LATVIJAS UNIVERSITĀTE

DATORIKAS FAKULTĀTE

AUTOMATIZĒTA UZŅĒMUMU TĪMEKĻA VIETŅU MEKLĒŠANA

BAKALAURA DARBS

Autors: **Kristaps Znotiņš**
Studenta apliecības Nr.: kz11113
Darba vadītājs: prof., Dr. dat. Guntis Arnicāns

RĪGA 2016

ANOTĀCIJA

Uzņēmumi savās tīmekļa vietnēs publicē informāciju, kas ir noderīga to klientiem, sadarbības partneriem un konkurentiem. Viens no pirmajiem soļiem šīs informācijas apguvei ir tīmekļa vietņu identificēšana, kurās uzņēmumi publicē ar tiem saistīto informāciju. Informāciju par savām tīmekļa vietnēm parasti ir ieinteresēti izplatīt arī paši uzņēmumu pārstāvji, un tā no dažādiem avotiem jau tiek apkopota vairākos publiski pieejamos repozitorijos, tomēr tās pievienošana, atjaunošana un kontrole var prasīt nozīmīgus cilvēkresursus.

Autors apskata uzņēmumu tīmekļa vietņu meklēšanas automatizācijas iespējas, izmantojot tīmekļa saturu un citus resursus, un piedāvā risinājumu, kas Latvijā reģistrētiem uzņēmumiem sasniedz vairāk kā 90% precizitāti ar līdz 70% pārklājumu tīmekļa vietņu piederības noteikšanai.

Atslēgvārdi: informācijas izgūšana, tīmekļa rasmošana, klasifikācija, tīmekļa vietņu klasifikācija

ABSTRACT

Automated discovery of company websites

Companies utilize their websites to publish information that is valuable for their clients, partners and competitors. One of the first steps in processing this information is to discover websites where companies publish their information. Companies are usually interested in making their website information available themselves and it is already been collected in multiple publicly available repositories; however submission, maintenance and control of this information require significant human resources.

Author describes methods for automated company website discovery by utilizing web content and other resources and proposes a website classification solution that exceeds 90% accuracy with up to 70% recall for companies of Latvia.

Keywords: information retrieval, web crawling, classification, website classification

SATURA RĀDĪTĀJS

APZĪMĒJUMU SARAKSTS

Atslēgas domēns – augstākā līmeņa privātais domēns, par kura saturu atbild konkrēta uzņēmuma pārstāvji un kurā ir atrodama uzņēmuma kontaktinformācija un cita klientiem un sadarbības partneriem noderīga informācija.

Atslēgas domēnu meklēšanas pārklājums – attiecība starp ar risinājumu atrasto atslēgas domēnu skaitu un kopējo atslēgas domēnu skaitu zelta standartā. Atslēgas domēnu meklēšanas pārklājums raksturo varbūtību tam, atslēgas domēns ar risinājumu tiks atrasts.

Atslēgas domēnu meklēšanas precizitāte – attiecība starp ar risinājumu pareizi atrasto atslēgas domēnu skaitu un kopējo ar risinājumu atrasto domēnu skaitu. Atslēgas domēnu meklēšanas precizitāte raksturo varbūtību tam, ka ar risinājumu atrasts domēns patiešām ir atslēgas domēns.

Augstākā līmeņa privātais domēns – interneta domēns, kas sastāv tikai no vienas neatkarīgas domēna vārda daļas (bez apakšdomēniem) un domēna publiskā sufiksa.

Domēna publiskais sufikss – domēna vārda beigu daļa, kas pieder Mozilla organizācijas uzturētajam publisko sufiksu sarakstam[1].

Interesanta kandidātdomēna lapa - kandidātdomēna lapa, kas ar lielu varbūtību satur uzņēmuma pamatdatus, ja kandidātdomēns ir atslēgas domēns.

Kandidātdomēns – augstākā līmeņa privātais domēns, kas ar lielu varbūtību varētu būt arī uzņēmuma atslēgas domēns.

Kandidātdomēnu atlases pārklājums - attiecība starp atslēgas domēnu skaitu uzņēmumiem atrastajos kandidātdomēnos un kopējo atslēgas domēnu skaitu zelta standartā.

Neatslēgas domēns – kandidātdomēns, kas nav atslēgas domēns.

[1] Pieejams: https://publicsuffix.org/list

Latvijas aktīvo SIA (sabiedrību ar ierobežotu atbildību) kopa – Latvijas Uzņēmumu reģistrā iekļauto sabiedrību ar ierobežotu atbildību kopa, kas ietver uzņēmumus, kam neeksistē pabeigts to likvidācijas vai reorganizācijas process.

Uzņēmumu korpuss – autora izveidota Latvijas aktīvo SIA kopas apakškopa, kurā iekļautajiem uzņēmumiem manuāli tika veikta to atslēgas domēnu meklēšana.

WHOIS - domēna vārdu reģistru uzturēts serviss, kas sniedz informāciju par domēna vārdu reģistrētājiem un izmantotājiem.

Zelta standarts – autora manuāli noteiktais atslēgas domēnu piekārtojums uzņēmumu korpusam.

IEVADS

Palielinoties tīmekļa resursos pieejamās informācijas apjomam un detalizācijas pakāpei, personas un uzņēmumi var gūt būtiskas priekšrocības, veicot tās apstrādi un analīzi. Uzņēmumu tīmekļa vietnes sniedz to klientiem, sadarbības partneriem un konkurentiem svarīgu informāciju un tiek plaši izmantotas gan komerciālos, gan izpētes nolūkos.

Uzņēmumu tīmekļa vietņu informācijas izmantošanai sākumā ir nepieciešams izgūt informāciju par to piederību konkrētiem uzņēmumiem. Nelielos apjomos uzņēmumu tīmekļa vietnes ir iespējams atrast informācijas portālos, meklētājdzinēju rezultātos, sociālajos tīklos un citos resursos. Tomēr uzņēmumu tīmekļa vietņu informācijas pievienošana, uzturēšana un kontrole lielos apjomos prasa nozīmīgus cilvēkresursus, jo tām ir raksturīga diezgan strauja mainība un daudzi to uzturēšanas procesi bieži tiek veikti manuāli. Rezultātā uzņēmumu tīmekļa vietņu informācija lielos apjomos parasti ir pieejama tikai komerciālā ceļā.

Īpaši noderīga brīvi pieejama uzņēmumu tīmekļa vietņu informācija lielos apjomos varētu kļūt pēc atvērto datu iniciatīvu īstenošanas, kuru rezultātā uzņēmumu pamatdati kļūst publiski pieejami lejupielādei. Piemēram, 2014. gadā savus pamatdatus publiski pieejamus lejupielādei padarīja arī Latvijas Uzņēmumu reģistrs.

Lielu daļu no uzņēmumu tīmekļa vietņu meklēšanas procesiem ir iespējams automatizēt, izmantojot tīmekļa resursu analīzi, ko komerciālos nolūkos veic arī daudzi uzņēmumi. Brīvāk pieejama uzņēmumu tīmekļa vietņu informācija varētu tikt izmantota arī dažādos pētījumos.

Darba ietvaros tiek apskatītas uzņēmumu tīmekļa vietņu meklēšanas automatizācijas iespējas. Konkrētāk, tiek apskatīta Latvijā reģistrēto uzņēmumu publiski pieejamās informācijas automatizēta papildināšana ar to tīmekļa vietņu informāciju.

Darba galvenais mērķis ir izveidot risinājumu, ar kura palīdzību varētu veikt Latvijas uzņēmumu tīmekļa vietņu automatizētu meklēšanu, pēc iespējas mazāk izmantojot citus centralizētus informācijas avotus.

Autors savā kursa darbā [Znotiņš, 2015] apskatīja salīdzinoši vienkāršāku līdzīgu risinājumu Vācijas uzņēmumu kontekstā. Šajā darbā tiek izmantoti daži kursa darbā aprakstītā risinājuma principi, bet piedāvātais risinājums ietver būtiskas izmaiņas un papildinājumus.

Darbs sastāv no vairākām nodaļām. Pirmajā nodaļā tiek precizēta risināmā problēma un dots ieskats tās risināšanas pieejā. Otrajā nodaļā tiek aprakstīts autora veidots uzņēmumu korpuss, kurā iekļautajiem uzņēmumiem tika veikta manuāla tīmekļa vietņu meklēšana. Trešajā nodaļā tiek analizēti un salīdzināti citi tīmekļa vietņu informācijas avoti. Ceturtajā, piektajā un sestajā nodaļā tiek aprakstīts piedāvātais risinājums. Septītajā un astotajā nodaļā tiek veikta risinājuma rezultātu analīze, tiek novērtētas tā raksturiezīmes, iespējamie pielietojumi un uzlabojumi.

1. PROBLĒMAS APSKATS

Pamatdati par uzņēmumiem ir publiski pieejami lejupielādei un apstrādei lielā daļā valstu, kas ir uzsākušas atvērto datu iniciatīvas. Kopsavilkumu par uzņēmumu datu pieejamību dažādās valstīs sniedz atvērto datu indekss[2]. Latvijā Uzņēmumu reģistrs savus pamatdatus[3] padarīja publiski pieejamus lejupielādei salīdzinoši nesen (2014. gada vidū).

Balstoties uz atvērtiem datiem, ir iespējams veikt uzņēmumu pamatdatu apstrādi, papildināšanu un jaunu pakalpojumu sniegšanu. Piemēram, OpenCorporates portāls[4], lielā mērā balstoties uz šiem datiem, piedāvā datubāzi, kurā ir iespējams meklēt vairāk kā 100 miljonus uzņēmumu no dažādām valstīm, ietverot arī Latvijas uzņēmumus. Autoram pagaidām nav zināmi daudzi citi projekti, kas līdzīgi izmantotu Latvijas uzņēmumu atvērtos datus.

Publiskotie uzņēmumu dati parasti nesatur informāciju par uzņēmumu tīmekļa vietnēm, jo uzņēmumi centralizēti nesniedz šādu informāciju valstu institūcijām. Darba ietvaros tiek apskatīta risinājuma izstrāde Latvijas uzņēmumu pamatdatu automatizētai papildināšanai ar informāciju par to tīmekļa vietnēm.

1.1. Problēmas nostādne

Darba galvenais uzdevums ir atrast uzņēmumu tīmekļa vietnes, izmantojot uzņēmumu pamatdatus. Ar uzņēmuma tīmekļa vietnēm darbā tiek saprastas tīmekļa vietnes, par kuru saturu atbild uzņēmuma pārstāvji un kurās ir atrodama uzņēmuma kontaktinformācija un cita klientiem un sadarbības partneriem noderīga informācija. Svarīgākais tīmekļa vietnes piederības kritērijs ir tās noderīgums personām, kas vēlas tīmeklī atrast uzticamu informāciju par uzņēmumu, kuru ir publicējuši uzņēmuma pārstāvji.

---

[2] Pieejams: http://index.okfn.org/dataset/companies

[3] Lejupielādei pieejamie Latvijas Uzņēmumu reģistra pamatdati. Pieejams: http://dati.ur.gov.lv

[4] Pieejams: https://opencorporates.com

1.2. Saistība ar esošiem informācijas avotiem

Latvijas uzņēmumu tīmekļa vietņu informācija ir publiski pieejama vairākos informācijas avotos, kas tiek apskatīti sīkāk darba nākamajās nodaļās. Kā piemēru var minēt uzņēmuma SIA Lursoft IT sniegto informāciju portālā zo.lv[5], kura informācijas lapās daļai uzņēmumu ir norādītas arī to tīmekļa vietnes.

Aktuālo tīmekļa vietņu informācijas pievienošanai un uzturēšanai tiek izmantotas dažādas metodes, kas nodrošina iespēju uzņēmuma pārstāvjiem, informācijas sniedzēju pārstāvjiem un trešajām pusēm pieteikt izmaiņas vai veikt tās tieši. Lielākā daļa šo darbību prasa būtiskus cilvēkresursus. Piedāvātajam risinājumam būtu būtiski jāsamazina nepieciešamo resursu apjoms uzņēmumu tīmekļa vietņu informācijas pievienošanai un uzturēšanai. Izveidotais risinājums varētu tikt izmantots līdzīgu informācijas avotu papildināšanai vai par pamatu no tiem neatkarīga informācijas avota izveidei.

Piedāvātajā risinājumā būtu vēlams paredzēt un atbalstīt tā pielāgošanas iespējas arī citu valstu uzņēmumu tīmekļu vietņu meklēšanai, jo daudzi pastāvošie informācijas avoti ir ierobežoti ar tīmekļa vietņu informācijas sniegšanu vienas valsts uzņēmumu ietvaros.

1.3. Atslēgas domēna jēdziens

Pirms uzņēmumu tīmekļa vietņu meklēšanas ir svarīgi precizēt uzņēmuma tīmekļa vietnes jēdzienu. Šim nolūkam darbā tiek definēts uzņēmuma atslēgas domēna jēdziens. **Uzņēmuma atslēgas domēns** darba ietvaros ir interneta domēns, par kura saturu atbild uzņēmuma pārstāvji un kurā ir atrodama uzņēmuma kontaktinformācija un cita klientiem un sadarbības partneriem noderīga informācija. Atslēgas domēns ir konkrēts augstākā līmeņa privātais domēns (piemēram, lu.lv vai cambridge.gov.uk). Minētajos piemēros .lv un .gov.uk ir **domēnu publiskie sufiksi**, kuros tiek reģistrēti privātie domēni. Augstākā līmeņa privātie domēni nesatur tā reģistrētāja veidotus apakšdomēnus.

Galvenie iemesli šādai tīmekļa vietnes definīcijai ir sekojoši:

[5] Pieejams: http://www.zo.lv

- augstākā līmeņa privāto domēna vārdu reģistrācija par maksu tiek veikta domēna vārdu reģistros, kas ierobežo to dinamiskumu;
- uzņēmumu īpašnieki savas tīmekļa vietnes parasti izvieto uz nodalītiem privātā līmeņa domēniem juridisku, drošības un tehnisku apsvērumu dēļ.

Atslēgas domēna noteikšanā būtiskākie kritēriji ir tajā atrodamās informācijas noderīgums uzņēmuma klientiem vai sadarbības partneriem un par tā saturu atbildīgās personas, kam būtu jāpārstāv uzņēmuma īpašnieki. Darba ietvaros netiek meklētas dažādos domēnos atrodamas atsevišķas tīmekļa lapas, kurās ir atrodama informācija par uzņēmumiem, jo šī informācija ir ļoti dinamiska un ir praktiski neiespējami noteikt, vai šo informāciju ir apstiprinājuši un izplatījuši uzņēmumu pārstāvji.

Viens augstākā līmeņa privātais domēns varētu būt atslēgas domēns vairākiem uzņēmumiem un vairāki augstākā līmeņa privātie domēni varētu būt atslēgas domēni vienam uzņēmumam. Darba ietvaros īpaša uzmanība gan tiek pievērsta uzņēmuma galvenā atslēgas domēna meklēšanai, jo šādi ir vieglāk salīdzināt iegūtos rezultātus ar citiem avotiem, kuros visbiežāk ir norādīta tikai viena tīmekļa vietne, un lielākajā daļā gadījumu lietotājus interesē tieši galvenā uzņēmuma tīmekļa vietne. Piedāvātais risinājums gan nodrošina iespēju piekārtot vairākus līdzīgas kvalitātes atslēgas domēnus vienam uzņēmumam, un to būtu iespējams paplašināt arī cita veida saistītu domēnu meklēšanai.

1.4. Uzņēmumu pamatdatu kopa

Darbā tiek apskatīti Latvijā reģistrētie uzņēmumi, kuru pamatdati ir pieejami lejupielādei Uzņēmumu reģistra tīmekļa vietnē[6]. Dati minētajā resursā tiek atjaunoti reizi dienā. Darbā veikto eksperimentu rezultāti attiecas uz datu versiju, kas tika izgūta 2016. gada 19. martā.

Ņemot vērā, ka vairums uzņēmumu pēc to darbības formas ir sabiedrības ar ierobežotu atbildību (skat. 1.1. tabula), tālākajos eksperimentos tiek apskatītas tieši aktīvās

[6] Pieejams: http://dati.ur.gov.lv

sabiedrības ar ierobežotu atbildību, kas veido **Latvijas aktīvo SIA kopu**, kurā datu izgūšanas brīdī ietilpa 163474 uzņēmumi. SIA aktivitāte tiek noteikta, pārbaudot, vai tā nav tikusi reorganizēta vai likvidēta. Aprakstītās metodes varētu izmantot arī cita veida organizācijām, bet ērtākai rezultātu novērtēšanai izvēlētais uzņēmumu veids tālākajos eksperimentos tiek apskatīts atsevišķi.

1.1. tabula

Uzņēmumu reģistrā un komercreģistrā iekļauto vienumu skaits pēc to formas (neiekļaujot likvidētus vai pārstrukturētus uzņēmumus)

Darbības forma	**Skaits**
Sabiedrība ar ierobežotu atbildību	163474
Zemnieku saimniecība	27886
Individuālais uzņēmums	13024
Individuālais komersants	12354
Kooperatīvā sabiedrība	1892
Akciju sabiedrība	1000
Pilnsabiedrība	573
Ārvalsts komersanta filiāle	532
Filiāle	465
Citi	309
Kopā	221509

1.5. Risinājuma raksturojums

Automatizēta uzņēmumu tīmekļa vietņu meklēšana plaši tiek veikta komerciālos nolūkos, lai apkopotu un analizētu uzņēmumu informāciju. Zinātniskajā literatūrā uzņēmumu tīmekļa vietņu meklēšanai autoram neizdevās atrast daudzus tieši atbilstošus pētījumus, bet problēmas risināšanai tika izmantotas daudzas metodes un modeļi, kas tiek plaši pielietoti līdzīga veida problēmu risināšanā.

