TADYSZAK Lucas BOMPARD Jules

Licence 3 Info

Projet Génie logiciel

Rapport WebMagic

Table des matières

[I. Présentation globale du projet 2](#_Toc65512547)

[1) Utilité du projet 2](#_Toc65512548)

[2) Configuration requise 2](#_Toc65512549)

[3) Quelques commandes Maven 3](#_Toc65512550)

[II. Description du projet 3](#_Toc65512551)

[III. Historique du Git 4](#_Toc65512552)

[1) Composition de l’équipe 4](#_Toc65512553)

[2) Activité sur le projet 4](#_Toc65512554)

[3) Utilisation des branches 4](#_Toc65512555)

[4) Utilisation des pull requests 4](#_Toc65512556)

[IV. Architecture logicielle 5](#_Toc65512557)

[1) Utilisation de bibliothèques extérieures 5](#_Toc65512558)

# Présentation globale du projet

## Utilité du projet

WebMagic est un framework crawler évolutif pour Java. Il couvre l'ensemble du cycle de vie du crawler : téléchargement, gestion des url, extraction de contenu et persistant. Il permet de simplifier le développement d'un crawler pour un site quelconque.

Un crawler est un logiciel qui a pour principale mission d'explorer le Web afin d'analyser le contenu des documents visités et les stocker de manière organisée dans un index. Les crawler parcourent en permanence, de manière autonome et automatique divers sites Web et pages Internet pour trouver de nouveaux contenus ou d'éventuelles mises à jour de contenu qui ont été explorées dans le passé. Derrière l'activité se cache une autre tâche : la fonction d'indexation des pages Web en fonction de la qualité du contenu (évalué selon les normes configurées en amont), aidant ainsi les moteurs de recherche à classer les pages Internet dans les résultats affichés.

## Configuration requise

Ajouter les dépendances écrites ci-dessous dans le fichier pom.xml.

WebMagic utilise slf4j avec l'implémentation de slf4j-log4j12. Si vous avez personnalisé votre implémentation slf4j, il faut exclure slf4j-log4j12. Simple Logging Facade for Java fournit une API de journalisation Java au moyen du design pattern Facade.

<dependency>

<groupId>us.codecraft</groupId>

<artifactId>webmagic-core</artifactId>

<version>0.7.4</version>

</dependency>

<dependency>

<groupId>us.codecraft</groupId>

<artifactId>webmagic-extension</artifactId>

<version>0.7.4</version>

<exclusions>

<exclusion>

<groupId>org.slf4j</groupId>

<artifactId>slf4j-log4j12</artifactId>

</exclusion>

</exclusions>

</dependency>

## Quelques commandes Maven

WebMagic est un projet purement Java, donc il sera très simple de télécharger et compiler le code source avec Maven. Vous trouverez ci-dessous la liste des commandes Maven les plus courantes à utiliser.

mvn validate : valide si le projet est correct et que toutes les informations nécessaires sont disponibles.

mvn compile : compile le code source du projet.

mvn test : test le code source compilé en utilisant un cadre de test unitaire approprié.

mvn package : prend le code compilé et le conditionne dans son format distribuable, comme un JAR.

mvn integration-test : traite et déploie le paquet si nécessaire dans un environnement où des tests d'intégration peuvent être effectués.

mvn verify : effectue tout contrôle pour vérifier que le paquet est valable et répond aux critères de qualité.

mvn install : installe le paquet dans le dépôt local, pour l'utiliser comme dépendance dans d'autres projets locaux.

mvn deploy : effectué dans un environnement d'intégration ou de publication, il copie le paquet final dans le dépôt distant pour le partager avec d'autres développeurs et projets.

mvn clean : nettoyer les artefacts créés par des constructions antérieures.

mvn site : génère la documentation du site pour le projet.

# Description du projet

Le projet contient un Readme qui explique brièvement l’outil WebMagic. On n’apprend pas dans le Readme comment utiliser le projet WebMagic avec les commandes pour compiler et exécuter ce dernier. Mais on sait à l’aide d’un badge que la compilation de la branche Master du projet a échoué. Ce projet est découpé en plusieurs parties peu détaillées.

1. Introduction : description du projet en 2 lignes ;
2. Caractéristiques : énumération des différentes fonctionnalités disponibles ;
3. Installation : dépendances nécessaires ;
4. Exemples : morceaux de codes qui montrent comment écrire un crawler ;
5. Documentation et autres : un lien qui redirige vers le site pour consulter la documentation plus détaillée, ainsi que des informations sur le projet WebMagic.

Le projet possède une documentation accessible via le site [WebMagic](http://webmagic.io/docs/) qui explique à quoi sert ce projet dans l’introduction de la documentation, comment installer le projet, récupérer le code source, compiler ce code source avec Maven ainsi que des exemples de crawler. La documentation n’est pas très détaillée, il faut avoir des connaissances en Java et Maven pour comprendre la documentation mais aussi mettre en place le projet. Cette documentation est incomplète, il manque des informations pour compiler et exécuter le projet sans avoir à faire des recherches sur d’autres sites.

# Historique du Git

## Composition de l’équipe

Le dépôt Github est composé de 40 contributeurs. On constate le créateur du projet a beaucoup contribué au début avec plus de 800 commits, les autres font très peu de commits (entre 1 et 5).

## Activité sur le projet

Le projet a une grosse activité du 21 avril 2013 au 21 juin 2014 puis ralentit jusqu’au 1er janvier 2018. Le projet ne percevra aucune modification pendant une année et demie. Du 23 juillet 2019 à aujourd’hui, le projet a connu de nouvelles modifications avec de nouveaux contributeurs mais ces dernières sont mineures et on peut dire qu’aujourd’hui le projet est stable.

## Utilisation des branches

Une branche est utilisée pour maintenir une version particulière du projet dans un but précis et fixé à l’avance. Elles peuvent servir à développer une nouvelle fonctionnalité sans polluer la branche supérieure avec des modifications non validées ou à garder une branche correspondant à une version stable définie.

Le dépôt du créateur code4craft contient une branche par défaut qui se nomme develop, deux branches actives master et revert-977-build et 20 autres branches inactives.

## Utilisation des pull requests

Un pull request est une méthode qui permet de soumettre des contributions à un projet utilisant un système de contrôle de version distribué comme Git. Un pull request se produit lorsqu'un développeur demande que les modifications apportées à un dépôt externe soient prises en compte pour être incluses dans le dépôt principal d'un projet.

Le projet utilise le mécanisme des pull requests avec un total de 144 pull requests dont 121 sont clos et 23 sont encore en attentes. Cette information montre une importante contribution active sur le projet WebMagic.

# Architecture logicielle

## Utilisation de bibliothèques extérieures

Nous avons vu dans la présentation du projet WebMagic que ce dernier utilisé Maven. Les bibliothèques externes utilisées par Maven sont indiquées dans le fichier pom.xml. On comptabilise