WEM : Web Mining
Laboratoire n°2
Application de techniques de *Data Mining* en utilisant le logiciel *RapidMiner*

23.03.2018

Objectifs

Ce laboratoire a comme objectif d'appliquer différentes techniques de data mining sur des ensembles de données issus du web en utilisant le logiciel *RapidMiner Studio*[1]. Il s'agit d'une plateforme de traitement, de modélisation et d'analyse de données permettant de réaliser des tâches de prétraitement (lecture, nettoyage, transformation, réduction, etc.) et de conceptualisation de systèmes permettant d'appliquer des algorithmes de data mining (clustering, règles d'association, classification, etc.) et d'évaluer les résultats obtenus. Ce logiciel existe dans une version communautaire, gratuite mais limitée et une version payante. Une licence académique gratuite est disponible pendant une année pour les étudiants.

Les points étudiés dans ce laboratoire seront :

- Prise en main de l'outil *RapidMiner*
- Modélisation d'un filtre anti-spam
- Market-basket analysis sur des achats d'un site de vente en-ligne
- Une analyse sur des commentaires à l'aide de *WordNet*[2]

Durée

- 6 périodes. A rendre le **20.04.2018** à 8h30 au plus tard.

Références

- Cours «Web Mining» de Nastaran Fatemi et Laura Raileanu
- Livre «Data Mining the Web: Uncovering Patterns in Web Content, Structure, and Usage» de Zdravko Markov et Daniel T. Larose
- Livre «Text Data Management and Analysis » de ChengXiang Zhai et Sean Massung

Donnée

Dans un premier temps, il vous faudra installer le logiciel *RapidMiner Studio*, vous devrez créer un compte et aller sur le site de l'éditeur pour activer votre licence académique. Le logiciel propose une

[1] https://rapidminer.com/
[2] http://wordnet.princeton.edu/

série de tutoriels (menu Help -> Tutorials), nous vous encourageons à suivre ceux des catégories « Basics » et « Modeling, Scoring and Validation ».

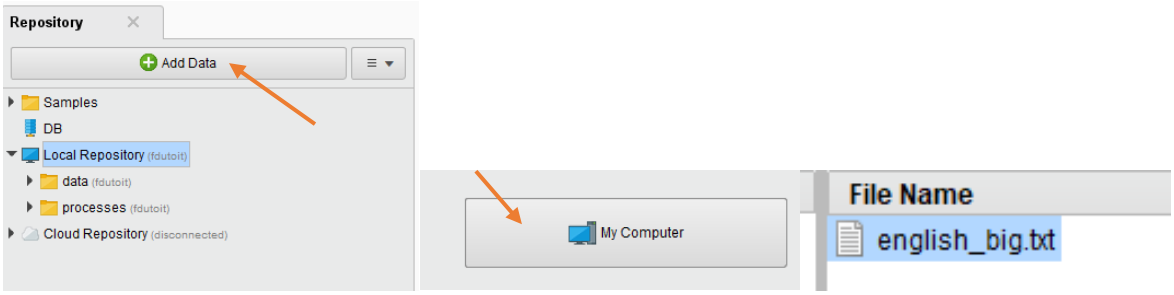
Avant de pouvoir commencer avec les manipulations de ce laboratoire, il vous faudra installer deux extensions pour *RapidMiner* :

- Wordnet extension
  https://marketplace.rapidminer.com/UpdateServer/faces/product_details.xhtml?productId=rmx_wordnet
- Text Processing
  https://marketplace.rapidminer.com/UpdateServer/faces/product_details.xhtml?productId=rmx_text

1. Classification de spam

Dans cette première partie nous allons modéliser un filtre anti-spam. Pour réaliser ceci nous vous mettons à disposition deux ensembles de données, un premier regroupant des SMS labélisés et un second des e-mails aussi labélisés. Nous allons vous guider pour la mise en place du premier *process* de *RapidMiner*.

