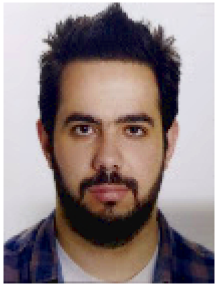
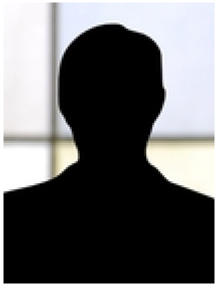
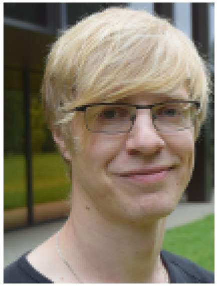
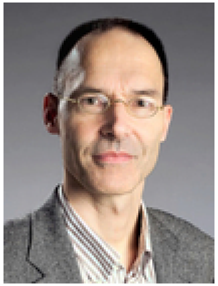
Wikipedia Text Reuse: Within and Without

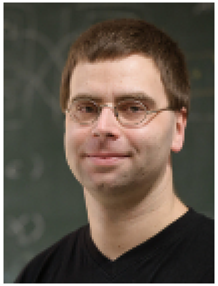

Milad Alshomary

Michael Völske


Tristan Licht


Henning Wachsmuth


Benno Stein


Matthias Hagen

Martin Potthast

PADERBORN UNIVERSITY

Bauhaus-Universität Weimar

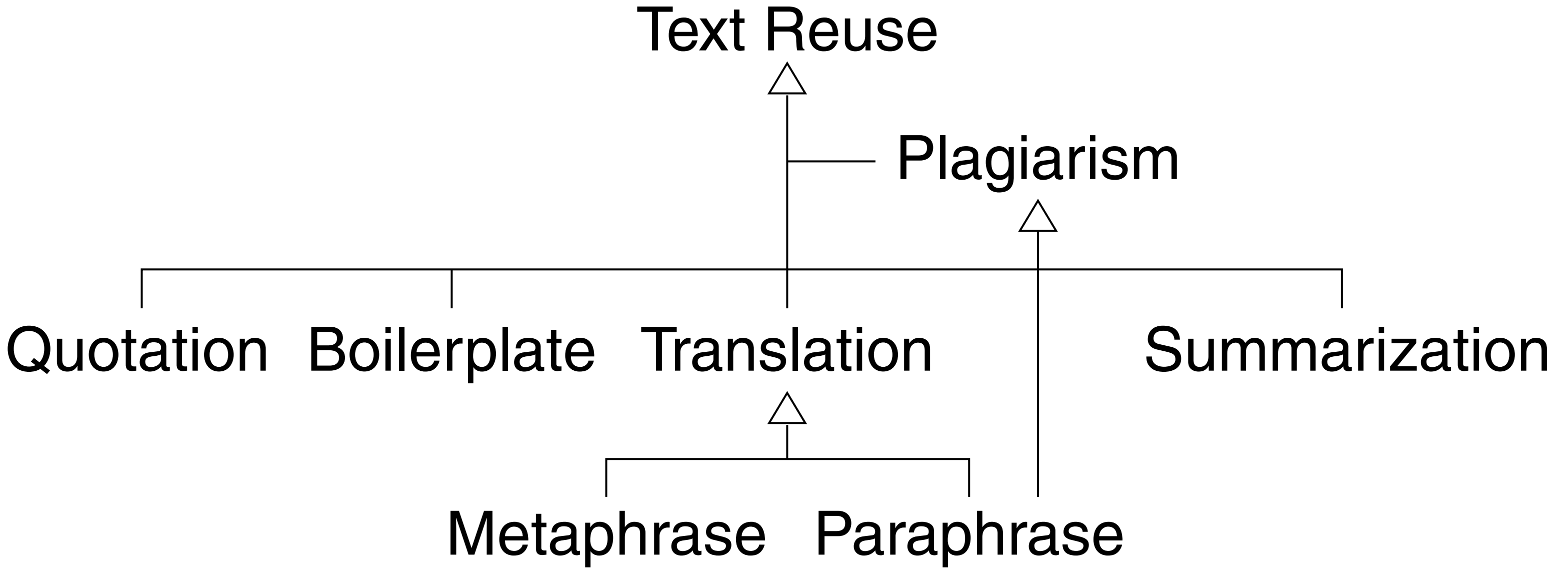
MARTIN-LUTHER-UNIVERSITÄT HALLE-WITTENBERG

UNIVERSITÄT LEIPZIG

[webis.de]

Introduction

Text Reuse



Text Reuse

Plagiarism

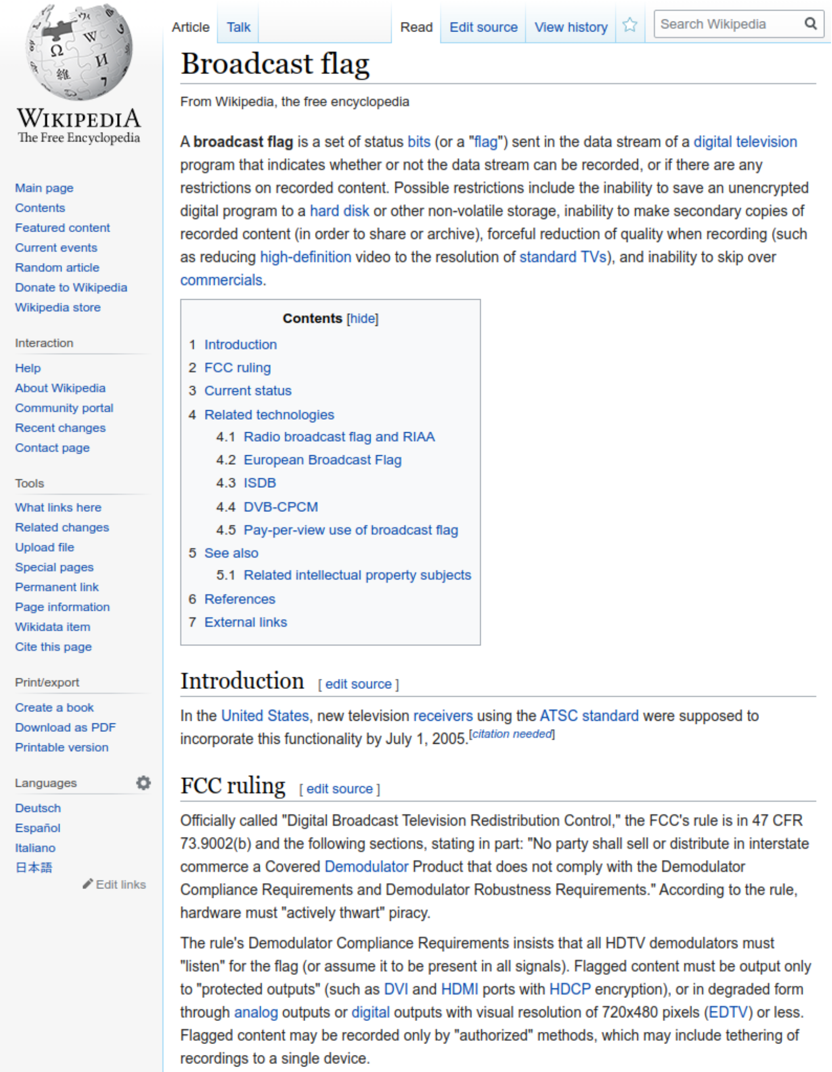
Quotation Boilerplate Translation Summarization

Metaphrase Paraphrase

- Depending on the circumstances, all kinds of text reuse can be plagiarism.
- Plagiarism may coincide with copyright infringement.
- “Boilerplate” is synonymous with “template”.
- Metaphrase and paraphrase are defined wrt. their ancient greek origins.
  Metaphrase means literal, word for word translation, paraphrase reproduction in own words.

Introduction

Wikipedia Text Reuse



Broadcast flag

From Wikipedia, the free encyclopedia

A **broadcast flag** is a set of status bits (or a "flag") sent in the data stream of a digital television program that indicates whether or not the data stream can be recorded, or if there are any restrictions on recorded content. Possible restrictions include the inability to save an unencrypted digital program to a hard disk or other non-volatile storage, inability to make secondary copies of recorded content (in order to share or archive), forceful reduction of quality when recording (such as reducing high-definition video to the resolution of standard TVs), and inability to skip over commercials.

