Analiza tekstova projektnih prijedloga primjenom postupaka statističke obrade prirodnoga jezika

Petric Maretić, Hermina

Master's thesis / Diplomski rad

2015

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **University of Zagreb, Faculty of Science / Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:217:603601

Rights / Prava: In copyright / Zaštićeno autorskim pravom.

Download date / Datum preuzimanja: **2024-03-28**

Repository / Repozitorij:

Repository of the Faculty of Science - University of Zagreb

SVEUČILIŠTE U ZAGREBU
PRIRODOSLOVNO–MATEMATIČKI FAKULTET
MATEMATIČKI ODSJEK

Hermina Petric Maretić

ANALIZA TEKSTOVA PROJEKTNIH PRIJEDLOGA PRIMJENOM POSTUPAKA STATISTIČKE OBRADE PRIRODNOGA JEZIKA

Diplomski rad

Voditelji rada:
doc. dr. sc. Jan Šnajder
prof. dr. sc. Mladen Vuković

Zagreb, srpanj 2015.

Ovaj diplomski rad obranjen je dana _____ pred ispitnim povjerenstvom u sastavu:

1. _____, predsjednik
2. _____, član
3. _____, član

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Sadržaj

Sadržaj **iii**

Uvod **2**

1 Pregled dosadašnjih istraživanja i korištenih metoda **3**
1.1 Grupno financiranje 3
1.2 Pregled dosadašnjih istraživanja 5
1.3 Metoda potpornih vektora 6
1.4 Modeliranje tema 9

2 Analiza tekstova projektnih prijedloga **14**
2.1 Skupljanje podataka 14
2.2 Predprocesiranje 14
2.3 Kategorizacija 15
2.4 Modeliranje tema 20

3 Rezultati **22**

Bibliografija **28**

Uvod

Ideja grupnog financiranja(eng. crowdfunding) preko interneta nastala je prije svega nekoliko godina, ali je vrlo brzo doživjela nagli rast popularnosti. Danas postoji vrlo veliki broj web stranica koje se bave grupnim financiranjem, a neke od popularnijih objavljuju desetke novih projekata dnevno. Takav bogati fond javno dostupnih projektnih prijedloga čini vrlo zanimljivu bazu za analizu tekstova, kategorizaciju i modeliranje tema u tekstovima.

Radi lakšeg snalaženja, projekti su na svakoj platformi raspoređeni u kategorije. Stoga npr. na platformi *Kickstarter* korisnici koje zanima umjetnost ne moraju prolaziti kroz sve projekte, već mogu izabrati kategoriju *Art*. Kategorizacija projekata vrlo je korisna i istraživačima. Naime, uz pomoć već definiranih kategorija vrlo je lako usporediti uspješnost, prosječnu zaradu projekta ili pak trajanje za različite kategorije. Osim što omogućuje usporedbu kategorija, omogućuje i restrikciju analize na samo nekoliko njih. Na popularno pitanje "Što trebam učiniti da bi moj projekt bio uspješan?" vrlo je teško odgovoriti u svakom slučaju, ali to je još teže ako istovremeno promatramo projekte kojima tematika i ciljana publika nisu nimalo slične. Također, uzmemo li u obzir velike razlike među temama pojedinih kategorija i činjenicu da su neke kategorije često osjetno mnogobrojnije od nekih drugih (čak i do 20 puta mnogobrojnije), vrlo se lako može dogoditi da, zanemarimo li kategorije, pravila uspješnosti koja naučimo puno bolje opisuju zastupljenije od manje zastupljenih kategorija.

Iz tog je razloga prirodno razdvojiti projekte na kategorije i uspoređivati projekte unutar iste klase. Problem nastaje u činjenici da svaka platforma definira svoje kategorije, a između kojih je nemoguće konstruirati preslikavanje. Time je usporedba projekata među platformama znatno otežana i gotovo je nemoguće odgovoriti na pitanje *"Na kojoj platformi projekti poput mojega imaju najveći uspjeh?"*.

U ovom ćemu radu napraviti klasifikator koji na temelju teksta projekta određuje kojoj on kategoriji pripada. Time je problem različitih kategorija kroz platforme premostiv i svojstva projekata mogu se uspoređivati kroz platforme. Nadalje, napravit ćemo modeliranje tema koje se pojavljuju u projektima, što otvara brojne mogućnosti za analizu. Osim analize zastupljenosti tema po kategorijama i usporedbe između platforma, koja je implementirana u ovom radu, moguća je i analiza

uspješnosti tema i projekata koji sadrže određene teme, zatim analiza svojstava tema i usporedba zastupljenosti tema kroz platforme, kao i kroz vrijeme.

Ovaj rad podijeljen je na tri dijela. Prvo poglavlje predstavlja teorijsku podlogu, drugo objašnjava proces analiziranja podataka, dok treći dio rada prezentira dobivene rezultate.

U prvom poglavlju dajemo pregled dosadašnjih istraživanja i metoda korištenih u ovom radu. Nakon objašnjavanja pojma grupnog financiranja, proći ćemo kroz kratak pregled istraživanja vezanih za grupno financiranje, a zatim objasniti metode koje se u radu koriste za klasifikaciju dokumenata i modeliranje tema u istima.

Drugo poglavlje opisuje proces analize tekstova projektnih prijedloga. Ukratko ćemo proći kroz sakupljanje podataka, a nakon toga ćemo objasniti pripremu podataka za analizu i potom samu analizu podataka.

Rezultate analize prezentirat ćemo u trećem poglavlju, u kojem ćemo promotriti vezu između kategorija i tema, a nakon kojega slijedi zaključak te popis referencija.

Poglavlje 1

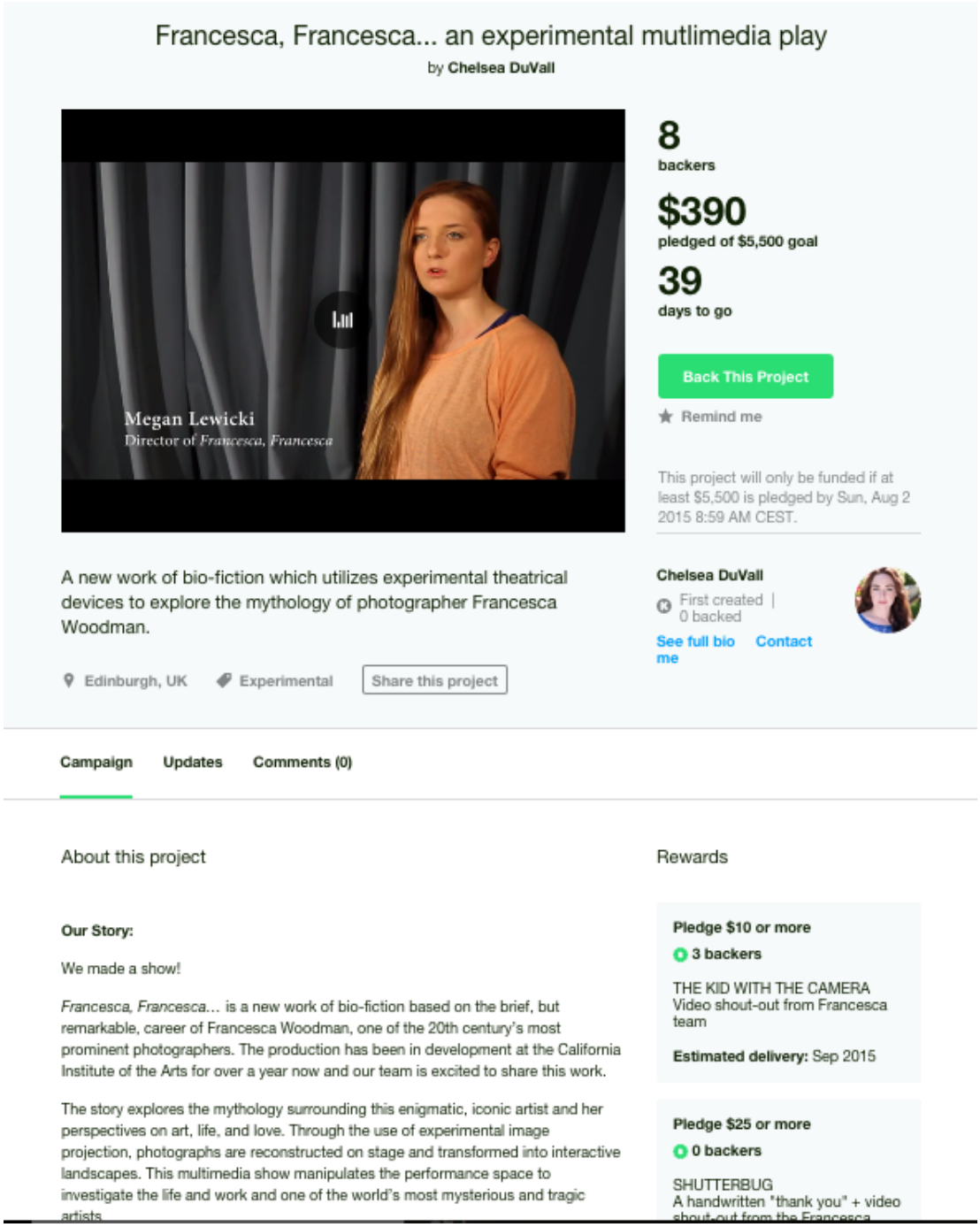
Pregled dosadašnjih istraživanja i korištenih metoda

Grupno financiranje još je uvijek nepoznato i novo područje te je, iako je to trenutno vrlo popularna tema istraživanja, dosadašnje istraživanje još uvijek relativno ograničeno. U ovom ćemo poglavlju objasniti što je uopće grupno financiranje i dati pregled dosadašnjih istraživanja, a nakon toga ćemo dati teorijsku pozadinu metoda korištenih za analizu projektnih prijedloga.

