

# 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

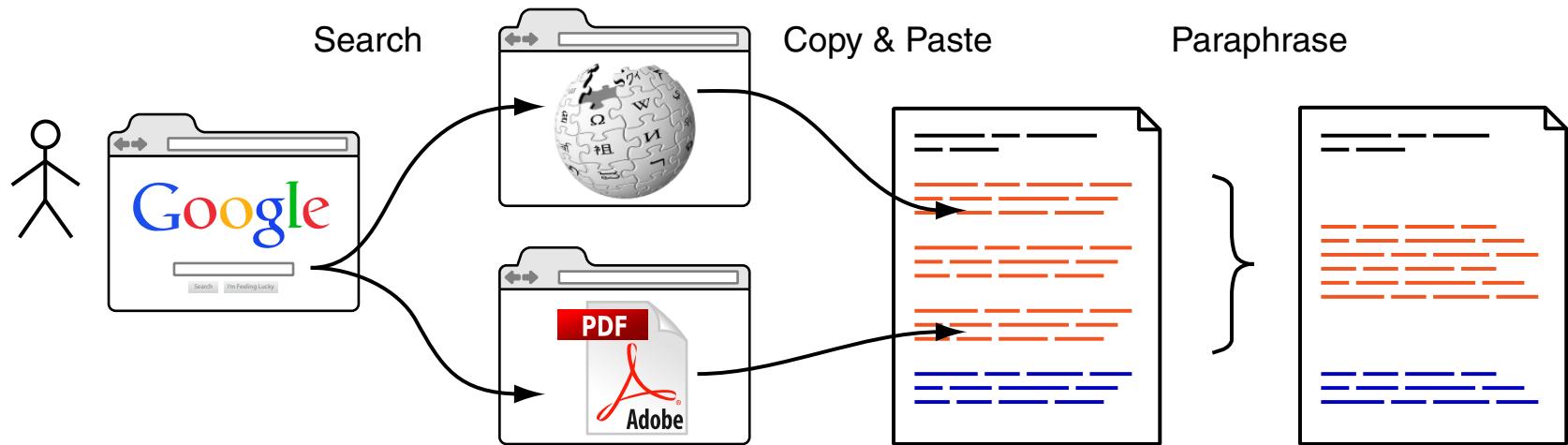
Master's Thesis

---

- Outline
- Introduction
  - The Webis-TRC-12 Dataset
  - Categorizing Crowdsourced Text Reuse
  - Search Missions For Source Retrieval
  - Summary

# Introduction

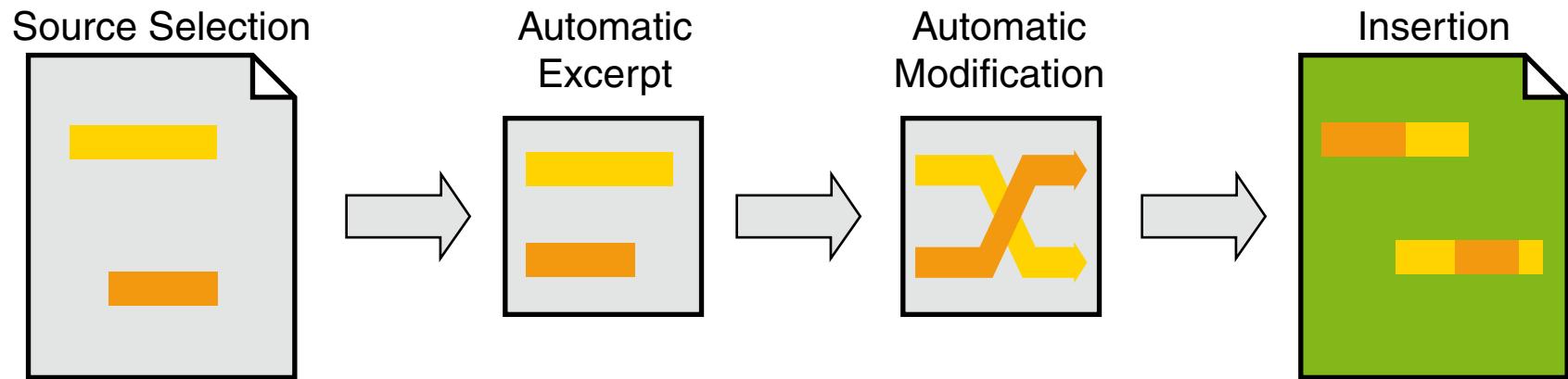
## Modeling Text Reuse From the Web



[Potthast 2011]

# Introduction

## Previous Plagiarism Corpora

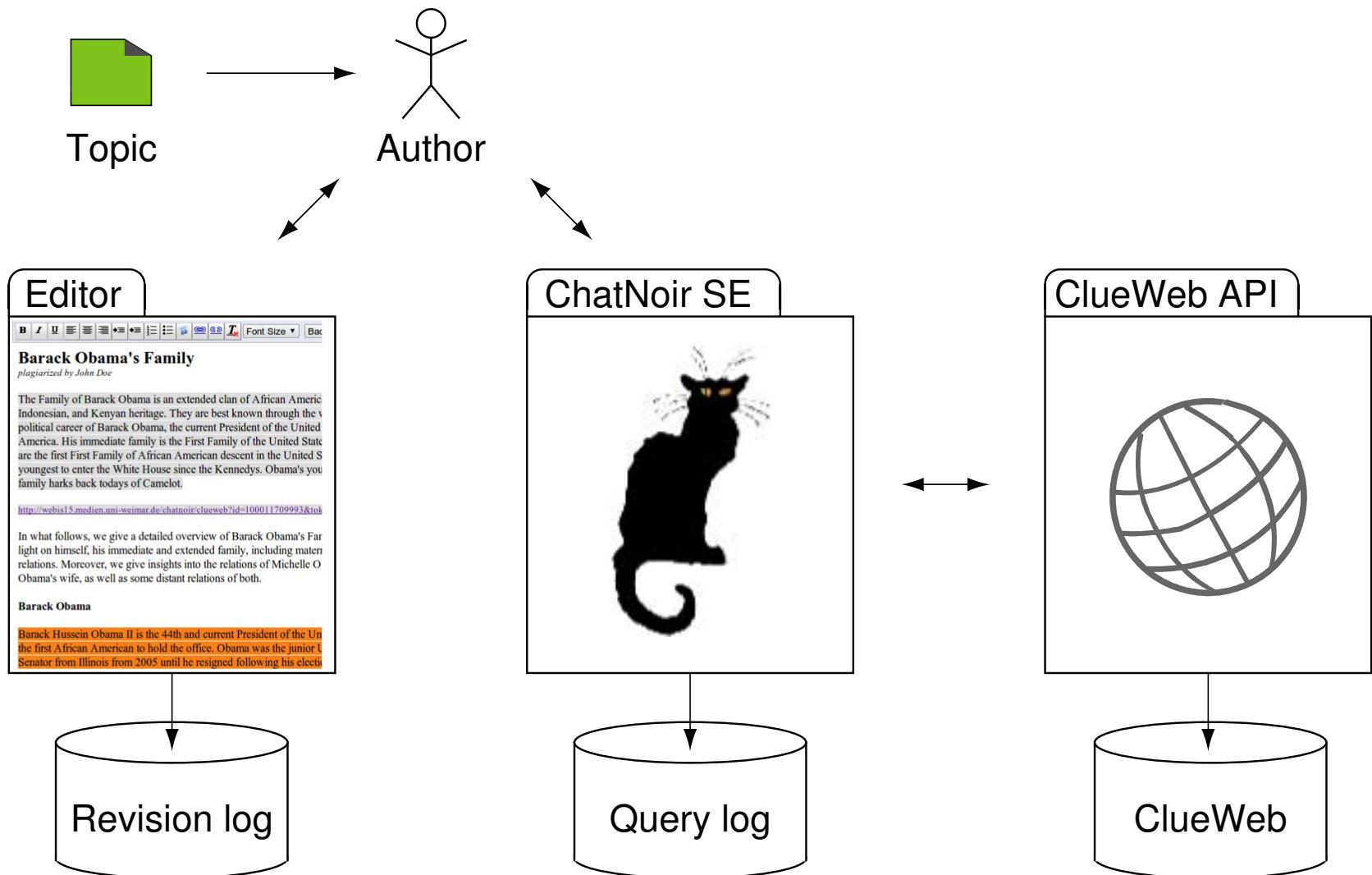


- Automatically generated *artificial plagiarism*
- PAN-PC-10/11: >25.000 documents; >60.000 plagiarism cases
- Automatic transformations do not preserve semantics

# The Webis-TRC-12 Dataset

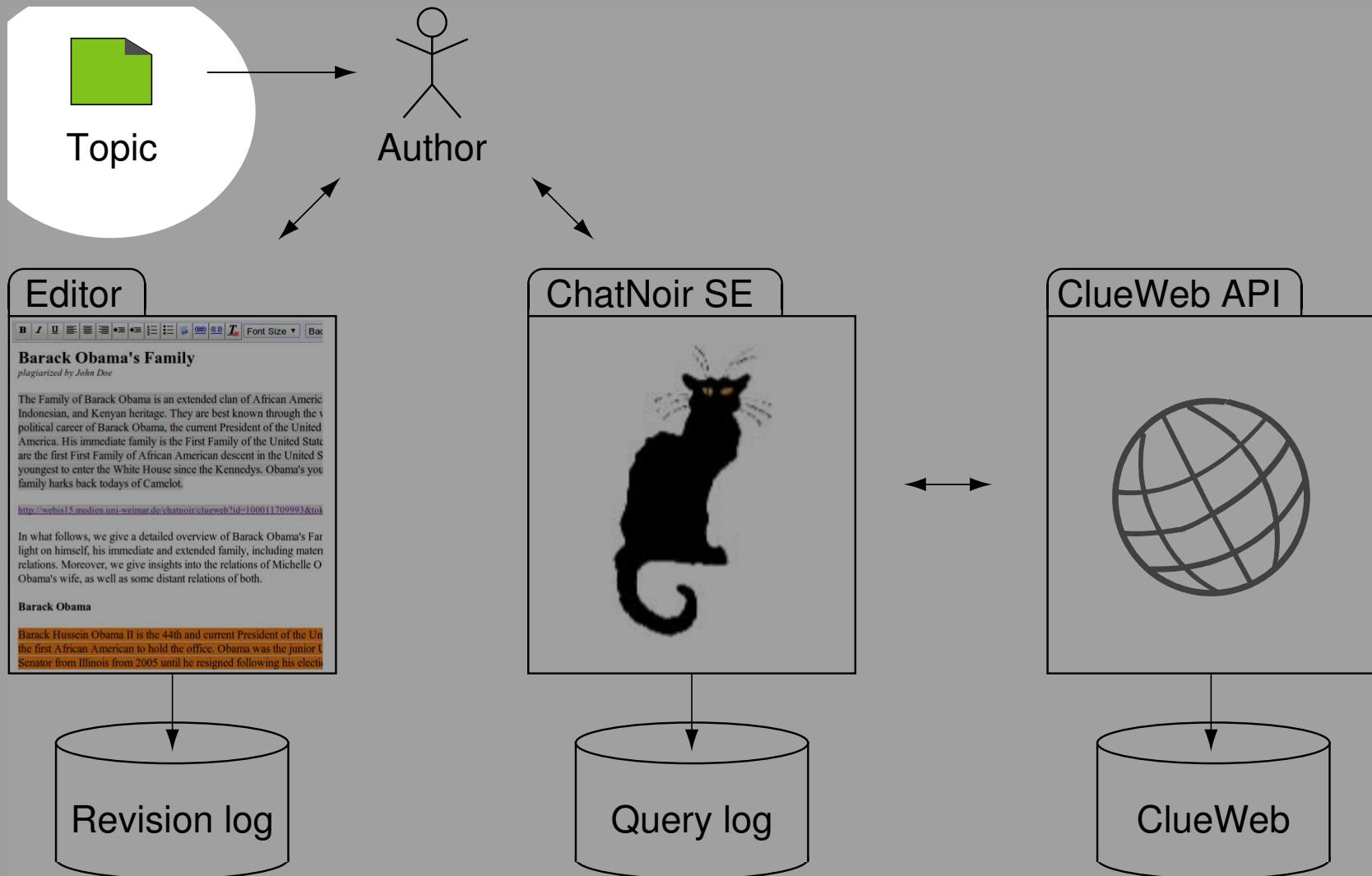
# The Webis-TRC-12 Dataset

## Construction Overview



# The Webis-TRC-12 Dataset

## Construction Overview: Topics



# The Webis-TRC-12 Dataset

## Construction Overview: Topics

Example topic:

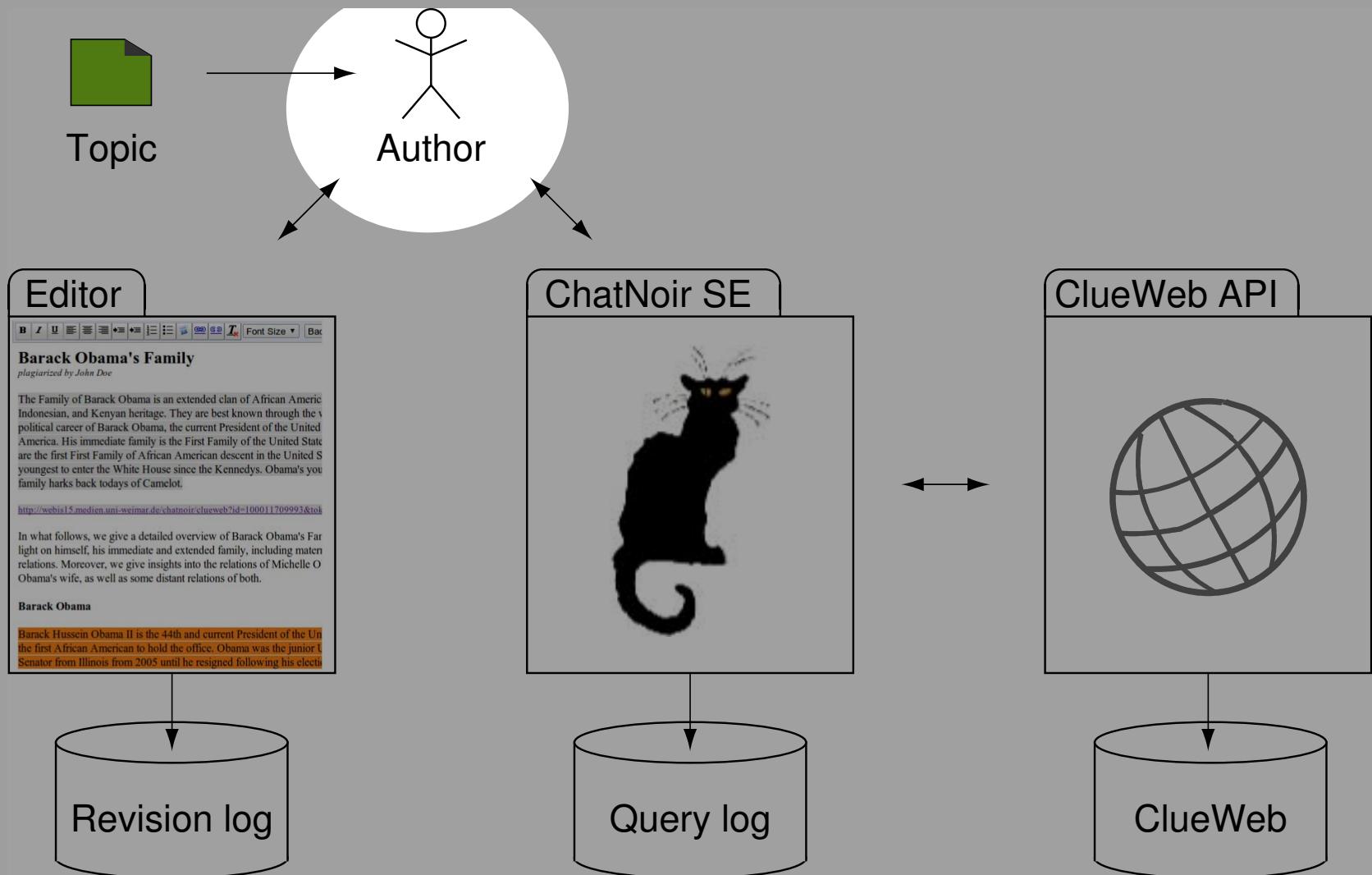
*Obama's family.*

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

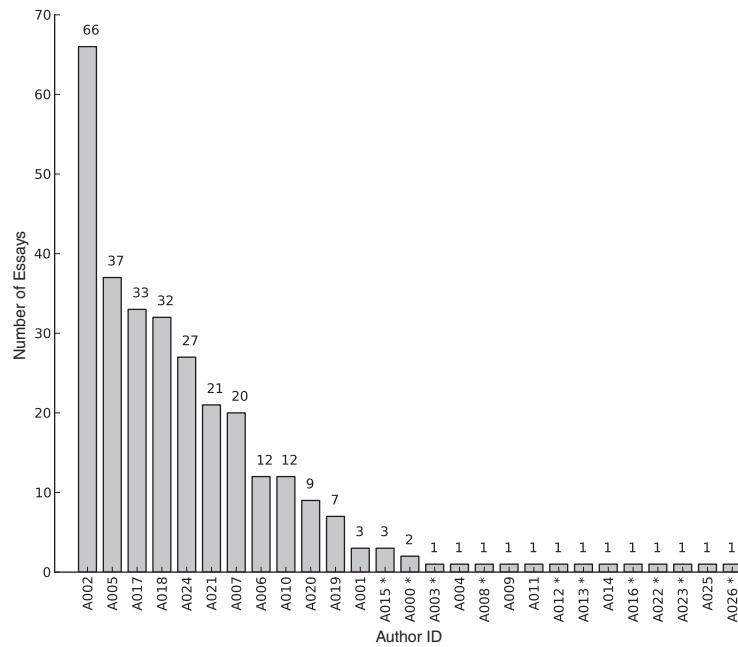
# The Webis-TRC-12 Dataset

## Construction Overview: Authors



# The Webis-TRC-12 Dataset

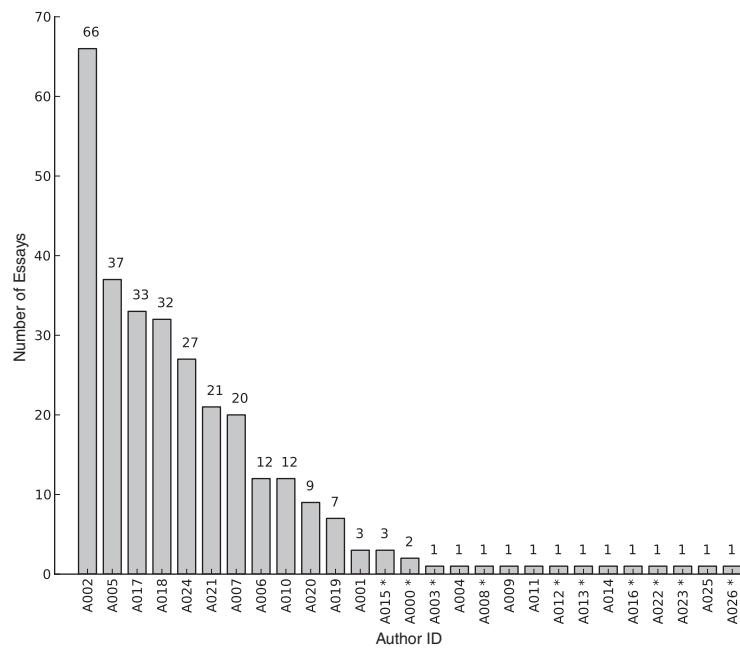
## Construction Overview: Authors



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

# The Webis-TRC-12 Dataset

## Construction Overview: Authors



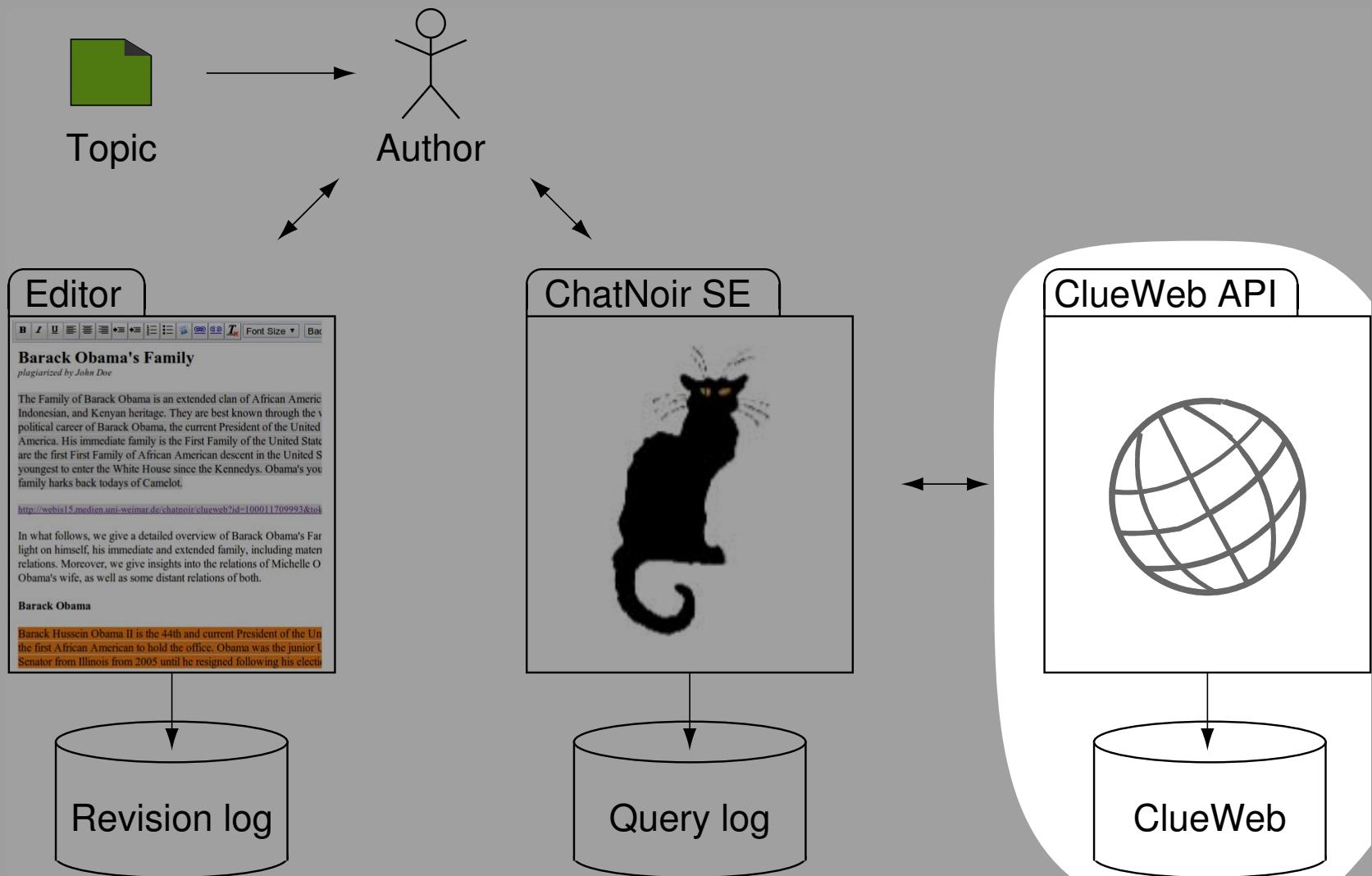
### Author Demographics (n=12)

Age (Median)	37
Years Writing (Median)	8
<i>Academic degree</i>	
Postgrad	33%
Undergrad	25%
<i>English</i>	
Native	67%
Second Language	33%