Uzņēmumu atslēgas domēnu meklēšanas problēma darbā tiek apskatīta kā augstākā līmeņa privāto domēnu bināras klasifikācijas uzdevums, kur augstākā līmeņa privātajam domēnam un uzņēmuma pamatdatiem tiek noteikts, vai augstākā līmeņa privātais domēns ir atslēgas domēns konkrētajam uzņēmumam.

Augstākā līmeņa privāto domēnu klasifikācijai par pamatu tiek izmantotas tīmekļa vietņu klasifikācijas metodes, ņemot vērā, ka konkrētajā gadījumā klasifikācija tiek veikta augstākā līmeņa privāto domēnu līmenī. Tīmekļa vietņu klasifikācija ir diezgan plaši pētīta problēma un [Qi, 2009] sniedz apkopojumu par vairākām darbā izmantotajām klasifikācijas pazīmēm.

Tīmekļa vietņu meklēšanai darba ietvaros tiek apskatīta informācijas izgūšana no tīmekļa vietņu satura, DNS ierakstiem, domēnu vārdu reģistriem, meklētājdzinējiem u.c. avotiem.

1.6. Rezultātu novērtēšana

Piedāvātā risinājuma novērtēšanai tiek izmantotas divas galvenās metodes:

1) Risinājuma rezultāti tiek salīdzināti ar autora izveidota uzņēmumu korpusa manuāli noteiktiem uzņēmumu atslēgas domēniem.
2) Risinājuma rezultāti tiek salīdzināti ar citiem publiski pieejamiem informācijas avotiem.

Kā galvenās rezultātu novērtēšanas metrikas attiecībā pret zelta standartu tiek izmantots atslēgas domēnu meklēšanas pārklājums un atslēgas domēnu meklēšanas precizitāte.

2. UZŅĒMUMU KORPUSS

Lai gan uzņēmumu tīmekļa vietņu informācija ir pieejama vairākos avotos, kas tiek sīkāk apskatīti nākamajā nodaļā, tajos salīdzinoši bieži tika novērota iztrūkstoša un neprecīza tīmekļa vietņu informācija. Tāpēc piedāvātā risinājuma un citu informācijas avotu sniegto rezultātu salīdzināšanai un risinājumā izmantoto algoritmu apmācībai tika izveidots **uzņēmumu korpuss**, kurā pēc gadījuma principa tika iekļauti 1000 uzņēmumi no iepriekš aprakstītās Latvijas aktīvo SIA kopas un kurā iekļautajiem uzņēmumiem manuāli tika meklēti to atslēgas domēni.

2.1. Atslēgas domēnu noteikšanas vadlīnijas

Korpusā iekļautajiem uzņēmumiem manuāli tika meklēti to atslēgas domēni, izmatojot:

- populārākos tīmekļa meklētājdzinējus;
- centralizētus informācijas portālus;
- sociālos tīklus;
- WHOIS ierakstus (domēna vārdu reģistru publiskotus ierakstus, kas satur informāciju par domēnu lietotājiem);
- citus informācijas avotus.

Uzņēmumiem atslēgas domēni tika meklēti saskaņā ar iepriekš aprakstīto atslēgas domēna definīciju. Katram uzņēmumam pamatā tika meklēts viens galvenais atslēgas domēns, kaut gan vēlāk korpusu būtu iespējams paplašināt ar vairākiem atslēgas domēniem un citiem atslēgas domēnu veidiem.

Aprakstītā atslēgas domēna definīcija sākumā var šķist diezgan neskaidra, tāpēc detalizētāk tiek apskatīti daži piemēri uzņēmumu atslēgas domēnu meklēšanai (skat. 2.1. tabula).

2.1. tabula

Piemēri atslēgas domēnu noteikšanai korpusa uzņēmumiem

Uzņēmuma Reģ. Nr.	**Uzņēmuma nosaukums**	**Atslēgas domēns**	**Paskaidrojums**
44103011884	SIA "DELVE 2"	delve2.lv	Domēna kontaktu lapā atrodams uzņēmuma nosaukums, reģistrācijas numurs un juridiskā adrese, meklētājdzinējos domēns tiek atgriezts pirmajos rezultātos, meklējot pēc tā nosaukuma, domēns ir norādīts portālos *zo.lv, 1188.lv, 1182.lv*. Ir saskatāma spēcīga atslēgas domēna atbilstība.
40103684181	"SPARE SPORT" SIA	sparesport.lv	Domēna kontaktu lapā atrodams uzņēmuma nosaukums un reģistrācijas numurs. Atslēgas domēns nav norādīts portālos *zo.lv, 1188.lv, 1182.lv, zl.lv*, bet tā *WHOIS* ierakstā ir atrodams uzņēmuma reģistrācijas numurs. Ir saskatāma vidēji spēcīga atslēgas domēna atbilstība.
40003661248	SIA "KVIDS PLUS"	neeksistē	Domēns nav norādīts lielākajos uzņēmumu informācijas portālos, meklētājdzinējos netiek atrasti atbilstoši domēni, neizdodas atrast arī uzņēmuma profilus sociālajos tīklos. Atslēgas domēns nav atrodams.

2.2. Atslēgas domēnu meklēšanas rezultāti

Atslēgas domēni kopumā tika atrasti 12.5% korpusa uzņēmumu (skat. 2.2. tabula).

2.2. tabula

Atslēgas domēnu skaits uzņēmumu korpusā

Kopējais korpusa uzņēmumu skaits	1000
Korpusa uzņēmumu ar atslēgas domēnu skaits	125
Korpusa uzņēmumu ar atslēgas domēnu īpatsvars (ar 95% ticamības intervālu)	12.50 ± 2.05%

Kā papildus novērojumu, balstoties uz proporcijas īpatsvara ticamības intervālu [Berenson, Levine, Krehbiel, 2011], var izdarīt spriedumu, ka ar 95% varbūtību 10.45% līdz 14.55% Latvijas aktīvajiem SIA ir pieejama tīmekļa vietne (izmantojot autora atslēgas domēna definīciju).

Jāņem vērā, ka uzņēmumu atslēgas domēni ir mainīgi, tāpēc to informāciju ir nepieciešams diezgan regulāri atjaunot.

3. ESOŠIE INFORMĀCIJAS AVOTI

Darba ietvaros tiek apskatīti vairāki informācijas avoti, kuros jau ir atrodama informācija par Latvijas uzņēmumu tīmekļa vietnēm. Nodaļā tiek aprakstīti tikai kvalitatīvākie identificētie Latvijas uzņēmumu tīmekļa vietņu informācijas avoti. Autors kursa darba ietvaros [Znotiņš, 2015] veica līdzīgu plašāku pieejamo informācijas avotu izpēti.

Informācijas avotu izvēlē tika ņemta vērā to:

- pieejamība (vai avots ir publiski pieejams un izmantojams);
- tvērums (vai avots aptver visus vai lielu daļu no Latvijas uzņēmumiem);
- norādīto tīmekļa vietņu precizitāte un pārklājums.

Izvēlētajiem informācijas avotiem tika salīdzināta tajos norādīto tīmekļa vietņu atbilstība izveidotajam atslēgas domēnu korpusam. Tie tika salīdzināti arī savā starpā un rezultātu nodaļā tie tiek salīdzināti ar autora piedāvāto risinājumu.

3.1. Izvēlētie informācijas avoti

Balstoties uz augstāk minētajiem kritērijiem, rezultātu salīdzināšanai tika izvēlēti vairāki uzņēmumu tīmekļa vietņu informācijas avoti:

- *SIA Lursoft IT* uzturētais portāls *zo.lv*[7]*;*
- *SIA Lattelecom BPO* uzturētais portāls *1188.lv*[8]*;*
- *SIA Latvijas Tālrunis* uzturētais portāls *zl.lv*[9]*;*
- *SIA Corporate Services* uzturētais *SIA Tele2* uzziņu portāls *1182.lv*[10].

[7] Pieejams: http://www.zo.lv

[8] Pieejams: http://www.1188.lv

[9] Pieejams: http://zl.lv

[10] Pieejams: http://www.1182.lv

Jāatzīmē, ka visi identificētie kvalitatīvākie informācijas avoti bija internetā pieejami uzņēmumu informācijas portāli. Korpusa uzņēmumiem tika salīdzināts augstāk minētajos avotos norādīto tīmekļa vietņu skaits (skat. 3.1. tabula).

3.1. tabula

Avotos norādīto tīmekļa vietņu skaita salīdzinājums

Informācijas avots	Uzņēmumu ar norādītu tīmekļa vietni skaits	Uzņēmumu ar norādītu tīmekļa vietni īpatsvars	Īpatsvara ticamības intervāls (95%)
zo.lv	63	6.30%	[4.79%, 7.81%]
1188.lv	21	2.10%	[1.21%, 2.99%]
zl.lv	101	10.10%	[8.23%, 11.97%]
1182.lv	73	7.30%	[5.69%, 8.91%]

3.2. Informācijas avotu savstarpējs salīdzinājums

Lai novērtētu norādīto tīmekļa vietņu atšķirības dažādos informācijas avotos, tika salīdzināti tajos norādīto tīmekļa vietņu augstākā līmeņa privātie domēni (skat. 3.2. tabula).

3.2. tabula

Norādīto augstākā līmeņa privāto domēnu salīdzinājums starp avotiem

Pirmais avots	Otrais avots	Sakrītošas	Atšķirīgas	Norādītas tikai pirmajā avotā	Norādītas tikai otrajā avotā
zo.lv	1188.lv	5	1	57	15
zo.lv	zl.lv	29	5	29	67
zo.lv	1182.lv	21	4	38	50
1188.lv	zl.lv	15	0	6	86
1188.lv	1182.lv	8	1	12	66
zl.lv	1182.lv	47	4	50	24
Kopā		125	15	192	308

Tika novērots liels skaits uzņēmumu, kam tīmekļa vietne ir norādītas tikai vienā informācijas avotā; bet gadījumos, kad tīmekļa vietnes ir norādītas abos avotos, tās sakrīt gandrīz 90% gadījumu. Tas apstiprina to, ka lielākajā daļā avotu izpratne par uzņēmumu tīmekļa vietnēm ir līdzīga, bet pastāv būtiskas atšķirības starp uzņēmumu kopām, kam tās ir norādītas dažādos avotos.

Gadījumiem, kuros tika novērotas atšķirīgas tīmekļa vietnes, tika apskatīti galvenie novēroto atšķirību iemesli (skat. 3.3. tabula).

3.3. tabula

Iemesli norādīto tīmekļa vietņu atšķirībām starp avotiem

Iemesls	**Biežums**
Saturiska rakstura kļūda vienā avotā	5
Gramatiska rakstura kļūda vienā avotā	4
Pārvirzīšana uz citu domēnu	4
Domēni ar vienādu saturu	1
Domēni ar līdzvērtīgu atbilstību	1
Kopā	15

Var novērot, ka lielākā daļa atšķirību pastāv novēršamu kļūdu dēļ. Tikai vienā gadījumā bija grūti noteikt, kurš no norādītajiem domēniem labāk atbilst konkrētajam uzņēmumam.

3.3. Informācijas avotu salīdzinājums ar zelta standartu

Zelta standarts ir labākais autora noteiktais atslēgas domēnu piekārtojums korpusa uzņēmumiem. Lai novērtētu avotu sniegto precizitāti un pārklājumu, avotos norādīto tīmekļa vietņu augstākā līmeņa privātie domēni tika salīdzināti ar zelta standartu (skat. 3.4. tabula).

3.4. tabula

Avotos norādīto tīmekļa vietņu salīdzinājums ar zelta standartu

Informācijas avots	**Sakrītošas**	**Atšķirīgas**	**Norādītas tikai avotā**	**Norādītas tikai zelta standartā**
zo.lv	35	7	21	84
1188.lv	18	0	3	108
zl.lv	65	13	23	48
1182.lv	42	9	14	75
Kopā	160	29	61	315

Lai novērtētu iemeslus avotos atrodamās informācijas atšķirībām no zelta standarta, tika pētītas atšķirīgās un tikai avotā norādītās tīmekļa vietnes. Sākumā tika noteikti galvenie atšķirīgi norādīto tīmekļa vietņu iemesli (skat. 3.5. tabula).

3.5. tabula

Iemesli avotos norādīto tīmekļa vietņu atšķirībām no zelta standartu

Iemesls	**Biežums**
Pārvirzīšana uz atslēgas domēnu	11
Domēni ar vienādu saturu	5
Gramatiska rakstura kļūda avotā	4
Saturiska rakstura kļūda avotā	6
Domēni ar līdzvērtīgu atbilstību	3
Kopā	29

Pārvirzīšana uz atslēgas domēnu un vienāda satura domēni ir salīdzinoši nebūtiskas nepilnības. Gramatiska un saturiska rakstura kļūdas ir nozīmīgas un nelielā skaitā gadījumu bija grūti noteikt, kurš no domēniem labāk būtu iekļaujams zelta standartā, bet statistiski nozīmīgus spriedumus par kļūdu sadalījumu atsevišķos avotos ir grūti veikt salīdzinoši nelielā datu apjoma dēļ.

Tika noskaidroti arī galvenie iemesli, kāpēc ir salīdzinoši daudz uzņēmumu, kam informācijas avotos ir norādītas tīmekļa vietnes, kas netika iekļautas zelta standartā (skat. 3.6. tabula).

3.6. tabula

Iemesli zelta standartā neiekļautu tīmekļa vietņu norādīšanai avotos

Iemesls	**Biežums**
Uz pieprasījumiem neatbildošs DNS serveris	28
Nepastāvoša, neredzama atbilstība	21
Citi	12
Kopā	61

Būtiskākā problēma šajā gadījumā bija nestrādājoši DNS serveri, ko varētu risināt, regulāri izgūstot domēnu DNS ierakstu informāciju un apstrādājot tīmekļa vietnes, kurām tie ilglaicīgi nesniedz atbildes.

3.4. Rezultāti

Apskatītajos avotos savstarpēji pretrunīgas tīmekļa vietnes bija norādītas apmēram 10% gadījumu, kas liecina par līdzīgu izpratni par uzņēmumu atslēgas domēniem. Tomēr pastāvēja būtiskas atšķirības starp uzņēmumu kopām, kam avotos ir norādītas tīmekļa vietnes.

Salīdzinot ar zelta standartu, pretrunīgas tīmekļa vietnes bija norādītas apmēram 15% gadījumu un būtisks skaits avotos norādītu tīmekļa vietņu nebija iekļautas zelta standartā, jo tās ilgstoši nebija pieejamas tehnisku iemeslu dēļ.

4. KANDIDĀTDOMĒNU ATLASE

Piedāvātais risinājums tīmekļa vietņu meklēšanai ietver divus galvenos posmus:

1. kandidātdomēnu atlasi;
2. kandidātdomēnu klasifikāciju.

Kandidātdomēnu atlases posma galvenais uzdevums ir atrast domēnus, kas ar lielu varbūtību varētu būt uzņēmuma atslēgas domēni. Praktisku apsvērumu dēļ sākotnējos veiktajos eksperimentos maksimālais kandidātdomēnu skaits vienam uzņēmumam tiek ierobežots līdz desmit domēniem. Kandidātdomēnu klasifikācijas posma galvenais uzdevums ir precīzi novērtēt kandidātdomēnu piederību uzņēmumiem.

Darbā vairāk uzmanības tiek pievērsts tieši kandidātdomēnu klasifikācijas posmam, bet šajā nodaļā tiek aprakstītas arī vairākas metodes kandidātdomēnu atlasei.

4.1. Kandidātdomēnu atlases metodes

Kandidātdomēnu atlases posma kvalitāti raksturo **kandidātdomēnu atlases pārklājums** *($R_{CD\_Selection}$, Recall for Candidate Domain Selection)*, kas ir attiecība starp atslēgas domēnu skaitu uzņēmumiem atrastajos kandidātdomēnos un kopējo atslēgas domēnu skaitu zelta standartā (4.1).

$$R_{CD\_Selection} = \frac{Kandidātdomēnu\ altlasē\ atrasto\ atslēgas\ domēnu\ skaits}{Zelta\ standartā\ esošo\ atslēgas\ domēnu\ skaits} \quad \textbf{(4.1)}$$

Uzņēmuma kandidātdomēnu atlasi ir iespējams veikt ar vairākiem paņēmieniem. Kandidātdomēnu sarakstā var iekļaut:

- manuāli pieteiktus domēnus;
- citos informācijas avotos norādītus domēnus;
- ar meklēšanas dzinējiem atrastus domēnus;
- sociālajos tīklos norādītus domēnus.

Kandidātdomēnu atlasei darba ietvaros par pamatu tika nolemts izmantot meklētājdzinējus, jo to darbību ir iespējams automatizēt un pielāgot lielam uzņēmumu skaitam.

4.2. Kandidātdomēnu atlase ar meklētājdzinējiem

Kandidātdomēnu atlasei var izmantot gan eksistējošus, gan uzdevumam specifiski pielāgotus meklētājdzinējus. Abiem meklētājdzinēju veidiem ir savas priekšrocības un trūkumi. Ar eksistējošiem meklētājdzinējiem ir iespējams ātri iegūt sākotnējos rezultātus, bet tie ir grūtāk pielāgojami un liela daļa no tiem nav brīvi pieejami apjomīgai izmantošanai automatizētā režīmā. Specifiski veidoti meklētājdzinēji sniedz lielas pielāgošanas iespējas, bet to izveidošana un uzturēšana prasa lielākus resursus, kaut gan Latvijas tīmekļa vietņu gadījumā apstrādājamo datu apjoms gan nav tik liels.

