

An Empirical Comparison of Web Page Segmentation Algorithms

ECIR 2021



**Johannes
Kiesel¹**



**Lars
Meyer¹**



**Florian
Kneist¹**



**Benno
Stein¹**



**Martin
Potthast²**

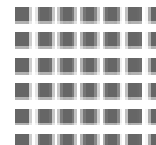


¹



UNIVERSITÄT
LEIPZIG

²



Webis

^{1,2}

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.

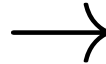


Photo Credit Getty Images



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 Content Guidelines](#)
Advertise on News Radio 430 WTVN, Get Results [Contact](#) [Philadelphia](#)
© 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 News Radio 430 WTVN, Get Results [Contact](#) [Philadelphia](#)
© 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 @HeartRadio

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 @HeartRadio

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 News Radio 430 WTTM, Get Results](#) [Contact](#) [Philanthropy](#)
© 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 Content Guidelines](#)
[Advertise on News Radio 430 WTTM, Get Results](#) [Contact](#) [Philanthropy](#)
© 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

Listen to Elton John on @HeartRadio

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

Listen to Elton John on @HeartRadio

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

@HeartRadio - Get the Free Radio App General Content Guidelines
Advertise on News Radio 430 WTTN, Get Results Contact Philadelphia
© 2017 HeartMedia, Inc.



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

@HeartRadio - Get the Free Radio App General Content Guidelines
Advertise on News Radio 430 WTTN, Get Results Contact Philadelphia
© 2017 HeartMedia, Inc.

Listen Now on @HeartRADIO

Listen Now on @HeartRADIO

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 @HeartRadio

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

Listen to Elton John on @HeartRadio

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

@HeartRadio - Get the Free Radio App General Content Guidelines
Advertise on News Radio 430 WTTN, Get Results Contact Philadelphia
© 2017 HeartMedia, Inc.



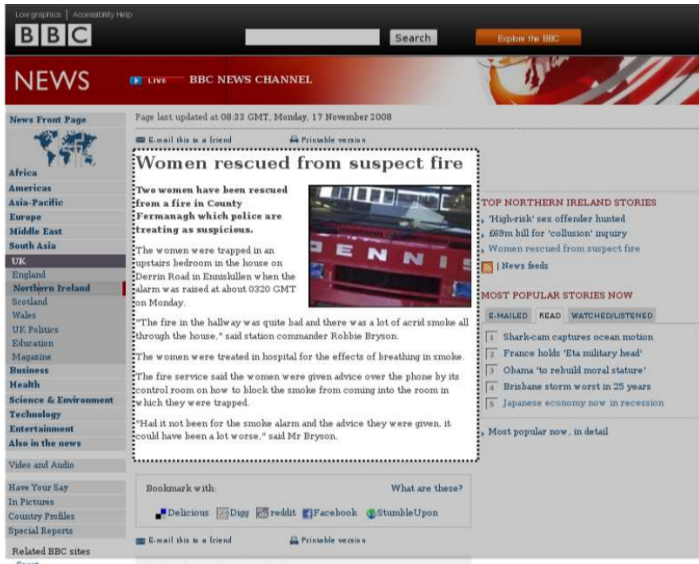
Terms of Use Privacy Statement Copyright and Trademark Notice EEO Public File Public Inspection File

@HeartRadio - Get the Free Radio App General Content Guidelines
Advertise on News Radio 430 WTTN, Get Results Contact Philadelphia
© 2017 HeartMedia, Inc.

Listen Now on @HeartRADIO

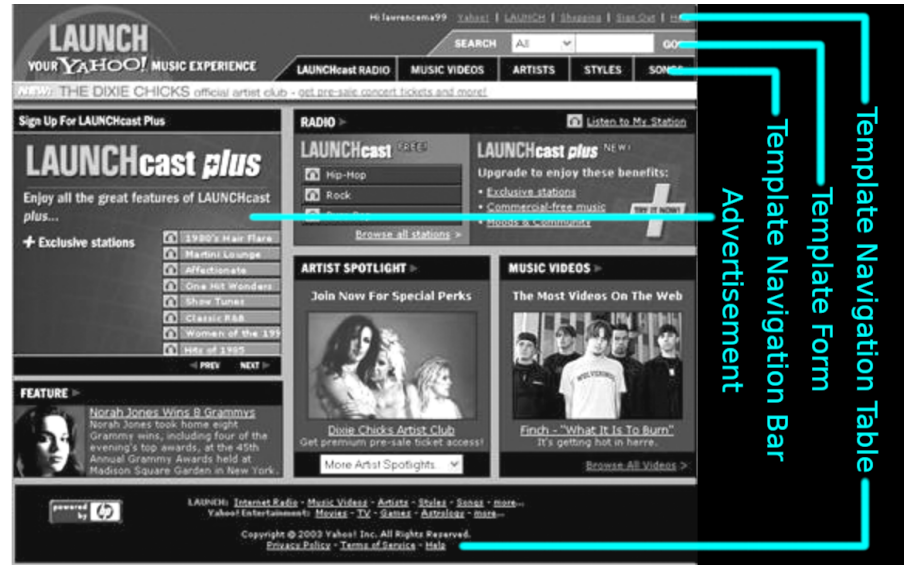
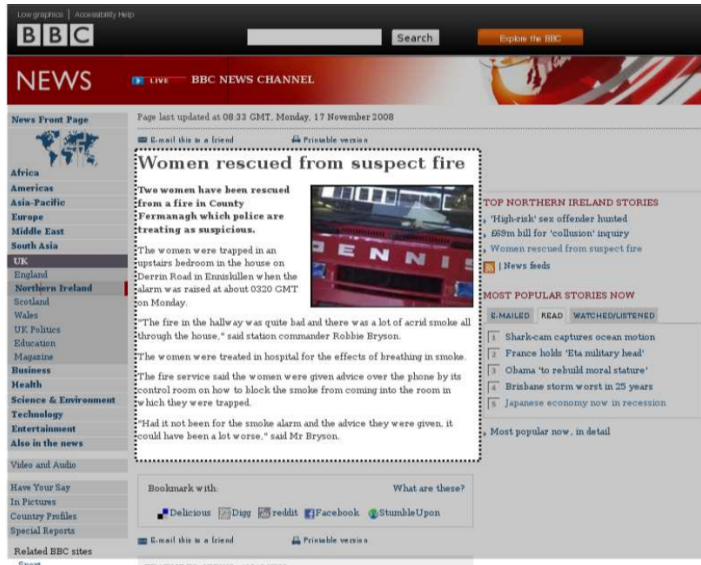
Listen Now on @HeartRADIO

Web Page Segmentation: Downstream Tasks (Examples)



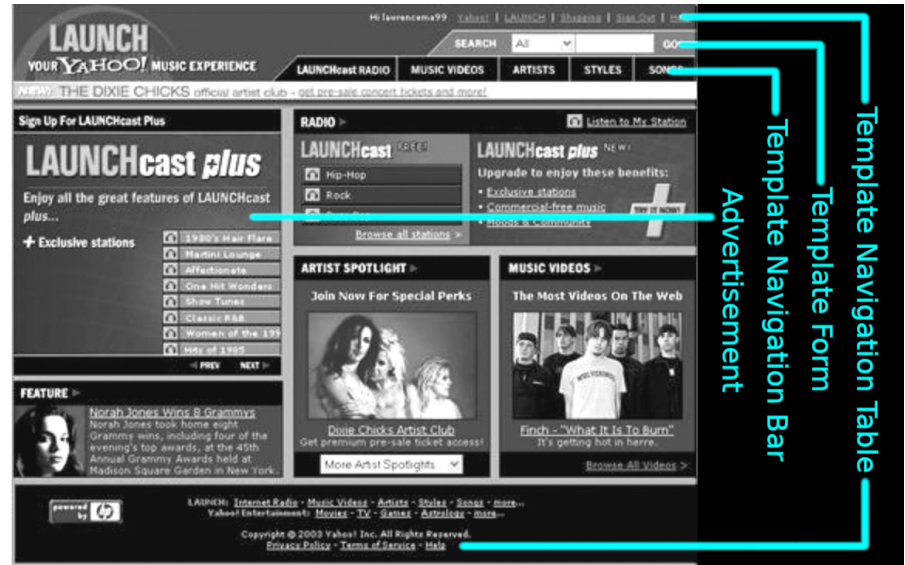
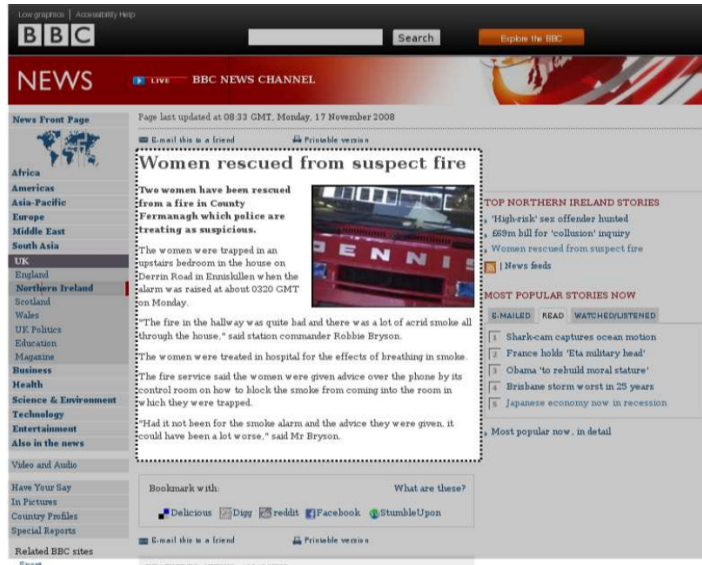
- **Content Extraction**
Image: *Language Independent Content Extraction from Web Pages*. Javier et al., DIR'09.

Web Page Segmentation: Downstream Tasks (Examples)

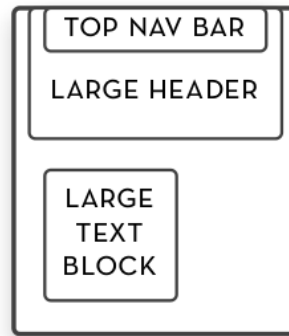


- ❑ Content Extraction
Image: *Language Independent Content Extraction from Web Pages*. Javier et al., DIR'09.
- ❑ Template Detection
Image: *Automatic Data Extraction From Template Generated Web Pages*. Ma et al., PDPTA'03.

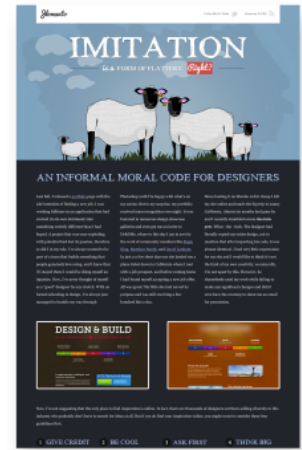
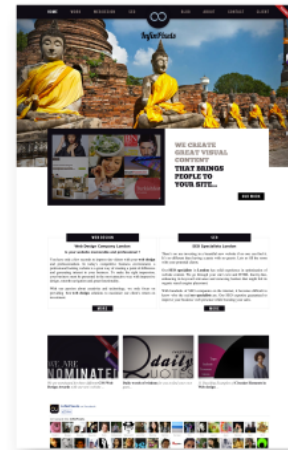
Web Page Segmentation: Downstream Tasks (Examples)



- ❑ **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

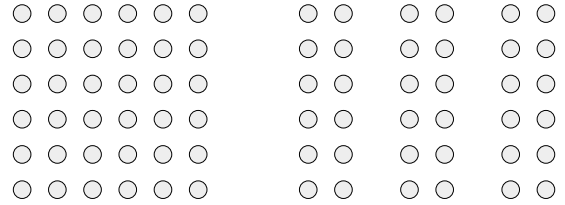


Concept Formation: Web Page Segment

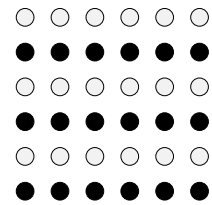
A web page segment is a part of a web page containing those elements that belong together as per agreement among a majority of viewers.

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

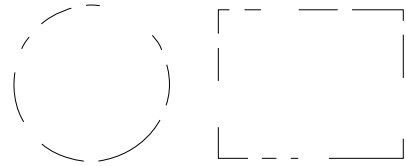
Gestalt Principles provide common ground



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

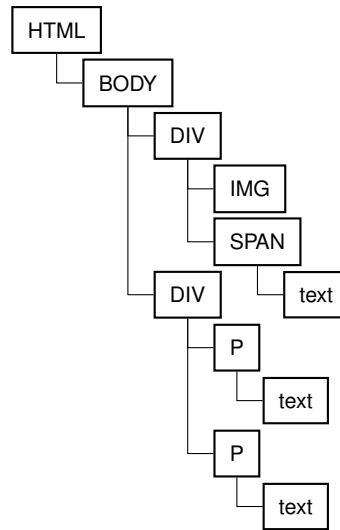
A web page segment is a part of a web page containing those elements that belong together as per agreement among a majority of viewers.

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

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

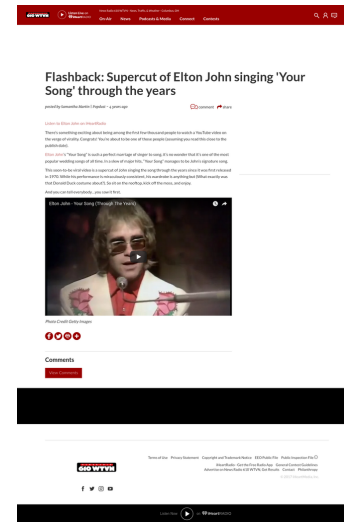
Suggested sets of elements:

Listen Live on iHeartRadio News Radio 610WTVN-News, Traffic, Weather - Columbus, OH 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

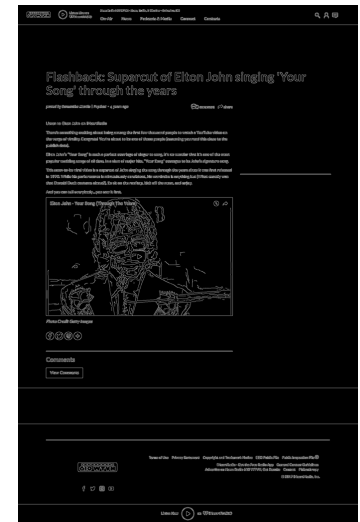


Characters

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

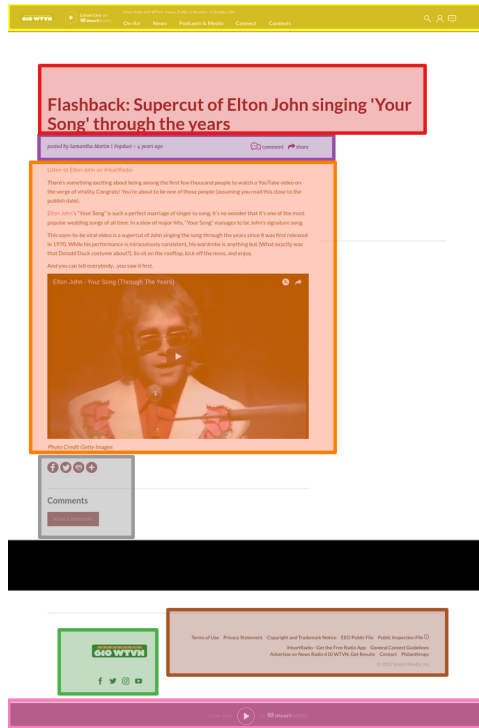
Note: $P_{B^3}(S, S') = R_{B^3}(S', S) \Rightarrow F_{B^3}(S, S') = F_{B^3}(S', S)$

Evaluation Framework for Web Page Segmentation

This screenshot shows a BBC News article titled "Flashback: Supercut of Elton John singing 'Your Song' through the years". The page is annotated with blue boxes highlighting various elements: the article title, the author's name and post time, the introductory text, the video player, the social sharing icons, the comments section, the footer navigation, and the footer text. A central graphic of a stylized '2' with a question mark is positioned between the two screenshots.

This screenshot shows the same BBC News article page as the left one, but with different blue segmentation boxes. The boxes highlight the article title, the author's name and post time, the introductory text, the video player, the social sharing icons, the comments section, the footer navigation, and the footer text. The segmentation highlights differ from the left screenshot, illustrating a different analysis or annotation approach.

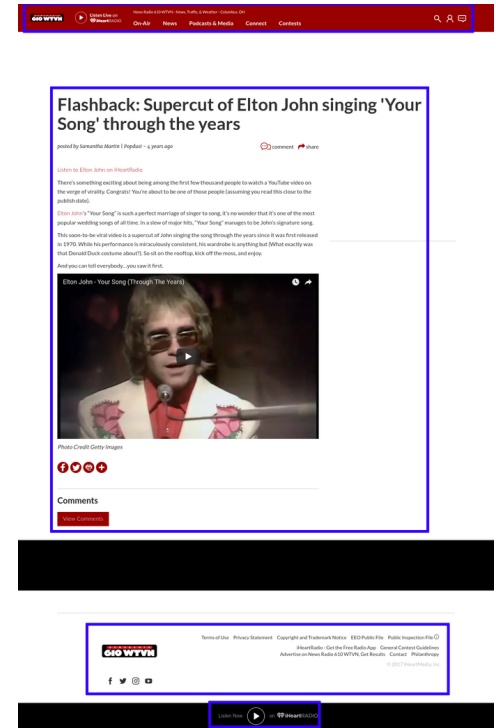
Evaluation Framework for Web Page Segmentation



S



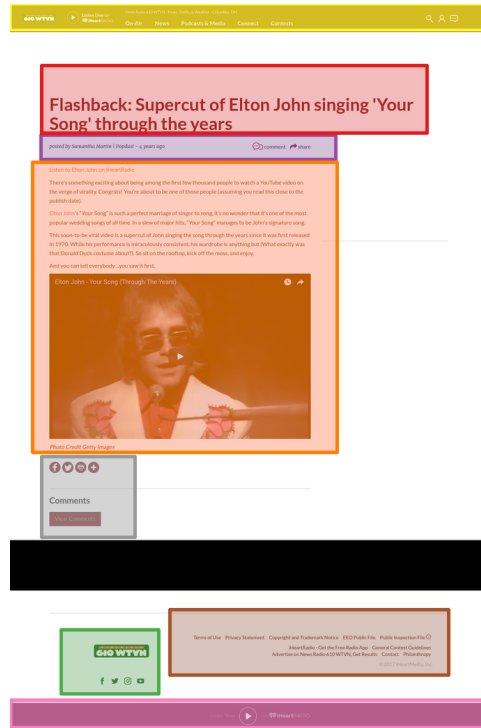
Characters of S in S^*



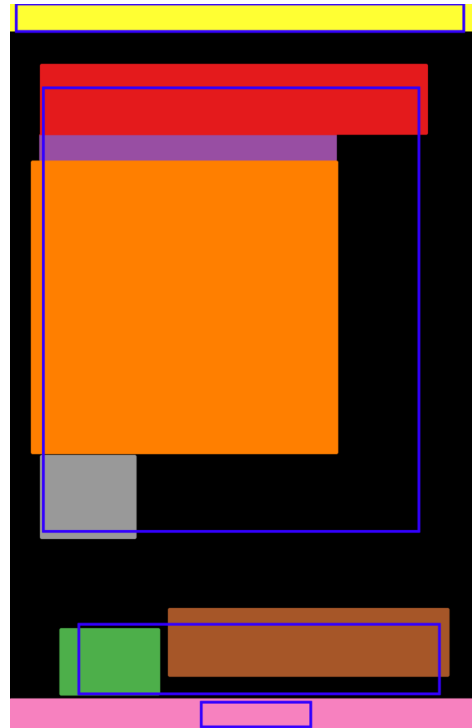
S^*

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

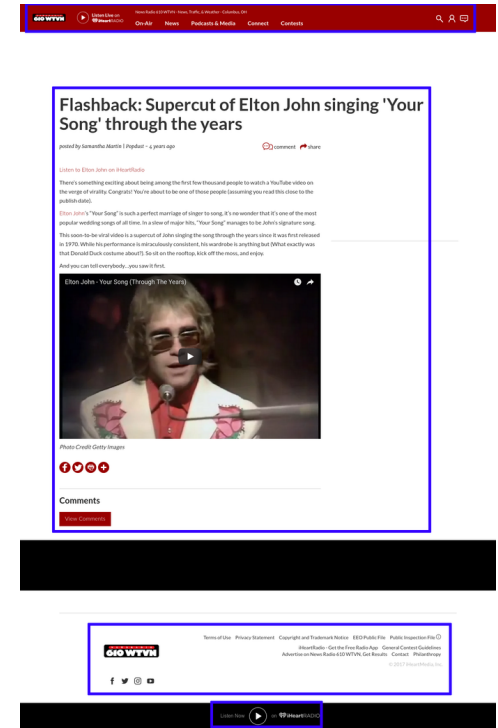
Evaluation Framework for Web Page Segmentation



