UNIVERSITÉ DE SHERBROOKE
DÉPARTEMENT D'INFORMATIQUE

IFT 870 BIN 710 - Forage de données

TP#2 : Prétraitement et représentation de données

Hiver 2020

Le but de ce devoir est de pratiquer le prétraitement et la représentation de données : auscultation, nettoyage, intégration, réduction.

Ce devoir est à faire en équipe de deux obligatoirement. Il devra être complété avant le vendredi 28 février 2020 à 23h59. Vous devez remettre, sur `opus.dinf.usherbrooke.ca`, un fichier Ipython notebook (nommé `tp2.ipynb`) contenant votre rapport et vos scripts Python pour ce devoir.

Description des tâches à réaliser : Base de données des codes de médicaments au É-U

On vous fournit un jeu de données composé de deux tables au format `csv` : `product.csv` et `package.csv`. Vous pouvez trouver la description des attributs de ces tables aux adresses `https://www.fda.gov/drugs/drug-approvals-and-databases/ndc-product-file-definitions` et `https://www.fda.gov/drugs/drug-approvals-and-databases/ndc-package-file-definitions`. Pour traduire ces pages en français, utilisez `https://translate.google.ca/`.

1. Auscultez les données et présentez un résumé de votre auscultation ;
2. Listez toutes les relations observées entre les attributs (informations communes, corrélations) ;
3. Détectez et corrigez les incohérences entre des valeurs d'attributs dans les deux tables ;
4. Complétez au maximum les données manquantes dans les deux tables (Attention, tout ne peut pas être complété !) ;
5. Détectez et retirez les objets dupliqués dans les deux tables ;
6. Intégrez les deux tables et nettoyez le résultat (données dupliquées, incomplètes, incohérentes, erronées etc.) ;
7. Proposez un nouvel ensemble d'attributs (représentation) qui élimine la redondance des informations dans les valeurs des attributs ;
8. À partir de la nouvelle représentation, proposez un ensemble d'attributs à utiliser pour prédire toutes les classes pharmaceutiques d'un médicament (attribut PHARM_CLASSES) ;
9. Appliquez un modèle de classification pour prédire les classes pharmaceutiques des médicaments pour lesquels l'information est manquante ;

10. Évaluez un échantillon de vos résultats à l'aide de connaissances d'experts (`Google` est notre expert !)

Remise du travail

Pour soumettre votre travail, connectez-vous, dans un fureteur, au serveur `http://opus.dinf.usherbrooke.ca` en utilisant votre CIP, puis choisissez le cours `IFT870 (BIN710)` et le projet `TP2`. Chargez votre fichier `tp2.ipynb` et soumettez-le. Le nom de votre fichier de remise doit être exactement `tp2.ipynb`. Indiquez bien les noms des deux membres de l'équipe dans le fichier. Ne faites qu'une seule soumission par équipe. Ne remettez pas d'autre fichier.