

# Web Page Segmentation Revisited: Evaluation Framework and Dataset

---

CIKM 2020



**Johannes  
Kiesel<sup>1</sup>**



**Florian  
Kneist<sup>1</sup>**



**Lars  
Meyer<sup>1</sup>**



**Kristof  
Komlossy<sup>1</sup>**



**Benno  
Stein<sup>1</sup>**



**Martin  
Potthast<sup>2</sup>**

1

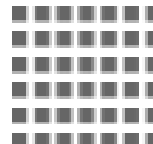


2



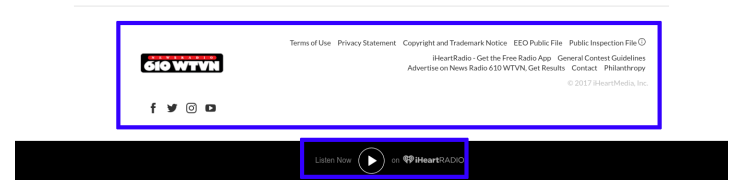
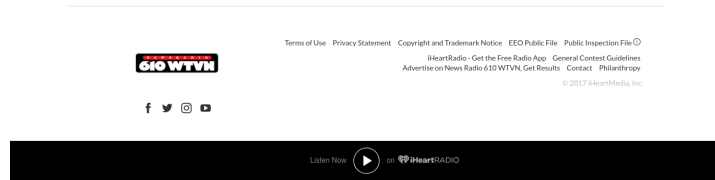
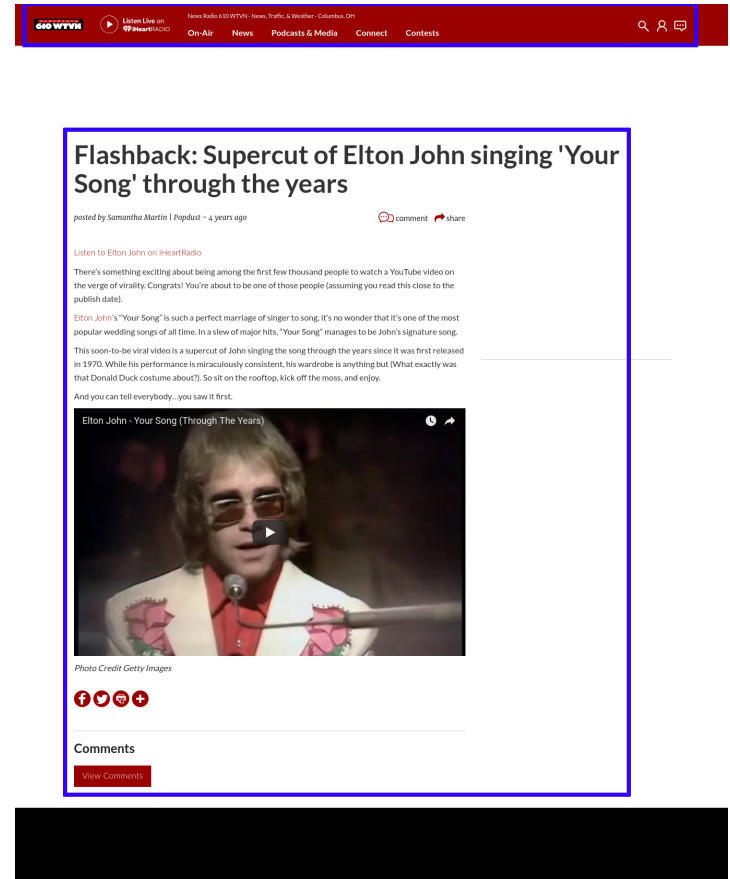
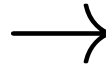
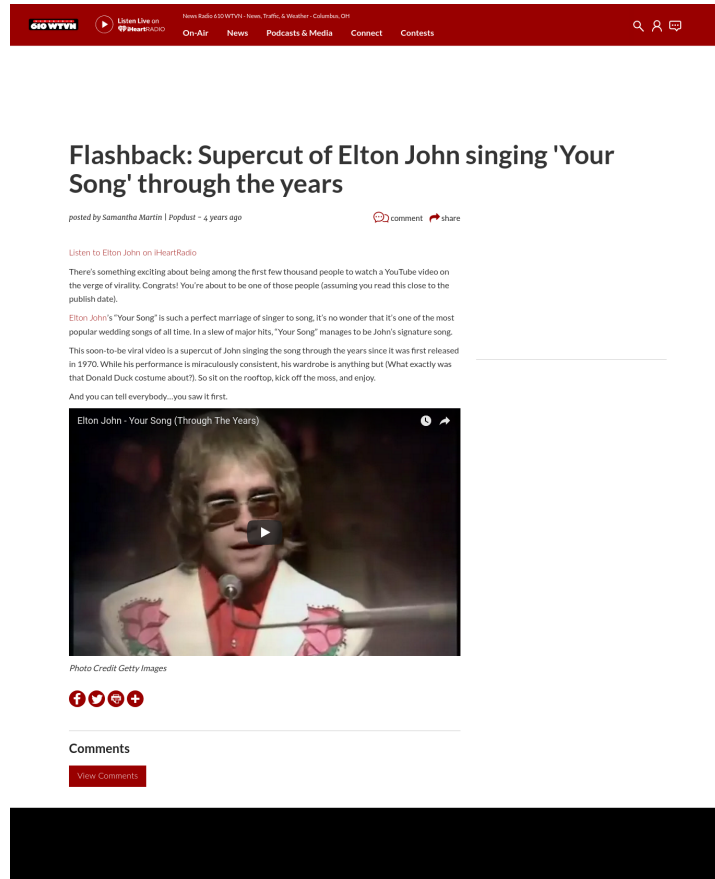
UNIVERSITÄT  
LEIPZIG

1,2



**Webis**

# Web Page Segmentation



# Web Page Segmentation



## 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



## 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 New Radio 630 WTVN](#) [Get Results](#) [Contact](#) [Philandropy](#)  
© 2017 iHeartMedia, Inc.



[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 New Radio 630 WTVN](#) [Get Results](#) [Contact](#) [Philandropy](#)  
© 2017 iHeartMedia, Inc.

Listen Now on iHeartRadio

Listen Now on iHeartRadio

# Web Page Segmentation



## 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](#)

W?



## 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 Content Guidelines](#)  
Advertise on New Radio 610 WFTN [Get Results](#) [Contact](#) [Philandropy](#)  
© 2017 iHeartMedia, Inc.

Listen Now on iHeartRadio



[Terms of Use](#) [Privacy Statement](#) [Copyright and Trademark Notice](#) [EEO Public File](#) [Public Inspection File](#)

iHeartRadio - Get the Free Radio App [General Content Guidelines](#)  
Advertise on New Radio 610 WFTN [Get Results](#) [Contact](#) [Philandropy](#)  
© 2017 iHeartMedia, Inc.

Listen Now on iHeartRadio

# Web Page Segmentation



## 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](#)



## 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 WFTV](#) [Get Results](#) [Contact](#) [Philandropy](#)  
© 2017 iHeartMedia, Inc.

Listen Now on [iHeartRADIO](#)



[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 WFTV](#) [Get Results](#) [Contact](#) [Philandropy](#)  
© 2017 iHeartMedia, Inc.