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

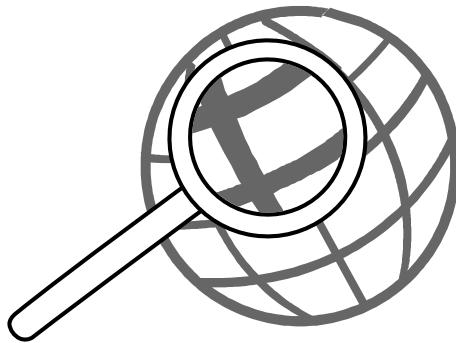
# The Webis-TRC-12 Dataset

## Construction Overview: Sources



# The Webis-TRC-12 Dataset

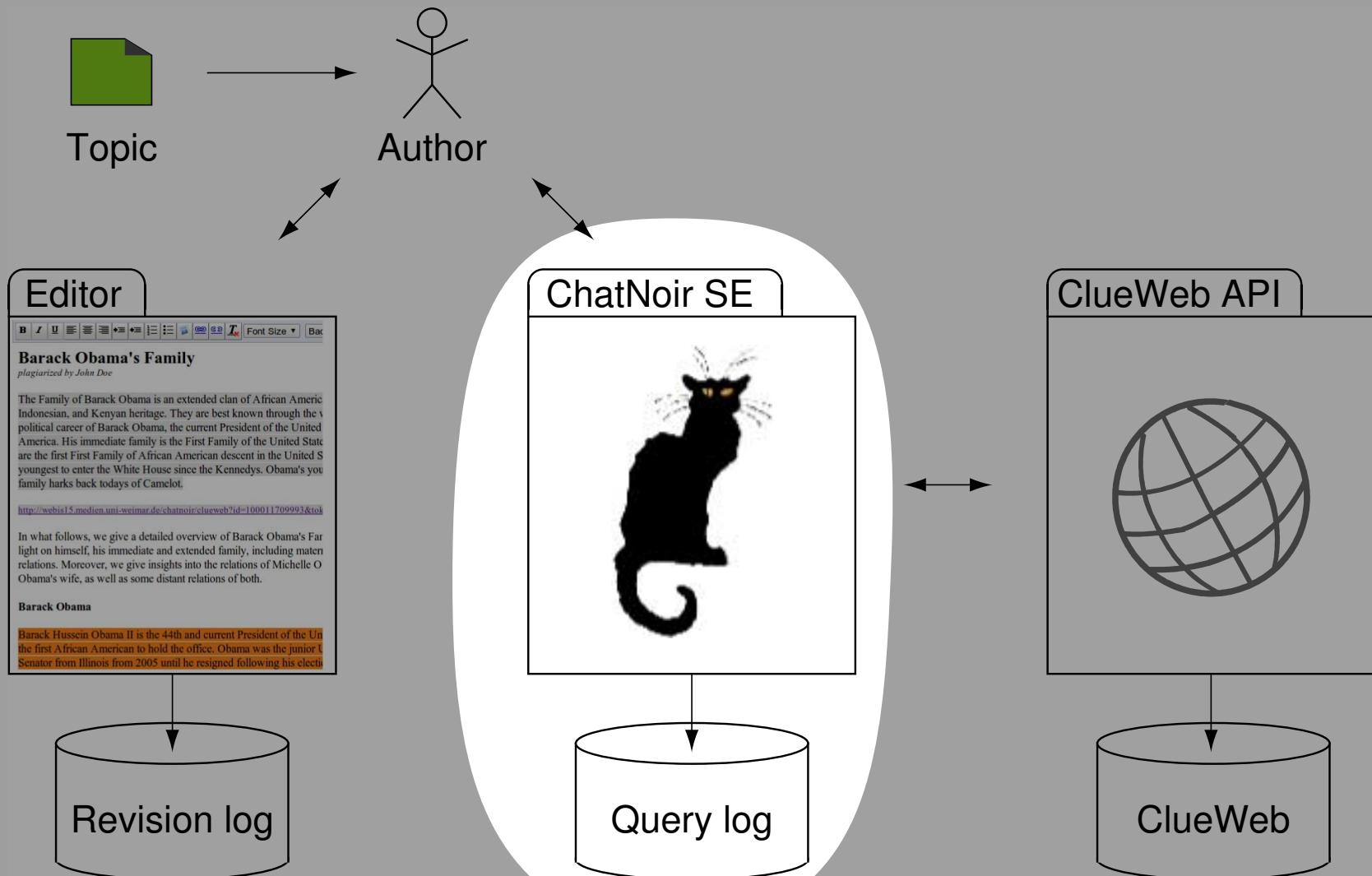
## Construction Overview: Sources



- ❑ 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



# The Webis-TRC-12 Dataset

## Construction Overview: Search Engine

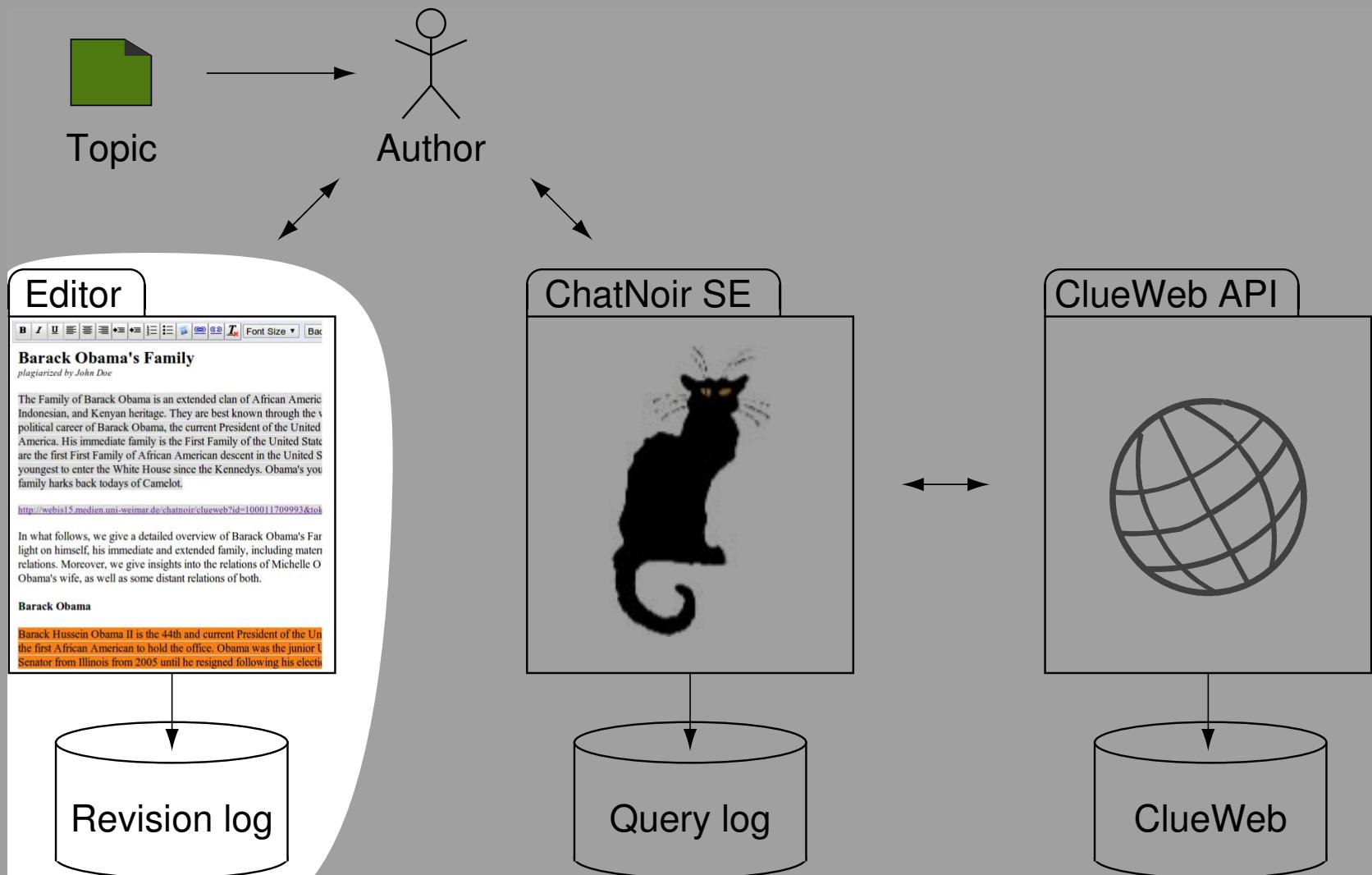
[Potthast et al. 2012a]



- ❑ Used for source retrieval
- ❑ Indexes ClueWeb
- ❑ Records fine-grained interaction log [example]

# The Webis-TRC-12 Dataset

## Construction Overview: Editor



# The Webis-TRC-12 Dataset

## Construction Overview: Editor

[Potthast et al. 2012b]

The screenshot shows a web-based editor titled "Plagiarism Editor". The main area contains a rich text editor with various styling tools (bold, italic, underline, etc.) and a color palette. A sample text is displayed:

**Barack Obama's Family**  
plagiarized by John Doe

The text describes Barack Obama's family background, mentioning his African American, English, Indonesian, and Kenyan heritage, and his role as the current President of the United States.

Below the text is a URL: <http://webis15.medien.uni-weimar.de/chatnoir/clueweb?tid=100011709993&token=w10911001-qrel>

The instructions on the right side of the editor provide steps for plagiarizing:

- Search for sources matching the topic using the [ChatNoir search engine](#). Do not use any other search engine!
- Once you found a passage of text to plagiarize, copy it into your text.
- Change the background color of the copied passage. Also, add a link to the source web page with the same background color. This is so we can follow up on your work.
- Modify the plagiarized passage so that an automatic plagiarism detector (like Turnitin) won't be able to detect it.
- Repeat these steps until your text is complete.

Remarks:

- The text shall be at least 5000 words long.
- It shall contain a couple of plagiarized passages.
- You shall also write some passages yourself.
- You may choose the text genre: an essay, a news article, a press release, a blog post, an advertisement etc.
- You may follow links on web pages found via the search engine.
- While modifying and rewriting a plagiarized passage, you may mix it with others, delete things, or add sentences.

Use the editor on the left to write your text. Do not use any other editor. Your text will be frequently saved on our servers. In case of errors, you will be notified in the status message below. Report errors back to us before you continue writing.

**Topic**

Obama's family tree. Write about President Barack Obama's family history, including genealogy, national origins, places and dates of birth, etc. For instance: where did Barack Obama's parents and grandparents come from; what did his mother work; etc.

**Links**      **Contact**

[ChatNoir Search](#)      [pan@webis.de](mailto:pan@webis.de)

**Status**      **Word Count**

Document saved      7955

**Color Key**

4: clueweb?tid=100011709993&token=w10911001-qrel  
5: clueweb?tid=1000010185058&token=w10911001-qrel  
6: clueweb?tid=0025990694&token=w10911001-qrel  
8: clueweb?tid=1000179002028&token=w10911001-qrel  
9: clueweb?tid=100007704609&token=w10911001-qrel  
10: clueweb?tid=100012205652&token=w10911001-qrel  
12: clueweb?tid=100013617161&token=w10911001-qrel  
17: clueweb?tid=00988616391&token=w10911001-qrel  
20: clueweb?tid=00010221241&token=w10911001-qrel  
24: clueweb?tid=100009959211&token=w10911001-qrel

# The Webis-TRC-12 Dataset

## Construction Overview: Editor

[Potthast et al. 2012b]

The screenshot shows the 'Plagiarism Editor' window. On the left, there's a rich text editor with a toolbar for bold, italic, underline, etc. Below it is a preview area containing text about 'Barack Obama's Family'. The text is attributed to 'plagiarized by John Doe'. The main workspace has a title 'Instructions' with the following text: 'Write a text about the topic specified below. The text shall contain passages which are plagiarized from different web pages.' Below this are three numbered steps: 1. Search for sources matching the topic using the Chat Noir search engine. Do not use any other search engine! 2. Once you found a passage of text to plagiarize, copy it into your text. 3. Change the background color of the copied passage. Also, add a link to the source web page with the same family harks back today of Camelot.

In the preview area, the text starts with '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 Obas 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 harks back today of Camelot.'

Below the preview, there's a URL: <http://webis15.medien.uni-weimar.de/chatnoir/clueweb?tid=100011709993&token=wto91100110>.

The text continues: 'In what follows, we give a detailed overview of Barack Obama's family. We shall focus on 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.'

A section titled 'Barack Obama' follows, with a detailed biography of his life and political career.

Another section discusses Barack Obama's birth and early life in Hawaii.

