

Tacita

Orodje za pomoč pri anonimizaciji sodb

Institut Jožef Stefan

Laboratorij za umetno inteligenco

Center za prenos znanja na področju
informacijskih tehnologij



Anonimizacija sodb

Pred javno objavo je treba sodbe zaradi varovanja zasebnosti anonimizirati – odstraniti iz teksta vse omembe podatkov, preko katerih bi lahko identificirali določene osebe vpletene v postopku.

Razvoj sistema, ki bo s tehnikami strojnega učenja pomagal sodnikom pri anonimizaciji z avtomatsko detekcijo delov besedila za anonimizacijo.



Podatki

Učna množica sestoji iz **2840 sodb**
Vrhovnega in Višjega sodišča v Ljubljani:

- **neanonimizirani** dokumenti v open office (.odt) formatu;
- **anonimizirani** dokumenti dostopni preko HTTP dostopa v JSON formatu (<http://sodnapraksa.si>)

Predprocesiranje:

- ekstrakcija besedila
- poravnava anonimizirane in neanonimizirane različice
- lematizacija



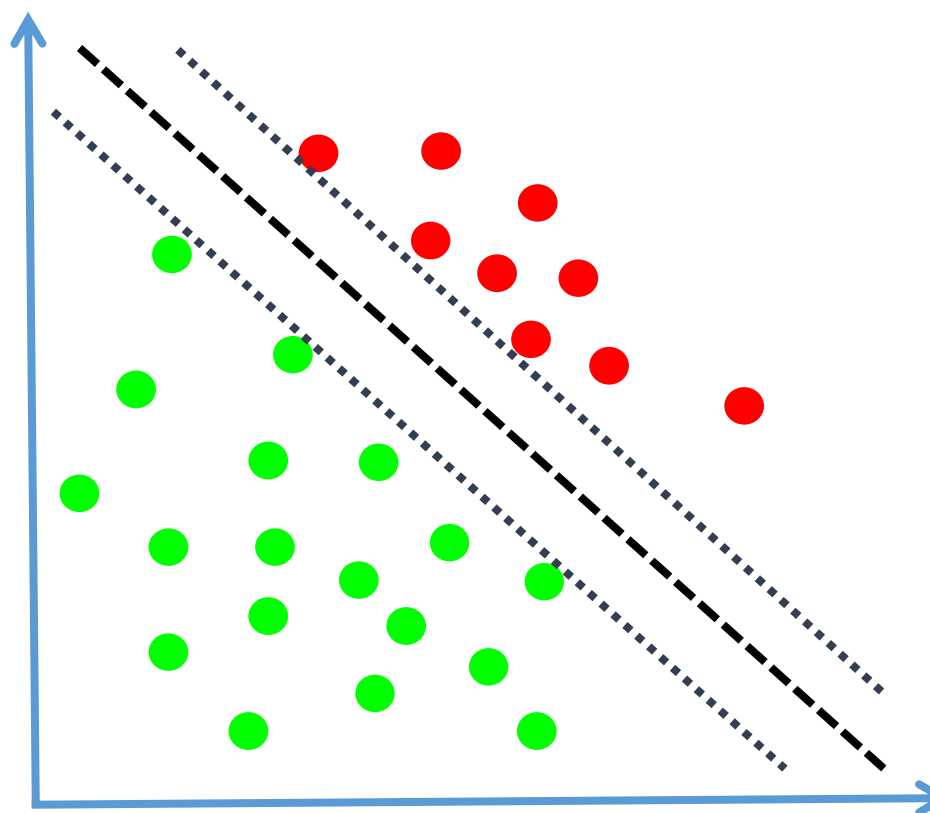
Pristop

Naučen statistični model, na podlagi katerega iz oblike in konteksta besede v besedilu napovemo ali jo je potrebno anonimizirati.

zaradi 12 kaznivih dejanj trgovine z ljudmi po drugem in prvem odstavku 175. člena v zvezi z 20. členom KZ-1 in kaznivega dejanja zlorabe prostitucije po tretjem in prvem odstavku 175. člena v zvezi z 20. členom KZ-1 ter zoper obdolženo [redacted] zaradi petih kaznivih dejanj trgovine z ljudmi po drugem in prvem odstavku 175. člena v zvezi z 20. členom KZ-1 in kaznivega dejanja zlorabe prostitucije po tretjem in prvem odstavku 175. člena v zvezi z 20. členom KZ-1 ter zoper obdolženo [redacted] zaradi dveh kaznivih dejanj trgovine z ljudmi po drugem in prvem odstavku 175. člena v zvezi z 20. členom KZ-1. Obdolženci so v priporočili prvega odstavka 201. člena Zakona o kazenskem postopku (v

