Corrigés-exercices-td-1-2

Voici la correction des exercices demandés dans le fichier :

1. **Enregistrement des pages Web dans une colonne `fetch` :**
   - Utilisez la fonction `Edit column > Add column by fetching URLs` pour chaque URL dans une colonne dédiée. Nommez la colonne `fetch` puis cliquez sur Ok.
2. **Extraction du titre depuis la balise `h1` dans une colonne `title` :**
   - Cliquez sur la colonne `fetch > "Edit column" > "Add column based on this column".`
   - Nommer la colonne `title`. Dans le champ Expression, entrez : `value.parseHtml().select("h1")[0].wholeText()`, puis cliquez sur Ok.
3. **Extraction du contenu des balises `p` dans la colonne `content` :**
   - Cliquez sur la colonne `fetch` > "Edit column" > "Add column based on this column".
   - Nommez la colonne `content`. Dans le champ Expression, entrez : `value.parseHtml().select("p").join("|")`.
   - Ensuite, séparez le contenu des balises `p` de la colonne `content` : `Edit cells > "Split multi-valued cells by separator"`, utilisez `|` comme séparateur.
   - Pour ne conserver que le texte des balises dans la colonne `content` : `Edit cell > Transform`, puis dans le champ Expression : `value.parseHtml().wholeText()`
4. **Retrait des caractères HTML et espaces en trop dans `title` et `content` :**
   - Utilisez : `Edit cells > "Common transforms" > "Unescape HTML entities"` pour retirer les caractères HTML.
   - Pour les espaces en trop : `Edit cells > "Common transforms" > "Trim leading and trailing whitespace``.
   - Appliquez ces opérations aux colonnes `title` et `content`
5. **Création de la colonne `texte_minuscule` :**
   - Cliquez sur la colonne `content > "Edit column" > "Add column based on this column"`. Nommez la colonne `texte_minuscule`. Dans le champ Expression, utilisez : `value.toLowercase()`. Appliquez les même méthodes que dans l'étape 4 pour supprimer les espaces et caractères spéciaux HTML
6. **Création d'une facette temporelle (Date facet) :**
   - Convertissez les valeurs de la colonne `date` en cliquant sur `Edit cells > Common transforms > To date`.
   - Pour générer la facette : `date > "Facet" > "Timeline facet"` .
   - En observant la facette, on peut voir que le corpus s'étend du 29/06/2024 au 14/09/2024

7. **Création de la colonne `unique_word` (mots uniques dans `content`) :**
   - Cliquez sur la colonne `content > "Edit column" > "Add column based on this column"`. Nommez la colonne `unique_word`. Dans le champ Expression, utilisez : `value.split(" ").uniques().join(",")`.
8. **Création de la colonne `unique_word_length` (nombre de mots uniques) et facette numérique :**
   - Cliquez sur la colonne `unique_word > "Edit column" > "Add column based on this column"`. Nommez la colonne `unique_word_length`. Utilisez l'expression : `value.split(",").length()` pour compter le nombre de mots uniques.
   - Pour générer la facette : `unique_word_length > "Facet" > "Numeric facet"`. Parmi les observations possibles, on peut voir qu'une majorité de lignes de texte contiennent entre 0 et 20 mots uniques.
9. **Création de la colonne `climat` pour vérifier la présence du mot "climat" dans `title` :**
   - Cliquez sur la colonne `title > "Edit column" > "Add column based on this column"`. Nommez la colonne `climat`. Utilisez l'expression : `value.contains("climat")`.
   - Pour générer la facette : `climat > "Facet" > "Text facet"`.
   - Le titre qui ne contient pas le mot "climat" est "Le monde "est en train d'échouer" à atteindre les objectifs de développement fixés en 2015 par l'ONU, alerte Antonio Guterres ""
10. **Réorganisation et export des colonnes :**
    - Utilisez `"All" > Edit columns > "Re-order / Remove columns"`, puis conservez seulement `id`, `date`, `title`, `content`. Pensez à supprimer toutes les facettes avant l'opération.
    - Exportez au format `.csv` sous le nom `exercices_td1-2_NOM_PRENOM.csv`.

N'oubliez pas de documenter chaque étape et d'envoyer les fichiers demandés.