UNIVERSIDADE FEDERAL DE PERNAMBUCO

CENTRO DE INFORMÁTICA

SISTEMAS DE INFORMAÇÃO

KAYQUE LUCAS SANTANA DOS SANTOS

DESENVOLVIMENTO DE UM SISTEMA DE BANCO DE QUESTÕES AUTOMÁTICO UTILIZANDO AS PROVAS DO ENEM

Recife

2022

KAYQUE LUCAS SANTANA DOS SANTOS

DESENVOLVIMENTO DE UM SISTEMA DE BANCO DE QUESTÕES AUTOMÁTICO UTILIZANDO AS PROVAS DO ENEM

Trabalho apresentado ao Programa de Graduação em Sistemas de Informação do Centro de Informática da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Bacharel em Sistemas de Informação.

Orientador: Alex Sandro Gomes

Recife

2022

KAYQUE LUCAS SANTANA DOS SANTOS

DESENVOLVIMENTO DE UM SISTEMA DE BANCO DE QUESTÕES AUTOMÁTICO UTILIZANDO AS PROVAS DO ENEM

Trabalho apresentado ao Programa de Graduação em Sistemas de Informação do Centro de Informática da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Bacharel em Sistemas de Informação.

Recife, 19 de maio de 2022.

BANCA EXAMINADORA

_____

Prof. Alex Sandro Gomes (Orientador)
UNIVERSIDADE FEDERAL DE PERNAMBUCO

_____

Prof. Filipe Carlos de Albuquerque Calegário (2º membro da banca)
UNIVERSIDADE FEDERAL DE PERNAMBUCO

AGRADECIMENTOS

Gostaria de agradecer a todas as pessoas que tiveram participação em minha jornada no curso de Sistemas de Informação.

Inicialmente a meus familiares, especialmente à minha maior incentivadora e apoiadora, a minha mãe Ana Paula. Às minhas duas irmãs, Camilla Tainá e Anna Cecília, que sempre se fizeram presentes nos momentos mais desafiadores. À minha tia Cintia que representa um papel importantíssimo em minha vida e à minha avó Neurielide que é uma imensa fonte de inspiração para mim.

Aos meus amigos de longa data que me acompanham desde o início da minha vida acadêmica e sempre se alegraram com minhas conquistas: Ana Beatriz, Andressa, Bianka, Jéssica, João Gabriel, Luana e Tainá. Aos amigos que fiz durante o curso, especialmente Lucas Burgos e Talyta Pacheco, cuja parceria e companheirismo foram facilitadores do meu desenvolvimento, me permitindo crescer pessoal e profissionalmente.

A todos os professores que tive a oportunidade de conhecer no Centro de Informática, especialmente o Prof. Fernando Neto. Aos colegas Carlos José e Leandro Marques que tiveram significativa colaboração neste projeto. E, por fim, ao meu orientador, Prof. Alex Sandro Gomes, cuja atenção, sensibilidade e experiência foram essenciais para a execução e o sucesso deste trabalho.

“Tente mover o mundo. O primeiro passo será mover a si mesmo”.

Platão

Resumo

A elaboração de exercícios e avaliações é uma atividade que faz parte da rotina dos professores da educação básica. Para preparar alunos para provas de vestibular, os professores do ensino médio se utilizam das pesquisas em apostilas, sites e blogs para encontrar questões provenientes dos vestibulares oficiais. Tal atividade gera uma demanda extra de tempo para estes profissionais. Este trabalho propõe o desenvolvimento de um sistema que permitirá aos professores otimizar o tempo gasto na busca por questões de provas de vestibulares na internet, especificamente as provas do ENEM. O sistema proposto utiliza a técnica de *Web Scraping* para coleta automática das provas e faz uma análise de texto baseada em expressões regulares para estruturar os dados contidos nos documentos, persistindo-os em um banco de dados. Para mensurar a efetividade e a qualidade do modelo proposto, a aplicação foi testada por 18 usuários, contendo profissionais da educação e da tecnologia da informação. Cada participante, após interagir com o sistema, respondeu a um questionário no modelo SUS (*System Usability Scale*, do inglês Escala de Usabilidade do Sistema) e contabilizou a quantidade de erros observados. Os resultados demonstraram uma aceitação acima da média pelos participantes e uma taxa de 68% de questões extraídas sem nenhum erro, sobre o conjunto de questões avaliadas.

Palavras-chave: Extração Automatizada de Dados, Recuperação da Informação, Tecnologias Educacionais

Abstract

The elaboration of exercises and evaluations is an activity that is part of the routine of basic education teachers. To prepare students for entrance exams, high school teachers use research in handouts, websites and blogs to find questions from official entrance exams. Such activity generates an extra demand of time for these professionals. This work proposes the development of a system that allows teachers to optimize the time spent searching for entrance exam questions on the internet, specifically ENEM exams. The proposed system uses the Web Scraping technique for automatic collection of entrance exams on the internet and performs a text analysis based on regular expressions to structure the data contained in the documents, persisting them in a database. To measure the effectiveness and quality of the proposed model, the application got tested by users. After interacting with the system, each participant answered a survey using the SUS model (System Usability Scale) and counted the number of errors observed. The results showed an above average acceptance by the participants and a 68% rate of questions that were extracted with no error, from the evaluated questions set.

Keywords: Automated Data Extraction, Information Retrieval, Educational Technologies

Lista de Ilustrações

Lista de Tabelas

Lista de Abreviaturas e Siglas

TIC	Tecnologia(s) da informação e Comunicação
CETIC	Centro Regional de Estudos para o Desenvolvimento da Sociedade da Informação
ENEM	Exame Nacional do Ensino Médio
PROUNI	Programa Universidade para Todos
SISU	Sistema de Seleção Unificada
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
PDF	Portable Document Format
HTML	Linguagem de Marcação de Hipertexto
SaaS	Software as a Service
NIST	National Institute of Standards and Technology
SUS	System Usability Scale
BPMN	Business Process Model and Notation
ENADE	Exame Nacional de Desempenho dos Estudantes

SUMÁRIO

1 INTRODUÇÃO

O planejamento é uma atividade pertencente à prática docente, cuja realização é prevista pela Lei de Diretrizes e Bases da Educação Nacional (BRASIL, 1996). A lei determina que os sistemas de ensino devem assegurar período reservado para tal atividade, incluído na carga horária de trabalho. Complementando a norma geral, a lei do estado de Pernambuco, que rege a carga horária de pessoal do Grupo Ocupacional Magistério, prevê a realização de ações que "se destinam ao desenvolvimento de atividades escolares e extra classe, à correção de provas, estudos, trabalhos escolares, preparação de aulas e outras atividades correlatas". (PERNAMBUCO, 1989). O termo aulas-atividade é utilizado para denominar essas ações. A lei determina ainda que, do total de aulas-atividade, "metade será obrigatoriamente cumprida pelo professor no recinto escolar". O local de realização da outra metade não é especificado, ou seja, é de livre escolha para o professor. Segundo o Estatuto do Magistério de Pernambuco, que dispõe sobre o Magistério Público de Pré-Escolar, Ensino Fundamental e Ensino Médio do Estado de Pernambuco:

> A hora-aula atividade compreende as ações de preparação, acompanhamento e avaliação de prática pedagógica e inclui:
> a) elaboração de planos de atividades curriculares, provas e correção de trabalhos escolares;
> b) participação em eventos, reflexão da prática pedagógica, estudos, debates, avaliações, pesquisas e trocas de experiências;
> c) aprofundamento da formação docente;
> d) participação em reuniões de pais e mestres e da comunidade escolar;
> e) atendimento pedagógico a alunos e pais.
>
> (PERNAMBUCO, 1996)

Conclui-se, portanto, que o planejamento e preparação de instrumentos pedagógicos são parte integrante e contínua do trabalho do professor. Dados da pesquisa TIC Educação

2020, do CETIC, sobre o uso das Tecnologias de Informação e Comunicação nas escolas brasileiras, apontam os desafios enfrentados pelas instituições de ensino, docentes, pais e professores durante a pandemia COVID-19. Dentre os principais está o aumento da carga de trabalho dos professores, percebido por 73% das instituições educacionais. Essa foi a questão mencionada com mais frequência pelos gestores de escolas estaduais, 83%, e a terceira mais mencionada pelos gestores de escolas particulares, 67% (CETIC.BR, 2020). No ano anterior, a pesquisa de mesma origem apontou que a pressão ou falta de tempo para cumprir com o conteúdo previsto foi uma barreira percebida por 79% dos professores de escolas públicas urbanas para o uso das TIC na escola (CETIC.BR, 2019).

Santos e Sobrinho (2011) constatam que a redução da carga de trabalho do professor é uma medida imprescindível para o fomento da melhoria de condições de trabalho deste profissional. Os autores revelam que a maioria costuma extrapolar sua carga horária máxima exigida ao trabalhar em casa com planejamentos, correções de provas e atividades. Constata-se, portanto, uma dificuldade dos professores no gerenciamento de seu tempo de trabalho. Isso cria a necessidade de ferramentas que otimizem o tempo de realização de suas atividades de planejamento e preparação de instrumentos pedagógicos.

Ainda segundo os dados da pesquisa TIC Educação 2020, ao elencarem os desafios para a continuidade das atividades educacionais durante o período da pandemia COVID-19, 61% dos gestores escolares citaram a falta de habilidade dos professores para utilizar recursos de tecnologia em atividades pedagógicas (CETIC.BR, 2020). O distanciamento social levou a comunidade escolar a depender mais de dispositivos tecnológicos e das redes de telecomunicações, sobretudo a internet, para o cumprimento dos objetivos de ensino-aprendizagem. A situação resultou por colocar uma lente de aumento sobre a deficiência das habilidades técnicas dos professores na utilização de TIC. O problema da baixa adesão dos profissionais da educação aos recursos de TIC, causado pela insuficiência de habilidades técnicas, pode ser minorado pelo desenvolvimento de ferramentas de TIC que ofereçam uma *interface* amigável ao usuário.

