Predlog projekata

Predlog projekta treba da bude kratak ali informativan (do 3 strane A4 formata). Predlog treba da bude jasan (jasan opis cilja projekta i opis plana realizacije tog cilja).

Projekat se radi u timovima u grupama od 2 ili 3 studenta.

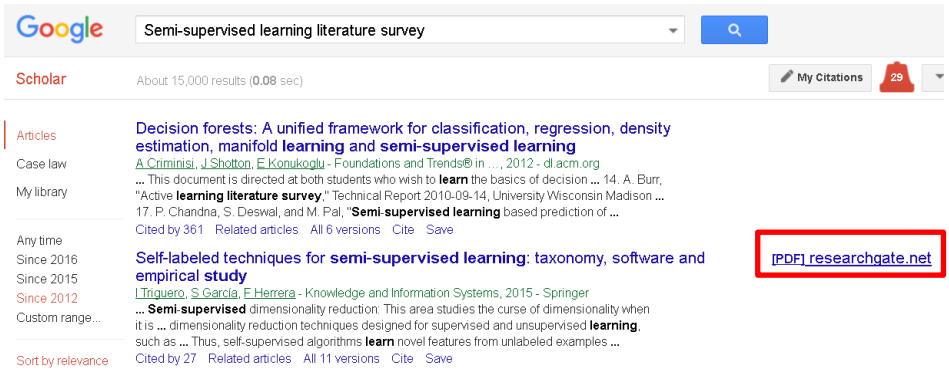
Za sva pravila i rokove vezane za predaju predloga projekta, pogledajte slajdove *Izrada projekta.pdf*. Za sve projekte koji nisu predani na vreme ili nisu ispoštovali neko od navedenih pravila, važi da će biti ocenjeni sa maksimalnom ocenom 6.

Potpun predlog projekta sadrži sve elemente koji su navedeni u poglavlju 2. Ako propustite da navedete nešto od ovih elemenata, vaš projekat automatski neće biti prihvaćen.

Da biste napisali predlog projekta potrebno je:

1 **Da ustanovite generalnu temu koja vas zanima, a zatim i da pronađete odgovarajući skup podataka**
U poglavlju 1 možete naći korisne smerinice za odabir podataka – prikazani su skupovi podataka koje možete da skinete i proučite, kao i tipične vrste projekata.

2 **Da konkretizujete problem koji ćete rešavati**
To može biti predikcija, analiza podataka putem klasterovanja, automatska preporuka proizvoda korisnicima,...
Najlakši način da ovo uradite jeste da potražite nekoliko radova u kojima se koristio isti ili sličan skup podataka i da vidite kako je tamo definisan problem i koju motivaciju su naveli za njegovo rešavanje.
Za pretragu literature preporučujemo Google Scholar akademsku bazu https://scholar.google.com/. Na slici 1 je prikazano kako možete pristupiti željenoj publikaciji. Ako neka publikacija nije javno dostupna, možete probati da joj pristupite sa akademske mreže (u Park City-ju ili u računarskom centru FTN-a) ili pišite asistentkinji na mail pa će ona pokušati da je nađe za vas. Korisno uputstvo o čitanju naučnih radova (i havtanje beležaka o pročitanom) možete naći na sajtu predmeta **/Projekat/Kako procitati naucni rad.pdf**.



Slika 1 Pristup željenoj publikaciji

3 **Da date predlog kako se problem može rešiti**
Ideju o ovome dobićete čitajući radove na sličnu temu. Ne tražimo da smislite nov metod za rešavanje nekog problema, moguće je i da samo ponovite postojeća istraživanja. Razmislite da li se neki pristupi koje ste videli mogu kombinovati.

4 **Da date predlog kako se rešenje može evaluirati**
Treba da definišete evaluacionu proceduru (podela na trening/test skup, unakrsna validacija,…) kao i meru evaluacije performansi (accuracy, R^2,...). Najlakše je da vidite kako je to rađeno u prethodnim publikacijama.

Sadržaj koji bi trebao da se nađe u vašem predlogu projekta je opisan u poglavlju 2.

