CRM 205 : Analyse Textuelle (Text Mining)

Cours du 4 Juin 2019 – Lefebvre Stéphanie

Définition

- Data mining : fouille de données est un ensemble de méthodes qui consistent à extraire un savoir ou une connaissance, à partir d'une base de donnée.
  Méthode la mieux adaptée à des données à grand volume et les informations récoltées doivent aider à prendre des décisions.

Exemples de modèle d'analyse de données

Association – chercher des patterns au sein desquelles un événement est lié à un autre événement.

Analyse de séquence – chercher des patterns au sein desquelles un événement mène à un autre événement plus tardif.

Prédiction – découvrir des patterns de données pouvant mener à des prédictions raisonnables sur le futur. Ce type de data mining est aussi connu sous le nom d'analyse prédictive.

Définition

- Text mining : sous catégorie du data mining. Fouille de données appliquées à des éléments textuels. Il s'agit de l'analyse de contenu (Mot, Syntaxe, etc.) pour en extraire des connaissances et tenter d'en comprendre le sens.

Etapes d'une étude de Data Mining

1. Définition des objectifs
2. Inventaires des données existantes
3. Collecte des données : établir des fichiers
4. Exploitation et préparation des données
5. Segmentation de la population
6. Élaboration et validation des modèles prédictifs
7. Déploiement des modèles
8. Formation des utilisateurs des modèles
9. Suivi des modèles
10. Enrichissement des modèles

Types de fichiers de données

- Fichiers « officiels » créés par un organisme, généralement structuré, répondant à des normes.
- Ex : https://www.data.gouv.fr/fr/datasets/prenoms-des-nouveaux-nes-1/

- mais des erreurs peuvent exister
- => des outils de vérifications sont développés. Ex : https://validata.fr/doku.php

Types de fichiers de données

- Fichiers créés « manuellement »
  - Faire des recherches pour trouver les informations;
  - Relever ce qui nous intéresse;
  - Créer un fichier .CSV contenant les données que nous réutiliserons.

Types de fichiers de données

- Fichiers créés de manière « automatisée »
  - Déterminer les données qui nous intéresse;
  - Trouver le ou les sites où elles sont accessibles;
  - Ecrire un script sous R qui va les collecter automatiquement = CRAWLING et SCRAPING

Définitions

- Crawling : Collecter automatiquement le contenu d'une page pour la traiter, la classer et fournir des informations ;
- Scraping : Extraction du contenu d'un site web dans le but de la transformer pour permettre son utilisation dans un autre contexte ;

A-t-on le droit d'utiliser ces pratiques ?

- Pénal : Vol de données https://www.haas-avocats.com/ecommerce/rgpd-aspiration-de-donnees-sur-un-site-web-quelles-sanctions/
- Civil : Concurrence déloyale (parasitisme) – Code de la propriété intellectuelle

- https://business.lesechos.fr/directions-numeriques/digital/transformation-digitale/030113062124 3-pagesjaunes-s-attaque-aux-robots-aspirateurs-de-donnees-318010.php

Double exception : consécration du droit au text et data mining par l'introduction d'une exception aux droits d'auteur et producteur de bases de données

- Le Code de la propriété intellectuelle est ainsi modifié :
- 1° Après le second alinéa du 9° de l'article L. 122-5, il est inséré un 10° ainsi rédigé :

« 10° Les copies ou reproductions numériques réalisées à partir d'une source licite, en vue de l'exploration de textes et de données incluses ou associées aux écrits scientifiques pour les besoins de la recherche publique, à l'exclusion de toute finalité commerciale. Un décret fixe les conditions dans lesquelles l'exploration des textes et des données est mise en œuvre, ainsi que les modalités de conservation et de communication des fichiers produits au terme des activités de recherche pour lesquelles elles ont été produites ; ces fichiers constituent des données de la recherche ; »

