XIX CONIC
Congresso de Iniciação Científica do IFPE

16 a 20 de dezembro de 2024 (evento remoto)

Recuperação de Informação Aplicada à Detecção Automática de Suspeitas de Cópias entre Programas de Computadores

Autor: Manoel Victor Oliveira da Silva

Orientador: Prof. Allan Diego Silva Lima

INSTITUTO FEDERAL
Pernambuco

Sumário

- Introdução
- Objetivos
- Trabalhos Relacionados
- Metodologia
- Resultados
- Conclusão e Trabalhos Futuros

Introdução

Contexto Geral:

- A detecção de plágio em códigos-fonte é um desafio crescente no ensino de programação.
- Bases de dados existentes são composta por exemplos em inglês e/ou gerados automaticamente.

Relevância do Trabalho:

- Contribuir com ferramentas ao cenário educacional local.
- Criar um corpus em português para pesquisas futuras.

Objetivos geral

Desenvolver com conjunto de atividades visando iniciar o desenvolvimento de um nova ferramenta para detecção automática de cópias capaz de superar as limitações da ferramenta desenvolvida previamente.

Objetivos específicos

1. Realizar uma revisão bibliográfica sobre o tema da detecção de plágio para em código-fonte;
2. Criar um corpus de testes para a ferramenta;
3. Identificar e priorizar as estratégias para detecção de plágio de código-fonte presentes na literatura que são capazes de melhorar a precisão e a cobertura da ferramenta;
4. Converter a ferramenta atual para Python;
5. Depositar no INPI um pedido de registro de software da nova versão da ferramenta.

Trabalhos relacionados

Towards a Definition of Source-Code Plagiarism (Cosma e Joy, 2008)

- Definem plágio de código-fonte e discutem as implicações da prática acadêmica.
- Destacam a importância de detectar o plágio para garantir a integridade acadêmica.
- Relevância: O trabalho de Cosma e Joy estabelece a base conceitual para a definição do plágio, contextualizando o objetivo da detecção no ambiente educacional.

Trabalhos relacionados

Source-code Similarity Detection and Detection Tools Used in Academia: A Systematic Review (Novak, Joy e Kermek, 2019)

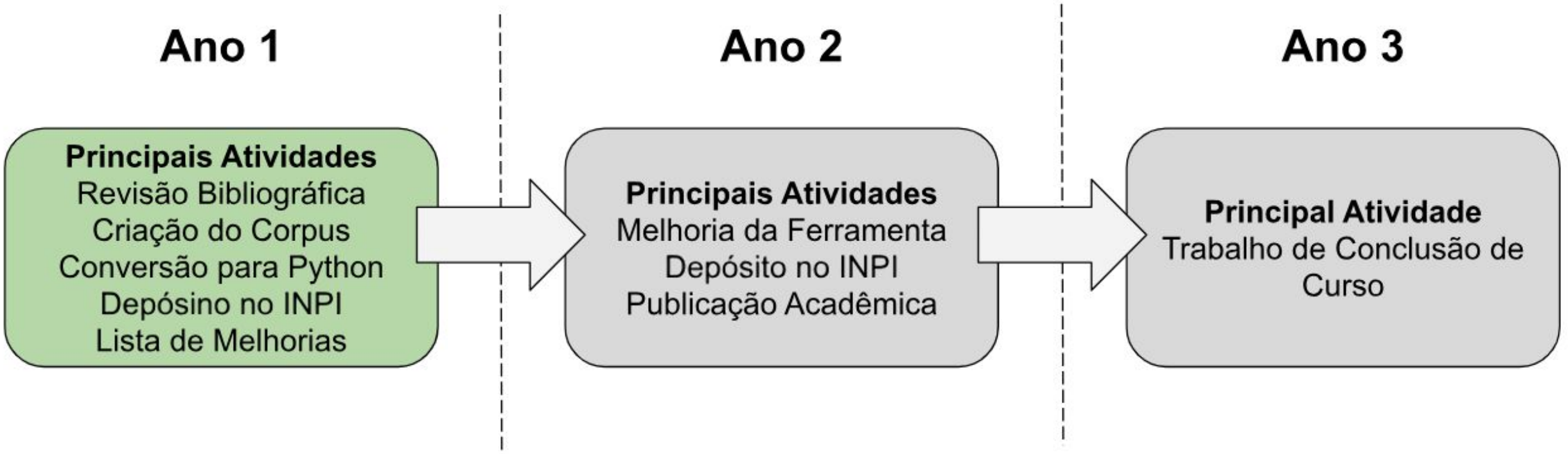
- Apresentam uma revisão sistemática sobre as ferramentas de detecção de plágio utilizadas na academia.
- Analisam diferentes métodos de comparação e ofuscação, além de conjuntos de dados e algoritmos.
- Limitação: O estudo de Novak et al. não fornece um corpus específico para testar o classificador de plágio, dificultando a validação de suas metodologias em ambientes reais de ensino.
- Relevância: A revisão de Novak et al. detalha as ferramentas existentes, fornecendo uma base para a comparação com o "Pega-Cópia".