Darba ietvaros tika nolemts sākumā par pamatu izmantot jau eksistējošus tīmekļa meklētājdzinējus vairāku iemeslu dēļ:

- daudziem meklētājdzinējiem ir pieejamas automatizētas datu izgūšanas saskarnes (kuru izmantošana gan var būt ierobežota vai pieejama tikai par samaksu lielu datu apjomu gadījumā);
- to grafiskās saskarnes bez būtiskiem ierobežojumiem parasti ir pieejamas bez maksas;
- ar eksistējošiem meklētājdzinējiem ir iespējams ātrāk iegūt pirmos rezultātus un novērtēt iespējamos ieguvumus no specializētu meklētājdzinēju izveides;
- nepieciešamības gadījumā vēlāk būtu iespējams izveidot arī pielāgotu vienkāršotu meklētājdzinēju Latvijas tīmekļa resursiem, balstoties uz sākotnēji gūtajiem rezultātiem.

Kandidātdomēnu atlases pārklājuma novērtēšanai tika izmēģināti vairāki meklēšanas dzinēji ar dažādiem meklēšanas iestatījumiem. Tika apskatīti pirmie 10 meklēšanas rezultāti, veicot kandidātdomēnu atlasi korpusa uzņēmumiem (skat. 4.1. tabula).

4.1. tabula

Kandidātdomēnu atlases pārklājums dažādiem meklētājdzinējiem

Kopējais korpusa atslēgas domēnu skaits	125
Google meklētājdzinēja[11] atrasto atslēgas domēnu skaits	104
Bing meklētājdzinēja[12] atrasto atslēgas domēnu skaits	99
Yahoo meklētājdzinēja[13] atrasto atslēgas domēnu skaits	86
Google meklētājdzinēja atrasto kandidātdomēnu atlases pārklājums	83.2%
Bing meklētājdzinēja atrasto kandidātdomēnu atlases pārklājums	79.2%
Yahoo meklētājdzinēja atrasto kandidātdomēnu atlases pārklājums	68.8%

Google un Bing meklētājdzinēji sasniedza līdzīgi augstu kandidātdomēnu atlases pārklājumu. Nelielus uzlabojumus sniedza arī vairāku meklētājdzinēju rezultātu apvienošana, bet tālākajos eksperimentos par pamatu tika izmantoti Google meklētājdzinēja rezultāti to lielākā kandidātdomēnu pārklājuma dēļ.

7. nodaļā tiek apskatīts arī alternatīvs risinājums kandidātdomēnu atlasei ar lokālu meklētājdzinēju, kas darba izstrādes laikā vēl tika mērogots, bet deva pozitīvus pirmos rezultātus.

4.3. Kandidātdomēnu filtrēšana

Veicot meklēšanu pēc uzņēmumu nosaukumiem, meklētājdzinēju rezultātos bieži parādās saites uz domēniem, kuros tiek apkopota informācija par uzņēmumiem (skat. 4.2. tabula).

[11] Pieejams: https://www.google.lv

[12] Pieejams: http://www.bing.com

[13] Pieejams: https://www.yahoo.com

4.2. tabula

Meklētājdzinēju rezultātos biežāk atrastie domēni korpusa uzņēmumiem

Domēns	Sk.	Domēns	Sk.	Domēns	Sk.
zl.lv	1687	infobits.lv	115	medicine.lv	44
firmas.lv	1487	balticexport.com	99	viss.lv	44
lursoft.lv	724	yello.lv	98	businessnetwork.lv	43
1188.lv	713	tsetso.info	98	news.lv	40
crediweb.lv	558	facebook.com	76	kurdarbs.lv	35
lvinfo.lv	405	parbaudi.lv	73	cv.lv	34
zo.lv	362	firmas.work	63	1189.lv	30
firmas.link	274	draugiem.lv	59	sudzibas.lv	30
1182.lv	270	db.lv	51	wikipedia.org	26
pilseta24.lv	212	visidarbi.lv	49	baltic-leads.com	24

Lai paātrinātu kandidātdomēnu pazīmju izguvi un klasifikāciju, atlasītajiem domēniem tiek pievienots biežāk atrasto domēnu filtrs, kas būtiski samazina apstrādājamo kandidātdomēnu skaitu (skat. 4.3. tabula), izslēdzot populārākos domēnus, kuros tiek apkopota informācija par lielu skaitu uzņēmumu.

4.3. tabula

Atrasto kandidātdomēnu skaits korpusa uzņēmumiem

Kopējais korpusa uzņēmumu skaits	1000
Ar Google meklētājdzinēju atrasto kandidātdomēnu skaits	3745
Ar Google meklētājdzinēju atrasto kandidātdomēnu skaits pēc to filtrēšanas	1387

5. KANDIDĀTDOMĒNU PAZĪMJU IZGUVE

Kandidātdomēnu pazīmju izguves posma galvenais uzdevums ir izgūt informāciju, kas varētu tikt izmantota kandidātdomēnu klasifikācijai atslēgas domēnos un neatslēgas domēnos. Informācijas izguvei tiek apskatīti vairāki informācijas avoti, kuru izmantošanu ir iespējams automatizēt.

Izgūtajām pazīmēm jābūt izmantojamām ar binārās klasifikācijas algoritmiem, nepieciešamības gadījumā tās nedaudz pielāgojot konkrētiem algoritmiem.

Izgūtās pazīmes pēc to veida var iedalīt vairākās grupās:

- bināras pazīmes;
- nepārtrauktas skaitliskas pazīmes;
- kategoriju pazīmes.

Šajā nodaļā apskatītās pazīmes tiek atsevišķi novērtētas pēc to sniegtās informācijas noderīguma, izmantojot vairākas metrikas. Ja pazīme tiek identificēta kā noderīga, tā tiek iekļauta nākamajā nodaļā aprakstītajā kandidātdomēnu klasifikācijas posmā.

5.1. Vispārīgi informācijas avoti

Apakšnodaļā tiek apskatīti vairāki vispārīgi un salīdzinoši brīvi pieejami informācijas avoti.

5.1.1. Kandidātdomēna nosaukums

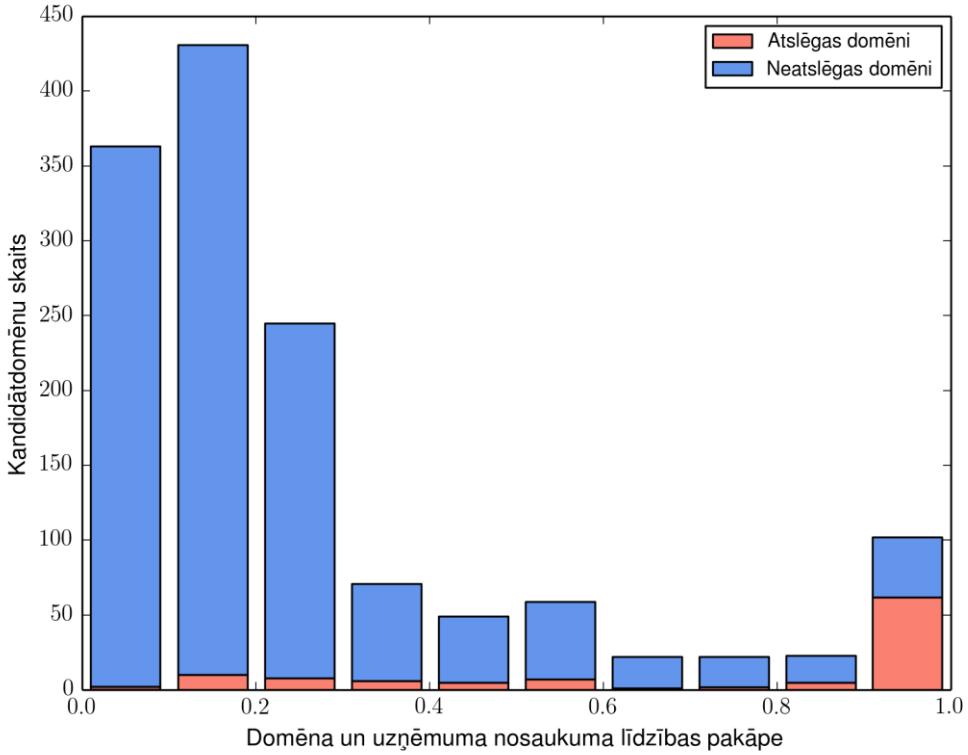
Vienkāršākā informācija, kas varētu tikt izmantota kandidātdomēnu klasifikācijai, ir atrodama to nosaukumos. Atslēgas domēna nosaukumam intuitīvi vajadzētu būt līdzīgam uzņēmuma nosaukumam. Domēna un uzņēmuma nosaukuma līdzībai tiek izmantota nepārtraukta skaitliska pazīme, jo šo līdzību var raksturot ar reālu skaitli, kas pieder intervālam [0; 1].

Kandidātdomēna un uzņēmuma nosaukuma līdzības novērtēšanai var izmantot dažādas simbolu virkņu līdzības novērtēšanas metodes [Navarro, 2001]. Sākumā tika

izmēģināts salīdzinoši vienkāršais Levenšteina attāluma algoritms. Pazīmes izgūšanai tika veiktas vairākās darbības:

1) tika normalizēti kandidātdomēnu nosaukumi, paturot tikai to augstākā līmeņa privātā domēna pirmo daļu;
2) tika normalizēti uzņēmumu nosaukumi, atmetot tukšumsimbolus un citus speciālos simbolus un pārveidojot akcentētus simbolus uz to neakcentētajiem variantiem;
3) tika novērtēts normalizēto simbolu virkņu Levenšteina attālums
4) iegūtais Levenšteina attālums tika normalizēts intervālā [0; 1], dalot garākās simbolu virknes garuma un Levenšteina attāluma starpību ar garākās simbolu virknes garumu.

Aprakstītās pazīmes vērtības tika izgūtas korpusa kandidātdomēniem un rezultātā iegūtajai pazīmei tika novērota vidēji spēcīga saistība ar atslēgas domēna pazīmi. Iegūtajai pazīmei tika izveidots grafisks attēlojums (skat. 5.1. attēls), kurā redzams kandidātdomēnu sadalījums atslēgas un neatslēgas domēnos atkarībā no uzņēmuma nosaukuma un domēna nosaukuma līdzības pakāpes.



5.1. attēls. Domēnu un uzņēmumu nosaukuma līdzības pakāpe atslēgas un neatslēgas domēniem

Pazīmei tika noteikti galvenie tās kvalitātes raksturlielumi (skat. 5.1. tabula).

5.1. tabula

Kandidātdomēna un uzņēmuma nosaukumu līdzības pazīmes galvenie raksturlielumi

Pīrsona korelācijas koeficienta vērtība (pret atslēgas domēna pazīmi)	0.507
Informācijas guvums *(information gain)*	0.146

Iegūtos rezultātus autors centās uzlabot, izmantojot citus tekstu salīdzināšanas algoritmus. Piemēram, ar Jaro–Winkler algoritmu [Cohen, Ravikumar, Fienberg, 2003] tika mēģināts augstāk novērtēt simbolu virknes, kas sakrīt to sākumdaļā, un mazāk sodīt simbolu dzēšanu no uzņēmumu nosaukumiem, kas parasti ir garāki par to domēnu nosaukumiem. Pazīmes raksturlielumus gan ar minētajiem eksperimentiem neizdevās būtiski uzlabot un tika nolemts tālāk izmantot iepriekš aprakstīto uz Levenšteina attālumu balstīto metodi.

5.1.2. Kandidātdomēna publiskais sufikss

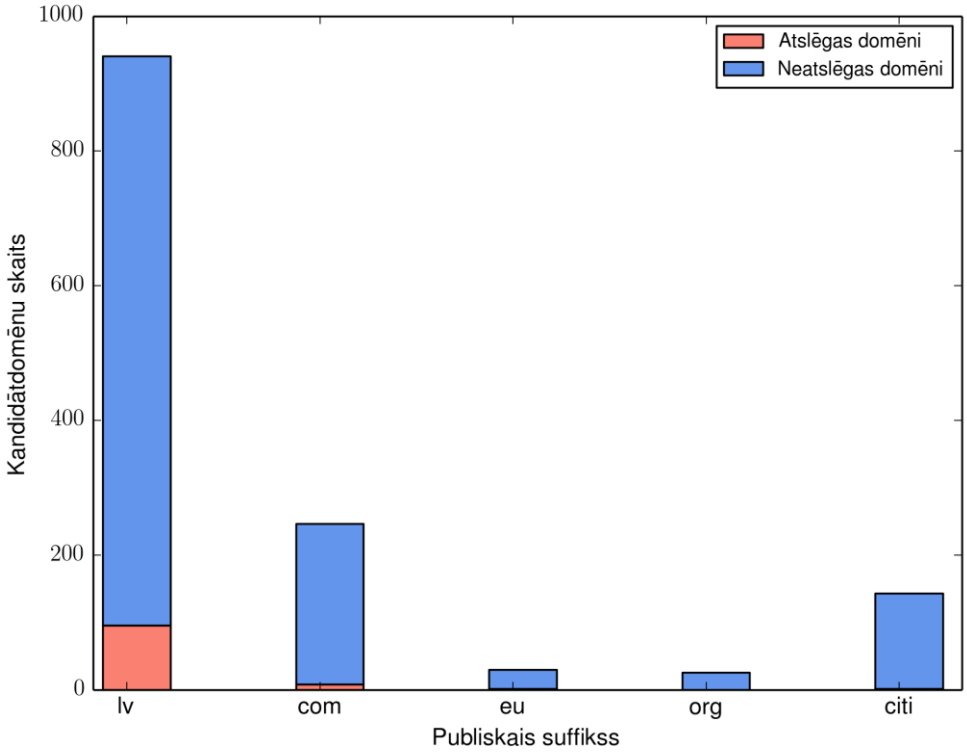
Publiskie sufiksi (piemēram .lv, .com, .co.uk) ir nosaukumi, kuros lietotāji var tieši reģistrēt savus privātos domēnus. Intuitīvi darba ietvaros apskatīto Latvijas uzņēmumu tīmekļa domēniem vajadzētu biežāk piederēt .lv sufiksam. Lai novērtētu šo hipotēzi, tika novērtēts korpusa atslēgas domēnu publisko sufiksu sadalījums (skat. 5.2. tabula).

5.2. tabula

Publisko sufiksu sadalījums atslēgas domēniem

Publiskais sufikss	**Atslēgas domēnu skaits**
.lv	108
.com	12
.eu	2
citi	3

No tabulas datiem var novērot, ka apmēram 85% atslēgas domēnu pieder .lv sufiksam. Lai novērtētu kandidātdomēnu sufiksa pazīmes noderīgumu, tika novērtēts atslēgas domēnu un neatslēgas domēnu īpatsvars biežāk atrastajiem kandidātdomēnu sufiksiem (skat. 5.2. attēls).



5.2. attēls. Publiskie sufiksi atslēgas un neatslēgas domēniem

No iegūtajiem rezultātiem var secināt, ka kandidātdomēni ar .lv sufiksu ir atslēgas domēni biežāk nekā kandidātdomēni ar citiem sufiksiem; bet novērotā sakarība nav ļoti spēcīga. Jāņem vērā, ka pazīmes noderīgums ir būtiski atkarīgs no atrasto kandidātdomēnu publisko sufiksu sadalījuma. Ja kandidātdomēnu skaits būtu lielāks un tie biežāk piederētu citiem publiskajiem sufiksiem, tad pazīmes noderīgums būtiski uzlabotos. Tāpēc tika nolemts pazīmi iekļaut nākamajos posmos apskatāmo pazīmju skaitā un tā tika modelēta ar kategoriju pazīmes palīdzību ar 3 iespējamām vērtībām:

- .lv sufikss;
- .com sufikss;
- cits sufikss.

Kandidātdomēna sufiksa kategorijas pazīmei tika noteikti tās galvenie kvalitātes raksturlielumi (skat. 5.3. tabula).

5.3. tabula

Kandidātdomēna publiskā sufiksa pazīmes galvenie raksturlielumi

Pīrsona korelācijas koeficienta vērtība (pret atslēgas domēna pazīmi)	0.116
Informācijas guvums (information gain)	0.015

5.1.3. Kandidātdomēnu sākumlapas nosaukums

Ar HTTP pieprasījumu izgūstot kandidātdomēna sākumlapas saturu, ir iespējams noskaidrot tīmekļa vietnes sākumlapas nosaukumu, ko satur HTML lapas nosaukuma tags. Līdzīgi kā kandidātdomēna nosaukumam tika izveidota nepārtraukta pazīme, kas raksturoja kandidātdomēna sākumlapas nosaukuma līdzību ar uzņēmuma nosaukumu. Iegūtā pazīme diezgan cieši korelēja ar uzņēmuma nosaukuma pazīmi un tās raksturlielumi bija nedaudz mazāk kvalitatīvi kā domēna nosaukuma pazīmei, tomēr tā tika iekļauta klasifikācijas posmā apskatāmo pazīmju sarakstā.

5.1.4. Citi informācijas avoti

Autors apskatīja vairākus citus vispārīgus informācijas avotus, no kuriem neizdevās izgūt pazīmes, kas varētu būtiski palīdzēt kandidātdomēnu klasifikācijai, bet daļa no izgūtās informācijas varētu tikt izmantota atslēgas domēnu informācijas uzturēšanai.

Kandidātdomēnu informācijas papildināšanai var izmantot to DNS ierakstu informāciju. Veicot DNS vaicājumus, ir iespējams noteikt:

- vai kandidātdomēna DNS serveris strādā korekti;
- kādus DNS serverus izmanto kandidātdomēns;
- kur ir izvietoti kandidātdomēna DNS serveri (pēc to IP adresēm).

[Mockapetris, 1987] un saistītie standarti sniedz detalizētu informāciju par datiem, kas ir pieejami, veicot DNS vaicājumus. Iepriekšējās nodaļās apskatītajiem informācijas avotiem bieža problēma bija tajos norādītajām tīmekļa vietnēm ilgstoši nestrādājoši DNS serveri. Piedāvātajā risinājumā ir paredzēts speciāli apstrādāt tīmekļa vietnes, kuru DNS serveri ilgstoši nesniedz atbildes.

