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

25.03.2022

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 des « pièges-à-clics »
- Analyse des sentiments sur des commentaires à l'aide de *WordNet*[2]
- Recommandation de films
- Règles d'association sur des achats d'un site de vente en ligne
- Regroupement (clustering) des applications sur *Google Play Store*

Durée

- 4 périodes encadrées. A rendre le **jeudi 07.04.2022** à **23h59** au plus tard.

Références

- Cours «Web Mining» de 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

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

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 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 trois 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
- Recommender system
  https://marketplace.rapidminer.com/UpdateServer/faces/product_details.xhtml?productId=rmx_irbrecommender

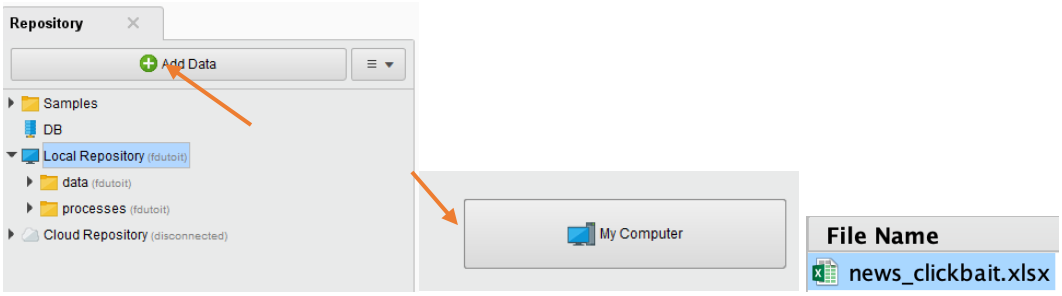
1. Classification des « pièges-à-clics »

L'objectif dans cette première partie est de modéliser un filtre pour des « pièges-à-clics » dans les médias d'information en ligne. Pour réaliser ceci, nous vous mettons à disposition un ensemble de données regroupant plus des 10'000 titres de la presse collectées en 2016. Chaque ligne de cet ensemble de données correspond à un seul titre, qui est décrit par les deux attributs suivants :

- *headline* : le texte représentant le titre
- *clickbait* : l'étiquette (label) identifiant le titre correspondant comme piège (1) ou non (0).

Nous allons vous guider pour la mise en place du premier *process* de *RapidMiner*.

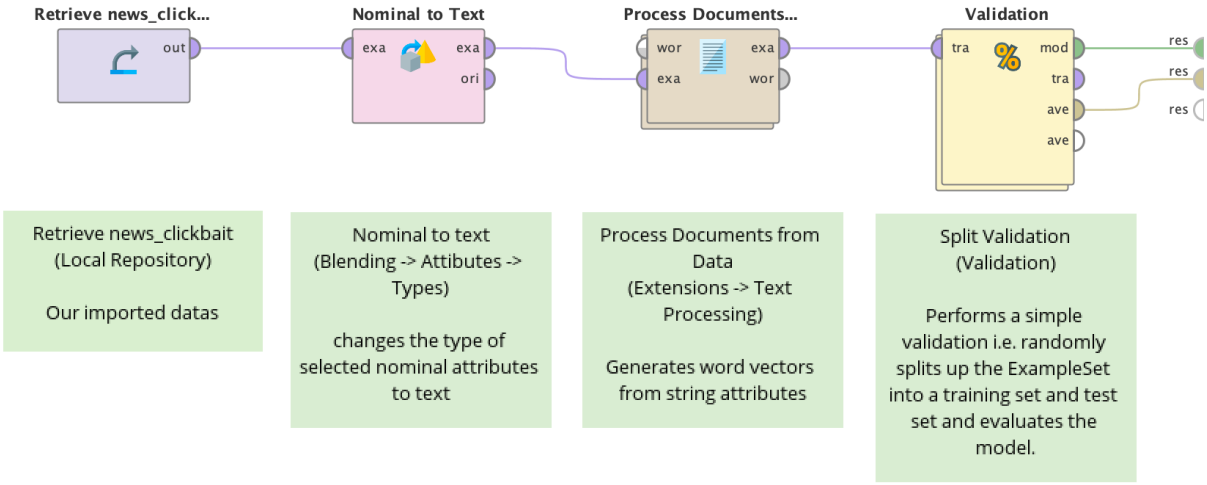
1. La première étape consiste en l'importation des données dans le logiciel.



