

Útoky na detekcie plagiátorstva

Autor: Ivan Gulis

Obsah prezentácie

- Ciele práce
- Útoky
- Metódy detekcie
- Úvodný experiment
- Návrh riešenia
- Implementácia
- Overenie riešenia - LCS substring
- Overenie riešenia - detekcie útokov
- Zhodnotenie
- Ciele do budúcnosti
- Otázky a diskusia



Obrázok 1

Ciele práce

- **analyzovať** jednotlivé metódy plagiátorských útokov a zistiť ich slabiny
- útoky aj metódy detekcie **kategorizovať**
- **navrhnúť** techniky na ochranu pred útokmi
- **vykonať experiment** s existujúcimi nástrojmi na detekciu
- navrhnúť a **implementovať** detekčné mechanizmy proti úspešným útokom
- **experimentálne preveriť** správnosť riešenia

Útoky

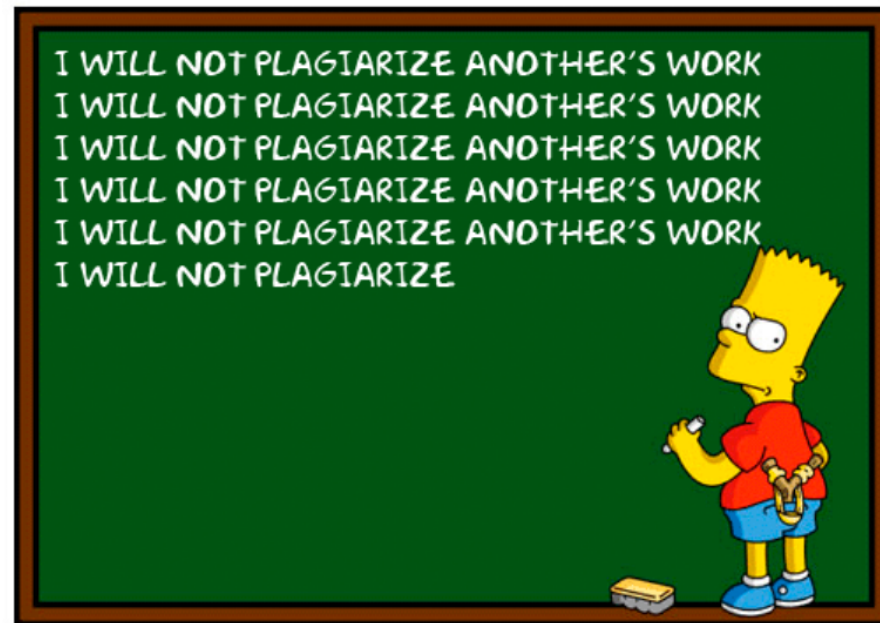
- synonymami
- zmena slovosledu
- zámena čísel za text
- homoglyfy
- biele znaky za medzery
- obrázky miesto textu
- využitie PDF vrstiev