1.1 Grupno financiranje

Grupno financiranje u svijetu je sve popularniji oblik prikupljanja početnog kapitala za raznovrsne projekte, startupe i sl., a porast popularnosti doživljava i u Hrvatskoj, što potvrđuje činjenica da su prošle godine hrvatski projekti na taj način prikupili čak $2,76$ milijuna kuna, odnosno tri puta više nego u svim prijašnjim godinama zajedno.

Kako grupno financiranje uopće funkcionira? Osoba koja želi provesti neki projekt, svoju ideju detaljno objasni i postavi na jednu od platformi za grupno financiranje. Taj opis, osim teksta, može uključivati i slike i videe, a svoj projekt možete i povezati sa svojim društvenim mrežama kako biste osigurali pozornost svojih poznanika. Nakon što postavite projekt, ljudi iz cijelog svijeta imaju mogućnost poduprti ga svojim ulaganjem, u zamjenu za što dobivaju jednu od unaprijed definiranih nagrada koje mogu biti raznovrsne (npr. ukoliko se radi o snimanju filma, nagrada može biti majica s naslovom filma, ali i večera s redateljem za velike ulagače). U trenutku postavljanja projekta potrebno je definirati novčani cilj i želite li biti financirani čak i ako se on ne dostigne. Naime, u slučaju da ste financirani, obavezali ste se na izdavanje nagrada svojim investitorima. Ukoliko se radi o skupom projektu, a ne dosegnete svoj novčani cilj, možda uopće nećete biti u mogućnosti izvesti projekt, pa



Slika 1.1: Primjer projekta na Kickstarter platformi

definirate da želite financiranje samo u slučaju dovoljne zainteresiranosti. Također, potrebno je definirati i vremenski period unutar kojeg su investicije omogućene, kao i vremenski okvir dodjele nagrada za investiranje. Primjer jednog takvog projekta možemo vidjeti na slici 1.1.

1.2 Pregled dosadašnjih istraživanja

Grupno financiranje svojim naglim porastom popularnosti u zadnjih nekoliko godina izaziva vrlo veliki interes i u znanstvenom svijetu. Dosadašnja istraživanja ovakvih projekata dolaze iz raznovrsnih znanstvenih grana, pa tako područja koja proučavaju ovakve projekte uključuju poslovnu ekonomiju s analizom vrste projekata i ciljane publike[1], do prava koje pokušava riješiti sve pravne zavrzlame u novom području [3].

Statistička analiza uključuje općenitu analizu projekata i deskriptivnu statistiku koja opisuje vezu između uspješnosti projekta, vrste projekta, geolokacije, aktivnosti projekta i mnogih drugih elemenata koji opisuju svaki projektni prijedlog[8], ali i predviđanje uspješnosti projekta kombinacijom složenih metoda strojnog učenja i vremenskih nizova koji opisuju razvoj projekta u prvih nekoliko sati od objavljivanja[4]. Činjenicu da statistička analiza projekata nije zanimljiva samo znanstvenicima potvrđuje i činjenica da je krajem 2014. godine lansirana prva web platforma (Krowdster - www.krowdster.co) koja služi analizi i poboljšanju projektnih prijedloga, a čija se većina funkcionalnosti bazira na vrlo opširnoj deskriptivnoj statistici dosad postojećih projekata kroz poznatije platforme.

Pokušaj predviđanja uspjeha projekta vrlo je aktualna tema u zadnje dvije godine, osim zbog znanstvenog izazova, i zbog praktične koristi. Uzimajući u obzir razne faktore koji okružuju projekt, pokazalo se da se uspješnost projekta može predvidjeti s točnošću od 68% [6]. Pokazuje se da analiza pojedinih fraza u projektnom prijedlogu te rezultate može značajno poboljšati [7]. Iako ti rezultati u komercijalnom smislu nisu dovoljno dobri, iz njih su proizašle vrlo korisne smjernice za izgradnju uspješnog projekta - poput činjenice da projekti koji kao dio opisa imaju video imaju mnogo veće šanse da postanu uspješni. Također, neki su se znanstvenici fokusirali samo na određene dijelove projekata, pa su tako pokazali da je aktivnost nosioca projekta nakon postavljanja projekta i ažuriranje istoga vrlo bitan faktor u postizanju dobrih rezultata[9].

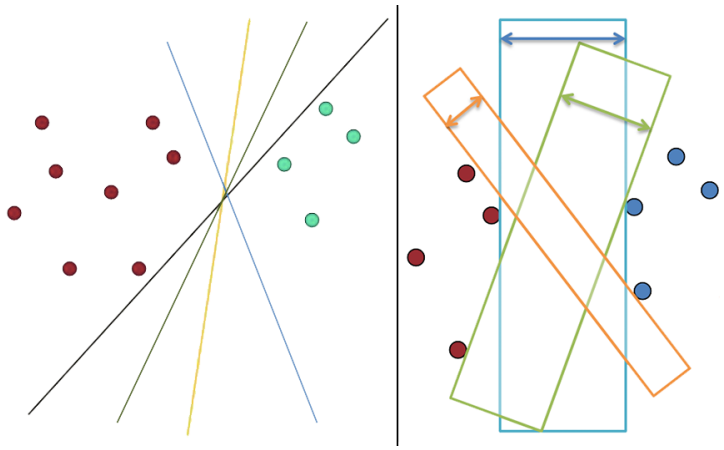
1.3 Metoda potpornih vektora

Metoda potpornih vektora je nadzirana metoda strojnog učenja koja se koristi pri klasifikacijskim problemima. Neka je S skup podataka kojem svaki podatak spada u jednu od dvije klase $c1$ i $c2$. Tada je cilj klasifikacijskog problema pronaći pravilo koje će na najbolji mogući način razdvojiti elemente skupa S na klase $c1$ i $c2$.

Linearni SVM

Prikažimo elemente skupa S u n-dimenzionalnom prostoru. Metoda potpornih vektora tada će tražiti (n-1)-dimenzionalnu hiperravninu koja će najbolje razdvajati te dvije klase.

Primjer 1.3.1. *Neka je klasifikacijski problem koji proučavamo dvodimenzionalan i separabilan. Tada je pravac koji će razdvajati klase moguće povući na više načina.*



Slika 1.2: Separabilan klasifikacijski problem

Metoda potpornih vektora pronalazi hiperravninu koja ima najveću marginu razdvajanja klasa, gdje je margina udaljenost između kritičnih točaka, odnosno točaka najbližih plohi razdvajanja (eng. decision boundary). Ideja je u tome da ako ne postoje točke blizu plohe razdvajanja, to znači da nemamo nesigurnih klasifikacijskih odluka.

Kao i kod svakog klasifikacijskog problema, trebamo definirati funkciju odluke. Funkciju odluke definirat ćemo preko podskupa primjera iz skupa za učenje, preko tzv. potpornih vektora (eng. support vectors).

Gotovo uvijek pokazat će se da problem nije potpuno linearno separabilan. Tada osnovni princip ostaje isti - razdvajajuća hiperravnina mora što bolje razdvajati klase, ali su dozvoljene greške uz penalizaciju.

Formalnije, neka su $x_i, i \in \{1, ..., n\}$ točke koje treba klasificirati i $y_i, y_i \in \{-1, 1\}$, pripadne klase i neka je naš problem linearno separabilan. Tada definiramo razdvajajuću hiperravninu pomoću normale w i pomaka b, sa $w^T x + b = 0$. Klasifikator je sada određen funkcijom $f(x_i) = sign(w^T x_i + b)$, pri čemu maksimiziramo marginu udaljenosti između kritičnih točaka. Primijetimo da se funkcija odluke vezana uz hiperravninu (w, b) ne mijenja ako hiperravninu skaliramo parametrom λ, na $(\lambda w, \lambda b)$, pri čemu je $\lambda \in \mathbb{R}^+$. Stoga parametre hiperravnine možemo skalirati tako da funkcijska margina iznosi 1. Tada vrijede dva ograničenja na skupu za učenje:

$$\begin{aligned} w^T x_i + b &\geq 1, \text{ ako je } y_i = 1 \\ w^T x_i + b &\leq 1, \text{ ako je } y_i = -1, \end{aligned} \tag{1.1}$$

gdje se ove nejednakosti pretvaraju u jednakosti u slučaju da je x_i potporni vektor. S obzirom na to da je sada udaljenost svake točke od hiperravnine jednaka

$$r = \frac{w^T x_i + b}{\|w\|}, \tag{1.2}$$

sada je margina ρ dana s:

$$\rho = \frac{2}{\|w\|}. \tag{1.3}$$

Kako bismo maksimizirali marginu, tražimo parametre w i b koji će minimizirati $\frac{1}{2}\|w\|$, još uvijek zadovoljavajući uvjete (1.1). To je kvadratni optimizacijski problem s linearnim ograničenjima te ga rješavamo pomoću Lagrangeovih multiplikatora [5].

