VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ
ÚSTAV BIOMEDICÍNSKÉHO INŽENÝRSTVÍ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION
DEPARTMENT OF BIOMEDICAL ENGINEERING

DETEKCE PLAGIÁTŮ TEXTOVÝCH DOKUMENTŮ

PLAGIARISM DETECTION OF TEXT DOCUMENTS

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE — Bc. RADEK LÍZAL
AUTHOR

VEDOUCÍ PRÁCE — Ing. LUKÁŠ SMITAL, Ph.D.
SUPERVISOR

BRNO 2015

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

Fakulta elektrotechniky a komunikačních technologií

Ústav biomedicínského inženýrství

Diplomová práce

magisterský navazující studijní obor
Biomedicínské a ekologické inženýrství

Student:	Bc. Radek Lízal	***ID:***	125525
Ročník:	2	***Akademický rok:***	2014/2015

NÁZEV TÉMATU:

Detekce plagiátů textových dokumentů

POKYNY PRO VYPRACOVÁNÍ:

1) Seznamte se s definicí plagiátorství z pohledu psaného textu a prostudujte současné metody jeho detekce. 2) Vytvořte vlastní databázi vhodnou pro detekci plagiátorství. 3) Nalezněte příznaky pro detekci plagiátorství a v prostředí Matlab separátně otestujte jejich vhodnost na vytvořené databázi. Dosažené výsledky diskutujte. 4) Realizujte detektor plagiátů textových dokumentů využívající kombinaci vhodných příznaků. 5) Navržený detektor otestujte a dosažené výsledky statisticky vyhodnoťte. 6) Algoritmy pro detekci opatřete vhodným grafickým uživatelským rozhraním.

DOPORUČENÁ LITERATURA:

[1] CHÝLA, R. Detekce plagiátorství. Ikaros [online]. 2009, roč. 13, č. 2. Dostupné z: http://ikaros.cz/node/5253

[2] SI, A., H.V. LEONG a R.W.H. LAU. CHECK: A Document Plagiarism Detection System. In Proceedings of ACM Symposium for Applied Computing. February 1997, s. 70–77.

Termín zadání:	9.2.2015	***Termín odevzdání:***	22.5.2015

Vedoucí práce: Ing. Lukáš Smital, Ph.D.
Konzultanti diplomové práce:

prof. Ing. Ivo Provazník, Ph.D.
Předseda oborové rady

UPOZORNĚNÍ:

ABSTRAKT

Tato diplomová práce se zabývá seznámením s definicí plagiátu, rozlišuje, jaké typy plagiátorství se v praxi často objevují a jak se texty podezřelé na plagiátorství vyhledávají. Způsoby detekce jsou při tomto vyhledávání zcela zásadní, tudíž se jim věnuje celá kapitola. V práci se rovněž objevují ukázky programů, které jsou v praxi již používány. Následující kapitola seznamuje s vybranými typy příznaků, které byly implementovány v prostředí Matlab k vytvoření detektoru plagiátů v textovém dokumentu. Vytvořený program je popsán v osmé kapitole. Použité příznaky a chování detektoru jsou otestovány v kapitole nazvané testování příznaků. Testováním byla zjištěna kvalita těchto příznaků. V závěru jsou pak diskutovány výsledky, zároveň s výhodami a nevýhodami detektoru.

KLÍČOVÁ SLOVA

Plagiátorství, detekce, korpus, databáze, příznak.

ABSTRACT

This diploma thesis introduces the definition of plagiarism, distinguishes the types of plagiaries which often take place in praxis and the ways of determining the suspected texts. The means of detection are essential; therefore a whole chapter is dedicated to those. For the detection purposes, it is vital to pre-process the data to reduce the demand factor of the program. There is a preview of some programs which are already being used for the detection of plagiarism. The following chapter introduces some selected indications which have been implemented in the Matlab environment to create a detector of plagiaries in text documents. The created program is described in chapter eight. The applied indications and the detector response are described in a chapter called Indications testing. The testing proved the quality of these indications. The results together with pros and cons of the particular methods are discussed in the conclusion.

KEYWORDS

Plagiarism, detection, corpus, database, indication

LÍZAL, R. *Detekce plagiátů textových dokumentů*. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií. Fakulta elektrotechniky a komunikačních technologií, 2014. 54 s. Diplomová práce. Vedoucí práce: Ing. Lukáš Smital, PhD.

PROHLÁŠENÍ

Prohlašuji, že svou diplomovou práci na téma Detekce plagiátů textových dokumentů jsem vypracoval samostatně pod vedením vedoucího diplomové práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

Jako autor uvedené diplomové práce dále prohlašuji, že v souvislosti s vytvořením této diplomové práce jsem neporušil autorská práva třetích osob, zejména jsem nezasáhl nedovoleným způsobem do cizích autorských práv osobnostních a/nebo majetkových a jsem si plně vědom následků porušení ustanovení § 11 a následujících zákona č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon), ve znění pozdějších předpisů, včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č. 40/2009 Sb.

V Brně dne ....

(podpis autora)

PODĚKOVÁNÍ

Děkuji vedoucímu diplomové práce Ing. Lukáši Smitalovi, PhD. za účinnou metodickou, pedagogickou a odbornou pomoc a další cenné rady při zpracování mé diplomové práce.

V Brně dne ....

(podpis autora)

OBSAH

SEZNAM OBRÁZKŮ

SEZNAM TABULEK

ÚVOD

Tato diplomová práce seznamuje s termínem plagiát a jasně jej definuje. Jelikož žijeme v době počítačů a informačních sítí, toto téma stále nabývá na aktuálnosti. Čtenář si po přečtení ujasní, co je definováno jako plagiátorství a kdy se může takového jednání dopustit. Práce také seznámí s typy plagiátorství a uvádí několik případů pro názornost. Následující kapitoly obsahují informace o způsobech detekce plagiátů v textových dokumentech. Další kapitola seznamuje s jednotlivými přístupy a vysvětluje rozdíly mezi nimi. Při práci s jednotlivými soubory a texty je zpracováváno velké množství informací a pro následující procesy je nutné data předpřipravit. Způsoby této přípravy se zabývá kapitola s názvem předzpracování dat. Po seznámení se s principy detekce je představeno několik nejznámějších softwarů pro vyhledávání plagiátů s krátkým popisem funkcí a jejich prostředí. Vlastní návrh detektoru je poté popsán v praktické části, kde jsou představeny jednotlivé příznaky použité ve vytvořeném uživatelském rozhraní. Každý příznak je pro detektor přínosem, ale zmíněny jsou také jednotlivé nevýhody. Samotný program je na ukázce v kapitole osm a seznamuje uživatele s prostředím programu a jeho výsledky. V práci se také nachází kapitola věnující se testování příznaků. Testováním byl získán přehled o kvalitě příznaků a také o hodnotách procentuální podobnosti pro plagiáty nebo originální texty. V příloze jsou následně uvedeny další možné výsledky programu a je podrobněji popsán samotný postup při práci s ním.

1 PLAGIÁTORSTVÍ

1.1 Definice plagiátu

Termínem plagiát se zabývá norma ČSN ISO 5127. Ta má usnadnit mezinárodní komunikaci v oblasti informací a dokumentace. Norma uvádí termíny a definice vybraných pojmů a také vymezuje vztahy mezi hesly. Plagiát definuje norma ČSN ISO 5127-2003 jako „*Představení duševního díla jiného autora půjčeného nebo napodobeného v celku nebo z části, jako svého vlastního*.“ [1].

Česká terminologická databáze knihovnictví a informační vědy (TDKIV) uvádí, že plagiát je „*Nedovolená napodobenina (přesná nebo částečná) uměleckého nebo vědeckého díla jiné osoby, která je bez uvedení předlohy vydávána za originál*.“ [2].

1.2 Definice plagiátorství

Plagiátorství tedy znamená „*Vydávání cizího literárního nebo jiného uměleckého nebo vědeckého díla za vlastní, popř. převzetí části cizí práce, bez uvedení použitých zdrojů*.“ [2]. Za plagiátorství se považuje užití jakékoli myšlenky v práci, kterou autor považuje za svou, ať už se s ní setkal v jakékoli podobě, bez uvedení jejího skutečného autora. Vytvořením plagiátu se porušuje autorský zákon, ale také základní pravidla vědecké etiky [1].

Za plagiátorství se považuje nejen zmíněné převzetí či okopírování díla někoho jiného, ale také nedbalé citování, neúmyslné opomenutí citace, která byla použita, či nesprávná práce s původním textem, kde se nejčastěji vyskytuje problém se špatným parafrázováním textu. Parafrázování textu (záměna textu původního, změna slovosledu) je také převzetím myšlenky a proto musí být citován autor textu. Plagiátorství se dopouští nejen ten, kdo text opisuje, ale také osoba, poskytující takové služby, které k plagiátorství vedou, či přímo nabádají. Jedná se o internetové stránky, které nabízejí texty zdarma nebo za úplatu [3].

2 TYPY PLAGIÁTORSTVÍ

Při detekování plagiátu se nejedná vždy o stejný plagiát. Roli hraje motivace plagiátora, zdroj informací či nevědomost.

2.1 Úmyslný plagiát

Za úmyslný plagiát se považuje:

- Zkopírování cizího díla či jeho doslovné opisování a vydávání tohoto díla za vlastní.
- Převzetí informací z práce, která ještě nebyla dokončena a publikována, a vydávání jí za svou (seminární, bakalářské, diplomové, disertační, habilitační práce aj.)
- Vydávání celku za svůj v případě, že jde o kompilaci cizích neoznačených děl.
- Kopírování grafických či jiných součástí práce bez citování skutečného autora těchto materiálů. Toto platí nejen pro celek textu či jeho částí, ale i pro strukturu, obsah nebo stylistické členění.
- Úmyslné neuvedení použitých zdrojů a získaných dat.
- Stažení práce z volně dostupných zdrojů nebo i její nákup a vydávání jí za svou [3].

2.2 Neúmyslný plagiát

K vytvoření plagiátu nedochází vždy vědomě, za plagiátorství se ale považuje také. K takovému jednání dochází nejčastěji z neznalosti či z nedbalosti.

Za neúmyslný plagiát se považuje:

- Špatné či nedostatečné citování či odkazování v textu.
- Nesprávně prováděné seskupení práce s použitím doslovných částí textu jiného autora bez jeho uvedení a necitování všech zdrojů (nesprávná kompilace).
- Text, ve kterém došlo pouze ke změnám slovosledu, pozměnění slov nebo vsunutí dalších slov, bez správné citace (nesprávná parafráze). Pokud autor práce převezme myšlenku od autora jiného, naformuluje ji svým vlastním

způsobem ve svém textu, ale řádně odkáže na skutečného autora myšlenky, jedná se o správnou parafrázi a k plagiátu nedochází.

- Nerozpoznání všeobecně známých skutečností a informací, které nemusí být doslovně citovány. Rozdíl mezi těmito informacemi je obtížné rozeznat. Za obecně známou informaci lze považovat např. matematickou formuli, Pythagorovu větu, která nemusí být citována, oproti tomu u matematických vět, které lze nalézt jen v odborných publikacích, je nutné uvést zdroj informací a citaci.
- Auto-plagiátorství (self-plagiarism) se také považuje za neúmyslný plagiát. Jedná se o necitování vlastních děl, které vznikly dříve, v práci nové.
- Použití cizí myšlenky, u které autor textu nezná zdroj nebo nemá dostatečné informace o zdroji a rozhodl se ji vydat za vlastní (kryptomnézie - skrytá paměť) [3].

3 ZPŮSOBY DETEKCE PLAGIÁTŮ

Detekce plagiátů v textových dokumentech se provádí více způsoby v závislosti na přístupu daného softwaru. Základní princip detekce spočívá ve srovnání dvou textů, kde se srovnává nový text s textem již publikovaným. Text práce, která již publikovaná byla, se musí nacházet v úložišti, aby bylo možné jej použít.

Jedním z takových přístupů odhalování plagiátů je **intrakorpální** přístup. Tento přístup využívá vlastní databázi (korpus) prací, se kterými je podezřelý text srovnáván. Další z možných přístupů se nazývá **extrakorpální**. Tento způsob detekce nevyužívá vlastní databáze, ale podezřelé texty srovnává s databázemi jinými. Tyto databáze jsou nejčastěji internetové vyhledávače nebo také sbírky elektronických dokumentů v rámci elektronických knihoven. Třetí způsob, který pomáhá najít podezřelé části v textu, se nazývá **intrinsický.** Tento způsob se ve svém přístupu liší, jelikož neumožňuje porovnání dvou textů pro nalezení plagiátů, ale je nápomocný tím, že je schopný odhalit podezřelé části práce díky změnám v její charakteristice psaní. Poslední způsob se nazývá **smíšený**, jelikož pracuje s kombinací alespoň dvou dříve zmíněných přístupů [4].