Obrázok 2

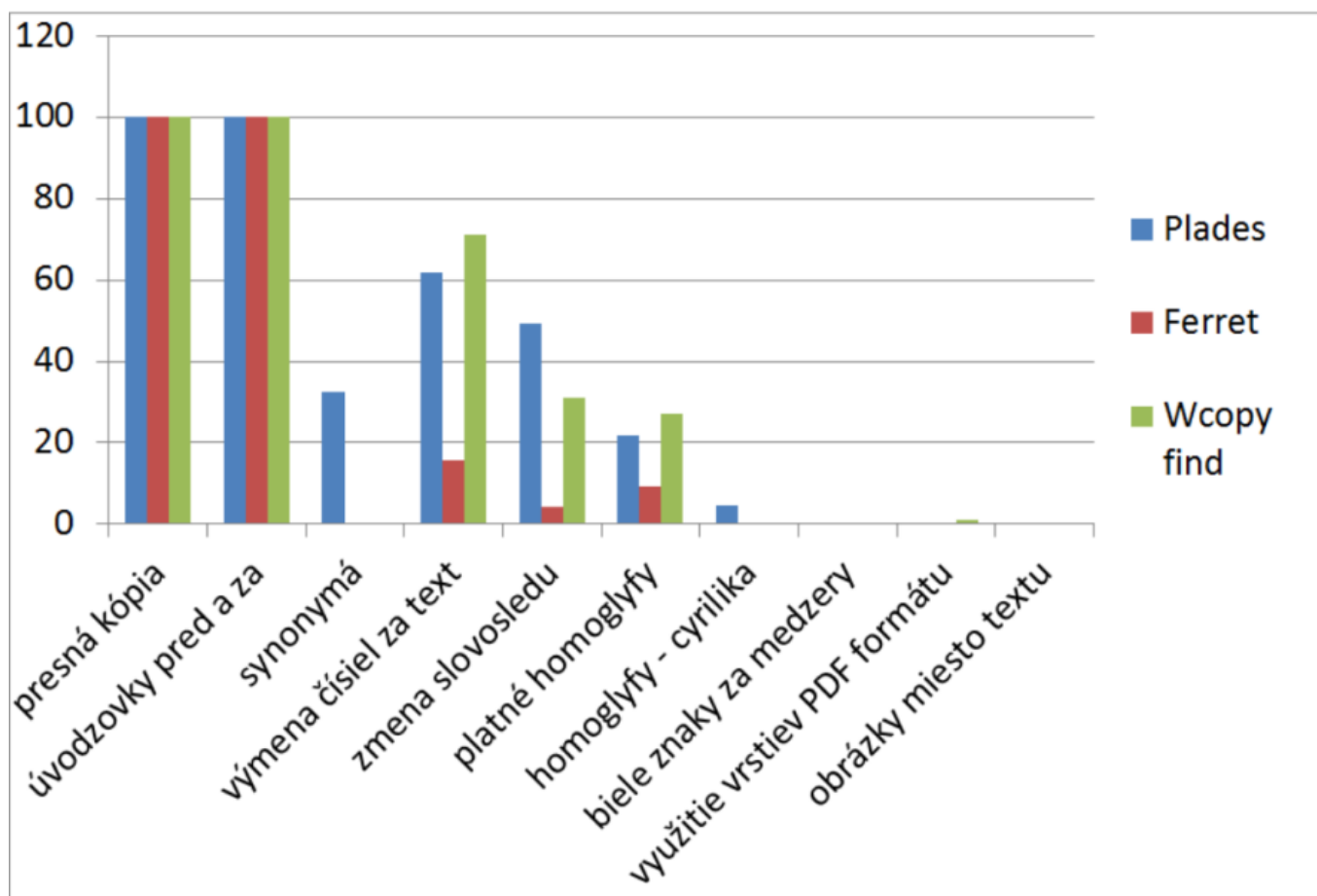
Metódy detekcie

- N-gramy
- LCS substring
- LCS subsequence
- Greedy-String-Tiling
- Levensteinova vzdialenosť
- Cosine similarity
- TF-IDF
- Metadáta



Obrázok 3

Úvodný experiment



Návrh riešenia

- Detekcia počtom podozrivých znakov
- Detekcia frekvenčnou analýzou
- Detekcia priemernou dĺžkou slova
- Detekcia analýzou objemu dát



Obrázok 5

Implementácia

- Vylepšenie LCS substring
- Tvorba štatistík *slov, znakov, objemov dát*
- Spracovanie štatistík



Obrázok 6

Overenie riešenia - LCS substring

PlaDes	LCS (1)	LCS (10)	LCS (100)	LCS (1000)
synonymá	3,67	19,77	52,37	53,61
výmena čísiel za text	2,14	15,72	58,45	59,31
zmena slovosledu	2,58	16,16	28,34	28,34
homoglyfy i za všetky L	1,08	8,13	37,84	41,59
homoglyfy i za všetky L, cl - d, rn - m	0,45	3,47	10,63	10,63
homoglyfy - cirilika (50% a,e)	0,97	3,92	16,25	16,25
homoglyfy - cirilika (100% a,e)	0,23	1,99	2,84	2,84
homoglyfy - cirilika (100% 8 pismen)	0,24	0,24	0,24	0,24
biele znaky za medzery (50% L)	7,1	9,68	9,68	9,68
biele znaky za medzery (100% L)	0,47	0,47	0,47	0,47
PDF vrstvy	0,74	1,3	1,3	1,3
obrázky	0	0	0	0

Tabuľka 1

Overenie riešenia - detekcie útokov

PlaDes	LCS (100)	Početom podozrivých znakov	Frekvenčnou analýzou	Priemernou dĺžkou slova	Analýzou objemu dát
homoglyfy i za všetky L	37,84	nie	áno	nie	nie
homoglyfy - cyrilika (50% a,e)	16,25	áno	nie	áno	nie
biele znaky za medzery (50% L)	9,68	nie	nie	áno	nie
biele znaky za medzery (100% L)	0,47	nie	áno	áno	nie
PDF vrstvy	1,3	nie	nie	nie	áno
obrázky	0	nie	nie	nie	áno






Tabuľka 2

Zhodnotenie

ZhodnotenieV prvej polovici práce sme **analyzovali známe metódy útokov na detekcie plagiátorstva**, popísali sme detekčné metódy, a experimentovali sme s aplikáciami *PlaDes*, *Ferret*, *WCopyFind*. V druhej polovici sme **navrhli riešenia nevyriešených problémov**, navrhli a implementovali detekcie a vylepšili sme metódu LCS - substring. Implementované riešenie sme overili.

