
XX CONIC
Congresso de Iniciação Científica do IFPE
INSTITUTO FEDERAL
Pernambuco

XX Congresso de Iniciação Científica do IFPE

Recuperação de Informação Aplicada à Detecção Automática de Suspeitas de Cópias entre Programas de Computadores - Ano 2: Publicação do Corpus de Programas e Proposição de um Modelo Baseado em IA

Autor: Manoel Victor Oliveira da Silva.

Orientador: Allan Diego Silva Lima.

Sumário

Introdução

Contexto Geral:

- A detecção de plágio em programação é crucial para garantir a **integridade acadêmica** no ensino de lógica e desenvolvimento de software.

Lacuna na Literatura:

- Existe uma lacuna significativa na disponibilidade de **corpus autênticos em português**.
- A maioria dos *corpus* disponíveis é composta por bases em outros idiomas e/ou são gerados automaticamente

Relevância do Trabalho:

- Desenvolvimento de um **corpus autêntico em português**, constituído por códigos-fonte escritos na linguagem **JavaScript** e produzidos por estudantes em atividades educacionais;
- O material foi coletado de forma orgânica, **sem geração automática**, preservando as características originais das produções.

Objetivos Geral e Específicos

Objetivo Geral:

Este plano de atividades tem como objetivo principal publicar um corpus de programas de computadores escritos por estudantes de componentes curriculares de introdução à programação, na língua portuguesa e na linguagem de programação JavaScript.

Objetivos Específicos:

1. **Documentar e organizar o Corpus** para publicação, facilitando a utilização por outros pesquisadores.
2. **Redação de Artigo Científico** sobre o corpus.
3. **Proposição de um Modelo baseado em IA**.

Trabalhos relacionados

Towards a Definition of Source-Code Plagiarism (COSMA; JOY, 2008)*:*

- Este é um artigo teórico introduz uma definição para o problema de plágio. Portanto, não se utiliza um corpus de código para experimentos.
- Linguagem de Programação: Não aplicável.
- Língua dos Programas: Não aplicável.

O Processo de definição:

O processo baseou-se em uma pesquisa online com acadêmicos de programação do Reino Unido.

Utilizou um questionário baseado em cenários para investigar as percepções dos acadêmicos sobre o que constitui plágio em contexto de graduação.

Definição de proposta:

- **Peso da tarefa:** Referente ao conteúdo da disciplina podendo gerar falsos positivos.
- **Reutilizar:** Copiar sem alterações, adaptar o código ou gerar automaticamente sem permissão explícita.
- **Obter:** Roubar o código de outro aluno ou colaborar de forma inapropriada, resultando em submissões similares quando o trabalho deveria ser individual.
- **Reconhecer inadequadamente:** Não citar a fonte (Em comentários ou documentação), fornecer referências faltas ou erradas.

Trabalhos relacionados

Artigo	Corpus Utilizado	Linguagens de Programação	Tamanho do Corpus	De onde os Arquivos Vêm	Como o Corpus foi Criado
CHEERS; LIN; SMITH, 2021	Conjuntos de dados de teste gerados (SPPlagiarise)	Java (foco da implementação atual do BPlag).	29 programas base usados para gerar 13.050 variantes no total.	Os programas base consistiram em 5 submissões de trabalhos de graduação em Java coletadas do GitHub e 24 amostras de algoritmos.	Os dados foram gerados usando a ferramenta SPPlagiarise.

