Crowdsourcing Interaction Logs to Understand Text Reuse from the Web

Martin Potthast Matthias Hagen Michael Völske Benno Stein

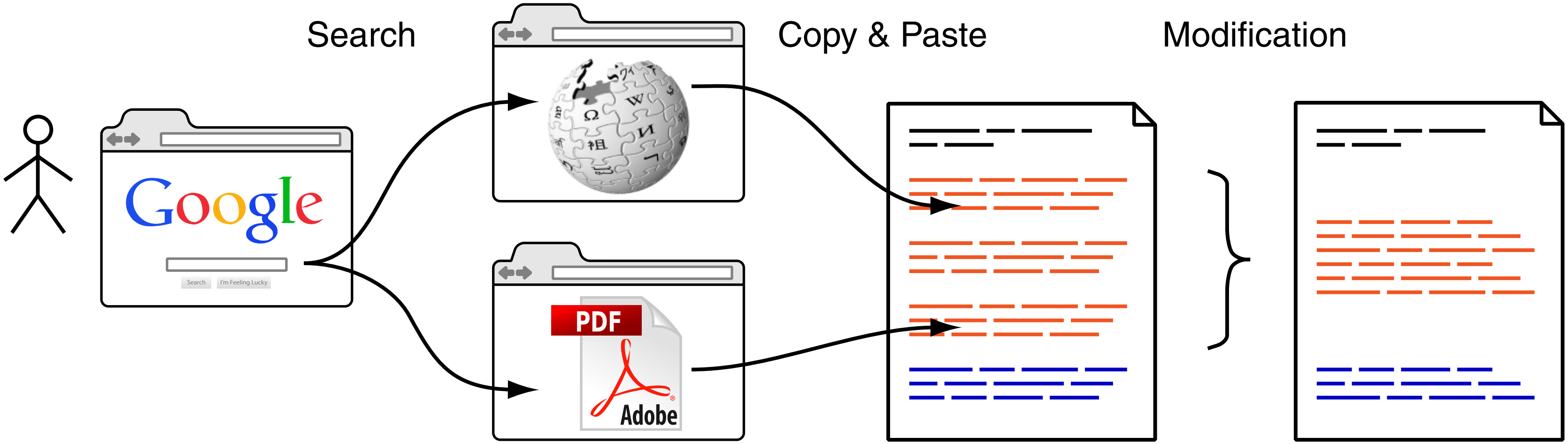
Bauhaus-Universität Weimar
www.webis.de

Outline

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

Introduction

Modeling Text Reuse From the Web



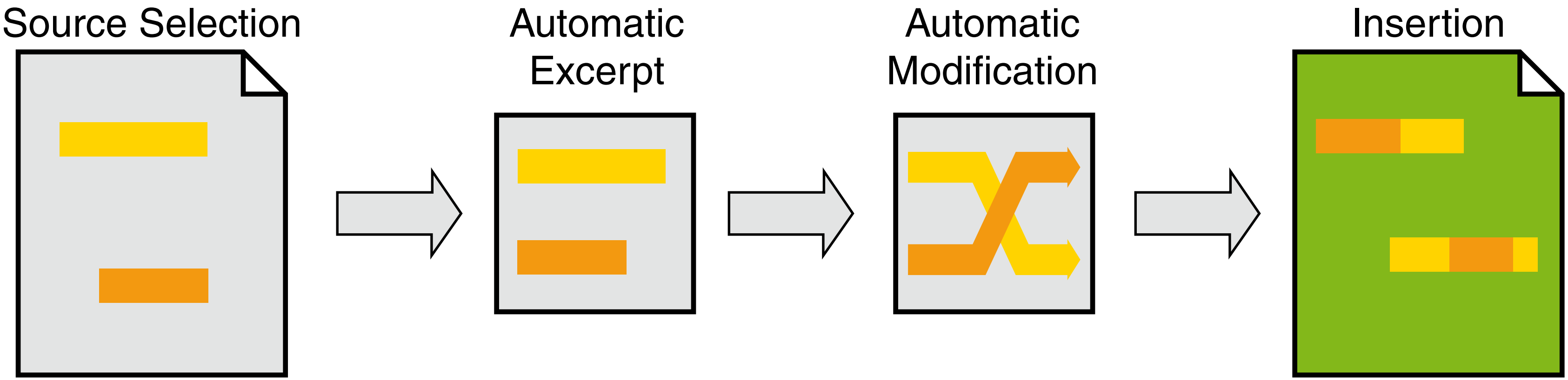
Search

Copy & Paste

Modification

Introduction

Previous Text Reuse Corpora



Source Selection

Automatic Excerpt

Automatic Modification

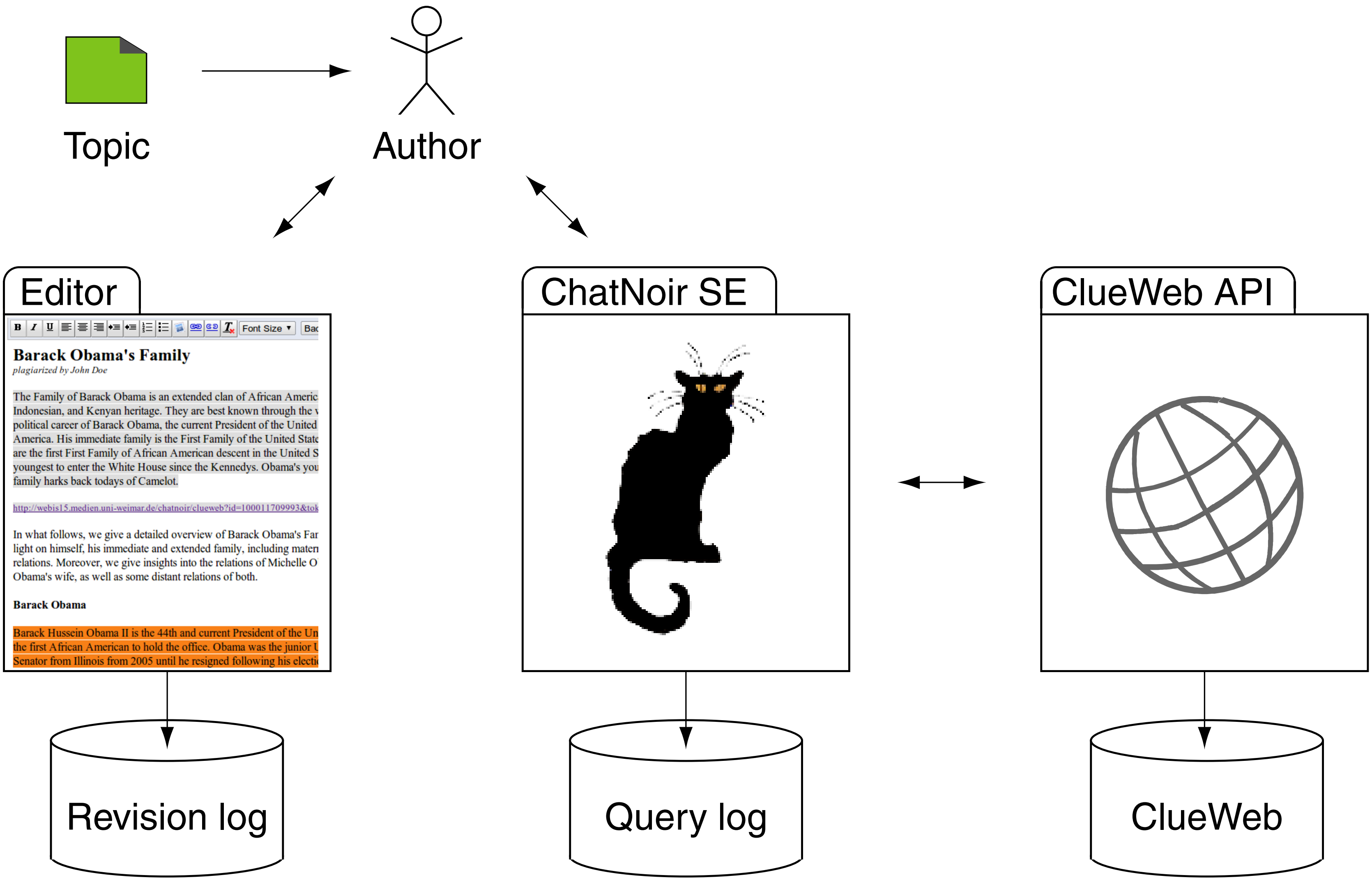
Insertion

- 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



Barack Obama's Family

plagiarized by John Doe

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

http://webis15.medien.uni-weimar.de/chatnoir/clueweb?id=100011709993&tok

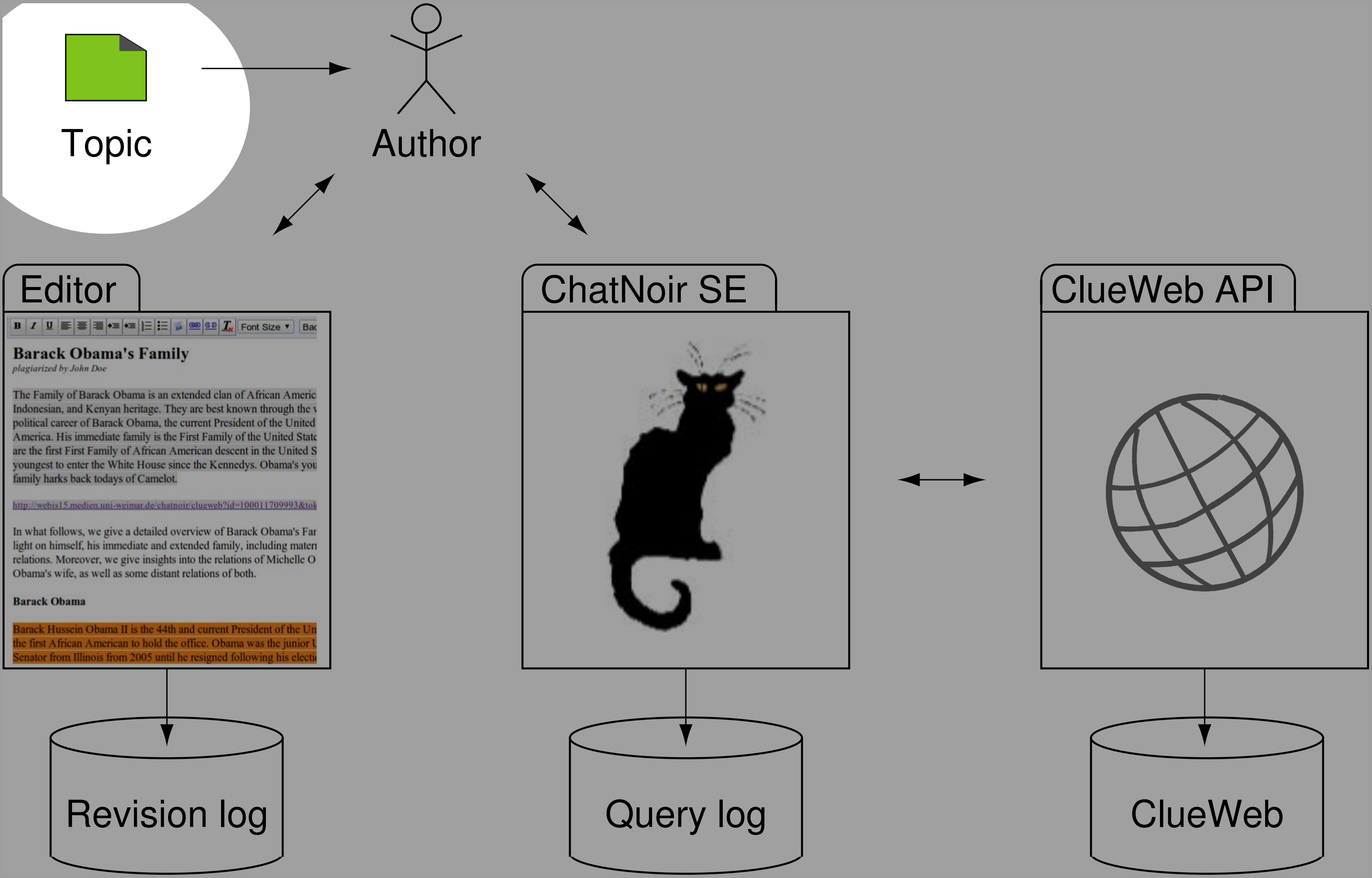
In what follows, we give a detailed overview of Barack Obama's Far
light on himself, his immediate and extended family, including matern
relations. Moreover, we give insights into the relations of Michelle O
Obama's wife, as well as some distant relations of both.

