Izrada projekta

Sistemi za istraživanje i analizu podataka

Rokovi

Obaveza	Datum	Način predaje
Predaja predloga projekta	5.12.2016.	Asistentkinji na mail (slivkaje@uns.ac.rs)
Predaja revidiranog predloga projekta	11.12.2016.	Asistentkinji na mail (slivkaje@uns.ac.rs)
Prezentacija predloga projekta	13.12.2016.	Prezentacija u trajanju od 5-10 min u terminu predavanja
Prva kontrolna tačka	U nedelji 16.1.2016. – 20.1.2016	Usmene konsultacije sa asistentkinjom u zakazanom terminu
Druga kontrolna tačka	Prva nedelja letnjeg semestra (27.02.2016. – 03.03.2017.)	Usmene konsultacije sa asistentkinjom i profesorom u zakazanom terminu
Predaja finalne verzije projekta	15.04.2017.	Asistentkinji na mail (slivkaje@uns.ac.rs)

Izrada projekta

- Sami (uz konsultacije sa profesorom i asistentom) istražujete i predlažete teme
  - Dobićete sugestije gde možete da pronađete skupove podataka i da se informišete o temama koje vas zanimaju
- Timski rad u grupama od 2-3 studenata
- Svaki student mora da da svoj doprinos – tražićemo da specificirate ko je šta radio

Predaja predloga projekta

- Kratak (ali informativan) izveštaj na 1-3 strane A4 formata
- Predlog projekta šaljete asistentkinji na mail do 5.12.2016.
  - Predlog možete poslati i pre ovog termina kako biste dobili sugestije šta treba dopuniti/izmeniti
- Predlozi će biti revidirani i biće vam date sugestije šta je potrebno dopuniti/izmeniti da bi vam projekat bio prihvaćen
- Revidirane predloge projekta šaljete asistentkinji na mail najkasnije 11.12.2016.
- Ako niste predali finalnu verziju predloga projekta 11.12.2016. ili je ona nepotpuna (niste ispoštovali sve sugestije koje smo vam dali), maksimalna ocene koju možete dobiti iz projekta je 6
- Sugestije kako sastaviti predlog projekta se nalaze na sajtu predmeta u /Projekat/Predlog projekta.pdf
- Primer lepo napisanog predloga projekata možete naći na sajtu predmeta u /Projekat/Primer dobrog predloga.pdf/

Prezentacija predloga projekta

- Potrebno je da pripremite kratku prezentaciju u trajanju od 5 – 10 minuta koju ćete održati pred profesorom, asistentom i svojim kolegama u terminu predavanja
- Cilj:
  - dobićete dodatne sugestije od profesora, asistenta i vaših kolega kako da projekat bude bolji
  - upoznaćete se sa radom vaših kolega, videti primere tema u istraživanju podataka i na koji način se (okvirno) rešavaju
  - potencijalno ćete uočiti druge timove koji se bave sličnom temom pa možete sarađivati u smislu razmene iskustava i literature
- Prezentacija predloga projekta je obavezna
  - Može i samo jedan član tima

Prva kontrolna tačka

- U ovom trenutku očekujemo da ste krenuli sa izradom projekta i da imate neke osnovne rezultate da nam pokažete
  - To može biti primenjen neki osnovni algoritam za klasifikaciju/regresiju/klasterovanje…
  - Imate neku meru performansi tog algoritma (npr. *accuracy*, R^2, analizirali ste klastere, …)
- Cilj je da nam pokažete:
  - da ste se dobro upoznali sa problemom koji rešavate
  - da imate skup podataka na kome radite
    - ako sami kreirate skup podataka, treba da ste ga formirali barem delimično tako da možete da isprobate neki osnovni model koji ste naveli u predlogu projekta
  - da razumete cilj projekta
    - Npr. ako se radi o predikciji neke vrednosti – šta je ciljna varijabla, koje vrednosti može da uzima, kako su te vrednosti zastupljene u skupu podataka, itd.
- Radi se o usmenoj odbrani kod asistentkinje u zakazanom terminu
  - Ne morate spremiti prezentaciju, ali ponesite fajlove koji su vam potrebni (možete i na sopstvenom laptopu)
  - Odbrana je obavezna – ako iz opravdanih razloga niste u mogućnosti da dođete u zakazanom terminu, molimo vas da pre isteka roka javite da zakažemo drugi termin

