Le module_browser_snt.py

(sources : python.org, crummy.com)

Ce module comporte des fonctions qui permettent d'aborder quelques notions de programmation au sujet des robots web (spiders, crawlers) et la recherche de mots clés dans une page (Term Frequency). Il est conçu à partir des modules **urllib** et **BeautifulSoup 4** ainsi que des fonctions de traitement des chaînes de caractères de Python. Il faut donc installer le module **BeautifulSoup 4** pour pouvoir l'utiliser.

> **python -m pip install beautifulsoup4** dans une invite de commande Windows
> **python pip install beautifulsoup4** dans une invite de commande Linux

Enfin, il suffit de copier le fichier **module_browser_snt.py** dans le même dossier que les programmes Python qui l'utilisent.

module_browser_snt.find_tags(page, tag) :
page est une chaîne de caractères. **tag** est une chaîne de caractères du type 'a' pour une balise <a>, 'img' pour une balise <img>, … La fonction retourne une liste. Elle permet permet d'obtenir toutes les balises du type considéré contenues dans la page.

module_browser_snt.get_attributes(tags_list, attribute) :
tags_list est une liste de balises du même type. **attribute** est une chaîne de caractères du type 'href', 'alt', …, spécifiant l'attribut à considérer. La fonction retourne une liste. Elle permet permet d'obtenir la valeur des attributs du type considéré, pour chaque balises.

module_browser_snt.get_page(url_name) :
url_name est une chaîne de caractères du type 'https://home.cern/fr', … La fonction retourne une chaîne de caractères. Elle permet d'obtenir le code html d'une page web d'adresse **url_name**.

module_browser_snt.get_text(page) :
page est une chaîne de caractères. La fonction retourne une chaîne de caractères. Elle permet d'obtenir tout le texte contenu dans une page html.

module_browser_snt.keep_attributes_with(attributes_list, term) :
attributes_list est une liste de valeurs d'attributs. **term** est une chaîne de caractères. La fonction retourne une liste. Elle permet permet de ne conserver que les valeurs d'attributs contenant le terme considéré.

module_browser_snt.keep_attributes_without(attributes_list, term) :
attributes_list est une liste de valeurs d'attributs. **term** est une chaîne de caractères. La fonction retourne une liste. Elle permet permet de ne conserver que les valeurs d'attributs ne contenant pas le terme considéré.

module_browser_snt.lowercase(text) :
text est une chaîne de caractères. La fonction retourne une chaîne de caractères. Elle permet de passer tout un texte en minuscules.

module_browser_snt.remove_duplicates_words(words_list) :
words_list est une liste de chaînes de caractères, correspondants à une liste de mots. La fonction retourne une liste de mots (chaînes de caractères). Elle permet de supprimer les doublons dans une liste de mots.

module_browser_snt.remove_small_words(words_list, taille)
words_list est une liste de chaînes de caractères, correspondants à une liste de mots. **taille** est un entier correspondant à la taille. La fonction retourne une liste de mots (chaînes de caractères). Elle permet de supprimer des mots dans une liste de mots.

module_browser_snt.remove_symbols(text, old_symbols) :
text est une chaîne de caractères. **old_symbols** est une chaîne de caractères contenant les caractères à enlever. La fonction retourne une chaîne de caractères. Elle permet de supprimer des caractères d'un texte. Les caractères sont en fait remplacés par des espaces.

module_browser_snt.remove_words(words_list, old_words) :
words_list est une liste de chaînes de caractères, correspondants à une liste de mots. **old_words** est une liste de chaînes de caractères contenant les mots à enlever. La fonction retourne une liste de mots (chaînes de caractères). Elle permet de supprimer des mots dans une liste de mots.

module_browser_snt.replace_symbols(text, old_symbols, new_symbols) :
text est une chaîne de caractères. **old_symbols** est une chaîne de caractères contenant les caractères à remplacer. **new_symbols** est une chaîne de caractères contenant les nouveaux caractères, dans le même ordre. La fonction retourne une chaîne de caractères. Elle permet de remplacer des caractères d'un texte.

module_browser_snt.split_into_words(text) :
text est une chaîne de caractères. La fonction retourne une liste de mots (chaînes de caractères). Elle permet de découper un texte en liste de mots.

module_browser_snt.uppercase(text) :
text est une chaîne de caractères. La fonction retourne une chaîne de caractères. Elle permet de passer tout un texte en majuscules.

module_browser_snt.words_frequencies(text, words_list) :
text est une chaîne de caractères. **words_list** est une liste de chaînes de caractères, correspondants à une liste de mots. La fonction retourne une liste de doublets (chaînes de caractères, entier). Elle permet d'obtenir le nombre de mots identiques contenus dans un texte. La liste est triée par ordre croissant d'occurrence des mots.