Teme za upravljanje digitalnim dokumentima

Dragan Ivanović
dragan.ivanovic@uns.ac.rs

Katedra za informatiku, Fakultet tehničkih nauka, Novi Sad

2015.

Hibernate i Lucene

- Implementacija repozitorijuma e-knjiga
- Indeksiranje se radi putem Hibernate anotacija
- Pretraživanje se radi pomoću Hibernate Search biblioteke
- Veb ili standalone aplikacija
- Uzeto i odbranjeno!

Web search engine

- Periodično indeksiranje sadržaja odgovarajućeg Internet domena (npr. uns.ac.rs) preuzetog odgovarajućim crawler-om
- Implementacija veb aplikacije na kojoj se
  - zadaju upiti
  - dobijaju odgovori
- Za crawler se mogu koristiti gotova rešenja (npr. Apache Nutch)
- Uzeto i odbranjeno!
- *deploy*-ovana aplikacija: link
- *source code* - link

Search engine optimization

- Kreiranje veb sajta nekog softverskog proizvoda - proizvolja arhitektura i programski jezik
- Veb sajt treba da je optimizovan za pronalaženje sa Veb search engine-a (Google, Yahoo!) upotrebom ključnih reči koji pripadaju domenu softverskog proizvoda
- Analiza legalnih i spam tehnika
- Želja da se softverski proizvod lako pronalazi i da ga ljudi kupuju i koriste
- Uzeto i odbranjeno!
- veb sajt DOSIRD UNS - link
- veb sajt OpenDLT - link

Izrada srpskog stemmera

- Snowball - Programski jezik za rad sa stringovima, paterni određuju dalji tok izvršavanja
- Kreiranje i implementacija pravila za steming srpskih reči upotrebom Snowball-a
- Dobijeni stemmer je potrebno transformisati u Java kod (postoji gotov alat za ovu namenu)
- Dobijeni stemmer je potrebno i verifikovati i prikazati indikatore performansi
- Unakrsna validacija, postoji anotirani data set

Multi-lingual search

- Implementacija pretprocesora (analizatora) teksta koji omogućuje bilingualno pretraživanje repozitorijuma tekstualnih digitalnih dokumenata pisanih na Srpskom i Hrvatskom jeziku
- Ovo bilingualno pretraživanje je značajno jer postoji preko 10 miliona ljudi koji razume oba ova jezika
- Potrebno je kreirati jedinstveni stemer za oba jezika koji imaju slična morfološka pravila (postoji neka verzija koju je potrebno doraditi)
- Zatim koristiti rečnik koji sadrži reči koje su različite u ova dva jezika (postoji, ali i ovo treba doraditi)
- Analizator implementirati kao proširenje Lucene IR biblioteke
- Izvršiti verifikaciju kreiranog analizatora i prikazati rezultate.
- Unakrsna validacija, postoji anotirani data set

Long-term preservation - File Fixity

- Program za backup digitalnih dokumenata
- Inicijalno se vrši backup, odnosno dupliranje sadržaja i postavljanje u odgovarajući backup folder, kao i računanje fingerprint-a (jedinstvene niz bitova koji predstavlja dokument) nekim fingerprinting algoritmom
- Fingerprint se čuva do sledeće sinhronizacije radne verzije i backup-a
- Sinhronizacija je periodična
- Ako je radna verzija digitalnog dokumenta promenjena određenim programom (legalna izmena), računa se nova fingerprint vrednost i vrši backup
- Ako radna verzija nije menjana a došlo je vreme sinhronizacije, računa se fingerprint i ako je isti kao prethodna njegova vrednost sve je u redu, a ako nije ista onda je fajl oštećen
  - vršiti odgovarajuću notifikaciju
  - iz backup-a vratiti dokument

Long-term preservation - LOCKSS

- Implementirati long-term preservation (prethodni slajd) upotrebom LOCKSS - link

MongoDB - text search

- Implementirati digitalni repozitorijum u MEAN stek tehnologijama
- Indeksiranje i pretragu implementirati pomoću MongoDB text search - link

Lucene analyzer tester

- Razvoj aplikacije za testiranje Lucene Analyzer-a
- Kreiranje indeksa
  - indeksiranje sadržaja čiji su metapodaci u XML zapisima
  - indeksiranje digitalnih dokumenata koji se nalaze u odgovarajućoj folderskoj strukturi
  - importovanje već pripremljenih indeksa
- Postavljanje upita
  - importovati već pripremljene upite
  - korisnicima sistema dozvoliti da unose upite
- Rad sa rezultatima
  - korisnik sistema u listi odgovora označava relevantne i nerelevantne rezultate
  - sistem pamti odgovore korisnika
  - sistem računa preciznost, povrat, f-meru, kappa slaganje, itd.
- Uzeto! Ima prostora i za dalja istraživanja
- *deploy*-ovana aplikacija - link; nalog- proba, proba
- *source code* - link

Pretraga georeferenciranih podataka

- Automatsko georeferenciranje teze bazirano na NLP tehnikama
- Machine learning techniques
- Slovenian natural language processing tools
- Slovenian geospatial catalogs
- KEYSTONE COST action
- GEOPOLO

Personalizacija pretrage teza i disertacija

- Pored reči u upitu, sistem treba da koristi istoriju korisnikovih upita, lične informacije o korisniku, lokaciju korisnika, koji je uređaj korišćen za pretragu, itd.
- Context information in the query
- Digitalna biblioteka doktorskih disertacija u Novom Sadu i njeni korisnici su izvor podataka za istraživanje
- 4.000 disertacija, nekoliko hiljada upita mesečno, preko 400 registrovanih korisnika
- Kipar je saradnik u ovom istraživanju
- KEYSTONE COST action

Automatic recommendation system

- Sistem treba da koristi istoriju korisnikovih preuzimanja i na osnovu toga da sugeriše korisniku sadržaje koji mu mogu biti interesantni
- Digitalna biblioteka doktorskih disertacija u Novom Sadu i njeni korisnici su izvor podataka za istraživanje
- 4.000 disertacija, preko 10 hiljada preuzimanja sadržaja mesečno (zapisano u logovima)
- Malta je saradnik u ovom istraživanju
- KEYSTONE COST action

Ostale teme

- Teme unutar DOSIRD UNS projekta - link
- OAI-PMH i Dublin core
- MG4J search engine
- JPlag, MOSS, Sherlock source code similarity detection tools
- Apache Tika content analysis toolkit
- Vaše ideje