Analyzing a Large Corpus of Crowdsourced Plagiarism

Master’s Thesis

Michael Völske
Fakultät Medien
Bauhaus-Universität Weimar
06.06.2013

Supervised by:
Prof. Dr. Benno Stein
Prof. Dr. Volker Rodehorst

Advisors:
Dr. Steven Burrows
Dr. Matthias Hagen
Dr. Martin Potthast
Analyzing a Large Corpus of Crowdsourced Plagiarism

Master’s Thesis

Outline

· Introduction
· The Webis-TRC-12 Dataset
· Categorizing Crowdsourced Text Reuse
· Search Missions For Source Retrieval
· Summary
Introduction
Modeling Text Reuse From the Web

Search

Copy & Paste

Paraphrase

[Potthast 2011]
Introduction

Previous Plagiarism Corpora

- Automatically generated *artificial plagiarism*
- PAN-PC-10/11: >25,000 documents; >60,000 plagiarism cases
- Automatic transformations do not preserve semantics
The Webis-TRC-12 Dataset
The Webis-TRC-12 Dataset

Construction Overview

Topic

Author

Editor

ChatNoir SE

ClueWeb API

Revision log

Query log

ClueWeb
The Webis-TRC-12 Dataset

Construction Overview: Topics

Editor

ChatNoir SE

ClueWeb API

Revision log

Query log

ClueWeb

Topic

Author

7 © Michael Völske 2013
The Webis-TRC-12 Dataset

Construction Overview: Topics

Example topic:

Obama's family.

Write about President Barack Obama's family history, including genealogy, national origins, places and dates of birth, etc. Where did Barack Obama’s parents and grandparents come from? Also include a brief biography of Obama’s mother.

- Based on TREC Web Track topics 2009–2011
- 150 topics, 297 essays
- Target essay length: 5000 words
The Webis-TRC-12 Dataset

Construction Overview: Authors

- Topic
- Author
- Editor
- ChatNoir SE
- ClueWeb API

ChatNoir SE: The Family of Barack Obama is an extended clan of African American, Indonesian, and Kenyan heritage. They are best known through the political career of Barack Obama, the current President of the United America. His immediate family is the First Family of the United States, the first First Family of African American descent in the United States.

In what follows, we give a detailed overview of Barack Obama's far left on himself, his immediate and extended family, including mutual relations. Moreover, we give insights into the relations of Michelle Obama's wife, as well as some distant relations of both.

Barack Obama

ChatNoir SE: Barack Obama is the 44th and current President of the United States, the first African American to hold the office. Obama was the junior Senator from Illinois from 2005 until he resigned following his election.
The Webis-TRC-12 Dataset

Construction Overview: Authors

- Crowdsourcing: 27 total
- Professional writers hired on oDesk + volunteers
- Fluent English speakers
The Webis-TRC-12 Dataset

Construction Overview: Authors

- Crowdsourcing: 27 total
- Professional writers hired on oDesk + volunteers
- Fluent English speakers

Author Demographics (n=12)

<table>
<thead>
<tr>
<th>Category</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (Median)</td>
<td>37</td>
</tr>
<tr>
<td>Years Writing (Median)</td>
<td>8</td>
</tr>
<tr>
<td>Academic degree</td>
<td></td>
</tr>
<tr>
<td>Postgrad</td>
<td>33%</td>
</tr>
<tr>
<td>Undergrad</td>
<td>25%</td>
</tr>
<tr>
<td>English</td>
<td></td>
</tr>
<tr>
<td>Native</td>
<td>67%</td>
</tr>
<tr>
<td>Second Language</td>
<td>33%</td>
</tr>
</tbody>
</table>
The Webis-TRC-12 Dataset

Construction Overview: Sources

Editor

Revision log

ChatNoir SE

Query log

ClueWeb API

ClueWeb

Author

Topic

Barack Obama's Family
plagiarized by John Doe

The Family of Barack Obama is an extended clan of African American, Indonesian, and Kenyan heritage. They are best known through the political career of Barack Obama, the current President of the United States. His immediate family is the First Family of the United States, the first First Family of African American descent in the United States to enter the White House since the Kennedy's. Obama's family looks back today as Camelot.

