Selvin Fehrić, dipl. ing. el. - Završni magistarski rad

Fakultet/Akademija	FAKULTET ELEKTROTEHNIKE
Tip Rada	Završni magistarski rad
Kandidat, zvanje	Selvin Fehrić, dipl. ing. el.
Naziv Teme	Sistemi za prikupljanje, obradu, ekstrakciju i klasifikaciju web sadržaja
Rezime/Abstract	Gomilanjem ogromnih količina podataka na webu javlja se potreba za njihovim strukturiranjem i klasifikacijom kako bi se ti podaci mogli iskoristiti za različite namjene. Najveće svjetske kompanije koriste razne vrste sistema za obradu ovih podataka i koriste ih za različite potrebe kao što su: kategorizacija podataka, sistemi za preporuke, indeksiranje i pretraga podataka, automatski alati za zaštitu i slično. U ovom radu je istražena i obrazložena aktuelna problematika izrade cjelokupnog sistema za prikupljanje obradu, ekstrakciju i klasifikaciju web sadržaja. Rad uključuje: teoretski dio i praktični dio - u kojem je dizajniran i implementiran jedan funkcionalan alat lako prilagodljiv za različite vrste problema vezane za iskorištavanje postojećeg web sadržaja, kao i iskustva sa različitim open source alatima. Magistarski rad daje ideje za realizaciju jednog ovakvog cjelokupnog sistema na osnovu iskustva pri implementaciji istog. Na početku rada su obrazloženi osnovni principe i osnovne arhitekture sistema za prikupljanje sadržaja sa weba, njihove osnovne osobine i različite implementacije. Zatim su opisani različiti načini skladištenja tih ogromnih količina podataka, njihova obrada, ekstrakcija bitnih informacija kao i finalna klasifikacija tih podataka. Za sve ove module navedene su najvažnije osobine, nekoliko implementacija kao i njihove prednosti i mane. Posebno su predstavljeni različiti načini ekstrakcije i klasifikacije sadržaja. Dizajnirana je i implementirana cjelokupna arhitektura jednog ovakvog kompletnog sistema za strukturiranje i klasfikaciju podataka koristeći različite open source komponente poput: - Apache HBase - distribuirana i skalabilna NoSQL baza podataka radi bržeg čitanja/snimanja podataka sa Hadoop fajl sistemom kao podlogom, - Apache ActiveMQ - sistem razmjene poruka čime se sistemu obezbjeđuje distributivnost i skalabilnost, - Boilerpipe - biblioteka za ekstrakciju sadržaja iz HTML-a, - WEKA - alati sa podrškom za klasifikaciju. Na osnovu iskustva pri implementaciji ovog sistema i integraciji postojećih alata predloženo je optimalno rješenje i navedene su prednosti i nedostaci različitih arhitektura te načini implementacije istih. Ključne riječi: web sadržaj, prikupljanje podataka, mašinsko učenje, ekstrakcija teksta, klasifikacija teksta, distributivnost, open-source alati.
Datum	02.07.2015
Predsjednik	Dr sc. Edin Pjanić, docent – Uža naučna oblast Računarstvo i informatika Fakultet elektrotehnike Univerziteta u Tuzli
Mentor	Dr sc. Amer Hasanović, vanredni profesor – Uža naučna oblast Računarstvo i informatika Fakultet elektrotehnike Univerziteta u Tuzli
Član komisije	Dr sc. Emir Mešković, docent – Uža naučna oblast Računarstvo i informatika Fakultet elektrotehnike Univerziteta u Tuzli
Član komisije	-
Član komisije	-
Zamjenski član	-
Dodatni detalji i lokacija	Dana 02.07. 2015. godine u 14,oo sati na Fakultetu elektrotehnike Univerziteta u Tuzli
Zavrsne Odredbe	Pristup javnosti je slobodan. Rad se može pogledati u Sekretarijatu fakulteta radnim danom od 08 do 14 sati.