Veicot HTTP pieprasījumus uz kandidātdomēnu sākumlapām, ir iespējams noskaidrot:

- vai kandidātdomēns atbild uz HTTP pieprasījumiem;
- vai kandidātdomēns tiek aktīvi izmantots (vai tajā atrodas tikai noklusētā domēnu reģistra informācija);
- kāda tipa un apjoma saturs ir izvietots tā sākumlapā.

Šī informācija netiek izmantota klasifikācijas posmā tieši, bet tā var tikt izmantota kandidātdomēnu un atslēgas domēnu informācijas uzturēšanai un prezentēšanai.

Kandidātdomēniem tika novērtēta arī to satura valoda, bet tā bija cieši saistīta ar iepriekš noteikto domēna publisko sufiksu un bija tehniski grūtāk izgūstama. Tāpēc tai netika veidota atsevišķa klasifikācijas pazīme. Līdzīgus rezultātus sniedza arī kandidātdomēnu serveru atrašanās vietas noteikšana pēc to IP adrešu informācijas, kas arī netika iekļauta klasifikācijas pazīmju sarakstā.

5.2. Kandidātdomēnu saturs

Kandidātdomēna piederības uzņēmumam noteikšanai var izmantot tā saturu. Piemēram, ja kandidātdomēna sākumlapā ir saite ar nosaukumu “Kontakti”, kas ved uz domēna lapu, kuras teksts satur uzņēmuma reģistrācijas numuru, tad intuitīvi var spriest, ka varbūtība tam, ka šis kandidātdomēns ir arī atslēgas domēns, ir diezgan augsta. Ar līdzīgiem paņēmieniem kandidātdomēna piederību uzņēmumam bieži mēģinātu noteikt arī cilvēks.

Uzņēmumu pamatdatos ietilpst vairāki atribūti, kuriem intuitīvi ar lielāku varbūtību vajadzētu atrasties kandidātdomēnos, kas ir arī atslēgas domēni:

- uzņēmuma nosaukums;
- uzņēmuma reģistrācijas numurs;
- uzņēmuma juridiskā adrese.

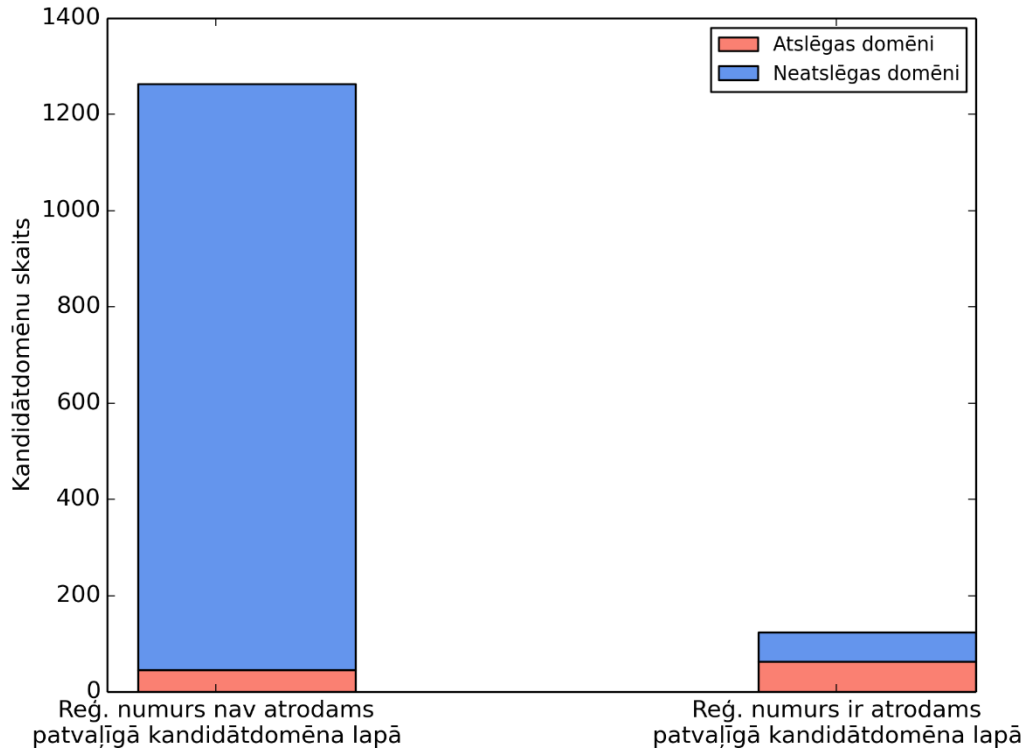
Apakšnodaļā tiek apskatītas pazīmes, kas var tikt izgūtas no kandidātdomēnu satura, automatizēti izgūstot to saturu.

5.2.1. Uzņēmuma reģistrācijas numurs

Sākumā tiek apskatīta uzņēmumu reģistrācijas numuru atrašanās kandidātdomēnu saturā, jo to atrašanās tīmekļa saturā ir tehniski vieglāk nosakāma.

Lai pārbaudītu hipotēzi, ka uzņēmuma reģistrācijas numurs biežāk ir atrodams atslēgas domēnā, sākumā tika veikta to meklēšana korpusa kandidātdomēniem. Lai ātrāk iegūtu pirmos rezultātus, sākumā tika izmantots Google meklētājdzinējs, kas nodrošina iespēju meklēt noteiktus atslēgas vārdus tikai konkrēta domēna saturā (visās tā lapās). Vēlāk tika apskatīta arī automatizēta satura izgūšana, kurai nav nepieciešams izmantot meklētājdzinējus

Tika novērota diezgan cieša saistība starp uzņēmuma reģistrācijas numura atrašanos kādā kandidātdomēna lapā un atslēgas domēniem (skat. 5.3. attēls).



5.3. attēls. Uzņēmuma reģistrācijas numura atrašanās patvaļīgā kandidātdomēna lapā atslēgas un neatslēgas domēniem

Reģistrācijas numura atrašanās kādā kandidātdomēna lapā tika modelēta ar bināru pazīmi, kurai tika noteikti tās galveni kvalitātes rādītāji (skat. 5.4. tabula). Iegūtās pazīmes

izgūšanu un kvalitāti gan ir iespējams būtiski optimizēt, kas tie apskatīts nākamajā apakšnodaļā.

5.4. tabula

Kandidātdomēna reģistrācijas numura atrašanās patvaļīgā kandidātdomēna lapā pazīmes galvenie raksturlielumi

Pīrsona korelācijas koeficienta vērtība (pret atslēgas domēna pazīmi)	0.503
Informācijas guvums (information gain)	0.103

5.2.2. Interesantas kandidātdomēnu lapas

Iepriekšējā apakšnodaļā tika apskatīta reģistrācijas numura atrašanās patvaļīgā kandidātdomēna lapā. Iegūtajai pazīmei gan ir saskatāmi vairāki trūkumi:

- ja reģistrācijas numurs tiek meklēts visās kandidātdomēnu lapās, ir nepieciešams pirms tam izgūt visu kandidātdomēna saturu, kas var ievērojami noslogot gan resursdatorus, gan informācijas izguvēju;
- visas kandidātdomēnu lapas tiek vērtētas vienādi, neņemot vērā to, ka ne visas lapas ar vienādu varbūtību satur uzņēmuma reģistrācijas numuru (uzņēmuma reģistrācijas numuram intuitīvi lielākai ietekmei būtu jābūt, ja tas atrodas domēna sākumlapā vai kontaktu lapā, nevis kādā grūti sasniedzamā domēna lapā).

Minēto trūkumu novēršanai var izmantot fokusētu tīmekļa apstaigāšanu [Chakrabarti, 1999], kas paredz tīmekļa satura izguvi pēc noteiktām pazīmēm, samazinot kopējo izgūstamo datu apjomu. Konkrētajā gadījumā minētās problēmas tika risinātas ar interesantu kandidātdomēnu lapu izgūšanu. Ar **interesantu kandidātdomēna lapu** darba kontekstā tiek saprasta kandidātdomēna lapa, kas ar lielu varbūtību satur uzņēmuma pamatdatus, ja kandidātdomēns ir atslēgas domēns. Intuitīvi interesanta kandidātdomēna lapa varētu būt, piemēram, tīmekļa lapa, uz kuru no domēna sākumlapas ved saite ar nosaukumu "Kontakti". Interesantas kandidātdomēnu lapas intuitīvi varētu ietvert uzņēmumu rekvizītu, kontaktinformācijas un vizītkartes (par mums) lapas.

Interesantu kandidātdomēnu lapu identificēšanu var uztvert kā bināru tīmekļa lapu klasifikācijas uzdevumu. Lai identificētu interesantas lapas ar saišu palīdzību, ir nepieciešams noteikt pazīmes, kas liecina, ka saite ved uz interesantu lapu. [Glove, 2002] apraksta kā saišu teksts, to konteksts (tuvumā esošie vārdi) u.c. pazīmes var tikt izmantotas tīmekļa lapu klasifikācijai. [Kan, 2005] detalizēti apraksta pazīmes, kas var tikt izmantotas tīmekļa lapu klasifikācijai, balstoties uz to vienoto resursu vietvārdi (URL).

Par pamatu to identificēšanai tika nolemts izmantot tieši kandidātdomēnu saišu un to struktūras informāciju, jo ar šo metodi interesantas lapas var identificēt, neizgūstot to saturu, kas būtiski samazina apstrādājamo datu apjomu. Konkrētajā gadījumā tika apskatīta salīdzinoši vienkārša interesantām kandidātdomēnu lapām raksturīgu simbolu virkņu meklēšana uz tām vedošo saišu tekstos un vietvāržos. Apskatot uzņēmumu reģistrācijas numurus saturošas kandidātdomēnu lapas, tika identificētas nepārtrauktu simbolu virknes, kuras visbiežāk saturēja uz tām vedošie saišu vietvārži un nosaukumi (skat. 5.5. tabula). Biežāko simbolu virkņu identificēšana tika balstīta uz pēc iespējas garāku un specifiskāku biežāk sastopamo kopīgo simbolu virkņu izvēli. Tam tika izmantots modificēts kopīgo n-gramu meklēšanas algoritms [Kim, 1994]

5.5. tabula

Biežāk sastopamās simbolu virknes uz interesantām lapām vedošu saišu vietvāržos un nosaukumos

Simbolu virkne	**Relatīvais biežums**
“kontakt”	0.62
“contact”	0.28
“rekviz”	0.19
“about”	0.13
“mums”	0.08

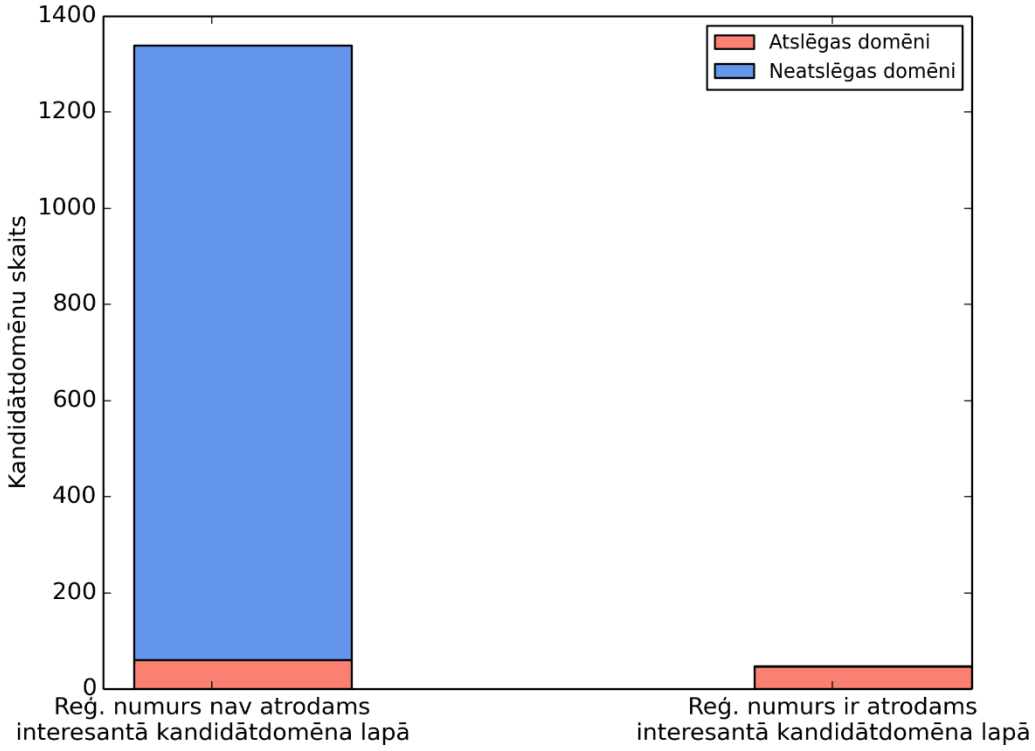
Interesantas lapas tika izgūtas, veicot vairākas darbības:

- izgūta un interesantu lapu vidū iekļauta tika kandidātdomēna sākumlapa

- izgūtas un interesantu lapu vidū iekļautas tika no kandidātdomēna sākumlapas sasniedzamās lapas, uz kurām veda saites, kuru teksts vai vietvārži saturēja iepriekš identificētās simbolu virknes.

Interesantu lapu izgūšanu vēlāk būtu iespējams uzlabot, izgūstot lielāku kandidātdomēnu satura un struktūras daļu.

Lai novērtētu interesantu kandidātdomēnu lapu izgūšanas ietekmi, tika izveidota binära pazīme, kas raksturoja uzņēmuma reģistrācijas numura atrašanos kādā kandidātdomēna interesantā lapā (skat. 5.4. attēls).



5.4. attēls. Uzņēmuma reģistrācijas numura atrašanās interesantā kandidātdomēna lapā atslēgas un neatslēgas domēniem

Salīdzinot ar iepriekš apskatīto pazīmi, kad reģistrācijas numurs tika meklēts visās kandidātdomēna lapās, interesantas lapas, kas saturēja uzņēmuma nosaukumu, praktiski vienmēr piederēja atslēgas domēnam. Nedaudz samazinājās atrasto reģistrācijas numuru kopējais skaits, bet tas ir saistīts arī ar to, ka Google meklētājdzinējs spēj atrast vairāk reģistrācijas numuru sava tehniskā pārākuma dēļ. Kopumā pazīmes kvalitatīvie

raksturlielumi būtiski palielinājās (skat. 5.6. tabula). Izgūstot tikai interesantas kandidātdomēnu lapas, no katra kandidātdomēna vidēji tika izgūtas mazāk kā 3 lapas, kas būtiski uzlabo risinājuma mērogošanas iespējas.

5.6. tabula

Kandidātdomēna reģistrācijas numura atrašanās interesantā kandidātdomēna lapā pazīmes galvenie raksturlielumi

Pīrsona korelācijas koeficienta vērtība (pret atslēgas domēna pazīmi)	0.637
Informācijas guvums (information gain)	0.132

5.2.3. Kandidātdomēnu satura izguves tehniskie aspekti

Tīmekļa rasmošana, kas ietver tā satura atklāšanu un izgūšanu, ir plaši pētīta problēma. Kandidātdomēnu satura izguvei konkrētajā gadījumā tika izmatots pielāgots vairākpavedienu tīmekļa apstaigātājs, kam par pamatu tika izmantota Crawler4j[14] bibliotēka. Tā konfigurācijā un pielāgošanā tika ņemti vērā efektīvas un drošas tīmekļa satura izguves principi, par kuriem diezgan detalizētu apkopojumu sniedz [Castillo, 2005]. Šie pamatprincipi tiek ievēroti, veicot vairākas darbības:

- tiek ņemts vērā robots.txt faila saturs, ar kura palīdzību tīmekļa vietņu īpašnieki var ierobežot tīmekļa apstaigātājiem paredzēto saturu
- starp lapu izgūšanām no viena kandidātdomēna un resursdatora tiek ievērota vismaz vienu sekundi ilga pauze;
- lai pārmērīgi nenoslogotu resursdatorus, tiek ierobežots maksimālais no viena kandidātdomēna izgūstamo lapu skaits;
- tiek pielāgots izmantoto pavedienu skaits, vienlaicīgo savienojumu skaits un kļūdu apstrāde;

[14] Tīmekļa satura izguves bibliotēka. Pieejams: https://github.com/yasserg/crawler4j

- kandidātdomēni tika apvienoti sērijās, lai nodrošinātu paralēlu apstrādi, pārāk nenoslogojot tos individuāli.

