Povzetek projekta Po kreativni poti do znanja 2017 – 2020, 2. odpiranje, za namen objave in predstavitve na spletni strani sklada

1. Polni naslov projekta:

Razvoj slovenskih jezikovno-tehnoloških rešitev za uporabo v poslovnih informacijskih sistemih

- **V katero področje na prvi klasifikacijski ravni KLASIUS-P-16 se uvršča projekt glede na vsebinsko zasnovo** (neustrezno področje izbrišite)**:**

06 - Informacijske in komunikacijske tehnologije (IKT)

2. V sodelovanju z: (navede se univerza oz. samostojni visokošolski zavod, ki je prijavil projekt in članica, ki je nosilka projekta ter partner/ja – podjetje/ji oz. organizacija, ki je/sta bilo/i vključeno/i v projekt)
Univerza v Ljubljani, Fakulteta za elektrotehniko
_____
Medius d.o.o.
_____

3. Besedilo:

- Opredelite problem, ki se je razreševal tekom izvajanja projekta

Visokotehnološka podjetja dandanes s pridom uporabljajo jezikovne tehnologije v produkcijskih poslovnih okoljih. Jezikovne tehnologije so vgrajene v aplikacije, ki omogočajo obdelavo velike količine podatkov. BBC na primer uporablja sistem za samodejno označevanje ključnih besed svojih prispevkov in tako na lažji in bolj pregleden način ponudi opis prispevka bralcem, AirBnB pa na svoji spletni platformi omogoča cenzuro telefonskih številk in naslovov, če bi gostitelj najemniku poskušal razkriti svoje kontaktne podatke ali lokacijo. Tovrstna cenzura je ključen element njihovega poslovnega modela. Trenutno je na trgu več priznanih odprtokodnih ogrodij, ki omogočajo uporabo jezikovnih tehnologij v enem od svetovnih jezikov. Slovenščina je zaradi relativno majhnega števila govorcev in posledično manjše tržne relevantnosti pogosto zapostavljena.

Trenutno morajo podjetja vložiti precej truda, da pridobijo podatke in jih priredijo za uporabo v danem ogrodju. To je pogosto dolgotrajen proces, morebitnih slovenskih kupcev pa žal ni dovolj, da bi se podjetjem finančno splačalo razviti programske pakete, ki so lahko dovolj splošno uporabni za obdelavo in procesiranje slovenskih besedil. S projektom smo skušali prenesti znanje ne samo študentom, temveč tudi gospodarskemu partnerju, ki bo lahko tudi po končanem projektu vzdrževal odprtokodni programski arhiv in imel boljše izhodiščne možnosti za vključitev jezikovnih tehnologij v svoje lastne projekte.

- Opišite potek reševanja problema oz. kratek povzetek projekta

Znano je, da je na področju strojnega učenja ena zahtevnejših nalog priprava podatkov in pravilno vrednotenje pridobljenih modelov, manj zahtevna pa je uporaba obstoječih naprednih sodobnih algoritmov za strojno učenje. Za uspešno vrednotenje in enostavno proženje učnih algoritmov ter za pregledovanje obsežnih besedil in hitro označevanje je bil v projektu razvit prosto dostopni spletni vmesnik. Njegova uporaba je bila dokumentirana v uporabniški dokumentaciji. Ta vmesnik, ki so ga razvili sodelujoči študenti, je pripomogel k bolj učinkovitemu pregledovanju vhodnih podatkov in pridobljenih besedilnih rezultatov. Razviti vmesnik smo povezali z ostalimi komponentami – mikrostoritvami, ki kot celota predstavljajo celotno aplikacijo na tak način, da je le-ta primerna za uporabo tudi v poslovnih okoljih. Vse komponente smo objavili na prosto dostopnem repozitoriju odprte kode github.com. Z razvojem spletnega vmesnika in pripadajočih dodatnih zalednih sistemov smo tako omogočili bolj enostavno uporabo že razvitih ogrodij za procesiranje jezika ter omogočili lažjo pripravo in realizacijo dodatnih popravkov označb - učnih podatkov. S tem smo tudi zmanjšali čas urejanja in na uporabniku bolj prijazen način omogočili preizkušanje in vrednotenje različnih tehnik strojnega učenja na področju procesiranja naravnega jezika.