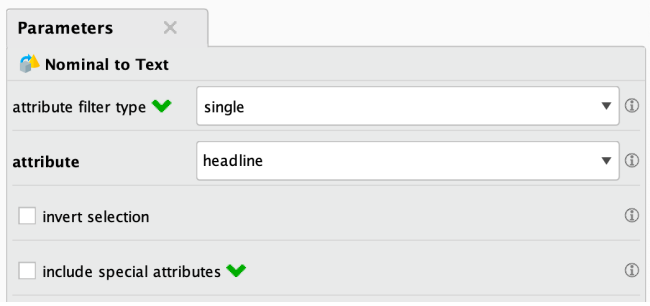
2. Sur la page suivante « Format your columns », vous changerez le rôle de la colonne *clickbait* en *label* et son type en *binomial*.
3. Vous sauverez ensuite vos données dans le dossier data du « Local Repository ».
4. Cliquez sur l'onglet « Design » pour retourner sur l'interface de création du processus.

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

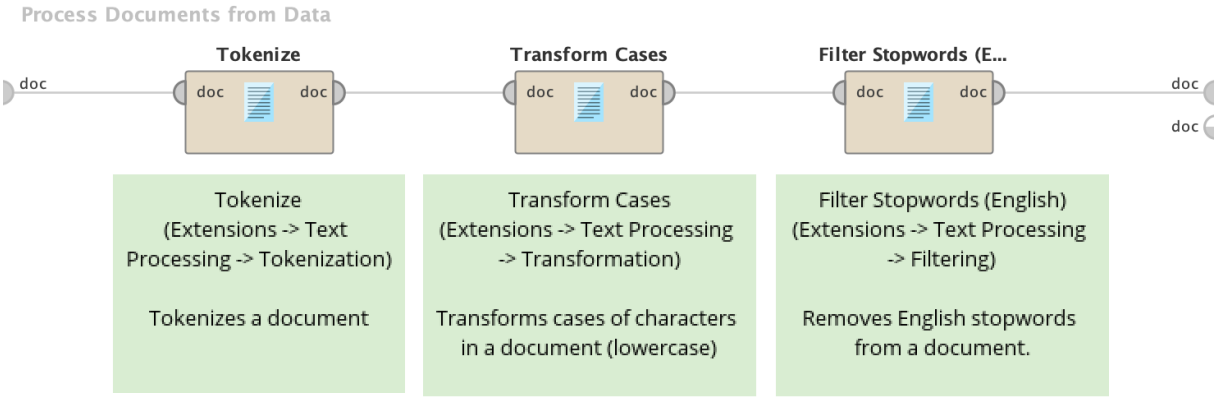
- Vue « générale » :



- Paramètres du bloc « **Nominal to Text** »
  On convertit le type de l'attribut *headline* en texte :