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 U
Senator from Illinois from 2005 until he resigned following his electi

The Webis-TRC-12 Dataset

Construction Overview: Topics



Barack Obama's Family

plagiarized by John Doe

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

http://webis15.medien.uni-weimar.de/chatnoir/clueweb?id=100011709993&tok

In what follows, we give a detailed overview of Barack Obama's Far
light on himself, his immediate and extended family, including matern
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 U
Senator from Illinois from 2005 until he resigned following his electi

The Webis-TRC-12 Dataset

Construction Overview: Topics

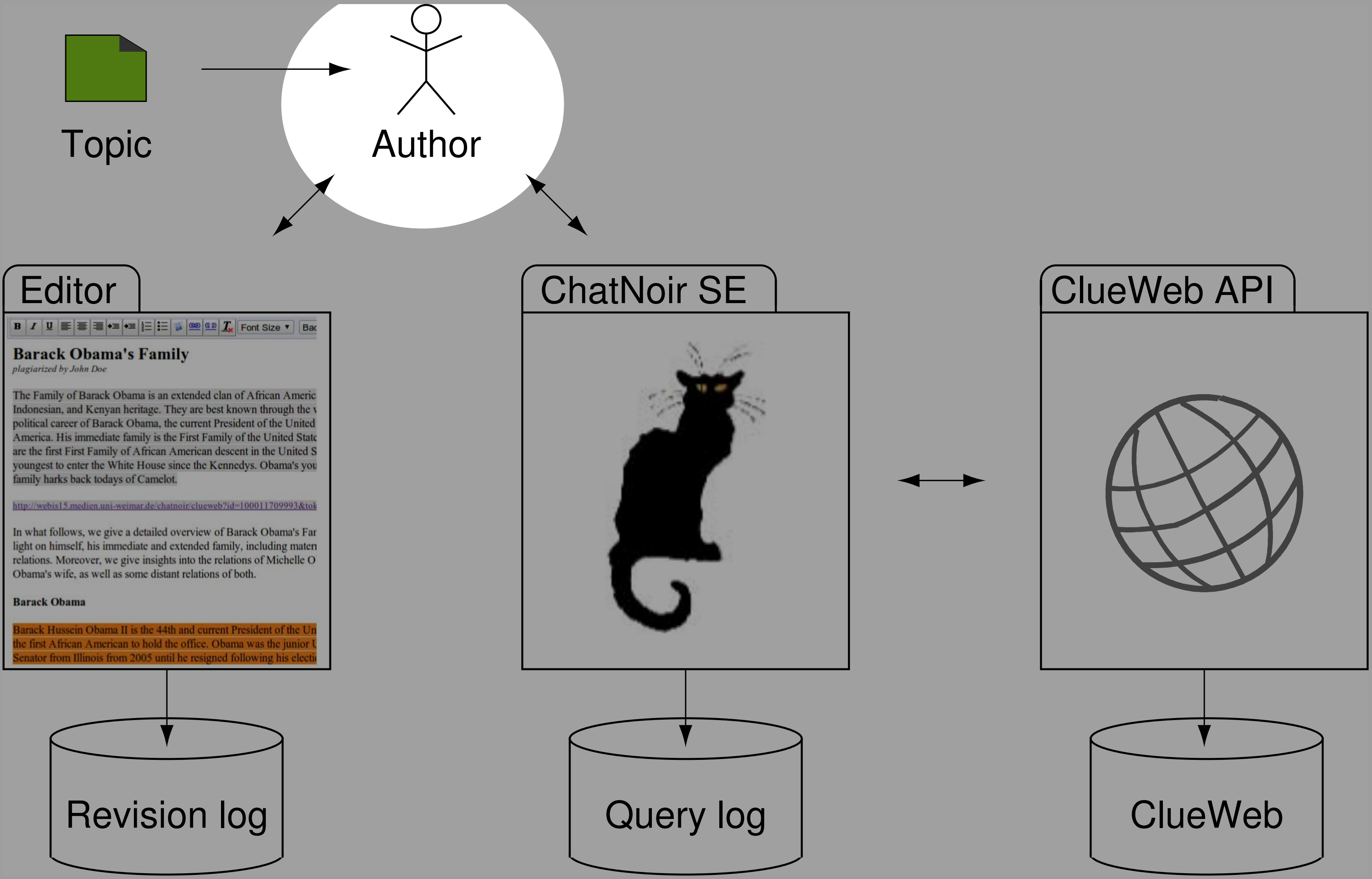
Example topic:

> *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



Barack Obama's Family

plagiarized by John Doe

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

http://webis15.medien.uni-weimar.de/chatnoir/clueweb?id=100011709993&tok

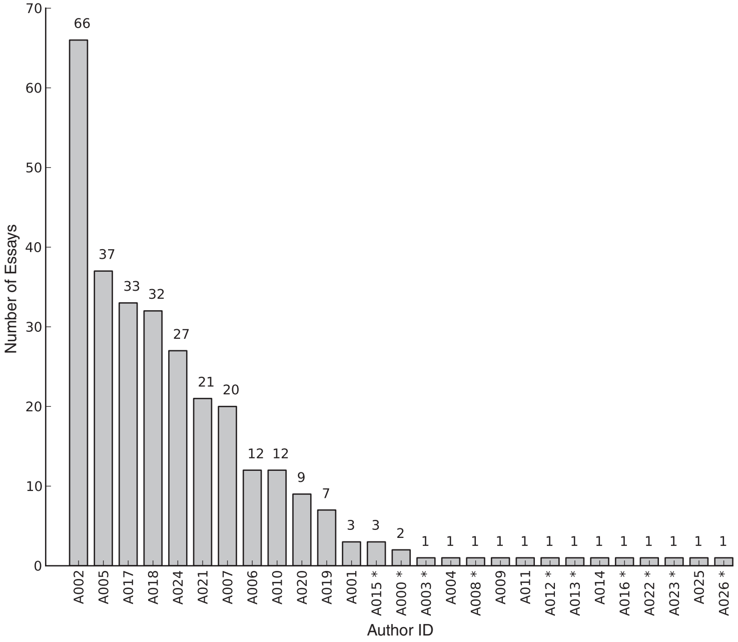
In what follows, we give a detailed overview of Barack Obama's Far
light on himself, his immediate and extended family, including matern
relations. Moreover, we give insights into the relations of Michelle O
Obama's wife, as well as some distant relations of both.

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 U
Senator from Illinois from 2005 until he resigned following his electi

The Webis-TRC-12 Dataset

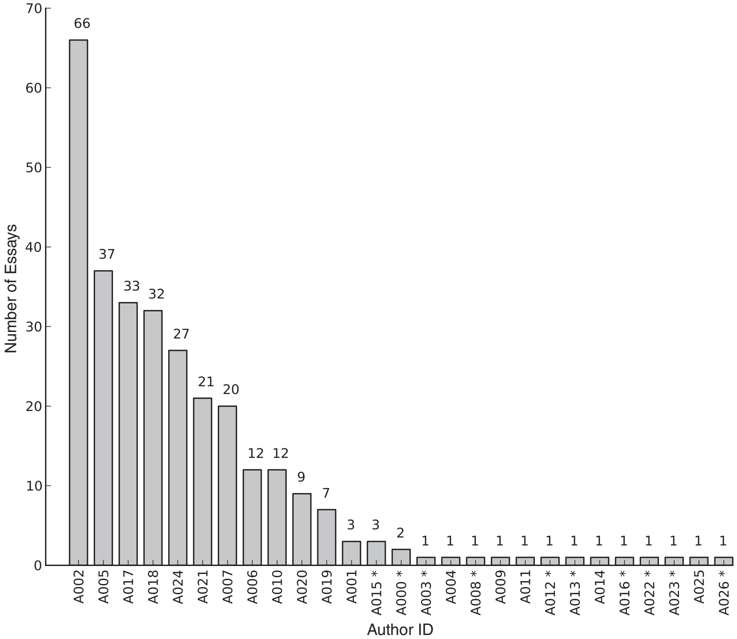
Construction Overview: Authors



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

The Webis-TRC-12 Dataset

Construction Overview: Authors

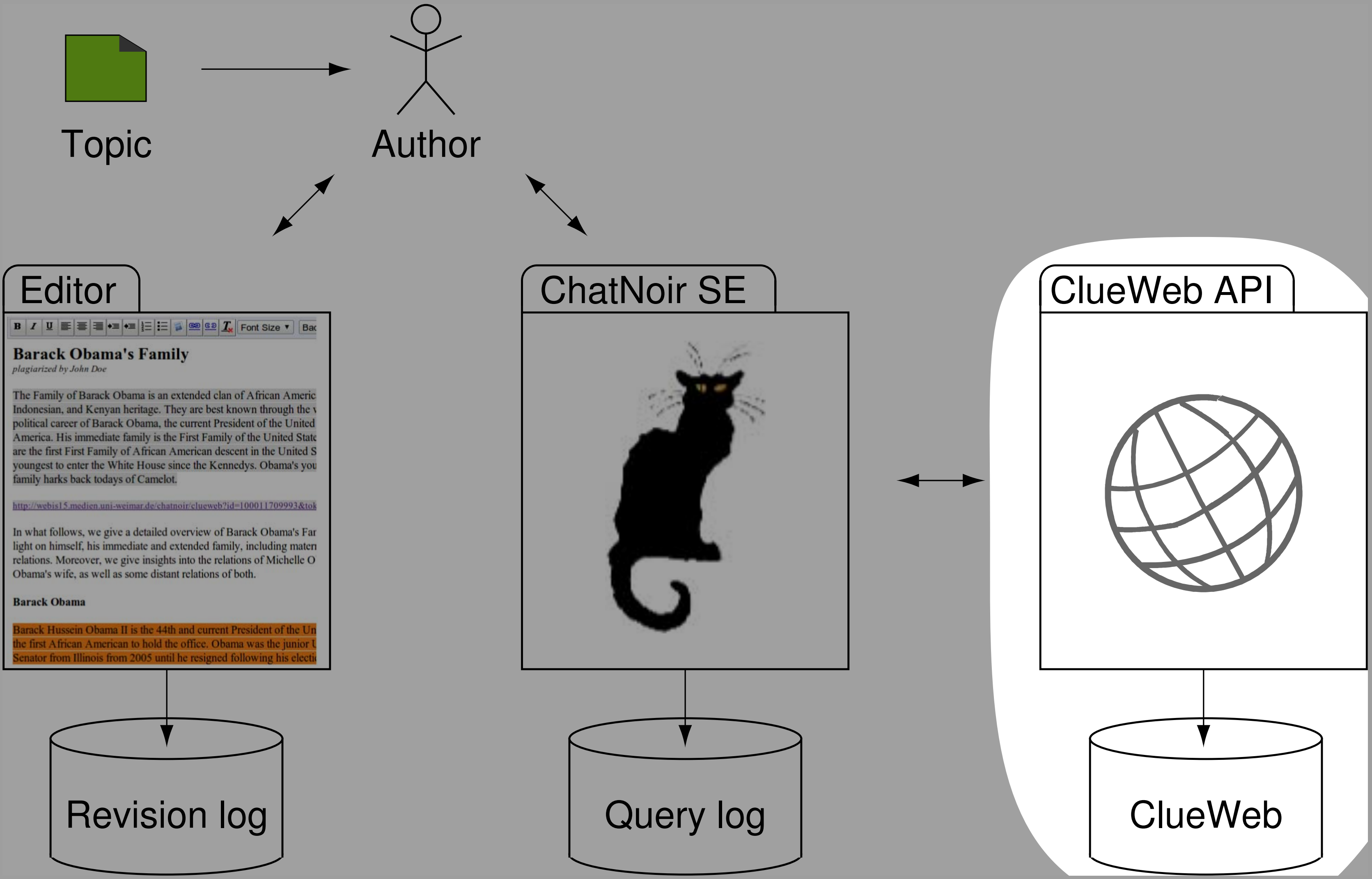


Author Demographics (n=12)	
Age (Median)	37
Years Writing (Median)	8
Academic degree	
Postgrad	41%
Undergrad	25%
None	17%
n/a	17%
English	
Native	67%
Second Language	33%

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

The Webis-TRC-12 Dataset

Construction Overview: Sources



Barack Obama's Family

plagiarized by John Doe

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

http://webis15.medien.uni-weimar.de/chatnoir/clueweb?id=100011709993&tok

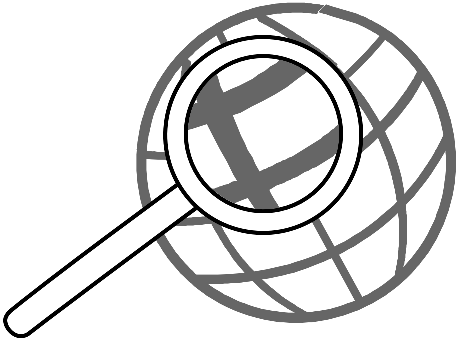
In what follows, we give a detailed overview of Barack Obama's Far light on himself, his immediate and extended family, including matern relations. Moreover, we give insights into the relations of Michelle O Obama's wife, as well as some distant relations of both.

