

Budapesti Műszaki és Gazdaságtudományi Egyetem

Villamosmérnöki és Informatikai Kar

Automatizálási és Alkalmazott Informatikai Tanszék

Deák Zsolt

automatizált adatgyűjtés

HTML adathalászat Powershellel

Konzulens

TBD

BUDAPEST, 2016

Összefoglaló

Adva van egy erőteljes eszköz, a PowerShell, aminek első számú célja a rendszeradminisztráció automatizálása, megkönnyítése. Szintén adott a probléma, hogy az Interneten strukturáltalanul jelen levő információhoz hozzáférhessünk. Ebben a munkámban egy konkrét esettanulmányon keresztül bemutatom, hogy hála a PowerShell sokszínűségének egy alapfeladatától távol eső területen is hatékony megoldást lehet vele készíteni.

A példában a Használtautó.hu autóhirdetéseinek adatait nyerem ki és dolgozom fel. A feldolgozás célja, hogy a site egy hiányosságát, az összehasonlító funkciót pótolja. Eredetileg egy az árukereső.hu azonos lehetőségéhez hasonló, a termékeket adataikkal együtt egymás mellett oszlopokban, táblázat formájában megjelenítő összehasonlítás volt a cél. Ezt kiegészítettem egy rangsorral, amit az autók tulajdonságaiból számított érték alapján állítok fel. Mivel különböző korú és állapotú járművek összehasonlítása lineáris módszerekkel, néhány tulajdonság kiválasztásával még megközelítőleg sem ad valós képet, így ennek alapját egy általam kidolgozott egyszerű (és determinisztikus) algoritmus adja. A nagyjából azonos korú és értékű autók összehasonlítása láthatóan értékes információval is szolgálhat ennek segítségével. A példa teljessége érdekében létrehoztam egy egyszerű weblapot is, hogy online elérhető legyen a szolgáltatás. Ez utóbbi nem PowerShell nyelven van írva, hanem egy szokványos PHP és JavaScript alapú website.

Abstract

There is a strong tool for system administration given, the PowerShell. There is the problem to be able to access unstructured data and information laying around on the Internet. In this work I made a proof of concept to prove that PowerShell can be used efficiently for tasks very much different from its original usage, thanks to the variety of features it has.

The proof of concept is about processing data of Használtautó.hu’s car pages. This processing is focused on the car comparison functionality that is not present on the site. The idea is coming from árukereső.hu’s similar functionality, a table based, side by side comparator of products’ details. I improved this idea by ranking the cars based on their main parameters. Due to the ineffectiveness of linear methodologies in comparing cars of varied ages and conditions by a handful of data, I needed to develop my own simple (and deterministic) algorithm. This gives the base of the car ranking that can produce valuable information about cars of similar ages and prices. For the sake of completeness, I created a webpage for the service to be available online. This is a user friendly abstraction in place of the PowerShell command line, written in PHP and JavaScript.