Harnessing Web Archives to Tackle Selected Societal Challenges

The Oral Exam of

Johannes Kiesel

To Obtain the Academic Degree of

Dr. rer. nat.

Web Technology & Information Systems Group
Bauhaus-Universität Weimar

www.uni-weimar.de www.webis.de

Harnessing Web Archives to Tackle Selected Societal Challenges

Societal challenges

Issues that concern most if not all members of a society, either now or in a likely future.

Well-known challenges:*

- ❑ Critical assessment of information
- ❑ Protection of the environment
- ❑ Preservation of culture
- ❑ Ensuring public health
- ❑ Security and privacy

*Taken from European Commission (Horizon 2020), World Economic Forum, Gesellschaft für Informatik

Harnessing Web Archives to Tackle Selected Societal Challenges

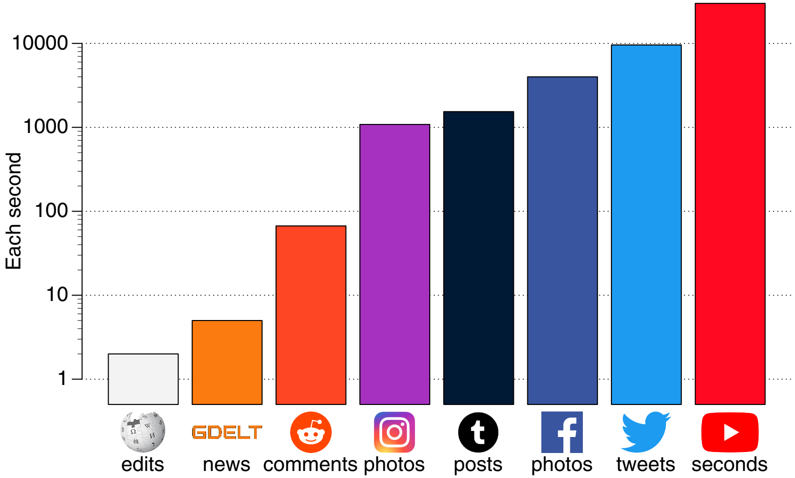
Societal challenges

Issues that concern most if not all members of a society, either now or in a likely future.

Well-known challenges:*

- ❑ Critical assessment of information
- ❑ Protection of the environment
- ❑ Preservation of culture
- ❑ Ensuring public health
- ❑ Security and privacy

Society → Web



Source: DOMO, Reddit, GDELT, Wikipedia

Web archives

- ❑ Allow for large-scale analyses
- ❑ Allow to trace changes
- ❑ Allow to replicate analyses

*Taken from European Commission (Horizon 2020), World Economic Forum, Gesellschaft für Informatik

Harnessing Web Archives to Tackle Selected Societal Challenges

Main contributions

1. Preservation of digital culture

- ❑ 10K pages high-fidelity archive (FAIRest dataset award)
- ❑ Reproduction assessment task
- ❑ 9K pages segmentation dataset
- ❑ Segmentation evaluation measures

2. Critical assessment of information

- ❑ Revert-based vandalism detection
- ❑ 30K edits Wiki vandalism dataset
- ❑ 1M hyperpartisan news dataset
- ❑ Style-based polarity detection
- ❑ Hyperpartisan news challenge (SemEval, 42 teams)

3. Online security and privacy

- ❑ 3B web sentences dataset
- ❑ Position-dependent language model
- ❑ Security estimate: mnemonic passwords
- ❑ Personal archiving tool

Tailored web archiving technology (Webis Web Archiver)

Harnessing Web Archives to Tackle Selected Societal Challenges

Main contributions

1. Preservation of digital culture

- ❑ 10K pages high-fidelity archive (FAIRest dataset award)
- ❑ Reproduction assessment task
- ❑ 9K pages segmentation dataset
- ❑ Segmentation evaluation measures

2. Critical assessment of information

- ❑ Revert-based vandalism detection
- ❑ 30K edits Wiki vandalism dataset
- ❑ 1M hyperpartisan news dataset
- ❑ Style-based polarity detection
- ❑ Hyperpartisan news challenge (SemEval, 42 teams)

3. Online security and privacy

- ❑ 3B web sentences dataset
- ❑ Position-dependent language model
- ❑ Security estimate: mnemonic passwords
- ❑ Personal archiving tool

Tailored web archiving technology (Webis Web Archiver)

→ New tasks

Harnessing Web Archives to Tackle Selected Societal Challenges

Main contributions

1. Preservation of digital culture

- ❑ 10K pages high-fidelity archive (FAIRest dataset award)
- ❑ Reproduction assessment task
- ❑ 9K pages segmentation dataset
- ❑ Segmentation evaluation measures

2. Critical assessment of information

- ❑ Revert-based vandalism detection
- ❑ 30K edits Wiki vandalism dataset
- ❑ 1M hyperpartisan news dataset
- ❑ Style-based polarity detection
- ❑ Hyperpartisan news challenge (SemEval, 42 teams)

3. Online security and privacy

- ❑ 3B web sentences dataset
- ❑ Position-dependent language model
- ❑ Security estimate: mnemonic passwords
- ❑ Personal archiving tool

Tailored web archiving technology (Webis Web Archiver)

→ New tasks → New or improved algorithms

Harnessing Web Archives to Tackle Selected Societal Challenges

Main contributions

1. Preservation of digital culture

- ❑ 10K pages high-fidelity archive (FAIRest dataset award)
- ❑ Reproduction assessment task
- ❑ 9K pages segmentation dataset
- ❑ Segmentation evaluation measures

2. Critical assessment of information

- ❑ Revert-based vandalism detection
- ❑ 30K edits Wiki vandalism dataset
- ❑ 1M hyperpartisan news dataset
- ❑ Style-based polarity detection
- ❑ Hyperpartisan news challenge (SemEval, 42 teams)

3. Online security and privacy

- ❑ 3B web sentences dataset
- ❑ Position-dependent language model
- ❑ Security estimate: mnemonic passwords
- ❑ Personal archiving tool

Tailored web archiving technology (Webis Web Archiver)

→ New tasks → New or improved algorithms → More adequate evaluation measures

Harnessing Web Archives to Tackle Selected Societal Challenges

Main contributions

1. Preservation of digital culture

- ❑ 10K pages high-fidelity archive (FAIRest dataset award)
- ❑ Reproduction assessment task
- ❑ 9K pages segmentation dataset
- ❑ Segmentation evaluation measures

2. Critical assessment of information

- ❑ Revert-based vandalism detection
- ❑ 30K edits Wiki vandalism dataset
- ❑ 1M hyperpartisan news dataset
- ❑ Style-based polarity detection
- ❑ Hyperpartisan news challenge (SemEval, 42 teams)

3. Online security and privacy

- ❑ 3B web sentences dataset
- ❑ Position-dependent language model
- ❑ Security estimate: mnemonic passwords
- ❑ Personal archiving tool

Tailored web archiving technology (Webis Web Archiver)

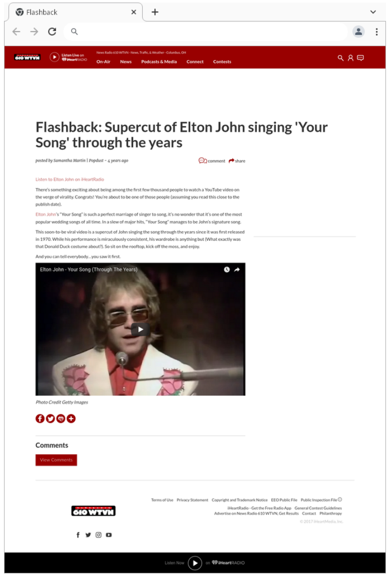
→ New tasks → New or improved algorithms → More adequate evaluation measures → Larger and more accurate datasets

Challenge 1
Preservation of Digital Culture

Web Page Segmentation

(highlighting reproducibility)

Web Page Segmentation


Flashback
610 WTVN
Listen Live on iHeartRADIO
News Radio 610 WTVN - News, Traffic, & Weather - Columbus, OH
On-Air
News
Podcasts & Media
Connect
Contests
comment
share
Elton John - Your Song (Through The Years)
View Comments
Listen Now on iHeartRADIO

Flashback: Supercut of Elton John singing 'Your Song' through the years

posted by Samantha Martin | Popdust - 4 years ago

Listen to Elton John on iHeartRadio

There's something exciting about being among the first few thousand people to watch a YouTube video on the verge of virality. Congrats! You're about to be one of those people (assuming you read this close to the publish date).

Elton John's "Your Song" is such a perfect marriage of singer to song, it's no wonder that it's one of the most popular wedding songs of all time. In a slew of major hits, "Your Song" manages to be John's signature song.

This soon-to-be viral video is a supercut of John singing the song through the years since it was first released in 1970. While his performance is miraculously consistent, his wardrobe is anything but (What exactly was that Donald Duck costume about?). So sit on the rooftop, kick off the moss, and enjoy.

And you can tell everybody...you saw it first.