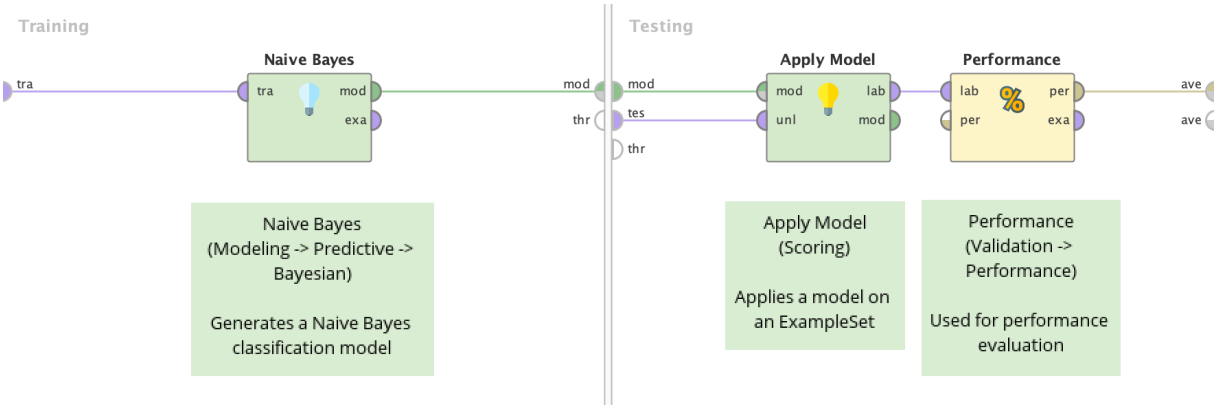


- Vue détaillée du bloc « **Process Documents from Data** »



Ce 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.

- Vue détaillée du bloc « **Split Validation** » :



Le bloc « Split Validation » est composé de deux parties, la première (à gauche) va générer un modèle à partir d'un ensemble d'apprentissage (*training set)*, la seconde (à droite) appliquera le modèle à l'ensemble de test (*test set)* et évaluera sa performance. Le paramètre *split ratio* détermine la proportion de données utilisées comme ensemble d'apprentissage.

Comme il s'agit d'une classification binaire, veuillez utiliser l'opérateur correspondant pour évaluer la performance du modèle. Outre l'*accuracy*, vous pourriez choisir des autres mesures comme le *recall* et *precision* par exemple.

Questions sur la classification :

1. Est-ce que le changement du paramètre *split ratio* influence la performance du modèle ? Veuillez expliquer votre réponse.

2. Veuillez changer le bloc « Split Validation » avec le bloc « Cross Validation » en utilisant le même classificateur pour l'apprentissage et les mêmes operateurs dans la partie d'évaluation. Qu'est-ce que vous pouvez constater des résultats correspondants ?

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

4. Jusqu'à présent nous avons utilisé un classificateur bayésien. Veuillez essayer d'autres familles de classificateurs, quel est l'impact sur le résultat obtenu ?

2. Analyse des sentiments sur des commentaires

Pour cette deuxième 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

Dans une première étape, nous allons travailler avec l'ensemble de données *CoronaTwitterComments_2labels.xlsx* contenant environ 3'000 commentaires liés au sujet COVID-19 issues de Twitter en Mars 2020. Ceux-ci sont déjà étiquetés avec 1 (commentaire positif) ou -1 (négatif). Vous devrez créer un processus 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 ?

Dans une deuxième étape, un ensemble de données *CoronaTwitterComments_3labels.xlsx* complété avec des commentaires neutres (étiqueté avec 0) va être considéré. Veuillez adapter le process créé précédemment pour 3 étiquettes (classes) et commenter les nouveaux résultats.

Remarque

Sur *Mac*, il se peut que vous rencontriez une erreur *IO* de type « too many files open in System », la raison est que si le bloc « Open WordNet Dictionnary » est placé dans le bloc « Process Document from Data » il sera exécuté pour chaque commentaire. La solution[3] est d'ouvrir une seule fois le dictionnaire à l'extérieur et de stocker sa sortie dans un bloc de type « Remember » et ensuite d'utiliser le bloc « Recall » pour reprendre le dictionnaire déjà chargé en mémoire. Dans le bloc « Recall » vous veillerez à décocher la case « remove from store ». Vous devriez aussi vous assurer que le stockage du dictionnaire Wordnet est exécuté avant le traitement du texte à l'aide de l'option « Change operator order » du menu « Process ».