A 'Color Key' table at the bottom lists 24 numbered items, each associated with a specific color and a URL:

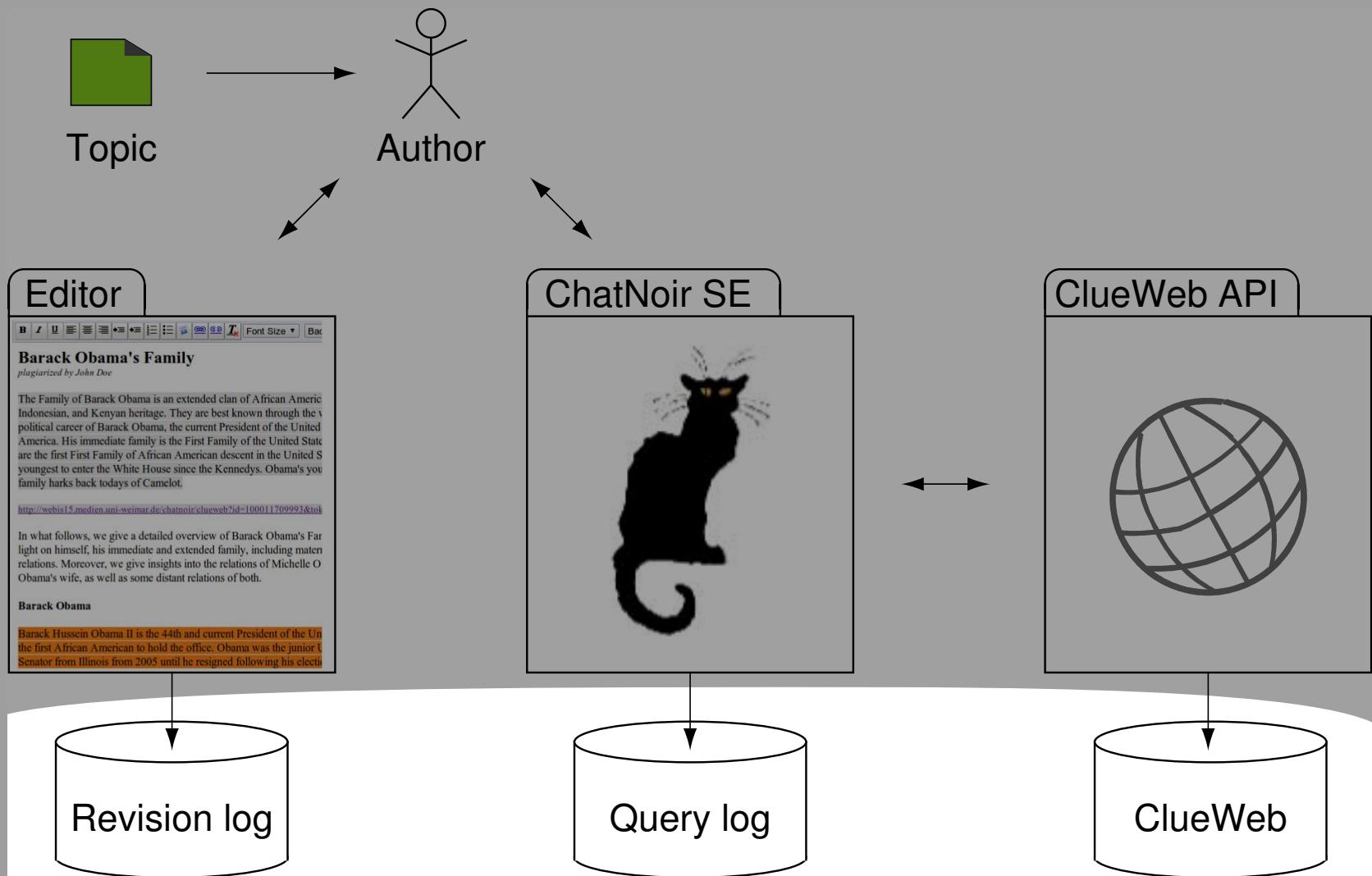
Status	Word Count
Document saved	7955

Color Key
4: clueweb?id=100011709993&token=wto911001-qrel
5: clueweb?id=100001018505&token=wto911001-qrel
6: clueweb?id=0025990694&token=wto911001-qrel
8: clueweb?id=1000179000208&token=wto911001-qrel
9: clueweb?id=100007704609&token=wto911001-qrel
10: clueweb?id=100012205624&token=wto911001-qrel
16: clueweb?id=100013617161&token=wto911001-qrel
17: clueweb?id=00988616391&token=wto911001-qrel
20: clueweb?id=00010221241&token=wto911001-qrel
24: clueweb?id=1000009050214&token=wto911001-qrel

- 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



# 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?

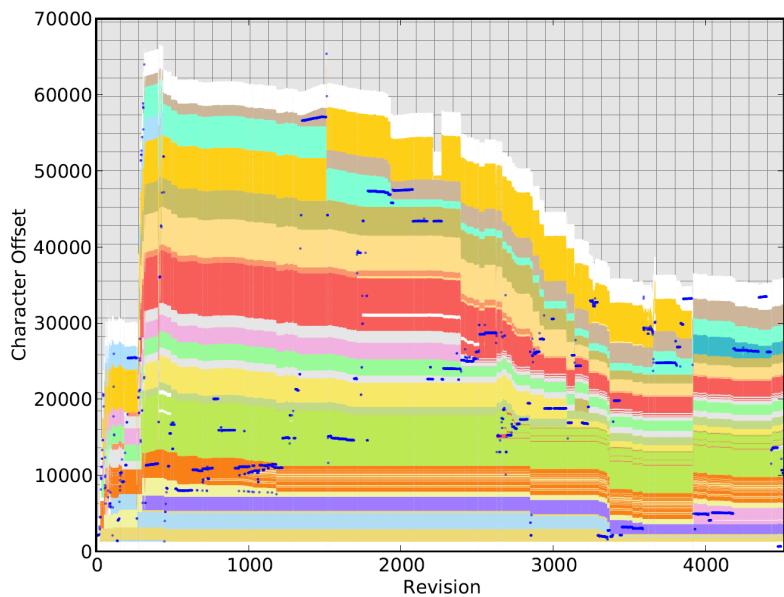
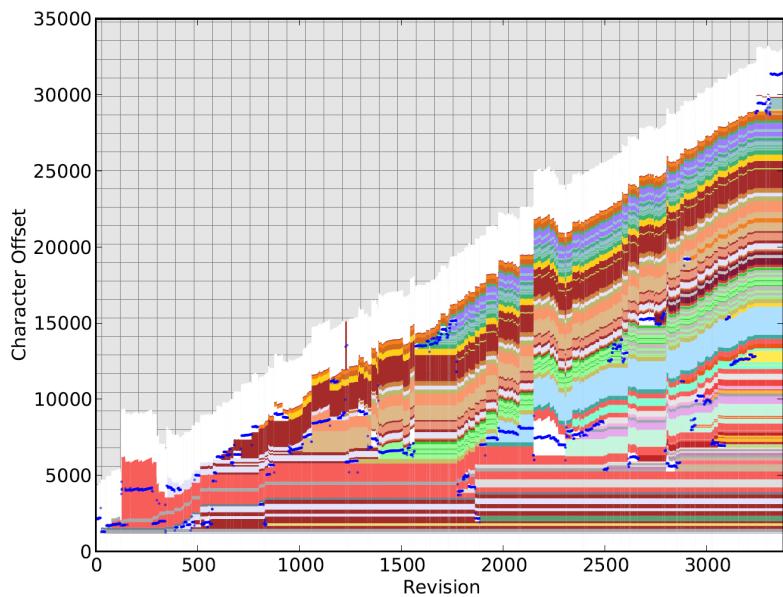
We expect new research insights and impacts to

- text reuse detection
- query formulation
- paraphrasing

# Categorizing Crowdsourced Text Reuse

# Categorizing Crowdsourced Text Reuse

## Build-Up Versus Boil-Down Reuse

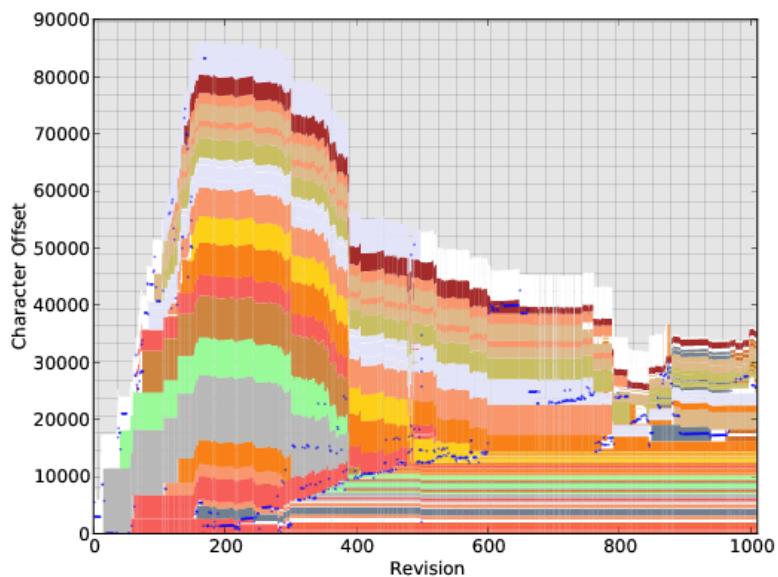
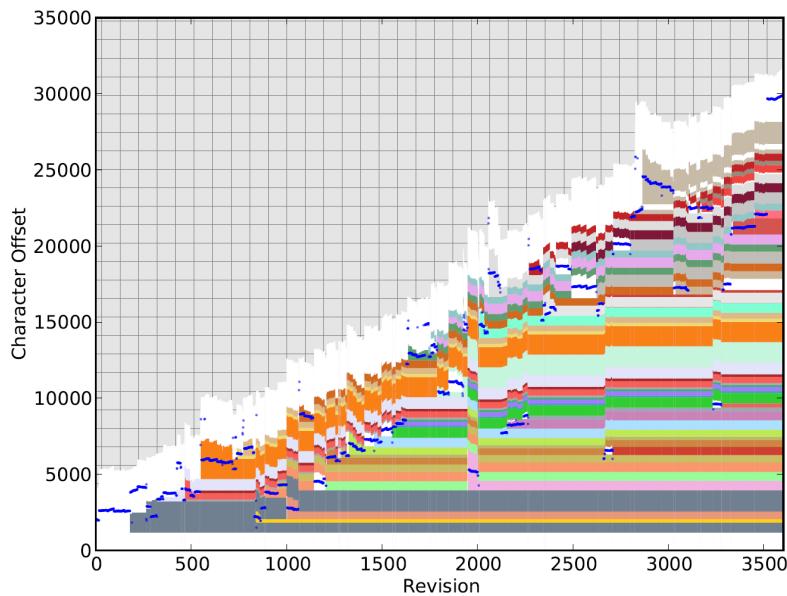


Build-up reuse (left) versus boil-down reuse (right).

- 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



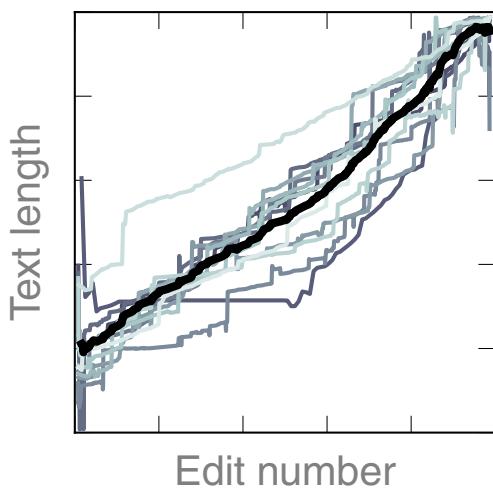
Build-up reuse (left) versus boil-down reuse (right).

- 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%

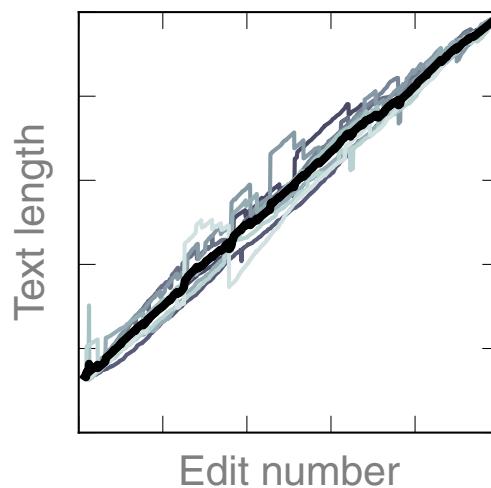
# Categorizing Crowdsourced Text Reuse

## Build-Up Versus Boil-Down Reuse

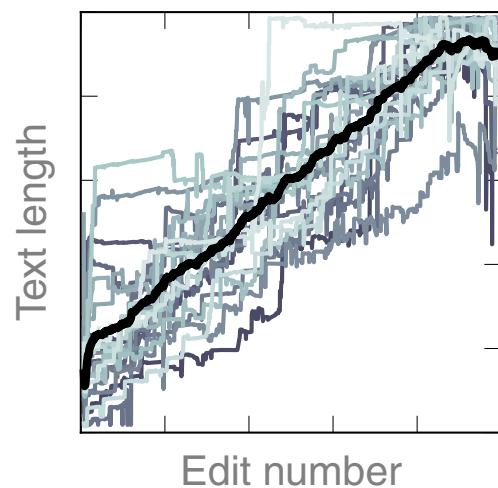
Author 6 (12 topics)



Author 20 (9 topics)



Author 21 (21 topics)



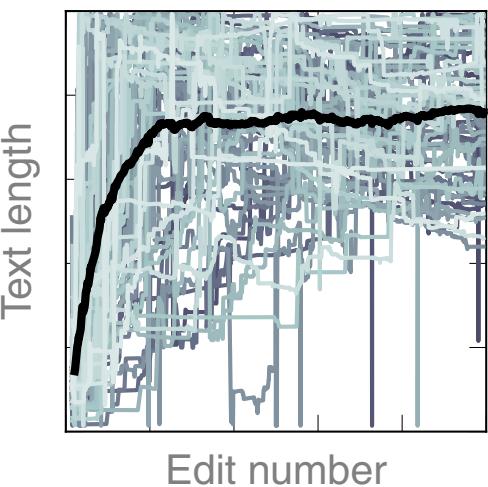
Build-up reuse: Averaged editing histories by authors.

