Traitement des questions ouvertes sur R : text mining

GADO Seman Giovanni Jocelyn
&
HABA Fromo Francis
&
SARR Abdou Karim

Elèves ingénieurs statisticiens économistes

Professeur : **M. HEMA Aboubacar** , *Analyste de recherches*

1er Juin 2024

1. Introduction
2. Généralités
3. Etapes du processus de text mining et packages utilisés sur R pour le text mining
4. Résultats de l'application sous R
5. Conclusion
6. Bibliographie et webographie

Sommaire

1 **Introduction**

2 Généralités

3 Etapes du processus de text mining et packages utilisés sur R pour le text mining

4 Résultats de l'application sous R

5 Conclusion

6 Bibliographie et webographie

Introduction

L'ère numérique a conduit à une prolifération de données textuelles disponibles : emails, articles de blogs, réseaux sociaux, documents d'entreprise, etc.
Les entreprises et les organisations génèrent et stockent des volumes massifs de texte non structurés, ne pouvant être utilisés directement et qui nécessitent des méthodes efficaces pour en extraire des informations pertinentes. Parmi tous les outils permettant de le faire, le **text mining** nous intéresse particulièrement.
Dans cette présentation, nous essayerons dans une première partie d'aborder les généralités du text mining, ensuite nous verrons les étapes de son déroulement ainsi que les packages de R dédiés au text mining et nous finirons avec les présentations des résultats d'un cas d'application.

Sommaire

1 Introduction

2 Généralités

- Qu'est ce que c'est que le text mining ?
- Quelques méthodes et techniques du text mining
- Text Mining vs Text Analytics : quelle est la différence ?
- Quelques applications du text mining que nous retrouvons dans la vie de tous les jours

3 Etapes du processus de text mining et packages utilisés sur R pour le text mining

4 Résultats de l'application sous R

5 Conclusion

6 Bibliographie et webographie

Qu'est ce que c'est que le text mining ?

Le **text mining** (traduit en français par **fouille de texte**) est un ensemble de techniques d'exploration de données qui repose sur le Machine Learning et qui permet de transfomer des données **textuelles** non structurées en données structurées pour ensuite procéder à leur analyse et en tirer des informations. C'est une branche du **data mining** qui regroupent les techniques d'exploitation des données de différents types.
Il a a émergé dans les années 1980 et 1990 avec le développement rapide des technologies.

Quelques méthodes et techniques du text mining

- **La technique de la fréquence de mots**
  - Identification des termes ou concepts les plus récurrents dans un ensemble de données.
  - Utile, notamment pour analyser les avis de clients ou les conversations sur les réseaux sociaux.
- **La méthode de la collocation**
  Répérage des séquences de mots apparaissant fréquemment à proximité l'une de l'autre. Certains mots apparaissent très souvent ensemble (bigrammes ou de trigrammes, des combinaisons de deux à trois mots)
- **La méthode de la concordance**
  Reconnaissance du 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.

- **Les systèmes « IR » (information retrieval)**
  Utilisation des différents algorithmes pour suivre les comportements des utilisateurs et identifier les données pertinentes (La tokenization)
- **L'analyse de sentiment**
  Analyser les émotions contenues dans un texte

Text Mining vs Text Analytics : quelle est la différence ?

Le **Text Mining** est souvent confondu avec le **Text Analytics**. En réalité, il s'agit de deux concepts légèrement différents permettant d'analyser automatiquement des textes.
Le Text Mining identifie les informations pertinentes dans un texte, tandis que le Text Analytics vise à découvrir des tendances à travers de larges ensembles de données.
En général, le Text Analytics est utilisé pour créer des tableaux, des diagrammes et des graphiques ou autres rapports visuels(analyses quantitatives). Le Text Mining quant à lui, combine les statistiques, la linguistique et le Machine Learning pour prédire automatiquement des résultats à partir d'expériences passées(analyses qualitatives). Ces deux notions se confondent assez souvent.

