Composition Originality Tools

Comparative Research Preliminary Report September 15, 2017 • http://adam.rusch.me

By: Adam Rusch PhD Candidate, College of Education Education Policy, Organization and Leadership University of Illinois at Urbana-Champaign

As online systems make it easier for learners to access a multitude of academic resources, it becomes harder for teachers and graders to differentiate original content from materials that are plagiarized or inappropriately cited. At best, a student may have made an innocent mistake by copying information without giving appropriate credit to the original source. At worst, a student may be deliberately plagiarizing another scholar's work to represent as his or her own. In response, various "originality tools" have arisen in the education technology marketplace to scan student papers and compare them against currently known works to determine if a paper contains original or unoriginal writing.

The purpose of this project is to compare five originality tools that are commonly integrated into Learning Management Systems (LMS) used in academic settings. After students submit an assigned paper through the LMS, the instructor is able to use the tool to check for originality as part of the assessment process. This allows for corrective action, such as asking for more appropriate citation, or punitive action, such as flagging the work for plagiarism, to be taken.

The five originality tools being investigated are: SafeAssign, Turnitin, VeriCite, UniCheck and Urkund. SafeAssign is a Blackboard product available exclusively on their platforms and was used on the Blackboard Learn system for this study. The other originality tools are available on most leading LMS platforms. Turnitin was used as a Moodle plugin. VeriCite was used as a Sakai plugin. UniCheck and Urkund were each used through their standalone web portals.

Methods

To compare the five tools, I ran a series of 50 sample papers through each system. The papers were 10 page selections from the Princeton University Doctoral Dissertations, 2011-2017 database¹ and came from across 10 different disciplines to account for variations in subjects and themes. This database was selected because it had not been indexed by any of the originality tools being investigated.

Of the papers, 25 were "unseeded" and only contained original content from the author. Another 25 of the papers were "seeded" with known non-original content coming from easily available web sources, such as Wikipedia, Encyclopedia Brittanica, Center for Disease

¹ http://dataspace.princeton.edu/jspui/handle/88435/dsp01td96k251d

Control, HowStuffWorks.com, and the EBSCO online journal database available to me as a student at the University of Illinois. See Appendix I for full list of test works.

The originality score and significant matching citations were recorded as the papers were run through each system. I had two major questions. Q1: Does the originality tool properly identify the seeded material as non-original content and the site I plagiarized? Q2: Does the tool find examples of the author's original work in any other places online? These were cases where the authors had used portions of their doctoral work in other scholarly pieces that had been published before or after the dissertation defense.

Results

For Q1, VeriCite and UniCheck were the best performers with Urkund close behind. As you see, these tools identified 22 and 21 of the 25 examples of seeded material, respectively. SafeAssign identified the site for our seeded material only 15 times and Turnitin identified the site 18 times. However, there are some interesting notes of divergence.

	Plagiarisn	n Analysis	Results	
SafeAssign	Turnitin	VeriCite	Urkund	UniCheck
15 *	18*	22	21	22
60%	72%	88%	84%	88%
	See App	endix II for f	ull list	

In one case, SafeAssign did identify portions of the seeded material as non-original. However, it listed the source as a site that had copied information from the website where I had obtained it. Essentially, it found a fellow plagiarizer while it did not find the original seeded material itself.

In four cases, Turnitin identified seeded material as non-original without identifying the site where I obtained it. Twice, like SafeAssign, it gave us websites that had copied the material and (in these cases) properly cited it. In two other cases, Turnitin told us that the materials were from their cross-institutional student paper database ("Submitted to [Institution] Student Paper").

For Q2, I did not know if there would be any additional publications but was interested in seeing what the tools would find. Turnitin found the most examples of original source material being published elsewhere, identifying 12 works. Urkund and Unicheck each found 11 works published elsewhere and VeriCite identified 9. SafeAssign did not find any.

	Source	Analysis Re	esults	
SafeAssign	Turnitin	VeriCite	Urkund	UniCheck
0	12	9	11	11
	See Appe	endix III for f	ull list	

In a few cases, there were close similarities for source papers and seeded materials that were not considered significant results and recorded as matches here, even if they gave high non-originality percentages. These generally fell into one of thee categories: 1) The author used common jargon for the field that we would expect to see repeated across multiple papers, 2) The author inserted long quotations that others had also included in their works, 3) The system identified a bibliographic reference as matching content because it used a standard citation style.

