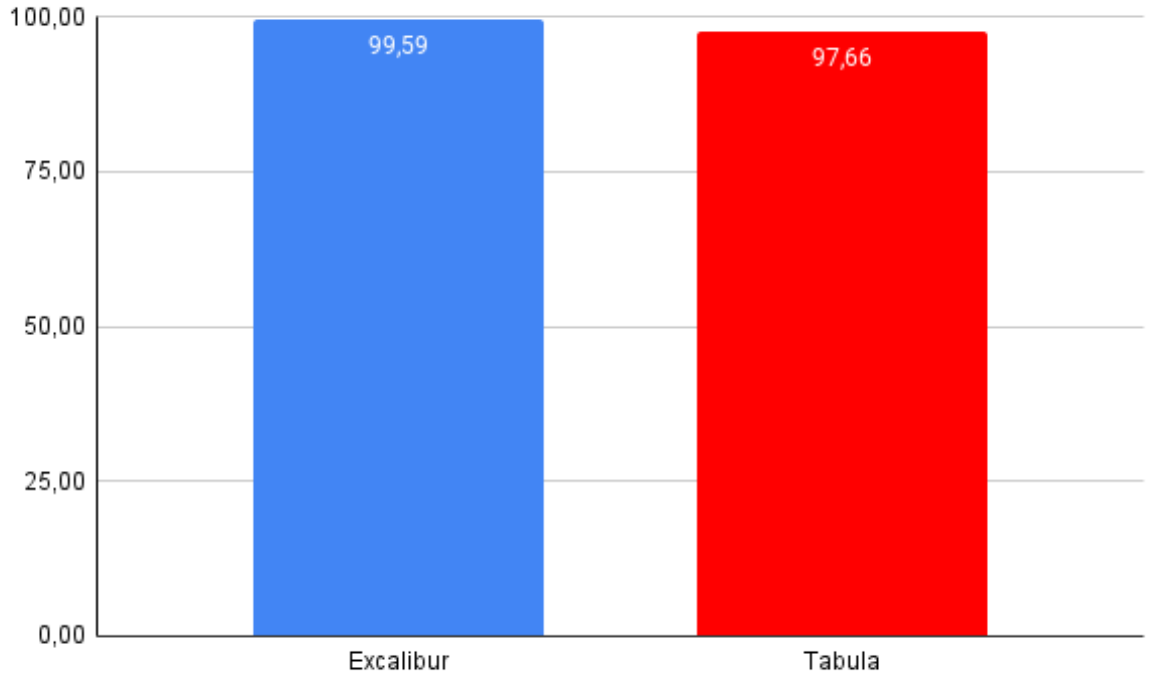
Tabela 5.3 – Visão geral dos resultados das ferramentas de extração de dados tabulares.

Ferramenta	QT
Excalibur	**99,59**
Tabula	97,66

Fonte: Autoria Própria.

O Gráfico 5.2 mostra detalhadamente que as ferramentas Excalibur e Tabula atingem um resultado de quantidade de extração bem próximos, no entanto, a ferramenta Excalibur apresenta um melhor desempenho. A quantidade total de respostas nas tabelas é de 14.475. O total extraído pelo Excalibur foi 14.415, ou seja, 99,59%. O total extraído pelo Tabula resultou em 14.136, equivalente a 97,66%.

Gráfico 5.2 – Comparação da extração de todos os anos do Excalibur e Tabula, quanto maior o resultado mais eficiente é a ferramenta.



Fonte: Autoria Própria.

5.6.2 Avaliação de Dados Textuais

A extração de texto desempenha uma função importante para fluxos de trabalho de processamento de dados. Formatos de arquivo complexos tornam o processo de extração sujeito a erros e tornam muito difícil verificar a exatidão dos componentes de extração. Com base em cenários de preservação digital e recuperação de informação, três requisitos de qualidade em termos de eficácia das ferramentas de extração de texto são identificados: 1) a questão é extraída corretamente (é mantido a integridade do texto ao lidar com várias colunas de texto); 2) o texto extraído aparece na ordem de leitura correta em relação a outra região/elemento; 3) a estrutura da questão é preservada (DURETEC; RAUBER; BECKER, 2017).

Os valores (1C, 2C, MC, 1Q, 1QV, VQP, QFG e QT) na Tabela 5.4, foram calculados com a regra de três simples. De acordo a tabela, as métricas para duas colunas (2C) têm uma taxa de recuperação relativamente baixa, pois as ferramentas não identificam que a página contém 2C, sendo assim, a extração acaba sendo realizada como se a página tivesse coluna única (1C). Isso ocorre de maneira parecida com as métricas de Coluna

Mista (MC). No entanto, as marcações “-”, na tabela, definidas na Seção 5.2, significam que no conjunto de dados não foram identificadas questões com essas métricas.

Identificou-se que o CyberPDF tem dificuldades em extrair questões em 2C. Isso pode ocorrer devido à maneira que foi configurada para ser utilizada neste trabalho, conforme explicado na Seção 5.1. O PDFMiner tem apenas uma métrica com valor acima de 50%, em 1C-VQP com 62,17% de extração, no entanto, a comparação geral com as métricas é a ferramenta que melhor extrai dados.