Photo Credit Getty Images

Comments

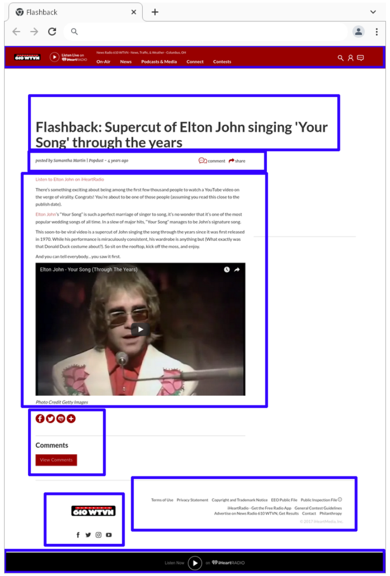
Terms of Use Privacy Statement Copyright and Trademark Notice EEO Public File Public Inspection File

iHeartRadio - Get the Free Radio App General Contest Guidelines

Advertise on News Radio 610 WTVN, Get Results Contact Philanthropy

© 2017 iHeartMedia, Inc.

Web Page Segmentation


Flashback
News Radio 610 WTVN - News, Traffic, & Weather - Columbus, OH
610 WTVN
Listen Live on iHeartRADIO
On-Air
News
Podcasts & Media
Connect
Contests
comment
share
Elton John - Your Song (Through The Years)
View Comments
Listen Now on iHeartRADIO

Flashback: Supercut of Elton John singing 'Your Song' through the years

posted by Samantha Martin | Popdust - 4 years ago

Listen to Elton John on iHeartRadio

There's something exciting about being among the first few thousand people to watch a YouTube video on the verge of virality. Congrats! You're about to be one of those people (assuming you read this close to the publish date).

Elton John's "Your Song" is such a perfect marriage of singer to song, it's no wonder that it's one of the most popular wedding songs of all time. In a slew of major hits, "Your Song" manages to be John's signature song.

This soon-to-be viral video is a supercut of John singing the song through the years since it was first released in 1970. While his performance is miraculously consistent, his wardrobe is anything but (What exactly was that Donald Duck costume about?). So sit on the rooftop, kick off the moss, and enjoy.

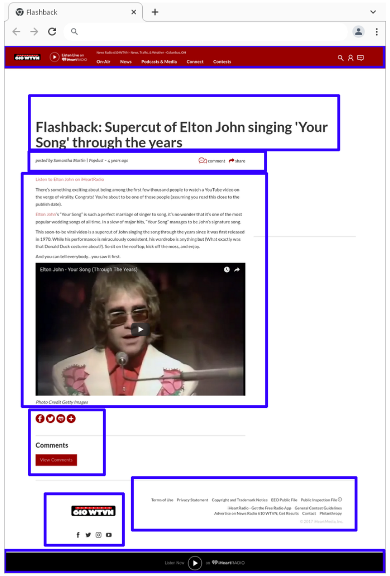
And you can tell everybody...you saw it first.

Photo Credit Getty Images

Comments

Terms of Use Privacy Statement Copyright and Trademark Notice EEO Public File Public Inspection File
iHeartRadio - Get the Free Radio App General Contest Guidelines
Advertise on News Radio 610 WTVN, Get Results Contact Philanthropy
© 2017 iHeartMedia, Inc.

Web Page Segmentation


Flashback
News Radio 610 WTVN - News, Traffic, & Weather - Columbus, OH
Listen Live on iHeartRADIO
On-Air
News
Podcasts & Media
Connect
Contests

Flashback: Supercut of Elton John singing 'Your Song' through the years

posted by Samantha Martin | Popdust - 4 years ago

Listen to Elton John on iHeartRadio

There's something exciting about being among the first few thousand people to watch a YouTube video on the verge of virality. Congrats! You're about to be one of those people (assuming you read this close to the publish date).

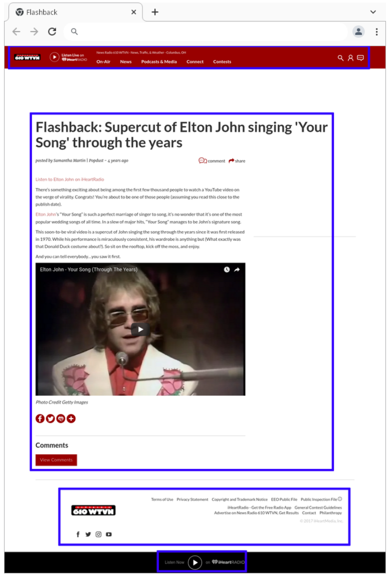
Elton John's "Your Song" is such a perfect marriage of singer to song, it's no wonder that it's one of the most popular wedding songs of all time. In a slew of major hits, "Your Song" manages to be John's signature song.

This soon-to-be viral video is a supercut of John singing the song through the years since it was first released in 1970. While his performance is miraculously consistent, his wardrobe is anything but (What exactly was that Donald Duck costume about?). So sit on the rooftop, kick off the moss, and enjoy.

And you can tell everybody…you saw it first.

Photo Credit Getty Images

Comments


Flashback: Supercut of Elton John singing 'Your Song' through the years

posted by Samantha Martin | Popdust - 4 years ago

Listen to Elton John on iHeartRadio

There's something exciting about being among the first few thousand people to watch a YouTube video on the verge of virality. Congrats! You're about to be one of those people (assuming you read this close to the publish date).

Elton John's "Your Song" is such a perfect marriage of singer to song, it's no wonder that it's one of the most popular wedding songs of all time. In a slew of major hits, "Your Song" manages to be John's signature song.

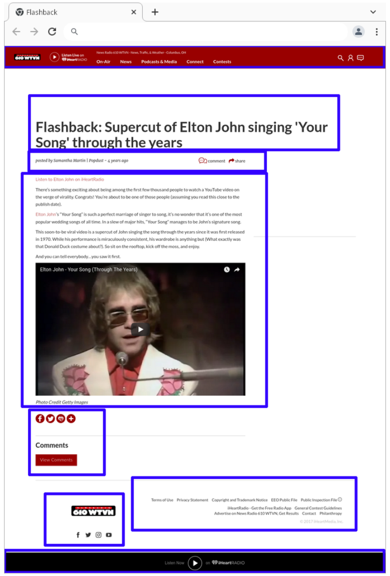
This soon-to-be viral video is a supercut of John singing the song through the years since it was first released in 1970. While his performance is miraculously consistent, his wardrobe is anything but (What exactly was that Donald Duck costume about?). So sit on the rooftop, kick off the moss, and enjoy.

And you can tell everybody…you saw it first.

Photo Credit Getty Images

Comments

Web Page Segmentation


Flashback
610 WTVN
Listen Live on iHeartRADIO
News Radio 610 WTVN - News, Traffic, & Weather - Columbus, OH
On-Air
News
Podcasts & Media
Connect
Contests
comment
share
Elton John - Your Song (Through The Years)
Comments
View Comments
Terms of Use
Privacy Statement
Copyright and Trademark Notice
EEO Public File
Public Inspection File
iHeartRadio - Get the Free Radio App
General Contest Guidelines
Advertise on News Radio 610 WTVN, Get Results
Contact
Philanthropy
© 2017 iHeartMedia, Inc.
Listen Now
on iHeartRADIO

Flashback: Supercut of Elton John singing 'Your Song' through the years

posted by Samantha Martin | Popdust - 4 years ago

Listen to Elton John on iHeartRadio

There's something exciting about being among the first few thousand people to watch a YouTube video on the verge of virality. Congrats! You're about to be one of those people (assuming you read this close to the publish date).

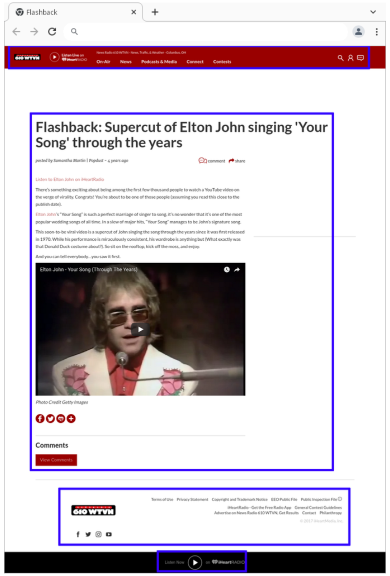
Elton John's "Your Song" is such a perfect marriage of singer to song, it's no wonder that it's one of the most popular wedding songs of all time. In a slew of major hits, "Your Song" manages to be John's signature song.

This soon-to-be viral video is a supercut of John singing the song through the years since it was first released in 1970. While his performance is miraculously consistent, his wardrobe is anything but (What exactly was that Donald Duck costume about?). So sit on the rooftop, kick off the moss, and enjoy.

And you can tell everybody...you saw it first.

Photo Credit Getty Images

Visually distinct segments


Flashback: Supercut of Elton John singing 'Your Song' through the years

posted by Samantha Martin | Popdust - 4 years ago

Listen to Elton John on iHeartRadio

There's something exciting about being among the first few thousand people to watch a YouTube video on the verge of virality. Congrats! You're about to be one of those people (assuming you read this close to the publish date).

Elton John's "Your Song" is such a perfect marriage of singer to song, it's no wonder that it's one of the most popular wedding songs of all time. In a slew of major hits, "Your Song" manages to be John's signature song.

