Crowdsourcing Interaction Logs to Understand Text Reuse from the Web

Martin Potthast      Matthias Hagen      Michael Völske      Benno Stein
Bauhaus-Universität Weimar
www.webis.de
Outline

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

Previous Text Reuse Corpora

- PAN-PC-09/10/11: >25,000 documents; >60,000 plagiarism cases each
- Automatically generated *artificial plagiarism*
- Automatic modifications do not preserve semantics
The Webis-TRC-12 Dataset
The Webis-TRC-12 Dataset

Construction Overview

Editor

ChainNoir 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
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 [details]
- 150 topics, 297 essays
- Target essay length: 5000 words
The Webis-TRC-12 Dataset
Construction Overview: Authors

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

Revision log
Query log
ClueWeb
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

<table>
<thead>
<tr>
<th>Author Demographics (n=12)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (Median)</td>
</tr>
<tr>
<td>Years Writing (Median)</td>
</tr>
<tr>
<td><strong>Academic degree</strong></td>
</tr>
<tr>
<td>Postgrad</td>
</tr>
<tr>
<td>Undergrad</td>
</tr>
<tr>
<td>None</td>
</tr>
<tr>
<td>n/a</td>
</tr>
<tr>
<td><strong>English</strong></td>
</tr>
<tr>
<td>Native</td>
</tr>
<tr>
<td>Second Language</td>
</tr>
</tbody>
</table>
The Webis-TRC-12 Dataset

Construction Overview: Sources

ClueWeb API

Revision log

Query log

Editor

ChatNoir SE

Author

ClueWeb

© Völske 2013
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

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

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 Stat
are the first First Family of African American descent in the United St
youngest to enter the White House since the Kennedys. Obama's you
family harks back today's of Camelot.

http://webis15.researcher.de/obamabio/obamabio?c1=12001179993&n3

In what follows, we give a detailed overview of Barack Obama's Far
light on himself, his immediate and extended family, including mate
relations. 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 Un
the first African American to hold the office. Obama was the junior/4
Senator from Illinois from 2005 until he resigned following his elec
The Webis-TRC-12 Dataset
Construction Overview: Search Engine

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

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, which is the first First Family of African American descent in the United States. The youngest of the Kennedy's, Obama's wife, was back then a part of Camelot.

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

Barack Obama

Barack Obama II 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.
Barack Obama's Family

plagiarized by John Doe

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

http://michi5.medien.uni-weimar.de/chaim/b/abschluss/10001170993@link=0911001.qpl

In what follows, we give a detailed overview of Barack Obama's Family. We shed light on himself, his immediate and extended family, including maternal and paternal relations. Moreover, we give insights into the relations of Michelle Obama, Barack 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 the junior United States Senator from Illinois from 2005 until he resigned following his election to the presidency. Obama is a graduate of Columbia University and Harvard Law School. He worked as a community organizer in Chicago prior to earning his law degree, and practiced as a civil rights attorney in Chicago before serving three terms in the Illinois Senate from 1997 to 2004. He also taught Constitutional Law at the University of Chicago Law School from 1992 to 2004. Following an unsuccessful bid for a seat in the U.S. House of Representatives in 2000, Obama was elected to the Senate in November 2004.

Barack Obama was born on 4 August 1961 at the Kapiolani Medical Center for Women & Children in Honolulu, Hawaii, to Ann Dunham of Wichita, Kansas and Barack Hussein Obama Sr. of Nyangoma-Kogelo, Nyanza Province, Kenya. His parents met while both were attending the East-West Center of the University of Hawaii at Manoa, where his father was enrolled as a foreign student. The couple married only a month earlier on February 2, 1961.

Childhood and Youth

When Barack Obama was two years old, his parents divorced and his father moved to Connecticut to continue his education before returning to Kenya. He saw his son only once more before dying in an automobile accident in 1982.

When Obama was six, his mother married Lolo Soetoro, an Indonesian oil manager. In 1970, when Soetoro's student visa was revoked because of political unrest in Indonesia, Dunham and Barack then in first grade, accompanied him to Jakarta.
The Webis-TRC-12 Dataset

Construction Overview: Editor

- Custom web-based rich text editor
- Records sources of re-used text passages
- New revision every 300ms of inactivity
- Detailed revision history

Barack Obama's Family

The Family of Barack Obama is an extended clan of African American, English, Indonesian, and Kenyan heritage. They are best known through the writings and political career of Barack Obama, the current President of the United States of America. His immediate family is the First Family of the United States. The Obi are the first First Family of African American descent in the United States and the youngest to enter the White House since the Kennedys. Obama's young, energetic family has drawn much attention from the media.

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 to the presidency. Obama is a graduate of Columbia University and Harvard Law School. He worked as a community organizer in Chicago prior to earning his law degree, and practiced as a civil rights attorney in Chicago before serving three terms in the Illinois Senate from 1997 to 2004. He also taught Constitutional Law at the University of Chicago Law School from 1992 to 2004. Following an unsuccessful bid for a seat in the U.S. House of Representatives in 2000, Obama was elected Senator in November 2004.