Discussion

The differences in the ways that each tool performed helps us understand some of the differences that may be part of their design. See Appendix IV for screen shots that give a sense of how each system works. In the first case, SafeAssign seems to be the tool with the most limited scans for outside sources. While it picked up 60% of the seeded materials, they all came from easily imagined research sites like Wikipedia, History.com, and the Center for Disease Control.

It looks like the SafeAssign web search has a very tight set of parameters for their web search. It missed quite a few general websites, found none of the journal articles used for seeded materials, and found no examples of the original material being published in journal articles elsewhere. It may be part of the SafeAssign design philosophy that they believe these gaps will be filled as an institutional database is built up of submitted student papers.

Turnitin does seem to have an extensive web cache built up for originality comparison. It is notable that this tool found the most examples of original work published elsewhere. On the other hand, there are some intriguing questions about how this web cache is handled. In four cases, the system correctly identified seeded material as being non-original content but did not give us a link to the original site.

In two cases, the non-original content was attributed to students from another institution. While the match was sufficient to determine likely plagiarism in this case, it is very likely that in another scenario with less of a match it would be less clear and I can't be sure whether Turnitin recognized that the seeded material also came from the World Wide Web. There is also considerable debate over the way that Turnitin co-mingles all institutional works into a common database because of privacy and intellectual property concerns. Opponents say it is wrong to appropriate this work that was only intended for a class assignment for a shared database or corporate endeavor.²

In two more cases, the non-original content was identified but attributed to a site that didn't match our seeded material. However, in both of those cases the link to the websites were dead, and I only had the Turnitin web cache for reference. Based on the ability that Turnitin showed for finding source material in Q2, I expected that they would also be able

 $^{^2\} https://www.insidehighered.com/news/2017/06/19/anti-turnitin-manifesto-calls-resistance-some-technology-digital-age$

to find the site for seeded material. But again, without having the correct site listed I can't credit Turnitin with finding the examples I had seeded.

Did the Turnitin web crawler come across these secondary sites first and then determine that they didn't need to mention the other sites (the correct sites) as the source of non-original content? If that is the case, then it is especially problematic that the Turnitin web cache is out of date and contains dead links. Did Turnitin decide that the sources it showed us were more valid? If an instructor cannot verify the status of allegedly non-original content it becomes more difficult to prepare a plan for correction.

The performance of the last three tools was nearly equivalent. VeriCite and UniCheck performed the best in the main question, with Urkund close behind. In the secondary question, Vericite was behind by just 2-3 samples. I found in examining these tools that the difference may be in the depth that the web crawl indexed sites. While all three found the sites that any student might access from a simple web search, the differences emerged as the tools dug into deeper layers. There were several instances of papers being found that had been saved in journal, organizational, or governmental archives, or on the personal web space of faculty members (i.e. spaces where they could give their students the link to download a PDF and read the article for a college course).

These results reveal the two different forms of originality checks. SafeAssign and Turnitin seem to be designed foremost for checking submitted content against the institutional databases of their clients and partner institutions. This is useful for discovering if students have submitted the same paper in more than one course, copied work from a friend, or bought a paper from a "term paper" database.

VeriCite, UniCheck and Urkund performed better in identifying web sources that may be the source of non-original content. While UniCheck and Urkund fared slightly better than VeriCite in the deep search, it was not enough to render a clear verdict as to any system being superior. All three found the most commonly searched elements that students would find when composing an academic work. This is important for using composition originality tools as a teaching resource to train students in the writing process.

Conclusion

With a variety of composition originality tools to select from, institutions looking to integrate a service into their LMS will be able to consider whatever criteria is best for them. There may good reasons why an institution would prefer the institutional database check or the web search check to be the main focus of their tool. There are also structural and bureaucratic considerations that will inform the decision.

Ultimately, these tools are intended to make the assessment process easier and more informative for both students and instructors. Freeing up instructor time that might be spent investigating papers for possible infractions gives more opportunities for other class activities. Anything that the tools can do to guide students in developing good writing practices is an obvious bonus that we should hope to be in the future of these systems.