É importante observar que a informatização dos processos de trabalho através do uso da Internet e de sistemas de informação já integra o exercício da atividade educadora. Já é

uma prática comum dos docentes a realização de pesquisas na internet à procura de material de apoio para compor suas aulas e instrumentos pedagógicos (SILVA, 2020). Um exemplo dessa prática é a busca na internet por questões de provas de vestibulares e concursos oficiais. Para melhor preparar seus alunos para esses exames, os professores se utilizam das questões buscadas para aplicar testes, simulados e provas. O ENEM é uma das provas de grande interesse.

O ENEM é uma avaliação realizada anualmente desde o ano de 1998 visando verificar a aprendizagem dos estudantes egressos do ensino médio no Brasil. Com a criação do PROUNI em 2004 e, posteriormente, do SISU em 2010, o ENEM começa a ser utilizado como meio de entrada nas universidades do país. Através dos anos posteriores, o ENEM tornou-se a prova de vestibular mais importante do país. O meio básico para ter acesso às provas passadas do ENEM é através do portal na internet do INEP. O INEP é um órgão federal mantido pelo Ministério da Educação, e seu site disponibiliza todas as provas do ENEM em formato PDF.

Buscar provas do ENEM e extrair questões para utilizá-las em instrumentos de ensino é uma atividade que contém uma série de limitações para o professor. Essas limitações são potencializadas especialmente se não há apoio de nenhuma ferramenta especializada. Uma vez obtido o acesso à prova pretendida, é necessário realizar uma pesquisa no documento para extrair as questões de interesse. A diversidade de modelos de organização das provas se mostram como um ponto dificultador na extração de informações desses documentos. As provas podem conter, além do enunciado e opções de resposta, tabelas, gráficos e imagens. O tempo gasto na busca pelas provas, na identificação das questões de interesse, e na extração, armazenamento, organização e reutilização dessas questões, varia conforme os aspectos socioculturais de cada professor.

A deficiência em habilidades técnicas pode gerar um aumento significativo do tempo e esforço empregados na elaboração de instrumentos pedagógicos. A automação no processo de análise, extração e migração de dados desses arquivos se revela como uma solução possível para a redução de tempo e esforço humano. Ela se torna possível a partir da disponibilidade das provas do ENEM em um portal online de acesso gratuito, e por estarem salvas em um

formato de arquivo digital comum, o PDF. A estruturação dos dados extraídos permite que se desenvolva um banco de questões automático. Barbosa, Silva e Freitas (2021) constataram que a ferramenta banco de questões se mostrou um instrumento relevante para auxiliar professores no desenvolvimento de instrumentos avaliativos de boa qualidade.

1.1 MOTIVAÇÃO

A motivação deste trabalho é encontrar formas de automatizar o processo que abrange: a coleta de arquivos de exames avaliativos educacionais disponíveis na internet no formato PDF, a extração dos dados das questões presentes nesses exames e o armazenamento destes dados em um banco de dados relacional. Serão tomadas como base de dados as provas do ENEM. Queremos otimizar o tempo gasto pelos professores na busca por questões para elaboração de instrumentos avaliativos e promover o aumento no uso de recursos tecnológicos pelos profissionais da educação.

1.2 OBJETIVOS

1.2.1 PERGUNTA DA PESQUISA

O presente trabalho tem como propósito responder a seguinte questão de pesquisa:

É possível desenvolver uma ferramenta capaz de realizar a extração de dados de provas em PDF e organizar estes dados estruturadamente de forma automática e com uma boa aceitação por parte dos profissionais de educação?

1.2.2 HIPÓTESES

Para responder à pergunta de pesquisa, parte-se da hipótese de que a combinação de soluções tecnológicas já existentes para a extração de dados de PDFs permita a automatização da extração das questões de provas.

Formulamos as seguintes hipóteses:

H_o: uma ferramenta capaz de realizar a extração de dados de provas em PDF e a organizar estes dados estruturadamente não teria uma boa aceitação por partes dos profissionais de educação.
H_1: uma ferramenta capaz de realizar a extração de dados de provas em PDF e a organizar estes dados estruturadamente teria uma boa aceitação por parte dos profissionais de educação.

1.2.3 OBJETIVO GERAL

O objetivo geral deste trabalho é verificar se criar uma plataforma para automatizar o processo de busca por questões de exames educacionais na internet tem uma boa usabilidade e uma boa aceitação por parte dos profissionais de educação. Este objetivo visa beneficiar os professores, diminuindo o tempo e o esforço empregados na busca, leitura, análise, identificação e extração das questões de provas da internet. O trabalho tomará como base as provas do ENEM.

1.2.4 OBJETIVOS ESPECÍFICOS

- Analisar soluções existentes e resultados recentes da literatura;
- Especificar uma arquitetura de solução;
- Implementar o código necessário para realizar a importação de itens com imagens;
- Capturar a percepção de potenciais usuários sobre a usabilidade (efetividade, eficiência e satisfação) da ferramenta através do questionário calibrado SUS, visando obter métricas sobre a aceitação por parte destes usuários;
- Mensurar a qualidade do processo de importação através de uma contabilização de erros realizada por potenciais usuários.

1.3 ESTRUTURA DE TRABALHO

A organização do trabalho se dá pela seguinte forma:

- Capítulo 1 — Introdução
- Capítulo 2 — Referencial Teórico
- Capítulo 3 — Método

- Capítulo 4 — Modelagem e Desenvolvimento da Solução
- Capítulo 5 — Testes com Usuários e Resultados
- Capítulo 6 — Referências Bibliográficas

2 REFERENCIAL TEÓRICO

Neste capítulo será apresentada, brevemente, uma base teórica para os temas técnicos abordados por este trabalho.

2.1 WEB SCRAPING

Web Scraping denota uma extração de dados da web a partir de uma raspagem (do inglês, *scrape*) da tela. É uma técnica baseada na extração de dados a partir da leitura do código-fonte de páginas na internet. As páginas web, como são chamadas, são construídas através da HTML e interpretadas pelos browsers. O *Web Scraping* consiste em ler a página de forma sistemática, coletando os dados contidos no seu código-fonte. Através desta técnica é possível minerar dados de qualquer página disponível na web, servindo de suporte para ferramentas de pesquisa. Essa técnica será utilizada por este trabalho para realizar a coleta automática das provas e gabaritos do ENEM contidos no portal do INEP.

2.2 PORTABLE DOCUMENTO FORMAT - PDF

O PDF é um formato de documento amplamente utilizado para salvar de maneira eficiente a representação digital de seus dados. A estrutura dos documentos em PDF abrange uma série de variáveis que interferem diretamente no processo de extração de seus dados. Destaca-se a disposição dos itens nas páginas, que não seguem uma regra específica, e a codificação dos arquivos, que varia segundo as ferramentas utilizadas para produzi-los. Wiechork (2021) diferencia o documento PDF nato digital do documento PDF digitalizado. O documento nativo nasce sendo escrito digitalmente, geralmente produzido a partir de ferramentas de escrita de texto, como Microsoft Office, LibreOffice, entre outros. O PDF digitalizado é aquele criado a partir de uma captura digital de um documento físico, que pode ser feita utilizando uma fotografia, salva no formato de documento digital. Essa diferença norteia as diferentes técnicas de extração de dados. Com os documentos em PDF nato digital, é possível extrair os dados a partir da decodificação dos arquivos, enquanto para os documentos PDF digitalizados é necessário usar ferramentas de visão computacional para pré-processar as imagens e transformá-las em texto. Todas as provas do ENEM encontradas

no portal do INEP estão no formato PDF nato digital.

2.3 EXPRESSÕES REGULARES

Uma expressão regular é um método formal de se especificar um padrão de texto, composta por símbolos e caracteres com funções especiais que, quando agrupados, formam uma sequência/expressão (JARGAS, 2016). A expressão funciona como uma regra que pode ser testada em qualquer conjunto de dados. O teste é bem-sucedido quando o conjunto casa com a regra, ou seja, obedece exatamente a todas as suas condições, constituindo um *match,* termo em inglês que tem sentido de combinar, corresponder, igualar. As expressões regulares não se limitam a verificar se um conjunto de dados atende ao padrão testado ou não. Elas podem capturar partes específicas de textos, definindo grupos. Quando um conjunto de dados é testado por uma expressão regular definida em grupos, todas as partes do conjunto de dados que atenderem a esses grupos serão retornadas. Uma mesma expressão pode conter inúmeros grupos. Essa característica torna as expressões regulares uma solução viável para segmentar conjuntos de textos que seguem um padrão específico, como provas e exames.

2.4 SOFTWARE COMO UM SERVIÇO — SaaS

Software como Serviço é um modelo de serviço pertencente à computação em nuvem. Conforme a definição proposta pelo *National Institute of Standards and Technology* (NIST, 2011), é um serviço fornecido ao consumidor através de uma infraestrutura da nuvem. Este pode ser acessado por dispositivos através de interface *thin client*, isto é, um cliente com poucos ou nenhum aplicativos instalados, sendo necessário apenas um browser web na maioria das vezes. O cliente não administra ou controla a infraestrutura básica, incluindo rede, servidores, sistemas operacionais, armazenamento, ou mesmo capacidades individuais da aplicação. Ao eliminar a necessidade de transferências e instalações de software na máquina do usuário, a arquitetura SaaS se revela uma boa opção para um sistema que será utilizado por usuários com baixo conhecimento tecnológico e pouco tempo disponível, o que é o caso dos profissionais da educação. Por essa razão, o sistema desenvolvido aqui se utilizará da mesma arquitetura para encapsular todo o processo de recuperação da informação e apresentar apenas os dados estruturados ao usuário final, estando disponível para uso na web sem a necessidade

de instalação na máquina do usuário.

3 MÉTODO

O trabalho consistiu de revisão de estado da arte, revisão de estado da técnica, desenvolvimento e testes.