Quelques applications du text mining que nous retrouvons dans la vie de tous les jours

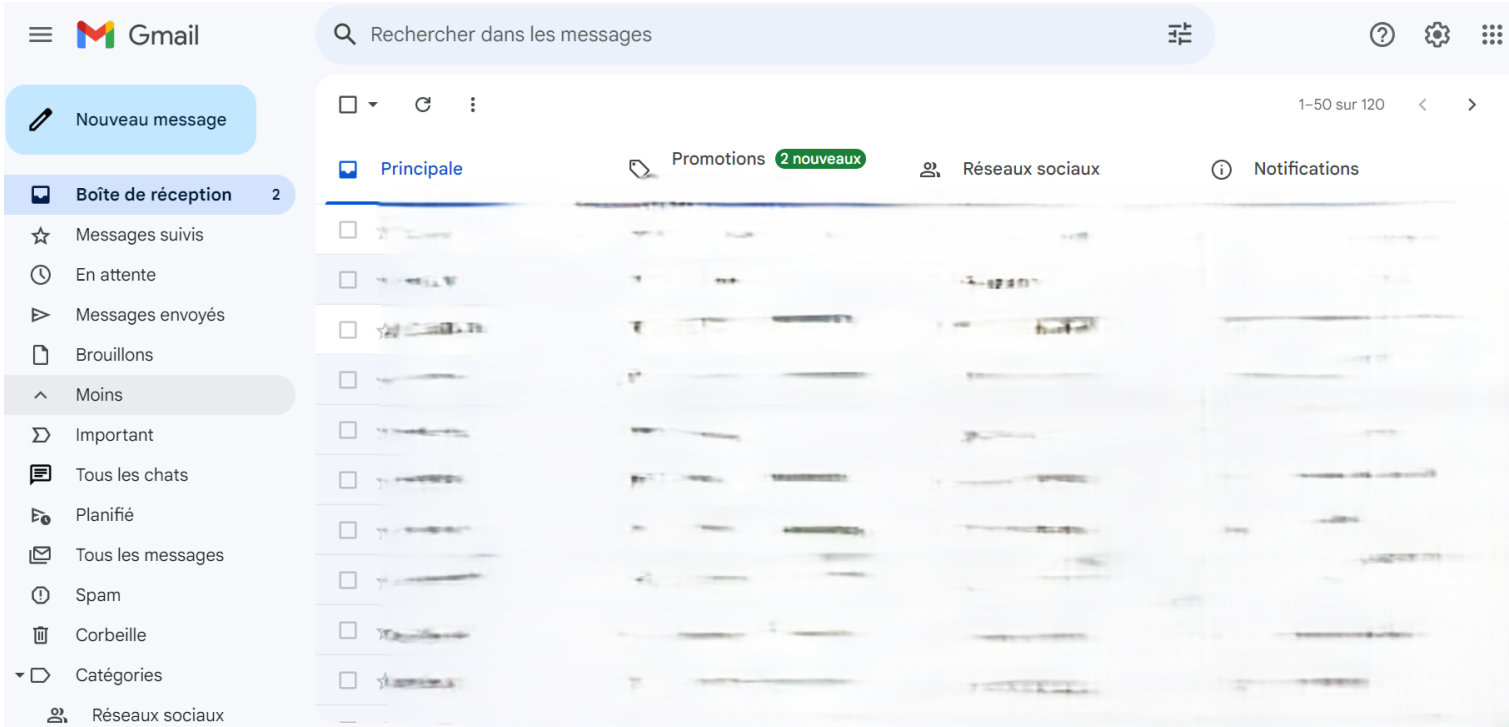


Figure 2: Gestion des spams

Sommaire

1 Introduction

2 Généralités

3 **Etapes du processus de text mining et packages utilisés sur R pour le text mining**