Tabela 5.4 – Visão geral dos resultados das ferramentas de extração de dados textuais.

		1Q	1QV	VQP	QFG	QT
CyberPDF	1C	**55,10**	**47,77**	**81,71**	N	N
	2C	1,81	0	7,63	N	N
	MC	-	-	26,67	N	N
PDFMiner	1C	46,38	40,13	62,17	N	N
	2C	**26,19**	**42,31**	**44,64**	N	N
	MC	-	-	**35,83**	N	N

Fonte: Autoria Própria.

A Tabela 5.5 apresenta uma comparação detalhada nas extrações com o CyberPDF, entre o ano que obteve a pior média com o ano com melhor média aplicando sobre as métricas contidas no conjunto de provas em cada ano, nem todas as provas possuem todas as métricas. Os valores com “ - ” não significam zero e sim que determinadas métricas não constam nas avaliações. Os anos de 2018, 2016 e 2013 resultaram com melhor média, empatados com 60% de extrações. Neste caso, optou-se por apresentar os valores do ano mais recente. O ano de 2009 teve uma média de 20% de extrações realizadas, sendo o ano com pior média.

Tabela 5.5 – Comparação dos resultados das extrações com o CyberPDF do ano com pior média em comparação com o ano que obteve melhor média de resultados.

		1Q	1QV	VQP	QFG	QT
2009	1C	36,04	-	38,41	N	N
	2C	5,51	-	8,57	N	N
	MC	-	-	11,11	N	N
2018	1C	100	100	100	N	N
	2C	0	-	0	N	N
	MC	-	-	-	N	N

Fonte: Autoria Própria.

A tabela 5.6 apresenta uma comparação detalhada nas extrações com o PDFMiner, entre o ano que obteve a pior média com o ano com melhor média aplicando sobre as métricas contidas no conjunto avaliado. O ano de 2009 teve uma média de 14% de extrações realizadas, sendo o ano com pior média. O ano de 2016 foi o ano com melhor média, 79%.

Tabela 5.6 – Comparação dos resultados das extrações com o PDFMiner do ano com pior média em comparação com o ano que obteve melhor média.

		1Q	1QV	VQP	QFG	QT
2009	1C	19,82	-	22,56	N	N
	2C	13,39	-	3,43	N	N
	MC	-	-	11,11	N	N
2016	1C	91,02	100	72,22	N	N
	2C	52,10	-	77,78	N	N
	MC	-	-	-	N	N

Fonte: Autoria Própria.

5.6.3 Avaliação de Regiões de Interesse

A Tabela 5.7 mostra os resultados das extrações realizadas com o ExamClipper, o qual identifica ROI e realiza recortes das questões nas provas em PDF. Ressaltando que não foi encontrada uma ferramenta concorrente ao ExamClipper. Decidiu-se aplicar a extração com a ferramenta Aletheia, pois a mesma apresenta opção para ROI.

Tabela 5.7 – Visão geral dos resultados da ferramenta de extração de ROI.

		1Q	1QV	VQP	QFG	QT
ExamClipper	1C	91,44	72,61	65,33	74,78	78,62
	2C	69,54	61,54	51,27	44,16	58,33
	MC	-	-	36,67	43,18	50

Fonte: Autoria Própria.

A comparação do Aletheia com o ExamClipper não é apresentada, visto que, a utilização do Aletheia para as extrações não condiz com a maneira utilizada com as demais ferramentas, pois os valores do Aletheia resultaram tendo em vista a utilização dos arquivos XML, com suas respectivas imagens, criados durante o desenvolvimento dos *ground truths*, ou seja, os mesmos arquivos foram reaproveitados para realizar as extrações. Isso acabou

favorecendo as extrações em 100% para todas as métricas avaliadas. A vantagem é que, todas as questões foram extraídas com o Aletheia e estão disponíveis no GitHub [17].

Os valores do ExamClipper foram obtidos com a própria detecção que a ferramenta disponibiliza na interface de recortes. Foi perceptível uma pequena dificuldade de extração quando o *layout* da página possui MC com o ExamClipper. Apenas três métricas não ficaram acima de 50%, MC-VQP, MC-QFG e 2C-QFG.

A tabela 5.8 apresenta uma comparação detalhada nas extrações com o ExamClipper, entre o ano que obteve a pior média com o ano com melhor média. O ano de 2005 teve uma média de 36% de extrações, sendo o ano com pior média. O ano de 2012 e 2010 foram os anos com melhor média, empatados com 79% de extrações. Optou-se por apresentar os valores do ano mais recente.

Tabela 5.8 – Comparação dos resultados das extrações com o ExamClipper do ano com pior média em comparação com o ano que obteve melhor média de resultados.

		1Q	1QV	VQP	QFG	QT
2005	1C	78,85	100	46,7	59,83	46,43
	2C	46,59	0	19,02	38,04	15
	MC	-	-	15,38	7,14	0
2012	1C	95,33	-	78,57	63,16	83,33
	2C	85,14	100	73,33	72	60
	MC	-	-	-	-	-

Fonte: Autoria Própria.