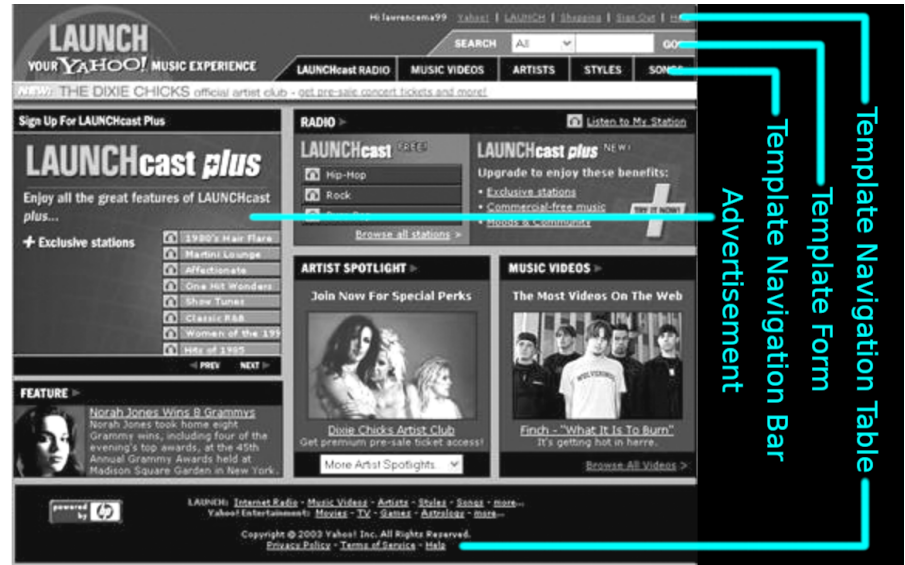
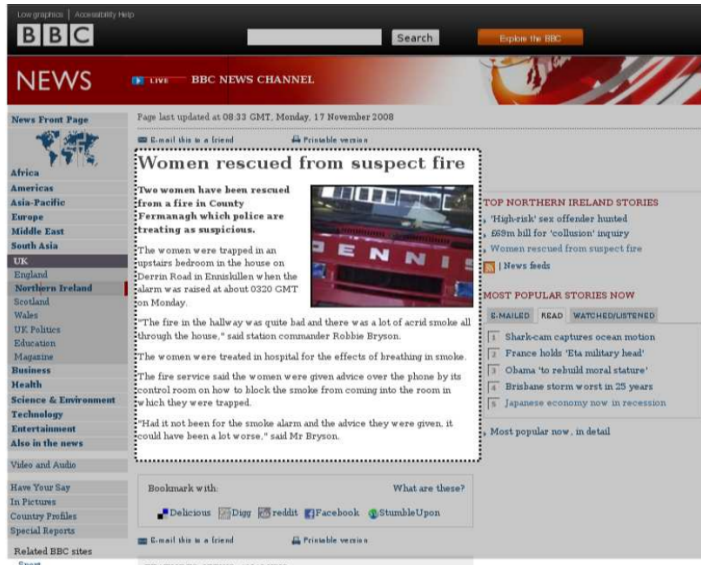
Listen Now on [iHeartRADIO](#)

# Web Page Segmentation: Downstream Tasks (Examples)



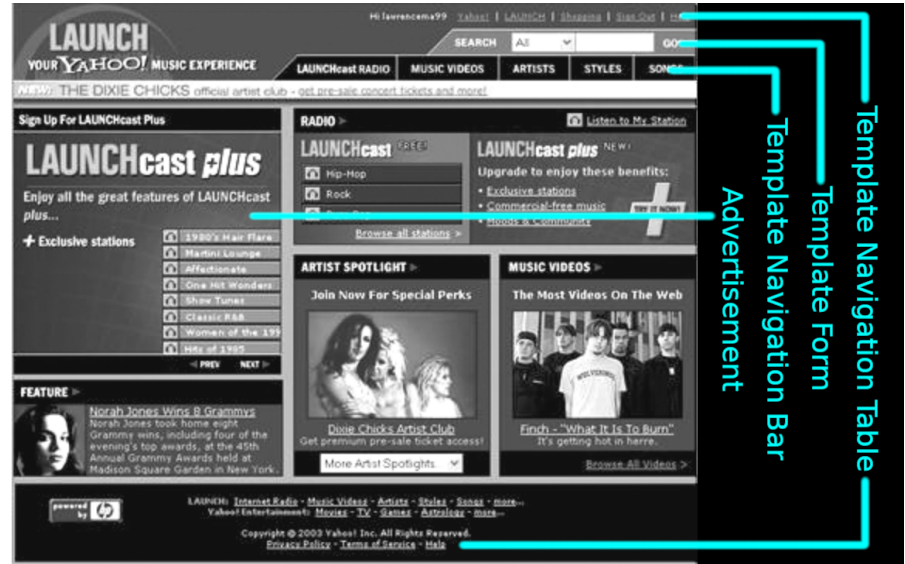
- ❑ **Content Extraction**  
Image: *Language Independent Content Extraction from Web Pages*. Javier et al., DIR'09.

# Web Page Segmentation: Downstream Tasks (Examples)



- ❑ Content Extraction  
Image: *Language Independent Content Extraction from Web Pages*. Javier et al., DIR'09.
- ❑ Template Detection  
Image: *Automatic Data Extraction From Template Generated Web Pages*. Ma et al., PDPTA'03.

# Web Page Segmentation: Downstream Tasks (Examples)



## Content Extraction

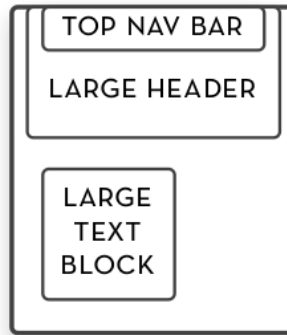
Image: *Language Independent Content Extraction from Web Pages*. Javier et al., DIR'09.

## Template Detection

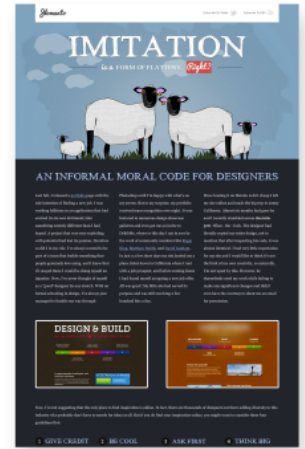
Image: *Automatic Data Extraction From Template Generated Web Pages*. Ma et al., PDPTA'03.

## Design Mining

Image: *Webzeitgeist: Design Mining the Web*. Kumar et al., CHI'13.



LAYOUT QUERY



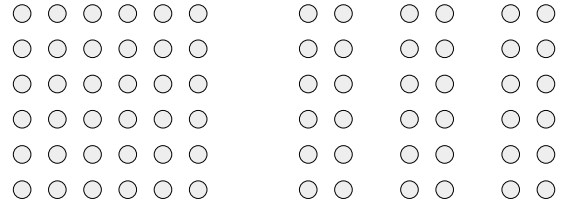


# Concept Formation: Web Page Segment

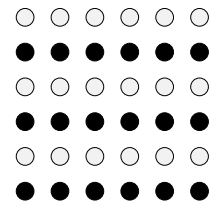
A web page segment is a part of a web page containing those elements that belong together as per agreement among a majority of viewers.

