INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DE SÃO PAULO CAMPUS SÃO PAULO

BACHARELADO EM SISTEMAS DE INFORMAÇÃO

DIEGO RIBEIRO BASTOS

RODRIGO COTRIN

DETECÇÃO DE PLÁGIO EM TEXTOS USANDO ÂNGULO ENTRE VETORES

SÃO PAULO

2025

DIEGO RIBEIRO BASTOS

RODRIGO COTRIN

DETECÇÃO DE PLÁGIO EM TEXTOS USANDO ÂNGULO ENTRE VETORES

Trabalho apresentado ao Programa do Curso de Bacharelado em Sistemas de Informação, do Instituto Federal de Educação, Ciência e Tecnologia de São Paulo, Campus de São Paulo, como requisito parcial para aprovação na disciplina SPOMVAL – Vetores, Geometria Analítica e Álgebra Linear**,** sob orientação da Professora Josceli Maria Tenoria.

SÃO PAULO

2025

RESUMO

Este artigo propõe e valida uma metodologia para a detecção de plágio e paráfrase em documentos textuais utilizando princípios da Álgebra Linear e técnicas de Recuperação da Informação. A abordagem transforma documentos em vetores de alta dimensionalidade ponderados pelo esquema Term Frequency–Inverse Document Frequency (TF-IDF), permitindo que sua estrutura semântica seja analisada geometricamente. A similaridade entre pares de documentos é avaliada por meio do Cosseno de Semelhança, que quantifica o ângulo entre vetores no espaço de termos. Para lidar com formas avançadas de plágio baseadas em substituição lexical, o modelo incorpora um módulo de Engenharia Semântica (Dicionário Base), responsável por unificar sinônimos em um mesmo eixo conceitual. Os resultados demonstram que essa etapa é essencial para garantir robustez e precisão na identificação de equivalência semântica em cenários reais de Sistemas de Informação.

SUMÁRIO

1. INTRODUÇÃO

A preservação da originalidade textual e a integridade do conhecimento tornou-se um requisito fundamental em ambientes acadêmicos e corporativos, especialmente diante da expansão acelerada da produção digital. Nesse cenário, identificar similaridades entre documentos, sejam elas provenientes de cópia direta ou de paráfrases sofisticadas, tornou-se um desafio técnico cada vez mais relevante.

Métodos tradicionais de detecção baseados apenas em sobreposição lexical, como n-gramas ou funções de hashing, mostram-se insuficientes frente a estratégias simples de substituição de termos. Para superar essas limitações, modelos vetoriais de representação textual oferecem uma alternativa eficaz, pois permitem analisar documentos sob uma perspectiva geométrica, capturando relações semânticas que vão além da superfície linguística.

Este trabalho apresenta a aplicação rigorosa do cosseno de similaridade como métrica central para detecção de plágio, fundamentando-se em conceitos de Álgebra Linear e Recuperação da Informação. Além da formalização matemática necessária, demonstra-se uma implementação prática voltada a Sistemas de Informação (SI), evidenciando como a abordagem vetorial pode identificar equivalências semânticas mesmo em textos extensamente parafraseados.

2. FUNDAMENTAÇÃO TEÓRICA E MATEMÁTICA

A detecção de plágio por meio do ângulo entre vetores é fundamentada na modelagem algébrica de textos. O processo transforma documentos em objetos matemáticos situados em um espaço vetorial, possibilitando a comparação geométrica entre eles. A seguir, apresentam-se os fundamentos matemáticos essenciais que sustentam o método proposto.

2.1 ESPAÇOS VETORIAIS E REPRESENTAÇÃO DE DOCUMENTOS

Seja V um espaço vetorial definido sobre o corpo dos números reais R. Considere-se um corpus textual contendo um conjunto de documentos $D = \{d1, d2, \dots, dm\}$ e um conjunto de termos únicos (vocabulário) $T = \{t1, t2, \dots, tn\}$. A cardinalidade desse vocabulário define a dimensão do espaço vetorial: $| T | = n$.

