

Twenty Newsgroups

Analyse&interprétation

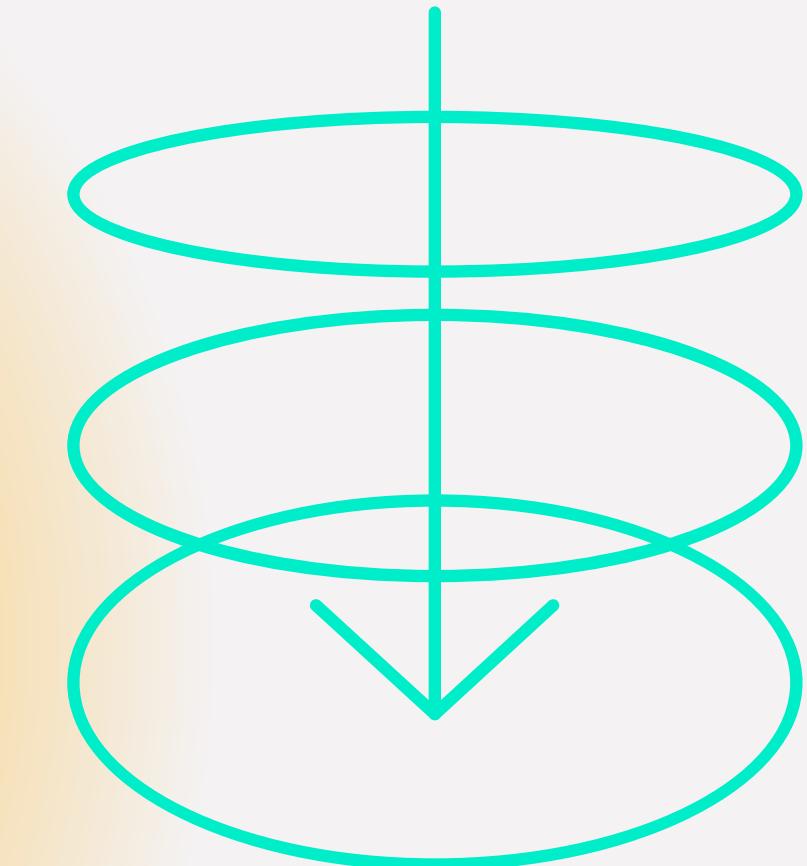
#TF-IDF#KNN#DT

présenté par:

Elouaddany Abdelkarim
Bouzid Abderrahim

Supervisé par :