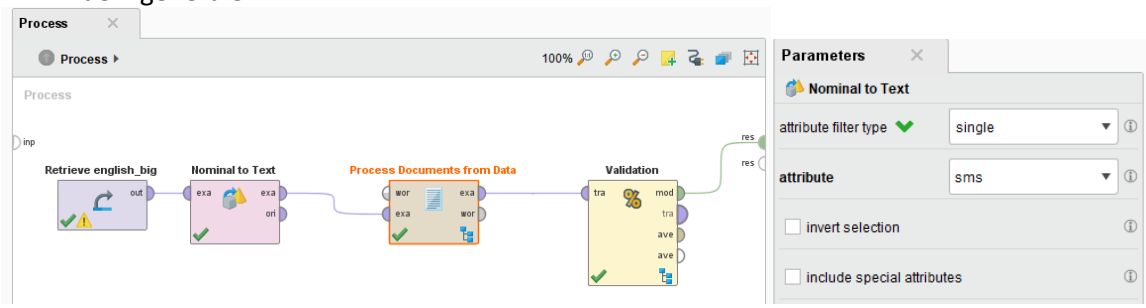
1. La première étape consiste en l'importation des données dans le logiciel.



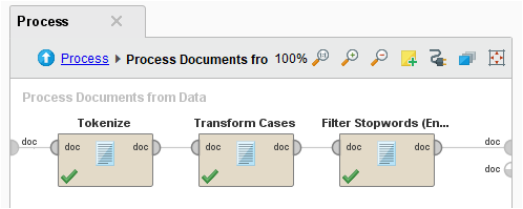
2. Sur la page « Format your columns », vous changerez le rôle de la colonne *class* en *label*.
3. Vous sauverez ensuite vos données dans le dossier data du « Local Repository ».
4. Une fois l'importation terminée, *RapidMiner* ouvrira un tableau comportant les données importées. Veuillez vérifier que seules 2 classes sont présentes dans la colonne *class*.
5. Cliquez sur l'onglet « Design » pour retourner sur l'interface de création du processus.

Nous allons à présent créer notre premier processus de classification. Vous trouverez ci-dessous les différents blocs permettant cette tâche de classification :

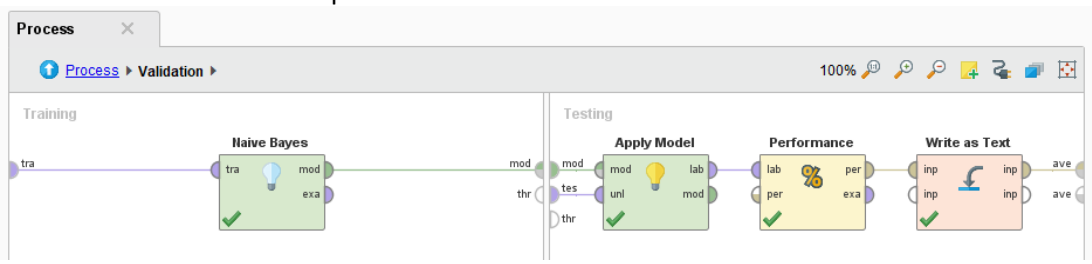
- Vue « générale » :



- Vue détaillée du bloc « Process Documents from Data » :



- Vue détaillée du bloc « Split Validation » :



Quelques remarques/informations supplémentaires :

- Le bloc « Nominal to Text » est configuré en mode « single », il va forcer le typage de la colonne *sms* en texte.
- Le bloc « Process Documents from Data » produira par défaut un vecteur de type TF-IDF, il est possible de le paramétrer pour avoir d'autres types de vecteurs.
- Le bloc « Split Validation » est composé de deux parties, la première (à gauche) va générer un modèle à partir du *training set*, la seconde à droite appliquera le modèle au *test set* et évaluera sa performance.
- Le bloc « Write as Text » doit être configuré, vous devrez indiquer un fichier de sortie dans lequel la précision du modèle ainsi que la matrice de confusion seront indiquées.

Questions sur la classification :

1. 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 ?
2. 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 ?
3. Finalement vous utiliserez la seconde source de données (*emails.zip*) sur laquelle vous appliquerez le même *process*. Que constatez-vous en comparant les 2 résultats ?

2. Market basket analysis

Dans cette seconde partie nous allons nous intéresser à un problème de *market basket analysis*, les données que nous vous proposons regroupent l'ensemble des ventes (transactions) durant une année complète d'un site de vente en ligne de cadeaux. Nous souhaitons générer des règles d'associations par rapport à ces ventes. Vous pouvez consulter les sources indiquées dans le README fourni avec les données pour plus d'informations.