S



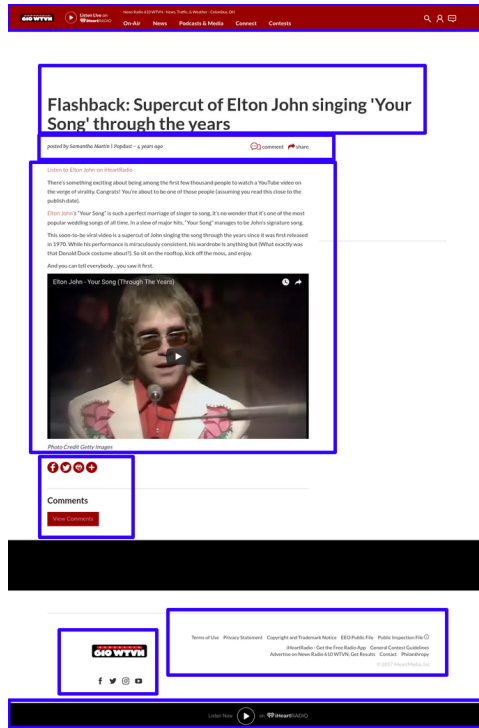
Pixels of S in S^*



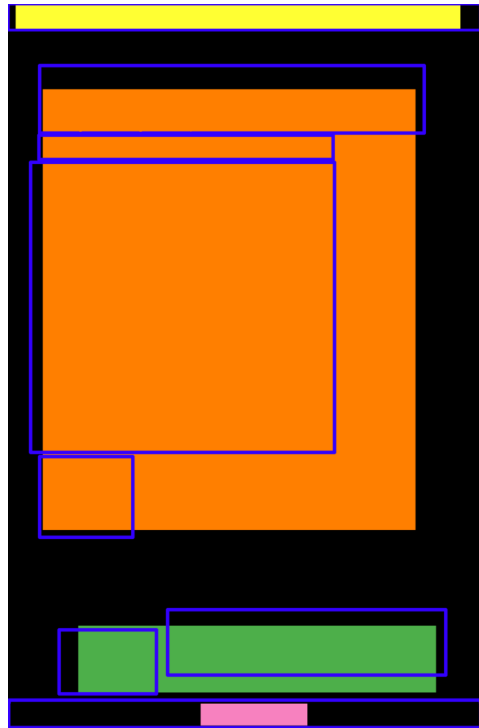
S^*

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

Evaluation Framework for Web Page Segmentation



S



Pixels of S^* in S

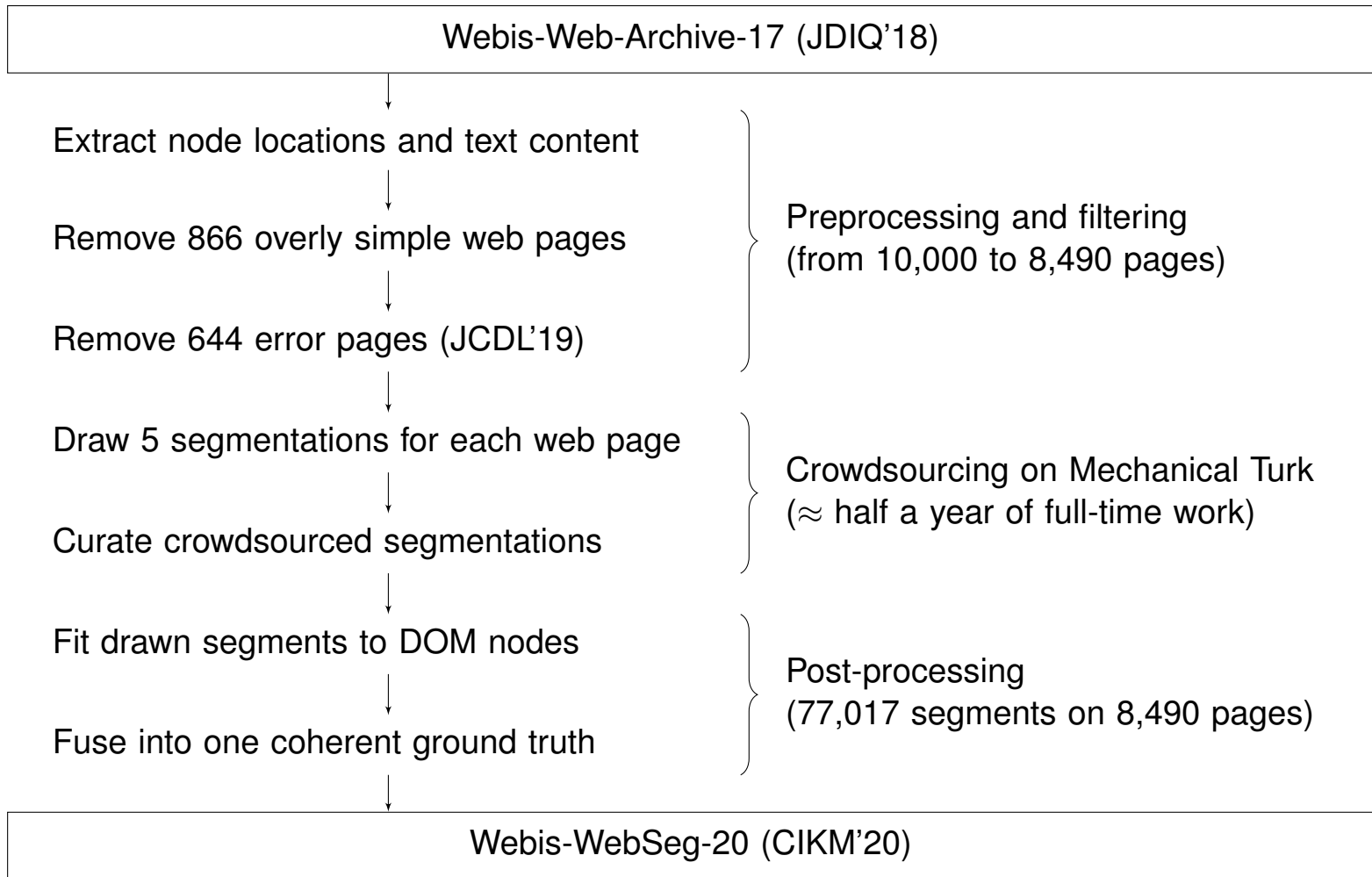


S^*

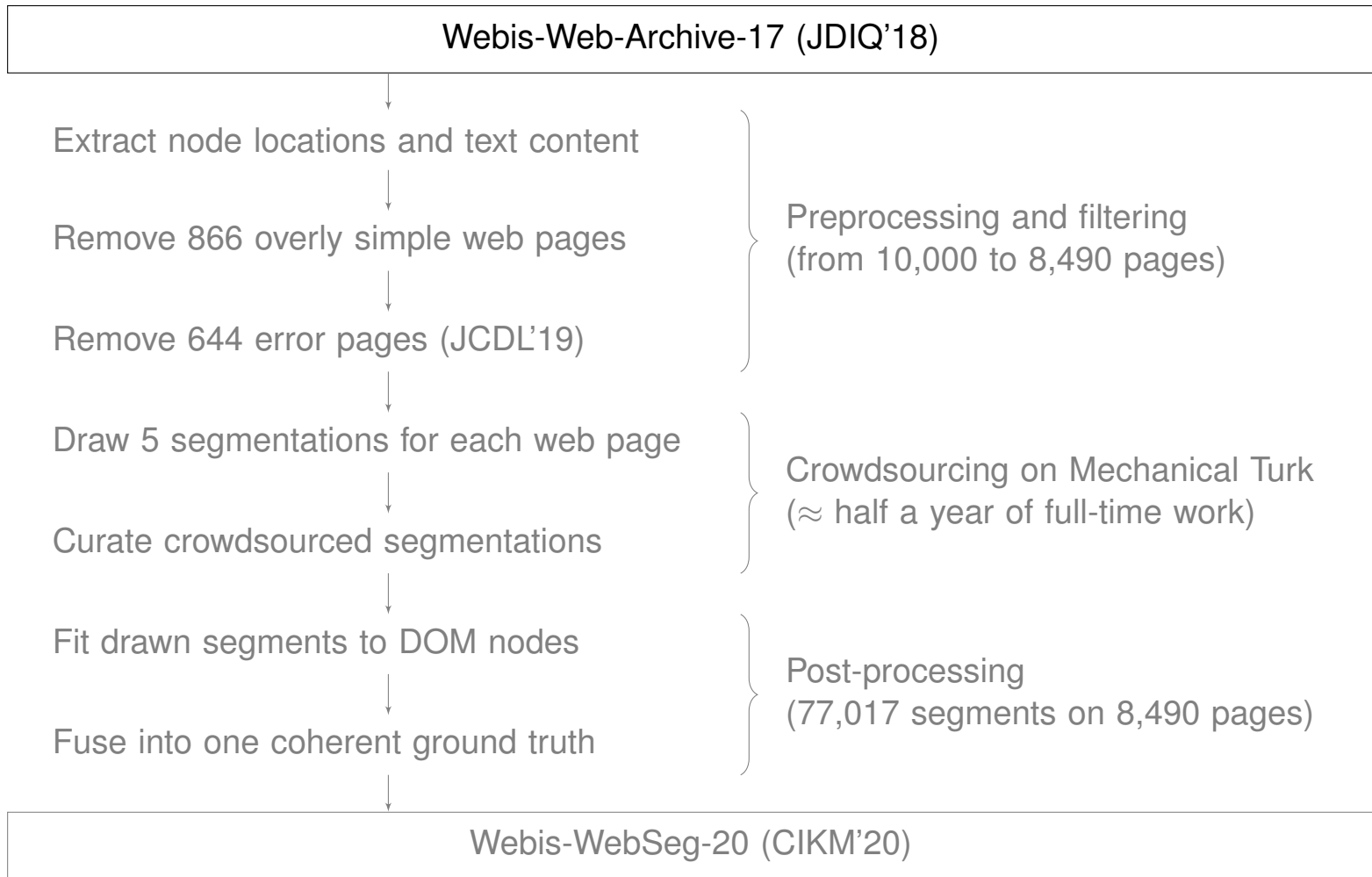
$$P_{B^3}(S, S^*) = \text{avg}_e \left(\frac{|\text{elements in same segment as } e \text{ in both } S \text{ and } S^*|}{|\text{elements in same segment as } e \text{ in } S|} \right)$$

