Eesti Advokatuur
Tartu Ülikool

Eesti ja EL õigusallikate tekstikorpuse rakendamine viidete otsinguks

Õigusrobootika konkursi raames loodud programmikoodi kirjeldus

Programmikoodi loojad: Mario Sepp, Roland Pihlakas

Visiooni ja kirjelduse loojad: Mario Sepp, Margot Maksing,
Mirell Krain, Roland Pihlakas

Tallinn 2017

Probleemi kirjeldus

Õigusküsimuste lahendamine on ajakulukas protsess, kuna hõlmab erinevatest veebikeskkondadest ja erineval kujul, nii elektroonsel kui ka paberkandjal, oleva õigusalase teabe otsimist. Sealjuures pole õigusallikatest otsimine piisavalt paindlik. Advokaatidel, prokuröridel, kohtunikel ja juristidel tekib töös sageli vajadus otsida mõne seaduse sätte tõlgendusi kohtupraktikast. Samuti on oluline seaduste omavaheliste viidete ja seoste leidmine.

Eesmärgid

Õigusalase teabe otsimise lihtsustamiseks on esmane eesmärk luua ühine platvorm nii Eesti kui ka Euroopa Liidu õigusallikate kokku koondamiseks. Õigusallikate integreerimine ühte keskkonda aitab kaasa nende efektiivsemale haldamisele ning hõlbustab kiiremat õigusküsimuse lahendamist, kuna hoiab kokku aega, mis kulub erinevatest keskkondadest õigusalase teabe otsimisele.

Teoreetilised lahendused

Eestis lahendavad kohtud vaidlusi kehtivate seaduste alusel. See tähendab, et esmatähtis on õigete õigusnormide äratundmine. Järgmise sammuna on oluline seaduse teksti mõttest aru saamine ehk õiguse tõlgendamine. Õigust aitab tõlgendada varasem kohtupraktika. Lisaks on argumenteerimist abistavateks materjalideks seaduste kommentaarid ja ka Juridica artiklid.

Programmi kõige olulisem funktsionaalsus peaks olema nö kaasuse lahendamine. Programm võiks esitada kasutajale küsimusi, millele vastates jõuab kasutaja lõpplahenduseni. Tsiviilõiguses on tavapäraselt igal nõudel seaduses konkreetne nõude alus ehk üks kindel paragrahv, mille järgi kõiki sarnaseid õigusvaidlusi lahendatakse (või vähemalt nende lahendamist alustatakse). Nõude alus on omakorda seotud teiste paragrahvidega, mille abil kontrollitakse ühe või teise eelduse täidetust.

Esimesena peaks programm töötama viisil, et kaasuseid on võimalik läbi lahendada vastavalt programmi poolt esitatud küsimustele, mis põhinevad nõude alustel ja eeldustel. Küsimused peaksid olema võimalikult lühikesed, kuid samas üheselt mõistetavad. Võimalusel tuleks eelistada märksõnade kasutamist lausetele, sest see on kasutajale kiirem ja mugavam.

Järgmisena tuleb lisada võimalus, et iga küsimuse ja vastusevariandi juures on viited kohasele kohtupraktikale ja õiguskirjandusele. Kasutajal peab olema võimalik tutvuda erinevate argumentide ja seisukohtadega enne, kui ta programmi esitatud küsimusele vastab. Siinkohal on ilmselt oluline, et arvuti suudaks ise kohtupraktikat nii palju mõista ja süstematiseerida, et ta pakuks kasutajale 10-20 kõige olulisemat lahendit (või isegi vähem). Ideaalis võiks leitud lahendis olla võimalik märgistada vajalik tekstilõik, et sellele ühe nupuvajutusega viidata, näiteks märgistatud lõik kopeerida valmivasse dokumenti (NB! koos korrektse viitega). Programm peaks "õppima" kasutaja tehtud valikutest, st pakkuma esimesena neid lahendeid, mida kasutajad ise sagedamini viitavad (see õpivõime ei peaks aga kasutaja jaoks ilmselge olema).

Kui programm sellisel viisi töötab, tuleb järgmisena leida võimalus lahendusskeemi lühendamiseks. Lihtsamate kaasuste puhul ei soovi kasutaja kindlasti alati kogu skeemi läbi lahendada. Seega peaks kasutajal olema võimalik alustada "skeemi keskelt". Üks võimalus on pakkuda kasutajale võimalust sisestada märksõna, millest ta soovib alustada, tekstina. Programm ise pakub kasutaja teksti põhjal kõige sarnasemad etapid pikast kaasuse lahenduse skeemist, kasutaja saab valida talle sobiva ja jätkata sealt skeemi lahendamist. Sel juhul peaks kasutaja kindlasti mõned korrad eelnevalt pikka skeemi lahendama, et ta teaks märksõnu, mida lühendatud skeemi jaoks sisestada.