Les données sont fournies sous la forme suivante : sur chaque ligne nous trouvons le détail de la vente d'un produit : `InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID, Country`

- InvoiceNo : n° de facture
- StockCode : n° de produit
- Description : nom du produit
- Quantity : quantité achetée

- InvoiceDate : date de facturation
- UnitPrice : prix unitaire
- CustomerID : n° de client
- Country : pays de résidence du client

Pour pouvoir travailler avec des règles d'associations, nous avons besoin de données transactionnelles composées uniquement de variables binaires. Nous ne pouvons donc pas utiliser ces données directement, votre première tâche consistera donc à les prétraiter.

1. Veuillez réaliser un simple logiciel qui prendra en entrée le fichier brut « data.csv » et produira en sortie un CSV utilisable pour la génération de règles d'associations.
2. La mise en forme minimale attendue est la suivante :

```
InvoiceNo, Product1, Product2, Product3, Product4, Product5, Product6, …, ProductN
Facture1, true, false, false, true, false, true, … , false
Facture2, false, true, false, true, false, false, … , false
Facture3, false, true, false, false, true, true, … ,true
```

   Vous regrouperez sur chaque ligne une facture avec un booléen à `true` pour tous les produits achetés. Vous devrez mettre en en-têtes tous les produits vendus par le magasin. Nous vous conseillons de garder le nom complet du produit comme nom de colonne afin de faciliter l'interprétation des résultats.

3. Ces données ne sont pas forcément très propres. Il existe par exemple des numéros d'articles qui correspondent à plusieurs noms différents (mal-orthographié, nom raccourci, taille différente, etc.). Il existe aussi des noms de produits qui ont visiblement été introduits manuellement : « ? », « ?? », « ???MISSING », « ?MISSING », « DAMAGED », « DAMAGES », « DAMAGES ? », etc. Veuillez identifier et nettoyer ce genre de cas dans les données. Nous ne vous demandons pas de corriger l'intégralité des données mais il faudra au minimum corriger les erreurs les plus fréquentes.

Le bloc « Create Association Rules » permet comme son nom l'indique de générer des règles d'associations. Vous devrez le précéder d'un bloc « FP-Growth » qui sélectionnera les éléments fréquents à partir du jeu de données. Une fois les règles d'association générées, vous commenterez votre paramétrage et les résultats obtenus.

Vous aurez noté que dans les données fournies, en plus du numéro de facture, nous trouvons le numéro de client de l'acheteur. Un client peut avoir commandé plusieurs fois sur le site des ensembles d'articles différents. Veuillez adapter votre programme de prétraitement des données afin de regrouper les ventes par client. La forme minimale attendue est la suivante :

```
ClientNo, Product1, Product2, Product3, Product4, Product5, Product6, …, ProductN
Client1, true, false, false, true, false, true, … , false
Client2, false, true, false, true, false, false, … , false
Client3, false, true, false, false, true, true, … ,true
```

On regroupera donc sur une ligne des articles achetés en plusieurs fois (différentes factures).

Questions sur les règles d'association

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. Utilisation d'un WordNet sur des commentaires d'utilisateurs

Pour cette dernière partie, vous aller définir vous-même le processus d'analyse. Le but de cette manipulation est de classifier (positivement ou négativement) des commentaires à l'aide de *WordNet*. Nous vous avons fait installer l'extension en début de laboratoire, vous trouverez plus d'informations sur la page de présentation de celle-ci: https://community.rapidminer.com/t5/RapidMiner-Text-Analytics-Web/Sentiment-Analysis-using-Wordnet-Dictionary/ta-p/31664

L'ensemble de données conseillé pour cette manipulation est composé de 3'000 commentaires issus de 3 sites Internet majeurs. Ceux-ci sont déjà étiquetés avec 1 (commentaire positif) ou 0 (négatif). Vous devrez créer un process complet sur *RapidMiner* permettant d'ajouter un attribut « sentiment_wordnet » aux données existantes. Vous commenterez ensuite 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 ?

Rendu/Evaluation

Vous remettrez sur *Moodle* un zip contenant les sources, les libraires utilisées, les éventuels fichiers d'entrée, etc. Vous ajouterez en plus dans votre rendu les fichiers de configuration utilisés dans *RapidMiner Studio* (menu View -> Show Panel -> XML) et un petit rapport dans lequel vous discuterez de vos différentes manipulations et des résultats obtenus, de vos choix d'implémentation et répondrez aux questions posées.

Vous pouvez discuter entre les groupes mais il est strictement interdit d'échanger du code.

Adresse E-Mail de l'assistant : fabien.dutoit@heig-vd.ch

Bonne chance !