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

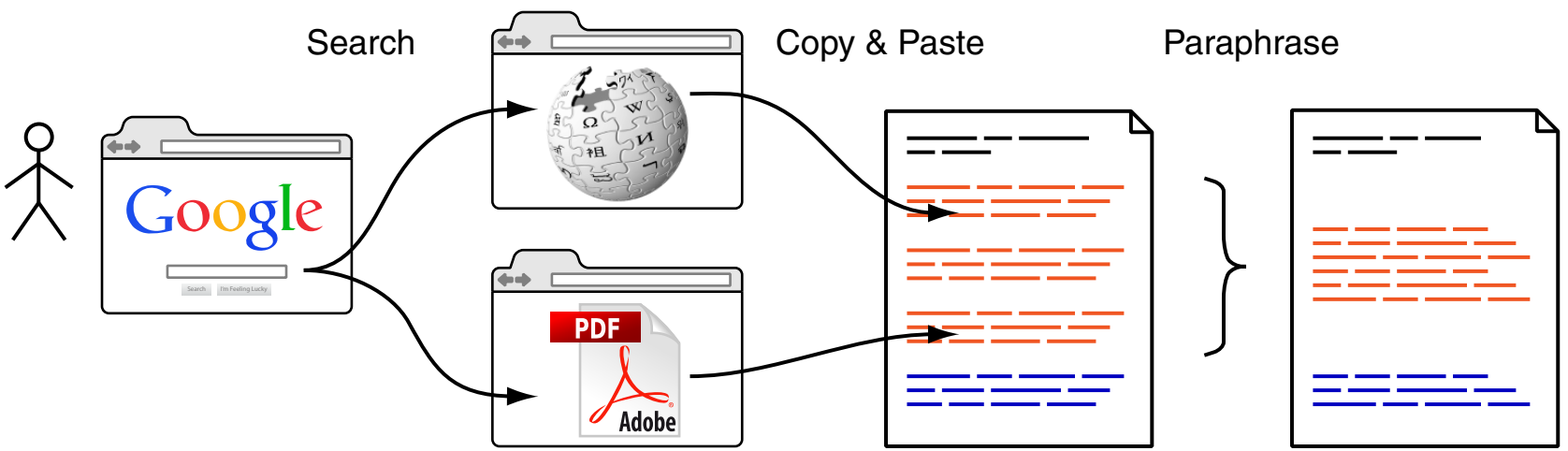
Analyzing a Large Corpus of Crowdsourced Plagiarism

Master's Thesis

Outline

Introduction

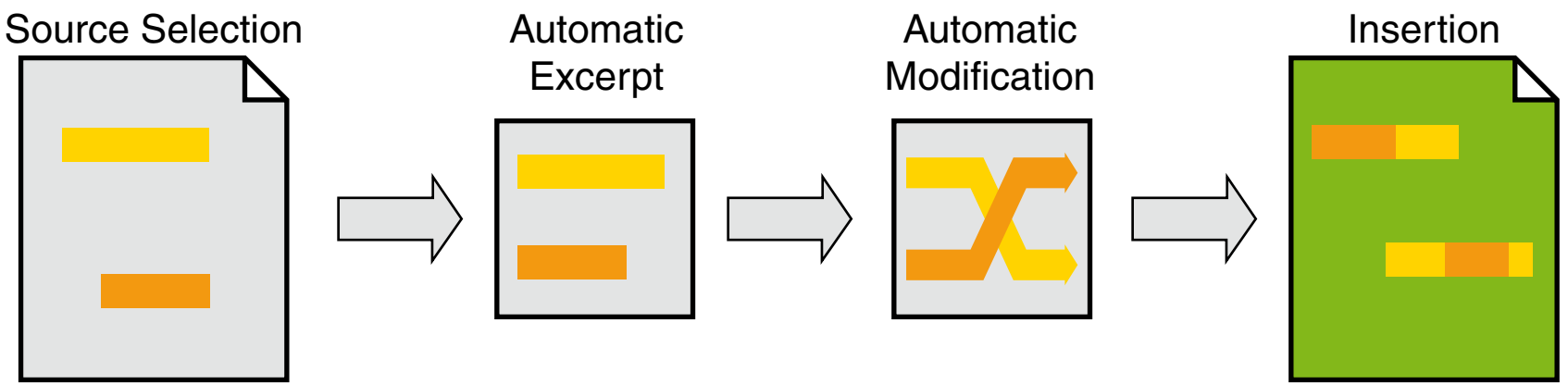
Modeling Text Reuse From the Web



[Potthast 2011]

Introduction

Previous Plagiarism Corpora

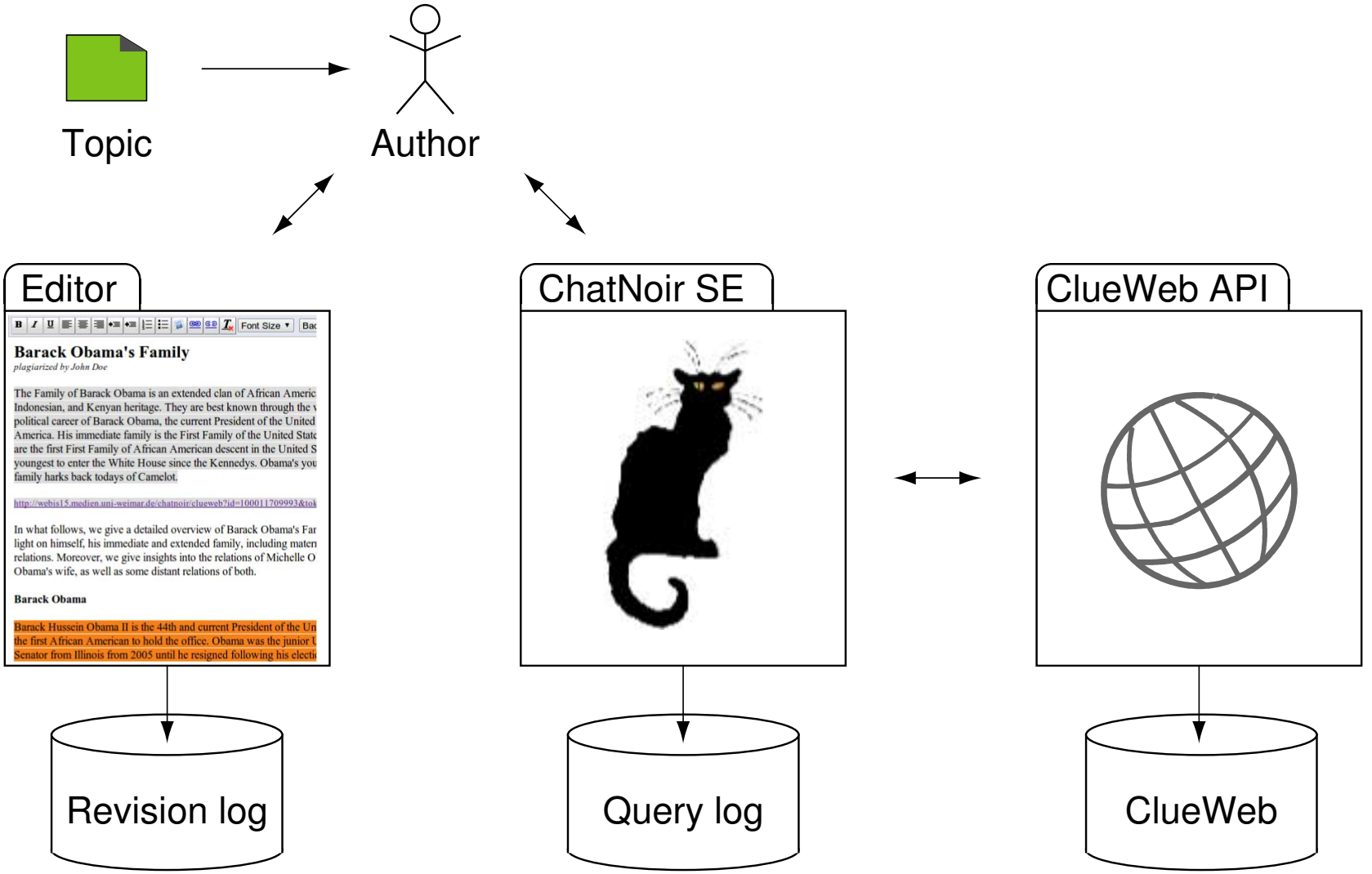


- Automatically generated *artificial plagiarism*
- PAN-PC-10/11: >25.000 documents; >60.000 plagiarism cases
- Automatic transformations do not preserve semantics