- 2° Après le 4° de l'article L. 342-3, il est inséré un 5° ainsi rédigé :
- « 5° Les copies ou reproductions numériques de la base réalisées par une personne qui y a licitement accès, en vue de fouilles de textes et de données incluses ou associées aux écrits scientifiques dans un cadre de recherche, à l'exclusion de toute finalité commerciale. La conservation et la communication des copies techniques issues des traitements, au terme des activités de recherche pour lesquelles elles ont été produites, sont assurées par des organismes désignés par décret. Les autres copies ou reproductions sont détruites. »

Scraping HTML et CSS

- Les sites internet sont codés en langages HTML, CSS et PHP.
- Ils contiennent des contenus visibles et du contenu invisible que l'on appelle Métadonnées.
- Le SCRAPING prend en compte la structure HTML pour collecter le contenu visible ainsi que les métadonnées.

Exemple de SCRAPING :

- Article du site l'internaute , mis à jour le 10 septembre 2018, sur les attentats de Paris https://www.linternaute.com/actualite/societe/1281824-attentat-a-paris-une-longue-liste-d-attaques-dans-la-capitale/
- Clique droit, puis choisir « code source de la page »
- La mise en forme du contenu du site se fait à partir de « balise »

Récupération dates et lieux attentats à partir d'un article de journal

```
# Lien vers les notions theoriques : https://medium.freecodecamp.org/an-introduction-to-web-scraping-using-r-40284110c848

# Attentat de Paris

# 1/ Charger l'URL contenant les données à scraper
URL <- "https://www.linternaute.com/actualite/societe/1281824-attentat-a-paris-une-longue-liste-d-attaques-dans-la-capitale/"

# 2/ lecture de la page à scraper
SiteWeb <- read_html(URL)
head(SiteWeb)

## $node
## <pointer: 0x00000000185190e0>
##
```

```
DateAttentat <- SiteWeb %>%
  html_nodes("strong") %>%
  html_text()
DateAttentat

##  [1] "Dimanche 9 septembre 2018"
##  [2] "Samedi 12 mai 2018 "
##  [3] "Jeudi 20 avril 2017 -"
##  [4] "Vendredi 3 février 2017 - "
##  [5] "Vendredi 13 novembre 2015 -"
##  [6] "Mercredi 7 janvier 2015 - "
##  [7] "25 juillet 1995 "
##  [8] "24 décembre 1994 - "
##  [9] "17 septembre 1986 -"
## [10] "9 août 1982 -"
## [11] "3 octobre 1980 - "
## [12] "15 septembre 1974 -"
## [13] "Attentat à Paris : une longue liste d'attaques dans la capitale"
## [14] "Elections européennes 2019"
```

```
> NomAttentat <- SiteWeb %>%
+   html_nodes("h2") %>%
+   html_text()
> NomAttentat
 [1] "Attaque au couteau quai de Loire "
 [2] "Attentat dans le quartier Opéra"
 [3] "Attentat des Champs-Elysées"
 [4] "Attentat du musée du Louvre"
 [5] "Attentats du 13 novembre 2015"
 [6] "Attentats de Charlie Hebdo et de l'Hyper Cacher"
 [7] "Attentat du RER Saint-Michel (1995)"
 [8] "Attentat dans un avion Air France (1994)"
 [9] "Attentat de la rue de Rennes (1986)"
[10] "Attentat de la rue des Rosiers (1982)"
[11] "Attentat de la synagogue de la rue Copernic (1980)"
[12] "Attentat du Drugstore Publicis de Saint-Germain (1974)"
> NomAttentat <- data.frame(NomAttentat)
> NomAttentat
                                                NomAttentat
1                          Attaque au couteau quai de Loire
2                            Attentat dans le quartier Opéra
3                                Attentat des Champs-Elysées
4                                Attentat du musée du Louvre
5                              Attentats du 13 novembre 2015
6           Attentats de Charlie Hebdo et de l'Hyper Cacher
7                        Attentat du RER Saint-Michel (1995)
8                  Attentat dans un avion Air France (1994)
9                        Attentat de la rue de Rennes (1986)
10                     Attentat de la rue des Rosiers (1982)
11       Attentat de la synagogue de la rue Copernic (1980)
12 Attentat du Drugstore Publicis de Saint-Germain (1974)
>
```

