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
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
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

\[
Label(ID_j) = \sum_{k=1}^{m_j} \ln(Label(IP_j))
\]

- Students are clustered

\[
Cluster(ID_j) = \begin{cases} 
0.00 & 0 \leq ID_j < M/5 \\
0.25 & M/5 \leq ID_j < 2M/5 \\
0.50 & 2M/5 \leq ID_j < 3M/5 \\
0.75 & 3M/5 \leq ID_j < 4M/5 \\
1.00 & 4M/5 \leq ID_j \leq M 
\end{cases}
\]
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
<table>
<thead>
<tr>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>K</th>
<th>L</th>
<th>M</th>
<th>N</th>
<th>O</th>
<th>P</th>
<th>Q</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>64</td>
<td>64</td>
<td>6</td>
<td>10</td>
<td>1,15</td>
<td>12210</td>
<td>3</td>
<td>1,0961</td>
<td><a href="http://www.ninds.nih.gov/disorders/cerebral_palsy/cerebral_palsy.htm">http://www.ninds.nih.gov/disorders/cerebral_palsy/cerebral_palsy.htm</a></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>65</td>
<td>66</td>
<td>6</td>
<td>12210</td>
<td>3</td>
<td>1,0961</td>
<td><a href="http://en.wikipedia.org/wiki/Cerebral_palsy/">http://en.wikipedia.org/wiki/Cerebral_palsy/</a></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>67</td>
<td>66</td>
<td>6</td>
<td>12210</td>
<td>1</td>
<td>0,0000</td>
<td><a href="http://childrenswebmd.com/cerebral-palsy-topic-overview">http://childrenswebmd.com/cerebral-palsy-topic-overview</a></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>68</td>
<td>67</td>
<td>6</td>
<td>12210</td>
<td>1</td>
<td>0,0000</td>
<td><a href="http://www.cdc.gov/mmwr/preview/mmwrhtml/mm5333a4.htm">http://www.cdc.gov/mmwr/preview/mmwrhtml/mm5333a4.htm</a></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>69</td>
<td>68</td>
<td>6</td>
<td>12210</td>
<td>7</td>
<td>1,34591</td>
<td><a href="http://www.livingwithcerebralpalsy.com/famous-cerebral.php">http://www.livingwithcerebralpalsy.com/famous-cerebral.php</a></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>70</td>
<td>69</td>
<td>6</td>
<td>12210</td>
<td>10</td>
<td>2,30259</td>
<td><a href="http://www.disabled-world.com/janetman/publish/c-famous.shtml">http://www.disabled-world.com/janetman/publish/c-famous.shtml</a></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>71</td>
<td>70</td>
<td>6</td>
<td>12210</td>
<td>3</td>
<td>1,3661</td>
<td><a href="http://en.wikipedia.org/wiki/Stephen_Hawking#Personal_life">http://en.wikipedia.org/wiki/Stephen_Hawking#Personal_life</a></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>73</td>
<td>72</td>
<td>6</td>
<td>12210</td>
<td>2</td>
<td>0,89315</td>
<td><a href="http://www.ucp.org/ucp_channel.cfm/1/1564/64-64-5054">http://www.ucp.org/ucp_channel.cfm/1/1564/64-64-5054</a></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>74</td>
<td>73</td>
<td>6</td>
<td>12210</td>
<td>7</td>
<td>1,34591</td>
<td><a href="http://www.cerebralpalsynetwork.com/famous-people-with-cerebral-palsy.htm">http://www.cerebralpalsynetwork.com/famous-people-with-cerebral-palsy.htm</a></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>75</td>
<td>74</td>
<td>6</td>
<td>12210</td>
<td>1</td>
<td>0,0000</td>
<td><a href="http://www.laughtergenealogy.com/biographies/founders/hopkins.html">http://www.laughtergenealogy.com/biographies/founders/hopkins.html</a></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>76</td>
<td>75</td>
<td>6</td>
<td>12210</td>
<td>3</td>
<td>1,3661</td>
<td><a href="http://en.wikipedia.org/wiki/Christy_Brown">http://en.wikipedia.org/wiki/Christy_Brown</a></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>77</td>
<td>76</td>
<td>6</td>
<td>12210</td>
<td>7</td>
<td>1,34591</td>
<td><a href="http://www.livingwithcerebralpalsy.com/famous-cerebral.php">http://www.livingwithcerebralpalsy.com/famous-cerebral.php</a></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>78</td>
<td>77</td>
<td>6</td>
<td>12210</td>
<td>4</td>
<td>1,38529</td>
<td><a href="http://en.wikipedia.org/wiki/Geri_Jewell">http://en.wikipedia.org/wiki/Geri_Jewell</a></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>79</td>
<td>78</td>
<td>6</td>
<td>12210</td>
<td>3</td>
<td>1,0961</td>
<td><a href="http://en.wikipedia.org/wiki/Eric_S._Raymond">http://en.wikipedia.org/wiki/Eric_S._Raymond</a></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>80</td>
<td>79</td>
<td>7</td>
<td>4</td>
<td>0,45</td>
<td>12408</td>
<td>1</td>
<td>0,0000</td>
<td><a href="http://www.mtc.gov/mk/forum/uploads/20070201_082141_e-inclusion_003.doc">www.mtc.gov/mk/forum/uploads/20070201_082141_e-inclusion_003.doc</a></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>81</td>
<td>80</td>
<td>7</td>
<td>12409</td>
<td>6</td>
<td>1,79176</td>
<td><a href="http://www.hearinglossweb.com/tech/tech.htm">http://www.hearinglossweb.com/tech/tech.htm</a></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>82</td>
<td>81</td>
<td>7</td>
<td>12409</td>
<td>1</td>
<td>0,0000</td>
<td><a href="http://www.tocks.com/mk/download_klip.php?file=tlp0w/e-moby_11(t).pdf">www.tocks.com/mk/download_klip.php?file=tlp0w/e-moby_11(t).pdf</a></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>83</td>
<td>82</td>
<td>7</td>
<td>12409</td>
<td>1</td>
<td>0,0000</td>
<td><a href="http://forum.idwiki.com.mk/forum_posts.asp?TID=14218&amp;FN=2">http://forum.idwiki.com.mk/forum_posts.asp?TID=14218&amp;FN=2</a></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>84</td>
<td>83</td>
<td>7</td>
<td>12409</td>
<td>3</td>
<td>1,0961</td>
<td><a href="http://about-cerebral-palsy.org/">http://about-cerebral-palsy.org/</a></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
An example of manual “investigation”
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

Learning Management System

Documents

Student logs

IP addresses

EDITING

- A1: 000 - 249
- A2: 250 - 499
- A3: 500 - 749
- A4: 750 - 999
- A5: > 1000

SIMILARITY

- B1: unique IP
- B2: similar IP
- B3: similar IPs
- B4: same IP
- B5: same IPs

JOINT INDICATORS

- C1: no conflicts
- C2: creation < first access
- C3: editing time > access interval
- C4: access interval < modification

EXTERNAL

INTERSECTION AND RE-CLUSTERING OF ALL CLUSTERS

REFERENCES

- D1: unique ref.
- D2 - D10: gradual similarity
- D11: same ref.

FORMATING

- E1: Usual
- E2: Few unusual
- E3: Many unusual

TYPOGRAPHY

- F1: no mistakes
- F2: punctuation
- F3: indentation
- F4: several mistakes

LINGUISTIC MINING: words, phrases, verbs, conjunctions
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 😊