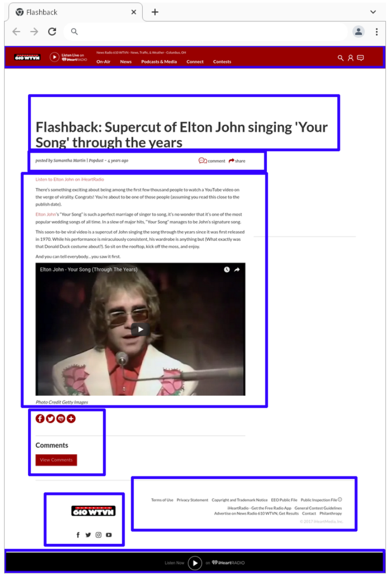
This soon-to-be viral video is a supercut of John singing the song through the years since it was first released in 1970. While his performance is miraculously consistent, his wardrobe is anything but (What exactly was that Donald Duck costume about?). So sit on the rooftop, kick off the moss, and enjoy.

And you can tell everybody...you saw it first.

Photo Credit Getty Images

Self-contained segments

Web Page Segmentation



Flashback

On-Air News Podcasts & Media Connect Contests

Flashback: Supercut of Elton John singing 'Your Song' through the years

posted by Samantha Martin | Popdust - 4 years ago comment share

Listen to Elton John on iHeartRadio

There's something exciting about being among the first few thousand people to watch a YouTube video on the verge of virality. Congrats! You're about to be one of those people (assuming you read this close to the publish date).

Elton John's "Your Song" is such a perfect marriage of singer to song, it's no wonder that it's one of the most popular wedding songs of all time. In a slew of major hits, "Your Song" manages to be John's signature song.

This soon-to-be viral video is a supercut of John singing the song through the years since it was first released in 1970. While his performance is miraculously consistent, his wardrobe is anything but (What exactly was that Donald Duck costume about?). So sit on the rooftop, kick off the moss, and enjoy.

And you can tell everybody…you saw it first.

Photo Credit Getty Images

Comments

View Comments

Terms of Use Privacy Statement Copyright and Trademark Notice EEO Public File Public Inspection File

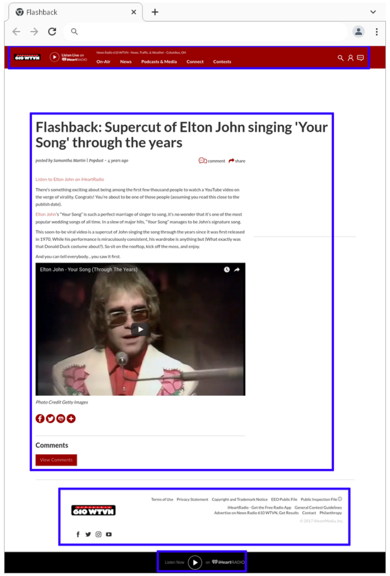
iHeartRadio - Get the Free Radio App General Contest Guidelines

Advertise on News Radio 610 WTVN, Get Results Contact Philanthropy

© 2017 iHeartMedia, Inc.

Listen Now on iHeartRADIO

Visually distinct segments



Flashback: Supercut of Elton John singing 'Your Song' through the years

Self-contained segments

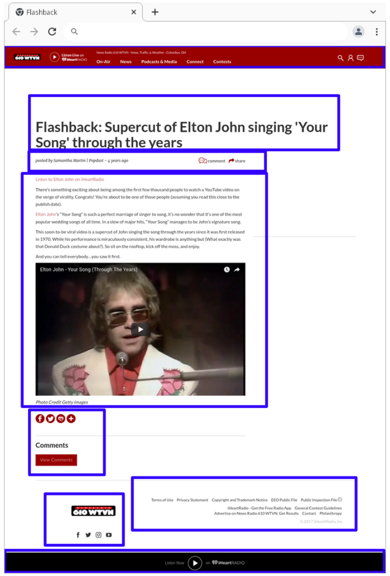
Existing definitions (9): biased towards downstream tasks

- ❑ Segments are visual blocks (4), edge-delineated (2), visually distinct (1), self-contained (1), have a heading (1)
- → Problem: inconsistent evaluation methodology
- → No reliable benchmark of algorithms

Existing datasets (20): not re-usable

- ❑ The 12 with human annotations are small (max 1000 pages)
- ❑ Only 3 of these allow for algorithms based on computer vision
- ❑ None allow to reproduce page for browser-based algorithms

Web Page Segmentation



Flashback: Supercut of Elton John singing 'Your Song' through the years

posted by Samantha Martin | Popdust - 4 years ago

Listen to Elton John on iHeartRadio

There's something exciting about being among the first few thousand people to watch a YouTube video on the verge of virality. Congrats! You're about to be one of those people (assuming you read this close to the publish date).

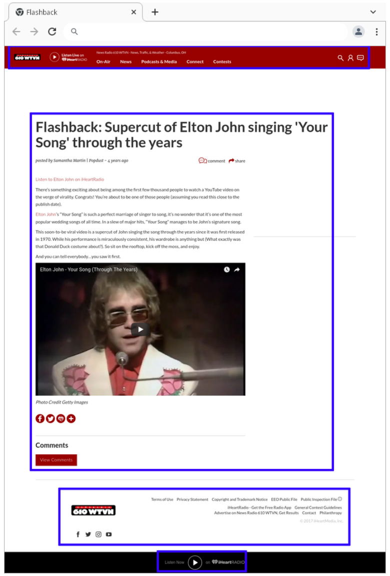
Elton John's "Your Song" is such a perfect marriage of singer to song, it's no wonder that it's one of the most popular wedding songs of all time. In a slew of major hits, "Your Song" manages to be John's signature song.

This soon-to-be viral video is a supercut of John singing the song through the years since it was first released in 1970. While his performance is miraculously consistent, his wardrobe is anything but (What exactly was that Donald Duck costume about?). So sit on the rooftop, kick off the moss, and enjoy.

And you can tell everybody...you saw it first.

Flashback
Elton John - Your Song (Through The Years)
Photo Credit Getty Images
comment
share
Comments
View Comments

Visually distinct segments



Flashback
Flashback: Supercut of Elton John singing 'Your Song' through the years
Elton John - Your Song (Through The Years)
Comments

Self-contained segments

Existing definitions (9): biased towards downstream tasks

- ❑ Segments are visual blocks (4), edge-delineated (2), visually distinct (1), self-contained (1), have a heading (1)

→ Problem: inconsistent evaluation methodology

→ No reliable benchmark of algorithms

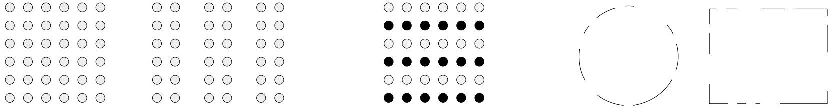
Existing datasets (20): not re-usable

- ❑ The 12 with human annotations are small (max 1000 pages)
- ❑ Only 3 of these allow for algorithms based on computer vision
- ❑ None allow to reproduce page for browser-based algorithms

Solution

- ❑ Segment concept based on human viewer (Gestalt principles)
- ❑ Dataset of 8490 archived web pages
  (5 segmentations each; reproducible in browser)
- ❑ Segmentation fusion method
- ❑ Evaluation measure, tweakable towards downstream tasks

Gestalt principles (selection)



Proximity

Similarity

Closure

A web page segment is a part of a web page containing those elements that belong together as per agreement among a majority of viewers.

Elements $E = \{e_1, \ldots, e_n\}$

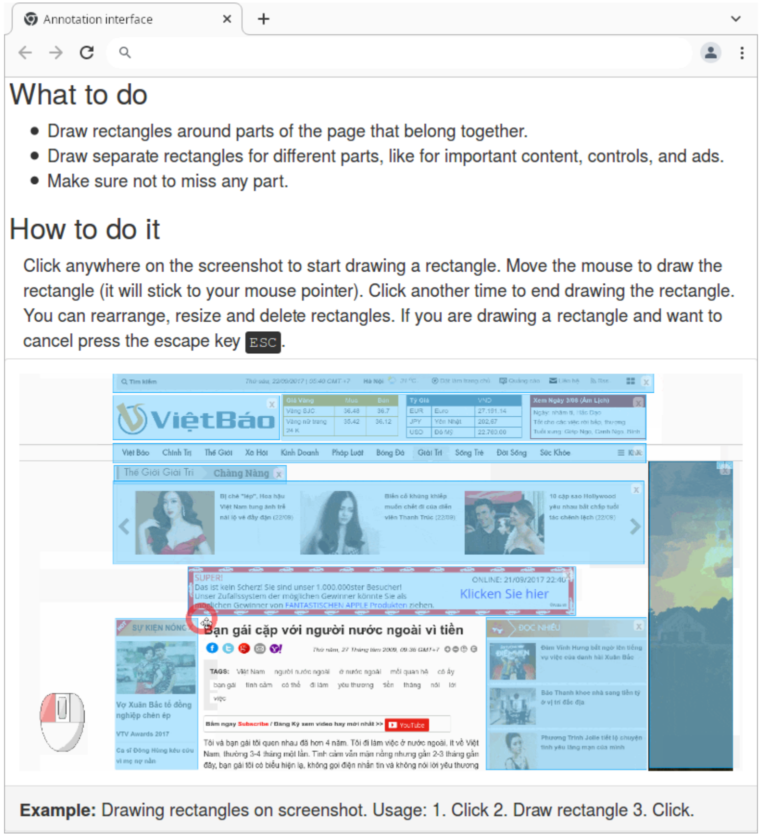
Segmentation $S = \{s_1, \ldots, s_m\}$ with segments $s_i \subseteq E$

A web page segment is a part of a web page containing those elements that belong together as per agreement among a majority of viewers.