Rationale: Web pages are created for human viewers, and so are segments

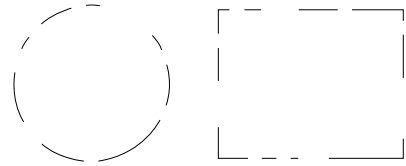
Gestalt Principles provide common ground



Proximity



Similarity



Closure



Symmetry

# Evaluation Framework for Web Page Segmentation

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  $E = \{e_1, \dots, e_n\}$

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

# Evaluation Framework for Web Page Segmentation

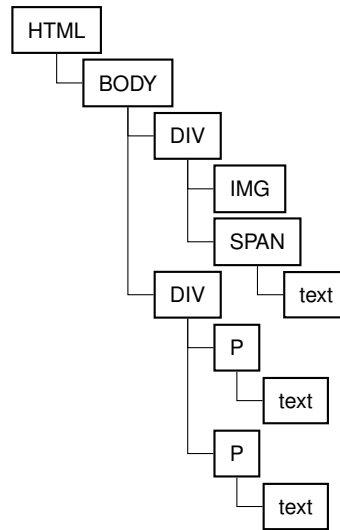
A web page segment is a part of a web page containing those elements that belong together as per agreement among a majority of viewers.

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

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

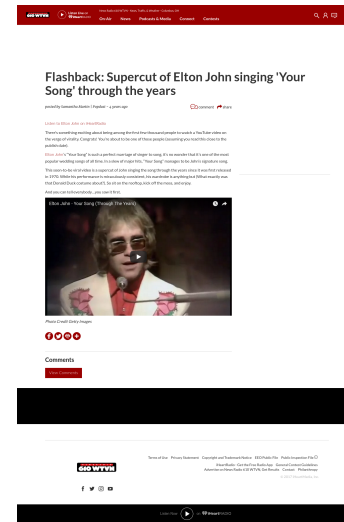
Suggested sets of elements:

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

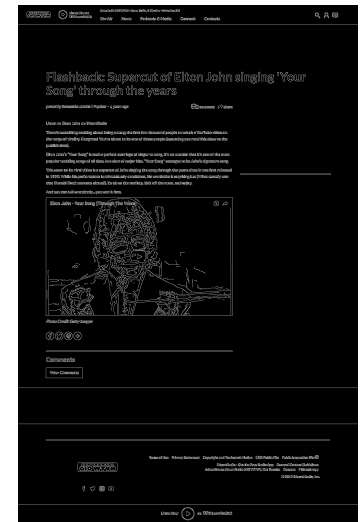


Characters

DOM nodes



Pixels



Edges

# Evaluation Framework for Web Page Segmentation

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  $E = \{e_1, \dots, e_n\}$

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

Precision

$$P_{B^3}(S, S^*) = \text{avg}_e \left( \frac{|\text{elements in same segment as } e \text{ in both } S \text{ and } S^*|}{|\text{elements in same segment as } e \text{ in } S|} \right)$$

Recall

$$R_{B^3}(S, S^*) = \text{avg}_e \left( \frac{|\text{elements in same segment as } e \text{ in both } S \text{ and } S^*|}{|\text{elements in same segment as } e \text{ in } S^*|} \right)$$

F-Measure,  $F_{B^3}$ , is defined as the harmonic mean of precision and recall as usual

Note:  $P_{B^3}(S, S') = R_{B^3}(S', S) \Rightarrow F_{B^3}(S, S') = F_{B^3}(S', S)$

# Evaluation Framework for Web Page Segmentation

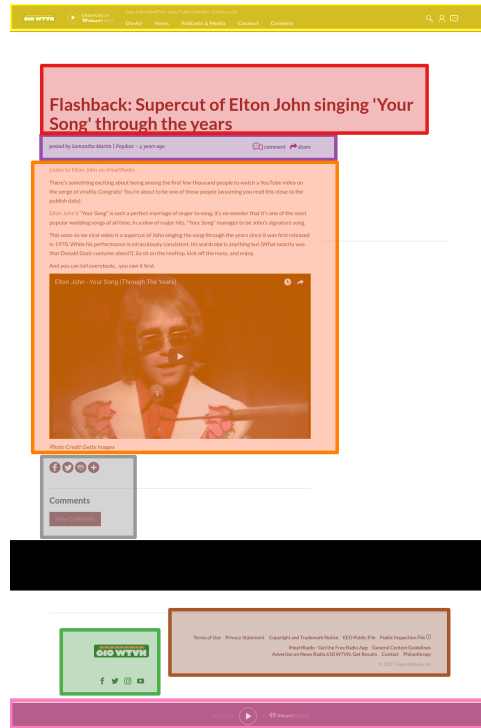
The screenshot shows a BBC News article titled "Flashback: Supercut of Elton John singing 'Your Song' through the years". The page is segmented with blue boxes around the following elements:

- The article title and sub-header.
- The author information: "written by Samantha Martin | Popular | a year ago".
- The main text paragraph: "There's something exciting about being among the first few thousand people to watch a YouTube video on the night of Friday, 13 August. You're about to be one of those people (assuming you read this close to the publication date)."
- The video player: "Elton John - 'Your Song' (Through The Years)".
- The social sharing icons: Facebook, Twitter, and Email.
- The comments section: "Comments" and "View Comments".
- The footer: BBC News logo, navigation icons, and legal/terms information.
- The video player controls at the bottom.

?

This screenshot is identical to the one on the left, showing the same BBC News article with the same blue segmentation boxes highlighting the title, author, text, video, social icons, comments, footer, and video player controls.

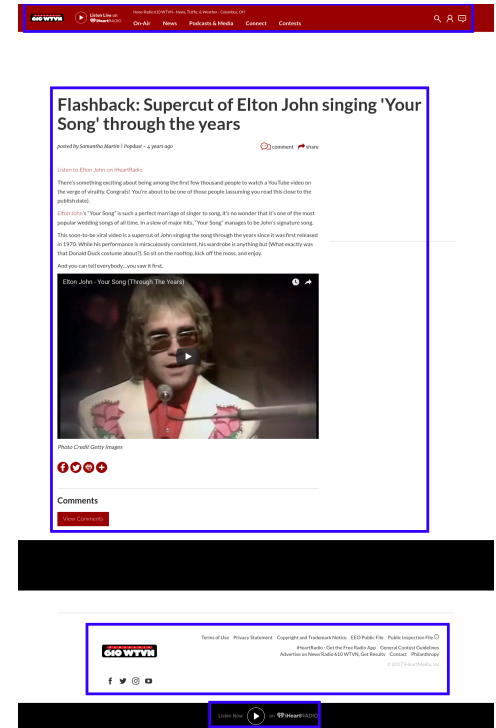
# Evaluation Framework for Web Page Segmentation