In what follows, we give a detailed overview of Barack Obama's family, his immediate and extended family, including related relations. Moreover, we give insights into the relations of Michelle Obama's wife, as well as some distant relations of both. Barack Obama

Barack Hussein Obama II is the 44th and current President of the United States. He is the first African American to hold the office. Obama was a junior senator from Illinois from 2005 until he resigned following his election in November 2008. He was then the Democratic nominee for President of the United States in 2012.
The Webis-TRC-12 Dataset

Construction Overview: Sources

- ClueWeb09: 500 million English pages
- Representative sample of the web
- Commonly used in search engine evaluation (TREC)
The Webis-TRC-12 Dataset

Construction Overview: Search Engine

Editor

Revision log

Query log

ClueWeb API

ChatNoir SE

Topic

Author

Barack Obama's Family

The Family of Barack Obama is an extended clan of African American
Indonesian, and Kenyan heritage. They are best known through the
political career of Barack Obama, the current President of the United
America. His immediate family is the First Family of the United States;
the first First Family of African American descent in the United States
youngest to enter the White House since the Kennedys. Obama's young
family harks back to a myth of Camelot.

http://webis15.mediaxl.uni-saarland.de/ChatNoir/revLog/10001709993.html

In what follows, we give a detailed overview of Barack Obama's Far
light on himself, his immediate and extended family, including maiden
relatives. Moreover, we give insights into the relations of Michelle O
Obama's wife, as well as some distant relations of both.

Barack Obama

Barack Hussein Obama II is the 44th and current President of the Uni-
the first African American to hold the office. Obama was the junior
Senator from Illinois from 2005 until he resigned following his elec
The Webis-TRC-12 Dataset
Construction Overview: Search Engine

[Potthast et al. 2012a]

- Used for source retrieval
- Indexes ClueWeb
- Records fine-grained interaction log  [example]
The Webis-TRC-12 Dataset
Construction Overview: Editor

Revision log

Query log

ClueWeb

ChatNoir SE

Editor

Author

Topic

ChatNoir SE

ClueWeb API

Barack Obama's Family

plagiarized by John Doe

The Family of Barack Obama is an extended clan of African American, Indonesian, and Kenyan heritage. They are best known through the political career of Barack Obama, the current President of the United America. His immediate family is the First Family of the United States are the first First Family of African American descent in the United States. The youngest is entering the White House since the Kennedys. Obama’s best friends have his family back together in Camelot.

In what follows, we give a detailed overview of Barack Obama’s Family, his immediate and extended family, including their relations. Moreover, we give insights into the relations of Michelle Obama’s wife, as well as some distant relations of both.

Barack Obama

Barack Hussein Obama II is the 44th and current President of the United States, the first African American to hold the office. Obama was the junior United States Senator from Illinois from 2005 until he resigned following his election.
The Webis-TRC-12 Dataset
Construction Overview: Editor

[Potthast et al. 2012b]
The Webis-TRC-12 Dataset
Construction Overview: Editor

[Potthast et al. 2012b]

- Custom web-based rich text editor
- Records sources of re-used text passages
- New revision every 300ms of inactivity
- Detailed revision history  [example]
The Webis-TRC-12 Dataset

Three Main Data Sources

- ClueWeb API
- Revision log
- ClueWeb
- Query log
- Editor
- ChatNoir SE
- Author

Revision log
Query log
ClueWeb

Editor
ChatNoir SE
ClueWeb API

Barack Obama's Family
plagiarized by John Doe

The Family of Barack Obama is an extended clan of African American, Indonesian, and Kenyan heritage. They are best known through the political career of Barack Obama, the current President of the United America. His immediate family is the First Family of the United States, the first First Family of African American descent in the United States. Obama's family traces back to their roots of Cameroon.

In what follows, we give a detailed overview of Barack Obama's Family background, his immediate and extended family, including their relations. Moreover, we give insights into the relations of Michelle Obama's family, as well as some distant relations of both.

