DroidsProposal

Droids, an intelligent standalone robot framework

Abstract

Droids aims to be an intelligent standalone robot framework that allows to create and extend existing droids (robots).

Proposal

As a standalone robot framework Droids will offer infrastructure code to create and extend existing robots. In the future it will offer as well a web based administration application to manage and controll the different droids which will communicate with this app.

Droids makes it very easy to extend existing robots or write a new one from scratch, which can automatically seek out relevant online information based on the user's specifications. Since the flexible design it can reuse directly all custom business logic that are written in java.

In the long run it should become umbrella for specialized droids that are hosted as sub-projects. Where an ultimate goal is to integrate an artificial intelligence that can control a swarm of droids and actively plan/react on different tasks.

Background

The initial idea for the Droids project was voiced in February 2007 from Thorsten Scherler mainly because of personal curiosity and developed as a labs project. The background of his work was that Cocoon trunk (2.2) did not provide a crawler anymore and Forrest was based on it, meaning we could not update anymore till we found a crawler replacement. Getting more involved in Solr and Nutch he saw the request for a generic standalone crawler.

For the first version he took nutch, ripped out and modified the plugin/extension framework. However the second version were not based on it anymore but was using Spring instead. The main reason was that Spring has become a standard and helped to make Droids as extensible as possible.

Soon the first plugins and sample droids had been added to the code based.

Rationale

There is ever more demand for tools that automatically do determinate tasks. Search engines such as Nuts are normally very focused on a specific functionality and are not focused on extensibility. Furthermore there are manly focused on crawling, requesting certain pages and extract links to other pages, which in our opinion is only one small area for automated robots. While there are a number of existing crawler libraries for various task, each of them comes with a custom API and there are no generic interface for automatically determining which crawler (droids) to use for a specific task.

The Droids project attempts to remove this duplication of efforts. We believe that by pooling the efforts of multiple projects we will be able to create a generic robot framework that exceeds the capabilities and quality of the custom solutions of any single project. The focus of Droids is not a single crawler but more to offer different reusable components that custom droids (robots) can use to automate certain tasks. An intelligent standalone robot framework project will not only provide common ground for the developers of crawler but as well for any other automated application (robots) libraries.

Initial Goals

The initial goals of the proposed project are:

- Viable community around the Droids codebase
- Active relationships and possible cooperation with related projects and communities (e.g. reusing Tika for text extraction)
- Generic robot API for crawling, extracting structured text content and/or new task, filtering task and handle the content
- Flexible extension and plugin development to create a wide range of functionality
- Fuel develop of various droids and bring the current wget style crawler to state-of-the-art level

Current Status

Meritocracy

All the initial committers are familiar with the meritocracy principles of Apache, and have already worked on the various source codebases. We will follow the normal meritocracy rules also with other potential contributors.

Community

There is not yet a clear Droids community. Instead we have a number of people and related projects with an understanding that an intelligent standalone robot framework project would best serve everyone's interests. The primary goal of the incubating project is to build a self-sustaining community around this shared vision.

Core Developers

The initial set of developers comes from various backgrounds, with different but compatible needs for the proposed project.

Alignment

As a generic robot framework Droids will likely be widely used by various open source and commercial projects both together with and independent of other Apache tools. Apache projects like Cocoon, Lenya and Forrest are potential candidates for using different droids as an embedded component.

Known Risks

Orphaned products

Till now only one company is known to use Droids in a productive environment however there is a constant interest in a generic robot framework expressed by various Apache committers. For many potential users the existing tools are to complicated or too much focused on a specific usecase which will help to gain a bigger user base.

Once the project gets started we can quickly build the wget style droids to a feature level of existing tools based on plugin development that reuses code from sources mentioned below. After that we believe to be able to quickly grow the developer and user communities based on the benefits of a generic framework offering reusable plugins and different droids over custom alternatives.

Inexperience with Open Source

All the initial developers have worked on open source before and many are committers and PMC members within other Apache projects.

Homogenous Developers

The initial developers come from a variety of backgrounds and with a variety of needs for the proposed toolkit.

Reliance on Salaried Developers

Some of the developers are paid to work develop certain functionality on this, but the proposed project is not the primary task for anyone.

Relationships with Other Apache Products

Droids is related to at least the following Apache projects. None of the projects is a direct competitor for Nutch, but there are many cases of potential overlap in functionality where droids will try to reuse this functionality in providing wrapper classes.

- http://lucene.apache.org/nutch/ - The Nutch project already contains a crawler/parser framework that does many of the things that Droids is designed to do but the focus is on a unique use case, where droids is aimed to be global re-usable.
- http://incubator.apache.org/tika/ - Apache Tika is a toolkit for detecting and extracting metadata and structured text content from various documents using existing parser libraries.
- http://hadoop.apache.org/core/ - Hadoop is a software platform that lets one easily write and run applications that process vast amounts of data.

A Excessive Fascination with the Apache Brand

All of us are familiar with Apache and we have participated in Apache projects as contributors, committers, and PMC members. We feel that the Apache Software Foundation is a natural home for a project like this.

Documentation

The main documentation is distributed with the code

- Docu
- DocuDeployed

Initial Source

Droids will start with the code base that have been developed in the Apache Labs project:

- code base

Source and Intellectual Property Submission Plan

All seed code and other contributions will be handled through the normal Apache contribution process.

We will also contact other related efforts for possible cooperation and contributions.

External Dependencies

Droids will mainly depend on the Spring core distribution.

Cryptography

Droids itself will not use cryptography, but it is possible that some of the external libraries will include cryptographic code to handle different features.

Required Resources

Mailing lists

- droids-dev@incubator.apache.org
- droids-commits@incubator.apache.org
- droids-private@incubator.apache.org

Subversion Directory

- https://svn.apache.org/repos/asf/incubator/droids

Issue Tracking

- JIRA Droids (DROIDS)

Other Resources

- none

Initial Committers

Name	Email	CLA
Thorsten Scherler	thorsten at apache dot org	yes
Ryan McKinley	ryan at apache dot org	yes
Grant Ingersoll	gsingers at apache dot org	yes
Oleg Kalnichevski	olegk at apache dot org	yes

Affiliations

Name	Affiliation
Thorsten Scherler	Freelance r

Sponsors

Champion

Grant Ingersoll

Nominated Mentors

- Ross Gardler <rgardler at apache dot org>
- Paul Fremantle <pzf at apache dot org>
- Grant Ingersoll <gsingers at a.o>

Sponsoring Entity

- Apache HttpComponents
- Apache Lucene