$S$



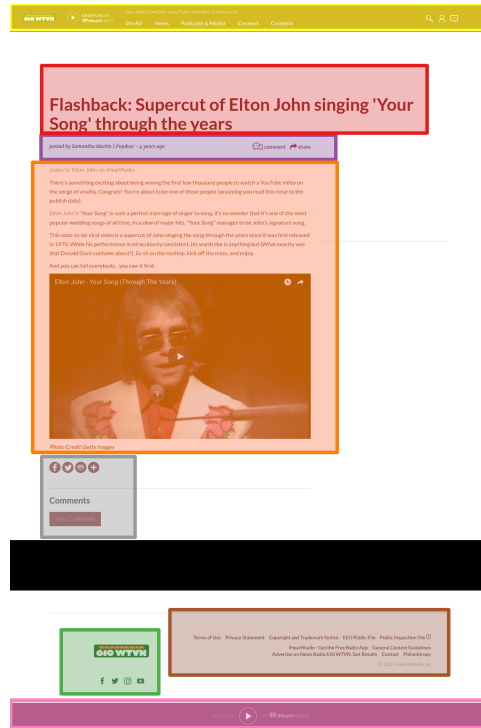
Characters of  $S$  in  $S^*$



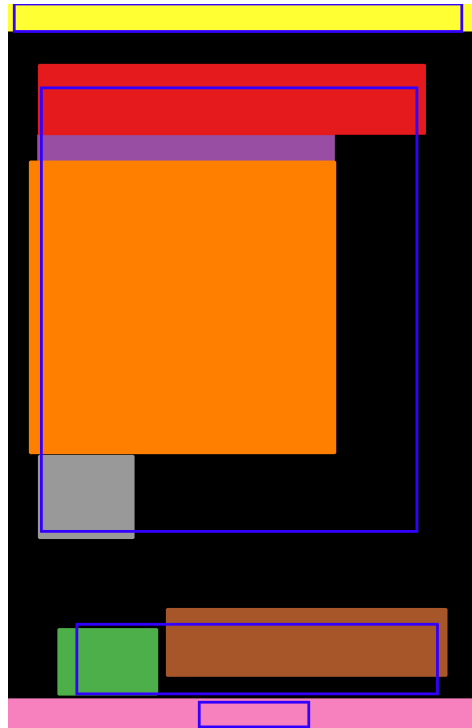
$S^*$

$$P_{B^3}(S, S^*) = \text{avg}_e \left( \frac{|\text{elements in same segment as } e \text{ in both } S \text{ and } S^*|}{|\text{elements in same segment as } e \text{ in } S|} \right)$$

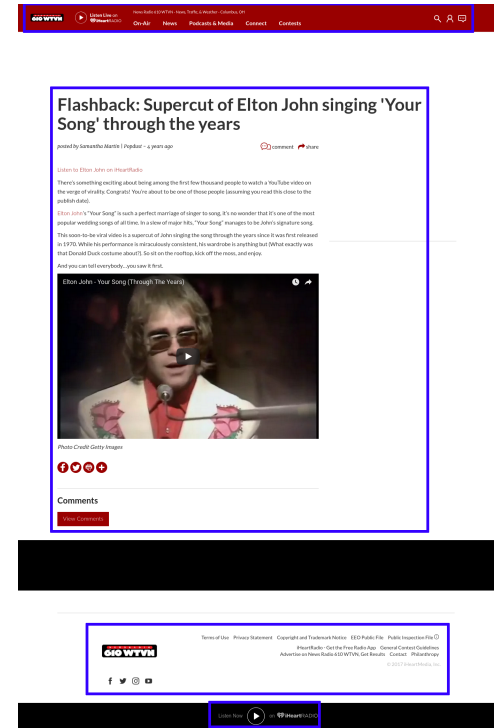
# Evaluation Framework for Web Page Segmentation



$S$



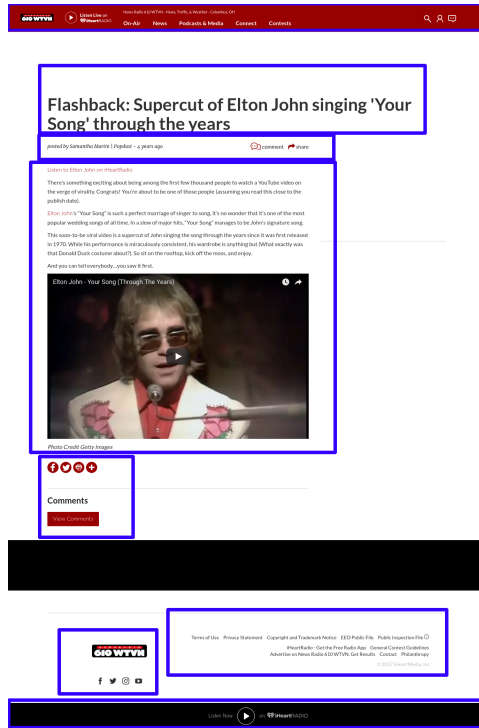
Pixels of  $S$  in  $S^*$



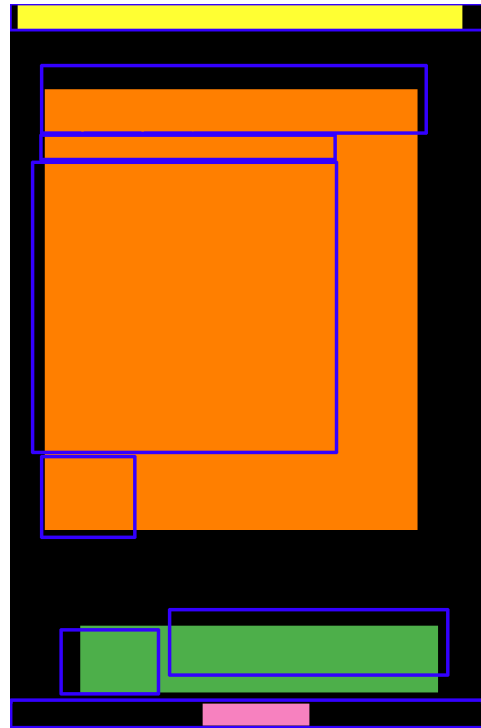
$S^*$

$$P_{B^3}(S, S^*) = \text{avg}_e \left( \frac{|\text{elements in same segment as } e \text{ in both } S \text{ and } S^*|}{|\text{elements in same segment as } e \text{ in } S|} \right)$$

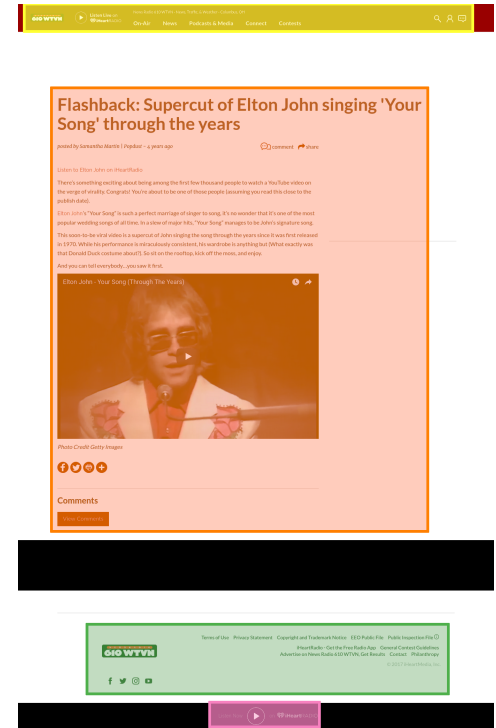
# Evaluation Framework for Web Page Segmentation