Contents [hide]

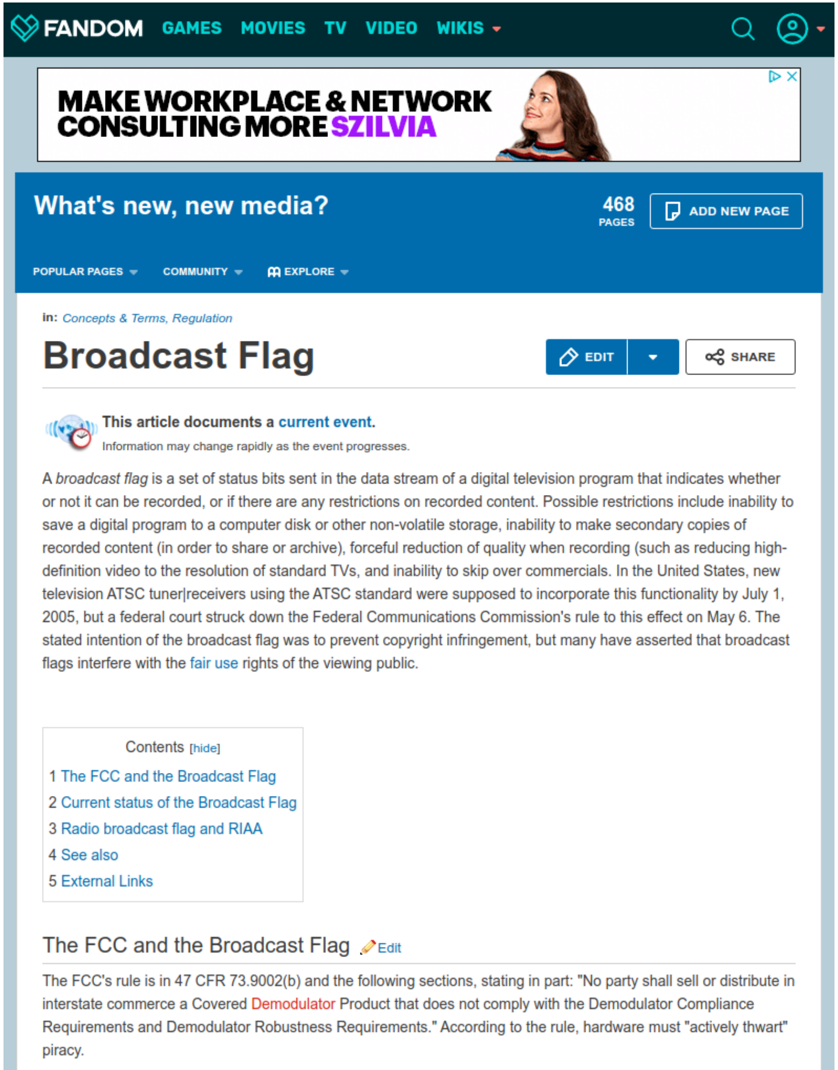
Introduction [edit source]

In the United States, new television receivers using the ATSC standard were supposed to incorporate this functionality by July 1, 2005.[citation needed]

FCC ruling [edit source]

Officially called "Digital Broadcast Television Redistribution Control," the FCC's rule is in 47 CFR 73.9002(b) and the following sections, stating in part: "No party shall sell or distribute in interstate commerce a Covered Demodulator Product that does not comply with the Demodulator Compliance Requirements and Demodulator Robustness Requirements." According to the rule, hardware must "actively thwart" piracy.

The rule's Demodulator Compliance Requirements insists that all HDTV demodulators must "listen" for the flag (or assume it to be present in all signals). Flagged content must be output only to "protected outputs" (such as DVI and HDMI ports with HDCP encryption), or in degraded form through analog outputs or digital outputs with visual resolution of 720x480 pixels (EDTV) or less. Flagged content may be recorded only by "authorized" methods, which may include tethering of recordings to a single device.



in: Concepts & Terms, Regulation

Broadcast Flag

This article documents a current event.

Information may change rapidly as the event progresses.

A *broadcast flag* is a set of status bits sent in the data stream of a digital television program that indicates whether or not it can be recorded, or if there are any restrictions on recorded content. Possible restrictions include inability to save a digital program to a computer disk or other non-volatile storage, inability to make secondary copies of recorded content (in order to share or archive), forceful reduction of quality when recording (such as reducing high-definition video to the resolution of standard TVs, and inability to skip over commercials. In the United States, new television ATSC tuner|receivers using the ATSC standard were supposed to incorporate this functionality by July 1, 2005, but a federal court struck down the Federal Communications Commission's rule to this effect on May 6. The stated intention of the broadcast flag was to prevent copyright infringement, but many have asserted that broadcast flags interfere with the fair use rights of the viewing public.

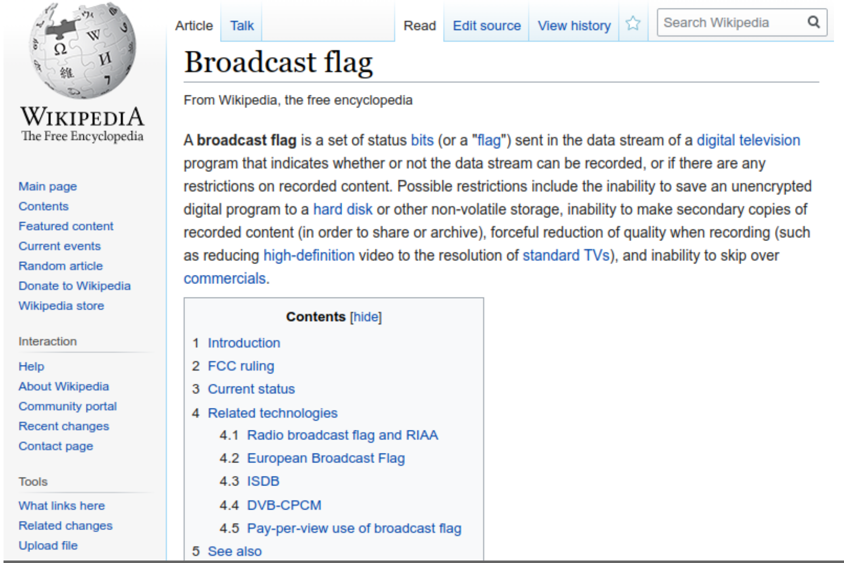
Contents [hide]

The FCC and the Broadcast Flag Edit

The FCC's rule is in 47 CFR 73.9002(b) and the following sections, stating in part: "No party shall sell or distribute in interstate commerce a Covered Demodulator Product that does not comply with the Demodulator Compliance Requirements and Demodulator Robustness Requirements." According to the rule, hardware must "actively thwart" piracy.

Introduction

Wikipedia Text Reuse



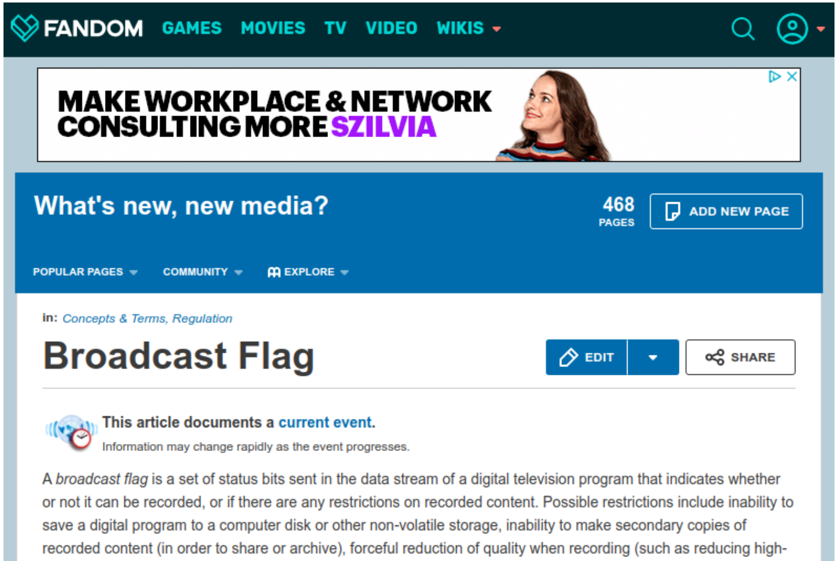
Broadcast flag

From Wikipedia, the free encyclopedia

A **broadcast flag** is a set of status bits (or a "flag") sent in the data stream of a digital television program that indicates whether or not the data stream can be recorded, or if there are any restrictions on recorded content. Possible restrictions include the inability to save an unencrypted digital program to a hard disk or other non-volatile storage, inability to make secondary copies of recorded content (in order to share or archive), forceful reduction of quality when recording (such as reducing high-definition video to the resolution of standard TVs), and inability to skip over commercials.

Contents [hide]

1 Introduction
2 FCC ruling
3 Current status
4 Related technologies
4.1 Radio broadcast flag and RIAA
4.2 European Broadcast Flag
4.3 ISDB
4.4 DVB-CPCM
4.5 Pay-per-view use of broadcast flag
5 See also



What's new, new media?

in: *Concepts & Terms*, *Regulation*

Broadcast Flag

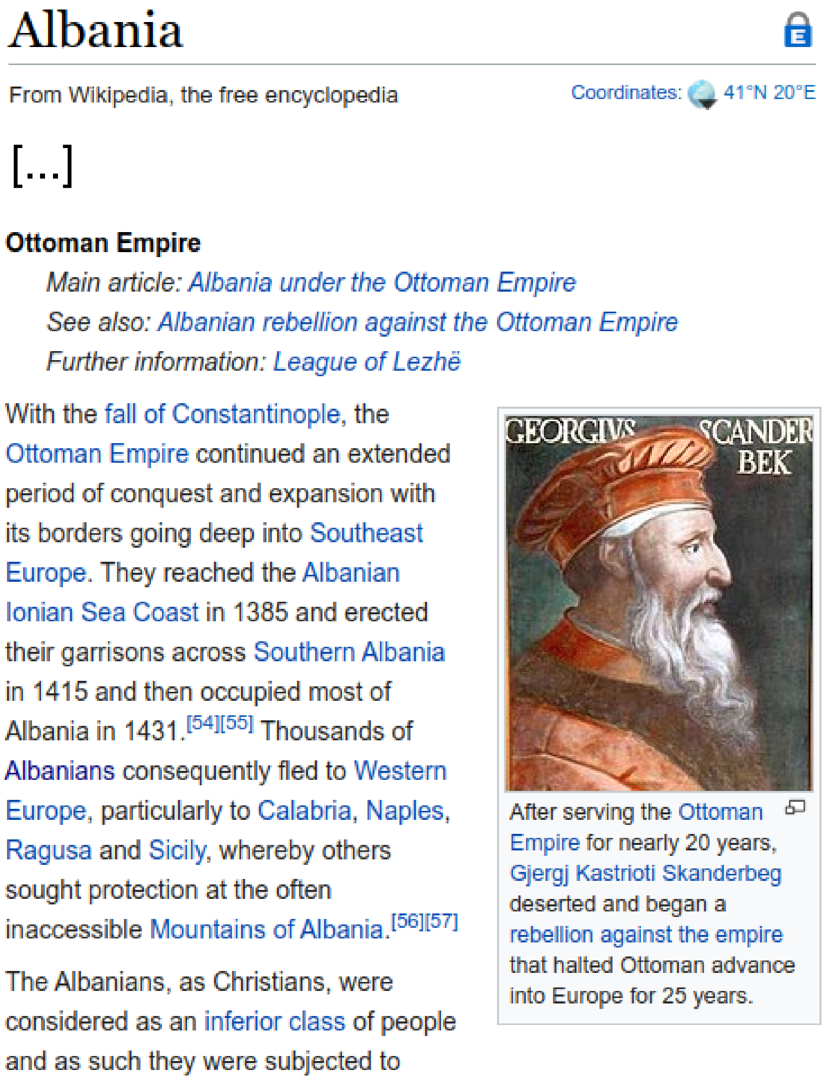
This article documents a current event.
Information may change rapidly as the event progresses.

A *broadcast flag* is a set of status bits sent in the data stream of a digital television program that indicates whether or not it can be recorded, or if there are any restrictions on recorded content. Possible restrictions include inability to save a digital program to a computer disk or other non-volatile storage, inability to make secondary copies of recorded content (in order to share or archive), forceful reduction of quality when recording (such as reducing high-

- ❑ Wikipedia articles are reused on commercial web pages.
- ❑ Reuse sometimes lacks attribution, a violation of Wikipedia's copyrights.
- ❑ Reused content is not necessarily updated along the original.
- ❑ What is the extent to which the web is comprised of Wikipedia text reuse?
- ❑ What is the added value of duplicating Wikipedia content?

Introduction

Wikipedia Text Reuse



Albania

From Wikipedia, the free encyclopedia

Coordinates: 41°N 20°E

[...]

Ottoman Empire

Main article: Albania under the Ottoman Empire

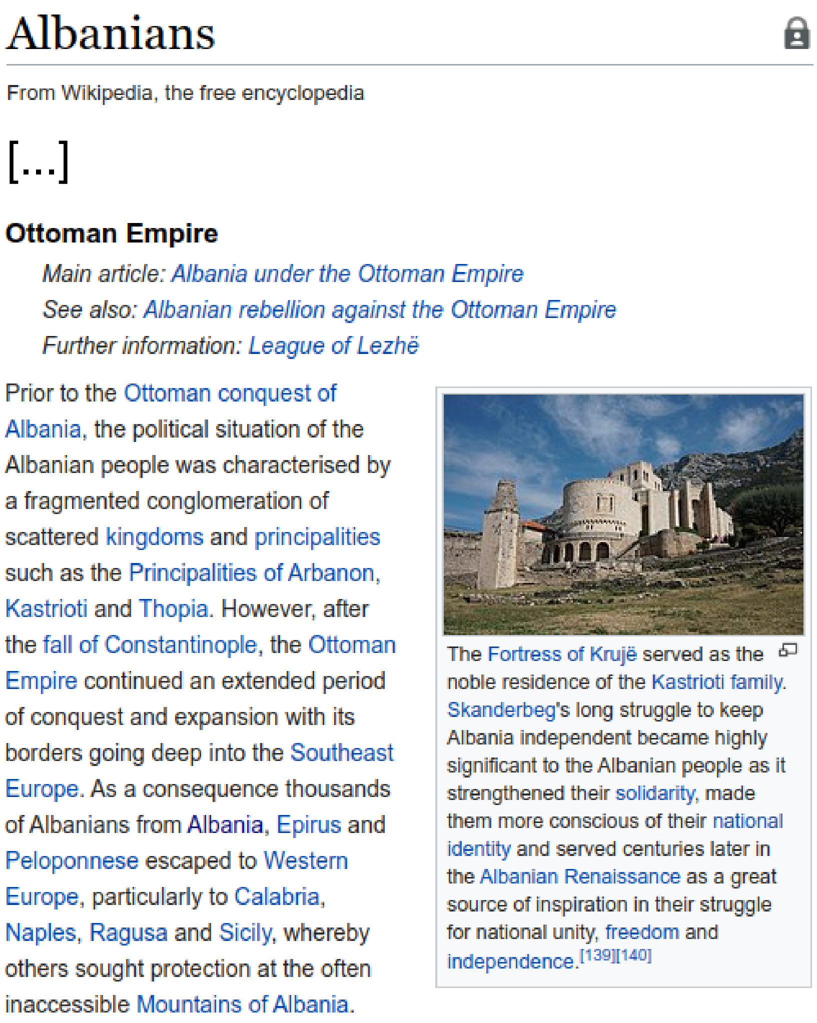
See also: Albanian rebellion against the Ottoman Empire

Further information: League of Lezhë

With the fall of Constantinople, the Ottoman Empire continued an extended period of conquest and expansion with its borders going deep into Southeast Europe. They reached the Albanian Ionian Sea Coast in 1385 and erected their garrisons across Southern Albania in 1415 and then occupied most of Albania in 1431.[54][55] Thousands of Albanians consequently fled to Western Europe, particularly to Calabria, Naples, Ragusa and Sicily, whereby others sought protection at the often inaccessible Mountains of Albania.[56][57]

After serving the Ottoman Empire for nearly 20 years, Gjergj Kastrioti Skanderbeg deserted and began a rebellion against the empire that halted Ottoman advance into Europe for 25 years.

The Albanians, as Christians, were considered as an inferior class of people and as such they were subjected to



Albanians

From Wikipedia, the free encyclopedia

[...]

Ottoman Empire

Main article: Albania under the Ottoman Empire

See also: Albanian rebellion against the Ottoman Empire

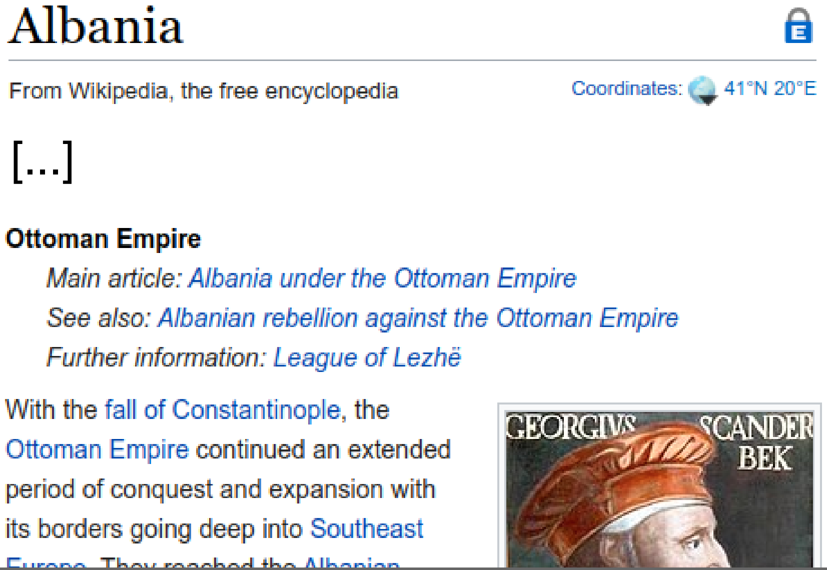
Further information: League of Lezhë

Prior to the Ottoman conquest of Albania, the political situation of the Albanian people was characterised by a fragmented conglomeration of scattered kingdoms and principalities such as the Principalities of Arbanon, Kastrioti and Thopia. However, after the fall of Constantinople, the Ottoman Empire continued an extended period of conquest and expansion with its borders going deep into the Southeast Europe. As a consequence thousands of Albanians from Albania, Epirus and Peloponnese escaped to Western Europe, particularly to Calabria, Naples, Ragusa and Sicily, whereby others sought protection at the often inaccessible Mountains of Albania.

The Fortress of Krujë served as the noble residence of the Kastrioti family. Skanderbeg's long struggle to keep Albania independent became highly significant to the Albanian people as it strengthened their solidarity, made them more conscious of their national identity and served centuries later in the Albanian Renaissance as a great source of inspiration in their struggle for national unity, freedom and independence.[139][140]

Introduction

Wikipedia Text Reuse



Albania

From Wikipedia, the free encyclopedia

Coordinates: 41°N 20°E

[...]

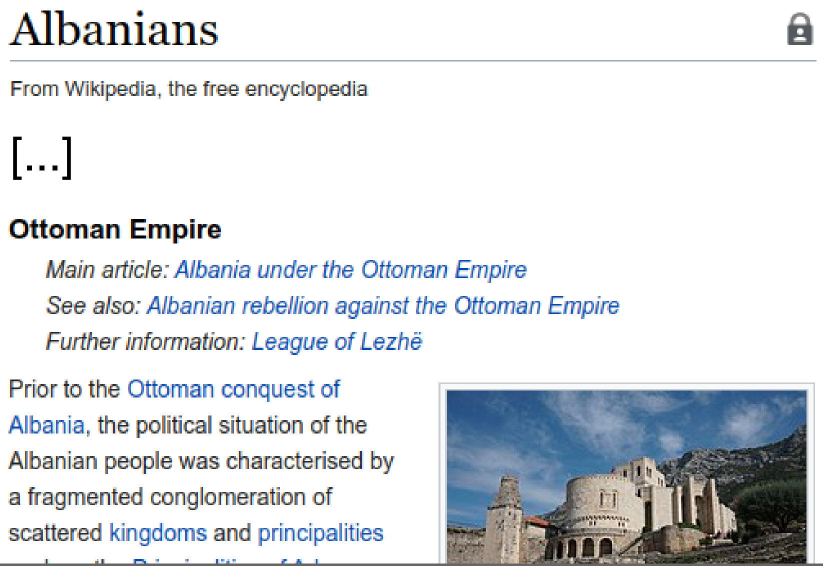
Ottoman Empire

Main article: Albania under the Ottoman Empire

See also: Albanian rebellion against the Ottoman Empire

Further information: League of Lezhë

With the fall of Constantinople, the Ottoman Empire continued an extended period of conquest and expansion with its borders going deep into Southeast Europe. They reached the Albanian



Albanians

From Wikipedia, the free encyclopedia

[...]

Ottoman Empire

Main article: Albania under the Ottoman Empire

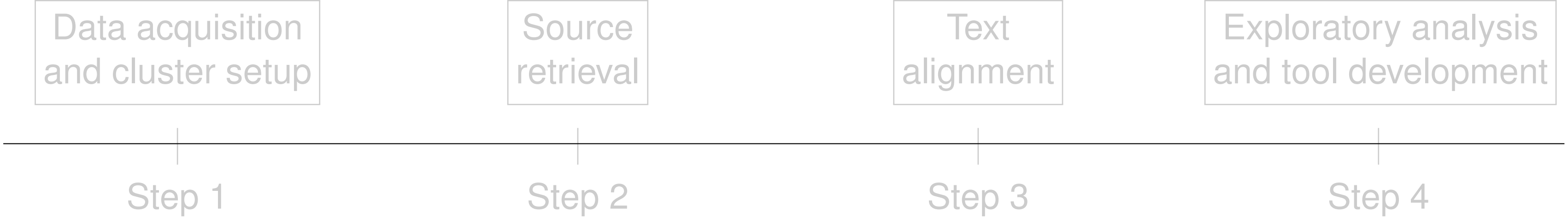
See also: Albanian rebellion against the Ottoman Empire

Further information: League of Lezhë

Prior to the Ottoman conquest of Albania, the political situation of the Albanian people was characterised by a fragmented conglomeration of scattered kingdoms and principalities

- ❑ Wikipedia articles reuse text from other Wikipedia articles.
- ❑ Different articles progress independently, giving rise to inconsistency.
- ❑ There are no tools to support text reuse; manual reuse is not tracked.
- ❑ What is the extent to which Wikipedia reuses itself?
- ❑ How can reuse and repair be supported?

Text Reuse Detection



Data acquisition and cluster setup

Source retrieval

Text alignment

Exploratory analysis and tool development

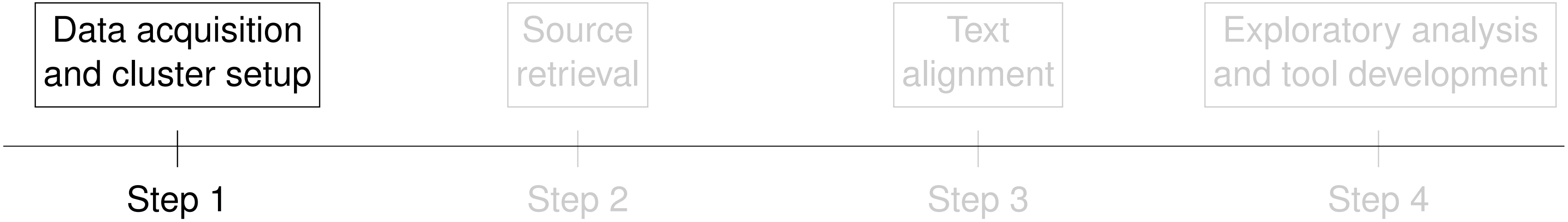
Step 1

Step 2

Step 3

Step 4

Text Reuse Detection


Data acquisition and cluster setup

Source retrieval

Text alignment

Exploratory analysis and tool development

Step 1

Step 2

Step 3

Step 4

Wikipedia: [dumps.wikimedia.org]

- English dump from May 2016
- 4.2m articles
- 11.4m paragraphs

Common Crawl: [commoncrawl.org]

- Release from April 2017
- 10% sample
- 1.4m websites
- 591m web pages
- 187m paragraphs

Text Reuse Detection

Data acquisition and cluster setup

Source retrieval

Text alignment

Exploratory analysis and tool development

Step 1

Step 2

Step 3

Step 4

	alphaweb [2009]	betaweb [2015]	gammaweb [2016]	deltaweb [2018]
Nodes	44	135	3	78
Disk (PB)	0.2	4.1	0.02	12
Cores	176 (3.2 TFLOPs)	1,740 (67.4 TFLOPs)	96 + 61,440 (206.7 TFLOPs)	1,248 (119.8 TFLOPs)
RAM (TB)	0.8	28	4.8	10

Text Reuse Detection

Data acquisition and cluster setup

Source retrieval

Text alignment

Exploratory analysis and tool development

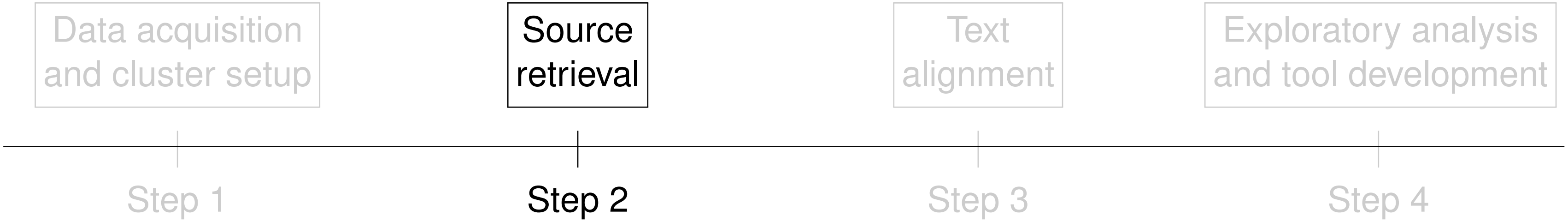
Step 1

Step 2

Step 3

Step 4

Text Reuse Detection


Data acquisition and cluster setup
Source retrieval
Text alignment
Exploratory analysis and tool development
Step 1
Step 2
Step 3
Step 4

Ranking of documents with respect to their likelihood of being source for text reuse.

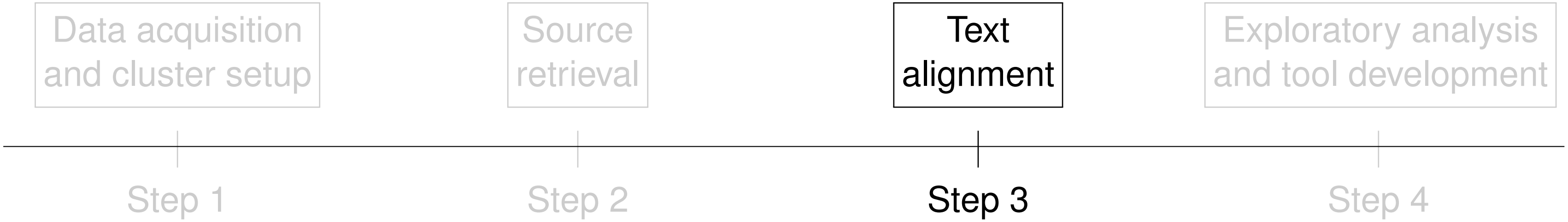
Scoring function ρ for two documents d_1 and d_2:

$$\underbrace{\exists c_i \in d_1, c_j \in d_2\colon \quad h(c_i) \cap h(c_j) \neq \emptyset}_{\text{Search pruning}} \quad \rightarrow \quad \rho(d_1, d_2) = \max_{\substack{c_i \in d_1 \\ c_j \in d_2}}(\varphi(c_i, c_j)),$$

where

- c is a paragraph-length chunk of a document d
- h is a locality-sensitive hash function producing a set of hash values
- φ the *tf* · *idf*-weighted cosine similarity

Text Reuse Detection



Data acquisition and cluster setup

Source retrieval

Text alignment

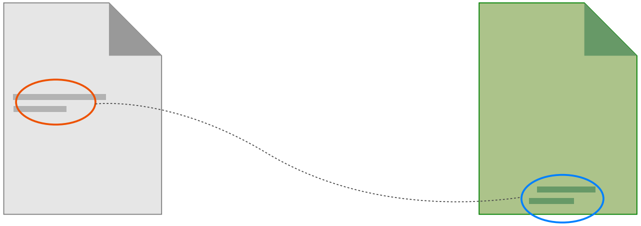
Exploratory analysis and tool development

Step 1

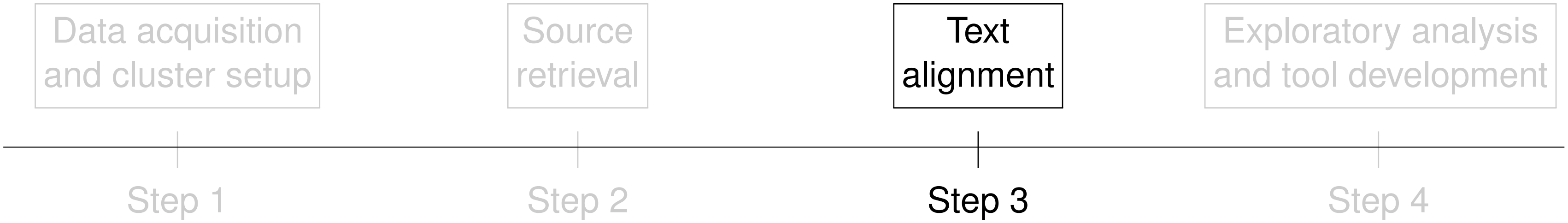
Step 2

Step 3

Step 4



Text Reuse Detection



Data acquisition and cluster setup

Source retrieval

Text alignment

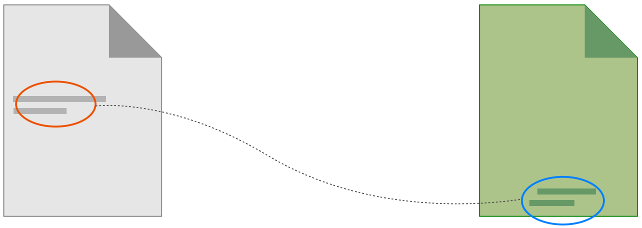
Exploratory analysis and tool development

