**Rapport Projet Python**

1. Description du projet

Le projet réalisé dans le cadre du cours de Python consiste en la création de deux outils distincts : l’un permettant l’aspiration de textes sur un site Web, l’autre permettant la classification de ces textes en fonction de leur thème. Pour la première partie du projet nous avons choisi de récolter une vingtaine d’articles de la rubrique « Culture » du journal en ligne Le Monde (en ligne : lemonde.fr). Dans la seconde partie du projet, nous avons décidé de les classer en fonction des cinq catégories : Art (sculpture, peinture, photographie etc.), Scène (théâtre, opéra etc.), Musique, Cinéma et Littérature.

1. Les packages utilisés

Pour la partie de crawling, nous avons utilisé la bibliothèque BeautifulSoup. Ce module peut être téléchargé via Pip. C’est ce module qui permet de récupérer le contenu html d’une page et de manipuler les différentes balises pour récupérer les informations nécessaire principales et utile pour notre tâche. Nous avons également utilisé le module requests afin de faire une requête à partir de l’url de la page que l’on souhaite récupérer. Enfin nous avons utilisé le moule os pour manipuler les dossiers (création des dossiers qui contiendront le corpus de la première partie et les textes classés pour la seconde partie du projet). Ces deux modules sont déjà disponibles avec la version de base de python 3.

Pour la seconde partie, nous avions tout d’abord comme objectif d’utiliser le module Scikit-learn. Nous avons abandonné l’idée d’utiliser ce module car il s’agit d’un module de Machine Learning et il nous aurait donc fallu beaucoup trop de données d’apprentissage annotées pour réaliser l’apprentissage du module et il aurait donc fallu beaucoup de temps pour mettre en œuvre cette partie. Nous avons donc décidé de rester sur un modèle utilisant un lexique pour faire une recherche de champ lexical dans les textes tokenisé avec le module spacy.

1. le découpage en modules, les chaines de traitements mises en œuvres (faites des schémas)
2. les choix algorithmiques et les principales structures de données (fichiers, tableaux, etc.)
3. une évaluation sommaire des sorties
4. les bogues constatés
5. les améliorations à apporter et extensions prévues.