Konačno, imajte u vidu da se vaš plan realizacije cilja projekta) može značajno izmeniti kada naučite više o podacima sa kojima ste planirali da radite ili nakon što pročitate više o tome kako su drugi ljudi rešavali slične probleme. Ovo je u redu sve dok je cilj projekta (zadatak koji rešavate i skup podataka za koji ste se odlučili) približno sličan onome što ste dali u predlogu projekta. Ne morate nas obaveštavati o manjim promenama, ali bi bilo dobro da nas obavestite ukoliko je došlo do većeg zaokreta u odnosu na ono šta ste planirali da uradite.

1 Smernice za odabir projekata

Cilj Vašeg projekta je da steknete i dokumentujete vaše iskustvo o primeni tehnika istraživanja podataka na jednom ili više skupova podataka. Tokom istraživanja trebalo bi da prođete kroz sledeće korake:

- Identifikacija skupa podataka i domena problema
- Odluka šta želite da postignete primenom tehnika istraživanja podataka
- Odabir odgovarajućih metoda i algoritama
- Implementacija i testiranje korišćenih/razvijenih metoda
- Evaluacija primenjenih tehnika na skupu podataka
- Izveštavanje o rezultatima

1.1 Tipovi projekata

Tipične vrste projekata su:

- Fokusiranje na podatke i određeni zadatak (npr. predikcija, klasterovanje, …) i korišćenje nekoliko različitih algoritama iz literature za rešavanje datog zadatka. Fokus je u ovom tipu projekta više na skupu podataka i zadatku nego na algoritmima
- Fokusiranje na algoritme: porediti jedan ili više osnovnih (*baseline*) algoritama sa novim (nedavno objavljenim) algoritmom. Fokus je da se proveri da li novi algoritam zaista radi kao što je napisano u radu u kome je algoritam objavljen
- Razvoj novog algoritma/metode i njegova implementacija. Algoritam je potrebno primeniti na skup podataka u cilju rešavanja određenog zadatka. Važno je evaluirati kakve su performanse Vašeg algoritma u poređenju sa postojećim osnovnim metodama (dovoljna je jedna metoda)

Za bilo koji projekat je potrebna pažljiva primena empirijskih evaluacija, npr. korišćenje odgovarajućih particija skupa podataka na trening i test podatke, unakrsna evaluacija, … Kod projekta se neće ocenjivati dobijeni rezultati (nije potrebno dostići visoke performanse ili pobediti postojeći algoritam). Sa druge strane, važno je obezbediti uvid u to zbog čega je neka tehnika rezultovala uspehom ili neuspehom.

1.2 Skupovi podataka

Skup podataka bi trebao da je dovoljno velik skup iz oblasti problema koji želite da rešavate ili o kome želite da naučite. Neki predlozi su:

- Jedan ili više dobro poznatih skupova podataka iz oblasti kategorizacije teksta (teksta obeleženog ciljnim obeležjem, odnosno labelom) http://disi.unitn.it/moschitti/corpora.htm
- Cela Wikipedia ili neki njen deo https://en.wikipedia.org/wiki/Wikipedia:Database_download – web stranice sa linkovima i oznakom kategorije (labelom) i ostalim meta podacima
- Skup *Million Song data set* (http://labrosa.ee.columbia.edu/millionsong/) koji sadrži audio obeležja i meta podatke za milion pesama (podaci zgodni za sistem za davanje preporuka u muzici)
- US patenti od 1980 do 2012 http://www.google.com/googlebooks/uspto-patents-assignments.html ili US trademark aplikacije od 1870 do 2012 http://www.google.com/googlebooks/uspto-trademarks.html (tekstualni skup podataka, meta podaci, vremenske serije)
- MIMIC II skup podataka o pacijentima sa intenzivne nege https://physionet.org/mimic2/ (analiza vremenskih serija, meta podaci)
- Google sintaktički n-grami posmatrani tokom vremena http://googleresearch.blogspot.co.il/2013/05/syntactic-ngrams-over-time.html (analiza teksta, vremenske serije)