The Webis-TRC-12 Dataset

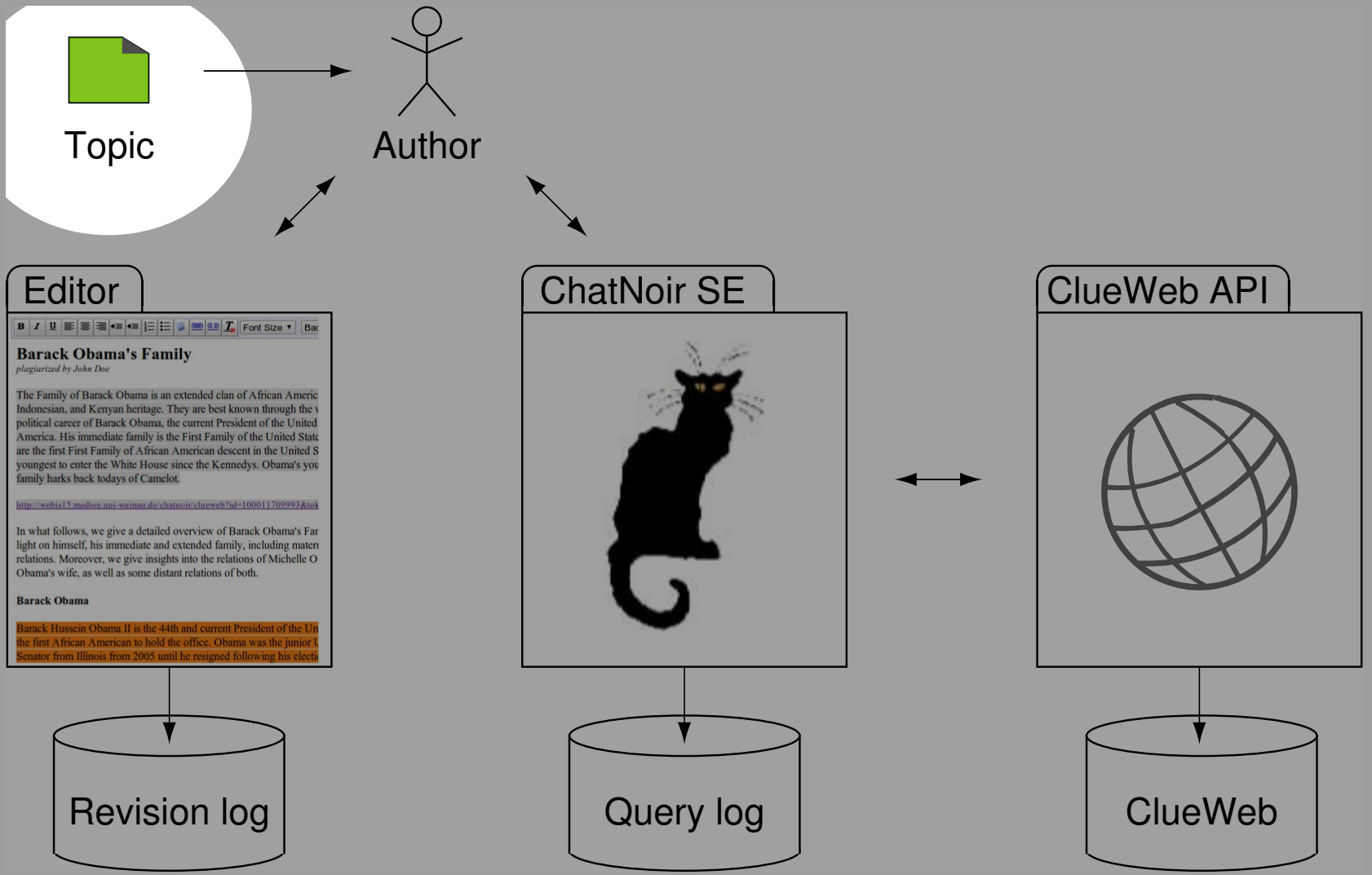
The Webis-TRC-12 Dataset

Construction Overview



The Webis-TRC-12 Dataset

Construction Overview: Topics



The Webis-TRC-12 Dataset

Construction Overview: Topics

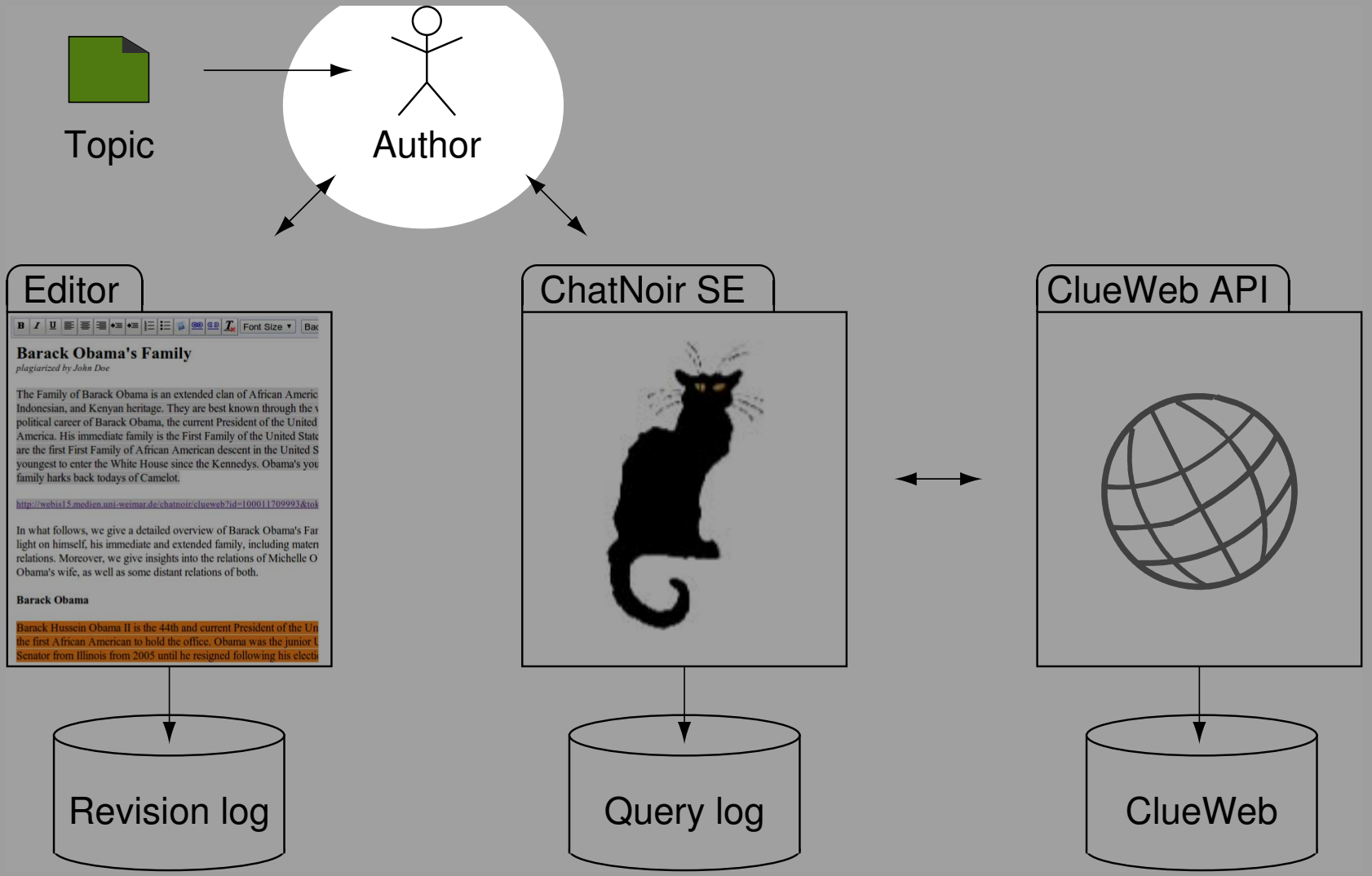
Example topic:

> *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

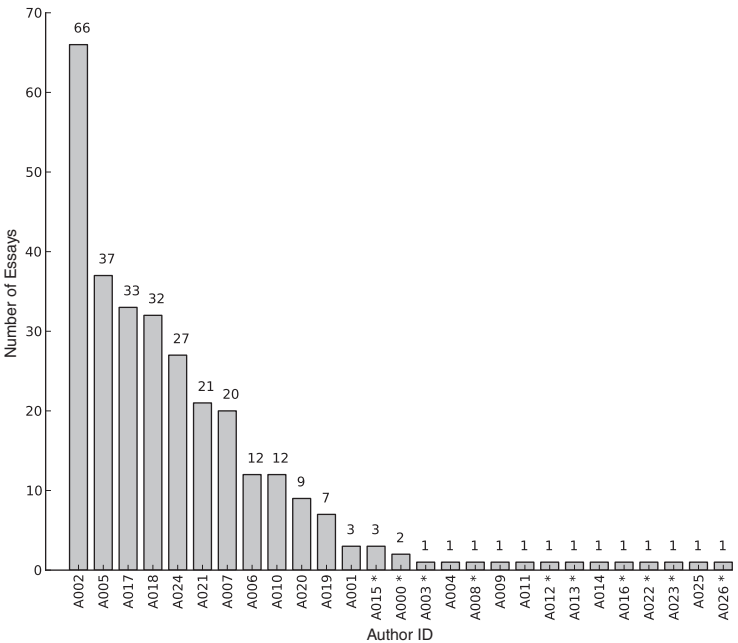
The Webis-TRC-12 Dataset

Construction Overview: Authors



The Webis-TRC-12 Dataset

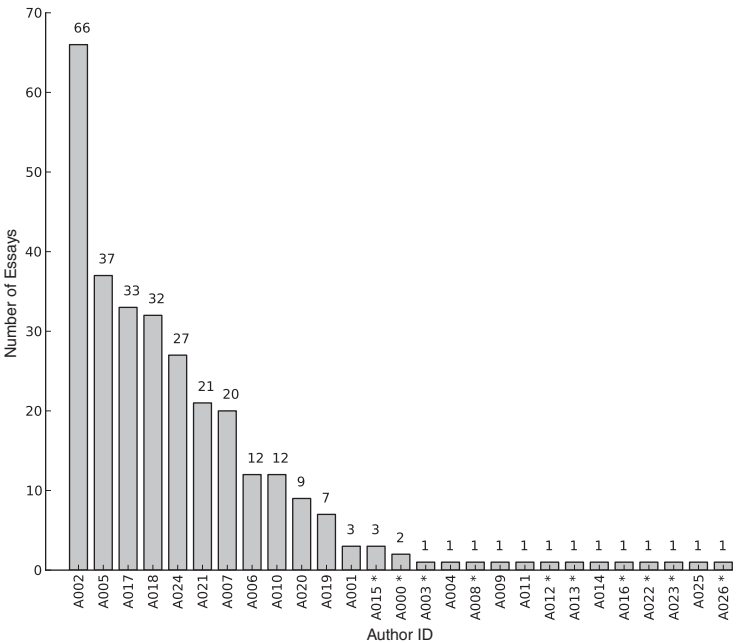
Construction Overview: Authors



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

The Webis-TRC-12 Dataset

Construction Overview: Authors

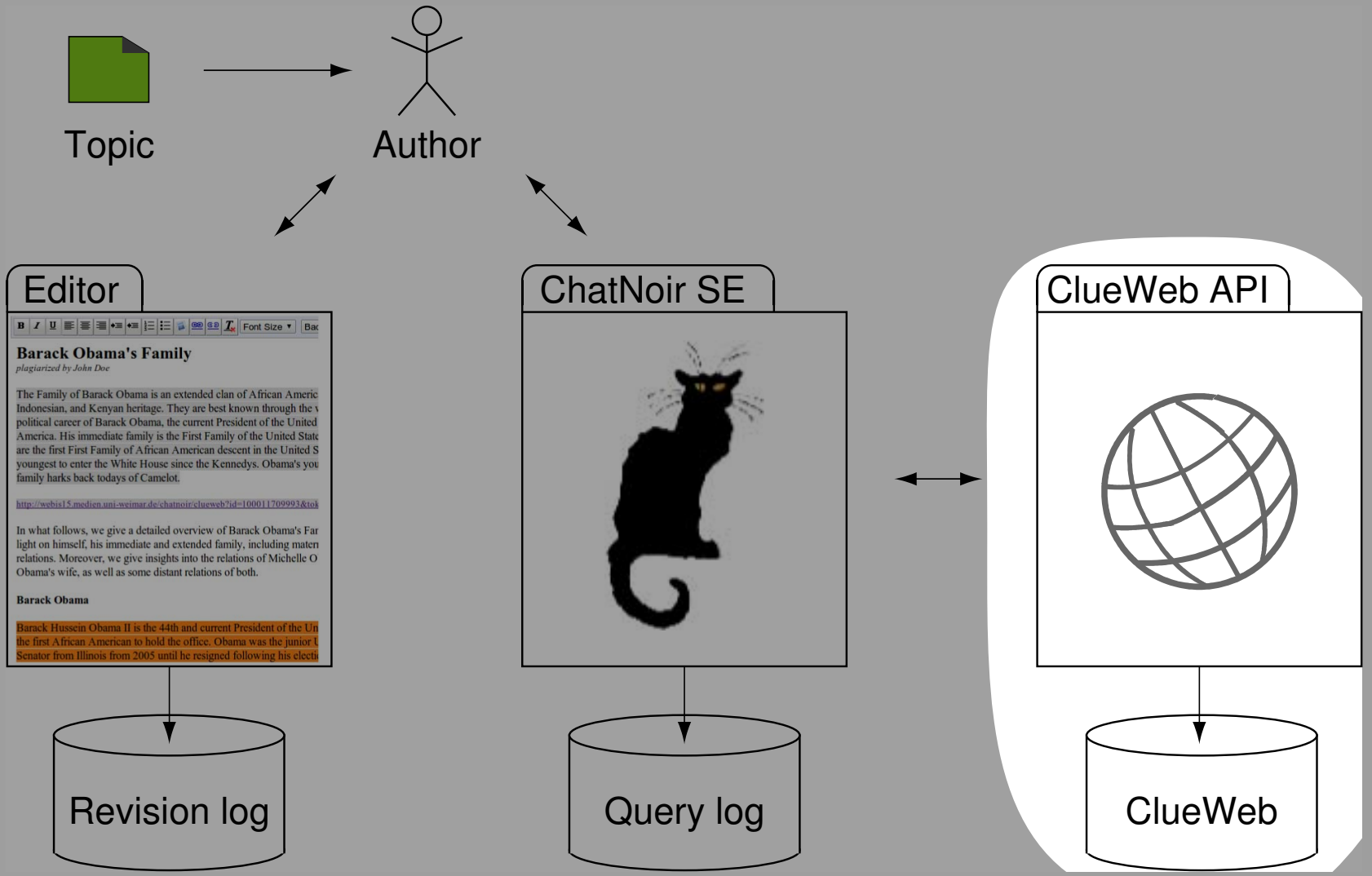


Author Demographics (n=12)	
Age (Median)	37
Years Writing (Median)	8
Academic degree	
Postgrad	33%
Undergrad	25%
English	
Native	67%
Second Language	33%