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 U Senator from Illinois from 2005 until he resigned following his electi

The Webis-TRC-12 Dataset

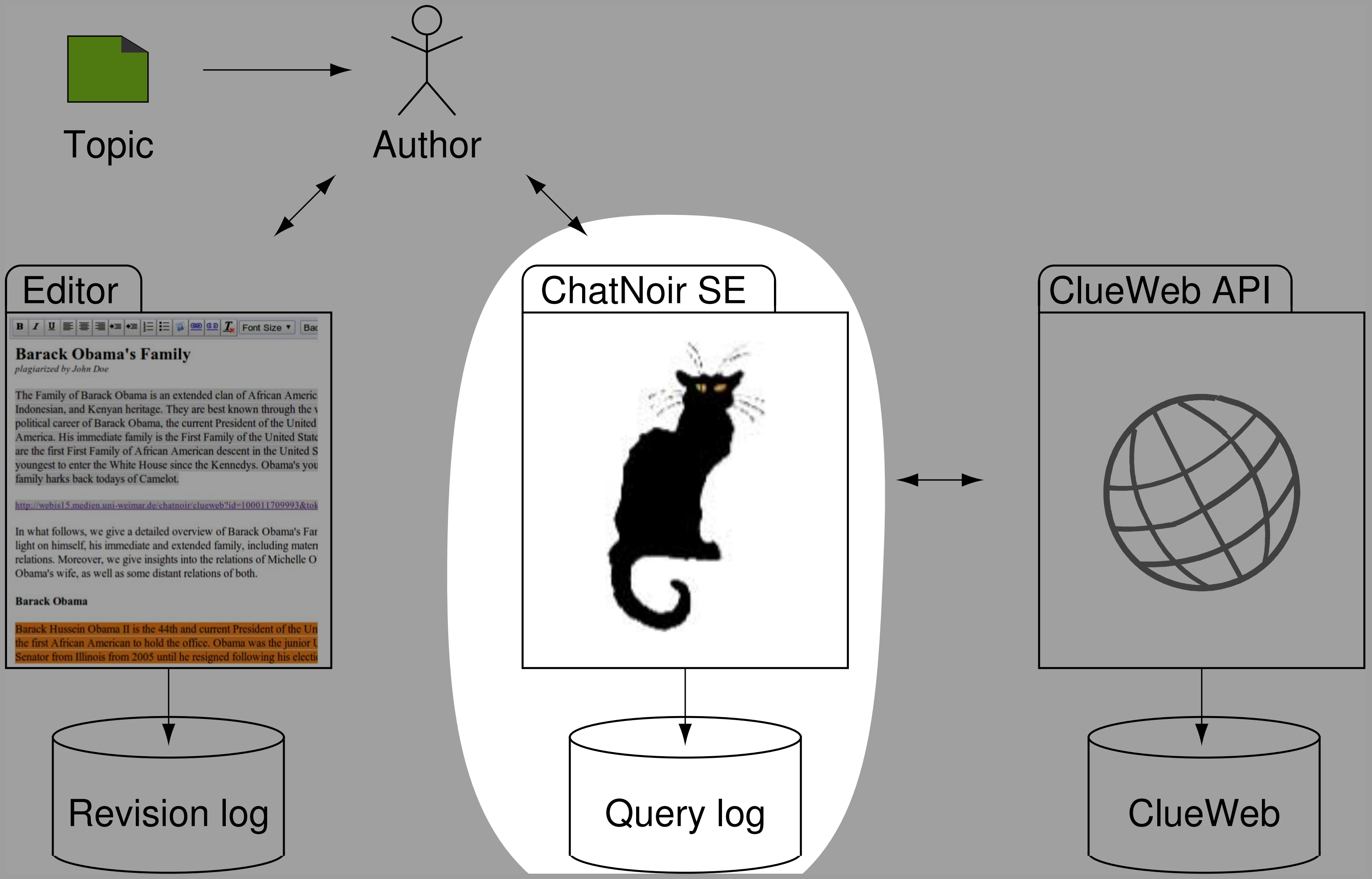
Construction Overview: Sources



- 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



Barack Obama's Family

plagiarized by John Doe

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

http://webis15.medien.uni-weimar.de/chatnoir/clueweb?id=100011709993&tok

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

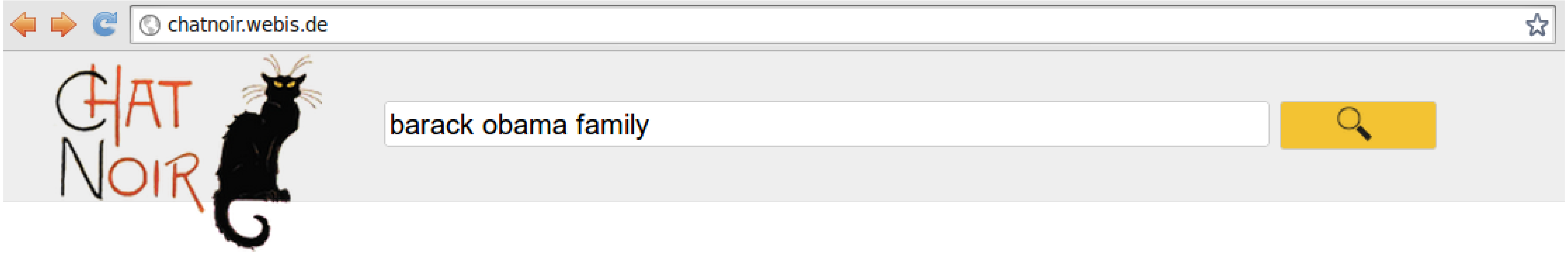
Barack Obama

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 U
Senator from Illinois from 2005 until he resigned following his electi

© Völske 2013

The Webis-TRC-12 Dataset

Construction Overview: Search Engine



chatnoir.webis.de

barack obama family

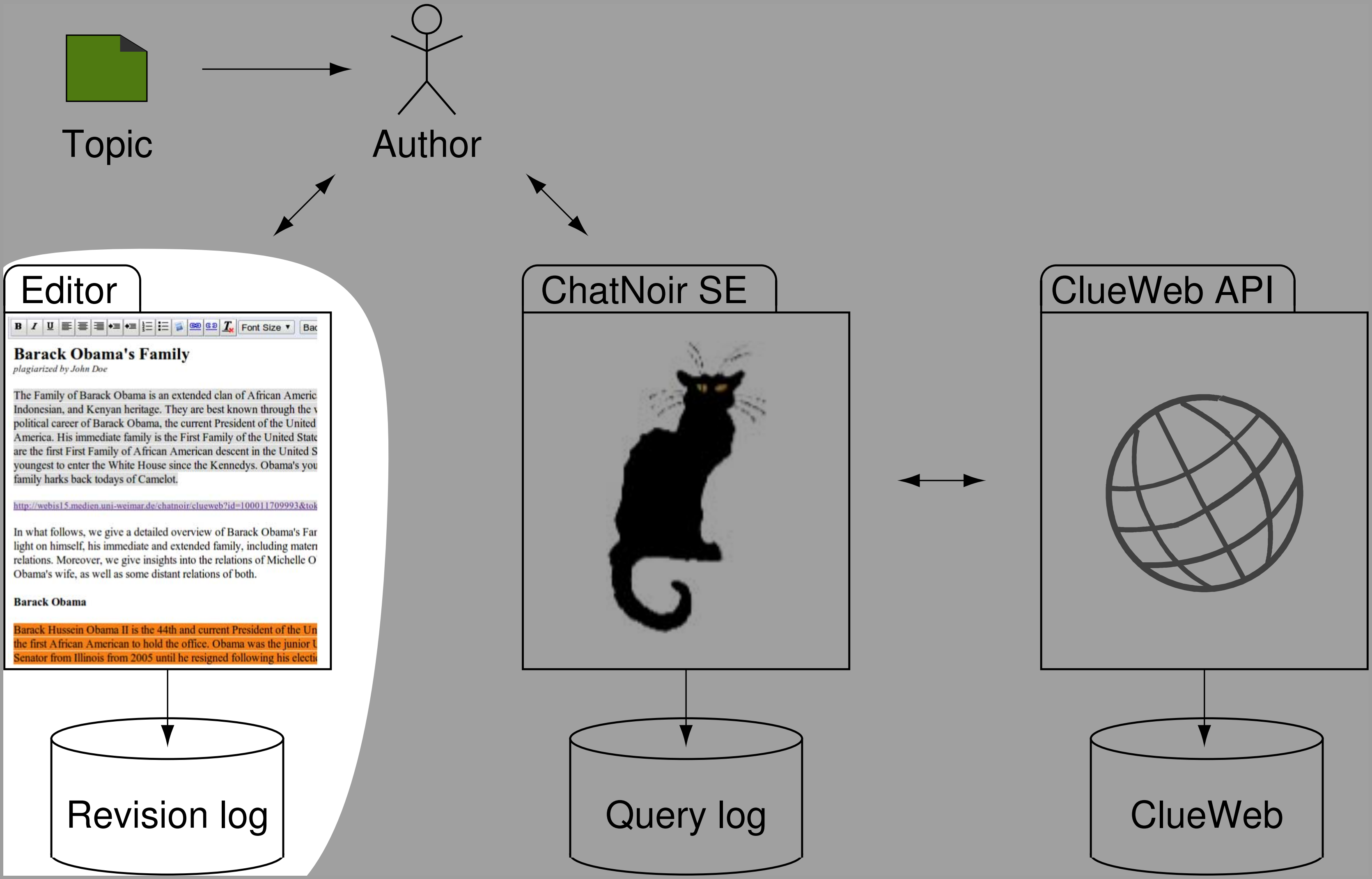
- Used for source retrieval [chatnoir.webis.de]

- 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 Americ
Indonesian, and Kenyan heritage. They are best known through the v
political career of Barack Obama, the current President of the United
America. His immediate family is the First Family of the United State
are the first First Family of African American descent in the United S
youngest to enter the White House since the Kennedys. Obama's you
family harks back todays of Camelot.

http://webis15.medien.uni-weimar.de/chatnoir/clueweb?id=100011709993&tok

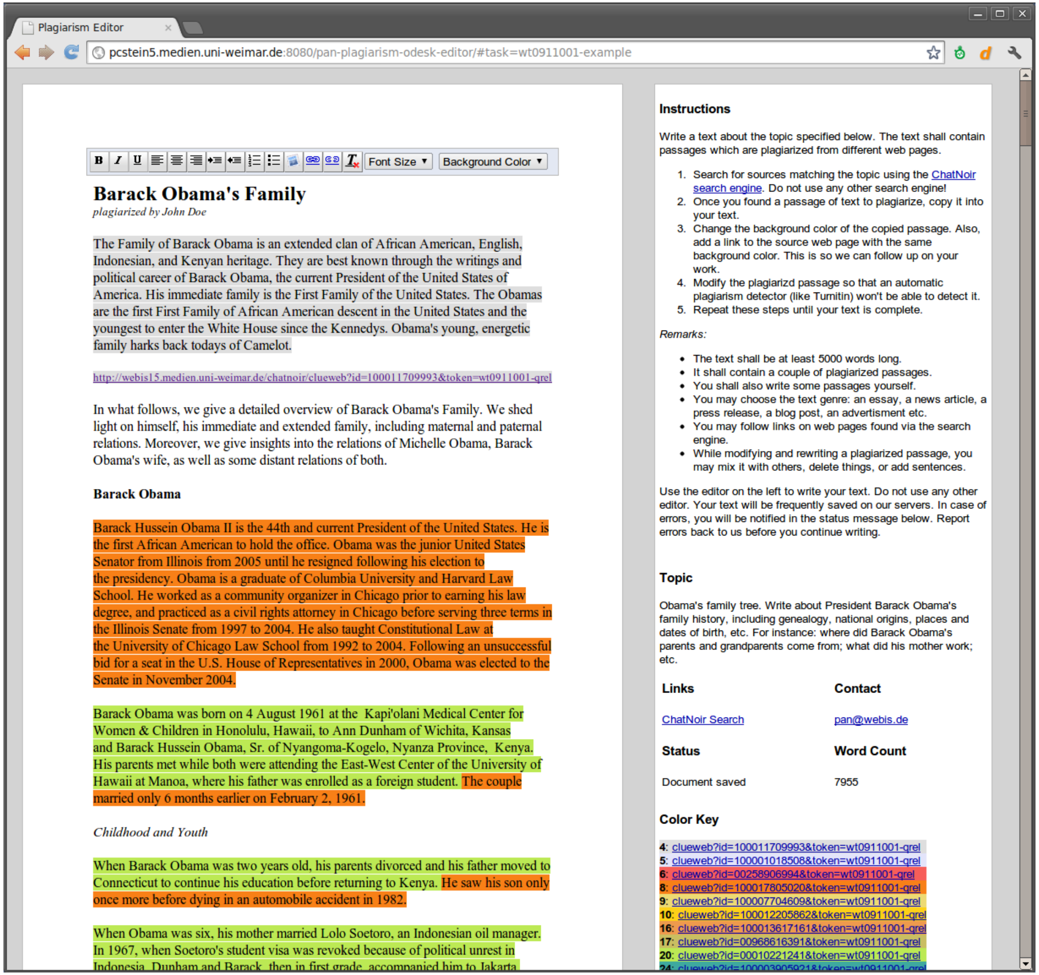
In what follows, we give a detailed overview of Barack Obama's Far
light on himself, his immediate and extended family, including matern
relations. Moreover, we give insights into the relations of Michelle O
Obama's wife, as well as some distant relations of both.

