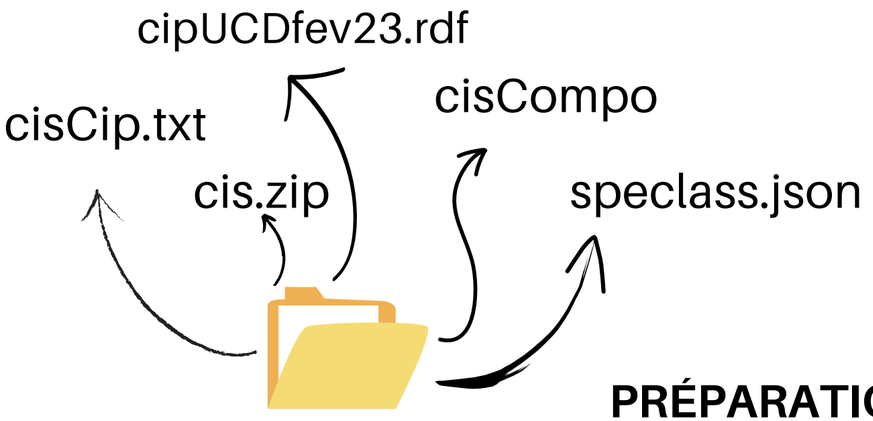
Données sur les médicaments du marché français

Introduction à la Data Science

1. INTRODUCTION
2. PROBLEMES
3. SOLUTIONS
4. REPARTITION DES TÂCHES
5. CONCLUSION

Introduction

Phases du projet



1

TRAITEMENT DES DONNÉES

- Modifier le format
- Extraire les données

2

PRÉPARATION DES DONNÉES

- Respecter les caractères accentués
- Respecter les contraintes sur les colonnes

3

INTEGRATION DES DONNEÉS

- Opérer les jointures des **DataFrame**

4

APPRENTISSAGE

- Utilisation de **Spark MLib** sur les données nettoyées

5

VISUALISATION

- Affichage de **l'arbre décisionnel**

Problèmes

PREPARATION DES DONNÉES

- **Charger** les données dans **DataBricks** (encodage)
- **Joindre** les colonnes obtenues à partir du **.nt** avec la colonne **EphMRA**

NETTOYER LES DONNÉES

- Définition des filtres pour **nettoyer** les jeux de données
- **Récupération** des lignes corrompues

MEA

- Comprendre l'**aspect fonctionnel** des données ingérées
- Choisir les **clés primaires**

SPARK MLIB

- Comprendre la structure d'un **pipeline ML**
- Comprendre l'approche par **arbre de décision**

Solutions

Chargement des données

Conversion en UTF-8

iconv

Nettoyage des données

Récupération des lignes corrompues

Nettoyage des jeux de données

filter, when, etc.

badRecordPath

Modélisation

CIP7 ⇔ CIP13
CIS ≠ CIP

Pipeline ML

Stages

Input DataFrame

Transformer/Estimator

Model

Annexe

CIS	CIP7	CIP13	RAISON
68049089	null	3400938135480	Corruption CIP7 null
66707752	3330784	3400933307844	Corruption prix de base négatifs
66770635	3798346	3400937983464	Corruption prix finaux négatifs

Répartition des tâches

- **Cédric**
  - cisCip
  - cisCompo
  - partie ML
- **Frédéric**
  - partie ML
  - cisCip
  - cisCompo
- **Amyr**
  - cipUDCfev23
  - cisCip
  - modélisation