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

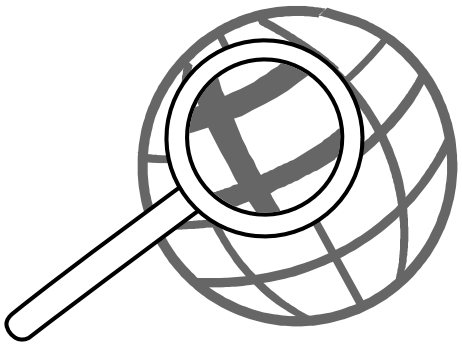
The Webis-TRC-12 Dataset

Construction Overview: Sources



The Webis-TRC-12 Dataset

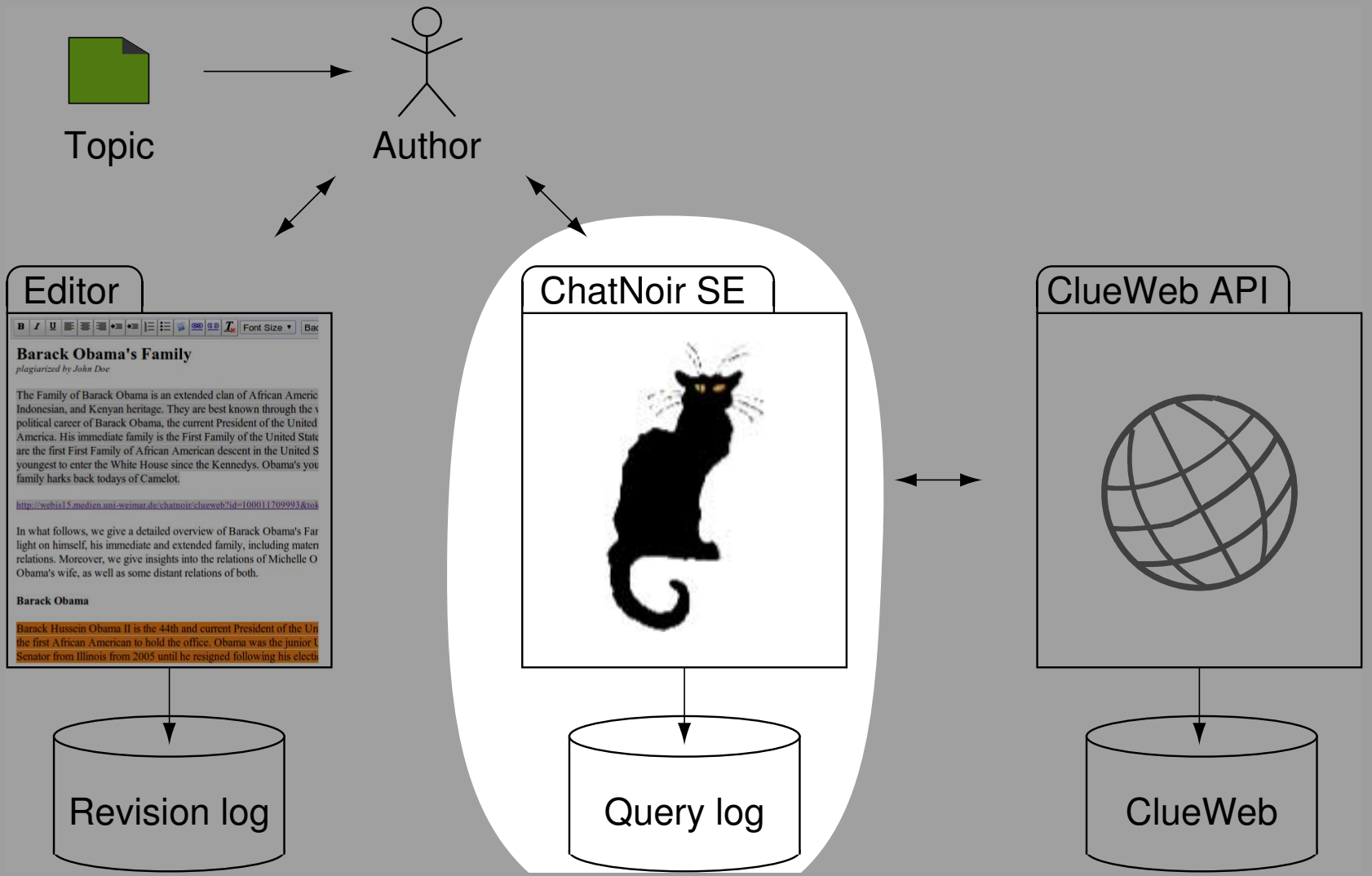
Construction Overview: Sources



- 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

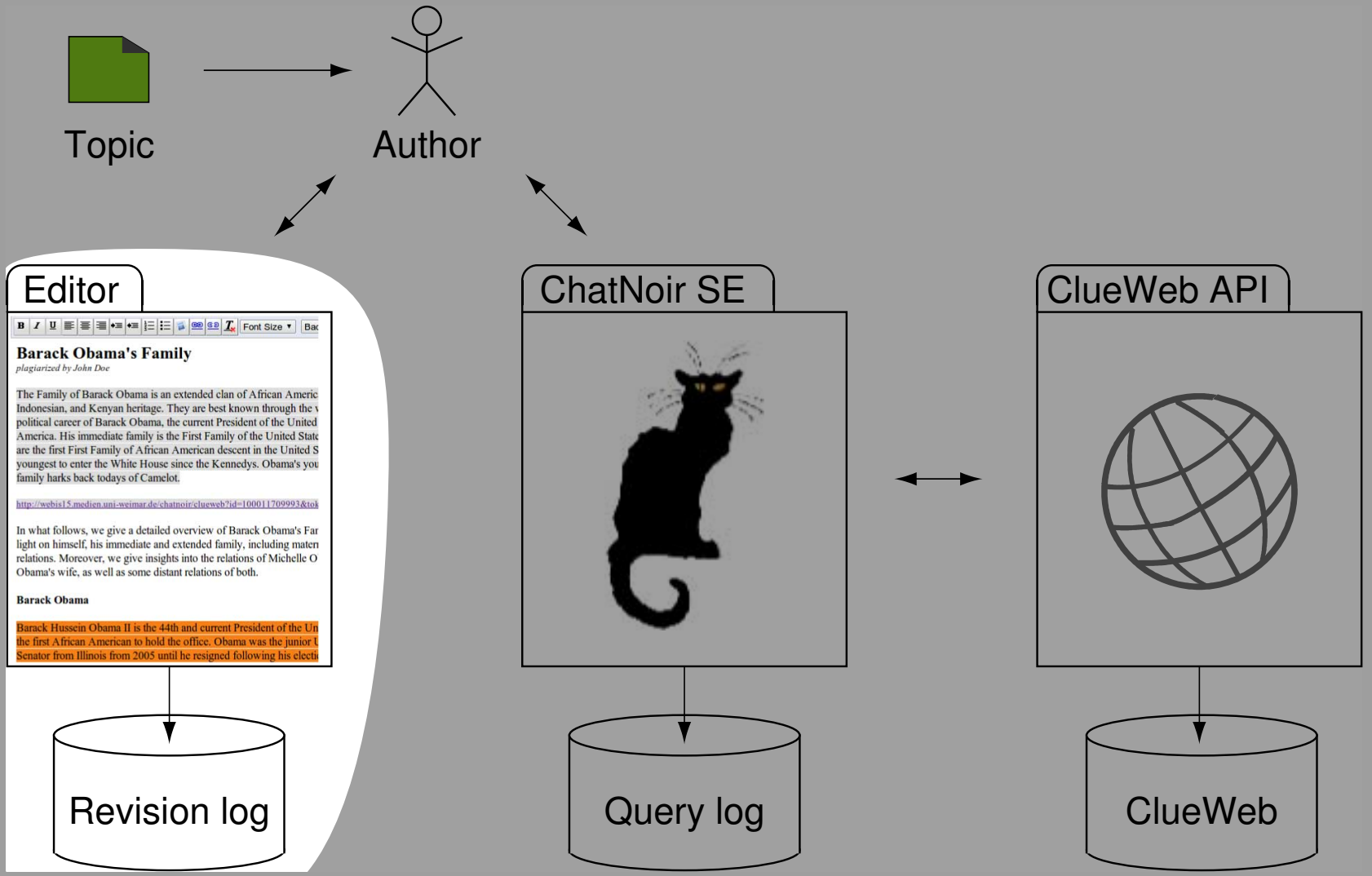
Construction Overview: Search Engine

[Potthast et al. 2012a]