$S$



Pixels of  $S^*$  in  $S$



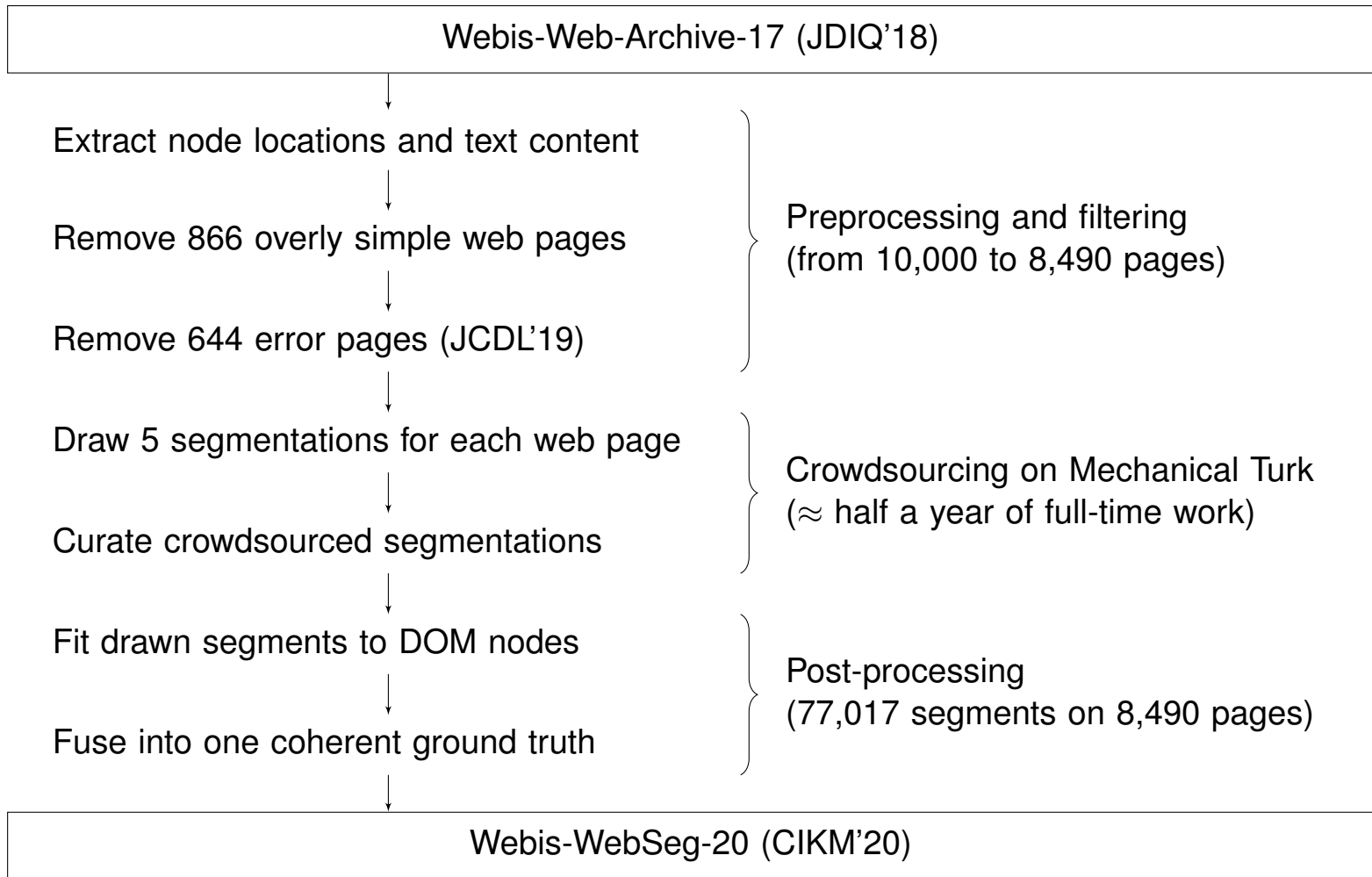
$S^*$

$$P_{B^3}(S, S^*) = \text{avg}_e \left( \frac{|\text{elements in same segment as } e \text{ in both } S \text{ and } S^*|}{|\text{elements in same segment as } e \text{ in } S|} \right)$$

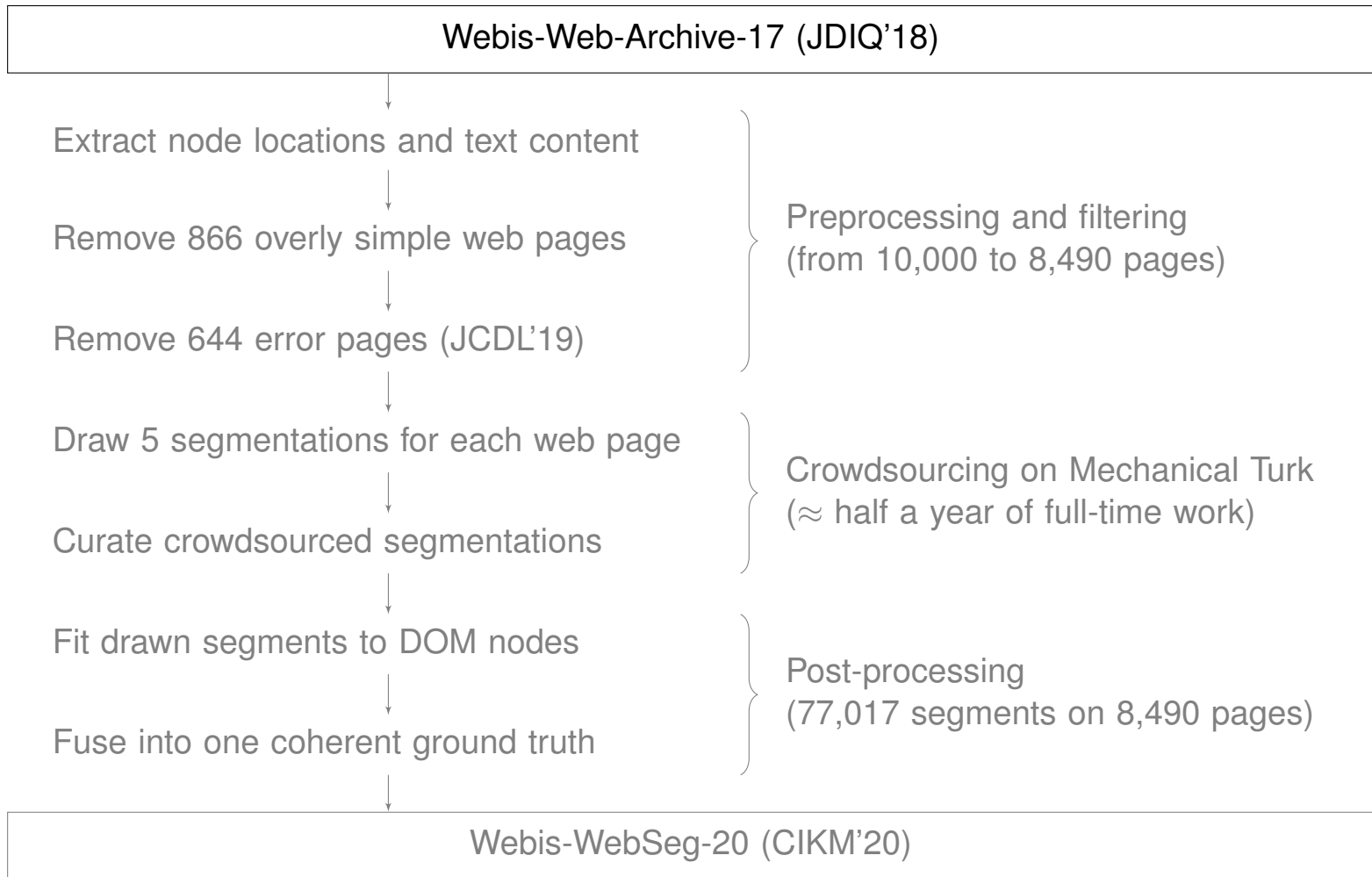
$$R_{B^3}(S, S^*) = \text{avg}_e \left( \frac{|\text{elements in same segment as } e \text{ in both } S \text{ and } S^*|}{|\text{elements in same segment as } e \text{ in } S^*|} \right)$$