- one author per plot
- gray lines: individual essays
- black line: average

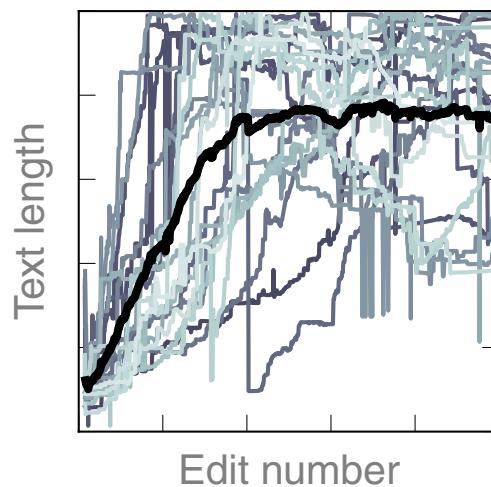
# Categorizing Crowdsourced Text Reuse

## Build-Up Versus Boil-Down Reuse

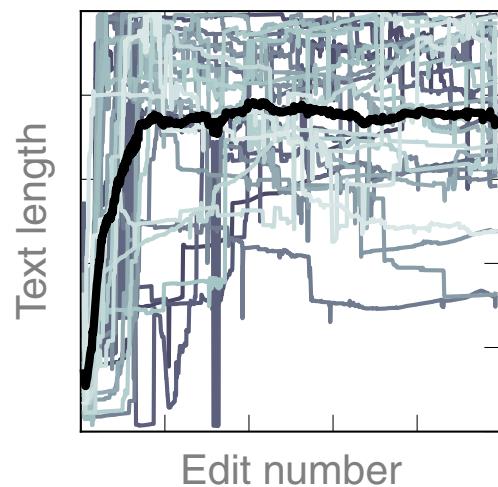
Author 2 (66 topics)



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

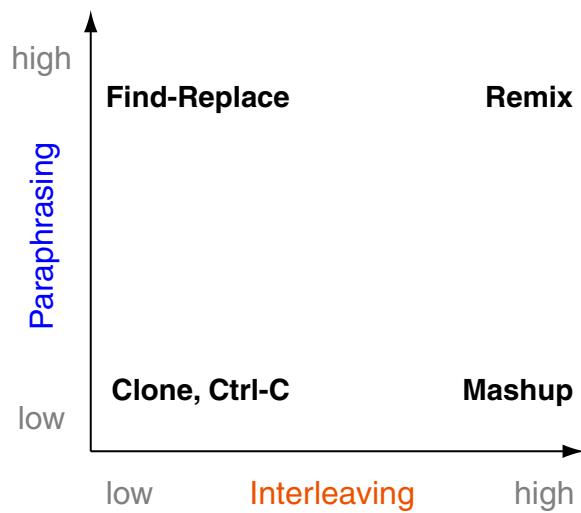
Mashup

## Classification Scheme for Text Reuse.

- ❑ types of plagiarism as distinguished by Turnitin [Turnitin 2012]

# Categorizing Crowdsourced Text Reuse

## Classification Scheme for Text Reuse

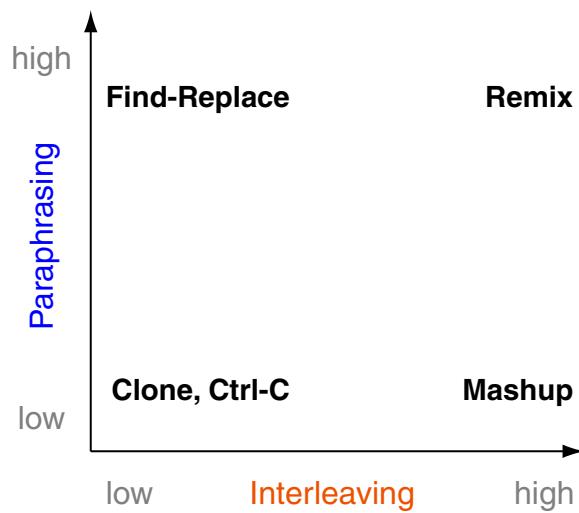


## Classification Scheme for Text Reuse.

- 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



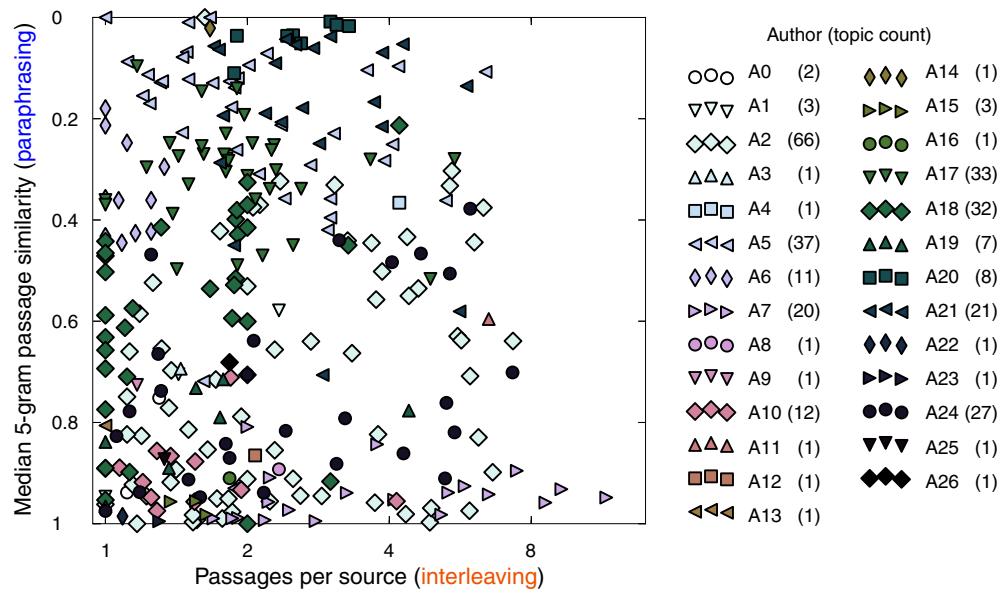
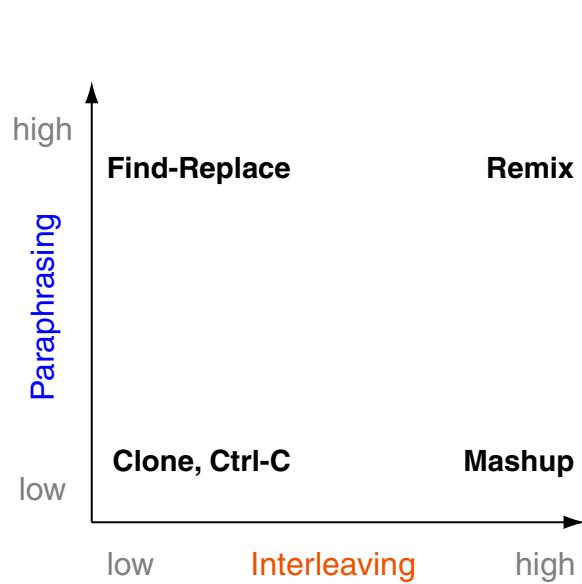
- Quantify: **N-Gram similarity** and **ratio of passages to sources**  
[details]
- Measure for all essays
- Hypothesis: will show evidence of authors' individual text reuse styles

