

11. VAJA

Plagiatorstvo











UVOD

Plagiatorstvo je na Wikipediji definirano kot neupravičeno prisvajanje, natančno posnemanje in kraja z objavljanjem govora, misli, idej in izrazov drugega avtorja ter njihovo predstavljanje kot svoje lastno izvirno delo. Odgovornost avtorja je, da spoštuje avtorsko pravico drugih [1]. Plagiatorstvo je goljufija, ki predstavlja ne le prisvajanje celotnega dela, temveč tudi kopiranje in uporabo delov, odlomkov in idej, na kar pogosto pozabljamo. Plagiatorstvo predstavljajo vse oblike prisvajanja tujega avtorskega dela, še posebej pa [2]:

- predstavljanje tujega dela kot lastno,
- kopiranje besedila ali idej nekoga drugega brez navedbe vira oz. avtorstva,
- opustitev narekovajev v navedbi,
- podajanje nepravilnih podatkov o viru navedbe,
- kopiranje misli, stavka in spreminjanje besed v njem brez navajanja vira oz. avtorstva,
- kopiranje tolikšne količine besedila ali idej iz drugega vira, da predstavljajo večji del novega dela, ne glede na to, ali je vir naveden ali ne.

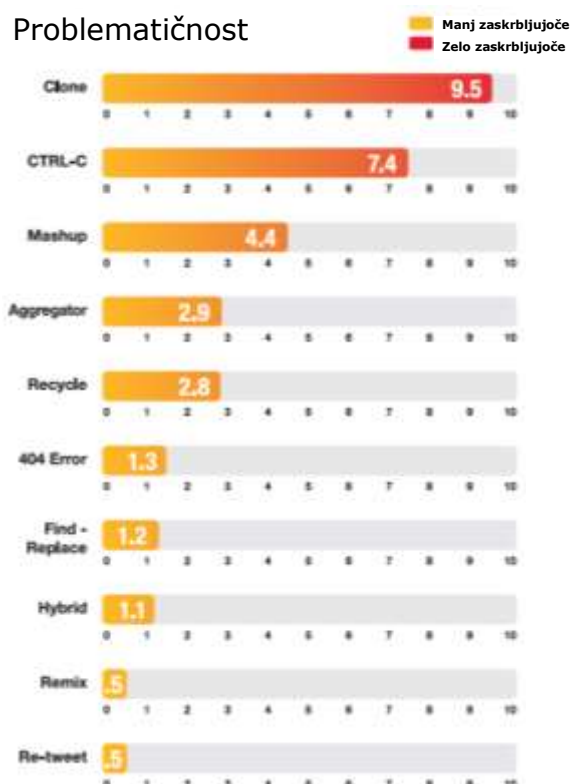
V največjem številu primerov bi se lahko izognili plagiatstvu že z ustreznim navajanjem vira. Obstaja več različnih razvrstitev plagiatstva. Eno podaja tabela 1, njihovo pogostost in problematičnost pa prikazujeta sliki 1 in 2.

Tabela 1: Različni tipi plagiatstva, razporejeni od najbolj do najmanj izrazitih in problematičnih [3–5]

1.		"Clone"	Predstavljanje celotnega tujega dela kot lastno.
2.		"CTRL-C"	Besedilo vsebuje velik delež besedila iz nekega vira brez citiranja, ki je lahko izveden z načinom kopiraj-prilepi ali s pretipkavanjem.
3.		"Find-Replace"	Besedilo je brez citatov in vsebuje velik delež tujega besedila, kjer so besede dodane, odstranjene ali zamenjane s sinonimi, dodane namerne črkovalne ali slovnične napake in spremenjena slovnica in slog pisanja.
4.		"Remix"	Besedilo vsebuje daljše dele besedil iz več različnih virov, tako da spada skupaj, vendar brez navedbe originalnih virov.
5.		"Recycle"	Besedilo vsebuje daljše dele avtorjevega prejšnjega dela brez navedbe citata, kar imenujemo tudi samoplagiatstvo.
6.		"Hybrid"	Besedilo vsebuje daljše dele besedil iz več različnih virov, vendar avtor citira le nekatere vire.
7.		"Mashup"	Besedilo vsebuje kopirano besedilo (odstavke, stavke ali le nekatere besedne zveze) iz različnih virov.
8.		"404 Error"	Besedilo vsebuje namerno nenatančno citiranje, npr. citate na neobstoječe ali nepravilne vire oz. literaturo.
9.		"Aggregator"	Vsebuje citate na pravilne vire, vendar besedilo praktično ne vsebuje nič novega oz. originalnega.
10.		"Re-tweet"	Vsebuje citate na pravilne vire, vendar je besedilo preveč podobno originalnemu besedilu in/ali strukturi.



