

## R

# R text mining – Analyse de documents

**REF : SIDS202**

**DUREE : 14h**

**Présentiel Classe virtuelle**

### **PUBLIC**

Chefs de projets statistiques, chargés d'étude, analystes ...

Modalités et délais d'accès : les inscriptions sont fermées 24h avant la 1ère journée de formation.

Accessibilité : Si vous avez des contraintes particulières liées à une situation de handicap, veuillez nous contacter au préalable afin que nous puissions, dans la mesure du possible, adapter l'action de formation.

### **PREREQUIS**

Connaissance des bases de la théorie statistique, avoir suivi la formation R niveau 1 ou avoir une utilisation avancée de R.

### **MODALITES PEDAGOGIQUES**

8 à 10 personnes maximum par cours

1 poste de travail par stagiaire

Remise d'une documentation pédagogique papier ou numérique pendant le stage

La formation est constituée d'apports théoriques, d'exercices pratiques, de réflexions et de retours d'expérience

### **MODALITES D'EVALUATION**

Auto-évaluation des acquis par le stagiaire via un questionnaire en ligne

Attestation de fin de stage remise au stagiaire

## **OBJECTIFS PEDAGOGIQUES**

L'objectif de cette formation est de maîtriser les techniques de Text mining et d'analyse de document à l'aide du logiciel R.

## **PROGRAMME**

### **Utilisation des données textuelles**

- Import de fichier
- Extraction des textes de pdf
- Extraction des textes de html

### **Manipulation et nettoyage des données textuelles**

- Lemmatisation : Regroupement par des mots
- Stemmatisation : Garder la racine des mots
- Suppression des stop-words
- Création de la table des mots

### **Analyse de la table des mots**

- Décomptes des termes
- Fréquence des mots
- Calcul de l'indicateur TF-IDF
- Utilisation de la Loi de Zipf et interprétation
- Création des n-gram des termes du texte
- Corrélation entre les mots

### **Analyse de sentiments des textes**

#### **Modélisation**

- LDA
- Régressions
- Classifications

#### **Visualisation**

- Fréquence des mots
- Wordcloud
- Mots les plus pertinents (avec la loi de zipf)
- Les réseaux de mots

### **Passerelle entre {tidytext} et {tm}**

Deux études de cas corrigés

Version du : 16/12/2021