Druga kontrolna tačka

- U ovom trenutku očekujemo da ste prilično napredovali sa projektom
  - Izvršili eksplorativnu analizu podataka
  - Isprobali većinu modela koje ste naveli u predlogu
  - Optimizovali ih
  - Možete da izvedete neke zaljučke iz svojih rezultata
- Cilj je da vidimo kako teče projekat i da konkretizujemo da li i šta je još potrebno da uradite kako biste ostvarili maksimalnu ocenu iz praktičnog dela
- Ne zaboravite da nako izrade ostavite sebi vremena i da napišete završni izvreštaj o projektu
- Radi se o usmenoj odbrani kod profesora i asistentkinje u zakazanom terminu
  - Ne morate spremiti prezentaciju, ali ponesite fajlove koji su vam potrebni (možete i na sopstvenom laptopu)
  - Odbrana je obavezna – ako iz opravdanih razloga niste u mogućnosti da dođete u zakazanom terminu, molimo vas da pre isteka roka javite da zakažemo drugi termin

Predaja finalne verzije projekta

- Potrebno je da do naznačenog roka asistentkinji na mail pošaljete završni izveštaj
  - Radi se o teksualnom izveštaju u kome opisujete
    - Problem koji ste rešavali i motivaciju za njegovo rešavanje
    - Pregled relevantne literature koju ste pročitali
    - Skup podataka
    - Metod/algoritme koje ste primenili
    - Rezultate
    - Zaključke
  - Na sajtu predmeta su okačene i detaljnije sugestije o sadržaju izveštaja /Projekat/Pisanje završnog izveštaja o projektu.pdf
  - Na sajtu predmeta se nalazi i propisan format (IEEE Template u dve kolone) /Projekat/Sablon_word_A4
  - Propisan obim: 6-8 strana za grupe od 2 člana i 8-10 strana za grupe od 3 člana

Predaja finalne verzije projekta

- Ako izveštaj završite pre 15.04.2017., možete ga poslati asistentkinji na mail kako biste dobili komentare o samom tekstu izveštaja
  - Pravo na ovo imate samo ako ste ispoštovali i prvu i drugu kontrolnu tačku
  - Ovo možete uraditi u najviše dva navrata. Zato se potrudite da kada dobijete sugestije ispravite što je moguće više, kako biste dobili još konstruktivnih komentara u narednoj iteraciji
- Primere lepo urađenih projekata možete naći na sajtu predmeta u /Projekat/Primer dobrog projekta.pdf

Šta se sve ocenjuje

- Poštovanje obaveza u zakazanim rokovima
- Koliko ste dobro upoznati sa problemom
  - Da li je literatura koju ste izložili dovoljno obimna i relevantna
  - Korisne sugestije za čitanje naučnih radova se nalaze na slajdovima /Projekat/ Kako procitati naucni rad.pdf
- Koliko toga je urađeno (timski i pojedinačno)
  - Procenu šta sve treba da uradite/doradite za maksimalnu ocenu dobićete na kontrolnim tačkama
  - Pored toga, u svakom trenutku (do predaje finalnog izveštaja) možete dogovoriti konsultacije kako biste proverili da li projekat ide u pravom smeru
- Koliko dobro je napisan završni izveštaj
  - Mora da poštuje propisan šablon
  - Mora da bude adekvatnog obima
  - Tema mora da bude izložena na adekvatan i razumljiv način (detaljan podsetnik možete naći na sajtu predmeta /Projekat/Pisanje završnog izveštaja o projektu.pdf)