Em algumas provas do ano de 2005, as avaliações ficaram próximas de zero com o ExamClipper. Isso ocorreu pela maneira que o *layout* da prova é formatado, mais especificamente para as provas que pertencem a formação geral 2, listadas na Seção 5.4. A Figura 5.6 mostra a proximidade do término da questão 8 com o início da questão 9. Essa falta de espaçamento dificulta para algumas ferramentas a identificação automática de duas questões, neste caso. Isso de fato ocorreu usando a ferramenta ExamClipper.

[17] https://github.com/karinawie/PDFExtraction/tree/master/extractions/aletheia

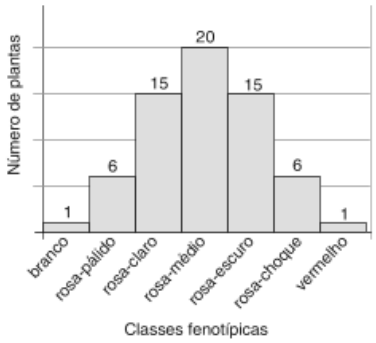
Figura 5.6 – *Layout* de questões da prova de biologia do ano de 2005.

Atenção: Responda às questões de números 8 e 9, somente se sua área de formação for Bacharelado.

BACHARELADO

Questão 8

Um floricultor cultiva uma espécie de planta diplóide que produz flores cujas cores variam do branco ao vermelho. O cruzamento de duas linhagens puras, uma com flores brancas e outra com flores vermelhas, originou indivíduos da geração F_1 que, cruzados entre si, geraram, em F_2, o resultado esquematizado no gráfico.



a. A cor da flor nessa espécie de planta segue que tipo de padrão de herança? Justifique. **(valor: 5,0 pontos)**

b. Qual o número provável de pares de genes envolvidos na cor da flor? **(valor: 2,5 pontos)**

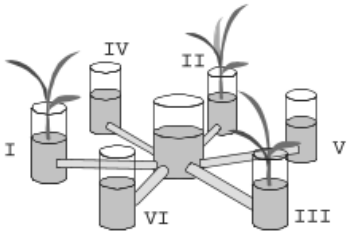
c. Sabendo-se que as plantas de tonalidade intermediária (rosa-médio) são as de maior valor comercial, que tipo de cruzamento o floricultor deve fazer para obter a maior proporção possível de flores dessa tonalidade? **(valor: 2,5 pontos)**

Questão 9

Considere o experimento abaixo esquematizado:

Figura 1

Nenhum



- Preparou-se um sistema com 7 potes cheios de areia umedecida, sendo um deles central e interligado aos demais por braços também cheios de areia (Figura 1).

Fonte: Autoria Própria.

5.6.4 Conclusão dos Resultados das Extrações

Os resultados das extrações foram prezados nas identificações automáticas que as ferramentas possibilitam sem interferência manual de especialistas humanos. Questões com partes muito afastadas das outras dificultam a identificação automática de toda a questão, que fica separada em blocos, mesmo que a ferramenta não seja específica para extrações de questões.

Observou-se que essa formatação mais “espaçosa” ocorre nos primeiros anos de aplicações das provas, entre 2004 e 2009. Várias questões foram desconsideradas de “extração correta”, pelo fato de uma das linhas não estarem na ordem de leitura correta.

É notável que, se as provas utilizassem um *layout* padrão para todos os cursos em todos os anos, as extrações seriam mais eficientes, ao menos utilizando a ferramenta CyberPDF, que se vale das coordenadas como padrão para os demais arquivos. No entanto, é perceptível que as ferramentas de extrações para dados textuais apresentam uma dificuldade em identificar questões com duas colunas e colunas mistas.

6 CONCLUSÃO

A ampla utilização de arquivos PDF promoveu pesquisas na análise de seu *layout* para fins de extração dessas informações. Este trabalho, propôs extrair automaticamente dados de provas em formato PDF natos digitais disponibilizados em um repositório de exames educacionais, considerando o *layout* do documento. Essas informações coletadas podem ser amplamente utilizadas para diversas áreas de ensino, fornecendo informações úteis e oportunas para estudantes, professores, administradores e pesquisadores.

Este trabalho avaliou o desempenho de cinco ferramentas de extração de dados em arquivos PDF: duas para dados textuais; duas em dados tabulares; e duas que extraem ROI. Os arquivos utilizados para a avaliação compreendem 343 provas do Enade, com 11.196 questões objetivas e discursivas dispostas em 6.834 páginas, e contemplando todos os 396 gabaritos com 14.475 alternativas extraídas das questões objetivas.

No decorrer do trabalho a seguinte pergunta de pesquisa foi respondida:

1. Como analisar o desempenho de ferramentas para a extração de dados de provas em PDF, aplicando em um conjunto de provas?
   Verificando as provas de cada ano para identificar características divergentes de uma prova para outra. Essas características devem ser anotadas para que, com base nisso, as métricas sejam definidas. Para esse trabalho as métricas utilizadas estão apresentadas na Seção 5.2. Além das métricas, é preciso definir qual medida aplicar nas extrações dos dados.