## 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

# Categorizing Crowdsourced Text Reuse

## Classification Scheme for Text Reuse

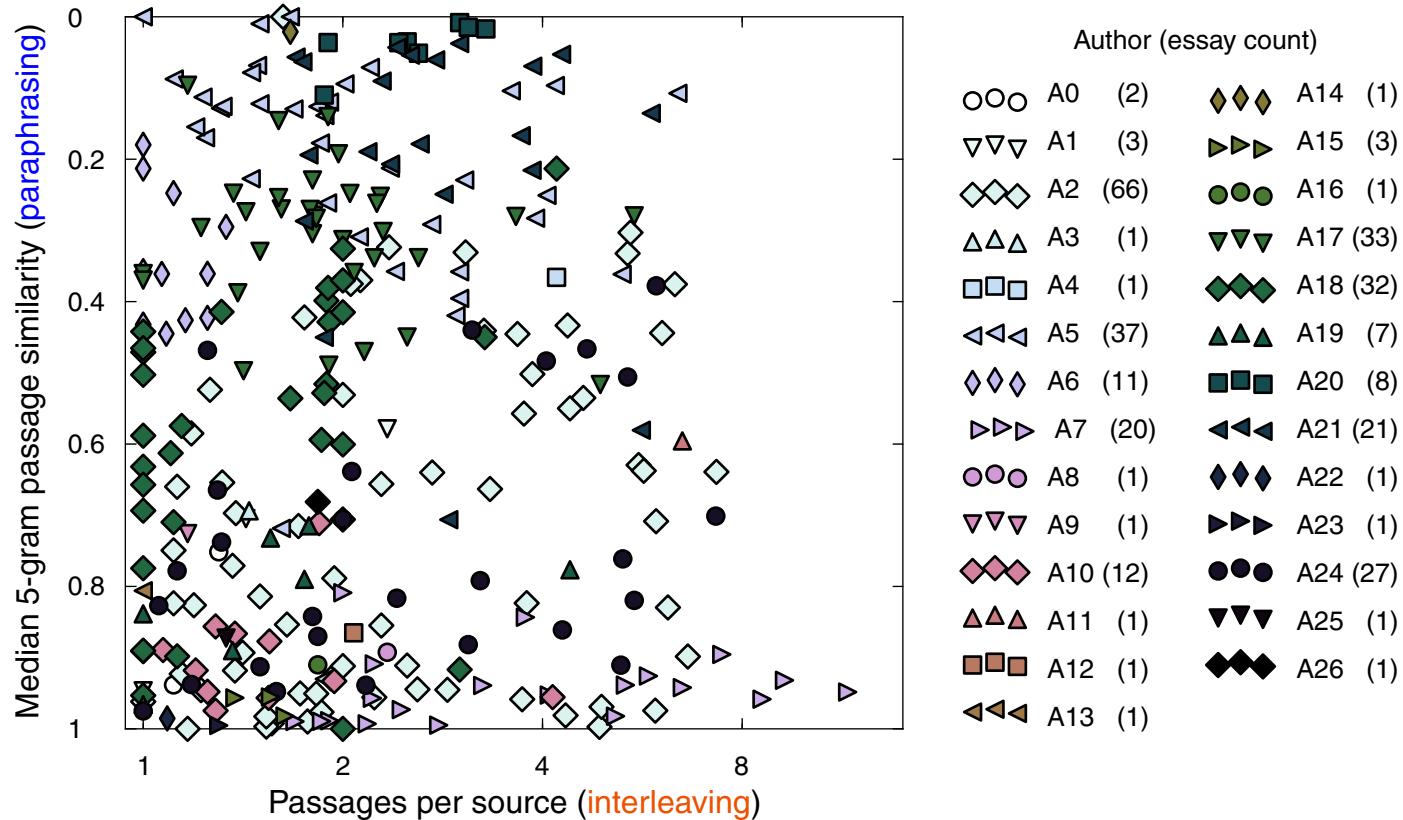


## Classification Scheme for Text Reuse.

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

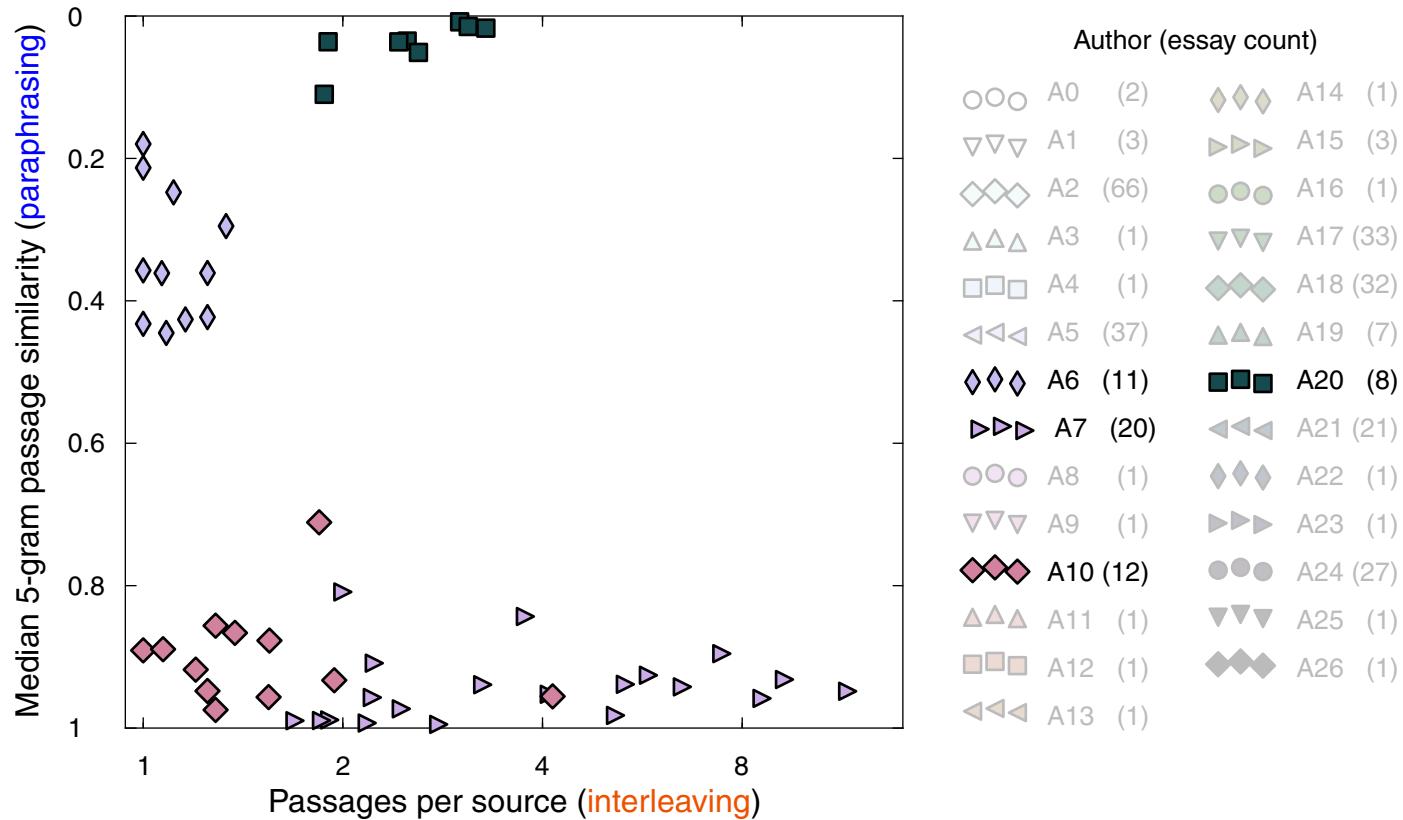
# Categorizing Crowdsourced Text Reuse

## Classification Scheme for Text Reuse



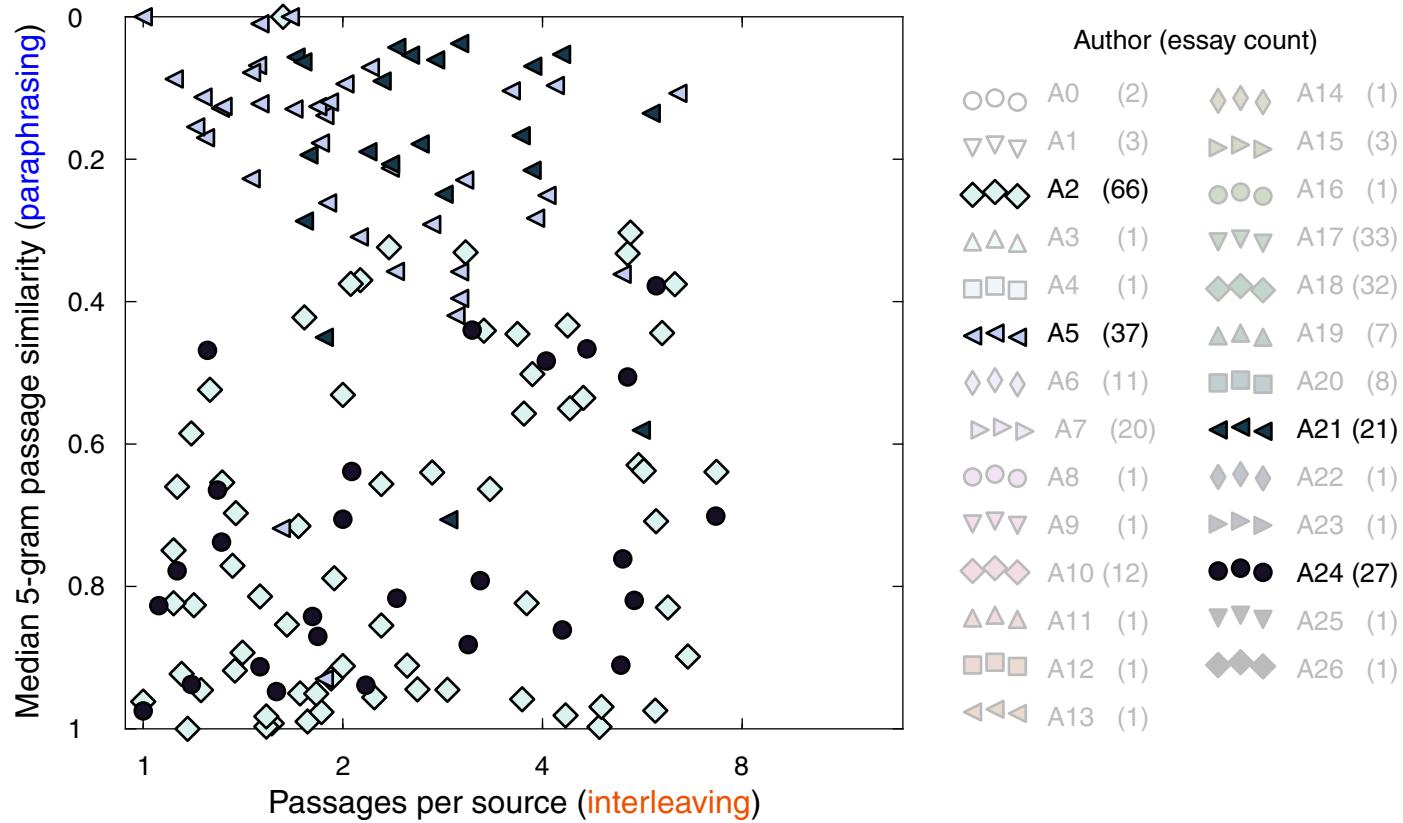
# Categorizing Crowdsourced Text Reuse

## Classification Scheme for Text Reuse



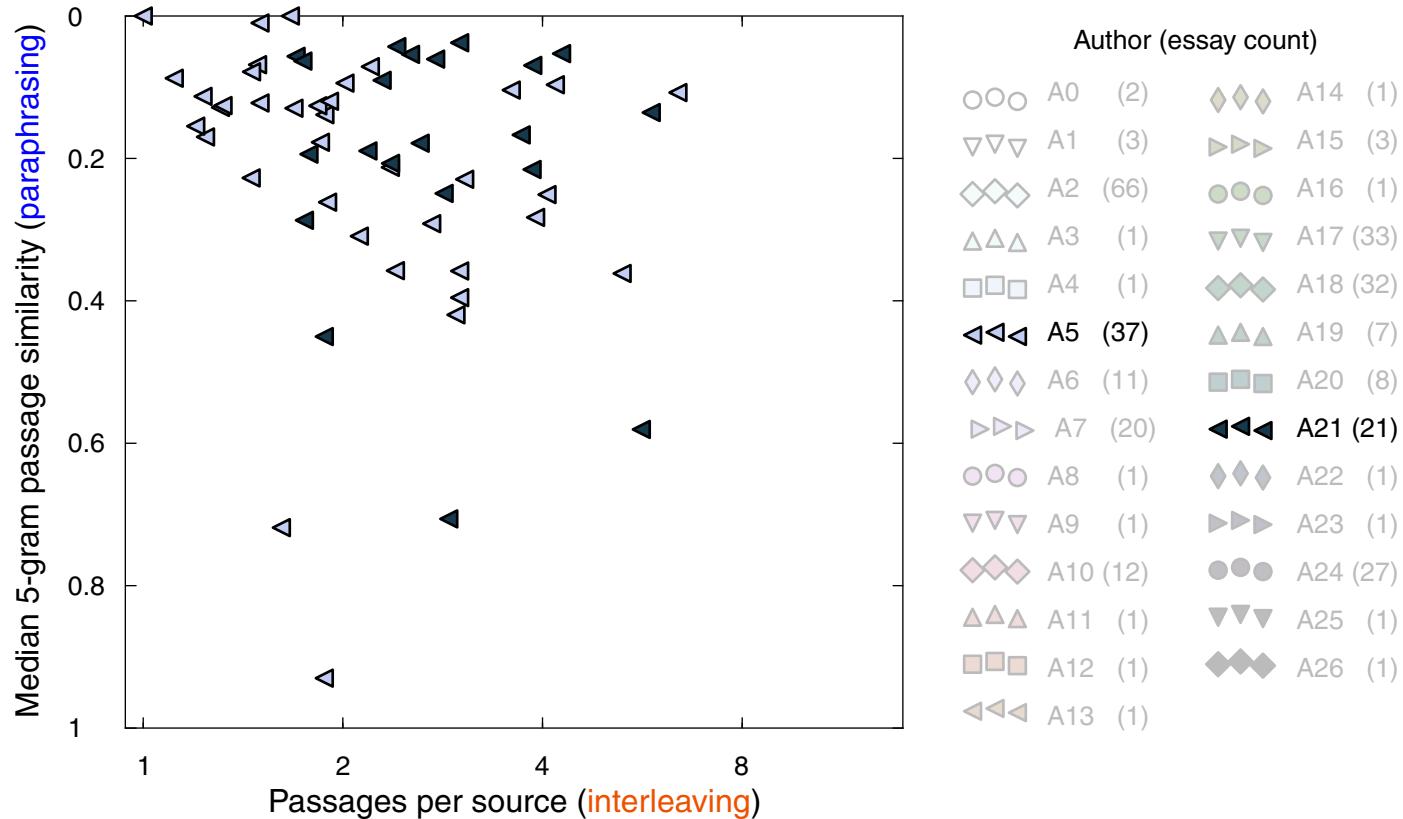
# Categorizing Crowdsourced Text Reuse

## Classification Scheme for Text Reuse



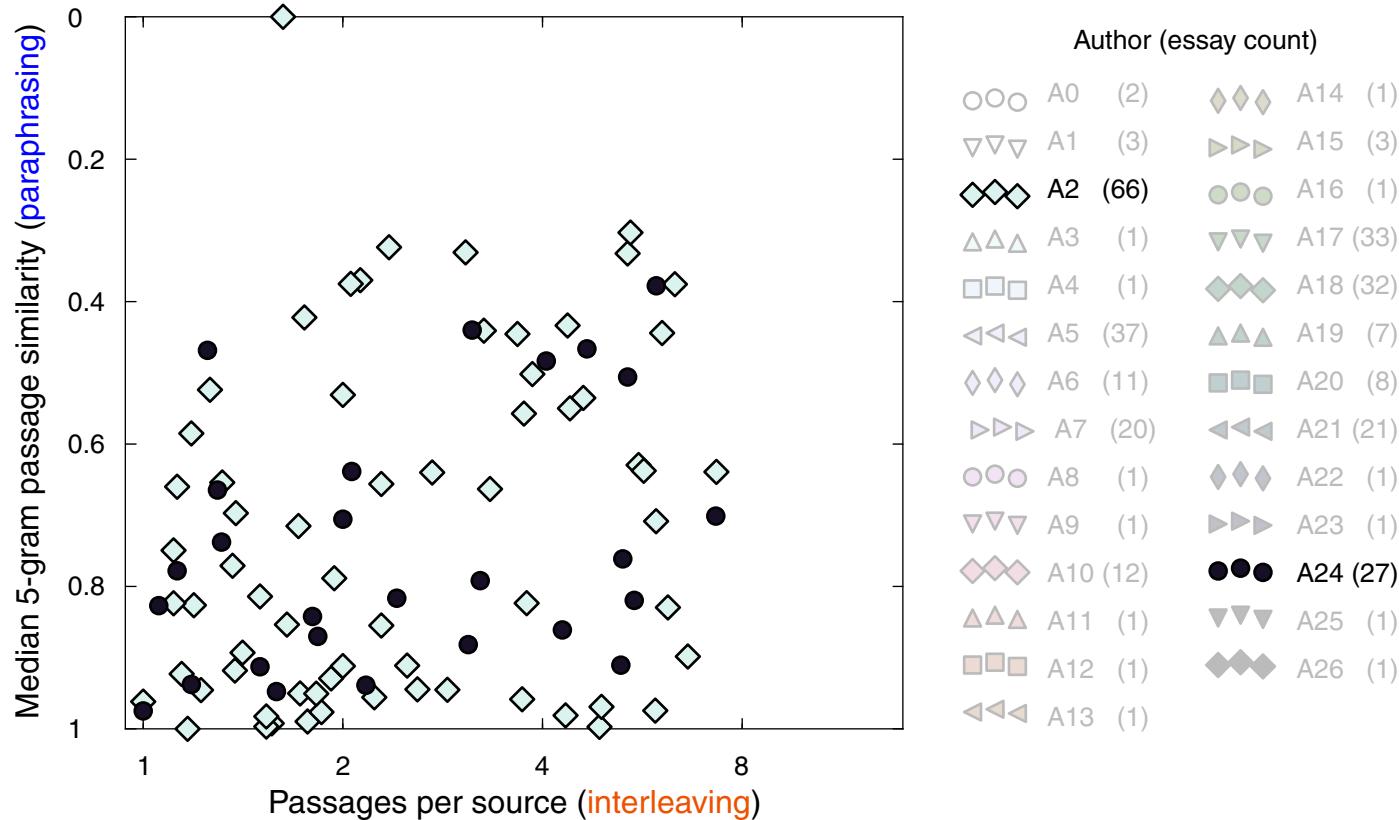
# Categorizing Crowdsourced Text Reuse

## Classification Scheme for Text Reuse



# Categorizing Crowdsourced Text Reuse

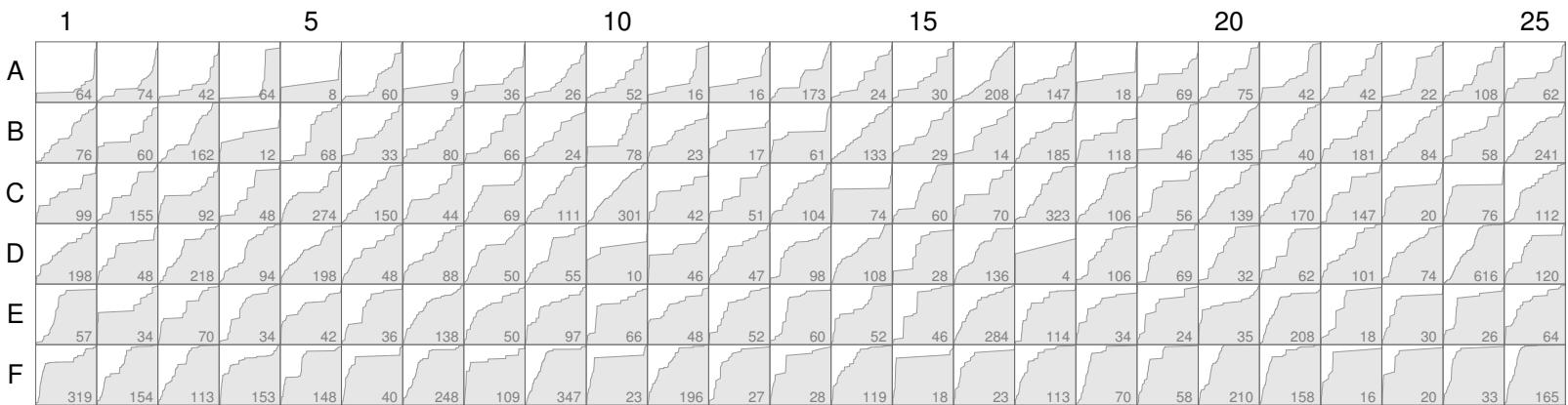
## Classification Scheme for Text Reuse



# Search Missions For Source Retrieval

# Search Missions For Source Retrieval

## Distribution of Queries Over Time

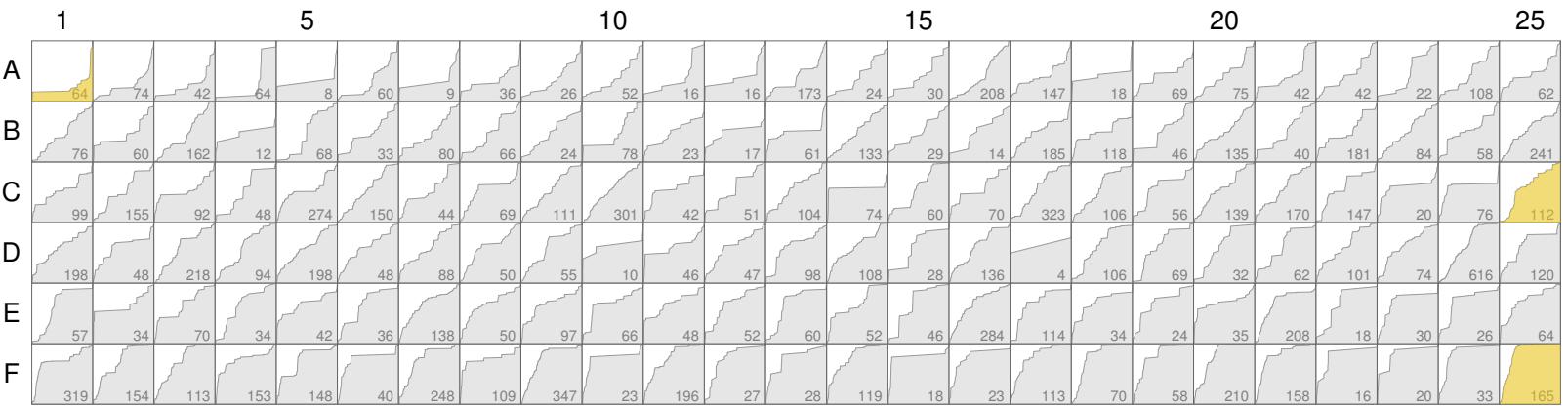


Distribution of queries over time.

- ❑ 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

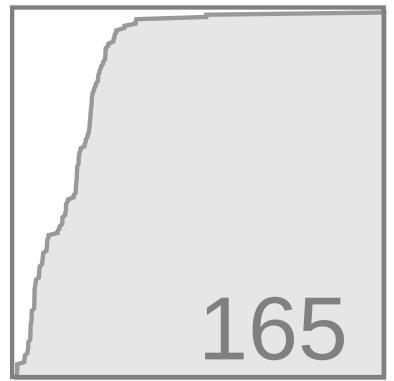
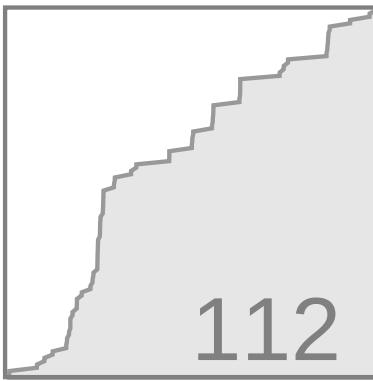
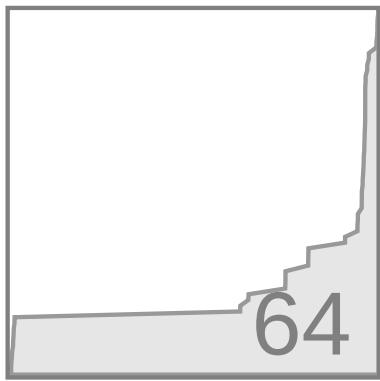


Distribution of queries over time.

- ❑ 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