Česti propusti u izradi projekta

- Niste obrazložili sve izbore parametara (zašto takva podela na trening/test skup, zašto ta mera evaluacije, zašto smatrate da je rejting od 3 pa naviše pozitivan, itd.). Mogući razlozi:
  - do njih došli empirijski (optimizacijom modela, isprobavanjem,…)
  - prilikom eksplorativne analize ste uočili neki šablon
  - na osnovu domenskog iskustva
  - drugi autori koriste slične postavke parametara
- Primenili ste model a zaboravili da optimizujete vrednosti njegovih parametara
- Trenirali ste model ili optimizovali parametre na istim podacima (ili na podskupu podataka) na kome evaluirate model. Napravite razliku između trening/validacionog/test skupa

Moguća pobojšanja kvaliteta projekta

- „Dodatne poene“ možete osvojiti analizom grešaka modela. Izdvojite podskup primera na kojima je vaš model pogrešio i pokušajte da razumete razlog. Npr. napravili ste model za automatsko prepoznavanje sentimenta teksta. Po inspekciji grešaka koje model pravi utvrdili ste da ne prepoznaje negaciju ili sarkastične komentare
- Razmislite da li možete da kombinujete postojeće pristupe, npr. da li možete da kombinujete glasove različitih klasifikatora

Primeri predloga projekata

(prethodne generacije studenata)

Fantasy football

- Fantasy Football je takmičenje originalno osmišljeno za Američki fudbal
- Takmičari sastavljaju svoj tim, predvidjajući na osnovu onoga što znaju o igračima (u tom trenutku) kako bi izgledao “dream team” u narednoj nedelji lige
- Postoje ograničenja i pravila
  - budžet za kupovinu i prodaju igrača
  - ograničenje broja igrača koji potiču iz istog tima
  - poeni kapitena tima se dupliraju
- U zavisnosti od toga kako se njihovi odabrani igrači pokažu u realnim utakmicama, takmičari dobijaju poene na osnovu kojih se rangiraju na nedeljnim, mesečnim i godišnjim listama

Fantasy football

- Cilj projekta:
  1. Prevideti broj poena koji će svaki od igrača osvojiti – problem predikcije
  2. Na osnovu predikcija odrediti tim koji će imati najbolji rezultat (uz postavljena ograničenja i pravila) – optimizacioni problem
- Motivacija:
  - Veliko interesovanje za takmičenje (milioni korisnika)
  - Forbes je estimirao da je Fantasy Football tržište vredno oko $70 milijardi, sa 32 miliona učesnika
- Skup podataka:
  - Na zvaničnoj stranici lige dostupni su podaci o svim mogućim preformansama igrača su dostupni sa zvanicne stranice lige
  - Prethodne perfromanse igrača
  - Uslovi igre („raspoloženje “ igrača zavisi od toga da li se igra na domaćem ili stranom terenu, ko je protivnik,...)
  - Uvid u fizičko stanje igrača (npr. povrede)

Fantasy football

- Plan
  1. Sakupljanje podataka
  2. Filtriranje/čišćenje podataka
  3. Eksplorativna analiza – cilj je odrediti koji podaci su relevantni za predikciju
  4. Implementacija postojećeg algoritma

     T. Matthews, S. Ramchurn, and G. Chalkiadakis. Competing with humans at fantasy football: Team formation in large partially-observable domains. 2012.
  5. Uporediti sa drugim dostupnim metodama (iz drugih publikacija)
  6. Istražiti da li se postojeći pristup može pobojšati
  7. Odrediti koji algoritmi su pogodni za ovu vrstu zadataka, a koji nisu i zbog čega
  8. Skidati podatke na nedeljnom nivou i pratiti kako algoritmi reaguju na različite događaje – na šta su osetljivi, koji je bolji u čemu,...
- Evaluacija
  - Međusobno poređenje algoritama
  - Predviđa se broj poena (kontinualna varijabla) – RMSE

Sport

- NBA liga: predviđanje učinka igrača na narednoj utakmici i određivanje pozicije koja najviše odgovara igraču na osnovu njegovog stila igre
  - Profesionalni košarkaški klubovi su počeli da uvode statistiku i razne vrste predviđanja, radi postizanja što boljeg uspeha: predviđanje povrede igrača, osvajača u narednoj sezoni, uspeha pojedinačnog igrača,...
  - Predviđanje učinka pojedinačnih igrača iz ekipe narednog protivnika, može u velikoj meri da utiče na taktiku trenera
  - Cilj trenera je da od igrača izvuče maksimum na poziciji koja mu najbolje odgovara
- Slične projekte možete osmisliti i za druge sportove poput tenisa, trka konja,...