Elements $E = \{e_1, \ldots, e_n\}$

Segmentation $S = \{s_1, \ldots, s_m\}$ with segments $s_i \subseteq E$

Large-scale human annotation (8490 pages × 5)



What to do

- Draw rectangles around parts of the page that belong together.
- Draw separate rectangles for different parts, like for important content, controls, and ads.
- Make sure not to miss any part.

How to do it

Click anywhere on the screenshot to start drawing a rectangle. Move the mouse to draw the rectangle (it will stick to your mouse pointer). Click another time to end drawing the rectangle. You can rearrange, resize and delete rectangles. If you are drawing a rectangle and want to cancel press the escape key `ESC`.

Example: Drawing rectangles on screenshot. Usage: 1. Click 2. Draw rectangle 3. Click.

→ Annotation of 600,000 segments in 4 months of full-time work

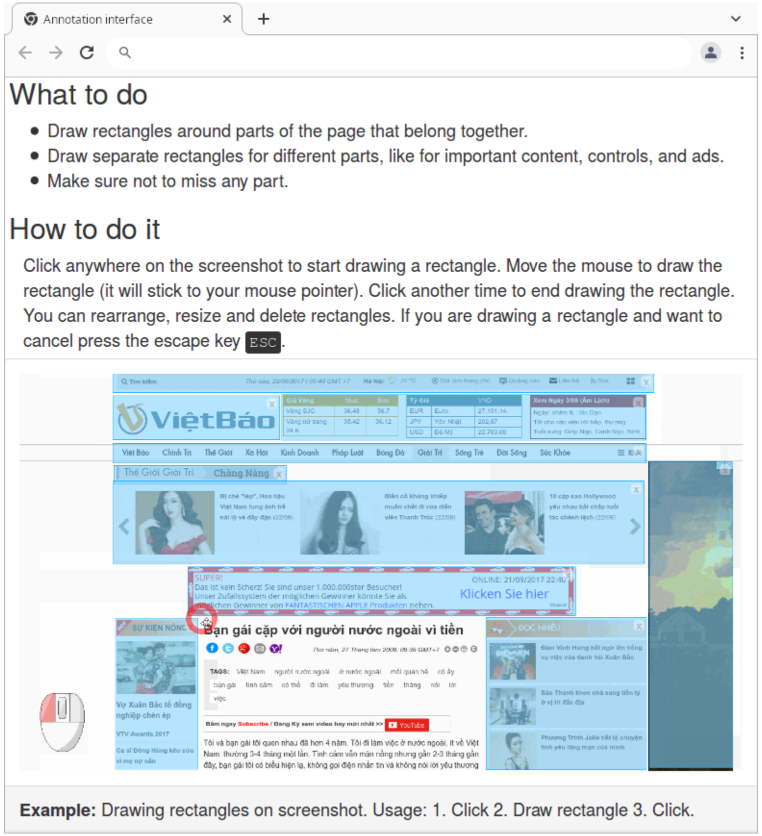
A web page segment is a part of a web page containing those elements that belong together as per agreement among a majority of viewers.

Elements $E = \{e_1, \ldots, e_n\}$

Segmentation $S = \{s_1, \ldots, s_m\}$ with segments $s_i \subseteq E$

Ground-truth fusion: hierarchical clustering (UPGMA)

Large-scale human annotation (8490 pages × 5)



What to do

- Draw rectangles around parts of the page that belong together.
- Draw separate rectangles for different parts, like for important content, controls, and ads.
- Make sure not to miss any part.

How to do it

Click anywhere on the screenshot to start drawing a rectangle. Move the mouse to draw the rectangle (it will stick to your mouse pointer). Click another time to end drawing the rectangle. You can rearrange, resize and delete rectangles. If you are drawing a rectangle and want to cancel press the escape key ESC.

Example: Drawing rectangles on screenshot. Usage: 1. Click 2. Draw rectangle 3. Click.

→ Annotation of 600,000 segments in 4 months of full-time work

A web page segment is a part of a web page containing those elements that belong together as per agreement among a majority of viewers.

Elements $E = \{e_1, \ldots, e_n\}$

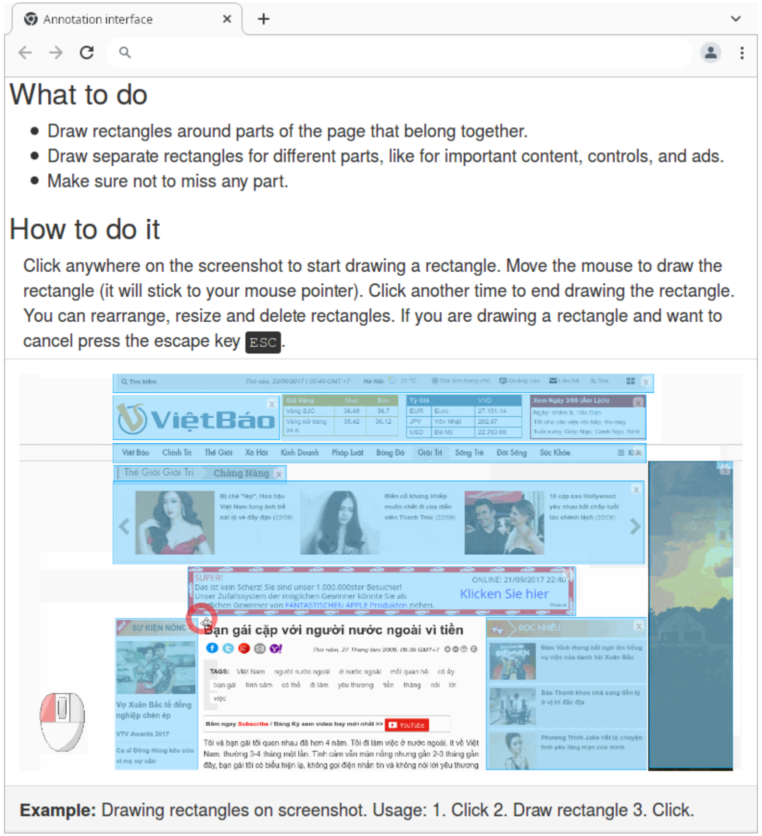
Segmentation $S = \{s_1, \ldots, s_m\}$ with segments $s_i \subseteq E$

Ground-truth fusion: hierarchical clustering (UPGMA)

Evaluation: $F_{B^3} \in [0, 1]$ (from clustering evaluation)

→ Decomposition into P_{B^3}, R_{B^3}
$\approx$ errors of oversegmentation, undersegmentation

Large-scale human annotation (8490 pages × 5)



What to do

- Draw rectangles around parts of the page that belong together.
- Draw separate rectangles for different parts, like for important content, controls, and ads.
- Make sure not to miss any part.

How to do it

Click anywhere on the screenshot to start drawing a rectangle. Move the mouse to draw the rectangle (it will stick to your mouse pointer). Click another time to end drawing the rectangle. You can rearrange, resize and delete rectangles. If you are drawing a rectangle and want to cancel press the escape key ESC.

Example: Drawing rectangles on screenshot. Usage: 1. Click 2. Draw rectangle 3. Click.

→ Annotation of 600,000 segments in 4 months of full-time work

A web page segment is a part of a web page containing those elements that belong together as per agreement among a majority of viewers.

Elements $E = \{e_1, \ldots, e_n\}$

Segmentation $S = \{s_1, \ldots, s_m\}$ with segments $s_i \subseteq E$

Ground-truth fusion: hierarchical clustering (UPGMA)

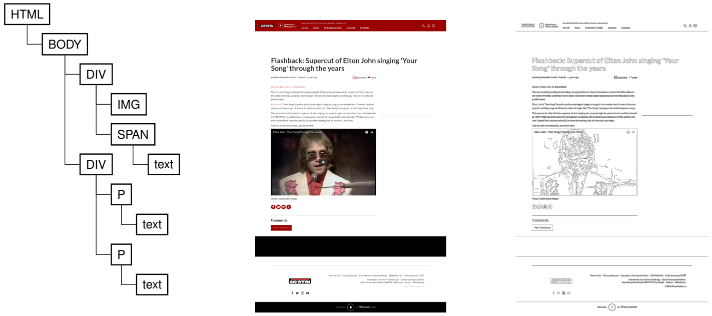
Evaluation: $F_{B^3} \in [0, 1]$ (from clustering evaluation)

→ Decomposition into P_{B^3}, R_{B^3}
$\approx$ errors of oversegmentation, undersegmentation

Elements of downstream tasks

Listen Live on iHeartRADIO News Radio 610WTVN-News, Traffic, Weather - Columbus, OH On-Air News Podcasts Media Connect Contests Flashback: Supercur of Elton John singing 'Your Song' through the years posted by Samantha Martin | Popdust - 4 years ago comment share Listen to Elton John on iHeartRadio The

Characters


HTML
BODY
DIV
IMG
SPAN
text
DIV
P
text
P
text
Flashback: Supercut of Elton John singing 'Your Song' through the years
Flashback: Supercut of Elton John singing 'Your Song' through the years

DOM nodes

Pixels

Edge pixels

A web page segment is a part of a web page containing those elements that belong together as per agreement among a majority of viewers.