## Distribution of Queries Over Time



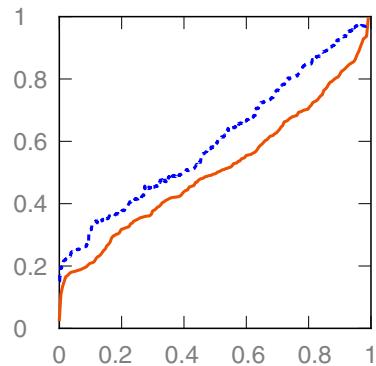
Distribution of queries over time.

- ❑ 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

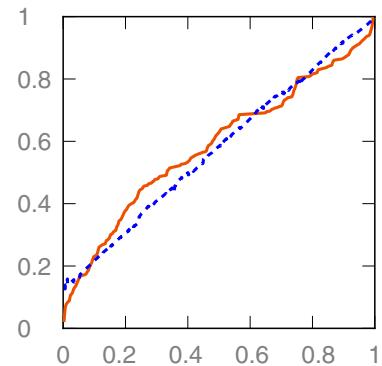
# Search Missions For Source Retrieval

## Correlation of Editing and Querying

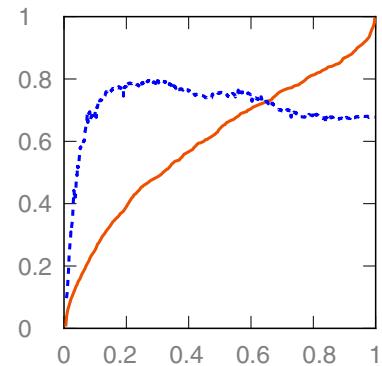
Author 5 (18 topics)



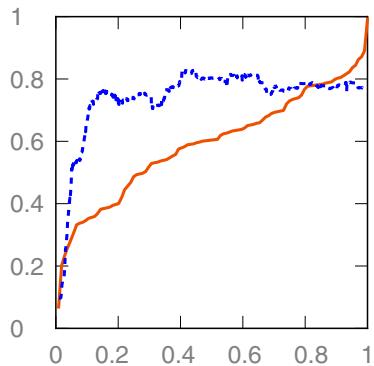
Author 20 (9 topics)



Author 2 (33 topics)



Author 24 (13 topics)



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

## 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

- [Potthast 2011] Martin Potthast. *Technologies for Reusing Text from the Web.* Dissertation, Bauhaus-Universität Weimar, December 2011.
- [Potthast et al. 2012a] Martin Potthast, Matthias Hagen, Benno Stein, Jan Graßegger, Maximilian Michel, Martin Tippmann, and Clement Welsch. *ChatNoir: A Search Engine for the ClueWeb09 Corpus.* SIGIR 2012, Portland, Oregon, August 2012.
- [Potthast et al. 2012b] Martin Potthast, Tim Gollub, Matthias Hagen, Jan Graßegger, Johannes Kiesel, Maximilian Michel, Arnd Oberländer, Martin Tippmann, Alberto Barrón-Cedeño, Parth Gupta, Paolo Rosso, and Benno Stein. *Overview of the 4th International Competition on Plagiarism Detection.* CLEF 2012 Evaluation Labs and Workshop – Working Notes Papers, September 2012.