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 U
Senator from Illinois from 2005 until he resigned following his electi

The Webis-TRC-12 Dataset

Construction Overview: Editor


Plagiarism Editor
pcstein5.medien.uni-weimar.de:8080/pan-plagiarism-odesk-editor/#task=wt0911001-example
B I U
Font Size ▾
Background Color ▾

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 harks back todays of Camelot.

http://webis15.medien.uni-weimar.de/chatnoir/clueweb?id=100011709993&token=wt0911001-qrel

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 Kapi'olani 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 6 months 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 1967, when Soetoro's student visa was revoked because of political unrest in Indonesia, Dunham and Barack, then in first grade, accompanied him to Jakarta

Instructions

Write a text about the topic specified below. The text shall contain passages which are plagiarized from different web pages.

1. Search for sources matching the topic using the ChatNoir 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 background color. This is so we can follow up on your work.
4. Modify the plagiarizd passage so that an automatic plagiarism detector (like Turnitin) won't be able to detect it.
5. 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 advertisment 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.

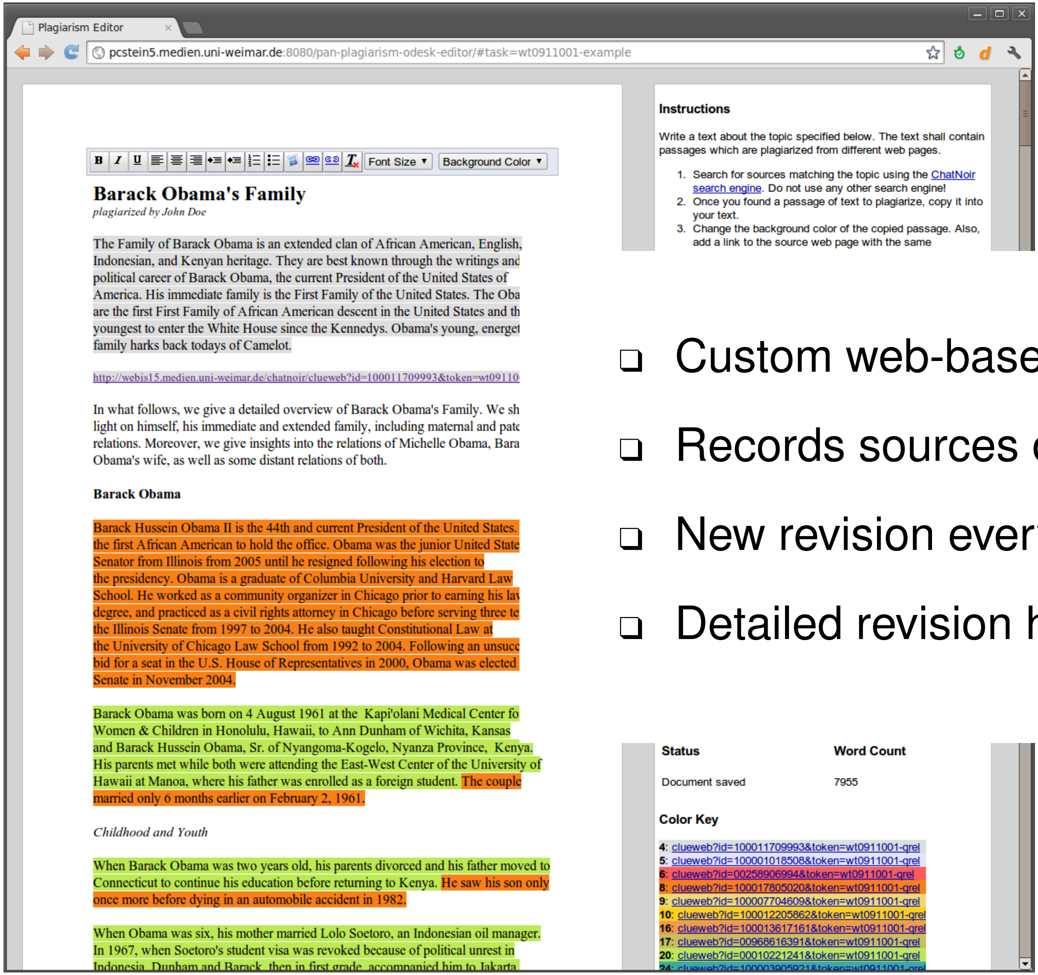
Links	**Contact**
ChatNoir Search	pan@webis.de
Status	**Word Count**
Document saved	7955

Color Key

4: clueweb?id=100011709993&token=wt0911001-qrel
5: clueweb?id=100001018508&token=wt0911001-qrel
6: clueweb?id=00258906994&token=wt0911001-qrel
8: clueweb?id=100017805020&token=wt0911001-qrel
9: clueweb?id=100007704609&token=wt0911001-qrel
10: clueweb?id=100012205862&token=wt0911001-qrel
16: clueweb?id=100013617161&token=wt0911001-qrel
17: clueweb?id=00968616391&token=wt0911001-qrel
20: clueweb?id=00010221241&token=wt0911001-qrel
24: clueweb?id=100003905921&token=wt0911001-qrel

The Webis-TRC-12 Dataset

Construction Overview: Editor



Plagiarism Editor
pcstein5.medien.uni-weimar.de:8080/pan-plagiarism-odesk-editor/#task=wt0911001-example
B I U
Font Size ▾
Background Color ▾

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 Oba
are the first First Family of African American descent in the United States and th
youngest to enter the White House since the Kennedys. Obama's young, energet
family harks back todays of Camelot.

http://webis15.medien.uni-weimar.de/chatnoir/clueweb?id=100011709993&token=wt09110

In what follows, we give a detailed overview of Barack Obama's Family. We sh
light on himself, his immediate and extended family, including maternal and pate
relations. Moreover, we give insights into the relations of Michelle Obama, Bara
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 State
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 lav
degree, and practiced as a civil rights attorney in Chicago before serving three te
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 unsucc
bid for a seat in the U.S. House of Representatives in 2000, Obama was elected
Senate in November 2004.

Barack Obama was born on 4 August 1961 at the Kapi'olani Medical Center fo
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 6 months 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 1967, when Soetoro's student visa was revoked because of political unrest in
Indonesia, Dunham and Barack, then in first grade, accompanied him to Jakarta.

Instructions

Write a text about the topic specified below. The text shall contain passages which are plagiarized from different web pages.

1. Search for sources matching the topic using the ChatNoir 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

Status	**Word Count**
Document saved	7955

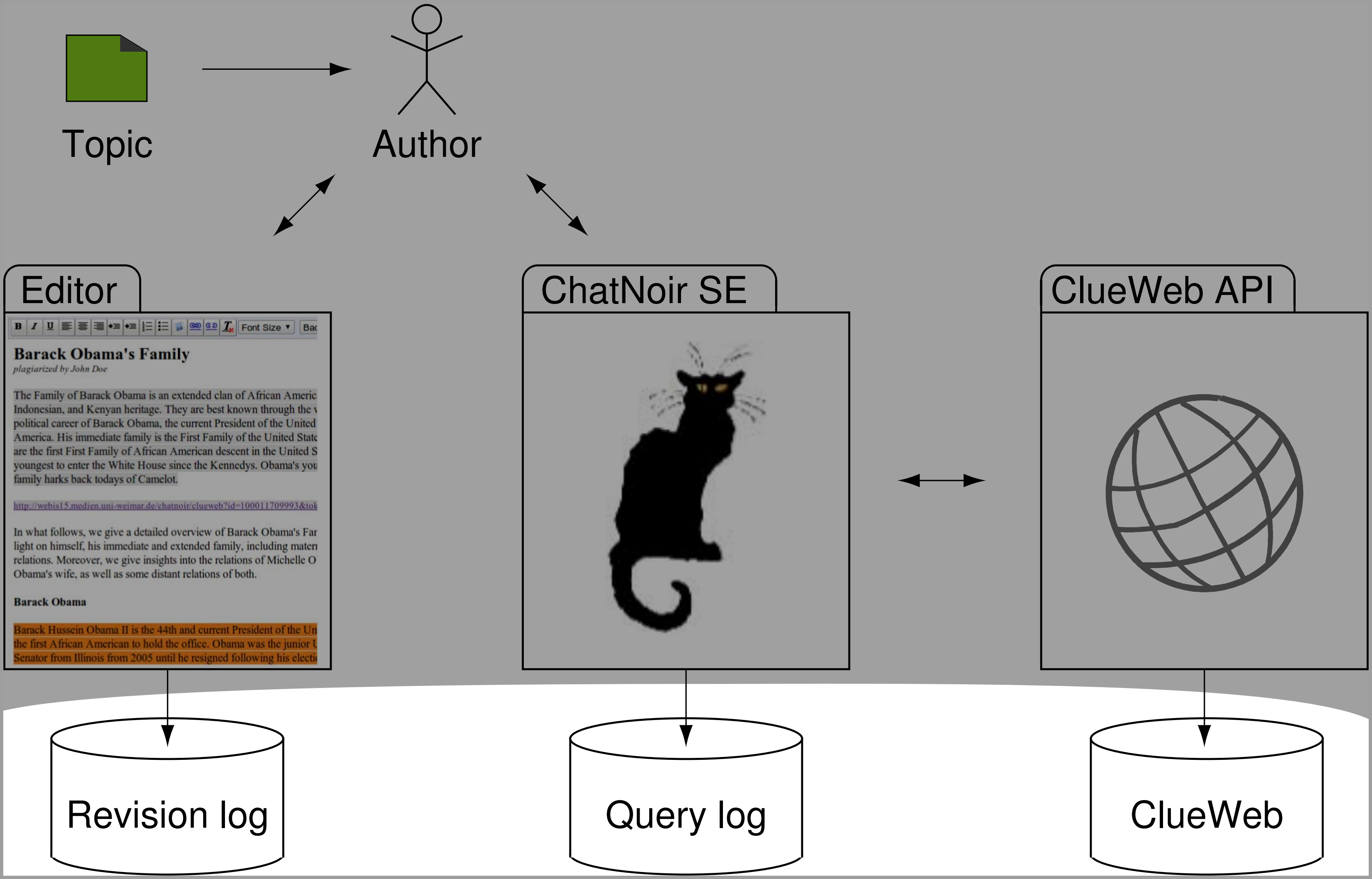
Color Key

4: clueweb?id=100011709993&token=wt0911001-qrel
5: clueweb?id=100001018508&token=wt0911001-qrel
6: clueweb?id=00258906994&token=wt0911001-qrel
8: clueweb?id=100017805020&token=wt0911001-qrel
9: clueweb?id=100007704609&token=wt0911001-qrel
10: clueweb?id=100012205862&token=wt0911001-qrel
16: clueweb?id=100013617161&token=wt0911001-qrel
17: clueweb?id=00968616391&token=wt0911001-qrel
20: clueweb?id=00010221241&token=wt0911001-qrel
24: clueweb?id=100003905921&token=wt0911001-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



Topic

Author

Editor

B I U Font Size ▾ Bac

Barack Obama's Family

plagiarized by John Doe

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

http://webis15.medien.uni-weimar.de/chatnoir/clueweb?id=100011709993&tok

In what follows, we give a detailed overview of Barack Obama's Far
light on himself, his immediate and extended family, including matern
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 U
Senator from Illinois from 2005 until he resigned following his electi

ChatNoir SE

ClueWeb API

Revision log

Query log

ClueWeb

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?