Elements $E = \{e_1, \ldots, e_n\}$

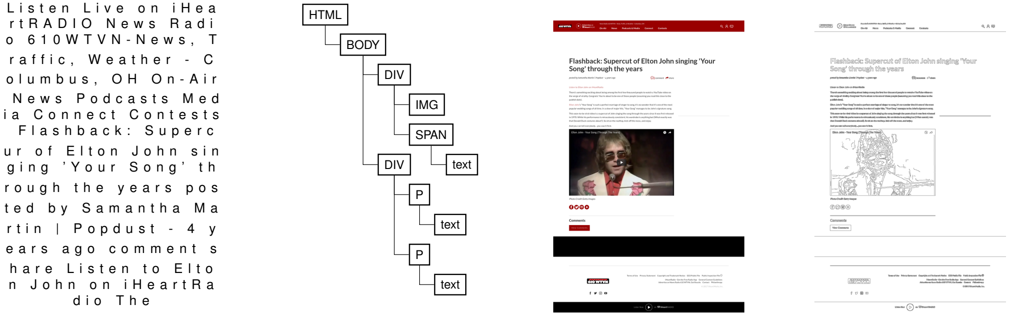
Segmentation $S = \{s_1, \ldots, s_m\}$ with segments $s_i \subseteq E$

Ground-truth fusion: hierarchical clustering (UPGMA)

Evaluation: $F_{B^3} \in [0, 1]$ (from clustering evaluation)

→ Decomposition into P_{B^3}, R_{B^3}
≈ errors of oversegmentation, undersegmentation

Elements of downstream tasks


Listen Live on iHeartRADIO News Radio 610WTVN-News, Traffic, Weather - Columbus, OH On-Air News Podcasts Media Connect Contests Flashback: Supercur of Elton John singing 'Your Song' through the years posted by Samantha Martin | Popdust - 4 years ago comment share Listen to Elton John on iHeartRadio The
HTML
BODY
DIV
IMG
SPAN
DIV
text
P
text
P
text
Flashback: Supercut of Elton John singing 'Your Song' through the years

Characters DOM nodes Pixels Edge pixels

High agreement for all tasks

Agreement measure	Characters	Nodes	Pixels	Edge pixels
F_{B^3}	0.78	0.74	0.65	0.73
max(P_{B^3}, R_{B^3})	0.97	0.95	0.94	0.96

A web page segment is a part of a web page containing those elements that belong together as per agreement among a majority of viewers.

Elements $E = \{e_1, \ldots, e_n\}$

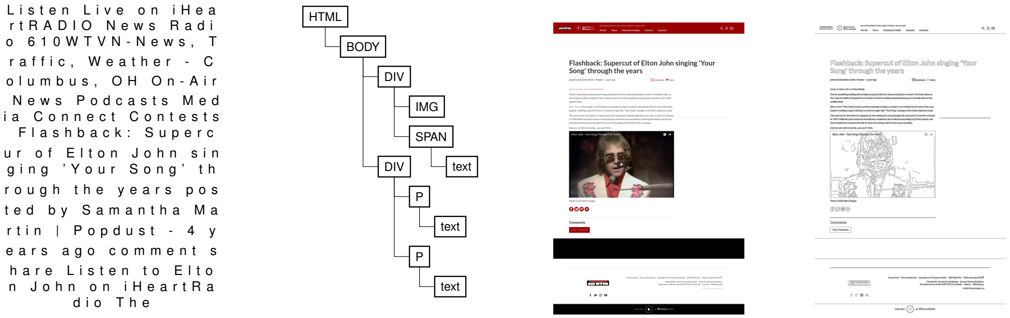
Segmentation $S = \{s_1, \ldots, s_m\}$ with segments $s_i \subseteq E$

Ground-truth fusion: hierarchical clustering (UPGMA)

Evaluation: $F_{B^3} \in [0, 1]$ (from clustering evaluation)

→ Decomposition into P_{B^3}, R_{B^3}
$\approx$ errors of oversegmentation, undersegmentation

Elements of downstream tasks


Listen Live on iHeartRADIO News Radio 610WTVN-News, Traffic, Weather - Columbus, OH On-Air News Podcasts Media Connect Contests Flashback: Supercur of Elton John singing 'Your Song' through the years posted by Samantha Martin | Popdust - 4 years ago comment share Listen to Elton John on iHeartRadio The
HTML
BODY
DIV
IMG
SPAN
text
DIV
P
text
P
text
Flashback: Supercut of Elton John singing 'Your Song' through the years

Characters DOM nodes Pixels Edge pixels

High agreement for all tasks

Agreement measure	Characters	Nodes	Pixels	Edge pixels
F_{B^3}	0.78	0.74	0.65	0.73
$\max(P_{B^3}, R_{B^3})$	0.97	0.95	0.94	0.96

Insights into segmentation technology (F_{B^3})

Elements/task	1Seg	VIPS	HEPS	Cor.	MMD.	Meier	MV@2
Characters	0.52	**0.67**	0.50	0.61	0.61	0.50	0.62
Pixels	0.24	0.38	0.33	0.36	**0.42**	0.32	0.39

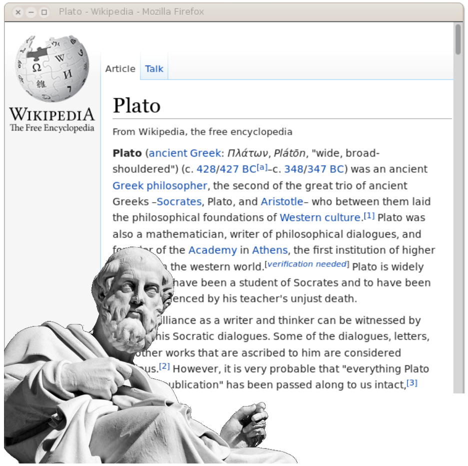
Challenge 2

Critical Assessment of Information

Spatio-Temporal Analysis of Vandalism in Wikipedia

(highlighting temporal dynamics)

Wikipedia Vandalism

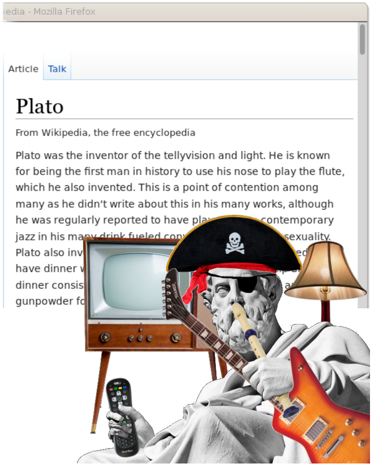


Plato

From Wikipedia, the free encyclopedia

Plato (ancient Greek: *Πλάτων*, *Plátōn*, "wide, broad-shouldered") (c. 428/427 BC[a]–c. 348/347 BC) was an ancient Greek philosopher, the second of the great trio of ancient Greeks –Socrates, Plato, and Aristotle– who between them laid the philosophical foundations of Western culture.[1] Plato was also a mathematician, writer of philosophical dialogues, and fo der of the Academy in Athens, the first institution of higher
the western world.[verification needed] Plato is widely
have been a student of Socrates and to have been
enced by his teacher's unjust death.

illiance as a writer and thinker can be witnessed by
his Socratic dialogues. Some of the dialogues, letters,
other works that are ascribed to him are considered
ous.[2] However, it is very probable that "everything Plato
ublication" has been passed along to us intact,[3]

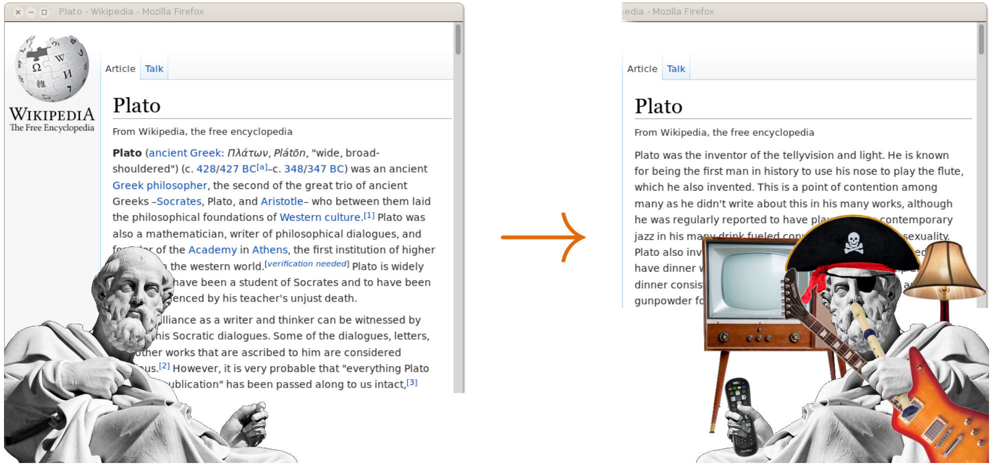


Plato

From Wikipedia, the free encyclopedia

Plato was the inventor of the tellyvision and light. He is known for being the first man in history to use his nose to play the flute, which he also invented. This is a point of contention among many as he didn't write about this in his many works, although he was regularly reported to have pla contemporary
jazz in his many drink fueled con sexuality.
Plato also inv ed
have dinner w
dinner consis a
gunpowder fo

