ThirdReport

Google Summer of Code 2014 Report 3

Project Name: NUTCH-841 Create a Wicket-based Web Application for Nutch 2.X

Report date: 30th July 2014

Student Name: Fjodor Vershinin

Mentor Name: Lewis John McGibbney (lewismc)

Development Codebase: https://bitbucket.org/feodorv/uinutch/

Project description

Main goal of this project is to create an Apache Wicket-based Web GUI for Apache Nutch 2.X.

Review of Previous Actions

- ~~Change entire build structure to Ant + Ivy as per existing 2.x codebase~~
- Implement seed information upload using REST API
- ~~Create embedded database for storing crawls, user credentials, and so on~~
- ~~Write tests and some documentation/javadoc~~

Objectives

Change build structure to Ant + Ivy

Project has been switched to Ant + Ivy build system and integrated into 2.x codebase. There were some issues concerning this move. One of them was library dependencies, for example nutch is tightly coupled with hbase and hadoop, and they are dependent from outdated jersey and jetty. So, I downgraded dependencies of my project and had rewritten relevant parts of application.

Add instances management

The main issue concerning this objective were implementation of connection status checks. I'd done it with a constant polling of nutch API servers. But what application should to do, if nutch server is going down?

Logs by REST API

Ability to get logs is very important in order to understand exceptional situations. However, I would propose to implement additional API point, which can answer, if connection with hbase and indexing server is established, because for now, it is not that easy to understand, if something is going wrong and recover from exceptional situation.

Fix seed management UI layout

UI of seed management component is not that good, we should decide how to improve it. For example, ability to add url regex filters would be nice.

Add some security support

The main problem concerning security in this project is how much security we want. Actually, we already have embedded database support, so it's possible to create system with a roles, users and so on. Or we can stick with just a simple username/password without separate users.

Run application on ASF dedicated virtual machine

TBC by lewismc once code has been committed to codebase and we are ready to deploy the server.

Cover application with unit tests

Tests are quite important. I had tried to develop this application with TDD approach, but I'd failed it because of continuous refactorings. Anyway, I would cover application with tests when codebase will be stable enough.

Contributions to Nutch community

Future Actions

To be completed by student

Mentors Comments

To be completed by mentor

Signed: