Une recherche sur Text mining

Préparé par : AKRAM EL BASRI

Encadre par : Monsieur EL GUEZAZ

MARRAKECH
جامعة القاضي عياض
UNIVERSITÉ CADI AYYAD

المدرسة العليا للتكنولوجيا- الصويرة
L'ÉCOLE SUPÉRIEURE DE TECHNOLOGIE – ESSAOUIRA

TABLE DES MATIERES

Fin.

Introduction

Dans cette recherche, nous avons parlé sur text mining ou la fouille de textes, Précisément les points suivants :

- La définition de text mining.
- Les différentes techniques utilisées dans Fouille de textes.
- Le processus de text mining.
- Les applications de text mining.
- La partie pratique:
    - Application de fouille de texte sur une dataset (Natural Language Processing).
    - Application algorithme de Random Forest sur le dataset.

Fin.

La définition de Fouille de textes (Text Mining)

La **fouille de textes** ou « l'extraction de connaissances » dans les textes est une spécialisation de la fouille de données et fait partie du domaine de l'intelligence artificielle. Cette technique est souvent désignée sous l'anglicisme *text mining*.

Elle désigne un ensemble de traitements informatiques consistant à extraire des connaissances selon un critère de nouveauté ou de similarité dans des textes produits par des humains pour des humains. Dans la pratique, cela revient à mettre en algorithme un modèle simplifié des théories linguistiques dans des systèmes informatiques d'apprentissage et de statistiques, et des technologies de compréhension du langage naturel.

Fin.

Il existe une large variété de techniques et méthodes de Text Mining. Voici les plus couramment utilisées.

1) Les techniques d'analyse :

La technique de la ” fréquence de mots “ consiste à identifier les termes ou concepts les plus récurrents dans un ensemble de données. Ceci peut s’avérer très utile, notamment pour analyser les avis de clients ou les conversations sur les réseaux sociaux.

La méthode de la concordance, quant à elle, est utilisée pour reconnaitre le contexte dans lequel un ensemble de mots apparaît dans un texte. Cette technique permet d’éviter l’ambiguïté et de comprendre le sens d’un terme dans le contexte spécifique.

La méthode de la collocation, quant à elle, consiste à repérer les séquences de mots apparaissant fréquemment à proximité l’une de l’autre. Certains mots apparaissent très souvent ensemble. Il peut s’agir de bigrammes ou de trigrammes, des combinaisons de deux à trois mots. En identifiant ces colocations, il est possible de mieux comprendre la structure sémantique d’un texte et d’obtenir des résultats de Text Mining plus fiables.

2) La récupération d'informations :

La récupération d’informations consiste à trouver des informations pertinentes à partir d’un ensemble préd-défini de requêtes ou de phrases. On utilise souvent cette approche dans les systèmes de catalogues de bibliothèques ou les moteurs de recherche web.
Les systèmes ” IR ” (information retrieval) utilisent différents algorithmes pour suivre les comportements des utilisateurs et identifier les données pertinentes.
La ” tokenization ” consiste à décomposer un long texte en phrases ou en mots appelés ” tokens ” (jetons). Ces jetons sont ensuite utilisés dans les modèles pour le clustering de texte ou les tâches visant à associer des documents.
Le ” stemming “, quant à lui, consiste à séparer les préfixes et les suffixes des mots pour en dériver le mot racine et sa signification. Cette technique permet de réduire la taille des fichiers d’index.

Fin.

3) La classification de texte :