3. Système de recommandation des films

Dans cette troisième partie nous allons réaliser un système de recommandation des films à l'aide de l'extension *Recommenders* que nous vous avons fait installer en début de laboratoire. Vous trouverez plus d'informations sur la page : http://bib.irb.hr/datoteka/596976.rcomm2012_recommenders.pdf

Nous allons nous intéresser sur l'approche collaborative pour effectuer la recommandation des films en utilisant l'ensemble des données proposé qui contient d'environ 9800 films notés par 610 utilisateurs de MovieLens.

Plus précisément, l'ensemble des données *movies.csv* qui décrit les films contient 3 attributs :

- *movieId* : identifiant du film
- *title* : le titre du film (avec l'année de sortie du film entre parenthèses)
- *genres* : les genres du film

L'ensemble des données *ratings.csv* contenant les notes des films données par les utilisateurs inclut 4 attributs :

- *userId* : identifiant de l'utilisateur
- *movieId* : identifiant du film

[3] https://community.rapidminer.com/discussion/46762/too-many-files-open-in-system-wordnet

- *rating* : la note de l'utilisateur pour le film correspondant
- *timestamp* : le timestamp de la note

Avant de commencer à travailler avec les opérateurs de recommandation, les deux ensembles des donnés *movies.csv* et *ratings.csv* doivent être joints (à l'aide du bloc « Join ») par rapport leur attribut commun – *movieId*. De plus, nous devons définir le rôle des attributs *movieId* et *userId* en tant que « item identification » et « user identification » respectivement.

La prochaine étape sera de séparer l'ensemble de données résultant (à l'aide du bloc « Split Data ») en deux parties, dont une va représenter ensemble d'apprentissage et l'autre l'ensemble de test. Construisez le premier système de recommandation sur l'ensemble d'entrainement à l'aide de l'opérateur « Item k-NN », qui appartient aux filtres collaboratifs pour la recommandation, et ensuite évaluez sa performance sur l'ensemble de test (en utilisant les blocs « Apply Model » et « Performance » de l'extension « Recommander »).

Questions :

1. Est-ce que le rapport de partition de l'ensemble des données initiales (bloc « Split Data ») influence la performance du modèle ? Veuillez expliquer votre réponse.

2. Qu'est-ce que vous pouvez constater si vous changez la valeur de k de l'opérateur « Item k-NN » ou si vous utilisé un k-nn pondéré ?

3. Veuillez changer le bloc « Item k-NN » avec le bloc « User k-NN ». Qu'est-ce que vous pouvez constater des résultats correspondants ? Que pouvez-vous constater en utilisant les autres approches de filtrage collaborative ?

4. Choisissez la méthode qui donne la meilleure performance, et, à l'aide du bloc « Model Combiner », combiner le avec la méthode « User k-NN ». Qu'est-ce que vous pouvez constater des résultats correspondants ? Les résultats changent-ils si « Item k-NN » est ajouté à la combinaison précédente ?

À l'aide des filtres collaboratives, nous pouvons également prédire la note d'un film que l'utilisateur donnerait en utilisant un processus similaire à celui construit précédemment. Toutefois, pour effectuer cette tâche de prédiction, il faut définir en plus le rôle de l'attribut *ratings* en tant que « label ». Après la répartition en ensembles d'apprentissage et ensemble de test avec un rapport de partition 90%-10%, construisez le modèle de prédiction d'entrainement à l'aide de l'opérateur « Item k-NN » qui appartient aux filtres collaboratifs pour la prédiction et ensuite évaluez sa performance sur l'ensemble de test en utilisant le bloc « Apply Model ».

Questions :

5. Qu'est-ce que vous pouvez constater sur les résultats obtenus ?

6. En évaluant la performance du modèle construit, quelles sont vos observations sur les résultats en changeant les paramètres du k-nn (notamment la valeur de k et le mode de corrélation) ?

7. Veuillez changer le bloc « Item k-NN » avec le bloc « User k-NN ». Qu'est-ce que vous pouvez constater des résultats correspondants ? Qu'est-ce que vous pouvez constater en utilisant les autres approches de filtrage collaborative pour la prédiction ?