Linearni SVM sa slabom marginom

Kao što smo već spomenuli, klasifikacijski problem često nije potpuno linearno separabilan, odnosno u podacima postoji šum. U tom slučaju linearni klasifikator mora dopustiti krivu klasifikaciju nekih varijabli i takav događaj mora penalizirati. U tu svrhu uvodimo tzv. slabu marginu (eng. soft margin), odnosno mijenjamo uvjet iz (1.1) u:

$$y_i(w^T x_i + b) \geq 1 - \xi_i, \ \xi_i \geq 0, \tag{1.4}$$

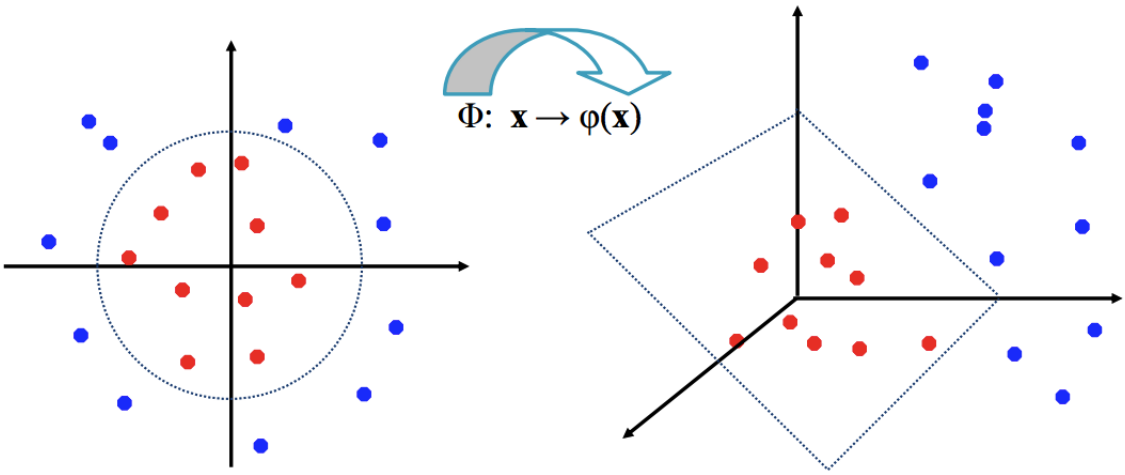
pri čemu minimiziramo funkciju

$$\frac{1}{2}\|w\| + \frac{1}{2}C \sum_{i=1}^{n} \xi_i, \tag{1.5}$$

gdje je C koji se određuje testiranjem modela na skupu za učenje (obično cross-validacijom).

Nelinearni SVM

Za probleme koji su linearno separabilni (uz razumni šum), linearni SVM (sa slabom marginom) bit će dovoljno uspješan kao klasifikator. Ipak, problemi često uopće nisu linearno separabilni. U tom slučaju uvodimo različite "jezgrene funkcije" koje originalni prostor varijabli mapiraju u novi, višedimenzionalni prostor, u kojem će naš problem biti linearno separabilan.



Slika 1.3: Mapiranje problema u višedimenzionalni prostor u kojem je on linearno separabilan

Zbog toga definiramo transformaciju $\psi : x \to \psi(x)$ kojom svaku točku mapiramo u neki novi prostor. U definiciji optimizacijskog problema, ta će se promjena očitovati na skalarnom produktu. U linearnom SVM-u je tako skalarni produkt između dva vektora bio definiran kao $K(x_i, x_j) = x_i^T x_j$, dok će skalarni produkt u nelinearnom SVM-u biti definiran kao $K(x_i, x_j) = \psi(x_i)^T \psi(x_j)$.

Jezgrena funkcija je funkcija K koja odgovara skalarnom produktu dvaju vektora u novom prostoru. Na taj smo način implicitno sve točke prebacili u novi prostor.

Najpoznatije jezgrene funkcije uključuju sljedeće:

$$\begin{aligned}
\text{linearna } K(x_i, x_j) &= x_i^T x_j \\
\text{polinomijalna } K(x_i, x_j) &= (1 + x_i^T x_j)^d \\
\text{RBF } K(x_i, x_j) &= \exp(-\frac{\|x_i - x_j\|}{2\sigma^2}) \\
\text{sigmoid } K(x_i, x_j) &= \tanh(\gamma x_i^T x_j + r)
\end{aligned}$$

Nebinarna klasifikacija

Sve verzije dosad objašnjenog algoritma klasificirale su točke u dvije klase. Kako SVM i jest metoda za binarne klasifikacijske probleme, uobičajeno je da se višeklasne probleme reducira na više binarnih klasifikacijskih problema. Ta se redukcija najčešće radi na jedan od sljedeća dva načina:

- Jedan-nasuprot-ostalih - (K binarnih klasifikatora)
- Svaki-protiv-svakog - ($\frac{K(K-1)}{2}$ binarnih klasifikatora)

U ovom ćemo radu koristiti redukciju višeklasnog problema na "svaki-protiv-svakog" binarnog.

1.4 Modeliranje tema

Modeli tema (eng. topic models) su statistički modeli koji kao cilj imaju otkrivanje apstraktnih "tema" koje se pojavljuju u skupini dokumenata. Intuitivno, ukoliko govorimo o nekoj temi, očekujemo da će se neke riječi pojavljivati češće od nekih drugih. Slično, u tekstovima se teme obično izmjenjuju i spominje se više njih odjednom i to različitom intenzivnošću. Cilj modeliranja tema je matematički opisati tu apstraktnu strukturu "teme", koju, ako je shvatimo kao podjelu "intenzivnosti" pojave riječi, možemo shvatiti kao distribuciju riječi. Istovremeno, kad izmodeliramo teme, dokument možemo opisati distribucijom tema i takav opis koristiti za uspoređivanje s drugim dokumentima.

Modeliranje tema relativno je novo područje i prva takva ideja pojavila se 1998. godine kad je grupa znanstvenika (Papadimitriou et al.) pokušala modelirati teme pomoću već poznate LSI metode (eng. latent semantic indexing), koja se dotad koristila u svrhu smanjenja dimenzionalnosti. Značajan doprinos pružio je Hofmann (1999.) modelom pLSI (eng. probabilistic LSI), na čijem su radu nastavili Blei, Ng i Jordan i 2003. predstavili latentnu Dirichletovu alokaciju (LDA - eng. latent Dirichlet allocation). Latentna Dirichletova alokacija još je uvijek najpopularnija metoda

za modeliranje tema te ćemo u ovom radu nju koristiti. U nastavku dajemo glavnu ideju i kratko objašnjenje metode[2].

Latentna Dirichletova alokacija

Osnovni pojmovi

Kako bismo se lakše snašli u generativnom modelu, definiramo osnovne pojmove:

- Riječ definiramo kao element rječnika, indeksiranog skupom $\{1, \ldots, V\}$. Riječ na v-toj poziciji u rječniku reprezentirana je vektorom w dimenzije V, takvim da je na v-tom mjestu jedinica, a na ostalim mjestima nula.
- Dokument je niz od N riječi. Označavamo ga kao $\mathbf{w} = w_1, w_2, \ldots, w_N$, gdje je w_n n-ta riječ u nizu.
- Korpus je skup M dokumenata, u oznaci $D = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_M\}$.

Model

Latentna Dirichletova alokacija (LDA) je generativni probabilistički model korpusa. Osnovna ideja je da su dokumenti reprezentirani kao slučajne kombinacije nad latentnim temama, gdje je svaka tema predstavljena kao distribucija nad riječima. LDA pretpostavlja da je svaki dokument $\mathbf{w}$ iz korpusa D nastao sljedećim generativnim procesom:

- Izaberemo $N \sim Poisson(\xi)$.
- Izaberemo $\Theta \sim Dir(\alpha)$, gdje je α k-dimenzionalni vektor.
- Za svaku od N riječi w_n:
  1. Izaberemo temu $z_n \sim Multinomial(\Theta)$
  2. Izaberemo riječ w_n s vjerojatnošću izbora $p(w_n|z_n, \beta)$.

Uvedimo nekoliko pretpostavki koje pojednostavljuju model. Za dimenziju k iz Dirichletove distribucije (a time i dimenziju teme z) pretpostavljamo da je fiksna i poznata. Nadalje, vjerojatnosti izbora riječi dane su $k \times V$ matricom β, gdje je $\beta_{ij} = p(w^j = 1|z^i = 1)$. Tu matricu smatramo fiksnom i pokušavamo je što točnije procijeniti. Također, pretpostavka Poissonove distribucije nije nužna ni za jedan dio procesa koji slijedi te se po potrebi mogu koristiti i drugačije distribucije ili čak neslučajne duljine dokumenata.

k-dimenzionalna Dirichletova slučajna varijabla Θ može poprimiti vrijednosti u $(k-1)$-dimenzionalnom simpleksu[1] i ima sljedeću funkciju gustoće na tom simpleksu:

$$p(\Theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \Theta_1^{\alpha_1 - 1} \cdots \Theta_k^{\alpha_k - 1}, \tag{1.6}$$

gdje je parametar α k-dimenzionalan vektor tako da vrijedi $\alpha_i > 0, \forall i$, a $\Gamma(.)$ gama funkcija. Dirichletova distribucija ima dobra svojstva na simpleksu; to je distribucija iz obitelji eksponencijalnih, ima konačno-dimenzionalnu dovoljnu statistiku te je konjugirana[2] multinomijalnoj distribuciji.

