

Application de techniques de Data Mining en utilisant le logiciel RapidMiner

Introduction

1 Classification de spam

Pour la création des fichiers CSV utilisé pour la génération de règles d'associations, un nettoyage des données a été fait.

Les lignes qui ont des données vides sont filtrées, comme par exemple le nom du produit. Quand la quantité (Quantity) est plus petite que zéro, les lignes sont aussi filtrées.

Un algorithme a aussi été fait pour détecter des noms de produit incohérent et pour filtrer les lignes. Voici la liste des noms incohérents trouvés : *wet pallet, sold in set ?, wet damaged, faulty, posy candy bag, michel oops, wet/mouldy, lost, damaged, daisy notebook, fba, taig adjust, water damage, ribbons purse, ?display ?, found in w/hse, damaged stock, samples, mixed up, ???missing, breakages, chilli lights, dotcom set, re-adjustment, packing charge, amazon sales, ?, mix up with c, owl doorstop, stock check, mia, wrap folk art, carriage, damages, ?lost, john lewis, found box, 20713, show samples, damages wax, website fixed, jumbo bag owls, wet boxes, frog candle, found, given away, mouldy, ???, broken, wrong code, damages ?, display, wet, bunny egg box, polkadot pen, ?missing, dotcom, amazon adjust, damaged, returned, toybox wrap, smashed, ?sold as sets ?, showroom, led tea lights, wet/rusty, missing ?, wrong barcode, lost in space, crushed ctn, popcorn holder, dotcom sales, lost ??, sold as 1, bingo set, thrown away., crushed boxes, test, check, crushed, wet ?, mailout, amazon, rain poncho, adjust, ebay, adjustment, spotty bunting, wet rusty, cracked, sale error, ???lost, water damaged, missing, garage key fob, dotcom stock, dotcom adjust, sold as 22467, ??, dotcom postage, jumbo bag toys, shoe shine box, ?? missing, on cargo order, wrong code ?, check ?, label mix up, postage, wrap carousel, can't find, wrongly marked, retrospot lamp, thrown away, counted, ???missing, cordial jug*

De plus, chaque donnée de colonne est systématiquement trimée pour éviter tout problème et pour faciliter la suite du traitement.

Voici des données qui sont calculées lors du filtrage :

- Nb ligne dans le fichier : 541909
- Nb ligne après nettoyage : 397924
- Nb produit : 3639

Nous avons donc filtré 143985 lignes.

Un autre algorithme a été fait pour standardiser la description des produits en liant la description avec le numéro du produit et l'on garde la description qui est la plus utilisée.

Pour voir une partie de l'algorithme, il faut regarder la fonction `Main.resolveProductDescription()` dans le code Java.

Voici d'autres informations que nous calculons lors de la création des nouveaux fichiers CSV

Création du fichier par pays (permet de tester rapidement le code) :

- Nb ligne in new file (byCountry.csv) : 37
- Processing Time : 00 :01.625
- Total time for create the file : 00 :01.656

Création du fichier par client :

- Nb ligne in new file (byCustomer.csv) : 4339
- Processing Time : 01 :44.361
- Total time for create the file : 01 :44.674

Création du fichier par facture :

- Nb ligne in new file (byInvoice.csv) : 18536

- Processing Time : 07 :26.901
- Total time for create the file : 07 :27.901

Afin de pouvoir générer ces données nous avons utilisé le multi threading ce qui nous permet de diviser par deux le temps de création des fichiers.

Pour lire le fichier, nous avons dû en créer un nouveau et copiant et collant son contenu. Nous pensons qu'il doit avoir un problème d'encodage.

Pour tester le programme il suffit de lancer la commande `mvn clean install`, de modifier le fichier `application.properties`, pour définir où se trouve le fichier de base, et où l'on veut créer les nouveaux fichiers.

La génération de ces fichiers n'est pas simple et nous a pris beaucoup de temps. Nous pourrions encore améliorer la génération, en tenant compte des produits retournés, par exemple. Il serait aussi bien d'avoir un meilleur descriptif de ces données.

1.1 Questions sur la classification

Dans le bloc "Process Documents from Data" nous n'avons pas mis d'étape de stemming. Est-ce que l'ajout de ce préprocessing a un impact sur les résultats obtenus ?

Dans l'exemple ci-dessus, nous avons utilisé un classificateur bayésien, veuillez essayer d'autres familles de classificateurs, quel est l'impact sur le résultat obtenu ?

Finalement vous utiliserez la seconde source de données (`emails.zip`) sur laquelle vous appliquerez le même processus. Que constatez-vous en comparant les 2 résultats ?

2 Market basket analysis

2.1 Questions about association rules

Constatez-vous des différences dans les règles d'associations obtenues entre les 2 regroupements différents (par facture/par client) ? Veuillez commenter vos résultats.

Est-il possible de générer une/des autre/s colonne/s à partir des données initiales qui produisent des règles intéressantes ?

3 Questions sur les règles d'association

Commentez les résultats obtenus par rapport aux étiquettes existantes sur le dataset.

Quelle est l'influence des différentes étapes de "text processing" sur le résultat que vous obtenez ?

4 Conclusion