Step 1

Step 2

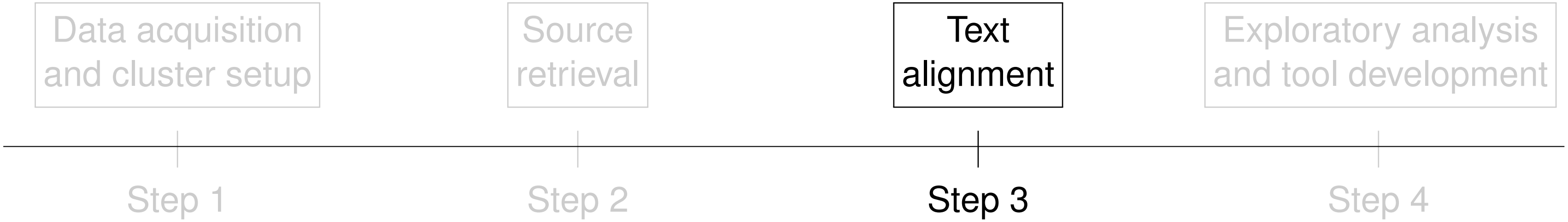
Step 3

Step 4



	Technology	What can be detected
basic	MD5 hashing	Identity analysis for paragraphs
	Hashed breakpoint chunking	Synchronized identity analysis for paragraphs
	Locality-sensitive hashing	Tolerant similarity analysis for paragraphs
complex	Dot plotting	Sequence analysis of word n-grams

Text Reuse Detection



Data acquisition and cluster setup

Source retrieval

Text alignment

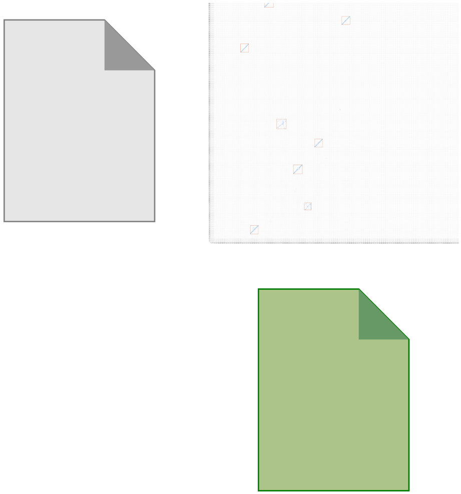
Exploratory analysis and tool development

Step 1

Step 2

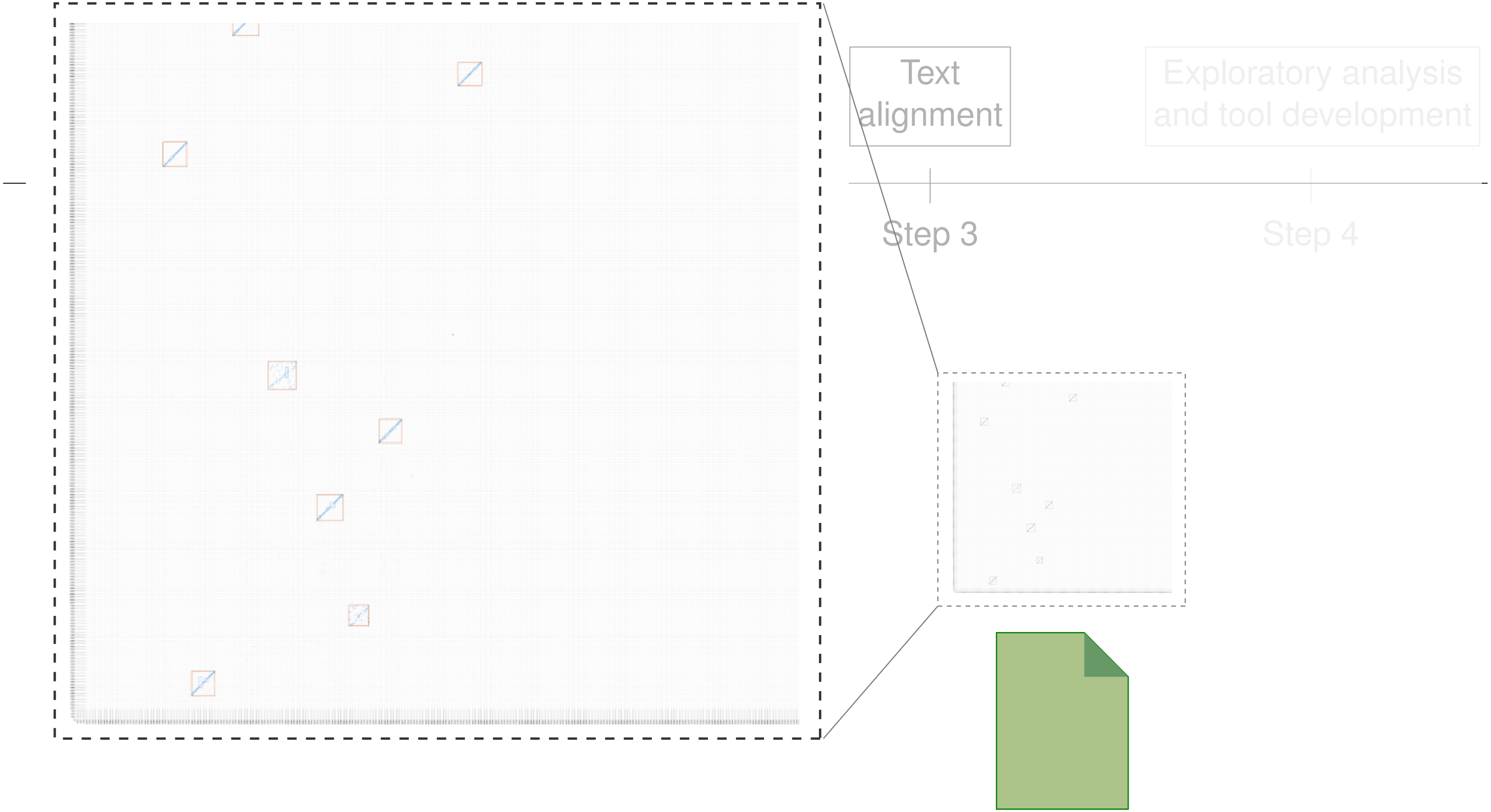
Step 3

Step 4



Geometric sequence analysis of all word 3-grams of two interesting documents.