3.1 DESENVOLVIMENTO DA FERRAMENTA

A etapa de desenvolvimento iniciou com uma revisão do estado da arte e da técnica. O estudo deu base à criação de um processo para a extração automática de questões de documentos no formato PDF. O processo foi documentado com o uso do BPMN, uma notação de gerenciamento de processos de negócio. A partir do processo gerado, criou-se um algoritmo capaz de coletar automaticamente as provas do ENEM e extrair os textos e imagens de cada arquivo, utilizando métodos das bibliotecas *Scrapy* e *PyMuPDF*. Fez-se uso, ainda, das bibliotecas *ghostscript*, para realizar a compressão dos arquivos em formato PDF coletado. Todo o código desenvolvido por este trabalho pode ser acessado no repositório *Github* [1].

3.2 COLETA E ANÁLISE DE DADOS

A última fase da pesquisa foi a realização de testes com usuários. Desenvolveu-se uma interface para utilização do algoritmo e visualização dos resultados. Cada participante utilizou a interface pesquisando por uma prova do ENEM e obtendo como resultado o total de 20 questões e gabaritos da edição pesquisada, selecionados de forma aleatória. A interface apresenta junto aos resultados coletados o documento PDF extraído, permitindo ao usuário comparar os resultados coletados com o conteúdo original. Ao finalizar a pesquisa e visualização dos resultados, cada participante respondeu a um questionário SUS (*System Usability Scale*, do inglês Escala de Usabilidade do Sistema).

O questionário SUS é um método de avaliação do nível de usabilidade de um sistema. Foi desenvolvido por John Brooke com o intuito de medir de forma rápida como a percepção das pessoas sobre a usabilidade dos sistemas. Se provou uma ferramenta simples e confiável,

[1] Disponível em: https://github.com/KayqueSantos/banco-questoes-enem

tendo sido já citado em mais de 1200 publicações e incorporado no uso comercial (BROOKE, 2013). O questionário foi escolhido para essa pesquisa por ser um método de realização rápida e por avaliar características como: efetividade (atingir o objetivo do usuário), eficiência (quanto esforço e recurso é necessário) e satisfação (o quão o usuário está satisfeito com o sistema).

O questionário SUS consiste em responder as 10 perguntas listadas abaixo:

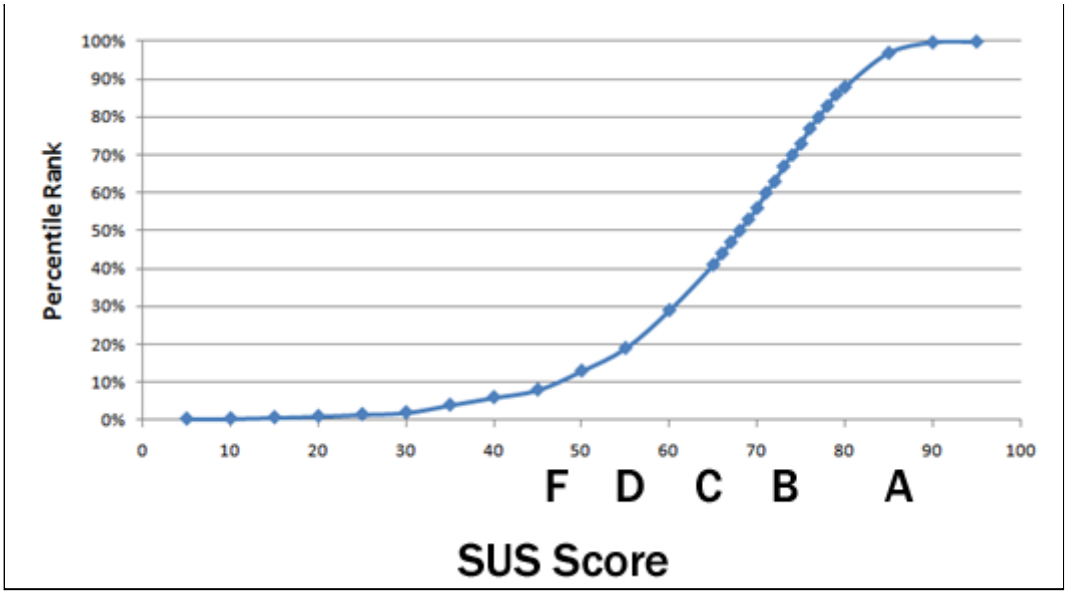
1. Eu penso que gostaria de usar esse sistema com frequência.
2. Eu acho o sistema desnecessariamente complexo.
3. Eu achei o sistema fácil de usar.
4. Eu acho que precisaria de ajuda de uma pessoa com conhecimentos técnicos para usar o sistema.
5. Eu acho que as várias funções do sistema estão muito bem integradas.
6. Eu acho que o sistema apresenta muita inconsistência.
7. Eu imagino que as pessoas aprenderão como usar esse sistema rapidamente.
8. Eu achei o sistema atrapalhado de usar.
9. Eu me senti confiante ao usar o sistema.
10. Eu precisei aprender várias coisas novas antes de conseguir usar o sistema.

As perguntas podem ser adaptadas ao contexto do sistema que se deseja avaliar, porém, não devem ter seu sentido ou ordem mudados. Para cada uma delas o usuário deve responder em uma escala de 1 a 5, onde 1 significa Discordo Completamente e 5 significa Concordo Completamente. Para obter o resultado da avaliação, é necessário calcular as respostas obtidas da seguinte forma:

1. Para cada pergunta de número ímpar (1, 3, 5, 7 de 9), subtrair 1 da pontuação que o usuário atribuiu à resposta.
2. Para cada pergunta de número par (2, 4, 6, 8 10), subtrair a pontuação que o usuário atribuiu de 5 (5-x).
3. Somar todos os valores das dez perguntas e multiplicar por 2,5.

O resultado é chamado *System Usability Score* (Pontuação da Usabilidade do Sistema). Pesquisadores do método SUS definiram alguns critérios para a sua utilização e a interpretação correta dos resultados. A figura 5.5 mostra uma classificação percentual das pontuações do SUS. Na escala podemos ver uma denotação de F a A, onde valores iguais ou inferiores a F representam uma usabilidade ruim, e valores iguais ou superiores a C representam uma usabilidade acima da média. A letra C está fixa no ponto 68, significando que a média do *System Usability Score* é 68.

Figura 3.1 — Classificação percentual das pontuações do SUS



Fonte: Brooke (2013)

Por último, cada participante dos testes informou o número de erros encontrados nas questões e nos gabaritos extraídos, o que forneceu uma métrica de quantidade de questões extraídas com erro.

4 MODELAGEM E DESENVOLVIMENTO DA SOLUÇÃO

4.1 ANÁLISE DO ESTADO DA ARTE

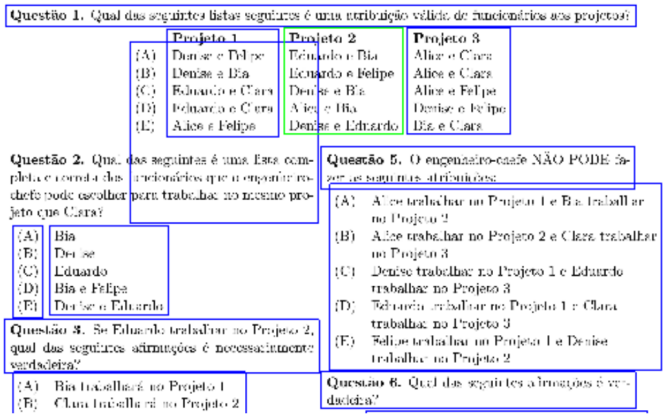
Para propor uma solução tecnicamente viável e atualizada, foi feita uma revisão do estado da arte. Foram buscadas pesquisas recentes que se propuseram a resolver o mesmo problema, a extração de questões de provas a partir de documentos em formato PDF.

Em Deon (2018) foi desenvolvida uma ferramenta para automatizar a extração de questões de provas a partir de documentos no formato PDF. Utilizaram-se bases de dados as provas das Olimpíadas Brasileiras de Informática. O estudo testou diversas formas de extração de dados a partir de PDF nato digital utilizando bibliotecas da linguagem de programação *Python*. A utilização da biblioteca *pdfoftext* apresentou um bom desempenho na extração do texto contido nos documentos, mas não para conteúdo presente em figuras e tabelas. Para contornar essa limitação e desenvolver uma solução eficaz, o autor utilizou o software de visão computacional LAREX para realizar a segmentação de documentos. A segmentação é o processo de decompor as páginas em diferentes regiões como textos, imagens, separadores e tabelas. Ela é realizada a partir de informações de *ground truth*, termo que se refere a uma área de geoprocessamento demarcada com auxílio humano ou de ferramentas de visão computacional (Wierchok, 2021). A construção de um *ground truth* é complexa e os softwares existentes, ainda que otimizem o trabalho, não dispensam a necessidade de um especialista humano para realizar ajustes. Foi desenvolvida uma interface amigável ao usuário para revisar e corrigir as marcações provenientes da segmentação automática feito pelo LAREX.

Em Wiechork (2021) um processo foi proposto utilizando como base as provas do ENADE. Também foi baseado em segmentação de documentos, sendo, portanto, semi automatizado. A etapa de segmentação foi feita por meio do software *Aletheia*. O software não oferece uma interface via linha de comando ou frameworks de programação, criando a necessidade de intervenção manual para sua instalação, configuração e execução.

A figura 4.1 ilustra a segmentação realizada por visão computacional.

Figura 4.1 — Segmentação por Visão Computacional



Fonte: Deon (2018)

Ambas as soluções, ainda que tenham apresentado bons resultados, possuem forte dependência em um profissional especialista. A dependência de um software externo que realize a segmentação do documento através de visão computacional faz com que o usuário final precise instalar o software em sua máquina para poder usá-lo. A necessidade de instalar um software pode ser um fator que afaste usuários com menor conhecimento tecnológico, como é o caso dos professores.

4.2 MATERIAL E FERRAMENTAS

Para desenvolver a solução, foram analisadas quais linguagens de programação, bibliotecas e *softwares* poderiam ser utilizados para obter melhores resultados. A linguagem de programação *Python* foi a mais utilizada no desenvolvimento das ferramentas. Por ser uma linguagem de alto nível, ela fornece maior facilidade ao desenvolvedor, otimizando o tempo de desenvolvimento. *Python* dispõe de bibliotecas robustas para a ferramenta que será desenvolvida. Bibliotecas são conjuntos de métodos que realizam uma determinada tarefa dentro de uma linguagem de programação. A biblioteca *Scrapy* contém um conjunto completo de métodos de *Web Scraping* e é amplamente utilizada.

