

Web Mining

Laboratoire 2

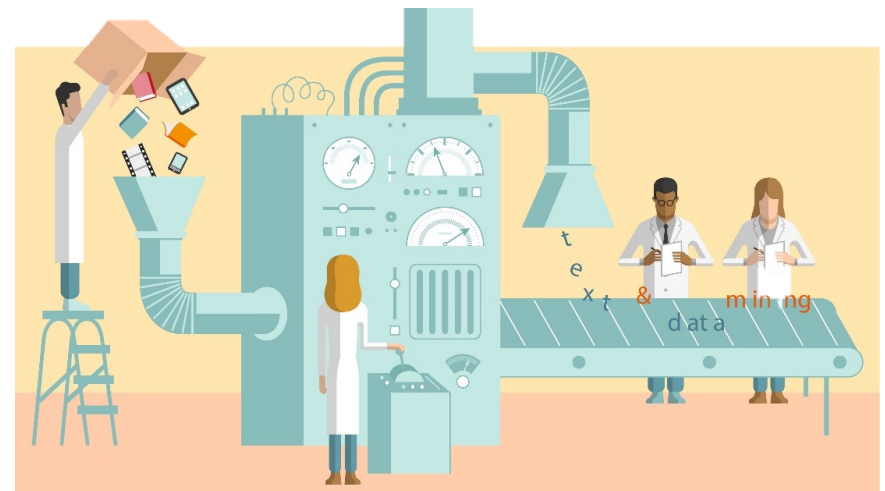
Antoine Rochat, Elena Najdenovska

Institut des Technologies de l'Information et de la Communication (IICT)

Application de techniques de Data Mining en utilisant le logiciel *RapidMiner*

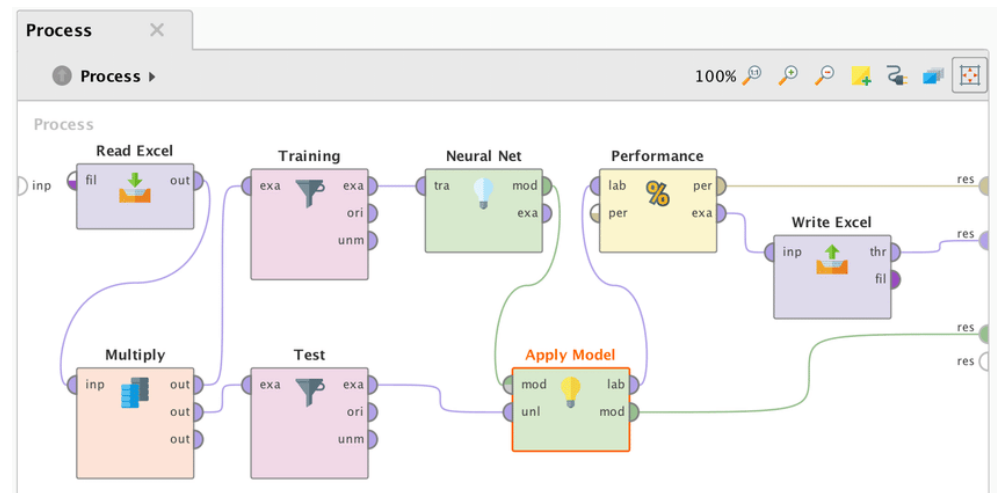
Les points étudiés:

- Prise en main de *RapidMiner*
 - Modélisation d'un filtre des « pièges-à-clics »
 - Analyse des sentiment sur des commentaires à l'aide de *WordNet*
 - Recommandation de films
 - Règles d'association sur des achats d'un site de vente en ligne
 - Clustering d'applications Google
-
- A rendre sur *Moodle* avant le **vendredi 07.04.2022 à 23h59**
 - Groupes de 3 personnes