Uz dane parametre α i β, zajednička distribucija kombinacije tema Θ, skupa od N tema $\mathbf{z}$ i skupa od N riječi $\mathbf{w}$ dana je sa:

$$p(\Theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) = p(z_n|\Theta) \prod_{n=1}^{N} p(z_n|\Theta) p(w_n|z_n, \beta), \tag{1.7}$$

gdje je $p(z_n|\Theta)$ upravo Θ_i za jedinstven i tako da je $z_n^i = 1$. Integriranjem po Θ i sumiranjem po z dobije se marginalna distribucija dokumenta:

$$p(\mathbf{w}|\alpha, \beta) = \int p(\Theta|\alpha) \left(\prod_{n=1}^{N} \sum_{z_n} p(z_n|\Theta) p(w_n|z_n, \beta) \right) d\Theta. \tag{1.8}$$

Gornji izraz nije moguće egzaktno izračunati, ali ga možemo aproksimirati pomoću mnogih algoritama, kao što su Laplaceova aproksimacija, varijacijska aproksimacija i Monte Carlo Markovljevi lanci.[2]

Konačno, množenjem marginalnih distribucija pojedinih dokumenata, dobivamo vjerojatnost korpusa:

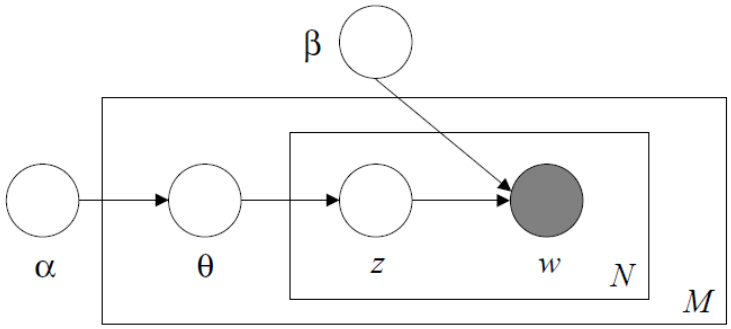
$$p(D|\alpha, \beta) = \prod_{d=1}^{M} \int p(\Theta_d|\alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn}|\Theta_d) p(w_{dn}|z_{dn}, \beta) \right) d\Theta_d. \tag{1.9}$$

Ideju LDA modela možemo prikazati i grafički, što predstavljamo na slici 1.4. Kao što se vidi sa slike, LDA se sastoji od 3 razine. Parametri α i β su parametri na razini

[1] za k-dimenzionalni vektor Θ kažemo da leži u $(k-1)$-dimenzionalnom simpleksu ako $\Theta_i \geq 0$, $\sum_{i=1}^{k} \Theta_i = 1$

[2] Ako je distribucija a posteriori $P(\theta|x)$ u istoj familiji distribucija kao i distribucija a priori $P(\theta)$, tada kažemo da su te dvije distribucije konjugirane.

korpusa za koje pretpostavljamo da se biraju jednom u procesu generiranja korpusa. Varijable Θ_d su varijable na razini dokumenta koje biramo jednom po dokumentu, a opisuju distribuciju tema u dokumentu. Napokon, varijable z_{dn} i w_{dn} su varijable na razini riječi i one se biraju jednom za svaku riječ u svakom dokumentu, a predstavljaju izbor teme kojoj riječ pripada, a zatim izbor riječi u ovisnosti o temi i parametru β.



Slika 1.4: Grafička reprezentacija LDA. Pravokutnici predstavljaju ponavljanje; vanjski pravokutnik predstavlja M dokumenata, dok unutarnji pravokutnik predstavlja N izbora tema i riječi unutar dokumenata.

Opravdanje modela - LDA i izmjenjivost varijabli

Definicija 1.4.1. *Za konačan skup slučajnih varijabli $\{z_1, \ldots, z_N\}$ kažemo da je izmjenjiv ako je zajednička distribucija invarijantna na permutacije. Drugim riječima, ako je π permutacija prirodnih brojeva od 1 do N, tada je*

$$p(z_1, \ldots, z_N) = p(z_{\pi(1)}, \ldots, z_{\pi(1)}).$$

Definicija 1.4.2. *Beskonačan niz slučajnih varijabli je beskonačno izmjenjiv ako je svaki njegov konačan podniz izmjenjiv.*

De Finnetijev teorem o reprezentaciji kaže da je zajednička distribucija beskonačno izmjenjivog niza slučajnih varijabli ekvivalentna nasumičnom izvlačenju parametra iz neke distribucije. Tada su te slučajne varijable nezavisne i jednako distribuirane uvjetno na odabrani parametar.[2]

U kategorizaciji teksta smo, koristeći tzv. *Bag of words* metodu za izbor značajki, zapravo zanemarili poredak riječi u dokumentu. Time smo implicitno pretpostavili da su riječi u dokumentu izmjenjive. Slično, u LDA-u pretpostavljamo da su riječi generirane temama (fiksiranim uvjetnim distribucijama) i da su te teme beskonačno

izmjenjive unutar dokumenta. Po de Finettijevom teoremu, vjerojatnost niza riječi i tema mora imati oblik:

$$p(\mathbf{w}, \mathbf{z}) = \int p(\Theta) \left(\prod_{n=1}^{N} p(z_n|\Theta) p(w_n|z_n) \right) d\Theta,$$

gdje je Θ parametar distribucije nad temama. LDA distribuciju nad dokumentima iz jednadžbe 1.8 dobijemo marginalizacijom varijabli tema i odabirom Dirichletove distribucije za Θ.

Validacija modela

Dokumenti u korpusu nisu ni na koji način označeni, a model uči na nenadzirani način. Stoga moramo definirati način validacije modela. Želimo postići visoku vjerodostojnost modela na skupu za testiranje. Perpleksitet (eng. perplexity) uzorka jednaka je inverzu geometrijske sredine vjerodostojnosti po riječima te monotono pada s obzirom na vjerodostojnost modela na skupu za testiranje. Niži perpleksitet pokazuje da se model bolje ponaša kod generalizacije. Formalno, perpleksitet definiramo kao:

$$perpleksitet(D_{test}) = exp\left\{ - \frac{\sum_{d=1}^{M} log(p(\mathbf{w}_d))}{\sum_{d=1}^{M} N_d} \right\}.$$

Tu ćemo mjeru koristiti za validaciju modela, kao i za određivanje optimalnog broja tema. S obzirom na to da će veći izbor tema obično značiti bolju generalizaciju, povećanjem broja tema očekuje se monotoni pad perpleksiteta. Zbog toga ćemo kao zaustavni kriterij kod izbora broja tema koristiti malenu razliku u promjeni perpleksiteta.

Poglavlje 2

Analiza tekstova projektnih prijedloga

Analiza tekstova projektnih prijedloga izvršena je nad svim projektima s Kickstartera (www.kickstarter.com) nastalima do veljače 2015. godine i uspoređena s projektima objavljenima na Indiegogou (www.indiegogo.com) do veljače 2015. godine. U ovom ćemo poglavlju detaljno opisati proces prikupljanja podataka te svaki dio analize podataka zasebno, kao i izazove s kojima smo se u tom procesu susreli.

2.1 Skupljanje podataka

Prikupljanje podataka implementirano je u programskom jeziku PHP. Skripta je prvo sa glavne stranice platforme (Kickstarter ili Indiegogo) prikupila url-ove svih do tog trenutka objavljenih projekata i spremila ih u MySQL bazu podataka. Zatim su pomoću baze url-ova prikupljene čitave stranice projekata i iz njih izvučeni bitni podaci za svaki projekt - naslov, kategorija, projektni prijedlog, datum početka i uspješnost. Prikupljeni podaci tada su spremljeni u MySQL bazu podataka, u tablicu u kojoj svaki redak predstavlja jedan projekt s navedenim poljima.

2.2 Predprocesiranje

Kako bismo dobili što bolje rezultate, potrebno je obraditi tekst da bismo iz njega dobili što jasniju informaciju. U tu svrhu, osim trivijalnih metoda predprocesiranja, poput pretvaranja svih slova u mala slova (kako bismo se osigurali da računalo zna da su riječi "Play" i "play" zapravo ista riječ) i odstranjivanja interpunkcije,

radimo i neke sofisticiranije koje ćemo objasniti u ovom poglavlju. Predprocesiranje implementirano je u programskom jeziku Python.

Eliminacija stop-riječi

Stop riječi su riječi koje nemaju bitno semantičko značenje, a često se pojavljuju u jeziku. Takve se riječi obično izbacuju iz teksta, jer nisu "korisne", a zbog čestog pojavljivanja mogu smetati. Naime, u klasifikacijskom problemu one ne nose informaciju niti o jednoj kategoriji, a zbog česte pojave mogu "zagušiti" klasifikator. U problemu modeliranja tema će se pak pojaviti kao najčešće riječi u vrlo velikom broju tema, što ne želimo zbog njihovog malog (nepostojećeg?) semantičkog značenja. U engleskom jeziku to su riječi poput *and, or, the, a....* U ovom je radu korišten popis engleskih stop-riječi ugrađen u paket *sklearn* (http://scikit-learn.org/) za Python.

