Универзитет „Св. Кирил и Методиј“ во Скопје
ФАКУЛТЕТ ЗА ИНФОРМАТИЧКИ НАУКИ И КОМПЈУТЕРСКО ИНЖЕНЕРСТВО

FIFTY WAYS TO DETECT A GHOSTWRITER

Katerina Zdravkova
Faculty of Computer Science and Engineering
University Ss. Cyril and Methodius in Skopje
Rudjer Boskovikj bb, 1000 Skopje, Macedonia
e-mail: katerina.zdravkova@finki.ukim.mk

Conference on Data Mining and Data Warehouses
Ljubljana, 10 October 2011

Contents

- Plagiarism and ghostwriting
- Mutually opposed sides
- Indicators
  - External
  - Internal
  - Joint
- Experimental results
- System architecture
- Conclusion and new project at FINKI

Plagiarism in the academic world

- Copying of another's work
- Borrowing other's ideas
- Implementing other's work without proper crediting
- In the recent years, facilitated by massive storage technologies and online machine translation
- Countermeasures:
  - search engines + Google Translate
  - tools: iThenticate, Turnitin, or WriteCheck

Ghostwriting

- A ghostwriter is a person who acts in the name of the official author
- Well-paid business
  - Example, El Dante, “completed 12 graduated theses of 50 pages or more”
- Specialized agencies (essay or paper mills)
- China: estimation that “university students spend up to half a billion yuan ($73 million) a year to have other people write their essays”

Measures against ghostwriting

- Ghostwriting is rarely detected and almost impossible to prove.
- Lucrative business
- Procedural concerns grow
  - USA intends to expand the federal rules to diminish its side effects.
  - Faculties minimize the contribution of individual essays in the final grade.

Mutually opposed sides

- Teachers: usually intuitively feel the cheat
  - no means to prove it with indisputable certainty.
- Students: do not hesitate to use all means to get a favorable grade
  - always have an accurate and very rational excuse for all teachers' accusations.
- Very few teachers have the courage to find material evidence of the cheat, and time to personally enquire the student.

