OpenNLPProposal

OpenNLP Proposal

The following is a proposal for a new top-level project within the ASF.

Abstract

OpenNLP is a Java machine learning toolkit for natural language processing (NLP).

Proposal

OpenNLP is a machine learning based toolkit for the processing of natural language text. It supports the most common NLP tasks, such as tokenization, sentence segmentation, part-of-speech tagging, named entity extraction, chunking, parsing, and coreference resolution. These tasks are usually required to build more advanced text processing services.

The goal of the OpenNLP project will be to create a mature toolkit for the abovementioned tasks. An additional goal is to provide a large number of pre-built models for a variety of languages, as well as the annotated text resources that those models are derived from.

Background

OpenNLP was started in 2000 by Jason Baldridge and Gann Bierner while they were graduate students in the Division of Informatics at the University of Edinburgh. OpenNLP, broadly speaking, was meant to be a high-level organizational unit for various open source software packages for natural language processing; more practically, it provided a high-level package name for various Java packages of the form opennlp.*. The first OpenNLP software package was the Grok natural language parsing toolkit, which was also the genesis of what is now called the OpenNLP Toolkit. The software released on the OpenNLP sourceforge site (started in 2000, along with Grok) was simply a set of interfaces defined in the package opennlp.common and referred to as the OpenNLP Java API. The actual implementations of natural language processing components were provided in Grok, along with code for sentence parsing with Combinatory Categorial Grammar. This code was used heavily in both Baldridge's and Bierner's dissertations. The first paper that used Grok, and especially the components that would become the OpenNLP Toolkit is Hockenmaier, Bierner and Baldridge (2000) (later updated as the journal article Hockenmaier, Bierner, and Baldridge (2004)).

In 2003, it was decided to remove the NLP infrastructure from Grok as there was a clear separation between the basic text processing components and the syntactic and semantic analysis components. At the same time, Grok was rebranded as OpenCCG (openccg.sf.net). The final release of the OpenNLP Java API was made in March 2003; the new OpenNLP Toolkit was created from the API and the Grok text processing components, with version 1.0 being released in April 2004. The OpenNLP Toolkit and OpenCCG have evolved independently since then and have mostly independent and active developer and user communities. OpenCCG is primarily used in the academic community, while OpenNLP has considerable use in both academia and industry. As in indication of the academic impact of OpenNLP, a search on Google scholar (done in March 2010) returned about 650 publications citing the package. Some of these include the OpenNLP website and a few non-publications plus some self-citations. Based on a scan of these results, we estimate that about 500 actual publications have used OpenNLP in their work, and there are an addition 50 or so quasi-publications like surveys and instruction manuals.

The activity level of the OpenNLP project has fluctuated over that past 10+ years, with a large uptick in the last two years especially. Most recently, due both to the availability of new documentation and the release of version 1.5 , there have been many more downloads and page views for the OpenNLP project. In fact, September 2010 had the most downloads (1,561) and project web hits (226,391) of any month since the project's beginning in 2000, and October is keeping pacing with that figure so far. As a result, OpenNLP has gone from being in the 2000th to 4000th ranked project (between January and May, 2010) to being ranked 570, 314, 181 and 439 for July, August, September, and October respectively. Full details are available on the Sourceforge statistics page for OpenNLP. (There are 240,000 projects hosted on SourceForge, though this figure includes many, many projects that never actually get started: it seems that about 7-10% of these are stable, active projects based on a review done in 2007.)

Rationale

OpenNLP fills a significant gap at the ASF in regards to human language processing tools. While Lucene/Solr, UIMA and Mahout all have some tools in this area, none of them are solely focused on tools specifically for working with natural language like OpenNLP.

Initial Goals

The initial goals of the proposed project are:

- Bring the community together at the ASF and make the development process transparent for them
- Write user documentation about all major components
- Automated build including train and evaluate regression tests
- Produce an Incubating release

Current Status

Meritocracy

Some of the initial committers are familiar with Apache's idea of meritocracy, others aren't. We will get everybody on the same level as part of the incubation process.