3.1 Intrakorpální přístup

V případě použití tohoto přístupu je nutné vyřešit několik základních problémů jako je uložení dokumentů do databáze, způsob zpracování algoritmu pro vyhledávání plagiovaných dokumentů a jejich porovnávání nebo redukce objemu dokumentů.

- **Uložení dokumentů do databáze** – při této operaci je nutné, aby byly dokumenty uchovávané ve stejné formě. Také je potřeba stanovit, zdali budou dokumenty nějakým způsobem předzpracovány, v jakém formátu budou uloženy a porovnávány.

- **Algoritmy pro porovnávání dokumentů** – způsob řešení tohoto algoritmu je pro detektory naprosto zásadní, jelikož závisí na tom, jak detektor pracuje a tedy jak je při detekci plagiátů spolehlivý a přesný. Jednodušší algoritmy porovnávají plagiáty, které jsou naprosto přesné (tzv. CTRL-C -> CTRL-V, copy & paste), lze ale zvolit i pokročilejší metody detekce.

Nevýhoda této metody spočívá v tom, že ve vlastní databázi musí obsahovat plagiovaný (kopírovaný) dokument. V případě, že dokument v databázi není, detekce není možná. Tento problém se po čase „řeší“ sám, jelikož při používání detektoru narůstá počet nahraných dokumentů a pravděpodobnost odhalení plagiátu se zvyšuje. Druhý problém tohoto přístupu je neschopnost odhalení originálního dokumentu. V případě, že detektor objeví 2 stejné dokumenty, jako plagiát označí ten, který byl do databáze nahrán později [4].

3.2 Extrakorpální přístup

Extrakorpální detekce plagiátů se odlišuje od předešlého přístupu tím, že nevyužívá vlastní databázi. Kontrolované práce ale také porovnává s jistým úložištěm. Korpus pro tuto detekci je nejčastěji internetový vyhledavač nebo sbírka elektronických dokumentů v rámci elektronických knihoven. Tento přístup je principiálně stejný jako intrakorpální, protože korpus uchovávající dokumenty je stejně přítomen. Tuto databázi spravuje poskytovatel dat a obsahuje specializované dokumenty, knihy nebo také vědecké články. Pokud je využíváno internetového vyhledávání, záleží na vyhledávacím algoritmu, který určí prohledávané stránky a dokumenty.

Detekce plagiátů poté není určena zpracovaným algoritmem, ale závisí na možnostech poskytovatele dat. Vyhledávání pomocí internetových průzkumníků je poté umožněno opakujícími se dotazy obsahujícími krátké úseky podezřelého dokumentu. Vzhledem k tomu, že je nutné využívat cizí služby (internet, elektronické knihovny), je tento druh detekce velmi pomalý a také značně nespolehlivý.

Tento typ detekce je možné využívat přes služby *Microsoft Bing* či *Yahoo*. Praktickým problémem při využití extrakorpálního přístupu jsou podmínky použití těchto služeb. V praxi lze totiž pro nekomerční účely provést maximálně několik tisíc dotazů za den. Při detekci plagiátů je toto číslo nedostatečné a výsledky detekce nemohou být přesné. Pro komerční použití je potom překážkou cena [4].

3.3 Intrinsická detekce plagiátů

Tato kategorie detekce je zcela speciální a od předchozích typů se zásadně liší. Metoda zkoumá charakteristické znaky textu, tedy jejich stylistické vlastnosti. Nazývá se stylometrie nebo také kvantitativní (korpusová) lingvistika.

Metodika vychází z předpokladu, že každý autor vykazuje různé hodnoty lingvistických charakteristik, mezi něž patří například průměrná délka vět a souvětí, slovní zásoba, relativní frekvence podstatných či přídavných jmen, používání frází a slovních spojení atd.

Při použití intrinsické detekce se objevuje zcela zásadní problém a to je neexistující databáze pro porovnávání textů. V případě označení textu, který je pravděpodobně plagiátem, jelikož vykazuje rozdílné stylometrické charakteristiky, neexistuje možnost zjištění konkrétního zdroje plagiovaného dokumentu. Také zde existuje určité riziko špatně definovaného podezřelého textu, kdy se může zdát, že jde o plagiát, ale ve skutečnosti autor psal text ve spěchu, bez času na korektury a proto se stylistika textu liší. Odlišení takto falešně detekovaných podezřelých textů je prakticky nemožné [4].

Intrinsickou metodu detekce plagiátů lze v praxi použít jako pomocnou metodu. Pro samotnou detekci se nepoužívá.

3.4 Smíšené metody detekce plagiátů

Ze zmíněných postupů vyplývá, že žádný typ detekce není zcela jednoznačný. V případě, že by se všechny metody zkombinovaly dohromady, detekce se více zpřesní. Lze si představit, že text je intrinsickými měřeními procházen a nalezne podezřelý úsek. Tento úsek by byl nejprve prověřen intrakorpálními metodami, proti již v systému uloženým dokumentům, a poté extrakorpálním přístupem, oproti otevřeným zdrojům na internetu.

V praxi se však intrinsické metody pro detekci plagiátu nepoužívají. Kombinují se pouze přístup extrakorpální s intrakorpálním. S vnější databází navíc až v případě, kdy není plagiování potvrzeno v přístupu s vlastní databází [4].

3.5 Detekce neviditelného značkování

Detekce neviditelného značkování spadá do kategorie nástrojů, které nepracují s obsahem textu, ale vyhledávají tzv. vodoznaky v dokumentech. Autor textu může pomocí nepatrných změn do své práce zadat okem nepostřehnutelné změny (prodloužení horní vodorovné čáry v písmenu „t“, změna mezer mezi řádky, střídání délek mezer mezi slovy, atd.). Detekování takových nepravidelností v textu znamená, že plagiátor text zkopíroval.

Výhoda detekce takto zakódovaných znaků poté umožňuje snadnější zjištění plagiátů u krátkých textů a také je možné rozlišit plagiátora a skutečného autora. Detektor s vysokou úspěšností nachází skutečné plagiáty, ne pouze podobné texty.

Nevýhoda ale může nastat v případě, že plagiátor kódování smaže, nebo zkopíruje jen část kódu, tudíž znehodnotí značkování. V takovém případě nelze rozpoznat původní text od opsaného.

4 PŘEDZPRACOVÁNÍ DAT

Zásadním problémem, který ovlivňuje výpočetní náročnost intrakorpální detekce je ve zjištění úseků textu, které spolu souvisí. Nemá smysl prohledávat texty, které jsou psány v českém jazyce s texty, které jsou psány v jiném jazyce. Také je zbytečné prohledávat texty, které mají jiná témata. V přístupu, který má vlastní databázi, je tedy nutné zjistit podobnost prohledávaného textu s dokumenty uloženými v korpusu. K takovému předzpracování slouží kosinová podobnost a také měření pomocí extrakce klíčových slov [4].

4.1 Kosinová podobnost

Kosinová podobnost je známá ve více oblastech vědy. V případě využití pro zjištění podobnosti textů pracuje na principu míry četnosti jednotlivých termínů (slov). Jedná se tedy o porovnání četnosti výskytu stejných slov ve dvou dokumentech. Kosinová podobnost je definována jako:

$$\cos im(A,B) = \frac{\sum_{i=1}^{n} a_i * b_i}{\sqrt{\sum_{i=1}^{n} (a_i)^2 * \sum_{i=1}^{n} (b_i)^2}} \tag{4.1}$$

kde A a B jsou jednotlivé vektory reprezentující dokumenty vytvořené ze slov tvořící oba dokumenty, tedy

$A = [a_1, a_2, a_3, ...a_n]$,

$B = [b_1, b_2, b_3, ...b_n]$.

Příklad řešení kosinové podobnosti:

Věta *A*: „ *Bez peněz do hospody nelez*.“

Věta *B*: „ *Člověk bez peněz je vlk bez zubů*.“

Pomocný řetězec *C* je poté složen z jedinečných slov obou vět, abecedně seřazený: „ *bez člověk do hospody je nelez peněz vlk zubů*.“ Hodnoty a_i resp. b_i jsou pak určeny výskytem slova v řetězci *C*.

A= [1,0,1,1,0,1,1,0,0],

B= [1,1,0,0,1,0,1,1,1],

$$\cos sim(A,B) = \frac{1*1+0*1+1*0+1*0+0*1+1*0+1*1+0*1+0*1}{\sqrt{(1+0+1+1+0+1+1+0+0)*(1+1+0+0+1+0+1+1+1)}} = \frac{2}{\sqrt{30}} = 0{,}365$$

Hodnota *cosim (A, B)* nabývá hodnot 0, kdy jsou dokumenty absolutně nezávislé, většinou jde o dokumenty, které jsou psané v jiném jazyce, nebo 1, kdy jde o naprosto totožné texty. Výpočet kosinové podobnosti je velmi jednoduchý, rychlý a informace o podobnosti textů je zcela jasná. Hodnota podobnosti, pohybující se nad 0,8, již v reálných případech ukazuje na dokumenty, které se nápadně podobají. Naopak hodnota pod 0,5 jednoznačně vyvrací podobnost textů [4]. Výsledky je poté možno využít pro zjištění místa, které může být potenciálně plagiované.

4.2 Extrakce klíčových slov

Mnohem složitější metoda pro předzpracování dat je extrakce klíčových slov z textu. Tuto metodu je vhodné provést již při ukládání dokumentu do databáze. Proces spočívá v označení klíčových slov, která se poté porovnávají. Jako klíčová slova se označují podstatná jména vyskytující se v textu nejčastěji. Tato slova jsou porovnávána s referenčním databázovým slovníkem za pomoci lexikální databáze WordNet, pro zjištění synonym [5].

Metoda je nejvhodnější pro předzpracování textu v anglickém jazyce, jelikož databáze WordNet obsahuje nejvíce anglických slov [4].

4.3 Metriky

Nástroje pro detekci plagiátů pracují ve velké většině na bázi porovnávání dokumentů a hledání shody mezi nimi. Samotné přístupy se od sebe mohou odlišovat například v tom, jak dokument před zpracováním upravit, jestli pracovat s dokumenty celými nebo se zjednodušenou verzí a také v tom, co považovat za základní jednotku pro porovnání. Základní jednotkou může být zvoleno slovo, věta, odstavec apod., výsledkem je ale vždy číslo, které vyjadřuje podobnost dokumentů [13].

4.3.1 Symetrická metrika podobnosti

Symetrická metrika podobnosti je základní metrika pro porovnání dvou dokumentů. V literatuře je nejčastěji nazývána podobnost (resemblance) a je aplikací Jaccardova koeficientu, který je standardně používaný při určování měr podobnosti. Symetrickou

podobnost lze popsat:

$$res(A,B) = \frac{|V(A) \cap V(B)|}{|V(A) \cup V(B)|} \tag{4.2}$$

Resemblance (podobnost) nabývá hodnot 0-1, kde nula znamená naprosto rozdílné dokumenty, jedna naopak dokumenty naprosto stejné [13].

Jak symetrická podobnost dokumentů pracuje, ukazuje následující příklad:

$$V(A) = \{1,2,3,4,5,6,7,8,9,0,a,b,c,d,e\} = 15$$
$$V(B) = \{w,3,x,7,c,0\} = 6$$
$$V(A) \cap V(B) = \{3,7,c,0\} = 4$$
$$V(A) \cup V(B) = \{1,2,3,4,5,6,7,8,9,0,a,b,c,d,e,w,x\} = 17$$
$$res(A,B) = \frac{|V(A) \cap V(B)|}{|V(A) \cup V(B)|} = \frac{4}{17} = 0{,}235$$

Výsledek znamená, že podobnost těchto dokumentů je 23,5%. Jelikož jde o symetrickou metodu, platí, že res(B, A)= 0,235.

Nevýhoda této metody je nepřesnost výsledků u dokumentů různé délky [14].

4.3.2 Nesymetrická metrika podobnosti

Jak bylo zmíněno v předchozí kapitole, nevýhoda symetrické metody je výskyt nepřesností u dokumentů různé délky. Zkreslené výsledky eliminuje nesymetrická metrika podobnosti. Metriku nezajímá pouze shoda dokumentů, ale také v jaké míře je jeden dokument obsažen v druhém. Z principu metody vyplývá i jiné pojmenování metody a to z anglického slova containment. Nesymetrickou metodu lze také zapsat do rovnice, jako v předchozím případě:

$$con(A,B) = \frac{|V(A) \cap V(B)|}{|V(A)|} \tag{4.3}$$

Containment jako výsledek obsahuje 2 proměnné a to, na kolik procent je dokument *A* obsažen v dokumentu *B* a také, z kolika procent je dokument *B* složen z dokumentu *A*. Je to opět číslo 0-1, kde hodnotu 1 nabývá v případě, že dokument *A* je podmnožinou dokumentu *B*, je v něm celý obsažen, nebo 0, kdy nemají dokumenty žádný společný obsah.