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

The Webis-TRC-12 Dataset

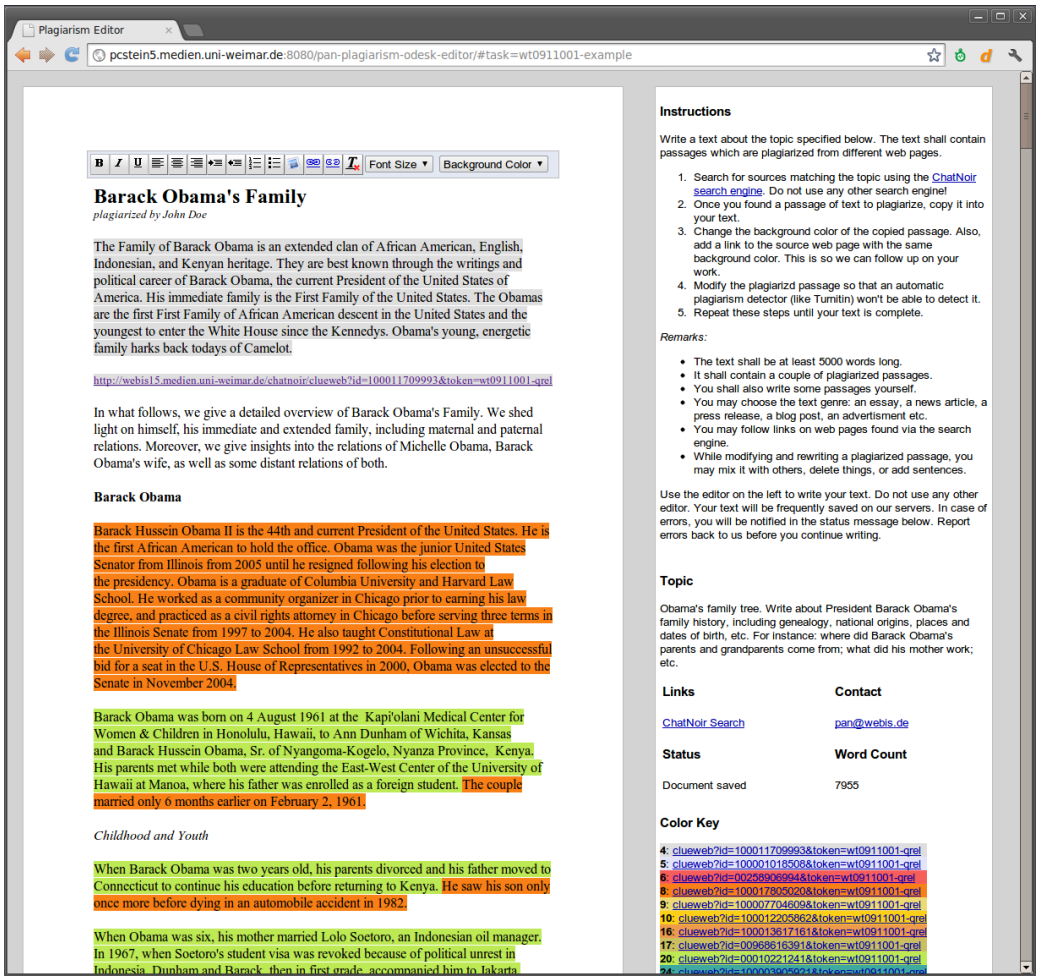
Construction Overview: Editor



The Webis-TRC-12 Dataset

Construction Overview: Editor

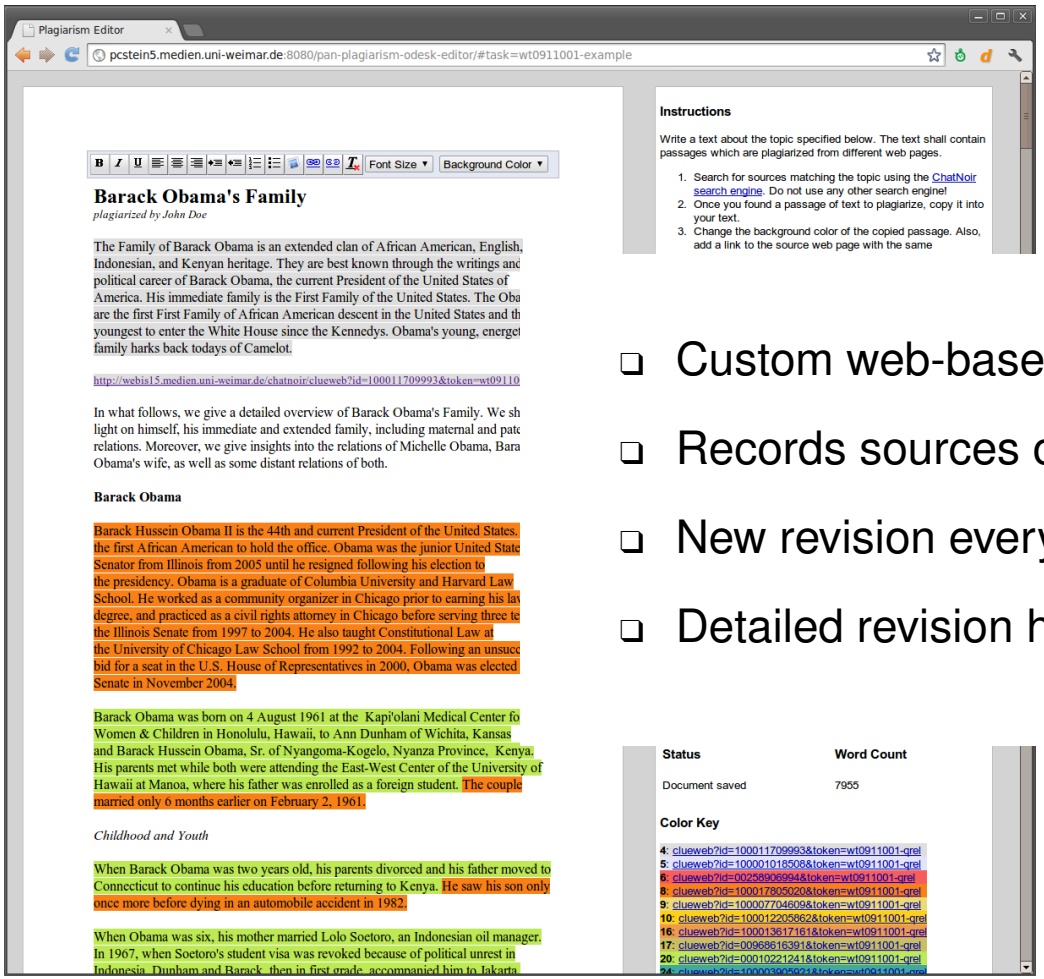
[Potthast et al. 2012b]



The Webis-TRC-12 Dataset

Construction Overview: Editor

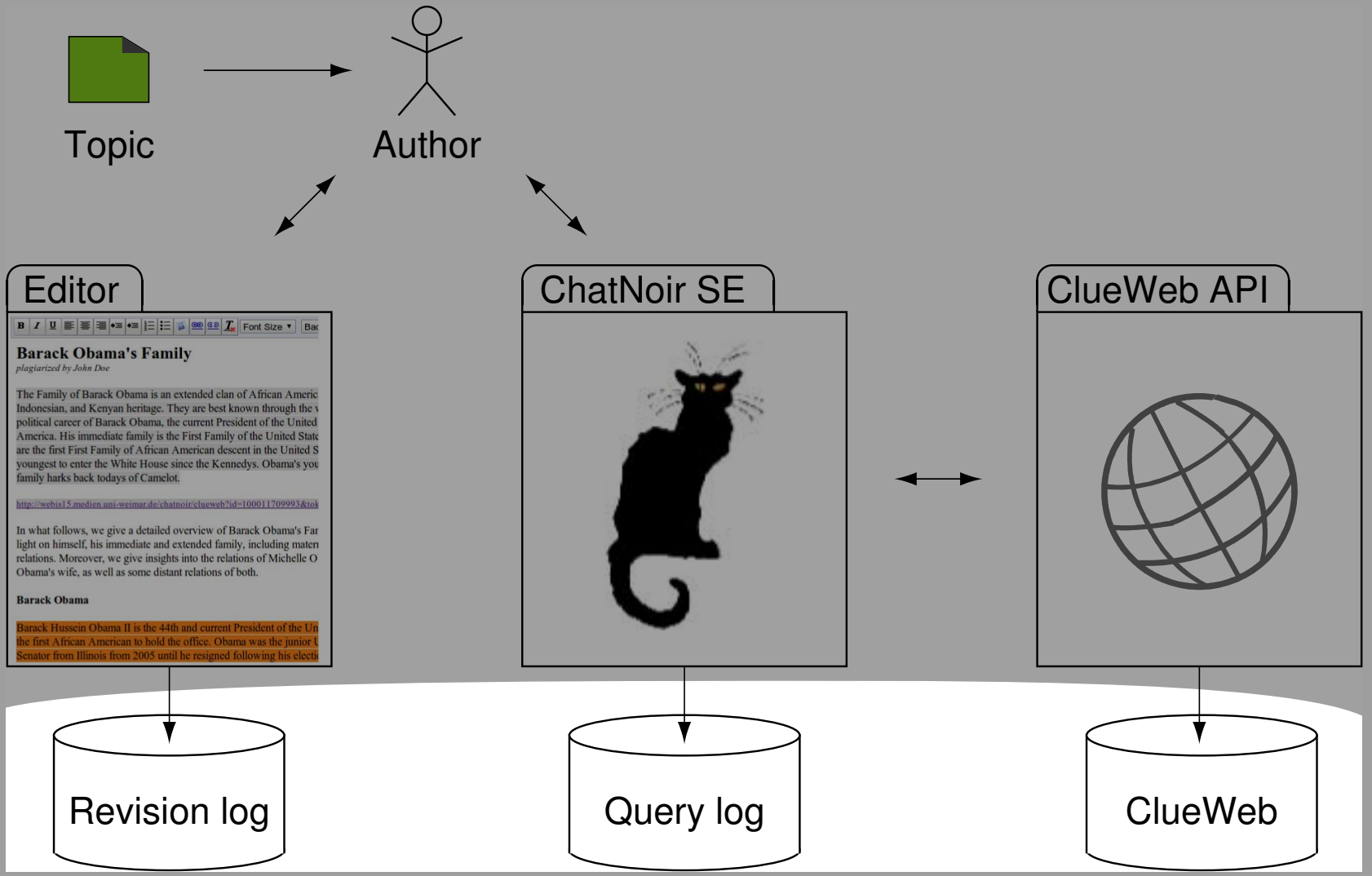
[Potthast et al. 2012b]



- 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



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?

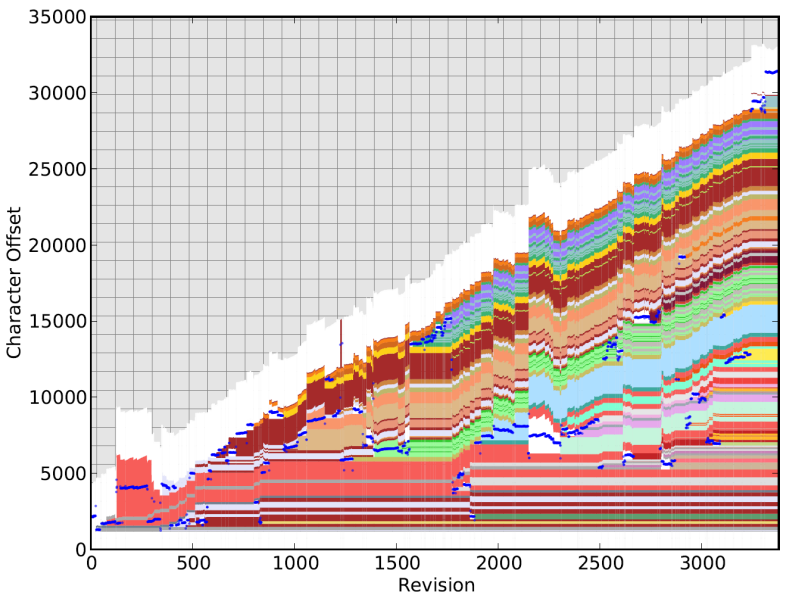
We expect new research insights and impacts to

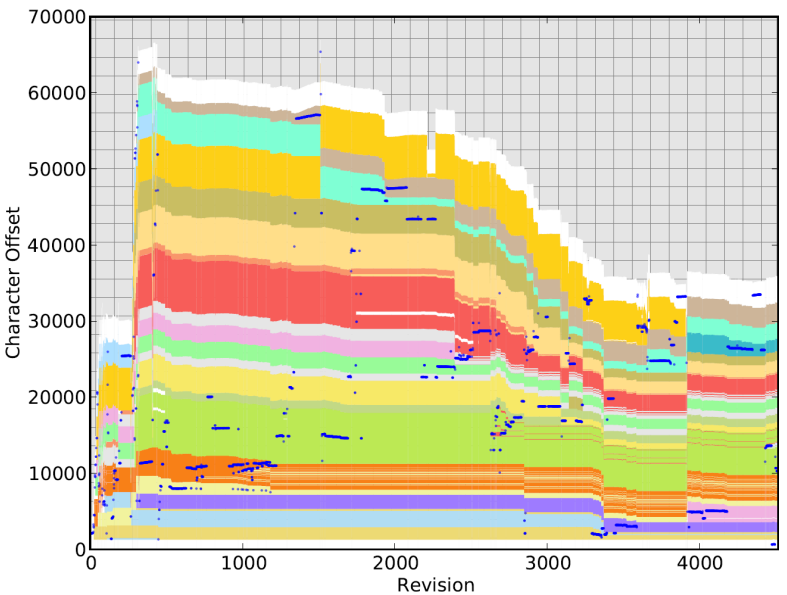
- text reuse detection
- query formulation
- paraphrasing

Categorizing Crowdsourced Text Reuse

Categorizing Crowdsourced Text Reuse

Build-Up Versus Boil-Down Reuse



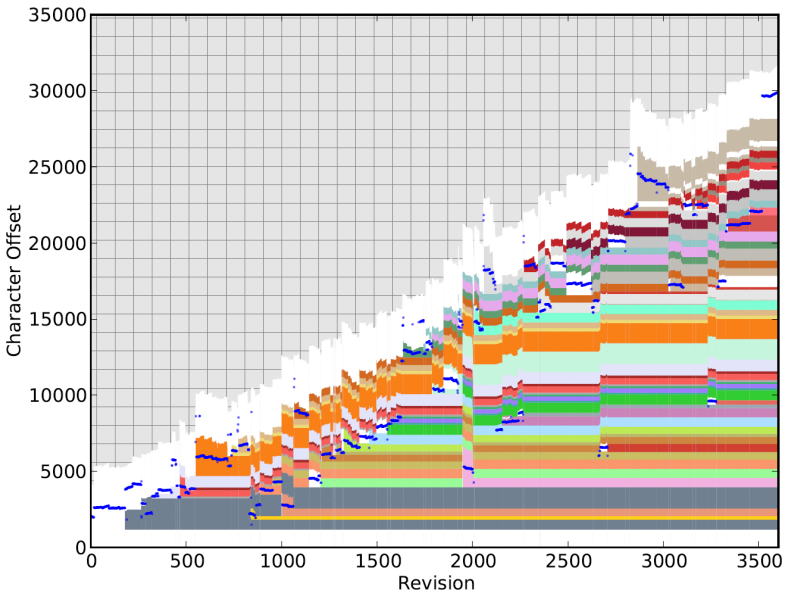


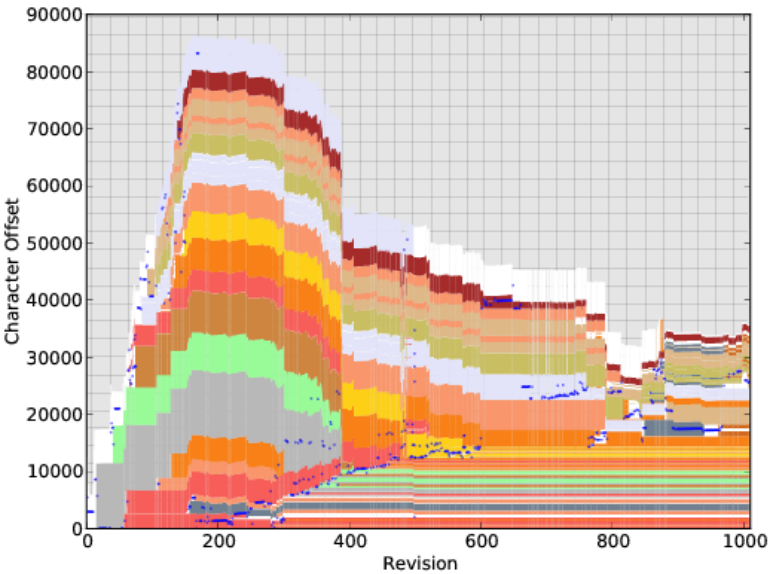
Build-up reuse (left) versus boil-down reuse (right).