Svođenje na korijen riječi

U tekstu se vrlo često ista riječ pojavljuje u različitim oblicima, kao npr. *playing, played, play.* S obzirom na to da su semantička značenja svih tih oblika riječi jednaka, bilo bi dobro na neki način računalu to dati do znanja. Zbog toga sve riječi svodimo na njihov korijenski oblik. Osim što smo time osigurali da računalo te riječi smatra jednakima, pa će ono "znati" da i tekst koji sadrži riječ "playing" i onaj koji sadrži riječ "played" govore o istoj stvari, time smo i značajno smanjili prostor riječi, pa će računanje biti mnogo brže i zahtijevati manje radne memorije. S obzirom na to da svođenje na korijen riječi nije uvijek ovako jednostavno, u tu se svrhu koriste napredni, već razvijeni paketi. U ovom je radu korišten *Snowball stemmer* (https://pypi.python.org/pypi/snowballstemmer) koji je sadržan u paketu *NLTK* za Python.

2.3 Kategorizacija

Projektni prijedlozi na Kickstarteru podijeljeni su u 15 različitih kategorija. Raspodjelu projekata po kategorijama za platformu Kickstarter možemo vidjeti u tablici 2.1. Na tim ćemo projektima naučiti klasifikator kako bismo projekte s Indiegogo-a mogli rasporediti u iste kategorije i zatim uspoređivati ove dvije platforme.

Već je samo pogledom na tablicu 2.1 jasno da bismo pri klasifikaciji mogli imati problema s nejednako raspoređenim kategorijama. Naime, uzimamo li za ocjenu klasifikatora samo ukupnu točnost procjene na skupu za testiranje, klasifikator koji sve projekte klasificira u jednu od pet najbrojnijih kategorija može postići točnost preko 64%, što za klasifikator na 15 klasa uopće nije loše.

Tablica 2.1: Raspodjela projekata po kategorijama

Kategorija	**Broj projekata**
Art	14993
Comics	4881
Crafts	3104
Dance	2335
Design	11367
Fashion	9153
Film & Video	39616
Food	11670
Games	13330
Journalism	1871
Music	33694
Photography	6122
Publishing	20946
Technology	10283
Theater	7022
Ukupno	**190387**

U svrhu ovog rada napravljena je klasifikacija pomoću Naive Bayes klasifikatora[5] i pomoću SVM klasifikatora. Osim što su, gledajući samo ukupnu točnost procjene na skupu za testiranje, rezultati s Naive Bayesom bili značajno lošiji (otprilike 30% točnosti, naspram 80% točnosti), Naive Bayes je u ovom primjeru bio i puno osjetljiviji na nejednake veličine kategorija te je većinu projektnih prijedloga klasificirao u jednu od zastupljenijih kategorija, dok je manje zastupljene kategorije ostavio gotovo praznima. S obzirom na to da je skup podataka vrlo velik, taj se problem mogao riješiti restrikcijom skupa za treniranje na skup u kojem su sve kategorije podjednako brojne (brojnosti malo manje od najmanje klase). Ipak, iako je taj pristup riješio problem preferiranja određenih klasa, sveukupna točnost algoritma se dodatno smanjila. Zbog tako velike razlike u uspješnosti algoritama, kao klasifikator smo izabrali SVM i rezultate Naive Bayesa nećemo prezentirati u daljnjem tekstu.

Izbor značajki

Kao značajke koje opisuju svaki dokument korištene su riječi koje su pojavljuju u dokumentu. Značajke su promatrane nezavisno, odnosno korištena je tzv. *bag of words* metoda.

Kako je broj svih riječi koje se pojavljuju u tekstu potencijalno vrlo velik, zbog

brzine računanja, ali i zbog uklanjanja šuma, treba razmotriti ideju biranja najboljih značajki.

Česte riječi bez najčešćih

Najjednostavnija ideja koja nam pada na pamet jest određivanje nekog broja n i izbor najfrekventnijih n riječi za značajke. Naime, s obzirom na to da se one pojavljuju puno češće od ostalih riječi, bit će češće prisutne u dokumentima i na taj način češće imati utjecaj na izbor klase. Iako je taj način razmišljanja donekle točan, kod takvog načina biranja riječi javlja se sličan problem kao i kod stop-riječi. Naime, unatoč činjenici da smo izbacili stop-riječi iz korpusa, i dalje postoje stop-riječi definirane korpusom (eng. corpus-specific stop words). Ovdje se radi o riječima koje su vrlo uobičajene za naš korpus te će se zbog toga pojaviti u vrlo velikom postotku dokumenata, bez obzira na kategoriju dokumenta. U našem korpusu to su riječi poput *project, fund, reward....* Da bismo izbacili takve riječi, izabrat ćemo n najčešćih, ali ćemo izabrati i parametar p i te riječi vaditi iz skupa riječi koje se pojavljuju u manje od $p\%$ dokumenata.

χ^2 test za izbor značajki

χ^2 test testira nezavisnost dvaju događaja. Napravimo li χ^2 test za događaje $A_i = \{\text{dokument sadrži i-tu riječ}\}$ i $B_k = \{\text{dokument pripada k-toj kategoriji}\}$ dobit ćemo informaciju o tome jesu li oni zavisni. Sada je jasno da su riječi koje su zavisne s nekom od kategorija potencijalno mnogo bolje značajke od onih koje to nisu. Iako bi uz ovakav izbor značajki računanje moglo trajati osjetno duže, implementacija testa u paketu *sklearn* za Python omogućava izbor značajki na ovakav način iz skupa dokumenata reprezentiranog rijetkom matricom, što čini tu implementaciju mnogo bržom i manje zahtjevnom za radnu memoriju od ostalih.

Transformacija značajki

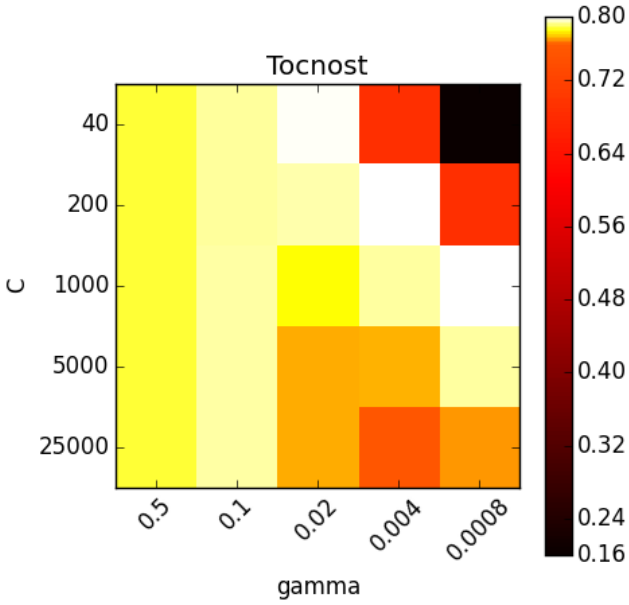
U prethodnom smo koraku izabrali najbolje značajke. Ipak, među njima nisu sve jednako diskriminativne, a ni na koji način nismo definirali koje su značajke "bitnije" od drugih.

Tf-Idf transformacija značajki (eng. Term frequency, inverse document-frequency) smanjuje utjecaj značajki koje se u dokumentima pojavljuju često, u korist onih koje se pojavljuju u manjem broju dokumenata. Ta transformacija, implementirana u paketu *sklearn* za Python, svakoj značajki svakog dokumenta pridružuje težinu koja raste frekvencijom pojave te riječi u dokumentu, ali pada frekvencijom pojave

riječi kroz cijeli korpus. Pokazuje se da upotreba ove transformacije donosi vidljivo poboljšanje rezultata.

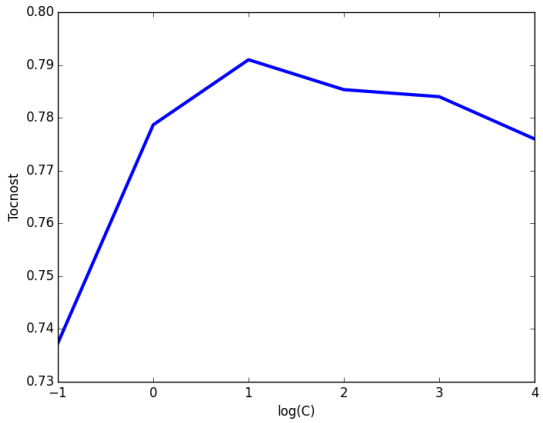
Izbor jezgre i optimalnih parametara

Za klasifikaciju teksta obično se uzima SVM s linearnom ili RBF jezgrom. Kako bismo odabrali što bolji model za naš problem, za svaku od ovih jezgara moramo pronaći optimalne parametre. U slučaju linearne jezgre optimiziramo parametar C, a u slučaju RBF jezgre, radi se o parametrima C i σ. Optimalne smo parametre tražili pretragom po rešetki (eng. grid search) uz cross-validaciju na smanjenom skupu za treniranje. Drugim riječima, za svaku točku u rešetki vrijednosti parametara (jednodimenzionalnu, odnosno dvodimenzionalnu, ovisno o jezgri) istreniramo model na dijelu skupa za treniranje i provjerimo njegovu točnost na ostatku. S obzirom na to da se skup za treniranje u ovu svrhu najčešće svede na mnogo manji od početnoga, na njemu radimo cross-validaciju, odnosno podijelimo ga na k dijelova i treniramo k puta, testirajući svaki put na k-tom skupu, a trenirajući na skupu za treniranje bez k-tog dijela.