Slika 1: Pogostost posameznega tipa plagiatorstva [4].



Slika 2: Problematicnost posameznega tipa plagiatorstva [4].

Zanimiv fenomen je **samoplagiatorstvo**, ki mu pogosto rečemo tudi "recikliranje", in je ponovna uporaba večjega dela povsem identičnega lastnega, že prej objavljenega besedila brez navedbe izvirnega dela. Pravno gledano ni s tem nič narobe, v akademskih in raziskovalnih sferah pa je neetično, saj mora biti znanstveno in raziskovalno delo izvirno v delih besedila, ki niso citirani [2].

NAČINI ODKRIVANJA PLAGIATOV [5]

Način kopiraj-prilepi je za avtomatizirano odkrivanje plagiatov najlažji. Za tak plagiat zadošča najosnovnejša in najstarejša metoda – **primerjava nizov znakov**. Več kot je skupnih nizov, večja je verjetnost plagiatu.

Pri parafraziranem besedilu je odkrivanje plagiatu zahtevnejše in osnovna metoda ne zadošča. Sodobnejše metode, ki sicer še vedno temeljijo na primerjavi nizov znakov, ugotavljajo delno ujemanje. Tako niso več vezane na popolno zaporedje besed in stavkov v besedilu, ampak najdejo ujemanja istih nizov besed tudi, če so v originalu in plagiatu različno razporejeni (drugačen besedni red, prerazporeditev odstavkov ali poglavij besedila).

Ena izmed najbolj razširjenih metod iz tega sklopa se zaradi smiselne povezave s prstnimi odtisi imenuje **besedilni odtis** (angl. *fingerprinting*). Med prvimi jo je že leta 1994 predstavil Manber. Pri tej metodi predpostavljamo, da dve datoteki vsebujeta pomembno število enakih delov besedila, ki niso premajhni, ob tem pa je lahko določeno število delov tudi različno; prav tako sta lahko datoteki različno veliki, če je vsebina ene datoteke samo manjši del vsebine druge datoteke. Metoda je uporabna tako za odkrivanje plagiatov besedil

kot programske kode, ni pa sposobna prepoznati podobnosti, če je ista vsebina parafrazirana z drugimi besedami. Postopek po Manberju je v osnovi sledeč: najprej določimo nek niz znakov, npr. akt (to imenujemo sidro), nato pa preiščemo celotno besedilo za temi nizi znakov (našli jih bomo recimo v besedah aktovka, aktiven, trakt ...). Tam, kjer jih najdemo, preverimo še dodatnih npr. 50 znakov, ki sledijo temu nizu. Tako dobimo niz 53 znakov, ki jih imenujemo podrobnost. Zbirka vseh podrobnosti nekega besedila predstavlja besedilni odtis. Enako naredimo za vse datoteke, ki jih pregledujemo. Več kot je med dvema besedilnima odtisoma enakih podrobnosti, večja je verjetnost, da je eno od besedil plagiat. Osnova metode je dobra določitev sider. Za večjo uspešnost se priporoča uporaba več sider, dobra sidra pa lahko določimo npr. tako, da analiziramo besedila več datotek in poiščemo niz znakov, ki je sicer pogost, vendar ne prepogost.

Značilna metoda, ki se prav tako lahko uporablja za odkrivanje plagiatov, je še t. i. **model vreče besed** (angl. *bag-of-words model*). Pri tej metodi ne upoštevamo besed v kontekstu, ampak gledamo na besedilo ali na posamezne dele besedila, recimo, odstavke ali povedi – ilustrativno – kot vrečo, v katero vržemo vse različne besede, iz katerih je bilo besedilo zgrajeno. Vsaki besedi dodamo podatek, kako pogosto se pojavlja. Več kot je skupnih točk med dvema datotekama, večja je verjetnost plagiata. To metodo uporablja tudi slovenski detektor plagiatov v Digitalni knjižnici Univerze v Mariboru.

