Borko Kovačević
Univerzitet u Beogradu
Filološki fakultet
Srbija
borko.kovacevic@fil.bg.ac.rs

Projekat *AVANTES*

Realizacija projekta *AVANTES – Advancing Novel Textual Similarity-based Solutions in Software Development* („Nova rešenja u razvoju softvera zasnovana na sličnosti tekstova“) otpočeta je u septembru 2020. godine. Projekat je deo Programa za razvoj projekata iz oblasti veštačke inteligencije i finansiran je od strane Fonda za nauku Republike Srbije. Partnerske institucije koje učestvuju na projektu su Elektrotehnički fakultet Univerziteta u Beogradu, Filološki fakultet Univerziteta u Beogradu i Inovacioni centar Elektrotehničkog fakulteta. Interdisciplinarni istraživački tim bavi se razvijanjem inteligentnog alata za prepoznavanje semantičke sličnosti između delova softverskog sistema ispisanih na programskim jezicima i komentara na prirodnim jezicima. Posebnu pažnju istraživači usmeravaju na rešavanje problema sličnosti između dva teksta različitih dužina, pre svega na srpskom, uz upoređivanje sa rezultatima dobijenim za engleski jezik.

Potreba za ovakvim projektom potiče iz prakse, odnosno bazira se na zahtevima tržišta rada. Obrada prirodnih jezika (*NLP – Natural Language Processing*) predstavlja jednu od najbrže rastućih oblasti veštačke inteligencije. Zbog svoje svakodnevne primene, ona postaje sve više prepoznata u društvu. Softverska industrija vezana za *NLP*, konkretno razvoj i upotreba softvera, od izuzetnog je značaja za srpsku i evropsku, ali i globalnu industriju i ekonomiju. Usled pomenutog značaja, neprestano raste broj ljudi zaposlenih u softverskoj industriji fokusiranoj na *NLP*, kao i količina kodova i obim softvera. Kao nusproizvod ovoga, sve je češća pojava ponovnog pisanja delova softvera koji su već primenjeni, odnosno korištenje tuđeg koda, podataka i modela, čineći tako softver sve komplikovanijim, to jeste težim za razumevanje, održavanje i testiranje.

U vezi sa tim, problem su i autorska prava i utvrđivanje da li su dva primerka softvera ista.

U kontekstu svega pomenutog, cilj projekta *AVANTES* je da pospešujući veštačku inteligenciju i *NLP* doprinese razvoju softvera, odnosno identifikovanju sličnih delova kodova, modela i baza podataka, sa namerom pojednostavljivanja procesa razumevanja, održavanja i testiranja softvera. Širi i sveobuhvatniji cilj projekta je i da odredi stepen u kojem semantika tekstualnog opisa reflektuje semantičke i strukturalne karakteristike softverskih objekata (kodova, baza podataka i modela), kao i da premosti raskorak između prirodnih i programskih jezika, u okviru kojih su pomenuti objekti definisani.

Naš istraživački tim razviće inteligentni alat za prepoznavanje semantičke sličnosti između delova softverskog sistema ispisanih na programskim jezicima i komentara na prirodnim jezicima. Posebnu pažnju istraživači će usmeriti na rešavanje problema sličnosti između dva teksta različitih dužina, pre svega na srpskom, oslanjajući se i na upoređivanje sa rezultatima dobijenim za engleski jezik. Takođe, realizovani sistem moći će da prepozna duplikate delova softvera. Za potrebe projekta koristiće se novi metodi za analizu programskog koda koji podrazumevaju upotrebu tehnika mašinskog učenja i veštačke inteligencije.

Pored alata za utvrđivanje sličnosti između delova softvera i unetih komentara, grupa softverskih inženjera i lingvista formiraće i novi algoritam za pretragu koda prema značenju, tačnije putem upita na prirodnom jeziku (srpskom i engleskom). Osim toga, razviće se i algoritam za prepoznavanje sličnosti tekstova različitih dužina, a kao rezultat dobiće se i skupovi podataka i modeli za automatsku obradu srpskog jezika.

Transdisciplinarni aspekat ovog projekta ogleda se u analizi jezičkih faktora koji kreiraju semantičku sličnost između jedinica različitih nivoa. Fenomen semantičke sličnosti meri se između jedinica različitog tipa – u pitanju su paragrafi, rečenice i fraze. Zadatak je da se odredi stepen u kojem je značenje veće jedinice obuhvaćeno manjom jedinicom. Posebna pažnja se posvećuje poređenju srpskih i engleskih primera, kao i poređenju između jezika komentara koda i opštijeg domena novinskih tekstova.

Projekat *AVANTES* od velikog je značaja za Srbiju – kako za naučnu zajednicu, tako i za srpsku (softversku) industriju i ekonomiju. Njegova glavna vrednost je to što će istraživači iz naizgled udaljenih i nepovezanih naučnih oblasti (softverskog inženjerstva i lingvi-

stike) formirati skupove anotiranih podataka i uvesti inovacije u postojeće tehnologije za obradu srpskog jezika. Ovo je posebno bitno imajući u vidu da je za srpski jezik trenutno dostupno daleko manje resursa nego za neke veće jezike, poput engleskog. To će olakšati rad softverskim inženjerima u našoj zemlji, ali i lingvistima koji se bave istraživanjem srpskog jezika.