Mr Bahassine Said



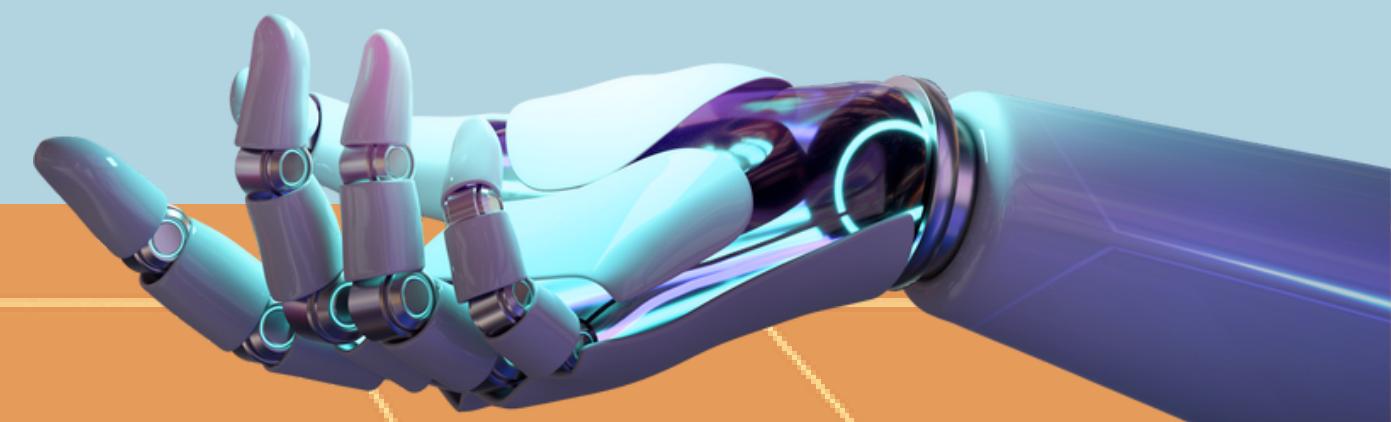
Plan

1 Introduction

2 Contexte

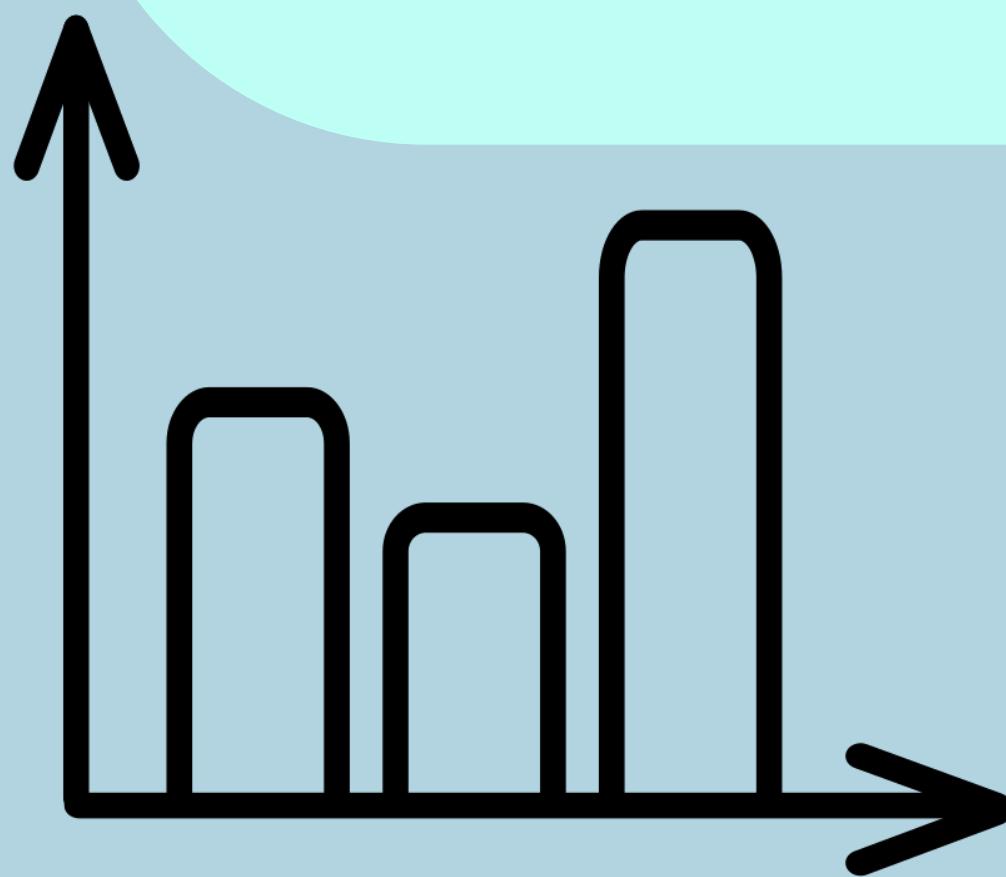
3 Techniques

4 Application



Introduction

Dans le domaine du traitement du langage naturel, la classification de texte est une tâche très répandue dans divers contextes commerciaux





Notre objectif dans ce projet est de catégoriser des documents textuels en utilisant une approche de classification. Pour ce faire, nous utilisons un jeu de données appelé "The 20 Newsgroups data". Ce jeu de données est constitué d'environ 20 000 documents provenant de différents groupes de discussion, regroupés en 20 catégories distinctes



• • •

Pour accéder à ces données, nous utilisons l'API scikit-learn, qui facilite la récupération et la manipulation de ce jeu de données spécifique pour notre tâche de classification



