Spam vadászat a Twitter adattengerében

Szerző: **SZABADOS Noémi**, III. évfolyam, nomisabi@gmail.com

Témavezető: **LENDÁK Imre** egyetemi tanár

Intézmény: Újvidéki Egyetem, Műszaki Tudományok Kara, Szoftvermérnök és információs tehnológiák, Újvidék

A dolgozat célja bemutatni egy web-alapú rendszert amely képes kiszűrni a Twitter szociális hálózat rendszerében tevékenykedő, ún. spam felhasználókat, akik a nem kívánt (spam) email-hez hasonló módon a Twitter felhasználóit elhalmozzák nem kért üzenetekkel. A dolgozat elemzi azokat a módszereket, amelyekel a spam felhasználók megtalálják “áldozatukat”, illetve bemutat egy lehetséges megoldást ezek a felhasználók azonosítására és tevékenységük akadályozására.

A kutatás során a jelentős kihívást jelentett, hogy az analízist a Twitter rendszeréből kinyert nagyméretű adathalmazon kellett lefuttatnunk. Az adatokat *big data* eszközök segítségével elemeztük, elsősorban a Hadoop keretrenszer és a Spark adatfolyam kezelő keretrendszer segítségével.

A kutatás kulcseredménye egy Python programozási nyelven és a Django keretrendszer alkalmazásával fejlesztett webalkalmazás, amely bejelentkezés után analizálja a belépett felhasználó adatait és tartalomelemzés segítségével kiszűri, majd megjeleníti azokat a Twitter felhasználókat, amelyek (valószínűleg) spam jellegű adatokat küldtek neki. A rendszer lehetővé teszi a felhasználónak, hogy döntsön arról, hogy az azonosított spam-gyanús profilokat feljelenti-e a Twitter a rendszerében, vagy teljességükben blokkolja őket.

Kulcsszavak: Twitter, spam, adattudomány, big data, Hadoop