Système d'IA pour le Filtrage et la Categorization des Proposition d'Activités de Startup

Group 5• Feb 22, 2025

Membres de l'équipe :

Team members	Phone Number	Email
AMMAR KHODJA Lilia	0779989558	lilia.ammarkhodja@ensia.edu.dz
ACHOURI Anfal	0559169390	anfal.achouri@ensia.edu.dz
ARAB Sarra	0540221667	sarra.arab@ensia.edu.dz
RAHMOUNI Rahil	0549673850	rahil.rahmouni@ensia.edu.dz
MAARAFI Imene Nour El Houda	0669334891	imene.nour.el.houda.maarfi@ensia.edu.dz

1. Bienvenue !

À une époque de progrès technologique rapide et d'innovation entrepreneuriale, notre projet vise à créer un système basé sur l'IA permettant de filtrer et de catégoriser efficacement les propositions d'activités de startup soumises par des auto-entrepreneurs. Ce système a pour objectif d'identifier de nouvelles idées d'affaires tout en excluant automatiquement les activités existantes et commerciales, favorisant ainsi un écosystème entrepreneurial plus dynamique.

Nous avons commencé avec un ensemble de données diversifié contenant des descriptions d'activités en plusieurs langues — arabe et français — organisé en quatre sous-ensembles distincts : Arabe-Arabe (AA), Français-Français (FF), Français-Arabe (FA) et Mixte (Français & Arabe dans la même cellule). Chaque sous-ensemble a subi un prétraitement et une normalisation approfondis, suivis de vérifications de similarité basées sur des modèles d'embedding afin de faire correspondre les activités existantes issues des listes officielles. Le résultat a été un ensemble de données affiné contenant uniquement des idées de startup uniques, qui ont ensuite été filtrées pour exclure les activités commerciales à l'aide d'une liste de mots-clés soigneusement sélectionnée.

Les activités validées ont ensuite été regroupées en fonction de leur similarité, générant ainsi des catégories potentielles à soumettre à l'examen des administrateurs. Afin d'améliorer l'engagement des utilisateurs et de simplifier le processus de soumission, nous avons développé une application web permettant aux entrepreneurs de proposer leurs idées de startup. Les administrateurs peuvent gérer facilement les soumissions, examiner les catégories et télécharger des fichiers pour un traitement par lots.

Ce projet illustre l'intégration de l'automatisation basée sur l'IA avec une supervision humaine, ouvrant la voie à des propositions d'entreprises innovantes susceptibles de contribuer à la croissance économique et à la diversification.

Pour plus de contexte, voici la démonstration vidéo de notre produit final :

FULLY AUTOMATED MINISTERY ACTIVITIES FILTERING SYSTEM.…

2. Description de la base de Données

Le base de données utilisée dans ce projet comprenait des activités et leurs descriptions provenant de divers auto-entrepreneurs. Il se caractérise par sa nature multilingue, avec des entrées en arabe, en français ou une combinaison des deux. Afin de faciliter notre analyse, nous avons divisé la base de données en quatre sous-ensembles :

- **AA (Arabe-Arabe)** : Activités décrites uniquement en arabe.
- **FF (Français-Français)** : Activités décrites uniquement en français.
- **FA (Français-Arabe)** : Activités avec une description en français et une autre en arabe.
- **Mixte** : Entrées contenant à la fois du français et de l'arabe dans la même cellule.

Cette catégorisation nous a permis d'appliquer des techniques de prétraitement adaptées à chaque sous-ensemble.

3. Méthodologie

3.1. Prétraitement et Normalisation

Chaque sous-ensemble de la base de données a subi un prétraitement pour standardiser le format du texte, y compris la normalisation (ex. : suppression de la ponctuation, conversion en minuscules) et la tokenisation. Cette étape a permis d'assurer la cohérence entre toutes les entrées, facilitant ainsi une analyse plus précise.