De acordo com as configurações utilizadas nas ferramentas para este trabalho, relatadas na Seção 5.4, foi possível avaliar as ferramentas de extrações. Com base nos dados extraídos, conclui-se que a ferramenta Excalibur reconhece mais tabelas comparado às utilizações do Tabula, no entanto, demora alguns segundos a mais para a extração. O PDFMiner consegue identificar automaticamente várias questões em todas as métricas definidas, enquanto o CyberPDF não identifica automaticamente questões que começam em uma página/coluna e terminam em outra e que estejam em duas colunas. A ferramenta PDFMiner também extrai mais rapidamente. As extrações de ROI utilizadas com o Aletheia ficaram todas em 100%. Esses resultados não são os ideais, visto que para isso foi utilizado um viés configurado na construção do *ground truth*, mesmo assim, foi possível obter todas as extrações das questões utilizadas nesta pesquisa. O ExamClipper oferece a opção de ajustar manualmente as regiões, demorando mais tempo. Se isso tivesse sido aplicado as extrações também ficariam em 100%, porém não seria automatizado.

Considera-se que comparar o tempo das extrações é importante, pois com isso é possível identificar qual ferramenta é mais eficiente em cada métrica e qual extrai maior quantidade em menos tempo. Esses resultados podem sofrer mudanças dentro de certos

limites, por exemplo, ajustando manualmente algumas identificações que as ferramentas selecionam, alterando configurações de entrada, entre outros.

6.1 TRABALHOS FUTUROS

Visto que é possível extrair os dados nesse tipo de exame, estima-se que o processo possa ser aplicado a outras provas, apenas reutilizando as ferramentas utilizadas neste trabalho.

O conjunto do *ground truth* ficará disponível ao público para fins de pesquisa acadêmica, pois trabalhos semelhantes utilizam esses tipos de conjuntos prontos para outras pesquisas. Propõe-se, ainda, disponibilizar um sistema com as extrações em bancos de questões. Para isso, podem ser utilizadas as extrações do Aletheia, também disponível no GitHub [1].

6.2 PUBLICAÇÕES

Durante a execução da pesquisa para esta dissertação, foram obtidas algumas contribuições:

- Publicação e apresentação de um trabalho Qualis B3 (previsão novo Qualis A4), 40 ° Congresso da Sociedade Brasileira de Computação (CSBC) 28 ° WEI Workshop sobre Educação em Computação, intitulado como “Explorando Resultados por Questão no Enade em Ciência da Computação para Subsidiar Revisão de Projeto Pedagógico de Curso”.
- Submissão aceita de um trabalho Qualis B2, 23rd International Conference on Enterprise Information Systems (ICEIS), intitulado como “Automated Data Extraction from PDF Documents: Application to Large Sets of Educational Tests”, aguardando a apresentação no momento da conclusão desta dissertação.

[1] Disponível em: https://github.com/karinawie/PDFExtraction/tree/master/extractions/aletheia

REFERÊNCIAS BIBLIOGRÁFICAS

Alaei, A.; Nagabhushan, P.; Pal, U. A benchmark kannada handwritten document dataset and its segmentation. In: **2011 International Conference on Document Analysis and Recognition**. [S.l.: s.n.], 2011. p. 141–145.

Antonacopoulos, A. et al. A realistic dataset for performance evaluation of document layout analysis. In: **2009 10th International Conference on Document Analysis and Recognition**. [S.l.: s.n.], 2009. p. 296–300. ISSN 2379-2140.

ANTONACOPOULOS, A.; KARATZAS, D.; BRIDSON, D. Ground truth for layout analysis performance evaluation. In: . [S.l.: s.n.], 2006. v. 3872, p. 302–311. ISBN 978-3-540-32140-8.

ANTONACOPOULOS, A.; MENG, H. A ground-truthing tool for layout analysis performance evaluation. In: **Proceedings of the 5th International Workshop on Document Analysis Systems V**. Berlin, Heidelberg: Springer-Verlag, 2002. (DAS 02), p. 236244. ISBN 3540440682.

Bast, H.; Korzen, C. A Benchmark and Evaluation for Text Extraction from PDF. In: **2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)**. [S.l.: s.n.], 2017. p. 1–10.

BEEL, J. et al. Docears PDF Inspector: Title Extraction from PDF Files. In: **Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries**. New York, NY, USA: Association for Computing Machinery, 2013. (JCDL 13), p. 443444. ISBN 9781450320771. Disponível em: <https://doi.org/10.1145/2467696.2467789>.

BERG Øyvind R. **High precision text extraction from PDF documents**. 2011. Tese (Thesis en Informatics) — UNIVERSITY OF OSLO, 2011.

BUDHIRAJA, S. S. **Extracting Specific Text From Documents Using Machine Learning Algorithms**. 2018. Tese (Thesis of Computer Science) — Lakehead University, Canada, 2018.

BUI, D. D. A. et al. Extractive text summarization system to aid data extraction from full text in systematic review development. **Journal of Biomedical Informatics**, v. 64, p. 265 – 272, 2016. ISSN 1532-0464. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1532046416301514>.

CHANDARANA, M. K. J. A review of optical character recognition. **International Journal of Engineering Research & Technology (IJERT)**, v. 02, p. 219–223, 12 2014. ISSN 2278-0181.

CHARãO, A. et al. Explorando resultados por questão no enade em ciência da computação para subsidiar revisão de projeto pedagógico de curso. In: **Anais do XXVIII Workshop sobre Educação em Computação**. Porto Alegre, RS, Brasil: SBC, 2020. p. 16–20. ISSN 2595-6175. Disponível em: <https://sol.sbc.org.br/index.php/wei/article/view/11121>.