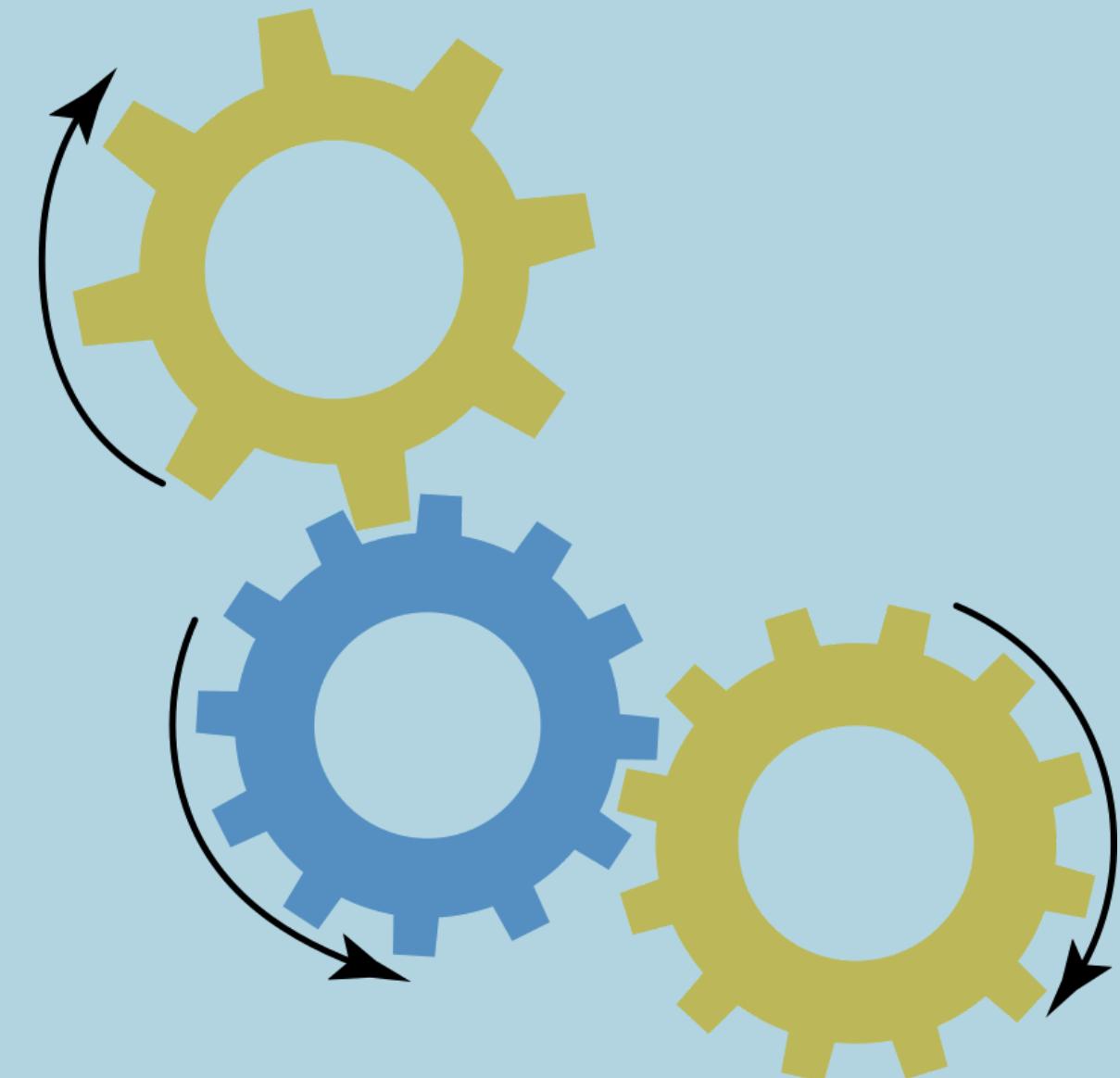
Introduction



Dans le domaine du traitement du langage naturel, la classification de texte est une tâche très répandue dans divers contextes commerciaux

Contexte

pour la classification de texte sur le jeu de données "20 Newsgroups", nous avons suivi les étapes suivantes :