Ciele do budúcnosti

- Lepšie určenie hraničných hodnôt
- Zrýchlenie spracovania .doc
- Zapojenie OCR
- Poškodené a zaheslované dokumenty
- Podozrivé znaky už z originálu dokumentu

Original	Thresholded	OCR
	66htv	66htv
	5n7pf	5n7pf
	qv s xp	qv s xp
	6x94d	6x94d
	jh78q	jmsq

Obrázok 7

Otázky a diskusia

Domov	VÝSLEDKY PODOBNOSTI DOKUMENTOV						
Spracovanie	Projekt Export podobnosti Výsledky Štatistika						
Výsledky	Hľadať dokument						
Nastavenie	Označ	Dokument 1	Typ	Dokument 2	Typ	Podobnosť v %	Porovnanie dokumentov
Odoslať názor	<input type="checkbox"/>	v006_MSI2007_okresa_plagiatcopy.doc		v005_MSI2007_okresa_original.doc		99,47	Zobraziť
Pomocník	<input type="checkbox"/>	v005_MSI2007_okresa_original.doc		v006_MSI2007_okresa_plagiatcopy.doc		99,39	Zobraziť
O aplikácii	<input type="checkbox"/>	v011_MSI2008_komorovsky_plagiatmodify.doc		v010_MSI2008_komorovsky_original.doc		96,72	Zobraziť
Zatvoriť	<input type="checkbox"/>	v010_MSI2008_komorovsky_original.doc		v011_MSI2008_komorovsky_plagiatmodify.doc		96,30	Zobraziť
	<input type="checkbox"/>	v009_MSI2008_jastrzemska_plagiatweb.doc		v008_MSI2008_jastrzemska_original.doc		52,59	Zobraziť
	<input type="checkbox"/>	v008_MSI2008_jastrzemska_original.doc		v009_MSI2008_jastrzemska_plagiatweb.doc		34,55	Zobraziť
	<input type="checkbox"/>	v011_MSI2008_komorovsky_plagiatmodify.doc		v006_MSI2007_okresa_plagiatcopy.doc		2,84	Zobraziť
	<input type="checkbox"/>	v011_MSI2008_komorovsky_plagiatmodify.doc		v005_MSI2007_okresa_original.doc		2,84	Zobraziť
	<input type="checkbox"/>	v010_MSI2008_komorovsky_original.doc		v006_MSI2007_okresa_plagiatcopy.doc		2,82	Zobraziť
	<input type="checkbox"/>	v010_MSI2008_komorovsky_original.doc		v005_MSI2007_okresa_original.doc		2,82	Zobraziť
	<input type="checkbox"/>	v006_MSI2007_okresa_plagiatcopy.doc		v011_MSI2008_komorovsky_plagiatmodify.doc		2,29	Zobraziť
	<input type="checkbox"/>	v006_MSI2007_okresa_plagiatcopy.doc		v010_MSI2008_komorovsky_original.doc		2,29	Zobraziť
	<input type="checkbox"/>	v005_MSI2007_okresa_original.doc		v011_MSI2008_komorovsky_plagiatmodify.doc		2,28	Zobraziť
	<input type="checkbox"/>	v005_MSI2007_okresa_original.doc		v010_MSI2008_komorovsky_original.doc		2,28	Zobraziť
	<input type="checkbox"/>	v009_MSI2008_jastrzemska_plagiatweb.doc		v007_MSI2008_aufriecht_original.doc		1,32	Zobraziť
	<input type="checkbox"/>	v009_MSI2008_jastrzemska_plagiatweb.doc		v011_MSI2008_komorovsky_plagiatmodify.doc		1,22	Zobraziť
	<input type="checkbox"/>	v009_MSI2008_jastrzemska_plagiatweb.doc		v010_MSI2008_komorovsky_original.doc		1,12	Zobraziť
	<input type="checkbox"/>	v009_MSI2008_jastrzemska_plagiatweb.doc		v006_MSI2007_okresa_plagiatcopy.doc		1,12	Zobraziť
	<input type="checkbox"/>	v009_MSI2008_jastrzemska_plagiatweb.doc		v005_MSI2007_okresa_original.doc		1,12	Zobraziť
	<input type="checkbox"/>	v009_MSI2008_jastrzemska_plagiatweb.doc		v012_MSI2008_varga_original.doc		1,02	Zobraziť
	<input type="checkbox"/> Označiť všetko Min. podobnosť (%) 0						
	ZOZNAM GRAF						

Obrázok 8

Zdroje

- 1. <http://www.quickmeme.com/img/b1/b1e9be76f623d290c09378a71114144ebef59a36f68586db2f95bd3718f3d24b.jpg>
- 2. <https://pbs.twimg.com/media/BNr6c8LCUAAmeY1.png>
- 3. <https://brendatobias.files.wordpress.com/2012/09/bart-simpson-plagiarize.png>
- 4. Vlastný obrázok
- 5. <http://elearningindustry.com/wp-content/uploads/2013/11/Top-10-FREE-Plagiarism-Detection-Tools-For-Teachers-1024x1024.jpg>
- 6. <http://s3.amazonaws.com/libapps/accounts/54646/images/statistics.png>
- 7. <https://ahm3dibrahim.files.wordpress.com/2011/06/dhiraagucaptchaocr.jpg>
- 8. <http://www2.fiit.stuba.sk/~chuda/plagiarism/img/zoznam.png>