Choudhury, S. R. et al. Figure metadata extraction from digital documents. In: **2013 12th International Conference on Document Analysis and Recognition**. [S.l.: s.n.], 2013. p. 135–139. ISSN 2379-2140.

Clark, C.; Divvala, S. Pdffigures 2.0: Mining figures from research papers. In: **2016 IEEE/ACM Joint Conference on Digital Libraries (JCDL)**. [S.l.: s.n.], 2016. p. 143–152.

Clausner, C.; Antonacopoulos, A.; Pletschacher, S. Icdar2017 competition on recognition of documents with complex layouts - rdcl2017. In: **2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)**. [S.l.: s.n.], 2017. v. 01, p. 1404–1410.

CONSTANTIN, A.; PETTIFER, S.; VORONKOV, A. PDFX: Fully-Automated PDF-to-XML Conversion of Scientific Literature. In: **Proceedings of the 2013 ACM Symposium on Document Engineering**. New York, NY, USA: Association for Computing Machinery, 2013. (DocEng 13), p. 177180. ISBN 9781450317894. Disponível em: <https://doi.org/10.1145/2494266.2494271>.

CRUZ, E.; MACHADO, R.-J.; SANTOS, M. On the rim between business processes and software systems. In: ____. [S.l.: s.n.], 2019. p. 170. ISBN ISBN13: 9781522572718|ISBN10: 1522572716|EISBN13: 9781522572725.

DENNY, P. et al. Peerwise: Students sharing their multiple choice questions. In: **Proceedings of the Fourth International Workshop on Computing Education Research**. New York, NY, USA: Association for Computing Machinery, 2008. (ICER 08), p. 5158. ISBN 9781605582160. Disponível em: <https://doi-org.ez47.periodicos.capes.gov.br/10.1145/1404520.1404526>.

DOERMANN David; ZOTKINA Elena; LI Huiping. GEDI - A Groundtruthing Environment for Document Images. In: **Ninth IAPR International Workshop on Document Analysis Systems (DAS 2010)**. [S.l.: s.n.], 2010. Submitted.

DURETEC, K.; RAUBER, A.; BECKER, C. A text extraction software benchmark based on a synthesized dataset. In: **Proceedings of the 17th ACM/IEEE Joint Conference on Digital Libraries**. [S.l.]: IEEE Press, 2017. (JCDL '17), p. 109118. ISBN 9781538638613.

EXCALIBUR. **Excalibur: PDF Table Extraction for Humans**. 2018. Acesso em 29 de novembro de 2020. Disponível em: <https://excalibur-py.readthedocs.io/en/master/>.

FAN, M.; KIM, D. S. Detecting Table Region in PDF Documents Using Distant Supervision. **arXiv: Computer Vision and Pattern Recognition**, 2015.

Fang, J. et al. Dataset, ground-truth and performance metrics for table detection evaluation. In: **2012 10th IAPR International Workshop on Document Analysis Systems**. [S.l.: s.n.], 2012. p. 445–449.

Fang Yuan; Bo Lu. A new method of information extraction from PDF files. In: **2005 International Conference on Machine Learning and Cybernetics**. [S.l.: s.n.], 2005. v. 3, p. 1738–1742 Vol. 3.

gov.br. **Exame Nacional de Desempenho de Estudantes (Enade)**. 2021. Acesso em 15 de janeiro de 2021. Disponível em: <https://www.gov.br/inep/pt-br/acesso-a-informacao/perguntas-frequentes/exame-nacional-de-desempenho-dos-estudantes-enade>.

____. **Exame Nacional de Desempenho dos Estudantes (Enade)**. 2021. Acesso em 16 de janeiro de 2021. Disponível em: <https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enade>.

____. **Site do Inep passa a integrar o portal único do Governo Federal e pode ser acessado em gov.br/inep**. 2021. Acesso em 15 de janeiro de 2021. Disponível em: <http://portal.inep.gov.br/artigo/-/asset_publisher/B4AQV9zFY7Bv/content/id/6950005>.

Hadjar, K. et al. Xed: a new tool for extracting hidden structures from electronic documents. In: **First International Workshop on Document Image Analysis for Libraries, 2004. Proceedings.** [S.l.: s.n.], 2004. p. 212–224.

HASSAN, T. Object-Level Document Analysis of PDF Files. In: **Proceedings of the 9th ACM Symposium on Document Engineering**. New York, NY, USA: Association for Computing Machinery, 2009. (DocEng 09), p. 4755. ISBN 9781605585758. Disponível em: <https://doi.org/10.1145/1600193.1600206>.

Hassan, T.; Baumgartner, R. Table recognition and understanding from pdf files. In: **Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)**. [S.l.: s.n.], 2007. v. 2, p. 1143–1147.

He, D. et al. Multi-scale multi-task fcn for semantic page segmentation and table detection. In: **2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)**. [S.l.: s.n.], 2017. v. 01, p. 254–261.

INEP. **Exame Nacional de Desempenho dos Estudantes (Enade)**. 2020. Acesso em 07 outubro 2020. Disponível em: <http://portal.inep.gov.br/enade>.