Community

OpenNLP already has a considerable user base, both in industry and academia.

Core Developers

See the initial committer list.

Alignment

OpenNLP has tie-ins with several existing Apache projects. We have been distributing wrappers for UIMA for some time now (two UIMA committers also contribute to OpenNLP). We expect this collaboration to strengthen further after our move to Apache.

Another obvious connection exists to some of the projects under the Lucene umbrella. On the one hand, projects like Solr may benefit from the OpenNLP analysis capabilities to create specialized search for particular domains. On the other, OpenNLP may benefit from the machine learning code that is being developed in Mahout, and maybe get some people from that community to lend a hand.

Known Risks

Orphaned products

The project has been around for quite a number of years already, it has a well-established user community and a diverse set of committers.

Inexperience with Open Source

OpenNLP has been an open source project for quite some time. Many of the developers are already familiar with both open source in general and the ASF in particular.

Homogenous Developers

The current group of developers is very diverse, no two developers work for the same organization.

Reliance on Salaried Developers

Most of the developers are not paid to work on OpenNLP, so there is little reliance on salaried developers.

Relationships with Other Apache Products

NLP is often used in search and other algorithms that work with unstructured data, thus OpenNLP is likely to be useful to the Lucene and Solr communities. It also aligns nicely with both Mahout and UIMA.

A Excessive Fascination with the Apache Brand

We think the project aligns nicely with the goals of the ASF to disseminate source code to the public free of charge. NLP has long been the subject of cutting edge research, but is often lacking in community and shared knowledge. We believe that by bringing OpenNLP to the ASF, the Apache brand will help deliver NLP capabilities to a much larger audience and likewise a cutting edge project like OpenNLP can further the ASF brand by providing users with tried and true, as well as new, natural language processing capabilities.

Documentation

- http://opennlp.sourceforge.net/README.html
- http://sourceforge.net/apps/mediawiki/opennlp/index.php?title=Main_Page

Initial Source

The source code is maintained in two CVS repositories on SourceForge.

OpenNLP Maxent: http://maxent.cvs.sourceforge.net/viewvc/maxent/

OpenNLP Tools and OpenNLP UIMA: http://opennlp.cvs.sourceforge.net/viewvc/opennlp/

Source and Intellectual Property Submission Plan

The OpenNLP source code is already open source under the AL 2.0.

External Dependencies

Library	License			Description
JWNL	BSD			Java Wordnet Library
JUnit	CPL			Unit Testing Framework
UIMA	AL 2.0			Unstructured Information Management Architecture

Cryptography

OpenNLP neither provides nor uses any cryptography.

Required Resources

Mailing lists

- opennlp-dev
- opennlp-private
- opennlp-user
- opennlp-commits

Subversion Directory

https://svn.apache.org/repos/asf/incubator/opennlp

Issue Tracking

Jira: OPENNLP

Other Resources

Initial Committers

Name	Email			CLA
Thilo Goetz	twgoetz@apache.org			yes
Grant Ingersoll	gsingers@apache.org			yes
Jörn Kottmann	joern@apache.org			yes
Thomas Morton	tsmorton@gmail.com			no
William Silva	william.colen@gmail.com			yes
Jason Baldridge	jasonbaldridge@gmail.com			yes
James Kosin	james.kosin@gmail.com			yes

Affiliations

Name	Affiliation	
Thilo Goetz	IBM	
Grant Ingersoll	Lucid Imagination	
Jörn Kottmann	Infopaq International A/S	
Thomas Morton	Comcast Corporation	
William Silva	São Paulo University	
Jason Baldridge	The University of Texas at Austin	
James Kosin	International Communications Group, Inc.	

Sponsors

Champion

Grant Ingersoll

Nominated Mentors

Isabel Drost

Grant Ingersoll

Benson Margulies

Sponsoring Entity

The Apache Incubator