Barack Obama

Barack Hussein Obama II is the 44th and current President of the US and the first African American to hold the office. Obama was the junior Senator from Illinois from 2005 until he resigned following his election to the Senate.
The Webis-TRC-12 Dataset

Research Questions
The Webis-TRC-12 Dataset

Research Questions

1. Different text reuse approaches distinguishable?

2. Relation to existing plagiarism categorizations?
The Webis-TRC-12 Dataset

Research Questions

1. Different text reuse approaches distinguishable?

2. Relation to existing plagiarism categorizations?

3. Influence of text reuse task on search engine interaction?

4. Detectable via query log analysis?
The Webis-TRC-12 Dataset

Research Questions

1. Different text reuse approaches distinguishable?
2. Relation to existing plagiarism categorizations?
3. Influence of text reuse task on search engine interaction?
4. Detectable via query log analysis?

We expect new research insights and impacts to

- text reuse detection
- query formulation
- paraphrasing
Categorizing Crowdsourced Text Reuse
Categorizing Crowdsourced Text Reuse
Build-Up Versus Boil-Down Reuse

Build-up reuse (left) versus boil-down reuse (right).

- text length (y-axis) over text revision (x-axis)
- colors: different source documents (original text is white)
- blue dots: position of the writer’s last edit
Categorizing Crowdsourced Text Reuse

Build-Up Versus Boil-Down Reuse

Build-up reuse (left) versus boil-down reuse (right).

- text length (y-axis) over text revision (x-axis)
- colors: different source documents (original text is white)
- blue dots: position of the writer’s last edit
- Build-up: 45%; boil-down: 40%; mixed: 12%
Categorizing Crowdsourced Text Reuse

Build-Up Versus Boil-Down Reuse

Build-up reuse: Averaged editing histories by authors.

- one author per plot
- gray lines: individual essays
- black line: average
Categorizing Crowdsourced Text Reuse
Build-Up Versus Boil-Down Reuse

Author 2 (66 topics)

Author 7 (20 topics)

Author 24 (27 topics)

**Boil-down reuse:** Averaged editing histories by authors.

- one author per plot
- gray lines: individual essays
- black line: average
Categorizing Crowdsourced Text Reuse

Classification Scheme for Text Reuse.

- Find-Replace
- Remix

- Clone, Ctrl-C
- Mashup

Classification Scheme for Text Reuse.

- types of plagiarism as distinguished by Turnitin [Turnitin 2012]
Categorizing Crowdsourced Text Reuse
Classification Scheme for Text Reuse

Classification Scheme for Text Reuse.

- types of plagiarism as distinguished by Turnitin [Turnitin 2012]
- interpret text reuse (plagiarism) as a combination of two factors: paraphrasing and interleaving
Categorizing Crowdsourced Text Reuse
Classification Scheme for Text Reuse

- Quantify: N-Gram similarity and ratio of passages to sources [details]
- Measure for all essays
- Hypothesis: will show evidence of authors’ individual text reuse styles

Classification Scheme for Text Reuse.

- types of plagiarism as distinguished by Turnitin [Turnitin 2012]
- interpret text reuse (plagiarism) as a combination of two factors: paraphrasing and interleaving
Classification Scheme for Text Reuse.

- types of plagiarism as distinguished by Turnitin [Turnitin 2012]
- interpret text reuse (plagiarism) as a combination of two factors: paraphrasing and interleaving
Categorizing Crowdsourced Text Reuse
Classification Scheme for Text Reuse

Author (essay count)

Passages per source (interleaving)
Median 5-gram passage similarity (paraphrasing)
Categorizing Crowdsourced Text Reuse
Classification Scheme for Text Reuse

Median 5-gram passage similarity (paraphrasing)

Passages per source (interleaving)

Author (essay count)

© Michael Völske 2013
Categorizing Crowdsourced Text Reuse
Classification Scheme for Text Reuse

![Graph showing the relationship between passages per source and median 5-gram passage similarity.]