Postup výpočtu nesymetrické podobnosti je naznačen na příkladu:

$$V(A) = \{1,2,3,4,5,6,7,8,9,0,a,b,c,d,e\} = 15$$
$$V(B) = \{w,3,x,7,c,0\} = 6$$
$$V(A) \cap V(B) = \{3,7,c,0\} = 4$$
$$V(A) \cup V(B) = \{1,2,3,4,5,6,7,8,9,0,a,b,c,d,e,w,x\} = 17$$

$$con(A,B) = \frac{|V(A) \cap V(B)|}{|V(A)|} = \frac{4}{15} = 0{,}267$$

Ze vzorce vyplývá, že dokument *A* je obsažen v dokumentu *B* téměř z 27%, respektive 27% obsahu z dokumentu *A* se také nachází v dokumentu *B*.

Druhý výsledek, kde se zjišťuje obsah *B* v dokumentu *A*, je získán pomocí vzorce:

$$con(B,A) = \frac{|V(A) \cap V(B)|}{|V(B)|} = \frac{4}{6} = 0{,}667 \qquad (4.4)$$

Dvě třetiny dokumentu *B* jsou tedy také v dokumentu *A*. V tomto případě jsou to čtyři písmena z šesti.

Výhoda nesymetrické metody podobnosti je tedy nezávislost na délce dokumentů [13].

5 ODHALOVÁNÍ PLAGIÁTŮ

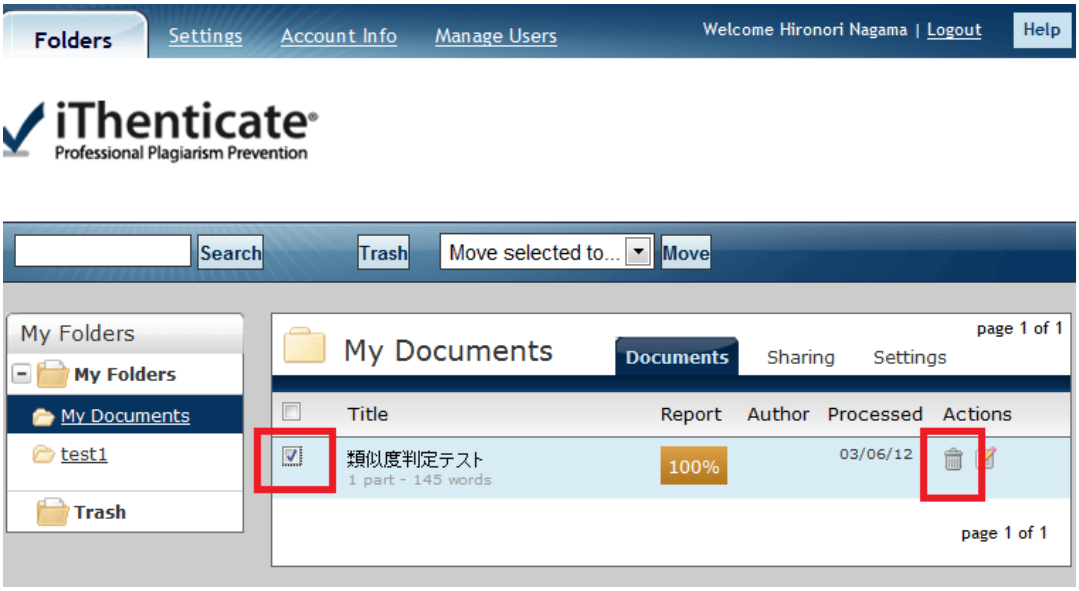
K odhalování plagiátů se využívají přístupy intrakorpální (s vlastní databází), extrakorpální (s externí databází), intrinsický (využívající stylistiku textu) a jejich kombinace. Softwary, které slouží k nalezení plagiátů, si způsob detekce vyhodnocování chrání jako své obchodní tajemství.

5.1 Softwary pro odhalování plagiátů

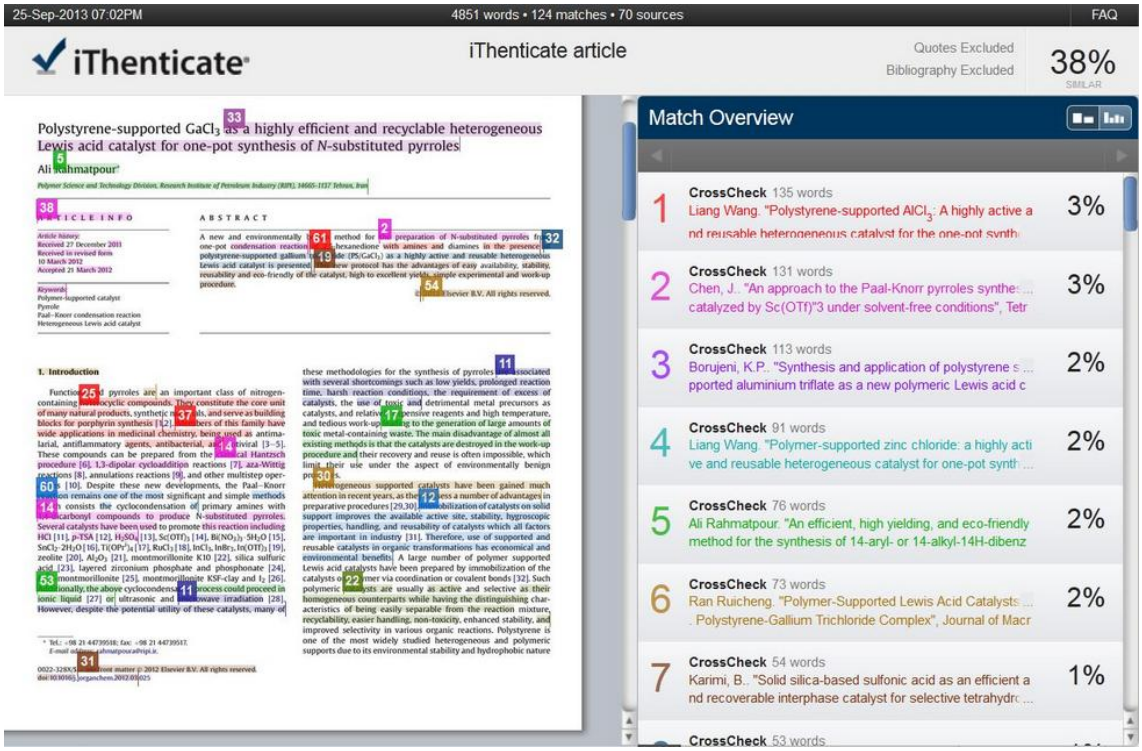
Kvůli rostoucímu počtu plagiátorských prací na všech úrovních bylo nutné tyto problémy řešit. Proto vzniklo několik softwarů, které se snaží o detekci takových děl.

Softwary pro odhalování plagiátů:

Ithenticate- software pro vyhledání a detekci plagiátů. Program umožňuje jednotlivým subjektům okamžité ověření původu - originality dokumentů. Ithenticate také zjišťuje, zdali dokumenty osobního vlastnictví uživatele nejsou neoprávněně používány [1]. Obr. 5.1 ukazuje pracovní prostředí zaregistrovaného uživatele, jeho práce a nakolik procent jsou podobné pracím jiným. Program také odkazuje na podobné práce a také podobné úryvky z textu zabarví, viz Obr. 5. 2. Program je placený.

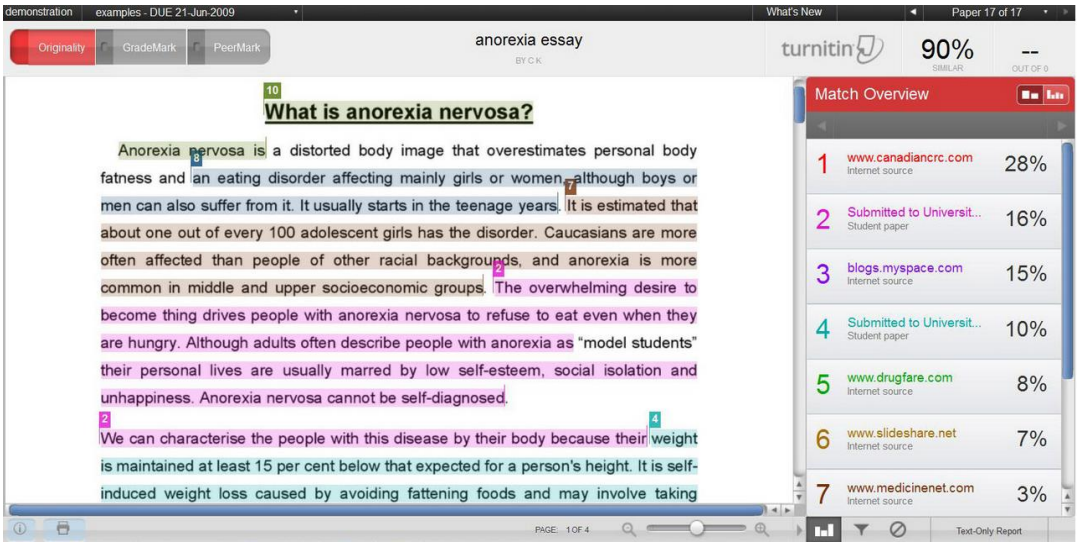


Obr. 5.1: Ukázka z programu Ithenticate. Převzato z [6].



Obr. 5.2: Ukázka z programu Ithenticate. Převzato ze [7].

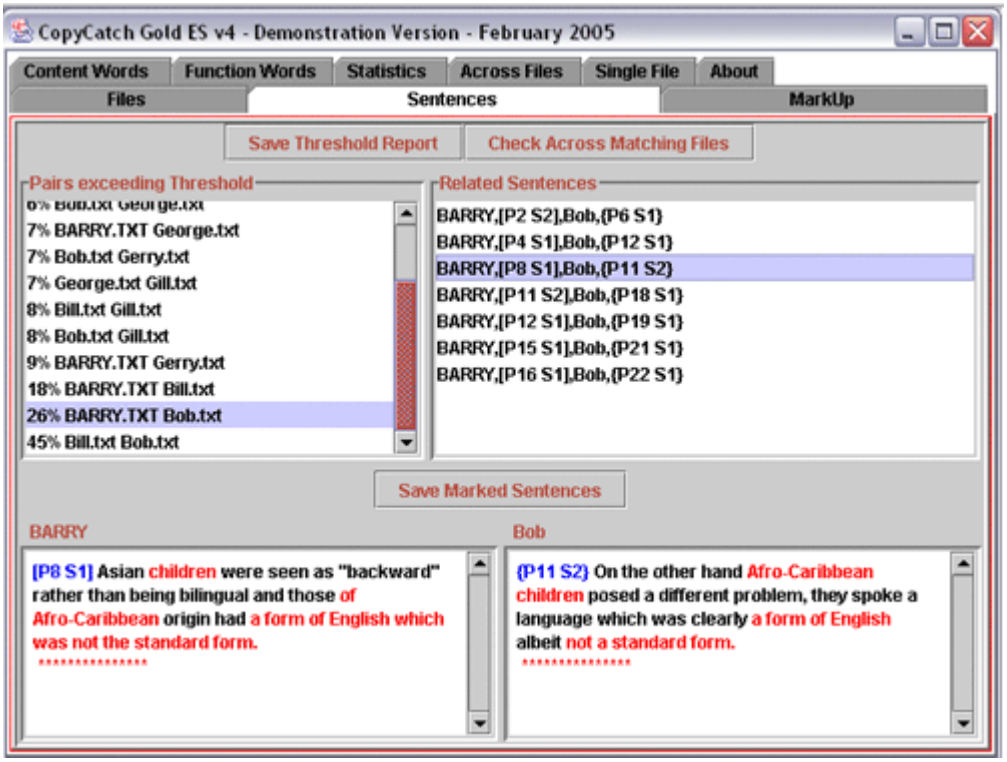
Turnitin - program, který je při odhalování plagiátů hojně využíván [1]. Tento program také procentuálně vyjadřuje shodnost textu s jinými, které také barevně označí a odkáže na podobnou práci, viz Obr. 5.3.



Obr. 5.3: Výsledná detekce textu v programu Turnitin. Převzato z [8].