3.2. Intégration d'Embeddings et Correspondance de Similarité

Nous avons utilisé différents modèles d'embeddings adaptés à chaque sous-ensemble linguistique : **"paraphrase-multilingual-MiniLM-L12-v2"** et **"Alibaba-NLP/gte-multilingual-base"** pour le reste. Ces modèles nous ont permis de transformer les données textuelles en représentations numériques,

facilitant ainsi des vérifications efficaces de similarité. En comparant les entrées traitées avec les listes officielles d'activités, nous avons pu identifier et supprimer les activités déjà reconnues par le ministère.

3.3. Filtrage des Activités Commerciales

Après avoir éliminé les activités existantes, nous avons affiné une liste de mots-clés commerciaux pour exclure les propositions d'affaires traditionnelles. Certaines exceptions, telles que les termes liés au commerce électronique, ont été prises en compte afin d'éviter une exclusion injustifiée d'activités valides. Cette approche a permis de ne conserver que les idées de startups innovantes dans la base de données.

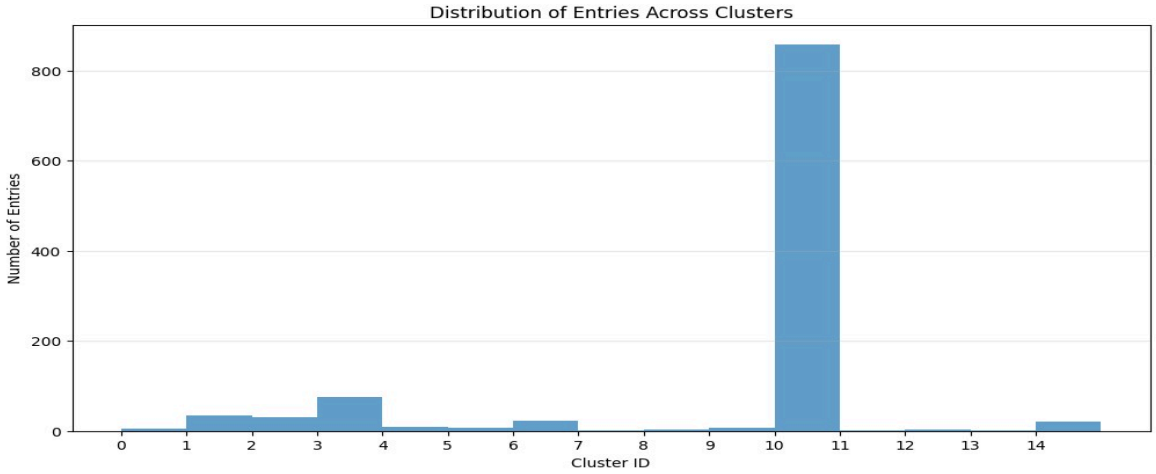
```
# Lists of traditional commercial activities (to exclude)
traditional_commercial_activities = [
    "restaurant", "vente", "achat", "café", "fast food", "boucherie", "poissonnerie", "téléphone",
    "bijouterie", "magasin", "boutique", "fournisseur", "détaillant", "grossiste", "revendeur",
    "importateur", "exportateur", "franchise", "vendeur", "commerçant", "négoce", "dépôt vente",
    "magasin alimentaire", "produits alimentaires", "produits de consommation", "liquidation",
    "bazar", "marché", "vente de voitures", "agents commerciaux", "services de livraison",
    "construction", "matériel de construction", "équipement industriel", "téléphonie mobile", "réparation",
    "location", "distribution", "importation", "commerce de détail", "distributeur", "commerçant de proximité",
    "agricole", "produits agricoles", "alimentation générale", "commerçant ambulant", "transports",
    "مطعم", "بيع", "شراء", "مقهى", "وجبات سريعة", "جزارة", "سمك", "هاتف", "مجوهرات", "متجر", "دكان",
    "مورد", "تاجر تجزئة", "تاجر جملة", "بائع", "مستورد", "مُصدر", "امتياز", "تجارة", "بيع بالتجزئة",
    "منتجات غذائية", "منتجات استهلاكية", "تصفية", "بازار", "سوق", "بيع السيارات", "وكلاء تجاريين",
    "خدمات التوصيل", "إنشاء", "معدات البناء", "تجهيزات صناعية", "هواتف محمولة", "صيانة", "إيجار",
    "توزيع", "استيراد", "تجارة التجزئة", "موزع", "تاجر محلي", "زراعي", "منتجات زراعية", "تجارة عامة",
    "تاجر متجول", "نقل"
]

# Allowed e-commerce & intermediary activities (to keep)
ecommerce_intermediary_activities = [
    "e-commerce", "plateforme de vente", "e-shop", "site marchand", "marketplace", "dropshipping",
    "affiliation", "vente en ligne", "webshop", "commerce virtuel", "commerce digital", "fournisseur e-commerce",
    "commerce électronique", "start-up", "entrepreneur digital", "services de paiement", "agent commercial",
    "courtier", "publicité en ligne", "consultant", "services en ligne", "freelance", "plateforme de freelancing",
    "commercialisation", "digital marketing", "plateforme B2B", "réseau de distribution",
    "تجارة إلكترونية", "منصة بيع", "متجر إلكتروني", "موقع تجاري", "سوق إلكتروني", "دروبشيبينغ",
    "تسويق بالعمولة", "متجر على الإنترنت", "تجارة افتراضية", "تجارة رقمية", "مورد تجارة إلكترونية",
    "شركة ناشئة", "ريادي رقمي", "خدمات الدفع", "وكيل تجاري", "سمسار", "إعلانات عبر الإنترنت",
    "مستشار", "خدمات عبر الإنترنت", "مستقل", "منصة عمل مستقل", "تسويق", "تسويق رقمي", "إدارة الحملات",
    "منصة إعلانات", "تسويق بالعمولة", "وسيط", "وكيل مبيعات", "حلول رقمية", "تخزين سحابي", "دعم العملاء",
    "إعلانات رقمية", "وسائل التواصل الاجتماعي", "منصة بي تو بي", "شبكة توزيع"
]
```

