**Code Similarity Calculator**

Mail studentima:

Dragi kolege, pozdrav

Najprije vam želim sve najbolje u Novoj 2008. godini, puno uspjeha, sreće, zdravlja i ljubavi (s obzirom na vaše godine, možda sam i invertirao poredak po važnosti ;-).

Zadnjih mjesec i po, dva dana smo imali „kreativnu stanku“ u radu na našem CSC (Code Similarity Calculator – ima li netko prijedlog za bolje ime ?) projektu. Mene je rasturio krah laptopa (3 tjedna za promjenu matične ploče), pa su naletili praznici, i onda za dobru mjeru tjedan dana gadne gripe.

Ali, to je prošlost a treba vidjeti što će biti s budućnošću.

A što se tiče budučnosti, mislim da je stvar sazrela da se malo konkretizira. Svaki projekt mora imati cilj, a s moje strane cilj za CSC se iskristalizirao i ukratko se može opisati kao „znanstvena produkcija“. Para s tim zaraditi nećemo (iako, nikad ne reci nikad, možda i fakultet zagrize ako napravimo dobar sustav ;-), a istraživačka samo-motivacija („ajmo malo vidjeti što se tu može“) rezultira nedovoljnom posvećenošću.

A „znanstvena produkcija“ podrazumijeva pisanje članaka. Na osnovu dosada napravljenog i raspoloživih podataka, imam ideju za pisanje jednog članka s naslovom otprilike „Code Plagiarism in Computer Science Classes at Croatian University“. Osnovna ideja je analizirati podatke koje imamo za prethodne tri godine (odnosno imati ćemo ih pred ljeto) i vidjeti koliko je bilo plagiranja koda i kakvi se zaključci mogu izvući. S obzirom da za svakog studenta imamo i njegov rang pri upisu, te prosječan rang na godini, otvaraju se mogućnosti za proračunavanje zanimljivih korelacija.

Uz malo „teorijske“ spike o Bolonjskom procesu, reformi kurikuluma i načina držanja nastave, to može ispasti jedan sasvim ozbiljan i fin članak. Možda čak i CC kalibra !!! CC članak je članak objavljen u časopisu iz Current Content liste a o „rangu“ tih časopisa dovoljno govori činjenica da je (su)autorstvo na 1 (JEDNOM) CC članku glavni i najteži uvjet za dobivanje docentskog mjesta (sad vam je jasan moj ultimativni motiv ;-)

Za vas kao studente, važnost pisanja članaka je značajno manja, osim u slučaju da pucate na eventualnu znanstvenu karijeru na FERu :-). U tom slučaju suautorstvo na članku (a pogotovo na CC članku !) značajno povećava šanse za dobivanje mjesta znanstvenog novaka (iliti, jackpota ;-).

Autora na članku može biti 3 (tri) da bi se svakome od njih doprinos brojao kao cijeli što je donekle ograničavajuće s obzirom na činjenicu da bi svakako trebalo uključiti kolegicu koja nam je dala podatke (a koju ćemo sigurno još puno trebati). Što znači da je otvoreno jedno mjesto :-).

A što bi trebalo raditi ?

Trebalo bi isprogramirati aplikaciju koja će na osnovu podataka iz baze, i korištenjem metrika koje već imamo implementirane (a koje su sasvim dovoljne za ovo istraživanje !) izračunati potrebne statistike i usput imati i određene vizualizacijske mogućnosti.

Naravno, najprije bi trebalo dobro promisliti i vidjeit što i kako računati – znanstvenim rječnikom, trebalo bi definirati hipoteze :-) a onda, kao što rekoh, implementirati sustav koji će sve to izračunati.

Baza: SQL Server 2005, okolina .NET Framework 3.5 i Visual Studio 2008, tehnologija pristupa podacima LINQ ! Ovo je jedan lijepi bonus :-) – upoznavanje sa state-of-the-art Microsoftovom tehnologijom !

S obzirom da ste vas tri mušketira, ova ponuda može izgledati isključiva, ali to je samo privid. Jer različitih tema za istraživanje ima koliko ti srce želi. Smišljanje različitih algoritama i uspoređivanje njihovih „performansi“ na različitim podacima (uz usporedbu s već postojećim sustavima (MOS) od kojih neki imaju i javno dostupna sučelja).

Razmišljali smo i o Abstract Syntax Trees kao mogućem smjeru istraživanja ... ? Pa smo malo stali :-)

Možda bi se moglo razmisliti i o „dizanju“ cijele stvari nivo više – iznad usporedbe file-to-file na system-to-system gdje bi se istraživala sličnost dizajna, povezanosti među objektima, ugovora (javnih sučelja) koje objekti/razredi pružaju, ...

Ja, eto, u sljedećem zimskom semestru imam predmet „Objektni dizajn“ na M.Sc. dijelu studija, i kako će biti prilično teško smisliti 50-100 specifičnih zadataka, sustav koji bi prepoznavao plagijarizam/sličnost na razini 10-15 klasa bi bio isto zanimljiv.

A sad mi pade i Zatemas napamet – naime, bolno sam svjestan koliko je rješenje za automatsko ocjenjivanje ugrađeno u AhyCo sustav loše (to je u biti napravio jedan student za diplomski rad – al nemojte to nikome reći ;-) pa ukoliko je Zatemas značajno bolji, i to je jedno potencijalno područje za „znastvenu produkciju“ (recimo, članak - „ZATEMAS - System for automatic evaluation of home-work assignements in computer science classes“)

Dakle, tema i prostora za istraživanje ima, a sad je pitanje ima li volje (i vremena ;-) ?

Gospodo, izjasnite se ...

Pozdrav

Zvone

* Daleko od toga da ja namjeravam primjeniti hands-off pristup, ali bi moja uloga primarno bila „voditeljska“ (definiranje plana, analiza i određivanje statističkih kriterija, dizajn aplikacije ...) a u konačnici i taj članak će trebati napisati
* S obzirom na složenost problema, moja jeprocjena da se aplikacija može izraditi za tri mjeseca, uz još tri mjeseca za rafiniranje podataka/tehnika/metrika i pisanje članka – deadline – godišnji odmor 2008 :-)
* Naravno, ja bih sve to mogao i sam, ali bi trajalo dulje a i kod samostalnog rada uvijek je tu problem motivacije :-(
* Poboljšanje i smišljanje novih algoritama također – pogotovo me još uvijek pinga mogućnost s AST stablima i uspoređivanjem progrmskog koda na većoj skali, kad u igru uaze i metrike dizajna i arhitekture – također jedno područje