- Skup podataka o globalnim događajima http://www.gdeltproject.org/data.html#rawdatafiles sa 250 miliona geo-referenciranih događaja od 1979 do danas (prostorni, temporalni, relacioni podaci o političkim događajima u celom svetu)
- Skupovi podataka o socijanim mrežama https://snap.stanford.edu/data/index.html (relacioni podaci o mrežama, temporalni, prostorni, ...)
- Microsoft-ov *Learning-to-rank* skup podataka http://research.microsoft.com/en-us/projects/mslr/ (rezultati pretrage (*search engine*) za različite upite sa ocenama relevantnosti)
- Neki skup podataka sa UCI repozitorijuma http://archive.ics.uci.edu/ml/datasets.html
- Neki od skupova podataka sa https://www.kaggle.com/
- http://sail.usc.edu/iemocap/ - za skup podataka se obratiti asistentkinji na mail

Predložene skupove podata je moguće i kombinovati sa nekim drugim skupovima podataka ukoliko smatrate da bi to bilo korisno. Na primer, interesantno bi bilo kombinovati onformacije dostupne na Wikipediji za pobojšanje algoritma za kategorizaciju teksta. Ili, moguće je obučiti algoritam na jednom domenu (skupu podataka), a zatim ga testirati na nekom drugom domenu u cilju evaluacije koliko dobro razvijeni pristup generalizuje na druge domene. Moguće je odabrati i neki skup podataka koji nije naveden u listi, kao i izvući podskup navedenog skupa ukoliko je previše velik za obradu.

1.3 Softver korišćen u izradi

Naše predloge za izbor softvera koji možete koristiti naći ćete na sajtu predmeta u **/Vežbe/Preporučeni softver.pdf**. Možete koristiti kodove postojećih algoritama ukoliko su vam dostupni, samo pazite da navedete reference i mesto odakle ste skinuli softver. Tipičan projekat bi mogao da sadrži *pipeline* koraka analize ili procesiranja podataka gde neki mogu biti napisani od strane vas, a neki mogu biti iskorišćeni delovi postojećeg softvera.

2 Predložen šablon i sadržaj predloga projekta

Vaš predlog projekta treba da sadrži sledeće elemente:

1. **Definicija problema/cilja projekta**
   Jasna definicija problema. Na primer, automatska definicija sentimenata tweet-ova, razvoj sistema za preporuku filmova, predikcija kretanja cene akcija,…
2. **Motivacija problema rešavanog u projektu**
   Ukratko objasniti zbog čega je vredno rešavati ovaj problem (npr. rešavanje važnog otvorenog istraživačkog problema, adresiranje važne praktične primene,…). Dakle, ne vaša lična motivacija, nego gde bi se vaše rešenje moglo praktično primeniti i koje bi probleme rešilo.
3. **Relevantna literatura (minimum 3 rada)**

Dajte kratak pregled onoga što se zna u o datom problemu u literaturi. Trebalo bi navesti barem 3 relevantne reference (naučni radovi). Za svaki od navedenih radova napisati jedan pasus koji obuhvata:

1. Šta je bio zadatak rada (šta je predviđano, kakvi su bili ciljevi i slično)
2. Koja se metodologija koristila (npr. primena Naivnog Bajesa, primena regresije, ...)
3. Kakav je bio skup podataka (naznaka kavi su atributi, referenca na rad u kome je skup podataka objavljen)
4. Kako je evaluirano rešenje (unakrsna validacija/podela na trening/test skup, ... i koja se mera koristila – accuracy/$R^{2...}$)
5. Koji su najvažniji rezultati **za vaš rad** u rečenici do dve (neki interesantni zaključci tipa „pokazano je da se metodom x može dobro predvideti y, ali ne i z“ / „dostignuta je tačnost od x%“ / „metoda x je za ovu primenu bolja od metode y“)
6. **OBAVEZNO: izvedite zaključak – šta ćete od navedenog u radu primeniti u vašem projektu (metod, skup, evaluacija,...), a po čemu će se vaš rad razlikovati**. Tj. ovim sumarizujete zašto smatrate da je dat rad relevantan za vaš projekat
7. (Opciono) šta po Vama nedostaje ovom rešenju? (“Autori nisu uzeli u obzir…”/ “bilo bi lepo dopuniti sa”/”Radi dobro na x, ali pitanje je kako radi na y”)

U ovoj fazi nije potrebno dati pun pregled relevantne literature koju ćete koristiti u završnom izveštaju, ali je potrebno da ste svesni o tome šta su drugi istraživači radili na sličnu temu. Na primer, proverite da li projekat koji želite da uradite nije neko već rešavao na potpuno isti način kao što ste planirali.

