Exercices-td-1-2

Ces exercices ont pour but de vous familiariser avec le logiciel OpenRefine. Vous documenterez chaque étape dans un document Word / Pages / LibreOffice. Vous enverrez ce document ainsi que le fichier produit à la dernière étape par mail à `nicolas.gutehrle@univ-fcomte.fr`.

Dans OpenRefine, créez un nouveau projet intitulé "ExercicesTD1" en important le fichier `exercice-td1-2.csv`. Une fois le projet prêt, vous réaliserez les actions ci-dessous.

1. Vous enregistrerez la page Web correspondant à chaque URL dans une colonne `fetch`.
2. Le titre de la page est contenu dans une balise `h1`. Extrayez ce titre depuis la colonne `fetch`. Vous l'enregistrerez dans une colonne `title`.
3. Le contenu de la page est situé dans des balises `p`. Collectez ces balises dans une colonne `content`. Vous ferez en sorte qu'il y ait une ligne par balise `p`. De plus, vous ferez également en sorte que seul le contenu des balises `p` soit stocké. Appuyez vous de ce support de cours pour réaliser cette action.
4. Assurez vous de retirer les caractères spéciaux du format HTML ainsi que les espaces en trop dans les colonnes `title` et `content`
5. Produisez une colonne `texte_minuscule`, qui contiendra le contenu de chaque colonne en minuscule. Supprimez les espaces avant ou après les chaînes de texte dans les cellules.
6. Produisez une facette temporelle (Timeline facet) à partir de la colonne `date`. Pouvez-vous dire à partir de cette facette la période sur laquelle s'étend le corpus ?
7. Créez une colonne `unique_word` qui contient une liste de mots uniques pour chaque ligne de la colonne `content`
8. Créez une colonne `unique_word_length` qui indique le nombre de mots uniques dans la colonne `unique_word`. Produisez ensuite une facette numérique à partir de cette colonne. Que pouvez-vous observer ?
9. Créez une colonne `climat` qui indique si la colonne `title` contient le mot "climat". A partir de cette colonne, créez une facette textuelle. A partir de cette facette, pouvez-vous dire quel titre ne contient pas le mot "climat" ?
10. A l'aide du menu `All > Re-order / Remove columns`, faites en sorte de ne conserver que les colonnes `id, date, title, content`. Exportez ensuite ce tableau au format .csv, en nommant le fichier `exercices_td1-2_NOM_PRENOM`, où `NOM` et `PRENOM` correspondent à votre nom et prénom. Pensez à supprimer toutes les facettes

avant de réaliser l'opération. Vous enverrez ce fichier, ainsi que le dossier explicatif de vos démarches à l'adresse mail indiquée ci-dessus.