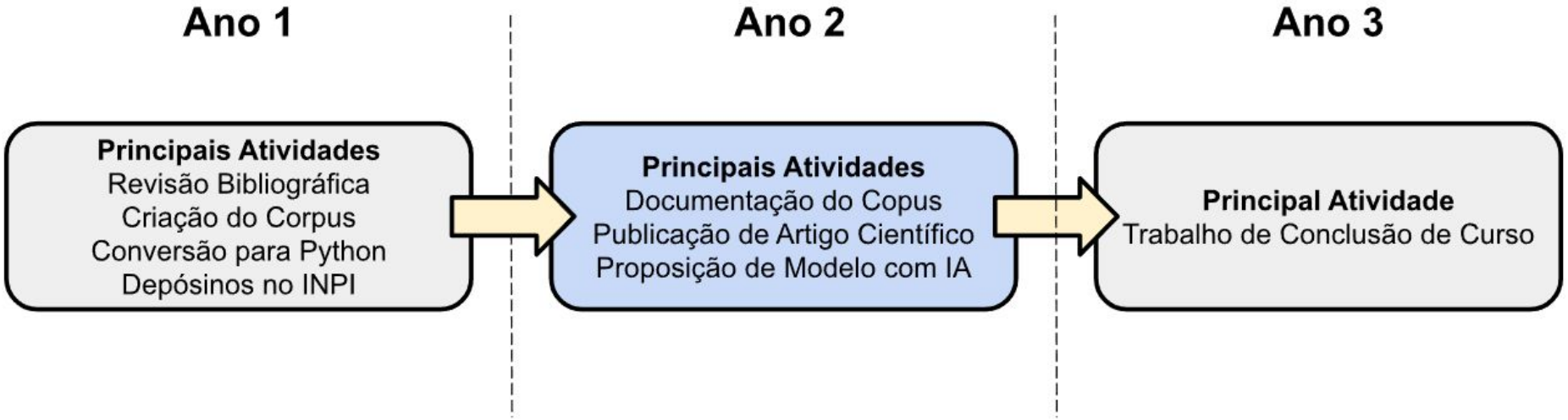
Trabalhos relacionados

Artigo	Corpus Utilizado	Linguagens de Programação	Tamanho do Corpus	De onde os Arquivos Vêm	Como o Corpus foi Criado
ULLAH, 2018	Trabalhos de programação de estudantes (4 estudos de caso: Fatorial, Bubble Sort, Busca Binária e Stack).	C++ e Java.	Pequeno, baseado em quatro estudos de caso.	O conjunto de dados foi de estudantes que submeteram suas tarefas de programação.	Informação não disponível nas fontes fornecidas.
WAN; LIU; GAO, 2018	Projetos de hardware	Verilog HDL	Informação não disponível nas fontes fornecidas.	Informação não disponível nas fontes fornecidas.	Informação não disponível nas fontes fornecidas.

Trabalhos relacionados

Artigo	Corpus Utilizado	Linguagens de Programação	Tamanho do Corpus	De onde os Arquivos Vêm	Como o Corpus foi Criado
Ljubovic & Pajic, 2020	Repositórios de código de granularidade ultra-fina (ultra-fine-grained repositories).	C (em um curso introdutório de programação).	Repositórios criados para 300 estudantes. Porém não diz números exatos sobre quantos exercícios os alunos fizeram ao todo	Submissões de trabalhos de casa de estudantes de um módulo universitário introdutório.	Criado monitorando a atividade do estudante no IDE baseado em nuvem, usando o recurso "autosave" para registrar alterações mínimas.

Metodologia



Metodologia

Resultados: Criação e Organização do Corpus

Processo de Coleta e Organização:

- Códigos-fonte de cinco semestres (2020.1, 2020.2, 2021.1, 2021.2 e 2022.1) foram analisados.
- Anonimização dos dados: Nomes dos alunos foram criptografados utilizando hash.
- Processamento inicial: Utilização do classificador copyFinder (categorizando em ‘Copy’, ‘Warning’ ou ‘Checked’).
- Verificação Manual: Códigos com warnings foram analisados em pares, sendo classificados em "suspeita de cópia".

Resultados

Quadro - Dados preliminares do corpus em desenvolvimento.

