

## **Rapport de TP - Extraction d'Information**

### **Introduction**

Ce rapport présente une synthèse des analyses effectuées dans le cadre du TP d'extraction d'information. L'objectif de ce travail est de se familiariser avec des outils de traitement automatique du langage pour analyser un corpus de données textuelles et en extraire des informations structurées.

### **Étape 1: Extraction de Mots-Clés**

L'extraction de mots-clés a été réalisée avec la bibliothèque YAKE, permettant de détecter les termes importants dans les documents sans nécessiter de supervision manuelle. Cette technique identifie les mots et expressions les plus significatifs, facilitant l'isolement des thèmes principaux du corpus.

### **Étape 2: Nuage de Mots**

Le nuage de mots a été généré à partir des mots-clés extraits, en excluant les termes peu informatifs à l'aide d'une liste de stopwords enrichie. Cette visualisation aide à repérer les thèmes récurrents grâce à la taille des mots, proportionnelle à leur fréquence d'occurrence dans les textes.

### **Étape 3: Identification des Entités Nommées**

L'identification des entités nommées (NER) a permis de repérer les principales personnes, organisations et lieux mentionnés dans les textes. Bien que cette étape n'ait pas utilisé de modèle avancé, une approche manuelle a permis de classer les entités selon leur catégorie en fonction des mots-clés identifiés.

### **Étape 4: Analyse de Sentiment**

L'analyse de sentiment a été effectuée en utilisant la bibliothèque `textblob`, permettant de calculer la polarité (positive, négative, neutre) et la subjectivité des phrases sélectionnées. Cette étape fournit une évaluation de la tonalité générale des documents, offrant des indications utiles sur les opinions et sentiments.

## Conclusion

Dans le cadre de ce TP, plusieurs techniques d'extraction d'information ont été appliquées pour analyser les données textuelles de manière automatique et méthodique. L'extraction de mots-clés avec YAKE a permis de repérer les termes les plus significatifs, représentant les thèmes récurrents dans les textes. La génération de nuage de mots a fourni une visualisation efficace de ces thèmes, rendant les sujets dominants immédiatement visibles.

L'identification des entités nommées a permis de distinguer les personnes, lieux et organisations les plus cités, offrant ainsi un aperçu des figures et entités influentes. Enfin, l'analyse de sentiment avec textblob a donné des indications sur la tonalité des textes, en identifiant la polarité et la subjectivité des phrases sélectionnées.

Chaque étape de ce TP a contribué à une meilleure compréhension des données textuelles, en exploitant des méthodes de traitement automatique du langage pour extraire des informations pertinentes de manière autonome et structurée.