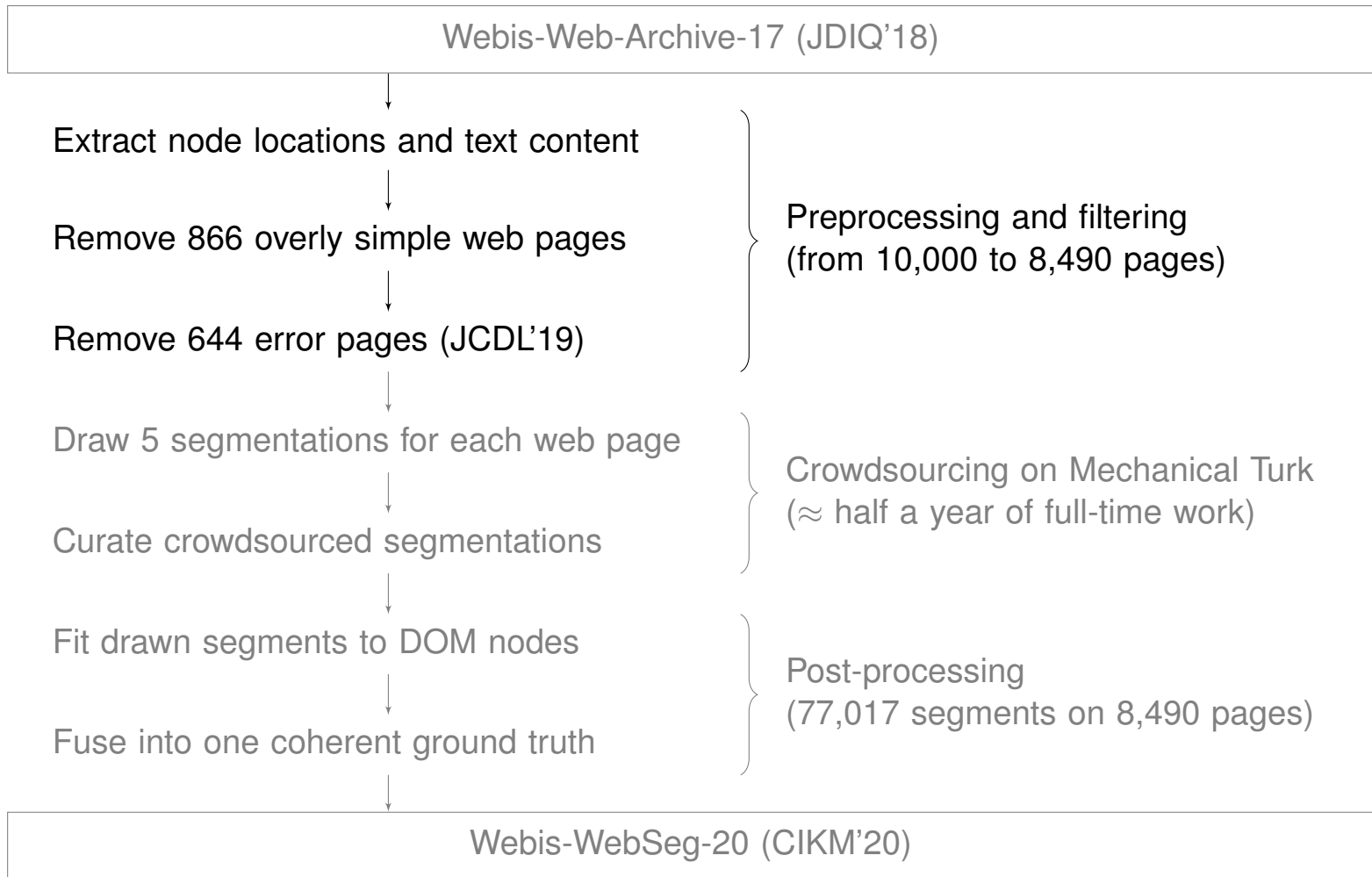
# The Webis-WebSeg-20 Dataset



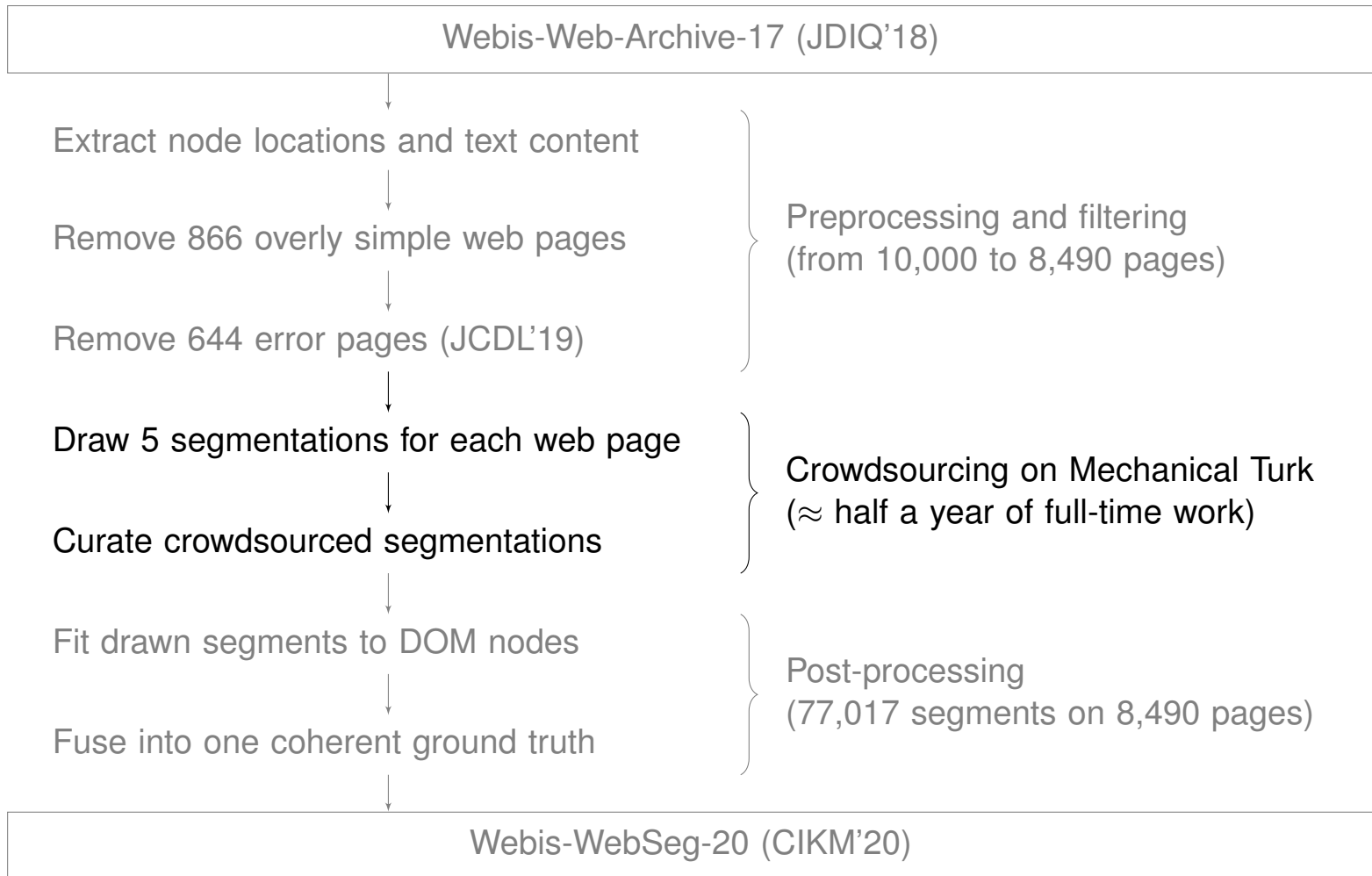
# The Webis-WebSeg-20 Dataset



# The Webis-WebSeg-20 Dataset



# The Webis-WebSeg-20 Dataset



# The Webis-WebSeg-20 Dataset

Webis-Web-Archive-17 (JDIQ'18)