SEMESTRES	ARQUIVOS	*WARNING*	*COPY*	*CHECKED*	RESSUBMISSÃO
2020.1	2271	344	17	1559	351
2020.2	495	73	13	347	62
2021.1	1401	256	60	820	265
2021.2	1333	277	86	632	338
2022.2	929	187	26	475	241

Resultados

	2020.1	2020.2	2021.1	2021.2	2022.1
Operadores, Tipos e Variáveis					
Exercícios	0	0	0	0	0
Mini-prova	0	0	254	218	247
Execução Condicional					
Exercícios	93	0	0	0	0
Mini-prova	71	89	211	175	268
Operadores Lógicos					
Exercícios	87	0	0	0	0
Mini-prova	107	87	179	198	0
Laços					
Exercícios	195	0	0	0	0
Mini-prova	199	58	168	159	0
Subprogramas					
Exercícios	70	0	0	0	0
Mini-prova	86	55	150	167	0

Resultados

	2020.1	2020.2	2021.1	2021.2	2022.1
Vetores					
Exercícios	63	0	0	0	0
Mini-prova	102	51	146	137	0
Arrays					
Exercícios	49	0	0	0	0
Mini-prova	64	47	103	129	227
Registros					
Exercícios	47	0	0	0	0
Mini-prova	60	50	106	116	0
Recursão					
Exercícios	0	0	0	0	0
Mini-prova	0	0	0	22	0
Prova					
Prova Unidade I	377	28	65	12	0
Prova Unidade II	396	28	14	187	0
Prova Final	205	2	5	0	0

Publicação e Disseminação

Disponibilização do Corpus:

- Após análise de diversas plataformas, o corpus estruturado será publicado na plataforma Zenodo.
- Essa publicação garantirá acessibilidade para a comunidade acadêmica, promovendo a disseminação do conhecimento gerado.

Publicação Científica:

- Os avanços e análises serão consolidados em um artigo científico.
- Submissão prevista para o Simpósio Brasileiro de Banco de Dados (SBBD) ou periódico, como a revista Inteligência Artificial (IBERAMIA).

Conclusão e Trabalhos Futuros

- A necessidade de verificação manual consumiu mais tempo que o previsto.
- A modelagem da IA foi adiada para o próximo plano de atividades.
- A revisão bibliográfica consolidou o embasamento teórico e a relevância de se criar um corpus autêntico em português.

Principal Desafio:

A **verificação manual** dos códigos-fonte para identificar suspeitas de cópia demonstrou ser trabalhosa e exige um tempo maior que o previsto. Essa etapa é essencial para garantir a autenticidade e qualidade dos dados.

Trabalhos Futuros

- Conclusão do processo de avaliação das suspeitas de cópia.
- Publicação definitiva do corpus no Zenodo.
- Submissão do artigo científico.
- Encaminhamento do estudante para a elaboração do Trabalho de Conclusão de Curso.

Referências

MANNING, C. D., RAGHAVAN, P., & SCHÜTZE, H. (2008). Introduction to Information Retrieval. Cambridge University Press.

COSMA, G.; JOY, M. Towards a Definition of Source-Code Plagiarism. IEEE Transactions on Education, v. 51, n. 2, p. 195-200, maio 2008.

NOVAK, M.; JOY, M.; KER MEK, D. Source-code Similarity Detection and Detection Tools Used in Academia: A Systematic Review. ACM Transactions on Computing Education, v. 19, n. 3, Article No.: 27, p. 1-37, Maio 2019.

WAN, H.; LIU, K.; GAO, X. Token-based Approach for Real-time Plagiarism Detection in Digital Designs. In: 2018 IEEE Frontiers in Education Conference (FIE). Outubro 2018.

ULLAH, F., Wang, J., Farhan, M. et al. Plagiarism detection in students' programming assignments based on semantics: multimedia e-learning based smart assessment methodology. Multimedia Tools and Applications, v. 79, p. 8581-8598, março 2018.

Ljubovic, Vedran, and Enil Pajic. "Plagiarism Detection in Computer Programming Using Feature Extraction from Ultra-Fine-Grained Repositories." IEEE Access, vol. 8, 2020, pp. 96505–96514, https://doi.org/10.21227/71fw-ss32.

INSTITUTO FEDERAL
Pernambuco