Our experience

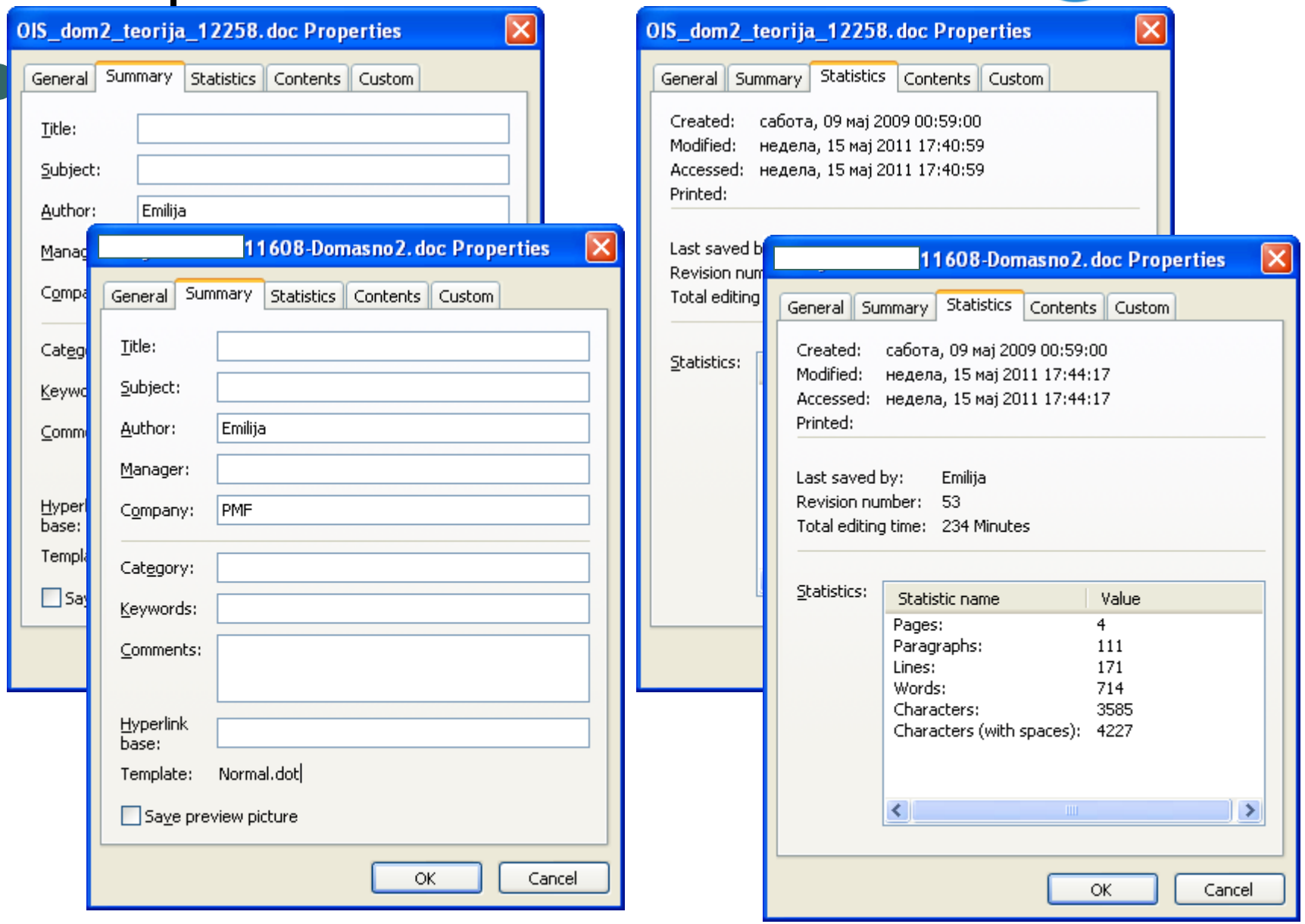
- Professional Ethics course
- In average 150 students
- Manual “investigation”
- Input:
  - Document archive (mainly .doc, .docx, .odt and pdf)
  - Reports of student activities (XLS file)
- Output: clusters of similar papers

External indicators

- Document metadata
- Activity reports
- Document properties + student access times = joint indicators
- IP address

Document properties

- Basic document metadata:
  - title of the paper
  - name of the first author
  - name of the user who last saved it,
  - time of paper creation,
  - revision number
  - total editing time


OIS_dom2_teorija_12258.doc Properties
General
Summary
Statistics
Contents
Custom
Title:
Subject:
Author: Emilija
11608-Domasno2.doc Properties
Author: Emilija
Manager:
Company: PMF
Category:
Keywords:
Comments:
Hyperlink base:
Template: Normal.dot
Save preview picture
OK
Cancel
Created: сабота, 09 мај 2009 00:59:00
Modified: недела, 15 мај 2011 17:40:59
Accessed: недела, 15 мај 2011 17:40:59
Printed:
Last saved by: Emilija
Revision number: 53
Total editing time: 234 Minutes
Statistics:
Statistic name
Value
Pages: 4
Paragraphs: 111
Lines: 171
Words: 714
Characters: 3585
Characters (with spaces): 4227

Activity report

- Time of every view to particular activity (assignment access, upload, submission view)
- IP address

Our calculation

- Each IP address is labeled

$$Label(IP_i) = n_i$$

- Each student ID is labeled

$$Label(ID_j) = \sum_{k=1}^{m_j} \frac{\ln(Label(IP_j))}{m_j}$$

- Students are clustered

$$Cluster(ID_j) = \begin{cases} 0{,}00 & 0 \le ID_j < M/5 \\ 0{,}25 & M/5 \le ID_j < 2M/5 \\ 0{,}50 & 2M/5 \le ID_j < 3M/5 \\ 0{,}75 & 3M/5 \le ID_j < 4M/5 \\ 1{,}00 & 4M/5 \le ID_j \le M \end{cases}$$