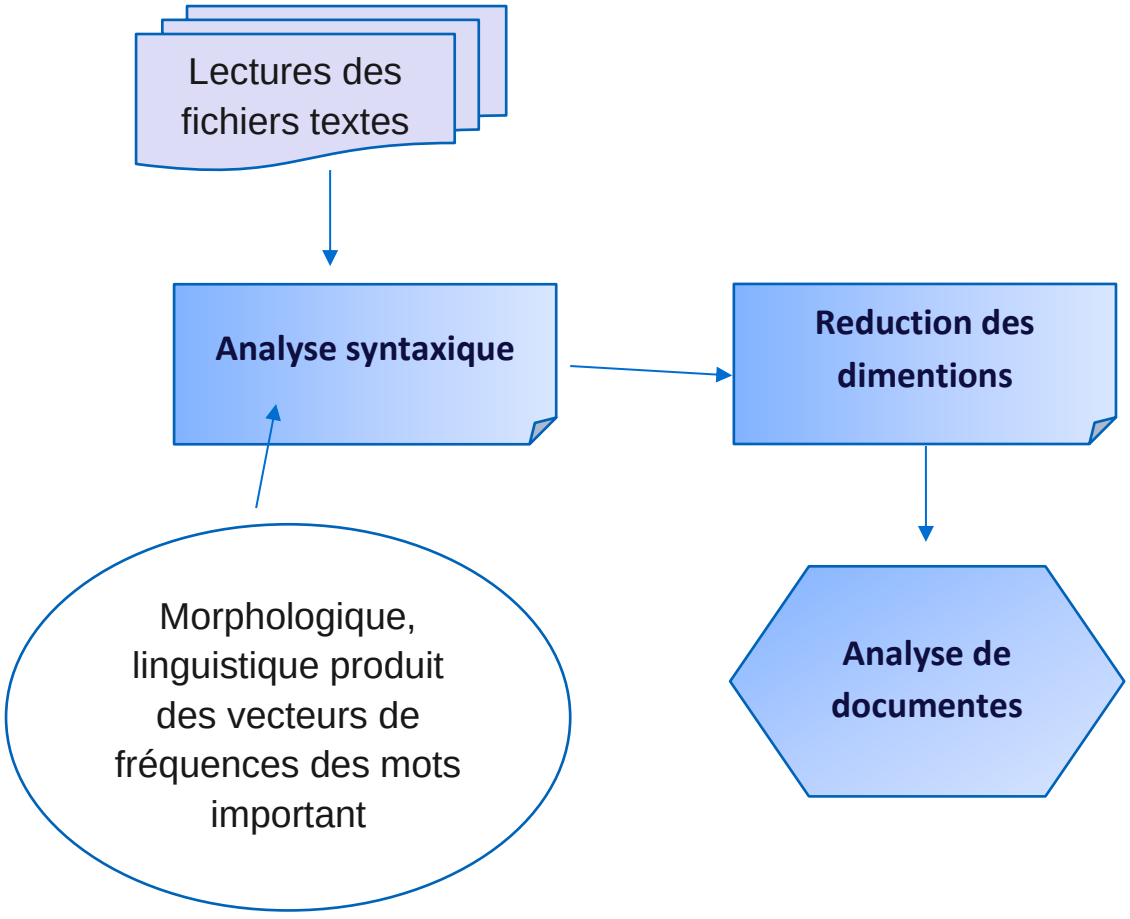
Il existe aussi des méthodes plus avancées de Text Mining. La classification de texte consiste à assigner des étiquettes aux données de texte non structurées. C'est une étape essentielle et indispensable pour le traitement naturel du langage (Natural Language Processing).
Elle permet en effet d'organiser et de structurer un texte complexe afin d'en dégager des données pertinentes. C'est grâce à cette technique, que les entreprises sont en mesure d'analyser toutes sortes d'informations textuelles afin d'en tirer de précieuses indications.
Il existe différentes formes de classification de texte. L'analyse de sujet (Topic Analysis) permet de comprendre les principaux thèmes ou sujets d'un texte. C'est l'une des principales façons d'organiser les données de texte.

Fin.

Le processus de text mining

Le processus de *Text Mining* consiste à analyser des ensembles de documents textuels afin de capturer les concepts et thèmes-clés, et de découvrir les relations et les tendances cachées. Il ne nécessite pas que vous connaissiez les mots ou les termes précis utilisés par les auteurs pour exprimer ces concepts. Bien qu'il s'agisse de processus très différents, l'exploration de texte est parfois confondue avec la récupération d'informations. Si l'extraction et le stockage précis des informations représentent un défi considérable, l'extraction et la gestion efficaces du contenu, de la terminologie et des relations compris dans ces informations jouent un rôle vital.

Processus de fouille de textes : vue simplifiée



Fin.

Les applicarions de text mining

Traitement automatique des messages, emails. Il est possible de « filtrer » automatiquement les « courriers indésirables » les plus indésirables. Cela est basé sur certains termes ou mots qui ne sont pas susceptibles d'apparaître dans des messages légitimes.