Najbolj napredne tehnike odkrivanja plagiatov vključujejo **metode procesiranja naravnega jezika**. Tako med drugim uporabljajo osnovna jezikovno-tehnološka orodja, kot so lematizator, ki vsaki besedi v besedilu pripiše njeno osnovno, slovarsko obliko; oblikoslovni označevalnik, ki vsaki besedi pripiše tudi njene oblikoslovne lastnosti (besedno vrsto, spol, sklon ...); skladijski razčlenjevalnik, ki stavek skladijsko razčleni (na osebek, povedek, predmet ...); ali še kakšne slovarske vire – zlasti se priporoča uporaba vira, kot je *Wordnet*, ki omogoča odkrivanje sinonimov. V pregibnih jezikih, kot je slovenski, veliko pomaga že, če namesto pregibnih oblik besed, ki so v besedilu, primerjamo ujemanje besed v osnovni obliki: namesto da bi iskali ujemanje med *gremo na Jamajko* in *grem na Jamajko*, iščemo ujemanje med *iti na Jamajka* in *iti na Jamajka*. Pri parafraziranih plagiatih metode, ki vključujejo orodja procesiranja naravnega jezika, omogočajo večjo uspešnost. Mnogi vidijo v tem tudi največji potencial za nadaljnje izboljšanje avtomatiziranega odkrivanja plagiatov.

Plagiat, ki vsebuje različne **tehnične trike**, bi bilo preprosto odkrivati, če bi avtorji programov za odkrivanje plagiatov trike poznali. Večina orodij takih načinov ne zazna dovolj uspešno. Tehnični triki, ki najpogosteje preliščijo programe za odkrivanje plagiatov so:

- vstavljanje znakov iz tujih abeced, ki imajo zelo podoben videz (npr. Unicode 004F, 039F ali 041E),
- vstavljanje belo obarvanih črk namesto presledkov,
- vstavljanje skeniranega besedila kot slike.

Nazadnje omenimo še prav poseben sklop metod za avtomatizirano odkrivanje plagiatov, ki ima velike možnosti za uporabo zaradi dveh razlogov: prvič, z njimi lahko odkrijemo skoraj vse tipe plagiatorstva, in drugič, zanje ne potrebujemo originala. Vendar se tem metodam pripisuje manjša zanesljivost kot tistim, ki primerjajo plagiat z originalom, zato so bistveno manj razširjene. Gre za postopke, v katerih **analiziramo avtorjev slog**. Pogoje je, da imamo na voljo najmanj deset različnih izvirnih besedil v skupnem obsegu recimo najmanj tisoč besed, ki jih je napisal avtor, ki ga želimo preveriti. S tem posegamo na področje, ki se ga je prijelo ime 'forenzično jezikoslovje' in dela svoje prve korake tudi v

Sloveniji. Gre za postopke, v katerih na osnovi izvirnih besedil avtorja določimo značilne parametre njegovega sloga. Elementi besedila, ki so lastni avtorju, so:

- *Besedne lastnosti*; izračunamo lahko vektorje besednih frekvenc, pri čemer so najpogostejše pomembne predvsem funkcijske besede (predlogi, zaimki, pomožni glagoli, vezniki, členki), od katerih lahko določene uporabljamo bolj pogosto, kot je običajno.
- *Znakovne lastnosti*; izdelajo se znakovni n-grami oz., poenostavljeno povedano, pogostosti zaporednih nizov črk v besedilu. Način je presenetljivo uspešen.
- *Skladenjske lastnosti*; izdajali naj bi nas tudi značilni skladenjski vzorci, se pravi struktura stavkov in povedi.
- *Semantične lastnosti*; analizirajo se s pomočjo iskanja sinonimov in hipernimov (npr. s pomočjo Wordneta).

PROGRAMI ZA ODKRIVANJE PLAGIATOV

Programi za odkrivanje plagiatorstva morajo imeti naslednje tri lastnosti [6]:

1. *Nedovzetnost na oblikovne znake* (presledki, tabulatorji, nove vrstice itd.) – s tem je mišljena predvsem sposobnost odkrivanja plagiata ne glede na to, koliko takšnih znakov vstavimo med posamezne besede.
2. *Nedovzetnost na šum* – pomeni, da morajo biti ujemajoči deli besedil ali izvirne kode dovolj dolgi, da lahko govorimo o plagiranju, saj besedilo ni plagiat, če se v njem naključno pojavi beseda ali besedna skupina kot v nekem drugem dokumentu.
3. *Neodvisnost od položaja* – pomeni, da mora program najti dele besedila, ki so plagiat, ne glede na to, kje v dokumentu se nahajajo.