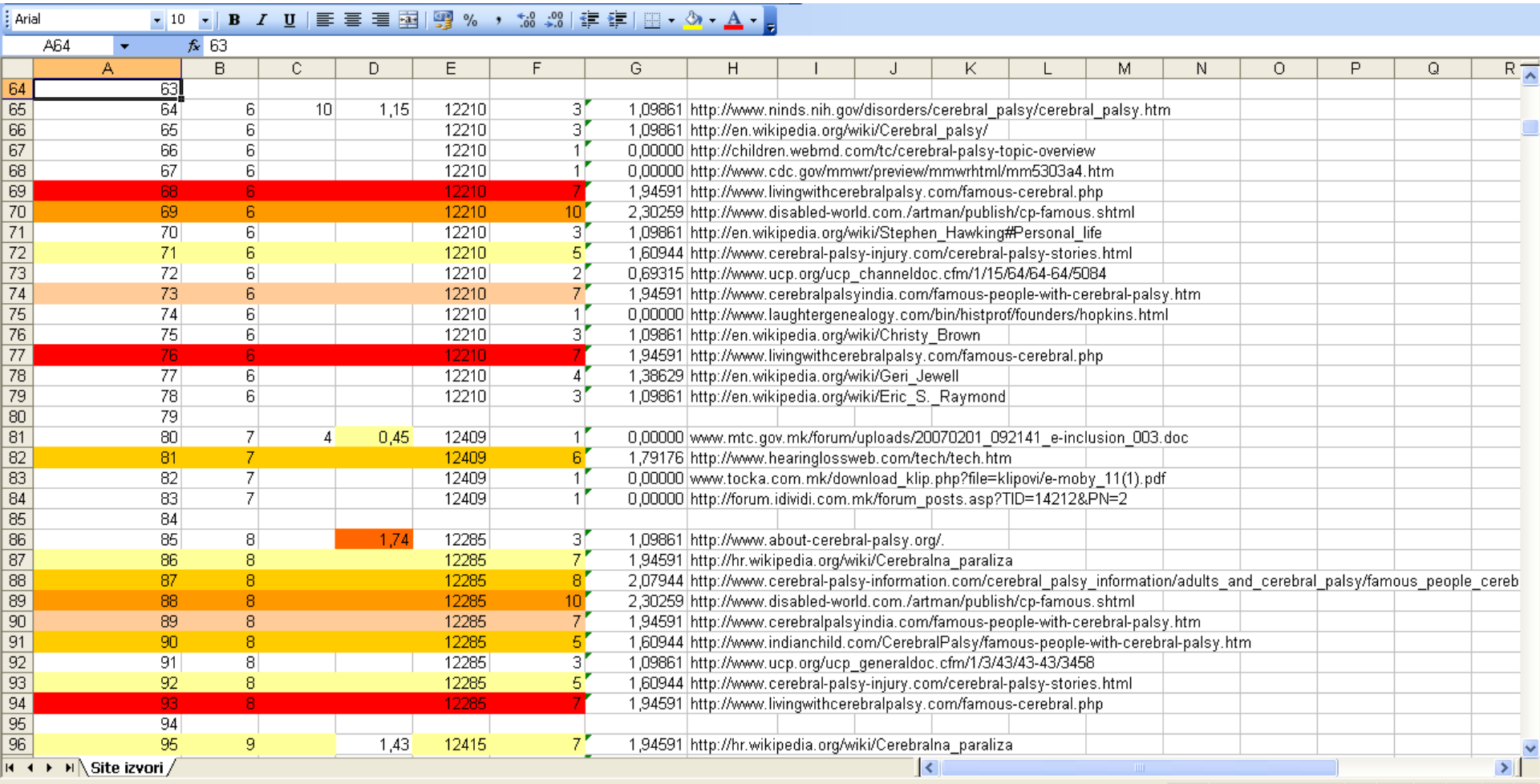
Joint indicators = difference between time of:

- document creation and the first access of the definition of its topic,
- document total editing time and the difference between first uploading and first access
- final uploading and document last modification.