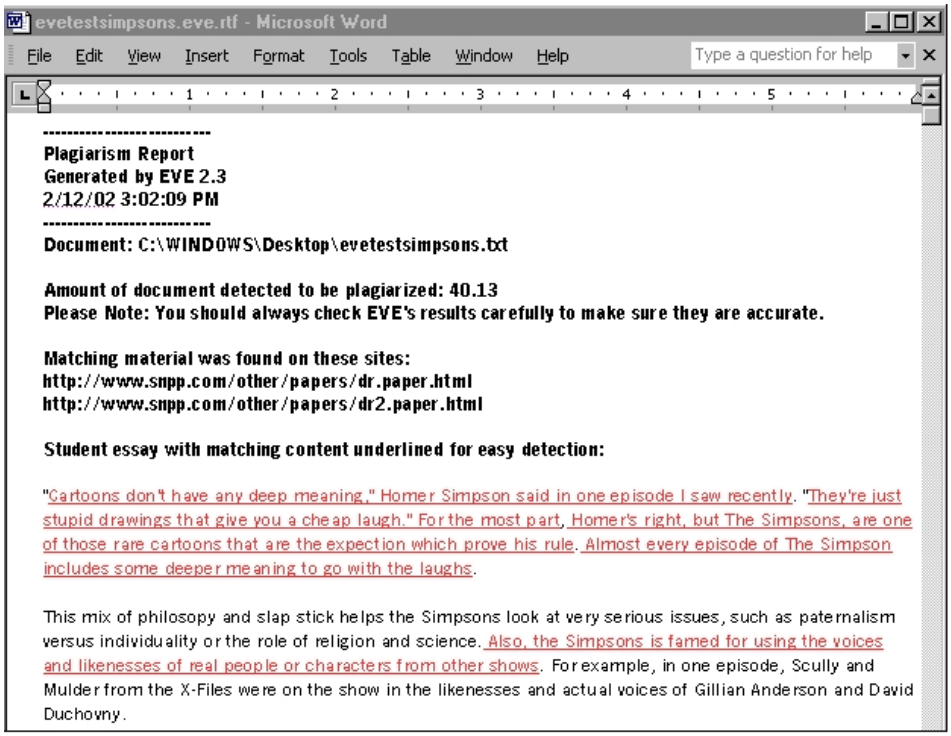
Moss (Measure of Software Similarity) - program, který lze využít při detekci zdrojového kódu. Detekuje v jazycích C, C++, Java, Pascal, Ada, ML, Lisp, Matlab aj. Tento program je zdarma a pro nekomerční použití [1].

Copy Catch Gold - systém pro odhalování plagiátů v elektronických materiálech. Tento software se využívá např. ve školách, jelikož porovnává vybrané dokumenty a zobrazí, jak se vybrané dokumenty v obsahu shodují, viz obr 5. 4 [1].



Obr. 5.4: Ukázka ze softwaru Copy Catch Gold, na které lze vidět detekování souboru. Podobnost se rovná 26%. Software také vyznačí stejná slova. Převzato z [9].

Essay Verification Engine (EVE2) - tento software provádí kontrolu textů a vyhodnocuje jejich shodnost. Program je placený, lze ale získat Trial verzi na 15 dnů. Po porovnání je programem vygenerována krátká zpráva, která pojednává o výsledku detekce. Shodu vyjadřuje procenty a shodný text vypíše, viz Obr. 5. 5 [1]. Program je určený pro PC s Windows 95/ 98/ ME/ NT4/ 2000/ XP.



Obr. 5.5: Výsledná zpráva po detekci v programu Essay Verification Engine. Převzato z [10].

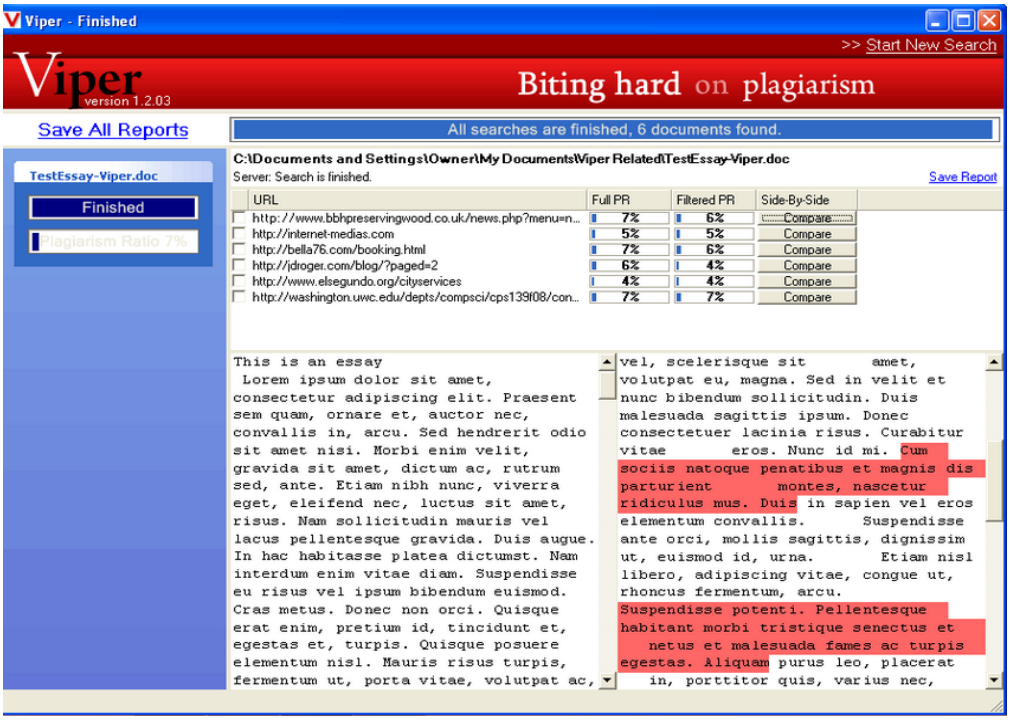
Plagiarism Finder - aplikace systému Windows. Lze spustit na každém PC, které má přístup na internet. Podobně jako předchozí software kontroluje shodnost dokumentů a poté vytvoří podrobnou zprávu, ve které nechybí informace o procentuální shodě i s odkazem na zdroj informací- viz Obr. 5. 6 [1]. Program je placený.

WcopyFind - program porovná vybrané části textů nebo slov ve frázích. Není tolik využíván, jelikož neumí pracovat s pdf formátem a neprochází internetové stránky.

Scan my Essay - systém, který porovná shodnost textů. Tento program je zdarma a také podává výslednou zprávu s procentuální shodností s jiným textem. Stejné pasáže také zvýrazní. Ukázka z programu je na Obr. 5.7.

Programy, které detekují shodnost textů, nejsou schopny samy rozhodnout, zda se jedná o plagiát. Pokud je shodnost vysoká, vždy o plagiátu musí rozhodnout lidský faktor.

Obr. 5.6: Výsledná zpráva o shodě dokumentu s jinými dostupnými dokumenty z programu Plagiarism Finder. Převzato z [11].



Obr. 5.7: Výsledná zpráva o shodě s jinými texty v programu Scan My Essay. Převzato z [12].

6 PRAKTICKÁ ČÁST

Praktická část této práce spočívala ve vytvoření programu, nacházejícího texty s podezřelou shodou, které mohou být potenciálními plagiáty. Pro tento program bylo využito prostředí Matlab, kde bylo vytvořeno grafické uživatelské rozhraní pro větší přehlednost a lepší použitelnost. Příznaky pro zjištění podobnosti textů byly na základě předchozích kapitol a rozborů zvoleny následovně:

- zjišťování počtu slov ve zvolených dokumentech
- symetrická podobnost dokumentů
- zjišťování shodnosti pěti nejpoužívanějších slov s počtem výskytů u zvolených dokumentů
- nesymetrická podobnost textů
- naprostá shodnost vět v textech
- kosinová podobnost textů
- velikost textů v bitech

Pro nahrávání dokumentů do Matlabu je nezbytné nahrát data ve stejném formátu. Jelikož se jedná o texty, je nutné převést dokumenty, které mají být zpracovány, do textového formátu označeného koncovkou „.txt“.

6.1 Zjišťování počtu slov ve zvolených dokumentech

První příznak v programu pracuje na porovnání počtu slov a jeho vyhodnocení. Text, který je nahrán do uživatelského prostředí, je rozdělen na slova. Počet těchto slov je poté porovnán se stejně zpracovaným dokumentem a vyhodnocen. Výsledek tohoto příznaku je vyjádřen v procentech.

Výhoda tohoto příznaku je nenáročnost a rychlost vyhodnocení. Opsané nebo mírně změněné texty tento příznak vyhodnotí jako nápadně podobné. Jeho výhoda se naplno projeví ve spojení s ostatními příznaky. Vytvořit detektor pouze na základě tohoto příznaku je však zcela irelevantní, jelikož shoda počtu slov může být náhodná a příznak by mohl vyhodnotit jako plagiát i texty s naprosto odlišnými tématy. V uživatelském rozhraní se tento příznak nazývá počet slov.

6.2 Symetrická podobnost dokumentů

Princip metody byl popsán v předchozí kapitole 4.3.1. Implementace v Matlabu tedy vycházela ze stejného postupu. Texty byly upraveny tak, aby jednotlivá slova byla oddělena. Pro zvýšení počtu detekovaných slov, byla všechna slova zbavena diakritiky, pro případ, kdy plagiátor záměrně nepoužije diakritiku nebo ve slově udělá chybu. Sestavení rovnice dle (4.2) se získá výsledek podobnosti dvou porovnávaných dokumentů.

Také u této metody je výhodou jednoduchost a rychlost výpočtů. Získané výsledky mohou naznačit nápadnou podobnost dokumentů. Při posuzování kratších odstavců, které jsou podezřelé z plagiátorství, je samotná metoda poměrně přesná. Nevýhoda metody se projeví u textů, které nemají stejnou délku. Na výsledku se pak projeví určitá nepřesnost. V uživatelském rozhraní se tento příznak nazývá symetrická podobnost.

6.3 Nesymetrická podobnost dokumentů

Stejně jako metoda symetrická, tak i tato metoda vychází z teorie popsané v kapitole 4.3.2. Data jsou na vstupu detektoru ošetřena stejným způsobem, jako u všech předchozích příznaků. Rozdíl oproti symetrické metodě je tedy ve finálním vzorci, kde je možné zjistit výskyt dokumentu A v dokumentu B a opačně.

Výhoda této metody je, že obsahuje dva výsledky, které nejsou zkresleny chybou v závislosti na délce dokumentu. Uživateli programu je vypsáno z kolika procent jeden dokument obsahuje druhý. V uživatelském rozhraní se tento příznak nazývá nesymetrická podobnost.

6.4 Zjišťování shodnosti nejpoužívanějších slov

Tato metoda ve své podstatě vypočítává výskyt jednotlivých slov, které podle jejich počtu opakování sestaví do nové matice. Dílčí výsledky obsahují pět nejpoužívanějších slov s počty jejich výskytů, které se porovnají mezi dvěma posuzovanými soubory, a pokud se stejné slovo nachází v obou dokumentech se stejným počtem výskytů, výsledek je vyhodnocen jako kladný. Tento příznak, stejně jako předchozí, na začátku upravuje data pro zpracování. Věty se rozdělí na slova neobsahující diakritiku. Také je nutné zbavit se předložek a spojek, které se u souvislých delších textů objevují, a jejichž shoda by mohla být zcela náhodná. Upravené dokumenty více charakterizují obsah

samotného textu a nezkreslují výsledek. Pokud se pět nejpoužívanějších slov u dvou dokumentů shoduje ve třech případech, s velkou pravděpodobností se jedná o plagiát.

Velkou výhodou tohoto příznaku je jeho spolehlivost. V uživatelském rozhraní se tento příznak nazývá klíčová slova.

6.5 Shodnost vět v textech

Příznak, který se zabývá naprostou shodností vět v textech, je jedním z nejspolehlivějších. Jedná se o porovnávání jednotlivých vět mezi dokumenty a v případě rovnosti je vyhodnocen výsledek jako kladný. Pokud plagiátor opisuje text a nepřidává žádnou svou myšlenku, detektor nalezne shodné úseky a určí shodu. Texty, se kterými se pracuje, jsou upraveny v prostředí Matlab tak, aby jednotlivé věty byly odděleny na řádcích. Dělícími znaky jsou tečka, otazník a vykřičník. Stejně jako ve všech předchozích metodách, i zde jsou slova zbavena diakritiky. Po těchto úpravách jsou věty mezi sebou porovnávány. Výsledkem detektoru je poznámka, na kolik procent jsou si věty v dokumentech podobné.

Za výhodu lze považovat spolehlivost a v principu jednoduchost metody. Detektor odhalí plagiátory, kteří původní dílo kopírují. Nevýhoda příznaku se projeví v případě, že plagiátor prokáže vlastní iniciativu a pozmění slova v každé jednotlivé větě. V uživatelském rozhraní se tento příznak nazývá podobnost vět.

6.6 Kosinová podobnost

Příznak nazvaný kosinová podobnost je také značně spolehlivý. V principu se i v tomto případě jedná a rozbor slov, se kterými se v dokumentech pracuje. Pokud jsou slova stejná, je opět vyhodnocena shoda jako kladný výsledek. Počet shod je pak dle rovnice (4.1) vyjádřen v procentech.

Jako ve všech ostatních případech, texty je nutné předzpracovat. Zde je nutné vyseparovat jednotlivá slova bez diakritiky. Ta se poté porovnávají a vyhodnocuje se míra shody výskytů stejných slov v rámci celého dokumentu.