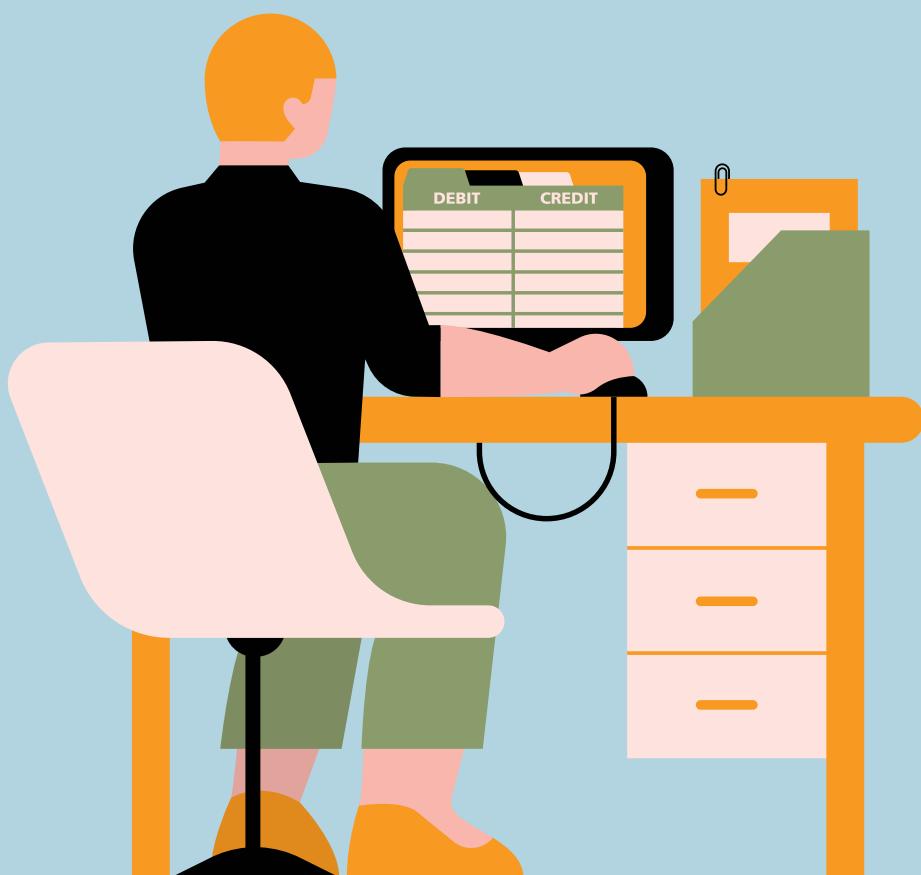


Contexte

Préparation des données

Nous avons chargé le jeu de données "20 Newsgroups" en utilisant la bibliothèque scikit-learn.

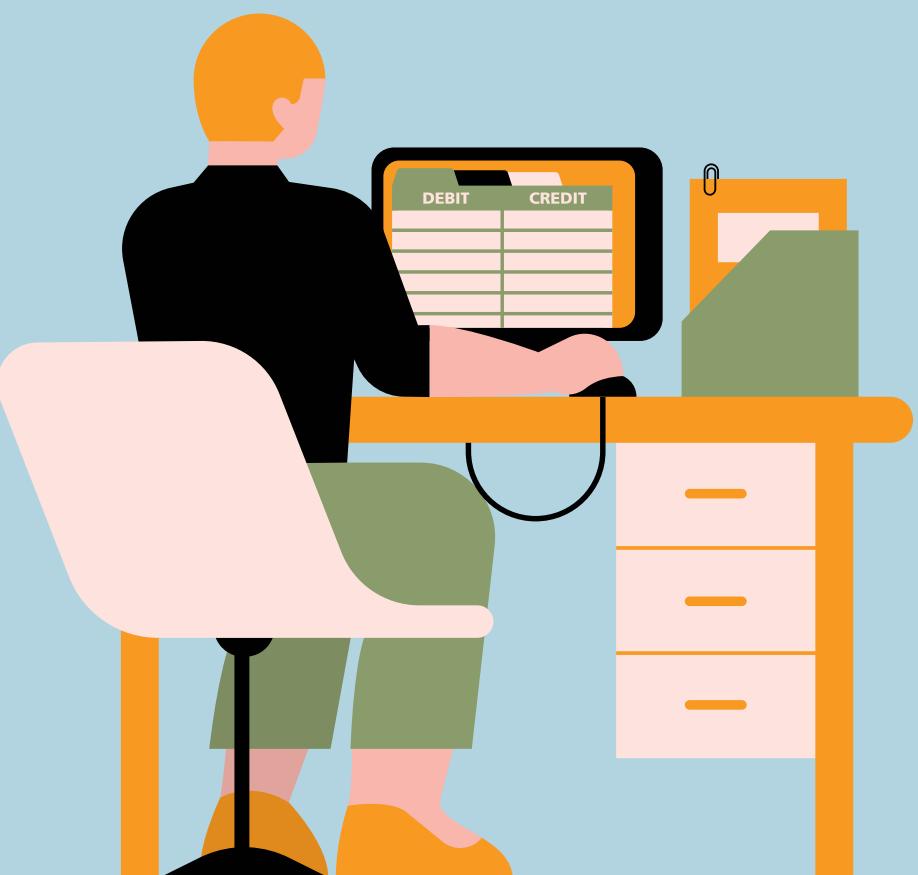
Nous avons effectué une analyse exploratoire des données (EDA) pour comprendre la structure, la distribution et les caractéristiques du jeu de données.