Pristop - klasifikacija

Klasifikacija besed v dva razreda z uporabo **metode podpornih vektorjev**



Linearno ločljiv primer

Besedo predstavimo kot m -dimenzionalen vektor **značilk** (v našem primeru $m \sim 600$)

Metoda poišče optimalno **hiperravnino**, ki loči med primeri v tem (preslikanem) prostoru.

Značilke - osnovne

1. Besedni vzorec:

i) kratek

Janez Novak: **Ua Ua**

12.4.1915: **d.d.d**

V8-216: **Ud-d**

SI56-1550-3554-29: **Ud-d-d-d**

ii) dolg

Janez Novak: **Uaa Uaa**

12.4.1915: **dd.d.dddd**

V8-216: **Ud-ddd**

SI56-1550-3554-29:

UUdd-dddd-dddd-dd

2. Predpone dolžine 1, 2 in 3

s.p.: {**s**, **s.**, **s.p**}

mag.: {**m**, **ma**, **mag**}

3. Pripone dolžine 1, 2 in 3

obdolženega: {**a**, **ga**, **ega**}

1205€: {**€**, **5€**, **05€**}

4. Slovar je_ime, je_priimek, je_lokacija

Janez: [**true**, **false**, **false**]

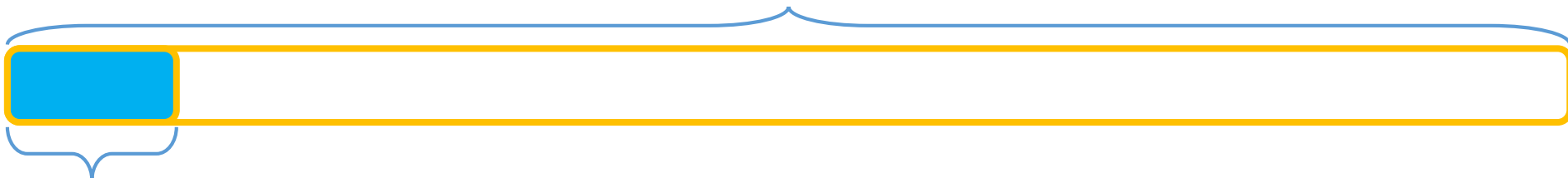
Novak: [**false**, **true**, **false**]

Muhačič: [**false**, **false**, **false**]

Brezovica: [**false**, **false**, **true**]

5. Ujemanje z pomembno besedo iz prvega odstavka (uvoda)

m – dimenzionalen vektor značilk



Osnovne značilke
(trenutna beseda)

Značilke – kontekst

Model **vreča besed** (bag-of-words):

Primer:

„... vrhovni tožilec in obdolženec Janez Goriški ...“:

{Goriški, in, Janez, obdolženec, tožilec, vrhovni}

„... vrhovni tožilec Janez Goriški in obdolženec ...“:

{Goriški, in, Janez, obdolženec, tožilec, vrhovni}

Model **n-gramov** (uni-grami in bi-grami):

Primer:

„... zoper obdolženega Janeza ...“: *{obdolženega, zoper obdolženega}*

m – dimenzionalen vektor značilk



**Kontekstne značilke
(trenutna beseda)**

Značilke – kontekst

Mera **TF-IDF** - Izbor relevantnih kontekstnih značilk

$$tfidf = tf \times idf$$

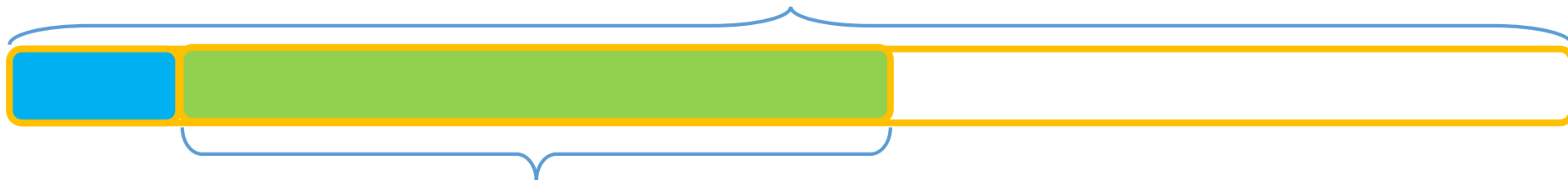
tf - *Term Frequency*

- pogostost besede v kontekstu anonimizirane besede

idf - *Inverse Document Frequency*

- inverz pogostosti besede v kontekstih vseh besed

m - dimenzionalen vektor značilk



**Kontekstne značilke
(trenutna beseda)**

Značilke – kontekst

Vreča besed:

Vzamemo okno besed (štiri pred in dve za)

Primer:

„... vrhovni tožilec in obdolženec Janez Goriški, rojen ...“ - anonimizirano

„... obdolženec, stanujoč na naslovu Poljanska ulica ...“ - anonimizirano

„... podjetje, na naslovu Poljanska ulica ...“ - neanonimizirano

N-grami:

Uni-grami za anonimizacijo:

„... obdolženi...“ „... EMŠO ...“

„... priča ...“ „... stranke ...“

„... priglašeni ...“ „... št. ...“

„... pokojni ...“ „... črpalke ...“

Bi-grami za anonimizacijo:

„... tožene stranke...“

„ ... nova upraviteljica ...“

„ ... izvedenškega mnenja ...“

„ ... program univerze ...“

m – dimenzionalen vektor značilk



**Kontekstne značilke
(trenutna beseda)**

Značilke – značilke sosedov

Standardna klasifikacija predpostavlja medsebojno **nepovezane** in **neodvisne** besede

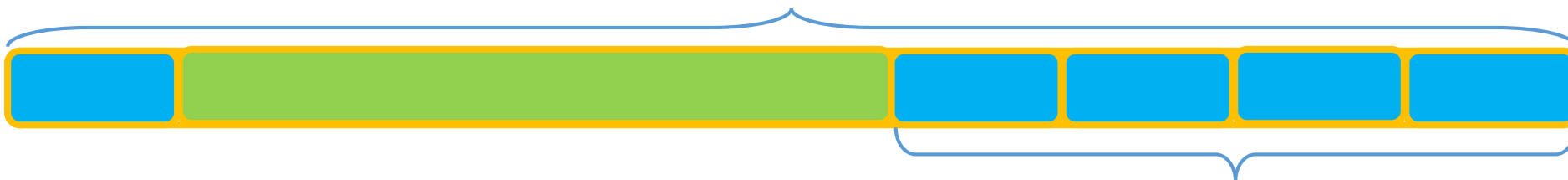
Razred besede je odvisen od njenih sosed:

Primer:

„ ... obdolženec *Erik Medved* ... ” ali „*Medved* je uplenil ... ”

Vključimo še osnovne značilke sosedov obravnavane besede (tri pred, eno za)

m – dimenzionalen vektor značilk



**Osnovne značilke
sosednjih besed**

Uspešnost sistema

5-kratno prečno preverjanje (5-fold cross validation)

Točnost: 72,8% od vseh primerov (besed), ki smo jih označili za anonimizacijo je pravih (pravilno pozitivni). Ostali so lažno pozitivni (jih ne bi smeli anonimizirati).

Priklic: 91,8% je delež resničnih pozitivnih primerov (besed) - se pravi smo (nepravilno) "pozabili" anonimizirati 8,2% primerov.

Primeri

- **uspeh:** »...je z uvodoma navedenim sklepom odredila pripor zoper osumljenega *Janeza Novaka* iz pripornih razlogov begosumnosti in ponovitvene nevarnosti...«
- **uspeh:** »da se obdolženim *Janezu Novaku, Micki Izmišljeni* in *Jožici Neobstaja* podaljša pripor«
- **nepotrebna anonimizacija:** »...na podlagi mednarodne pravne pomoči z *Republiko Srbijo*«
- **napaka:** »...najel varnostnika po imenu *Josip*«
- **napaka?:** »*Vrhovno sodišče* se ne strinja z zagovornikom obdolžene...«

Primeri - uspešna anonimizacija

Vrhovno sodišče Republike Slovenije v senatu, ki so ga sestavljali vrhovni sodniki in sodnik predsednik ter in kot zapisnikarja, v kazenski zadevi zaradi kaznivih dejanj po drugem in prvem odstavku 175. člena Kazenskega zakonika Specializiranega državnega tožilstva RS št. Zoper obdolženega se obdolženo in pa iz pripornega razloga po 1. točki prvega odstavka 201. člena Zakona o kazenskem postopku podaljša pripor še za dva meseca, do vključno 27. 6.