Kosinová podobnost má velkou výhodu v případě, že plagiátor parafrázuje text a snaží se alespoň o minimální iniciativu. Pokud ale mění jen některá slova ve větách, detektor vyhodnotí jako vysoce nápadnou shodu i takový text a upozorní na plagiát. V uživatelském rozhraní se tento příznak nazývá kosinová podobnost.

6.7 Porovnávání velikosti textů

Porovnávání velikosti textů je metoda, která detekuje velikost souborů, porovná je mezi sebou a vyhodnotí. Samotný příznak nehledá texty, které mohou být plagiáty, ale pouze podezřelé shody ve velikosti.

Výhodou příznaku je jeho jednoduchost a rychlost vyhodnocení. Porovnávání velikosti textů má nevýhodu např. v detekování podezřelé shody i u dvou naprosto rozdílných dokumentů, kde je shodná velikost čistě náhodná. Jako samotný příznak pro detekci se použít nedá. V uživatelském rozhraní se tento příznak nazývá velikost souboru.

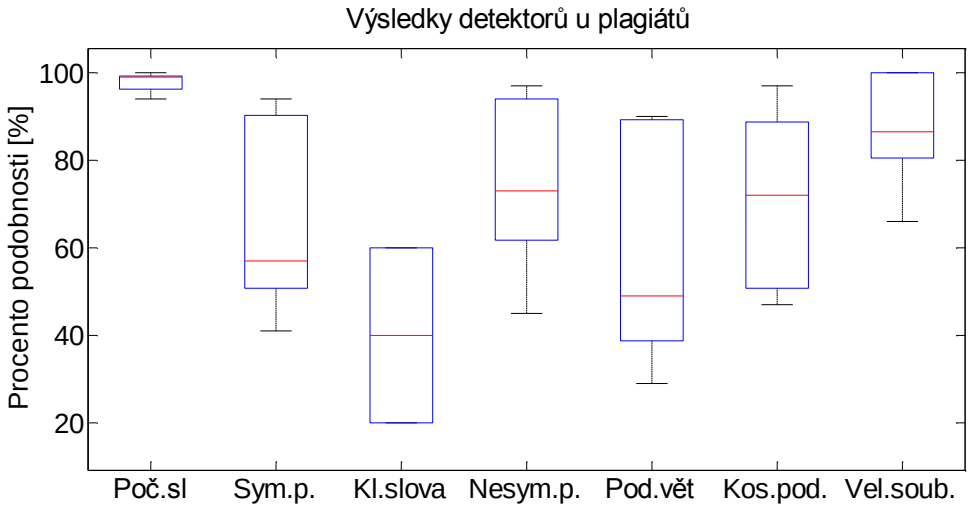
7 TESTOVÁNÍ PŘÍZNAKŮ

Pro testování příznaků byla použita data ze školní databáze studentských projektů. Dokumenty se svým rozsahem lišily v závislosti na zpracovávaném tématu. Jejich rozmezí se pohybovalo od 6-20 stran.

7.1 Výsledky testování

Testování odezvy příznaků bylo provedeno na třech typech dokumentů. Byly nalezeny a označeny texty, kde se jednalo o plagiáty, texty, které měly stejné téma, ale plagiáty nepředstavovaly a také texty s rozdílným tématem. Jako plagiát bylo označeno 18 dokumentů, tedy 9 originálů a 9 plagiátů těchto textů. Při testování příznaků s texty, které byly zpracovány na stejné téma, ale originálně, bylo použito také 18 dokumentů. Při testování příznaků u rozdílných textů byl počet také 18.

Chování příznaků u textů, které byly označeny jako plagiáty, zobrazuje obr. 7.1. Jedná se o příkaz boxplot vytvořený v prostředí Matlab. Pro takové statistické vyhodnocení bylo použito 18 dokumentů, které byly jako plagiát vyhodnoceny. Boxplot je krabicový diagram, který graficky vizualizuje numerická data získaná při testování. Pro tento případ byl použit boxplot, který vyznačuje minima a maxima dat a hodnoty mezi nimi. Horizontální červená linie v grafech značí medián [15]. Vzhledem k menší čitelnosti příznaků, je posloupnost popsána zde (zleva): počet slov, symetrická podobnost, klíčová slova, nesymetrická podobnost, podobnost vět, kosinová podobnost a velikost souboru. Všechny zobrazené boxploty mají stejnou posloupnost.



Obr. 7.1: Výsledky testování příznaků u plagiátů.

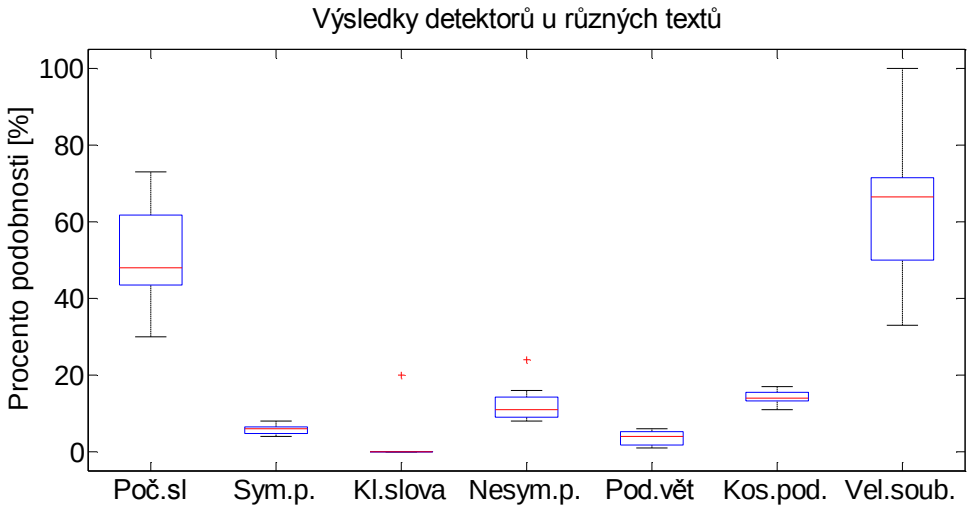
Výsledky z boxplotů ukazují, že rozptyl hodnot je velký. Vzhledem k tomu, že byla testována data, která jsou skutečnými plagiáty, a vždy se nejedná o naprosto stejnou kopii ve všech ohledech, je tento výsledek očekávaný. Pokud by byly dokumenty naprosto totožné, boxploty všech příznaků budou na sto procentech. Chování příznaků v případě, že se jedná o texty se stejným tématem, ale zpracované autory originálně, je zobrazeno na obr. 7.2.

Obr. 7.2: Výsledky testování příznaků u textů se stejným zadáním, které jsou originální.

Je patrné, že kvalita příznaku s názvem počet slov, který porovnává počet slov mezi dvěma dokumenty a procentuálně je vyhodnocuje, není jako samotný detektor použitelný, jelikož i u rozdílných dokumentů vykazuje vyšší rovnosti. Velikost souboru vykazuje také vyšší hodnoty podobnosti. Podobnost ve velikosti je ale také zcela náhodná a o plagiátu tedy tyto samotné funkce rozhodovat nemohou, jelikož je zde

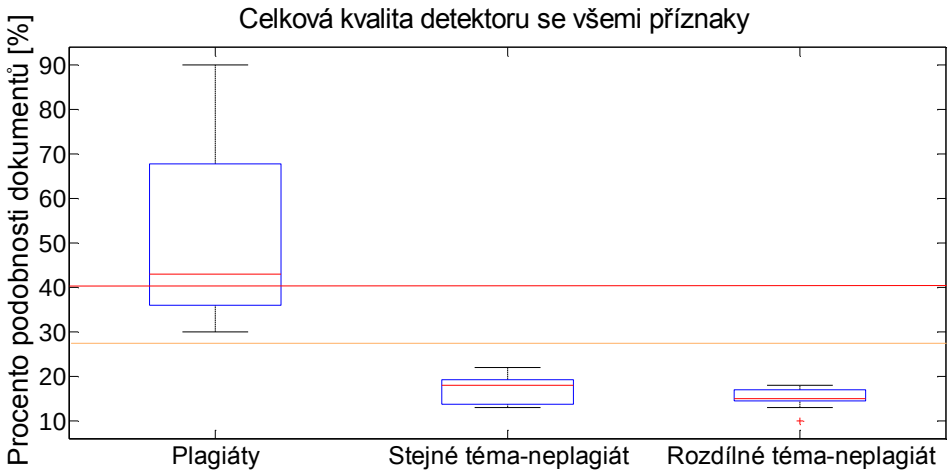
podobnost zcela náhodná. Naopak příznaky symetrická podobnost a klíčová slova nemají rozsah hodnot velký. Znamená to, že vyhodnocují shody v malém množství a na plagiáty neupozorňují. Nesymetrická podobnost, podobnost vět a kosinová podobnost jsou také kvalitními příznaky. Takové hodnoty byly u dokumentů, které jsou originální, očekávané.

Příznaky také prošly testováním u dokumentů, kde se shoda neočekávala, jelikož byly z témat naprosto odlišných. Získané výsledky jsou zobrazeny na obr. 7.3. Výsledky se příliš neliší od dokumentů, které jsou zpracovány na stejné téma, ale plagiáty nejsou. Potvrzuje se pouze zjištěná kvalita příznaků u předchozího testu.



Obr. 7.3: Výsledky z testování příznaků u dvou rozdílných dokumentů.

Při použití všech příznaků a vyhodnocení celkové podobnosti dokumentů pro zjištění kvality vytvořeného detektoru plagiátů vykazuje program výsledky zobrazené na obr. 7.4. Vlevo jsou výsledy z plagiátů, uprostřed jsou dokumenty, které mají stejné zadání, ale jsou originály a napravo jsou dokumenty s rozdílným zadáním. Oranžová linie vyznačuje rozsah, pod kterým detektor vyhodnotí dokumenty jako originály, mezi oranžovou a červenou linií je rozsah, kdy detektor vyhodnotí dokumenty jako nápadně si podobné a od červené linie k vyšším hodnotám jsou dokumenty vyhodnoceny jako plagiáty, jelikož vykazují nápadně vysokou podobnost.



Obr. 7.4: Výsledek detekování programu v závislosti na vložených datech.

Zobrazené výsledky v této kapitole potvrzují kvalitu příznaků symetrické a nesymetrické podobnosti, příznaku, který počítá shodnost nejpoužívanějších slov, zvaného klíčová slova, příznaku podobnost vět, který vykazuje kladné výsledky v případě, že se shodují celé věty a také kosinové podobnosti, která pracuje se slovy a poradí si tak se změnami slovosledu ve větách. Nízkou kvalitu mají příznaky počet slov a velikost souboru. Celkový výsledek vytvořeného programu naznačuje, že rozdíl při zkoumání plagiátů a originálních textů je značný a program je tedy poměrně přesný. Vždy je ale důležité, aby pro maximální jistotu, zdali se jedná o plagiát či nikoliv, již detekované výsledky vyhodnotil člověk.

7.2 Nastavení vah k příznakům

Pro správné počítání detektoru, který jako finální výsledek vypisuje celkovou podobnost dokumentů, je nutné nastavit váhy ke každému příznaku zvlášť na základě provedeného testování. Pro výpočet těchto vah byla použita metoda pořadí, která vyžaduje jen ordinální informaci stanovení pořadí jednotlivých příznaků dle spolehlivosti. Nejdůležitějšímu příznaku bylo přiřazeno číslo k (k= počet příznaků), druhému k-1 až k nejméně spolehlivému příznaku, kterému byla přiřazena důležitost 1. Poté je i-tému příznaku přiřazeno přirozené číslo b_i [17]. Váha v_i je poté vypočítána dle vzorce 7.1:

$$v_i = \frac{b_i}{\sum_{i=1}^{k} b_i}, \qquad \sum_{i=1}^{k} b_i = k\frac{(k+1)}{2}, \qquad pro\ i=1,2,\ldots k. \tag{7.1}$$

Stanovení důležitosti příznaků je zobrazeno v tabulce 7. 1. Vzhledem k tomu, že příznak nesymetrická podobnost vykazuje dva výsledky, je počítáno s osmi příznaky.