3.4. Regroupement et Catégorisation

La base de données filtrée a ensuite été soumise à des algorithmes de regroupement (*AgglomerativeClustering* avec distance cosinus) pour regrouper les activités similaires. Ce processus a permis de générer des noms de catégories potentielles (générés par un LLM *Groq*), qui ont été présentés à un administrateur pour

validation. Le processus de validation de l'administrateur garantit que seules les catégories pertinentes sont acceptées.



4. Résultats

La mise en œuvre de notre système a abouti à une base de données affinée contenant des activités de startup uniques, accompagné de catégories générées regroupant des idées similaires. Les administrateurs peuvent examiner et approuver ces catégories, garantissant ainsi un système bien organisé pour la gestion des soumissions. Le résultat final met non seulement en avant des idées d'affaires innovantes, mais illustre également l'efficacité de notre approche basée sur l'intelligence artificielle.

---

5. Développement de l’Application Web

Pour compléter notre méthodologie, nous avons développé une application web conviviale. Cette plateforme permet aux auto-entrepreneurs de soumettre leurs idées de startup via un formulaire intuitif. Les administrateurs ont accès à un tableau de bord où ils peuvent consulter les soumissions, valider les catégories et télécharger des fichiers pour un traitement par lots. Cette intégration de l'IA avec

une validation humaine offre une solution complète pour gérer efficacement les propositions de startup.

6. Conclusion

Notre projet démontre avec succès le potentiel de la technologie de l'IA pour améliorer le processus de soumission des propositions de startups. En combinant le filtrage et la catégorisation automatisés avec une supervision humaine, nous avons créé un système robuste qui non seulement identifie des idées d'entreprise innovantes, mais favorise également un environnement propice aux entrepreneurs en herbe. À l'avenir, nous comptons affiner davantage notre modèle et explorer de nouvelles fonctionnalités qui pourraient renforcer l'engagement des utilisateurs et améliorer l'expérience globale.

7. Drive

Lien vers le drive du projet contenant tous les documents et fichiers utilisés :

ANAE_Hackathon_group5