Contexte

Prétraitement des données

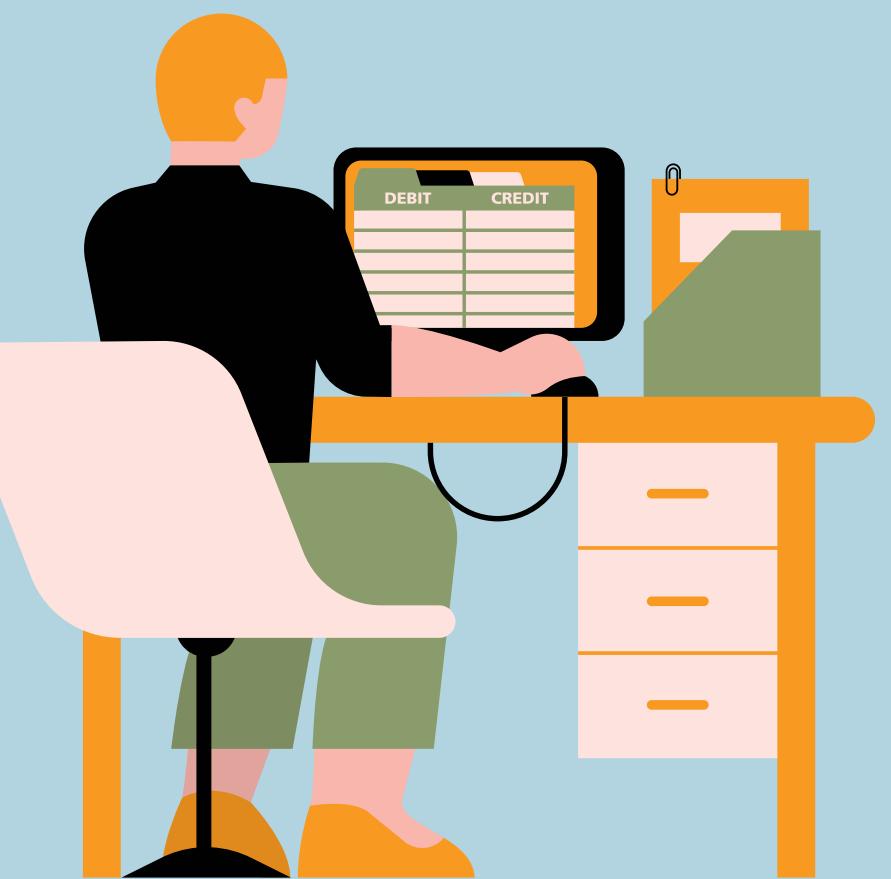
Nous avons effectué des étapes de prétraitement générales, y compris la conversion en minuscules, la suppression des mots vides, la suppression des caractères alphanumériques et la suppression de la ponctuation.



Contexte

Division des données

Nous avons divisé le jeu de données en deux parties : un ensemble d'entraînement et un ensemble de test