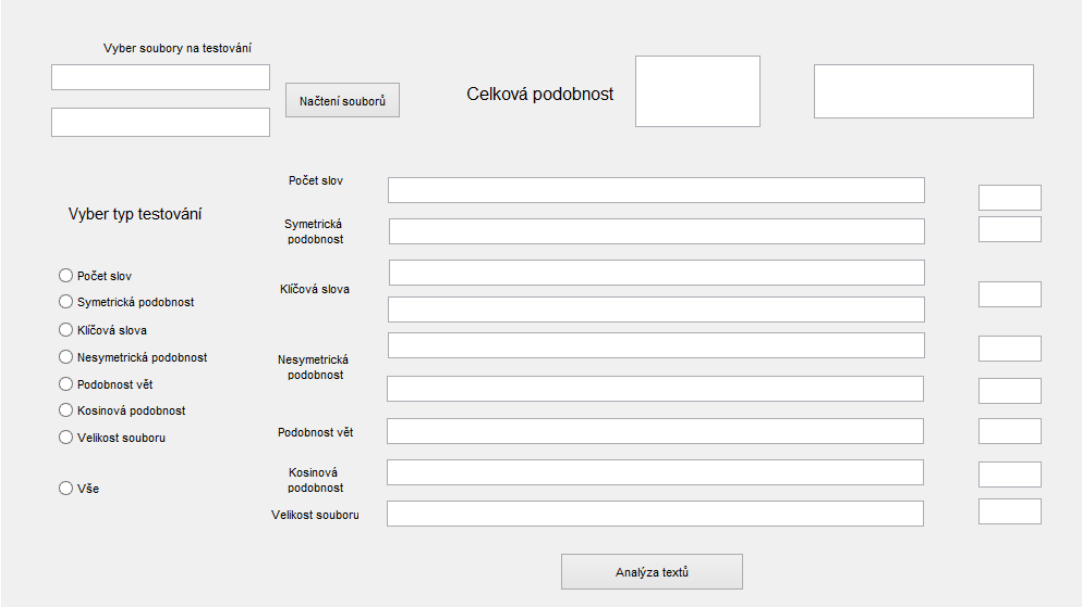
Tabulka 7. 1: Rozřazení příznaků dle jejich spolehlivosti a nastavení vah.

	Název příznaku							
	Počet slov	Sym. podobnost	Klíčová slova	Nesym. pod. 1	Nesym. Pod. 2	Podobnost vět	Kosinová podobnost	Velikost souboru
Pořadí	7	5	6	3	4	1	2	8
Hodn. b_i	2	4	3	6	5	8	7	1
Váha v_i	0,05	0,08	0,11	0,16	0,16	0,22	0,14	0,03

Na základě výsledků z boxplotů získaných v Matlabu byla sestavena tabulka s příznaky, které jsou více spolehlivé a méně spolehlivé. Pokud je výsledný boxplot u příznaku roztažený při testování dokumentů, které nejsou plagiáty, je vyhodnocen jako méně spolehlivý než takový, který má rozptyl menší. U testování plagiátů zase příznak, který má výsledek ve vyšších hodnotách a nepodléhá náhodnému detekování, je vyhodnocen jako spolehlivý. Tyto výsledky dopomohly k sestavení pořadí jednotlivých detektorů dle jejich spolehlivosti. Jako spolehlivé příznaky byly určeny „podobnost vět“, „kosinová podobnost“ nebo „nesymetrická podobnost“ vykazující dva výsledky. Naopak příznaky s názvem „velikost souboru“ či „počet slov“ byly vyhodnoceny jako méně spolehlivé, tudíž mají malý význam na celkovém výsledku, který informuje o procentuální podobnosti zkoumaných dokumentů.

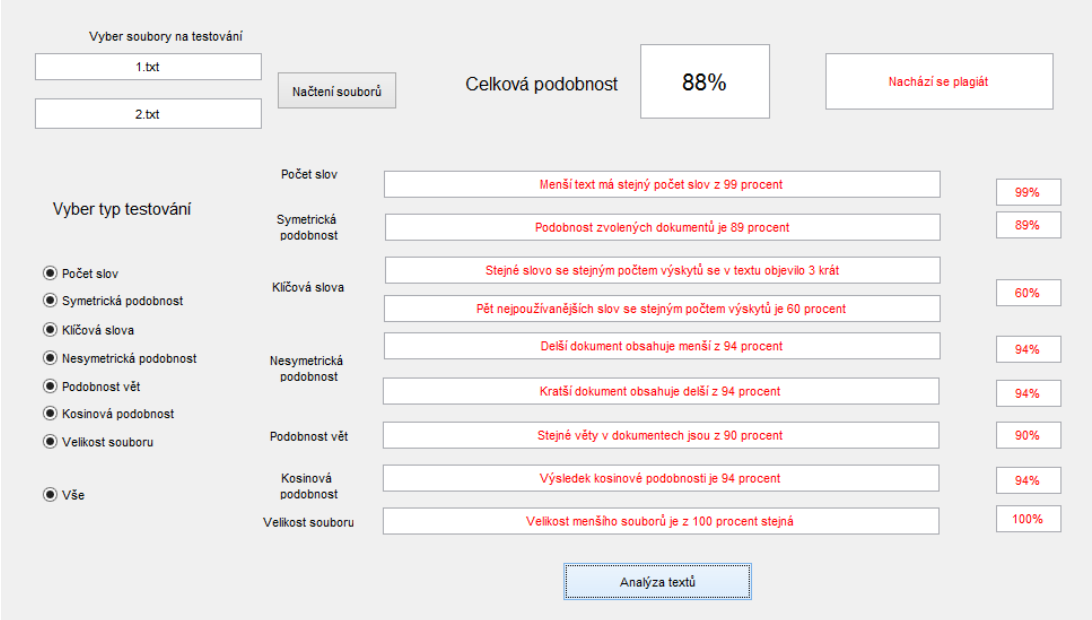
8 PROGRAM PRO DETEKCI PLAGIÁTŮ

V programu Matlab byly implementovány všechny zmíněné příznaky a jako konečná podoba programu bylo zvoleno grafické uživatelské rozhraní pro větší přehlednost a jednoduchost ovládání. Po spuštění programu je na uživateli, aby vybral soubory, které chce podrobit analýze, viz obr 8.1 vlevo nahoře. Po kliknutí na tlačítko s názvem „Načtení souborů“ se objeví okno pro výběr souborů. Data je nutné před nahráváním upravit do podoby textového editoru, jelikož Matlab s takovými daty umožňuje práci. V dalším kroku je na uživateli programu, aby si zvolil příznaky, pomocí kterých chce testovat dokumenty. Rychlejší výběr umožňuje kliknutí na možnost „Vše“, která zvolí všechna nabízená pole, viz obr 8.1 vlevo. Takto vybrané texty se již mohou porovnávat, k čemuž slouží tlačítko s názvem „Analýza textů“, jenž se nachází v dolní části programové obrazovky. Po dokončení analýzy, která se vykoná automaticky, si uživatel může přečíst, jak jednotlivé příznaky vyhodnotily podobu dokumentů. Každý příznak na svém řádku obsahuje větu s celkovým vyhodnocením, za níž se nachází další vyhodnocovací okno, kde lze nalézt pouze procentuální podobnost dokumentů pro rychlejší orientaci. Na každém řádku se navíc nachází práh, individuálně zvolený na základě provedeného testování, který při podezřele vysoké podobnosti zabarví text do červena. V horní části uživatelského prostředí se nachází výsledné vyhodnocovací okno, obsahující celkovou procentuální podobnost dvou dokumentů, která se počítá z dílčích výpočtů. K celkovému vyhodnocení ještě přispívá textová zpráva, která v závislosti na podobnosti vypíše oznámení o podezření na plagiát. Vzhled programu lze vidět na obr. 8.1.



Obr. 8.1: Konečná verze programu před načtením dokumentů.

V případě, že texty jsou si podobné z méně než 28 %, se v kolonce objeví: „Plagiát se nenachází“ a text je zabarven do zelena. V rozmezí celkové podobnosti od 29 % - 40 % je možné vidět vyhodnocení s textem: „Texty jsou si nápadně podobné“ s oranžovým zabarvením, a při vyšší podobnosti dokumentů program vyhodnotí testované soubory jako plagiáty textem: „Nachází se plagiát“, který je červený. Vzhled programu po vyhodnocení dokumentů jako plagiáty ilustruje obr. 8.2.



Obr. 8. 2: Konečná podoba programu po vyhodnocení dokumentů. Plagiát je nalezen s celkovou shodou dokumentů 88 %.

Program, který je použitelný pro detekování podezřele si podobných dokumentů, vychází z teoretické předlohy popsané v kapitolách 4 a 5. Jednotlivé funkce byly v Matlabu vytvořeny jako vlastní funkce, které se v hlavním programu postupně volají v případě jejich zvolení. Po provedení analýzy je ke každému příznaku vypsána poznámka o podobnosti. Každý příznak má nastavený limit, kde se změní barva v případě, je-li tohoto prahu dosaženo. Nastavení velikosti prahu bylo zvoleno na základě testování, a to tak, aby se podezřelé texty vždy detekovaly. Konkrétní hodnoty lze vidět v Tab. 8. 1.

Tab. 8. 1: Tabulka obsahuje informace o nastaveném prahu pro jednotlivé příznaky

Název příznaku	**Nastavený limit pro detekci podobnosti dokumentů**
Podobnost vět	25 %
Procento shody nejpoužívanějších slov	40 %
Porovnávání velikostí dokumentů	95 %
Shoda počtu slov	90 %
Kosinová podobnost	45 %
Symetrická podobnost	40 %
Nesymetrická podobnost	40 %

Slabší příznaky, kde je možná shoda náhodná, konkrétně příznaky s názvy „Počet slov“ a „Velikost souboru“ mají prahy vysoké, aby uživatele informovaly o podobnosti v případě, kdy je podobnost skutečně podezřelá. Nejnižší práh má naopak příznak „Podobnost vět“, jelikož se jedná o příznak, který vykazuje shodu v případě, kdy jsou věty v dokumentech naprosto přesně stejné. V případě, že u dvou dokumentů je shoda vět z 25 %, jedná se o vysokou shodu, která se u originálních dokumentů nevyskytuje.

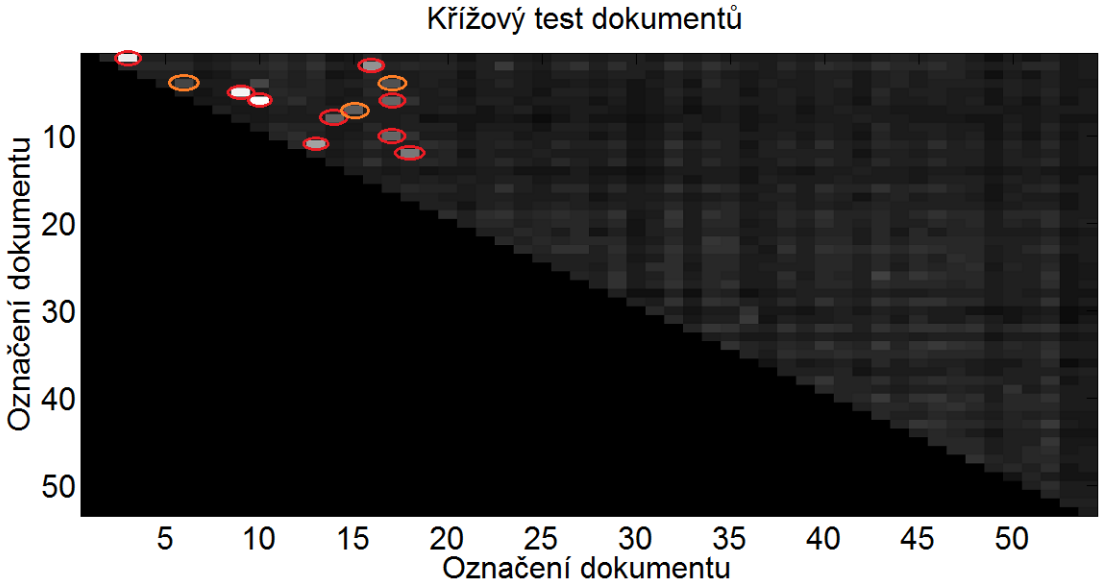
Ukázka nalezeného plagiátu, který byl detektorem zachycen, je v tabulce 8. 2. Jedná se o originální text z dokumentu, který nebyl žádným způsobem upraven. Červený text značí shodná slova v obou dokumentech.

Tab. 8.2: Úryvek z detekovaného dokumentu.

Dokument 1	**Dokument 2**
Program obsahuje celkem tři na sebe navazující funkce. Funkce batoh kontroluje hodnoty vstupních parametrů a používá další dvě funkce k řešení problému. Funkce pocitej_jakost se vrací výstupní vektor jakost. Funkce reseni_problemu pracuje s maticemi. Seřazováním a třízením jednotlivých prvků matice se snaží najít optimální řešení problému. Jednotlivé funkce jsou popsány v kapitolách.	Program obsahuje 3 funkce, které jsou mezi sebou provázány. První z nich funkce batoh kontroluje hodnoty vstupních parametrů a využívá zbylé dvě funkce k řešení problému. Funkce pocitej_jakost vrací výstupní vektor jakost. Poslední funkce reseni_problemu pracuje s maticemi. Seřazováním a třízením jednotlivých prvků matice se snaží najít optimální řešení problému. Podrobnější popis jednotlivých funkcí bude popsán v dalších kapitolách.

