PLDAC

Analyse d'un corpus du CV, apprentissage de représentation sur des documents structurés et démêlage des facteurs explicatifs

Comptes rendus des réunions et suivi

AUTEURS:

AKNOUCHE ANIS
HADDADI HACENE

ENCADREUR:

VINCENT GUIGUE

Année Universitaire: février 2021

- **Réunion 1:** 4/02/2021
    - Points discutés:
        - Présentation de la problématique

    - Objectifs:
        - Prise en main de la base de données
        - Implémenter le countVectorizer
        - Utilisation d'un classifieur pour la prédiction du domaine industriel

- **Réunion 2:** 26/02/2021

    - Points discutés:
        - Prise en main des données,
        - Utilisation du countVectorizer de scikit-learn et de la liste Stop-Words de nltk
        - Implémentation de classifieurs : K-means avec une accuracy de 32%, SVM avec une accuracy de 40%
    - Objectifs:
        - Edition d'un document de suivi et comptes rendus (GoogleDocs)
        - Edition du carnet de bord
        - Statistiques sur les mots, fréquence d'apparition…
        - Estimation de l'accuracy en la comparant à celles des algorithmes de prédiction de base tels que la prédiction avec la classe majoritaire (cas de classes désiquilibrées), la prédiction aléatoire (cas de classes équilibrées), naive bayes…
        - Ajouter un classifieur naive bayes multinomiale
        - Recherche documentaire sur le PLSA (Probabilistic Latente Semantic Analysis)

- **Réunion 3:**