Analyse des réclamations de garantie ou d'assurance, entretiens de diagnostic :
Par exemple : Les demandes de garantie ou les entretiens médicaux initiaux peuvent se résumer en brefs récits. De plus en plus, ces notes sont collectées par voie électronique. Donc, ces types de récits sont facilement disponibles pour l'entrée.

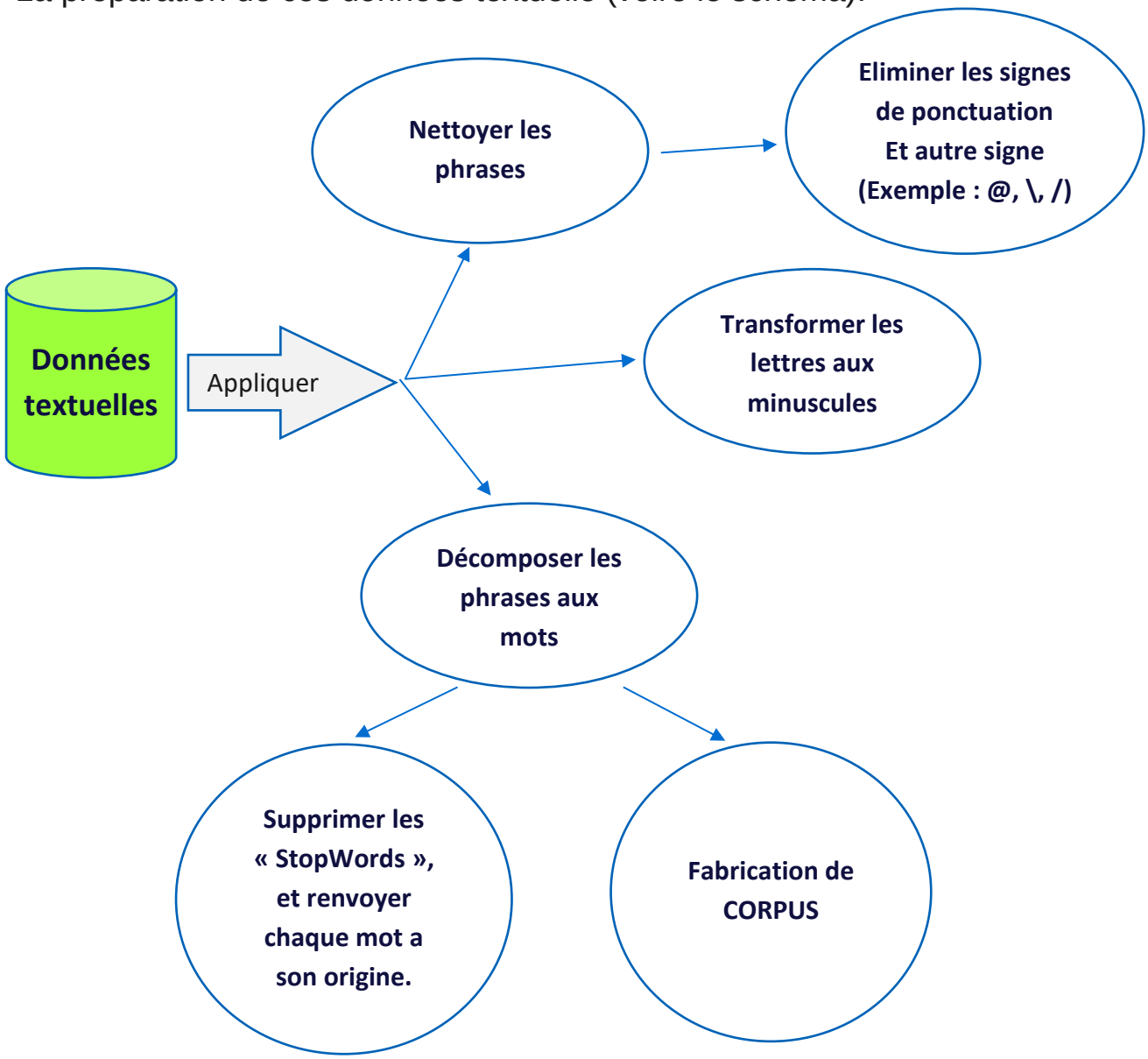
Enquêter sur les concurrents en explorant leurs sites Web :
Exemple de pratique : Vous pouvez accéder à une page Web et commencer à "explorer" les liens que vous y trouvez pour traiter toutes les pages Web qui y sont référencées. De cette manière, vous pourriez dériver une liste de termes et de documents disponibles sur ce site. Par conséquent, déterminez les termes et caractéristiques les plus importants qui sont décrits.