Predvsem za uporabo v akademskem okolju je bilo razvitih več sistemov za odkrivanje in s tem preprečevanje plagiatorstva. Vsi temeljijo na statistični obdelavi besedil in njihovi primerjavi z obstoječo bazo znanja, to je v preteklosti napisanimi besedili, ki služijo za primerjavo. Programi, s katerimi primerjamo različna besedila, ne določijo, katero besedilo je original in katero plagiat, torej plagiatorstva sami ne odkrivajo, pač pa opozarjajo na podobne dele besedil in izračunajo odstotek podobnosti [2]. Presoja in odločitev, ali gre za plagiat ali ne, pa je prepuščena človeku, npr. profesorju oz. mentorju ali uredniku, ki preverja članke pred objavo. Določiti je potrebno mejne vrednosti o podobnosti dveh datotek. Pri tem je treba razsoditi npr. ali je prišlo do naključne enakosti dela besedila, ali je prišlo do prepisovanja enega avtorja od drugega, ali sta oba avtorja prepisovala iz tretjega vira, ali gre pri enakem delu besedila za citate in dejstva, ki jih ni mogoče drugače zapisati ... [6]

Obstaja precej orodij, s katerimi si lahko pomagamo pri odkrivanju plagiatorstev. Bojan Butolen, avtor diplomske naloge o plagiatorstvu [6], izpostavi predvsem aplikacije *DOC Cop*, *WCopyFind*, *Plagiarism Detector*, *Plagiarism Finder* in *Anti-plagiarist*. Omeniti pa moramo še Turnitin-ov *OriginalityCheck*, ki je ta čas morda eden izmed najbolj znanih in ga uporabljamo tudi na Univerzi v Ljubljani. Nekatere izmed njih so prosto dostopne, nekatere pa plačljive. Razlike med njimi pa niso le v plačljivosti, ampak še npr. v načinu delovanja, kakšne možnosti dajejo uporabniku, katere datotečne formate podpirajo in kakšne rezultate vračajo.

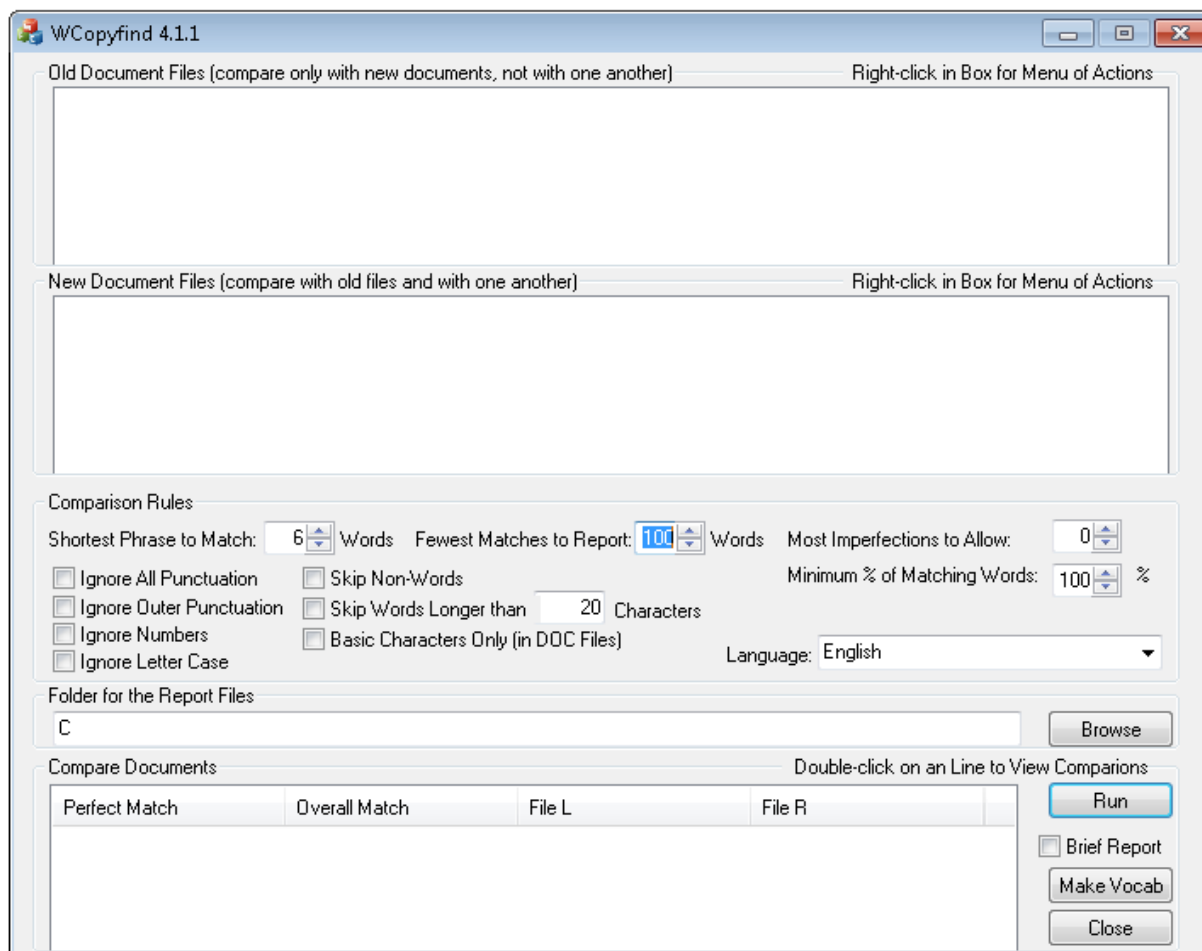
Pri tej vaji se bomo osredotočili na aplikaciji *WCopyFind* in *Plagiarism Detector*, s katerima lahko za plagiatorstvo preverjamo tudi slovenska besedila.