8.1 Křížové porovnávání dokumentů

Po vytvoření detektoru bylo možné testovat dokumenty i jiným způsobem a to křížovým porovnáváním dokumentů mezi sebou. Data, která byla k testování použita, obsahovala velké množství dokumentů, které mohly obsahovat další plagiáty. Pro testování bylo nutné upravit dokumenty do textové podoby, ukončené koncovkou txt. Takto upravené byly zpracovány v Matlabu. Křížovému porovnání bylo podrobeno 54 dokumentů, mezi kterými bylo nalezeno 12 si podobných textů. Obrázek 8. 3. obsahuje výsledek vygenerovaný po testování. Světlejší hodnoty vykazují vyšší shodnost dokumentů, nežli tmavší. Červeně vyznačené body jsou ty, které svou podobností přesáhly práh spadající do plagiátů, který je nastavený na 41 %. Oranžové body jsou ty, které detektor vyhodnotí jako nápadně si podobné. Svou podobností jsou v rozmezí od 28 % do 41 %.



Obr. 8. 3: Křížové porovnávání dokumentů.

Křížovým porovnáváním bylo prozkoumáno 54 dokumentů, mezi kterými bylo nalezeno 12 podezřele si podobných textů. Po prozkoumání těchto dokumentů bylo přímo jako plagiát označeno 9 z nich, zbylé detekuje detektor jako nápadně si podobné. V tomto případě se jednalo o plagiáty ve všech případech, některé jsou ale kopií druhého textu z menší části. Detektor tedy zachytil plagiáty ve 100 % případů a přímo jako plagiát vyhodnotil 75 % z nich.

9 ZÁVĚR

Diplomová práce se zabývá plagiátorstvím v textovém dokumentu. Seznamuje s jeho definicí a také popisuje jednotlivé typy plagiátorství. Vysvětluje, na jakém principu mohou pracovat vyhledávací softwary tvůrců. Není však možné zjistit, na jakém principu skutečně softwary pracují, jelikož je to know how každého tvůrce. Tímto se chrání před dalším prolomením detektorů, aby nebylo jednoduché se těmto metodám vyhnout a dopouštět se vědomě plagiátorství. Ukázky principů uvedených softwarů mohou naznačit způsob porovnávání a zjišťování podezřelých textů.

Kapitola předzpracování dat se zabývá způsoby úprav dat, které mohou dopomoci k rychlejší práci počítače a k jednoduššímu připravení dat pro detekci. Způsob předzpracování dat je pro detekci velmi důležitý, jelikož se mimo usnadnění výpočetní náročnosti nastaví také základní prvek pro porovnání, což je při detekci naprosto zásadní. Mezi důležité příznaky, popsané ve čtvrté kapitole, v této práci jistě patří kosinová podobnost a symetrická a nesymetrická podobnost, které jsou velice silnými příznaky a jsou schopny si poradit i s pracemi, které se snaží detektoru plagiátu vyhnout způsobem změny slovosledu či vybraných slov. Pokud by se tyto detektory použily na vybrané části textu, kde se spíše očekávají opsané řádky, tyto příznaky budou ještě silnější. Nebudou totiž ovlivněny slovy, kde i plagiátor musí být originální, například u osobních údajů nebo v úvodu, či závěru, kde vyjadřuje vlastní názor či hodnocení.

Kapitola následující představuje všechny příznaky, které byly použity při tvorbě programu, aby byl vytvořen co nejspolehlivější detektor. Popisuje jejich charakteristiku, jak fungují a jaké mají výhody a nevýhody.

V sedmé kapitole je představen samotný program a popis jeho ovládání. Program je navržen tak, aby byl intuitivní a aby uživatel byl schopen program ovládat i bez přečtení dokumentace.

Následující kapitola se věnuje testování a statistickému zhodnocení všech příznaků, ze kterých je patrné, které příznaky jsou kvalitní a tedy jejich výsledky lze brát vážně. Testování bylo prováděno na dokumentech, které byly prokázanými plagiáty, i na těch, které byly zpracovány na stejné téma, ale o plagiáty se nejednalo. Otestovány byly také dokumenty, které neměly společné téma. Příznaky zjišťující počet slov a porovnávající velikost souboru vyšly z testů nejhůře, jelikož detekovaly shody i u originálních textů. Naopak příznaky zjišťující podobnost vět, výskyt pěti nejpoužívanějších slov, symetrickou a nesymetrickou podobnost a kosinovou podobnost, které pracují se slovy, vycházely jako příznaky silné. Po porovnání výsledků celkové podobnosti, která je

počítána z jednotlivých příznaků, které jsou naváhovány dle statistických vyhodnocení vychází, že detektor je poměrně přesný. V příloze je poté ukázka více možných výsledků detektoru.

Výsledky detektoru jsou dle testování dobré a spolehlivé. Program má ale i nevýhody. Bohužel se nelze vyhnout problémům se zjišťováním, který z podobných dokumentů je originál a který je opsaný. Také nelze zachytit podobu u dokumentů, které jsou například opsané z cizího jazyka. Nevýhoda vytvořeného programu je bohužel nutnost data před testováním upravit. Tu je ale možné řešit stažením programu pro konvertování dokumentů, který jednoznačně urychlí převody do textového formátu.

LITERATURA

[1] *Infogram: Portál pro podporu informační gramotnosti* [online]. 2014 [cit. 2014-10-15]. Dostupné z: http://www.infogram.cz/findInSection.do?sectionId=1115&categoryId=1165

[2] *Národní knihovna* [online]. 2014 [cit. 2014-10-15]. Dostupné z:http://aleph.nkp.cz/F/PK2GXX1DPE1HRMYEHKMP3BPE9X9H4RSNX5J6B5RT1VVE7P31EB-01587?func=find-b&find_code=WTD&x=0&y=0&request=plagi%C3%A1t&adjacent=N-

[3] NĚMEČKOVÁ, Lenka. *Plagiátorství* [online]. Ústřední knihovna ČVUT, 2009 [cit. 2014-12-31]. Dostupné z:http://knihovna.cvut.cz//administrace/upload_dir/files/92d3b80c1ab35ca5a6cba67ff8dceaf9c9931380.pdf. ČVUT.

[4] PŘIBIL, Jiří. *Efektivní metody detekce plagiátů v rozsáhlých dokumentových skladech* [online]. Jindřichův Hradec, 2010 [cit. 2014-12-31]. Dostupné z: https://isis.vse.cz/lide/clovek.pl?zalozka=7;id=56906;studium. Doktorská dizertační práce. Vysoká škola ekonomická v Praze. Vedoucí práce Prof. Radim Jiroušek, DrSc.

[5] Keyword extraction in open-domain multilingual textual resources. *Keyword extraction in open-domain multilingual textual resources* [online]. 2005, č. 1 [cit. 2014-12-30]. DOI: 10.1109/AXMEDIS.2005.29. Dostupné z:http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=1592096&url=http%3A%2F%2Fieeexplore.ieee.org%2Fstamp%2Fstamp.jsp%3Ftp%3D%26arnumber%3D1592096

[6] Waseda University. *Waseda University* [online]. 2014 [cit. 2014-11-12]. Dostupné z:http://www.waseda.jp/navi/services/system/doc/ithenticate05l.png

[7] Reviewers Update. *Reviewers Update* [online]. 2013 [cit. 2014-11-12]. Dostupné z:http://www.elsevier.com/reviewers/reviewers-update/how-crosscheck-can-combat-the-perils-of-plagiarism

[8] University of Wolverhampton. *University of Wolverhampton* [online]. 2009 [cit. 2014-11-12]. Dostupné z:http://www2.wlv.ac.uk/celt/turnitin/quotations.jpg

[9] Ad (Words and Sense). *Ad (Words and Sense)* [online]. 2005 [cit. 2014-11-12]. Dostupné z:http://www.adwordsadsensetools.com/IMG/cache-500x377/2_CCCopyCatchSent-500x377.gif?

[10] The Bedford/ St. Martin's workshop on plagiarism. *St. Martin's workshop on plagiarism* [online]. 2002 [cit. 2014-12-31]. Dostupné z: http://bcs.bedfordstmartins.com/plagiarism/content/cat_040/scrnshot/evetest.jpg

[11] Softonic. *Softonic* [online]. 2002 [cit. 2014-11-12]. Dostupné z:http://screenshots.en.sftcdn.net/en/scrn/69655000/69655151/scr_1406222392-632x535.jpg

[12] Viper: The Anti-plagiarism scanner. *Viper: The Anti-plagiarism scanner* [online]. 2014 [cit. 2014-11-12]. Dostupné z:http://www.scanmyessay.com/images/scan-example1.gif

[13] HAUZÍREK, 2007. *Možnosti automatické detekce plagiátů* [online]. V Praze [cit. 2015-05-11]. Dostupné z: http://trochu.kvalitne.cz/diplomka/text/DipText4.xhtml#toc35. Diplomová práce. Vysoká škola ekonomická v Praze.

[14] FLÉGL, Jan. 2009. *Sofistikované metody pro kontrolu elektronických textů* [online]. Brno [cit. 2015-05-11]. Dostupné z: https://www.vutbr.cz/www_base/zav_prace_soubor_verejne.php?file_id=18302. Bakalářská práce. Vysoké učení technické v Brně.

[15] Robert McGill, John W. Tukey and Wayne A. Larsen *The American Statistician* Vol. 32, No. 1 (Feb., 1978), pp. 12-16. Dostupné z: http://www.jstor.org/discover/10.2307/2683468?uid=3737856&uid=2134&uid=2&uid=70&uid=4&sid=21106823515133

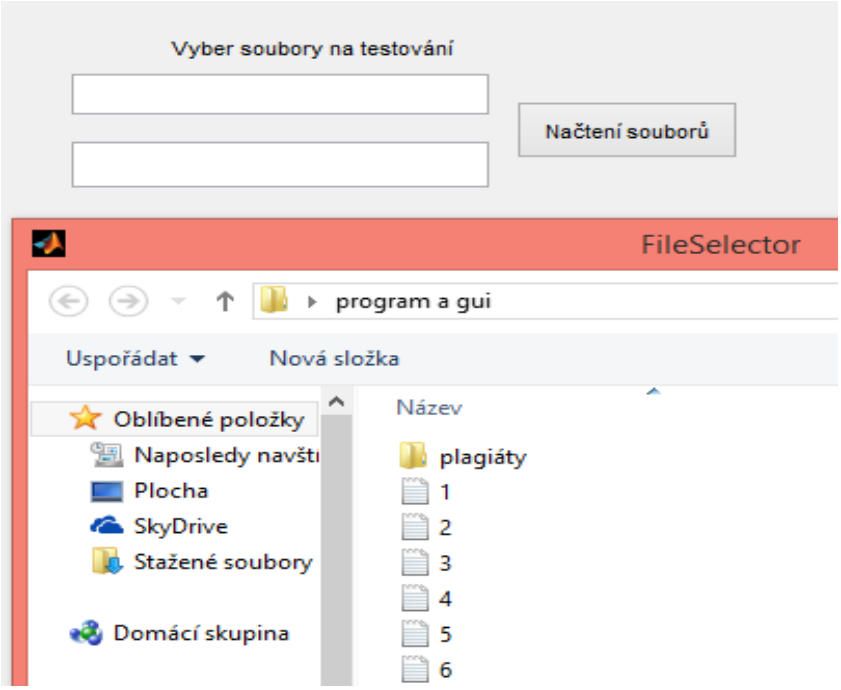
[16] MATHWORKS. *MathWorks* [online]. [cit. 2015-04-1]. Dostupné z: http://www.mathworks.com/?refresh=true

[17] *Rozhodovací procesy* [online]. [cit. 2015-05-18]. Dostupné z: http://www.rozhodovaciprocesy.cz/vicekriterialni-rozhodovani/2-1-metody-stanoveni-vah-kriterii.html

SEZNAM PŘÍLOH

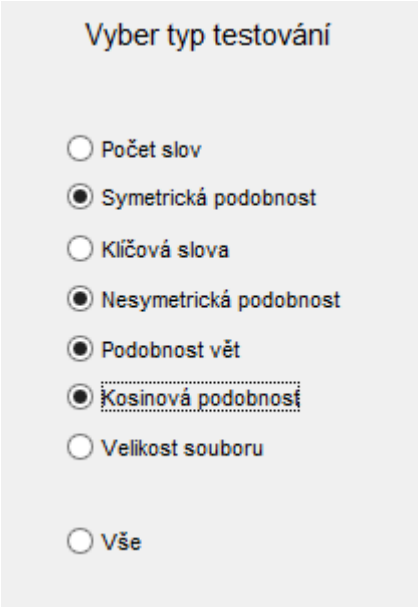
A UKÁZKY Z PROGRAMU

Po převedení dokumentů do txt. formy je možné data nahrát do programu. Po kliknutí na tlačítko „Načtení souborů“ se objeví file selector, který vybídne uživatele k výběru dokumentů, viz obr. A. 1.