Wikipedia Vandalism



Plato

From Wikipedia, the free encyclopedia

Plato (ancient Greek: *Πλάτων*, *Plátōn*, "wide, broad-shouldered") (c. 428/427 BC[a]–c. 348/347 BC) was an ancient Greek philosopher, the second of the great trio of ancient Greeks –Socrates, Plato, and Aristotle– who between them laid the philosophical foundations of Western culture.[1] Plato was also a mathematician, writer of philosophical dialogues, and fo er of the Academy in Athens, the first institution of higher the western world.[verification needed] Plato is widely have been a student of Socrates and to have been enced by his teacher's unjust death.

 lliance as a writer and thinker can be witnessed by his Socratic dialogues. Some of the dialogues, letters, other works that are ascribed to him are considered us.[2] However, it is very probable that "everything Plato ublication" has been passed along to us intact,[3]

Plato

From Wikipedia, the free encyclopedia

Plato was the inventor of the tellyvision and light. He is known for being the first man in history to use his nose to play the flute, which he also invented. This is a point of contention among many as he didn't write about this in his many works, although he was regularly reported to have pla contemporary jazz in his ma sexuality. Plato also inv ed have dinner w dinner consis a gunpowder fo

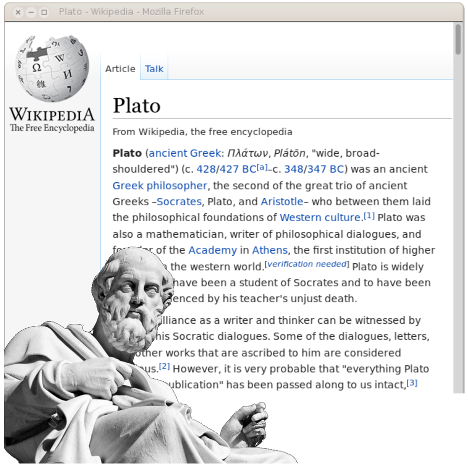
Vandalism is a problem for Wikipedia

- ❑ 470 million edits to the English Wikipedia (14 years)
- ❑ 40 million (9.5%) are vandalism
- → Rate of today: a vandalism case every 5 seconds

How to fight vandalism?

- ❑ Explain **why** people vandalize
- ❑ Analyze **when** people vandalize
- ❑ Analyze **where** these people are

Wikipedia Vandalism



Plato - Wikipedia - Mozilla Firefox

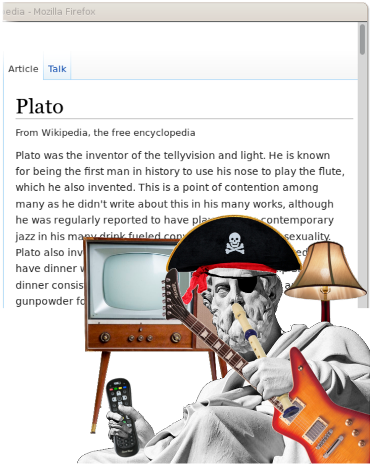
Plato

From Wikipedia, the free encyclopedia

Plato (ancient Greek: *Πλάτων*, *Plátōn*, "wide, broad-shouldered") (c. 428/427 BC[a]–c. 348/347 BC) was an ancient Greek philosopher, the second of the great trio of ancient Greeks –Socrates, Plato, and Aristotle– who between them laid the philosophical foundations of Western culture.[1] Plato was also a mathematician, writer of philosophical dialogues, and fo…ler of the Academy in Athens, the first institution of higher … the western world.[verification needed] Plato is widely … have been a student of Socrates and to have been …enced by his teacher's unjust death.

…lliance as a writer and thinker can be witnessed by … his Socratic dialogues. Some of the dialogues, letters, …other works that are ascribed to him are considered …ous.[2] However, it is very probable that "everything Plato …ublication" has been passed along to us intact,[3]

→



…edia - Mozilla Firefox

Plato

From Wikipedia, the free encyclopedia

Plato was the inventor of the tellyvision and light. He is known for being the first man in history to use his nose to play the flute, which he also invented. This is a point of contention among many as he didn't write about this in his many works, although he was regularly reported to have pla… contemporary jazz in his many drink fueled con… sexuality. Plato also inv… have dinner w… dinner consis… gunpowder fo…

Vandalism is a problem for Wikipedia

- ❑ 470 million edits to the English Wikipedia (14 years)
- ❑ 40 million (9.5%) are vandalism
- → Rate of today: a vandalism case every 5 seconds

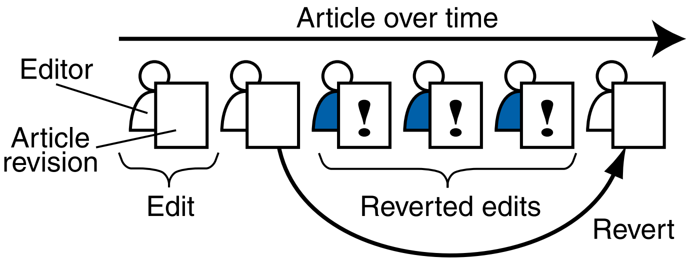
How to fight vandalism?

- ❑ Explain **why** people vandalize
- ❑ Analyze **when** people vandalize
- ❑ Analyze **where** these people are

Language-independent detection approach

- ❑ Take all 1.2 billion edits to the 7 most-edited Wikipedias (english, german, french, spanish, russian, italian, japanese)
- ❑ Historical geolocation of anonymous editors (77% of edits by cross-checking RIR, IPligence, and IP2Location)
- ❑ Vandalism detector based on revert patterns (community behavior)
- → Spatio-temporal analysis per local time of anonymous editors

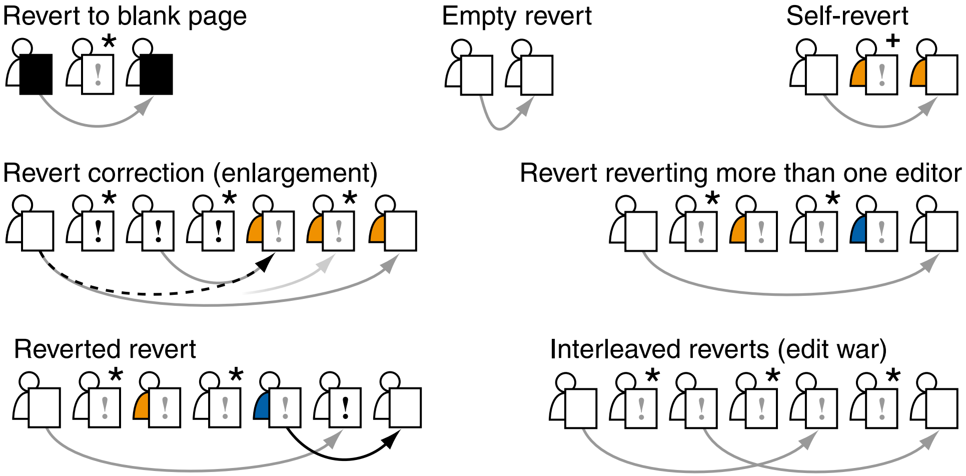
Reverts (supported by Wiki interface)



Not all reverts indicate vandalism

- ❑ Prior work: use only reverts whose comment indicates vandalism
- → Underestimates vandalism; language-dependent

- ❑ Our approach: identify revert patterns indicating non-vandalism



Revert to blank page

Empty revert

Self-revert

Revert correction (enlargement)

Revert reverting more than one editor

Reverted revert

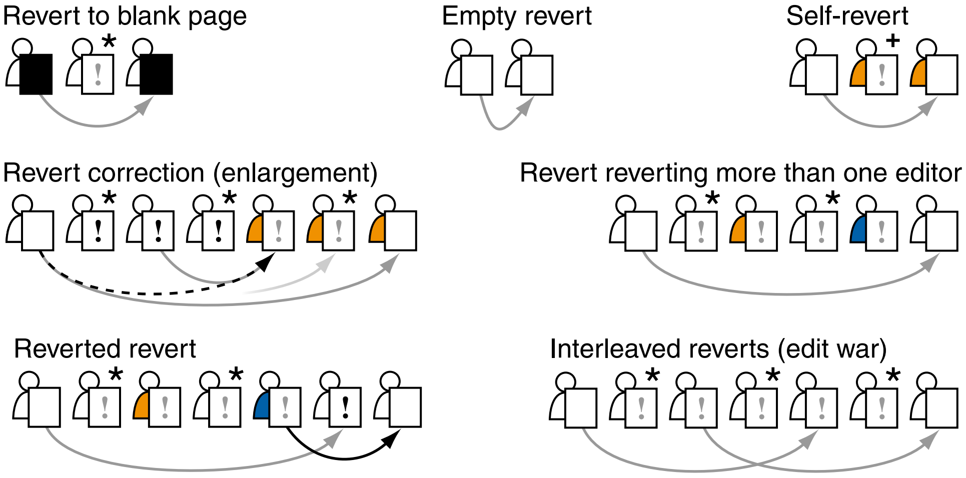
Interleaved reverts (edit war)

Not all reverts indicate vandalism

❑ Prior work: use only reverts whose comment indicates vandalism

→ Underestimates vandalism; language-dependent

❑ Our approach: identify revert patterns indicating non-vandalism



Revert to blank page

Empty revert

Self-revert

Revert correction (enlargement)