WCopyFind [6–8]

WCopyFind je namizna aplikacija, ki je brezplačna in prosto dostopna na spletu [7,8]. Deluje na operacijskih sistemih Windows in Linux. Namen aplikacije *WCopyFind* je primerjava zbirke dokumentov, pri katerih sumimo na plagiatstvo. Program iz dokumentov vzame besedilo in primerja dele besedila med dokumenti za natančno ujemanje. O opravljeni primerjavi se ustvari poročilo v obliki dokumenta v formatu *.html*, v katerem si lahko ogledamo, kateri deli dokumentov se ujemajo ter odstotek podobnosti med vsakim parom dokumentov. V programu lahko nastavimo nekaj parametrov, s katerimi določamo, na kakšen način bomo dokumente med seboj primerjali. Nastavimo lahko (slika 3):

- minimalno dolžino dela besedila, za katero želimo preveriti podobnost,
- minimalno dovoljeno število ujemaajočih delov besedila,
- minimalno dolžino besedila v znakih,
- ignoriranje posebnih znakov (števila, ločila, obravnavanje malih in velikih črk kot enake),
- preskakovanje skupin znakov, ki niso besede,
- preskakovanje dolgih besed,
- določanje maksimalne različnosti med deloma besedila.

S temi nastavitvami lahko odločilno vplivamo na rezultat primerjave, saj lahko odstranimo veliko motečih faktorjev in s tem dobimo natančnejšo primerjavo. Z nastavitvami pa vplivamo tudi na trajanje same primerjave.