Internal examination

- References
- Formatting styles
- Typographic similarities
- Linguistic similarities

References

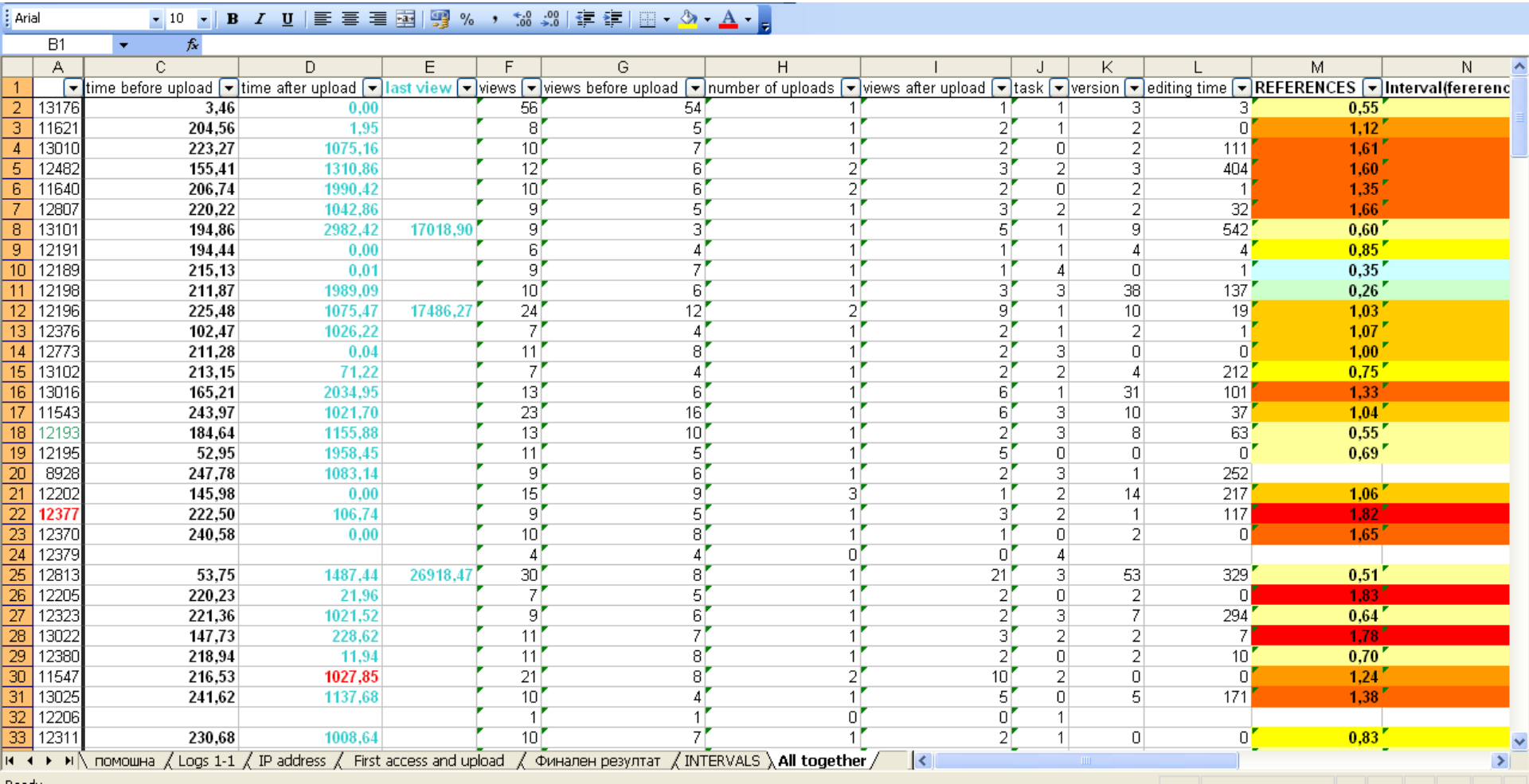


Arial 10 | A64 | fx 63

	A	B	C	D	E	F	G	H
64	63							
65	64	6	10	1,15	12210	3	1,09861	http://www.ninds.nih.gov/disorders/cerebral_palsy/cerebral_palsy.htm
66	65	6			12210	3	1,09861	http://en.wikipedia.org/wiki/Cerebral_palsy/
67	66	6			12210	1	0,00000	http://children.webmd.com/tc/cerebral-palsy-topic-overview
68	67	6			12210	1	0,00000	http://www.cdc.gov/mmwr/preview/mmwrhtml/mm5303a4.htm
69	68	6			12210	7	1,94591	http://www.livingwithcerebralpalsy.com/famous-cerebral.php
70	69	6			12210	10	2,30259	http://www.disabled-world.com./artman/publish/cp-famous.shtml
71	70	6			12210	3	1,09861	http://en.wikipedia.org/wiki/Stephen_Hawking#Personal_life
72	71	6			12210	5	1,60944	http://www.cerebral-palsy-injury.com/cerebral-palsy-stories.html
73	72	6			12210	2	0,69315	http://www.ucp.org/ucp_channeldoc.cfm/1/15/64/64-64/5084
74	73	6			12210	7	1,94591	http://www.cerebralpalsyindia.com/famous-people-with-cerebral-palsy.htm
75	74	6			12210	1	0,00000	http://www.laughtergenealogy.com/bin/histprof/founders/hopkins.html
76	75	6			12210	3	1,09861	http://en.wikipedia.org/wiki/Christy_Brown
77	76	6			12210	7	1,94591	http://www.livingwithcerebralpalsy.com/famous-cerebral.php
78	77	6			12210	4	1,38629	http://en.wikipedia.org/wiki/Geri_Jewell