- Navedite in opišite rezultate projekta ter njihov doprinos k družbeni koristnosti

Sodelujoči v projektu so se spoznali z delom na področju razvoja poslovnih aplikacij v interdisciplinarni skupini. Ideja projekta je vključevala razvoj spletnega vmesnika za manipulacijo in upravljanje enega svetovno priznanih ogrodij (OpenNLP) za procesiranje naravnega jezika in njegovo priredbo za uporabo slovenskega jezika. Aplikacijo smo razvili kot prosto dostopno orodje in jo objavili na spletu v treh projektih:

1. https://github.com/MediusInc/slotex-nlp-core
2. https://github.com/MediusInc/slotex-nlp-web
3. https://github.com/MediusInc/slotex-nlp-entity

Dodatno smo objavili tudi krovni projekt, ki vsebuje navodila za uporabo in omogoča enostavno namestitev medsebojno odvisnih komponent:
https://github.com/MediusInc/slotex-nlp

Dodali smo tudi enostaven program za hitro manipulacijo modelov, ki omogoča direkten odriv velike količine podatkov v spletno aplikacijo:
https://github.com/MediusInc/slotex-nlp-pwc

Prosto dostopna programska oprema, objavljena na javno dostopni Github platformi za hranjenje odprtokodne programske opreme, omogoča, da k razvoju zasnovane aplikacije pristopijo tudi drugi

Dunajska cesta 20, 1000 Ljubljana t 01 434 10 81 e info@sklad-kadri.si www.srips-rs.si

razvijalci programske opreme in tem lahko razširimo možnosti dograditev. Prav tako objava na omenjeni platformi omogoča namestitev aplikacije komurkoli, ki jo želi preizkusiti ali uporabiti celo v produkcijskem okolju.

Poleg angleškega jezika je trenutno podprt tudi slovenski jezik z dodanimi modeli, inicialno naučenimi na šolski podatkovni zbirki SSJ500k. Prav tako smo si prizadevali, da bi bila namestitev dokaj zapletenega sistema na podlagi mikrostoritvene arhitekture enostavna – z uporabo krovnega projekta, ki vsebuje le dokumentacijo in skriptirano namestitev.

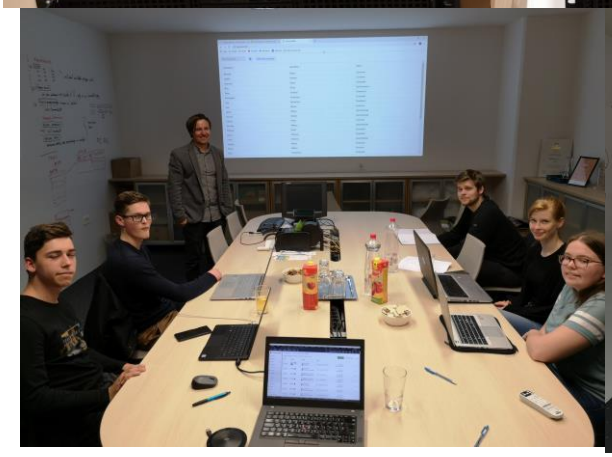
Za potrebe obveščanja javnosti smo na osnovi objavljene dokumentacije tudi izdelali posebno spletno stran pod domeno https://slotex.si, kjer predstavljamo projekt javnosti in spodbujamo k nadaljnjemu prostovoljnemu udejstvovanju pri razvoju novih idej, izboljšavi dokumentacije, izboljšavi aplikacije in algoritmov za procesiranja naravnega jezika ter k splošni uporabi aplikacije.

Največji doprinos k družbenem razvoju in napredku je sama odprtokodna aplikacija. Programska koda, ki je javno objavljena na platformi github.com nudi vsem interesentom vključitev v projekt in nadaljnji razvoj ali pa samo njeno uporabo.
S tako zasnovo skušamo tudi spodbuditi razvijalce programske opreme in raziskovalce jezikovnih tehnologij k ohranjanju slovenskega jezika v hitro razvijajoči se digitalni dobi. Prav tako pa ostale morebitne uporabnike navdušiti nad aplikacijo in jim pokazati, da lahko z uporabo obstoječih algoritmov in nekaj znanja o programiranju hitro naredimo robusten sistem za procesiranje naravnega jezika tudi za slovenski jezik.

4. Priloge:

- Slikovno gradivo: Priložite vsaj dve sliki npr. sliko končnega produkta, sliko študentov pri delu na projektu, sliko s sestankov ipd. Pri pošiljanju slik bodite pozorni, v kolikor gre za končni produkt, da bo zadoščeno zahtevam glede informiranja in obveščanja (ustrezni logotipi itd.).

Dunajska cesta 20, 1000 Ljubljana t 01 434 10 81 e info@sklad-kadri.si www.srips-rs.si



Dunajska cesta 20, 1000 Ljubljana t 01 434 10 81 e info@sklad-kadri.si www.srips-rs.si