CoDeinE: artificial text COrpus DEsIgNed Ethically

automatic synthesis of clinical documents

Abstract: Machine learning methods have become prevalent in language technologies. They rely on annotated corpora to train models and evaluate algorithms. The CoDeinE project proposes to address the lack of shareable corpora in sensitive domains such as health or banking. The key idea of the project is to use confidential corpora to automatically generate synthetic texts that mimic the linguistic properties of real documents while preserving confidentiality. The project addresses important issues in natural language processing and is also concerned with defining confidentiality criteria to ensure that no original confidential information is found in the generated synthetic texts. We will use clinical documents in electronic patient records as a case study. Furthermore, the project will rely on Games With A Purpose and crowd sourcing to validate and annotate the synthesized texts.

Résumé: L'apprentissage automatique est un levier important des technologies du langage qui nécessite des corpus annotés pour entraîner et évaluer des algorithmes. Le projet CoDeinE propose de pallier au manque de corpus partageables dans des domaines sensibles (santé, finance, ...). L'idée clé du projet est d'utiliser des corpus confidentiels pour générer automatiquement des textes synthétiques anonymes capables d'émuler des documents réels du point de vue de leurs caractéristiques linguistiques. Le projet se positionne dans la thématique du traitement automatique de la langue, mais s'intéresse également à la définition de critères de confidentialité permettant de garantir qu'aucune information confidentielle originale ne se retrouve dans les textes synthétiques générés. Notre cas d'étude sera celui de documents cliniques présents dans les dossiers électroniques patient. Le projet s'appuiera sur la ludification et les sciences participatives pour valider puis annoter les textes synthétisés.

- Dates: 2021-2025
- Financement: ANR-PRC 2020
- Partenaires: CRC, CEA List, LISN, LORIA