Appendix I

Paper / Subject	Year	Seed	Seed Source	Paper / Subject	Year	Seed	Seed Source
Anthopology				East Asian Studies			
Gordon, Gwendolyn	2014			Bridges IV, William H.	2012		
Polk, Daniel	2014			Gregory, Scott Wentworth	2012		
Robinson, Mark	2014			Hunter, Michael	2012	2 Page	AsiaSociety.org
MoranThomas, Amy	2012	2 Page	Wikipedia	Ro, Sang-ho	2012	2 Page	Today Translations
Savova, Nadezhda Dimitrova	2013	4 Page	EBSCO Journal	Compton, Eno	2013	4 Page	New World Encyclopedia
Architecture				Economics			
Buckley, Craig	2013			Alvarez, Jorge Alejandro	2016		
Efrat, Zvi	2014			Ge, Qi	2016		
Hsieh, Lisa L.	2013			Ravit, Jason Gregory	2016		
Campbell, Mark	2014	2 Page	EBSCO Journal	Zeltzer, Dan	2016	2 Page	Amer Soc of Mech Engr
Sunwoo, Irene	2013	4 Page	Encyclopedia Brittanica	Feng, Xiaochen	2016	4 Page	BeBusinessed.com
Chemical and Biological Engine	eering			German			
Girardi, Matthew	2015			Attanucci, Timothy J.	2012		
Bozym, David	2015			Christian, Margareta Ingrid	2012		
Davis, Raleigh Lloyd	2015	2 Page	EBSCO Journal	Eldridge, Sarah Vandegrift	2012		
Dsilva, Carmeline Joan	2015	2 Page	Wolfram Mathworld	Spies, Petra	2012	2 Page	Wikipedia
Hiszpanski, Anna Maria	2015	4 Page	HowStuffWorks.com	King, Alana Jane	2014	4 Page	Christian Cyclopedia
Chemistry				Neuroscience			
Fortmeyer, Ivy Camille	2016			Coen, Philip	2015		
Ganguly, Aahana Nibedita	2016			Silbert, Lauren	2014		
Terrett, Jack Alexander	2016	2 Page	Phys.org	Solway, Alec	2014	2 Page	Encyclopedia of Philosophy
Hone, Graham	2016	2 Page	Wikipedia	Eldar, Eran	2014	4 Page	Wikipedia
Digianantonio, Katherine	2016	4 Page	Center for Disease Control	Opendak, Maya	2015	4 Page	Scopus Journal
Classics				Public and International Affairs	}		
Jones, Madeleine Kersti	2013			Coffey, Diane	2015		
Clark, Virginia Emily	2014			Palmer, John RB	2013		
Meinrath, Danielle	2015			Collins, Liam	2014	2 Page	Small Wars Journal
Oswald, Simon	2014	2 Page	History.com	Kanter, David	2014	4 Page	Wikipedia
Sirois, Martin	2014	4 Page	Scopus Journal	Lim, Darren James	2014	4 Page	CIA Worldbook