Video igre

- Predikcija ishoda meča u igrici Dota2 i predlog heroja sa kojim igrač ima veće šanse za pobedu
  - RPG igra u kojoj su igrači podeljeni na dve frakcije
  - Na početku meča igrači biraju jednog (od preko 100 mogućih) heroja, pri čemu heroji imaju neke svojstvene karakteristike (strength, agility, intelligence)
  - Motivacija: nagradni fond za takmičenje *The International Dota 2 Championships* je preko 18 miliona dolara

Capital bikesharing

- Cilj: predviđanje lokacija novih stanica za bicikle i uklanjanje nepotrebnih stanica
- Motivacija:
  - Ekspanzija sistema za iznajmljivanje bicikala u velikim gradovima
  - Stanice sa većim prometom su popularnije i donose veći prihod – uvećenje profita i zadovoljstva korisnika
  - Za otvaranje ili zatvaranje stanice kompanija mora da uloži finansijska sredstva

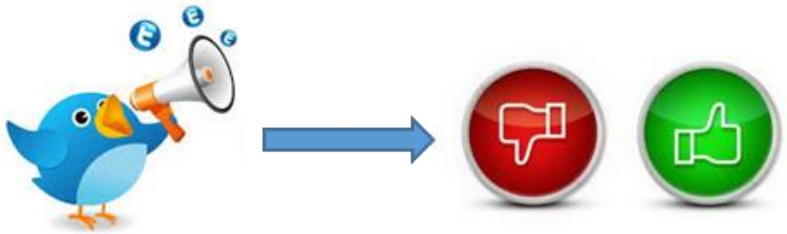
Capital bikesharing

- Skup podataka
  - *http://www.capitalbikeshare.com/trip-history-data* - informacije o trajanju putovanja (datum, vreme početka i kraja putovanja), početna i krajnja stanica (i geografske koordinate), tip članarine,... Takođe se na osnovu datih podataka može odrediti popularnost stanica (ciljna varijabla)
  - http://wiki.openstreetmap.org/wiki/ - preuzeti podaci o objektima u blizini stanica: stanice metroa, velike kompanije, znamenitosti, restorani,... Uticaj objekata je rangiran prema udaljenosti od stanice (u minutima šetnje i km)
  - Podaci su ručno filtrirani i pročišćeni
- Problem je definisan kao problem klasifikacije – stanice su razvrstane u 5 kategorija popularnosti
- Primenjen je model stabla odlučivanja
- Evaluacija: unakrsna validacija, merenje preciznosti i odziva