Contexte

Choix du modèle

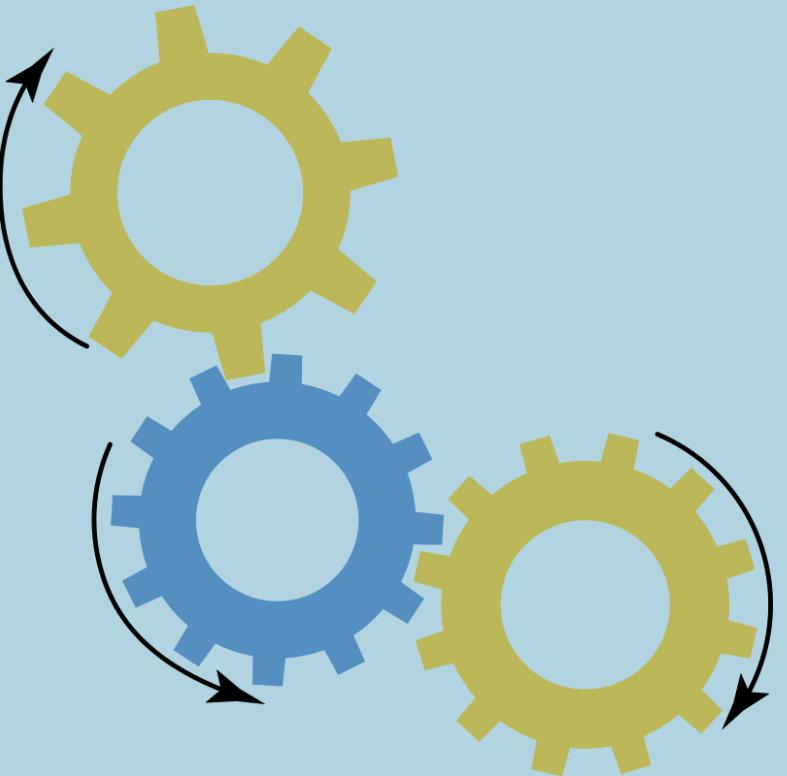
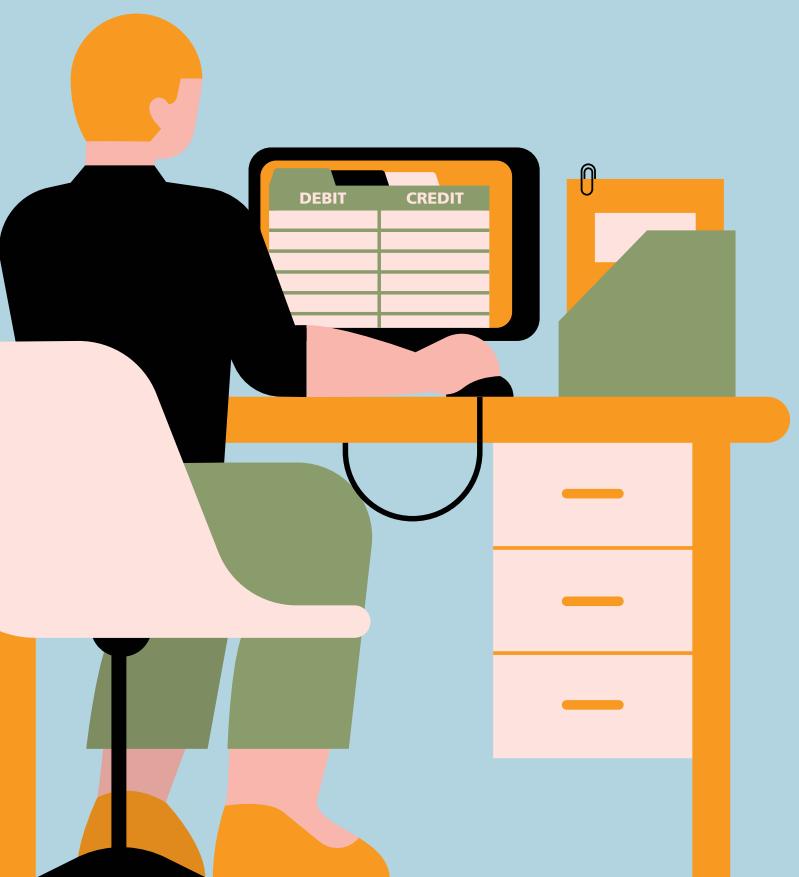
Nous avons choisi d'utiliser des modèles de classification pour attribuer des étiquettes catégoriques aux documents textuels.



