ESI
مدرسة علوم المعلومات
ⵜⵉⵎⵣⵔⵜ ⵜⴰⵙⴷⴰⵡⵉⵜ ⵏ ⵜⵣⵔⴼⵜ
ECOLE DES SCIENCES DE L'INFORMATION
ISCHOOL.MA
Génie des Données et de Veille

PROMOTION
2018 - 2021

GÉNÉRATION AUTOMATISÉE DE BASES DE DONNÉES ÉVOLUTIVES SUR LE MARCHÉ DE L'EMPLOI ET D'INDICATEURS POUR L'ANALYSE DE L'ÉVOLUTION DES OFFRES ET DES DEMANDES D'EMPLOIS

DESCRIPTIF DU CODE SOURCE

Réalisé par :

NDAM II Christophe Andy Weber

PRÉREQUIS

- PyCharm (Environnement de développement)
- Scrapy (framework)

Modules (à avoir au moins) :

- Pandas (librairie)
- Time (librairie)
- Requests (librairie)

PRINCIPALES LIBRAIRIES CRÉÉES ASSOCIÉES AUX DIFFÉRENTES TÂCHES DES ROBOTS CRAWLERS

Librairies	Descriptions / rôles
data_cleaning.py	Comporte plusieurs fonctions pour les différents types de nettoyages sur chacune des données collectées
string_transform.py	Créée pour les différentes transformations éventuelles à effectuer sur les chaines de caractères
persist.py	Assure la sauvegarde (la persistance) des données collectées, nettoyées et structurées
horo_format.py	Permet la gestion des dates, des heures, des minutes, des secondes et l'horodatage des données collectées au cours de l'ensemble des opérations
combine_export.py	Gère la combinaison des données collectées en gérant les éventuelles redondances

DÉPENDANCES

- data_cleaning.py est **indépendant**
- string_transform.py se requiert de **data_cleaning.py**
- persist.py se requiert **pandas de Python** et éventuellement de **horo_format.py**
- horo_format.py requiert de **time de Python** et éventuellement de **data_cleaning.py**
- combine_export.py requiert de **persist.py** et de **horo_format.py**

Remarque :

Les noms des variables, fonctions et scripts crées ont été définis de telle sorte qu'ils soient clairs quant à leur utilisation dans les contextes.

TESTS DE CAS (scripts de cas particuliers)

Librairies	Descriptions / rôles
test_emploima.py	Test de cas d'extraction d'offres sur emploi.ma
test_profilema.py	Test de cas d'extraction de demandes sur emploi.ma
test_rekrute.py	Test de cas d'extraction d'offres sur rekrute.ma

ROBOTS CRAWLERS (scripts généraux)

Librairies	Descriptions / rôles
my_crawler_em.py	EMPSmartSpider, qui dans le cadre de ce travail est le robot crawler crée pour rassembler, nettoyer, structurer et sauvegarder les données relatives à l'offre en matière d'emploi, en particulier sur la plateforme Emploi.ma.
my_crawler_rek.py	REKSmartSpider, qui est dans le cadre de ce travail est le robot crawler crée pour rassembler, nettoyer, structurer et sauvegarder les données relatives à l'offre en matière d'emploi, en particulier sur la plateforme Rekrute.com.
my_crawler_pro.py	PROSmartSpider qui dans le cadre de ce travail est le robot crawler créé pour rassembler, nettoyer, structurer et sauvegarder les données relatives à la demande en matière d'emploi, en particulier sur la plateforme Emploi.ma.

Note :

Leur utilisation optimale nécessite l'ensemble des librairies créées à cet effet.

Quelques CODES DE TESTS élaborés à toutes fins utiles

(dans le zip **annexes_AME_Christophe**)

- dataframe.py (exemple de dataframe)
- nlp.py (utilisation de NLTK)
- tests_empty_words.py (suppression de mots vides avec NLTK)
- page_simple.py (cas d'une page simple)

MES RECOMMANDATIONS

- Disposer d'une connexion internet active avant même d'ouvrir le projet lors de l'utilisation de PyCharm pour éviter toute erreur incompréhensible, pourtant liée uniquement à l'absence de connexion internet au démarrage.

- Pour éviter au maximum des erreurs d'exécution il est TOUJOURS préférable de tester des cas particuliers avant de généraliser et d'automatiser l'ensemble du processus sur toutes les pages d'une plateforme donnée.

NDAM II Christophe Andy Weber