Author (essay count)
- A0 (2)
- A1 (3)
- A2 (66)
- A3 (1)
- A4 (1)
- A5 (37)
- A6 (11)
- A7 (20)
- A8 (1)
- A9 (1)
- A10 (12)
- A11 (1)
- A12 (1)
- A13 (1)
- A14 (1)
- A15 (3)
- A16 (1)
- A17 (33)
- A18 (32)
- A19 (7)
- A20 (8)
- A21 (21)
- A22 (1)
- A23 (1)
- A24 (27)
- A25 (1)
- A26 (1)

Passages per source (interleaving)

Median 5-gram passage similarity (paraphrasing)
- 0
- 0.2
- 0.4
- 0.6
- 0.8
- 1
Categorizing Crowdsourced Text Reuse

Classification Scheme for Text Reuse

![Graph showing the relationship between passages per source (interleaving) and median 5-gram passage similarity (paraphrasing).]

Key:
- ○○○ A0 (2)
- ▽▽▽ A1 (3)
- ◇◇◇ A2 (66)
- ▲▲▲ A3 (1)
- □□□ A4 (1)
- ▲▲▲ A5 (37)
- ◇◇◇ A6 (11)
- ▽▽▽ A7 (20)
- ◇◇◇ A8 (1)
- ▽▽▽ A9 (1)
- ◇◇◇ A10 (12)
- ▲▲▲ A11 (1)
- ▲▲▲ A12 (1)
- ▲▲▲ A13 (1)
- ▲▲▲ A14 (1)
- ▽▽▽ A15 (3)
- ◇◇◇ A16 (1)
- ▲▲▲ A17 (33)
- □□□ A18 (32)
- ▲▲▲ A19 (7)
- ◇◇◇ A20 (8)
- ▽▽▽ A21 (21)
- ◇◇◇ A22 (1)
- ▽▽▽ A23 (1)
- ◇◇◇ A24 (27)
- ▲▲▲ A25 (1)
- ▽▽▽ A26 (1)

© Michael Völske 2013
Categorizing Crowdsourced Text Reuse
Classification Scheme for Text Reuse

Author (essay count)

A0 (2)  
A1 (3)  
A2 (66) 
A3 (1)  
A4 (1)  
A5 (37) 
A6 (11) 
A7 (20) 
A8 (1)  
A9 (1)  
A10 (12)  
A11 (1)  
A12 (1)  
A13 (1)  
A14 (1)  
A15 (3)  
A16 (1)  
A17 (33) 
A18 (32) 
A19 (7)  
A20 (8)  
A21 (21) 
A22 (1)  
A23 (1)  
A24 (27) 
A25 (1)  
A26 (1)  
A27 (1)  
A28 (1)  
A29 (1)  
A30 (1)  
A31 (1)  
A32 (1)  
A33 (1)  
A34 (1)  
A35 (1)  
A36 (1)  
A37 (1)  
A38 (1)  
A39 (1)  
A40 (1)  
A41 (1)  
A42 (1)  
A43 (1)  
A44 (1)  
A45 (1)  
A46 (1)  
A47 (1)  
A48 (1)  
A49 (1)  
A50 (1)  
A51 (1)  
A52 (1)  
A53 (1)  
A54 (1)  
A55 (1)  
A56 (1)  
A57 (1)  
A58 (1)  
A59 (1)  
A60 (1)  
A61 (1)  
A62 (1)  
A63 (1)  
A64 (1)  
A65 (1)  
A66 (1)  
A67 (1)  
A68 (1)  
A69 (1)  
A70 (1)  
A71 (1)  
A72 (1)  
A73 (1)  
A74 (1)  
A75 (1)  
A76 (1)  
A77 (1)  
A78 (1)  
A79 (1)  
A80 (1)  
A81 (1)  
A82 (1)  
A83 (1)  
A84 (1)  
A85 (1)  
A86 (1)  
A87 (1)  
A88 (1)  
A89 (1)  
A90 (1)  
A91 (1)  
A92 (1)  
A93 (1)  
A94 (1)  
A95 (1)  
A96 (1)  
A97 (1)  
A98 (1)  
A99 (1)  
A100 (1) 