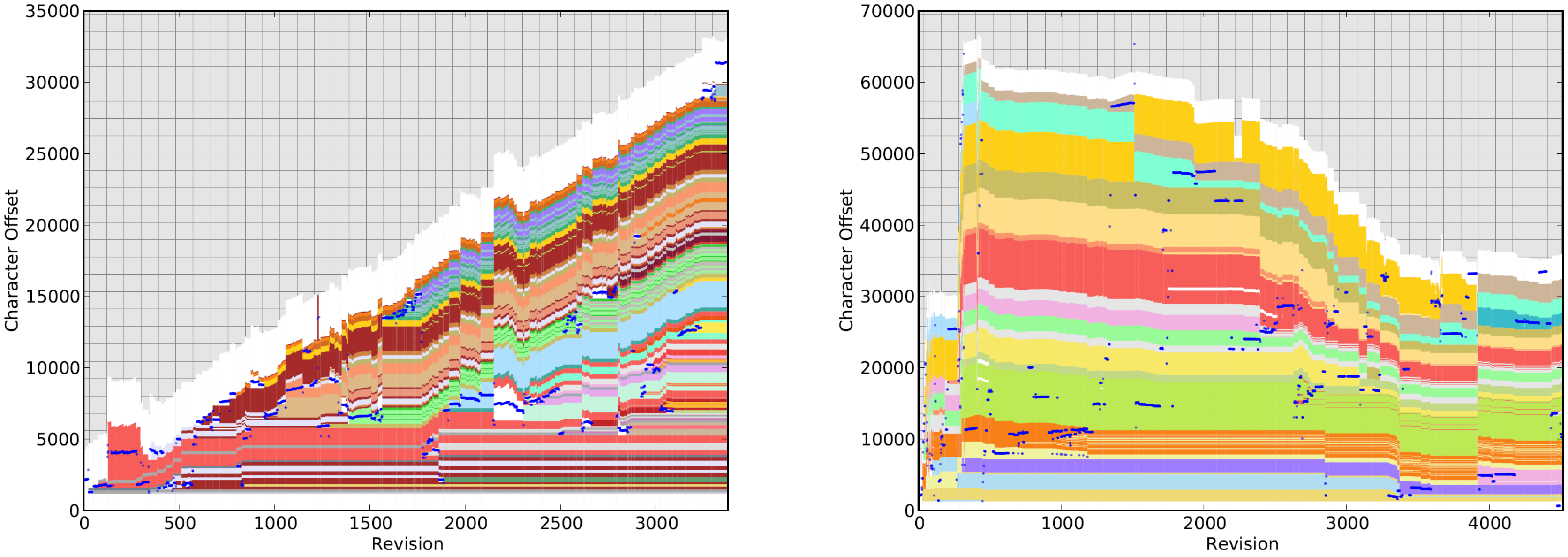
We expect new research insights and impacts to

- text reuse detection
- query formulation
- paraphrasing

Categorizing Crowdsourced Text Reuse

Categorizing Crowdsourced Text Reuse

Build-Up Versus Boil-Down Reuse

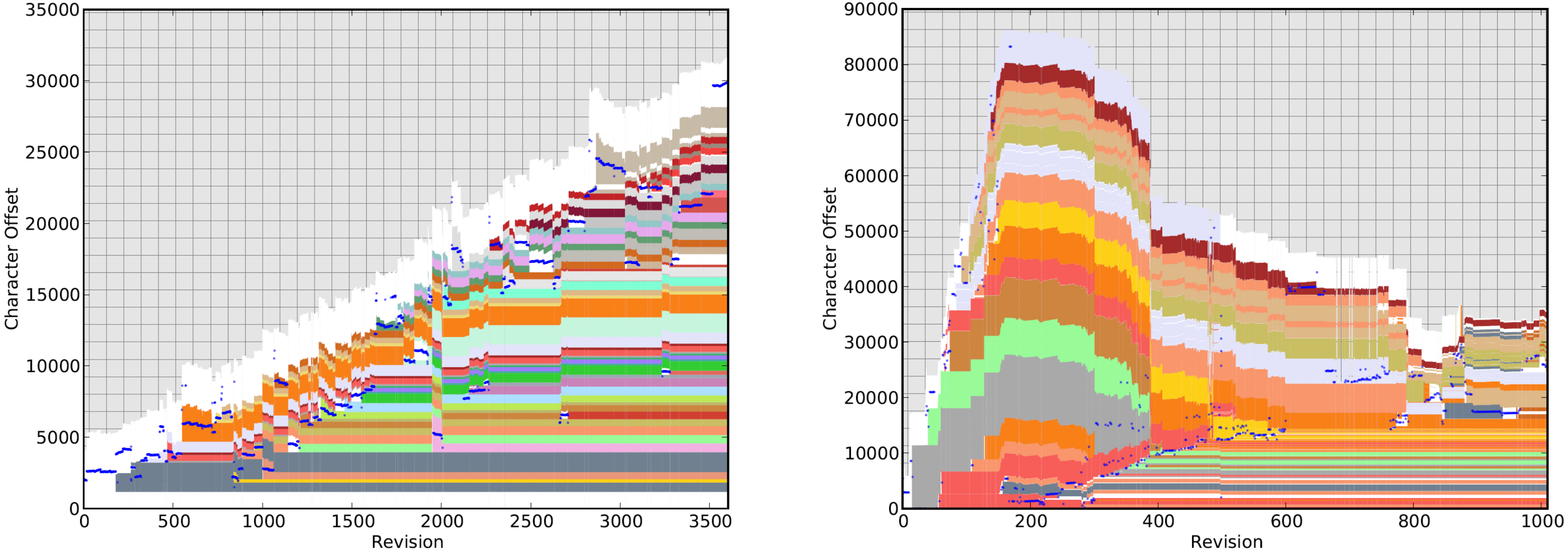


Build-up reuse (left) versus boil-down reuse (right).

- ❑ 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

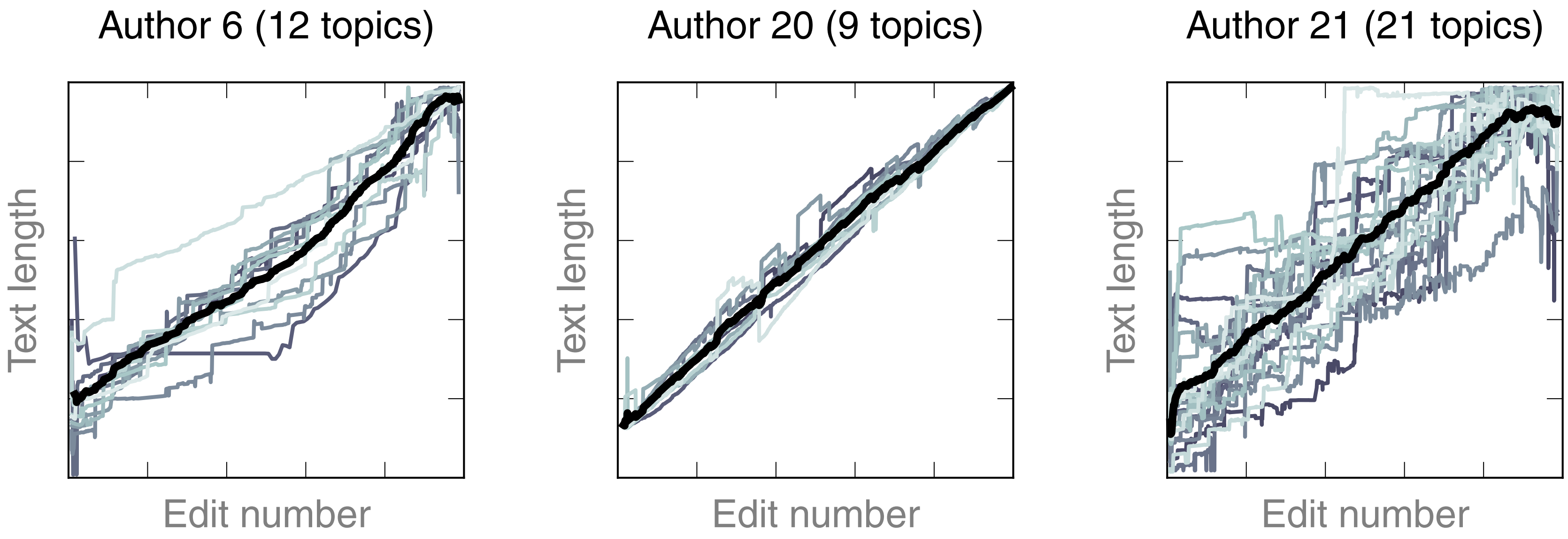


Build-up reuse (left) versus boil-down reuse (right).

- 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

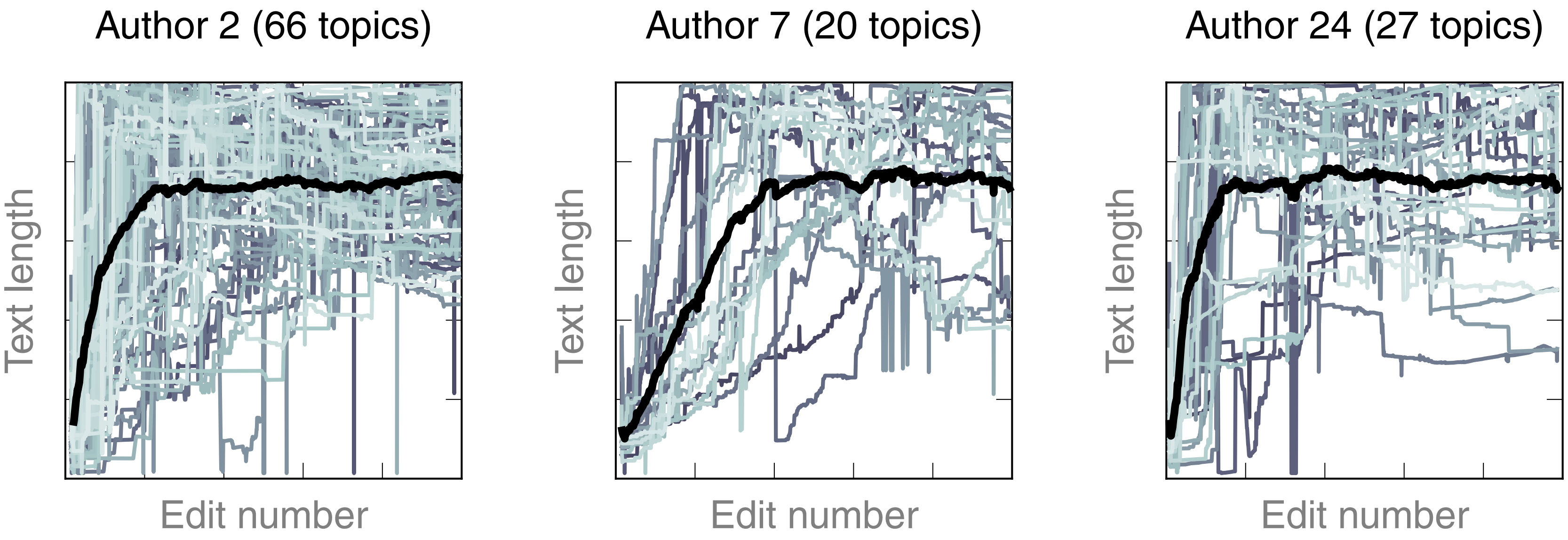


Build-up reuse: Averaged editing histories by authors.

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

Categorizing Crowdsourced Text Reuse

Build-Up Versus Boil-Down Reuse



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

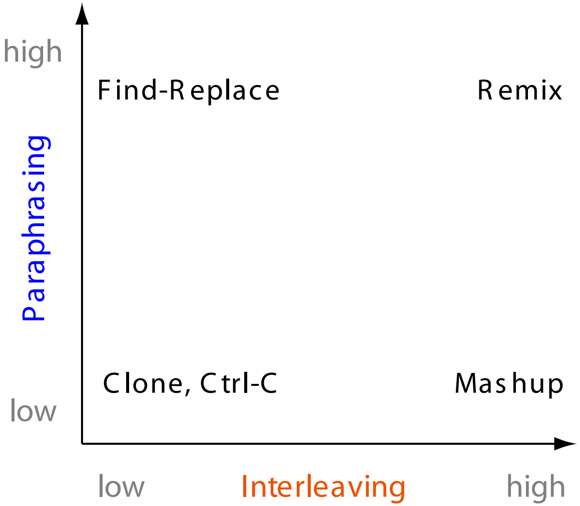
Mashup

Classification Scheme for Text Reuse.