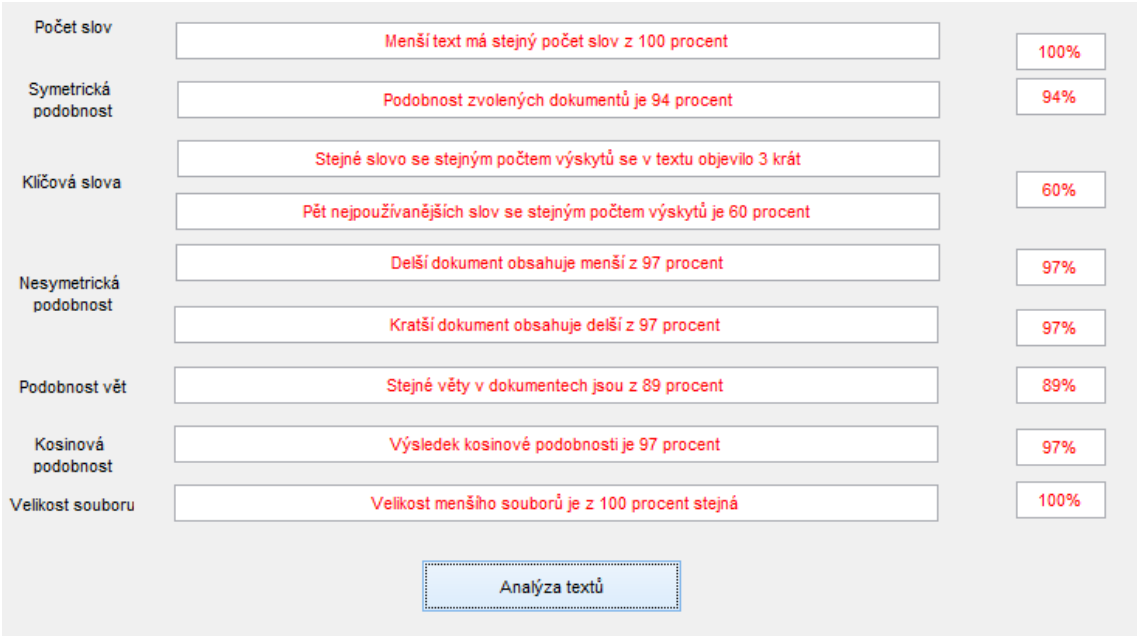
Obr. A. 1. Výřez z programu. Část, kde je možné nahrát soubory do vytvořeného uživatelského prostředí s file selectorem.

Poté si uživatel vybere příznaky, kterými chce dokumenty testovat. Pokud chce vybrat všechny nabízené, lze zvolit rychlejší volbu po zakliknutí okna u textu „Vše“. Popsaný postup je na obr A. 2.



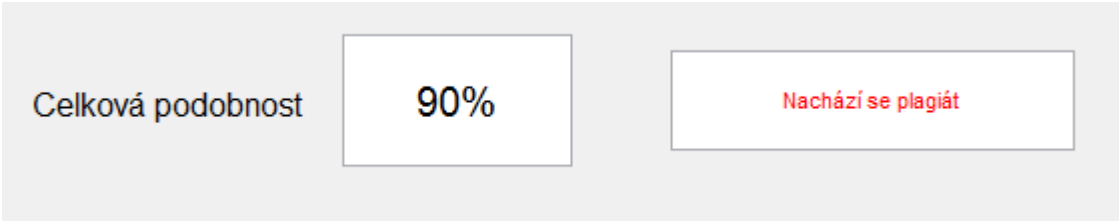
Obr. A. 2. Možnosti výběru příznaků pro analýzu.

Pro start programu slouží tlačítko „Analýza textů“ uprostřed dole na obr. A. 3. Vyhodnocování programu trvá od 3-10 sekund, poté se vypíší výsledky. V případě, že se detekují plagiáty, vypadá výsledek jako na obr. A. 3.



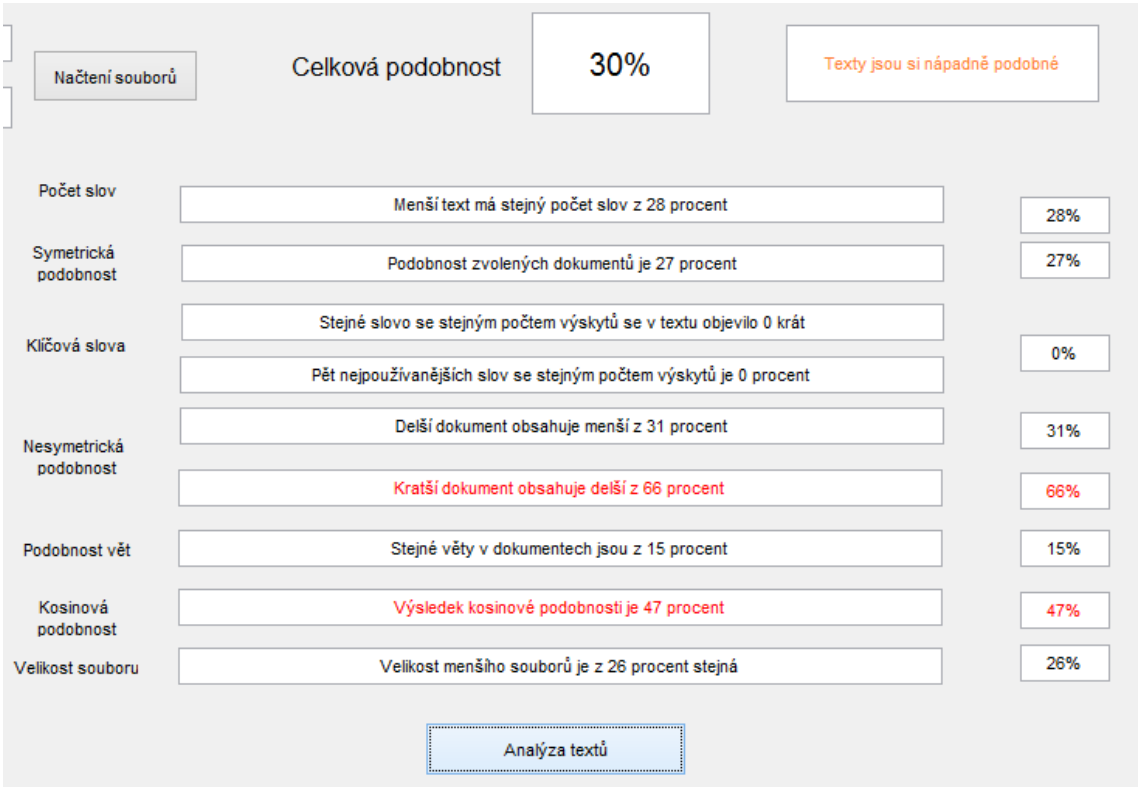
Obr. A. 3. Výsledky získané analýzou dokumentů.

Vše ještě shrnuje text s celkovým výsledkem a procentuální podobností, viz obr. A. 4.



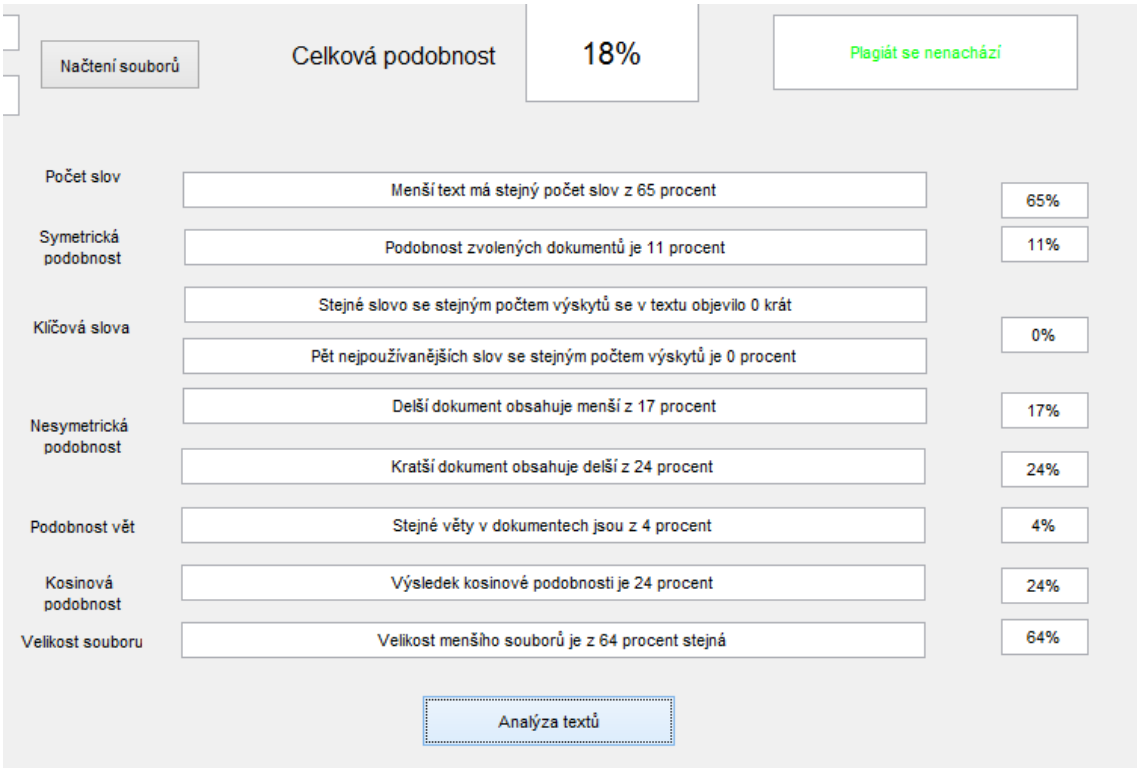
Obr. A. 4 Shrnutí celkové podobnosti dokumentů vzhledem ke zvoleným příznakům podpořený textem, který uživatele upozorní na nacházející se plagiát.

Výsledek, který vykazuje program po vyhodnocení méně si podobných dokumentů, vypadá jako na obrázku A. 5.



Obr. A. 5. Zachycuje vysokou podobnost u příznaků „Nesymetrická podobnost“ a „Kosinová podobnost“, které poukazují na to, že jeden ze zvolených textů je kopie druhého. Výsledné hlášení je, že texty jsou si nápadně podobné.

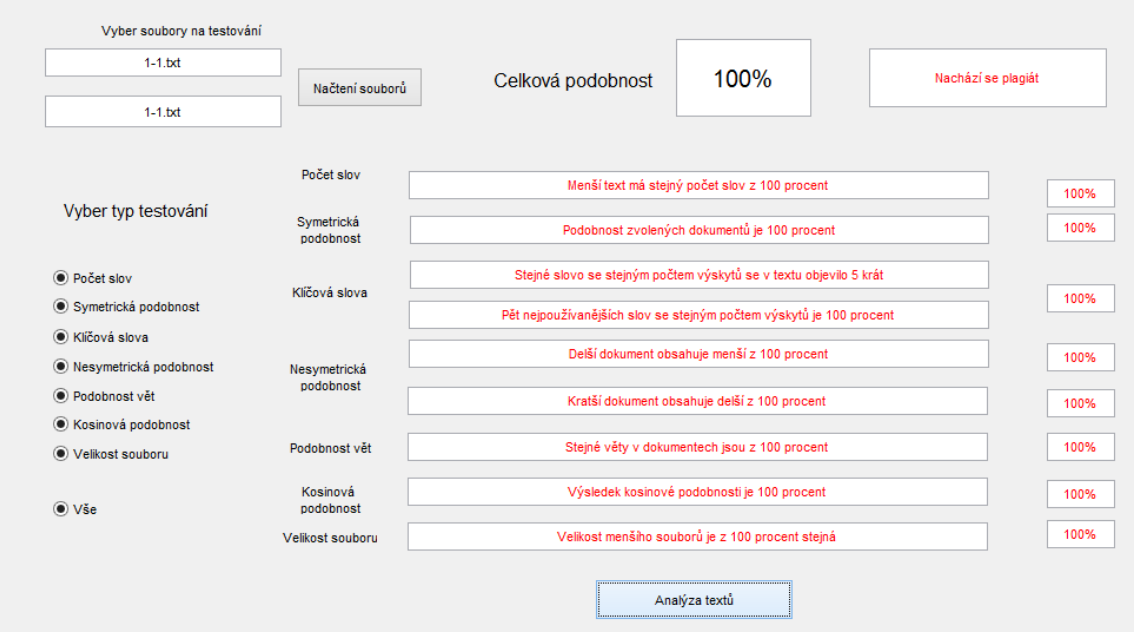
V případě, že texty nejsou plagiáty, program vypadá jako na obrázku A. 6.



Obr. A. 6. Verze výsledků v případě, že se o plagiáty nejedná.

Žádný z příznaků nedetekoval vyšší podobnost, než je zvolený práh a celková podobnost je také menší, než zvolený práh, proto je výsledné hlášení programu: „Plagiát se nenachází“.

Za předpokladu, že se do programového prostředí nahrají naprosto stejné dokumenty, všechny použité detektory vykážou shodnost v 100% a každé výsledné oznámení o shodě bude červené, viz obrázek A. 7.



Obr. A. 7. Podoba programu při detekci naprosto totožných dokumentů. Podobnost je vyhodnocena jako 100%.