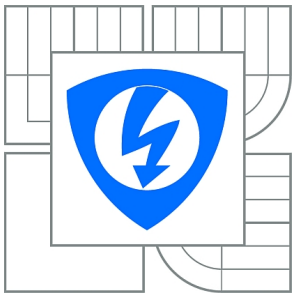
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

ÚSTAV AUTOMATIZACE A MĚŘICÍ TECHNIKY

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

DEPARTMENT OF CONTROL AND INSTRUMENTATION

DETEKCE PLAGIÁTŮ

PLAGIARISM DETECTION

BAKALÁŘSKÁ PRÁCE

BACHELOR'S THESIS

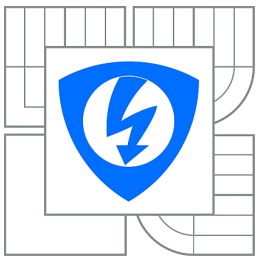
AUTOR PRÁCE — MARTIN KOBATH

AUTHOR

VEDOUCÍ PRÁCE — Ing. PETR HONZÍK, Ph.D.

SUPERVISOR

BRNO 2015



VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

Fakulta elektrotechniky a komunikačních technologií

Ústav automatizace a měřicí techniky

Bakalářská práce

bakalářský studijní obor
Automatizační a měřicí technika

Student:	Martin Kobath	***ID:***	158635
Ročník:	3	***Akademický rok:***	2014/2015

NÁZEV TÉMATU:

Detekce plagiátů

POKYNY PRO VYPRACOVÁNÍ:

Cílem práce je vytvořit sw, který na základě předloženého dokumentu v pdf vyhodnotí, zda jsou jeho součástmi i texty s největší pravděpodobností vykopírované z jiných veřejně dostupných zdrojů. Hlavní odbornou náplní je automatizování výběru fragmentů textu, u kterých je předpokládána zvýšená pravděpodobnost, že nejsou psány autorem většiny textu. Jedná se o úlohu umělé inteligence spadající do kategorie učení bez učitele.

1. Zpracujte rešerši na téma plagiátorství.
2. Posuďte využitelnost dostupných programů určených pro detekci plagiátů.
3. Nastudujte a popište algoritmy používané pro detekci plagiátů.
4. Zaměřte se na intrinsické metody, navrhněte způsob výběru fragmentů textu vhodných pro ověření např. ve vyhledávači Google.
5. Systém realizujte s cílem umožnit práci se závěrečnými pracemi vzniklými na vaší Alma mater.
6. Proveďte dva typy experimentů. S uměle vytvořenými plagiáty (vyhodnoťte míru úspěšnosti jejich detekce) a dále s minimálně 100 náhodně vybranými pracemi vzniklými na fakultě. Případné pozitivní nálezy konzultujte s vedoucím. Výsledky prezentované v práci anonymizujte.

DOPORUČENÁ LITERATURA:

M. Potthast et al.: Overview of the 5th International Competition on Plagiarism Detection, CLEF 2013 Evaluation Labs and Workshop 2013.

Termín zadání:	9.2.2015	***Termín odevzdání:***	25.5.2015

Vedoucí práce: Ing. Petr Honzík, Ph.D.
Konzultanti bakalářské práce:

doc. Ing. Václav Jirsík, CSc.
Předseda oborové rady

UPOZORNĚNÍ:

Autor bakalářské práce nesmí při vytváření bakalářské práce porušit autorská práva třetích osob, zejména nesmí zasahovat nedovoleným způsobem do cizích autorských práv osobnostních a musí si být plně vědom následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č.40/2009 Sb.

Bibliografická citace práce:

KOBATH, M. *Detekce plagiátů.* Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, 2015. 59 s. Vedoucí bakalářské práce Ing. Petr Honzík, Ph.D..

PROHLÁŠENÍ

Prohlašuji, že jsem svoji bakalářskou práci na téma Detekce plagiátů vypracoval samostatně pod vedením vedoucího bakalářské práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou v práci citovány a uvedeny v seznamu použité literatury.

Jako autor uvedené bakalářské práce dále prohlašuji, že v souvislosti s vytvořením této bakalářské práce jsem neporušil autorská práva třetích osob, zejména jsem nezasáhl nedovoleným způsobem do cizích autorských práv osobnostních a jsem si plně vědom následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č. 40/2009 Sb.

Poděkování

Děkuji vedoucímu bakalářské práce Ing. Petru Honzíkovi, Ph.D. za účinnou pedagogickou pomoc a další cenné rady při zpracování mé bakalářské práce.

V Brně dne . . . .

podpis autora

ABSTRAKT

Bakalářská práce se zaměřuje na detekci plagiátorství u psaných textů, specificky u závěrečných prací na autorově Alma mater. V práci je zhodnocena současná situace automatické detekce plagiátorství česky psaných textů a je proveden teoretický rozbor zaměřený na klasickou detekci pomocí hledání shod v dokumentu s využitím externích zdrojů, zvláště internetu.

Práce je ukončena vlastním návrhem automatického detekčního systému a experimentem na 100 náhodně vybraných prací z Fakulty elektrotechniky a komunikačních technologií.

KLÍČOVÁ SLOVA

plagiátorství, detektor plagiátů, shoda textů, závěrečná práce, vyhledání textu na internetu

ABSTRACT

The Bachelor‘s thesis concentrates on plagiarism detection in written text, mainly final theses in author’s Alma mater. The current situation of automatic plagiarism detection of text in czech language is evaluated and a theoretic analysis directed on classic detection by finding matches in a text with use of external sources, mainly the internet, is made.

Project ends with custom concept of automatic plagiarism detection system and with an experiment made on 100 randomly chosen theses from The Faculty of Electrical Engineering and Communication.

KEY WORDS

plagiarism, plagiarism detector, text sameness, final thesis, text search on internet

OBSAH

SEZNAM OBRÁZKŮ

SEZNAM TABULEK

1. Úvod

Práce se zabývá hledáním zdrojů předloženého textu na internetu za účelem odhalení plagiátorství. Náplň práce je rozdělena do tří částí: zhodnocení současných možností automatické detekce plagiátorství, teoretický rozbor problematiky automatické detekce plagiátorství a vlastní návrh detekčního systému.

Prvním bodem je zhodnocení funkčnosti některých volně dostupných detekčních systémů v případě práce s českými texty. Pro potřeby testů je vytvořen umělý plagiovaný dokument a je sledována schopnost detektorů nalézt původní zdroj textu.

Další částí je teoretický rozbor, jehož součástí je i úvodní definice pojmů. Samotný teoretický rozbor se zaměřuje na detekci plagiátorství pomocí hledání shod mezi dvěma texty s ohledem na použití internetu jako zdroje dokumentů. Kromě samotné detekce shod se zde řeší i předzpracování dokumentů před hledáním shod, způsob hledání možných zdrojových dokumentů na internetu a některé známé způsoby, kterými se plagiátoři snaží automatické detekční systémy obcházet.

Poslední část je věnována popisu vlastního návrhu automatického detekčního systému plagiátorství a konečným výsledkem je i realizace vlastního detekčního systému a jeho použití na vzorek 100 náhodně vybraných závěrečných prací.
Vytvořený detekční systém je součástí práce a může být kýmkoliv volně použit na kontrolu libovolné veřejně přístupné práce.

2. DEFINICE POJMŮ TÝKAJÍCÍCH SE PLAGIÁTORSTVÍ

Nelze začít hovořit o způsobech, jak detekovat plagiátorství, aniž bychom napřed nedefinovali co je a co není plagiát.

2.1. Základní pojmy

2.1.1. Plagiát

Norma ČSN ISO 5127-2003 definuje plagiát jako „představení duševního díla jiného autora půjčeného nebo napodobeného v celku nebo z části, jako svého vlastního" (Infogram.cz, [2008]). Tato definice je vhodná pro pojem plagiátorství jako činnosti. Nevztahuje se však už na produkt plagiátorství, který lze nazvat jako plagiát.

Definice se totiž odvolává na proces vzniku plagiátu. Plagiát jako takový, například ve formě dokumentu, nemá žádné znaky, které by ho odlišovali od jiných dokumentů stejného druhu. Odlišuje ho jedině fakt, že informace v něm obsažené autor neprávem vydává jako své vlastní dílo. Proto lze plagiát definovat pouze jako dílo, u kterého bylo prokázáno, že vzniklo procesem plagiátorství.

V následujícím textu se zabýváme pouze díly textovými, tedy dokumenty a to výhradě v českém jazyce.

2.1.2. Kompilace

Význam slova kompilace je sbírání, shromažďování nebo sestavování. Knihovna ČVUT definuje kompilaci jako „text vzniklý složením myšlenek a závěrů sebraných z více jiných původních textů, ne však kopírování celých doslovných pasáží textu. Kompilace neobsahuje žádný nový tvůrčí poznatek k tématu, není výsledkem výzkumné činnosti autora, je pouze složením již známých a publikovaných faktů a podává ucelený pohled na danou problematiku. Použité zdroje se řádně citují a odkazují, výsledná práce je prezentována jako kompilace, nevydává se za originál.“ (Němečková, 2009).

Kompilací může být celý dokument nebo jen jeho část. Je-li splněna podmínka uvedení použitých zdrojů, nelze kompilaci označit za plagiát. Tato podmínka je také jediná věc odlišující kompilaci od plagiátu.

2.1.3. Parafráze

O parafrázi hovoříme jako o vyjádření obsahu původního díla jinou formou (Němečková, 2009). V případě textu jde často o použití jiných slov. Podobně jako v případě kompilace se nejedná o plagiátorství, pokud je parafrázované dílo řádně ocitováno. Právě kvůli použití vlastních slov dochází u parafrází k největším sporům ohledně určení, jestli je text původní nebo přejatý.

2.2. Právní pozadí

Otázkou plagiátorství se v českém právním řádu zabývá Předpis č. 121/2000 Sb. Zákon o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon), který vstoupil v platnost 1. prosince roku 2000.

Zákon se nezabývá přímo plagiátorstvím, ale místo toho definuje autorské dílo a práva autora k jeho autorskému dílu.

2.2.1. Citace dokumentů

Citace je definována v autorském zákoně v § 31 (zakonyprolidi.cz, [2010]):

Citace

(1) Do práva autorského nezasahuje ten, kdo

a) užije v odůvodněné míře výňatky ze zveřejněných děl jiných autorů ve svém díle,

b) užije výňatky z díla nebo drobná celá díla pro účely kritiky nebo recenze vztahující se k takovému dílu, vědecké či odborné tvorby a takové užití bude v souladu s poctivými zvyklostmi a v rozsahu vyžadovaném konkrétním účelem,

c) užije dílo při vyučování pro ilustrační účel nebo při vědeckém výzkumu, jejichž účelem není dosažení přímého nebo nepřímého hospodářského nebo obchodního prospěchu, a nepřesáhne rozsah odpovídající sledovanému účelu;

vždy je však nutno uvést, je-li to možné, jméno autora, nejde-li o dílo anonymní, nebo jméno osoby, pod jejímž jménem se dílo uvádí na veřejnost, a dále název díla a pramen.

(2) Do práva autorského nezasahuje ani ten, kdo výňatky z díla nebo drobná celá díla citovaná podle odstavce 1 písm. a) nebo b) dále užije; ustanovení odstavce 1 části věty za středníkem platí obdobně.

Způsob provedení citací je definován normou ČSN ISO 690 (01 0197) platnou od 1. dubna 2011. Jedná se o českou verzi mezinárodní normy ISO 690:2010.

Norma nabízí tři způsoby zápisu odkazu na bibliografickou citaci v textu (Biernátová, Skůpa, 2011): Harvardský systém, Forma číselného odkazu a Průběžné poznámky.

Je však třeba dodat, že stylů citace se používá daleko více, což dále komplikuje rozhodování o správnosti citace. Navíc, všeobecně známá fakta není nutné citovat. Protože však citace přiznává zásluhy původnímu autorovi myšlenky a je tedy jediným způsobem jak zabránit tomu, aby byl dany text považován za plagiát, je správné rozeznání citace kritické pro určení plagiovaného textu.

V současnosti neexistuje systém, který by dokázal s jistotou určit, zdali je podezřelý úsek textu skutečně plagiovaný. Proto je konečné rozhodnutí ponecháváno na člověku.

2.3. Typy plagiátorství

Dr. C. Barnbaum ([2009]) rozděluje plagiátorství na pět různých druhů:

1. Copy & Paste (přímé)
2. Záměna slov
3. Plagiátorství stylu
4. Použití metafor
5. Plagiátorství myšlenky

Copy & Paste plagiátorství

Jedná se o přímé přepsání původního textu bez jakékoliv změny. Podle pravidel bibliografické citace je nutné uvést zdroj a odlišit citovaný text od ostatního textu použitím uvozovek nebo jinou metodou.

Záměna slov

V podstatě je tento způsob stejný jako přímé plagiátorství, pouze je změněno pořadí slov v původním textu, nebo jsou některá slova zaměněna za svá synonyma nebo je změněn jejich tvar. S tímto typem plagiátorství se u skutečných plagiátů potkáváme nejčastěji. Plagiátor se bude snažit využít veškeré jemu známé možnosti jazyka, aby se jeho verze textu co nejvíce odlišovala od svého zdroje.

Plagiátorství stylu

Jedná se o formu plagiátorství, kdy autor přímo nepřejímá text ze zdroje, ale sleduje ve svém dokumentu myšlenkový postup a strukturu textu svého zdroje. Pokud bychom vytvořili výtah z obou textů, byly by totožné. Tohoto typu plagiátorství se může autor dopustit i nechtěně, bez toho aniž by si to uvědomil.

Použití metafor

Metafory jsou používány k objasnění myšlenky nebo k poskytnutí analogie, která nabízí lepší způsob vysvětlení. Metafory jsou významnou součástí stylu psaní autora a v případě jejich použití ze zdrojového textu, je třeba tento zdroj uvést.

Plagiátorství myšlenky

Jedná se o případ, kdy je přejata původní myšlenka nebo nápad bez uvedení jejího původního autora. Zde může také dojít k nechtěnému plagiátorství v případě, že autor textu nesprávně rozezná původní myšlenku jako všeobecnou znalost. Mezi původní myšlenky spadají hlavně nové nápady a řešení týkající se specifických problémů. Určení však může být obtížné a liší se v závislosti na tom, co je v daném oboru považováno za všeobecné znalosti.

Sebeplagiátorství (autoplagiátorství)

Databáze Národní knihovny ČR ([2009]) definuje sebeplagiátorství jako: „Publikování, resp. kopírování vlastních dřívějších uměleckých nebo vědeckých prací bez uvedení jejich citací včetně autorství.“

Tento typ plagiátorství se objevuje často v případě, kdy autor netuší, že musí uvádět citace svých vlastních prací stejně jako prací jiných autorů. V horším případě může jít o snahu vydávat pouze trochu pozměněnou starou práci jako práci novou.

3. REŠERŠE NÁSTROJŮ PRO DETEKCI

Tato kapitola poskytuje přehled několika konkrétních nástrojů detekce plagiátorství. Pro zhodnocení funkčnosti těchto nástrojů byly vytvořeny 3 PDF dokumenty obsahující úsek textu vyjmutý z diplomové práce ing. Kubínyi, Automatizace linky pro defektoskopii železničních kol:

Významnou metodou v defektoskopickém oboru jsou zkoušky ultrazvukem [1], na jehož využití poprvé upozornil sovětský fyzik Sokolov. Jeho praktické využití však muselo počkat až do padesátých let, kdy došlo k významného rozvoji elektroniky. Ultrazvuk [7], [8], stejně jako zvuk a hluk, je mechanické kmitání částic (hmotných elementů) kolem rovnovážné polohy, šířící se ultrazvukovou vlnou v pružném prostředí ve frekvenčním rozsahu nad 20 kHz. Pro defektoskopické účely se běžně pracuje ve frekvenčním rozsahu od 100 kHz do 50 MHz, výjimečně až do 200 MHz.

První dokument obsahoval citovaný text v původní podobě. V druhém dokumentu bylo zaměněno pořadí některých slov a v třetím dokumentu došlo na celkovou změnu jednotlivých vět. Zbytek dokumentu byl doplněn obsahově nesouvisejícím textem.

