M2 LID – UGE
Projet Introduction à la Data Science

Le projet se déroule lors des séances de TP (soit 6 séances). Il est possible que pour certains groupes les 12 heures de TP ne soient pas suffisantes, il vous faudra alors travailler en dehors de séances de TP. Vous travaillerez en groupe de 4 étudiants max. Vous devez constituer les groupes lors de la première séance de TP et élire un responsable du projet.

Il faudra me fournir le lien du git de votre groupe (de préférence gitlab, mon id est @ocure). Vous devez me fournir les droits de *maintainer*. J'attends que chaque séance de TP donne lieu à au moins un commit git. L'absence de commit sur une séance entrainera une pénalité sur la note finale du projet. Vous allez travailler, en partie, sur Databricks Community edition. Il faudra publier le workspace du projet à la fin de chaque séance et faire un commit comportant l'URI de la publication à la fin de chaque séance.

Le README.md du git devra contenir: les noms et prénoms des membres du groupe, l'état du projet (son bon fonctionnement, les fonctionnalités supportées, le choix des méthodes de ML adoptées, etc.), les difficultés rencontrées et les solutions apportées, les réponses aux différentes questions posées dans l'énoncé du projet, une conclusion, une évaluation interne de l'implication de chacun des membres (entre -2 et +2).

Soutenance du projet: le 29 mars à partir de 13:45. Chaque membre du groupe devra mettre en évidence ses contributions. Des notes individuelles seront attribuées à chaque membre du groupe. Chaque soutenance durera environ 15'.

Contexte:
Le domaine du projet concerne des données sur les médicaments du marché français. Vous allez effectuer des tâches relatives aux traitements des données, du type préparation et intégration de données, apprentissage depuis ces données et visualisation des résultats.
Sur la page du cours, vous trouverez plusieurs jeux de données.
Développement principalement en Scala sur la plateforme Apache Spark. D'autres outils peuvent être utilisés pour certaines transformations de données et autres tâches. La partie de programmation concernant Spark devra être, en partie, réalisée sur Databricks (community Edition). En plus de cela, vous pouvez fournir du code Spark qui tournerait sur Spark-shell ou bien du code dans un IDE.

Séances ETL
Le fichier `cipUCDfev23.rdf` est sérialisé en RDF/XML. A partir de ce fichier, vous devez extraire les données suivantes : codes CIP13 et UCD13 (des identifiants de médicaments), le nom du médicament, le nom du laboratoire titulaire et le code EphMRA.
Le traitement depuis la sérialisation RDF/XML n'est pas évident. Il serait judicieux de passer par un autre format, plus facile à traiter par Spark. La transformation d'une sérialisation RDF à une autre peut être réalisée à l'aide d'un outil extérieur. Depuis le format RDF de sortie que vous aurez sélectionné, vous allez extraire des données qui prendront la forme de l'extrait de la figure ci-dessous:

cip13	nom	labo	ephMRE	ucd13
3400930088197	"FEMELIS CPR BT60"@fr	SERP	G02X9	3400894316701
3400958903458	"WEGOVY 0,25MG INJ STYLO0,5ML 4"@fr	NOVO NORDISK	A10S	3400890020947
3400932805433	"TOBREX 0,3% COLLY FL5ML"@fr	NOVARTIS PHARMA	S01A	3400891255362
3400936039865	"PIROXICAM IVX 10MG GELU BT30"@fr	TEVA SANTE	M01A1	3400892432533
3400955715696	"DAFALGAN CODEINE CPR BT100"@fr	UPSA SAS	N02B	3400891479287
3400930224649	"ATORVASTATINE BGA 40MG CPR F30"@fr	BIOGARAN	C10A1	3400894547488
3400933308445	"GLUCOSE 10% BLZ INJ P.500ML 1"@fr	BIOLUZ	K01B3	3400891289251
3400949998685	"MONTELUKAST EG 10MG CPR BT28"@fr	EG LABO	R03J2	3400893891001
3400930030974	"ALPHACUTANEE EXT FL+FL 5"@fr	LEURQUIN MEDIO.	V03H	3400890034319
3400949386253	"ATORVASTATINE GNR 20MG CPR B90"@fr	SANDOZ	C10A1	3400893695401
3400949508792	"ARAVA 10MG CPR FL30 ADP DFD"@fr	DIFARMED	M01C	3400893962121
3400930143988	"ETORICOXIB BGA 60MG CPR BT28"@fr	BIOGARAN	M01A3	3400894407522
3400936523708	"PREDNISONE SDZ 20MG CPR BT20"@fr	SANDOZ	H02A2	3400892657059
3400930297773	"DESACE INJ AMP2ML 4"@fr	ETHYPHARM LABOR	C01A1	3400890255103
3400932026494	"MINIPHASE CPR 21 X3"@fr	BAYER HEALTHCAR	G03A3	3400890591768
3400930159538	"BRONPAX AD. SIR FL180ML"@fr	BIOCODEX	R05D2	3400890138147
3400938112597	"ERCEFURYL 200MG GELU BT12"@fr	OPELLA HEALTH F	A07A	3400890319768
3400934088360	"KETREL 0,050% CR TB30G"@fr	BAILLEUL	D10A	3400891880571
3400927579776	"MEMANTINE SDZ 10MG CPR BT56"@fr	SANDOZ	N07D9	3400894242857
3400930203767	"ACID.URSODESOX.TVC 500MG CPR60"@fr	TEVA SANTE	A05A2	3400890008099