- 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



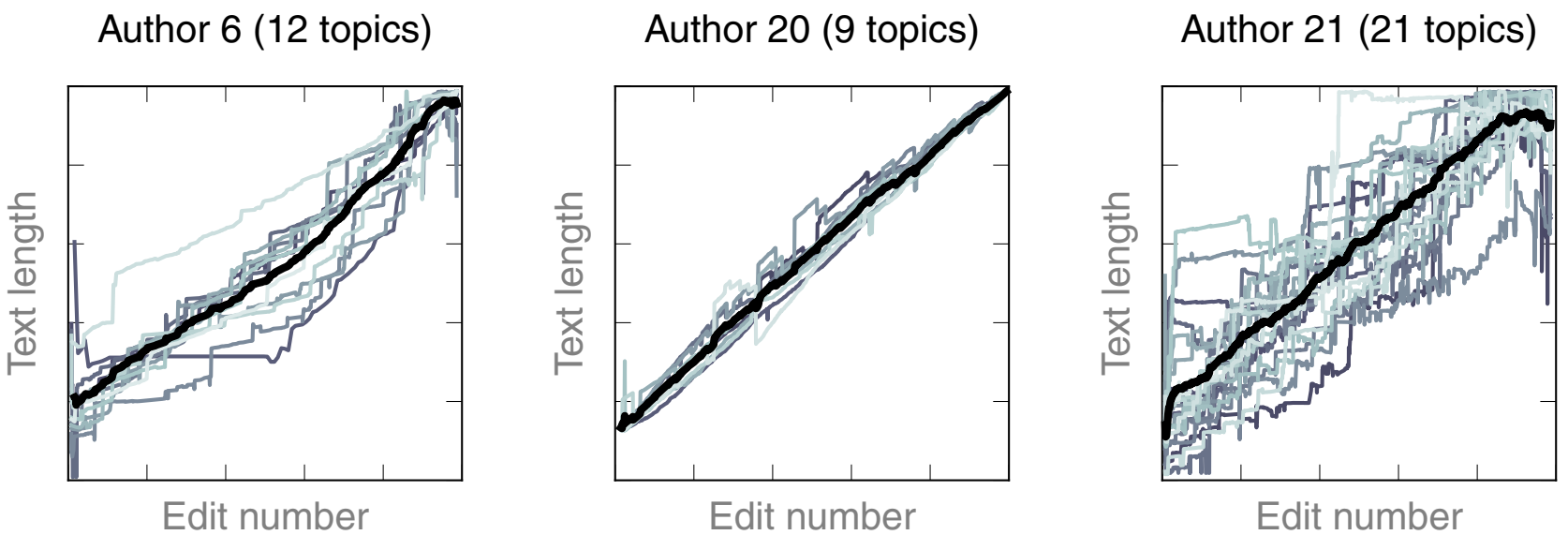


Build-up reuse (left) versus boil-down reuse (right).

- 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

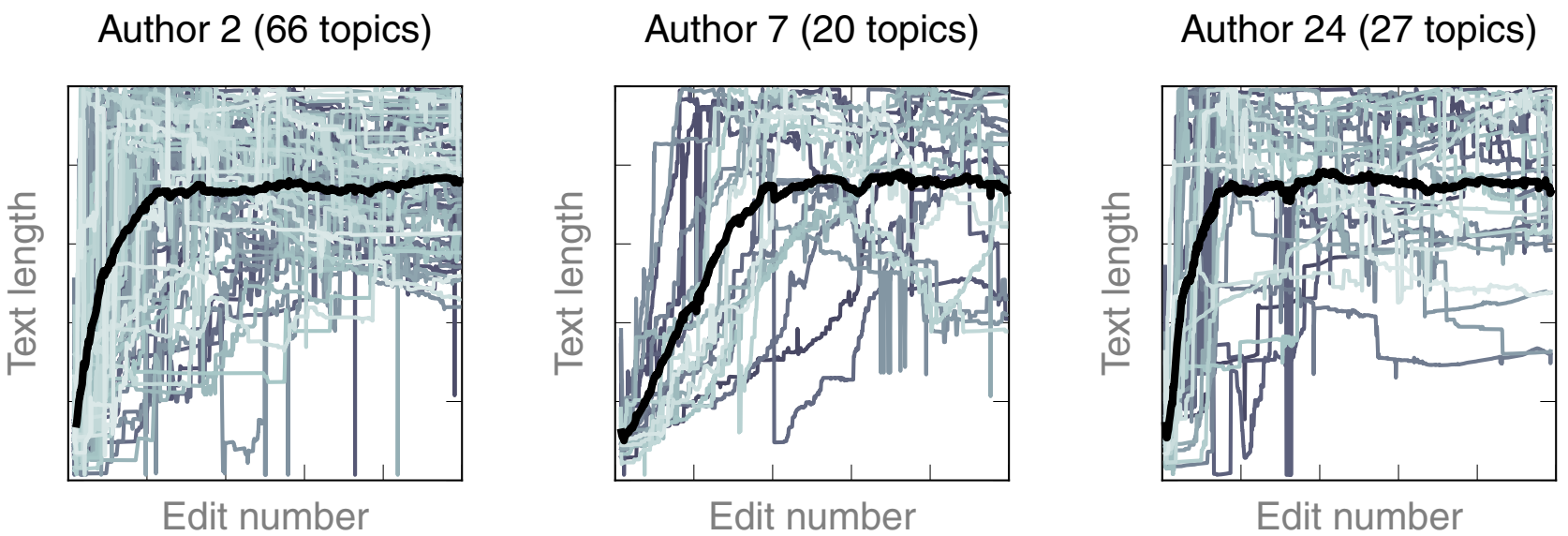


Build-up reuse: Averaged editing histories by authors.

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

Categorizing Crowdsourced Text Reuse

Build-Up Versus Boil-Down Reuse



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**

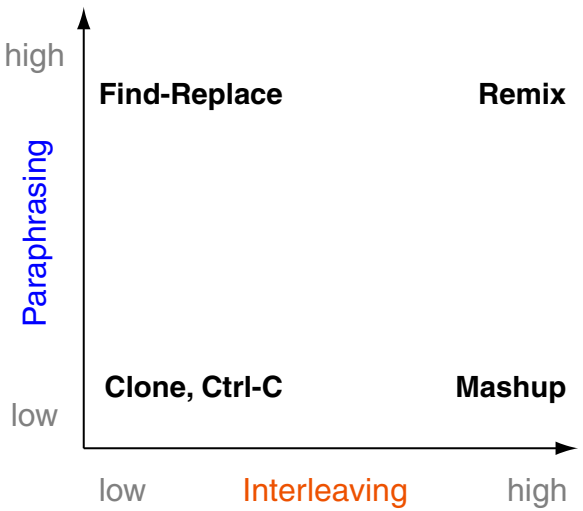
Clone, Ctrl-C **Mashup**

Classification Scheme for Text Reuse.

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

Categorizing Crowdsourced Text Reuse

Classification Scheme for Text Reuse

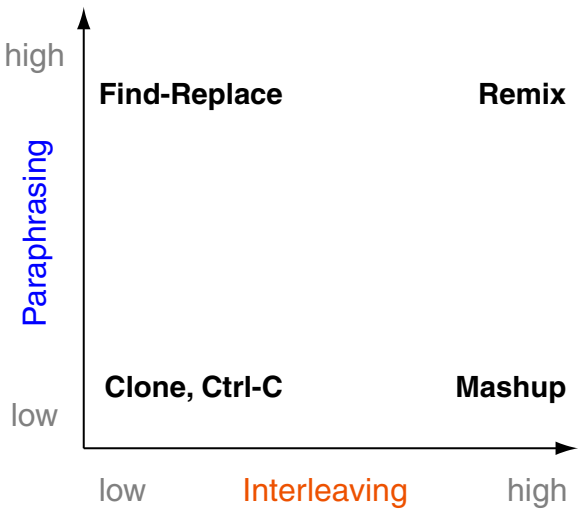


Classification Scheme for Text Reuse.

- 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



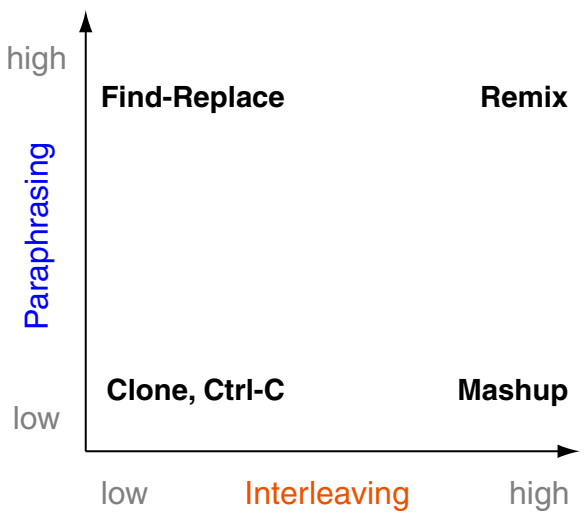
- Quantify: N-Gram similarity and ratio of passages to sources
  [details]
- Measure for all essays
- Hypothesis: will show evidence of authors' individual text reuse styles

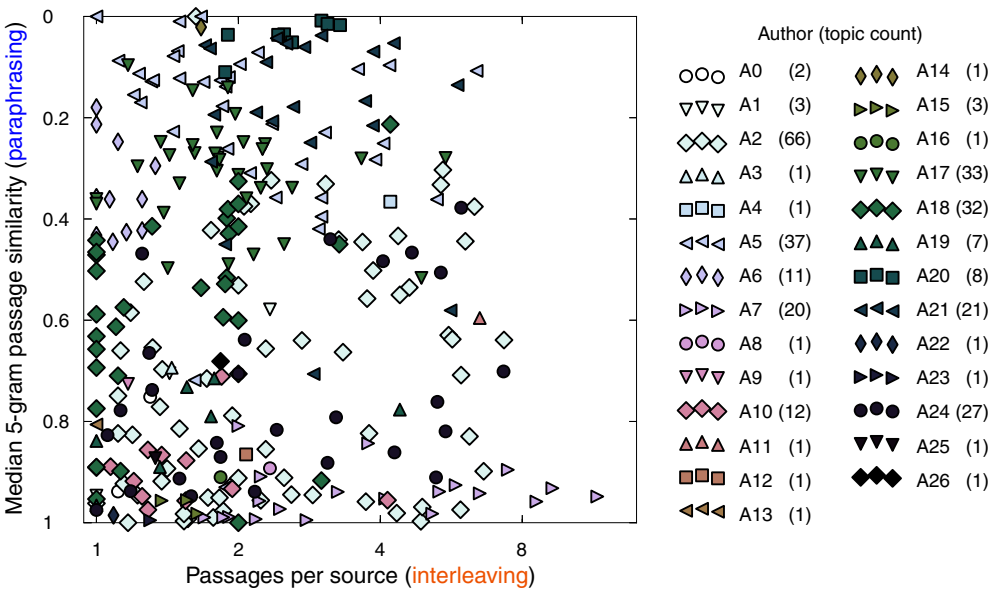
Classification Scheme for Text Reuse.