Podle možností testovaného detektoru byly použity buď celé dokumenty, část s plagiovaným textem nebo části plagiovaného textu. Mnoho ze zdarma dostupných nástrojů má omezení počtu vyhledání, proto nebylo možné vyzkoušet je na větším množství vzorků. Přestože tyto zkoušky neposkytují dostatek informací na přesné zhodnocení efektivity zkoušeného nástroje, dávají nám rychlý náhled na schopnosti nástroje a umožňují nám tyto nástroje porovnat.

Tabulka 1: Rozdělení nástrojů pro detekci plagiátorství

Nástroj	cena	vyhledávač	způsob vyhledávání	možnost nahrát soubory
PlagiarismChecker.com	zdarma	Google, Yahoo	proti internetu	ne
Plagiarisma.Net	zdarma / premium	Google, Yahoo, Babylon	proti internetu	ano
Plagiarism Software	zdarma	Google	proti internetu	ano
Plagium	zdarma / premium	vlastní	proti internetu	ano (premium)
Dupli Checker	zdarma	vlastní	proti internetu	ano
PlagScan	zdarma	vlastní	proti internetu	ano
Plagiarism Detector	zdarma	vlastní	proti internetu	ano
Viper	zdarma	vlastní	proti internetu / dokumentu	ano
WCopyfind	zdarma	vlastní	proti dokumentu	ano
PaperRater	zdarma / premium	vlastní	proti internetu	ne

Tabulka 2: Úspěšnost detektorů ve vyhledání zdroje plagiovaného textu

Název detektoru	Zdroj detektoru	Přímé kopírování	Změna pořadí	Změna slov
PlagiarismChecker.com	http://www.plagiarismchecker.com/	ANO	NE	NE
Plagiarisma.Net	http://plagiarisma.net/	NE	NE	NE
Plagiarism Software	http://www.plagiarismsoftware.net/	ANO	ANO	NE
Plagium	http://www.plagium.com/	NE	NE	NE
Dupli Checker	http://www.duplichecker.com/	NE	NE	NE
SeeSources	http://www.plagscan.com/seesources/analyse.php	NE	NE	NE
Plagiarism Detector	http://plagiarismdetector.net/	ANO	ANO	NE
Viper	http://www.scanmyessay.com/	NE	NE	NE
WCopyfind	http://plagiarism.bloomfieldmedia.com/z-wordpress/software/wcopyfind/	ANO	ANO	ANO
PaperRater	http://www.paperrater.com/	NE	NE	NE

3.1. Popis nástrojů

PlagiarismChecker.com

Mnoho na internetu volně dostupných detekčních nástrojů funguje jako prostředník mezi uživatelem a běžně používaným internetovým vyhledávačem, jako je například Google nebo Yahoo. PlagiarismChecker.com není v tomto ohledu výjimkou.

Uživatel je vyzván, aby do nástroje napsal slova, která chce vyhledat. Nástroj požaduje minimum 6 slov na řádek. Vepsané řetězce slov poté nástroj předá vyhledávači Google (nebo Yahoo, podle výběru uživatele), kde každý řádek slov uzavře do uvozovek a jednotlivé řádky oddělí příkazem OR. Uživatel je poté přesměrován na stránku vyhledávače, kde si může prohlédnout výsledky vyhledávání.

Přestože má vyšší úspěšnost než některé jiné uvedené nástroje, je z nich jednoznačně nejhorší. Uzavření textu do uvozovek znemožní dohledat zdroj v případě nesprávného pořadí slov a uživatel nemá možnost vyhledávat bez nich. To, spolu s omezením na minimum 6 slov pro vyhledávání, činí z PlagiarismChecker.com spíše zeď, kterou musí uživatel překonat, aby

se dobral uspokojivých výsledků. Proto je lepší volbou i práce přímo se samotným vyhledávačem.

Je dobré také upozornit, že vyhledávač Yahoo nebyl schopen nalézt ani dokument s přímým plagiátorstvím.

Plagiarisma.Net

Pracuje jako prostředník mezi uživatelem a vyhledávačem. Umožňuje nahrát PDF soubory, ale vyhledávání je omezeno na 2000 znaků a počet vyhledávání na den je také omezen.

Nahraný text nástroj rozdělí a jednotlivé části nechá vyhledat pomocí vyhledávače Google. Samotné vyhledávání pomocí Google dokáže nalézt přímé plagiátorství i s odlišným pořadím slov. Bohužel, Plagiarisma.Net výsledky vyhledávání špatně vyhodnocuje a odkazuje na jiné zdroje než je skutečný zdroj plagiátu.

Plagiarism Software

Tento nástroj zadaný text rozdělí na části a vyhledává je pomocí vyhledávače Google s příkazem allintext. Uživatel si poté může nechat zobrazit výsledky vyhledávání u jednotlivých částí. Je možné také nahrát text ze souboru, ale nástroj nebyl schopen správně přečíst text ve vytvořených PDF souborech.

Plagiarism Software používá poněkud nespolehlivý způsob rozdělení textu na části, který způsobuje neúspěšné vyhledávání, pokud je zadán kompletní text z plagiovaného dokumentu. Při vyhledávání pouze plagiované části textu dokumentu je však úspěšný.

Plagium

Použití zdarma je omezeno na 25 000 znaků a je omezeno i množství použití na dvě za den. Placená verze poskytuje i další nástroje jako možnost nahrávání souboru a alternativní verzi vyhledávání, kde je text rozdělen na odstavce a ty jsou vyhledávány jednotlivě.

Plagium používá vlastní řešení pro vyhledávání na internetu, při kterém používá celý dokument. U jednotlivých výsledků vyhledávání poté zobrazuje procento využití zadaného textu na nalezené stránce. Z výsledků zkoušky lze usoudit, že tento nástroj není příliš vhodný pro vyhledávání textu v českém jazyce.

Dupli Checker

Vyhledávání je omezeno na 1500 slov a na jedno vyhledání za den pro neregistrované uživatele. Umožňuje nahrát text ze souboru ve formátech txt a docx. Výslednou stránku lze zobrazit a porovnat shody s hledaným textem.

Neúspěch i při přímém plagiátorství naznačuje jazykové omezení nástroje.

SeeSources

Omezené vyhledávání na den, maximum 1000 slov a možnost nahrát text ze souborů ve formátech doc, docx, htm a txt. Z textu extrahuje unikátní slova pro vyhledávání.

Nástroj je zřejmě opět omezen jazykovou bariérou, jak naznačuje jeho úspěšnost.

Plagiarism Detector

Tento nástroj je funkčně totožný s Plagiarism Software.

Viper

Nástroj je k dispozici zdarma a vyžaduje instalaci. Je však stále schopný vyhledávat i na internetu, dokáže však vyhledávat jen omezeně dlouhý text. Jeho jedinou slabinou je, že nedokáže zpracovat češtinu.

WCopyfind

Porovnává pouze soubory proti sobě, poskytuje však pokročilé nastavení pro vyhledávání a zobrazuje oba dokumenty vedle sebe pro snadné porovnání nalezených shod. Se správným nastavením bylo dosaženo i přesvědčivé shody pro plagiovaný dokument se změněnými slovy. Výsledek však silně závisel na okolních slovech, která byla nezměněna. Lze proto předpokládat, že pečlivější snahu o skrytí původního textu by nebyl schopen odhalit.

PaperRater

Tento nástroj je určen výhradně pro anglický text. Jedná se spíše o nástroj ke kontrole správné gramatiky a pravopisu. Je však schopný kontrolovat text i na plagiátorství a obsahuje i některé užitečné nástroje jako je analýza stylu, která zjistí průměrný počet slov na větu, celkový počet slov a další statistické údaje, které lze zjistit u textu v libovolném jazyce.

3.2. Závěr

Na základě uvedeného vzorku nástrojů pro detekci plagiátorství, lze tvrdit, že žádný z volně dostupných nástrojů neposkytuje dostatečné možnosti k serióznějším kontrolám dokumentů. V případě pořízení placené verze nástroje sice vymizí omezení uživatele v možnostech vyhledávání, ale stále zůstává problém ve formě minimálního zpracování textu, které tak prakticky omezuje detekci plagiátorství pouze na jeho přímou formu.

V případě některých profesionálních nástrojů mohou zmizet i tyto překážky, ale objevuje se nová překážka ve formě omezení nástroje na několik nejpoužívanějších jazyků.

Jak se zdá, jediným skutečně použitelným detektorem plagiátorství pro dokumenty v češtině je systém Theses Masarykovy univerzity v Brně. Jiný systém na podobné úrovni zatím neexistuje.

4. STROJOVÉ ODHALOVÁNÍ PLAGIÁTORSTVÍ

Tato kapitola se zabývá podrobněji jednotlivými částmi procesu odhalování plagiátorství v psaném textu. Kromě zpracování textu před hledáním shodnosti a samotného hledání shodností, je zde zahrnuto i vyhledávání potenciálních zdrojů na internetu nebo v jiné velké databázi a způsoby, kterými se plagiátoři snaží zmást automatické detekční systémy.

4.1. Typy detekčních nástrojů

Detekční nástroje pro odhalování plagiátorství textu lze rozdělit podle způsobu, jakým se snaží plagiáty detekovat, na:

1. Nástroje pracující s obsahem dokumentu
   - Intrakorpální
   - Extrakorpální
   - Intrinsické
2. Alternativní přístupy
3. Smíšené

Jako smíšené lze označovat detektory, které kombinují dvě nebo více výše uvedených metod.

4.1.1. Intrakorpální detekce

Intrakorpální nástroj porovnává pouze dokumenty v rámci korpusu (databáze dokumentů). Je možné označit jeden dokument jako podezřelý a porovnat ho vůči korpusu, ve kterém jsou dokumenty považovány za možné zdroje. Nebo může nástroj porovnávat všechny dokumenty v korpusu mezi sebou.
Tento způsob detekce lze snadněji implementovat a není-li korpus příliš rozsáhlý, dosahuje rychlých výsledků. Výraznou nevýhodou je ovšem nemožnost detekovat plagiátorství z jiných dokumentů než těch, které se nachází v korpusu. Tento problém se zmírňuje s rostoucím korpusem, ale je třeba brát ohled na skutečnost, že pokud se v korpusu nachází plagiovaný dokument, který je tak považován za originál, může dojít k tomu, že nástroj později označí skutečný zdrojový dokument za plagiát (Přibil, 2010, s. 46).

4.1.2. Extrakorpální detekce

Extrakorpální nástroj se liší od intrakorpálního v tom, že nemá vlastní korpus. Místo toho využívá databáze třetích stran, ať už jsou to elektronické sbírky dokumentů nebo

internetové vyhledávače.
Účinnost a efektivnost takovéto detekce se pak odvíjí od možností poskytovatele dat. Velké množství dokumentů poskytuje dobré šance pro odhalení možného plagiátu, ale také vytváří problém s časovou náročností detekce. Při praktickém použití není možné porovnávat podezřelý dokument s každým dokumentem v databázi, ale je nutné vybrat jen určitou část dokumentů, u kterých má smysl je porovnávat. Účinnost detekce pak závisí na způsobu, jakým tyto dokumenty vybíráme (Přibil, 2010, s. 47).

4.1.3. Intrinsické metody

Intrinsická metoda detekce plagiátorství neporovnává dokument s databází, ale analyzuje text pouze podezřelého dokumentu. Vychází se z předpokladu, že části textu pocházející od jiného autora budou vykazovat různé hodnoty některých lingvistických charakteristik. Mezi tyto charakteristiky patří množství použitých slov, slabik nebo vět; rozsáhlost slovní zásoby, gramatika; použití slov vyznačující nejistotu, expresivnost a použitá kategorie osoby (Afroz, 2011, s. 5).

Protože tato metoda není závislá na existenci korpusu, je možné detekovat plagiátorství i u dokumentů, které není s čím porovnávat. Navíc, samotná analýza je rychlá, protože stačí zpracovat text pouze jednoho dokumentu.
Na druhou stranu, pokud je nějaká část textu touto metodou označena za plagiovanou, není to dostatečným důkazem pro dokázání plagiátorství, protože na rozdíl od předchozích metod nemáme původní zdroj textu. Je zde také riziko chyby v případě, kdy autor psal část textu ve spěchu a tato část je poté detektorem označena, jelikož ji nelze odlišit od případu skutečného plagiátorství. Dalším problémem je pak případ, kdy je dokument plagiovaný celý pomocí různých textů od jednoho autora. V tomto případě pak intrinsická metoda selhává (Přibil, 2010, s. 48).

Přesto je tato metoda díky své rychlosti vhodná ke kombinaci s jinými metodami, kde může dále zúžit množství textu, které je nutné porovnat. V praxi se také často používá pro zjištění autorství různých textů.

4.1.4. Alternativní přístupy

Mimo hledání shod ve vlastním textu dokumentů existují i jiné přístupy k detekci plagiátu, které se často snaží ovlivnit způsob, jakým dokumenty vznikají, aby získali informaci, podle které mohou poté spolehlivě odhalit plagiátorství a určit jeho zdroj.

Neviditelné značkování

Tato metoda každému originálnímu dokumentu přiřadí unikátní identifikační kód nebo značku. Porovnáním identifikačních značek lze pak v případě plagiovaného dokumentu snadno zjistit jeho původní zdroj. Pro dosažení preventivního účinku může být označení jasně viditelné, nebo může zůstat skryté.
Tento způsob prevence proti plagiátorství poskytuje jasné důkazy, je nezávislý na obsahu dokumentu a je velmi rychlý, má však určité nevýhody. Jakmile je odhalen způsob, jakým tento systém pracuje, je velmi jednoduché ho obejít tím, že se plagiátor vyhne kopírování značky. Navíc není možné zaručit, že dokument, kterému je přidán identifikační kód, je skutečně originální a neobsahuje plagiovaný text (Hauzírek 2007, s. 24 – 26).

Editor neumožňující plagiátorství

Princip spočívá v omezení autorů na použití vybraného nástroje, který obsahuje ochrany zamezující plagiátorství. Tyto ochrany mohou být například ve formě ukládání historie práce s dokumentem, omezení kopírování textu do editoru a kontroly správnosti editorem vložených metadat. Takovéto ochrany samozřejmě nikdy nemohou zcela vyloučit možnost plagiátorství, ale mohou alespoň tento proces ztížit dostatečně na to, aby bylo plagiátorství stejně obtížné, ne-li obtížnější než samostatná práce (Hauzírek 2007, s. 27).

Přestože je tenhle přístup účinný, na první pohled je vidět, že tento přístup je pro uživatele až příliš restriktivní. Praktické použití by takový systém mohl najít snad pouze ve školním prostředí.

Doplňování textu

Nápadem vcelku prostý systém, který vyjme z dokumentu části textu a autor dokumentu má poté text doplnit. Předpokládá se, že autor zná svůj styl psaní a bude schopný správně doplnit části, které psal sám a naopak bude dělat chyby v částech textu, které pochází od jiných autorů. Poté se vyhodnocuje správnost vyplnění chybějících slov, čas, který autor stráví doplňováním, a jiné měřitelné faktory. Na tomto principu pracuje například komerční program Glatt Plagiarism Screening Program (Hauzírek 2007, s. 28).

Úskalím této metody je potřeba přítomnosti autora textu v kontrolovaném prostředí a také možnost zkreslení výsledků detekce momentálním stavem autora, kde stres a únava mohou způsobit větší chybovost.

Následující text se zabývá výhradně extrakorpální detekcí.

4.2. Předzpracování porovnávaných dokumentů

Samotné hledání shod v textu vždy spadá na prosté přímé porovnání stejných slov v blízké vzdálenosti od sebe. Skloňování, časování a další operace se slovem však mohou snadno způsobit, že dvě významově shodná slova rozezná detektor jako dvě různá originální slova. Z toho důvodu je důležité zpracovat text před samotnou detekcí tak, aby byly tyto problémy co nejlépe odstraněny.

Míra zpracování textu se liší podle metody, která bude text vyhodnocovat. V případě porovnávání dvou textů je žádoucí zjednodušit slova na jejich základní formu a zbavit se nadbytečných informací jako je interpunkce a formátování.

Intrinsická metoda naopak pracuje s textem v původní podobě, ve které je rozeznatelný styl psaní autora, který se při zjednodušování textu ztrácí. V případě vyhledávání dokumentů pro porovnání pomocí vhodných klíčových slov je pro určení klíčových slov užitečný dokument zpracovaný, ale samotná slova se použijí v původním tvaru.

