**Synthèse des travaux de thèse pour le problème d’identification des destinataires - Année 2024**

**Objectifs du projet :** La réalisation d’un système d’identification des destinataires des adresses EDI reçues en temps réel appelées aussi adresses de livraison.

Description des travaux réalisés

1. **Principe de la solution envisagée :** le problème d’identification des destinataires à partir des adresses EDI reçues est un problème de recherche d’adresses similaires et s’inscrit dans la branche des algorithmes de recherche d’informations et les algorithmes de correspondance d’adresses. Le principe est de réussir à récupérer le destinataire qui correspond à l'adresse EDI dans la base de données dédiée. Les étapes envisagées pour arriver à ce but passent par deux grands principes:
   1. La recherche sémantique d’adresses pour la récupération de destinataires similaires à une requête EDI (nom et adresse postale d’une d’une commande) : Cette étape passe par la mise en place d’une base de représentation sémantique d’adresses généralement réalisée par la définition de modèles d’encodage pour la conversion des adresses en vecteurs sémantiques tels que les vecteurs de phrases. La base d’encodage permettrait par la suite l’utilisation d’algorithmes de recherche tel que le KNN pour la récupération des adresses similaires. Cette similarité se décrit naturellement par le calcul d’un score de proximité vectorielle à l’aide de mesures de similarité comme le cosinus ou autres.
   2. La correspondance sémantique d’adresses : Cette étape vient compléter la première et a pour but de trouver exactement le destinataire correspondant à l’adresse EDI. Divers modèles existent pour la correspondance sémantique. Ces modèles réduisent principalement le problème en une tâche de classification binaire. Cependant, la venue des modèles de langage large (LLM) et la montée en puissance des applications de l’IA générative, nous permet de proposer une solution de correspondance à travers la création d’agents intelligents spécialisés qui pourront venir choisir le bon destinataire correspondant dans une liste de candidats potentiels.
2. **Phases de recherches pour l’année 2024:**

* Rédaction d’article 01-02/2024 :
  + Rédaction et soumission d’un article pour le système de récupération d’adresses (Solution de vote).
* Sentence Transformers 02-05/2024:
  + Réentrainement, ajustement et ajout de nouveaux modèles d’embeddings pré-entrainés.
  + Création d’un nouveau dataset augmenté de paires d’adresses pour l’entraînement (nom + adresse EDI, nom + adresse destinataire).
  + Réajustement des modèles d’embeddings sur le nouveau dataset.
  + Modification d’une classe d’évaluation pour l’ajout de mesures supplémentaires.
  + Tests et évaluation des nouveaux modèles.
* Système de vote 03-06/2024 :
  + Intégration des nouveaux modèles entraînés dans la solution de vote.
  + Implémentation de nouvelles méthodes de vote pour l'expérimentation.
  + Tests et évaluation des différentes méthodes de vote.
* Rédaction d’article 06-07/2024 :
  + Retravailler la rédaction de l’article pour répondre aux critiques des reviewers.
  + Modification et soumission de l'article après un premier refus dans une autre conférence.
* Modèle LLM-RAG 09-12/2024:
  + Réflexion et tests de création de prompts pour des LLM afin de définir la possibilité de les utiliser dans l’identification des destinataires.
  + Création de datasets personnalisés à l’aide du système de récupération d’adresses pour l'ajustement de modèles LLM.
  + Première expérimentation d’ajustement des LLM pour la tâche d’identification de destinataires.
  + Design et création de prompts augmentés de type (Instruction prompting).
  + Ajustement de différentes versions de LLM suivant le prompt utilisé.
  + Tests et évaluation des résultats.
  + Recherche est mise en pratique de techniques de calcul de scores de confiance pour la validation des réponses LLM.