Včasih zaradi napake ne upoštevamo dela besedila. Napaka se tipično pojavi, kadar je bilo besedilo med opombami pod črto, ali ko gre za tipkarsko napako.

Sivo besedilo v obstoječem sistemu ni objavljeno

Pred Okrožnim sodiščem v Ljubljani teče zoper obdolžene zaradi 12 kaznivih dejanj trgovine z ljudmi po drugem in prvem odstavku 113. člena v zvezi z 20. členom in kaznivega dejanja zlorabe prostitucije po tretjem in prvem odstavku 175. člena Kazenskega zakonika (v nadaljevanju KZ-1), zoper obdolženo zaradi petih kaznivih dejanj trgovine z ljudmi po drugem in prvem odstavku 113. člena v zvezi z 20. členom KZ-1 in kaznivega dejanja zlorabe prostitucije po tretjem in prvem odstavku 175. člena KZ-1 ter zoper obdolženo zaradi dveh kaznivih dejanj trgovine z ljudmi po drugem in prvem odstavku 113. člena v zvezi z 20. členom KZ-1. Obdolženci so v priporu iz pripornih razlogov po 2. in 3. točki prvega odstavka 201. člena Zakona o kazenskem postopku (v nadaljevanju ZKP), obdolženi in pa iz pripornega razloga po 1. točki prvega odstavka 201. člena ZKP. Pripor zoper obdolžence teče od 27. 1. 2015 dalje. Obdolžencem je bil pripor p Okrožnega sodišča I Ks 3724/2015 z dne 24. 2. 2015 do 27. 4. 2015, obdolženima do 12.51 ure, obdolženi pa do 13.05 ure. Okrožni državni tožilec j odstavka 205. člena ZKP dne 9. 4. 2015 predlagal, da se obdolženim podaljša pripor. V predlogu za podaljšanje pripora navaja, da obstaja zakonski utemeljen sum, da so obdolženci storili očitana jim kazniva dejanja. Navaja, da so še vedno podani priporni razlogi ter navaja okoliščine, na podlagi katerih ocenjuje, da sta za obdolženega še vedno podana pripora razloga po 2. in 3. točki prvega odstavka 201. člena ZKP, za obdolženi in pa priporni razlog po 1. točki prvega odstavka 201. člena ZKP. Obrazlaga sorazmernost pripora in njegovo neogibnost za varnost ljudi ter potek kazenskega postopka. Kot razlog za podaljšanje pripora navaja, da preiskave iz objektivnih razlogov ni

Zelena barva označuje pravilno anonimizirano besedilo

Sistem je pravilno zaznal in anonimiziral tipkarsko napako

Primeri - nepotrebna anonimizacija

Vrhovno sodišče Republike Slovenije je v senatu, ki so ga sestavljali vrhovni sodniki in sodnik _____, kot predsednik, ter _____ in _____ kot članici, v pravdni zadevi tožeče stranke: _____, ki jo zastopa Odvetniška družba _____, v _____, zoper toženo stranko: _____, ki jo zastopa Odvetniška pisarna _____ v _____ zaradi plačila 22.214,56 EUR s pp,o predlogu tožene stranke za dopustitev revizije zoper sodbo Višjega sodišča v Ljubljani I Cp 1689/2011 z dne 4. 1. 2012, v zvezi s sodbo P 111/1995-I z dne 16. 11. 2010, na seji 18. aprila 2013

Predlog se zavrne.