Praktiskām apstaigātāja modifikācijām Java programmēšanas valodā tika izmantoti arī ieteikumi no [Zeinalipour-Yazti, Dikaiakos, 2001]. Pirms katras lapas (izņemot kandidātdomēna sākumlapas) satura izgūšanas tiek pārbaudīts, vai uz to vedošā saite atbilst interesantas tīmekļa lapas definīcijai. Izgūtās kandidātdomēnu lapas un to metadati tika saglabāti Apache Solr[15] meklēšanas serverī, kur tiek veikta to satura indeksācija. Par katru izgūto lapu tiek saglabāta informācija, kas ietver:

- lapas vietvārdi un nosaukumu;
- lapas tekstuālo saturu;
- uz to vedošo saišu informāciju;
- tās dziļumu kandidātdomēnā (attālumu no sākumlapas);
- saņemtās HTTP galvenes informāciju;
- izgūšanas datumu un laiku;

5.2.4. Uzņēmuma juridiskā adrese

Uzņēmumu juridisko adrešu pazīmju izguvi apgrūtina to dažādie pieraksta formāti, kas var būt īpaši atšķirīgi tīmekļa vidē. Lai risinātu šo problēmu, uzņēmumu juridiskās adreses tika sadalītas vairākās daļās:

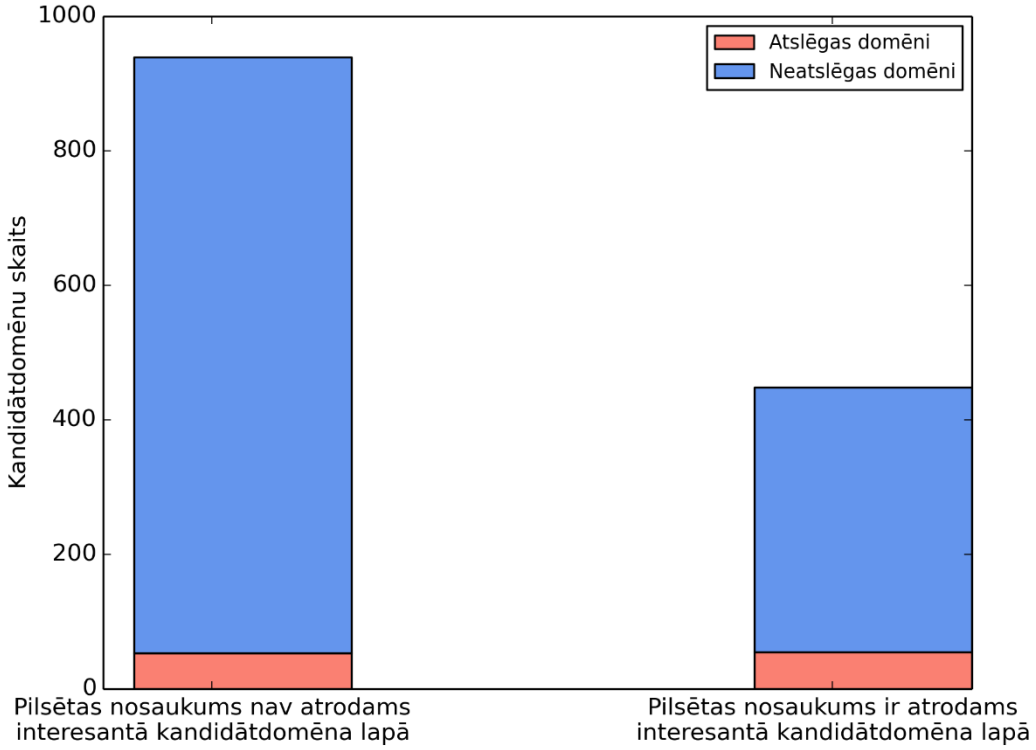
- pilsētas vai novada nosaukums;
- pasta indekss;
- ielas nosaukums un ēkas numurs vai ēkas nosaukums.

Katrai no šīm daļām tika izveidota atsevišķa klasifikācijas pazīme. Kopumā, apskatot tikai interesantas kandidātdomēnu lapas, tika gūti līdzīgi uzlabojumi kā iepriekš apskatītajai reģistrācijas numura pazīmei, tāpēc arī šīm pazīmēm tika izgūtas tikai interesantas

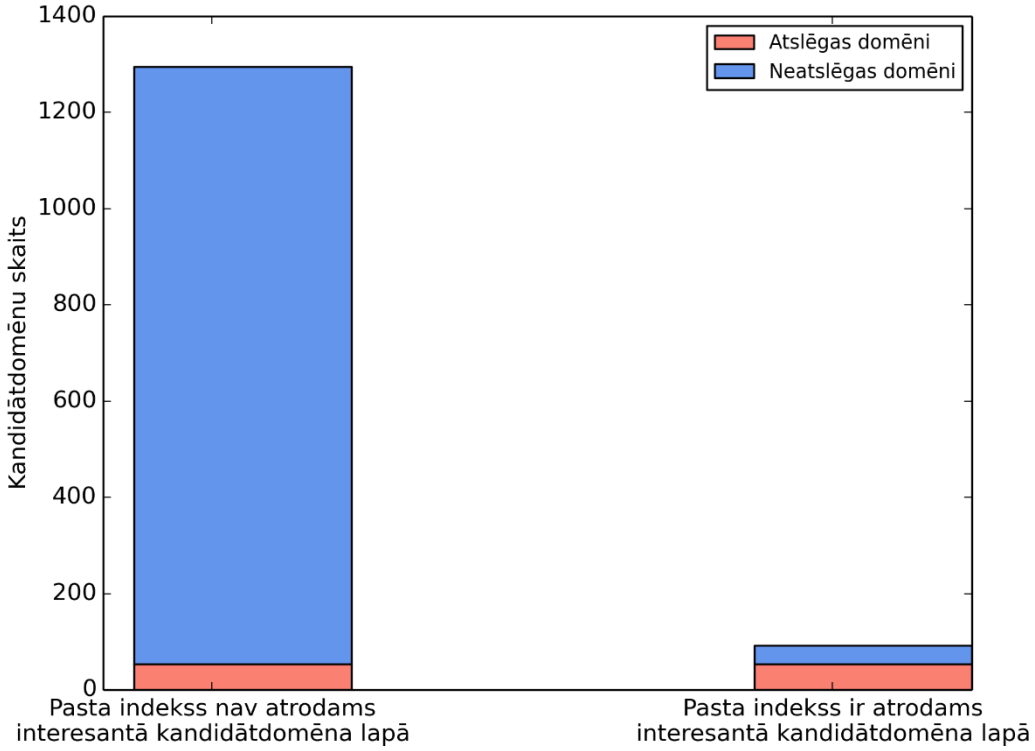
[15] Tekstuālas informācijas indeksēšanas un meklēšanas serveris. Pieejams: http://lucene.apache.org/solr/

kandidātdomēnu lapas. Visām minētajām juridiskās adreses daļām tika veidotas bināras klasifikācija pazīmes.

Starp pilsētas nosaukuma atrašanos kādā interesantā kandidātdomēna lapā un atslēgas domēniem tika novērota salīdzinoši vāja saistība (skat. 5.5. attēls). Lielā daļā gadījumu uzņēmuma pilsētas nosaukumu saturēja arī neatslēgas domēni. Spēcīgāka saistība tika novērota starp pasta indeksa atrašanos atslēgas un neatslēgas domēnos (skat 5.6. attēls).



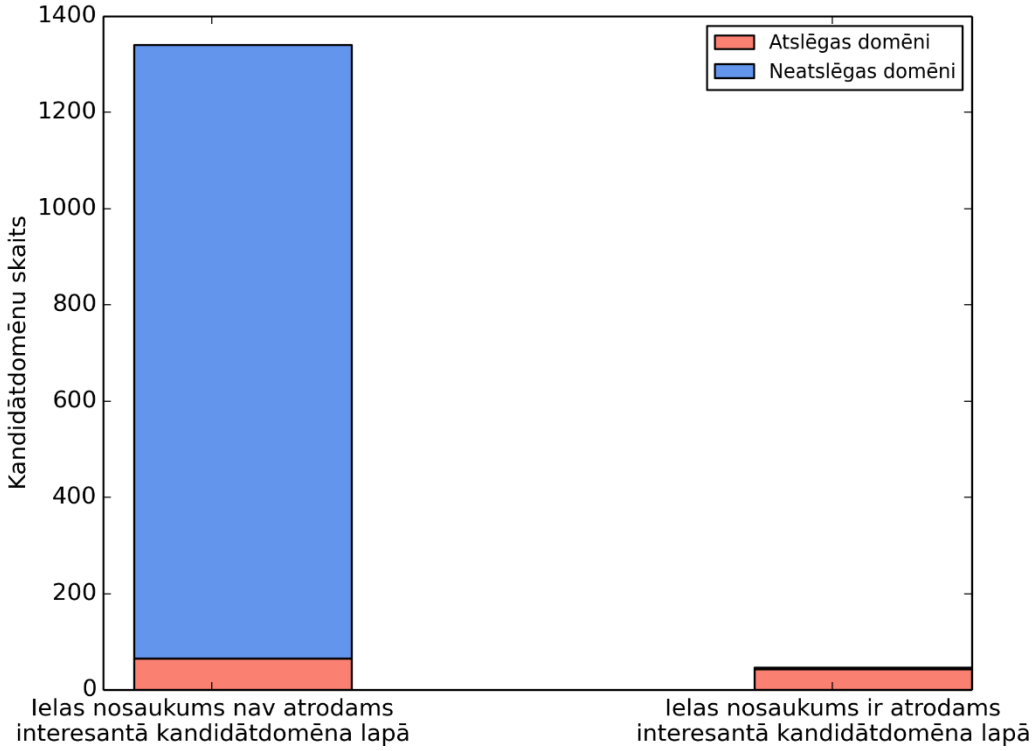
5.5. attēls. Uzņēmuma pilsētas nosaukuma atrašanās interesantā kandidātdomēna lapā atslēgas un neatslēgas domēniem



5.6. attēls. Uzņēmuma pasta indeksa atrašanās interesantā kandidātdomēna lapā atslēgas un neatslēgas domēniem

Ielas nosaukuma un ēkas numura meklēšanai tika paredzēti vairāki to pieraksta varianti. Alternatīvo pierakstu ģenerēšanai, ielas adrese sākumā tika sadalīta atsevišķās daļās, kas ietvēra tās nosaukumu, veidu, ēkas numuru, korpusa nosaukumu un dzīvokļa numuru.

Starp uzņēmuma ielas nosaukuma atrašanos kādā interesantā kandidātdomēna lapā un atslēgas domēniem tika novērota spēcīga saistība (skat. 5.7. attēls).



5.7. attēls. Uzņēmuma ielas nosaukuma atrašanās interesantā kandidātdomēna lapā atslēgas un neatslēgas domēniem

Kopumā no juridiskās adreses pazīmēm lielāko ieguvumu deva tieši ielas nosaukuma meklēšana, bet atsevišķos gadījumos arī pasta indeksa vai pilsētas nosaukuma pazīme varētu palīdzēt klasifikācijas posmā; tāpēc arī tās tika iekļautas sīkākai izskatīšanai kandidātdomēnu klasifikācijas posmā.

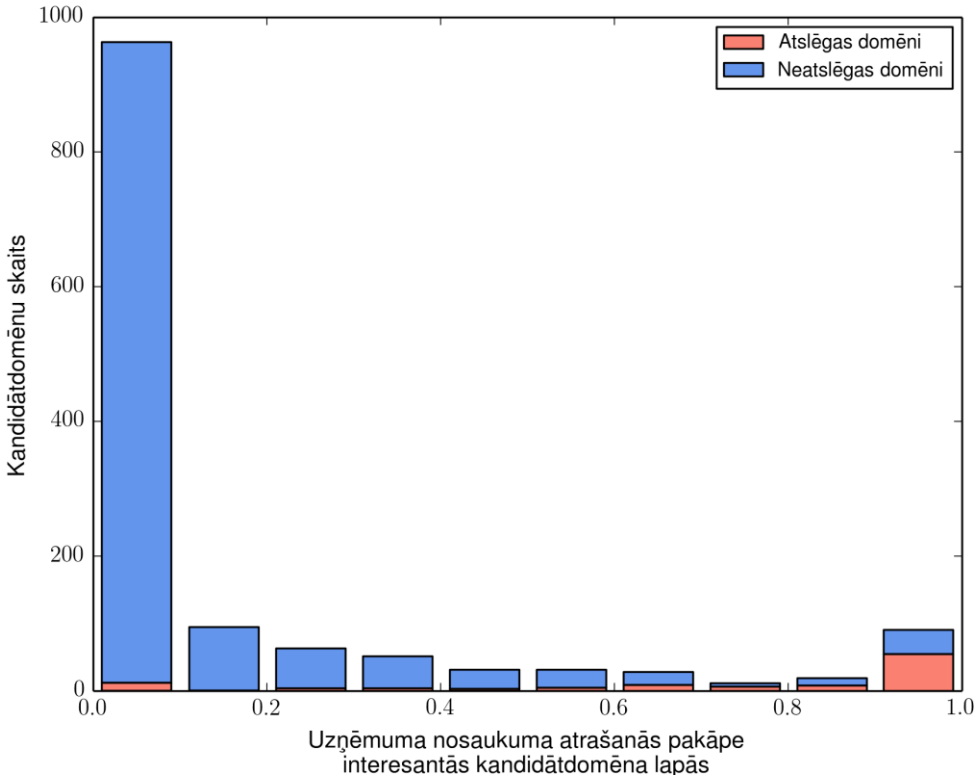
5.2.5. Uzņēmuma nosaukums

Atslēgas domēna lapām intuitīvi vajadzētu saturēt uzņēmuma nosaukumu biežāk nekā neatslēgas domēna lapām. Uzņēmumu nosaukumu pieraksti var būtiski atšķirties, tāpēc tika izveidota nepārtraukta pazīme, kas raksturotu kandidātdomēna nosaukuma atrašanās pakāpi tā saturā.

Arī šai pazīmei tika apskatītas tikai interesantas kandidātdomēnu lapas. Katrai lapai tika novērtēta uzņēmuma nosaukuma atrašanās pakāpe tajā, salīdzinot ar citām kandidātdomēnu lapām, izmantojot Apache Solr meklēšanas servera sniegtos rezultātus.

Kandidātdomēna novērtējums tika iegūts, ņemot tā lapu novērtējumu vidējo vērtību. Iegūtais novērtējums tika normalizēts intervālā [0, 1].

Starp uzņēmuma nosaukuma atrašanās pakāpi interesantās kandidātdomēnu lapās un atslēgas domēniem tika novērota salīdzinoši spēcīga saistība (skat. 5.8. attēls).



5.8. attēls. Uzņēmuma nosaukuma atrašanās pakāpe interesantās kandidātdomēna lapās atslēgas un neatslēgas domēniem

5.3. Centralizēti informācijas avoti

Kandidātdomēnu klasifikācijai teorētiski var izmantot gandrīz jebkādu informāciju, kas ir pieejam citos informācijas avotos. Tomēr piekļuve šiem avotiem bieži ir ierobežota dažādu iemeslu dēļ. Šajā apakšnodaļā tiek apskatīti vairāki centralizēti informācijas avoti, no kuriem ir iespējams izgūt pazīmes, kas var tikt izmantotas to klasifikācijai.

5.3.1. WHOIS ieraksti

Domēnu reģistru uzturētajos WHOIS ierakstos [Daigle, VeriSign, 2004] parasti ir atrodama informācija par uzņēmumiem, kas ir to lietotāji. Par Latvijas domēniem šo

informāciju sniedz Latvijas Universitātes Matemātikas un informātikas institūta Tīkla risinājumu daļa (NIC)[16]. Šī informācija Uzņēmuma informācijas atrašanās kandidātdomēna WHOIS ierakstā gan negarantē to, ka kandidātdomēns ir uzņēmuma atslēgas domēns. Uzņēmumi var uzticēt domēna vārdu reģistrēšanu trešajām pusēm, to informācija var būt novecojusi vai nepilnīga.

Daudziem WHOIS serveriem ir noteikti ierobežojumi šīs informācijas izgūšanai un izmantošanai, kas var atšķirties gan valstu, gan atsevišķu serveru līmenī. Praksē gan šo ierobežojumu robežas nav tik viegli nosakāmas un WHOIS informāciju lielos apjomos sniedz daudzi trešo pušu pārstāvji.

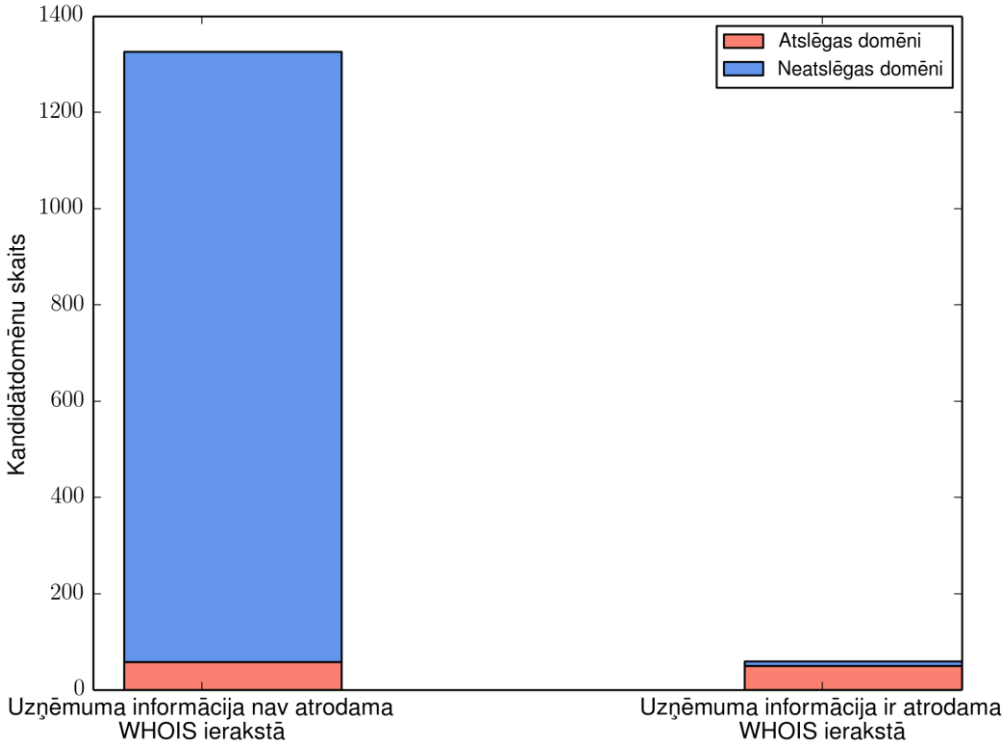
Neraugoties uz iepriekš minētajiem apstākļiem, darbā tiek apskatīts šis informācijas avots vairāku iemeslu dēļ:

- ja konkrētā gadījumā (piemēram, valsts ietvaros) šī informācija nav pieejama izmantošanai, tad atšķirīgo ierobežojumu dēļ tā varētu būt pieejama izmantošanai citos gadījumos;
- konkrētajā pielietojumā tiek apskatīta tikai uzņēmumu pamatinformācija, neinteresējoties par citu ierakstos atrodamo informāciju.