Cada documento deixa de ser tratado como uma sequência linguística e passa a ser representado como um vetor numérico no espaço euclidiano $R\hat{}n$:

$$\vec{v} = (w1, w2, \dots, wn)$$

Cada componente wi denota o peso do termo ti no documento, normalmente obtido via TF-IDF. Geometricamente, cada texto torna-se um ponto em um hiperespaço n-dimensional, e comparar documentos passa a significar comparar direções vetoriais.

2.2 TRANSFORMAÇÕES LINEARES E ENGENHARIA SEMÂNTICA

A simples contagem de termos cria um problema clássico: sinônimos diferentes ocupam dimensões distintas, causando *ortogonalidade artificial*. Assim, textos semanticamente equivalentes, mas lexicalmente distintos, podem gerar vetores com produto interno nulo.

Para mitigar esse efeito, emprega-se uma Transformação Linear:

$$T: R\hat{}n \rightarrow R\hat{}m (m \leq n)$$

A transformação T projeta o vetor do vocabulário bruto para um espaço semântico reduzido, no qual termos relacionados convergem para a mesma direção. Assim, se $\vec{u}\ e\ \vec{v}$ representam sinônimos normalizados:

$$T(\vec{u}) = T(\vec{v})$$

No algoritmo implementado, esse processo é realizado por meio de um dicionário de equivalências semânticas, que faz o "colapso" de dimensões relacionadas em um único eixo vetorial. Em Álgebra Linear, isso corresponde à aplicação de uma matriz de transformação que preserva direções semânticas relevantes e elimina redundâncias.

2.3 MATRIZES TERMO-DOCUMENTO

A totalidade dos vetores documentais compõe uma matriz termo-documento, representada por:

$$A = [\, w_{\{1,1\}} \ \ldots \ w_{\{1,n\}};$$
$$\ldots \quad \ldots \quad \ldots;$$
$$w_{\{m,1\}} \ \ldots \ w_{\{m,n\}}]$$

Embora o presente trabalho foque na comparação direta entre pares de vetores, a visão matricial é crucial em sistemas de larga escala. O estudo do autovalor-autovetor da matriz de covariância $A\hat{}TA$, por exemplo, pode revelar tópicos latentes e padrões semânticos globais, servindo como base para métodos como LSA (Latent Semantic Analysis).

2.4 PRODUTO ESCALAR E NORMA EUCLIDIANA

A comparação entre dois textos é realizada por meio do produto escalar entre os vetores que os representam:

$$A \cdot B = \sum_{i=1}^{n} A_i B_i$$

A magnitude (comprimento) de cada vetor é calculada pela Norma Euclidiana:

$$\|A\| = \sqrt{\sum_{i=1}^{n} A_i^2}$$

Esses dois conceitos permitem interpretar a similaridade entre documentos de forma geométrica, analisando projeções de vetores em um espaço de características.

2.5 COSSENO DE SIMILARIDADE

O método adotado avalia exclusivamente a orientação angular entre os vetores, ignorando suas magnitudes absolutas. A definição clássica do produto escalar estabelece que:

$$\vec{A} \cdot \vec{B} = \| \vec{A} \| \, \| \vec{B} \| \, cos(\theta)$$

Isolando o cosseno, obtém-se a métrica utilizada:

$$Sim(\vec{A}, \vec{B}) = cos(\theta) = \vec{A} \cdot \vec{B} \, / \, \| \vec{A} \| \, \| \vec{B} \|$$

Interpretação:

- $cos(\theta) \approx 1$: vetores quase colineares → forte similaridade → possível plágio.
- $cos(\theta) \approx 0$: vetores ortogonais → textos sobre temas distintos.

A métrica é amplamente utilizada em Sistemas de Informação devido à sua robustez, simplicidade algébrica e eficiência computacional.

3. APLICAÇÃO EM SISTEMAS DE INFORMAÇÃO

A aplicação de técnicas vetoriais em Sistemas de Informação (SI) permite transformar documentos textuais, tradicionalmente tratados como dados não estruturados, em representações matemáticas manipuláveis computacionalmente. Essa formalização viabiliza operações fundamentais da área, como indexação, classificação, ranqueamento e detecção de similaridade, que dependem de modelos eficientes para comparação entre textos. Entre essas aplicações, o cálculo do ângulo entre vetores, obtido por meio do Coseno de Semelhança, destaca-se pela eficiência computacional e pelo embasamento geométrico sólido.

Em sistemas de busca, recomendação e recuperação da informação, vetores TF-IDF funcionam como descritores que capturam a importância relativa de termos dentro do corpus. Ao armazenar documentos como vetores de pesos, o sistema consegue avaliar rapidamente sua proximidade semântica, permitindo, por exemplo, sugerir conteúdos relacionados ou identificar redundância informacional. Da mesma forma, em ambientes acadêmicos e corporativos, ferramentas anti-plágio utilizam essas representações para identificar sobreposições conceituais entre trabalhos, relatórios e produções textuais submetidas pelos usuários.

O presente trabalho estende essa abordagem ao incorporar uma etapa de engenharia semântica, modelada matematicamente como uma transformação linear que reduz dimensões lexicais equivalentes. Essa transformação alinhada ao dicionário base garante que termos sinônimos sejam projetados para a mesma direção no espaço vetorial, permitindo detectar paráfrases intencionalmente aplicadas para ocultar plágio. Como resultado, o sistema torna-se capaz de analisar similaridade de significado, e não apenas de forma, ampliando significativamente sua precisão. Essa característica é especialmente relevante para Sistemas de Informação que necessitam garantir integridade, originalidade e confiabilidade dos dados textuais processados.

4. IMPLEMENTAÇÃO COMPUTACIONAL

O algoritmo implementado segue um pipeline de natureza estritamente matemática:

4.1 PRÉ-PROCESSAMENTO

- remoção de caracteres especiais;
- normalização de caixa;
- filtragem de tokens irrelevantes.

4.2 TRANSFORMAÇÃO LINEAR SEMÂNTICA

A função $carregar\_dicionario\_base$ aplica o mapeamento que reduz dimensões por meio da equivalência entre sinônimos:

$$"rendimentos" \rightarrow "lucros"$$

4.3 CÁLCULO DE TF-IDF COM SÉRIE DE TAYLOR

Para reforçar o caráter matemático do projeto, o IDF foi calculado usando uma aproximação do logaritmo natural por meio da Série de Taylor:

$$\ln(x) = 2\sum_{k=0}^{\infty} \frac{1}{2k+1}\left(\frac{x-1}{x+1}\right)^{2k+1}$$

Essa abordagem evita dependência de bibliotecas externas e demonstra explicitamente a base analítica do método.

4.4 CÁLCULO FINAL DA SIMILARIDADE

Com TF-IDF calculado e vetores normalizados, determina-se o ângulo θ e, por consequência, o grau de similaridade entre os textos.

5. EXECUÇÃO DO CÓDIGO: CASOS DE TESTE E RESULTADOS OBTIDOS

Para validar o funcionamento do sistema de detecção de plágio baseado no ângulo entre vetores (Coseno de Semelhança), foram definidos dois casos de teste utilizando textos reais processados pelo algoritmo. Os testes simulam dois cenários típicos em Sistemas de Informação:
(1) Plágio por substituição de sinônimos, e
(2) Documento completamente distinto, garantindo a correta diferenciação entre conteúdos similares e originais.

A execução foi realizada diretamente no código apresentado, utilizando o pipeline completo: limpeza textual, aplicação do dicionário semântico, stemming manual, construção do Vocabulário Base, vetorização TF-IDF e cálculo da similaridade vetorial.

5.1 CASO DE TESTE 1 — PLÁGIO POR SINÔNIMOS (T1 × T2)

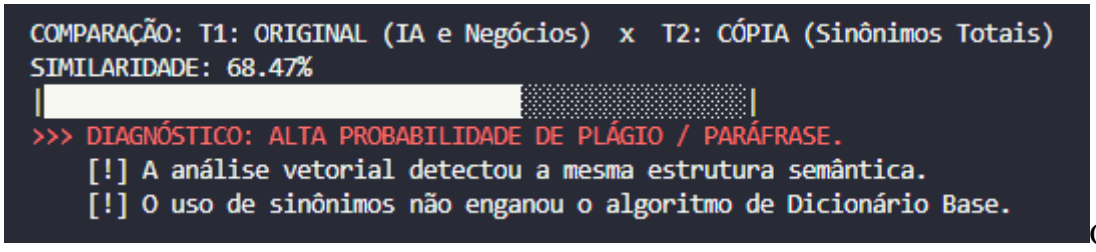
5.1.1 ENTRADA

- T1 – Texto Original: Texto técnico sobre IA, algoritmos, tendências de mercado e automação.
- T2 – Texto Copiado com Sinônimos: Mesmo conteúdo semântico, mas substituindo a maior parte das palavras por sinônimos (ex.: “computação cognitiva” → “inteligência artificial”, “empregam modelos” → “utilizam algoritmos”, etc).