- 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



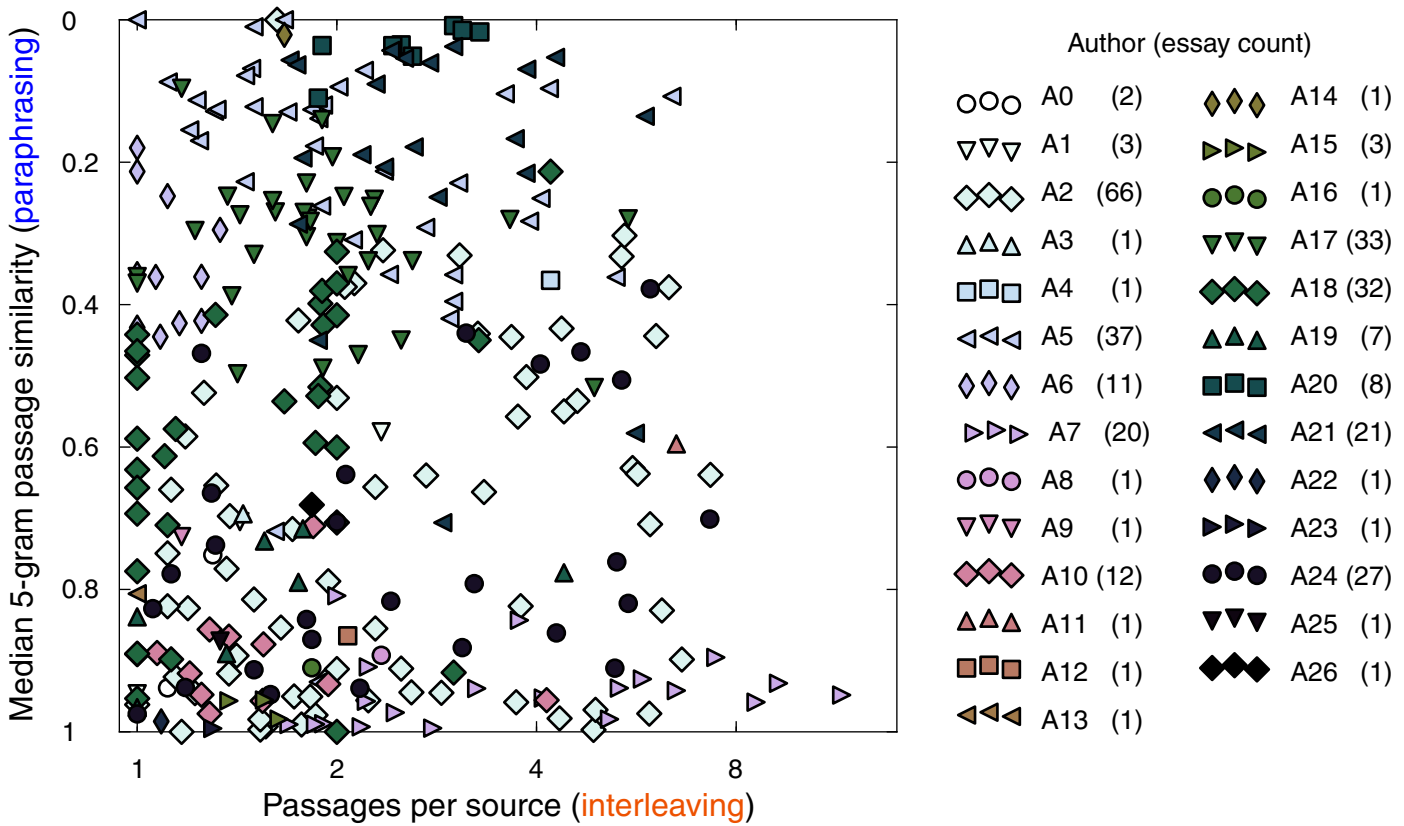


Classification Scheme for Text Reuse.