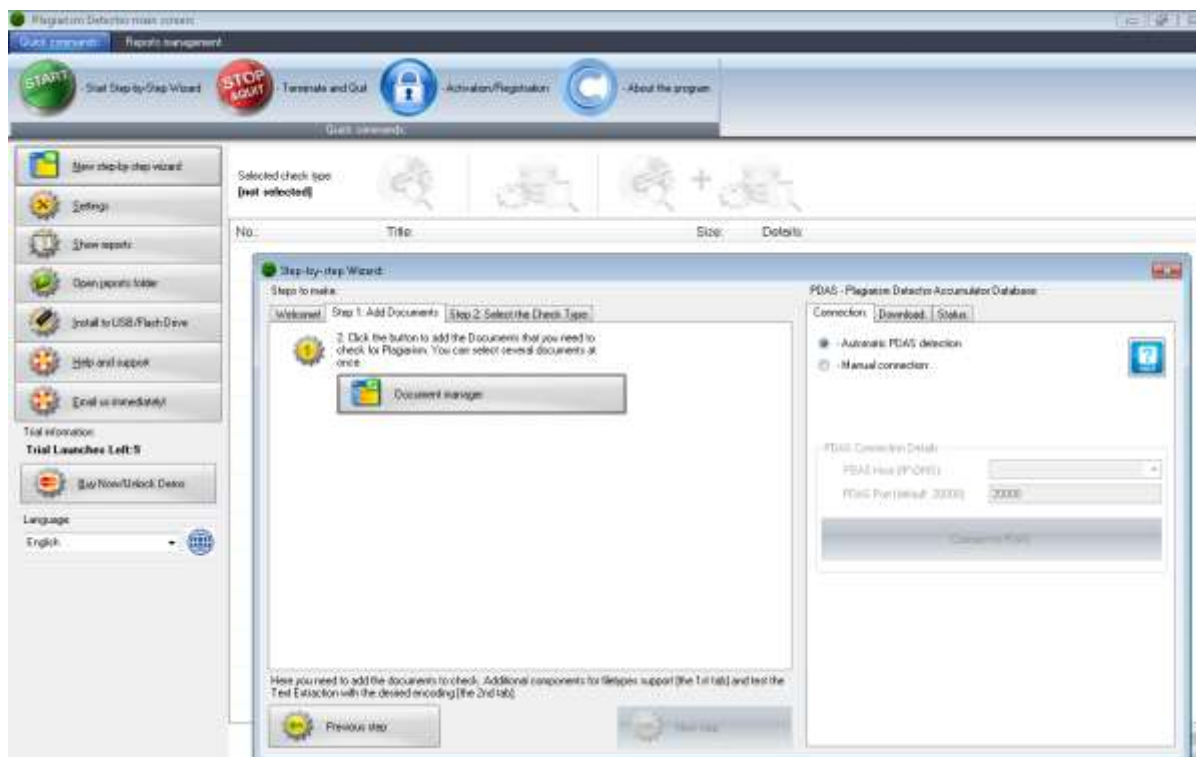
Slika 3: Program *WCopyFind* s prikazanimi možnimi nastavitvami parametrov.

WCopyFind podpira datotečne formate *.docx*, *.doc*, *.txt*, *.pdf* in *.html*. Program ne podpira neposredne primerjave z internetnimi viri. To slabost lahko delno odpravimo tako, da primerjavo izvedemo z bližnjicami do spletnih strani z viri, a moramo spletni naslov, s katerim želimo dokument primerjati, poznati. Program deluje tako, da izbere dva dokumenta, ki ju je potrebno primerjati, ter določi enega kot levega in drugega kot desnega. V levem dokumentu izbere prvo besedo, ki je daljša od treh znakov in se pojavi v obeh dokumentih. Ko najde tako besedo, okolico besede, ki je izbrana v levem dokumentu, primerja z okolico iste besede v desnem dokumentu. Ker se lahko posamezna beseda v dokumentu pojavi večkrat, program preveri vse enake besede v desnem dokumentu na enak način. Če se okolici ujemata in ujemanje preseže maksimalno dovoljeno dolžino, potem se del besedila označi kot enak in primerjava se nadaljuje.

Plagiarism Detector [6,9]

Plagiarism Detector je ena izmed namiznih aplikacij za primerjanje datotek in odkrivanje plagiatorstva. Program deluje na operacijskem sistemu Windows in uporablja internetne iskalnike *Google*, *Yahoo* in *Altavista* za primerjavo datotek z viri na internetu. Aplikacija podpira datotečne formate *.docx*, *.doc*, *.txt*, *.pdf*, *.html*, *.htm*, *.rtf*, *.ppt* in *.php*. Aplikacija je plačljiva, a jo lahko testiramo v brezplačni demo različici (slika 4) [9].

Slabost aplikacije je, da v času primerjave lahko zasede tudi do 100 % procesorske moči, prednost pa, da je hitrost primerjave zato velika, saj npr. aplikacija lahko okrog 4 strani besedila preveri v 4 do 5 minutah, kljub temu da preverja dokumente s kar tremi podatkovnimi bazami največjih internetnih iskalnikov. Program uporabniku nudi več statistike in informacij v poročilu primerjave kot *WCopyFind*, vendar manj možnosti za nastavljanje parametrov. Nastavimo lahko, kako dolgi naj bodo deli besedila, ki jih želimo primerjati, in kakšen naj bo razmik med posameznimi deli besedila.



Slika 4: Program *Plagiarism Detector*.