Copyright © Michael Völske 2013
Search Missions For Source Retrieval
## Search Missions For Source Retrieval

### Distribution of Queries Over Time

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>5</th>
<th>10</th>
<th>15</th>
<th>20</th>
<th>25</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>64</td>
<td>74</td>
<td>42</td>
<td>64</td>
<td>8</td>
<td>60</td>
</tr>
<tr>
<td>B</td>
<td>76</td>
<td>60</td>
<td>162</td>
<td>12</td>
<td>24</td>
<td>80</td>
</tr>
<tr>
<td>C</td>
<td>99</td>
<td>155</td>
<td>92</td>
<td>48</td>
<td>274</td>
<td>150</td>
</tr>
<tr>
<td>D</td>
<td>198</td>
<td>48</td>
<td>218</td>
<td>94</td>
<td>198</td>
<td>48</td>
</tr>
<tr>
<td>E</td>
<td>57</td>
<td>34</td>
<td>70</td>
<td>34</td>
<td>42</td>
<td>36</td>
</tr>
<tr>
<td>F</td>
<td>319</td>
<td>154</td>
<td>113</td>
<td>153</td>
<td>148</td>
<td>40</td>
</tr>
</tbody>
</table>

**Distribution of queries over time.**

- fraction of posed queries (y-axis) over elapsed time (x-axis) between the first query until essay completion
- each cell represents one of 150 essays
- the numbers denote the total amount of posed queries
- the cells are sorted by area under the curve
Search Missions For Source Retrieval

Distribution of Queries Over Time

Distribution of queries over time.

- fraction of posed queries (y-axis) over elapsed time (x-axis) between the first query until essay completion
- each cell represents one of 150 essays
- the numbers denote the total amount of posed queries
- the cells are sorted by area under the curve
Search Missions For Source Retrieval

Distribution of Queries Over Time

Distribution of queries over time.

- fraction of posed queries (y-axis) over elapsed time (x-axis) between the first query until essay completion
- each cell represents one of 150 essays
- the numbers denote the total amount of posed queries
- the cells are sorted by area under the curve
Correlation of editing and querying behavior.

- *dotted* averaged editing histories by authors  
- *solid* distribution of queries over time
Summary

1. Novel quality of the Webis-TRC-12 dataset of crowdsourced text reuse
2. Evidence of two fundamental editing strategies: build-up & boil-down
3. New classification scheme for documents in a text reuse corpus
4. Relationship between editing behavior and search engine use
Summary

1. Novel quality of the Webis-TRC-12 dataset of crowdsourced text reuse
2. Evidence of two fundamental editing strategies: build-up & boil-down
3. New classification scheme for documents in a text reuse corpus
4. Relationship between editing behavior and search engine use

Future Work

1. Interleaving and paraphrasing in the time dimension
2. Authors’ text reuse strategies across multiple documents
3. Paraphrasing study: track individual passages over time
Summary

1. Novel quality of the Webis-TRC-12 dataset of crowdsourced text reuse
2. Evidence of two fundamental editing strategies: build-up & boil-down
3. New classification scheme for documents in a text reuse corpus
4. Relationship between editing behavior and search engine use

Future Work

1. Interleaving and paraphrasing in the time dimension
2. Authors’ text reuse strategies across multiple documents
3. Paraphrasing study: track individual passages over time

Thank you for your attention!
References


Example topic:

*Obama’s family.*

Write about President Barack Obama’s family history, including genealogy, national origins, places and dates of birth, etc. Where did Barack Obama’s parents and grandparents come from? Also include a brief biography of Obama’s mother.
Example topic:

*Obama's family.*

Write about President Barack Obama's family history, including genealogy, national origins, places and dates of birth, etc. Where did Barack Obama's parents and grandparents come from? Also include a brief biography of Obama's mother.

Original topic 001 of the TREC Web Track 2009:

*Query.* obama family tree

*Description.* Find information on President Barack Obama’s family history, including genealogy, national origins, places and dates of birth, etc.

*Sub-topic 1.* Find the TIME magazine photo essay “Barack Obama’s Family Tree.”