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

Categorizing Crowdsourced Text Reuse

Classification Scheme for Text Reuse

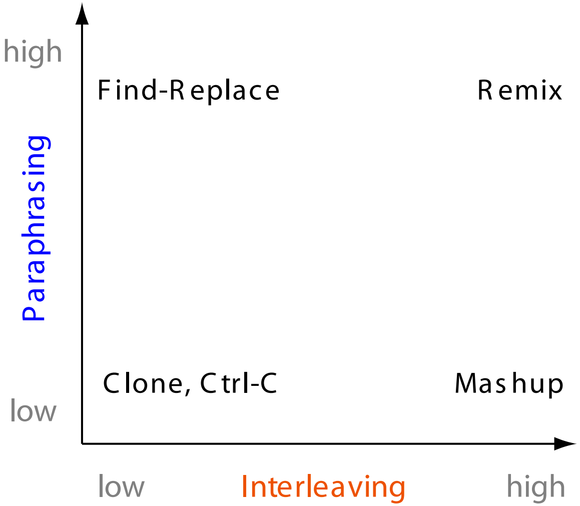


Classification Scheme for Text Reuse.

- ❑ 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



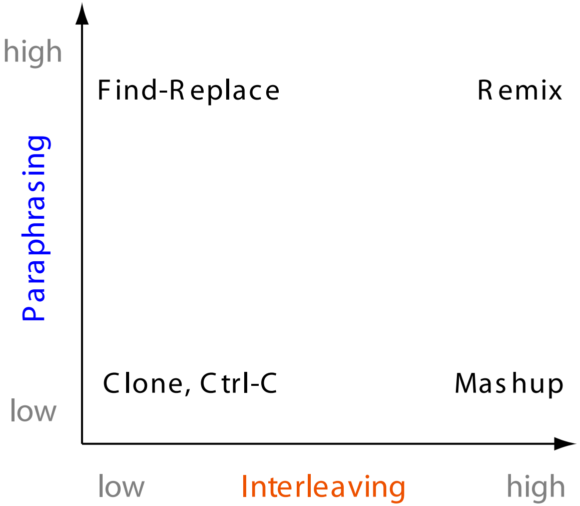
- ❑ Quantify: N-Gram similarity and ratio of passages to sources
  [details]
- ❑ Measure for all essays
- ❑ Hypothesis: will show evidence of authors' individual text reuse styles

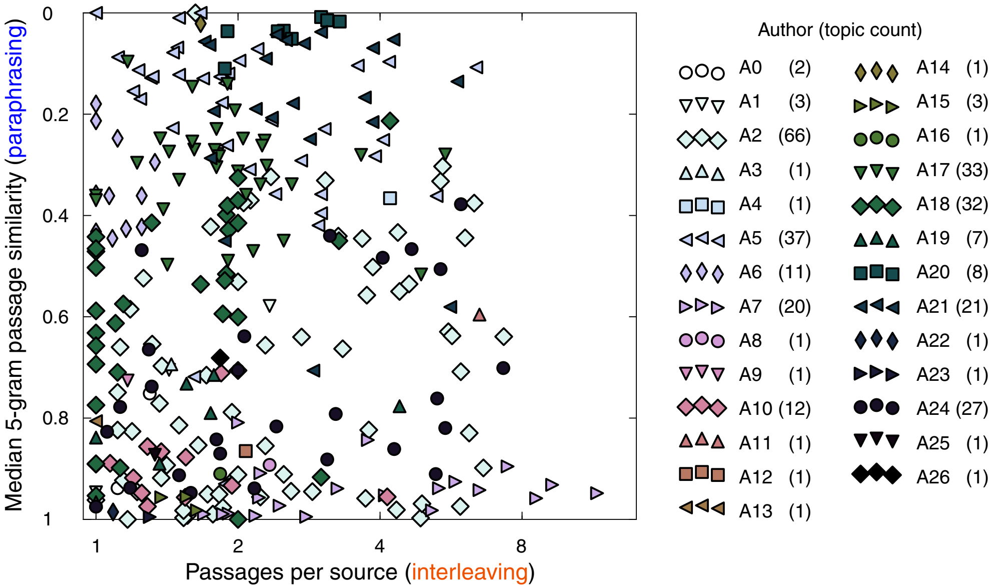
Classification Scheme for Text Reuse.

- ❑ 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



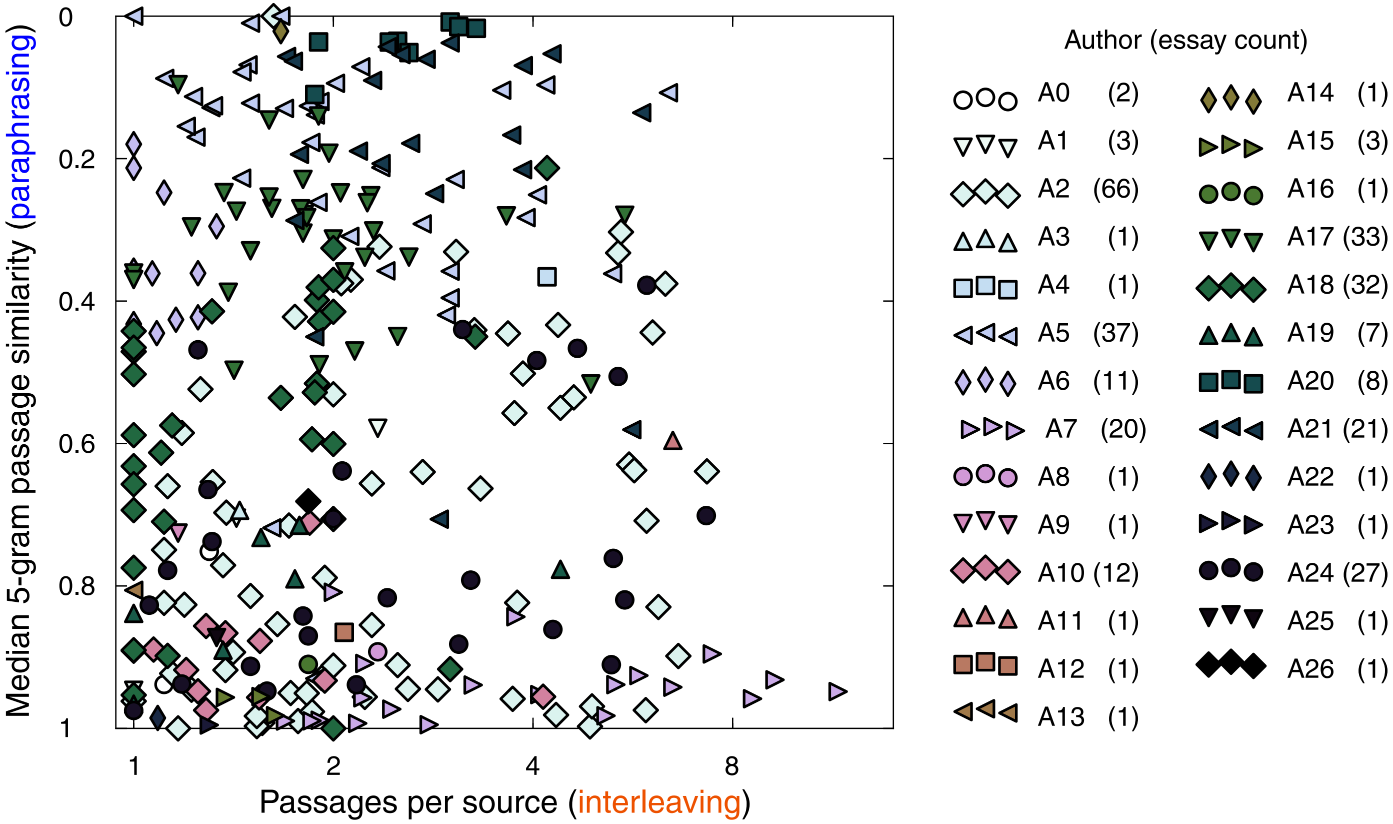


Classification Scheme for Text Reuse.

