

## FORMATION TEXT MINNING

Comprendre les méthodes de la statistique textuelle

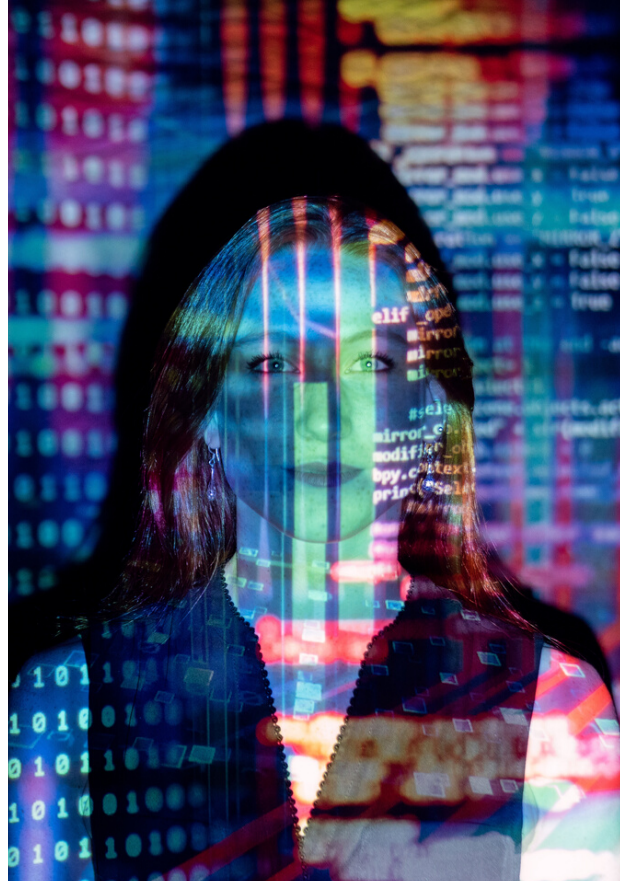
*Machine Learning et Deep Learning pour les données textuelles s'inscrivent dans le cadre du traitement statistique et de la valorisation des données dans tout projet Big Data.*

*Ce cours pratique vous présentera toute la chaîne de conception appliquée au Machine Learning dans un contexte Big Data batch et streaming.*

### OBJECTIFS DE LA FORMATION

- Mettre en œuvre l'extraction des caractéristiques de données textuelles
- Créer des sélections et des classements dans de grands volumes de données textuelles
- Choisir un algorithme de classification
- Évaluer les performances prédictives d'un algorithme

### COURS PRATIQUE SUR 3 JOURS



### FORMATION

#### PUBLIC CONCERNÉ

Ingénieurs/chefs de projet IA, consultants IA  
Toute personne souhaitant découvrir le Text Mining pour le Machine Learning et le Deep Learning

#### PRÉ-REQUIS

Bonnes connaissances en statistiques, du Machine Learning et du Deep Learning  
Expérience requise

#### DURÉE

3 jours

#### FRAIS DE PARTICIPATION

1950 € HT

#### INTERVENANT

Albeiro ESPINAL  
Doctorant Text Mining/NLP

#### INSCRIPTIONS

formation@group-dsi.com

# DÉROULÉ

## Jour 1

### LES APPROCHES TRADITIONNELLES EN TEXT MINING

- Les API pour récupérer des données textuelles
- La préparation des données textuelles en fonction de la problématique
- La récupération et l'exploration du corpus de textes
- La suppression des caractères accentués et spéciaux
- Stemming, Lemmatization et suppression des mots de liaison
- Tout rassembler pour nettoyer et normaliser les données

*Travaux pratiques : La recherche des documents, la préparation, la transformation et la vectorisation des données en DataFrame.*

### FEATURE ENGINEERING POUR LA REPRÉSENTATION DE TEXTE

- Comprendre la syntaxe et la structure du texte
- Le modèle Bag of Words et Bag of N-Grams
- Le modèle TF-IDF, Transformer et Vectorizer
- Le modèle Word2Vec et l'implémentation avec Gensim
- Le modèle GloVe
- Le modèle FastText

*Travaux pratiques : Mise en place des opérations d'extraction des caractéristiques de données textuelles afin d'effectuer des classifications*

[www.group-dsi.com](http://www.group-dsi.com)

## Jour 2

### LA SIMILARITÉ DES TEXTES ET CLASSIFICATION NON SUPERVISÉE

- Les concepts essentiels de similarité
- Analyse de la similarité des termes : distances Hamming, Manhattan, Euclidienne et Levenshtein
- Analyse de la similarité des documents
- Okapi BM25 et le palmarès de classement
- Les algorithmes de classification non supervisée

*Travaux pratiques : Construire un système de recommandation des produits similaires sur la base de la description et du contenu des produits que vous avez choisi.*

### LA CLASSIFICATION SUPERVISÉE DU TEXTE

- Prétraitement et normalisation des données.
- Modèles de classification
- Multinomial Naïve Bayes
- Régression logistique. Support Vector Machines
- Random Forest. Gradient Boosting Machines
- Évaluation des modèles de classification

*Travaux pratiques : Mise en œuvre des classifications supervisées sur plusieurs jeux de données.*

## Jour 3

### NATURAL LANGUAGE PROCESSING ET DEEP LEARNING

- Les bibliothèques NLP : NLTK, TextBlob, SpaCy, Gensim, Pattern, Stanford CoreNLP
- Les bibliothèques Deep Learning : Theano, TensorFlow, Keras
- Natural Language Processing et Recurrent Neural Networks
- RNN et Long Short-Term Memory. Les modèles bidirectionnels RNN
- Les modèles Sequence-to-Sequence
- Questions et réponses avec les modèles RNN

*Travaux pratiques : Construire un RNN pour générer un nouveau texte.*