Distributed Machine Learning for Web Crawling

Master thesis work for two students - 30 hec

Objective

The objective of this thesis is to develop a system capable of detecting and learning from patterns in how product information is presented on the web in a distributed fashion.

Background

Product information, such as name, category, brand and price, is used extensively by a variety of services on the web. Still, this information is difficult to get unless there is structured data feeds prepared with public access. This is rare and creates problems for services which makes use of product information, such as price comparison sites and bid optimization software for online advertising.

adomi.co is solving this problem by using innovative web crawling technologies. The current challenges include making use of large data sets created and associated manual input.

Tasks

• Identify and validate methods and systems that can be used for this purpose
• Implement these to be used with the web crawler, using MapReduce
• Develop test methodology to be used for testing the methods and systems on real data
• Document the results in a written report in English

Project Structure

The project can be done from anywhere, but we would prefer you to be based in Göteborg, Sweden.

The thesis is suitable for two students and consists of 30 hec credits.

Required Knowledge

• Fluent in programming, preferably Python
• Strong passion for algorithms and statistics
• Knowledge in machine learning techniques
• Experience with distributed systems and scalability problems is a great plus

How To Apply

Send your CV to alan@adomi.co

Feel free to include your GitHub, papers and projects. Anything you're proud of!

About adomi.co

The adomi.co team consists of ex-Googlers, software engineers and e-commerce professionals. We are driven by the desire to change how online advertising industry works. If you are looking to join a traditional, big and boring company, look elsewhere. We are dedicated to staying lean, small and fun!

We are a startup based in Göteborg, Sweden.

To learn more, visit www.adomi.co