Zpracování dokumentu stojí určitý čas. Je proto nutné zvážit při vytváření korpusu v jaké formě se do něj budou dokumenty ukládat. Zpracovaný dokument je oproti originálu méně náročný na paměť a není třeba ho znovu zpracovávat. Není však již příliš vhodný pro zobrazení na výstup uživateli aplikace.

Důležitou úlohu může hrát i způsob reprezentace zpracovaného dokumentu. Text lze ukládat v podobě znaků, ale i v podobě číselných kódů. Dokument v číselné reprezentaci bude strojově zpracován rychleji a dokument, který prošel níže uvedenými stupni zpracování, vyžaduje méně paměti. Na druhou stranu takto zpracovaný dokument již nelze zobrazit v čitelné textové podobě.

4.2.1. Odstranění nevýznamných částí dokumentu

V různých typech dokumentů, hlavně u vysokoškolských prací, se objevují strany nebo části textu, které jsou v těchto typech dokumentů vždy stejné, nebo obsahují nevýznamné informace. Reálná možnost, že je taková část dokumentu plagiovaná, je mizivá, a proto je zbytečné, aby tyto části dokumentů procházeli zpracováním v detektoru. Jedná se o:

- Standardní části dokumentu (titulní strana, zadání práce apod.)
- Části bez významných informací (obsah, seznam obrázků apod.)
- Jasně rozeznatelný citovaný text

Odstraněním těchto nadbytečných stran před dalším zpracováním se celý proces znatelně urychlí bez dopadu na kvalitu detekce. Nedbale provedené odstranění však může před detekcí skrýt část plagiovaného textu. Nejvíce to hrozí při odstraňování citací, kde je potřeba správně rozeznat jaký způsob citací je v dokumentu použit a záleží i na tom, jak důsledně je tento styl citace dodržován.

4.2.2. Převod na čistý text

Dokumenty jsou běžně dostupné v elektronické podobě v různých formátech a obsahují mimo textu i grafy a obrázky. Základním zpracováním dokumentu je tedy jeho převod na čistý text, kdy dochází ke ztrátě všech informací mimo textu samotného a jeho struktury.

Převod na čistý text je různě náročným procesem v závislosti na formátu, ve kterém je původní dokument uložen. V případě textových dokumentů se nejčastěji setkáváme s formáty PDF, dokumentu Word (doc, docx) a v případě porovnávání dokumentu vůči zdrojům na internetu i formát htm.

V případě odborných prací je v současnosti v drtivé většině používán formát PDF. Naneštěstí je z uvedených formátů nejobtížnější na získání čistého textu, ačkoliv jsou problémy extrakce textu omezeny téměř vždy na specifické české znaky. Existuje řada nástrojů, které pracují odlišnými způsoby a lze narazit na problémy typu rozdělení písmene s diakritikou na dva znaky, detekce mezer před českými znaky a jiné.

V případě dokumentu, který byl převeden do elektronické podoby jeho naskenováním, je extrakce textu ještě obtížnější.

4.2.3. Tokenizace

Procesem tokenizace se rozumí rozdělení textu na jednotlivá slova. Dojde tedy k odstranění formátování, interpunkce a převodu všech velkých písmen na malá (Přibil, 2010, s. 69).

S výjimkou intrinsických metod je toto zpracování textu pro detekci plagiátorství výhodné. Při porovnání částí textu nás zajímá hlavně shoda obsahu, který spočívá ve významu slov a interpunkce by tedy proces porovnání pouze komplikovala.
Nerozlišování velkých a malých písmen také zjednodušuje porovnávání, ale mohlo by způsobit potíže v případě zkratek názvů, které by mohly být zaměněny za slova a způsobovat shodu tam, kde shoda ve skutečnosti není.

4.2.4. Odstranění stopslov

Významným krokem ke zrychlení a zjednodušení porovnávání je odstranění stopslov z prohledávaných textů. Jako stopslova označujeme slova, která nenesou pro potřeby detekce žádnou užitečnou informaci. Jsou to často slova krátká a v jazyce velmi často používaná, také je sem možné zahrnout některá přídavná jména, kterými se plagiátor může snažit plagiovaný text vyplnit a tím ho odlišit.
Mezi stopslova lze tedy zařadit spojky, předložky, sloveso být a jeho obdoby a některá další slova podle potřeb detekce. Odstraněním těchto slov dosáhneme informačního zahuštění textu, čímž se sníží množství výrazů, které je třeba porovnávat a v případě ukládání

dokumentů ve zpracované formě se i sníží množství potřebného místa v paměti (Přibil, 2010, s. 69 – 70, 86).

4.2.5. Lematizace

Lematizace spočívá v převedení slov do jejich základního tvaru. Tento krok zpracování velmi ovlivňuje kvalitu detekce.
Český jazyk umožňuje jedno slovo vyjádřit v různých pádech a časech a tím se mění tvar slova. Význam slova zůstává prakticky stejný, ale z pohledu algoritmu se často jedná o zcela odlišná slova. Převedením slov v textu na základní tvar se těchto variací v tvaru slov zbavíme a detektor poté pracuje s textem, kde každé odlišné slovo má i odlišný význam.
Lze uvážit i možnost nahrazení slov za jejich antonyma. Je možné plagiovat zdrojový text negací původního významu a tak dosáhnout stejného významu s použitím odlišných slov (Přibil, 2010, s. 71, 86).

4.2.6. Vážení

Vážení se snaží dosáhnout podobného efektu jako odstraňování stopslov, ale místo odstranění nepotřebných slov přiděluje příliš obecným slovům nižší váhu, která je tak učiní pro detekci méně významnými (Přibil, 2010, s. 71, 87).

Jedná se o náročný proces, který lze v případě porovnávání textu nahradit rozšířením seznamu stopslov. Měl by však velký význam v případě, kdy by měl detektor sám rozhodovat o tom, jestli je dokument plagiát nebo ne.

4.3. Výběr potenciálních zdrojových dokumentů

V případě extrakorpální detekce je nemyslitelné porovnávat podezřelý dokument s každým dokumentem v korpusu. Je proto nutné vybrat ty dokumenty, u kterých lze předpokládat, že mohly být použity jako zdroj textu pro podezřelý dokument. Úspěšnost detekce pak závisí na citlivosti výběru dokumentů. V případě, že je příliš slabá, mohou uniknout detekci plagiované části, které pochází ze zdrojů, které se tématu dokumentu dotýkají pouze okrajově. A naopak příliš citlivý výběr nadbytečně zvyšuje množství dokumentů, které je nutno zpracovat, a tím také celkový čas detekce.

K výběru zdrojového dokumentu lze přistoupit dvěma způsoby:

- Výběr na základě podobnosti
- Výběr na základě obsahu neobvyklých slov

Seznam zdrojových dokumentů je možné dále zkrátit využitím seznamu použité literatury podezřelého dokumentu. Přestože to neplatí vždy, je možné předpokládat, že zdroje plagiátorství nebude plagiátor v seznamu uvádět a proto lze tyto literární zdroje z procesu detekce vyloučit.

4.3.1. Výběr na základě podobnosti

Kosinová podobnost

Zdroje pro porovnání se vybírají podle míry podobnosti s podezřelým textem. Jedním z možných způsobů, jak tuto podobnost určit je kosinová podobnost.

Kosinová podobnost je míra podobnosti dvou vektorů, která se získá výpočtem kosinu úhlu těchto vektorů. V případě podobnosti dokumentů reprezentují vektory četnost jednotlivých slov (nebo jiných úseků textu). Kosinová podobnost je definována jako:

$$\mathrm{cosim}(A,B) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} a_i \times b_i}{\sqrt{\sum_{i=1}^{n}(a_i)^2 \times \sum_{i=1}^{n}(b_i)^2}} \quad (1)$$

Kde A a B jsou vektory reprezentující dokumenty, tvořené počty jednotlivých unikátních slov v obou dokumentech. Kosinová podobnost nabývá hodnot od 0 do 1, kde 0 značí dva naprosto odlišné dokumenty a 1 značí, že jde o dvě zcela totožné kopie.

Před vytvořením vektorů A, B je potřeba nejprve vytvořit pomocný vektor, který obsahuje unikátní slova z obou dokumentů. Vektory A a B jsou pak tvořeny počty, kolikrát jsou daná slova v dokumentu obsažena.

Příklad:

Věta 1: „Výběr na základě podobnosti.“
Věta 2: „Výběr na základě obsahu neobvyklých slov.“

Pomocný vektor C = [výběr, na, základě, podobnosti, obsahu, neobvyklých, slov]
Vektor A = [1, 1, 1, 1, 0, 0, 0]
Vektor B = [1, 1, 1, 0, 1, 1, 1]

$$\text{cosim}(\text{A}, \text{B}) = \frac{1 \cdot 1 + 1 \cdot 1 + 1 \cdot 1 + 1 \cdot 0 + 0 \cdot 1 + 0 \cdot 1 + 0 \cdot 1}{\sqrt{(1 + 1 + 1 + 1 + 0 + 0 + 0) \cdot (1 + 1 + 1 + 0 + 1 + 1 + 1)}} = 0{,}612$$

V případě vlastního korpusu lze připravit vektory vždy při vkládání nového dokumentu a lze tak snadno vyhledávat vhodné zdroje podle kosinové podobnosti. Je vhodné měřit podobnost na dokumentech zpracovaných pomocí lematizace, kterou se množství unikátních slov značně sníží (Přibil, 2010, s. 68).

Klíčová slova

V případě, že používáme externí korpus poskytnutý třetí stranou, máme často omezený výběr způsobů jak v této databázi vyhledávat dokumenty. Totéž platí i pro vyhledávání na internetu. Vyhledávání je běžně omezeno na několik málo slov, ke kterým se vyhledají dokumenty, které tato slova obsahují nejčastěji. Z toho důvodu je vhodné využít klíčových slov.

Klíčová slova definují obsah dokumentu. Obvyklý postup při vytváření klíčových slov je najít určitý počet nejvíce používaných podstatných jmen v dokumentu. Slova se vyhledávají v dokumentu po zpracování lematizací, aby se zamezilo identifikaci stejného slova v různých tvarech jako několik unikátních slov.

Jiný způsob extrakce klíčových slov je použit v metodě KeyGraph, která shlukuje prvky, které spolu souvisí, do skupin, aby určila slova reprezentativní pro obsah dokumentu. Metoda vytváří graf termínů vyskytujících se v dokumentu, které jsou propojeny, pokud se dva termíny často vyskytují spolu. V grafu se identifikují podgrafy jako shluky maximálně propojených termínů a vyberou se termíny propojující dva podgrafy jako kandidáti na klíčová slova. Klíčová slova se určí podle pravděpodobnosti, že k propojení podgrafů bylo použito právě toto slovo (Lott, 2012, s. 6).

Ať už určíme podobnost dokumentů jakkoliv, jedná se pouze o podobnost, ze které není možné určit, jestli nalezený dokument obsahuje text, který je podezřelým dokumentem plagiován. Dokument, který má velkou podobnost s podezřelým dokumentem, používá mnoho stejných slov, ale může se týkat jiného tématu a naopak dokument, který má malou podobnost, může obsahovat jednu kapitolu, která je věnovaná stejnému tématu, jaké je rozebíráno v podezřelém dokumentu.

4.3.2. Výběr na základě obsahu neobvyklých slov

Jiným způsobem jak nalézt vhodné dokumenty k porovnání, je hledat v nich určité úseky textu, o kterých si myslíme, že mohou být plagiované.

Každý člověk má různě rozsáhlou slovní zásobu a je tedy pravděpodobné, že výrazy, které běžně nepoužívá, převzal z nějakého jiného zdroje. Pokud bychom vyhledávali dokumenty k porovnání na základě těchto slov, jsme schopni získat menší množství dokumentů než v případě hledání podle podobnosti a úspěšnost nezávisí na tom, jak moc se daný dokument tématu týká. Úspěšnost celého hledání však stojí na vhodném výběru výrazů k vyhledání.

Vhodnými výrazy jsou slova cizí nebo odborná, dále také zkratky nebo různá data. K nalezení vhodných slov je však nutné porozumění významu textu, což vyžaduje implementaci strojového učení.

V nejjednodušším případě lze slepě vybrat velké množství slov z dokumentu a po skupinách k nim hledat dokumenty, které je obsahují. Tímto způsobem se lze prakticky zbavit možnosti přehlédnutí plagiovaných výrazů při výběru za cenu vyšší časové náročnosti.

Zajímavé by mohlo být v některých případech i vyhledávání vzorců. Problémem je, že vzhled vzorců je závislý na formátu písma a extrakce textu ze vzorce může dávat v různých dokumentech odlišné výsledky.

4.4. Měření shodnosti v textu

K měření shodnosti se používá více algoritmů, které jsou použitelné na plný text, nebo vyžadují text zpracovaný do určité podoby. Dále se pak liší svojí náročností na čas, a jak odolné jsou proti snahám plagiátora skrýt svůj čin.

4.4.1. Vyhledání znakových řetězců

Algoritmy vyhledávající a porovnávající znakové řetězce jsou schopné odhalit přímé plagiátorství. Pracují s plným textem a jejich účinnost lze zvýšit vhodným předzpracováním textu, ale stále nejsou schopné poradit si s odlišným pořadím výrazů. Jejich výhodou je však rychlost s jakou přímé plagiátorství odhalují.
Typicky používanými algoritmy je naivní vyhledávání, algoritmus Rabin-Karp, algoritmus Knuth-Morris-Pratt, algoritmus Boyer-Moore a vyhledávání pomocí konečného automatu. Existují i algoritmy pro přibližné vyhledávání, které nalézají i řetězce podobné hledanému řetězci. Tyto algoritmy však nejsou vhodné, kvůli slovům, které vypadají podobně, ale mají různé významy. Časová náročnost jednotlivých algoritmů je uvedena v tabulce (Přibil, 2010, s. 102 - 107):

Tabulka 3: Časová náročnost algoritmů pro vyhledávání řetězců

Algoritmus	**Čas předzpracování**	**Čas vyhledávání**
Naivní vyhledávání	0	Θ ((n-m) m)
Rabin-Karp	Θ (m)	Θ (n+m) průměr Θ ((n-m)m) nejhůř
Knuth-Morris-Pratt	Θ(m)	Θ(n)
Boyer-Moore	Θ(m + |Σ|)	Ω(n/m), O(nm)
Konečný automat	Θ(m |Σ|)	Θ(n)

Kde **m** je délka řetězce a **n** je délka textu.

4.4.2. Algoritmy měřící vzdálenost řetězců

Oproti algoritmům vyhledávání znakových řetězců jsou tyto algoritmy již schopny nalézt shody i v řetězcích, které mají změněné pořadí slov nebo jim některá slova chybí. Mají však i své slabé stránky.

Jsou výpočetně náročnější a zaberou tak více času. Potřebují také vhodně určit hranici povolené vzdálenosti mezi řetězci, jinak může dojít k chybnému označení shody tam, kde na první pohled žádná shoda není.
Algoritmů existuje větší množství a každý má uplatnění jiné. Mezi algoritmy použitelné na

text patří hojně používaná Levenshteinova vzdálenost, ze které jsou odvozené další metriky (Přibil, 2010, s. 108).

Levenshteinova vzdálenost

Měří rozdíl mezi dvěma řetězci znaků, který je roven množství operací záměny znaku, odstranění znaku a přidání znaku, které jsou potřeba k přeměně jednoho řetězce na druhý.

Příklad:

„Algoritmus“ „Alkoholismus“

Alkoritmus → Alkolitmus → Alkolismus → Alkohlismus → Alkoholismus

Levenshteinova vzdálenost je 5 (3 záměny, 2 přidání).

Podobnými algoritmy je Damerau-Levenshteinova vzdálenost, která umožňuje mimo záměny, přidání a odstranění i výměnu dvou sousedících znaků.
Dále pak hledání nejdelšího společného podřetězce, které povoluje pouze vložení a odstranění, a Hammingova vzdálenost, která pracuje s řetězci o stejné délce a umožňuje pouze záměnu znaků.
Je možné také jednotlivým operacím přiřadit určitou váhu. Tím se ovšem výpočet dále prodlužuje. V našem případě zpracování textů však není základní jednotkou jeden znak, ale místo toho se použijí slova. Jinak je proces stejný (Přibil, 2010, s. 109 - 110).

4.4.3. Měření shodnosti dokumentů reprezentovaných n-gramy

Jako n-gram je označována řada po sobě jdoucích položek posloupnosti. V případě textu se obvykle jedná o posloupnost slov, kde **n** označuje počet slov v řadě a platí, že **n << m**, kde **m** je celkový počet slov v textu.