RapidMiner

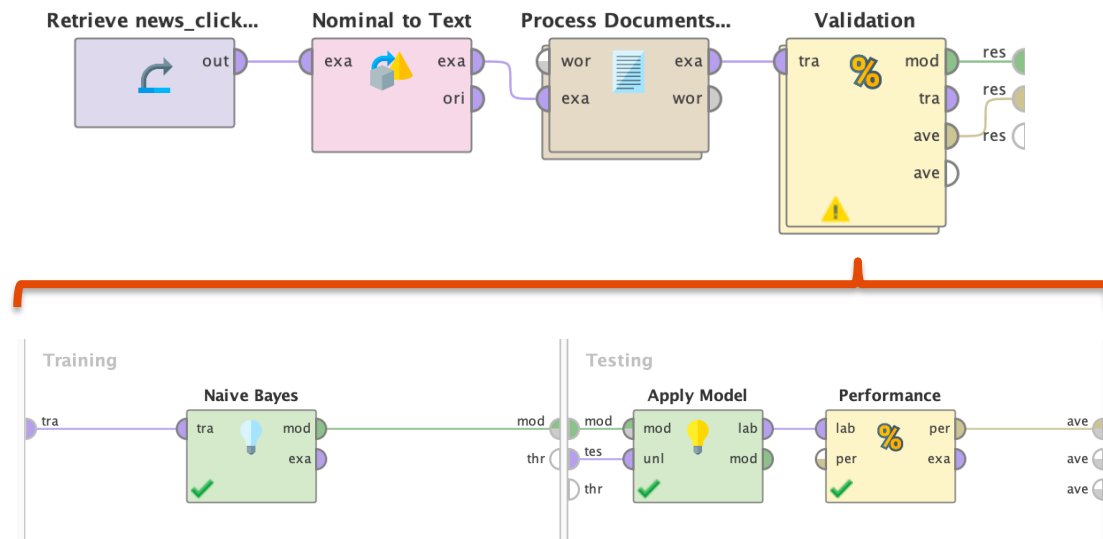
- Plateforme logicielle de traitement, de modélisation et d'analyse de données
 - Prétraitement
 - Application d'algorithmes et techniques de data mining
 - Evaluation
- Version communautaire gratuite mais limitée
- Version professionnelle payante
- ✓ Version académique gratuite et non limitée pendant 1 an



Partie 1

Filtre des « pièges-à-clics »

- Prise en main guidée de *RapidMiner*
- Filtrage des 10'000 titres des médias d'information en ligne
- Séparation du jeu de données: *ensemble d'apprentissage* et *ensemble de test*
- Construction d'un modèle par apprentissage supervisé, puis évaluation de celui-ci
- Quelques questions



Partie 2

Analyse des sentiments

- *WordNet* <https://wordnet.princeton.edu/>
- Base de données lexicale composée de noms, de verbes, d'adjectifs et d'adverbes en anglais groupés dans des ensembles de synonymes (synsets)
- L'extension de *RapidMiner* permet de d'associer une valeur numérique correspondant au sentiment par rapport à des données textuelles:
<0 sentiment négatif et >0 positif
- On va utiliser un ensemble d'environ 3'000 commentaires d'internautes issus de *Twitter liés au sujet COVID-19*. Ceux-ci sont déjà étiquetés
- Vous devrez faire une petite analyse comparant les sentiments issus du module *Wordnet* et ceux des étiquettes

WordNet Search - 3.1

Word to search for:

Noun

- *S: (n)* [stopping point](#), [finale](#), [finis](#), [finish](#), [last](#), [conclusion](#), [close](#) (the temporal end; the concluding time)
- *S: (n)* [conclusion](#), [end](#), [close](#), [closing](#), [ending](#) (the last section of a communication)
- *S: (n)* [finale](#), [close](#), [closing curtain](#), [finis](#) (the concluding part of any performance)