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

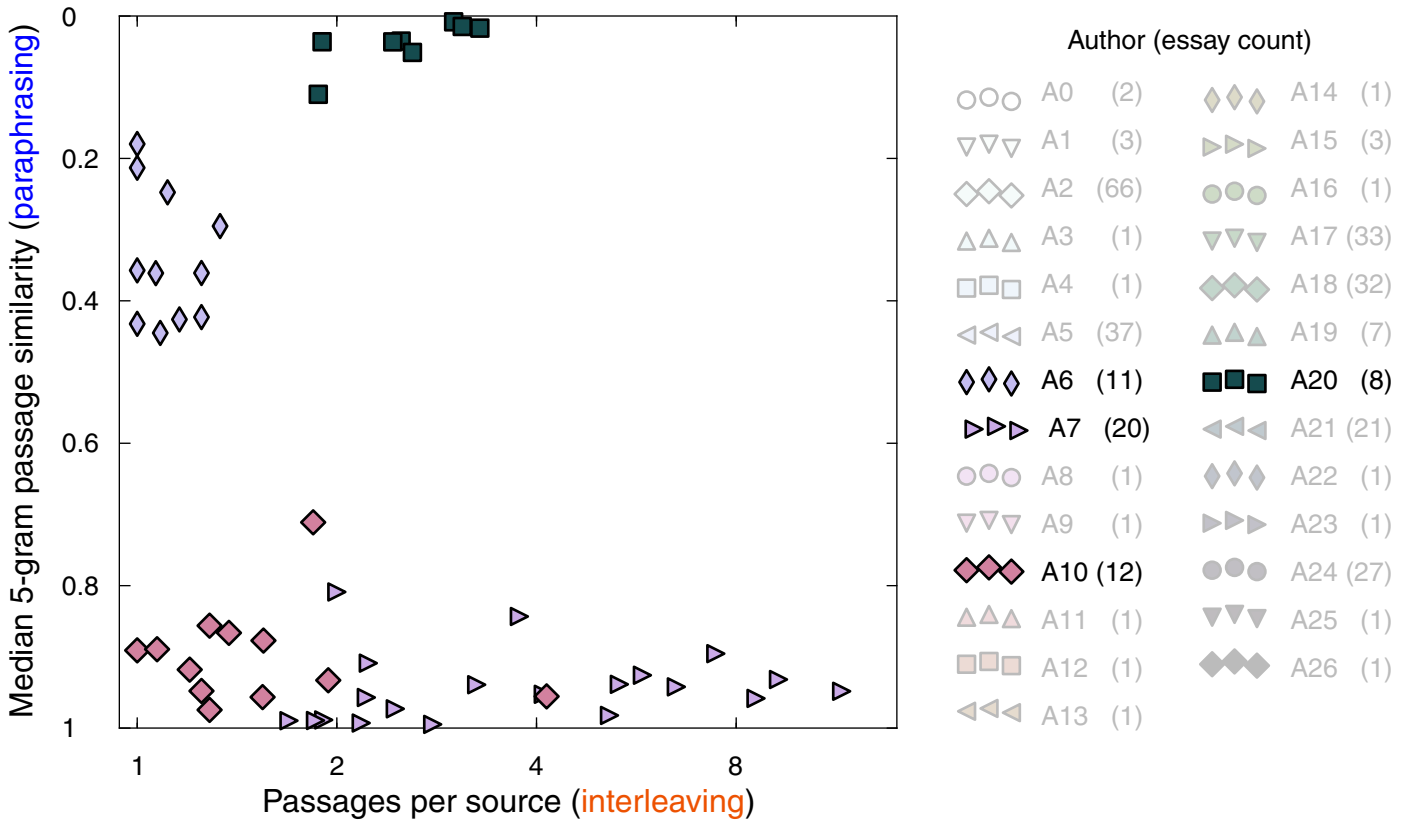
Categorizing Crowdsourced Text Reuse

Classification Scheme for Text Reuse



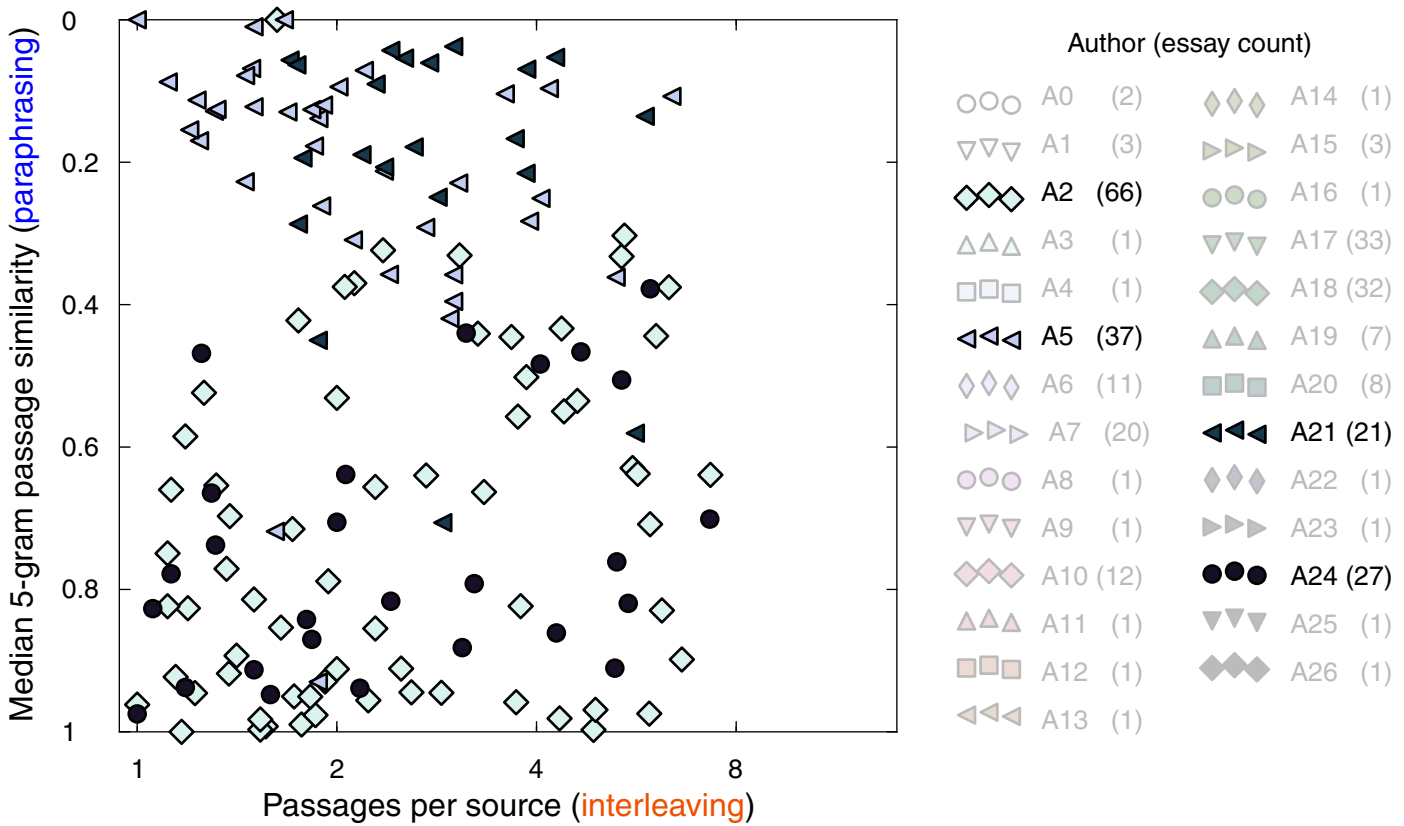
Categorizing Crowdsourced Text Reuse

Classification Scheme for Text Reuse



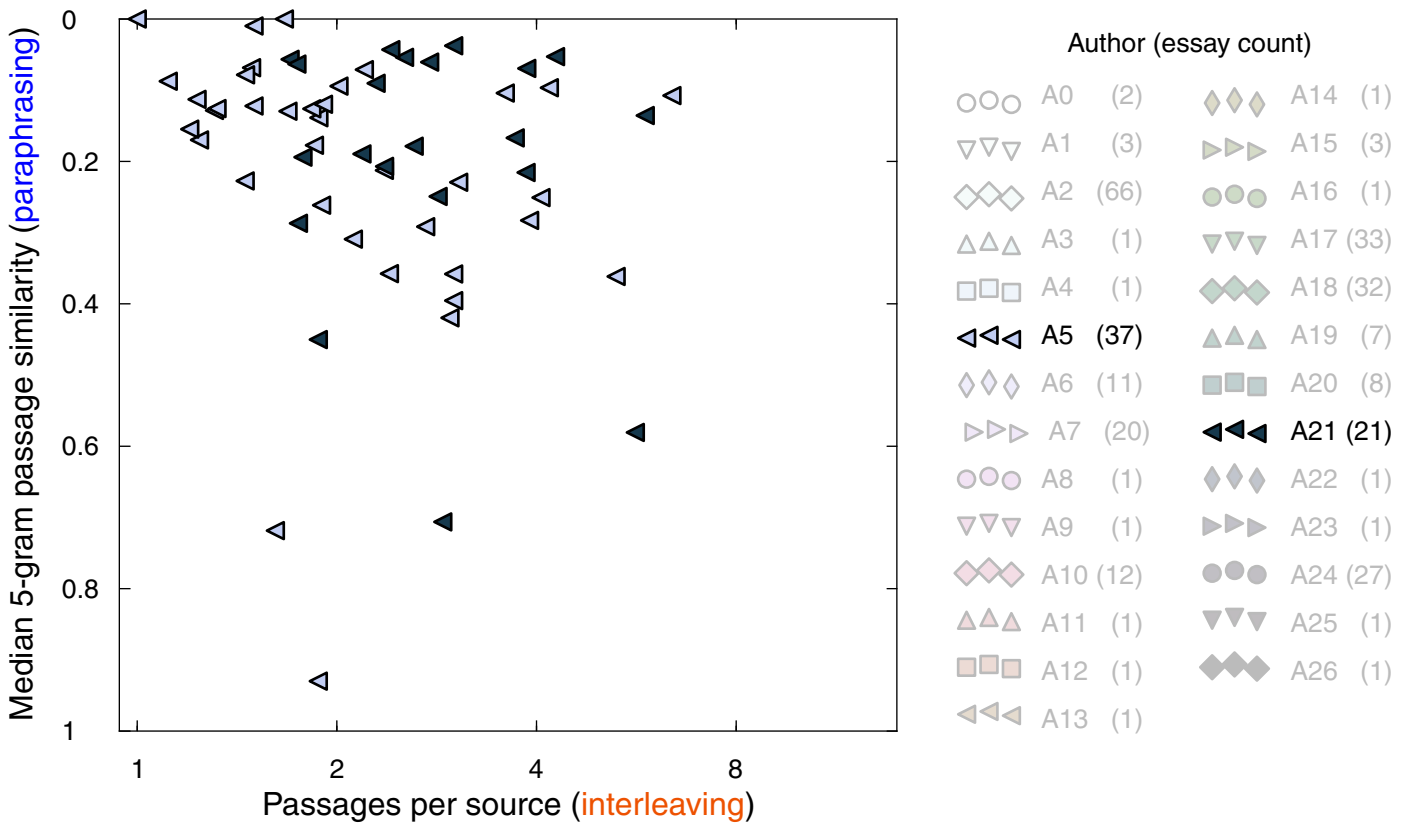
Categorizing Crowdsourced Text Reuse

Classification Scheme for Text Reuse



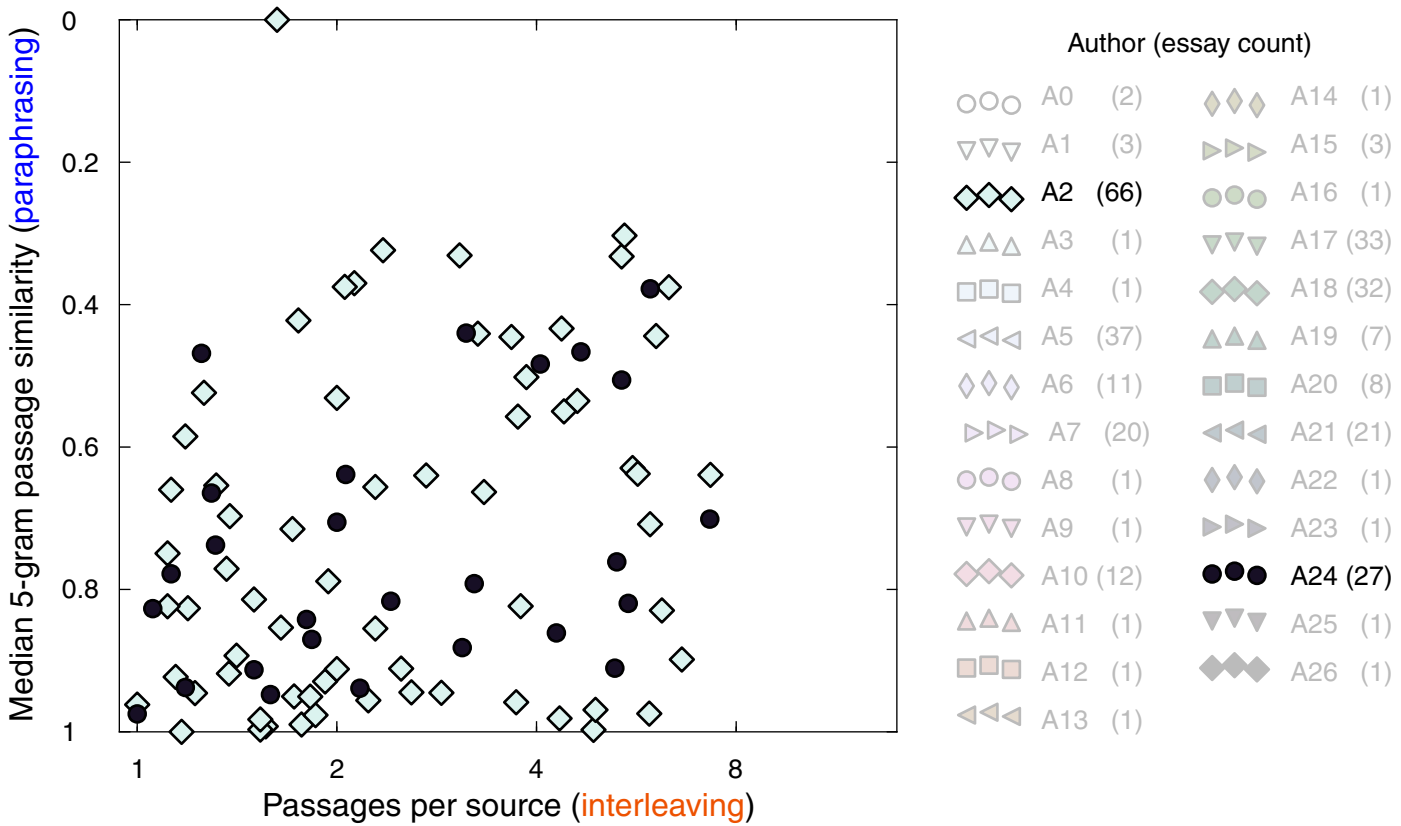
Categorizing Crowdsourced Text Reuse

Classification Scheme for Text Reuse



Categorizing Crowdsourced Text Reuse

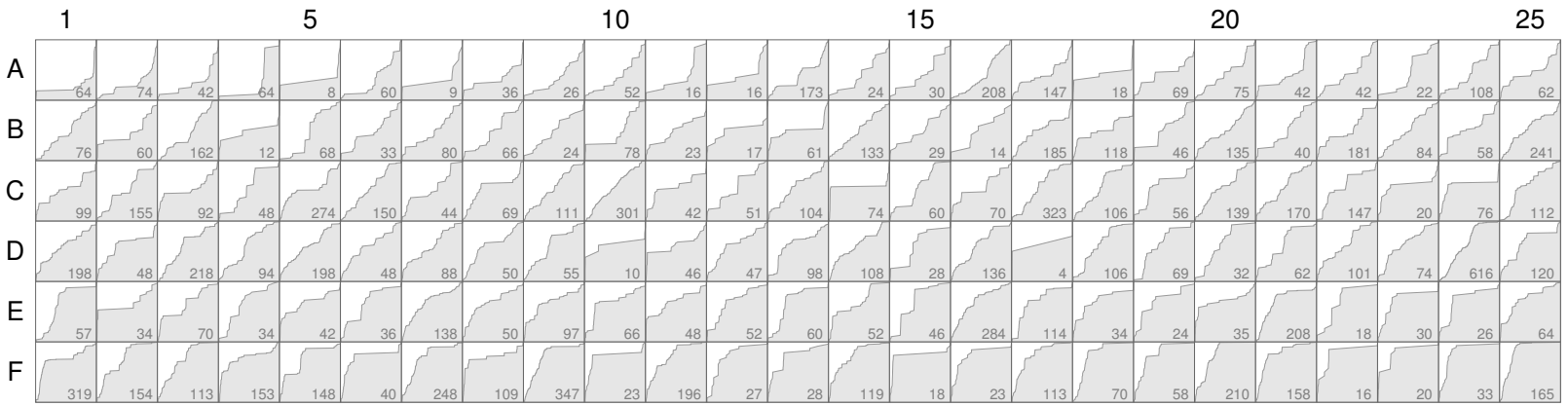
Classification Scheme for Text Reuse



Search Missions For Source Retrieval

Search Missions For Source Retrieval

Distribution of Queries Over Time

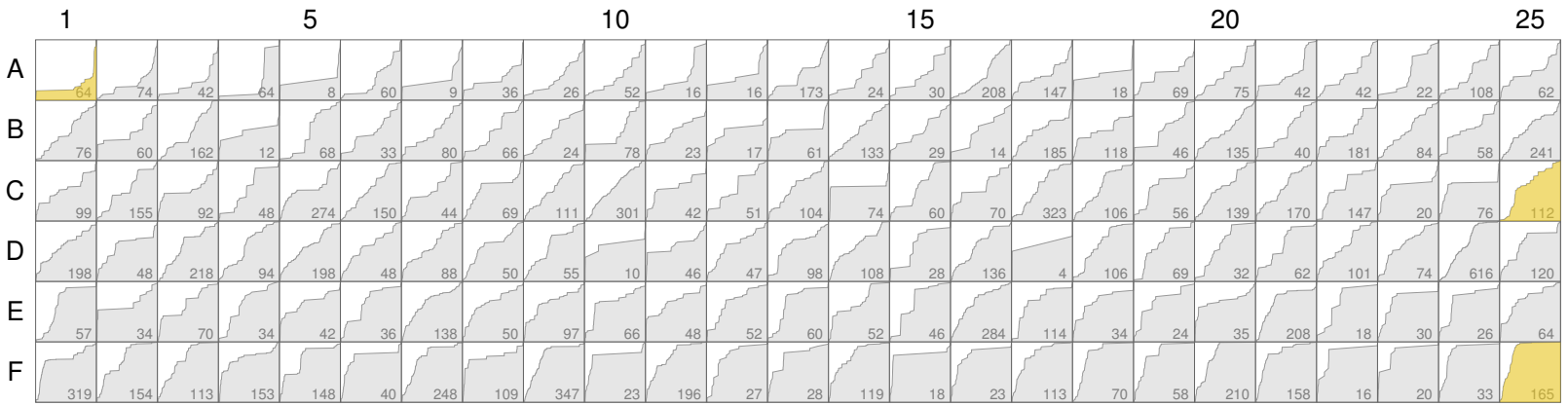


Distribution of queries over time.