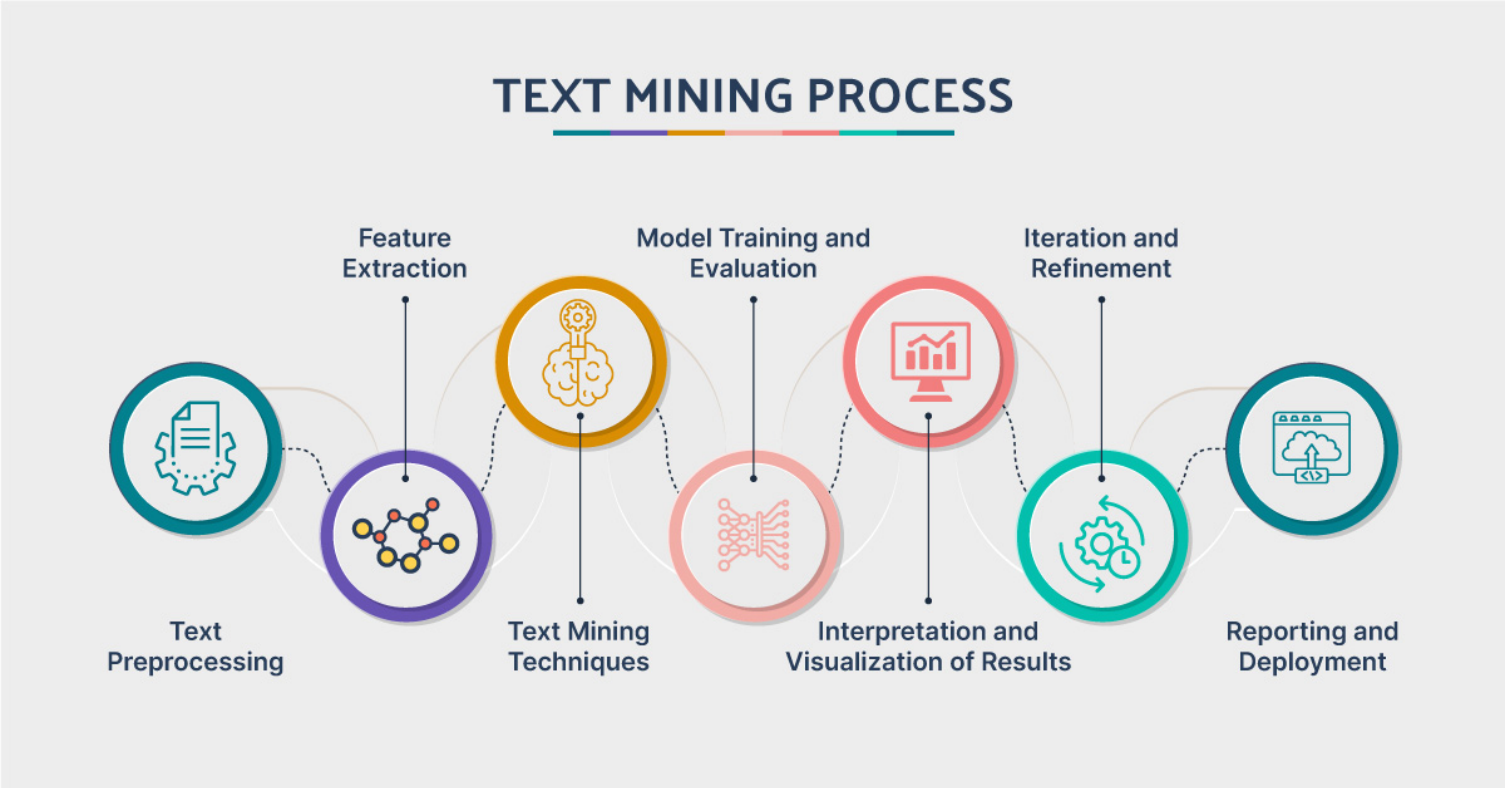
- **Etapes du processus de text mining**
- **Packages utilisés sur R pour le text mining**

4 Résultats de l'application sous R

5 Conclusion

6 Bibliographie et webographie

Etapes du processus de text mining



Les différentes étapes du processus peuvent être résumées en **4** grandes étapes :

- **Collecte des Données**
  Rassembler les textes à analyser à partir de diverses sources : web scraping, bases de données, fichiers internes, etc.