Na rozdíl od předchozích metod, vyžaduje měření shodnosti pomocí n-gramů zpracovat porovnávané texty do podoby sledu jednotlivých n-gramů. Příklad reprezentace věty pomocí 3-gramů:

„Jeho praktické využití však muselo počkat až do padesátých let.“

{ Jeho, praktické, využití }, { praktické, využití, však }, { využití, však, muselo }, { však, muselo, počkat }, { muselo, počkat, až }, { počkat, až, do }, { až, do, padesátých }, { do, padesátých, let }

Dokument reprezentován pomocí n-gramů tedy obsahuje **m-n+1** n-gramů, kde každý n-gram se skládá z **n** slov. Text dokumentu se tak podstatně zvětší a redukce počtu slov pomocí předzpracování textu zde má větší váhu.
Hlavní výhodou tohoto způsobu je, že porovnávání n-gramů probíhá stejným způsobem jako v případě vyhledávání znakových řetězců. V jednom z dokumentů se hledají n-gramy, které se shodují alespoň s jedním n-gramem v druhém dokumentu. Samotný proces je tedy v porovnání s komplexnějšími algoritmy, jako je měření vzdálenosti řetězců, rychlý. Navíc,

v případě, že ignorujeme pořadí jednotlivých prvků n-gramu, jsme schopni nalézt shody i u n-gramů, která obsahují stejná slova, ale v jiném pořadí (Přibil, 2010, s. 84 - 85).

Stejně jako v případě měření vzdálenosti řetězců je nutné určit vhodnou velikost parametru vyhledávání, který je v tomto případě velikost n-gramů. Malé hodnoty n dávají mnoho falešných shod a naopak příliš velké hodnoty n zhoršují schopnost odhalit krátké úseky plagiovaného textu. Na základě měření provedených na Polytechnické univerzitě ve Valencii, se jako optimální hodnoty velikosti n-gramu jeví hodnoty nízké, konkrétně n = { 2, 3} (Barrón-Cedeño a Rosso, 2009).

4.5. Metriky měření shodnosti

V případě detekce plagiátorství je výstupem množství zpracovaných dokumentů s nalezenými a označenými shodami. Ne každý z těchto dokumentů však obsahuje dostatek shod na to, aby mělo smysl ho zobrazovat uživateli. Proto je potřeba použít vhodnou metriku, kterou zhodnotíme míru shody s podezřelým dokumentem.

Metrik použitelných pro měření shodnosti existuje samozřejmě více. V našem případě je potřeba, aby daná metrika nebyla závislá na poměru velikostí porovnávaných textů. Zdroje plagiátorství zahrnují mimo rozsáhlých prací i krátké texty z internetových stránek a mohlo by tak snadno dojít ke zkreslení zhodnocení výsledků detekce.
Pro tyto účely je použitelná Kosinová podobnost, popsaná v kapitole 4.3.1. Zde uvedeme dvě další použitelné párové metriky:

- Asymetrická
- Symetrizovaná asymetrická

4.5.1. Asymetrická metrika

Asymetrická metrika určuje, jakou měrou je jeden dokument obsažen v druhém. Dává tedy různé hodnoty pro případ, kdy je porovnáván dokument A proti dokumentu B a naopak.

$$\mathrm{con}(\mathrm{A},\mathrm{B}) = \frac{|\mathrm{V}(\mathrm{A}) \cap \mathrm{V}(\mathrm{B})|}{|\mathrm{V}(\mathrm{A})|} \tag{2}$$

$$\mathrm{con}(\mathrm{B},\mathrm{A}) = \frac{|\mathrm{V}(\mathrm{B}) \cap \mathrm{V}(\mathrm{A})|}{|\mathrm{V}(\mathrm{B})|} \tag{3}$$

$|\mathrm{V}(\mathrm{A}) \cap \mathrm{V}(\mathrm{B})|$ představuje počet shodných slov (nebo jiných prvků) mezi dvěma dokumenty. Tento počet se dělí celkovým počtem slov jednoho z dokumentů. Výsledkem je číslo v rozsahu < 0; 1 >. V našem případě nás zajímá hodnota shody v poměru k podezřelému dokumentu. Přesáhne-li výsledná hodnota námi určenou hranici, je dokument označen jako zdroj plagiátorství a bude zobrazen uživateli (Hauzírek, 2007, s. 36).

4.5.2. Symetrizovaná asymetrická metrika

Symetrizovaná asymetrická metrika je výsledkem snahy o odstranění nevýhody asymetrické metriky, kterou jsou dva různé výsledky. Tím jednodušším způsobem je prostý průměr obou výsledků asymetrické metriky, který ale částečně vnáší zpět problém s citlivostí na velikosti dokumentů.
Lepším přístupem je použít maximum z obou hodnot. Tento přístup sice také neodstraní zcela

citlivost na délku dokumentu, ale pravděpodobnost vyskytnutí případu, kdy by délka dokumentu měla na výsledek negativní vliv, je velmi malá. Takový případ by vyžadoval dokument, který by byl tak krátký, že by ono zanedbatelné množství shodných slov tvořilo výraznou část tohoto dokumentu (Hauzírek, 2007, s. 38).

$$\mathrm{maxcon}(A, B) = \max\left(\frac{|V(A) \cap V(B)|}{|V(A)|}, \frac{|V(B) \cap V(A)|}{|V(B)|}\right) \quad (4)$$

4.5.3. Volba vhodné hranice

Volbě hranice shodnosti je třeba věnovat pozornost, protože rozděluje dokumenty na ty, kterými má smysl se zabývat a kterými ne. Při volbě je třeba dbát na citlivost metody, která hledá shody v dokumentu a také na velikost dokumentu.

Citlivá metoda dává více falešných shod, které jsou často izolovány od ostatních shod a tak je na první pohled jasné, že se jedná jen o náhodu. Tento problém lze kompenzovat vyšší hranicí.

Velikost dokumentu také hraje určitou roli. V případě hranice 0,01 a podezřelého dokumentu o 500 slovech, představuje zdroj každý dokument s více než 5 shodnými slovy. Zde se velmi snadno může jednat o náhodu. V případě podezřelého dokumentu o 20 000 slovech se však už jedná o 200 slov, které mohou představovat izolované případy nebo i celý plagiovaný odstavec.

Přesnost hranice lze také dodatečně zlepšit tak, že algoritmus ignoruje shody, které nemají ve své blízkosti další shodná slova a je tedy pravděpodobné, že se jedná o falešná pozitiva.

4.6. Snahy o obcházení automatické detekce plagiátorství

V současné době má již automatická detekce plagiátorství určitou historii a existuje množství nástrojů, které se této problematice věnují. Nikoho tedy nepřekvapí, že plagiátoři začali hledat způsoby, jak tyto nástroje obejít. Zde je uvedeno několik těchto metod.

Změna znakové sady

Snahou je zabránit správné extrakci textu ze souboru bez ovlivnění vzhledu tištěného obsahu. Lze měnit dokument jako celek, nebo nahradit určité znaky vzhledově podobnými znaky z jiných jazyků, které jsou reprezentovány jiným kódem. Změnou znakové sady tak lze dosáhnout toho, že extrahovaný text pro potřeby detekce obsahuje změť nesmyslných znaků, které nebudou mít shodu v žádném z dostupných zdrojů (Beall, 2013).

S touto metodou lze bojovat rozlišením, jaká sada je použita, nebo detekcí smysluplnosti textu za pomocí slovníku. V nejhorším případě to zjistí uživatel detektoru ve fázi, kdy detektor zobrazuje výsledky. Je však lepší kontrolovat soubor již na začátku, aby nedošlo ke zbytečné ztrátě času.

Nahrazení bílých znaků

Snaha zmást detektor nahrazením bílých znaků, například mezer, jinými znaky. Nahrazené znaky jsou pak vizuálně změněny tak, aby v tištěné podobě nebyly rozeznatelné (například změna barvy na barvu pozadí nebo změna velikosti) (Beall, 2013).

Tento problém lze řešit několika způsoby:

- Detekce smysluplnosti slov za pomocí slovníku.
- Kontrola délky slov. Nahrazení bílých znaků vytvoří nadměrně dlouhá slova, která by měla být snadno detekovatelná.
- Detekce počtu bílých znaků v dokumentu nebo počtu slov. Je použitelná pouze v případě, že je nahrazena větší část mezer.
- Nahrazení zástupných znaků mezerami. V případě, že by algoritmus dokázal rozeznat formátování těchto náhradních znaků, lze je přímo nahradit mezerami a zpracovat dokument normálně nebo upozornit uživatele detektoru.

Nahrazení textu grafikou

Formáty dokumentů zobrazující text i grafické objekty mohou text obsahovat pouze ve formě obrázků, detekční systém pak nemá text, který by mohl zpracovat (Beall, 2013).

V případě hojného použití v dokumentu lze tento způsob odhalit zjištěním délky dokumentu. V případě krátkých úseků textu převedených do grafiky je nutné zobrazit skutečný text v dokumentu pro porovnání s originálním dokumentem uživatelem. Jelikož text

v grafické formě musí být na pohled k nerozeznání od skutečného textu, bylo by možné využít nástroje pro detekci textu z grafiky a upozornit uživatele detektoru.

Použití zdroje v jiném jazyce

Plagiátor použije zdroj v jiném jazyce a vloží ho přeložený do svého jazyka do dokumentu (Beall, 2013).

Tuto metodu lze detekovat v případě strojového překladu, kde mohou být dostatečně velké nesrovnalosti i pro automatickou detekci. Pokud byl však proveden překlad ručně, bylo by nutné použít vícejazyčný detektor. Vzhledem k současnému stavu strojového překladu je navíc pravděpodobnost detekce tohoto plagiátorství znatelně menší.

Nechat si práci napsat někým jiným

Originální dokument napsaný jedním autorem, vydávaný jiným člověk jako jeho vlastní, nelze prakticky detekovat. V případě, že se původní autor nedopustí žádného plagiátorství, neexistuje zdroj plagiátu a protože jde o jednoho autora, nepomohou s odhalením dokumentu ani intrinsické metody.

V tomto případě je jedinou možností detekce otestování znalostí člověka, který se za autora vydává.

5. VLASTNÍ NÁVRH DETEKTORU PLAGIÁTŮ

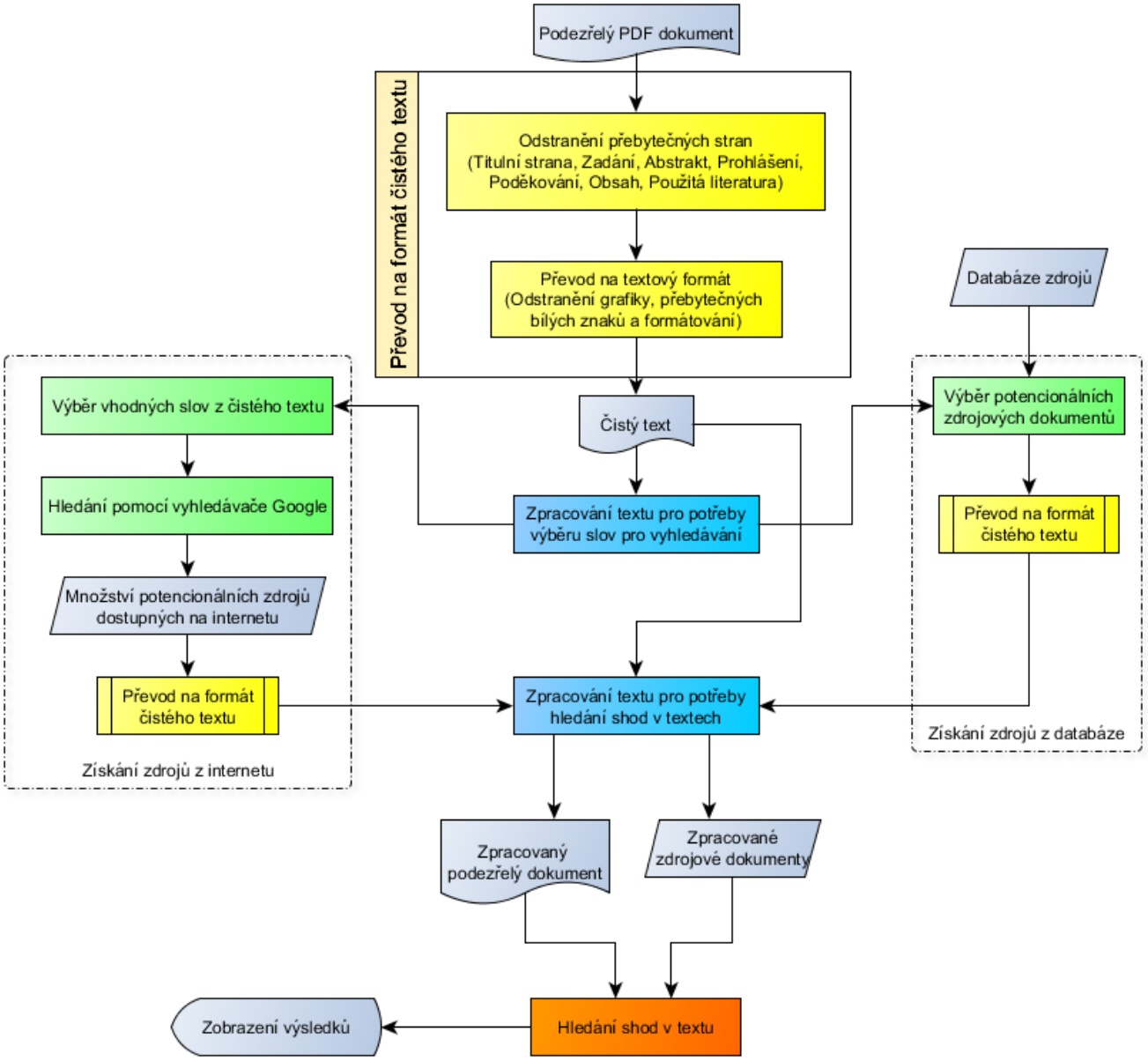
Proces detekce plagiátu od přijmutí podezřelého dokumentu až po zobrazení výsledků, lze rozdělit na několik částí:

- Převod dokumentu na formát čistého textu
- Zpracování textu pro vyhledávání zdrojů
- Vyhledávání zdrojů v databázi
- Vyhledávání zdrojů na internetu
- Zpracování textu pro hledání shod
- Hledání shod v textu
- Zobrazení výsledků

Celý proces je shrnut na následujícím funkčním schématu.

Realizace vlastního detektoru plagiátorství je provedena v jazyce C# v prostředí Visual Studio 2012. Jedná se o Windows Forms aplikaci.

Vytvořený software je přílohou práce ve formě projektu Visual Studio 2012. Součástí přílohy je i návod na použití a umělý plagiát ve formátu PDF. Software může být kýmkoliv volně použit na kontrolu libovolné veřejně přístupné práce.



Obrázek 1: Funkční schéma detektoru plagiátorství

5.1. Převod na formát čistého textu

5.1.1. Převod textu

Cílem je převést vstupní dokument z původního formátu na formát čistého textu, bez zvláštního formátování textu nebo objektů (např. obrázky), který se použije v dalším zpracování. Tento proces je nutno provést pro každý dokument, se kterým bude detektor pracovat.

Detektor musí rozeznat formát dokumentu a použít na něj odpovídající metodu převodu na čistý text. V případě psaného textu postačí, když budeme počítat s nejpoužívanějšími formáty pro tento typ dokumentů. Tyto formáty jsou PDF a DOC / DOCX, ale je třeba počítat i s formáty TXT a hlavně HTM / HTML, který tvoří značnou část výsledků vyhledávání na internetu.

K převodu dokumentů ve formátu PDF na formát TXT jsou v detektoru použity open source nástroje Xpdf.
Převod mezi MS Word dokumenty a textem je proveden pomocí C# .NET knihovny Code7248.word_reader.dll.
Převod HTM / HTML do textu je proveden odstraněním značek používaných při vytváření webových stránek.

Odkazy na zdroje:

Xpdf:	http://www.foolabs.com/xpdf/home.html
Code7248.word_reader.dll:	http://sourceforge.net/projects/word-reader/?source=dlp

5.1.2. Redukce textu

Při převodu na čistý text lze také text zredukovat odstraněním nepotřebných částí textu. Filtraci textu lze nejlépe provést u odborných prací, kde je často jasně daná struktura práce a prvních několik stran obsahuje pro nás nepodstatné informace jako je obsah nebo seznamy tabulek či obrázků.
Naopak internetové stránky nemají žádnou ucelenou strukturu a filtrovat jejich obsah by příliš často vedlo k ignoraci textu, který bychom testovat chtěli.

