GoogleSummerOfCode Giving HTML5 support for Apache Nutch 2.x

Giving HTML5 support for Apache Nutch 2.x

Description

The project is aimed at giving Html5 support to Apache Nutch 2.x with using a java library. With this project two goals is aimed. First one is implementation of a new parser which has to follow WHATWG HTML5 specification. Second one is implementation of a new plugin which uses newly implemented parser and extracts new elements of HTML5.

Reports

Reports will be added here.

Documentation

Documents will be added here.

Jira Issues

Issues will be added here.