The Journal of Open Source Software

An Application for Detecting Plagiarism in University Theses

Elyah Frisco Andriantsialo[1], **Volatiana Marielle Ratianantitra**[1], **and Thomas Mahatody**[1]

1 Laboratory for Mathematical and Computer Applied to the Development Systems, University of Fianarantsoa, Madagascar

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

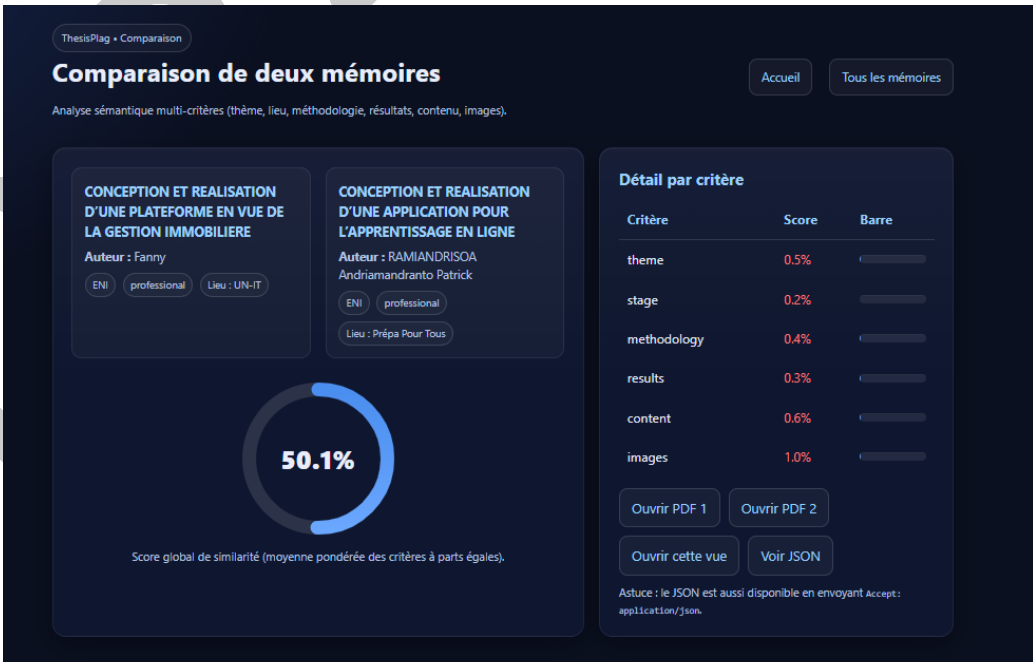
Editor:

Submitted: 26 December 2025
Published: unpublished

Summary

Academic plagiarism has evolved beyond simple copy-paste text to include complex paraphrasing and the reuse of visual elements like figures and diagrams. To address this, we present a hybrid, multimodal web application designed for contextualized plagiarism detection. The system utilizes a multi-criteria approach, analyzing documents based on six distinct dimensions: Theme, Location, Methodology, Results, Global Content, and Images (THLME-Gre schema).

Built with **Flask**, the application leverages advanced semantic models—specifically **Sentence-BERT** (Reimers & Gurevych, 2019) for textual analysis and **CLIP** (Contrastive Language-Image Pre-training) (Radford et al., 2021) for visual analysis. It employs a vector database (ChromaDB) to perform efficient Approximate Nearest Neighbor (ANN) searches across large repositories of university theses.



Comparaison de deux mémoires

Analyse sémantique multi-critères (thème, lieu, méthodologie, résultats, contenu, images).

CONCEPTION ET REALISATION D'UNE PLATEFORME EN VUE DE LA GESTION IMMOBILIERE

Auteur : Fanny

ENI · professional · Lieu : UN-IT

CONCEPTION ET REALISATION D'UNE APPLICATION POUR L'APPRENTISSAGE EN LIGNE

Auteur : RAMIANDRISOA Andriamandranto Patrick

ENI · professional · Lieu : Prépa Pour Tous

Score global de similarité (moyenne pondérée des critères à parts égales).

Détail par critère

Critère	Score	Barre
theme	0.5%	
stage	0.2%	
methodology	0.4%	
results	0.3%	
content	0.6%	
images	1.0%	

Astuce : le JSON est aussi disponible en envoyant `Accept: application/json`.

Figure 1: Screenshot of the application interface showing the dashboard and analysis results.