Appendix II

Plagiarism Analysis Results

Paper / Subject	SafeAssign	Turnitin	VeriCite	Urkund	UniCheck	Seed Source
Anthopology						
Gordon	2%	17%	11%	9%	4.3%	
Polk	2%	10%	10%	2%	0.0%	
Robinson	1%	10%	10%	2%	0.0%	
MoranThomas 2P	13%	25%	19%	19%	14.7%	Wikipedia
Savova 4P	1%	10%	53%	47%	42.1%	EBSCO Journal
340044	170	1070	3370	4770	42.170	EBSCO Journal
Architecture						
Buckley	1%	10%	10%	4%	0.0%	
Efrat	1%	17%	17%	13%	9.8%	
Hsieh	2%	12%	10%	2%	0.0%	
Campbell 2P	1%	17%	10%	1%	2.1%	EBSCO Journal
Sunwoo 4P	28%	54%	48%	46%	46.4%	Encyclopedia Brittanica
Chemical and Biological E Girardi	ingineering 11%	52%	47%	4%	37.3%	
Bozym	4%	19%	10%	4% 6%	4.7%	
Davis 2P	10%	37%	31%	1%	24.6%	EBSCO Journal
Dsilva 2P	9%	49%	42%	24%	60.0%	Wolfram Mathworld
Hiszpanski 4P	28%	56%*	41%	41%	43.5%	HowStuffWorks.com
riiszpanski 4F	2876	30%	41/0	41/0	43.370	nowstan works.com
Chemistry						
Fortmeyer	1%	16%	10%	4%	0.0%	
Ganguly	4%	28%	10%	6%	2.1%	
Terrett 2P	16%	77%	77%	83%	66.8%	Phys.org
Hone 2P	21%	41%	18%	24%	18.7%	Wikipedia
Digianantonio 4P	41%	63%	60%	44%	56.1%	Center for Disease Control
Classics						
Jones	1%	7%	10%	4%	0.0%	
Clark	8%	12%	22%	15%	9.9%	
Meinrath	2%	23%	10%	2%	0.0%	
Oswald 2P	20%	26%*	23%	22%	21.1%	History.com
Sirois 4P	2%	45%	10%	3%	2.2%	Scopus Journal
East Asian Studies						
Bridges	7%	15%	10%	4%	6.5%	
Gregory	4%	14%	10%	5%	0.0%	
Hunter 2P	26%	32%	27%	27%	21.2%	AsiaSociety.org
Ro 2P	9%	25%	25%	24%	23.5%	Today Translations
Compton 4P	11%	56%	55%	50%	44.4%	New World Encyclopedia
·						, ,
Economics						
Alvarez	2%	96%	10%	94%	74.3%	
Ge	1%	8%	65%	71%	55.4%	
Ravit	11%	14%	10%	2%	0.0%	
Zeltzer 2P	20%	49%	41%	35%	29.8%	Amer Soc of Mech Engr
Feng 4P	32%	94%	34%	96%	30.8%	BeBusinessed.com
German						
Attanucci	2%	23%	10%	9%	3.3%	
Christian	1%	17%	10%	5%	2.7%	
Eldridge	1%	17%	10%	1%	7.9%	
Spies 2P	13%	30%	24%	23%	22.3%	Wikipedia
King 4P	45%	51%*	39%	52%	34.8%	Christian Cyclopedia
Neuroscience						
Coen	2%	30%	10%	74%	0.0%	
Silbert	9%	62%	47%	38%	23.7%	
Solway 2P	10%	94%*	88%	95%	74.6%	Encyclopedia of Philosophy
Eldar 4P	34%	61%	52%	84%	53.5%	Wikipedia
Opendak 4P	20%	48%	50%	4%	16.3%	Scopus Journal
Public and International	Affairs					
Coffey	2%	93%	94%	96%	63.3%	
Palmer	1%	9%	38%	38%	2.5%	
Collins 2P	15%	50%	33%	28%	19.0%	Small Wars Journal
Kanter 4P	40%	65%	51%	51%	59.3%	Wikipedia
Lim 4P	33%*	58%	48%	44%	43.5%	CIA Worldbook
	, - <u></u>					
Seeded Material Found	15*	18*	22	21	22	Red: Seeded Material was not identified
Total Samples = 25	60%	72%	88%	84%	88%	Green: Seeded Material was identified