- **Prétraitement**
  - Nettoyage : Enlever les éléments non pertinents (ponctuations inutiles)
  - Tokenisation : Diviser le texte en unités linguistiques (mots, phrases)
  - Lemmatisation/Stemming : Réduire les mots à leur forme de base ou racine

- **Transformation**
  - Vectorisation : Convertir le texte en une forme numérique utilisable par les algorithmes (par exemple, TF-IDF, bag-of-words)
  - Réduction de dimensionnalité : Utiliser des techniques comme LSA (Latent Semantic Analysis) pour simplifier les données textuelles
- **Interprétation et Visualisation**
  Présenter les résultats de manière claire et compréhensible, souvent à l'aide de visualisations

Packages utilisés sur R pour le text mining

Voici quelques de R utilisés pour le text mining

- **tm (Text Mining)**
  - Importation de textes à partir de diverses sources (PDF, CSV, etc.)
  - Création de corpus et nettoyage de texte (élimination des stopwords, mise en minuscule, suppression des ponctuations)
  - Transformation des textes en Document-Term Matrices (DTM)
  - Analyse de fréquence des termes, nuages de mots, etc.
- **text2vec**
  - Tokenisation, création de vocabulaire et encodage de documents
  - Implémentation d'algorithmes de modélisation de texte tels que Latent Dirichlet Allocation (LDA) pour l'analyse de sujets
  - Fonctions pour l'apprentissage supervisé et non supervisé sur des données textuelles

- **tidytext**
  - Conversion de textes en formats de données ordonnées
  - Fonctions pour l'analyse des sentiments, la fréquence des termes, la modélisation de sujets
  - Intégration avec d'autres packages **tidyverse** pour une manipulation et une visualisation facile des données textuelles
- **quanteda**
  - Création et manipulation de corpus, nettoyage et prétraitement du texte
  - Conversion en Document-Feature Matrices (DFM)
  - Analyse de la fréquence des termes, modèles de mots-clés, analyse de similarité, etc.
  - Support pour l'analyse des sentiments et la modélisation des sujets
- **topicmodels**
  - Création de modèles de sujets pour extraire des thèmes à partir d'un corpus
  - Visualisation et interprétation des résultats

Sommaire

1 Introduction

2 Généralités

3 Etapes du processus de text mining et packages utilisés sur R pour le text mining

4 **Résultats de l'application sous R**

5 Conclusion

6 Bibliographie et webographie

Résultats de l'application sous R

paragraph	date	President	Party	text
1	1790-01-08	George Washington	Other	I embrace with great satisfaction the opportunity which now...
2	1790-01-08	George Washington	Other	In resuming your consultations for the general good you ca...
3	1790-01-08	George Washington	Other	Among the many interesting objects which will engage your...
4	1790-01-08	George Washington	Other	A free people ought not only to be armed, but disciplined; t...
5	1790-01-08	George Washington	Other	The proper establishment of the troops which may be deem...
6	1790-01-08	George Washington	Other	There was reason to hope that the pacific measures adopte...
7	1790-01-08	George Washington	Other	The interests of the United States require that our intercours...
8	1790-01-08	George Washington	Other	Various considerations also render it expedient that the ter...
9	1790-01-08	George Washington	Other	Uniformity in the currency, weights, and measures of the Un...
10	1790-01-08	George Washington	Other	The advancement of agriculture, commerce, and manufactu...
11	1790-01-08	George Washington	Other	Nor am I less persuaded that you will agree with me in opini...
12	1790-01-08	George Washington	Other	To the security of a free constitution it contributes in various...
13	1790-01-08	George Washington	Other	Whether this desirable object will be best promoted by affor...
14	1790-01-08	George Washington	Other	Gentlemen of the House of Representatives:
15	1790-01-08	George Washington	Other	I saw with peculiar pleasure at the close of the last session t...
16	1790-01-08	George Washington	Other	It would be superfluous to specify inducements to a measur...
17	1790-01-08	George Washington	Other	Gentlemen of the Senate and House of Representatives:
18	1790-01-08	George Washington	Other	I have directed the proper officers to lay before you, respect...