Contexte

Entraînement du modèle

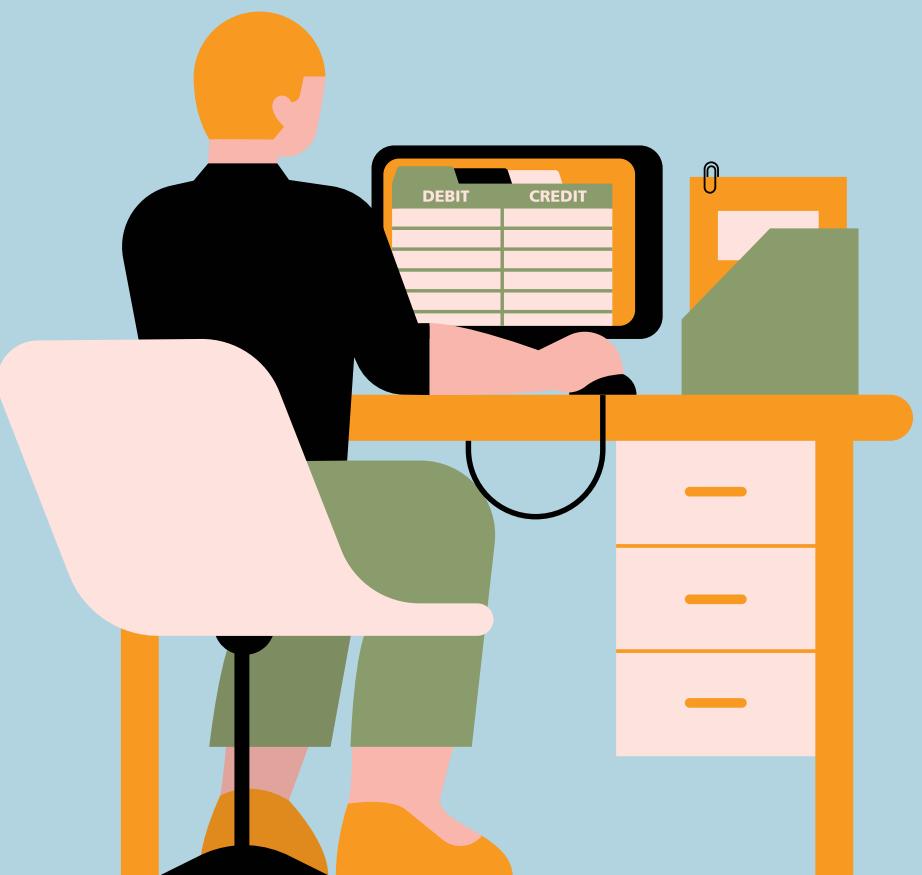
Nous avons formé les modèles sur l'ensemble d'entraînement en utilisant les vecteurs de texte et les étiquettes correspondantes.



Contexte

Évaluation du modèle

Nous avons évalué les performances des modèles en utilisant l'ensemble de test.