Lisaks peaks programm töötama ka otsingumootorina. See funktsionaalsus peaks olema eraldi valik võrreldes kaasuste lahendamisega. Kasutajal on võimalik sisestada üks või mitu märksõna ja valida, millistest allikatest ta nende märksõnade kohta teavet soovib (seadus, esimese/teise astme kohtupraktika, Riigikohtu praktika, õiguskirjandus). Esitatud märksõnade järgi kuvab programm 10-20 vastet, mis võiks taas olla viidatavad (+programm salvestab kasutajate valikuid).

Märksõnade otsing töötaks temaatilise kattumise põhimõttel, ehk et dokumentides leiduv sõnavara ei pea täpselt kattuma otsingus kasutatud sõnadega, piisab teema ühtimisest. Samuti on võimalik esitada "negatiivseid" märksõnu teatud teemade välistamiseks tulemustest. Viimaks on võimalik ka märkida, millised varasemad otsingutulemused ei paistnud käesoleval hetkel kohased, et otsingumootor teaks järgmisena pakkuda võimalikult erinevat sorti tulemusi.

Eespool kirjeldatud programm aitaks õiguse valdkonnas tegutsevatel inimestel paremini orienteeruda suures hulgas õigusteabes ning tagaks õiguse ühetaolise kohaldamise rohkematel sarnastel juhtudel.

Teostatud osa

Eelmises punktis kirjeldatud idee teostamiseks on vaja aga oluliselt rohkem aega ja suuremat meeskonda, kui Õigusrobootika konkurss seda võimaldas. Konkurssi käigus on valminud väiksema mahuga prototüüp, mis lahendab eespool kirjeldatud otsingumootori probleemi. Seda on võimalik täiendada ning edasi arendada.

Demo

Otsingud toimuvad tekstide temaatilise sarnasuse järgi märksõnadele. Leitavad tekstid ei pea otsingus olevaid märksõnu tingimata sisaldama, kui leidub sünonüüme või muud moodi temaatilist kattumist.
Otsida saab nii eesti kui inglise keelseid tekste.
Praegune demo sisaldab nii Eesti kui Euroopa Liidu seadusi, seisuga oktoober 2004. Kasutame hetkel vanu seadusi, kuna nende maht on väiksem, tänu millele demo katsetajal on vähem ootamist programmi esmase käivituse järel ning teiseks oli nende vanade korpuste formaat demo projekti raames lihtsamini kasutatav.
Lisaks on programmi sisse ehitatud võimalus otsida isiklikke tekstiarhiive, kui päringus vastav kaust ära määrata. Sellega on kasutajal võimalik lisada isiklikke andmekorpuseid.

Kasutusjuhend

Windowsi all kasutamiseks sobiv kompileeritud demo on allalaetav siit:
www.simplify.ee/legalsearch/LegalSearchText_31.01.2017.zip
Hiljem uuendatud versioonid on saadaval siin:
www.simplify.ee/legalsearch/LegalSearchText.zip

Arhiiv tuleks lahti pakkida ning tekkinud kaust kopeerida “Allalaadimiste” kaustast “Dokumentide” alla, vastasel korral ei ole võimalik programmi käivitada.

Tarkvaral on kaks versiooni, üks salvestab kõvakettale tekstifaili logi kõigist otsingupäringutest ning tulemustest. Teine versioon ei salvesta.
Vastavalt tuleks käivitada (saab käivitada topeltklikiga, ei pea konsooli avama):
`_LegalSearchText_with_logging.bat`
või
`_LegalSearchText.bat`

Peale käivitamist esimese päringu järel koostab tarkvara otsinguks vajaliku andmebaasi. See võtab mõned minutid aega. Hilisemate otsingute jaoks andmebaasi enam uuesti ei koostata, välja arvatud, kui otsitakse uues keeles (siis toimub samuti andmebaasi koostamine ainult üks kord keele kohta ning eelmise keele andmebaas jääb samuti alles).