Sodišče prve stopnje je toženki naložilo, da _____, ki. Sodišče je presodilo, da _____, toženka (matična družba) odgovarja za ob _____, ki je _____, in pri kateri je tož _____ deponirala devizna sredstva. Sodišče druge stopnje je pritožbi tožene stranke glede obresti delno ugodilo _____, je pritožbo o glavni stvari zavrnilo in potrdilo sodbo sodišča prve stopnje. Zoper sodbo sodišča druge stopnje _____ tožena stranka predlog za dopustitev revizije, v katerem predlaga, da se dopusti revizija glede naslednjih vpraš _____ ali je sodišče zmotno uporabilo pravo, ko je odločilo, da se uporablja slovensko pravo in da je pristojno slovensko _____ sodišče (po sedežu toženke, čeprav gre za razmerje, sklenjeno z **banko v Sarajevu**); ali je sodišče zmotno _____ uporabilo pravo, ko je v zvezi s preoblikovanjem LB Osnovne **banke Sarajevo** v LB Glavno filijalo **Sarajevo** ugodilo _____ tožbenemu zahtevku zgolj na podlagi formalne registracije te **banke** kot podružnice toženke, ne da bi upoštevalo _____ naravo deviznih vlog, da postopek sanacije LB Glavne filijale **Sarajevo** ni bil dokončan, da prenos sredstev, pravic _____ in obveznosti na toženko ni bil izveden, podpis in sprejem Sporazuma o vprašanih nasledstva ter ukrepe BiH ob _____ razpadu SFRJ in vojni, vključno s preoblikovanjem LB Glavne filijale **Sarajevo** v Ljubljansko **banko** d. d. **Sarajevo** _____ leta 1993 s strani BiH, pri čemer so delnice novo oblikovane Ljubljanske **banke** d. d. **Sarajevo** po odločitvi države _____ BiH pripadle BiH, zaradi česar vsega je bilo onemogočeno poslovanje Ljubljanske **banke** Glavne filijale **Sarajevo** in _____ kakršenkoli vpliv toženke nanjo; ali je sodišče zmotno uporabilo materialno pravo s tem, ko ni upoštevalo _____ sklicevanja toženke na načelo „branch compensating deposit principle“, po katerem matična banka odgovarja za _____ obveznosti podružnice le, če je za deponirana sredstva v podružnici prejela kritje v matični banki, kot tudi ne _____ določbe 1040. člena Zakona o obligacijskih razmerjih (v nadaljevanju ZOR) o tem, da je hranilna vloga razmerje _____ med deponentom in enoto **banke**, pri kateri je bila vloga odprta; ali je sodišče zagrešilo bistveno kršitev določb _____ pravnega postopka in zmotno uporabilo pravo, ko je že na podlagi vpogleda v fotokopijo hranilne knjižice _____ zaključilo, da ima tožeča stranka na hranilni knjižici pri toženi stranki deponirana denarna sredstva v vtoževani _____ višini in da se tožena stranka ne more sklicevati na dejstvo, da sodba sodišča prve stopnje nima ustrezne dokazne _____ podlage, ker tožeča stranka ni predložila hranilni knjižic v izvirniku, čeprav je tožena stranka to ugovarjala in _____ predlagala. Predlog ni utemeljen. Ker pogoji za dopustitev revizije iz prvega odstavka 367. a člena Zakona o _____ pravnem postopku (v nadaljevanju ZPP) niso podani, je Vrhovno sodišče predlog zavrnilo (drugi odstavek 367. c _____ člena ZPP). Vrhovno sodišče je stališča v zvezi z izpostavljenimi vprašanji že zavzelo v več odločbah (primerjaj II _____ Ips 49/2012 z dne 21. 3. 2013, II Ips 79/2012 z dne 17. 1. 2013, II Ips 307/2011 z dne 5. 4. 2012, II Ips 490/1997 _____ z dne 21. 1. 1999), izpodbijana sodba pa od njih ne odstopa. Ljubljana, 18. april 2013 Predsednik senata: _____ l.r.

V konkretnem primeru je bila tožena stranka banka.

Modra barva označuje nepotrebno anonimizirano besedilo

Primeri - naše napake pri anonimizaciji

Vrhovno sodišče Republike Slovenije je v senatu, ki so ga sestavljali vrhovni sodniki in sodnik [REDACTED], kot predsednik, ter [REDACTED] in [REDACTED], kot članici, v pravnih zadevah tožeče stranke: [REDACTED], ki jo zastopata [REDACTED], odvetnik v [REDACTED] in [REDACTED], odvetnik v [REDACTED], zoper toženo stranko: [REDACTED], ki jo zastopa [REDACTED], odvetnica v [REDACTED] zaradi plačila 24.928,75 EUR s pp, o predlogu tožeče stranke za dopustitev revizije zoper sodbo Višjega sodišča v Kopru I Cp 432/2014 z dne 10. 6. 2014, v zvezi s sodbo Okrožnega sodišča v Kopru P 371/2013 z dne 27. 2. 2014, na seji 13. novembra 2014

Predlog se zavrne.