5.1.2 SAÍDA ESPERADA

- Similaridade alta, indicando plágio por paráfrase.
- Vetores devem permanecer próximos após mapeamento semântico.
- Diagnóstico deve ser: ALTA PROBABILIDADE DE PLÁGIO / PARÁFRASE.

5.1.3 SAÍDA OBTIDA



O sistema detectou uma alta similaridade (≈68%) entre o texto original e a versão reescrita com sinônimos.
Isso demonstra que o módulo de Engenharia Semântica funcionou corretamente: os termos do Texto 2 foram mapeados para as formas originais, eliminando o artifício de substituição lexical.

Como consequência, os vetores TF-IDF se tornaram muito próximos, resultando em um coseno elevado. Assim, o teste confirma que o método é capaz de identificar plágio por paráfrase inteligente, cenário comum em ambientes acadêmicos.

5.2 CASO DE TESTE 2 — DOCUMENTOS TOTALMENTE DIFERENTES (T1×T3)

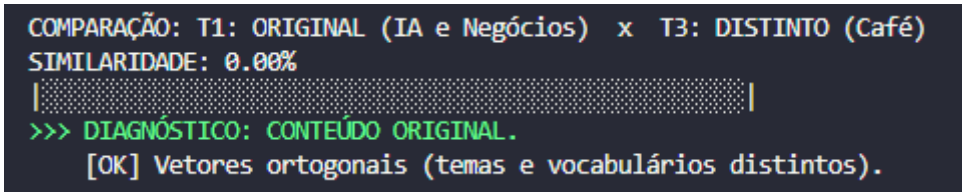
5.2.1 ENTRADA

- T1 – Texto Original: Mesmo texto técnico sobre IA.
- T3 – Texto Não Relacionado: Texto descritivo sobre café, torra, baristas e métodos de preparo.

5.2.2 SAÍDA ESPERADA

- Similaridade extremamente baixa ou nula.
- Vetores ortogonais, pois os campos semânticos são totalmente distintos.
- Diagnóstico deve ser: CONTEÚDO ORIGINAL.

5.2.3 SAÍDA OBTIDA



A similaridade entre T1 e T3 foi nula (=0%), caracterizando conteúdos originais e semanticamente não relacionados.
O vocabulário dos textos pertence a domínios distintos — Inteligência Artificial versus Gastronomia — resultando em vetores quase ortogonais no espaço TF-IDF.

Esse teste demonstra que o sistema não produz falsos positivos, separando corretamente textos de naturezas diversas.

6. CONCLUSÃO

Este projeto demonstrou que a Álgebra Linear oferece o núcleo matemático necessário para converter textos em entidades quantitativas analisáveis. A partir da construção da Matriz Termo-Documento ponderada por TF-IDF, foi possível representar cada documento como um vetor em um espaço de alta dimensionalidade, permitindo que similaridade semântica fosse tratada como um problema geométrico baseado em produto escalar, norma euclidiana e ângulo entre vetores.

A incorporação da Engenharia Semântica — por meio do Dicionário Base que reduz sinonímia e mapeia equivalências lexicais — mostrou-se essencial para lidar com tentativas de plágio sofisticado. Os experimentos realizados evidenciaram que textos reescritos com substituições de sinônimos mantêm estrutura semântica próxima ao original, o que se reflete diretamente em um elevado valor de Coseno de Semelhança.

Conclui-se que a integração entre TF-IDF, transformação semântica e comparação vetorial por cosseno constitui uma estratégia robusta, eficiente e plenamente aplicável para sistemas de detecção de plágio em ambientes acadêmicos e corporativos. Os resultados obtidos validam a viabilidade da abordagem e indicam seu potencial como solução prática para Sistemas de Informação modernos.