Text Reuse Detection



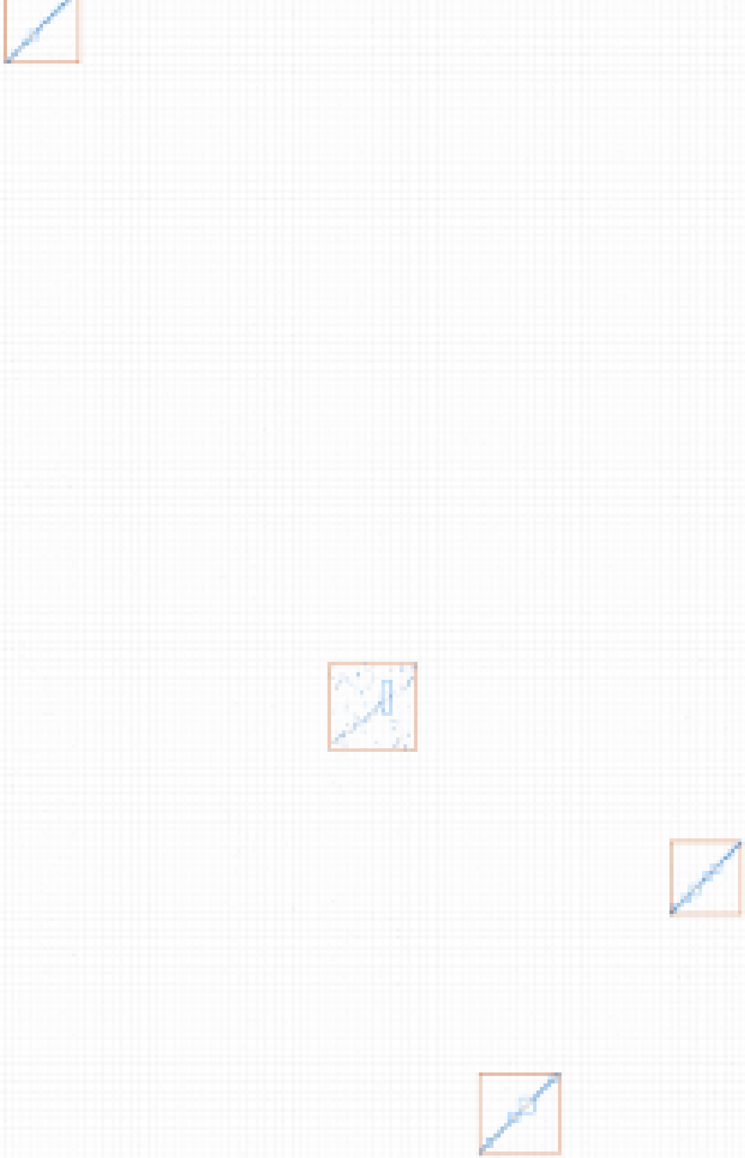
Text alignment

Exploratory analysis and tool development

Step 3

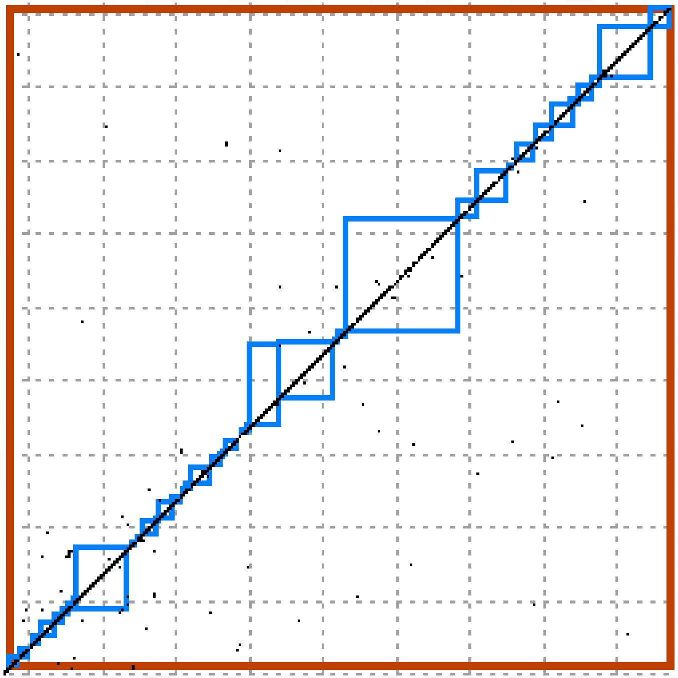
Step 4

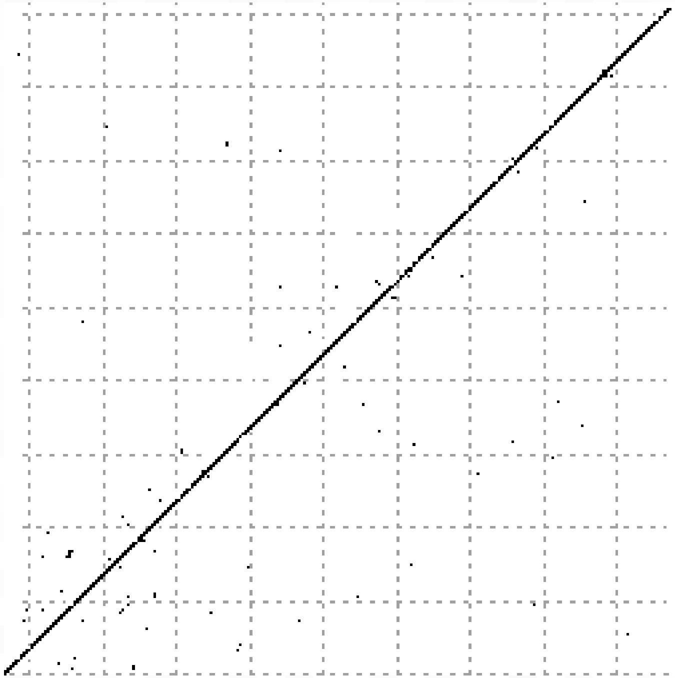
Geometric sequence analysis of all word 3-grams of two interesting documents.



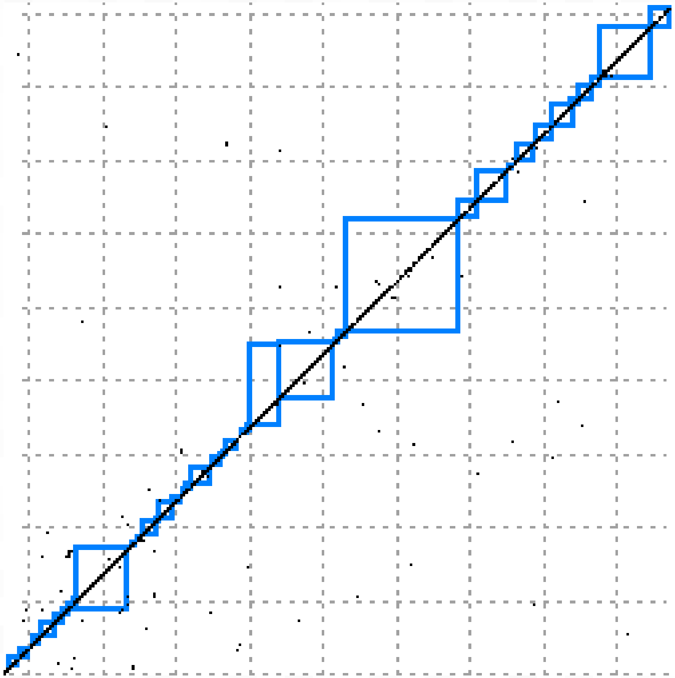
(two chapters)

(two pages)

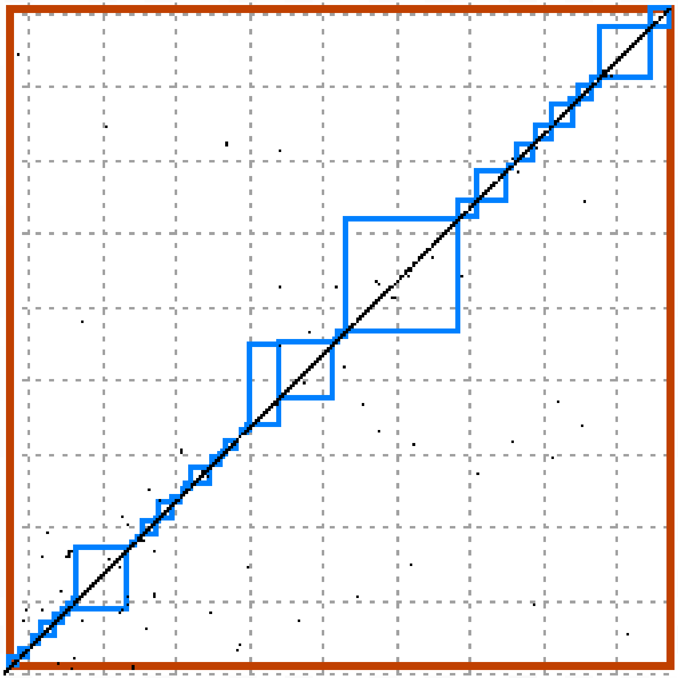




Level 1 (black): each dot indicates a common word 3-gram, i.e. a hash collision.

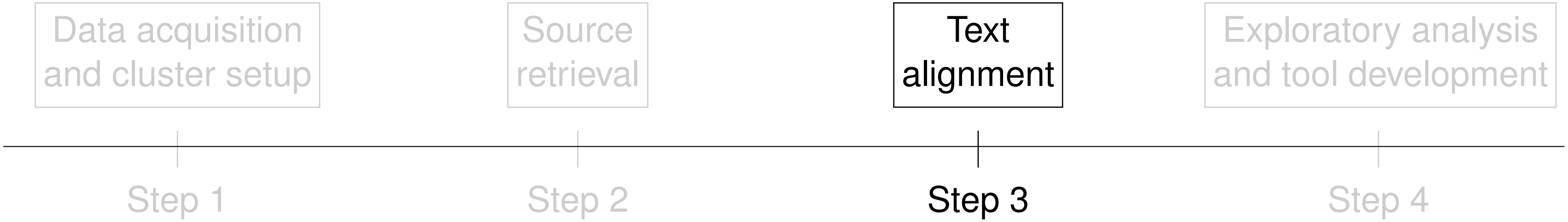


Level 2 (blue): neighbored common 3-grams are heuristically comprised.



Level 3 (red): blue groups are merged by a cluster analysis.

Text Reuse Detection



Data acquisition and cluster setup

Source retrieval

Text alignment

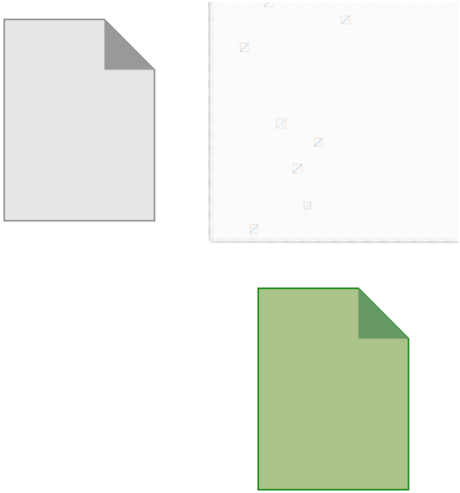
Exploratory analysis and tool development

Step 1

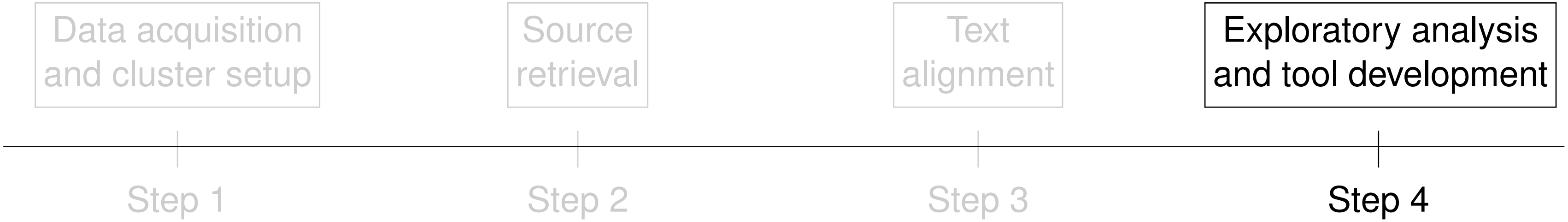
Step 2

Step 3

Step 4



Text Reuse Detection



Data acquisition and cluster setup

Source retrieval

Text alignment

Exploratory analysis and tool development

Step 1

Step 2

Step 3

Step 4

- ❑ Webis Wikipedia Text Reuse Corpus
- ❑ Download: [webis.de/data.html]
- ❑ Processing time: 2 months

➜ Pilot analysis

➜ Tool development

➜ Visual analytics

Reuse	**Within**	**Without**
Cases	110m	1.6m
Documents with Reuse Cases		
Articles	360,000	1.0m
Pages	–	15,000
Words in Reuse Cases		
Min.	17	23
Avg.	78	252
Max.	6200	1960

Pilot Analysis

Wikipedia Text Reuse on the Web

- 4898 out of 1.4m sampled sites reuse at least once from Wikipedia

  Presumption: Wikipedia's editors successfully avoid reuse from third parties altogether.
  Top three: wikia.com (563 pages), rediff.com (55), un.org (28).