79	78	6			12210	3	1,09861	http://en.wikipedia.org/wiki/Eric_S._Raymond
80	79							
81	80	7	4	0,45	12409	1	0,00000	www.mtc.gov.mk/forum/uploads/20070201_092141_e-inclusion_003.doc
82	81	7			12409	6	1,79176	http://www.hearinglossweb.com/tech/tech.htm
83	82	7			12409	1	0,00000	www.tocka.com.mk/download_klip.php?file=klipovi/e-moby_11(1).pdf
84	83	7			12409	1	0,00000	http://forum.idividi.com.mk/forum_posts.asp?TID=14212&PN=2
85	84							
86	85	8		1,74	12285	3	1,09861	http://www.about-cerebral-palsy.org/.
87	86	8			12285	7	1,94591	http://hr.wikipedia.org/wiki/Cerebralna_paraliza
88	87	8			12285	8	2,07944	http://www.cerebral-palsy-information.com/cerebral_palsy_information/adults_and_cerebral_palsy/famous_people_cereb
89	88	8			12285	10	2,30259	http://www.disabled-world.com./artman/publish/cp-famous.shtml
90	89	8			12285	7	1,94591	http://www.cerebralpalsyindia.com/famous-people-with-cerebral-palsy.htm
91	90	8			12285	5	1,60944	http://www.indianchild.com/CerebralPalsy/famous-people-with-cerebral-palsy.htm
92	91	8			12285	3	1,09861	http://www.ucp.org/ucp_generaldoc.cfm/1/3/43/43-43/3458
93	92	8			12285	5	1,60944	http://www.cerebral-palsy-injury.com/cerebral-palsy-stories.html
94	93	8			12285	7	1,94591	http://www.livingwithcerebralpalsy.com/famous-cerebral.php
95	94							
96	95	9		1,43	12415	7	1,94591	http://hr.wikipedia.org/wiki/Cerebralna_paraliza