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

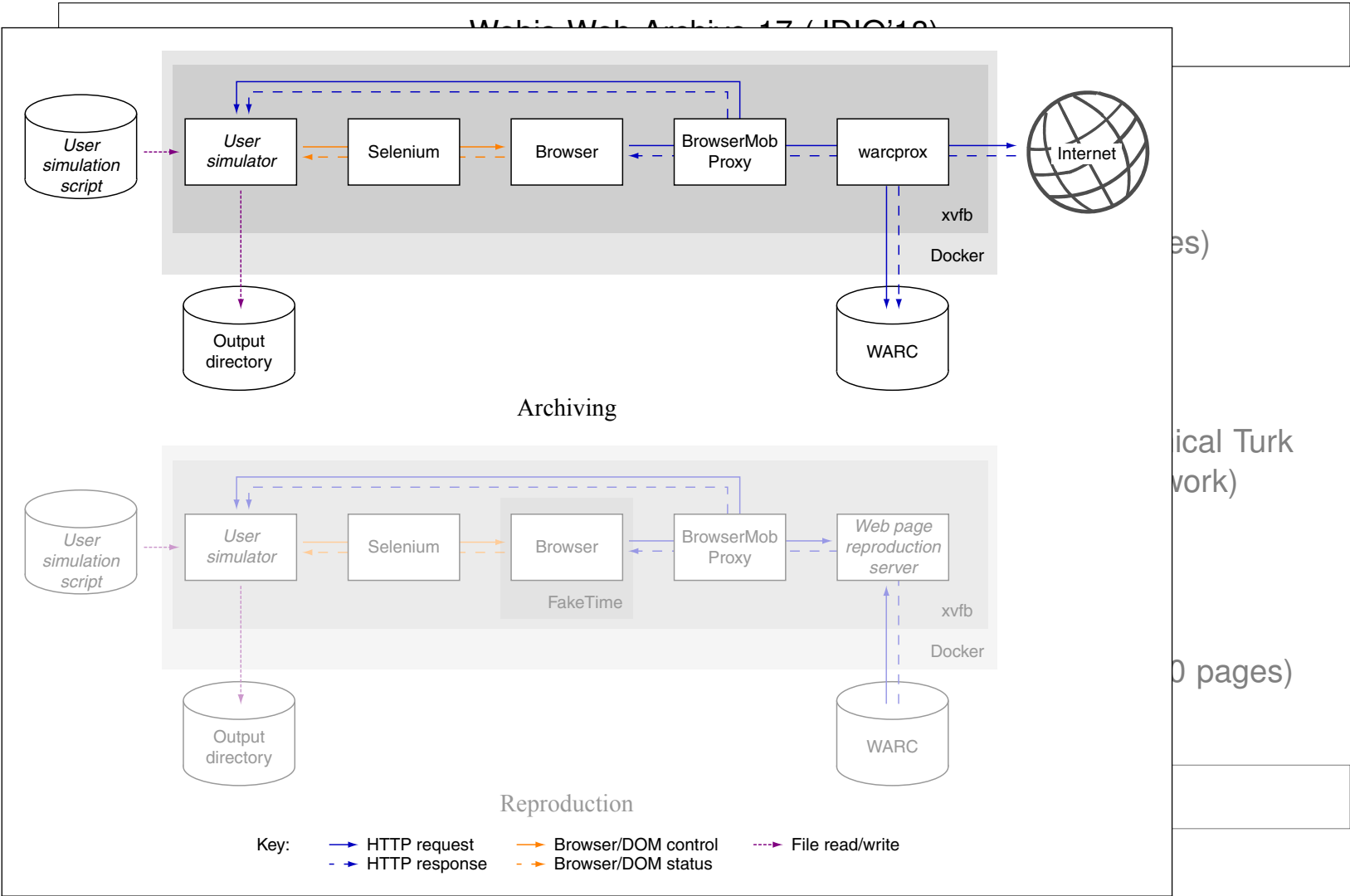
The Webis-WebSeg-20 Dataset



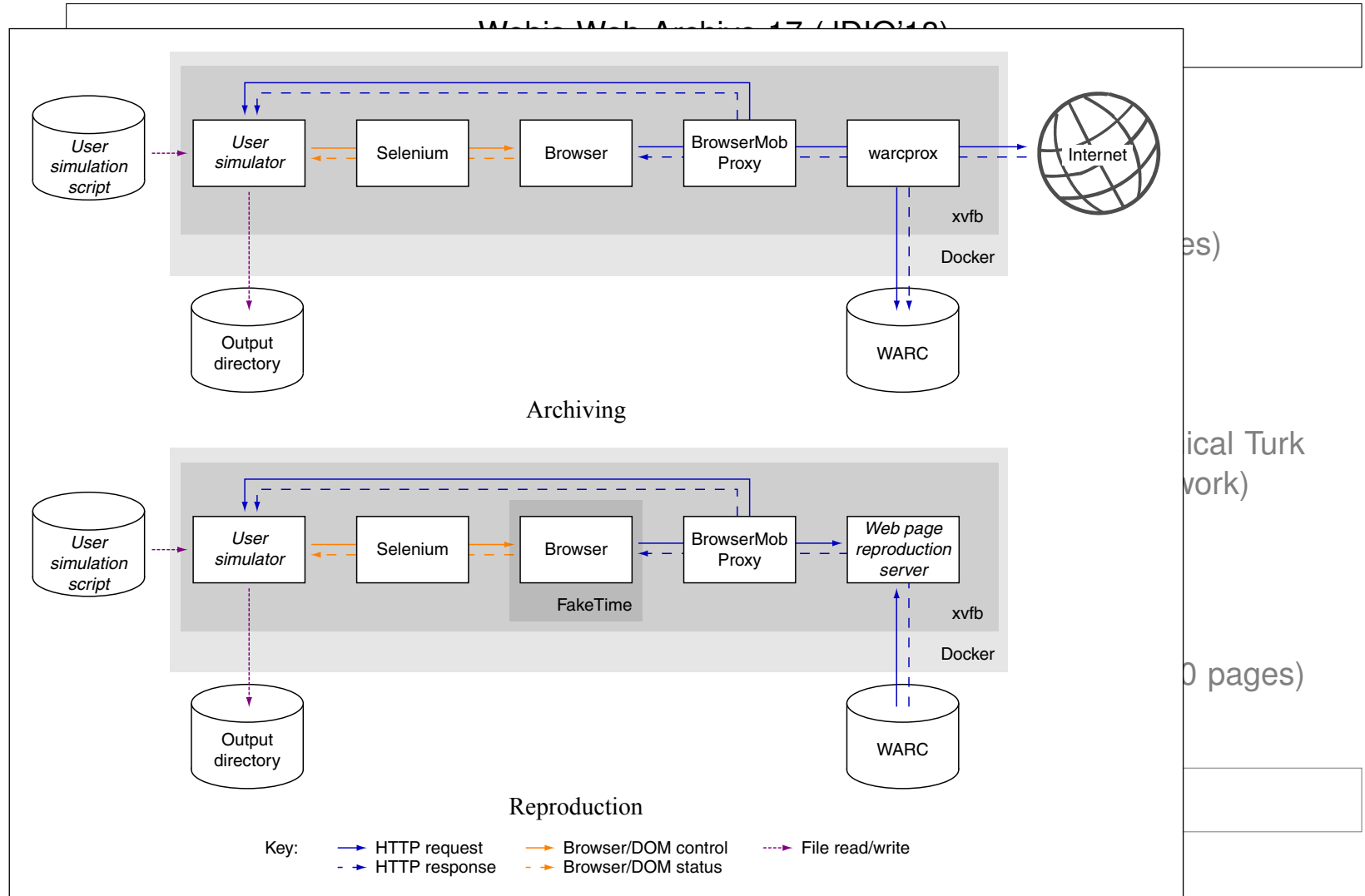
The Webis-WebSeg-20 Dataset



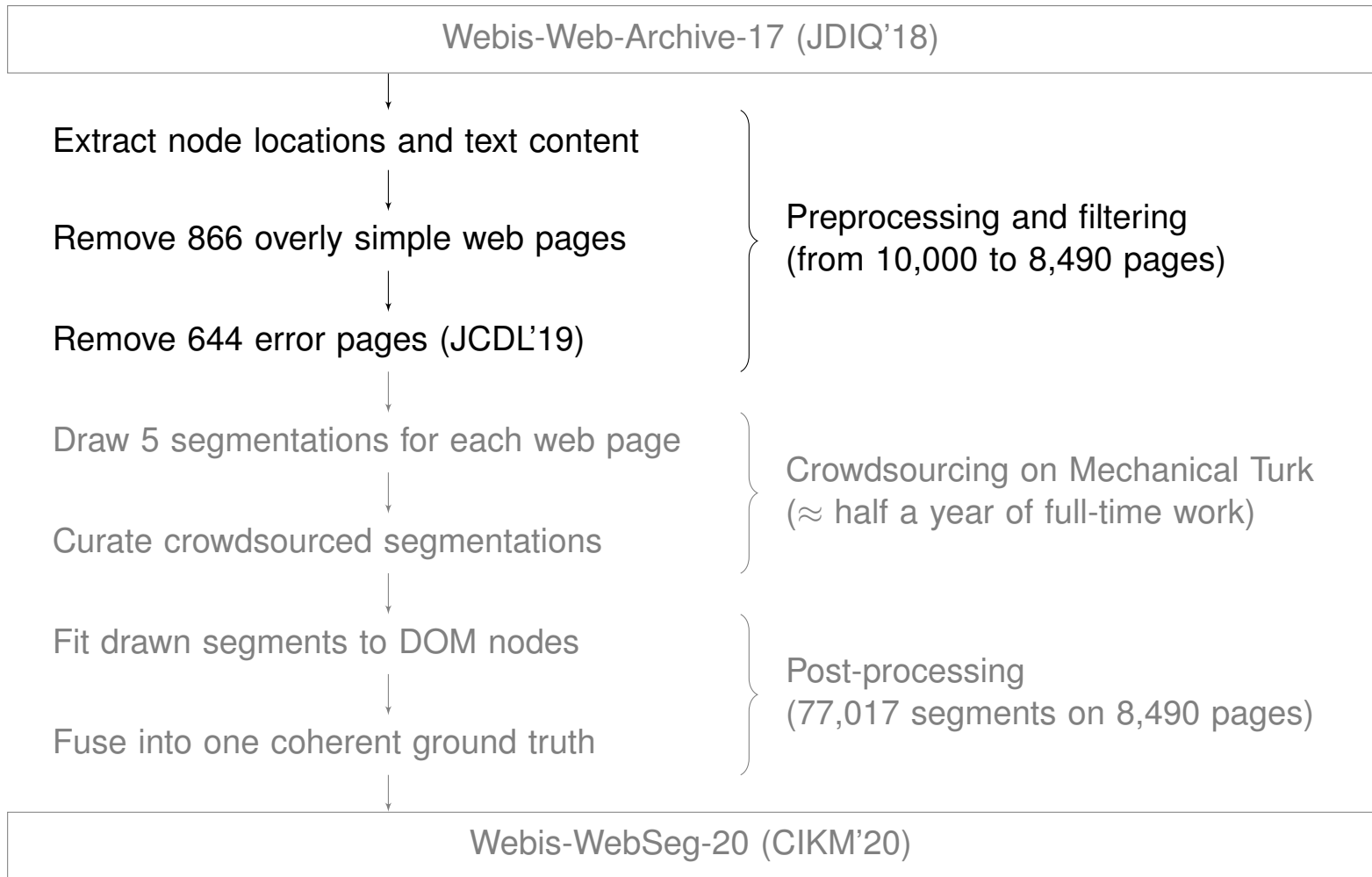
The Webis-WebSeg-20 Dataset



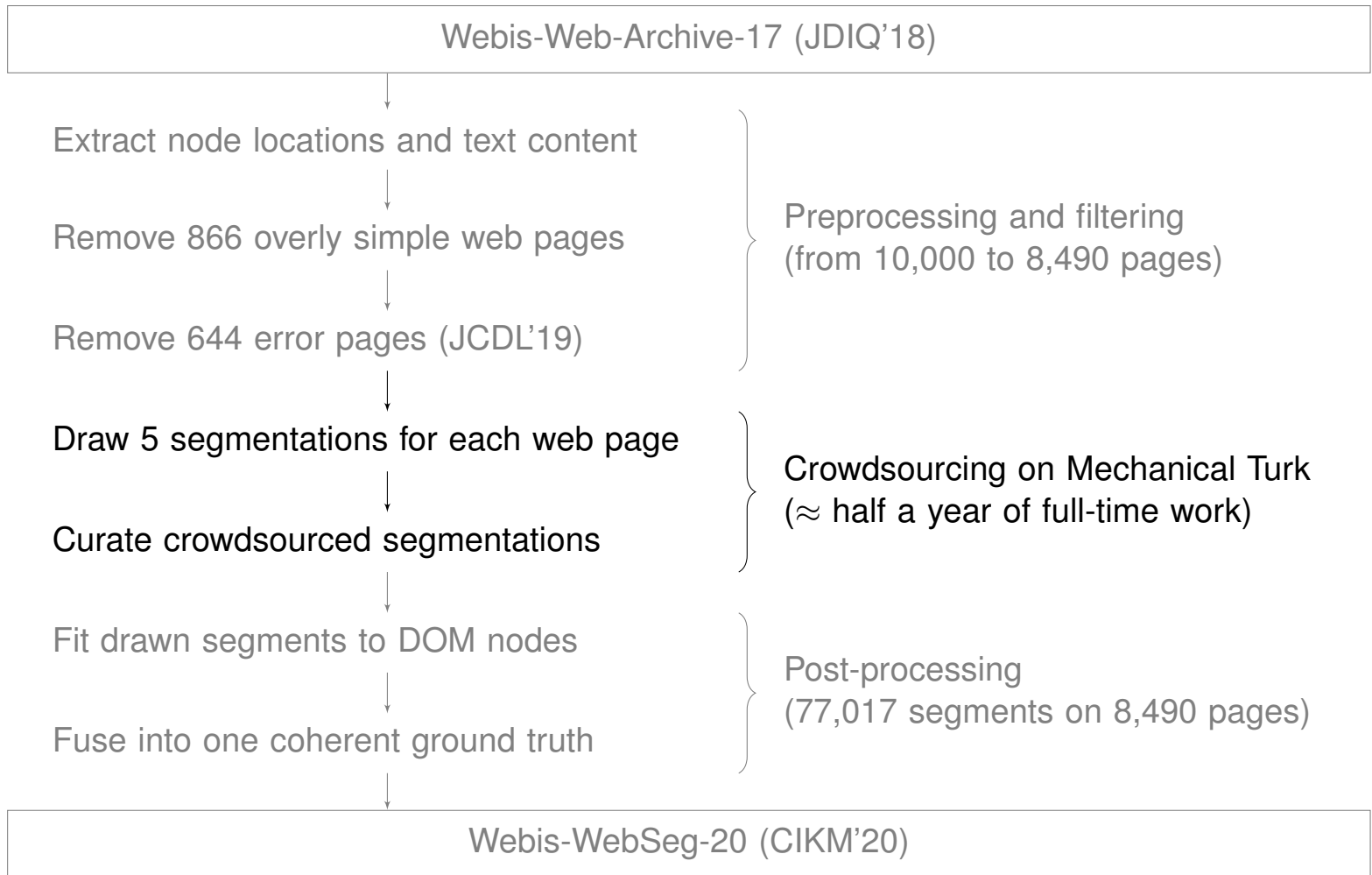
The Webis-WebSeg-20 Dataset



The Webis-WebSeg-20 Dataset



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 sections. A prominent article titled 'Bạn gái cặp với người nước ngoài vì tiền' (Girlfriend couples with foreigner for money) is highlighted with a red circle around its title and a red arrow pointing to it. Other visible elements include a financial table, a lottery announcement, and various news snippets. The page is overlaid with a semi-transparent map on the right side.