Contexte

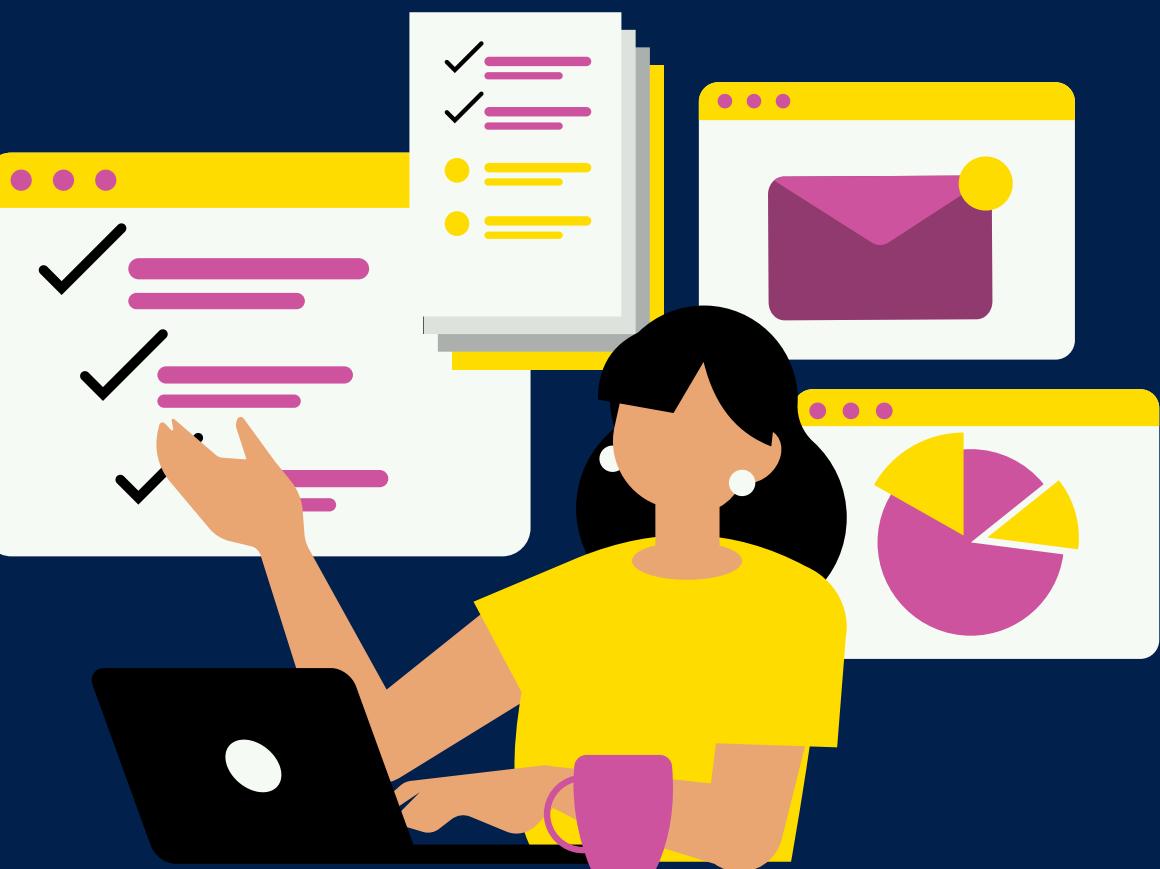
Sélection du modèle final

Nous avons choisi le modèle qui présentait les meilleures performances en termes de précision sur l'ensemble de test.



Technique

KNN
(K-Nearest Neighbors)



Tf-idf

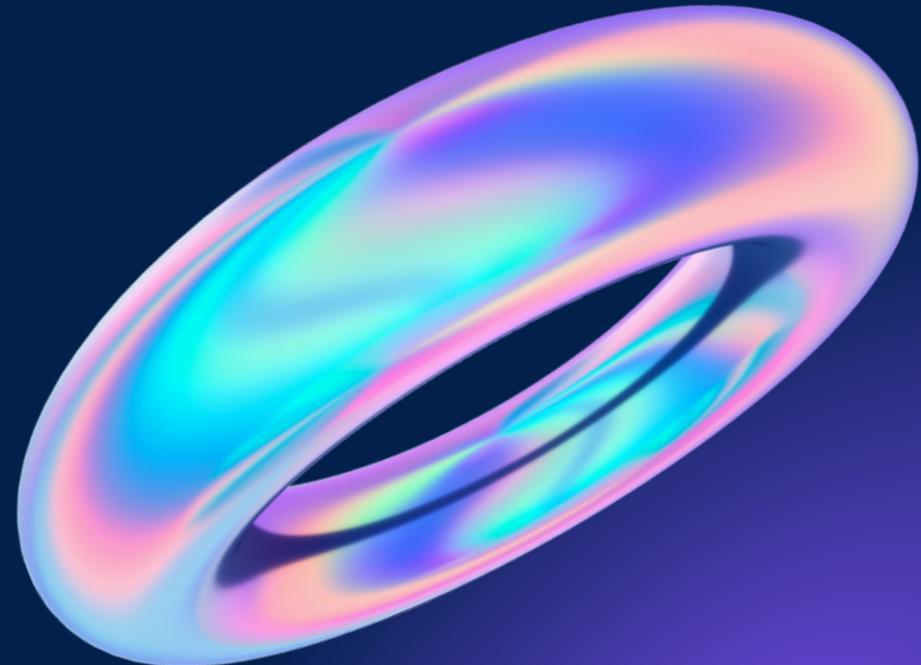
DT
(Decision Trees)

bibliothèque

- Scikit-learn
- pandas
- numpy



Réalisation



CONCLUSION ET PERSPECTIVE



Un grand Merci

Pour votre présence, vos délicates
attentions, qui ont rendu cette journée
parfaite et innoubliable.