Verb

- *S: (v)* [close](#), [shut](#) (move so that an opening or passage is obstructed; make shut)
- *S: (v)* [close](#), [shut](#) (become closed)
- *S: (v)* [close up](#), [close](#), [fold](#), [shut down](#), [close down](#) (cease to operate or cause to cease operating)
- *S: (v)* [close](#) (finish or terminate (meetings, speeches, etc.))
- *S: (v)* [conclude](#), [close](#) (come to a close)
- *S: (v)* [close](#) (complete a business deal, negotiation, or an agreement)
- *S: (v)* [close](#) (be priced or listed when trading stops)
- *S: (v)* [close](#) (engage at close quarters)
- *S: (v)* [close](#) (cause a window or an application to disappear on a computer desktop)
- *S: (v)* [close](#) (change one's body stance so that the forward shoulder and foot are closer to the intended point of impact)
- *S: (v)* [close](#), [come together](#) (come together, as if in an embrace)
- *S: (v)* [close](#) (draw near)
- *S: (v)* [close](#) (bring together all the elements or parts of)
- *S: (v)* [close](#) (bar access to)
- *S: (v)* [close](#), [fill up](#) (fill or stop up)
- *S: (v)* [close up](#), [close](#) (unite or bring into contact or bring together the edges of)
- *S: (v)* [close](#) (finish a game in baseball by protecting a lead)

Adjective

- *S: (adj)* [close](#) (at or within a short distance in space or time or having elements near each other)
- *S: (adj)* [close](#) (close in relevance or relationship)
- *S: (adj)* [near](#), [close](#), [nigh](#) (not far distant in time or space or degree or circumstances)
- *S: (adj)* [close](#) (rigorously attentive; strict and thorough)
- *S: (adj)* [close](#), [faithful](#) (marked by fidelity to an original)
- *S: (adj)* [close](#), [tight](#) ((of a contest or contestants) evenly matched)
- *S: (adj)* [close](#), [confining](#) (crowded)
- *S: (adj)* [airless](#), [close](#), [stuffy](#), [unaired](#) (lacking fresh air)
- *S: (adj)* [close](#), [tight](#) (of textiles)
- *S: (adj)* [close](#) (strictly confined or guarded)
- *S: (adj)* [close](#) (confined to specific persons)
- *S: (adj)* [close](#), [snug](#), [close-fitting](#) (fitting closely but comfortably)
- *S: (adj)* [close](#) (used of hair or haircuts)
- *S: (adj)* [cheeseparing](#), [close](#), [near](#), [penny-pinching](#), [skinny](#) (giving or spending with reluctance)
- *S: (adj)* [close](#), [closelipped](#), [closemouthed](#), [secretive](#), [tightlipped](#) (inclined to secrecy or reticence about divulging information)

Adverb

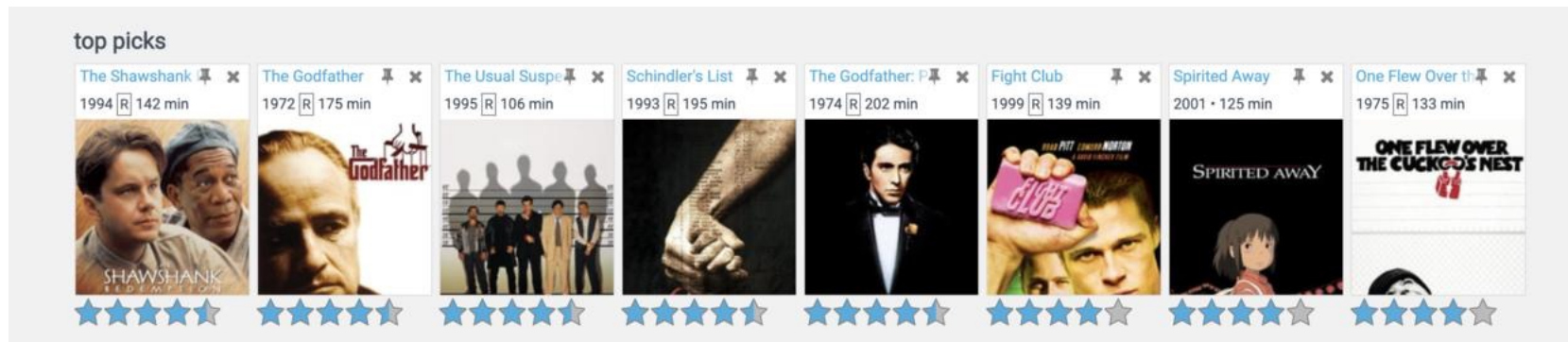
- *S: (adv)* [near](#), [nigh](#), [close](#) (near in time or place or relationship)
- *S: (adv)* [close](#), [closely](#), [tight](#) (in an attentive manner)