Site izvori

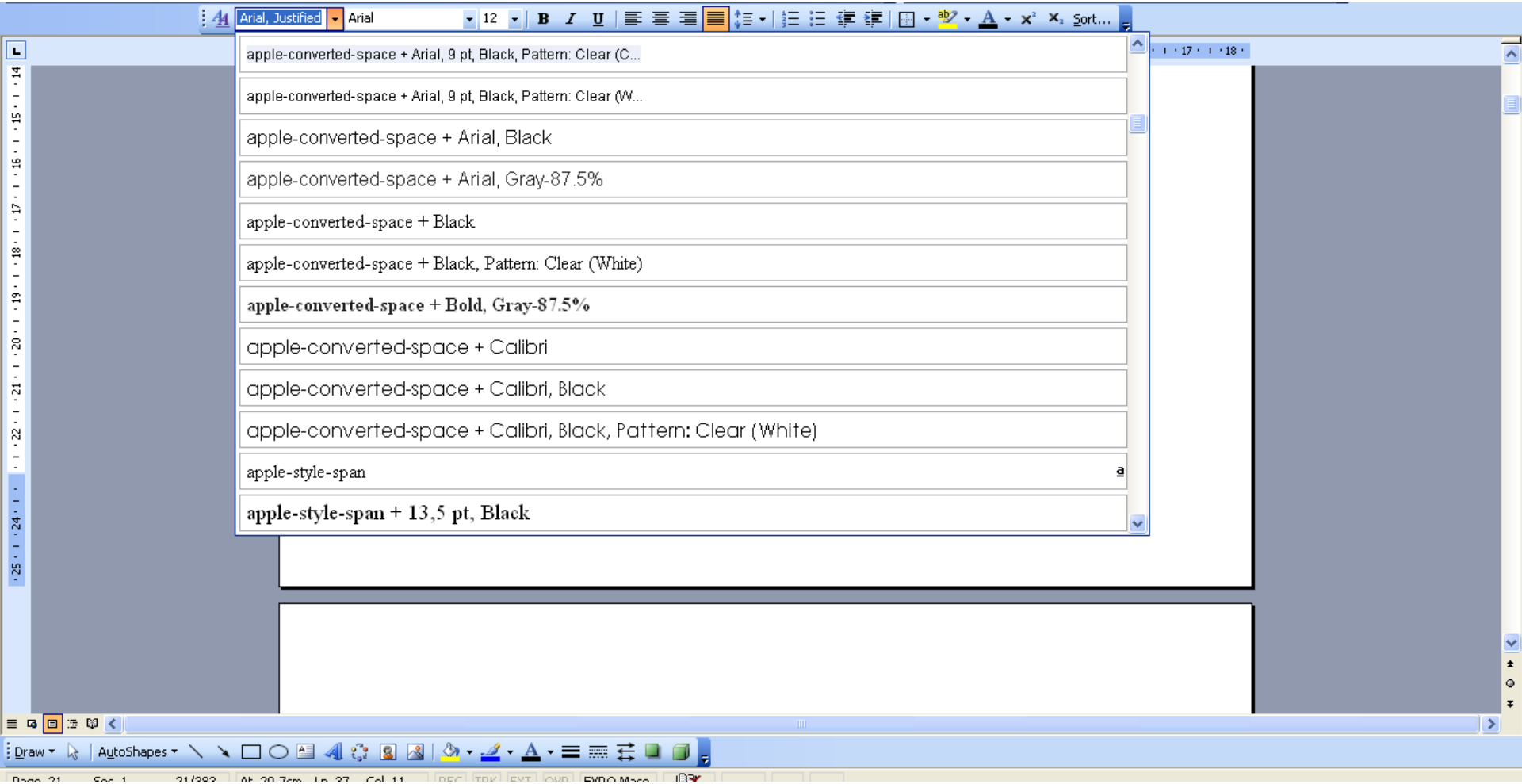
An example of manual "investigation"