Extract node locations and text content

The screenshot shows the homepage of the Vietnamese newspaper website VietBao.com. The page features a search bar at the top left, a navigation menu with categories like 'Chính Trị', 'Thể Giới', and 'Xã Hội', and several content blocks. A prominent block in the center contains three articles with images and headlines. A red circle highlights a specific node in the 'SỰ KIẾN NÔNG' section, which is a video player for a story titled 'Bạn gái cặp với người nước ngoài vì tiền'. The video player includes social media sharing icons, a 'Subscribe' button, and a YouTube logo. To the right, there is a 'ĐỌC NHIỀU' (Most Read) section with several article thumbnails. The page also displays a currency exchange table and a lottery announcement.

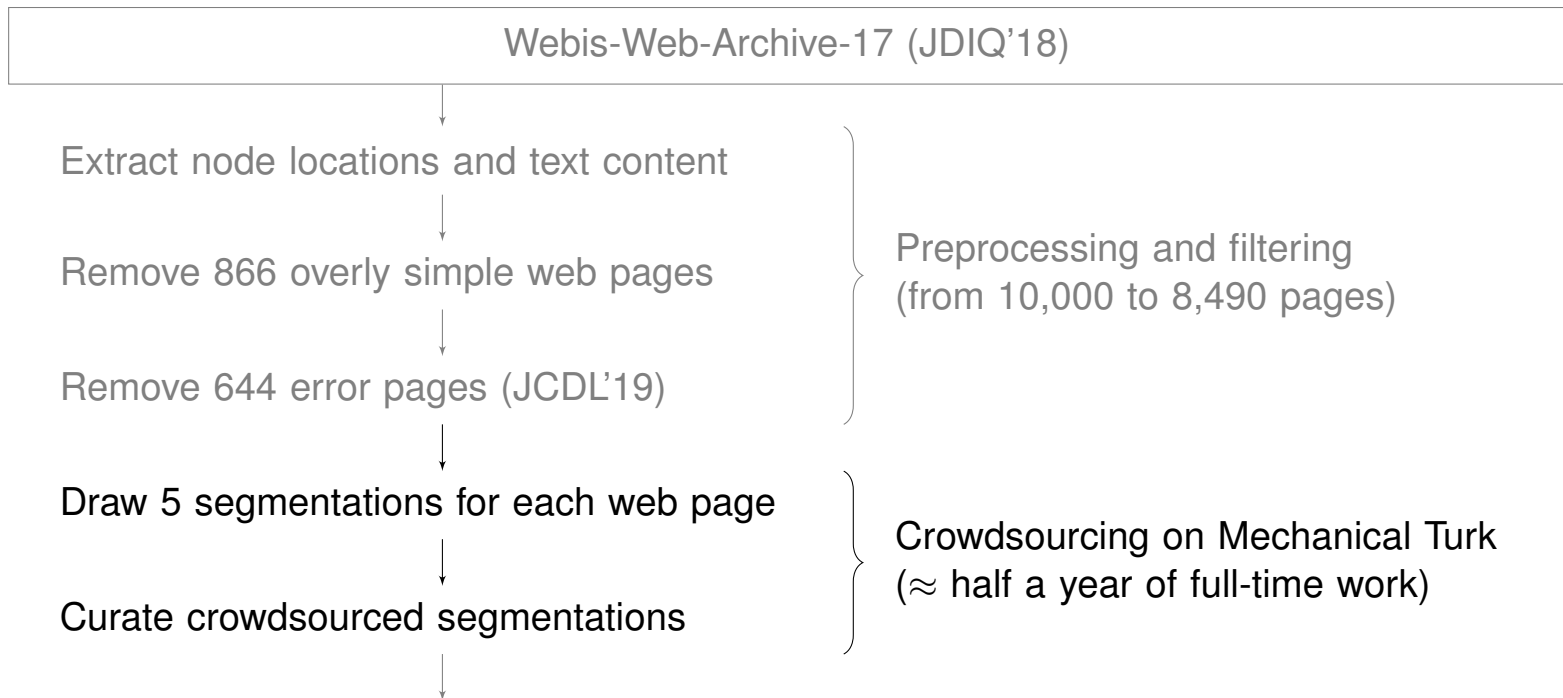
Loại Vàng	Mua	Bán	Ty Giá	VND
Vàng SJC	36.48	36.7	EUR Euro	27.191.14
Vàng nữ trang 24 K	35.42	36.12	JPY Yên Nhật	202.67
			USD Đô Mỹ	22.780.00

**SUPER!**  
Das ist kein Scherz! Sie sind unser 1.000.000ster Besucher!  
Unser Zufallssystem der möglichen Gewinner könnte Sie als möglichen Gewinner von FANTASTISCHEN APPLE Produkten ziehen.  
ONLINE: 21/09/2017 22:40  
[Klicken Sie hier](#)

**SỰ KIẾN NÔNG**  
**Bạn gái cặp với người nước ngoài vì tiền**  
Thứ năm, 27 Tháng Năm 2009, 09:36 GMT+7  
TAGS: Việt Nam người nước ngoài ở nước ngoài mối quan hệ cô ấy bạn gái tình cảm có thể đi làm yêu thương tiền tháng nói lời việc  
Bấm ngay [Subscribe](#) / Đăng Ký xem video hay mới nhất >>

**ĐỌC NHIỀU**  
Đám Vinh Hưng bất ngờ lên tiếng vụ việc của danh hài Xuân Bắc  
Bảo Thanh khoe nhà sang tiền tỷ ở vị trí đắc địa  
Phương Trinh Jolie tiết lộ chuyện tình yêu lãng mạn của mình