Katram kandidātdomēnam tika pārbaudīts, vai tā WHOIS ieraksts satur uzņēmuma reģistrācijas numuru vai nosaukumu. Šādi tika izveidota bināra pazīme, kas raksturo to, vai kandidātdomēna WHOIS ieraksts atbilst uzņēmuma pamatinformācijai.

Starp WHOIS ierakstu atbilstību un kandidātdomēna piederību atslēgas domēniem tika novērota spēcīga saistība (skat. 5.9. attēls).

[16] Pieejams: https://www.nic.lv/whois



5.9. attēls. Kandidātdomēnu *WHOIS* ierakstu atbilstība atslēgas un neatslēgas domēniem

Ja WHOIS ierakstā ir atrodama kandidātdomēna informācija, tad diezgan droši var apgalvot, ka kandidātdomēns ir arī atslēgas domēns. WHOIS ieraksta atbilstības pazīmei tika noteikti tās galvenie raksturlielumi (skat. 5.7. tabula).

5.7. tabula

Kandidātdomēna publiskā sufiksa pazīmes galvenie raksturlielumi

Pīrsona korelācijas koeficienta vērtība (pret atslēgas domēna pazīmi)	0.699
Informācijas guvums (information gain)	0.161

5.3.2. Meklētājdzinēji

Būtisku informāciju par kandidātdomēnu piederību uzņēmumiem var sniegt populārākie vispārīgie meklētājdzinēji vai speciāli veidoti meklētājdzinēji. Darba ietvaros tika apskatīti ar Google meklētājdzinēju atlasīti korpusa kandidātdomēni; tāpēc var uzskatīt, ka šajā gadījumā tā ir visiem apskatītajiem kandidātdomēniem piemītoša pazīme. Kā atsevišķas pazīmes varētu apskatīt arī kandidātdomēnu parādīšanos citu meklētājdzinēju rezultātos, veicot meklēšanu pēc to nosaukumiem, reģistrācijas numuriem vai citiem pamatdatiem.

Jāņem vērā, ka populārāko meklētājdzinēju izmantošanai lielā apjomā pastāv ierobežojumi. Tāpēc ir vērts apskatīt arī atvērtākus un manuāli veidotus meklētājdzinēju risinājumus. Lielos apjomos automatizētai tīmekļa satura meklēšanai ir pieejama FARO programmsaskarne[17]. Jāatzīmē, ka latviešu valodai tā sniedza salīdzinoši zemas kvalitātes rezultātus.

Autors darba beigās sāka strādāt arī pie vienkārša meklētājdzinēja izveides kandidātdomēnu atlasei, kas deva pozitīvus sākotnējos rezultātus. Sīkāk šie eksperimenti ir aprakstīti darba 7. nodaļā.

[17] Pieejams: http://developer.faro.com

6. KANDIDĀTDOMĒNU KLASIFIKĀCIJA

Kandidātdomēnu klasifikācijas posma uzdevums ir klasificēt atrastos kandidātdomēnus atslēgas un neatslēgas domēnos. Katra kandidātdomēna klasifikācijas rezultāts ir bināra vērtība, kas norāda, vai kandidātdomēns ir atslēgas domēns. Izgūtie kandidātdomēni tika automātiski anotēti, izmantojot iepriekš izveidotā korpusa uzņēmumu informāciju (skat. 6.1. tabula).

6.1. tabula

Korpusa kandidātdomēnu kopu raksturojoši rādītāji

Korpusa uzņēmumu skaits	1000
Korpusa kandidātdomēnu skaits	1387
Korpusa kandidātdomēnu skaits, kas ir atslēgas domēni	106
Korpusa kandidātdomēnu skaits, kas nav atslēgas domēni	1281

Kandidātdomēnu klasifikācijai tika izmantotas iepriekš izgūtās kandidātdomēnu pazīmes. Binārās klasifikācijas problēmas ir plaši pētītas un [Hastie, Tibshirani, Friedman, 2009] sniedz apkopojumu par biežāk izmantotajām klasifikācijas metodēm. Sākumā tika izmēģināti vienkāršāki modeļi, kas ir vieglāk interpretējami, bet varbūt nesasniedz labākos rezultātus. Pēc tam tika apskatīti un salīdzināti arī citi klasifikācijas modeļi.

Izvēlēto modeļu apmācībai tika izmantotas iepriekš izgūtās kandidātdomēnu pazīmes, paredzot iespēju tās nedaudz pielāgot konkrētu algoritmu specifikai. Klasifikācijas algoritmu novērtēšanai tika izmantotas scikit-learn[18] un Weka[19] datu analīzes un mašīnmācīšanās bibliotēkas.

[18] Datu analīzes un mašīnmācīšanās bibliotēka. Pieejams: http://scikit-learn.org

[19] Datu analīzes un mašīnmācīšanās bibliotēka. Pieejams: http://www.cs.waikato.ac.nz/ml/weka

6.1. Klasifikācijas pazīmes

Pirms konkrētu algoritmu pielietošanas tika apkopota informācija par iepriekš izgūtajām kandidātdomēnu pazīmēm (skat. 6.2. tabula).

6.2. tabula

Izgūto kandidātdomēnu klasifikācijas pazīmju raksturojums

Pazīmes apraksts	**Pazīmes tips**
Uzņēmuma nosaukuma un kandidātdomēna nosaukuma līdzības pakāpe	nepārtraukta
Uzņēmuma nosaukuma un kandidātdomēna sākumlapas nosaukuma līdzības pakāpe	nepārtraukta
Kandidātdomēna publiskais sufikss	kategoriju
Uzņēmuma nosaukuma atrašanās pakāpe kandidātdomēna saturā	nepārtraukta
Uzņēmuma reģistrācijas numura atrašanās kandidātdomēna saturā	binära
Uzņēmuma pilsētas nosaukuma atrašanās kandidātdomēna saturā	binära
Uzņēmuma pasta indeksa atrašanās kandidātdomēna saturā	binārā
Uzņēmuma ielas, ēkas nr./nos. atrašanās kandidātdomēna saturā	binära
Uzņēmuma informācijas atrašanās kandidātdomēna *WHOIS* ierakstā	binära

6.2. Loģistiskā regresija

Kā viena no pirmajām klasifikācijas metodēm tika izvēlēta loģistiskā regresija, kurai ir vairākas priekšrocības attiecībā uz risināmo problēmu, jo tā:

- atbalsta binäras, kategoriju un nepārtrauktas pazīmes;
- sniedz salīdzinoši vienkāršus modeļus, kurus noteiktās robežās ir iespējams arī interpretēt;
- ļauj diezgan vienkārši veikt ne tikai bināro klasifikāciju, bet arī novērtēt iegūto paredzējumu varbūtības;
- ļauj strādāt ar nestandartizētām pazīmēm, kas var nebūt normāli sadalītas;
- nodrošina vairākas korelējošu un nenozīmīgu pazīmju izņemšanas un regularizācijas iespējas.

Ņemot vērā, ka anotēto datu apjoms nebija liels, apmācībai un validācijai pamatā tika izmantota šķērsvalidācija. Salīdzinoši neliela daļa datu (30%) tika izmantota rezultātu testēšanai. Testēšanas datu apjoms gan ir salīdzinoši neliels un tās galvenais mērķis ir noteikt, vai šķērsvalidācijas posmos nav pieļautas būtiskas kļūdas un vai tiek iegūti līdzīgi rezultāti kā šķērsvalidācijā. Iegūto rezultātu novērtēšana lielākā apjomā, salīdzinot tos ar citiem informācijas avotiem, tiek veikta nākamajā nodaļā.

Loģistiskajai regresijai visas pazīmes tika atstātas praktiski nemainīgas; pielāgota tika tikai publiskā sufiksa kategorijas pazīme, kuras 3 iespējamās vērtības tika aizstātas ar divām binārām pazīmēm:

- kandidātdomēns pieder *.lv* sufiksam;
- kandidātdomēns pieder *.com* sufiksam.

Šķērsvalidācijas paraugi tika sadalīti 5 daļās. Rezultātu novērtēšanas metriku izvēlē tika ņemts vērā, ka rezultātu klašu īpatsvari paraugos ir būtiski nevienlīdzīgi un saskaņā ar [Davis, Goadrich, 2006] šādos gadījumos kā galveno kvalitātes rādītāju nav ieteicams izmantot kopējo rezultātu precizitāti. Nevienlīdzīgu klašu gadījumā ir ieteicams izmantot kvalitātes rādītājus, kurus būtiski ietekmē arī pozitīvās klases rezultāti, piemēram, laukumu zem precizitātes un pārklājuma līknes pozitīvo rezultātu klasei (PR AUC). Konkrētajā gadījumā kā galveno metriku tika nolemts izmantot F1-mēru pozitīvo rezultātu klasei, jo tas sniedz līdzīgus rezultātus un vairāk optimizē rezultātus pie augstāka precizitātes līmeņa.

Loģistiskās regresijas trūkumi var parādīties gadījumos, kad starp pazīmēm ir novērojamas būtiskas korelācijas saites. Ņemot to vērā, tika nolemts pielietot loģistiskās regresijas regularizācijas algoritmus, kas palielina iegūto modeļu stabilitāti. Izmantotā scikit-learn mašīnmācīšanās bibliotēka atbalstīja regularizācijas algoritmu un to parametru novērtēšanu. Tāpēc šķērsvalidācijas posmā, optimizējot iegūtā F1-mēra vērtību, tika pielāgoti vairāki regularizācijas parametri:

- izmantotais regularizācijas algoritms;
- izmantotā regularizācijas algoritma ietekmes pakāpe uz modeli.

Tika apskatītas trīs galvenās modeļu regularizācijas metodes. L1 regularizācija dod priekšroku modeļiem, kuru koeficienti pēc to absolūtajām vērtībām ir mazāki un kuros būtiski korelējošām un mazāk vērtīgām pazīmēm koeficienti ir vienādi ar nulli. L2 regularizācija līdzīgi dod priekšroku modeļiem, kuru koeficienti pēc to absolūtajām vērtībām ir mazāki, bet pilnībā neizslēdz arī būtiski korelējošas pazīmes. Elastic net regularizācija kombinē abas iepriekš minētās metodes [Zou, Hastie, 2005].

6.2.1. Šķērsvalidācija

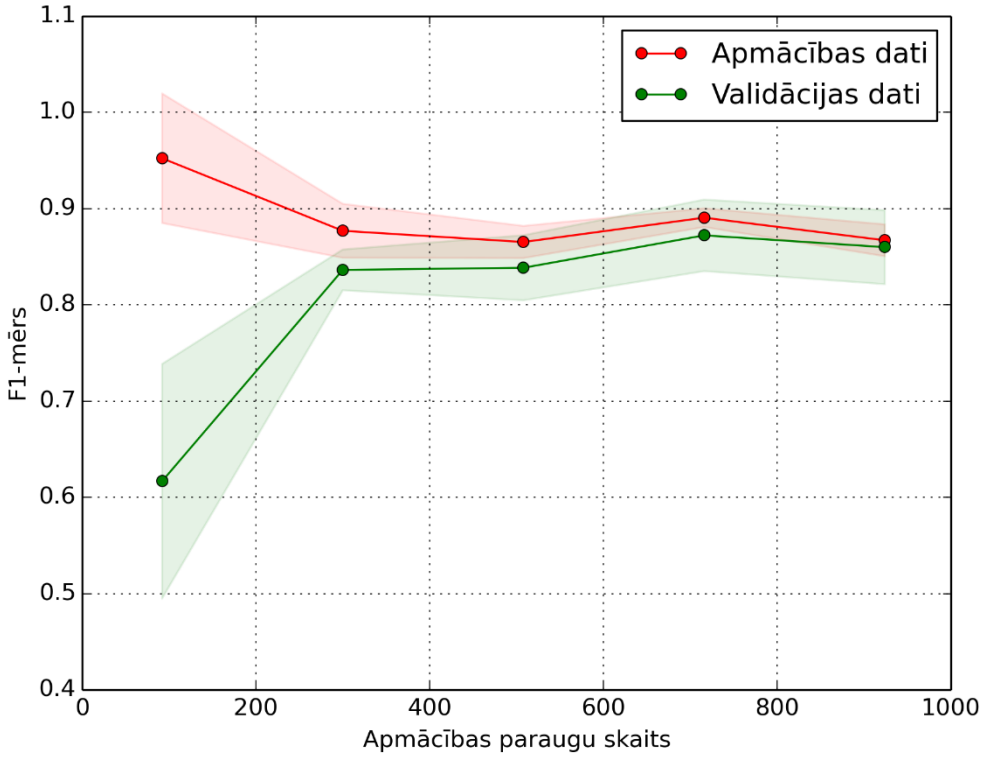
Šķērsvalidācijas kopas tika veidotas, stratificēti sadalot visu apmācības datu kopu 5 līdzīga apjoma daļās, lai tajās būtu līdzīgs pozitīvās rezultātu klases paraugu skaits (skat. 6.3. tabula).

6.3. tabula

Šķērsvalidācijas galvenie raksturlielumi

Šķērsvalidācijā iekļauto paraugu kopējais skaits	970
Vidējais apmācības paraugu skaits	776
Vidējais validācijas paraugu skaits	194
Labākais atrastais regularizācijas algoritms	L2
Labākā regularizācijas algoritma ietekmes pakāpe	4.0

Lai novērtētu šķērsvalidācijas procesu, papildus tika salīdzināts iegūtais F1-mērs pie dažādiem apmācības un validācijas datu apjomiem (skat. 6.1. attēls).



6.1. attēls. F1-mērs šķērsvalidācijas apmācības un validācijas datiem pie dažāda apmācības paraugu skaita

6.1. attēlā ir redzams F1-mērs apmācības un validācijas datiem, ja pie dažādiem datu apjomiem tiek veikta šķērsvalidācija. Iekrāsotie laukumi apzīmē iegūto rezultātu novirzi no vidējās vērtības 1 standartnovirzes attālumā.

Kopumā grafikā ir novērojamas vairākas pozitīvas tendences:

- atšķirības starp apmācības un validācijas kopām kļūst salīdzinoši nelielas jau pie apmēram 500 paraugiem; no tā var secināt, ka nenotiek pārāk liela modeļa pielāgošanās apmācības paraugiem;
- iegūtie rezultāti ir salīdzinoši stabili un iegūtā F1-mēra vērtība validācijas datiem pakāpeniski pieaug, palielinoties apmācības datu apjomam, kamēr apmācības datiem vērtība samazinās.

Šķērsvalidācijai tika noteikti galvenie tās rezultāti (skat. 6.4. tabula). Neatslēgas domēnu klasei tika novēroti augsti precizitātes un pārklājuma rādītāji, jo lielākā daļa

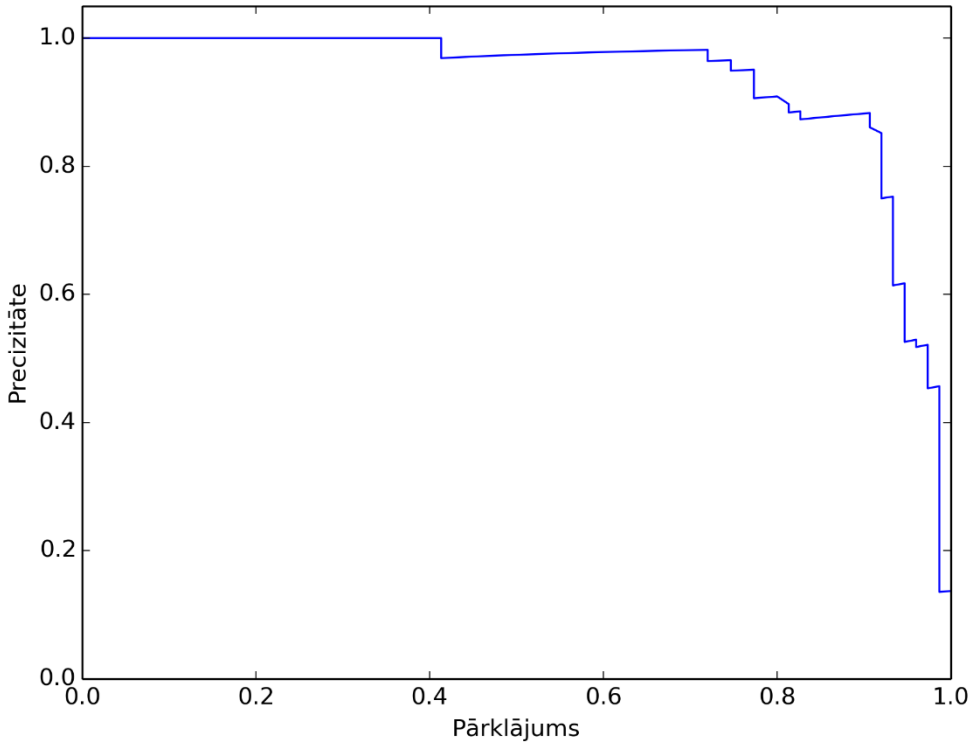
elementu pieder šai klasei. Atslēgas domēniem optimizētais F1-mērs sasniedza 87.8% ar 95.3% precizitāti un 81.3% pārklājumu.