Le traitement du format RDF permettant d'obtenir la table ci-dessus doit être effectué avec Apache Spark, avec l'abstraction de votre choix.
Vous devrez fournir un schéma du type `cip13 Long, nom String, labo String, ephMRA String et ucd13 Long.` Concernant ce schéma, faire en sorte que les données de colonnes cip13 et ucd13 ne puissent être nulles.
Question **Q1**: Motiver le format RDF sélectionné et le choix de l'abstraction Spark.
Question **Q2**: combien avez-vous de tuples dans votre résultat.

Vous disposez d'un autre fichier, `cisCip.txt,` qui doit respecter le schéma suivant:
- Code CIS
- Code CIP7 (Code Identifiant de Présentation à 7 chiffres
- Libellé de la présentation
- Statut administratif de la présentation
- Etat de commercialisation de la présentation tel que déclaré par le titulaire de l'AMM
- Date de la déclaration de commercialisation (format JJ/MM/AAAA)
- Code CIP13 (Code Identifiant de Présentation à 13 chiffres)
- Agrément aux collectivités ("oui", "non" ou « inconnu »)
- Taux de remboursement (avec un séparateur « ; » entre chaque valeur quand il y en a plusieurs)
- Prix de base du médicament en euro
- Prix final du médicament en euro
- Montant de la taxe du médicament en euro
- Information sur le médicament

Vous devez effectuer un travail de garantie de la qualité des données sur ce fichier. Par exemple, vous devez vous assurer du respect du format des dates, les champs cis, cip7 et cip13 ne doivent pas être nuls et sont des entiers, le prix des médicaments ne doit pas être négatifs, de respect des caractères accentués, correction du prix final, etc.
La validation d'un schéma (et donc des types des colonnes) est une première approche pour garantir cette qualité.
Vous pouvez tester les différentes approches pour l'option '*mode*' lors du chargement de votre abstraction Spark. Vous pouvez également expérimenter avec les options `badRecordsPath` et `columnNameOfCorruptedData`.
Vous devez fournir en annexe, les enregistrements qui ne satisfont pas vos règles de qualité en fournissant l'identifiant pertinent suivant la situation (par le cip7, cip13 ou cis) et le type de l'anomalie.

Question **Q3**: en fonction de la répartition des données, est-il possible de supprimer le symbole '%' de la colonne taux de remboursement. Présenter clairement votre méthode.

Vous devez effectuer plusieurs filtres:
- ne garder que les tuples dont le statut correspond à une présentation active
- ne garder que les médicaments dont l'état correspond à une déclaration de commercialisation
- ne garder que le prix final

On s'intéresse maintenant à un troisième fichier contenant la composition des médicaments. Le fichier se nomme `cisCompo` et présente le schéma suivant:
```
Code CIS
Désignation de l'élément pharmaceutique
Code de la substance
Dénomination de la substance
Dosage de la substance
Référence de ce dosage
Nature du composant
Numéro de liaison SA/FT
```

Vous ne devez considérer qu'un sous-ensemble des compositions: celui comportant les médicaments qui ont été sélectionnés depuis le fichier `cisCip.txt`, aux médicaments qui ne présentent pas une voie d'injection, perfusion au sens large (attention, réaliser une sélection fine) et avec une nature 'SA'.
Pour vous aider dans la démarche de suppression des médicaments par voie d'injection, vous pourrez vous aider du contenu du fichier `cis.zip`.

Question **Q4**: Combien de tuples garder vous depuis la source `cisCompo`?

Le dernier fichier (`speclass.json`) a intégrer se présente sous le format d'un fichier JSON. Il regroupe deux colonnes: un code cip7 et un identifiant d'une classe thérapeutique. Vous devez charger ce fichier dans votre environnement.

Vous disposez maintenant de cinq fichiers. Vous devez réaliser l'intégration des données de ces cinq fichiers.

Question **Q5**: Proposez une modélisation (sans perte d'information) sous la forme d'un diagramme entité association des données des différentes tables.