Studentu aktivitāte [rezultāts O1-A-4, lv, licence CC BY 4.0, 2018.gada 23.augustā]

# Kā darbojas teksta plaģiātisma atpazīšanas sistēmas

Datums: 2018.gada 23.augustā

Informācija par šī materiāla izmantošanu:



Šis darbs ir licencēts saskaņā ar Creative Commons Attribution 4.0 Starptautisko licenci.

Jūs varat brīvi koplietot, kopēt un izplatīt materiālu jebkurā vidē vai formātā. Jūs varat brīvi pielāgot, pārkārtot, transformēt un balstīties uz materiālu jebkādam nolūkam. Jums ir jādod atbilstoša atsauce uz oriģinālmateriālu, jānorāda saite uz licenci un jānorāda, vai materiālā ir veiktas izmaiņas. Jūs to varat darīt jebkādā saprātīgā veidā, neprasot licences turētājam apstiprināt materiāla izmantošanas atļauju.

 Papildinformācija par CC licencēšanu: <https://creativecommons.org/licenses/?lang=lv>

Atsauce:

[autors] Foltýnek, Tomáš, Mendel University in Brno, Czechia [nosaukums] How Textual Plagiarism Detection Works

[datums] 2018-08-23

[avots] <http://www.academicintegrity.eu/wp/all-materials> [piekļuves datums]

[tulkojums] Tatjana Odineca, Alla Anohina-Naumeca

# Kā darbojas teksta plaģiātisma atpazīšanas sistēmas

Mēs runāsim tikai par tekstu līdzību. Attēlu, skaņu vai videomateriālu līdzība tiek atpazīta ar pilnīgi citām metodēm, un mēs šos jautājumus neapskatīsim.

Labai teksta atpazīšanas programmatūrai (vai plaģiātisma atpazīšanas sistēmai) būtu jātiek galā ne tikai ar burtisko kopēšanu, bet arī ar teksta pārkārtošanu, parafrāzēšanu, apkopošanu vai pat tulkojumu citā valodā. Tā joprojām ir problēma, ko uzskata par sarežģītu, un profesionālo sistēmu izstrādātājiem bieži vien ir jāatrod kompromiss starp ātrumu un precizitāti.

Mums ir dokuments un mēs paredzam, ka tas ir cita(u) dokumenta(u) plaģiātisms. Sauksim to par aizdomīgu dokumentu, kas potenciāli varētu būt plaģiātisms. Mūsu uzdevums ir atrast iespējamos avota dokumentus, ko izmantoja aizdomīgā dokumenta autors.

Visam procesam ir trīs fāzes (Stamatatos et al., 2015):

1. fāze – **kandidātu izgūšana**. Protams, nav jēgas salīdzināt aizdomīgu dokumentu ar visu, kas tika uzrakstīts agrāk. Tādējādi mums ir ļoti svarīgi “sašaurināt” mērķavotus. Tieši tāpēc mēs izvēlamies tikai dokumentus, ko ir vērts pārbaudīt, vai nu no plašas datubāzes saskaņā ar metadatiem, vai ar (tīmekļa) meklētājprogrammu, izmantojot mērķtiecīgi veidotus vaicājumus.

2. fāze – **detalizēta analīze**. No pirmās fāzes mums ir pieejams ierobežots kandidātdokumentu saraksts. No tiem mums ir jāizvēlas tikai tie, kas tiešām var būt plaģiātisma avoti, un jāidentificē atkārtoti izmantoti teksta fragmenti.

3. fāze – **pēcapstrāde**. Otrā fāze var dot pārāk daudz rezultātu, daži no tiem pārklājas, un citi ir maznozīmīgi. Tas ir iemesls, kāpēc ir vajadzīga trešā fāze. Tās mērķis ir filtrēt un vizualizēt rezultātus. Rezultāts ir aizdomīgs dokuments, kurā atkārtoti izmantoti teksta fragmenti tiek atspoguļoti ar krāsu un ir dotas saites uz avota dokumentiem.