Slika 2.1: Točnost klasifikatora s RBF jezgrom s obzirom na izbor parametara

Rezultate pretrage po rešetki možemo prikazati grafički kao na slikama 2.1 i 2.2.



Slika 2.2: Točnost klasifikatora s linearnom jezgrom s obzirom na izbor parametara

Tablica 2.2: Točnosti klasifikatora s obzirom na izbor jezgre i parametara

		RBF jezgra, gamma					**Linearna jezgra**	**C**
		0.5	**0.1**	**0.02**	**0.004**	**0.0008**	0.737	**0.5**
C	**40**	0.778	0.784	0.789	0.689	0.105	0.779	**1**
	200	0.778	0.784	0.785	0.79	0.688	0.791	**2**
	1000	0.778	0.784	0.775	0.784	0.79	0.785	**4**
	5000	0.778	0.784	0.767	0.768	0.784	0.784	**8**
	25000	0.778	0.784	0.767	0.751	0.765	0.776	**16**

Pretraga je izvršena na 3000 podataka pomoću paketa *sklearn* i funkcije *GridSearchCV*, uz trostruku cross-validaciju (eng. 3-fold crossvalidation). Razlog zbog kojeg je skup podataka za treniranje tako smanjen je trajanje izvršavanja. Naime, jasno je da se radi o procesorski zahtjevnom problemu, a već je i za ovaj broj podataka pretraga po rešetci za RBF jezgru, izvršavajući se paralelno na 4 jezgre, trajala preko 30 minuta.

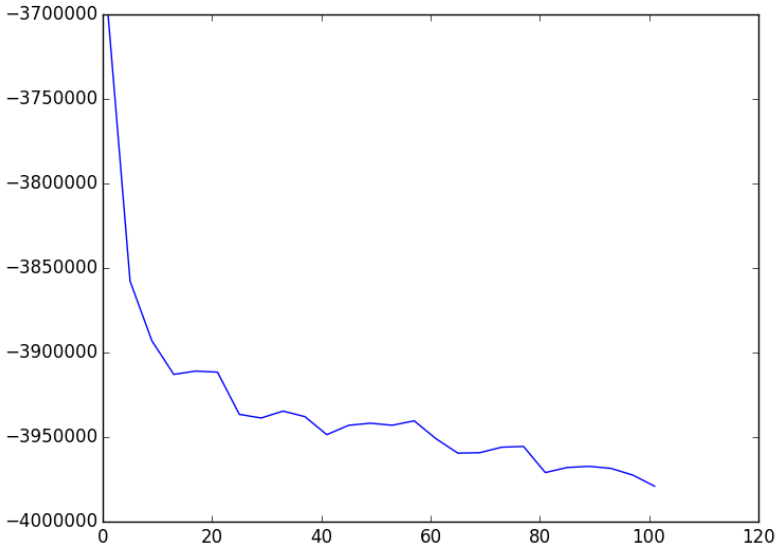
Osim grafičke vizualizacije, rezultate možemo prikazati i tablicom iz koje možemo isčitati točne numeričke vrijednosti. U tablici 2.2 vrijednosti su zaokružene na tri decimalna mjesta, a crvenom bojom označene su najbolje vrijednosti za svaku od jezgara. Vidimo da su najbolji postignuti za obje jezgre slični, a radi se o rezultatima s parametrima $C = 200$ i $\gamma = 0.004$ za klasifikator s RBF jezgrom te $C = 2$ za nešto bolji klasifikator s linearnom jezgrom. U nastavku ćemo trenirati SVM s linearnom jezgrom i parametrom $C = 2$.

2.4 Modeliranje tema

Iako su projektni prijedlozi raspoređeni po kategorijama, jedna kategorija za svaki projekt ne može u potpunosti opisati čime se projekt bavi. Iz toga razloga, s ciljem boljeg opisivanja projekata, modeliramo teme koje se protežu dokumentima i pomoću njih opisujemo projekte. Osim činjenice da će tema biti više nego kategorija, značajan doprinos u reprezentaciji dokumenta temama daje i činjenica da ne moramo svakom dokumentu pridružiti točno jednu temu, već mu pridružujemo kombinaciju tema i njihove zastupljenosti. Modeliranje tema implementirano je pomoću paketa *lda* za Python.

Izbor broja tema

Generativni latentni Dirichletov model kod biranja riječi koja će doći na n-to mjesto u dokument, prvo izabire jednu od k tema. Kako bismo model prilagodili podacima, moramo odrediti koji je broj tema iz kojeg se riječi mogu birati. U tu ćemo svrhu smanjenom skupu podataka prilagoditi više modela s različitim brojem tema, odnosno ponovno ćemo napraviti jednodimenzionalno pretraživanje po rešetci. Slika 2.3 prikazuje vrijednosti perpleksiteta za različite vrijednosti parametra k. Modeli su trenirani na 2000 dokumenata, za sve brojeve tema između 1 i 102 s korakom 4.



Slika 2.3: Perpleksitet modela u odnosu na broj tema

Ovakva nam je pretraga dala okvirnu ideju što bi mogao biti dobar broj tema. Naime, smatra se da je broj tema dobar ukoliko razlika između perpleksiteta modela za susjedne brojeve tema nije velika. Ipak, povećavanjem skupa podataka mogao bi se povećati i broj tema. Zato ponavljamo pretragu po rešetci, ovaj put na mnogo manjem rasponu tema (25 - 50), na većem skupu za treniranje - 10000 projekata. a kao uvjet zaustavljanja biramo kriterij sličan onome u članku [2], da razlika između perplextyja za dva broja tema bude manja od 0.001%

Konačno, radit ćemo model sa 40 tema. Model ćemo prilagoditi projektima s obje platforme kako bismo mogli usporediti teme između njih, a dobivene teme, kao i rezultate usporedbi, prikazat ćemo u sljedećem poglavlju.

Poglavlje 3

Rezultati

Linearni SVM klasifikator s parametrom $C = 2$ istreniran je na 40000 projekata s Kickstartera te potom testiran na 10000 projekata. Izbačene su predefinirane stopriječi i riječi s frekvencijom pojavljivanja većom od 80% te je zatim χ^2 testom izabrano 8000 najboljih značajki. Te su značajke transformirane Tf-Idf transformacijom te je tada na njima treniran model. Matricu konfuzije rezultata na skupu za testiranje možemo vidjeti na slici 3.1

	art	comics	crafts	dance	design	fashion	film & video	food	games	journalism	music	photography	publishing	technology	theater
art	<545>	12	40	9	32	12	17	13	8	3	17	34	42	18	33
comics	14	<216>	1	.	.	.	6	1	1	.	1	.	18	1	.
crafts	14	.	<93>	.	19	9	.	3	2	.	.	1	3	6	.
dance	3	.	.	<98>	1	1	4	1	.	.	1	.	1	.	6
design	39	1	28	.	<368>	50	3	14	17	1	1	7	10	49	1
fashion	16	1	22	1	45	<373>	2	3	1	2	.	1	2	8	1
film & video	32	8	2	17	4	4	<1848>	21	20	15	44	23	35	12	18
food	11	.	7	.	7	3	3	<553>	1	4	.	.	4	8	.
games	9	4	6	1	17	1	2	4	<667>	2	3	.	11	23	.
journalism	9	1	.	.	.	2	11	2	.	<33>	.	10	26	4	.
music	35	1	1	4	6	6	25	4	1	.	<1615>	.	8	12	6
photography	23	1	1	.	1	2	5	1	.	4	1	<240>	19	1	.
publishing	26	17	8	2	8	4	21	12	13	41	14	37	<883>	22	5
technology	14	.	6	.	59	7	6	7	14	4	4	7	6	<383>	1
theater	22	1	.	7	2	1	13	1	1	1	12	.	6	.	<300>

(row = reference; col = test)

Slika 3.1: Matrica konfuzije rezultata na skupu za testiranje

Vrlo su slični rezultati ($82,15\%$ točnosti naprama $82,01\%$ točnosti) dobiveni i s izborom 8000 najčešćih značajki bez riječi s frekvencijom pojavljivanja većom od 60%.