Okrajno sodišče v Ljubljani, Oddelek za verodostojno listino, je s sklepom VL 12627/ [REDACTED] predlog družbe [REDACTED] d.o.o. dovolilo zoper [REDACTED] predlagano izvršbo in im [REDACTED] predlogu za izvršbo navedeno terjatev. Po ugovoru dolžnice je izvršilno sodišče sklep [REDACTED] katerem je bila izvršba dovoljena ter določen izvršitelj. O zahtevku in o stroških je na [REDACTED] Prvostopenjsko sodišče je sklep o izvršbi razveljavilo še v preostalem delu in tožbeni [REDACTED] zavrnilo. Drugostopenjsko sodišče je zavrnilno sodbo razveljavilo glede zneska 2.777,31 EUR, ker je [REDACTED] delu zahtevka pravnomočno odločeno že v prejšnjem sojenju, sicer pa je zavrnilno sodbo potrdilo. Tožeča stranka je zoper sodbo pritožbenega sodišča vložila predlog za dopustitev revizije. Navaja, da je plačilo, ki je predmet tožbenega zahtevka, uveljavljala za delo, ki ga je opravila na toženkinem objektu v [REDACTED] ulica 7. Obseg in vrednost gradbenih del sta pravdni stranki dogovorili z rekapitulacijo z dne 31. 1. 2007 v znesku 93.920,00 EUR, dodatna dela so znašala 6.410,00 EUR, po tretjem končnem obračunu v znesku 27.706,05 EUR, ki je predmet tožbenega zahtevka, pa je tožeča stranka toženki izstavila račun z dne 5. 7. 2007, ki ga toženka (za razliko od predhodnih računov) ni plačala. Ugovarjala je aktivno legitimacijo tožeče stranke, češ da z njo ni bila v poslovnem razmerju, saj je gradbeno pogodbo z dne 12. 9. 2006 sklenila s samostojnim podjetnikom [REDACTED] sicer lastnikom in direktorjem tožeče stranke. Nasprotuje zaključku pravnomošne sodbe o pomanjkanju svoje aktivne legitimacije in pritožbenemu sodišču očita, da je zmotno uporabilo določbe 337. in 362. člena Zakona o pravdnem postopku (v nadaljevanju ZPP), ker pritožbene trditve, da je bila s podpisom rekapitulacije gradbena pogodba prenesena od samostojnega podjetnika [REDACTED] na družbo [REDACTED] d.o.o., niso dejanske narave in jih prekluzija ne zadene, pač pa pomenijo le obrazložitev nove pravne kvalifikacije razmerja med pravnima strankama. O tej pritožbeni trditvi tudi razlogi niso ustrezni (14. točka drugega odstavka 339. člena ZPP). Zato nedelana dopustitev revizije glede vračanja ali predstavlja navedenih dodatnih pravnih podlag, ki potrjujejo

Rdeča barva označuje besedilo, ki bi ga morali anonimizirati, pa ga nismo

Delovanje sistema

Procesiranje ene sodbe je **hitro** (1s).

Ko so na voljo novi podatki, je sistem mogoče **ponovno učiti** in izboljšati njegovo delovanje.

Proces ponovnega učenja traja nekaj ur.

Konsistentna raba sistema lahko pripomore k **poenoteni praksi anonimiziranja**, kar še dodatno izboljšuje delovanje sistema.

Delovanje sistema

Sistem bi lahko enostavno posplošili na **druge tipe dokumentov** (npr. revizijska poročila).

Večina značilk ni vezana na tip dokumenta

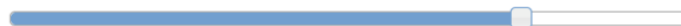
- izjema je preverjanje pomembnosti v prvem odstavku

Relevanten kontekst anonimizirane besede je vezan na tip dokumenta. Velika raznolikost dokumentov tako lahko bistveno oteži nalogo anonimizacije.