6.4. tabula

Šķērsvalidācijas galvenie rezultāti atslēgas un neatslēgas domēnu klasēm

Rezultātu klase	**Precizitāte**	**Pārklājums**	**F1-mērs**	**Paraugu skaits**
Neatslēgas domēnu klase	98.5%	99.7%	99.1%	895
Atslēgas domēnu klase	**95.3%**	**81.3%**	**87.8%**	**75**
Kopā	98.2%	98.2%	98.2%	970

Lai šķērsvalidācijas rezultātus būtu vieglāk interpretēt, tiem tika izveidota precizitātes un pārklājuma līkne atslēgas domēnu rezultātu klasei (skat. 6.2. attēls).



6.2. attēls. Šķērsvalidācijas rezultātu precizitātes un pārklājuma līkne atslēgas domēnu rezultātu klasei

No iegūtās precizitātes un pārklājuma līknes var secināt, ka risinājums sniedz salīdzinoši augstu precizitāti līdz tiek sasniegts apmēram 80% pārklājums. Pēc tam sākas straujš precizitātes kritums un 100% pārklājums tiek sasniegts ar mazāk kā 20% precizitāti.

Var spriest, ka lielākajā daļā gadījumu būtu vērts apskatīt, rezultātus 75% - 85% pārklājuma līmenī un 85% - 95% precizitātes līmenī, ko apstiprina arī iegūtais F1-mērs.

Tālākajās nodaļās prognozētais kandidātdomēnu atlases pārklājums atslēgas domēnu klasei tiek apzīmēts ar $R_{CD\_Classification}$ un precizitāte ar $P_{CD\_Classification}$.

6.2.2. Modeļa raksturojums

Iegūto loģistiskās regresijas modeli var aprakstīt ar formulas palīdzību un būtiska modeļa priekšrocība ir tā, ka to iespēju robežās ir var arī interpretēt (skat. 6.5. tabula).

6.5. tabula

Iegūtie loģistiskās regresijas koeficienti un iespējamību attiecības rezultātu atslēgas domēnu klasei

Pazīmes nosaukums	**Mainīgā nosaukums**	**Koeficients**	**Iespējamību attiecība**
Kandidātdomēna nosaukums	*DOM*	3.363	28.886
Kandidātdomēna sākumlapas nosaukums	*TITLE*	2.880	17.814
Kandidātdomēna *.lv* sufikss	*SUFLV*	0.036	1.037
Kandidātdomēna *.com* sufikss	*SUFCOM*	-0.314	0.730
Uzņēmuma nosaukums	*NOS*	2.158	8.654
Reģistrācijas numurs	*NUM*	4.965	143.329
Pilsētas nosaukums	*PIL*	0.002	1.002
Pasta indekss	*IND*	0.441	1.555
Ielas nosaukums	*IEL*	2.721	15.193
WHOIS ieraksts	*WHO*	3.817	45.449
	Brīvais loceklis	-5.876	

No tabulas 6.5. iespējamību attiecības kolonnas var nolasīt, cik apmēram reizes palielinās kandidātdomēna piederības atslēgas domēniem iespējamība, ja par vienu vienību palielinās atbilstošā mainīgā vērtība, pārējiem mainīgajiem paliekot nemainīgiem. Piemēram, dotais modelis prognozē, ka gadījumā, ja uzņēmuma reģistrācijas numurs parādās kādā tā interesentā lapā, tad iespējamība tam, ka kandidātdomēns ir atslēgas domēns, palielinās apmēram 143 reizes.

Kopumā iegūtie koeficienti un iespējamību attiecības diezgan labi atbilst iepriekš aplūkotajiem pazīmju noderīgumiem. Visvairāk informācijas sniedz tieši reģistrācijas numura, WHOIS ierakstu un ielas adreses pazīmes, bet salīdzinoši maz informācijas sniedz pilsētas, domēna sufiksu un pasta indeksa pazīmes. Šo iemeslu dēļ pilsētas pazīme netika iekļauta beigu modelī. Ievietojot iegūtos koeficientus loģistiskās regresijas formulā, iegūst izteiksmi, ar kuru var aprēķināt varbūtību tam, ka kandidātdomēns ir arī atslēgas domēns (6.1, 6.2).

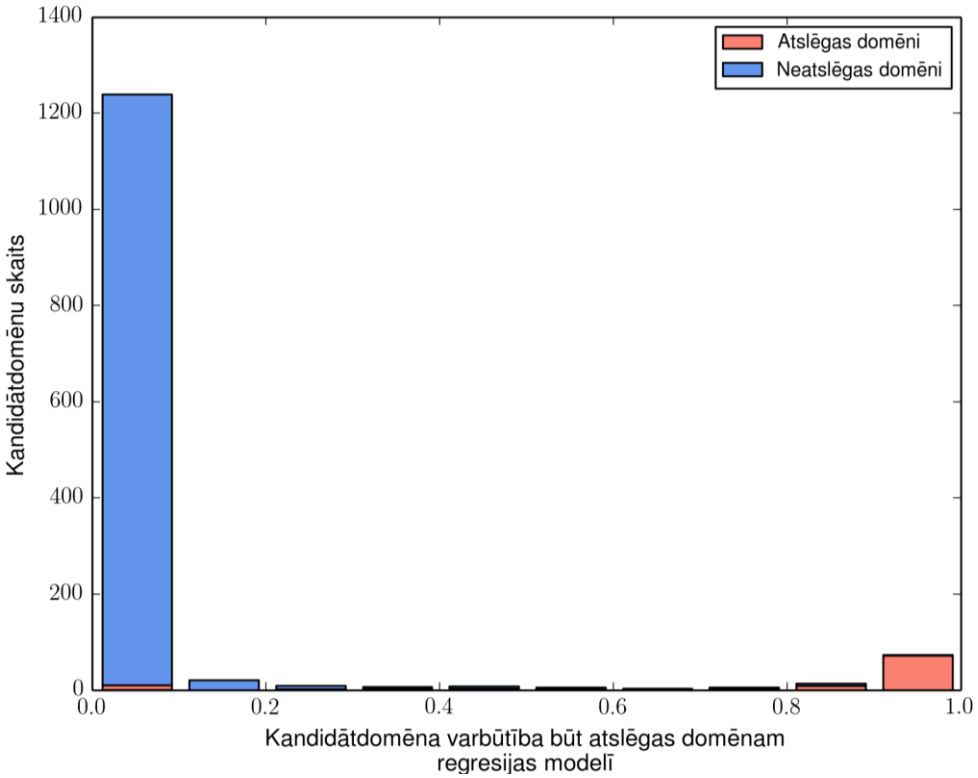
$$K \approx -5.87 + 3.36 * DOM + 2.88 * TITLE + 1.04 * SUFLV - 0.31 * SUFCOM + \\ + 2.15 * NOS + 4.96 * NUM + 0.44 * IND + 2.72 * IEL + 3.81 * WHO \quad \textbf{(6.1)}$$

$$P(atslēgas\ domēns) = \frac{1}{1 + e^{-K}} \quad \textbf{(6.2)}$$

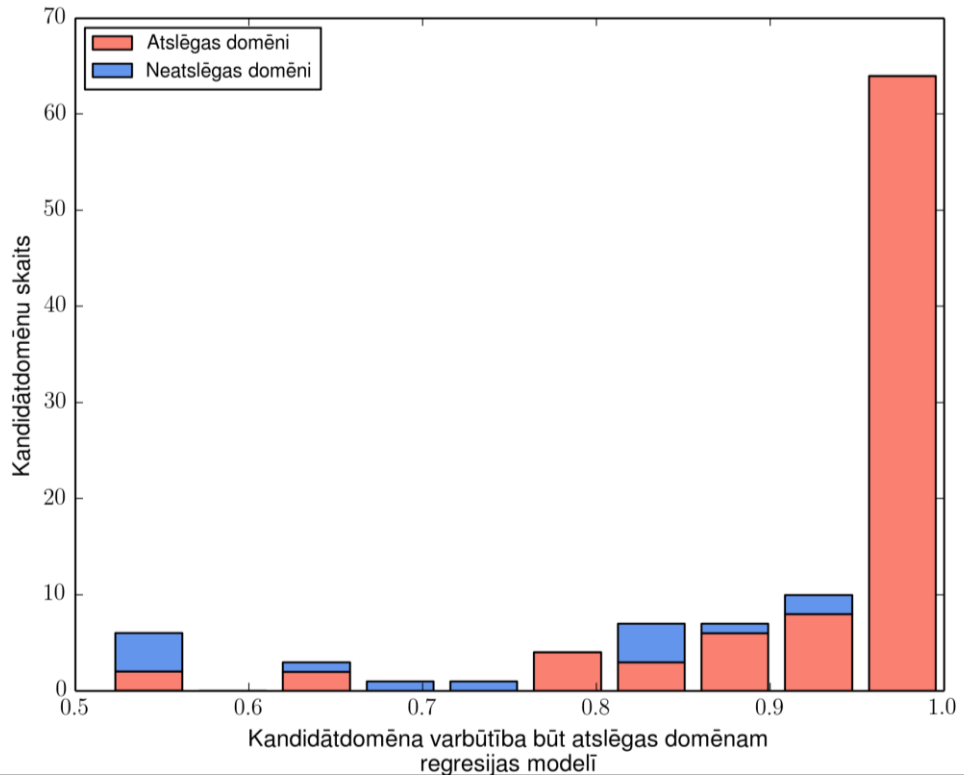
Piemēram, ja kandidātdomēns ir ar *.lv* sufiksu, tā nosaukuma līdzības pakāpe ar uzņēmuma nosaukumu ir 0.87, interesantā tā lapā ir atrodams uzņēmuma reģistrācijas numurs un pārējās pazīmes ir vienādas ar nulli, tad izveidotajā loģistiskās regresijas modelī varbūtība tam, ka kandidātdomēns ir atslēgas domēns, ir apmēram 88.6%. (6.3).

$$P(atslēgas\ domēns) \approx \frac{1}{1 + e^{-(3.36*0.87+0.04*1+4.96*1-5.87)}} = 88.6\% \quad \textbf{(6.3)}$$

Jāņem vērā, ka iegūtās varbūtības ir tikai modeļa sniegti novērtējumi, bet, izvēloties varbūtību slieksni, ir iespējams kontrolēt konkrētajā lietojumā nepieciešamo precizitāti. Korpusa kandidātdomēniem izveidotā modeļa sniegto varbūtību sadalījums atslēgas un neatslēgas domēniem ir redzams 6.3. attēlā. Sīkāks izveidotā modeļa sniegto varbūtību sadalījums korpusa kandidātdomēniem, kad noteiktā varbūtība būt atslēgas domēnam ir lielāka par 50%, ir redzams 6.4. attēlā.



6.3. attēls. Loģistiskās regresijas modeļa paredzēto varbūtību sadalījums visiem korpusa kandidātdomēniem



6.4. attēls. Loģistiskās regresijas modeļa paredzēto varbūtību sadalījums visiem korpusa kandidātdomēniem, kam novērtētā varbūtība ir lielāka par 50%

No attēlos redzamās informācijas var secināt, ka loģistiskās regresijas modeļa sniegtās varbūtības diezgan labi to patiesajām vērtībām.

6.3. Citas klasifikācijas metodes

Tika apskatīti citi sarežģītāki klasifikācijas algoritmi, kas potenciāli varētu gūt labākus rezultātus, ņemot vērā samērā lielās saistības starp kandidātdomēnu pazīmēm. Izvēlētajiem algoritmiem līdzīgi kā loģistiskajai regresijai tika veikti daži to pamata pielāgojumi.

Lai algoritmus varētu labāk salīdzināt savā starpā, katram no tiem šķērsvalidācija tika atkārtota 100 reizes un tika salīdzināts vidējais iegūtais F1-mērs un tā standartnovirze (skat. 6.6. tabula).

6.6. tabula

Klasifikācijas algoritmu salīdzinājums

Klasifikācijas algoritms	**F1-mēra vidējā vērtība**	**F1-mēra standartnovirze**
Loģistiskā regresija	86.4%	0.013
Atbalsta vektoru mašīna	86.7%	0.014
Vairākslāņu perceptrons	83.4%	0.037

Kopumā atbalsta vektoru mašīnas izmantošana sniedza nelielus uzlabojumus, salīdzinot ar loģistisko regresiju, kas gan varētu būt izskaidrojums ar novēroto rezultātu variāciju. Vairākslāņu perceptrons [Zuters] sniedza būtiski zemākus rezultātus un tā rezultātu variācija bija ievērojami augstāka, ņemot vērā iegūto F1-mēru standartnovirzi. Jāpiebilst, ka vairākslāņu perceptrons sniedza nedaudz labākus rezultātus pie lielām pārklājuma vērtībām, kas varētu būt noderīgi atsevišķos lietojumos.

Par pamatu kandidātdomēnu klasifikācijai tika nolemts izmantot izveidoto loģistiskās regresijas modeli, jo tas ir salīdzinoši vienkāršs, ātri aprēķināms un sniedz pietiekami augstus rezultātu, salīdzinot ar citiem algoritmiem.

6.4. Izvēlētā modeļa testēšana

Lai vēlreiz pārbaudītu izveidotā loģistiskās regresijas modeļa darbību, tika veikta tā pārbaude ar 30% korpusa paraugu, kas netika izmantoti šķērsvalidācijas posmā (skat. 6.7. tabula).

6.7. tabula

Šķērsvalidācijas galvenie rezultāti atslēgas un neatslēgas domēnu klasēm

Rezultātu klase	**Precizitāte**	**Pārklājums**	**F1-mērs**	**Paraugu skaits**
Neatslēgas domēnu klase	97.9%	99.5%	98.7%	384
Atslēgas domēnu klase	**92.6%**	**75.8%**	**83.3%**	**33**
Kopā	97.5%	97.6%	97.5%	417

Jāņem vērā, ka testa datu apjoms ir salīdzinoši neliels un, neveicot šķērsvalidāciju, iegūtie rezultāti var variēt lielākā intervālā. Galvenais šī posma uzdevums ir novērtēt, vai šķērsvalidācijas posmā netika pieļautas būtiskas kļūdas. Kopumā iegūtie rezultāti krasi neatšķiras no šķērsvalidācijas posmā iegūtajiem rezultātiem. Sīkāka izveidotā risinājuma pārbaude tiek veikta nākamajā nodaļā, kur risinājuma rezultāti tiek salīdzināti ar citiem informācijas avotiem.

7. LOKĀLAS KANDIDĀTDOMĒNU ATLASES EKSPERIMENTI

Iepriekšējās nodaļās aprakstītais risinājums sniedz salīdzinoši labus rezultātus, bet tā kandidātdomēnu atlases posmā tika izmantoti centralizēti meklētājdzinēji, kuru izmantošana lielos apjomos var būt ierobežota. Ņemot to vērā, šajā nodaļā tiek sniegts ieskats alternatīvā risinājumā, kas varētu tikt izmantots kandidātdomēnu atlasei. Alternatīvā risinājuma rezultāti gan pagaidām nav iekļauti rezultātu analīzes nodaļā, jo tā darbība vēl tiek mērogota.

7.1. Domēna vārdu repozitorija uzturēšana

Kandidātdomēnu atlasē ar meklētājdzinējiem 4. nodaļā tika aprakstīta ārēju meklētājdzinēju izmantošana, kas sasniedza apmēram 80% kandidātdomēnu atlases pārklājumu.

Latvijas uzņēmumu gadījumā var ievērot vairākas būtiskas iezīmes:

- lielākā daļa Latvijas uzņēmumu atslēgas domēnu pieder .lv augstākā līmeņa domēnam (apmēram 85% pēc atslēgas domēnu sufiksu analīzes 5. nodaļā un citos informācijas avotos norādītās atslēgas domēnu informācijas);
- .lv augstākā līmeņa domēnam piederošo domēnu kopējais skaits ir salīdzinoši neliels (pēc NIC statistikas[20] kopējais reģistrēto domēna vārdu skaits ir apmēram 120000, no kuriem liela daļa varētu nebūt aktīvi izmantoti).

No augstāk minētā var secināt, ka, veicot kandidātdomēnu atlasi tikai no .lv sufiksa domēniem, labākajā gadījumā būtu iespējams sasniegt apmēram 85% kandidātdomēnu atlases pārklājumu, kas ir salīdzinoši augsts rādītājs.

Lai veiktu kandidātdomēnu atlasi bez meklētājdzinēju palīdzības, ir nepieciešams uzturēt pēc iespējas aktuālu sarakstu ar Latvijas augstākā līmeņa privātajiem domēniem. Liela daļa vispārīgo domēna vārdu (gTDL) reģistru centralizēti sniedz visu tajos reģistrēto

[20] Pieejams: https://www.nic.lv/lv/statistika.html

domēna vārdu sarakstu, piemēram, ar CZDS (Centralized Zone Data Service) sistēmas[21] palīdzību (pēc reģistrācijas); bet visu Latvijas domēna vārdu saraksts publiski nav pieejams lejupielādei, kas ir raksturīgi konkrētu valstu reģistriem, kas atbild par valstīm specifisko domēna vārdu (ccTDL) reģistrāciju. Praksē valstīm specifisko aktīvo domēna vārdu sarakstu ir iespējams uzturēt, veicot tīmekļa apstaigāšanu vai papildinot to informāciju no citiem tīmekļa pakalpojumu sniedzējiem (parasti par simbolisku atlīdzību).

Jāatzīmē, ka gadījumā, ja domēna vārdu saraksts tiek uzturēts manuāli, ir svarīgi sekot līdzi izgūto datu apjomiem un izmantot efektīvas tīmekļa apstaigāšanas metodes, lai pārmērīgi nenoslogotu resursdatorus.