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

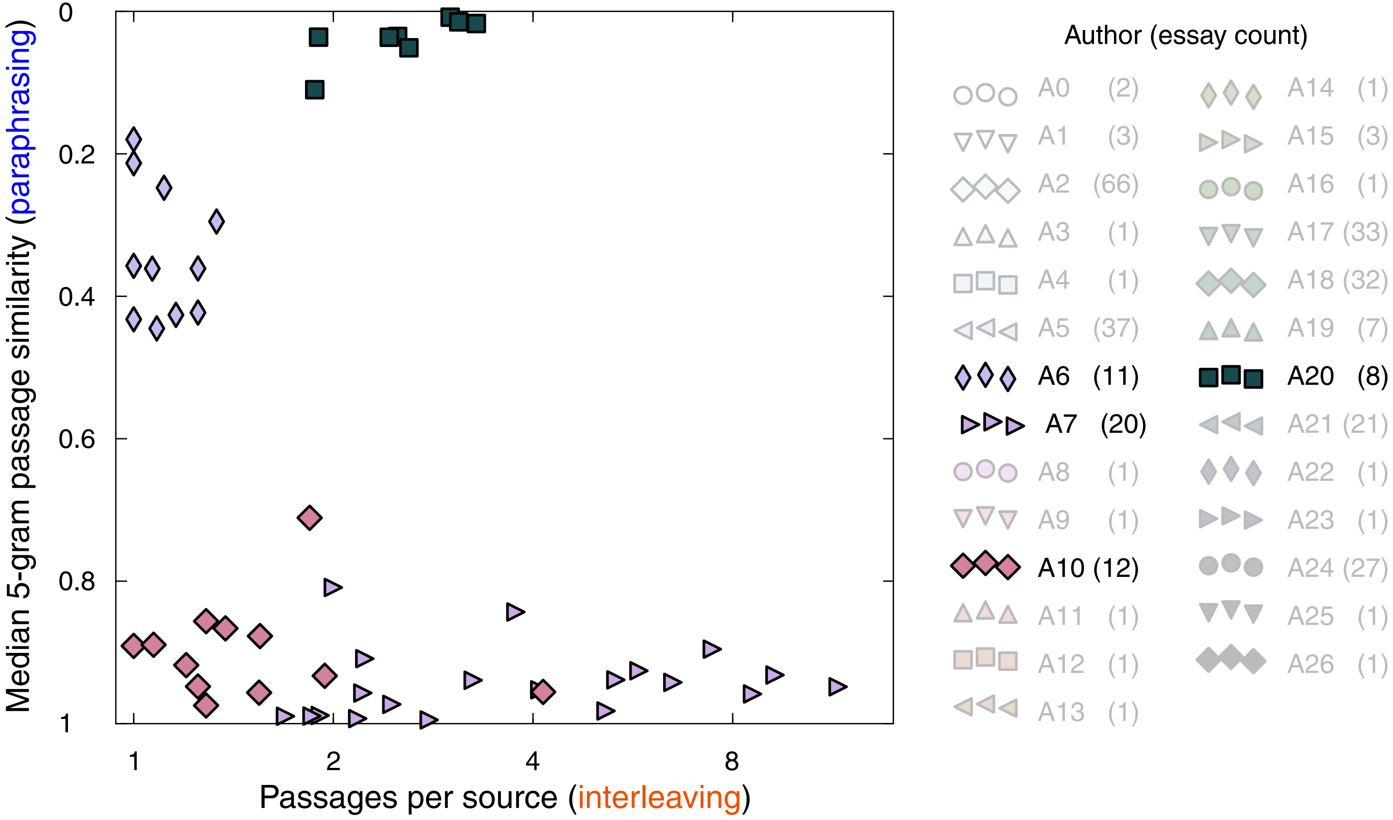
Categorizing Crowdsourced Text Reuse

Classification Scheme for Text Reuse



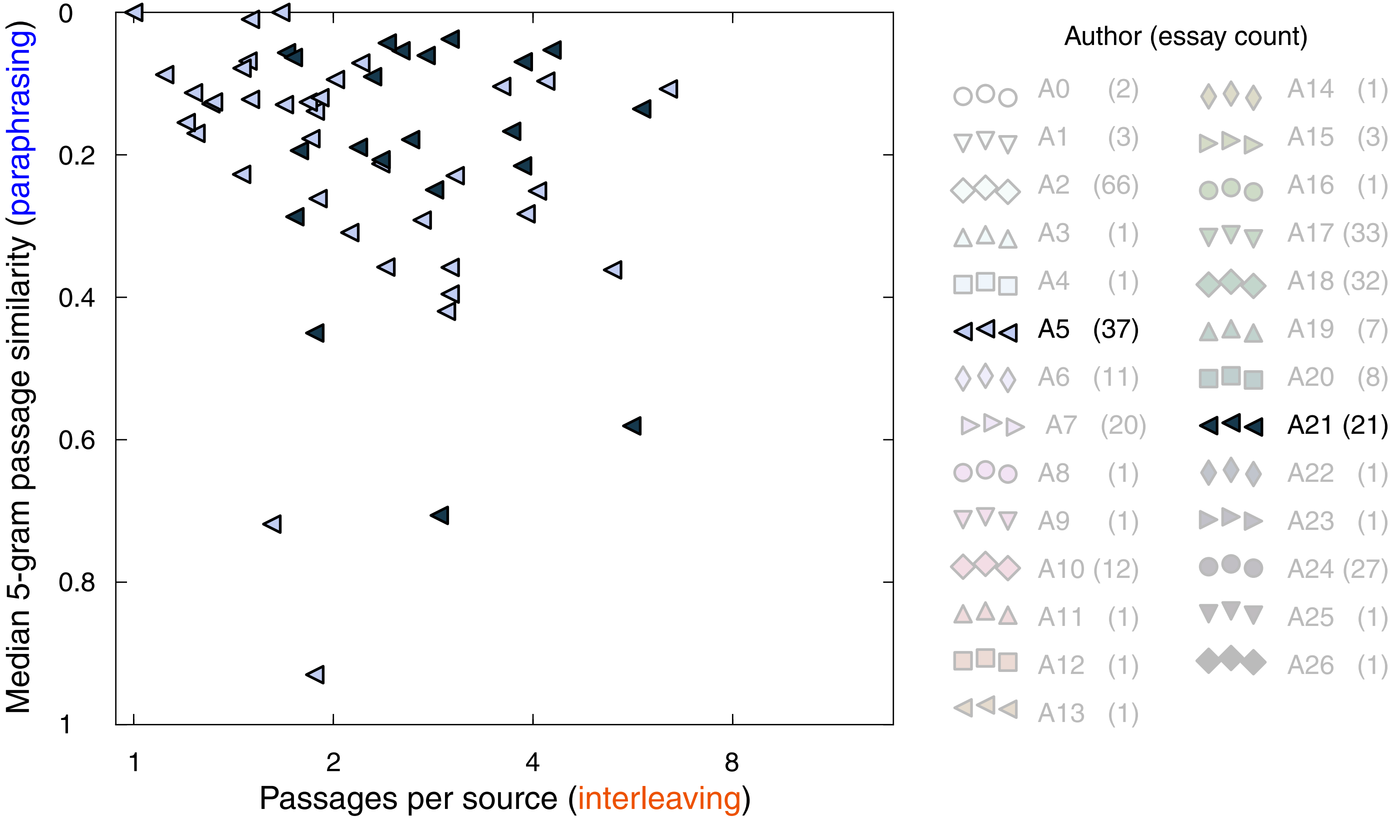
Categorizing Crowdsourced Text Reuse

Classification Scheme for Text Reuse



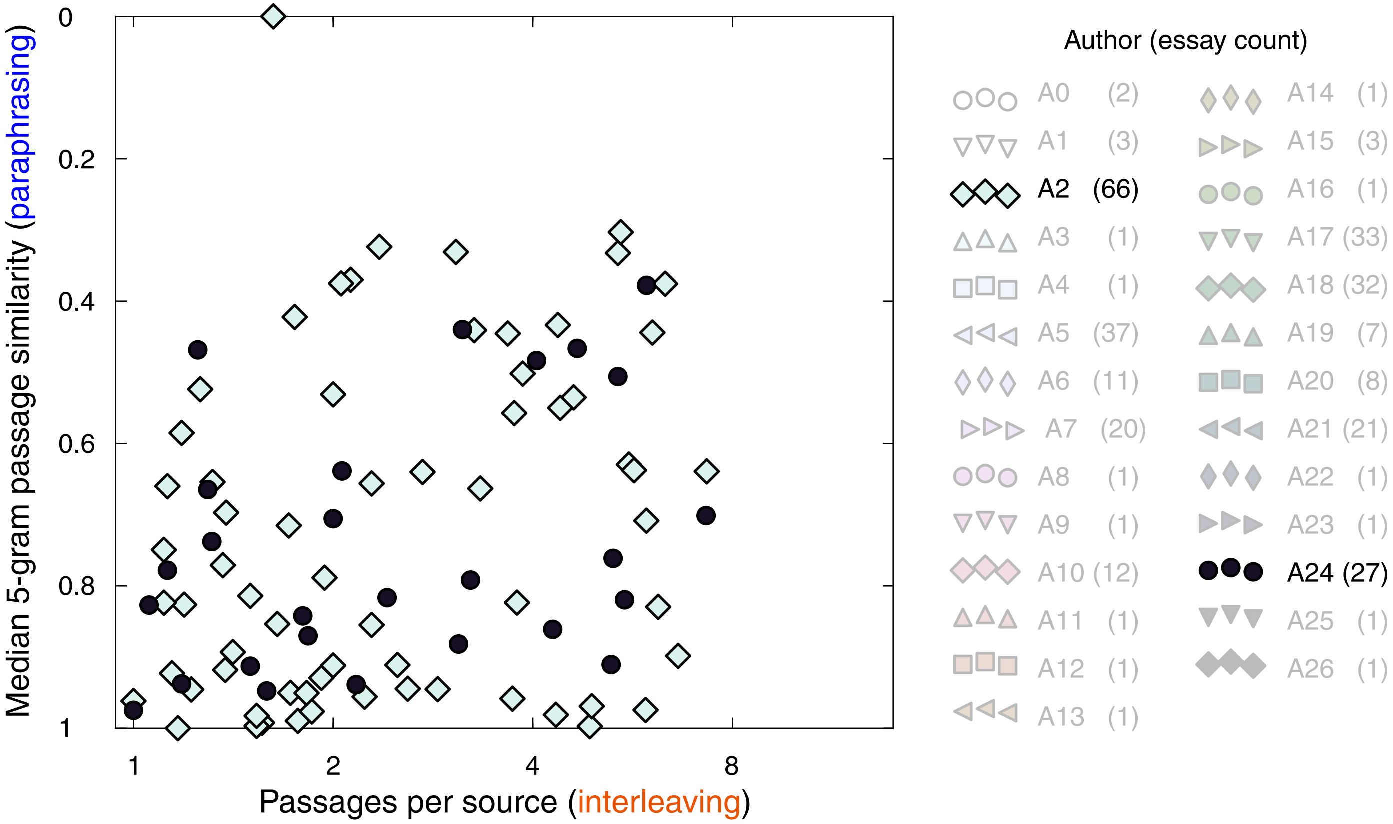
Categorizing Crowdsourced Text Reuse

Classification Scheme for Text Reuse



Categorizing Crowdsourced Text Reuse

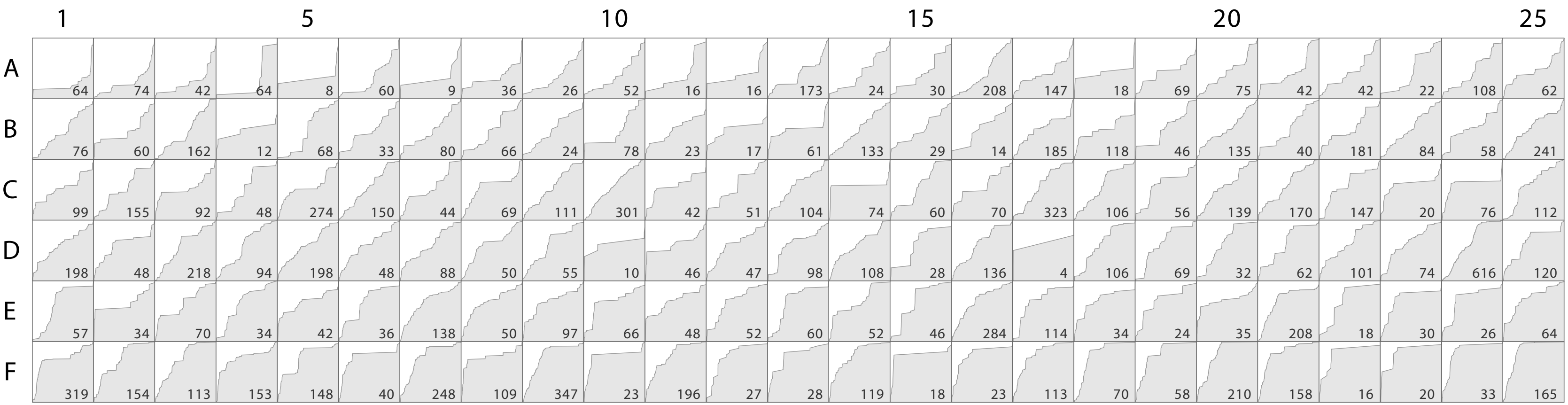
Classification Scheme for Text Reuse



Search Missions For Source Retrieval

Search Missions For Source Retrieval

Distribution of Queries Over Time

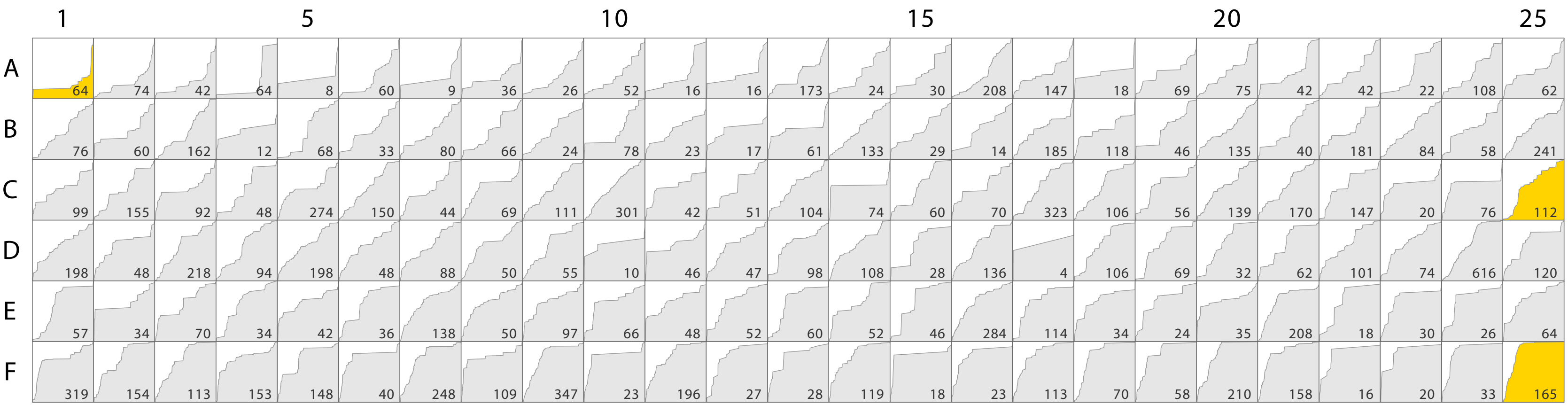


Distribution of queries over time.

- 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

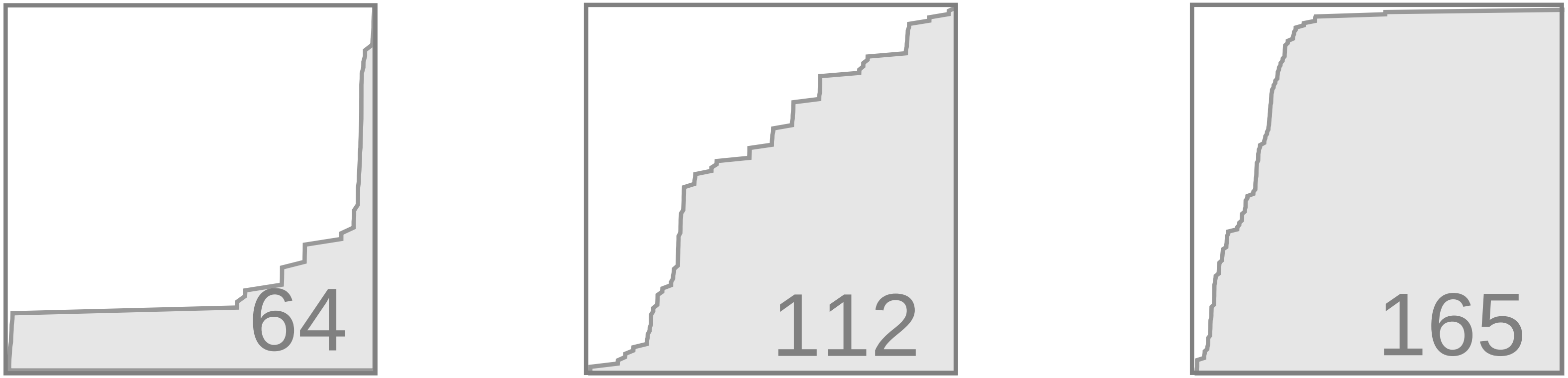


Distribution of queries over time.