Prilikom pretrage literature se trudite da radovi budu što skorašnjiji (idealno od 2014 pa naviše, a neka najdonja granica bude 2010). Dobar sajt za pretragu je *google scholar* *https://scholar.google.com/*. Ako rad nije javno dostupan, možete ga potražiti putem http://sci-hub.tw/. Za knjige je dobar sajt https://libgen.io/.

4. **Skup podataka**

   Opišite skup/skupove podataka koje planirate da koristite u projektu.

   Postojeći (javno dostupan) skup podataka:

   - Stavite referencu na rad u kome skup podataka konstruisan i link gde se skup podataka može skinuti
   - Naglasite koji atribut predstavlja ciljno obeležje i šta je njegov sadržaj (klasifikacija – koje klase postoje i koja je raspodela podataka po klasama, regresija – koji je opseg u pitanju)
   - Koji atributi/grupe atributa postoje u skupu podataka (na osnovu čega predviđate)

   Ako sami konstruišete skup podataka:

   - naglasite odakle skidate podatke (link)
   - koje atribute želite da napravite (na osnovu čega predviđate)
   - kako planirate da ga anotirate (dođete do ciljnog obeležja).
   - Koje vrednosti bi sadržalo ciljno obeležje (klasifikacija – koje klase postoje, regresija – koji je opseg u pitanju)

5. **Metodologija**

   Kratak nacrt metodologije koju planirate da primenite (npr. primena modela Naivnog Bayesa ili implementacija nekog algoritma). **Metodologija mora biti bazirana na radovima navedenim u relevantnoj literaturi.**

6. **Metod evaluacije**

   morate definisati kako planirate da evaluirate/izmerite/testirate vaše rezultate ili primenjene tehnike. Za probleme predikcije (klasifikacija ili regresija) ovo je prilično jednostavno – mere poput tačnosti (accuracy) ili R^2 su dobro definisane i dobre indikacije performansi koje postiže Vaš metod. Za tehnike poput klasterovanja ili pronalaženja patterna nije baš toliko jasno kako meriti uspeh. Jedna moguća tehnika je da pokušate da primenite algoritam za klasterovanje na skupu podataka gde su labele poznate ali uklonjene (ta informacija nije dostupna prilikom klasterovanja), a zatim klasterovanjem pokušate da ih povratite. Za ovakve tipove problema bi bilo dobro da se konsultujete sa literaturom kako bi ste procenili koje se tehnike evaluacije tipično koriste. Za projekte vezane za vizuelizaciju bi bilo dobro da se obratite nekolicini vaših kolega da probaju da koriste vaš sistem i daju subjektivnu evaluaciju (npr. uporede vaš metod sa drugim pristupom "na slepo" – testeri dobijaju vaš metod i konkurentski metod i ocenjuju ih ne znajući koji je koji).

   Obavezno definisati:

   - **Postupak evaluacije (eksperiment)** – npr. unakrsna validacija, podela na trening/validacioni/test skup (definisanti odnos)
   - **Meru evaluacije** – npr. tačnost, R^2,…

7. **(Opciono) Softver**

   Navedite i ukratko opišite softver i algoritme koje planirate da primenite u vašem projektu. Ukoliko planirate da sami razvijete novi algoritam, pokušajte da skicirate kako bi on izgledao i kako mislite da će raditi. Ukoliko ste se odlučili za postojeći algoritam, ukratko ga opišite i navedite relevantne reference. Takođe, ukoliko planirate da iskoristite nečiju tuđu implementaciju softvera, navedite reference, gde je taj softver dostupan (npr. link na web sajt) i navedite opis tog softvera (detalje koji su Vam dostupni).

8. **(Opciono) Plan**

   Skicirajte šta bi bile veće prekretnice (milestones) koje planirate da dostignete prilikom rada na vašem projektu. Navedite nekoliko tačaka. To ne mora da bude svaki mali detalj ali dajte generalnu ideju onoga što planirate da uradite do termina predaje projekta u aprilu (a po mogućnosti i do demonstracije prvih rezultata koja je planirana na kraju semestra). Ovi planovi se naravno mogu menjati tokom procesa rada sa podacima i algoritmima. Ne bi bilo loše da identifikujete potencijalne rizike u vašem predlogu, odnosno stavke koje mogu izazvati odlaganje ili probleme.