#!/usr/bin/env python3

# -\*- coding: utf-8 -\*-

"""

SCRIPT :

- parcourir les balises html avec BeautifulSoup

(pour télécharger librairie : Anaconda > Environments > Not installed > BeautifulSoup> Apply)

- extraire les balises : language name, upsidname, alternate name

- extraire leur contenu

SCRIPT (de ce qu’il reste à faire) :

- parcourir tous les fichiers

- extraire leurs balises et contenu

- enregistrer ces données dans un json "nomalisationLangues.Json"

"""

from bs4 import BeautifulSoup

nom\_repertoire = "Langues/"

nom\_file = nom\_repertoire+"FRENCH"

file = open(nom\_file).read()

soup = BeautifulSoup(file, "html.parser")

"""recherche des clefs/valeurs"""

for balise in soup.find\_all("td"):

if balise.string == "Language name:":

key\_name\_language = balise.string

value\_name\_language = balise.find\_next().string

if balise.string == "UPSID number:":

key\_upsid = balise.string

value\_upsid = balise.find\_next().string

if balise.string == "Alternate name(s):":

key\_alt\_name = balise.string

if balise.find\_next().string == None:

value\_alt\_name = "No alternate name(s)"

else:

value\_alt\_name = balise.find\_next().string

print(key\_name\_language, value\_name\_language, key\_upsid, value\_upsid, key\_alt\_name,value\_alt\_name)