Web Crawling

Scrapy “bien documuentado”

Ionut-Gabriel Radu ha desarrollado su propio crawler para la identificación de palabras en Rumano llamado RWScraper, has been designed around the open-source Scrapy (www.scrapy.org) framework, which can be used to build custom web crawlers. Scrapy provides the basic infrastructure necessary to extract web content based on several user-defined rules.

Mikhail Galkin ha realizado un trabajo acerca de identificación de contenido web en tablas, utilizando ScraperWiki is a powerful tool based on subscription model that is suitable for software engineers and data scientists whose work is connected with processing of large amounts of data. Además afirma Scrapy5 is a fast high-level framework written in Python for web-scraping and data extraction. Scrapy is spread under BSD license and available on Windows, Linux, MacOS and BSD. Merging performance, speed, extensibility and simplicity Scrapy is a popular solution in the industry. A lot of services are based on Scrapy, such as ScraperWiki or PriceWiki.

Almacenamiento

ha utilizado MongoDB para desarrollar a platform for large-scale design mining comprising a repository of over 100,000 Web pages and 100 million design elements. **Webzeitgeist.**

[Floratou et al.] comparan MongoDB y Have frente a otras bases SQL realizando diferentes benchmarks. Han demostrado que mientras las bases de datos relacionales pueden funcionar mejor, los sistemas NoSQL tienen ventajas en la usabilidad como módelos de datos flexibles, auto-sharding, tolerancia a fallos y balanceo de carga. Dichas ventajas han sido evaluadas y ratificadas por [Elfin Dede], provando MongoDB junto con Hadoop, la implementación más popular de MapReduce.

Elif Dede ha realizado una evaluación de rendimiento, escalabilidad y tolerancia a fallos de MongoDB con Hadoop, the most popular open source implementation of MapReduce,

Análisis y clustering

*El de Machine Learning (aprendizaje automático) es* una rama de la [inteligencia artificial](http://es.wikipedia.org/wiki/Inteligencia_artificial)cuyo objetivo es desarrollar técnicas que permitan a las computadoras *aprender*. De forma más concreta, se trata de crear programas capaces de generalizar comportamientos a partir de una información no estructurada suministrada en forma de ejemplos. Es, por lo tanto, un proceso de inducción del conocimiento. *Scikit-learn es una libreria de Machine Learning* *desarrollada en Python.[*Lars Buitinck] ha contruido un ejemplo de API (Aplication Programming Interface) con *scikit-learn, en la que implementa todas sus características. Establece que Scikit-learn es* *cada vez más popular, está diseñado para ser simple y eficiente, accesible a los no expertos, y reutilizable en distintos contextos. Por otra parte, [Fabian Pedregosa] conlcuye que scikit-learn expone una amplia variedad de algoritmos de aprendizaje automático, tanto supervisados como no supervisados, utilizando una interfaz consistente y orientado a las tareas, lo que permite una fácil comparación de métodos para una aplicación determinada.*