- 94% of reusing pages violate Wikipedia's copyrights.

  The term "Wikipedia" does not occur on the page.

- Nearly all pages exclusively reuse text.

  Redundant, affecting Wikipedia's ranking (?), quickly outdated, but still way better than making stuff up at random (hello GPT-2).

Pilot Analysis

Wikipedia Text Reuse on the Web

- ❑ 4898 out of 1.4m sampled sites reuse at least once from Wikipedia

  Presumption: Wikipedia's editors successfully avoid reuse from third parties altogether.
  Top three: wikia.com (563 pages), rediff.com (55), un.org (28).

- ❑ 94% of reusing pages violate Wikipedia's copyrights.

  The term "Wikipedia" does not occur on the page.

- ❑ Nearly all pages exclusively reuse text.

  Redundant, affecting Wikipedia's ranking (?), quickly outdated, but still way better than making stuff up at random (hello GPT-2).

- ❑ Nearly all pages show display ads.

- ❑ A conservative estimation of monthly ad revenue:

  – Simplifying assumptions: 1 ad per page at revenue per mille of 1.4 USD

  – Monthly views est. at 10% of the monthly page views of corresponding Wikipedia article

  – About 45,000 USD monthly ad revenue is generated by our sample

- ❑ Projection to the entire web:

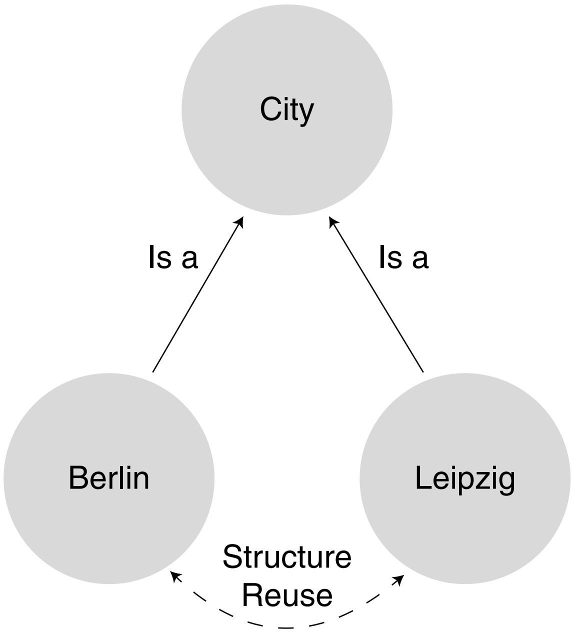
  – 600,000 reusing sites out of 180m (netcraft.com) ➜ 5.5m USD monthly ad revenue

  – 72% of Wikipedia's annual fundraising in the fiscal year 2016-2017.

Pilot Analysis

Wikipedia Text Reuse on Wikipedia

Structure reuse:

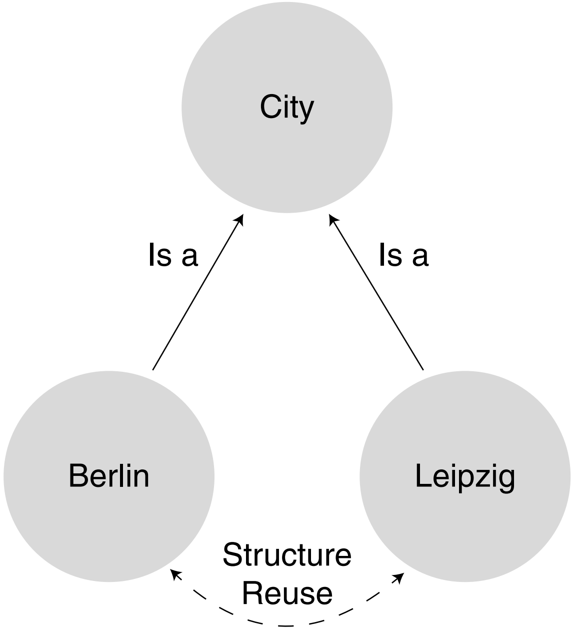


- ❑ Template-based articles
- ❑ Estimate: 87% (95.5m) of cases
- ❑ Classification difficult

Pilot Analysis

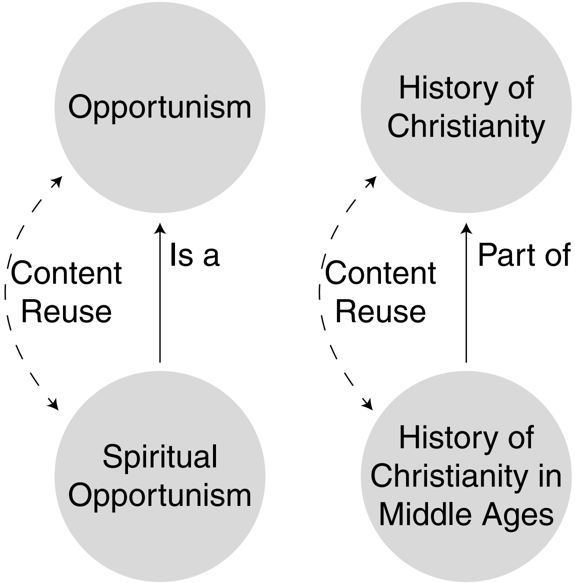
Wikipedia Text Reuse on Wikipedia

Structure reuse:



- ❑ Template-based articles
- ❑ Estimate: 87% (95.5m) of cases
- ❑ Classification difficult

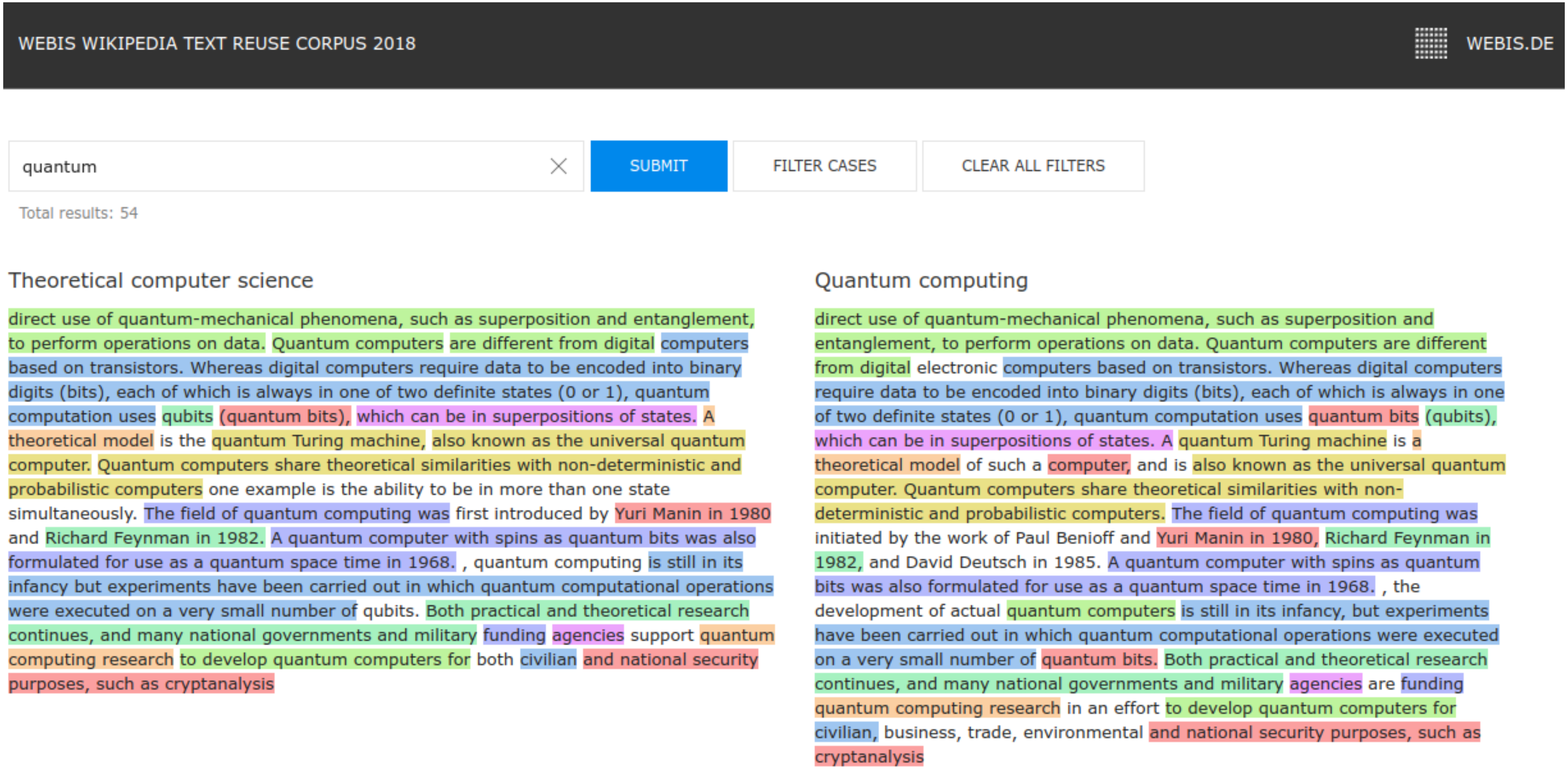
Content reuse:



- ❑ Concept hierarchies
- ❑ Shared concepts
- ❑ Occasionally part of structure reuse

Exploratory Analysis Tools

Search Engine



Total results: 54

Theoretical computer science

direct use of quantum-mechanical phenomena, such as superposition and entanglement, to perform operations on data. Quantum computers are different from digital computers based on transistors. Whereas digital computers require data to be encoded into binary digits (bits), each of which is always in one of two definite states (0 or 1), quantum computation uses qubits (quantum bits), which can be in superpositions of states. A theoretical model is the quantum Turing machine, also known as the universal quantum computer. Quantum computers share theoretical similarities with non-deterministic and probabilistic computers one example is the ability to be in more than one state simultaneously. The field of quantum computing was first introduced by Yuri Manin in 1980 and Richard Feynman in 1982. A quantum computer with spins as quantum bits was also formulated for use as a quantum space time in 1968. , quantum computing is still in its infancy but experiments have been carried out in which quantum computational operations were executed on a very small number of qubits. Both practical and theoretical research continues, and many national governments and military funding agencies support quantum computing research to develop quantum computers for both civilian and national security purposes, such as cryptanalysis

Quantum computing

direct use of quantum-mechanical phenomena, such as superposition and entanglement, to perform operations on data. Quantum computers are different from digital electronic computers based on transistors. Whereas digital computers require data to be encoded into binary digits (bits), each of which is always in one of two definite states (0 or 1), quantum computation uses quantum bits (qubits), which can be in superpositions of states. A quantum Turing machine is a theoretical model of such a computer, and is also known as the universal quantum computer. Quantum computers share theoretical similarities with non-deterministic and probabilistic computers. The field of quantum computing was initiated by the work of Paul Benioff and Yuri Manin in 1980, Richard Feynman in 1982, and David Deutsch in 1985. A quantum computer with spins as quantum bits was also formulated for use as a quantum space time in 1968. , the development of actual quantum computers is still in its infancy, but experiments have been carried out in which quantum computational operations were executed on a very small number of quantum bits. Both practical and theoretical research continues, and many national governments and military agencies are funding quantum computing research in an effort to develop quantum computers for civilian, business, trade, environmental and national security purposes, such as cryptanalysis

[demo.webis.de/wikipedia-text-reuse]

Exploratory Analysis Tools

Visual Analytics

Visualization of pairwise Wikipedia article similarities.

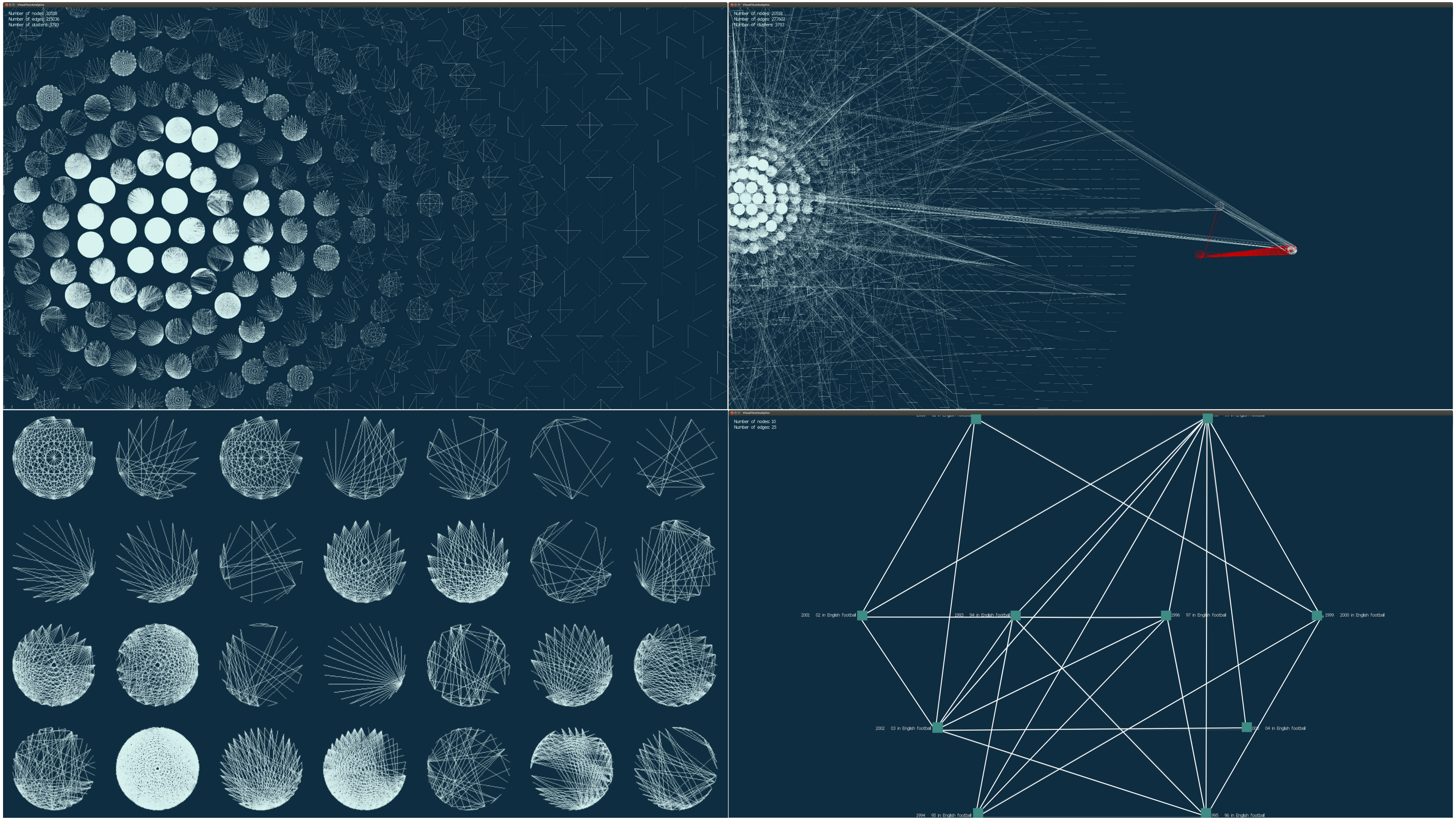
[Riehmann et al., EuroVis 2016]

Patrick
Riehmann

Bernd
Fröhlich

VR Group

Bauhaus-
Universität
Weimar



VisualTextAnalytics

Number of nodes: 20518
Number of edges: 215036
Number of clusters: 3793

VisualTextAnalytics

Number of nodes: 20518
Number of edges: 277602
Number of clusters: 3793

VisualTextAnalytics

Number of nodes: 10
Number of edges: 25

2000 01 in English football
99 in English football
2001 02 in English football
1993 94 in English football
1996 97 in English football
1999 2000 in English football
2002 03 in English football
2003 04 in English football
1994 95 in English football
1995 96 in English football

Conclusion

Take-away Messages

- Text reuse is second nature to Wikipedia.
- Content reuse should be actively unified.
- Tool support for reuse within MediaWiki is needed.
- Reuse outside Wikipedia can be a risk as well as an opportunity.
  Taraborelli's "Paradox of Reuse" vs. opposing randomly generated text as per GPT-2.
- Reuse is another indicator of Wikipedia's influence on the web at large.

Conclusion

Take-away Messages

- ❑ Text reuse is second nature to Wikipedia.
- ❑ Content reuse should be actively unified.
- ❑ Tool support for reuse within MediaWiki is needed.
- ❑ Reuse outside Wikipedia can be a risk as well as an opportunity.
  Taraborelli's "Paradox of Reuse" vs. opposing randomly generated text as per GPT-2.
- ❑ Reuse is another indicator of Wikipedia's influence on the web at large.

Future Work

- ❑ Categorizing reuse: content vs. structure reuse.
- ❑ Article template induction.
- ❑ Further scaling up to the entire Common Crawl.
- ❑ Visual analytics tools to explore reuse in context.

Conclusion

Take-away Messages

- Text reuse is second nature to Wikipedia.
- Content reuse should be actively unified.
- Tool support for reuse within MediaWiki is needed.
- Reuse outside Wikipedia can be a risk as well as an opportunity.
  Taraborelli's "Paradox of Reuse" vs. opposing randomly generated text as per GPT-2.
- Reuse is another indicator of Wikipedia's influence on the web at large.

Future Work

- Categorizing reuse: content vs. structure reuse.
- Article template induction.
- Further scaling up to the entire Common Crawl.
- Visual analytics tools to explore reuse in context.

Resources

- Paper, Code, Data, Demo: [webis.de]

Conclusion

Take-away Messages

- ❑ Text reuse is second nature to Wikipedia.
- ❑ Content reuse should be actively unified.
- ❑ Tool support for reuse within MediaWiki is needed.
- ❑ Reuse outside Wikipedia can be a risk as well as an opportunity.
  Taraborelli's "Paradox of Reuse" vs. opposing randomly generated text as per GPT-2.
- ❑ Reuse is another indicator of Wikipedia's influence on the web at large.

Future Work

- ❑ Categorizing reuse: content vs. structure reuse.
- ❑ Article template induction.
- ❑ Further scaling up to the entire Common Crawl.
- ❑ Visual analytics tools to explore reuse in context.

Resources

- ❑ Paper, Code, Data, Demo: [webis.de]

Thank you!