## 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?

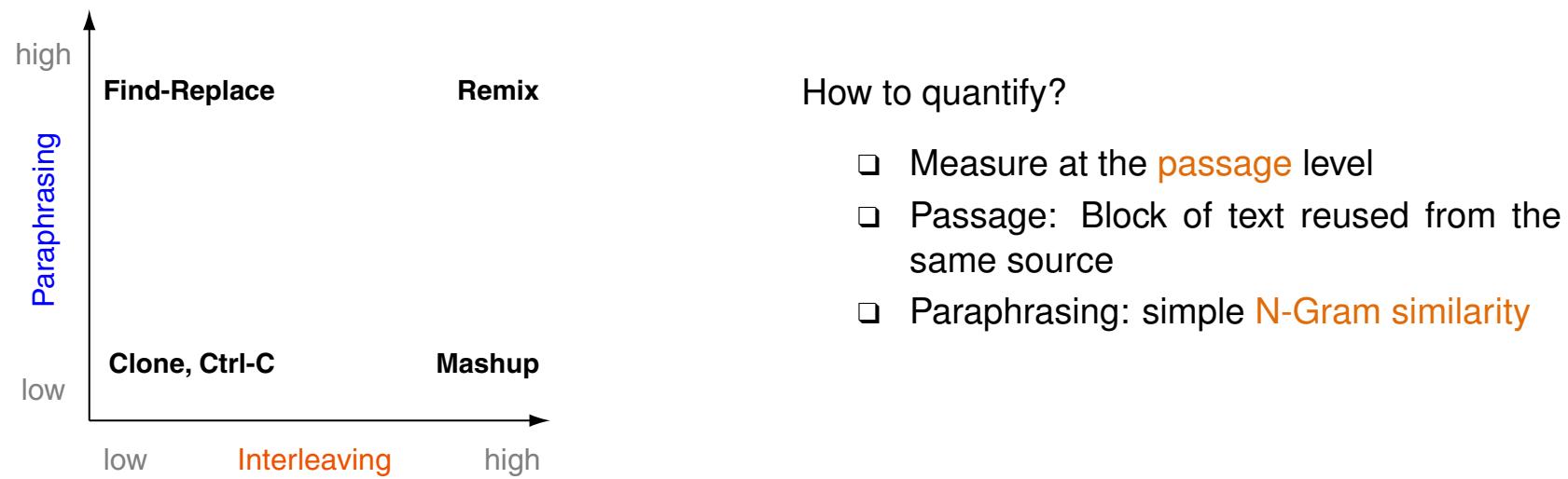
*Sub-topic 3.* Find biographical information on Barack Obama's mother.

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

[Turnitin 2012]

# Details: Paraphrasing & Interleaving

## Classification Scheme for Text Reuse

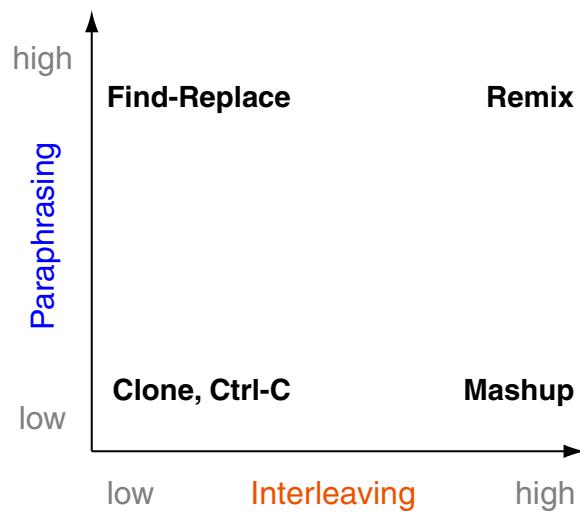


## Classification Scheme for Text Reuse.

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

# Details: Paraphrasing & Interleaving

## Classification Scheme for Text Reuse



### Three passages:

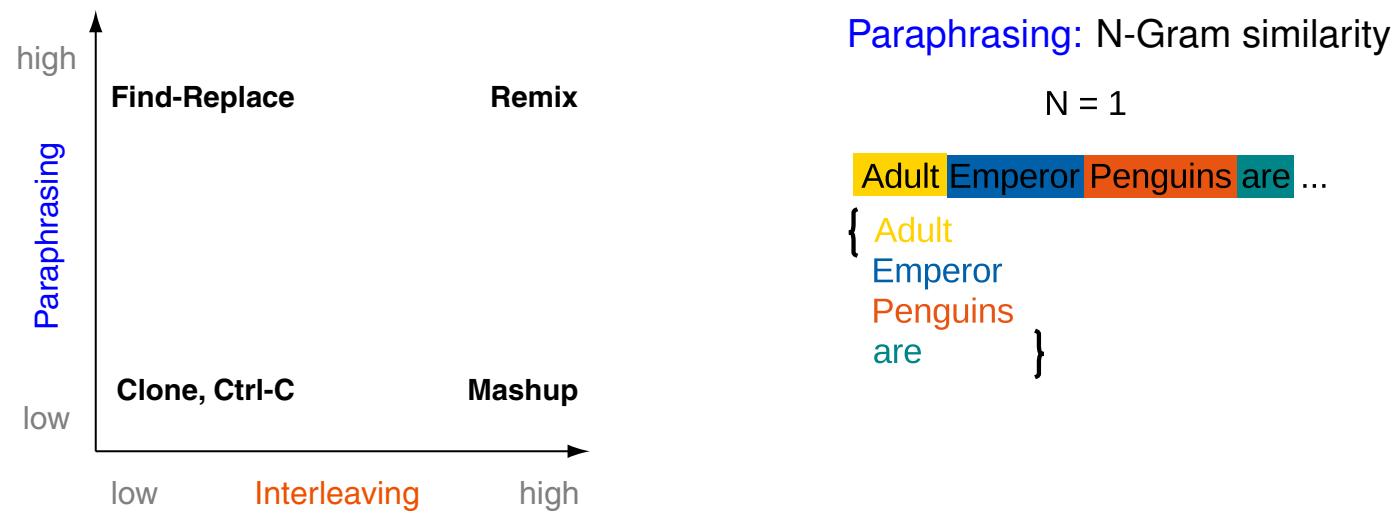
The Emperor penguins are the only penguins that inhabit the Antarctic continent and are the largest of all penguins. Adult Emperor penguins are typically 1.2 meters tall. Juveniles are slightly shorter, only about 90cm to 1m. Emperors weigh around 30 to 40 kg and their weight varies a great deal during 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

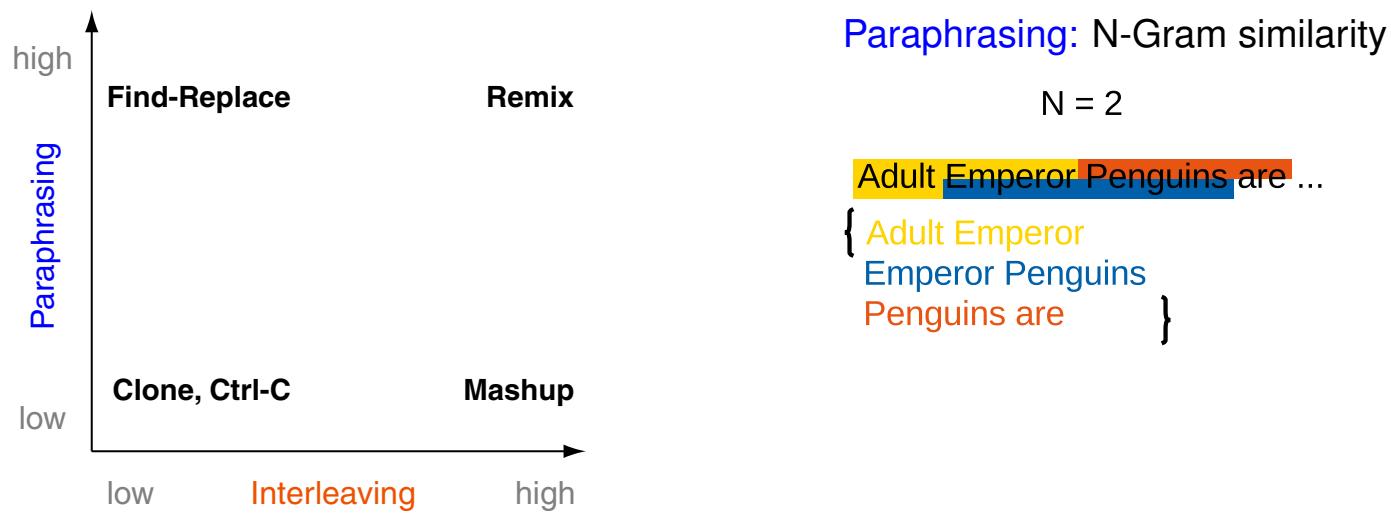


## Classification Scheme for Text Reuse.

- 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

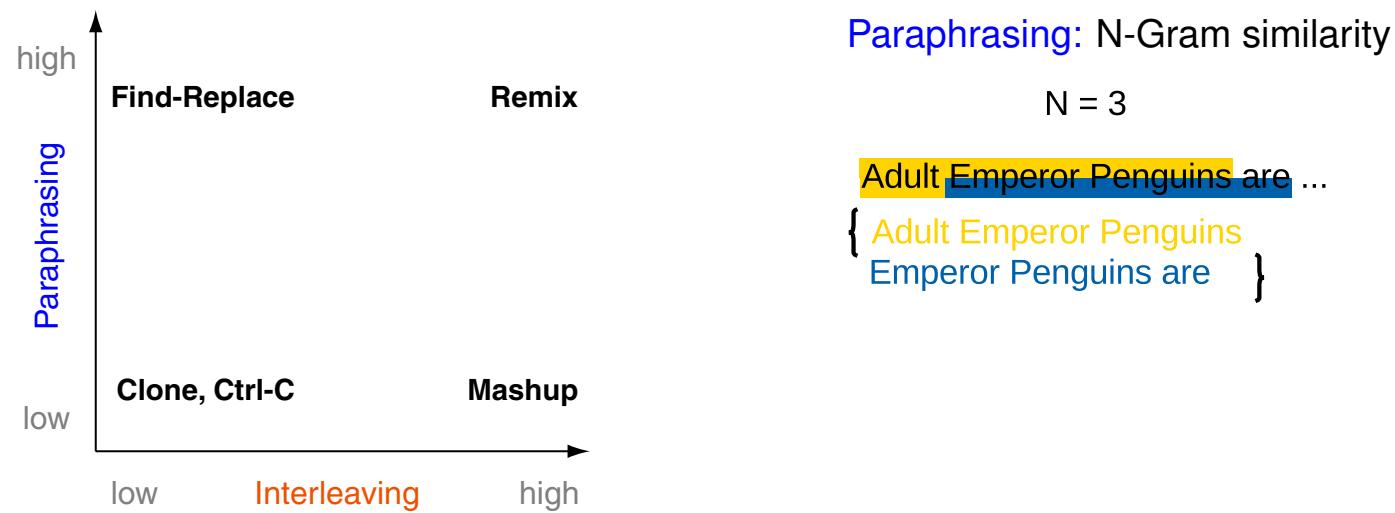


## Classification Scheme for Text Reuse.

- 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

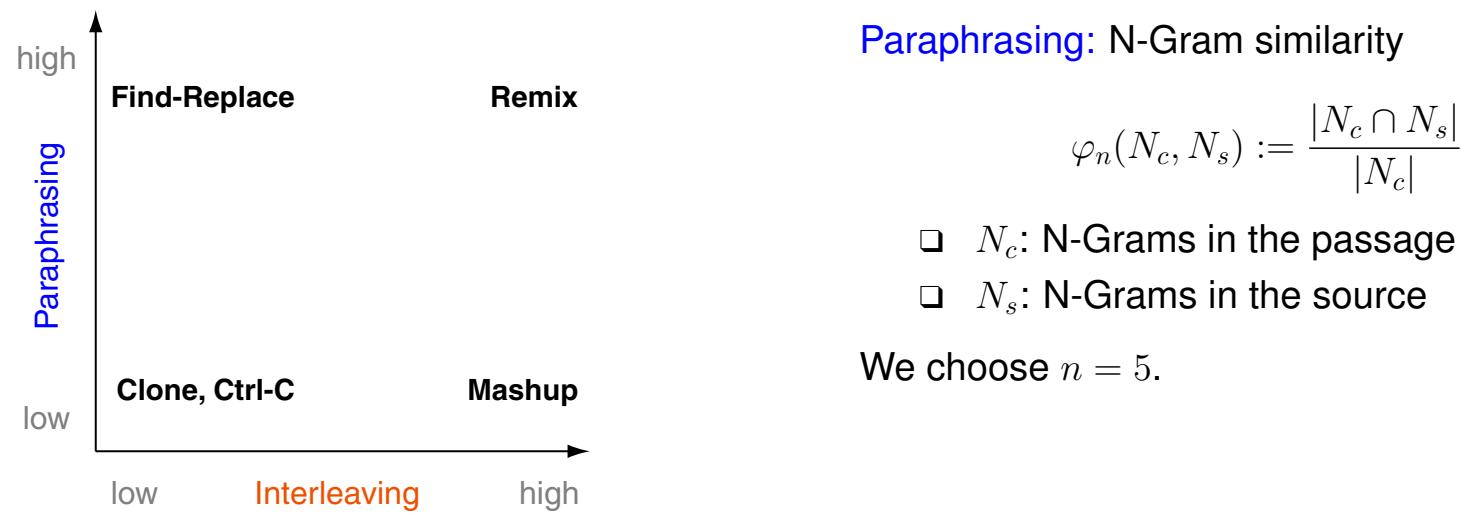


## Classification Scheme for Text Reuse.

- 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



## 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



## 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