Dentre as diversas bibliotecas que realizam extração de dados de documentos em formato PDF, a biblioteca *PyMuPDF* foi a escolhida para o desenvolvimento deste trabalho por ser uma ferramenta que contém módulos que lidam tanto com extração de texto quanto com a extração de imagens. Além disso, *PyMuPDF* não tem dependência em módulos externos. Não depender de módulo externo significa que a biblioteca não depende de que nenhum software seja instalado na máquina além da linguagem de programação base. Na tabela 4.1 vemos uma comparação entre a biblioteca *PyMuPDF* e a biblioteca *pdftotext*, que foi a qual obteve o melhor resultado em extração de texto nos testes realizados por Deon (2018) com diversas bibliotecas. Visando agilizar o processo de desenvolvimento, não houveram comparações com outras bibliotecas de *Python*.

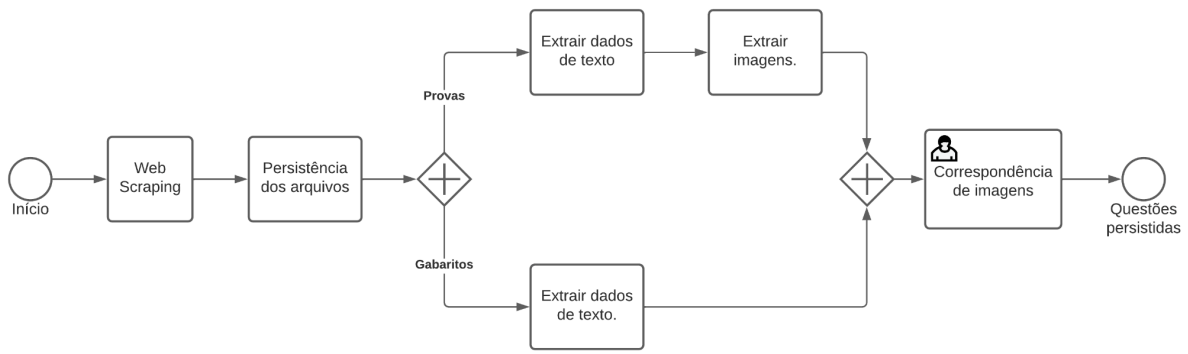
Tabela 4.1 — Comparação entre bibliotecas de *Python* de extração de dados de PDF

Nome	Extrai Texto	Extrai Figuras	Dependência de módulo externo
PyMuPDF	Sim	Sim	Não
pdftotext	Sim	Não	Sim

4.3 PROCESSO PROPOSTO

Nesta seção apresentaremos o processo executado pela ferramenta desenvolvida. A Figura 4.2 fornece uma visão geral do processo, apresentado em BPMN. O processo é composto de 5 etapas:

Figura 4.2 — Documentação do Processo



Fonte: Autoria própria

1. Web Scraping: O processo inicia com a coleta automática dos arquivos de provas e gabaritos do ENEM a partir do site do INEP. O site oferece a acesso a todas as provas e gabaritos de cada edição do ENEM.
2. Persistência dos arquivos: As provas e gabaritos coletados são salvos no armazenamento local da máquina, separados por pastas conforme o ano de aplicação da prova.
3. Análise das Provas:
    a. Segmentação de texto: O processo de extração dos dados textuais das provas começa obtendo todo o texto embutido no PDF. A segmentação do texto extraído é realizada por uma expressão regular que captura o enunciado e alternativas das questões. Os dados de edição, caderno e página onde a questão se encontra também são capturados. Esses dados serão cruzados com os dados das imagens na fase de correspondência de imagens.
    b. Extração de imagens: As imagens contidas no arquivo PDF são extraídas e salvas em pastas individuais para cada documento. Os dados de edição, caderno e página onde a imagem se encontra também são persistidos. Esses dados serão cruzados com os dados das questões na fase de correspondência de imagens.
4. Análise de Gabaritos:
    a. Segmentação de texto: O texto embutido no PDF é extraído da mesma forma que na análise das provas. O texto é filtrado por uma expressão regular, que captura e secciona os dados dos gabaritos.
5. Correspondência de imagens: Nesta etapa, o sistema apresenta uma interface para que o usuário possa realizar a correspondência entre as imagens coletadas na fase de extração de imagens e as questões coletadas na fase de segmentação de texto das provas. O usuário precisará escolher a questão à qual cada imagem pertence. Para otimizar o tempo e esforço do usuário, cada imagem coletada é previamente associada às questões da mesma prova e da mesma página. Visto que cada página contém até, no máximo, três questões, esse é o maior número de alternativas que o usuário terá por questão. O sistema permite, ainda, que o usuário associe mais de uma imagem a uma mesma questão e defina se ela pertence ao enunciado ou a alguma das alternativas.
6. Questões Persistidas: Ao fim do processo é esperado que todas as questões extraídas das provas em PDF sejam persistidas em um banco de dados.

4.4 CONJUNTO DE DADOS

O conjunto de dados para esta pesquisa é composto de provas e gabaritos de avaliações do ENEM. Todos os arquivos foram coletados no endereço de avaliações do ENEM. O site do INEP integra o portal único do Governo Federal e pode ser acessado em gov.br/inep (gov.br, 2021).

Algumas restrições sobre o conjunto de dados foram estabelecidas para otimizar o estudo. As provas de edições anteriores a 2009 possuem um leiaute muito diferente do leiaute das provas realizadas desse ano em diante. Visto que a segmentação por expressões regulares é estritamente baseada nos padrões textuais do documento, expressões adjacentes teriam que ser desenvolvidas para suportar as diferenças de leiaute. As edições a partir do ano 2009 até a última realizada, a de 2021, vêm seguindo o mesmo padrão, sendo necessário apenas uma expressão regular para todo o conjunto. Por esse motivo, as edições anteriores a 2009 foram desconsideradas. Foram desconsiderados, ainda, casos extraordinários de reaplicações que ocorreram em algumas edições e as provas do ENEM Digital que iniciaram na edição de 2020. As provas das edições de 2009, 2011, 2014 e 2015 foram excluídas devido a problemas encontrados para baixar seus respectivos arquivos no site do INEP. O corpo constitui-se, portanto, das provas de caráter impresso de edições dos períodos 2011 – 2013 e 2016 – 2021, totalizando 9 edições e 18 provas, 2 por edição. Cada prova contém 180 questões, totalizando 1620 questões.

Sobre o texto extraído de cada arquivo, nas provas a página inicial é sempre desconsiderada por não conter nenhuma questão, apenas instruções de realização da prova. As páginas referentes à redação também não são analisadas. As provas e gabaritos baixados possuem redundância devido ao formato de aplicação do ENEM. Para cada dia de prova, 4 exemplares diferentes são aplicados para fins de segurança e mitigação de fraudes entre candidatos. Todos os exemplares aplicados em um mesmo dia contém as mesmas questões, divergindo apenas a ordenação delas. Esse formato é vigente desde a edição de 2009. Por motivos de otimização, o algoritmo escolhe apenas um dos 4 exemplares para analisar por dia de prova.

4.5 COLETA DOS DOCUMENTOS PDF

O sistema se inicia com a coleta automática de todas as provas do ENEM no portal do INEP. O algoritmo desenvolvido consegue detectar os links correspondentes às provas e gabaritos de cada edição do exame e, para cada um deles, executa uma requisição HTTP do tipo GET, transferindo o documento para o armazenamento local. Os arquivos que não se referem à aplicação regular são ignorados.

Após a transferência de todos os arquivos, a leitura é iniciada. A biblioteca *PyMuPDF* é utilizada para extrair os dados textuais de cada documento. Inicialmente o algoritmo percorre cada arquivo PDF extraindo o texto da primeira página. O texto das primeiras páginas é filtrado para identificar os seguintes dados: dia e caderno. Esses dois identificadores são cruciais para o chaveamento dos arquivos. As edições se dividem em dois dias de realização e cada dia possui variados cadernos. Os cadernos de um mesmo dia possuem as mesmas questões, porém estas são ordenadas de forma diferente. Portanto, é necessário eleger apenas um caderno para cada dia da edição, tanto de prova quanto de gabarito, para que as informações sejam corretamente persistidas. Após a filtragem, são excluídos as provas que não contém gabaritos correspondentes, e vice-versa. Isso acontece porque, durante as requisições de transferência, alguns arquivos podem apresentar problemas. Após selecionados os arquivos que serão analisados, o algoritmo percorre-os e extrai todo o conteúdo de texto de cada um deles. Os textos das páginas são obtidos como texto corrido, sem nenhum tratamento. O conteúdo de texto será segmentado através das expressões regulares, tratadas na seção a seguir.

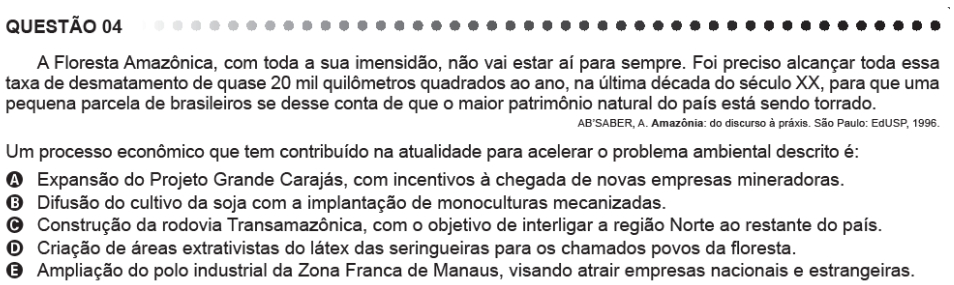
4.6 SEGMENTAÇÃO UTILIZANDO EXPRESSÕES REGULARES

A segmentação dos documentos utilizados como base para esse trabalho foi efetuada com o uso de expressões regulares. Expressões regulares são úteis para buscar ou validar um padrão de texto que pode ser variável (JARGAS, 2016). Para realizar o desenvolvimento dessas expressões, é necessário inicialmente uma observação dos padrões textuais seguidos pelo texto alvo.