Analizu točnosti po kategorijama možemo vidjeti na slici 3.2. Prvi stupac predstavlja preciznost klasifikatora za svaku od kategorija, drugi osjetljivost, a treći vrijednost $f1$, odnosno harmonijsku sredinu preciznosti i osjetljivosti. Četvrti stupac prikazuje broj dokumenata po kategoriji u skupu za testiranje. Kao što smo mogli i očekivati, primjećujemo da su dokumenti u rjeđim kategorijama, poput *journalism*, češće krivo klasificirani od onih u zastupljenijim kategorijama, poput *Film & video* ili *Music*. Razlog tome je i činjenica da je klasifikator za zastupljenije kategorije imao mnogo više dokumenata u skupu za treniranje.

	precision	recall	f1-score	support
art	0.67	0.65	0.66	835
comics	0.82	0.83	0.83	259
crafts	0.43	0.62	0.51	150
dance	0.71	0.84	0.77	116
design	0.65	0.62	0.64	589
fashion	0.79	0.78	0.78	478
film & video	0.94	0.88	0.91	2103
food	0.86	0.92	0.89	601
games	0.89	0.89	0.89	750
journalism	0.30	0.34	0.32	98
music	0.94	0.94	0.94	1724
photography	0.67	0.80	0.73	299
publishing	0.82	0.79	0.81	1113
technology	0.70	0.74	0.72	518
theater	0.81	0.82	0.81	367
avg / total	0.83	0.82	0.82	10000

Slika 3.2: Analiza rezultata na skupu za testiranje

Kako bismo omogućili uspoređivanje platformi, klasificirat ćemo projekte s platforme Indiegogo u već navedene kategorije, one s platforme Kickstarter. Raspodjelu projekata po prilagođenim kategorijama prikazujemo tablicom 3.1. Iako je *Film & video* i dalje najbrojnija kategorija, možemo primijetiti da je raspored projekata na ovoj platformi znatno drugačiji. Tako je kategorija *Journalism* s najmanje zastupljene kategorije na Kickstarteru sa samo 1871 projekata, prešla u jednu od zastupljenijih kategorija na Indiegogo-u, sa brojem projekata većim za čitav red veličine. Porast popularnosti vidimo i u kategoriji *Art*, dok kategorija *Comics* na Indiegogo-u ima manje projekata nego na Kickstarteru unatoč mnogo većem broju projekata sveukupno.

Ove razlike mogu se objasniti nešto drugačijom prirodom platformi. Npr., na Indiegogo-u je jedna od najpopularnijih kategorija *Community*, čiji ekvivalent na

Tablica 3.1: Raspodjela projekata po kategorijama za Indiegogo

Kategorija	**Broj projekata**
Art	43795
Comics	2080
Crafts	10306
Dance	3483
Design	14544
Fashion	12197
Film & Video	64106
Food	19514
Games	9683
Journalism	14767
Music	23384
Photography	9231
Publishing	24981
Technology	41102
Theater	10958
Ukupno	**304131**

Kickstarteru ne postoji. Radi se o projektima koji na neki način doprinose zajednici, a često imaju veze s umjetnošću ili novinarstvom. Ove ćemo razlike pokušati bolje objasniti raspodjelom tema po kategorijama.

Promotrimo sada teme dobivene modelom Latentne Dirichletove alokacije s unaprijed određenim brojem tema - $k = 50$. Teme su modelirane nad 20000 projekata s platforme Kickstarter i 20000 projekata s platforme Indiegogo. Kao što je komentirano ranije, teme smo modelirali nad projektima s obje platforme kako bismo mogli usporediti frekvenciju pojavljivanja tema između platformi.

U tablici 3.2 možemo vidjeti popis glavnih riječi po temama. Podsjetimo, svaka tema predstavljena je kao distribucija riječi, a u ovoj tablici možemo vidjeti najfrekventnije riječi unutar odgovarajuće teme, odnosno po 9 riječi koje je najbolje određuju. Prisjećamo, riječi su u predprocesiranju svedene na korijenski oblik. Osim nekih tema, poput teme broj 2 za koju je jasno da govori o glazbi ili teme broj 25 za koju je jasno da govori o razvoju novih programa i aplikacija, postoje i teme koje nisu u potpunosti jasne. Tema broj 3 govori nam da, unatoč tome što postoje posebna polja rezervirana samo za to, korisnici vrlo često unutar teksta projektnog prijedloga stavljaju web stranice vezane uz svoj projekt. Slično, teme 13 i 33 govori nam da postoje projekti koji nisu isključivo na engleskom jeziku (već i na španjolskom, odnosno francuskom). Iako su projekti na stranim jezicima izbačeni iz skupa za učenje,

Tablica 3.2: Glavne riječi u temama

Tema 0	want make just like peopl time know ve thing
Tema 1	pledg reward print kickstart goal card ship backer includ
Tema 2	music record album song band project studio cd releas
Tema 3	com www http 00 amp gt 000 watch 10
Tema 4	travel trip world countri peopl experi tour bike citi
Tema 5	project photograph imag photo work histori photographi new time
Tema 6	event communiti citi year donat local support rais fund
Tema 7	school student educ program learn help year colleg class
Tema 8	book stori publish comic print page write read cover
Tema 9	project develop social communiti creat peopl world support media
Tema 10	life love help world peopl dream live make way
Tema 11	help need money work year time pay famili job
Tema 12	film festiv produc product documentari movi stori short filmmak
Tema 13	la en el que los para las del es
Tema 14	film product crew make project cast stori need help
Tema 15	build hous home space room place open wall need
Tema 16	food make product coffe beer local cook kitchen flavor
Tema 17	help children communiti need provid support donat organ live
Tema 18	world stori man dark zombi charact dead horror time
Tema 19	use product water project energi power cost develop need
Tema 20	game play player card charact level new set goal
Tema 21	busi market compani help need product start fund 000
Tema 22	art artist paint work creat print piec project make
Tema 23	help peopl need contribut make fund amp campaign way
Tema 24	new work amp produc play director year includ john
Tema 25	use app develop user design softwar creat devic comput
Tema 26	perform art artist danc work new communiti music theatr
Tema 27	women stori peopl american war state right nation countri
Tema 28	design shirt product fashion cloth line brand collect color
Tema 29	farm garden natur plant water land local project communiti
Tema 30	famili children kid help dog parent child love year
Tema 31	help donat com facebook thank make need campaign money
Tema 32	video seri episod product tv anim project produc youtub
Tema 33	la et di le il en les pour des
Tema 34	design product use prototyp manufactur case make light need

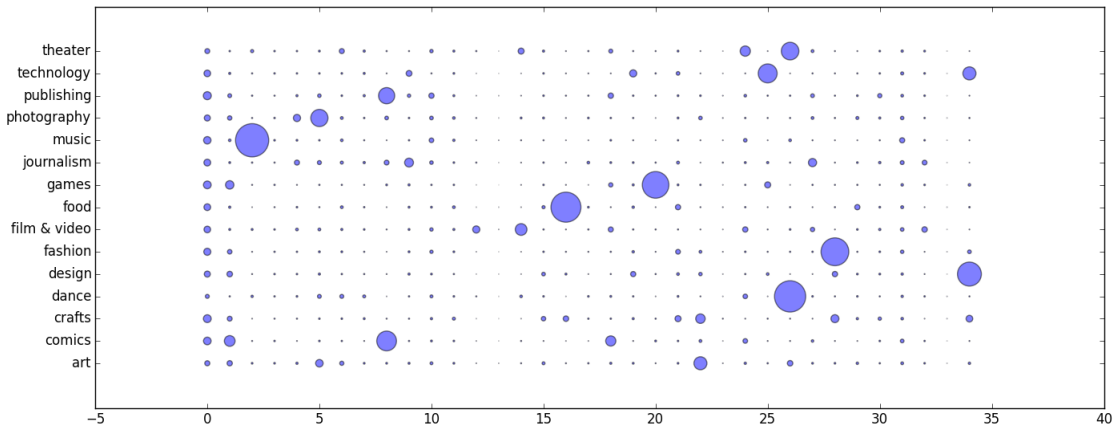
Tablica 3.3: Glavne teme nekih projektnih prijedloga

The Procurator	Tema 20
MWR Collection - Startup Fundraiser - PART 1	Tema 28
Food (Fun)damentals - creativity through animal free cooking	Tema 32
Exploring Visual Documentation for Social Uprising Spaces	Tema 27
Ten-Thousand Miles: The Undiluted Memoirs	Tema 5

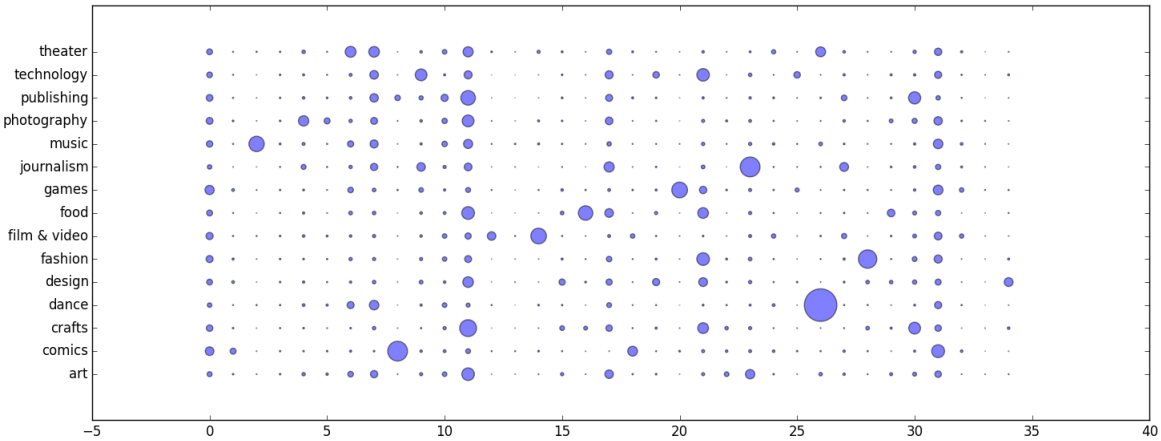
relativno je česta pojava da projektni prijedlozi imaju samo dio teksta napisan na stranom jeziku. Takvi bi tekstovi mogli vrlo nepovoljno utjecati na kvalitetu klasifikatora. Ipak, takvi projekti većinom dolaze s Indiegogo-a, pa oni nisu činili šum pri učenju klasifikatora. Iako je zanimljivo vidjeti da su neki strani jezici toliko zastupljeni na stranici da su formirali svoje teme, jedan od koraka u daljnjoj analizi svakako bi bio izbaciti sve takve riječi prije ponavljanja svih postupaka.

