INSTITUT SUPERIEUR POLYTECHNIQUE
DE MADAGASCAR

Examen de fin d'études

TP Intelligence Artificielle

Thème : Éditeur de Texte Augmenté par l'IA pour le Malagasy

Vous avez une liberté totale dans vos choix de technologiques, mais une application web est recommandée.
Le mode console n'est pas accepté.

Contexte : Traitement et manipulation d'une langue Low Resource

Les éditeurs de texte classiques (Word, Google Docs) sont optimisés pour l'anglais ou le français, qui disposent de milliards de données. Le Malagasy, lui, est une Langue à Faibles Ressources (Low Resource Language) sur le plan numérique.

Votre mission est de combler ce vide technologique : vous devez concevoir un Éditeur Intelligent capable d'assister un rédacteur malagasy, en trouvant des solutions astucieuses combinant les approches symboliques, algorithmiques et data driven pour pallier le manque de Big Data.

La course à l'Intelligence

Vous avez une totale liberté sur le design. Votre travail sera également évalué vis-à-vis de la quantité et de la qualité des modules d'IA intégrés.

Ci-dessous, une proposition de fonctionnalités pour votre inspiration :

- **Correcteur Orthographique :** Utilisation de dictionnaires (scraping de *tenymalagasy.org* ou *Wikipedia*). Application de distance de Levenshtein et/ou des tables de hachage.
- **Vérification à base de règles :** Utiliser les automates/REGEX pour détecter les erreurs classiques sur la formation des mots (par exemple : le 'nb' ou le 'mk' n'existe pas en Malagasy).
- **Lemmatisation :** Retrouver la racine d'un mot (ex: *manosika* -> *tosika*). En malagasy, le fandrasan-teny a son importance propre.
- **Autocomplétion (Next Word Prediction) :** Modèles Markov ou N-grams entraînés sur des corpus de textes malgaches (Bible, journaux, lois).
- **Traducteur Mot-à-Mot :** « clic droit » sur un mot pour voir sa traduction via API ou dictionnaire local.
- **Explorateur Sémantique (Knowledge Graph) :** Utiliser une ontologie (ex: Fianakaviana, Sakafo) pour suggérer des concepts liés. Ex: Écrire "Razana" (Ancêtre) suggère "Famadihana" via une relation sémantique dans le graphe.
- **Analyse de Sentiment :** Classification simple (Positif/Négatif) basée sur des listes de mots-clés (*Bag of Words*).
- **Synthèse Vocale (TTS) :** Lire le texte avec un accent local.
- **Reconnaissance d'Entités (NER) :** Détecter automatiquement les villes (*Antsirabe*) ou les personnalités.
- **Chatbot Assistant :** Un "Co-pilote" pour demander des synonymes ou des conjugaisons.

N'essayez pas d'entraîner un modèle GPT, ni de RNN de zéro, vous n'aurez pas assez de données ni de temps. Privilégiez les approches hybrides. Soyez astucieux !

Déroulement Conseillé (Start-up Mode)

Pour optimiser votre productivité, divisez-vous en "Squads" :

- **Squad Web/UI (2 personnes) :** UX Design, intégration d'un éditeur riche (ex: *Quill.js, Draft.js, CKEditor*) pour gérer le soulignement et les popups.
- **Squad Data/Scraping (2 personnes) :** Constitution du Dataset. Scraper des articles, nettoyer le texte, construire des lexiques.
- **Squad Algo/GLCIA (3 personnes) :** Implémentation des logiques NLP (Python/NLTK/Spacy) et création des APIs pour le Frontend.

Ce n'est qu'une proposition, vous êtes libres dans votre organisation. Il est conseillé de développer un MVP fonctionnel pour commencer. Dans le cas où une fonctionnalité très intéressante n'a pas pu être intégrée dans l'application mais qui fonctionne. Vous pouvez la présenter séparément dans la vidéo.

Livrables (Avant 16h30)

Tout doit être déposé sur le repository que vous avez indiqué à l'inscription.

1. **Code Source :** Complet et fonctionnel.
2. **Lien de Démo (Optionnel) :** URL de déploiement (Vercel, Streamlit, Ngrok...).
3. **Vidéo de Présentation (3 min max) :**
    - Montrez concrètement comment votre outil gère les spécificités du Malagasy.
    - Expliquez vos stratégies pour contourner le manque de données.
4. **README.md** :
    - Les membres du groupe avec leurs rôles respectifs
    - Documentation technique
    - Liste et brève description des fonctionnalités IA
    - Bibliographie (les sources de données, les articles et documentations, …)

Critères d'Évaluation

Critère	Poids	Description
Fonctionnalités IA	**40%**	Avez-vous réussi à faire du NLP pertinent malgré le manque de données ?
Expérience Utilisateur (UX)	**20%**	L'intégration est-elle fluide pour un rédacteur ?
Qualité Technique	**20%**	Qualité architecture et conception.
Présentation (Vidéo)	**20%**	Capacité à vendre la solution.

Soyez ambitieux. Créez l'outil que Madagascar attend !

Bonne Chance 😊

Annexe : Proposition de corpus, APIs et Bibliothèques

1. Corpus de textes malagasy

Source	URL	Intérêt
Wikipedia MG	mg.wikipedia.org (~90k articles)	API MediaWiki, langue moderne
Teny Malagasy	tenymalagasy.org	Dictionnaire avec définitions

2. Bibliothèques Python NLP

Bibliothèque	Installation	Usage
rapidfuzz	pip install rapidfuzz	Distance de Levenshtein (rapide)
nltk	pip install nltk	Tokenisation, n-grams
spacy	pip install spacy	NER, POS tagging (modèles custom)
BeautifulSoup	pip install beautifulsoup4	Scraping web
requests	pip install requests	Requêtes HTTP / APIs
Flask / FastAPI	pip install flask / fastapi	Backend API REST
gTTS	pip install gtts	Text-to-Speech basique

3. Éditeurs Rich Text (Frontend)

Éditeur	Framework	Remarque
Quill.js	Vanilla JS / React	Simple, personnalisable, recommandé
Draft.js	React	Facebook, très flexible
TipTap	Vue / React	Moderne, basé sur ProseMirror
CKEditor 5	Tous	Complet mais plus lourd

4. Rappels linguistiques malagasy

Combinaisons interdites : nb, mk, nk (début), dt, bp, sz — utiles pour la validation phonotactique.

Préfixes courants : mi-, ma-, man-, mam-, maha-, mpan-, mpam-, fi-, fan-, fam-

Suffixes courants : -ana, -ina, -na — utiles pour la lemmatisation.

Ordre des mots : VSO (Verbe-Sujet-Objet) — différent du français (SVO).

5. APIs et services externes

Service	URL / Endpoint	Usage
Wikipedia API	mg.wikipedia.org/w/api.php	Recherche, extraits d'articles
Google Translate	cloud.google.com/translate (payant)	Traduction (fallback)
LibreTranslate	libretranslate.com (gratuit/self-host)	Alternative open-source