Detektor před převodem vstupního PDF dokumentu na text umožňuje ignorovat určitý počet stránek od začátku a konce dokumentu. Počet lze určit manuálně, ale detektor ho dokáže určit i automaticky detekcí standartních částí prací jako je titulní strana, obsah apod. na základě určitých slov.

Text extrahovaný z původního souboru se neukládá do paměti, ale je ukládán do textového souboru ve složce detektoru.

5.2. Zpracování textu pro potřeby výběru slov pro vyhledávání

Ať už hodláme vyhledávat jakoukoliv metodou, prvním krokem při zpracování textu bude odstranění interpunkce a dalších přebytečných znaků z textu.

V našem případě toho dosahujeme prostou extrakcí jednotlivých slov z textu a jejich ukládání do textového souboru ve tvaru jednotlivých slov oddělených pomocí čárek. Všechna písmena jsou převedena na malá písmena.

Z textu se dále odstraňují stop slova, která nenesou žádnou významnou informaci a lze je proto ignorovat. Jejich odstranění je provedeno porovnáním slova se seznamem stop slov.

Podle nastavení detektoru pak může proběhnout lemmatizace slov a extrakce klíčových slov, která se použijí pro vyhledávání.

Pro lemmatizaci slov v textu je použit nástroj open source projektu LemmaGen.
Odkaz: http://lemmatise.ijs.si/

Extrakce klíčových slov je provedena výběrem nejčastěji se vyskytujících slov. Výběr je proveden po lemmatizaci, aby se snížil počet unikátních slov v dokumentu.

5.3. Získání zdrojů z databáze

Zdroje ve vlastní databázi lze vyhledávat několika způsoby, mezi jinými se jedná o kosinovou podobnost, klíčová slova a hledání na základě určitých vybraných slov.

V případě, že vyhledáváme zdroje ve vlastní databázi, nic nebrání využití kosinové podobnosti, protože můžeme dokumenty zpracovat pro snazší vyhledání před vložením do databáze.

Náš detektor je zaměřen na hledání zdrojů na internetu, hledání v databázi je zde uvedeno jen pro úplnost.

5.4. Získání zdrojů z internetu

Logický přístup k vyhledávání zdrojů na internetu je použití stejných prostředků pro snadné získání zdrojů informací, které používá většina studentů tvořících závěrečné práce, a tím jsou internetové vyhledávače, především však Google.

V našem detektoru je vyhledávání řešeno zadáním slov vybraných z podezřelého dokumentu do vyhledávače Google, čímž získáme data ze stránky s výsledky vyhledávání, ze které vybereme http a https odkazy, z nichž vyfiltrujeme stránky související s vyhledávací službou Google a výsledný seznam zdrojů stáhneme.

Detektor umožňuje vyhledávat klíčová slova z dokumentu, vyhledat manuálně zadaný text nebo postupně hledat po skupinách slov od začátku dokumentu až na jeho konec a pro každé hledání stáhnout zdroje.

Vyhledávání jednotlivých skupin slov z celého textu je nejpřesnější, ale také nejpomalejší z možností. Proces lze však značně urychlit nastavením počtu slov na jedno vyhledání, omezení na množství stažených zdrojů na jedno vyhledání a použitím jen každého n-tého slova.

Podrobněji vyhledávací algoritmus popisuje vývojový diagram vyhledávání (Obr. 7 a Obr. 8) na konci této kapitoly.

5.5. Zpracování textu pro potřebu hledání shod v textech

Pro hledání shod v textech detektor používá porovnání pomocí neuspořádaných n-gramů.

Pro porovnání je podezřelý dokument i nalezený zdroj převeden na formát čistého textu, ze kterého jsou odstraněna stop slova a následně je lemmatizován. Z lemmatizovaného textu jsou pak vytvořeny n-gramy, kde n je hodnota zvolená v nastavení programu.
Z důvodu rychlejšího porovnávání jsou slova při vytváření n-gramů nahrazena čísly, pro které je tvořena tabulka přiřazující slova k číslům.

Podrobněji je proces zpracování textu popsán ve vývojovém diagramu hledání shod (Obr. 9) na konci této kapitoly.

5.6. Hledání shod v textu

Samotné hledání shod je provedeno porovnáním vytvořených neuspořádaných n-gramů. Ke každému n-gramu podezřelého dokumentu je ve zdroji hledán n-gram, který obsahuje stejné prvky v libovolném pořadí.

Ze seznamu nalezených shod mezi n-gramy je pak počítána míra podobnosti ve formě kosinové podobnosti.

Protože však může být procentuální míra podobnosti mnohdy zavádějící při výběru zdrojů, kterým bychom měli věnovat bližší pozornost, je z výsledku porovnání počítána i tzv. míra shluků shod.

5.6.1. Míra shluků shod

Jedná se o autorem navržený vzorec, který vyhodnocuje míru rozptýlení shod v textu. Stejný počet shod má vyšší míru shluků shod, pokud jsou mezery mezi shodami malé, než když jsou od sebe vzdáleny více. Vychází se z předpokladu, že shody blízko u sebe naznačují, že se zde nachází celá shodná věta nebo i odstavec, což je daleko pravděpodobnější případ možného plagiátorství než několik náhodně rozmístěných shod, které spolu nesouvisí.

$$\text{Míra shluku shod} = \prod_{i=1}^{m}[n + 1 - (a_{i+1} - a_i)] \qquad (5)$$

$$a \in N, \qquad a > 0, \qquad a_{i+1} - a_i \leq n, \qquad a_i < a_{i+1}$$

Kde a_i jsou pozice shod v textu a n je počet prvků n-gramu.

Hodnota míry shluků shod s každou navazující shodou roste nelineárně. Jeden velký shluk je tedy ohodnocen významněji než několik menších shluků.

Výsledky porovnání jsou na konci ukládány do textových souborů pro pozdější zobrazení. Podrobněji je proces porovnání n-gramů popsán ve vývojovém diagramu porovnání n-gramů (Obr. 10 a Obr. 11) na konci této kapitoly.

5.7. Zobrazení výsledků

Jakmile máme k dispozici výsledky vyhledávání, je třeba je vhodně prezentovat uživateli. Protože uživatel detekčního nástroje má poslední slovo při rozhodnutí o tom, jestli se jedná o plagiát nebo ne, je nutné zobrazit porovnávané dokumenty v čitelné a přehledné formě.

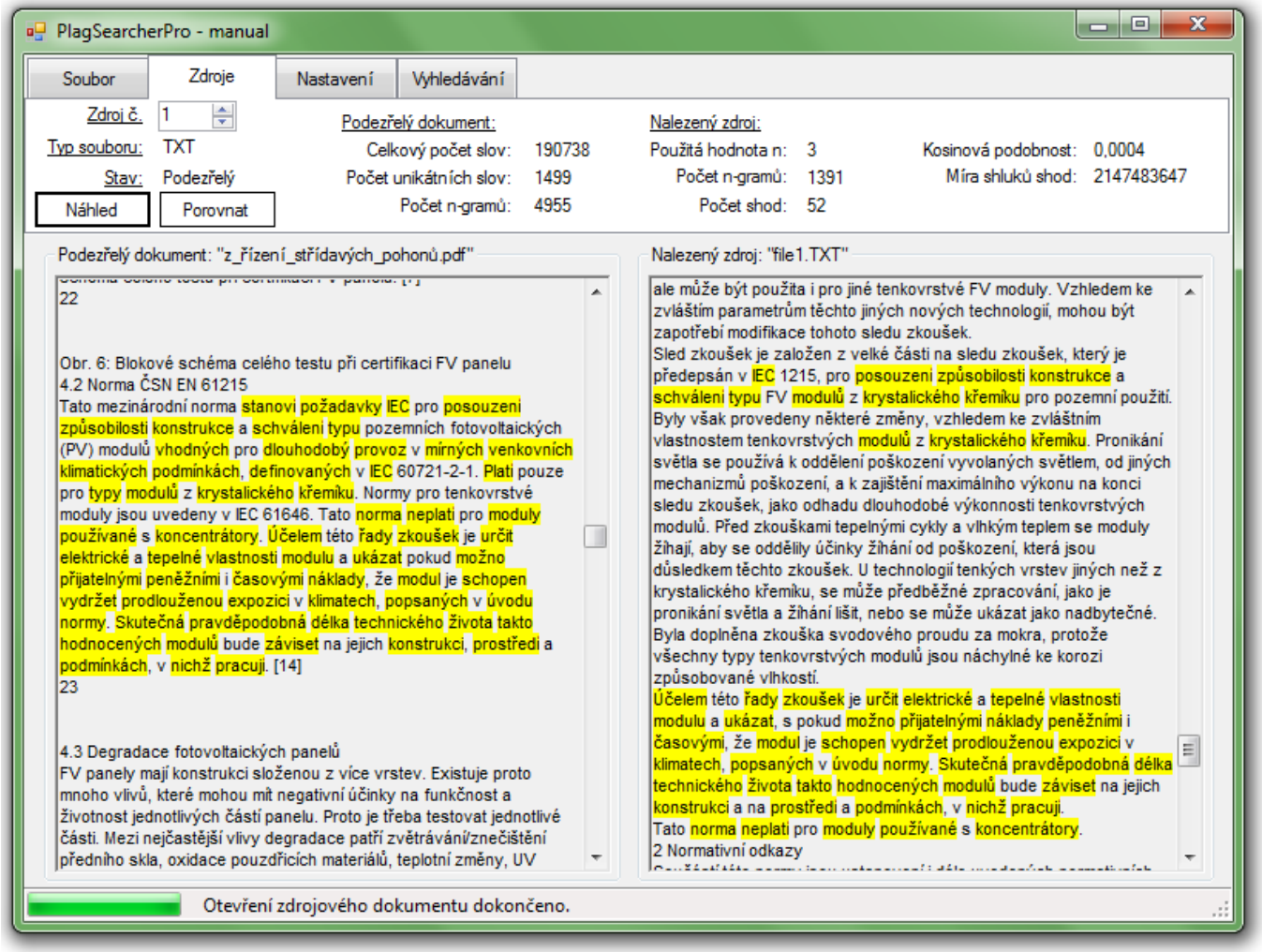
Po ukončení vyhledávání si může uživatel nechat zobrazit libovolný porovnaný zdroj. Detektor vedle sebe zobrazí text vstupního podezřelého dokumentu a vybraného zdroje a v obou zvýrazní nalezené shody. Uživateli pak stačí zaměřit se na vyznačené části textu a posoudit, jestli se skutečně jedná o plagiátorství.

Při procházení zdrojů jsou zobrazovány některé statistiky, mezi nimi také kosinová podobnost se vstupním dokumentem a míra shluků shod. Na základě těchto informací může uživatel věnovat pozornost jen těm zdrojům, které vykazují nezvykle vysoké míry těchto hodnot.

V případě vyhledávání a porovnání více podezřelých dokumentů za sebou ukládá detektor informace o nalezených zdrojích do zvláštního textového souboru a ukládá i část nalezených zdrojů na základě hodnoty jejich kosinové podobnosti a míry shluků shod.

5.8. Ovládání programu

Na následujícím snímku je zobrazeno okno detektoru se zobrazenými výsledky.



Obrázek 2: Snímek okna prototypu detektoru plagiátorství

Okno detektoru je z větší části tvořeno dvěma texty, levý extrahovaný ze vstupního podezřelého dokumentu a pravý z vybraného zdroje. Slouží pro rychlé porovnání nalezených shod, které jsou vyznačeny žlutě.

Spodní lišta okna obsahuje indikátor průběhu a zprávy o stavu aplikace. Zobrazují se zde zprávy o průběhu procesů i zprávy o chybách.

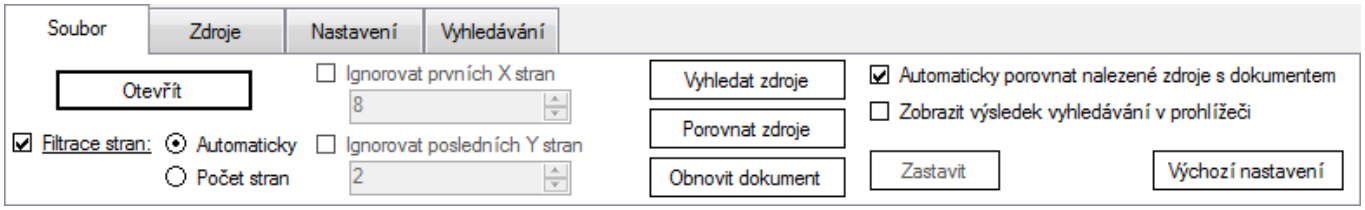
Vrchní část okna je rozdělena do několika záložek a slouží k ovládání aplikace.

5.8.1. Záložka Soubor

Tato záložka slouží k samotnému ovládání detektoru.

Tlačítkem **Otevřít** lze vybrat PDF dokument, pro který bude detektor hledat zdroje. Pokud je povolena **Filtrace stran**, tak při otevírání dokumentu je ignorován určitý počet stran od začátku a konce dokumentu určený **Automaticky** nebo ručně nastavením hodnot **X** a **Y**.

Pokud je vybrán více než jeden vstupní dokument, detektor bude postupně vyhledávat všechny vybrané dokumenty automaticky. Titulek okna se změní z „manual“ na „auto“.

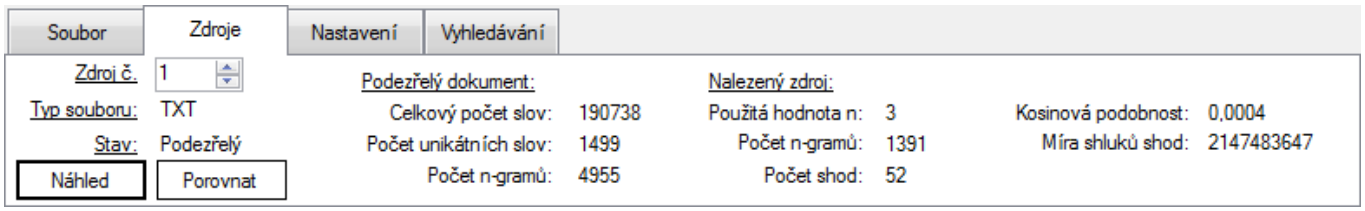


Obrázek 3: Záložka Soubor

Tlačítko **Vyhledat zdroje** spustí vyhledávání zdrojů na internetu přes vyhledávač Google hledáním slov vyskytujících se v dokumentu podle současného nastavení detektoru. Je možné i nastavit, aby detektor nejen vyhledal zdroje ale hned je i porovnal se vstupním dokumentem a popřípadě na konci zobrazil ve vyhledávači stránky s výsledky každého hledání vyhledávačem Google.
Porovnat zdroje slouží ke spuštění porovnání všech již stažených zdrojů se vstupním dokumentem.
Obnovit dokument umožňuje zobrazit výsledky posledního zpracovávaného dokumentu, pokud potřebné soubory stále existují ve složce detektoru.
Tlačítko **Zastavit** ukončí probíhající proces vyhledávání nebo porovnání.

Výchozí nastavení obnoví původní doporučené nastavení.

5.8.2. Záložka Zdroje

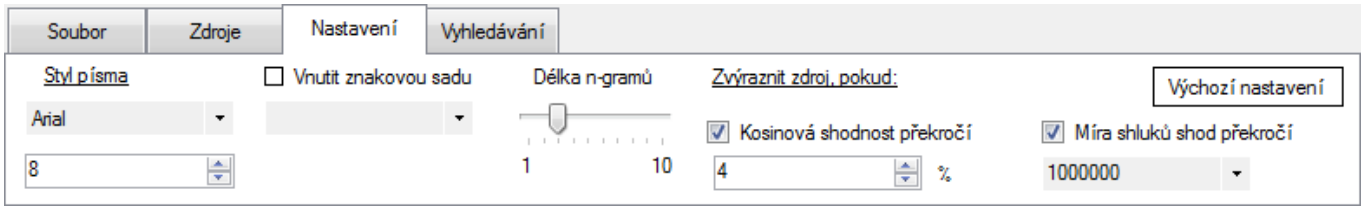


Obrázek 4: Záložka Zdroje

Tato záložka slouží k procházení nalezených zdrojů, rozlišených pomocí pořadového čísla. O vybraném zdroji a vstupním podezřelém dokumentu jsou zobrazeny základní informace včetně typu souboru zdroje a stavu, který vyhodnocuje pravděpodobnost, že byl zdroj použit při plagiátorství, na základě kosinové podobnosti a míry shluků shod podle nastavení detektoru.