Kao primjer, u tablici 3.3 možemo vidjeti naslove nekih projektnih prijedloga i glavnu temu koja se u njima pojavljuje. Možemo vidjeti da je *The procurator* vjerojatno naslov neke kartaške igre, dok je *MWR Collection - Startup Fundraiser - PART 1* vjerojatno projekt vezan uz modu. Prikazivanjem više tema koje opisuju projekt, dobili bismo jasniju sliku projekta, ali zbog preglednosti zapisa, takvu analizu ostavljamo daljnjem radu.

Konačno, promotrimo distribuciju tema po kategorijama za svaku kategoriju zasebno i pogledajmo jesu li slične. Distribuciju tema po kategorijama aproksimirat ćemo računajući prosječnu distribuciju tema svih projekata koji pripadaju istoj kategoriji. Na slikama 3.3 i 3.4 veličina svakog kružića proporcionalna je zastupljenosti odgovarajuće teme u odgovarajućoj kategoriji. Osim očekivanih stvari, poput velike zastupljenosti teme broj 2 u kategoriji *Music* ili teme broj 26 u kategoriji *Dance*, možemo vidjeti da je tema broj 1 mnogo zastupljenija na platformi *Kickstarter*, jer koristi općeniti rječnik koji se na toj platformi veže uz projekte. Ista tako, možemo vidjeti da je tema broj 11 mnogo zastupljenija na platformi *Indiegogo*, jer ta platforma ima nešto humanitarniji pristup, kao i mogućnost postavljanja društveno korisnih projekata koji pomažu zajednici. S obzirom na to da takvi projekti mogu biti od umjetničkih, preko objavljivanja priča ili tekstova, pa sve do akcija poput prodavanja kolača u humanitarne svrhe, takvi su se projekti na Indiegogo-u rasporedili kroz sve kategorije.



Slika 3.3: Distribucija tema po kategorijama na Kickstarteru



Slika 3.4: Distribucija tema po kategorijama na Indiegogo-u

Bibliografija

[1] Paul Belleflamme, Thomas Lambert i Armin Schwienbacher, *Crowdfunding: Tapping the right crowd*, Journal of Business Venturing **29** (2014), br. 5, 585–609.

[2] David M Blei, Andrew Y Ng i Michael I Jordan, *Latent dirichlet allocation*, the Journal of machine Learning research **3** (2003), 993–1022.

[3] C Steven Bradford, *Crowdfunding and the federal securities laws*, Columbia Business Law Review **2012** (2012), br. 1.

[4] Vincent Etter, Matthias Grossglauser i Patrick Thiran, *Launch hard or go home!: predicting the success of kickstarter campaigns*, Proceedings of the first ACM conference on Online social networks, ACM, 2013, str. 177–182.

[5] Jerome Friedman, Trevor Hastie i Robert Tibshirani, *The elements of statistical learning*, sv. 1, Springer series in statistics Springer, Berlin, 2001.

[6] Michael D Greenberg, Bryan Pardo, Karthic Hariharan i Elizabeth Gerber, *Crowdfunding support tools: predicting success & failure*, CHI'13 Extended Abstracts on Human Factors in Computing Systems, ACM, 2013, str. 1815–1820.

[7] Tanushree Mitra i Eric Gilbert, *The language that gets people to give: Phrases that predict success on kickstarter*, Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing, ACM, 2014, str. 49–61.

[8] Ethan Mollick, *The dynamics of crowdfunding: An exploratory study*, Journal of Business Venturing **29** (2014), br. 1, 1–16.

[9] Anbang Xu, Xiao Yang, Huaming Rao, Wai Tat Fu, Shih Wen Huang i Brian P Bailey, *Show me the money!: An analysis of project updates during crowdfunding campaigns*, Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, 2014, str. 591–600.

Sažetak

U ovom je radu napravljena analiza projektnih prijedloga metodama obrade prirodnog jezika. Projektni prijedlozi preuzeti su s dvije najpoznatije platforme za grupno financiranje - Kickstarter i Indiegogo, a postupak prikupljanja podataka ukratko je opisan u radu. Nakon toga, objašnjen je proces pretvorbe tekstova dohvaćenih direktno s internetskih platformi u oblik pogodan za analizu - tzv. predprocesiranje.

Analiza napravljena u ovom radu sastoji od dvije smislene cjeline. U prvoj je cjelini napravljena klasifikacija projektnih prijedloga na odgovarajuće kategorije. Dana je motivacija za taj problem i opisana metoda potpornih vektora, koja se u ovakvim problemima pokazuje vrlo uspješnom. Nakon toga je detaljno opisan proces izbora najboljih značajki za klasifikaciju, kao i izbora najboljih parametara za metodu. Na poslijetku je klasifikator, istreniran na velikom broju projekata s Kickstartera, podijelio projekte s Indiegogo-a u odgovarajuće kategorije. To će omogućiti daljnju analizu i usporedbi dviju platformi.

U drugoj je cjelini napravljeno modeliranje tema projektnih prijedloga s obje platforme. Opisan je proces izbora broja tema te je tada model s 35 tema prilagođen podacima, a te su teme prikazane popisom njihovih najfrekventnijih riječi.

Na poslijetku su usporedbom tema po kategorijama povezane dvije smislene cjeline koje se protežu kroz rad i dana je usporedba dviju najpoznatijih platformi.

Za potrebe rada u prvom su poglavlju objašnjene metode korištene za analizu. Također, dan je i kratak opis grupnog financiranja, kao i pregled dosadašnjih rezultata.

Summary

This thesis describes an analysis of crowdfunding projects with Natural language processing methods. Projects were downloaded from the two biggest crowdfunding platforms - Kickstarter and Indiegogo and the downloading process has been described in the thesis in short. After that, the preprocessing that transforms raw text into an analysis-friendly form was described.

The analysis consists of two parts. The first part is project classification into categories. After the motivation and theoretical background of the SVM method, we have described feature extraction methods and parameters tuning. The classification was then used to classify Indiegogo projects into categories from Kickstarter. This will make further analysis and platforms comparison possible.

The second part describes topic modelling on both platforms together. After choosing the number of topics, a model was fitted and the topics were represented by their most frequent words.

FInally, the comparison of topics throughout categories connected both parts of the thesis.

For the purposes of this thesis, the first section provides methods used for analysis. Additionally, a short description of crowdfunding, as well as an overview of current scientific results.

Životopis

Hermina Petric Maretić rođena je 30. svibnja 1991. godine u Zagrebu, gdje pohađa Osnovnu školu Ivana Filipovića i Osnovnu školu Silvija Strahimira Kranjčevića. Istovremeno pohađa i Osnovnu glazbenu školu Pavla Markovca, gdje maturira na odjelu gitare. 2009. godine završava Klasičnu gimnaziju u Zagrebu i u istom gradu, na Matematičkom odsjeku Prirodoslovno-matematičkog fakulteta, upisuje preddiplomski studij inženjerske matematike. Za vrijeme studija, kao bivši natjecatelj u debati, počinje voditi debatne sastanke u Klasičnoj gimnaziji.

Po završetku preddiplomskog studija Matematike, 2012. godine, dobiva priznanje za izniman uspjeh na preddiplomskom studiju. Te godine stažira kao programer u Ericssonu Nikoli Tesli i volontira u Ljetnoj tvornici znanosti u Samoboru, gdje drži radionicu iz kriptografije. Iste godine upisuje Diplomski studij Matematike i Računarstva. 2013. godine ljeto provodi na EPFL-u (Ecole polytechnique fédérale de Lausanne) kao stažist u Laboratoriju za računalnu biologiju i bioinformatiku te upisuje paralelni studije Matematičke statistike na PMF-u u Zagrebu. Iste godine počinje suradnju s prof. dr. sc. Mariom Štorgom, koja rezultira rektorovom nagradom za rad "Dinamička analiza mreža - evolucija istraživačkog područja temeljem radova objavljenih u sklopu DESIGN konferencije 2002-2014".

U toku studija bila je demonstrator iz kolegija Linearna algebra 1 i 2, te Mjera i integral. Pred kraj diplomskog studija Matematike i Računarstva, od Matematičkog odsjeka dobiva priznanje, a od Prirodoslovno-matematičkog fakulteta nagradu za izniman uspjeh na studiju. Iste godine sudjeluje na eStudentovom natjecanju Mozgalo u timu s Antom Čabrajom i Anamarijom Fofonjka, gdje osvajaju drugu nagradu, a u rujnu diplomira na studiju Računarstvo i matematika. 2015. godine prijavljuje se na doktorsku školu na EPFL-u te će u rujnu iste godine tamo započeti doktorat u području računarskih znanosti.