7.2. Kandidātdomēnu atlase

Ja ir pieejams aktuāls domēna vārdu saraksts, ir nepieciešams regulāri atjaunot informāciju, kas tiek izmantota to klasifikācijai. Atkal ir svarīgi ievērot efektīvas tīmekļa apstaigāšanas metodes, jo šajā gadījumā atjaunojamo datu apjoms ir salīdzinoši lielāks (praktiski gan tikai pāris reizes lielāks) nekā veicot kandidātdomēnu atlasi ar meklētājdzinējiem. [Garcia-Molina, Cho, 2003] apraksta metodes izgūstamo datu apjomu samazināšanai ar efektīvu indeksa atjaunošanas algoritmu palīdzību, uzturot patstāvīgu tīmekļa indeksu, kas var nozīmīgi palīdzēt konkrētajā gadījumā, jo daudziem nelieliem uzņēmumiem to tīmekļa vietņu saturs tiek atjaunots salīdzinoši reti.

Ja ir pieejama atjaunota domēnu informācija, to var izmantot kandidātdomēnu atlasei. Piemēram, no visu domēnu repozitorija konkrētam uzņēmumam var atlasīt kandidātdomēnus, kuru saturs vislabāk atbilst uzņēmuma nosaukumam vai satur tā reģistrācijas numuru.

[21] Pieejams: https://czds.icann.org

7.3. Lokālas kandidātdomēnu atlases rezultāti

Minētais risinājums tika izmēģināts korpusa uzņēmumiem, veicot kandidātdomēnu atlasi no vairākiem tūkstošiem domēnu. Būtiskākā atšķirība bija lielākais atlasīto kandidātdomēnu skaits katram uzņēmumam. Tomēr, atlasot pat vairākus tūkstošus kandidātdomēnu, to klasifikācijas precizitāte būtiski nemainījās un tika sasniegti negaidīti pozitīvi rezultāti.

Risinājuma efektīvai darbībai:

- ir nepieciešams uzturēt aktuālu informāciju par salīdzinoši lielu skaitu domēnu;
- ir nepieciešams izstrādāt efektīvu domēnu informācijas indeksēšanu, lai varētu veikt efektīvu kandidātdomēnu atlasi no lielākā skaita domēnu;
- kandidātdomēnu klasifikācija katram uzņēmumam ir jāveic ar lielāku kandidātdomēnu skaitu.

Kopumā visas minētās problēmas ir risināmas un pirmie eksperimenti deva pozitīvus rezultātus. Vienīgais risinājuma trūkums ir ierobežotās iespējas atrast atslēgas domēnus, kas nepieder .lv sufiksam. Aprakstītā alternatīvā kandidātdomēnu atlases risinājuma izmantošana piedāvāto risinājumu padarītu praktiski neatkarīgu no citiem centralizētiem informācijas avotiem.

8. REZULTĀTI

Nodaļā tiek apkopti ar risinājumu sasniegtie rezultāti, tie tiek salīdzināti ar zelta standartu un citiem informācijas avotiem, tiek aprakstīts risinājuma praktiskais pielietojums un informācijas atjaunošanas iespējas.

8.1. Piedāvātā risinājuma rezultātu salīdzinājums ar zelta standartu

Izveidotā risinājuma sniegtos rezultātus attiecībā pret zelta standartu ietekmē gan kandidātdomēnu atlases posma, gan kandidātdomēnu klasifikācijas posma kvalitātes rādītāji. Iepriekšējās nodaļās tika aprakstītas vairākas metrikas:

- kandidātdomēnu atlases pārklājums $R_{CD\_Selection}$;
- kandidātdomēnu klasifikācijas pārklājums $R_{CD\_Classification}$;
- kandidātdomēnu klasifikācijas precizitāte $P_{CD\_Classification}$.

Izmantojot šīs metrikas, ir iespējams novērtēt arī kopējā risinājuma kvalitāti attiecībā pret zelta standartu, kam tiek izmantotas divas galvenās metrikas:

- atslēgas domēnu meklēšanas pārklājums;
- atslēgas domēnu meklēšanas precizitāte.

Atslēgas domēnu meklēšanas pārklājums *(R, Recall)* ir attiecība starp ar risinājumu atrasto atslēgas domēnu skaitu un kopējo atslēgas domēnu skaitu zelta standartā. Atslēgas domēnu meklēšanas pārklājums raksturo varbūtību tam, atslēgas domēns ar risinājumu tiks atrasts.

Atslēgas domēnu meklēšanas precizitāte *(P, Precision)* ir attiecība starp ar risinājumu pareizi atrasto atslēgas domēnu skaitu un kopējo ar risinājumu atrasto domēnu skaitu. Atslēgas domēnu meklēšanas precizitāte raksturo varbūtību tam, ka ar risinājumu atrasts domēns patiešām ir atslēgas domēns.

Aprakstītās kopējā risinājuma novērtēšanas metrikas ir iespējams prognozēt, izmantojot iepriekšējās nodaļās iegūtos kandidātdomēnu atlases un klasifikācijas rādītājus (7.1, 7.2).

$$R = R_{CD\_Selection} * R_{CD\_Classificataion} \quad \textbf{(7.1)}$$

$$P = P_{CD\_Classificataion} \quad \textbf{(7.2)}$$

Veicot kandidātdomēnu atlasi ar meklētājdzinējiem, kandidātdomēnu atlases pārklājums sasniedza apmēram 83.2%. Un pie 85% varbūtības sliekšņa loģistiskās regresijas modelim kandidātdomēnu klasifikācija sasniedza 82.4% pārklājumu ar 91.7% precizitāti šķērsvalidācijas laikā.

Ievietojot vērtības formulās (7.1, 7.2), iegūst risinājuma kvalitātes rādītājus pie 85% varbūtības sliekšņa (7.3, 7.4).

$$R = 0.832 * 0.824 \approx 68.6\% \quad \textbf{(7.3)}$$

$$P = 91.7\% \quad \textbf{(7.4)}$$

Līdzīgi kopējo risinājuma atslēgas domēnu pārklājumu un precizitāti var prognozēt arī pie citām varbūtības sliekšņa vērtībām.

8.2. Piedāvātā risinājuma salīdzinājums ar citiem informācijas avotiem

Piedāvātais tīmekļa vietņu meklēšanas risinājums, kas ietver gan kandidātdomēnu atlases, gan to klasifikācijas posmus, tika salīdzināts ar citiem informācijas avotiem, izmantojot uzņēmumu testa izlasi, kurā tika iekļauti pēc gadījuma principa izvēlēti citi 1000 uzņēmumi no Latvijas aktīvo SIA kopas (skat. 7.1. tabula).

7.1. tabula

Piedāvātā risinājuma rezultāti uzņēmumu testa izlasei

Testa izlases uzņēmumu skaits	1000
Domēnu atbilstības slieksnis (minimālā loģistiskās regresijas modeļa varbūtība domēna atrašanai)	85%
Ar risinājumu atrasto domēnu skaits	85
Ar risinājumu atrasto atslēgas domēnu skaits	77
Ar risinājumu atrasto domēnu īpatsvars (ar 95% ticamības intervālu)	8.50 ± 1.73%
Ar risinājumu atrasto domēnu precizitāte	90.58%

Kopumā testa izlasē pie 85% varbūtības sliekšņa ar risinājumu tika atrasti 85 domēni, no kuriem vairāk nekā 90% bija atslēgas domēni. Salīdzinot ar korpusa uzņēmumiem, kam pie 85% varbūtības sliekšņa tika atrasti 97 domēni, no kuriem vairāk kā 91% bija atslēgas domēni, iegūtie rezultāti ir līdzīgi. No tā var secināt, ka galvenie kvalitātes rādītāju novērtējumi ir tuvi to patiesajām vērtībām.

Ar piedāvāto risinājumu atrastās tīmekļa vietnes testa izlasei tika salīdzinātas ar citos informācijas avotos norādītajām tīmekļa vietnēm (skat. 7.2. tabula).

7.2. tabula

Piedāvātā risinājuma atrasto tīmekļa vietņu salīdzinājums ar citiem avotiem

Avots	**Sakrītošas**	**Atšķirīgas**	**Norādītas tikai risinājumā**	**Norādītas tikai avotā**
zo.lv	26	0	59	43
1188.lv	6	0	79	5
zl.lv	51	3	31	63
1182.lv	33	2	50	51
Kopā	116	5	219	162

No iegūtajiem rezultātiem var secināt, ka tikai 5 gadījumos piedāvātajā risinājumā un kādā citā avotā uzņēmumiem tika atrastas pretrunīgas tīmekļa vietnes. Un lielākajai daļai pretrunīgo gadījumu piedāvātajā risinājumā atrastās tīmekļa vietnes domēns labāk atbilda atslēgas domēna definīcijai.

Līdzīgi kā starp citiem informācijas avotiem, ar piedāvāto risinājumu tika atrastas salīdzinoši daudzas tīmekļa vietnes, kas nebija norādītas citos informācijas avotos. Kopumā 33.0% no tīmekļa vietnēm, kas tika atrastas ar piedāvāto risinājumu, nebija norādītas nevienā citā informācijas avotā. Tas nozīmē, ka piedāvāto risinājumu varētu izmantot, lai būtiski papildinātu citus informācijas avotus.

Ar piedāvāto risinājumu atrasto tīmekļa vietņu skaits bija mazāks tikai par portālā *zl.lv* norādīto tīmekļa vietņu skaitu. Precizitātes novērtējumi ir subjektīvāki, jo tie ir atkarīgi no izvēlētās atslēgas domēna definīcijas, saskaņā ar kuru piedāvātais risinājums sasniedza ievērojami augstāku precizitāti nekā portāli *zo.lv*, *zl.lv* un *1182.lv*. Tomēr šajos portālos

lielākie atšķirību cēloņi bija norādītajām tīmekļa vietnēm nestrādājoši DNS serveri, gramatiskas kļūdas domēnu pierakstā un domēnu pārvirzīšanas, kas ir salīdzinoši viegli novēršamas tehniskas problēmas.

Neraugoties uz iepriekš minēto, atšķirību skaits starp ar piedāvāto risinājumu atrastajām un pārējos avotos norādītajām tīmekļa vietnēm ir salīdzinoši mazāks par atšķirību skaitu starp pārējiem avotiem, kas tika apskatīts 3. nodaļā; un tas ir no atslēgas domēna definīcijas neatkarīgs mērs, kas netieši norāda uz augstāku precizitāti piedāvātajā risinājumā.

8.3. Sistēmas veiktspēja un mērogošana

Pēc risinājuma izstrādes sistēma tika mērogota, lai ar to būtu iespējams veikt atslēgas domēnu meklēšanu arī pārējiem Latvijas uzņēmumiem. Mērogošanas posmā tīmekļa vietņu meklēšana ar līdzīgiem rezultātiem tika veikta apmēram 20000 uzņēmumu un tuvākajā laikā to ir paredzēts veikt arī pārējiem Latvijas aktīvo SIA kopas uzņēmumiem.

Mērogošanas posmā vairāk uzmanības un laika prasa tieši informācijas izguves kontrole, jo ir svarīgi sekot līdzi izgūto datu apjomam un pārāk nenoslogot atsevišķus resursdatorus, bet nepieciešamie skaitļošanas resursi ir salīdzinoši nelieli. Visi skaitļošanas uzdevumi tika veikti ar vienu lokālu datoru un vienu Amazon Elastic Compute Cloud (EC2) t2.medium instanci, kuras tehniskie parametri ir atrodami EC2 tīmekļa vietnē[22].

Pēc sākotnējās atslēgas domēnu meklēšanas būtu iespējams izstrādāt procesus atslēgas domēnu informācijas regulārai atjaunošanai. Lai izdarītu spriedumus par datu atjaunošanas biežumu un principiem ir sīkāk jānovērtē dažādu resursu informācijas mainīgums, kuru pagaidām ir grūti precīzi paredzēt.

[22] Pieejams: https://aws.amazon.com/ec2

8.4. Risinājuma pielietojums citu valstu uzņēmumiem

Autora izstrādāts līdzīgs risinājums tika praktiski izmantots automatizētai Vācijas uzņēmumu tīmekļa vietņu meklēšanai lielā apjomā. Kopumā tīmekļa vietnes tika meklētas vairākiem simtiem tūkstošu līdz pāris miljoniem Vācijas uzņēmumu.

Jāatzīmē, ka lielākā risinājuma daļa nav atkarīga no izmantotās valodas un ir salīdzinoši viegli pielāgojama darbam ar dažādu valstu uzņēmumiem.

NOBEIGUMS

Darba ietvaros tika izstrādāta sistēma Latvijas uzņēmumu tīmekļa vietņu automatizētai meklēšanai. Tās sniegto rezultātu precizitāte un pārklājums ir salīdzināms ar autora identificēto kvalitatīvāko informācijas avotu rezultātiem, un tā varētu tikt izmantota arī to būtiskai papildināšanai.

Salīdzinot ar citiem informācijas avotiem, piedāvātais risinājums:

- ļauj veikt praktiski visus tīmekļa vietņu meklēšanas procesus automātiski;
- ļauj norādīt konkrētam lietojumam nepieciešamo rezultātu precizitāti (mainot ar loģistiskās regresijas modeli iegūtās varbūtības slieksni);
- paredz iespēju vienam uzņēmumam meklēt vairākus atbilstošus atslēgas domēnus.

Risinājumā izmantotās metodes ļauj to salīdzinoši viegli pielāgot arī citu valstu uzņēmumu informācijas izguvei.

Darba ietvaros tika veikti eksperimenti arī ar lokālu kandidātdomēnu atlasi, kas deva pozitīvus sākotnējos rezultātus un varētu tikt izmantota, lai novērstu praktiski visas risinājuma atkarības no citiem centralizētiem informācijas avotiem.

IZMANTOTĀ LITERATŪRA UN AVOTI

M.L. Berenson, D.M. Levine, T.C. Krehbiel (2011). *Basic business statistics: Concepts and applications*. Pearson Higher Education AU, 12th ed., 2011, pp. 266-269.

C. Castillo (2005). *Effective web crawling.* ACM SIGIR Forum, vol. 39, no. 1, 2005, pp. 55-56.

S. Chakrabarti, M. Berg, B. Dom (1999). *Focused crawling: a new approach to topic-specific Web resource discovery.* Proc. of the eighth international conference on World Wide Web, vol. 31, no. 11-16, 1999, pp. 1623-1640.

W.W. Cohen, P. Ravikumar, S.E. Fienberg (2003). *A comparison of string metrics for matching names and records.* Kdd workshop on data cleaning and object consolidation, vol. 3, 2003, pp. 73-78.

J. Davis, M. Goadrich (2006). *The relationship between Precision-Recall and ROC curves.* Proc. of the 23rd international conference on Machine learning, 2006, pp. 233-240.

L. Daigle, VeriSign, Inc. (2004). *WHOIS Protocol Specification* IETF RFC 3912. IEEE, September, 2004, www.ietf.org/rfc/rfc3912.txt

H. Garcia-Molina, J. Cho (2003). *Effective page refresh policies for Web crawlers.* ACM Transactions on Database Systems, vol. 28, no. 4, 2003, pp. 390-426.

E.J. Glove, K. Tsioutsiouliklis, S. Lawrence, D.M. Pennock, G.W. Flake (2002). *Using Web Structure for Classifying and Describing Web Pages.* Proc. of the 11th international conference on World Wide Web, 2002, pp. 562-569.

T. Hastie, R. Tibshirani, J. Friedman (2009). *The Elements of Statistical Learning.* Springer, 2009.

M.Y. Kan, H.O.N. Thi (2005). *Fast webpage classification using URL features.* Proc. of the 14th ACM international conference on Information and knowledge management, 2005, pp. 325-326.

J.Y. Kim, J.S. Taylor (1994). *Fast string matching using an n-gram algorithm.* Software – Practice & Experience, vol. 24, no. 1, 2009, pp. 79-88.

P. Mockapetris (1987). *Domain Names – Implementation and and Specification.* IETF RFC 1035. IEEE, November, 1987, www.rfc-editor.org/rfc/rfc1035.txt

G. Navarro (2001). *A guided tour to approximate string matching.* ACM Computing Surveys, vol. 33, no. 1, 2001, pp. 31-88.

X. Qi, B.D. Davison (2009). *Web page classification: Features and algorithms.* ACM Computing Surveys, vol. 6, no. 1, 2009.

L. Wasserman (2013). *All of statistics: a concise course in statistical inference.* Springer Science & Business Media, 2013.

Demetrios Zeinalipour-Yazti, M.D. Dikaiakos (2001). *High-Performance Crawling and Filtering in Java.*

K. Znotiņš (2015). *Uzņēmumu tīmekļa vietņu meklēšana un monitorings.* Kursa darbs. Latvijas Universitāte.

H. Zou, T. Hastie (2005). *Regularization and variable selection via the elastic net.* Journal of the Royal Statistical Society: Series B, vol. 67, no. 2, 2005, pp. 301-320.

J. Zuters. *Neironu tīkli*. Available: http://home.lu.lv/~janiszu/courses/eanns/eanns.pdf

Bakalaura darbs „Automatizēta uzņēmumu tīmekļa vietņu meklēšana” izstrādāts LU Datorikas fakultātē.

Ar savu parakstu apliecinu, ka pētījums veikts patstāvīgi, izmantoti tikai tajā norādītie informācijas avoti un iesniegtā darba elektroniskā kopija atbilst izdrukai.

Autors: _____ Kristaps Znotiņš

Rekomendēju/nerekomendēju darbu aizstāvēšanai

Vadītājs: prof., Dr. dat. Guntis Arnicāns _____ 30.05.2016.

Recenzents: Dr.dat. Uldis Bojārs

Darbs iesniegts Datorikas fakultātē 30.05.2016.
Dekāna pilnvarotā persona: vecākā metodiķe Ārija Sproģe _____

Darbs aizstāvēts bakalaura gala pārbaudījuma komisijas sēdē __.06.2016.

_____. prot. Nr. _____.

Komisijas sekretārs: _____ _____