**Rapport sur l'approche d'extraction utilisant des expressions régulières  et l'analyse textuelle**

**Projet : E-Consignation RCAR**

**Introduction**

**Champs à extraire**

Les champs considérés dans cette analyse sont les suivants :

1. **LOI** : Numéros de référence au format "2.XX.XXX".
2. **LOI Date** : La date la plus proche par indice associée à un numéro de LOI.
3. **Journal (Aljarida) Numéro** : Le numéro entier le plus proche du mot-clé "الجريدة".
4. **Journal (Aljarida) Date** : La date la plus proche associée au mot-clé "الجريدة".
5. **Années** : Années comprises entre 1980 et l'année actuelle.
6. **Mots-clés en arabe** :
   * "الجريدة" (Journal officiel).
   * "الرسمية" (Officiel).
   * "درهم" (Monnaie).
   * "متر" (Unité de mesure).
   * "قرار" (Décision).
   * "مقرر" (Résolution).
   * "حكم" (Jugement).
7. **Contexte** : Une sélection de mots avant et après chaque champ extrait.

 Ce rapport présente une approche systématique pour traiter un corpus textuel, en combinant des règles d'analyse basées sur des regex prédéfinies, des correspondances floues ( Fuzzy Matching ) et des références à des dictionnaires de domaines spécifiques. Cette méthode est illustrée par l'extraction d'informations clés telles que des numéros spécifiques, des dates et des termes en langue arabe.

**Steps :**

1. Identifier des motifs spécifiques dans un corpus textuel, tels que des numéros, des années ou des mots clés.
2. Extraire des informations pertinentes en fonction des besoins d'analyse.
3. Associer les données extraites à des régions ou des catégories en utilisant un dictionnaire prédéfini.
4. Générer un contexte précis autour des éléments identifiés pour une meilleure compréhension.
5. Utiliser une seconde couche d'extractions combinant regex et SQL pour associer des champs selon des indices proximaux.

**Méthodologie**

La méthode adoptée repose sur les étapes suivantes :

1. **Lecture et prétraitement des données**
   * Les textes sont lus à partir de fichiers bruts, divisés en mots individuels et nettoyés des ponctuations inutiles.
   * Expression régulière utilisée pour supprimer la ponctuation :
   * [^\w\s]
2. **Détection de motifs spécifiques**
   * **Numéros de référence** : Les numéros au format « 2.XX.XXX » sont identifiés.
     + Expression régulière :
     + ^2\.\d{2}\.\d{3}$
   * **Années** : Les années comprises entre 1980 et l'année actuelle sont extraites dynamiquement.
     + Génération de l'expression régulière :
     + fr"^(19(8[0-9]|9[0-9])|20(0[0-9]|1[0-9]|2[0-{current\_year % 10}]))$"
   * **Entiers** : Le premier entier dans une chaîne est identifié.
     + Expression régulière :
     + \b\d+\b
3. **Recherche de mots-clés**
   * Des mots-clés en langue arabe tels que « الجريدة » (« journal officiel ») ou « الرسمية » (« officiel ») sont identifiés.
   * Les correspondances floues sont appliquées lorsque nécessaire pour capturer des variations orthographiques.
4. **Extraction de contexte**
   * Pour chaque motif identifié, un contexte est généré autour du mot correspondant. Par exemple :
     + 3 mots avant et 12 mots après pour les mots-clés en arabe.
     + 3 mots avant et après pour les numéros et années.
5. **Association aux régions**
   * Les mots-clés extraits sont comparés à un dictionnaire de localités (également organisées par région). La première correspondance trouvée est utilisée pour déterminer la région associée.
6. **Deuxième couche d’extraction avec regex et SQL**
   * Une analyse supplémentaire est appliquée aux données extraites pour enrichir les champs d’intérêt.
     + La **date de la loi** (« LOI date ») est déterminée comme la date la plus proche par indice du numéro de loi.
     + Le **numéro de l'aljarida (« journal »)** est détecté comme le nombre entier le plus proche par indice du mot-clé correspondant.
     + La **date de l'aljarida** (« journal date ») est identifiée de la même manière.
   * Ces règles sont implémentées à l'aide de scripts SQL appliqués sur la base de données contenant les résultats extraits.
7. **Stockage et analyse des résultats**
   * Les données extraites sont stockées dans une structure tabulaire (DataFrame) pour faciliter l'analyse.
   * Les champs clés incluent :
     + Motif identifié (e.g., numéro, année, mot-clé).
     + Contexte (en termes de mots).
     + Région associée.
     + Extraction finale (si une extraction supplémentaire est nécessaire).
     + Champs enrichis (date loi, numéro aljarida, date aljarida).