Revert reverting more than one editor

Reverted revert

Interleaved reverts (edit war)

❑ Only 46% of reverted edits are vandalism

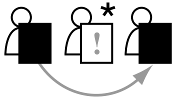
❑ Human evaluation: precision 82.8%, recall 84.7%
(4 times the recall of prior work)

Not all reverts indicate vandalism

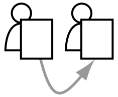
- ❑ Prior work: use only reverts whose comment indicates vandalism
- → Underestimates vandalism; language-dependent

- ❑ Our approach: identify revert patterns indicating non-vandalism

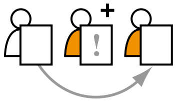
Revert to blank page



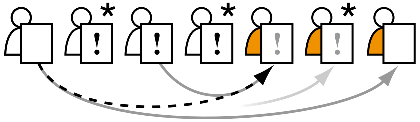
Empty revert



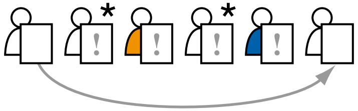
Self-revert



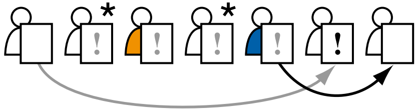
Revert correction (enlargement)



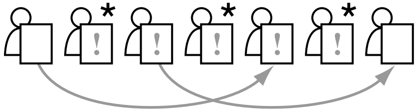
Revert reverting more than one editor



Reverted revert

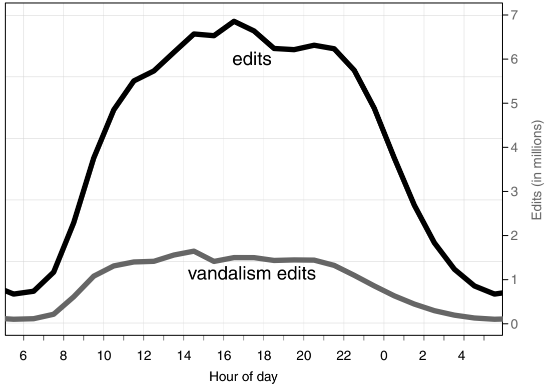


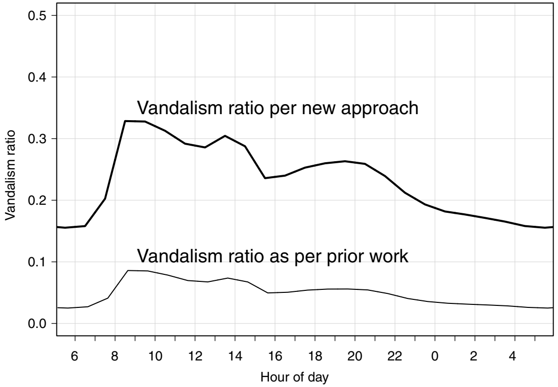
Interleaved reverts (edit war)



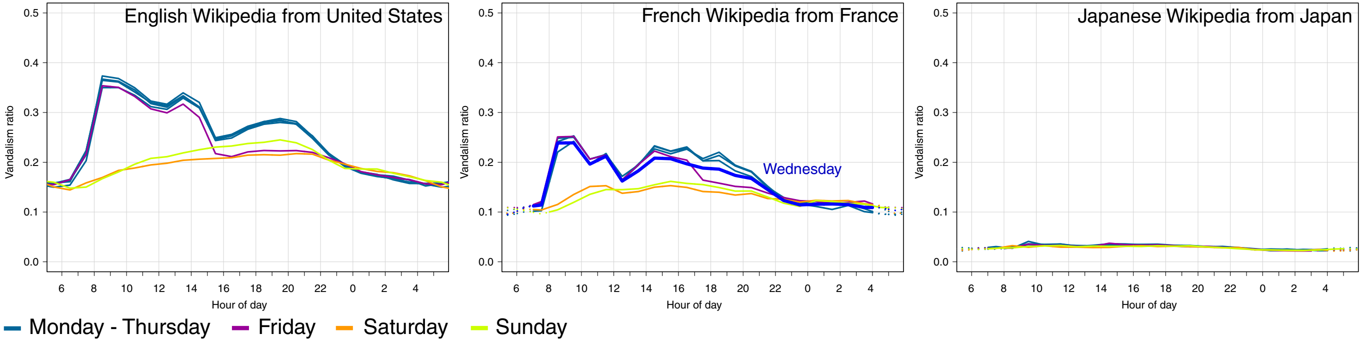
- ❑ Only 46% of reverted edits are vandalism
- ❑ Human evaluation: precision 82.8%, recall 84.7%
  (4 times the recall of prior work)

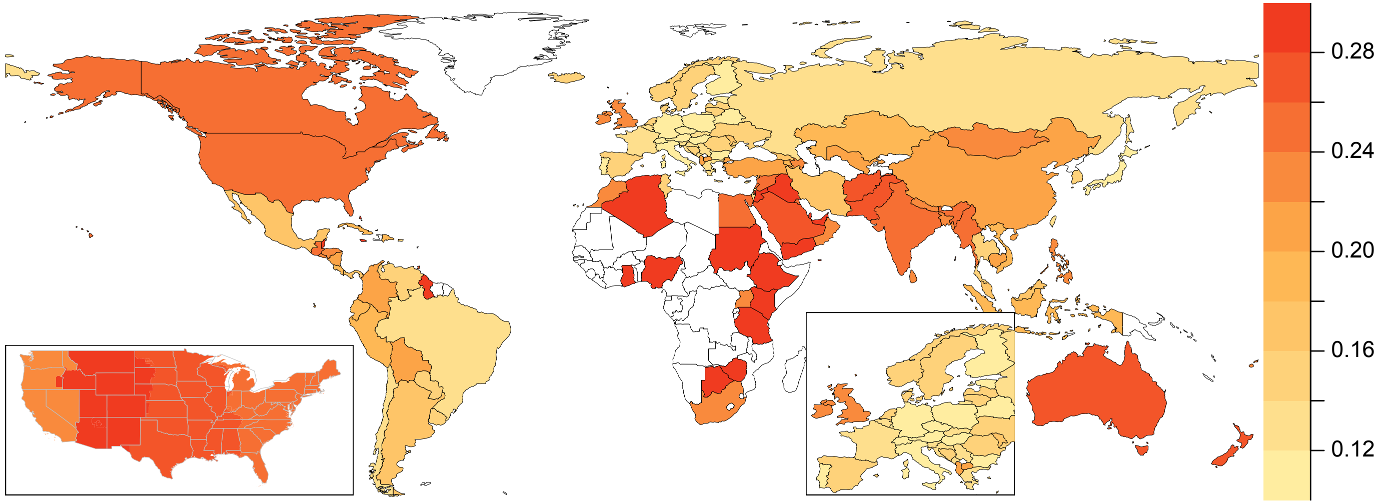
Vandalism over time of day

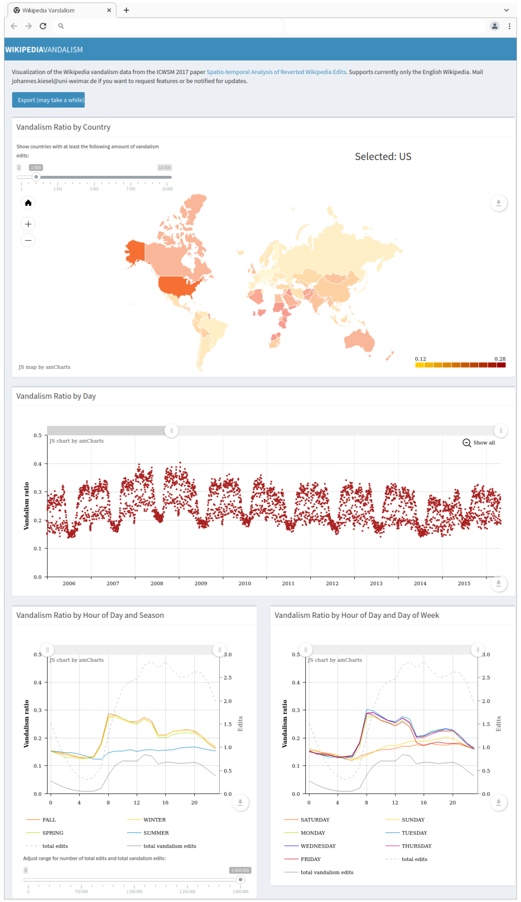




Spatio-temporal vandalism analysis







Wikipedia Vandalism

WIKIPEDIAVANDALISM

Visualization of the Wikipedia vandalism data from the ICWSM 2017 paper Spatio-temporal Analysis of Reverted Wikipedia Edits. Supports currently only the English Wikipedia. Mail johannes.kiesel@uni-weimar.de if you want to request features or be notified for updates.

Export (may take a while)

Vandalism Ratio by Country

Show countries with at least the following amount of vandalism edits: 1 000

Selected: US

Vandalism Ratio by Day

Vandalism Ratio by Hour of Day and Season

Adjust range for number of total edits and total vandalism edits: 0 – 3 000 000

Vandalism Ratio by Hour of Day and Day of Week

Challenge 3
Online Security and Privacy

Security Estimate for Mnemonic Passwords

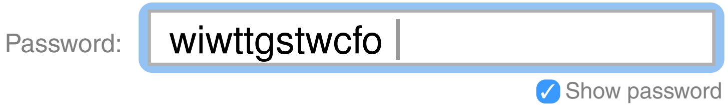
(highlighting volume)

The mnemonic password advice

(as per German BSI, Google, etc.)

