**Rapport du projet NoSQL : Scrapping du site Allociné**



****

Sayah MEDELLEL  
Moussa CAMARA  
Oussama ABDELHEDI

**Le rapport final**

**L’équipe du projet :**

**Sayah MEDELLEL, Moussa CAMARA, Oussama ABDELHEDI**

**Période :**

**Semestre 2 (2021)**

**Établissement :**

**ESGI, École Supérieure de Génie Informatique**

**Sujet :**

**Scraping du site Allociné**

**Professeur en charge :**

**Mr. Jean-Baptiste KOUEK**

**Sommaire**

**Ce rapport est composé :**

* **D’une partie présentation générale du projet, avec une introduction, des points non résolus, des difficultés rencontrées et d’un bilan global**
* **D’une annexe technique composée des dossiers concepteur et utilisateur**
* **D’une annexe informelle composée d’une bibliographie et d’un glossaire**

**Le dossier concepteur présente la conception du projet ainsi qu’une analyse de l’application regroupant plusieurs éléments :**

* **Une description de la structure du projet**
* **Une description des méthodes principales**
* **Une description des choix d’implémentation et d’autres détails techniques**

**Le dossier utilisateur regroupe l’ensemble des informations permettant de comprendre l’installation et l’utilisation de l’application.**

**La bibliographie regroupe les références des documents qui nous ont aidés pour réaliser le projet et le glossaire regroupe le vocabulaire spécifique et technique utilisé.**

**Introduction**

**Dans le cadre de la réalisation d’un projet en NoSql, nous avons réalisé une application qui permet de récupérer les données du site** [Allociné](https://www.allocine.fr/) **et de les analyser. Allociné est un site de référence en terme de séries TV, rassemblant plus de 10 millions de visiteurs par mois. Le site étant assez vaste, nous nous sommes focalisés sur un thème qui nous intéressait tous : les séries télévisées. En outre, un très grand nombre d’informations concernant les séries est disponible sur Allociné, nous avons donc limité les données à récupérer :**

* Aux informations principales de la séries (titre, date de sortie …)
* Aux avis de la série (commentaires, notes …)

Le projet est scindé en trois parties :

* Data Scraping : automatisation d’un processus permettant de recueillir les données de plusieurs séries et de les regrouper dans plusieurs fichiers structurés
* Data Processing : mise en place de plusieurs processus permettant d’extraire et/ou de rassembler des données brutes des fichiers structurés et des données stockées dans une base de données en vue de produire des informations utilisables
* Data Visualisation : représentation graphique des données afin de communiquer les informations de manière claire et efficace

**Pour réaliser ce projet, l’équipe est composée de trois membres :**

* **Sayah MEDELLEL**
* **Moussa CAMARA**
* **Oussama ABDELHEDI**

**Points non résolus**

**Globalement, les objectifs du projet ont été atteints. Pour rappel, le projet est scindé en trois parties […]**

**Difficultés rencontrées**

**Nous avons rencontré plusieurs difficultés (humaines et techniques) pour la réalisation du projet. La crise sanitaire de la Covid-19 et le confinement imposé par le Gouvernement ont eu un vrai impact, empêchant les sorties et donc les rencontres à l’extérieur. Nous étions donc forcés de collaborer à distance, ce qui a provoqué quelques désagréments.**

**Tout d’abord, on peut évoquer le manque de communication, parfois nous ne recevions pas les messages en temps et en heure, ce qui est un inconvénient lorsqu’il faut modifier le travail. L’absence physique et les échanges à distance ont favorisé les pertes d’informations et ont pu créer une sorte de manque de cohésion. Les idées étaient plus difficiles à exprimer et il s’ensuivit des incompréhensions entre les membres de l’équipe.**

**Enfin, nous avons fait face à plusieurs problèmes techniques (notamment au niveau de la conception). Concernant les installations, nous avons également eu quelques problèmes à installer et configurer correctement certains outils. Il était nécessaire de comprendre tous les éléments qui permettent l’installation et la configuration de ces outils, dont le fait de trouver les dernières libs à jour (précompilés).**

**Afin de surmonter toutes ces difficultés nous avons dû nous entraider, communiquer plus régulièrement entre nous à travers les systèmes de messageries et nous aider des différents documents qui étaient à notre disposition (cours, sites web).**

**Bilan**

**La réalisation de ce projet a permis au groupe d’acquérir des compétences sur la programmation en python […]**

**Annexe technique**

**I) Dossier concepteur**

**Introduction**

**Ce document explique la phase d’avant-projet, phase qui va permettre de définir les objectifs clairs de l’application et notamment la manière dont celle-ci va être réalisée. Cette phase est nécessaire afin de comprendre au mieux le sujet posé et de l’aborder convenablement. Il est clair que cela va permettre de simplifier les étapes de réalisation par la suite. Globalement, elle va permettre de structurer, organiser et planifier le projet.**

**1) Spécification détaillée de la structure du système**

Afin de comprendre la méthode de conception appliquée, nous avons effectué une description détaillée du système, réparti en un aspect technique et un aspect complémentaire.

**1) A) Aspect technique**

L’application est composée d’un ensemble de fonctions qui permettent la vérification de la DTD et donc de la validation d’un fichier XML. Globalement, les fonctions vont permettre à un utilisateur :

- De saisir le chemin ainsi que le nom des fichiers

- De vérifier le contenu de la DTD

- De valider le XML

**I) B) Aspect complémentaire**

**Globalement, le but d’une telle structure est de fournir un programme qui répond à plusieurs critères essentiels : flexibilité, durabilité et adaptabilité. D’ailleurs le programme conçu permet d’assurer complètement l’aspect de la dynamicité. En outre, le fonctionnement du programme et les résultats s’adaptent par rapport aux présentes dans le site. La date n’influe pas sur le comportement du programme, ce qui encore une fois est intéressant pour l’utilisateur.**

**I) C) Comportement**

**Nous allons détailler le fonctionnement du programme et des fonctions.**

**II) Dossier utilisateur**

**Introduction**

**Le dossier utilisateur permet à l’utilisateur de lui apporter tous les renseignements nécessaires et suffisants pour une bonne utilisation et compréhension de l’application. Il explique à l’utilisateur comment utiliser le programme, corriger les possibles erreurs et comprendre les résultats du programme.**

**II) A) Présentation de l’application**

**Le produit final est une application qui permet de scraper le site Allociné, les formatter et enfin les analyser efficacement à l’aide d’outil de visualisation. Pour obtenir les informations sur les séries, l’utilisateur lance le programme qui affichera les résultats détaillés ou bien les possibles erreurs.**

**II) B) Fonctionnalités**

**Annexe informelle**

**Bibliographie**

* **Cours de Mr. KOUEK**
* **https://stackoverflow.com**

**Glossaire**