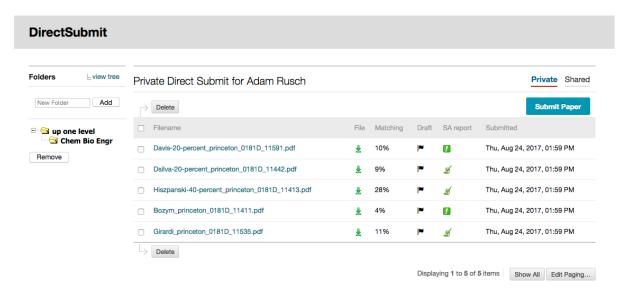
Appendix III

Source Analysis Results

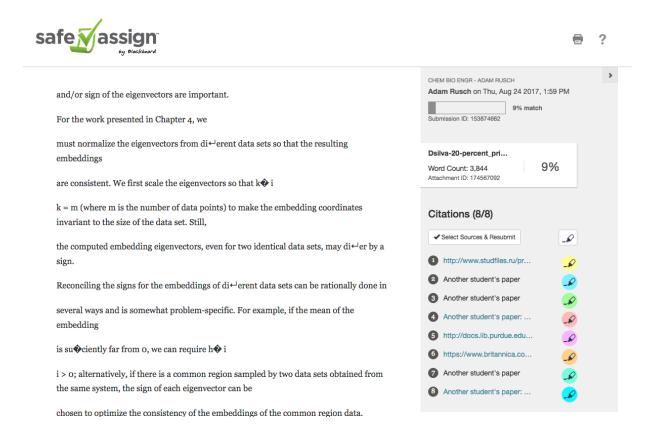
aper / Subject	SafeAssign	Turnitin	VeriCite	Urkund	UniCheck
Anthopology					
Gordon	2%	17%	11%	9%	4.3%
olk	2%	10%	10%	2%	0.0%
obinson	1%	10%	10%	2%	0.0%
IoranThomas 2P	13%	25%	19%	19%	14.7%
avova 4P					
vova 4P	1%	10%	53%	47%	42.1%
chitecture					
ckley	1%	10%	10%	4%	0.0%
at	1%	17%	17%	13%	9.8%
ieh	2%	12%	10%	2%	0.0%
mpbell 2P	1%	17%	10%	1%	2.1%
nwoo 4P	28%	54%	48%	46%	46.4%
emical and Biological	Fngineering				
ardi	11%	52%*	47%*	4%	37.3%*
zym	4%	19%	10%	6%	4.7%
vis 2P	10%	37%	31%	1%	24.6%
lva 2P	9%	49%	42%	24%	60%*
panski 4P	28%	56%	41%	41%	43.5%
mistry :meyer	1%	16%	10%	4%	0.0%
nguly	4%	28%	10%	6%	2.1%
rett 2P	16%	77%*	77%*	83%*	66.8%*
rett 2P ne 2P					
	21%	41%*	18%	24%	18.7%
anantonio 4P	41%	63%*	60%*	44%	56.1%*
ssics					
es	1%	7%	10%	4%	0.0%
·k	8%	12%	22%	15%	9.9%
ırath	2%	23%	10%	2%	0.0%
ald 2P	20%	26%	23%	22%	21.1%
s 4P	2%	45%	10%	3%	2.2%
: Asian Studies					
ges	7%	15%	10%	4%	6.5%
-					
ory	4%	14%	10%	5%	0.0%
er 2P	26%	32%	27%	27%	21.2%
Р	9%	25%	25%	24%	23.5%
pton 4P	11%	56%	55%	50%	44.4%
omics					
rez	2%	96%*	10%	94%*	74.3%*
	1%	8%	65%*	71%*	55.4%*
	11%	14%	10%	2%	0.0%
er 2P	20%	49%	41%	35%	29.8%
4P	32%	94%*	34%	96%*	30.8%
nan nucci	2%	23%	10%	9%	3.3%
tian	1%	17%	10%	5%	2.7%
lge	1%	17%	10%	1%	7.9%
2P	13%	30%	24%	23%	22.3%
IP .	45%	51%	39%	52%*	34.8%
oscience					
n	2%	30%*	10%	74%*	0.0%
rt	9%	62%*	47%*	38%*	23.7%*
ay 2P	10%	94%*	88%*	95%*	74.6%*
4P	34%	61%*	52%	84%*	53.5%*
idak 4P	20%	48%*	50%*	4%	16.3%*
c and Internation -	l Affaira				
ic and International ey	I Attairs 2%	93%*	94%*	96%*	63.3%*
ner	1%	9%	38%*	38%*	2.5%
ins 2P	15%	50%	33%	28%	19.0%
ter 4P 4P	40% 33%	65%* 58%	51% 48%	51% 44%	59.3% 43.5%
nal Material Found	0	12		11	
shed elswhere)		Blue:	Original mate	erial found el	sewhere

Appendix IV

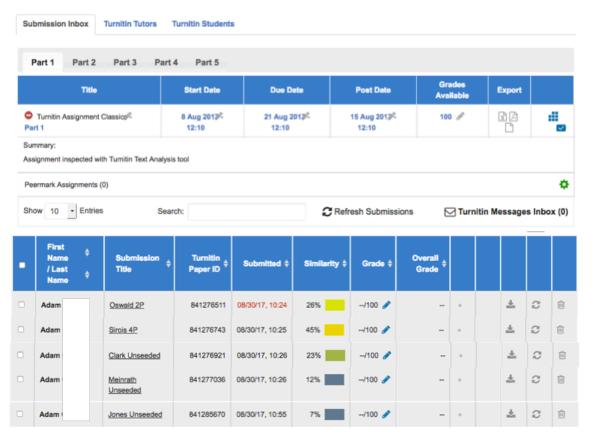
Screen shots to show system differences:



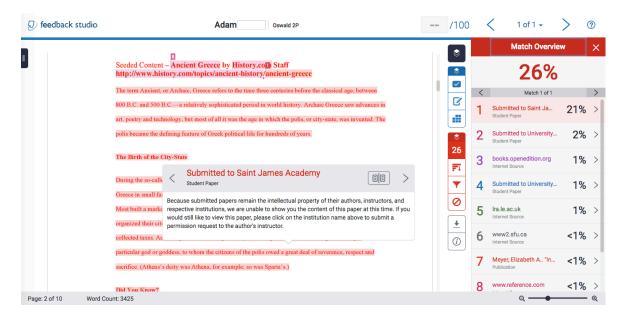
SafeAssign summary of Chemical and Biological Engineering works.