# The Webis-WebSeg-20 Dataset

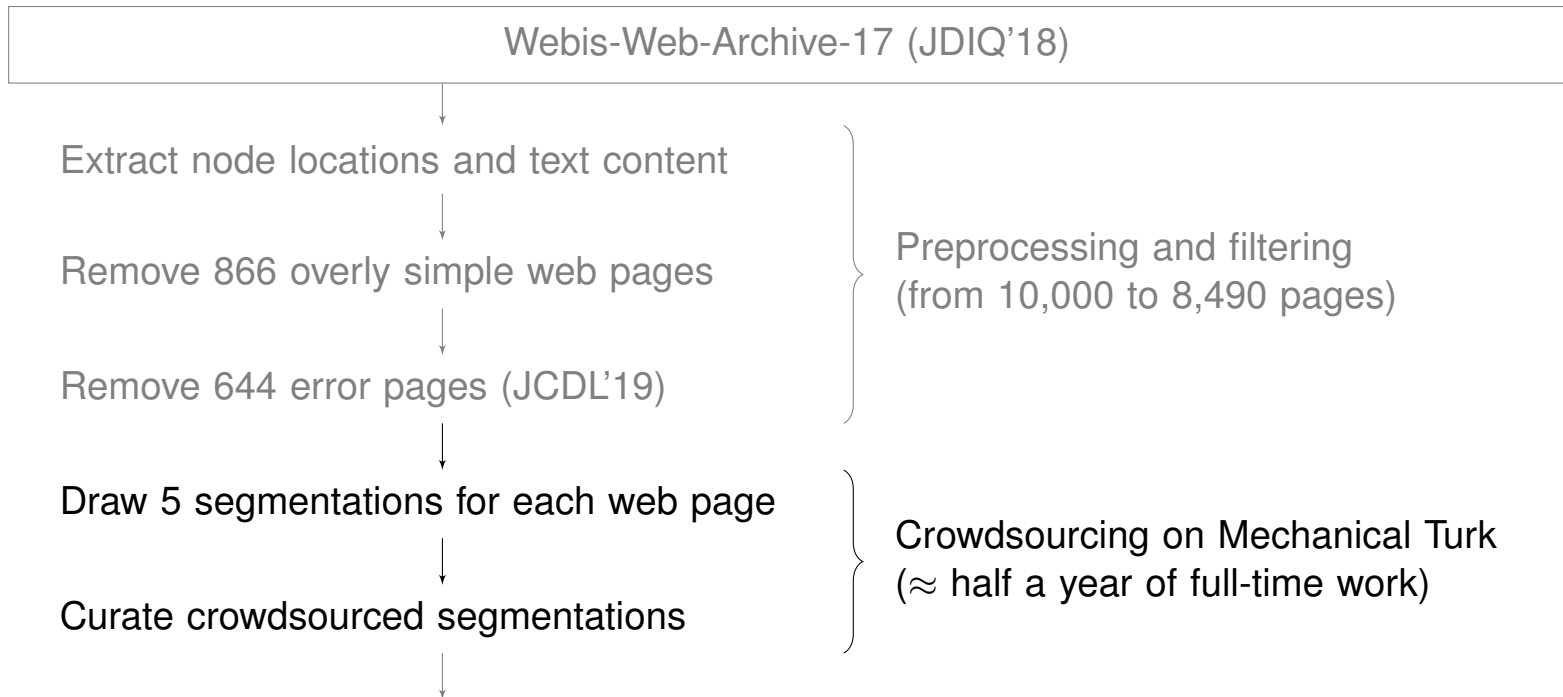


## Agreement measure

## Elements

	Characters	Nodes	Pixels	Edges
$F_{B^3}$	0.78	0.74	0.65	0.73

# The Webis-WebSeg-20 Dataset

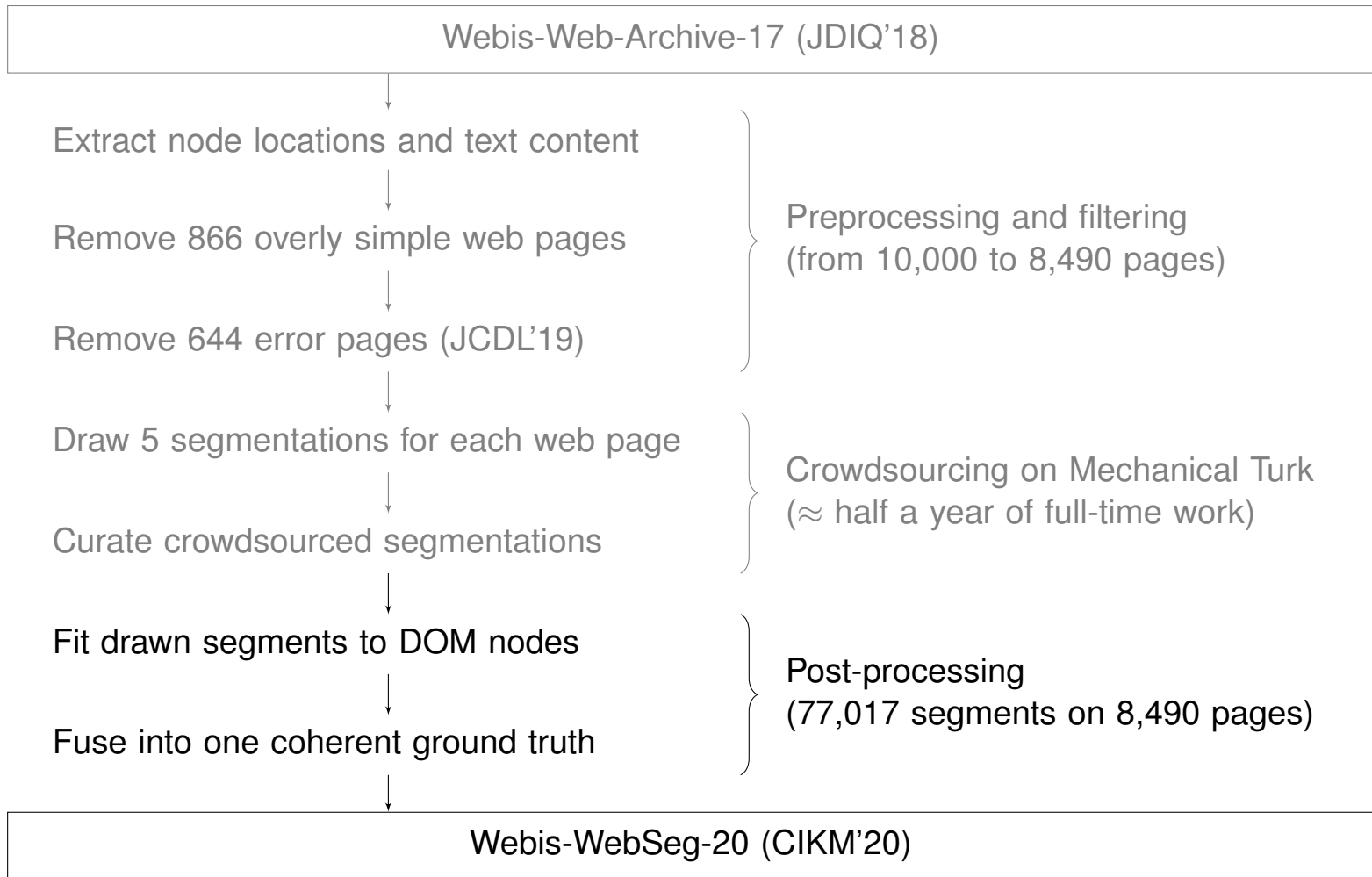


## Agreement measure

## Elements

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

# The Webis-WebSeg-20 Dataset





# Web Page Segmentation Revisited: Evaluation Framework and Dataset

## Evaluation Framework for Web Page Segmentation

- ❑ Segmentation similarity, quality, and fusion
- ❑ Comparison of level of detail of segmentations
- ❑ Adjustable for different downstream tasks

## Webis-WebSeg-20

- ❑ 8,490 pages from 4,824 sites
- ❑ 5 human annotators each page
- ❑ Segments in “Simple Feature Access” standard
- ❑ Web pages provided in several representations:
  - HTML file
  - Screenshot
  - Screenshot coordinates of DOM nodes
  - Webis-Web-Archive-17 WARC file



[https://webis.de/publications.html?q=johannes+kiesel+web+archive#stein\\_2020w](https://webis.de/publications.html?q=johannes+kiesel+web+archive#stein_2020w)

Paper, browser, code, data