Uporabniški vmesnik

Anonimizator

Izberite stopnjo anonimizacije



Zahtevi za varstvo zakonitosti se ugotovi in se kazenski postopek zoper obdolženega **Janeza Novaka**, rojenega **3. 9. 1975** v Celju, EMŠO **0309975500123**, Slovenca, državljan Republike Slovenije, ustavi, ker je v stanju neprištevnosti, ko zaradi paranoidne shizofrenije ni mogel razumeti pomena svojih dejanj in ni mogel imeti v oblasti svojega ravnanja, z drugimi grdo ravnal, jih z grožnjo z neposrednim napadom na življenje in telo preganjal in jih kako drugače z nasilnim omejevanjem njihovih enakih pravic spravljal v podrejen položaj s tem, da je: 1. v času od neugotovljenega dne v letu **2010** do decembra 2013 v Trgu pri Novi vasi, hodil po vasi s plinsko masko na obrazu in z nožem v rokah ter grozil, da bo vse požgal, pobil in poklal ter z veliko hitrostjo z avtomobilom znamke Passat divjal po vasi, če zlasti na dne **7. 7. 2013** se je z veliko hitrostjo s svojim avtomobilom znamke P **Označi besedo** na naslovu **Trg 22** v Novi vasi, se večkrat ustavil ob hiši in gr **Označi vse pojavitve besede** l, da bo pobil vse otroke in zažgal hišo, zaradi česar je Kr **ato je** zaklenil vsa vrata in si ni upal zaspati ter je celo nič čakal, če bo **Novak** res prišel do hiše, nakar se je okoli 22.30 ure **Novak** ponovno pripeljal do njihove hiše in pričel ponovno groziti, zato je **Kranjec** poklical policiste, ki so z Novakom opravili razgovor, ko pa sta se policista vrnila k **Kranjcu** pa je **Novak** ponovno prišel iz svoje hiše in se pričel dreti, da bo vse poklal, hodil je okoli hiše in vpil, zato sta ga policista odpeljala; dne 11. 7. 2013 je **Novak** z nožem napadel Miho Kranjca v bližini njegovega doma na naslovu **Trg 22** v Novi vasi, ko se je **Kranjec** s traktorjem peljal proti domu je na vogal hiše na cesti z nožem v rokah prišel **Janez Novak** in vpil nanj, ter grozil, da bo vse pobil, požgal, zato se je **Kranjec** s traktorjem umaknil vzvratno in po drugi strani peš odšel do hiše, od koder je poklical policiste, ko pa so prišli policisti pa je **Novak** še vedno imel nož v rokah in je vpil tudi na policiste in šele v njihovem spremstvu je lahko **Kranjec** pripeljal traktor do doma, ko pa so policisti odšli, pa je **Novak** pričel z avtomobilom divjati po vasi in je zaradi neprilagojene hitrosti tudi zletel s ceste in uničil vozilo, zaradi takšnega njegovega ravnanja in ker se je tudi po dvajsetkrat na dan z veliko hitrostjo vozil mimo njihove hiše pa se je **Kranjec** počutil ogroženega in se je bal zase in za svojo družino, zaradi njegovih izpadov pa so se **Kranjec** in njegova družina pred njim zaklepali v hišo ter si niso upali zapuščati hiše in hoditi po vasi, s čimer jih je spravljal v podrejen položaj; 2. v času od decembra 2013 do 13. 8. 2014 na Neobstoječi ulici 15 v Celju, obtoževal sosede, da mu vlamljajo v stanovanje in kradejo ter spuščajo vodo v stanovanje, v nočnih in zgodnjih jutranjih urah je razbijal po stanovanju in po ulici hodil s plinsko masko na obrazu in z nožem ali verigo v rokah ter grozil sosedom, z neprilagojeno hitrostjo obračal vozilo na makadamskem parkirnem prostoru, da je kamenje letelo po ostalih avtomobilih in osebah ter iz svojega stanovanja metal kamenje po avtomobilih in sosednjih stavbah, še zlasti pa: dne **17. 7. 2014** je v podhodu pred pisarno na naslovu Neobstoječa ulica 15 v Celju, pričakal **Marijo Kovač** s plinsko masko na

0309975500123

3. 9. 1975

Janez Novak

Janeza Novaka

Kovačeva

Kranjcu

Kranjec

Marijo Kovač

Mihe Kranjca

Mizar

Novak

Novaka

Roka Mizarja

Trg 22

Ročno označene
besede

2010

Ročno odoznačene
besede

17. 7. 2014