____. **Inep adia aplicação do Enade 2020**. 2021. Acesso em 27 de março de 2021. Disponível em: <http://portal.inep.gov.br/artigo/-/asset_publisher/B4AQV9zFY7Bv/content/inep-adia-aplicacao-do-enade-2020/21206>.

ISO32000-2:2017. **Document management Portable document format Part 2: PDF 2.0**. 2017. Acesso em 29 de novembro de 2020. Disponível em: <https://www.iso.org/standard/63534.html>.

JIANG, D.; YANG, X. Converting PDF to HTML Approach Based on Text Detection. In: **Proceedings of the 2nd International Conference on Interaction Sciences: Information Technology, Culture and Human**. New York, NY, USA: Association for Computing Machinery, 2009. (ICIS 09), p. 982985. ISBN 9781605587103. Disponível em: <https://doi.org/10.1145/1655925.1656103>.

KONDERMANN, D. Ground truth design principles: An overview. In: **Proceedings of the International Workshop on Video and Image Ground Truth in Computer Vision Applications**. New York, NY, USA: Association for Computing Machinery, 2013. (VIGTA 13). ISBN 9781450321693. Disponível em: <https://doi.org/10.1145/2501105.2501114>.

KRIG, S. Ground Truth Data, Content, Metrics, and Analysis. In: ____. **Computer Vision Metrics: Survey, Taxonomy, and Analysis**. Berkeley, CA: Apress, 2014. p. 283–311. ISBN 978-1-4302-5930-5. Disponível em: <https://doi.org/10.1007/978-1-4302-5930-5_7>.

LI, P.; JIANG, X.; SHATKAY, H. Extracting figures and captions from scientific publications. In: **Proceedings of the 27th ACM International Conference on Information and Knowledge Management**. New York, NY, USA: Association for Computing Machinery, 2018. (CIKM 18), p. 15951598. ISBN 9781450360142. Disponível em: <https://doi.org/10.1145/3269206.3269265>.

LIMA, R.; CRUZ, E. F. Extraction and multidimensional analysis of data from unstructured data sources: A case study. In: **ICEIS**. [S.l.: s.n.], 2019.

LIPINSKI, M. et al. Evaluation of header metadata extraction approaches and tools for scientific pdf documents. In: . New York, NY, USA: Association for Computing Machinery, 2013. (JCDL '13), p. 385386. ISBN 9781450320771. Disponível em: <https://doi.org/10.1145/2467696.2467753>.

LIU, Y. et al. Tableseer: Automatic table metadata extraction and searching in digital libraries. In: **Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries**. New York, NY, USA: Association for Computing Machinery, 2007. (JCDL 07), p. 91100. ISBN 9781595936448. Disponível em: <https://doi.org/10.1145/1255175.1255193>.

Manuel Aristarán, Mike Tigas, Jeremy B. Merrill, Jason Das, David Frackman and Travis Swicegood. **Tabula is a tool for liberating data tables locked inside PDF files**. 2018. Acesso em 20 de julho de 2020. Disponível em: <https://tabula.technology/>.

Oro, E.; Ruffolo, M. Xonto: An ontology-based system for semantic information extraction from pdf documents. In: **2008 20th IEEE International Conference on Tools with Artificial Intelligence**. [S.l.: s.n.], 2008. v. 1, p. 118–125.

Parizi, R. M. et al. 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)**. [S.l.: s.n.], 2018. p. 106–111.

PATEL, C.; PATEL, A.; PATEL, D. Optical character recognition by open source ocr tool tesseract: A case study. **International Journal of Computer Applications**, v. 55, p. 50–56, 10 2012.

REUL, C.; SPRINGMANN, U.; PUPPE, F. Larex: A semi-automatic open-source tool for layout analysis and region extraction on early printed books. In: . [S.l.: s.n.], 2017. p. 137–142.

SASIREKHA, D.; CHANDRA, E. Article: Text extraction from pdf document. **IJCA Proceedings on Amrita International Conference of Women in Computing - 2013**, AICWIC, n. 3, p. 17–19, January 2013. Full text available.

SILVA, A. Costa e; JORGE, A.; TORGO, L. Design of an end-to-end method to extract information from tables. **Document Analysis and Recognition**, v. 8, p. 144–171, 01 2006.

Strecker, T. et al. Automated ground truth data generation for newspaper document images. In: **2009 10th International Conference on Document Analysis and Recognition**. [S.l.: s.n.], 2009. p. 1275–1279.

Tao, X. et al. Ground-truth and performance evaluation for page layout analysis of born-digital documents. In: **2014 11th IAPR International Workshop on Document Analysis Systems**. [S.l.: s.n.], 2014. p. 247–251.

The Apache Software Foundation. **Apache PDFBox - A Java PDF Library**. 2020. Acesso em 20 maio 2020. Disponível em: <https://pdfbox.apache.org/>.

TKACZYK, D.; SZOSTEK, P.; BOLIKOWSKI, . GROTOAP2 - The Methodology of Creating a Large Ground Truth Dataset of Scientific Articles. **D-Lib Magazine**, v. 20, 11 2014.