Sentiment analiza tweetova

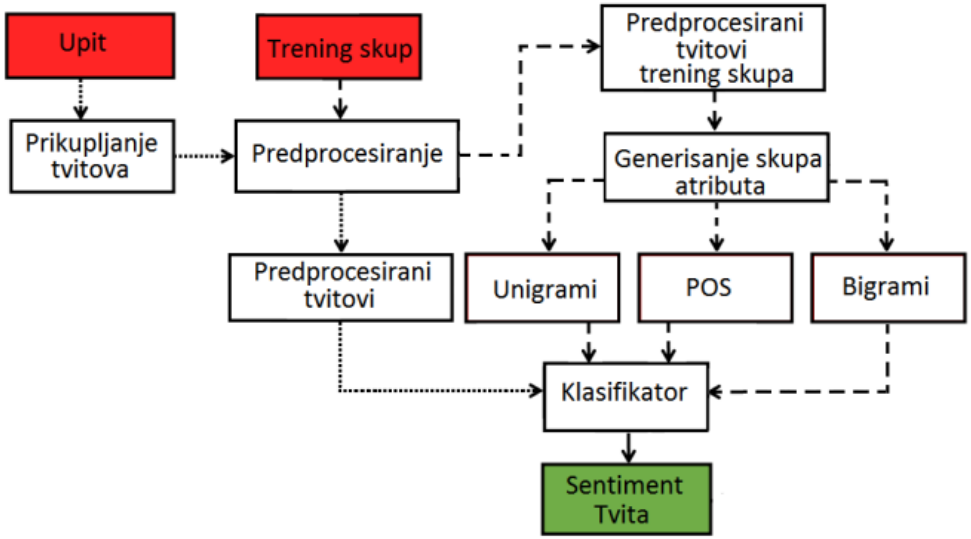
- Cilj: automatsko određivanje sentimenta tweet-a (pozitivno ili negativno)
- Motivacija:
  - Automatsko određivanje mišljenja ciljne grupe korisnika o događajima, proizvodima, poznatim ličnostima, kompanijama, ...
  - Primena u društvenim naukama (sociološke, ekonomske, istorijske, pravne), novinarstvu i reklamnim kampanjama
  - Kompanije mogu na ovaj način da ispituju javno mnjenje o svom proizvodu
- Skup podataka:
  - Sakupljani su tweet-ovi kojima je automatski dodeljivan sentiment na osnovu emotikona
  - Za potrebe evaluacije modela korišćen je ručno anotiran skup podataka

Emotikon	*Osećaj*
:-D, 8-D, 8D, x-D,xD, X-D, XD, =-D, =D	very happy
:-), :), :D, :o), :], :3, :c), :>, =], 8), =), :}, :^), :⁀)	happy
;-), ;), *-), *), ;-], ;], ;D, ;^), :-,	wink
>:P, :-P, :P, X-P, x-p, xp, XP, :-p, :p, =p, :-Þ, :Þ	cheeky
>:[, :-(, :(, :-c, :c, :-<, :⁀C, :<, :-[, :[, :{	sad



Sentiment analiza tweetova

- Formiran je pipeline za procesiranje tweet-ova



- Ispitivane su performanse različitih klasifikacionih algoritama (NB, SVM, Random forest,...) u kombinaciji sa različitim obeležjima skupa podataka (bez/sa POS tagova, formiranih bigrama,…)

Sentiment analiza

- Predikcija rejtinga restorana/usluge/servisa na osnovu sentimenta komentara
  - Xu, Y., Wu, X. and Wang, Q., 2015. Sentiment Analysis of Yelp‘s Ratings Based on Text Reviews – korišćeni su komentari sa Yelp-a
  - Dragan Vidaković
    - Sakupljeni komentari sa sajta http://www.donesi.com/
    - Korisnički nickname, naslov komentara, vreme postavljanja, tekst komentara, opis restorana, rejting,...
    - Ručno čišćenje komentara (ćirilica, ch-> č, non-ASCII),...
    - Izazov: nema puno resursa za srpski jezik
    - Eksplorativna analiza
      - Po zemljama
      - Kako izgleda prosečan pozitivan/negativan komentar (pridevi, često korišćene reči, velika/mala slova, dužina rečenica, interpunkcija,...)
      - Koji atributi će imati uticaja,...
  - Slična ideja mogla bi se sprovesti na nekom drugom sajtu, npr. http://oceniprofesora.com/

Sentiment analiza

- Slično, možete raditi analizu sentimenata revizija filmova, novinskih članaka (i komentara novinskih članaka), blogova, youtube komentara,...
- Npr. predviđanje rejtinga osnovu komentara

- Automatsko sortiranje muzike po žanrovima
  - Motivacija: razvoj softvera za asistenciju prilikom izbora muzike