Autre exemple complexe : Récupération de Tweets

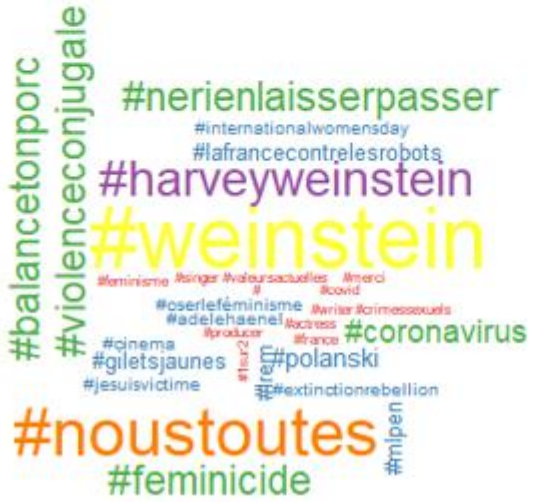
- Créer un compte tweeter utilisateur;
- Créer un compte tweeter développeur;
- Spécifier dans le script de R les identifiants et mots de passe Tweeter;
- Récupérer et analyser les tweets (voir fichier pdf “RécupérationTweets”)

Récupération et analyse de 4000 tweets à partir du #MeToo

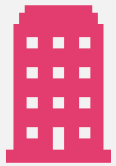
```
# Affichage des 10 Hashtags les plus populaires
print(tritheme[1:10])
```

```
## theme
##                     #metoo                  #weinstein
##                        194                          19
##                #noustoutes #violencesfaitesauxfemmes
##                         15                          12
##           #harveyweinstein           #balancetonporc
##                         11                           9
##                #feminicide      #nerienlaisserpasser
##                          9                           9
##         #violenceconjugale              #coronavirus
##                          9                           7
```

```
# 6 - Affichage des donnÃ©Â©es
library(wordcloud)

## Loading required package: RColorBrewer

opar <- par(mar = c(0.1,0.1,2.1,0.1))

wordcloud(names(tritheme)[-1],tritheme[-1],
          scale = c(3,0),
          colors = brewer.pal(6,"Set1"))

## Warning in wordcloud(names(tritheme)[-1], tritheme[-1], scale = c(3, 0), :
## #violencesfaitesauxfemmes could not be fit on page. It will not be
plotted.
```



Exemple d'applications d'analyse de textes



http://www.lefigaro.fr/secteur/high-tech/en-estonie-une-intelligence-artificielle-va-rendre-des-decisions-de-justice-20190401

https://entrepreneur-interet-general.etalab.gouv.fr/defis.html

Vidéos / Autres exemples d'applications

HTTPS://INHESJ.FR/EVENEMENTS/VIDEOTHEQUE

HTTPS://WWW.APPVIZER.FR/RECHERCHE?IDS=SC419,SC453&SELECTED=24826

Obligations légales

- Loi numérique impose « la transparence des algorithme »
- Art. L311-3-1 du Code des relations entre le public et l'administration :citoyen qu'il a le droit d'obtenir la communication des « règles » et « principales caractéristiques » de mise en œuvre du traitement algorithmique utilisé
- Ex : algorithme parcoursup 2018 discrimination ville /2019 discrimination lycée https://www.liberation.fr/checknews/2018/05/30/parcoursup-les-eleves-sont-ils-discrimines-en-fonction-de-leur-origine-geographique_1654734
- https://www.nextinpact.com/news/105098-transparence-algorithmes-loi-numerique-ignoree-nombreuses-administrations.htm

Dérives ?
Problèmes ?

- Les analyses se basent sur le passé => Uniformisation
- Fin de la jurisprudence qui jusqu'à présent fait évoluer le droit