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

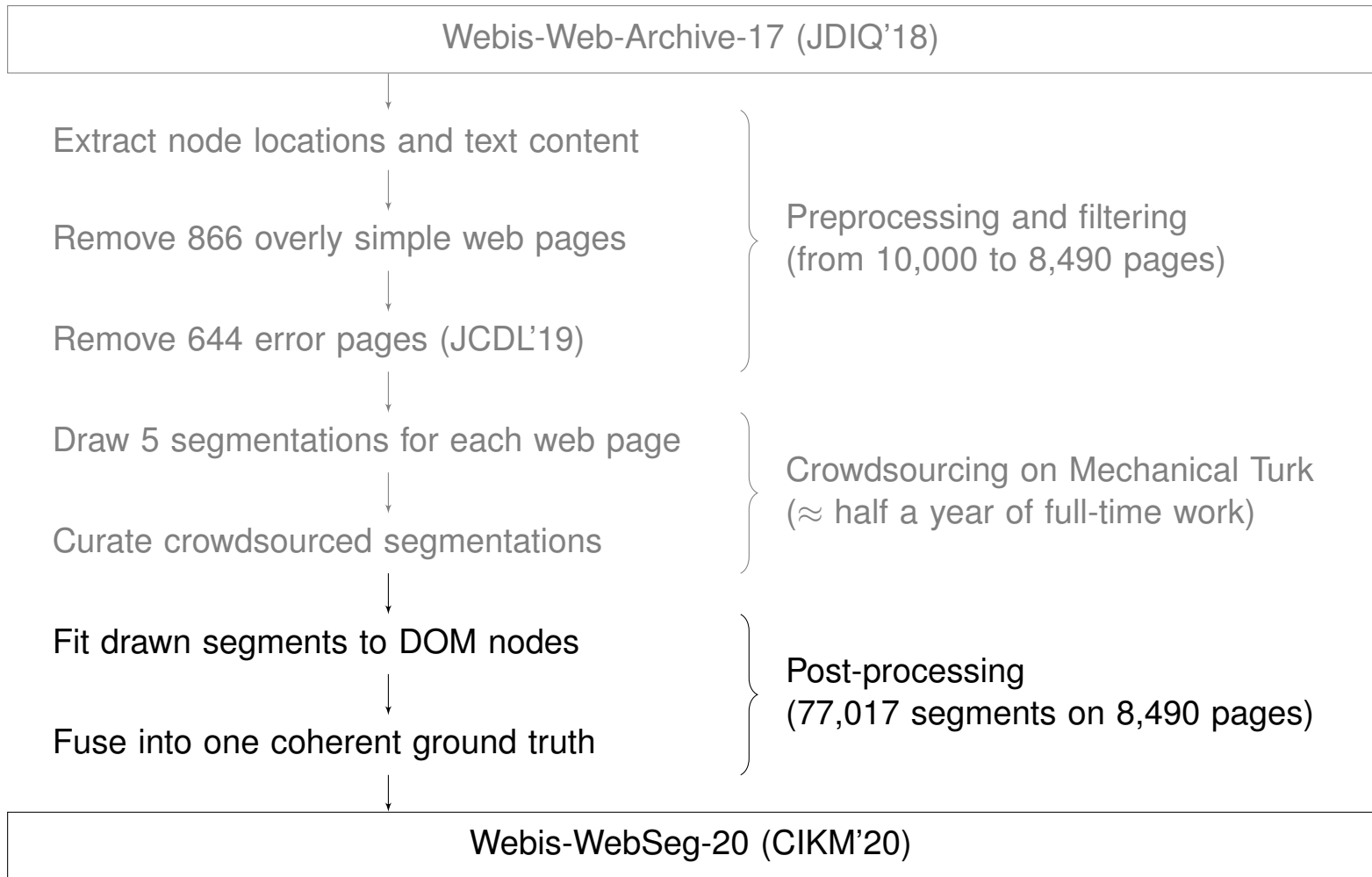
SUPER!
Das ist kein Scherz! Sie sind unser 1.000.000ster Besucher!
Unser Zufallssystem der möglichen Gewinner könnte Sie als möglichen Gewinner von FANTASTISCHEN APPLE Produkten ziehen.
ONLINE: 21/09/2017 22:40
[Klicken Sie hier](#)

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

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

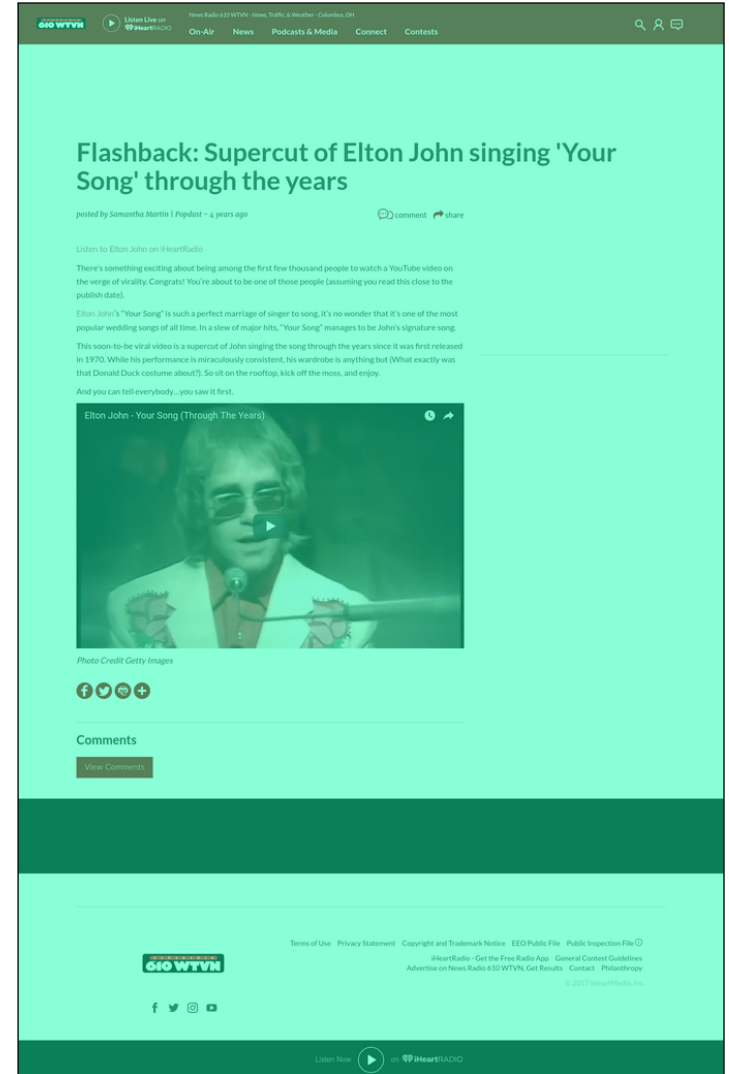


Algorithms

Name	Reference	Document	Features	Output
VIPS	Cai et al., 2003	Web page	Tree, style, location	Rectangle tree
HEPS	Manabe and Tajima, 2015	Web page	Tree, style	Node set
Cormier et al.	Cormier et al., 2017	Web page	Screenshot	Rectangle tree
MMDetection	Chen et al., 2019	Photo	Screenshot	Pixel masks
Meier et al.	Meier et al., 2017	Article page	Screenshot, text-mask	Mask

Baseline

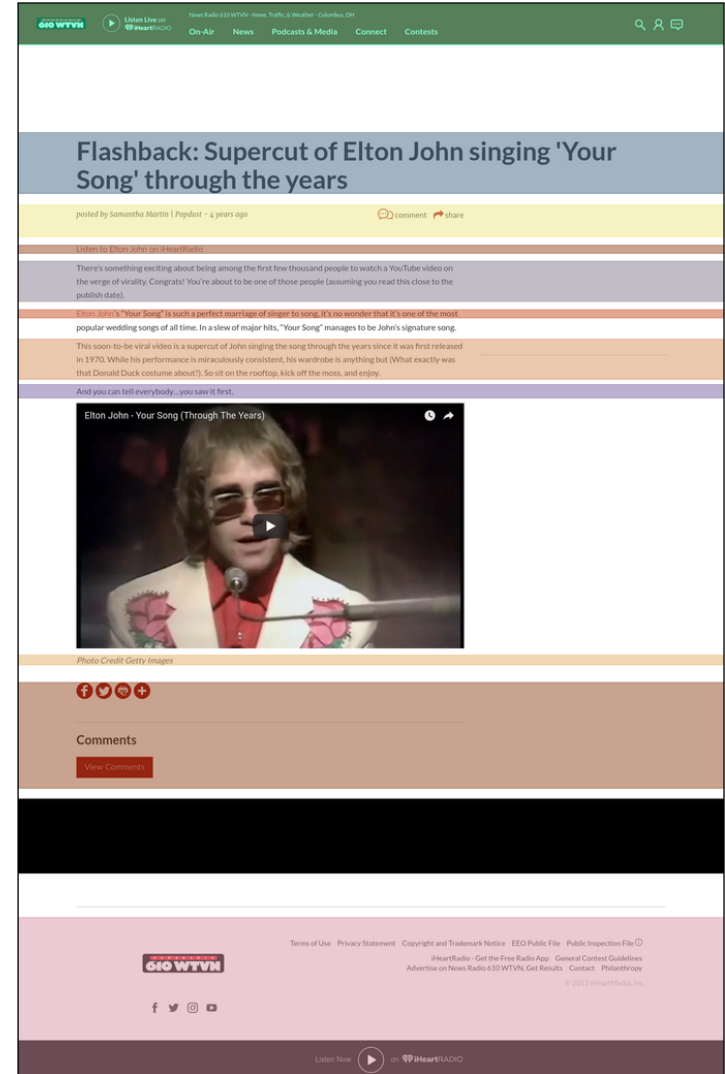
- ❑ One segment that covers the whole page
- ❑ Always achieves a recall of 1



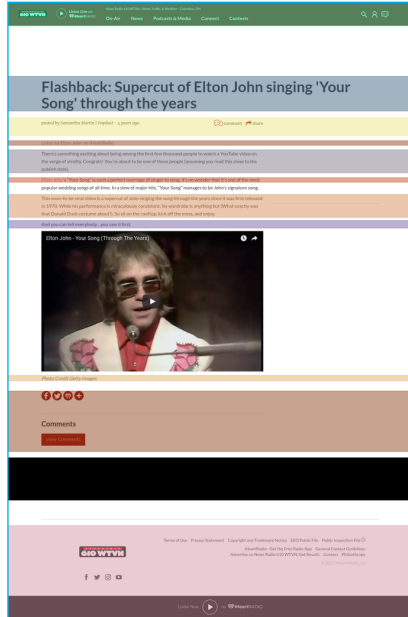
“Vision-based Page Segmentation algorithm” by Cai et al., 2003

- ❑ Starts with one segment that covers the whole page
- ❑ Computes the "degree of coherence" of each segment through heuristic rules
- ❑ Splits segments if their degree of coherence is less than the permitted degree (PDoC)

We re-implemented the algorithm to run in a modern browser

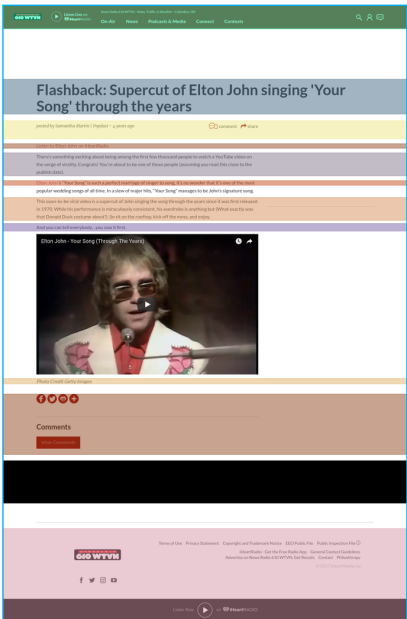


VIPS: Optimization for Permitted Degree of Coherence (PDoC)