Text mining

- Prepoznavanje plaćenih komentara
- Skinuti su komentari sa B92 i Blica i anotirani od strane korisnika kao “bot” ili “nije bot’ http://startit.rs/lovac-na-sendvice-bot/
- Izazov: ne postoji mnogo alata za procesiranje srpskog jezika

Sentiment analiza

- IEMOCAP baza podataka http://sail.usc.edu/iemocap/
  - Radi se o detekciji emocija (bes, sreća, gađenje,...) na osnovu audio-vizuelnih podataka
  - Baza podataka je snimljena pomoću 10 glumaca koji su izražavali emocije na osnovu zadatih scenarija, a ima i improvizacija
  - Sadrži audio, video i audio transkripte (tekst)
  - Ideja: pomoću ovih podataka formirati i kombinovati tri različita klasifikatora radi maksimizacije postignute tačnost klasifikacije emocija
  - Mogla bi se isprobati neka varijanta sa polu-nadgledanim obučavanjem, npr. co-training bi bio dobar kandidat s obzirom da ima više nezavisnih izvora podataka
  - Moglo bi se pokušati da se spoji ova baza sa drugim slobodnim emotivnim audio bazama kako bi se maksimizovala tačnost
  - Obratiti se asistentkinji za podatke

Još neke ideje – predviđanje cena akcija

- Predviđanja budućih cena akcija Apple na osnovu istorijskih cena
  - Poređenje postojećih matematičkih modela sa tehnikama istraživanja podataka
- Na osnovu sadržaja novinskih članaka (teme i sentimenta) predvideti promenu berzanskih vrednosti
  - Pokazano je da postoji korelacija između raspoloženja twitter korisnika i berzanskog indeksa
  - Pokazano je da postoji korelacija između objavljenih novinskih članaka i vrednosti akcija

Prepoznavanje ljudske aktivnosti

- Prepoznavanje tipa ljudske aktivnosti (šetanje, penjanje uz stepenice,...) na osnovu senzora mobilnog telefona
  - Primena: kućna rehabilitacija za ljude koje pate od traumatičnih povreda mozga, detekcija pada kod starijih osoba,...
  - Tipično se za merenje pokreta tela koriste razni senzori i uređaji (akcelerometar, GPS,...) – nezgodno i skupo
  - Mobilni telefon: rasprostranjen, pristupačan, neosetan za nošenje. Još jedna moguća primena bi bila mogućnost prilagođavanja režima rada prema trenutnoj aktivnosti vlasnika – preusmeriti pozive na govornu poštu ukoliko vlasnik džogira

Još neke ideje...

- Poređenje learning-to-rank (LTR) algoritama
  - Data je lista stavki – rangirati ih prema relevantnosti
  - Primena: sistemimi za pronalaženje informacija, sistemi za preporuku proizvoda, analiza emocija, izdvajanje ključnih reči, mašinsko prevođenje teksta,...
- Automatsko klasifikovanje životnih osiguranika u grupe rizika
  - Motivacija: proces kreiranja polise životnog osiguranika je složen i troši vreme i resurse – automatizacija bi ubrzala i olakšala ovaj proces i time donela poslovnu korist kompanijama
- Predviđanje popularnosti stranica na Wikipediji
  - Motivacija: praćenje, analiza i predviđanje popularnosti ideja, događaja, lokacija,...
  - Imalo bi primenu u sistemima za davanje preporuka
  - Ukazalo bi kako modifikovati stranicu u cilju povećanja njene popularnosti

Još neke ideje...

- Predviđanje da li će avionski let kasniti ili ne
  - za koji period dana je najmanja verovatnoća otkaza/kašnjenja leta?
  - šta je najčešći uzrok kašnjenja/otkaza?
  - ima li kašnjenje veze sa tipom aviona?
  - na kojim rutama je kašnjenje najčešće?
  - koliki uticaj imaju vremenske nepogode naspram ostalih uzroka?
  - Motivacija: stručnjaci procenjuju da će do 2025. godine procenat letova koji kasne ili su otkazani skočiti za 30%, što će rezultovati u velikim gubicima ne samo avio kompanija već i čitave američke privrede