*Sub-topic 2.* Where did Barack Obama’s parents and grandparents come from?

*Sub-topic 3.* Find biographical information on Barack Obama’s mother.
<table>
<thead>
<tr>
<th>Type</th>
<th>Rank</th>
<th>Frequency</th>
<th>Severity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clone</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Exact copy of another author's work</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mashup</td>
<td>2</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>A mix of material copied verbatim from several sources</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ctrl-C</td>
<td>3</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Significant portions of text copied from a single source</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Remix</td>
<td>4</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>Paraphrasing from several sources and making the content fit together seamlessly</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Recycle</td>
<td>5</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Self-plagiarism</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Re-Tweet</td>
<td>6</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Proper citation, but closely follows a single source</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Find-Replace</td>
<td>7</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>Near copy of a single source, with key phrases changed</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aggregator</td>
<td>8</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Proper citation, but (almost) no original work</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>404 Error</td>
<td>9</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Citations to non-existent or inaccurate information about sources</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hybrid</td>
<td>10</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>Combining properly cited sources with plagiarism in one paper</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

[Turnitin 2012]
Details: Paraphrasing & Interleaving

Classification Scheme for Text Reuse

How to quantify?

- Measure at the passage level
- Passage: Block of text reused from the same source
- Paraphrasing: simple N-Gram similarity

Classification Scheme for Text Reuse.

- types of plagiarism as distinguished by Turnitin [Turnitin 2012]
- interpret text reuse (plagiarism) as a combination of two factors: paraphrasing and interleaving
Three passages:

The Emperor penguins are the only penguins that inhabit the Antarctic continent and are the largest of all penguins. Adult Emperor penguins are typically 1.2 meters tall. Juveniles are slightly shorter, only about 90cm to 1m. Emperors weigh around 30 to 40 kg and their weight varies a great deal during the year. They can easily be recognized by their black cap, blue-grey neck, orange ear-patches and bills and yellow breasts. There is a thick layer of blubber under the Emperor’s skin which is covered by a dense layer of woolly down where an overlapping coat of feathers grows over. The outer feathers, however, are covered in a greasy waterproof coating.

Classification Scheme for Text Reuse.

- types of plagiarism as distinguished by Turnitin [Turnitin 2012]
- interpret text reuse (plagiarism) as a combination of two factors: paraphrasing and interleaving
Details: Paraphrasing & Interleaving

Classification Scheme for Text Reuse.

- types of plagiarism as distinguished by Turnitin [Turnitin 2012]
- interpret text reuse (plagiarism) as a combination of two factors: paraphrasing and interleaving
Classification Scheme for Text Reuse.

- types of plagiarism as distinguished by Turnitin [Turnitin 2012]
- interpret text reuse (plagiarism) as a combination of two factors: paraphrasing and interleaving
Classification Scheme for Text Reuse.

- types of plagiarism as distinguished by Turnitin [Turnitin 2012]
- interpret text reuse (plagiarism) as a combination of two factors: paraphrasing and interleaving
Details: Paraphrasing & Interleaving

Classification Scheme for Text Reuse

Paraphrasing: N-Gram similarity

\[ \phi_n(N_c, N_s) := \frac{|N_c \cap N_s|}{|N_c|} \]

- \( N_c \): N-Grams in the passage
- \( N_s \): N-Grams in the source

We choose \( n = 5 \).

Classification Scheme for Text Reuse.

- types of plagiarism as distinguished by Turnitin [Turnitin 2012]
- interpret text reuse (plagiarism) as a combination of two factors: paraphrasing and interleaving
Details: Paraphrasing & Interleaving
Classification Scheme for Text Reuse

- Find-Replace Remix
- Clone, Ctrl-C Mashup
- Interleaving

**Interleaving:** Passages per source

$$pps(C, S) := \frac{|C|}{|S|}$$

- $C$: passages
- $S$: sources

Classification Scheme for Text Reuse.

- types of plagiarism as distinguished by Turnitin [Turnitin 2012]
- interpret text reuse (plagiarism) as a combination of two factors: paraphrasing and interleaving