Tagad pievērsīsimies otrajai fāzei. Ir dažādas pieejas (Meuschke & Gipp, 2013):

*Leksiskā līdzība*

To sauc arī par teksta virkņu līdzības analīzi. Tas sākas ar nenozīmīgas informācijas izņemšanu no dokumenta un dažiem citiem pirmapstrādes soļiem, kas tiek veikti, lai palielinātu analīzes ātrumu un precizitāti. Mērķis ir atklāt vārdu kārtības maiņu un dažādas vārdu formas, pārrakstot veselu teikumus, paragrāfus vai nodaļas. Iespējams, visbiežāk sastopamā metode ir "sagriešana", kurā dokuments tiek sadalīts gabalos (secīgu vārdu secībās), šie gabali tiek marķēti un saglabāti datubāzē. Bieži sastopamo gabalu esamība kalpo kā līdzības mērs.

*Semantiskā analīze*

Semantiskā analīze ir teksta nozīmes analīze. Aizstājot sinonīmus, veicot dažādus statistikas analīzes veidus un sarežģītu semantisko attālumu aprēķinu, mēs iegūstam dokumentu kopu, kas visvairāk ir saistīti ar aizdomīgo tekstu.

Viens no iespējamiem veidiem, kā to izdarīt, ir izmantot vektora telpas modeli. Šajā modelī mēs apskatām dokumentu kā vektoru n-dimensionālā vektora telpā [kur n ir to vārdu skaits, kas parādās abos dokumentos (kopas apvienojums)]. Katras koordinātes vērtība atspoguļo konkrētā vārda sastopamības biežumu dokumentā, kas ir svērts ar šāda vārda kopējo biežumu. Lai paaugstinātu semantiskās analīzes precizitāti, var pielietot dažādas svaru piešķiršanas shēmas, dimensiju pārkārtošanu un samazināšanu. Pēc tam divu dokumentu vai fragmentu līdzība tiek aprēķināta kā divu vektoru kosinuss, kas izsaka leņķi starp tiem. Jo līdzīgākie ir divi fragmenti, jo mazāks ir leņķis starp atbilstošajiem vektoriem. Jāpiebilst, ka šāda metode var novest mūs pie potenciāliem avotiem, taču tās nesniedz skaidrus pierādījumus tam, ka aizdomīgā dokumenta autors tiešām nokopēja tekstu.

*Stilometriskā analīze*

Stilometriskās analīzes pamatā ir teikumu sastāvs, izmantotās izteiksmes, frāzes, izteikumi, personas, skaitļi utt., mēs varam atpazīt autora rakstīšanas stilu un identificēt tās vietas tekstā, kur rakstīšanas stils mainās. Šīs vietas, iespējami, norāda uz autora maiņu. Ja mums ir pietiekami daudz viena autora rakstīto dokumentu, mēs varam apstiprināt vai noliegt aizdomīga teksta autorību. Tomēr šī metode joprojām nav ļoti droša un nesniedz skaidrus pierādījumus par plaģiātismu.

Izmantotie informācijas avoti:

Stamatatos, E., Potthast, M., Rangel, F., Rosso, P., & Stein, B. (2015). Overview of the PAN/CLEF 2015 Evaluation Lab. In J. Mothe et al. (Eds.), *Experimental IR Meets Multilinguality, Multimodality, and Interaction. 6th International Conference of the CLEF Initiative (CLEF 15)* (pp. 518–538). Berlin Heidelberg New York: Springer. https://doi.org/<http://dx.doi.org/10.1007/978-3-319-24027-5_49>

Meuschke, N., & Gipp, B. (2013). State-of-the-art in detecting academic plagiarism. *International Journal for Educational Integrity*, *9*(1), 50–71. Retrieved from <http://www.ojs.unisa.edu.au/index.php/IJEI/article/view/847>

Piezīmes mācībspēkiem

Šajā dokumentā ir sniegta īsa informācija par teksta salīdzināšanas metodēm. Mērķis ir iegūt vispārīgu priekšstatu. Detalizētas informācijas iegūšanai ir ieteicams izmantot Meuschke & Gipp 's (2013) rakstu.