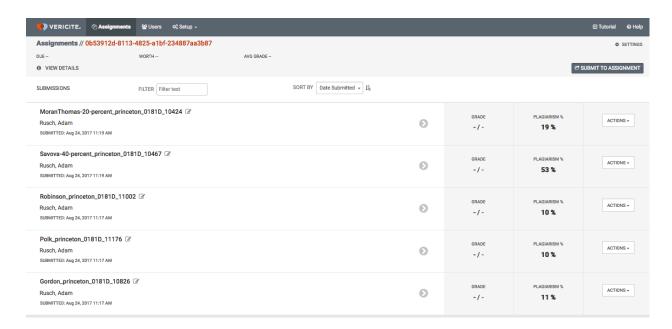
SafeAssign Report for Dsilva paper. Note that the seeded material source, Wolfram Mathworld, is not identified.



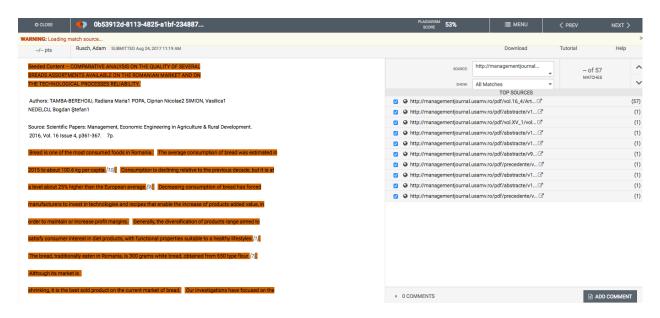
Turnitin summary of Classics works.



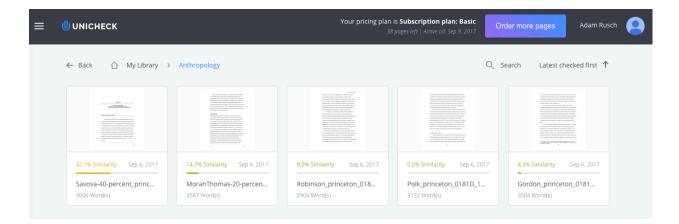
Turnitin results Oswald paper. Note the incorrect source for seeded material identified and that this result comes from the cross-institutional database.



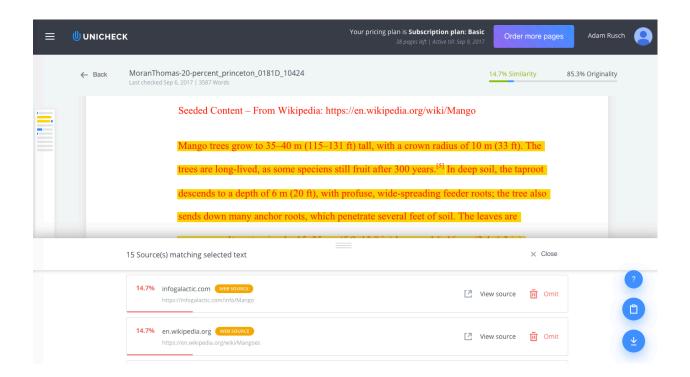
Vericite summary of Anthropology works.



VeriCite results Savova paper. Note how seeded content is highlighted and corresponds to source listed in side panel.



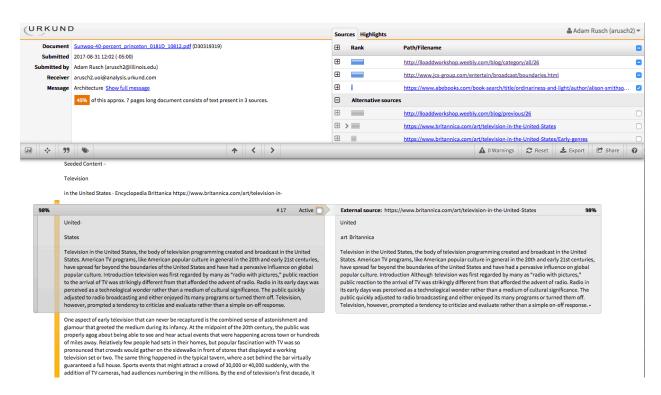
UniCheck summary of Anthropology works.