Os padrões textuais observados servem como base para o desenvolvimento de expressões regulares que capturam segmentos de interesse. Resumidamente, as expressões regulares permitem definir regras de captura baseadas na categoria de caractere (letra, dígito, pontuação, espaço em branco), na ordem dos grupos de captura (por exemplo, se um padrão P1 é seguido por um padrão P2, ou se um padrão P3 precisa estar no início do texto) e por repetições dos grupos (zero, no máximo uma vez, pelo menos uma vez, entre outras opções). Tais possibilidades permitem que a busca realizada seja completamente personalizada ao padrão do texto alvo.

No caso das provas do ENEM, o padrão observado foi: número da questão, enunciado e alternativas. O número da questão é sempre precedido pela palavra “questão”. O enunciado é um bloco de texto livre. Todas as questões são de múltipla escolha e possuem cinco alternativas numeradas da letra (a) à letra (e). O padrão descrito pode ser observado na figura 4.3.

Figura 4.3 — Padrão das questões do ENEM (1).



QUESTÃO 04

A Floresta Amazônica, com toda a sua imensidão, não vai estar aí para sempre. Foi preciso alcançar toda essa taxa de desmatamento de quase 20 mil quilômetros quadrados ao ano, na última década do século XX, para que uma pequena parcela de brasileiros se desse conta de que o maior patrimônio natural do país está sendo torrado.

AB'SABER, A. **Amazônia**: do discurso à práxis. São Paulo: EdUSP, 1996.

Um processo econômico que tem contribuído na atualidade para acelerar o problema ambiental descrito é:

A Expansão do Projeto Grande Carajás, com incentivos à chegada de novas empresas mineradoras.
B Difusão do cultivo da soja com a implantação de monoculturas mecanizadas.
C Construção da rodovia Transamazônica, com o objetivo de interligar a região Norte ao restante do país.
D Criação de áreas extrativistas do látex das seringueiras para os chamados povos da floresta.
E Ampliação do polo industrial da Zona Franca de Manaus, visando atrair empresas nacionais e estrangeiras.

Fonte: ENEM 2011 — Caderno Azul, 1.º dia — INEP

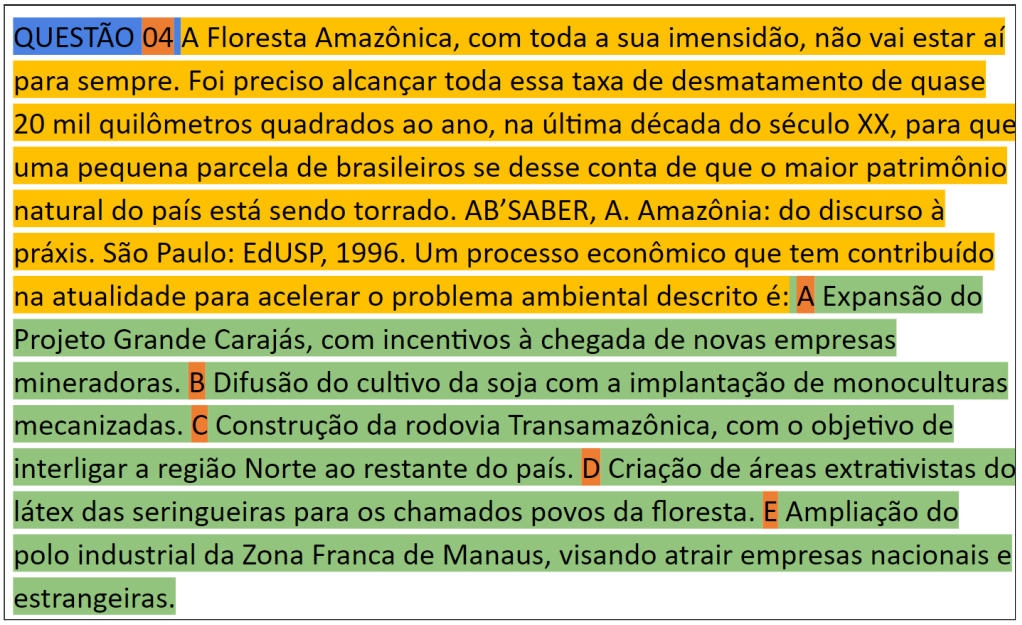
No contexto da pesquisa realizada por este trabalho, sabendo que todas as questões se iniciam com a palavra “Questão”, seguida do número correspondente, é possível obter esse segmento de enumeração das questões através de uma expressão regular que busca todos os segmentos do texto que sigam o padrão “Questão (n)”, onde (n) é qualquer número, representando o número da questão. Sabendo que, logo após a enumeração de cada questão, existe um bloco de texto correspondente ao corpo da questão (enunciado e alternativas), é possível capturar, com uma expressão regular, todos os blocos de texto compreendidos entre os seguimentos de enumeração das questões. Com essa expressão regular obtém-se, para cada

questão da prova, o seu número e o seu corpo textual, mas é possível refiná-la ainda mais para capturar e separar os enunciados das alternativas.

O corpo textual de cada questão é composto por um enunciado (bloco de texto livre), seguido de um conjunto de alternativas, frases que se dispõem em ordem alfabética da letra (a) à letra (e). Esse padrão torna possível realizar a separação entre cada segmento, e relacioná-los com a questão a qual pertence.

A figura 4.4 representa como as expressões regulares desenvolvidas para esse trabalho segmentaram os dados textuais extraídos a partir dos PDFs das provas.

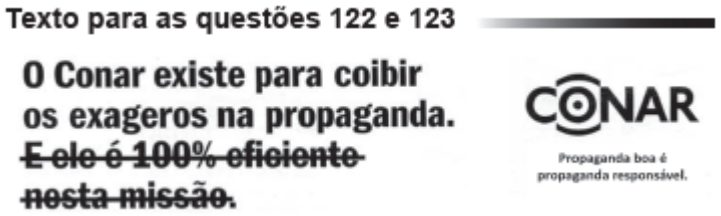
Figura 4.4 — Segmentação por Expressão Regular.



Fonte: Autoria própria

Um elemento presente em apenas algumas das questões é o texto compartilhado. Consiste em um bloco de texto que se aplica às duas questões que se localizam logo após, conforme mostra a figura 4.5. Perceba que o texto compartilhado é precedido pela frase "Texto para as questões", e há a especificação das duas questões às quais ele se aplica.

Figura 4.5 — Padrão das questões do ENEM (2).



Fonte: ENEM 2013 — Caderno Azul, 1.º dia — INEP

O princípio para criar a regra de identificação das questões é o mesmo utilizado para identificação dos textos compartilhados. O segmento de texto que os inicia segue um padrão com a frase “Texto para as questões” seguida de dois números, separados pela palavra “e”. Configura-se, portanto, um padrão variável que pode ser utilizado para capturar a presença de texto compartilhado, os números das questões à qual ele se refere e o bloco de texto que o compreende, vindo logo após o segmento inicial e encerrando com o início da questão.

Os gabaritos das provas também seguem um padrão específico, disposto em tabelas que relacionam os números das questões com a letra da alternativa correta, conforme mostrado na figura 4.6.

Figura 4.6 — Padrão das questões do ENEM (3).

Ciências Humanas e suas Tecnologias

Questões	Gabaritos
1	E
2	D
3	E
4	B
5	C
6	E
7	A
8	E

Ciências da Natureza e suas Tecnologias

Questões	Gabaritos
46	D
47	C
48	E
49	A
50	C
51	D
52	B
53	E

Fonte: ENEM 2011 — Gabarito Caderno Azul, 1.º dia — INEP

O processo de desenvolvimento das expressões regulares utilizadas para os gabaritos é semelhante ao descrito para as provas. Observa-se um padrão na disposição do texto: número da questão seguido do gabarito. O número da questão é representado por dígitos, e os gabaritos são representados por uma letra pertencente ao conjunto [A, B, C, D, E] ou pela palavra “anulado”, que representa uma questão anulada. A repetição de segmentos formados

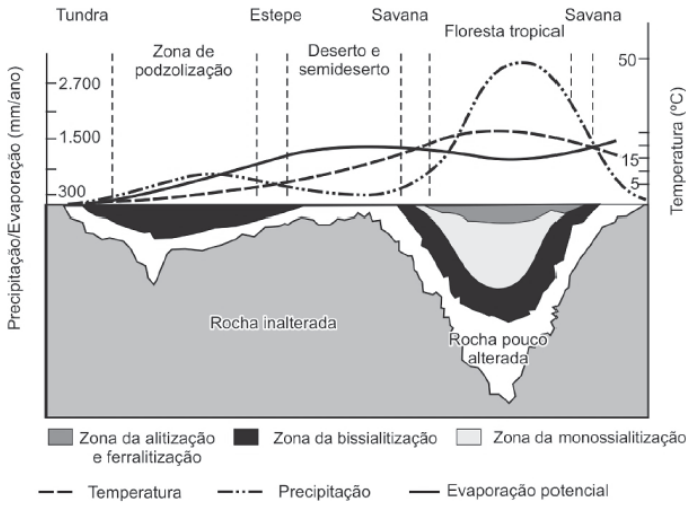
por duplas de texto onde o primeiro é da categoria dígito e o segundo da categoria letra dá base para uma expressão regular que reconheça essas repetições e capture os valores enquadrados nela. Dessa forma, obtém se o gabarito de cada questão.

4.7 EXTRAÇÃO DAS IMAGENS

Além de texto, as questões das provas também podem conter imagens. A figura 4.7 mostra uma questão que contém uma imagem.

Figura 4.7 — Padrão das questões do ENEM (4).

QUESTÃO 06



TEIXEIRA, W. *et al*. **Decifrando a Terra**. São Paulo: Nacional, 2009 (adaptado).