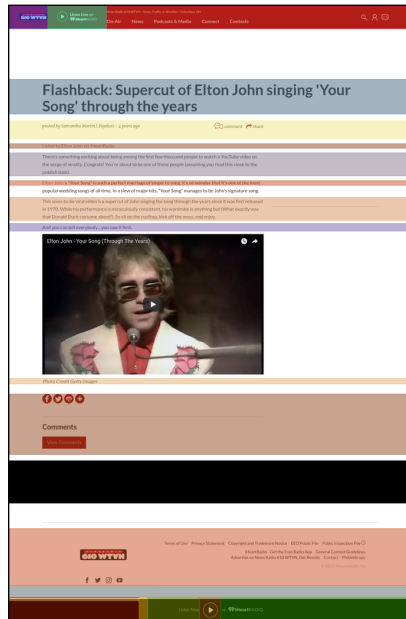


$$\text{PDoC} \in [1, 6]$$

VIPS: Optimization for Permitted Degree of Coherence (PDoC)

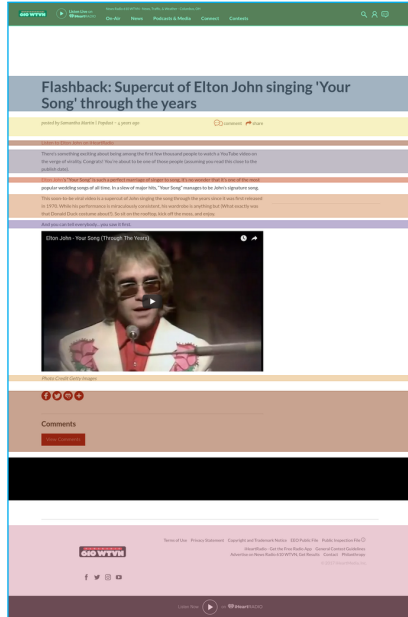


$PDoC \in [1, 6]$

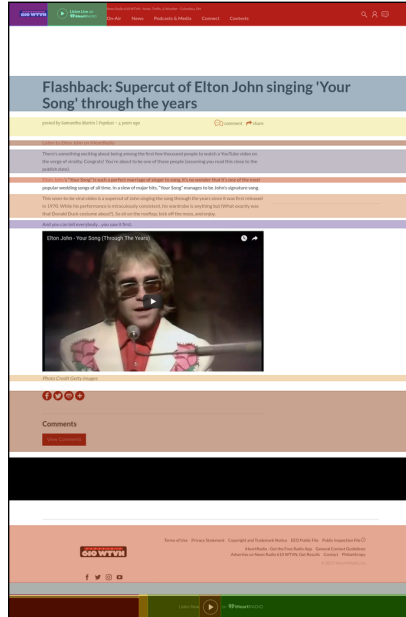


$PDoC = 7$

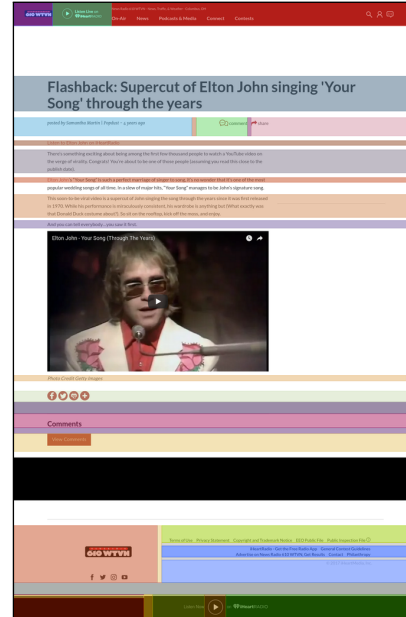
VIPS: Optimization for Permitted Degree of Coherence (PDoC)



$PDoC \in [1, 6]$

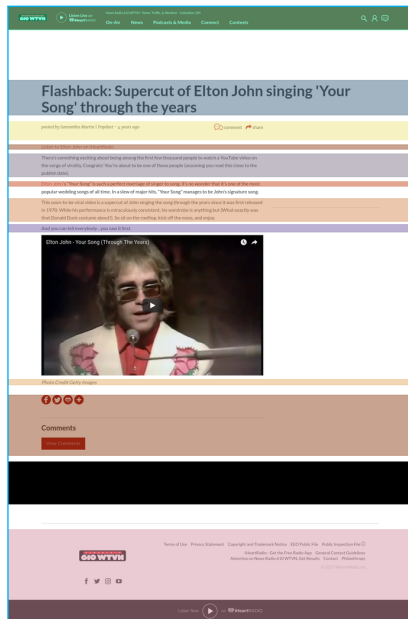


$PDoC = 7$

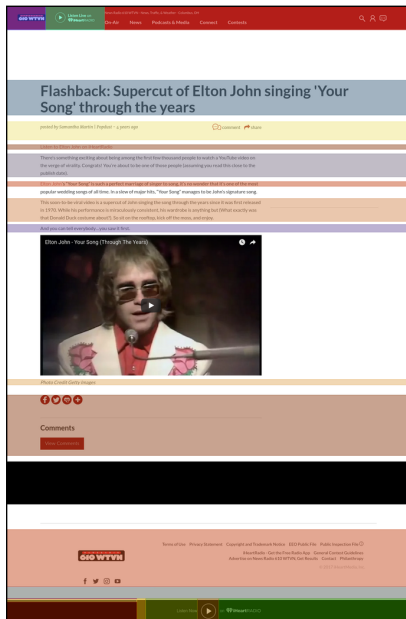


$PDoC \in [8, 9]$

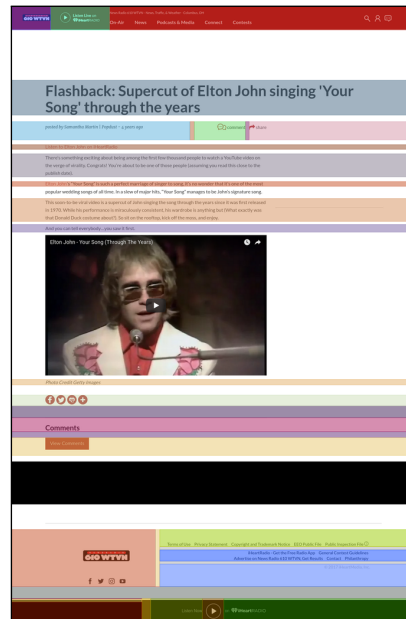
VIPS: Optimization for Permitted Degree of Coherence (PDoC)



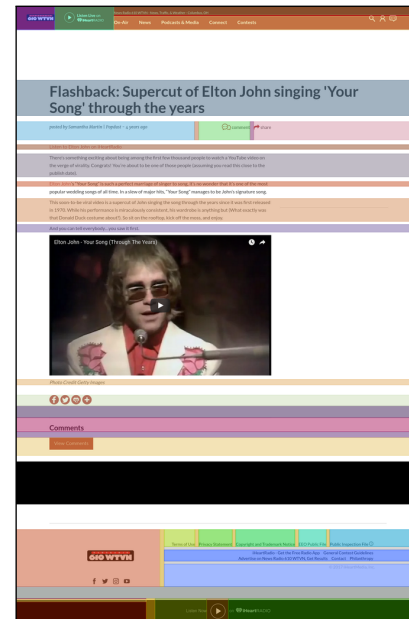
PDoC $\in [1, 6]$



PDoC = 7

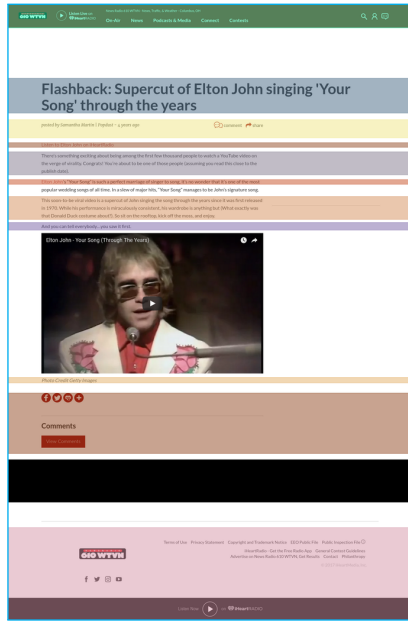


PDoC $\in [8, 9]$

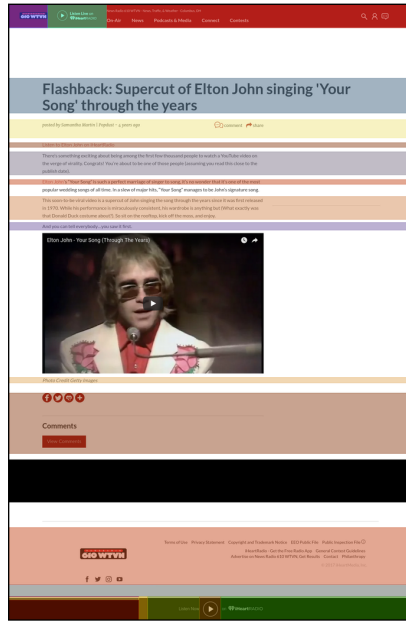


PDoC $\in [10, 11]$

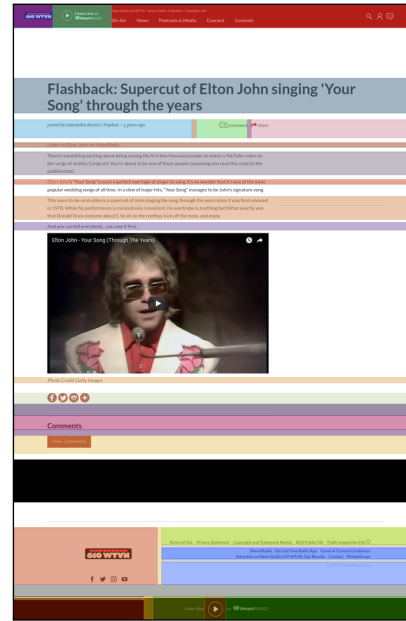
VIPS: Optimization for Permitted Degree of Coherence (PDoC)