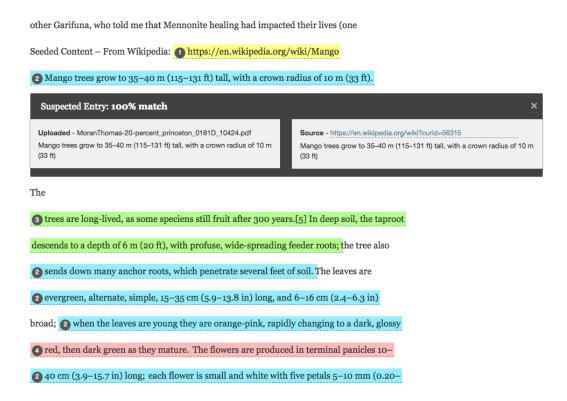
UniCheck results page of MoranThomas work.



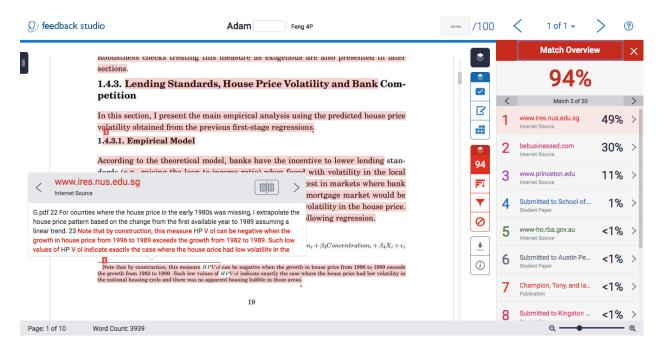
Urkund summary of Architecture works.



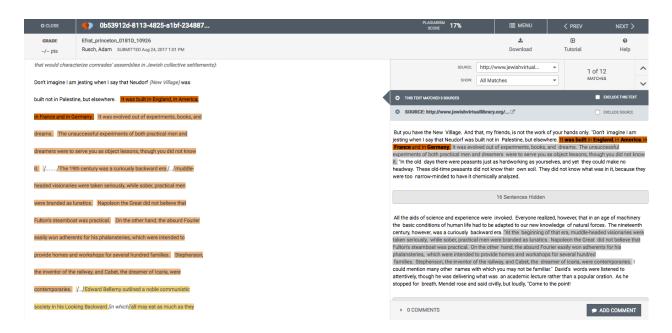
Urkund results of Sunwoo paper. Note how both a site that copied this seeded material as well as the legitimate site of seeded material are listed.



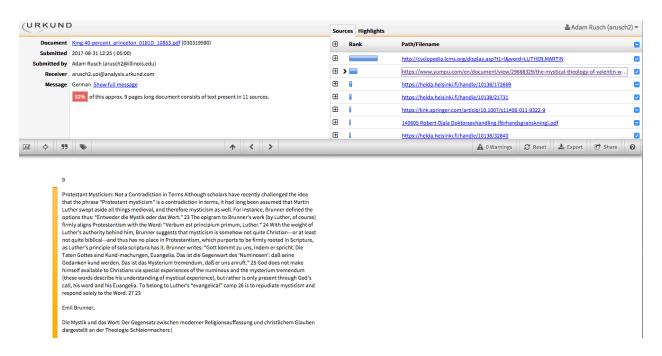
Example of SafeAssign correctly identifying a source of seeded material in MoranThomas paper from Anthropology. The text highlighted in yellow is identified as being in another student paper and the green highlight appears in another version of Wikipedia.



Example of Turnitin finding original work published in another place in Feng paper from Economics. In this case it is a PDF stored on a university web space in Singapore.



Example of Vericite identifying the original work that a block quote came from in Efrat paper from Architecture. Upon inspection, it can be seen that this selection is properly cited, which is good for an instructor to know.



Example of Urkund identifying seeded material and finding another place where the original material was published in King paper from German.