TKACZYK, D. et al. Cermine: automatic extraction of structured metadata from scientific literature. **International Journal on Document Analysis and Recognition (IJDAR)**, Springer Berlin Heidelberg, v. 18, n. 4, p. 317–335, 2015. ISSN 1433-2833. Disponível em: <http://dx.doi.org/10.1007/s10032-015-0249-8>.

TRAN, T. A.; NA, I. S.; KIM, S. H. Page segmentation using minimum homogeneity algorithm and adaptive mathematical morphology. **Int. J. Doc. Anal. Recognit.**, Springer-Verlag, Berlin, Heidelberg, v. 19, n. 3, p. 191209, set. 2016. ISSN 1433-2833. Disponível em: <https://doi.org/10.1007/s10032-016-0265-3>.

WU, J. et al. Pdfmef: A multi-entity knowledge extraction framework for scholarly documents and semantic search. In: **Proceedings of the 8th International Conference on Knowledge Capture**. New York, NY, USA: Association for Computing Machinery, 2015. (K-CAP 2015). ISBN 9781450338493. Disponível em: <https://doi.org/10.1145/2815833.2815834>.

YILDIZ, B.; KAISER, K.; MIKSCH, S. pdf2table: A Method to Extract Table Information from PDF Files. In: . [S.l.: s.n.], 2005. p. 1773–1785.

YUAN, F.; LIU, B.; YU, G. A study on information extraction from pdf files. In: **Proceedings of the 4th International Conference on Advances in Machine Learning and Cybernetics**. Berlin, Heidelberg: Springer-Verlag, 2005. (ICMLC05), p. 258267. ISBN 3540335846. Disponível em: <https://doi.org/10.1007/11739685_27>.

Yusuke Shinyama. **Python PDF parser and analyzer**. 2014. Acesso em 21 maio 2020. Disponível em: <http://www.unixuser.org/~euske/python/pdfminer/>.

APÊNDICE A – CURSOS AVALIADOS NO ENADE ENTRE 2004-2019

A seguir, a listagem apresenta os cursos avaliados de acordo com o ano da aplicação da prova:

2004 Agronomia, Educação Física, Enfermagem, Farmácia, Fisioterapia, Fonoaudiologia, Medicina, Medicina Veterinária, Nutrição, Odontologia, Serviço Social, Terapia Ocupacional e Zootecnia.

2005 Arquitetura e Urbanismo, Biologia, Ciências Sociais, Computação, Engenharias, Filosofia, Física, Geografia, História, Letras, Matemática, Pedagogia e Química.

2006 Administração, Arquivologia, Biblioteconomia, Biomedicina, Ciências Contábeis, Ciências Econômicas, Comunicação Social, Design, Direito, Formação de Professores, Música, Psicologia, Secretariado Executivo, Teatro e Turismo.

2007 Agronomia, Biomedicina, Educação Física, Enfermagem, Farmácia, Fisioterapia, Medicina, Medicina Veterinária, Nutrição, Odontologia, Serviço Social, Tecnologia Agroindústria, Tecnologia Radiologia, Terapia Ocupacional e Zootecnia.

2008 Arquitetura e Urbanismo, Biologia, Ciências Sociais, Computação, Engenharias, Filosofia, Física, Geografia, História, Letras, Matemática, Pedagogia, Química, Tecnologia em Alimentos, Tecnologia em Análise e Desenvolvimento de Sistemas, Tecnologia em Automação Industrial, Tecnologia em Construção de Edifícios, Tecnologia em Fabricação Mecânica, Tecnologia em Gestão e Produção Industrial, Tecnologia em Manutenção Industrial, Tecnologia em Processos Químicos, Tecnologia em Redes de Computadores e Tecnologia em Saneamento Ambiental.

2009 Administração, Arquivologia, Biblioteconomia, Ciências Contábeis, Ciências Econômicas, Comunicação Social, Design, Direito, Estatística, Música, Psicologia, Relações Internacionais, Secretariado Executivo, Teatro, Tecnologia em Design de Moda, Tecnologia em Gastronomia, Tecnologia em Gestão de Recursos Humanos, Tecnologia em Gestão de Turismo, Tecnologia em Gestão Financeira, Tecnologia em Marketing, Tecnologia em Processos Gerenciais e Turismo.

2010 Agronomia, Biomedicina, Educação Física, Enfermagem, Farmácia, Fisioterapia, Fonoaudiologia, Medicina, Medicina Veterinária, Nutrição, Odontologia, Serviço Social, Tecnologia em Agroindústria, Tecnologia em Agronegócio, Tecnologia em Gestão Ambiental, Tecnologia em Gestão Hospitalar, Tecnologia Radiologia, Terapia Ocupacional e Zootecnia.

2011 Arquitetura e Urbanismo, Artes Visuais, Biologia, Ciências Sociais, Computação, Educação Física, Engenharias, Filosofia, Física, Geografia, História, Letras, Matemática, Música, Pedagogia, Química, Tecnologia em Alimentos, Tecnologia em Análise e Desenvolvimento de Sistemas, Tecnologia em Automação Industrial, Tecnologia em Construção de Edifícios, Tecnologia em Fabricação Mecânica, Tecnologia em Gestão e Produção Industrial, Tecnologia em Manutenção Industrial, Tecnologia em Processos Químicos, Tecnologia em Redes de Computadores e Tecnologia em Saneamento Ambiental.

