perldoop

Perldoop is a new tool developed by researchers from University of Santiago de Compostela (Spain) as part of the project "High Performance Computing for Natural Language Processing -- HPCNLP". This tool automatically translates Hadoop-ready Perl scripts into its Java counterparts, which can be directly executed on a Hadoop cluster while improving their performance significantly.

Hadoop provides an utility to execute applications written in languages different from Java, known as Hadoop Streaming. To use this tool the only requirement is that applications should read from stdin and write to stdout. Even though Hadoop Streaming is a very useful tool, important degradations in the performance were detected using Hadoop Streaming with respect to Hadoop Java codes. Only for computational intensive jobs whose input/output size is small, the performance of Hadoop Streaming is sometimes better because of using a more efficient programming language.

Therefore, the best choice in terms of performance is to develop Hadoop applications using Java. Nevertheless, translating Perl codes into Java is a long and tedious task, especially when the applications consist of many regular expressions. For this reason, we developed the Perldoop tool, which allows to automate the translation process increasing the performance and efficiency as well as the productivity.

The general case of automatically translating an arbitrary Perl code into its Java equivalent is a very hard problem, due to the characteristics of both languages. Note that our objective in this work was not to develop a powerful tool that allows to automatically translate any existent Perl code to Java, but a simple and easy-to-use tool that takes as input Perl codes written for Hadoop Streaming, and produces Hadoop-ready Java codes.

If you use Perldoop, please cite this article:

J. M. Abuin, J. C. Pichel, T. F. Pena, P. Gamallo and M. Garcia. "Perldoop: Efficient Execution of Perl Scripts on Hadoop Clusters", IEEE International Conference on Big Data, pp. 766-771, 2014.

How to use

The main file of the tool is Perldoop.py, located in *src/*, which has two input parameters:

- The route to the Perl script to translate. The Perl code should be labeled and programmed following the rules detailed in the User Manual (located in the *doc/* folder).
- The route to the Java template where the translated code will be inserted.

Therefore, the correct syntax to execute Perldoop is:

```
python Perldoop.py [perl-file] [java-template-file] > [output-java-file]
```

Repository Contents

After downloading the Perldoop source code, the user will find three directories containing:

- *src/* Source code of Perldoop, implemented using Python.
- *examples/* Simple examples to illustrate the use of Perldoop. In the current version it includes HelloWorld and WordCount applications written in Perl.
- *applications/* More complex Perl applications. The current version includes three natural language processing modules written in Perl. In particular, the modules process plain text to perform the following tasks: Named Entity Recognition (NER), Part-of-Speech Tagging and Named Entity Classification (NEC). All the modules process text in Spanish language.

All the examples and applications have been tested using Hadoop 2.2.0.

INFORMACIÓN

Investigadores
José Manuel Abuín Mosquera
Juan Carlos Pichel Campos
Tomás Fernández Pena
Pablo Gamallo Otero
Marcos Garcia González

DESCARGAR

Repositorio Gitlab

Descargar de Gitlab

Repositorio Github

PUBLICACIÓNS

Perldoop: Efficient Execution of Perl Scripts on Hadoop Clusters
IEEE International Conference on Big Data, 2014

PROXECTOS DE INVESTIGACIÓN

HPCNLP: Computación de Altas Prestacións para o Procesamento da Linguaxe Natural