Delovanje algoritma aplikacije je odvisno od nastavitve uporabnika, ki iz besedila izbere dele besedila določene dolžine na vsakih nekaj besed. Izbrane dele besedila sestavi v seznam in za vsak del pošlje zahtevo za iskanje na internetne iskalnike. Program seznam strani, ki ga iskalniki vrnejo, uporabi pri nadaljnji primerjavi. Vsak zaznan vir, ki bi lahko predstavljal vir plagiatorstva, program najde na spletu in ga primerja. Program najprej izvede primerjavo besedo za besedo. Po končani primerjavi vseh virov dobimo rezultat, ki opisuje, kolikšen del besedila predstavlja potencialno plagiatorstvo, kolikšen del je originalen in kateri viri so v največji meri uporabljeni.

NAMEN VAJE

- Seznaniti se s tem, kaj je plagiatorstvo.
- Seznaniti se z uporabo dveh različnih programov za preverjanje plagiatorstva.
- Preveriti besedilo dokumenta, če in v kolikšnem deležu je plagiat.

DOMAČA NALOGA

- **1. del:** Na računalnik naložite aplikacijo *WCopFind* [7,8] in preučite, kakšne vrednosti parametrov bi bilo najbolj smotrno uporabiti. S to aplikacijo preverite besedilo svoje seminarske naloge za plagiatorstvo. Za primerjavo uporabite vse spletne strani in članke v slovenskem jeziku, ki ste jih uporabljali pri izdelavi svoje seminarske naloge. Če pri izdelavi naloge niste uporabljali literature v slovenskem jeziku, na spletu poiščite vsaj pet virov v slovenskem jeziku in jih uporabite za primerjavo z vašim besedilom seminarja. V poročilu o podobnosti dokumentov podajte podatke o ujemanju vašega seminarja s primerjanimi dokumenti in katere nastavitve oz. vrednosti posameznih parametrov ste tem uporabili. Poročilo shranite z imenom **vaja11_WCop_ime priimek**.
- **2. del:** Na računalnik naložite demo različico aplikacije *Plagiarism Detector* [9] in preučite, kakšne vrednosti parametrov bi bilo najbolj smotrno uporabiti. S to aplikacijo preverite besedilo svoje seminarske naloge za plagiatorstvo. V poročilu o podobnosti dokumentov podajte podatke o ujemanju vašega seminarja s primerjanimi dokumenti in katere nastavitve oz. vrednosti posameznih parametrov ste pri tem uporabili. Poročilo shranite z imenom **vaja11_PlagD_ime priimek**.
- Dokumenta oddajte **do četrтка, 5. 1. 2017, do 8.00**.

LITERATURA

- [1] Zakon o avtorski in sorodnih pravicah. Uradni list Republike Slovenije. Uradno prečiščeno besedilo (ZASP-UPB3), št. 16, str. 1805-1830, 23. 2. 2007. (www.uradni-list.si)
- [2] Biblioblog: Plagiatorstvo in njegovo odkrivanje: <http://www.biblioblog.si/2011/03/plagiatorstvo-in-njegovo-odkrivanje.html>, (dostop: jan 2016).
- [3] What is plagiarism? <http://www.plagiarism.org/plagiarism-101/what-is-plagiarism/>, (dostop: jan 2016).
- [4] White paper: The plagiarism spectrum – Instructor insights into the 10 types of plagiarism: https://www2.nau.edu/d-elearn/support/tutorials/academicintegrity/pdf/Turnitin_WhitePaper_PlagiarismSpectrum.pdf, (dostop: jan 2016).

- [5] Verdonik D. *Išče se original ...* Življenje in tehnika. št. 5, str. 12-19, maj 2013.
- [6] Butolen B. *Plagiatorstvo in programi za detekcijo plagiatov*. Diplomsko delo. Fakulteta za elektrotehniko, računalništvo in informatiko, Univerza v Mariboru, Maribor, maj 2009.
- [7] The plagiarism resource site, New release: *WCopYFind 4.1.1*, Charlottesville, Virginia:
<http://plagiarism.bloomfieldmedia.com/z-wordpress/2012/03/05/new-release-wcopyfind-4-1-1/>,
(dostop: dec 2016).
- [8] The plagiarism resource site, *WCopYFind Instructions*: <http://plagiarism.bloomfieldmedia.com/z-wordpress/software/wcopyfind/>, (dostop: dec 2016).
- [9] Plagiarism Detector, Software to search and detect plagiarism:
<http://www.plagiarism-detector.com/>, (dostop: dec 2016).