	A	C	D	E	F	G	H	I	J	K	L	M	N
1		time before upload	time after upload	last view	views	views before upload	number of uploads	views after upload	task	version	editing time	REFERENCES	Interval(fererenc
2	13176	3,46	0,00		56	54	1	1	1	3	3	0,55	
3	11621	204,56	1,95		8	5	1	2	1	2	0	1,12	
4	13010	223,27	1075,16		10	7	1	2	0	2	111	1,61	
5	12482	155,41	1310,86		12	6	2	3	2	3	404	1,60	
6	11640	206,74	1990,42		10	6	2	2	0	2	1	1,35	
7	12807	220,22	1042,86		9	5	1	3	2	2	32	1,66	
8	13101	194,86	2982,42	17018,90	9	3	1	5	1	9	542	0,60	
9	12191	194,44	0,00		6	4	1	1	1	4	4	0,85	
10	12189	215,13	0,01		9	7	1	1	4	0	1	0,35	
11	12198	211,87	1989,09		10	6	1	3	3	38	137	0,26	
12	12196	225,48	1075,47	17486,27	24	12	2	9	1	10	19	1,03	
13	12376	102,47	1026,22		7	4	1	2	1	2	1	1,07	
14	12773	211,28	0,04		11	8	1	2	3	0	0	1,00	
15	13102	213,15	71,22		7	4	1	2	2	4	212	0,75	
16	13016	165,21	2034,95		13	6	1	6	1	31	101	1,33	
17	11543	243,97	1021,70		23	16	1	6	3	10	37	1,04	
18	12193	184,64	1155,88		13	10	1	2	3	8	63	0,55	
19	12195	52,95	1958,45		11	5	1	5	0	0	0	0,69	
20	8928	247,78	1083,14		9	6	1	2	3	1	252		
21	12202	145,98	0,00		15	9	3	1	2	14	217	1,06	
22	12377	222,50	106,74		9	5	1	3	2	1	117	1,82	
23	12370	240,58	0,00		10	8	1	1	0	2	0	1,65	
24	12379				4	4	0	0	4				
25	12813	53,75	1487,44	26918,47	30	8	1	21	3	53	329	0,51	
26	12205	220,23	21,96		7	5	1	2	0	2	0	1,83	
27	12323	221,36	1021,52		9	6	1	2	3	7	294	0,64	
28	13022	147,73	228,62		11	7	1	3	2	2	7	1,78	
29	12380	218,94	11,94		11	8	1	2	0	2	10	0,70	
30	11547	216,53	1027,85		21	8	2	10	2	0	0	1,24	
31	13025	241,62	1137,68		10	4	1	5	0	5	171	1,38	
32	12206				1	1	0	0	1				
33	12311	230,68	1008,64		10	7	1	2	1	0	0	0,83	

помошна | Logs 1-1 | IP address | First access and upload | Финален резултат | INTERVALS | All together