Tlačítko **Náhled** zobrazí vybraný zdroj v pravém textovém sloupci a tlačítko **Porovnat** provede nové porovnání zdroje s podezřelým dokumentem podle současného nastavení.

5.8.3. Záložka Nastavení



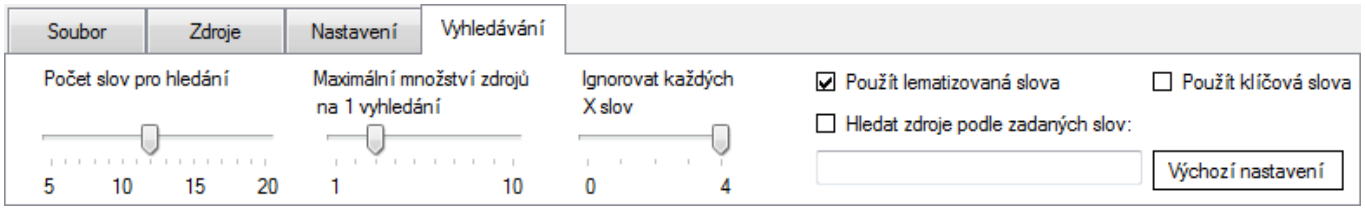
Obrázek 5: Záložka Nastavení

Zde je možné nastavit styl a velikost písma zobrazovaného textu, hodnotu **n** pro porovnávání pomocí n-gramů a hranice kosinové shodnosti a míry shluků shod, podle kterých se posuzuje pravděpodobnost, že byl zdroj použit při plagiátorství.

Nastavení **Vnutit znakovou sadu** umožňuje ručně nastavit znakovou sadu, kterou má detektor použít pro zobrazení textových a HTML souborů, které se ve výchozím nastavení zobrazují nesprávně.

Výchozí nastavení obnoví původní doporučené nastavení.

5.8.4. Záložka Vyhledávání



Obrázek 6: Záložka Vyhledávání

Tato záložka slouží k nastavení parametrů ovlivňujících průběh vyhledávání.

Počet slov pro hledání je množství slov, které budou spojeny do jednoho řetězce, který je poté vyhledáván na internetu pomocí vyhledávače Google. Vyšší množství slov znamená menší počet vyhledávání, ale také menší přesnost.

Maximální množství zdrojů na jedno vyhledání omezuje počet zdrojů, které je možné stáhnout pro jednotlivá hledání. Toto nastavení ovlivňuje celkový čas potřebný na kontrolu dokumentu nejvíce.

Ignorovat každých X slov nastavuje počet slov, která se ignorují mezi dvěma slovy vybranými pro vyhledávání. Vyšší hodnota způsobí řidší výběr slov, čímž mohou být vynechány některé podstatné výrazy.

Použít lematizovaná slova určuje, jestli detektor při vyhledávání používá slova v původní formě nebo v lematizované formě.

Použít klíčová slova dovoluje rychlé vyhledání pouze za pomocí klíčových slov, tedy těch nejčastěji se vyskytujících slov. Pro hledání jsou použita všechna slova s výskytem nad

1 %. **Počet slov pro hledání** určuje počet klíčových slov použitých na jedno vyhledávání. Množství stažených zdrojů není omezeno.

Hledat zdroje podle zadaných slov umožňuje vyhledat zdroje na základě uživatelem vepsaných výrazů.

Výchozí nastavení obnoví původní doporučené nastavení.

V případě přímých forem plagiátorství se často kopírují celé bloky textu bez větších úprav. Pro odhalení této formy postačí použít jen každé páté slovo a stáhnout pouze několik prvních zdrojů z každého vyhledávání, pokud použijeme přibližně asi 10 slov na jedno vyhledávání. Tak dosáhneme i dobré rychlosti vyhledávání a porovnání aniž by byla významně ovlivněna přesnost. Výchozí nastavení odráží tyto případy.

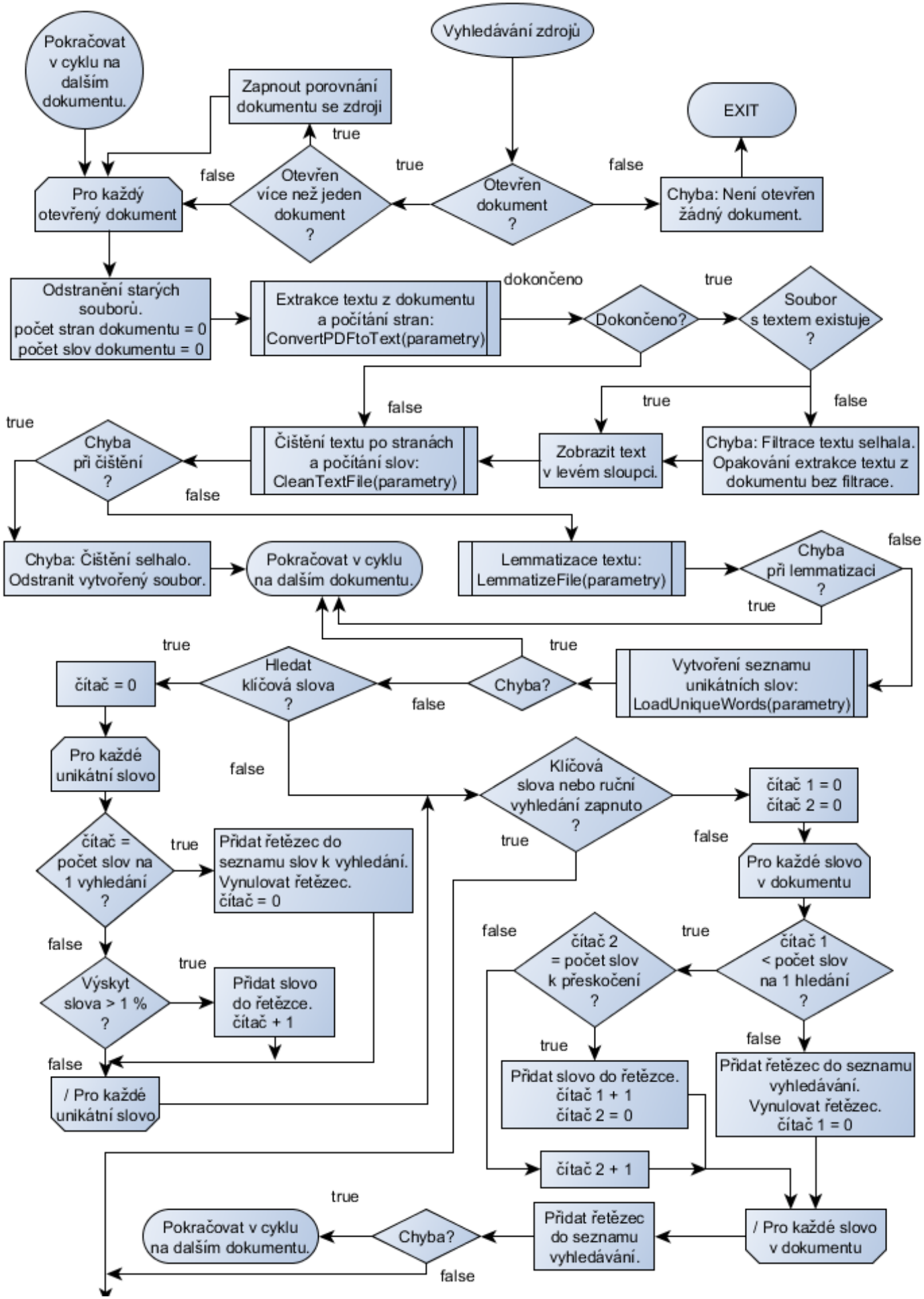
Pokud se ale snažíme o odhalení rafinovanějších forem plagiátorství, je třeba brát v úvahu prakticky každé slovo, což značně zvýší časovou náročnost kontroly dokumentu.

5.8.5. Ruční a automatický režim

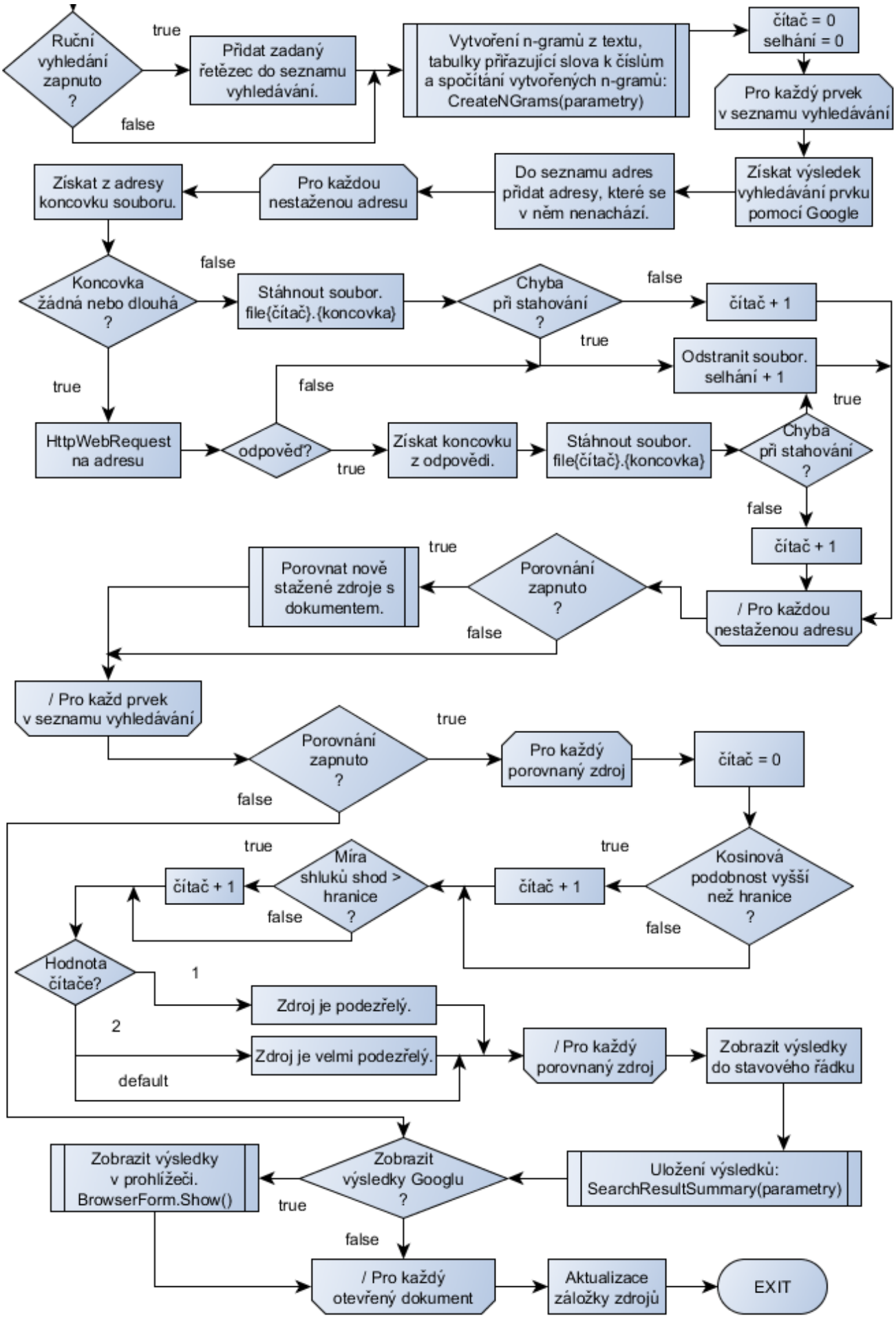
V případě, že je otevřen pouze jeden vstupní dokument, detektor se nachází v ručním režimu. V tomto režimu na konci vyhledávání pouze zobrazí informace o počtu stažených zdrojů, selhaných stažení, porovnaných zdrojů a z toho podezřelých zdrojů.

V případě, že je otevřeno více dokumentů, detektor přejde do automatického režimu. V tomto případě po ukončení vyhledávání s povoleným porovnáním uloží do textového souboru **Result Summary.txt** informace o uplynulém času, počtu stažení, počtu selhaných stažení, počtu porovnání a hodnotě n. Následuje seznam porovnaných zdrojů, který obsahuje výsledky porovnání jako je počet shod, kosinová podobnost nebo míra shluků shod a také odkaz, ze kterého byl zdroj stažen.
Dále, jsou do složky **results** uloženy zpracované soubory zdrojů, u kterých míra shluků shod překročila hodnotu 1000. Překopírováním těchto souborů do hlavní složky detektoru je poté možné dokument obnovit a prohlédnout si shody u vybraných zdrojů.

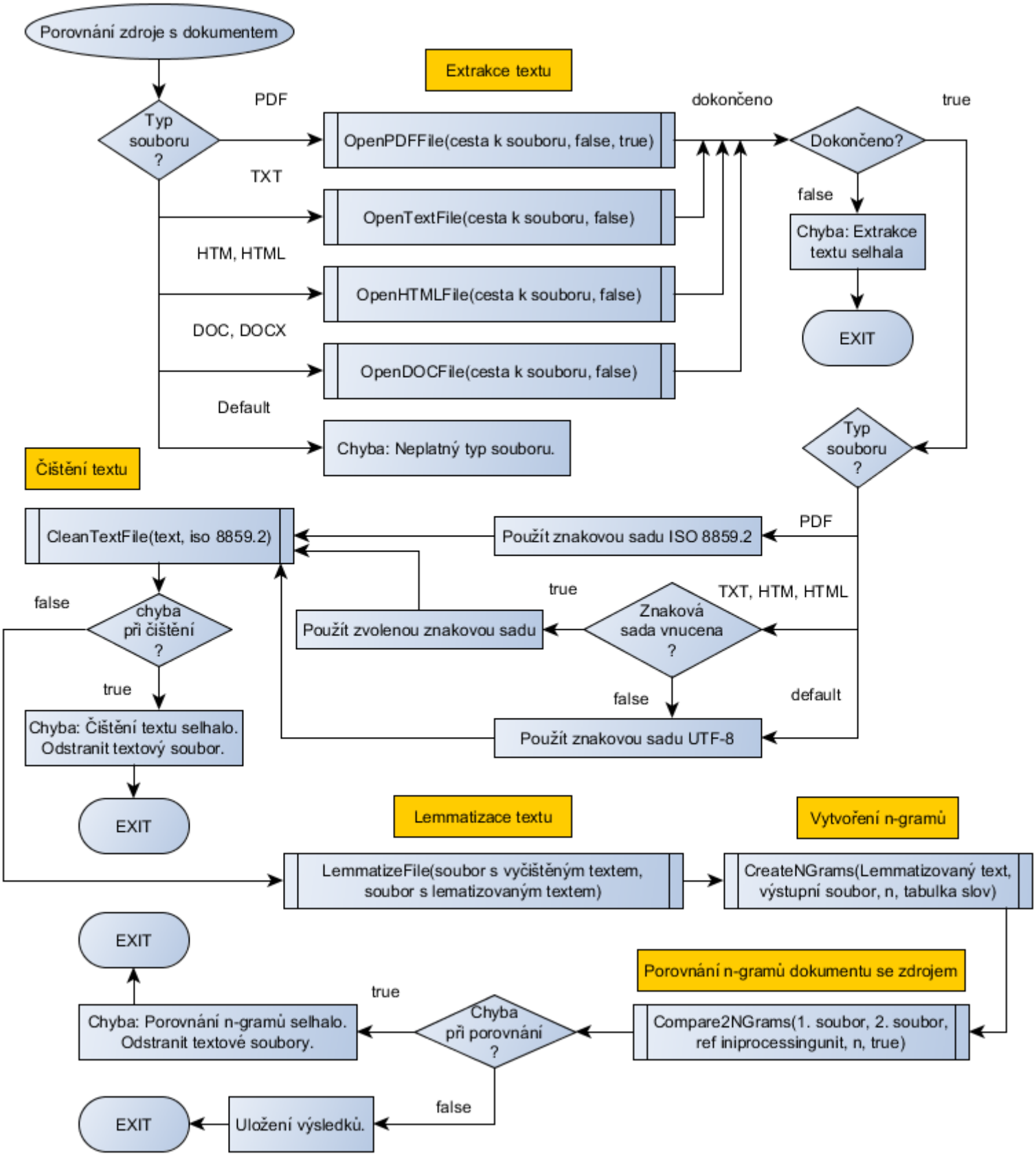
Detektor takto postupně vyhledá, porovná a uloží výsledky všech dokumentů, které byly vybrány k otevření.