La publicité numérique : est un domaine d'application modérément nouveau et en pleine croissance pour l'analyse de texte. Par rapport à l'approche traditionnelle basée sur les cookies, la publicité contextuelle offre une meilleure précision, préserve complètement la vie privée de l'utilisateur.

Fin.

La partie pratique

- Cette partie a réservé pour appliquer **un modèle supervisé** sur un dataset nommée ‘Restaurant_Reviews.ts’, pour le traitement de la langage naturel (Natural Language Processing).
- L’idée c’est que : On va appliquer **un apprentissage supervisé** sur le fichier précèdent qui contenants les commentaires des clients **« Feedback clients »** d’un restaurant américain.
- Les étapes à suivre pour implémenter ce model sont :
  La préparation de ces données textuelle (voire le schéma).



Fin.

1. Description de la base de données :

Notre datset es contient deux colonne (Review, et Liked), et 1000 lignes, la première colonne représente un point de vue ce format de texte sur la qualité de service d'un restaurant, et la deuxième c'est champe binaire possède deux modalités : '1' : si le feedback du client est positif, et '0' si non,(Figure 1)

	Review	Liked
0	Wow... Loved this place.	1
1	Crust is not good.	0
2	Not tasty and the texture was just nasty.	0

Figure 1 : description de dataset

2. L'implémentation de modèle :

A) La phase 01 : La préparation de données :

- Etape 01 :
  - ✓ Utiliser les libriries ' re ' (Regular Expression) et 'nltk' (Natural Language processing) de python, en suite charger les « stopwords » (Figure 2).

```
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'oursel
ves', 'you', "you're", "you've", "you'll", "you'd", 'you
r', 'yours', 'yourself', 'yourselves', 'he', 'him', 'hi
s', 'himself', 'she', "she's", 'her', 'hers', 'herself',
'it', "it's", 'its', 'itself', 'they', 'them', 'their',
'theirs', 'themselves', 'what', 'which', 'who', 'whom',
'this', 'that', "that'll", 'these', 'those', 'am', 'is',
```

Figure 2 : partie de : stopwords

  - ✓ Etape 02 : Supprimer les caractères inutiles des phrases dans notre dataset : (Figure 03), en suite permute les phrases en minuscule.

Fin.

```
- La phrase initial: wow... loved this place.
- Phrase regulariser :wow    loved this place
```

Figure 3: Le processus de la régularisation des phrases

- Etape 03 :

✓ Décomposer les phrases en entités (mots, lettre...) : (Figure 04).
✓ En supprime les « Stopwords » (Figure 04).
✓ Fabrication de CORPUS (construction de corpus) (Figure 04).

```
Les entités : ['Wow', 'Loved', 'this', 'place']
Suppression des stopwords: ['wow', 'love', 'place']
Origine des mots : ['wow', 'love', 'place']
Corpus : wow love place
```

Figure 4: L'étapes d'obtenir le corpus d'une phrase.

- Application de cette procédure sur le dataset, et affichage des résultats (Figure 06) :

Review	Liked
Wow... Loved this place.	1
Crust is not good.	0
Not tasty and the texture was just nasty.	0
Stopped by during the late May bank holiday of...	1
The selection on the menu was great and so wer...	1

Figure 5 : Avant les procédures de nettoyage

Fin.

```
['wow love place', 'crust good', 'tasti textur nasti', 'stop late may bank
holiday rick steve recommend love', 'select menu great price', 'get angri
want damn pho', 'honeslti tast fresh', 'potato like rubber could tell made
ahead time kept warmer', 'fri great', 'great touch', 'servic prompt', 'wou
ld go back', 'cashier care ever say still end wayyy overpr', 'tri cape cod
ravoli chicken cranberri mmmm', 'disgust pretti sure human hair', 'shock s
ign indic cash', 'highli recommend', 'waitress littl slow servic', 'place
```

Figure 6 : Apres le nettoyage

B) La phase 02 : L'encodage de données :

L'objectif de cette phase est l'utilisation de la méthode CountVictorizer (utiliser pour encoder les formats textuelles) de la librairie sklearn de python, pour encoder la liste des phrases qu'on a préparé précédente.
Apres l'application de cette méthode en obtenir une data frame pris pour appliquer notre algorithme de machine Learning (Figure 07).