- 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

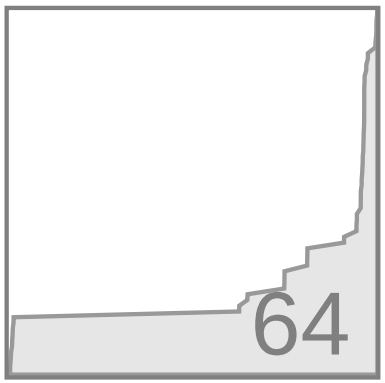


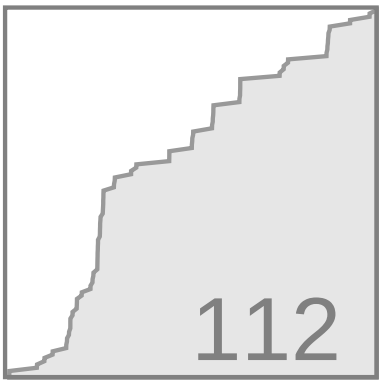
Distribution of queries over time.

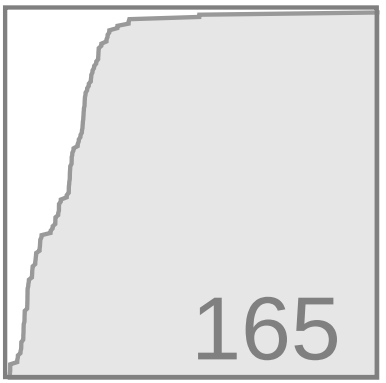
- fraction of posed queries (y-axis) over elapsed time (x-axis) between the first query until essay completion
- each cell represents one of 150 essays
- the numbers denote the total amount of posed queries
- the cells are sorted by area under the curve

Search Missions For Source Retrieval

Distribution of Queries Over Time





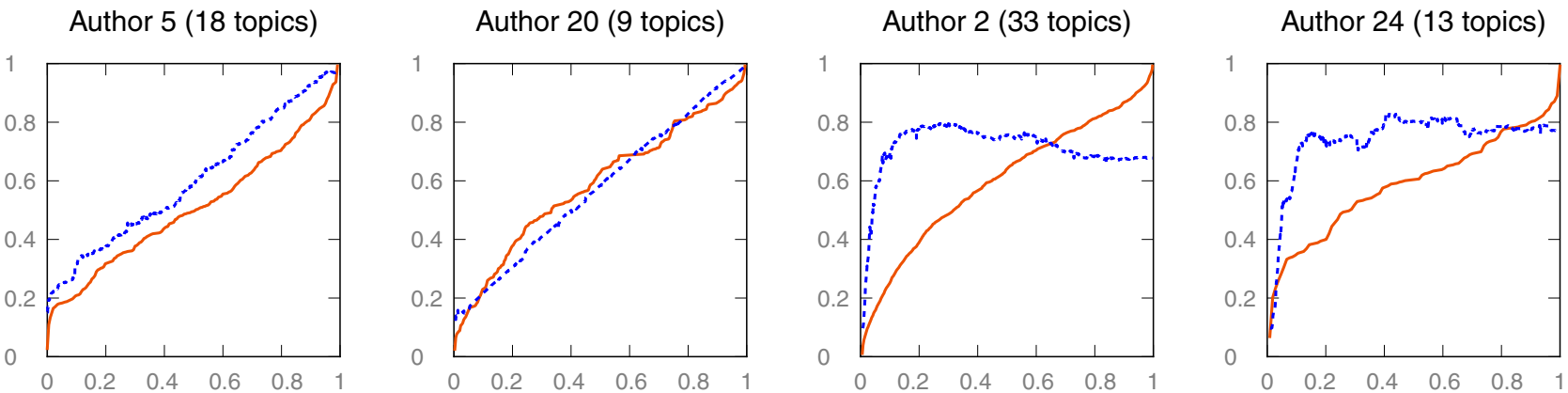


Distribution of queries over time.

- 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



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?

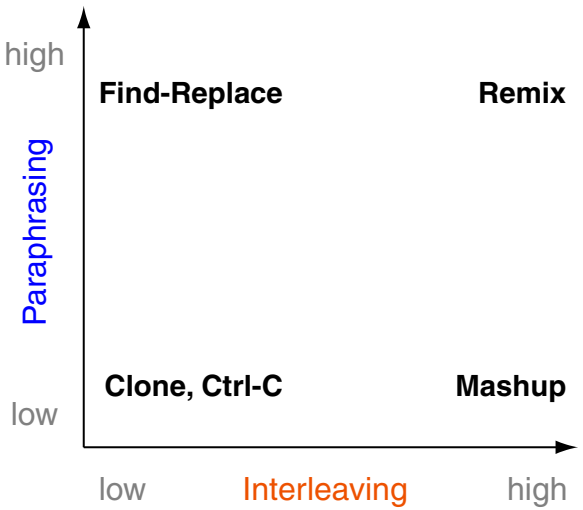
Sub-topic 3. Find biographical information on Barack Obama's mother.

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

[Turnitin 2012]

Details: Paraphrasing & Interleaving

Classification Scheme for Text Reuse



How to quantify?

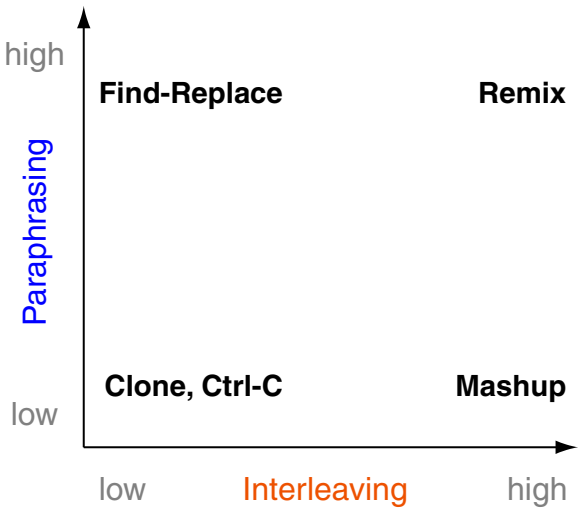
- Measure at the passage level
- Passage: Block of text reused from the same source
- Paraphrasing: simple N-Gram similarity

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

Details: Paraphrasing & Interleaving

Classification Scheme for Text Reuse



Three passages:

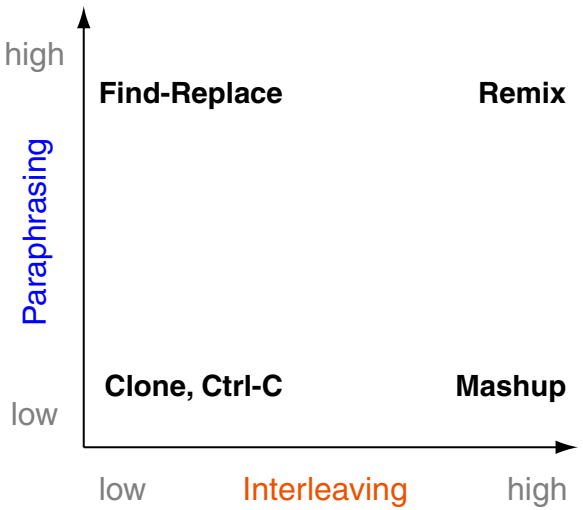
The Emperor penguins are the only penguins that inhabit the Antarctic continent and are the largest of all penguins.Adult Emperor penguins are typically 1.2 meters tall. Juveniles are slightly shorter, only about 90cm to 1m. Emperors weigh around 30 to 40 kg and their weight varies a great deal during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



Paraphrasing: N-Gram similarity

N = 1

Adult Emperor Penguins are ...

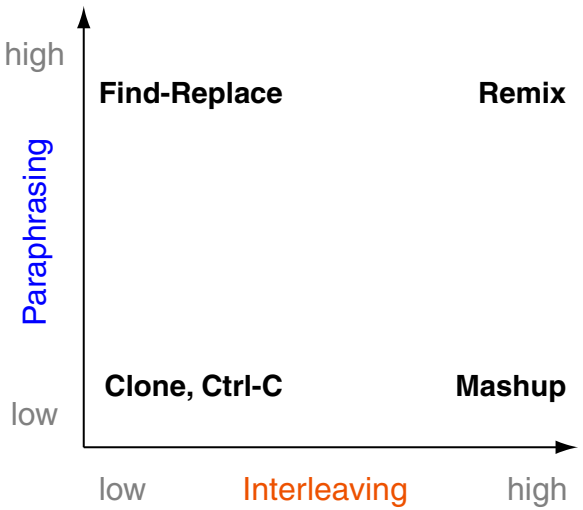
{ Adult
Emperor
Penguins
are }

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 = 2

Adult Emperor Penguins are ...

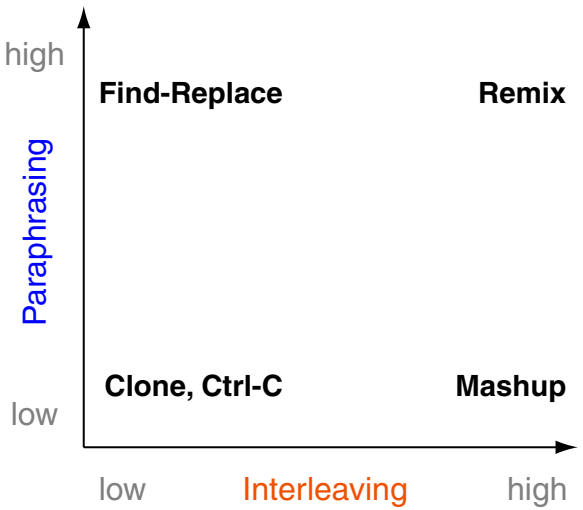
{ Adult Emperor
Emperor Penguins
Penguins are }

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~~ ...

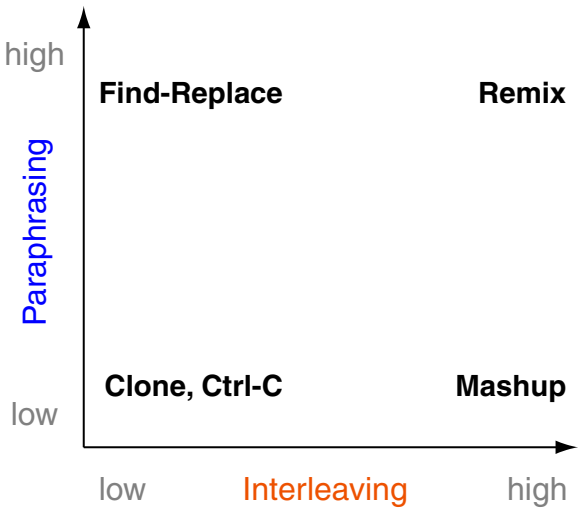
{ Adult Emperor Penguins
Emperor Penguins are }

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

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

- N_c: N-Grams in the passage
- N_s: N-Grams in the source

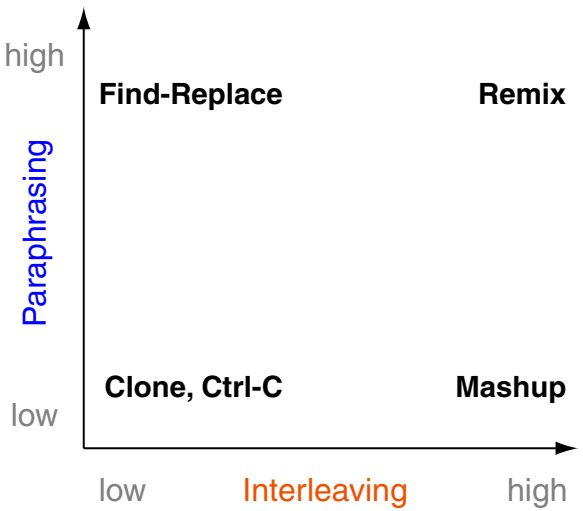
We choose $n = 5$.

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

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

- 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