4. Règles d'association sur des achats en ligne

Dans cette quatrième partie nous allons nous intéresser à un problème de *Market Basket Analysis*. Les données que nous vous proposons regroupent l'ensemble des ventes (transactions) en ligne d'un site de vente durant une année. Nous souhaitons générer des règles d'associations par rapport à ces ventes. Vous pouvez consulter la source indiquée dans le fichier 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, comme :

- *InvoiceNo* : identifiant de la facture/ vente
- *StockCode* : identifiant du produit
- *Description* : produit acheté
- *Quantity* : quantité
- *InvoiceDate* : date de la commande/ paiement
- *UnitPrice* : prix du produit
- *CustomerID*: identifiant du client
- *Country* : pays de résidence du client

Avant de pouvoir appliquer les règles d'associations, la première tâche consistera à prétraiter les données. Dans un premier temps, nous voudrons regrouper tous les achats effectués par un même client à l'aide du bloc « Aggregate ». Veuillez choisir la *Description* comme attribut d'agrégation et *concatenation* comme fonction d'agrégation. La colonne *CustomerID (*l'attribut selon lequel le groupement s'effectuera) doit ensuite prendre le rôle *id*.

Dans Rapidminer, 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. Veuillez noter que conformément au prétraitement des données appliqué, « *item list in a column »* doit être choisit comme *input format* du bloc « FP-Growth ».

Questions sur les règles d'association

1. Constatez-vous des changements pour des différentes paramètres des blocs « FP-Growth » et « Create Association Rules » ? Veuillez commenter le paramétrage choisit et les résultats obtenus.

2. Est-il possible d'utiliser une/des autre/s colonne/s à partir des données initiales pour produire des règles intéressantes ?

5. Clustering

Dans cette dernière partie nous allons effectuer un regroupement (clustering) des applications sur *Google Play Store*. Vous pouvez consulter la source indiquée dans le fichier README, fourni avec les données, pour plus d'informations.

Parmi les attributs figurant dans le fichier « googleplaystore.xlsx » mis en disposition, nous allons utiliser :

- *Rating* : l'évaluation globale de l'application par les utilisateurs
- *Reviews* : le nombre d'avis d'utilisateurs
- *Size* : la taille de l'application
- *Installs* : nombre d'utilisateurs qui ont installé l'application
- *Price* : prix de l'application

Pour faciliter l'interprétation des résultats obtenus nous allons aussi utiliser les attributs : *App* (nom de l'application) avec le rôle de *id* et *Category* (le genre) comme *label*. Veuillez noter que ces deux attributs ne vont pas être prises en compte pour effectuer le clustering.

Avant réaliser le clustering, les attributs numériques doivent être normalisés à l'aide de l'opérateur « Normalize ». Dans un premier temps vous allez utiliser l'algorithme k-means en testant différents paramètres. Les résultats fournis par l'algorithme vont également être évalués par l'opérateur « Cluster Distance Performance ». Veuillez-vous assurer que les trois sorties de ce bloc sont connectées.

Questions sur le clustering

1. Constatez-vous des changements pour des différentes nombre d'itérations (*max runs*) de k-means ?

2. Comment l'ensemble des données est partitionné en imposant seulement 2 clusters ? Comment les résultats changent-ils en augmentant le nombre des clusters ? Veuillez commenter les résultats obtenus en termes des *centroïdes* (centre de masse de chaque cluster), la moyenne des distances du chaque centroïde et, si pertinent, le genre des applications regroupées dans les clusters.

3. Comment les résultats changent-ils en utilisant des autres algorithmes de clustering ?

Rendu/Evaluation

Vous remettrez sur *Moodle* un zip contenant :

- Vos processus exportés au format XML
- Un rapport dans lequel vous discuterez de vos différentes manipulations et des résultats obtenus, de vos choix et répondrez aux questions posées.

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

Adresses E-Mail des assistants : antoine.rochat@heig-vd.ch et elena.najdenovska@heig-vd.ch