Figure 3: Discours des chefs d'Etat américains de 1790 à 2014

```
Corpus consisting of 23,469 documents and 4 docvars.
text1 :
"I embrace with great satisfaction the opportunity which now ..."

text2 :
"In resuming your consultations for the general good you can ..."

text3 :
"Among the many interesting objects which will engage your at..."

text4 :
"A free people ought not only to be armed, but disciplined; t..."

text5 :
"The proper establishment of the troops which may be deemed i..."

text6 :
"There was reason to hope that the pacific measures adopted w..."
```

Figure 4: Création des corpus

docs	embrac	great	satisfact	opportun	now	present	congratul	favor	prospect	public
text1	1	1	1	1	1	2	1	1	1	1
text2	0	0	0	0	0	1	0	0	0	0
text3	0	0	0	0	0	0	0	0	0	0
text4	0	0	0	0	0	0	0	0	0	0
text5	0	0	0	0	0	0	0	0	0	0
text6	0	0	0	0	0	0	0	0	0	0

Figure 5: Tokénisation du corpus

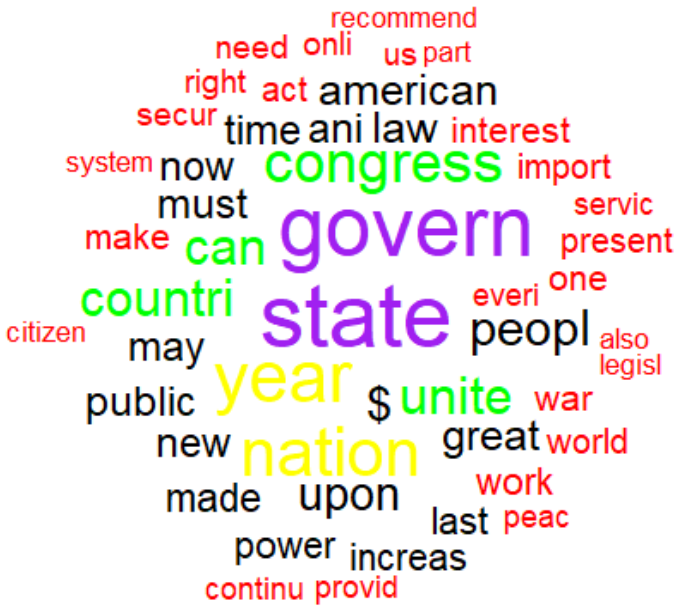


Figure 6: Nuage de mots

Figure 7: Nuage d'un président en particulier


economy
terrorism
military

as far as consists with	freedom	of sentiment its dignity may
spread for the blessings of	freedom	and equal laws.
the United States its legitimate	freedom	. The instructions to our
materials and subsistence, the	freedom	of labor from taxation with
the public debt whenever the	freedom	and safety of our commerce
a force proportioned to its	freedom	, and that the union
, the guardian of the	freedom	and safety of all and
purity of elections, the	freedom	of speech and of the
connected with it with that	freedom	and candor which a regard
new force and a greater	freedom	of action within its proper

Figure 9: Aperçu des phrases comportant le mot 'freedom'

Sommaire

1 Introduction

2 Généralités

3 Etapes du processus de text mining et packages utilisés sur R pour le text mining

4 Résultats de l'application sous R

5 **Conclusion**

6 Bibliographie et webographie

Conclusion

Le traitement des questions ouvertes en utilisant le text mining avec R offre une approche puissante pour analyser et extraire des informations précieuses à partir de données textuelles non structurées. Grâce à des packages comme tm, text2vec, tidytext, et quanteda. Les analystes peuvent efficacement nettoyer, transformer et explorer des corpus de textes. Ces outils permettent d'identifier des tendances, des sentiments et des motifs récurrents, aidant ainsi à une meilleure compréhension des opinions et des perceptions des répondants. En somme, le text mining en R représente une solution robuste et flexible pour le traitement des questions ouvertes, facilitant l'extraction d'insights exploitables et la prise de décisions informées.

Sommaire

1 Introduction

2 Généralités

3 Etapes du processus de text mining et packages utilisés sur R pour le text mining

4 Résultats de l'application sous R

5 Conclusion

6 **Bibliographie et webographie**

Bibliographie et webographie

- Julia Silge and David Robinson, *Text mining with R* (bookdown.org)
- François Husson, *R pour la statistique et la science des données*
- https://sites.google.com/site/rgraphiques/5–applications/textmining-en-langage-r