Päringute formaat:

- Üldiselt, otsingu tekstis leiduvad sõnad on otsingusõnad. Otsingusõnad on terminid, mille teemadega kattuvaid dokumente üritatakse leida. Otsingusõnade ette võib lisada erilist tähendust omavaid märksõnu, mis on kirjeldatud hiljem.
  Näide:
  Enter search query or command: `insurance`
- Kui otsingusõna ette kirjutada miinusmärk - siis mõjub see sõna negatiivse sõnana ehk et selle sõnaga seonduvaid dokumente või teemasid üritatakse tulemustes vältida. Miinusmärk peaks asuma vahetult sõna ees nõnda, et tühikut vahel ei ole.
  Näide:
  Enter search query or command: `border customs -marine`
- Päringu teksti esimesed märksõnad määravad mõningaid nüansse otsingu käitumises (need märksõnad ei ole kohustuslikud ning nende puudumisel kasutatakse vaikekäitumist, on lubatud esitada ka suvaline alamosa otsingut juhtivatest märksõnadest, kuid lubatud järjekord on fikseeritud). Märksõnad ei ole tõstutundlikud. Kokkuvõtlikult on nende märksõnade formaat järgnev:
  `[level: text|line (text)] [language: eng|est|engest|esteng|none (eng)] [scope: ee|eu|eeeu|custom:folder (ee)] [num_results: (20)]` *`[num_dims: (152|252)]`* `[notlike #1 #3 #8 ...] query words ... -negative -words ...`
- Need märksõnad ning nende kasutamine on inimeste keeles lahti seletatud järgnevalt. Päring koosneb tekstist, mida tõlgendatakse vasakult paremale:
  - `text|line`
    tohib kasutada ühte neist kahest märksõnast. Need määravad, kas otsida terviktekste (`text`) või ridasid neis tekstides (`line`). Vaikimisi otsitakse terviktekste. Ridade tasemel otsing ei toimi veel kuigi hästi, selle põhjused ja edaspidised lahendused on mainitud “edasiarenduste” peatükis.
  - `eng|est|engest|esteng|none`
    määrab otsingusõnade ning tulemuste kuvamise keele. `engest` ning `esteng` tähendavad seda, et otsing toimub ühe keele märksõnade järgi, kuid tulemused kuvatakse otsingutulemustele vastavatest tõlgetest. Kuna korpus on tõlgete paralleelkorpuse kujul, siis on selline võte võimalik. Vaikimisi otsitakse inglisekeelsete märksõnadega ning tulemused kuvatakse samuti ingliskeelsest korpusest.
    `none` variant tähendab, et tekst on tavaline toortekst, mis ei sisalda `<eesti>` ja `<inglise>` tagisid. Seda on vaja kasutada juhul, kui on soov otsida isiklikke tekstiarhiive (vaata järgnevat lõiku).
  - `ee|eu|eeeu|custom:folder`
    määrab, milliseid seadusi otsida. `ee` puhul otsitakse ainult Eesti seadusi, `eu` puhul ainult Euroopa Liidu seadusi, ning `eeeu` puhul otsitakse mõlemast seadusekogust. Vaikimisi otsitakse ainult Eesti seadustest.

`custom:folder` tähendab, et korpusena kasutatakse mõnda isiklikku tekstiarhiivi kaustas, mille nimi tuleb kirjutada sõna "`folder`" asemele ülal olevas päringus.
Näide:
`text none folder:demo_custom_book 5 secrets property`

- Järgnev number määrab, kui mitu otsingutulemust kuvada. Vaikimisi kuvatakse 20 otsingutulemust.
- *Edasijõudnutele: Järgnev number määrab, kui mitu tähenduse dimensiooni tekstidest tuvastatakse ning sõnadele omistatakse. Vaikimisi tuvastatakse dokumendi tasemel otsingu jaoks 152 dimensiooni ning ridade tasemel otsingu jaoks 252 dimensiooni. Tavakasutaja tõenäoliselt ei soovi selle numbri muutmisega katsetada.*
- `notlike ... numbrid`
  märksõna notlike järel olevad numbrid määravad järjenumbrite kaudu, millised otsingutulemused eelmises otsingus ei meeldinud ning mille sarnaseid võiks käesolevas otsingus vältida.
- Näide:
  Enter search query or command: `text engest eeeu 10 notlike 1 3 8 social care`

Arendusplaanid:

- Kaheastmeline otsing rea otsingu jaoks: kõigepealt otsitakse välja temaatiliselt sobivad dokumendid, seejärel neist dokumentidest kõige enam sobivad read. Hetkel toimib rea otsing sedasi, et käsitleb iga rida sõltumatu dokumendina. Praegusel moel on igal reaga seotud liiga vähe infot, et seda adekvaatselt temaatiliselt töödelda. Analoogselt võib teha otsitavaks paragrahve ning peatükke. Kõige keerukam on hästi tööle saada rea tasemel otsingut, suuremate ühikute otsimine on lihtsam.
- Otsingu sisendiks saab määrata ka mõne olemasoleva seaduse või kohtulahendi otsingu enda korpusest ning otsida teisi temaatiliselt seonduvaid seadusi ja lahendeid. See on lihtsalt teostatav.
- Otsinguga seonduvate alternatiivsete märksõnade soovitamine. See on lihtsalt teostatav.
- Riigi teataja arhiivide tugi
- Riigikohtu lahendite arhiivide tugi
- Juridica arhiivide tugi
- Otsingu sisendiks on võimalik anda tekstifaile, millele vastavaid temaatilisi tulemusi soovitakse leida. Ehk et tekstifaili sisu ongi otsingu märksõnadeks. Analoogselt on võimalik anda tekstifaile negatiivsete märksõnade alla. See on lihtsalt teostatav.
- Parem eesti keele tugi kasutades morfoanalüsaatorit sõnatüvede leidmiseks ning ühtestajat mitmetimõistetavate sõnade ning asesõnade tähenduse selgitamiseks. Samuti fraaside tuvastamiseks. Uusim meile teadaolev eesti keele tarkvaralise töötlemise tööriist asub aadressil https://estnltk.github.io Analoogne vanem tööriist leidub aadressil http://www.eki.ee/tarkvara/analyys/
- Parem inglise keele tugi mõne põhjalikuma morfoanalüsaatori kasutamisega, kui praegu koodi sissekirjutatud prototüüp. See võimaldab paremini tuvastada sõnatüvesid ning lisaks eemaldada osa stopp-sõnu (sõnad, mis on liiga sagedased, et tähendust omada). Ka praegune algoritm suudab stopp-sõnu ignoreerida, kuid seda statistilistel meetoditel.
- Erinevad otsingualgoritmid. Hetkel on kasutuses “Doc2vec” algoritm. Alternatiivina on võimalik kasutada veel LSI (Latent semantic analysis), LDA (Latent Dirichlet allocation) ning KCCA (Kernel Canonical Correlation Analysis) statistilisi algoritme. KCCA on huvitav selle poolest, et võimaldab sõnade täiendavat tähendusruumi statistiliselt tuvastada teksti tõlgete ehk paralleelkorpuse abiga. LSI ning LDA on mõlemad saadaval GenSem tarkvarapaketi sees. Mujal leidub ka KCCA implementatsioone ning siinse koodi autor on seda algoritmi ka ise ühe korra varasemalt kasutanud.
- Kui mõni otsingusõna on liiga harvaesinev, et statistiline lähenemine seda leida võiks (statistiline lähenemine ei indekseeri sõnu, mille tähendust tuvastada ei suuda), siis on potentsiaalselt tegu mingi lühendi või märksõna või nimega. Sel juhul aitaks täiendavalt tavateksti otsingu tugi, mis otsib otsingusõna otse algfailidest ning mitte statistilisest mudelist.

Viited

Käivitatav kompileeritud demo:
www.simplify.ee/legalsearch/LegalSearchText_31.01.2017.zip
Hiljem uuendatud versioonid on saadaval siin:
www.simplify.ee/legalsearch/LegalSearchText.zip
Demo arhiiv sisaldab ülal mainitud seadusekorpuseid

Demos kasutatav lähtekood (sisaldab versioonide ajalugu ning ka kõik uuendused tulevad siia):
https://github.com/levitation/LegalSearch
Lähtekood seisuga 31.01.2017 on saadaval ka siin:
www.simplify.ee/legalsearch/LegalSearchSources_31.01.2017.zip

Täiendav koodivaramu:
https://github.com/mariosepp/riigiteataja-converter
Sisaldab Riigi Teataja arhiivide parsimise loogikat.

Kasutatavate tekstikorpuste arhiiv eraldi (sisaldub ka kompileeritud koodi arhiivis ning lähtekoodi arhiivis):
www.simplify.ee/legalsearch/LegalSearchCorpuses.zip
Seal sisalduvad failid pärinevad originaalis siit:
http://www.cl.ut.ee/korpused/paralleel/