- 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

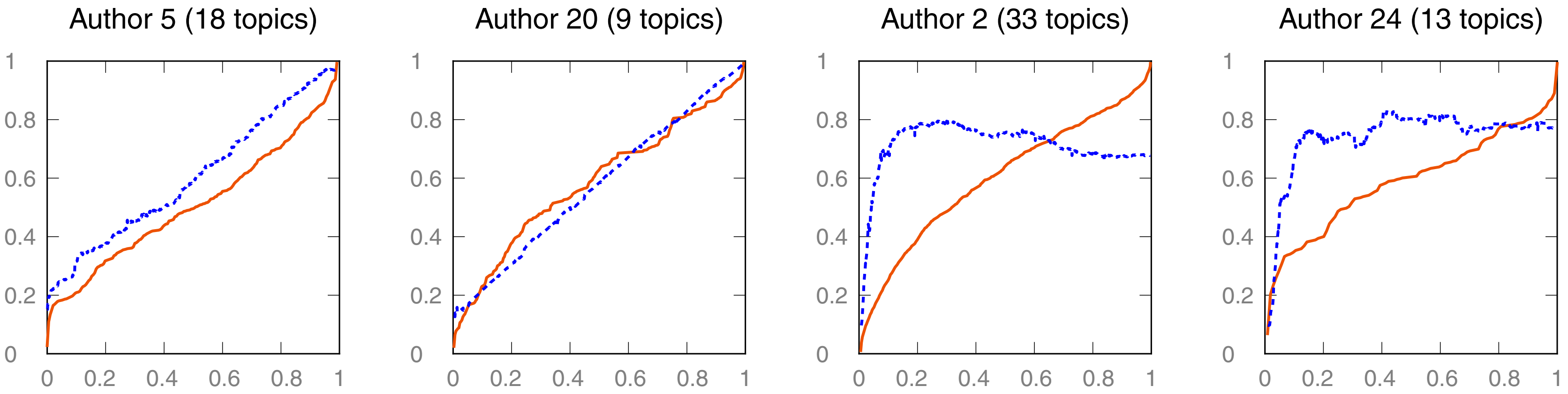


Distribution of queries over time.

- 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

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.

[<]

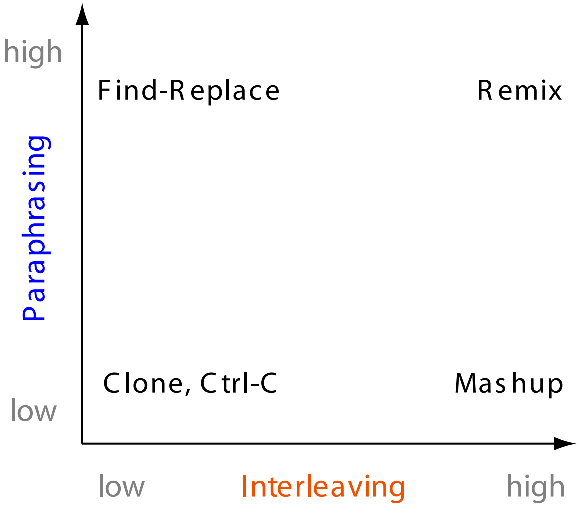
Type	**Rank**	
	Frequency	Severity
Clone Exact copy of another author’s work	1	1
Mashup A mix of material copied verbatim from several sources	2	3
Ctrl-C Significant portions of text copied from a single source	3	2
Remix Paraphrasing from several sources and making the content fit together seamlessly	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

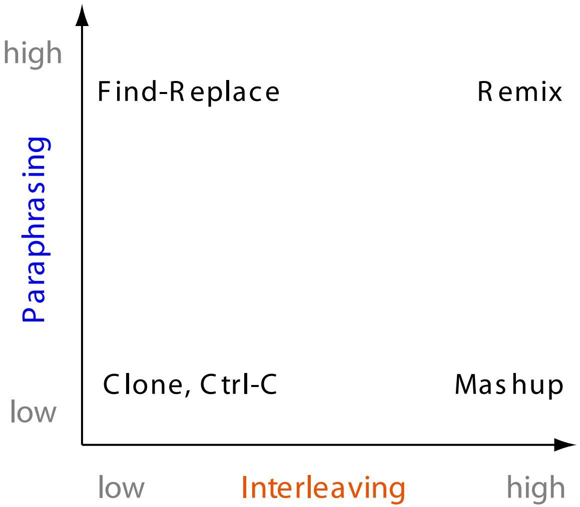
Classification Scheme for Text Reuse.

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

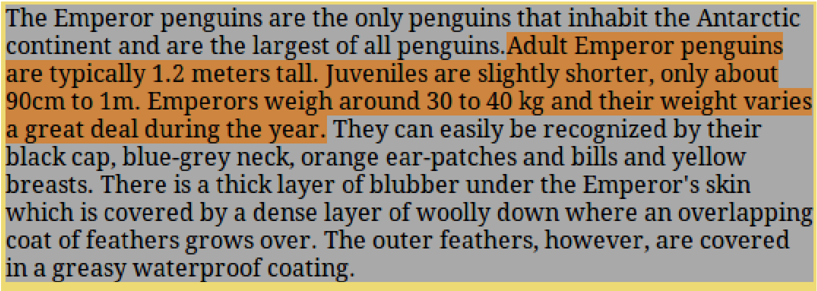
[<]

Details: Paraphrasing & Interleaving

Classification Scheme for Text Reuse



Three passages:



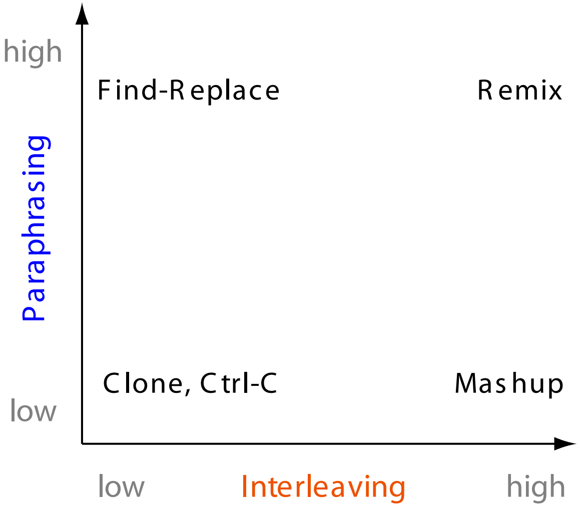
The Emperor penguins are the only penguins that inhabit the Antarctic continent and are the largest of all penguins. Adult Emperor penguins are typically 1.2 meters tall. Juveniles are slightly shorter, only about 90cm to 1m. Emperors weigh around 30 to 40 kg and their weight varies a great deal during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



high

Find-Replace

Remix

Paraphrasing

Clone, Ctrl-C

Mashup

low

low

Interleaving

high

Paraphrasing: N-Gram similarity

N = 1

Adult Emperor Penguins are ...

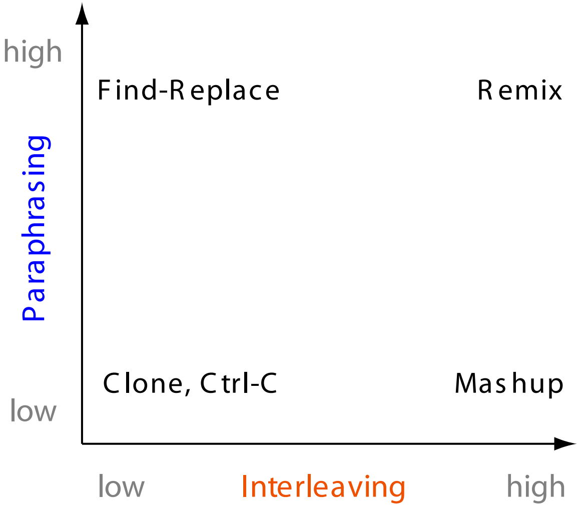
{ Adult
Emperor
Penguins
are }

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



high

Find-Replace

Remix

Paraphrasing

Clone, Ctrl-C

Mashup

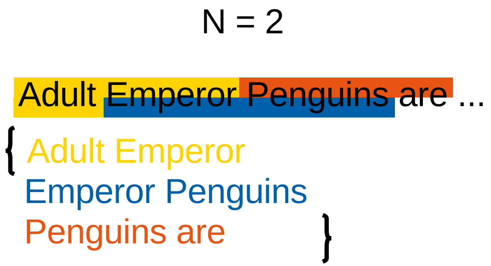
low

low

Interleaving

high

Paraphrasing: N-Gram similarity



N = 2

Adult Emperor Penguins are ...

{ Adult Emperor

Emperor Penguins

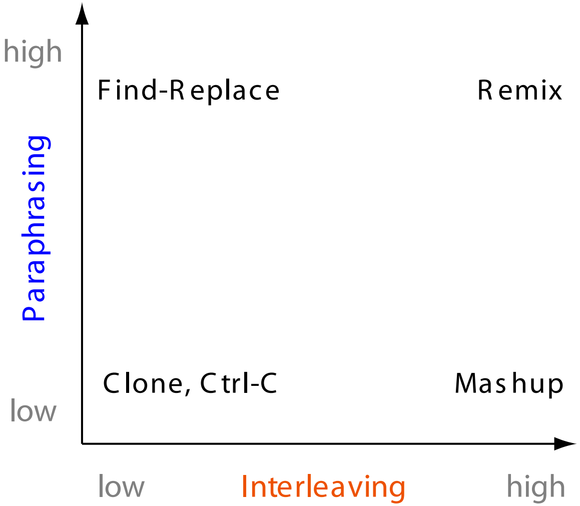
Penguins are }

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 ...

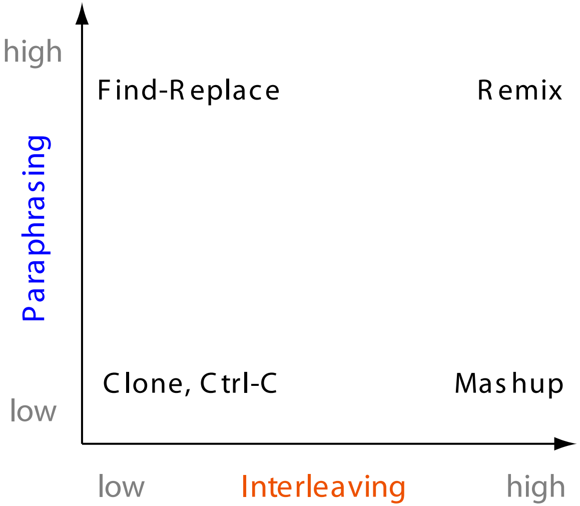
{ Adult Emperor Penguins
Emperor Penguins are }

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

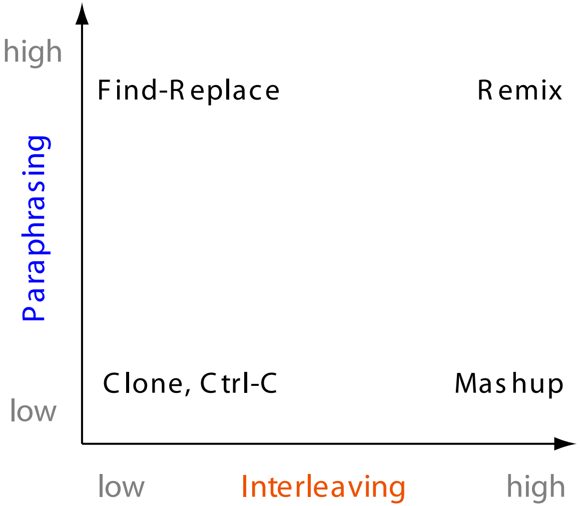
We choose $n = 5$.

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