Partie 3

Système de recommandation

3 types:

- Systèmes basés sur un **filtrage collaboratif**
 - informations sociales, des préférences et des expériences d'autres utilisateurs
 - hypothèse : les utilisateurs ayant des goûts similaires pourraient apprécier et consommer des articles similaires
- Systèmes basés sur un **filtrage de contenu**
 - des informations sur les activités passées de l'utilisateur
 - hypothèse : l'utilisateur pourrait être intéressé par des éléments similaires à ceux pour lesquels un intérêt a déjà été montré
- Systèmes **hybrides**
 - combinent à la fois plusieurs approches afin de renforcer leurs avantages



Partie 3

Système de recommandation

- Approche collaborative pour la recommandation et la prédiction des notes
- Données : environ 9800 films notés par 610 utilisateurs de MovieLens
- 2 ensembles des données:
 - *movies.csv*
 - *ratings.csv*
- Prétraitement des données pour joindre les films avec les notes des utilisateurs (attribut en commun: *movieID*)
- Quelques questions



				
John 	5	1	3	5
Tom 	?	?	?	2
Alice 	4	?	3	?

Partie 4

Market Basket Analysis

- Une technique de data mining qui permet de trouver les groupes d'articles qui ont la tendance à apparaître ensemble lors d'une transaction
- Construction du modèle basé sur des règles conditionnelles:
 - Si **condition** alors **résultat**
 - « Si un client achète du beurre, alors il achète aussi du lait »
 - « Si un client achète du lait et du sel, alors il achète aussi du fromage »



Partie 4

Market Basket Analysis

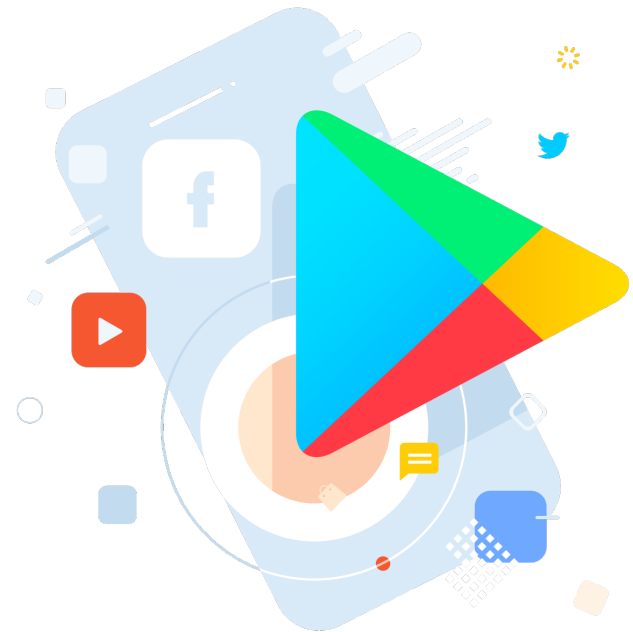
- Articles achetés sur un site web pendant 1 année
- Forme des données fournies:
 - *InvoiceNo* : identifiant de la facture/ vente
 - *StockCode* : identifiant du produit
 - *Description* : produit acheté
 - *Quantity* : quantité
 - *InvoiceDate* : date de la commande/ paiement
 - *UnitPrice* : prix du produit
 - *CustomerID*: identifiant du client
 - *Country* : pays de résidence du client
- Prétraitement des données pour organiser les achats par client
- Quelques questions



Partie 5

Clustering

- Clustering d'applications sur Google Play Store
- Différents attributs pour chaque applications
 - *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
- Et pour interpréter les résultats (ne pas prendre en compte pour le clustering)
 - *App*: nom de l'application
 - *Category*: le genre
- Quelques questions



Rendu

- Sur la page *Moodle* du cours
- Une archive *zip* contenant
 - Votre rapport
 - Vos *process RapidMiner*
- Dans votre rapport, en plus des réponses aux questions posées vous discuterez du fonctionnement de votre programme et de vos choix d'implémentation