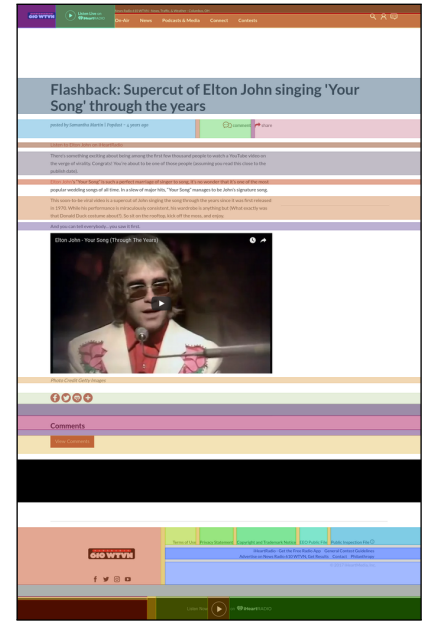
PDoC $\in [1, 6]$



PDoC = 7

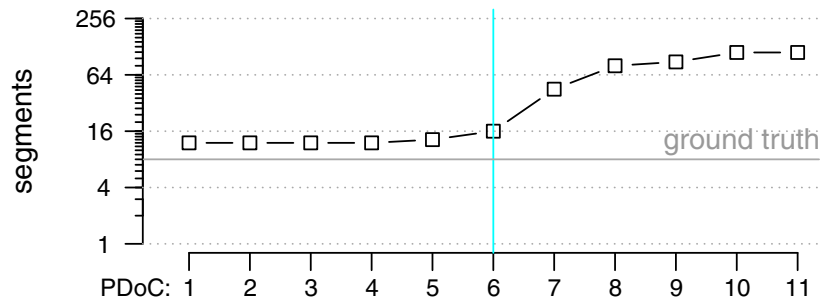


PDoC $\in [8, 9]$



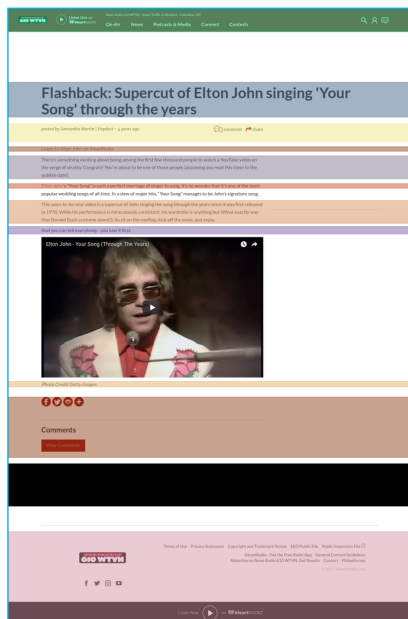
PDoC $\in [10, 11]$

Number of segments

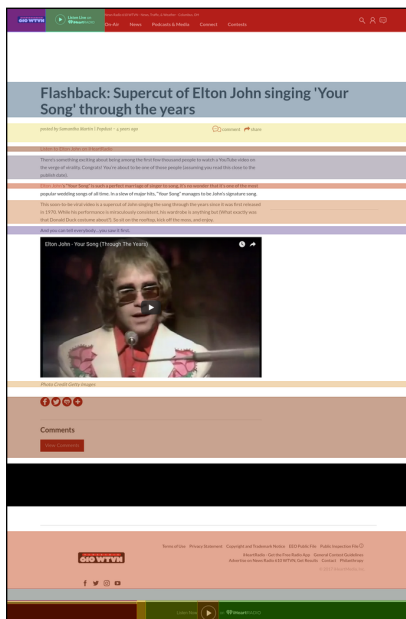


Legend: □ segments ○ F_B^3 ▼ P_B^3 △ R_B^3 ■ / ● / ▼ / ▲ fitted

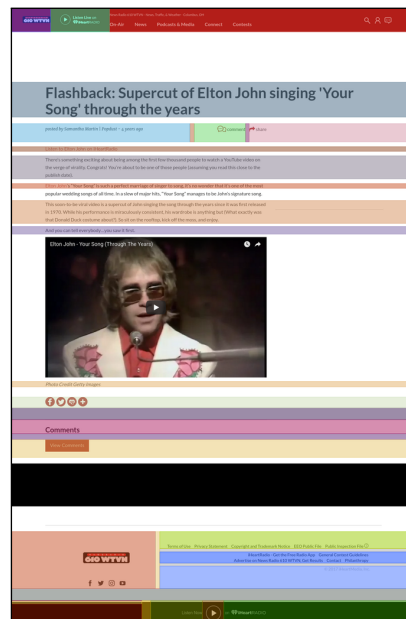
VIPS: Optimization for Permitted Degree of Coherence (PDoC)



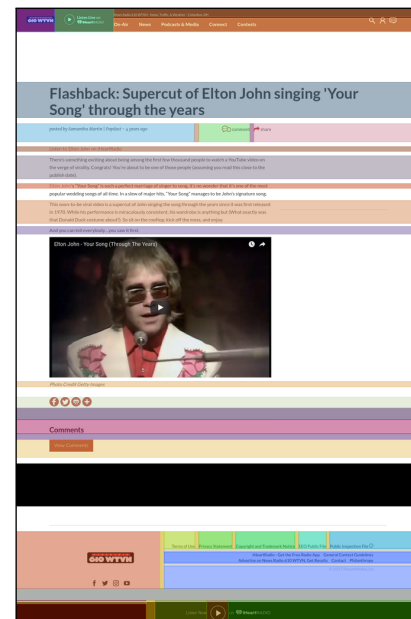
PDoC $\in [1, 6]$



PDoC = 7

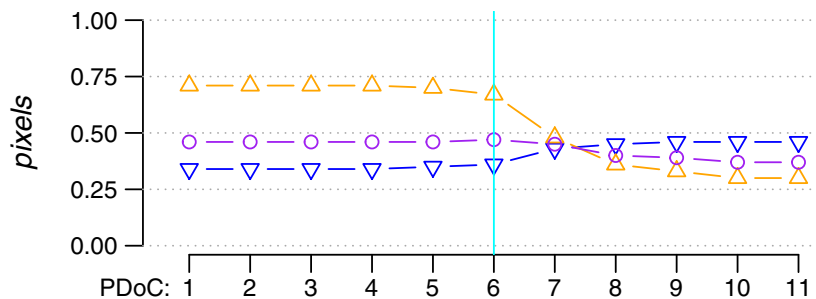


PDoC $\in [8, 9]$

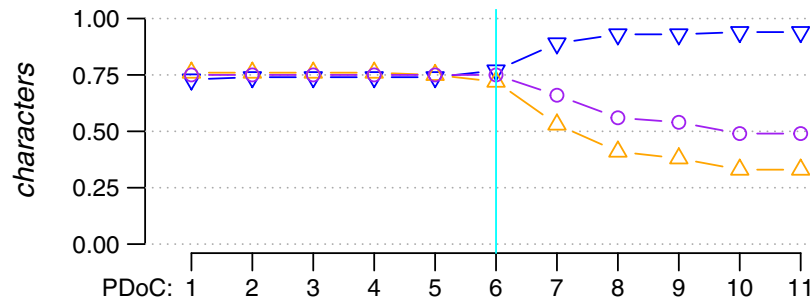


PDoC $\in [10, 11]$

Comparison with ground-truth



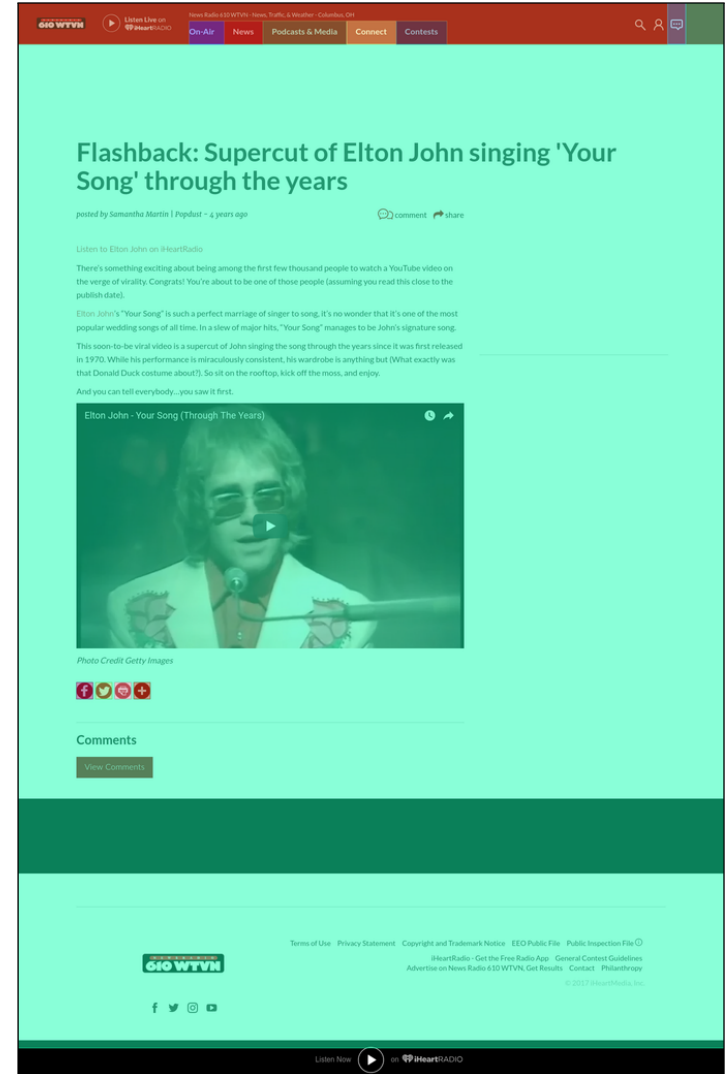
Legend: \square segments \circ F_B^3 ∇ P_B^3 \triangle R_B^3 $\blacksquare/\bullet/\blacktriangledown/\blacktriangle$ fitted



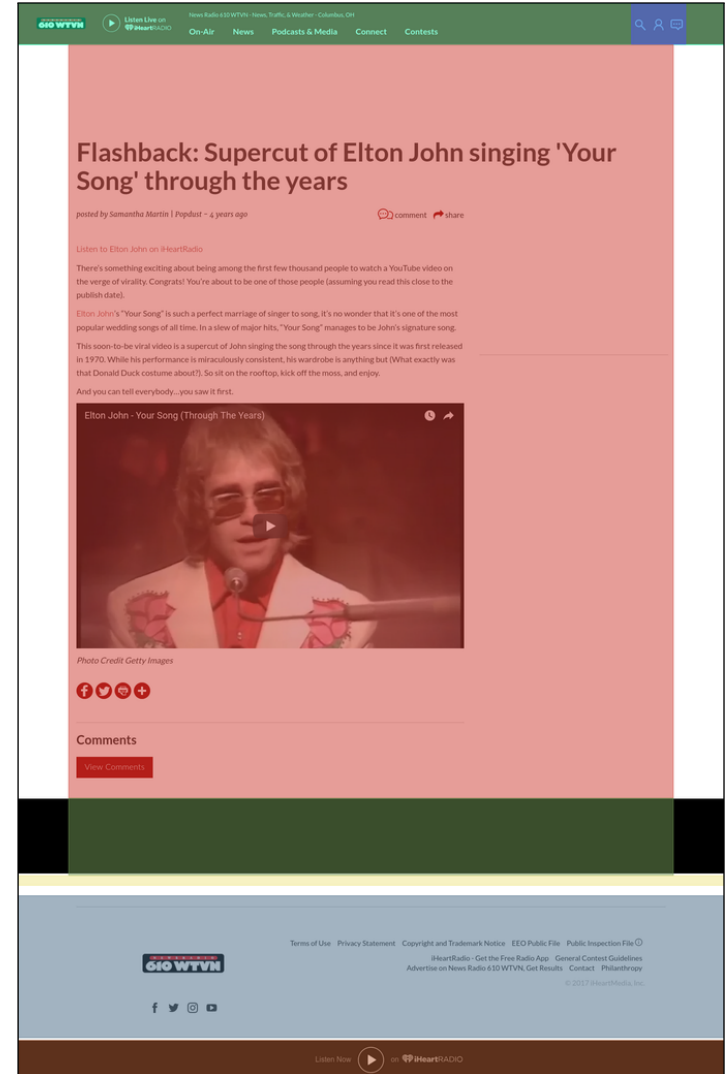
“HEading-based Page Segmentation algorithm” by Manabe and Tajima, 2015

- ❑ Identifies headings and their segments by heuristic rules
- ❑ A heading is “both visually prominent and described the topic of a segment”

We slightly adopted the author’s original implementation




- ❑ Uses the web page screenshot as sole input
- ❑ Identifies locally significant horizontal and vertical edge pixels
- ❑ Identifies horizontal and vertical “semantically significant” lines of such pixels
- ❑ Recursively splits segments by most semantically significant line



Popcash
The Popunder Network

Home Publishers Advertisers FAQ About Us Contact Login Register



Publishers

Maximize your revenue with PopCash.Net

HOW TO START?

Learn More about Our Advantages!

Sign Up Now!

Start earning money in less then 10 minutes!

Full Name

Email

Create Account

Popcash
The Popunder Network

Home Publishers Advertisers FAQ About Us Contact Login Register

Publishers

Maximize your revenue with PopCash.Net

HOW TO START?

Learn More about Our Advantages!