Formatting styles



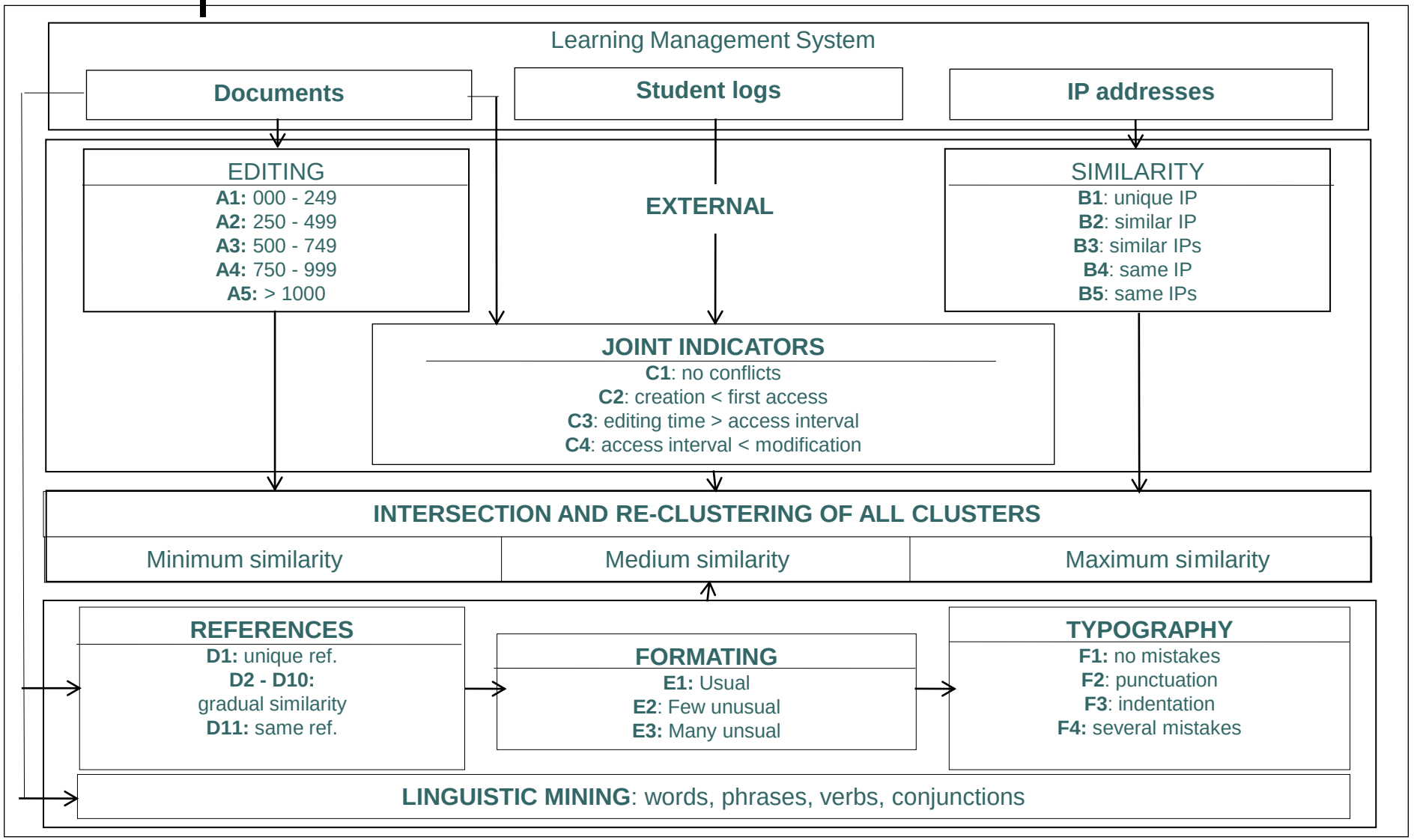
Typographic similarities

- No space after punctuation mark (17 out of 185), most with no editing time
- Reference [xxx]. (5, all from the same IP address)
- Indentation with spaces (12 out of 185), most in pdf

Linguistic similarities

- “Used sources” 37 / 185
- “References” 17 / 185
- “Many people … exist” 47 / 185
- “God / nature / humanity / life was not fair / honest” 12/ 185
  - Coincide with indentation with several spaces
- Words from particular dialects – discovered afterwards

System architecture



Conclusion

- Automatic tool is under construction
- It will only suggest potential presence of a ghostwriter
- Indicators are sensitive to student deficiencies
- But, the professional outsourcer was never caught in the net.
- Solution: oral examinations

- Thank you for your attention ☺