2012 Administração, Ciências Contábeis, Ciências Econômicas, Comunicação Social Jornalismo, Comunicação Social Publicidade e Propaganda, Design, Direito, Psicologia, Relações Internacionais, Secretariado Executivo, Tecnologia em Gestão Comercial, Tecnologia em Gestão de Recursos Humanos, Tecnologia em Gestão Financeira, Tecnologia em Gestão Logística, Tecnologia em Marketing, Tecnologia em Processos Gerenciais e Turismo.

2013 Agronomia, Biomedicina, Educação Física, Enfermagem, Farmácia, Fisioterapia, Fonoaudiologia, Medicina, Medicina Veterinária, Nutrição, Odontologia, Serviço Social, Tecnologia em Agronegócio, Tecnologia em Gestão Ambiental, Tecnologia em Gestão Hospitalar, Tecnologia Radiologia e Zootecnia.

2014 Arquitetura e Urbanismo, Artes Visuais, Ciência da Computação, Ciências Biológicas, Educação Física, Engenharias (Alimentos, Ambiental, Civil, Computação, Controle e Automação, Elétrica, Florestal, Mecânica, Química, Produção), Filosofia, Física, Geografia, História, Letras (Português, Espanhol e Inglês), Matemática, Música, Pedagogia, Química, Sistemas de Informação, Tecnologia em Análise e Desenvolvimento de Sistemas, Tecnologia em Automação Industrial, Tecnologia em Produção Industrial e Tecnologia em Redes de Computadores.

2015 Administração, Administração Pública, Ciências Contábeis, Ciências Econômicas, Comunicação Social Jornalismo, Comunicação Social Publicidade e Propaganda, Design, Direito, Psicologia, Relações Internacionais, Secretariado Executivo, Tecnologia em Comércio Exterior, Tecnologia em Design de Interiores, Tecnologia em Design de Moda, Tecnologia em Design Gráfico, Tecnologia em Gastronomia, Tecnologia em Gestão Comercial, Tecnologia em Gestão de Qualidade, Tecnologia em Gestão de Recursos Humanos, Tecnologia em Gestão Financeira, Tecnologia em Gestão Pública, Tecnologia em Logística, Tecnologia em Marketing, Tecnologia em Processos Gerenciais, Teologia e Turismo.

2016 Agronomia, Biomedicina, Educação Física, Enfermagem, Farmácia, Fisioterapia, Fonoaudiologia, Medicina, Medicina Veterinária, Nutrição, Odontologia, Serviço Social, Tecnologia em Agronegócio, Tecnologia em Estética e Cosmética, Tecnologia em

Gestão Ambiental, Tecnologia em Gestão Hospitalar, Tecnologia Radiologia e Zootecnia.

2017 Arquitetura e Urbanismo, Artes Visuais, Ciência da Computação, Ciências Biológicas, Ciências Sociais, Educação Física, Engenharias (Alimentos, Ambiental, Civil, Computação, Controle e Automação, Elétrica, Florestal, Mecânica, Química, Produção), Filosofia, Física, Geografia, História, Letras (Espanhol, Inglês e Português), Matemática, Música, Pedagogia, Química, Sistemas de Informação, Tecnologia em Análise e Desenvolvimento de Sistemas, Tecnologia em Automação Industrial, Tecnologia em Gestão de Tecnologia da Informação, Tecnologia em Produção Industrial e Tecnologia em Redes de Computadores.

2018 Administração, Administração Pública, Ciências Contábeis, Ciências Econômicas, Comunicação Social Jornalismo, Comunicação Social Publicidade e Propaganda, Design, Direito, Psicologia, Relações Internacionais, Secretariado Executivo, Serviço Social, Tecnologia em Comércio Exterior, Tecnologia em Design de Interiores, Tecnologia em Design de Moda, Tecnologia em Design Gráfico, Tecnologia em Gastronomia, Tecnologia em Gestão Comercial, Tecnologia em Gestão de Qualidade, Tecnologia em Gestão de Recursos Humanos, Tecnologia em Gestão Financeira, Tecnologia em Gestão Pública, Tecnologia em Logística, Tecnologia em Marketing, Tecnologia em Processos Gerenciais, Teologia e Turismo.

2019 Agronomia, Arquitetura e Urbanismo, Biomedicina, Educação Física, Enfermagem, Engenharias (Alimentos, Ambiental, Civil, Computação, Controle e Automação, Elétrica, Florestal, Mecânica, Química, Produção), Farmácia, Fisioterapia, Fonoaudiologia, Medicina, Medicina Veterinária, Nutrição, Odontologia, Tecnologia em Agronegócio, Tecnologia em Estética e Cosmética, Tecnologia em Gestão Ambiental, Tecnologia em Gestão Hospitalar, Tecnologia Radiologia, Tecnologia em Segurança do Trabalho e Zootecnia.

2020 O Inep adiou para 2021 a aplicação do Enade de 2020. O motivo são as restrições impostas devido à pandemia de Covid-19, com impacto no cronograma de aulas das instituições de ensino superior em todo o país. De acordo com o presidente do Inep, Alexandre Lopes, a nova data será redefinida conforme os ajustes dos calendários acadêmicos (INEP, 2021).