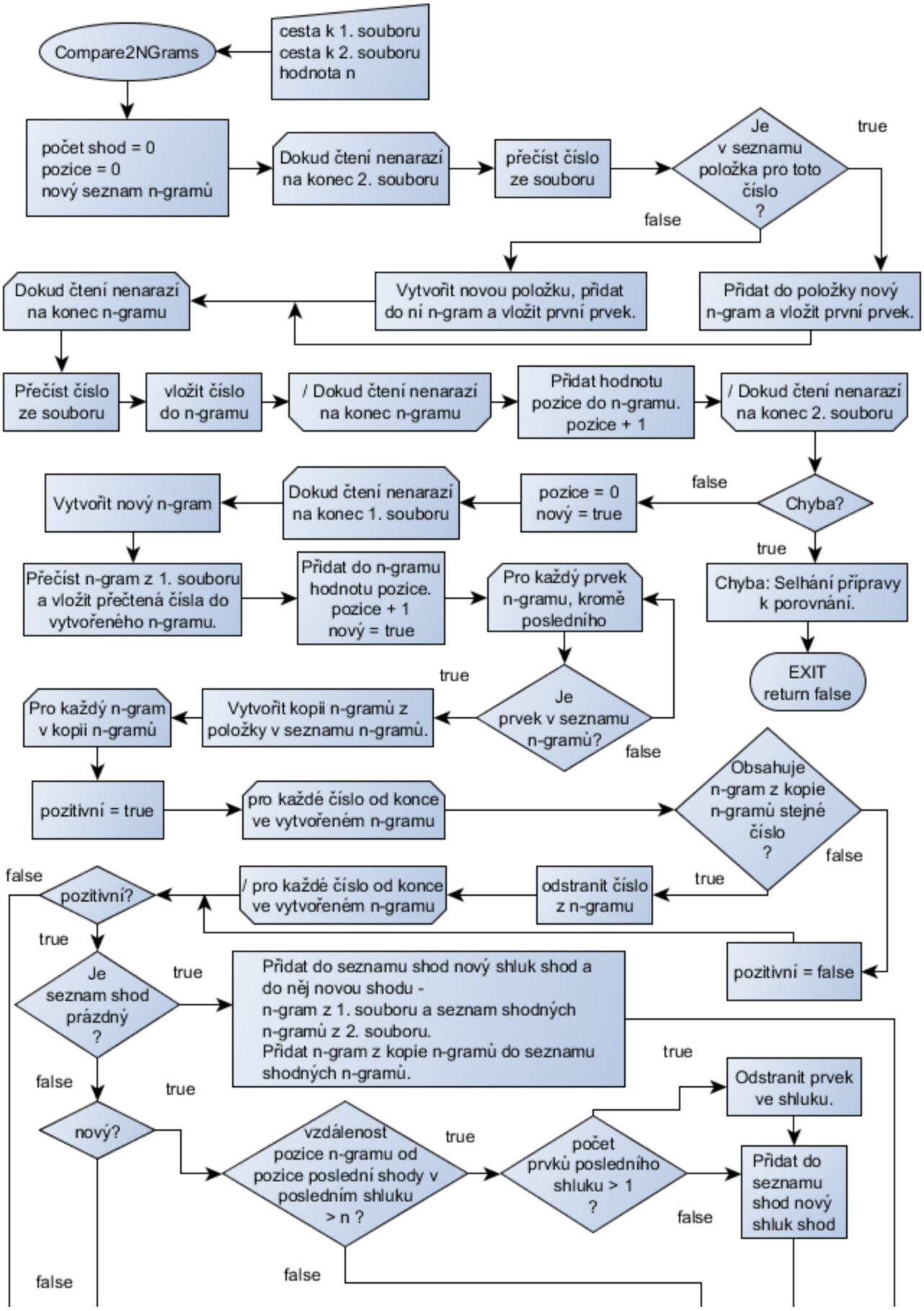
Obrázek 7: Vývojový diagram vyhledávání, 1. část



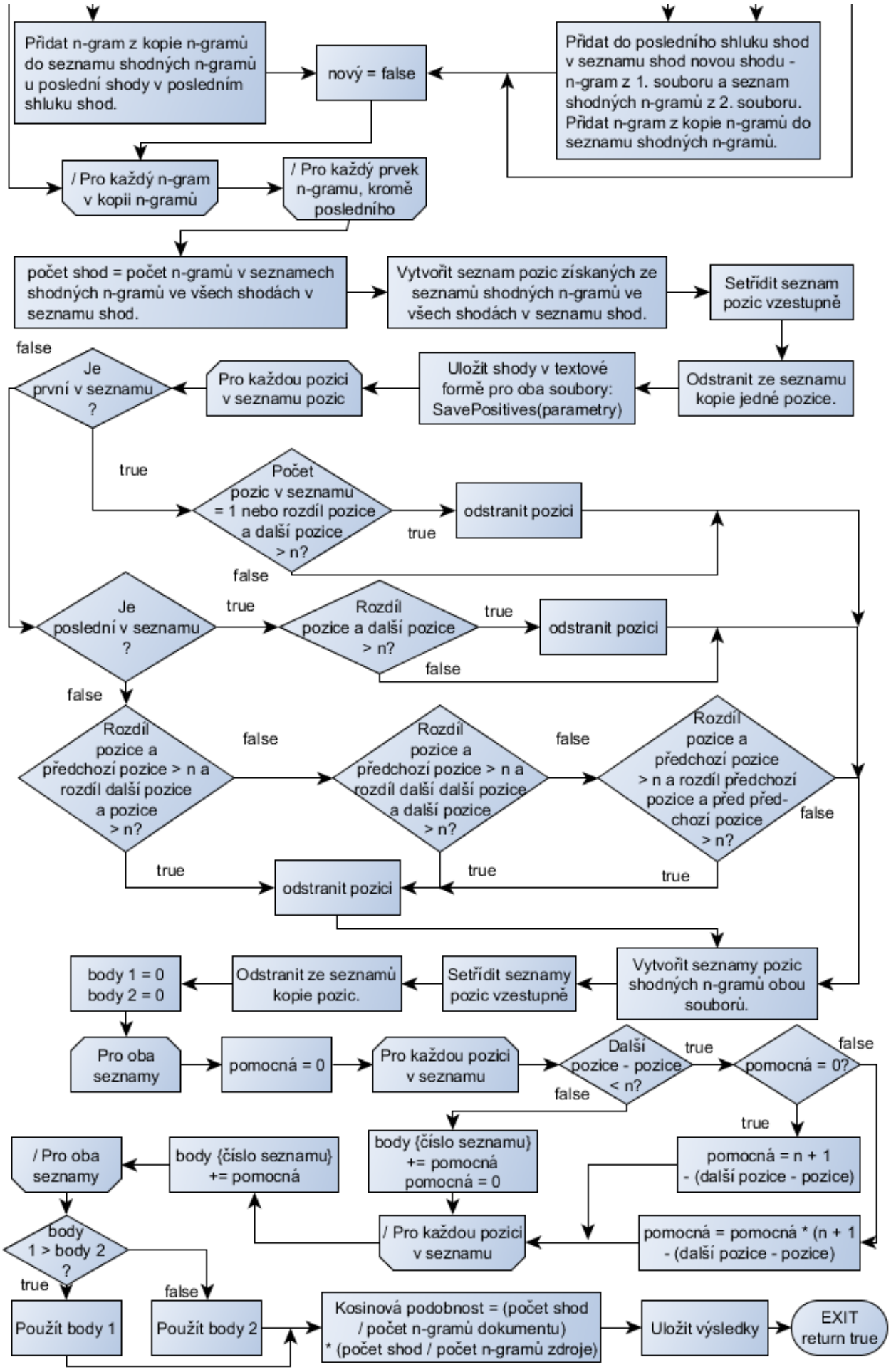
Obrázek 8: Vývojový diagram vyhledávání, 2. část



Obrázek 9: Vývojový diagram porovnání zdroje s dokumentem



Obrázek 10: Vývojový diagram porovnání dvou souborů n-gramů, 1. část



Obrázek 11: Vývojový diagram porovnání dvou souborů n-gramů, 2. část

6. VÝSLEDKY

6.1. Úspěšnost detekce plagiátorství pomocí detektoru

6.1.1. Umělý plagiát

Pro test úspěšnosti detekce byl vytvořen umělý plagiát – PDF dokument obsahující přímou kopii textu, originální text a upravenou kopii textu.

Přímá kopie byla získána ze článku **Elektrický proud v plynech** na webových stránkách cs.wikipedia.org. Použitý text:

Silné elektrické pole způsobí vytrhávání elektronů z atomů a molekul plynu (ionizaci plynu). Elektrický proud za této podmínky se nazývá elektrický výboj a je tvořen směsí volných elektronů a kladných, příp. záporných iontů v plynu.
Vysoká teplota znamená velkou kinetickou energii částic plynu, při jejichž nárazech může docházet k vyrážení elektronů z atomů nebo molekul. Elektrický proud v plynu za vysoké teploty se nazývá elektrický oblouk a je tvořen směsí elektronů a iontů. Vyznačuje se velmi jasným světelným zářením, které se využívá v obloukových lampách.

Upravená kopie byla získána z PDF dokumentu **Digitální nízkofrekvenční zesilovač s univerzálními vstupy**, který je dostupný na adrese www.elektrorevue.cz/cz/download/digitalni-nizkofrekvencni-zesilovac-s-univerzalnimi-vstupy/.
Původní text:

Jádrem tohoto bloku, který slouží pro řízení činnosti celého zesilovače, je 8bitový mikrokontrolér ATmega32 [17] vybavený řídícím firmware. Tento blok plní v zesilovači funkci řídícího zařízení (Master). Všechny ostatní bloky zesilovače jsou bloky řízené (Slave). Řídící mikrokontrolér je doplněn převodníky napěťových úrovní pro řízené bloky využívající 3,3V logiku a obvody pro spínání podsvícení LCD displeje a piezosirény. Tato piezosiréna realizuje akustickou zpětnou vazbu stisku tlačítka.
Celý program mikrokontroléru se vykonává v nekonečné smyčce, kde je testován příznak přerušení od uživatelského rozhraní (tj. stisku tlačítka). Vývojový diagram hlavního jádra programu je uveden na obrázku 2.
Po zapnutí zesilovače dojde k inicializaci všech dílčích bloků zesilovače. Po provedení jednotlivých inicializací program vstupuje do nekonečné smyčky, kde je testován příznak přerušení od uživatelského rozhraní. Pokud je tento příznak nastaven, dojde k obsloužení stisku daného tlačítka. Běh nekonečné smyčky může přerušit interní a externí přerušení.

Upravený text:

Pro řízení činnosti zesilovače slouží 8bitový mikrokontrolér ATmega32 vybavený řídícím firmware. Mikrokontrolér je jádrem bloku, který plní v zesilovači řídící funkci. Ostatní bloky

jsou řízené tímto blokem. Mikrokontrolér je dále doplněn převodníky úrovně napětí pro řízené bloky využívající 3,3V logiku, obvody spínající podsvícení LCD displeje a piezosirény. Ta slouží jako akustická zpětná vazba stisku tlačítka. Program mikrokontroléru probíhá v nekonečné smyčce, ve které probíhá test na příznak přerušení od uživatelského rozhraní. Na obrázku 2 je uveden vývojový diagram hlavního jádra programu.
Po zapnutí zesilovače se napřed inicializují všechny dílčí bloky zesilovače. Po inicializaci vstupuje program do nekonečné smyčky, kde je testován příznak přerušení od uživatelského rozhraní. Je-li tento příznak nastaven, proběhne obsloužení stisku daného tlačítka. Běh nekonečné smyčky může být přerušen interním i externím přerušením.

6.1.2. Test detektoru

Pro kontrolu umělého plagiátu bylo použito následující nastavení:

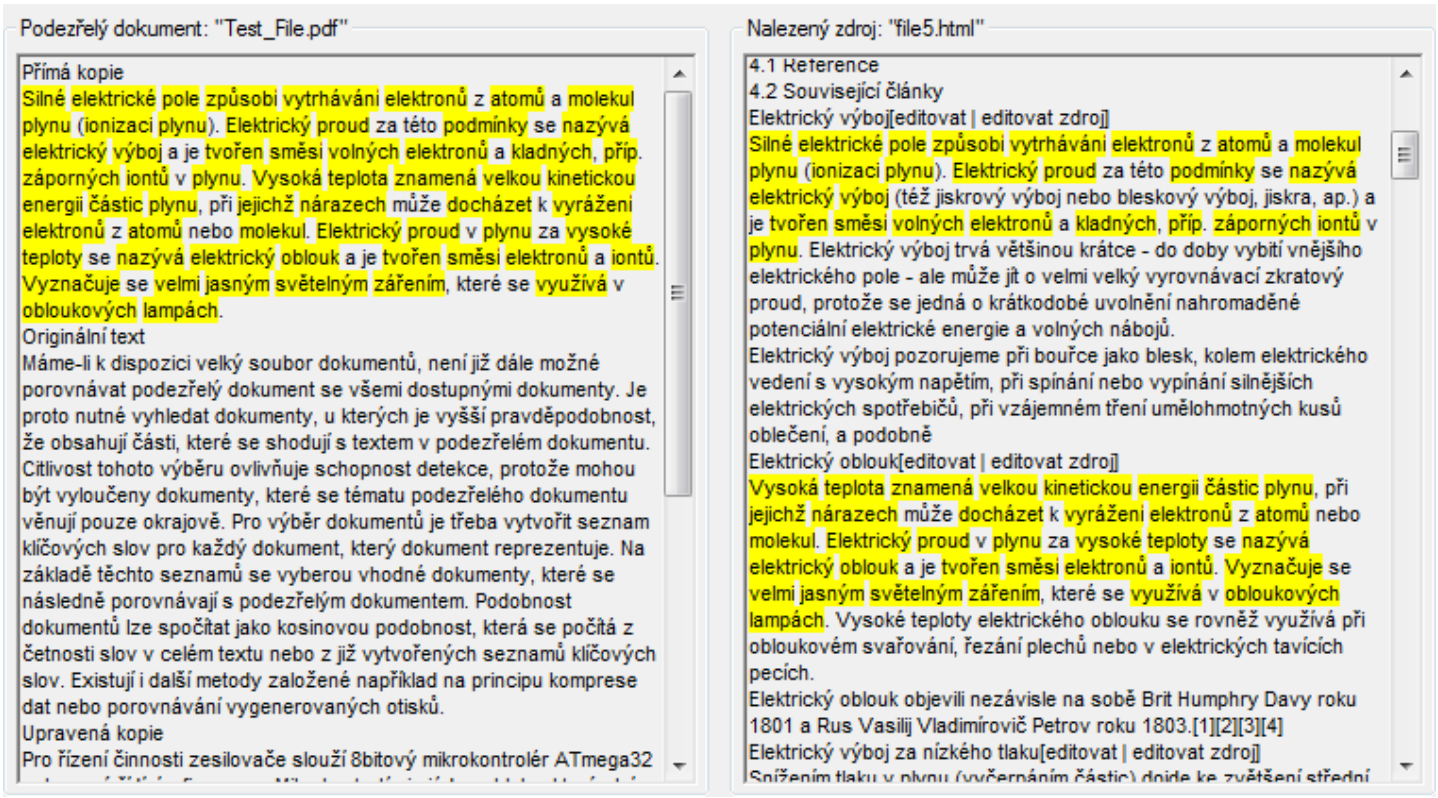
- Filtrace: NE
- Automatické porovnání: ANO
- Počet slov pro hledání: 10
- Maximální množství zdrojů na 1 vyhledání: 3
- Ignorovat každých X slov: 1
- Použít lemmatizovaná slova: ANO
- Použít klíčová slova: NE
- Hledat zdroje podle zadaných slov: NE

Při vyhledávání bylo staženo 35 zdrojů a 34 zdrojů bylo porovnáno. Výsledky porovnání ukazuje následující tabulka.