O gráfico relaciona diversas variáveis ao processo de formação de solos. A interpretação dos dados mostra que a água é um dos importantes fatores de pedogênese, pois nas áreas

A de clima temperado ocorrem alta pluviosidade e grande profundidade de solos.
B tropicais ocorre menor pluviosidade, o que se relaciona com a menor profundidade das rochas inalteradas.
C de latitudes em torno de 30° ocorrem as maiores profundidades de solo, visto que há maior umidade.
D tropicais a profundidade do solo é menor, o que evidencia menor intemperismo químico da água sobre as rochas.
E de menor latitude ocorrem as maiores precipitações, assim como a maior profundidade dos solos.

Fonte: ENEM 2011 — Gabarito Caderno Azul, 1.º dia — INEP

As imagens contidas nas questões servem de apoio para o seu entendimento, sendo, portanto, de crucial importância para a extração correta de questões. O algoritmo desenvolvido por este trabalho utiliza a biblioteca *PyMuPDF* para coletar todas as imagens contidas nos documentos, e salvá-las no armazenamento local. Em soluções que utilizam de visão computacional para identificar e extrair informações, as imagens são identificadas por fazerem parte do leiaute da questão. Na abordagem utilizada por este trabalho, não existe uma

forma de relacionar diretamente as imagens com as questões às quais pertencem, pois, os dados de texto e os dados de imagens são extraídos separadamente. As imagens extraídas são persistidas e chaveadas pela prova à qual pertencem, e a página do documento em que se localizam. O dado da página é utilizado para otimizar a relação entre as questões e suas respectivas imagens. Essa correlação é feita na fase de Correspondência de Imagens, a última fase do sistema.

Na fase de correspondência de imagens, todas as questões das provas e seus respectivos gabaritos já foram analisados, estruturados e persistidos. As imagens já foram coletadas e salvas. Uma tela contendo todas as questões é mostrada ao usuário. Para cada questão, o usuário final pode escolher entre as imagens coletadas para importar. O algoritmo relaciona automaticamente as imagens e as questões localizadas na mesma página do documento, diminuindo o esforço e o tempo do usuário nas revisões. Uma vez realizada a associação de imagens, o usuário final tem à sua disposição um banco de questões pronto para utilização.

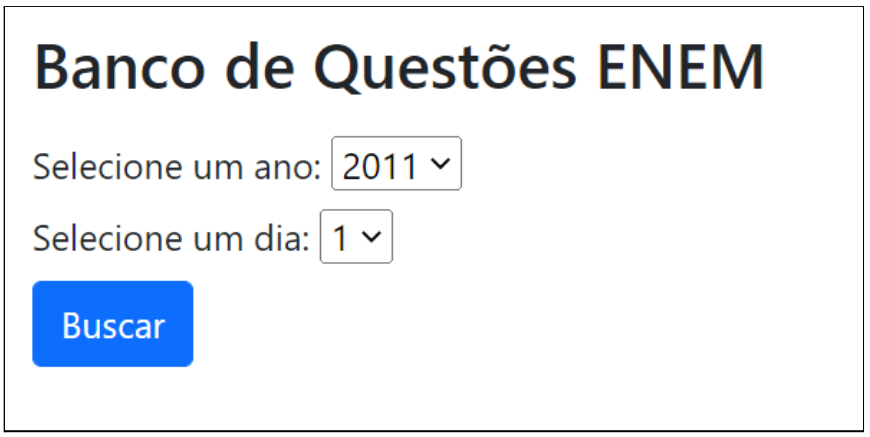
5 TESTES COM USUÁRIOS E RESULTADOS

Esse capítulo descreve os testes realizados junto a usuários e os resultados obtidos por meio deles.

5.1 DESENVOLVIMENTO DA INTERFACE

Para testar o algoritmo foi desenvolvida uma aplicação interativa para permitir que usuários o executem e visualizem as questões extraídas. A interface foi testada com base no processo proposto por esse trabalho sendo, portanto, aplicável aos usuários finais, os profissionais da área de educação. A principal funcionalidade oferecida é a de pesquisar por uma edição do ENEM. A figura 5.1 mostra a tela inicial da aplicação.

Figura 5.1 — Tela de busca

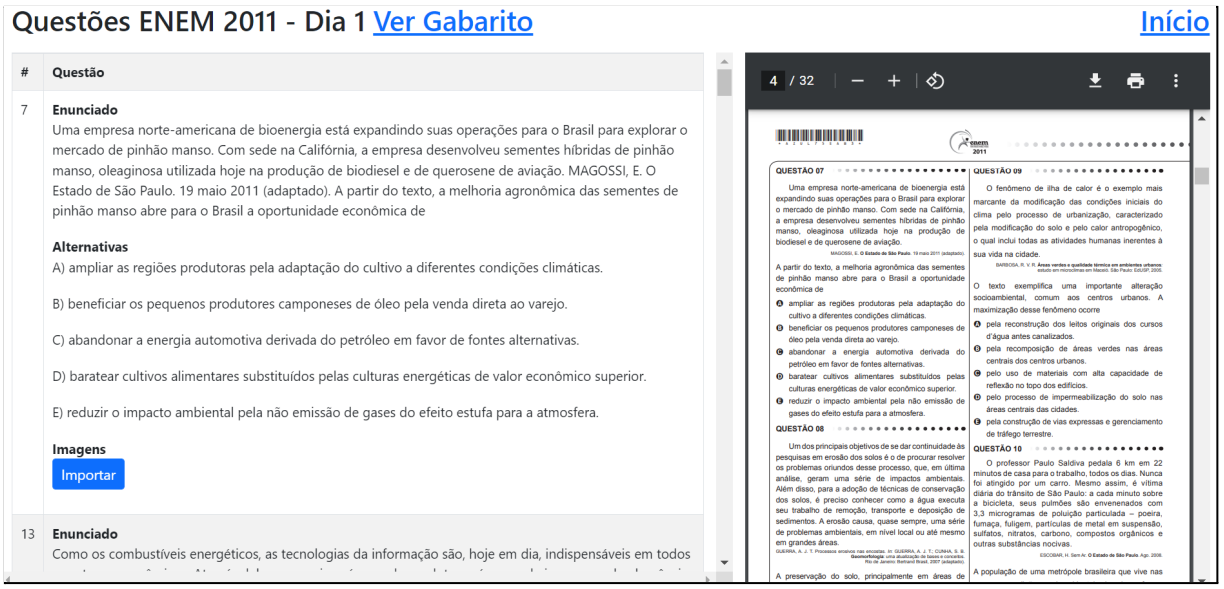


Fonte: Autoria própria

Ao realizar a busca, o algoritmo é executado para o ano e dia selecionados. Ao finalizar a coleta e extração dos dados, o usuário é direcionado para uma tela que mostra as das questões extraídas, junto a uma visualização do PDF original coletado. O usuário pode, então, revisar cada questão em relação ao documento original e verificar se a extração foi realizada corretamente. As questões são formatadas de acordo com seus atributos: texto compartilhado (se houver); enunciado; alternativas; imagens. Essa funcionalidade corresponde às fases de *Web Scraping*, Persistências dos Arquivos, Análise das Provas e Análise dos Gabaritos no processo proposto. Todos os dados são persistidos em um banco de dados, de

forma que essas rotinas são realizadas apenas na primeira vez que uma prova é pesquisada. A partir da segunda pesquisa de uma prova, o algoritmo apenas retorna os dados foram persistidos na extração inicial. A figura 5.2 mostra a tela de visualização das questões.

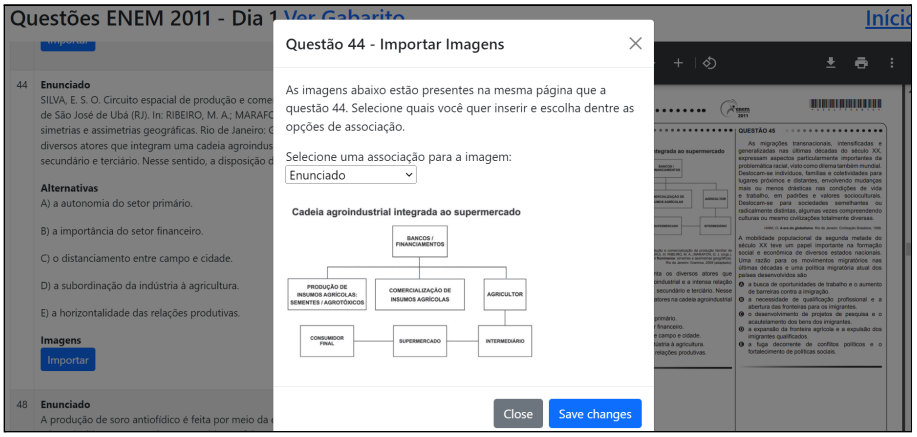
Figura 5.2 — Tela de visualização das questões



Fonte: Autoria própria

Ao clicar em Imagens — Importar, o usuário pode visualizar as possíveis imagens correspondentes à questão. Ele tem a opção de associar as imagens localizadas na mesma página que a questão analisada ao seu enunciado, texto compartilhado ou alternativas. Esta funcionalidade corresponde à etapa de Correspondência de Imagens no processo proposto. A Figura 5.3 mostra a funcionalidade de importar imagens.

