REPUBLIQUE DU SENEGAL

Un Peuple -Un But -Une Foi

****

Ministère de l'Economie du Plan et de la Coopération

****

Agence nationale de la Statistique et de la Démographie

Ecole nationale de la Statistique et de l'Analyse économique Pierre NDIAYE

TRAITEMENT DES DONNÉES

THEME : Nettoyage et structuration de données textuelles (logs, tweets, réponses ouvertes)

Rédigé par :

Ousseynou GUEYE

Dieynaba KA

Nicolas LOUGUE

Jean Pierre Adiouma NDIAYE

Amadou Ibrahim SOULEYMANE

Elèves Ingénieurs Statisticiens-Economistes

Sous la Supervision de :

M. Mamadou MBODJI

Ingénieur des Travaux statistiques

Enseignant à l'ENSAE

juin 2025

RAPPORT TECHNIQUE : TRAITEMENT DE DONNEES TEXTUELLES

Problématique

À l'ère du numérique, les données textuelles sont omniprésentes : réseaux sociaux, forums, articles de presse, réponses ouvertes dans les enquêtes, entre autres. Cependant, ces données sont généralement non structurées, bruyantes et difficiles à analyser telles quelles. Elles comportent souvent des redondances, des fautes de frappe, des conjugaisons variées ou des expressions idiomatiques. Pour les exploiter efficacement dans une analyse statistique ou dans une démarche de fouille de texte, il est nécessaire de les transformer en données exploitables. C'est tout l'enjeu du traitement automatique du langage naturel (TAL ou NLP en anglais), une discipline à l'interface entre l'informatique, la linguistique et les statistiques.

L'objectif de cette étude est d'illustrer, de manière simplifiée, les étapes fondamentales de traitement des données textuelles à l'aide du langage R. L'approche que nous proposons repose sur trois fondements théoriques essentiels : la tokenisation, la lemmatisation et la vectorisation.

Les Etapes du traitement

a. Tokenisation

Le processus d'analyse textuelle commence par la tokenisation. Cette étape consiste à découper un texte en unités de sens appelées tokens. Ces tokens peuvent être des mots, des groupes de mots (comme les bigrammes ou trigrammes), ou encore des phrases entières. La tokenisation permet de structurer le texte, autrement dit de passer d'un contenu continu et difficilement exploitable à une liste d'éléments manipulables individuellement. Par exemple, la phrase "Les politiques publiques évoluent rapidement" sera découpée en ["les", "politiques", "publiques", "évoluent", "rapidement"].

b. Lemmatisation

Une fois le texte découpé, la lemmatisation vise à réduire chaque mot à sa forme canonique, ou lemme. À la différence du stemming (ou racinisation), qui se contente de tronquer les mots sans considération grammaticale, la lemmatisation respecte les règles de la langue. Par exemple, les formes "étudié",

"étudient" et "étudiants" sont toutes ramenées à "étudier". Cette étape permet d'unifier les formes fléchies d'un mot, ce qui améliore la qualité des analyses statistiques.

c. Vectorisation

Après la lemmatisation, la vectorisation permet de convertir les textes en représentations numériques exploitables par des algorithmes statistiques ou de machine learning. La méthode la plus courante est le sac de mots (Bag-of-Words), où chaque document est représenté par la fréquence des mots qu'il contient. Une approche plus raffinée est la pondération TF-IDF (Term Frequency-Inverse Document Frequency), qui tient compte à la fois de la fréquence d'un mot dans un document et de sa rareté dans l'ensemble du corpus. Ces représentations permettent ensuite d'appliquer des techniques d'analyse comme la classification, la détection de thématiques ou le clustering.

d. Suppression des stop words (mots vides)

En parallèle, on procède souvent à d'autres opérations de nettoyage : suppression des mots vides (comme "le", "de", "et", qui n'apportent pas de valeur analytique), mise en minuscules, suppression de la ponctuation, des chiffres et des caractères spéciaux. L'ensemble de ces transformations vise à préparer le corpus pour les étapes d'analyse ultérieures, comme la détection de thématiques, l'analyse de sentiment ou la classification automatique.

Conclusion

Le traitement des données textuelles constitue une étape cruciale pour toute analyse qualitative ou exploratoire fondée sur des réponses ouvertes ou des documents non structurés. Grâce à des étapes bien définies comme la tokenisation, la lemmatisation et la vectorisation, il est possible de transformer un texte brut en données exploitables. Bien que les outils simples de R présentent certaines limites, ils offrent une base solide pour débuter. Les perspectives offertes par les modèles de langage avancés et les visualisations dynamiques promettent des analyses plus fines et plus adaptées aux enjeux actuels de la recherche en sciences sociales, marketing ou intelligence artificielle.