1. **Polni naslov projekta: SEAN**

2. **V sodelovanju z: Fakulteta za informacijske študije v Novem mestu, Nevtron & Company podjetje za proizvodnjo, trženje, turizem in svetovanje, d.o.o.**

3. **Povzetek projekta:**

Svetovni splet nam iz dneva v dan ponuja nova znanja in nove možnosti. Vse od njegovega pojava smo priča skokoviti rasti objavljenih in dostopnih vsebin. Z izjemnim povečanjem obsega spletnih informacij raste tudi zanimanje za hitro in učinkovito integracijo novih informacij, med drugim tudi na področju preučevanja zbirk besedilnih virov. Glavni povod in razlog za izvedbo projekta tiči v potrebi po pridobivanju relevantnih informacij preko spleta, saj v praksi obstaja izredno zanimanje na različnih področjih, vse od trženja pa do varnosti, kako in na kakšen način zaznati občutke, mnenja in zadovoljstvo ljudi o določenih vsebinah, produktih in storitvah na spletu.

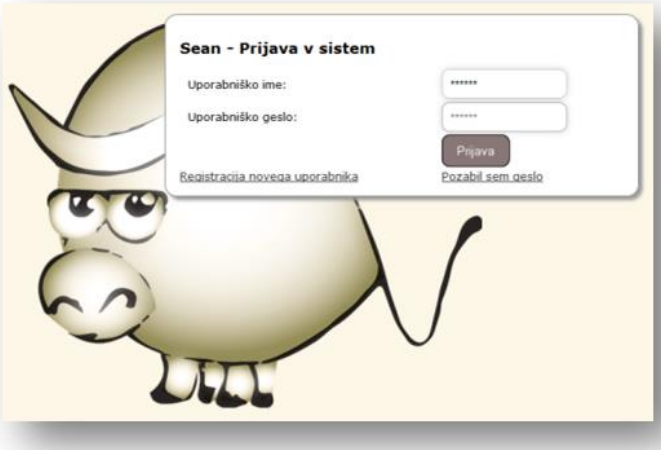
Cilj projekta SEAN je tako bil razviti prototip spletne aplikacije, ki bo omogočala zajem ter ocenjevanje (določitev sentimenta) več tisoč spletnih besedil z gospodarsko, ekonomsko in politično vsebino v slovenskem jeziku z različnih spletnih virov. Skupaj s študenti smo to tudi naredili in ustvarili obsežno bazo podatkov, v katero smo shranili podatke in meta podatke o vsebini zajetih spletnih besedil.

Študenti so se najprej seznanili s področjem analize sentimenta v besedilih ter spoznali algoritme za strojno učenje. Pregledali in analizirali so obstoječe prakse in tehnologije na področju analize sentimenta, preučili obstoječe tehnologije in njihovo praktično uporabo. Njihova naloga je bila, da pridobijo članke in novice iz petih znanih slovenskih spletnih portalov z gospodarsko vsebino (Rtvslo, 24ur, Dnevnik, Finance inŽurnal24), pridobijo vsebino člankov in novic ter njihove metapodatke (datum, avtor, ključne besede, url, …), jih shranijo v kompleksno bazo podatkov, izdelajo prototip spletne aplikacije, ki jim služi za označevanje pridobljenih spletnih besedil, pridobijo označen korpus besedil v slovenskem jeziku ter testirajo različne metode strojnega učenja za potrebe napovedne točnosti pri klasifikaciji spletnih dokumentov. Zato so študenti najprej preučili načine, kako učinkovito filtrirati, zajemati informacije in besedila s spleta z gospodarsko, ekonomsko in politično vsebino. Zaradi problemov, ker je vsak spletni medij drugačen (struktura HTML kode), so sprogramirali spletne pajke ločeno za vsak spletni medij ter pridobili približno 200.000 spletnih besedil. Zaradi velike količine podatkov so pridobili naključni vzorec spletnih besedil, približno 2.000 člankov in novic za vsak spletni medij (skupaj torej dobrih 10.000 spletnih besedil z metapodatki), ki so jih vnesli v bazo podatkov. Nato so razvili spletno aplikacijo, ki deluje na različnih spletnih brskalnikih. Študenti so določili sentiment spletnim besedilom, določili stopnje sentimenta spletnim besedilom (5 nivojsko ocenjevanje sentimenta, kjer 1 pomeni zelo negativno, 2 negativno, 3 nevtralno, 4 pozitivno in 5 zelo pozitivno) ter izvedli analizo sentimenta spletnih besedil. Vsak članek je bil označen s strani vsaj dveh označevalcev.

Projekt se izvaja v okviru Operativnega programa razvoja človeških virov za obdobje 2007–2013, 1. razvojne prioritete »Spodbujanje podjetništva in prilagodljivosti ter prednostne usmeritve« 1.3. »Štipendijske sheme«, v okviru potrjene operacije »Po kreativni poti do praktičnega znana«.

Rezultati projekta so tako analiza obstoječih praks na področju analize sentimenta (preučitev obstoječe tehnologije in njihove praktične uporabe), obsežna in kompleksna baza podatkov (in meta podatkov) o spletnih besedilih, korpus spletnih besedil, prototip spletne aplikacije, ki omogoča zajemanje spletnih besedil s spletnih virov, shranjevanje v bazo podatkov in podporo na različnih spletnih brskalnikih ter ocena zastopanosti (pozitivnega, negativnega, nevtralnega) sentimenta v izbranih slovenskih spletnih medijih, izbranih spletnih medijev ter predstavitev možnih vzrokov za te ocene.

Aplikacija SEAN – vstopna stran:



Projekt se izvaja v okviru Operativnega programa razvoja človeških virov za obdobje 2007–2013, 1. razvojne prioritete »Spodbujanje podjetništva in prilagodljivosti ter prednostne usmeritve« 1.3. »Štipendijske sheme«, v okviru potrjene operacije »Po kreativni poti do praktičnega znana«.