Statement of Need

Ensuring academic integrity is a growing challenge for higher education institutions in Madagascar, particularly at the **University of Fianarantsoa**, which manages over 30,000 students across various doctoral schools and departments. Currently, the university lacks a centralized, automated institutional tool for plagiarism detection. Faculty members often rely on manual verification or commercial tools that primarily focus on English content and surface-level text matching.

These existing solutions present two major limitations for our context: 1. **Language and Context:** The majority of student theses are written in **French**. Generic tools often struggle to distinguish between legitimate thematic overlap (e.g., multiple students working on "Web Design" or "Digitalization") and actual plagiarism. Our system addresses this by explicitly modeling the "Study Location" and "Methodology" as separate semantic criteria, reducing false positives caused by common academic jargon or shared internship locations. 2. **Multimodality:** Traditional tools frequently miss "visual plagiarism," where students might rewrite the text but copy diagrams, charts, or results directly. By integrating CLIP, our application detects similarities in visual content that text-only tools overlook (Chowdhury & Chellappa, 2016).

This software provides a robust, scalable, and locally deployable solution to enforce academic honesty, specifically tailored to the linguistic and structural needs of Malagasy university research.

Implementation and Architecture

The application follows a modular architecture. The core processing pipeline handles PDF extraction, separating text and images. - **Text** is encoded into dense vectors using `SentenceTransformer` to capture deep semantic meaning (Devlin et al., 2019). - **Images** are processed via `CLIP` to project visual data into a shared embedding space. - **Data Storage** is hybrid: metadata and structured criteria (Theme, Location, etc.) are stored in a Relational DBMS, while high-dimensional embeddings are indexed in a Vector Database for real-time retrieval.

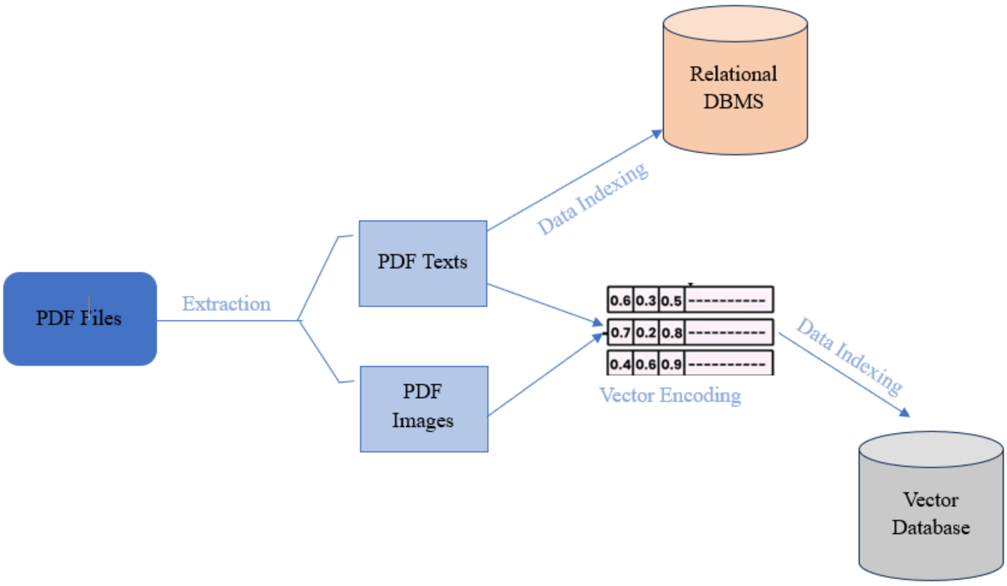


Figure 2: System Architecture: Data flow from PDF extraction to hybrid storage (Relational and Vector Database).

The global similarity score (S_{global}) is computed using an egalitarian weighting model, aggregating cosine similarities from the six defined criteria. This allows for a nuanced assessment, providing decision support thresholds (e.g., >80% for high suspicion) rather than a simple binary judgment.

References

Chowdhury, A. K., & Chellappa, R. (2016). Visual plagiarism: A new challenge in multimedia forensics. *IEEE Transactions on Information Forensics and Security*, *11*(8), 1709–1724.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*, 4171–4186. https://arxiv.org/abs/1810.04805

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., & others. (2021). Learning transferable visual models from natural language supervision. *International Conference on Machine Learning*, 8748–8763. https://arxiv.org/abs/2103.00020

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese BERT-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. https://arxiv.org/abs/1908.10084