Barack Obama was born on 4 August 1961 at the Kapiolani Medical Center to Women & Children in Honolulu, Hawaii, to Ann Dunham of Wichita, Kansas and Barack Hussein Obama Sr. of Nyangoma-Kogelo, Nyanza Province, Kenya. His parents met while both were attending the East-West Center of the University of Hawaii at Manoa, where his father was enrolled as a foreign student. The couple married only a month earlier on February 3, 1961.

Childhood and Youth

When Barack Obama was two years old, his parents divorced and his father moved to Connecticut to continue his education before returning to Kenya. He saw his son only once more before dying in an automobile accident in 1982.

When Obama was six, his mother married Lolo Soetoro, an Indonesian oil manager. In 1967, when Soetoro's student visa was revoked because of political unrest in Indonesia, Dunham and Barack then in first grade, accompanied him to Indonesia.
The Webis-TRC-12 Dataset

Three Main Data Sources

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

Revision log
Query log
ClueWeb

Barack Obama's Family
plagiarized by John Doe

The Family of Barack Obama is an extended clan of African American 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 Kennedys. Obama's family harks back to days of Camels.

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

Barack Obama

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

Research Questions
The Webis-TRC-12 Dataset
Research Questions

1. Different text reuse approaches distinguishable?

2. Relation to existing text reuse categorizations?

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

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

- 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
Categorizing Crowdsourced 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

<table>
<thead>
<tr>
<th>Passages per source (interleaving)</th>
<th>Median 5-gram passage similarity (paraphrasing)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.2</td>
<td>0.2</td>
</tr>
<tr>
<td>0.4</td>
<td>0.4</td>
</tr>
<tr>
<td>0.6</td>
<td>0.6</td>
</tr>
<tr>
<td>0.8</td>
<td>0.8</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Author (essay count)

<table>
<thead>
<tr>
<th>Author</th>
<th>Essay Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>A0</td>
<td>(2)</td>
</tr>
<tr>
<td>A1</td>
<td>(3)</td>
</tr>
<tr>
<td>A2</td>
<td>(66)</td>
</tr>
<tr>
<td>A3</td>
<td>(1)</td>
</tr>
<tr>
<td>A4</td>
<td>(1)</td>
</tr>
<tr>
<td>A5</td>
<td>(37)</td>
</tr>
<tr>
<td>A6</td>
<td>(11)</td>
</tr>
<tr>
<td>A7</td>
<td>(20)</td>
</tr>
<tr>
<td>A8</td>
<td>(1)</td>
</tr>
<tr>
<td>A9</td>
<td>(1)</td>
</tr>
<tr>
<td>A10</td>
<td>(12)</td>
</tr>
<tr>
<td>A11</td>
<td>(1)</td>
</tr>
<tr>
<td>A12</td>
<td>(1)</td>
</tr>
<tr>
<td>A13</td>
<td>(1)</td>
</tr>
<tr>
<td>A14</td>
<td>(1)</td>
</tr>
<tr>
<td>A15</td>
<td>(3)</td>
</tr>
<tr>
<td>A16</td>
<td>(1)</td>
</tr>
<tr>
<td>A17</td>
<td>(33)</td>
</tr>
<tr>
<td>A18</td>
<td>(32)</td>
</tr>
<tr>
<td>A19</td>
<td>(7)</td>
</tr>
<tr>
<td>A20</td>
<td>(8)</td>
</tr>
<tr>
<td>A21</td>
<td>(21)</td>
</tr>
<tr>
<td>A22</td>
<td>(1)</td>
</tr>
<tr>
<td>A23</td>
<td>(1)</td>
</tr>
<tr>
<td>A24</td>
<td>(27)</td>
</tr>
<tr>
<td>A25</td>
<td>(1)</td>
</tr>
<tr>
<td>A26</td>
<td>(1)</td>
</tr>
</tbody>
</table>

© Völske 2013
Categorizing Crowdsourced Text Reuse

Classification Scheme for Text Reuse

Median 5-gram passage similarity (paraphrasing)

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

Passed per source (interleaving)

32 [V]
Categorizing Crowdsourced Text Reuse
Classification Scheme for Text Reuse

Passages per source (interleaving)

Median 5-gram passage similarity (paraphrasing)

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)
Search Missions For Source Retrieval
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
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
Search Missions For Source Retrieval

Correlation of Editing and Querying

Correlation of editing and querying behavior.

- —— averaged editing histories by authors  [plots]
- —— distribution of queries over time  [plots]
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
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!
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>3</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>2</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>9</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>5</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>10</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>7</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>4</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>6</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>8</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

Classify plagiarism as a combination of:
- Paraphrasing
- Interleaving

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

Classify plagiarism as:
- Find-Replace
- Remix
- Clone, Ctrl-C
- Mashup

[Turnitin 2012]
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
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

  \[ N = 2 \]

  ```plaintext
  Adult Emperor Penguins are ...
  { Adult Emperor
  Emperor Penguins
  Penguins are }
  ```

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

\[ N = 3 \]

\{ Adult Emperor Penguins are ... \}

Interleaving

- low
- high

Paraphrasing
- low
- high

Find-Replace
Remix

Clone, Ctrl-C
Mashup

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

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

- \( N_c \): 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

Interleaving: Passages per source

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

- $C$: 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