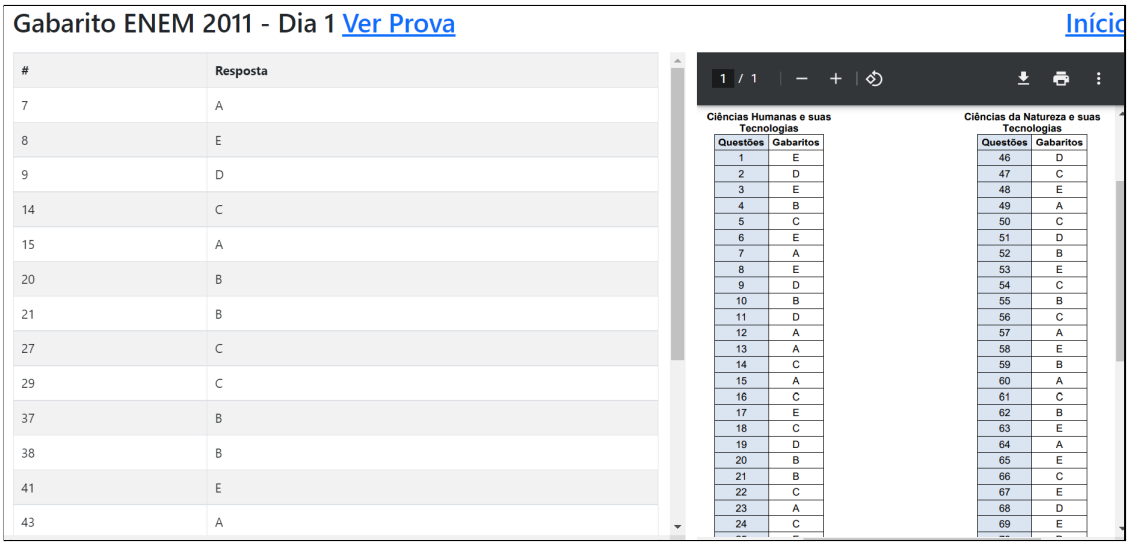
Figura 5.3 — Tela de importar imagens



Fonte: Autoria própria

A última tela é a de visualização dos gabaritos extraídos. Assim como na tela de questões, o usuário pode visualizar os gabaritos coletados e analisá-los imediatamente com base no documento original. A figura 5.4 mostra a tela de visualização dos gabaritos.

Figura 5.4 — Tela de visualização dos gabaritos



Fonte: Autoria própria

A interface desenvolvida foi implantada em um servidor web, tornando-se disponível para qualquer usuário através da internet, cumprindo o objetivo de que a solução dispensasse instalações de software por parte do usuário para utilizá-la.

5.2 EXECUÇÃO DOS TESTES

A aplicação foi testada por 18 participantes. O nível educacional de todos os participantes é de ensino superior completo ou cursando. Deste conjunto, 5 pessoas eram da área de Educação, 12 pessoas eram da área de Tecnologia da Informação e 1 pessoa era de outras áreas. A faixa etária dos participantes da área de Educação foi entre 24 e 50 anos. A faixa etária entre os participantes da área de Tecnologia da Informação foi entre 20 e 26 anos. A faixa etária dos participantes de outras áreas foi de 18 anos. Pessoas das duas áreas foram priorizadas para participar dos testes para que as respostas refletissem a visão de ambos a respeito da qualidade da solução, os profissionais da educação enquanto usuários finais, e os profissionais da tecnologia enquanto desenvolvedores de soluções tecnológicas. Cada

participante realizou um ciclo de utilização da aplicação para uma das provas, cobrindo as 18 provas do conjunto de dados. Para otimizar o tempo e não gerar sobrecarga aos participantes, cada teste consistiu na análise de 20 questões da prova, selecionadas aleatoriamente pelo algoritmo da aplicação. O teste consistiu em acessar a aplicação utilizando um navegador web, buscar a prova determinada, aguardar o carregamento dos resultados, visualizar os resultados das questões e dos gabaritos, e contabilizar a quantidade de questões e gabaritos errados. Ao finalizar o teste, o participante era direcionado a um formulário onde respondia o questionário SUS e informava a quantidade de erros encontrados.

5.3 RESULTADOS DO QUESTIONÁRIO SUS

Ao final dos testes, o *System Usability Score* foi calculado sobre as respostas dos 18 participantes. A tabela 5.1 mostra as médias das pontuações sobre o grupo geral, e sobre os grupos profissionais separadamente.

Tabela 5.1 — Pontuação do SUS

Grupo	Total de Participantes	Média
Todos	18	87.08
Educação	5	85.50
Tecnologia da Informação	12	86.66
Outras áreas	1	100

A pontuação total dada pelos usuários foi acima da média de 68. Isso significa que houve uma boa aceitação dos participantes à solução desenvolvida por este trabalho. Para observar se a avaliação sobre o uso e a experiência obteriam variação condicionada ao perfil tecnológico do usuário final, analisarmos as pontuações dos grupos separadamente por área profissional. Vê-se que, entre os dois principais grupos, não existiu variação expressiva no score. A diferença entre a nota concedida pelos profissionais da Educação e a nota dos profissionais da Tecnologia foi de 1,16. Embora menor, a aceitação por parte dos profissionais da Educação manteve-se alta e acima da média. Partindo do pressuposto de que os profissionais da Educação dispõem de menor conhecimento tecnológico em relação aos

profissionais da área de Tecnologia, pode-se inferir que a solução desenvolvida conseguiu oferecer uma facilidade de utilização equivalente para usuários de diversos perfis tecnológicos. Para confirmar essa interpretação, seria necessário ter um aprofundamento maior sobre o perfil tecnológico dos participantes.

5.4 RESULTADOS DA CONTABILIZAÇÃO DE ERROS

O total de questões do conjunto de 9 provas foi de 1620. Cada participante teve acesso a 20 questões. No total, 360 questões foram revisadas. Isso representa 22% do total de questões. Durante os testes do sistema, os usuários foram solicitados a comparar as questões e gabaritos extraídos pelo sistema e contabilizar quantos apresentaram erro. Foram considerados erros:

1. Enunciado vazio, incompleto ou diferente do documento original;
2. Alternativas vazias, incompletas ou diferentes do documento original;
3. Questão que possui texto compartilhado: texto compartilhado vazio, incompleto ou diferente do documento original;
4. Questão que contém imagens: Tela de importar não aparece nenhuma imagem;
5. Gabarito vazio ou diferente do documento de referência;

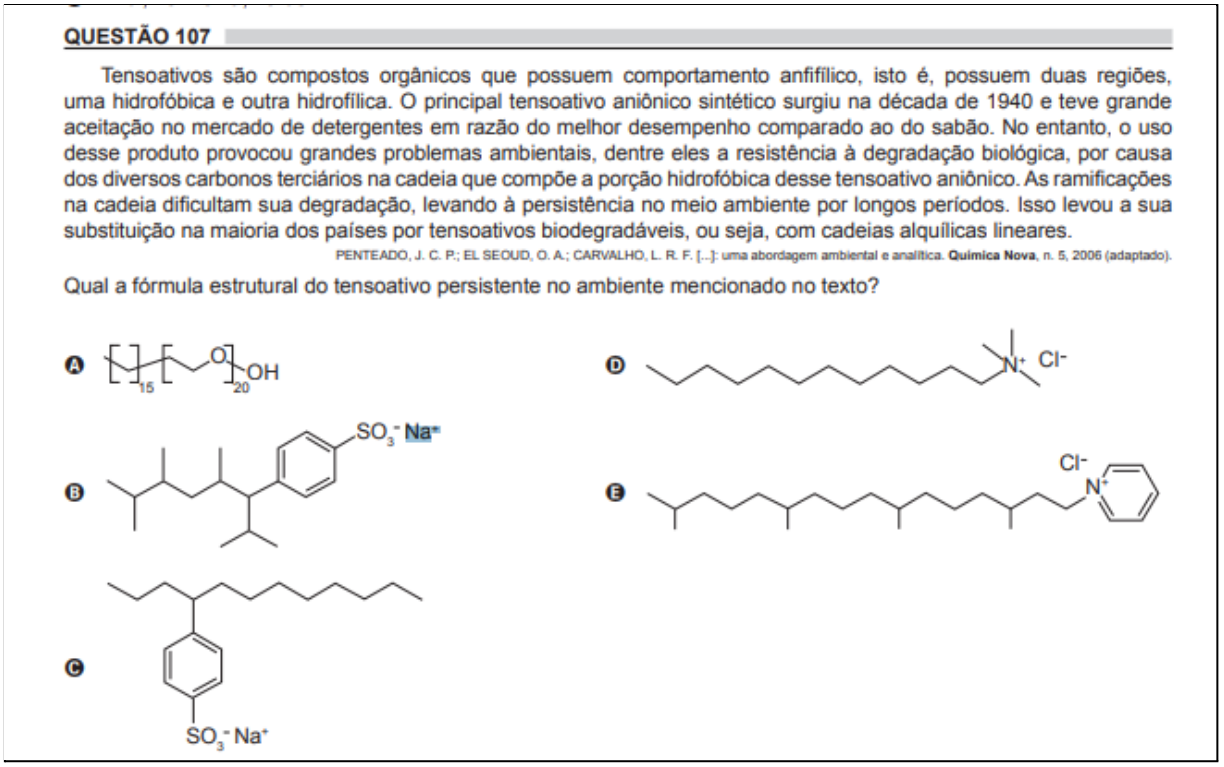
No mesmo formulário onde responderam ao questionário SUS, os participantes inseriram o total de erros encontrados. A tabela 5.2 mostra os resultados dos erros encontrados pelos participantes. Esses erros se limitam aos dados de: enunciado, alternativas, textos compartilhados e imagens, isto é, os dados extraídos dos documentos do tipo Prova.

Tabela 5.2 — Resultados das questões extraídas das provas

Grupo	Total do Conjunto	Total Avaliado	Acertos	Erros	% Acertos
Todas	1620	360	248	112	68%
Provas dia 1	810	180	163	17	90.5%
Provas dia 2	810	180	85	95	47%

Mais da metade das questões avaliadas estavam corretas, obtendo uma taxa de acerto de 68%. Buscando entender as causas para os possíveis erros encontrados pelos participantes, observou-se que as provas aplicadas no 2º dia de cada edição obtiveram uma menor taxa de acertos menor que 47%, o que pode ser considerado um resultado insatisfatório. As provas do dia 1, por outro lado, obtiveram uma taxa de aceitação quase ótima, com uma taxa de acertos de 90.5%. Ao investigar individualmente a extração das questões das provas do 2º dia, observou-se que essas provas concentravam o maior número de figuras que representavam gráficos ou fórmulas, conforme mostra a figura 5.5.