Tabulka 4: Výsledky porovnání zdrojů s umělým plagiátem

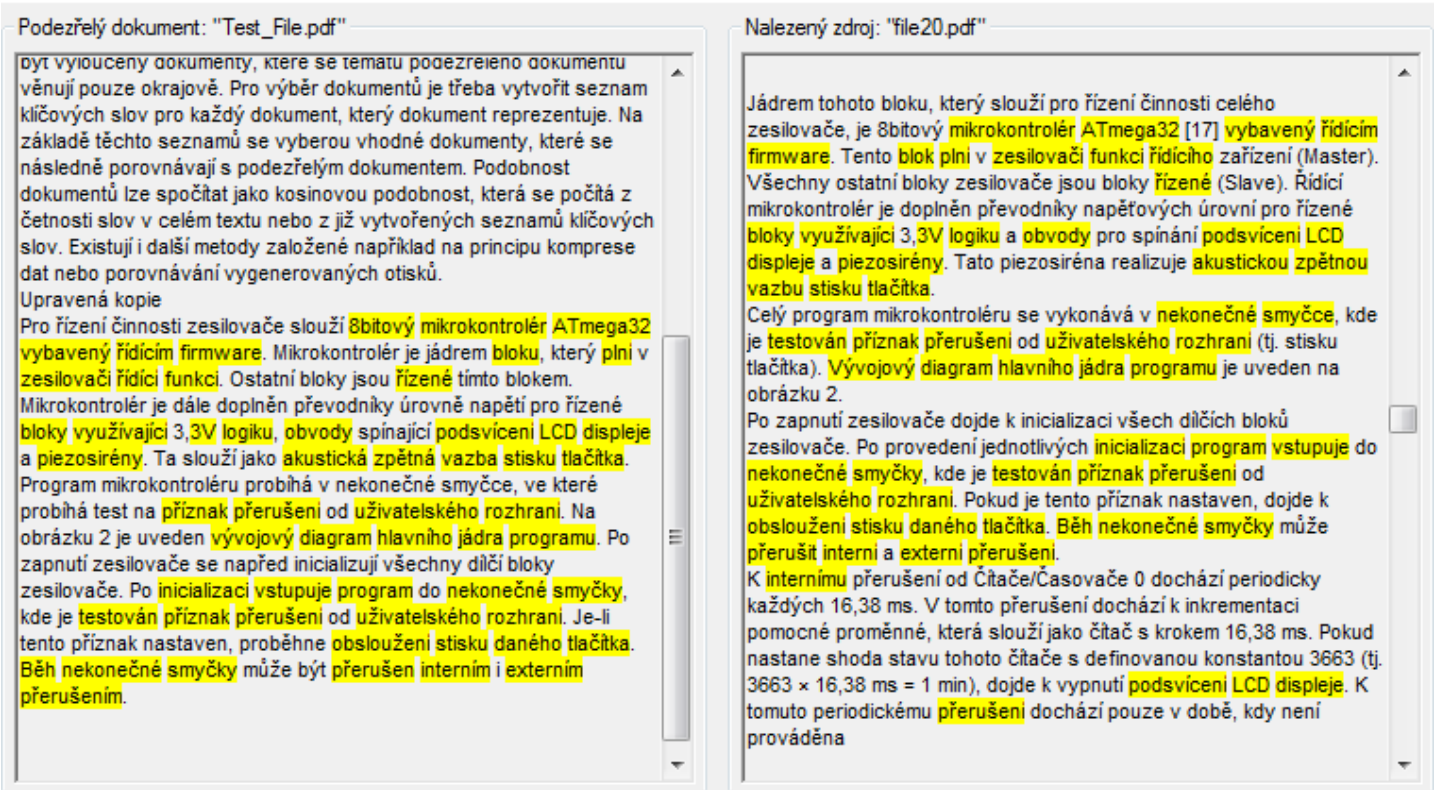
Číslo zdroje	typ souboru	počet n-gramů	počet shod	kosinová podobnost	míra shluků
0	pdf	2144	0	0,0000	0
1	html	680	0	0,0000	0
2	html	659	0	0,0000	0
3	htm	28511	0	0,0000	0
4	htm	39360	0	0,0000	0
5	html	647	56	0,0199	2147483647
6	pdf	1862	3	0,0000	0
7	pdf				
8	html	1040	0	0,0000	0
9	html	8471	0	0,0000	0
10	html	1095	0	0,0000	0
11	html	4848	0	0,0000	0
12	html	425	0	0,0000	0
13	pdf	10816	0	0,0000	0
14	html	1575	0	0,0000	0
15	html	5626	0	0,0000	0
16	html	1084	0	0,0000	0
17	pdf	6003	0	0,0000	0
18	pdf	12289	0	0,0000	0
19	pdf	8291	1	0,0000	0
20	pdf	3222	36	0,0016	20327
21	pdf	9350	0	0,0000	0
22	pdf	9947	1	0,0000	0
23	pdf	12086	1	0,0000	0
24	pdf	11310	1	0,0000	0
25	pdf	5203	0	0,0000	0
26	pdf	1120	0	0,0000	0
27	pdf	6433	0	0,0000	0
28	pdf	8576	0	0,0000	0
29	html	2534	0	0,0000	0
30	pdf	1356	3	0,0000	3
31	pdf	48654	0	0,0000	0
32	html	11486	0	0,0000	0
33	html	3549	0	0,0000	0
34	pdf	6109	0	0,0000	0

Z nalezených zdrojů do popředí jasně vystupují zdroje číslo 5 a 20, které mají zvýšenou hodnotu míry shluků shod. Abychom však mohli rozhodnout, je třeba si tyto zdroje nechat zobrazit a posoudit nalezené shody.
Zdroj č. 7 byl chráněn proti úpravám a proto nemohl být detektorem zpracován.



Obrázek 12: Nalezený zdroj přímé kopie textu

Na Obr. 12 je vidět jasně označená shoda mezi umělým plagiátem a nalezeným zdrojem. Text je zcela shodný a není proto pochyb, že detektor úspěšně nalezl zdroj přímé kopie textu.



Obrázek 13: Nalezený zdroj upravené kopie textu

Na Obr. 13 lze vidět označený text pro zdroj č. 20. Zde není shoda tak jasná jako u přímé kopie, přesto detektor díky lemmatizaci textu před porovnáním dokázal najít shodu i mezi některými slovy, které mají oproti originálu pozměněný tvar.
Při pozornějším prozkoumání vyjde najevo, jak velmi podobné si tyto dva texty jsou a lze tedy prohlásit, že detektor nalezl zdroj i pro upravenou část textu.

6.1.3. Zhodnocení

Na základě provedeného testu lze prohlásit, že vytvořený detekční nástroj je schopný nalézt a jasně označit případy přímého plagiátorství a v omezené míře nalézt a označit i případy plagiátorství, ve kterých se pachatel snažil zakrýt své stopy.
Detekční nástroj však není schopný nalézt shodu se zdrojem, který je v jiném než českém jazyce a nedokáže zpracovat dokumenty, které jsou chráněny proti úpravám heslem nebo jinou metodou.

6.2. Testy na náhodném vzorku závěrečných prací

Pro účely těchto testů bylo namátkou vybráno sto závěrečných prací, které byly v blízké minulosti vypracovány na Fakultě elektrotechniky a komunikačních technologií. V souladu se zadáním práce a rozhodnutím vedoucího práce jsou zveřejněná data anonymizována.

Na každém dokumentu ze vzorku bylo provedeno hledání zdrojů a jejich porovnání s dokumentem pomocí vytvořeného detekčního systému s následujícím nastavením:

- Filtrace: ANO
- Automatická filtrace ANO
- Automatické porovnání: ANO
- Počet slov pro hledání: 12
- Maximální množství zdrojů na 1 vyhledání: 3
- Ignorovat každých X slov: 4
- Použít lemmatizovaná slova: ANO
- Použít klíčová slova: NE
- Hledat zdroje podle zadaných slov: NE

Výsledky testů vytvořené detekčním systémem byly poskytnuty vedoucímu práce k prozkoumání. V době odevzdání práce bylo zpracováno a rozhodnuto o 28 dokumentech ze 100 testovaných, z toho 8 dokumentů bylo vedoucím práce označeno za plagiát.
Po konzultaci s vedoucím bylo rozhodnuto o rozdělení plagiátů do 3 skupin:

1. absence grafického odlišení textu od citátu, ale zdroj uveden na místě
2. absence grafického odlišení textu od citátu, ale zdroj uveden na konci práce v literatuře
3. zdroj neuveden

Nalezené plagiáty:

Typ 1:	6
Typ 2:	1
Typ 3:	1

Z výsledků je patrné, že dominantním typem plagiátu je typ 1, tedy nedostatečně zvýrazněný citát ze zdroje. Tento typ plagiátorství je ten nejméně závažný z námi definovaných typů, je důsledkem omylu nebo neznalosti a může být přehlížen nebo prominut. Teprve plagiát typu 3 se shoduje s tím, co si běžný člověk představí pod pojmem plagiátorství.

Z těchto výsledků však nelze vyvodit vlastnosti detektoru jako je například přesnost, protože neznáme skutečný počet plagiátů ve vzorku dokumentů. Lze však určit některé jiné vlastnosti týkající se procesu detekce.

Následující data jsou platná jen pro nastavení detektoru uvedené na předchozí straně a vychází z výsledků testů na vzorku 100 závěrečných prací.
Hodnoty jsou vztažené k vyhledávání dokumentu o 100 slovech (po filtraci a odstranění stopslov).

Množství slov v dokumentu

Průměr: 6339 slov

Časová náročnost

Průměr: 2,04 minut / 100 slov
Medián: 1,90 minut / 100 slov
Maximum: 3,67 minut / 100 slov
Minimum: 0,63 minut / 100 slov

Množství zdrojů s mírou shluků shod nad 1000

Průměr: 0,52 zdrojů / 100 slov
Medián: 0,36 zdrojů / 100 slov
Maximum: 1,76 zdrojů / 100 slov
Minimum: 0,02 zdrojů / 100 slov

Při kontrole dokumentu průměrné závěrečné práce s výše uvedeným nastavením detektoru lze tedy očekávat výsledky přibližně za 2 hodiny. K tomu je třeba přidat čas potřebný na prozkoumání průměrně 33 nalezených zdrojů shod a rozhodnutí o jejich závažnosti.

7. ZÁVĚR

Zjistili jsme, že současná nabídka detektorů je vyplněna především detektory pracující s anglickými texty. Mezi nimi se jako vždy objevují detektory schopné, ale i detektory, které neposkytují uživateli prakticky žádné benefity. Běžný člověk tedy nemá mnoho možností, pokud potřebuje zkontrolovat nějaký dokument, ale nechce kupovat drahou licenci, zvláště pak pokud je dokument v českém jazyce.

Problematika detekce plagiátorství je velmi široká. Vyžaduje dobré znalosti jazyka, zasahuje do porozumění jazyka strojem a hledání informací na internetu. V práci jsou uvedeny běžně používané metody jak hledat shody mezi texty, jako je použití n-gramů.

Na základě získaných znalostí jsme navrhli vlastní detekční systém, využívající služeb vyhledávače Google k nalezení možných zdrojů textu testovaného dokumentu. Detektor je stavěný specificky pro účel kontroly závěrečných prací a je schopný v určité míře ignorovat standartní části textu jako jsou například titulní strany. Dokáže sám stahovat možné zdroje hledaných úryvků textu, hledat shody mezi zdroji a dokumentem a hodnotit jejich závažnost.

Detektor jsme poté použili při dvou testech.

V prvním jsme hodnotili jeho schopnost poradit si s různými stupni plagiátorství. Ukázalo se, že detektor dokáže spolehlivě nalézt zdroj k přímé kopii textu a jednoznačně tento zdroj označit. Dokáže si v určité míře poradit i s textem, který byl oproti své původní podobě pozměněn.

V druhém testu jsme za pomocí detektoru hledali zdroje ke vzorku 100 náhodně vybraných závěrečných prací z Fakulty elektrotechniky a komunikačních technologií. Výsledky byly poskytnuty vedoucímu práce, který je pro potřeby tohoto testu rozdělil na práce čisté a plagiáty.

Navržený detekční systém je tedy použitelný pro detekci hlavně přímého plagiátorství u závěrečných prací. Metody v něm obsažené však tvoří pouze funkční základ a je zde proto místo pro zlepšení a hlavně zefektivnění celého procesu detekce.

8. POUŽITÁ LITERATURA

AFROZ, Sadia, Michael BRENNAN a Rachel GREENSTADT. *Detecting Hoaxes, Frauds, and Deception in Writing Style Online*. Drexel University, Philadelphia, [2011], 15 s. Dostupné z: https://www.cs.drexel.edu/~sa499/papers/oakland-deception.pdf

ANON. Autorský zákon: č. 121/2000 Sb. - Aktuální znění. AION CS, s.r.o. *Zákony pro lidi* [online]. [2000] [cit. 2015-01-08]. Dostupné z: http://www.zakonyprolidi.cz/cs/2000-121

ANON. Autoplagiátorství. NK ČR. *KTD: Česká terminologická databáze knihovnictví a informační vědy* [online]. [2012] [cit. 2015-01-08]. Dostupné z: http://aleph.nkp.cz/publ/ktd/00001/46/000014610.htm

BARNBAUM, C. PLAGIARISM: A Student's Guide to Recognizing It and Avoiding It. VALDOSTA STATE UNIVERSITY. *Valdosta* [online]. [cit. 2015-01-08]. Dostupné z: http://ww2.valdosta.edu/~cbarnbau/personal/teaching_MISC/plagiarism.htm

BARRÓN-CEDEÑO, Alberto a Paolo ROSSO. UNIVERSIDAD POLITÉCNICA DE VALENCIA, Spain. *On Automatic Plagiarism Detection Based on n-Grams Comparison*. 2009, 5 s. Dostupné z: http://users.dsic.upv.es/~prosso/resources/BarronRosso_ECIR09.pdf

BEALL, Jeffrey. Five Ways to Defeat Automated Plagiarism Detection. BEALL, Jeffrey. *Scholarly Open Access: Critical Analysis of scholarly open-access publishing* [online]. 2013 [cit. 2015-01-08]. Dostupné z: http://scholarlyoa.com/2013/02/07/five-ways-to-defeat-automated-plagiarism-detection/

BIERNÁTOVÁ, Olga a Jan SKŮPA. *Bibliografické odkazy a citace dokumentů: dle ČSN ISO 690 (01 0197) platné od 1. dubna 2011*. 2011, 27 s. Dostupné z: http://www.citace.com/soubory/csniso690-interpretace.pdf

ČESKÁ ZEMĚDĚLSKÁ UNIVERZITA V PRAZE; ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE; ZÁPADOČESKÁ UNIVERZITA V PLZNI. Plagiátorství - Infogram: Definice plagiátu. *Infogram: Portál pro informační gramotnost* [online]. [2008] [cit. 2015-01-08]. Dostupné z: http://www.infogram.cz/findInSection.do?sectionId=1115&categoryId=1168

HAUZÍREK, Michal. *Možnosti automatické detekce plagiátů* [online]. Praha, 2007 [cit. 2015-01-08]. Dostupné z: http://www.vse.cz/vskp/show_file.php?soubor_id=1230179. Diplomová práce. Vysoká škola ekonomická v Praze. Vedoucí práce Ing. Luboš Pavlíček.

LOTT, Brian. *Survey of Keyword Extraction Techniques*. 2012, 11 s. Dostupné z: http://www.cs.unm.edu/~pdevineni/papers/Lott.pdf

NĚMEČKOVÁ, Lenka. ÚSTŘEDNÍ KNIHOVNA ČVUT. *Plagiátorství*. Praha, 2009, 8 s. Dostupné z: http://knihovna.cvut.cz//administrace/upload_dir/files/92d3b80c1ab35ca5a6cba67ff8dceaf9c9931380.pdf

PŘIBIL, Jiří. *Efektivní metody detekce plagiátů v rozsáhlých dokumentových skladech* [online]. Jindřichův Hradec, 2010 [cit. 2015-01-08]. Dostupné z: http://www.vse.cz/vskp/show_file.php?soubor_id=1237269. Doktorská dizertační práce. Fakulta managementu v Jindřichově Hradci. Vedoucí práce Prof. Radim Jiroušek.