Projet 2 – Collecte de données en ligne

Dans ce projet, vous ferez la saisie de données provenant de diverses sources en ligne afin de constituer une collection de **corpus de textes**.

Description du problème:

Les données Web sont offertes dans une variété de formats et de langues. Votre tâche consiste à construire une collection de 5 corpus textuels, chacun composé de documents rédigés dans une langue différente (français, anglais, espagnol, italien, et autre). Les documents textuels seront recueillis sur Wikipedia, sur le site web du gouvernement britannique, sur Twitter, dans un document PDF, et à partir d'autres sources.

L'ensemble de données final comprendra toutes les observations de texte, placées en lignes, chaque ligne étant associée à un code de langue spécifique ("Fra", "Ang", "Esp", "Ita", "Aut").

1. Français : le texte des entrées (françaises) Wikipédia de toutes les actrices françaises dont le nom de famille commence par "L".
2. Anglais : le texte de tous les communiqués de presse du gouvernement britannique publiés en 2018.
3. Espagnol : 700 tweets (total) provenant de @realmadrid, @PaulinaRubio, @Armada_esp, et de 2 autres tweeters de votre choix.
4. Italien : le texte du Tutto don Camillo de Giovannino Guareschi (I racconti del Mondo piccolo) - Volume 1 di 5 (PDF), 1 page par ligne.
5. Autre : 500 autres documents textuels, dans une autre langue qui utilisent un alphabet latin.

Conseils pour les textes en français

Modules R suggérés: rvest, tidyverse

1. Visiter la page "Actrices françaises" sur Wikipedia.
2. Jouer avec la page jusqu'à ce que vous identifiez l'URL qui mène aux actrices françaises dont le nom de famille (ou le prénom) commence par un "L".
3. Saisir les URL de toutes les entrées Wikipédia requises (il devrait y en avoir plus de 400... vous devrez réfléchir à la manière de ne retenir que les entrées commençant par un "L").
4. Télécharger les articles Wikipedia dans un dossier local.
5. Extraire le texte principal de chaque article et sauvegarder-le dans une variable de type “chr”.

Conseils pour les textes en anglais

Modules R suggérés: XML, RCurl, stringr

1. Visiter la page "News and Communications" du gouvernement britannique.
2. Identifier l'URL qui donne les communiqués de presse du 1er janvier 2018 au 31 décembre 2018.
3. Saisir les URL de tous les communiqués de presse requis.
4. Télécharger les communiqués de presse dans un dossier local.
5. Extraire le texte principal de chaque communiqué de presse et sauvegarder-le dans une variable de type “chr”.

Conseils pour les textes en espagnol et en italien

Utiliser des modules R spécialisés pour twitter et pour l’extraction du texte d’un fichier PDF. Vous pouvez trouver une copie du PDF en cherchant "Tutto don Camillo" sur Google.