Figura 5.5 — Questão contendo fórmulas químicas



QUESTÃO 107

Tensoativos são compostos orgânicos que possuem comportamento anfifílico, isto é, possuem duas regiões, uma hidrofóbica e outra hidrofílica. O principal tensoativo aniônico sintético surgiu na década de 1940 e teve grande aceitação no mercado de detergentes em razão do melhor desempenho comparado ao do sabão. No entanto, o uso desse produto provocou grandes problemas ambientais, dentre eles a resistência à degradação biológica, por causa dos diversos carbonos terciários na cadeia que compõe a porção hidrofóbica desse tensoativo aniônico. As ramificações na cadeia dificultam sua degradação, levando à persistência no meio ambiente por longos períodos. Isso levou a sua substituição na maioria dos países por tensoativos biodegradáveis, ou seja, com cadeias alquílicas lineares.

PENTEADO, J. C. P.; EL SEOUD, O. A.; CARVALHO, L. R. F. [...]: uma abordagem ambiental e analítica. **Química Nova**, n. 5, 2006 (adaptado).

Qual a fórmula estrutural do tensoativo persistente no ambiente mencionado no texto?

A) $[\_]_{15}[\_\_O\_\_]_{20}OH$

B) $SO_3^-Na^+$

C) $SO_3^-Na^+$

D) N^+ Cl^-

E) Cl^- N^+

Fonte: Autoria própria

A questão apresentada na figura 5.5 contém fórmulas em todas as suas alternativas. O algoritmo criado não conseguiu detectar essas fórmulas como texto e nem como imagens, o que significa que elas representam uma diferente categoria de dado a ser tratado, o qual não foi considerado dentro dessa pesquisa. A razão para as provas do segundo dia conterem maior quantidade de dados dessa categoria é devido à organização das disciplinas nas provas do ENEM. Desde 2009, as provas de Matemática são realizadas no segundo dia de provas. Desde

2017, as provas de Ciências Exatas e da Natureza são realizadas no segundo dia de provas. Houve, portanto, uma maior incidência de questões contendo essa categoria de dado não suportada pela ferramenta, e isso acarretou a diminuição da taxa de acertos. Devido a restrições de tempo do projeto, não foi possível obter um aprofundamento maior sobre as resoluções para este problema. O principal caminho é investigar novas bibliotecas suportem a identificação e extração de gráficos e tabelas de documentos PDF.

Como os documentos dos gabaritos são disponibilizados separadamente, os dados extraídos dos gabaritos foram visualizados pelos participantes em uma tela separada e os itens também foram contabilizados. A taxa de acertos dos gabaritos foi de 100%. Alguns fatores contribuíram para o ótimo resultado: a consistência entre os leiautes dos gabaritos, que pouco difere entre si; a ausência de imagens e figuras; a quantidade de conteúdo presente nesses documentos é menor em relação aos documentos das provas.

6 CONCLUSÃO

Neste trabalho, procurou-se apresentar uma proposta de solução para um problema enfrentado pelos profissionais da educação: a sobrecarga de trabalho. Para resolver este problema, foram pesquisadas formas de automatizar uma rotina dos professores, a busca por conteúdos para compor seus artefatos de ensino. A pesquisa focou na extração de questões das provas do ENEM.

A solução proposta por esse trabalho se diferenciou dos trabalhos voltados ao mesmo problema por focar em: redução da complexidade de implementação, diminuição da dependência em usuário especialista e facilidade de distribuição da solução. Para avaliar a ferramenta desenvolvida, foram realizados testes por usuários da área de Educação e Tecnologia.

A pergunta de pesquisa foi respondida através dos resultados obtidos, pois, demonstrou-se que é possível desenvolver um algoritmo capaz de extrair questões a partir de provas hospedadas em repositórios na internet, utilizando uma abordagem semi-automática, que dispensa a presença de usuário especialista para ser executada e facilmente distribuível. A solução teve boa aceitação pelos usuários finais, com uma boa média de aprovação de usabilidade do sistema entre os profissionais de diferentes perfis tecnológicos.

O método utilizado apresentou limitações, dentre as quais destaca-se a ausência de suporte a algumas categorias de dados contidas nos documentos, restringindo a quantidade de questões que pode ser extraída com sucesso. Outra limitação clara é a necessidade de estar sempre atualizando o algoritmo para que este suporte novos padrões de prova. A segmentação feita com base em expressões regulares é totalmente dependente do padrão contido no documento para o qual ela se destina. Se houver alguma mudança no leiaute de novas provas, deverão ser implementadas novas expressões regulares que analisem corretamente os novos padrões. As duas limitações, no entanto, não impedem que o trabalho possa ser continuado e aprimorado.

6.1 TRABALHOS FUTUROS

Espera-se que trabalhos futuros possam continuar investigando as técnicas existentes em extração de dados de documentos no formato PDF para desenvolver uma solução capaz de

extrair as categorias de dados que não foram suportadas por este trabalho. Além disso, o uso de expressões regulares para segmentação dos dados textuais extraídos de documentos PDF deve ser estudado mais profundamente, considerando outras bases de dados e investigando formas de parametrizar os padrões aceitos, para que o método se torne cada vez mais abrangente.

7 REFERÊNCIAS BIBLIOGRÁFICAS

BARBOSA, P. L. S.; DA SILVA, R.; FREITAS, M. de L. Banco de questões do Instituto Federal do Ceará: um sistema web para auxiliar o processo de avaliação do estudante no ensino superior. **Revista Internacional de Educação Superior**, Campinas, SP, v. 8, n. 00, p. e022006, 2021. DOI: 10.20396/riesup.v8i00.8657696. Disponível em: https://periodicos.sbu.unicamp.br/ojs/index.php/riesup/article/view/8657696. Acesso em: 16 maio. 2022.

BRASIL. **Lei Nº 9.394, de 20 de dezembro de 1996.** Estabelece as diretrizes e bases da educação nacional. Disponível em: http://www.planalto.gov.br/ccivil_03/leis/l9394.htm. Acesso em: 16 maio. 2022.

BROOKE, John. SUS: a retrospective. **Journal of Usability Studies**, Bloomingdale, IL, v. 8, n. 2, p. 29-40, 1 fev. 2013. DOI: 10.5555/2817912.2817913. Disponível em: https://dl.acm.org/doi/10.5555/2817912.2817913. Acesso em: 16 maio. 2022.

CENTRO REGIONAL DE ESTUDOS PARA O DESENVOLVIMENTO DA SOCIEDADE DA INFORMAÇÃO (CETIC). **TIC Educação 2019.** [S.l.]: CETIC; NIC; CGI, 21 nov. 2020. Disponível em: https://cetic.br/pt/publicacao/pesquisa-sobre-o-uso-das-tecnologias-de-informacao-e-comunicacao-nas-escolas-brasileiras-tic-educacao-2019/. Acesso em: 16 maio. 2022.

CENTRO REGIONAL DE ESTUDOS PARA O DESENVOLVIMENTO DA SOCIEDADE DA INFORMAÇÃO (CETIC). **TIC Educação 2020 — edição COVID-19, metodologia adaptada.** [S.l.]: CETIC; NIC; CGI, 25 nov. 2021. Disponível em: https://cetic.br/pt/publicacao/pesquisa-sobre-o-uso-das-tecnologias-de-informacao-e-comunicacao-nas-escolas-brasileiras-tic-educacao-2020/. Acesso em: 16 maio. 2022.

DEON, Otávio Oliveira. **Automatizando a exportação de questões de provas da olimpíada brasileira de informática por meio de ferramentas de extração de texto e visão**

computacional. 2018. 49 p. Trabalho de Conclusão de Curso (Bacharel em Ciência da Computação) — Universidade Federal de Santa Maria, Santa Maria, RS, 2018.

JARGAS, Aurélio Marinho. **Expressões Regulares**: uma abordagem divertida. 4. ed. rev. e ampl. São Paulo, SP: Novatec Editora, 2012. 220 p. ISBN 978- 85- 7522- 337- 6.

MELL, Peter; GRANCE, Timothy. **The NIST definition of cloud computing.** National Institute of Standards and Technology Special Publication, 800-145, 2011. Disponível em: https://csrc.nist.gov/publications/detail/sp/800-145/final. Acesso em: 16 maio. 2022.

PERNAMBUCO. **Lei nº 10.335, de 16 de outubro de 1989.** Modifica a carga horária de pessoal do Grupo Ocupacional Magistério, incentiva o aperfeiçoamento docente e dá outras providências. Disponível em: https://legis.alepe.pe.gov.br/texto.aspx?tiponorma=1&numero=10335&complemento=0&ano=1989&tipo=&url=. Acesso em: 16 maio. 2022.

PERNAMBUCO. **Lei nº 11.329, de 16 de janeiro de 1996.** Dispõe sobre o Estatuto do Magistério Público de Pré-Escolar, Ensino Fundamental e Ensino Médio do Estado de Pernambuco. Disponível em: https://legis.alepe.pe.gov.br/texto.aspx?tiponorma=1&numero=11329&complemento=0&ano=1996&tipo=&url=. Acesso em: 16 maio. 2022.

SANTOS, Alaíde Almeida dos; SOBRINHO, Carlito Lopes Nascimento. Revisão sistemática da prevalência da síndrome de Burnout em professores do ensino fundamental e médio. **Revista Baiana de Saúde Pública**, Salvador, BA, v. 35, n. 2, p. 299-319, 2011. DOI: 10.22278/2318-2660.2011.v35.n2.a307. Disponível em: https://rbsp.sesab.ba.gov.br/index.php/rbsp/article/view/307. Acesso em: 16 maio. 2022

SILVA, Carlos José Pereira da. **Design de um sistema de informação para apoiar a atividade de planejamento de aulas:** uma abordagem situada. 2020. 88 p. Dissertação

(Mestrado em Ciência da Computação) — Universidade Federal de Pernambuco, Recife, PE, 2020.

WIECHORK, Karina. **Extração automatizada de dados de documentos em formato PDF**: aplicação a grandes conjuntos de exames educacionais. 2021. 73 p. Dissertação (Mestrado em Ciência da Computação) — Universidade Federal de Santa Maria, Santa Maria, RS, 2021.