Trabalhos relacionados

Plagiarism detection in students' programming assignments based on semantics: multimedia e-learning based smart assessment methodology (ULLAH, 2018)

- Proposta de técnica semântica para detecção de plágio em códigos-fonte, realizando a comparação de tokens.
- A análise semântica é aplicada para detectar similaridades que outras técnicas poderiam deixar passar.
- Limitação: O artigo de Ullah et al. também não disponibiliza um corpus para a validação da técnica proposta, o que limita a reprodutibilidade e a aplicabilidade de seus resultados.
- Relevância: A metodologia semântica proposta por Ullah et al. complementa a abordagem do "Pega-Cópia".

Metodologia



Metodologia

Metodologia de criação do Corpus

Processo:

- **Coleta de Dados:** Códigos-fontes submetidos por alunos durante o período da pandemia.
- **Criptografia:** Hash nos nomes.
  - Período-Tipo-Assunto-Questão-Hash.
  - Exemplo: 2020.2-Mini-Prova-Array-Q1-1a4a4599c1fde3681b5ed35d5c786071.
- **Análise Automática:** Uso inicial do *Pega-Cópia*.
- **Verificação Manual:** Esta etapa foi realizada por mim, em conjunto com o professor Allan Diego Silva Lima. Atualmente, a análise ainda está em andamento pelo professor.

Resultados

1. Refatoração do Verificador Original

- **Resultados:** Nova versão do *Pega-Cópia.*
- **Registro INPI:** BR512024002861-7.

2. Porte para Python

- **Resultados:** Nova versão do classificador em Python.
- **Registro INPI:** BR512024002862-5.

3. Criação do Corpus

- **Metodologia Aplicada:**
  - Processamento inicial com *Pega-Cópia*.
  - Verificação manual.

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
Subprograma					
Exercícios	70	0	0	0	0
Mini-prova	86	55	150	167	0

Resultados

	2020.1	2020.2	2021.1	2021.2	2022.1
Vetor					
Exercícios	63	0	0	0	0
Mini-prova	102	51	146	137	0
Array					
Exercícios	49	0	0	0	0
Mini-prova	64	47	103	129	227
Registro					
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

Resultados

Quadro 2 - 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

Conclusão e Trabalhos Futuros

Contribuições do Trabalho:

- Criação de corpus estruturado em português, utilizando o contexto educacional brasileiro, com foco em códigos-fonte simples de estudantes iniciantes.
- Desenvolvimento e adaptação de ferramentas para detecção de plágio.

Trabalhos Futuros:

- Publicação científica dos resultados e disponibilização do corpus para a comunidade acadêmica, promovendo a disseminação do conhecimento gerado e permitindo a reprodução de estudos similares.

Referência

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.; KERMEK, 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.

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.

INSTITUTO FEDERAL
Pernambuco