	0	1	2	3	4	5	6	7	8	9	...	990	991	992	993	994	995	996	997	998	999
absolut	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
absolutley	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
accid	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
accommod	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
accomod	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
yukon	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
yum	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
yummi	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
zero	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
Target	1	0	0	1	1	0	0	0	1	1	...	0	0	0	0	0	0	0	0	0	0

1566 rows × 1000 columns

Figure 7 : Data frame représente l'encodage de texte

- ⇨ **Chaque colonne représente les valeurs de codage d'une certaine phrase.**
- ⇨ **En charger 80% pour le training set, et 20% pour le test set.**

Fin.

C) La phase 03 : Application de l'algorithme random forest sur notre data frame :

- **Les résultats :**
  - **Le score de modèle :**

```
# score de modele
score_model = model.score(X_train , Y_train)
score_model
```

```
0.99625
```

Figure 8 : Le score de modèle

  - ***Le score de prédictions :***

```
score = model.score(X_test , Y_test)
```

```
0.755
```

Figure 9 : Le score de modèle

Fin.

➢ ***La matrice de confusion :***

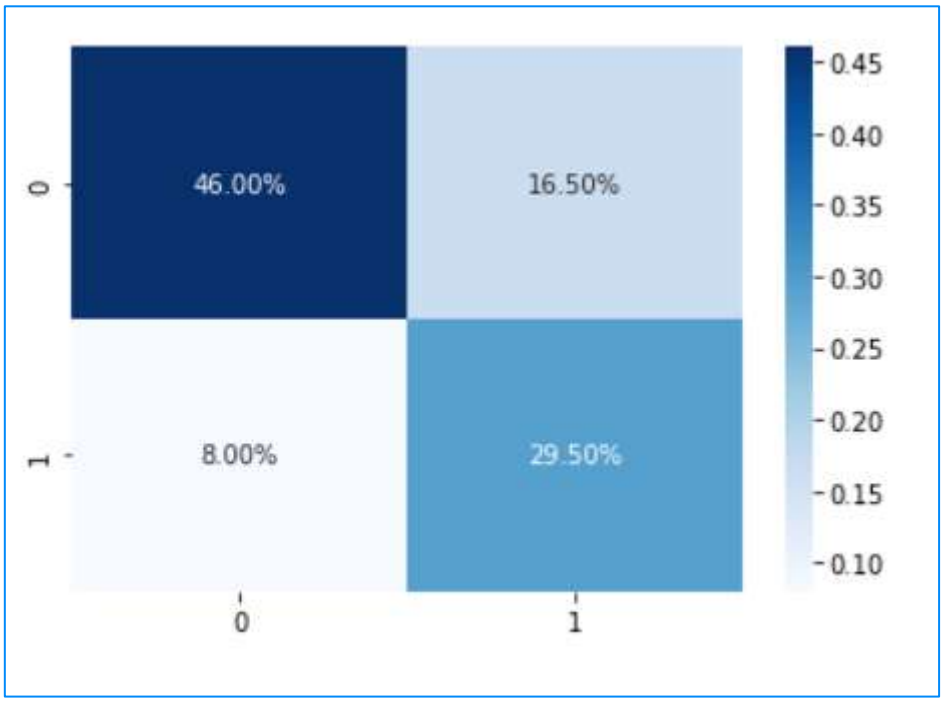


Figure 10 : heatmap de la matrice de confusion

Conclusion :

Ce rapport a réservé pour le devoir de recherche sur text mining, pour cette raison nous avons attaqué le sens de texte de fouille, et leurs methodes utiliser pour analyser les fichiers textuels, en suite discuter un peu sur le processus général de text mining, aussi les devers applications de text mining, et finalement travailler sur cas pratique.

Fin.

Liste de figure

Fin.