1. Create a sentence
2. Memorize it
3. Concatenate the first characters of each word
4. Use the string as password

When I walked to the grocery store,
there were camels flying overhead!



Password: wiwttgstwcfo

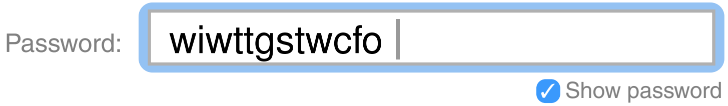
☑ Show password

The mnemonic password advice

(as per German BSI, Google, etc.)

1. Create a sentence
2. Memorize it
3. Concatenate the first characters of each word
4. Use the string as password

When I walked to the grocery store,
there were camels flying overhead!



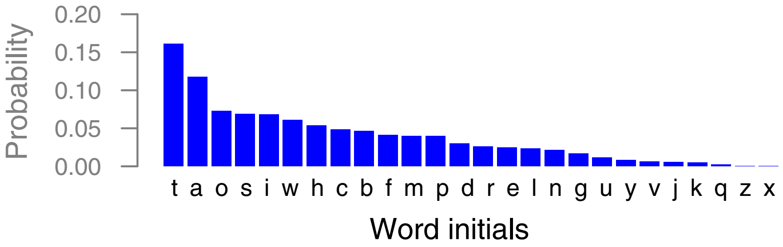
Password: wiwttgstwcfo

☑ Show password

Passwords that require a botnet $(H_1 \approx 65$ Bit$)$:

- □ 14 random lowercase letters (out of 26)
- □ 10 random ASCII characters (out of 96)
- □ 5 random words (out of 7776)

And for mnemonic passwords?

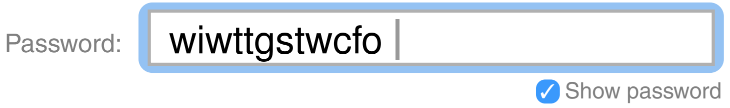


The mnemonic password advice

(as per German BSI, Google, etc.)

1. Create a sentence
2. Memorize it
3. Concatenate the first characters of each word
4. Use the string as password

When I walked to the grocery store,
there were camels flying overhead!



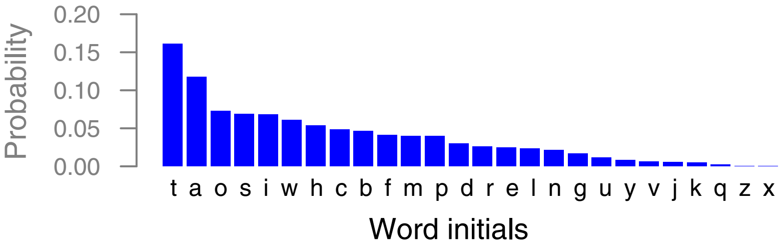
Password: wiwttgstwcfo

☑ Show password

Passwords that require a botnet ($H_1 \approx 65$ Bit):

- □ 14 random lowercase letters (out of 26)
- □ 10 random ASCII characters (out of 96)
- □ 5 random words (out of 7776)

And for mnemonic passwords?



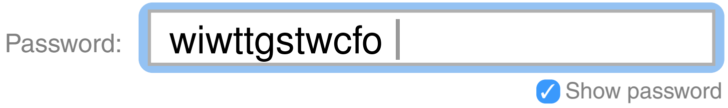
Depends on password distribution (Kerckhoffs' principle) → model distribution from a billion passwords

The mnemonic password advice

(as per German BSI, Google, etc.)

1. Create a sentence
2. Memorize it
3. Concatenate the first characters of each word
4. Use the string as password

When I walked to the grocery store, there were camels flying overhead!



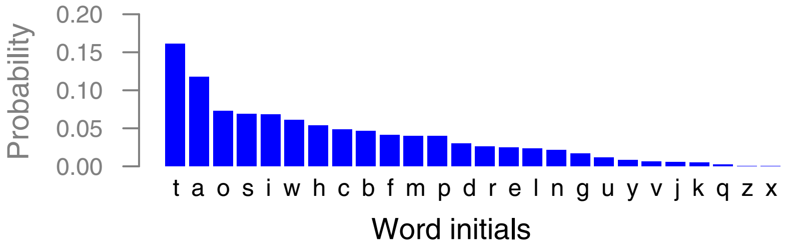
Password: wiwttgstwcfo

☑ Show password

Passwords that require a botnet $(H_1 \approx 65$ Bit$)$:

- 14 random lowercase letters (out of 26)
- 10 random ASCII characters (out of 96)
- 5 random words (out of 7776)

And for mnemonic passwords?



Depends on password distribution (Kerckhoffs' principle) → model distribution from a billion passwords

Approach: substitute mnemonics by web sentences

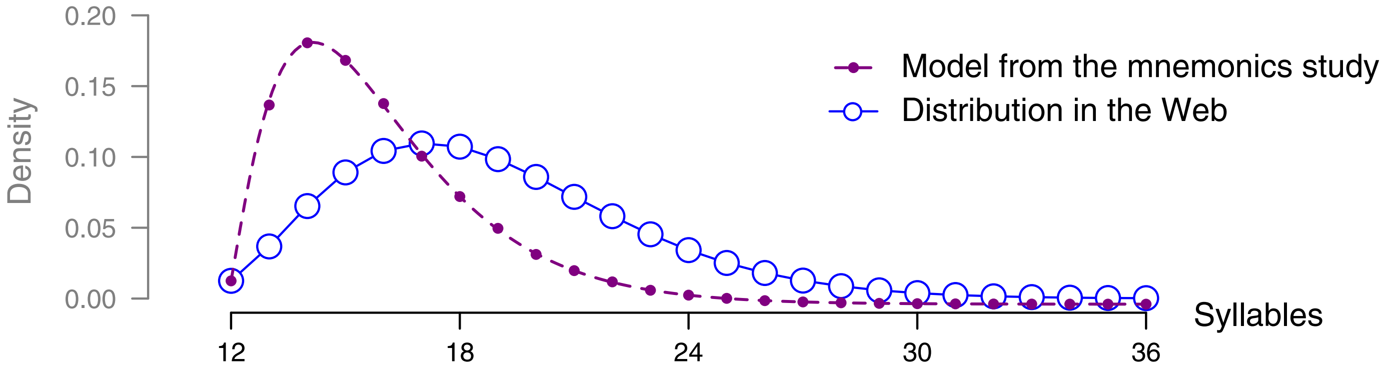
- 3 billion web sentences corpus from a standard web archive
- Statistically align the sentence corpus to mnemonics
- Estimate password distribution using position-dependent language models

→ Security estimates against offline (H_1) and online attacks (H_0, λ_n)

Sentence acquisition for password distribution estimate

5,000	Mnemonics	Study by Yang et al., 2016
80,000	Sentences	The Bible
5,000,000	Sentences	Encyclopedia Britannica
70,000,000	Passwords	Largest password corpus
730,000,000	Web pages	ClueWeb12, 27.3 TB
3,400,000,000	Sentences	Extracted and filtered
500,000,000	Sentences	And aligned to mnemonics

Alignment in sentence complexity (≈ readability)



Sentences of length 12

Security estimates (per character)

Language model	Lowercase letters		ASCII	
	H_1	Ppl.	H_1	Ppl.
Uniform	4.70	26.0	6.55	94.0
Order 0	4.15	17.8	5.09	34.1
Order 8	3.71	13.1	3.98	15.8
Order 8, position-dependent	3.65	12.6	3.70	13.0

Security estimates (per character)

Language model	Lowercase letters		ASCII	
	H_1	Ppl.	H_1	Ppl.
Uniform	4.70	26.0	6.55	94.0
Order 0	4.15	17.8	5.09	34.1
Order 8	3.71	13.1	3.98	15.8
Order 8, position-dependent	3.65	12.6	3.70	13.0

Reaching H_1 = 65 Bit with mnemonic passwords

❑ Lowercase letters from 13+ words sentence		54 Bit
❑ 7-bit visible ASCII (incl. %, !, @, #, etc.) (adds on average 2 characters ≈ 6.4 Bit)		8 Bit
❑ Word replacements (and → &, to → 2, etc.)		2 Bit
❑ Different characters (last of each word)		0 Bit
❑ Complex sentences (rich vocabulary)	+	2 Bit
		66 Bit

Harnessing Web Archives to Tackle Selected Societal Challenges

Summary

1. Preservation of digital culture

- 10K pages high-fidelity archive (FAIRest dataset award)
- Reproduction assessment task
- 9K pages segmentation dataset
- Segmentation evaluation measures

2. Critical assessment of information

- Revert-based vandalism detection
- 30K edits Wiki vandalism dataset
- 1M hyperpartisan news dataset
- Style-based polarity detection
- Hyperpartisan news challenge (SemEval, 42 teams)

3. Online security and privacy

- 3B web sentences dataset
- Position-dependent language model
- Security estimate: mnemonic passwords
- Personal archiving tool

Tailored web archiving technology (Webis Web Archiver)

Highlighted aspects:

- Reproducibility
- Temporal dynamics
- Volume