Sign Up Now!

Start earning money in less than 10 minutes!


Full Name

Email

Create Account

Popcash
The Popunder Network

Home Publishers Advertisers FAQ About Us Contact Login Register



Publishers

Maximize your revenue with PopCash.Net

HOW TO START?

Learn More about Our Advantages!

Sign Up Now!

Start earning money in less then 10 minutes!


Full Name

Email

Create Account

Popcash
The Popunder Network

Home Publishers Advertisers FAQ About Us Contact Login Register



Publishers

Maximize your revenue with PopCash.Net

HOW TO START?

Learn More about Our Advantages!

Sign Up Now!

Start earning money in less then 10 minutes!

Full Name

Email

Create Account

Popcash
The Popunder Network

Home Publishers Advertisers FAQ About Us Contact Login Register

Publishers

Maximize your revenue with PopCash.Net

HOW TO START?

Learn More about Our Advantages!

Sign Up Now!

Start earning money in less than 10 minutes!

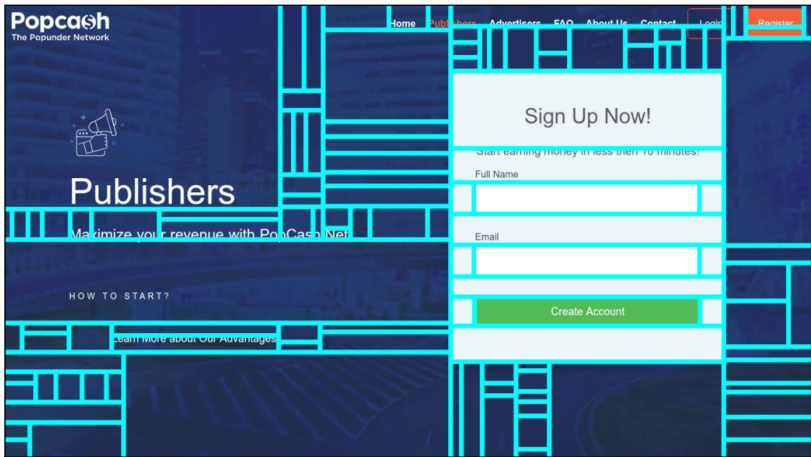
Full Name

Email

Create Account

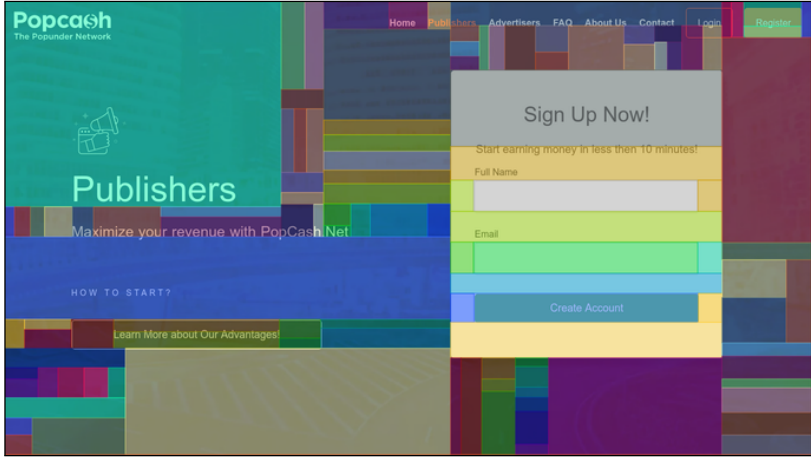
The image is a screenshot of the Popcash website, which is a platform for publishers. The website has a dark blue background with white and light blue text. The top navigation bar includes links for Home, Publishers, Advertisers, FAQ, About Us, Contact, Login, and Register. The main content area is divided into two columns. The left column features the Popcash logo, a megaphone icon, and the heading 'Publishers' with the subtext 'Maximize your revenue with PopCash Net'. Below this is a section titled 'HOW TO START?' with a link to 'Learn more about Our Advantages'. The right column is a registration form titled 'Sign Up Now!' with the subtext 'Start earning money in less than 10 minutes!'. The form includes input fields for 'Full Name' and 'Email', and a green 'Create Account' button. A red box highlights the 'Login' and 'Register' links in the top navigation bar.

Cormier et al.: Fitting to DOM Nodes

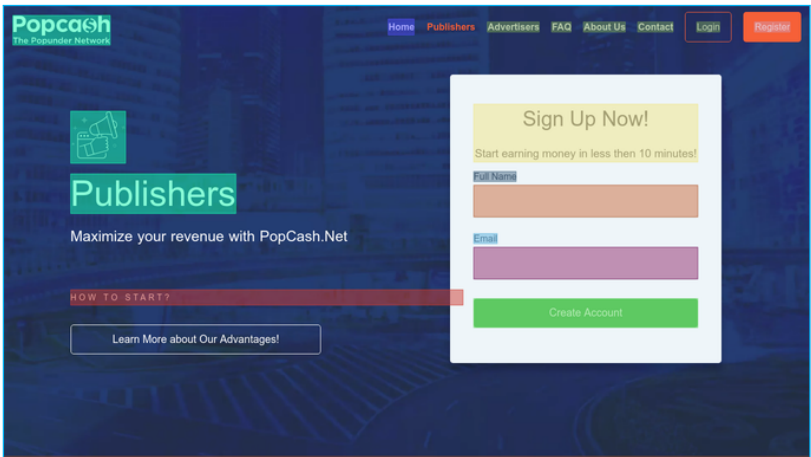


Original (borders)

- Segments are fitted to DOM nodes like the human annotations for the ground-truth

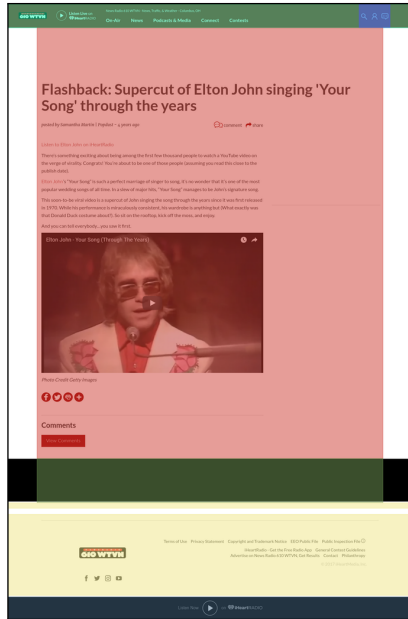


Original

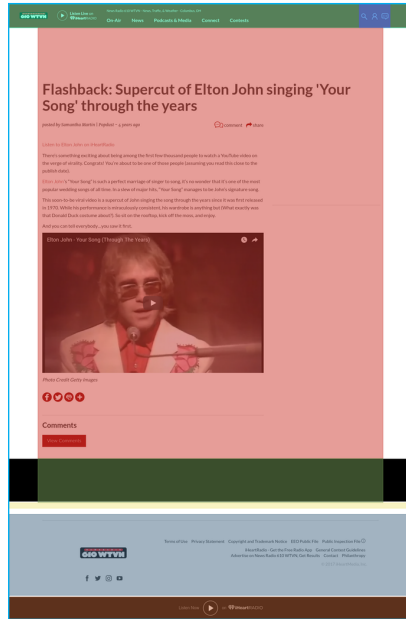


Fitted

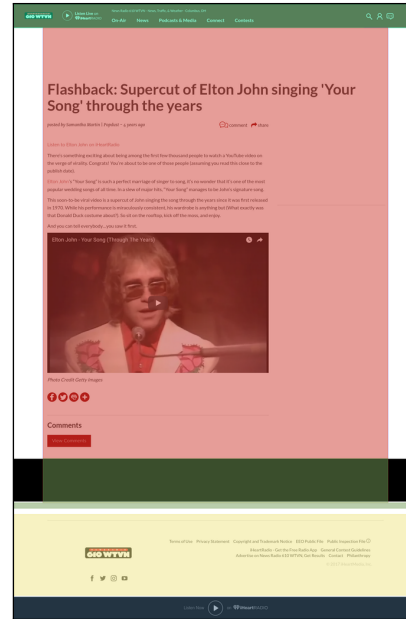
Cormier et al.: Optimization



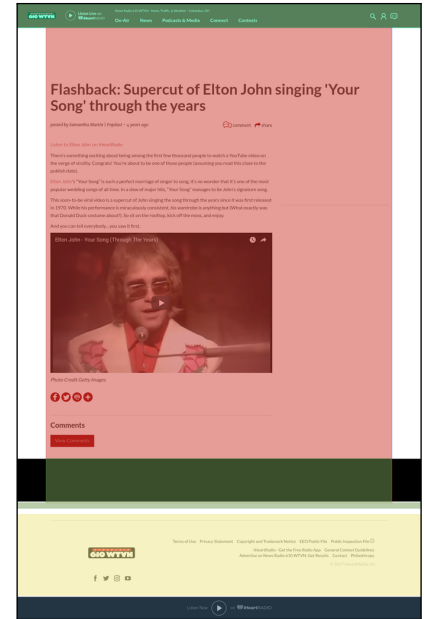
$$t_l = 256; s_{\min} = 45$$



$$t_l = 512; s_{\min} = 45$$

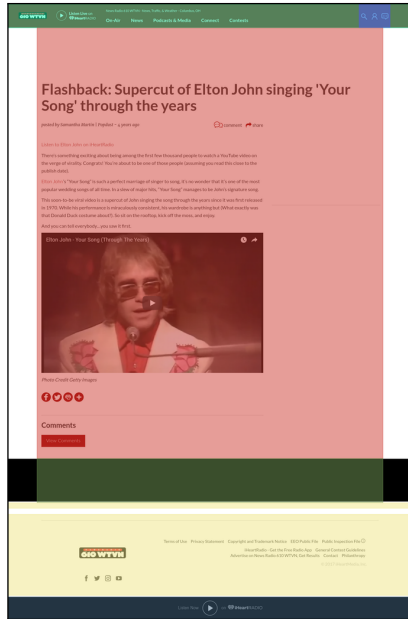


$$t_l = 256; s_{\min} = 90$$

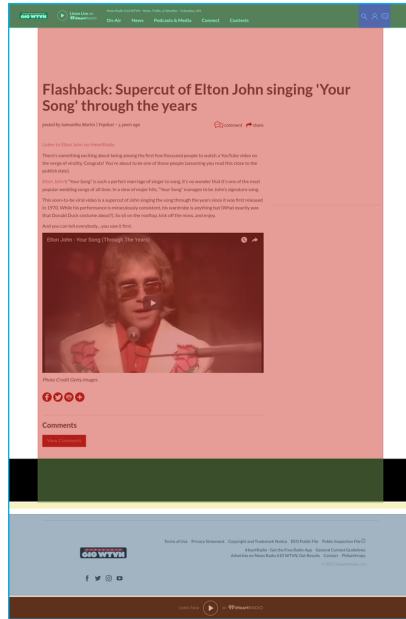


$$t_l = 512; s_{\min} = 90$$

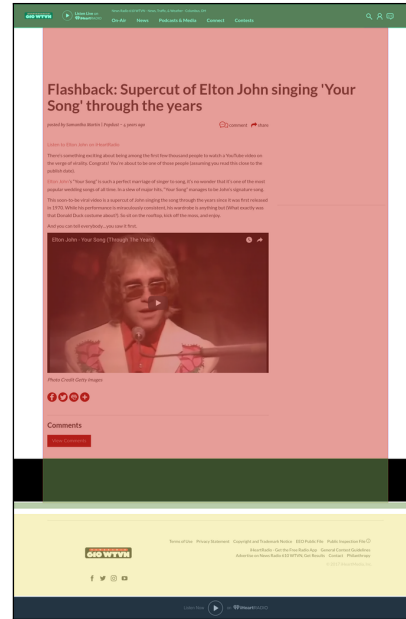
Cormier et al.: Optimization



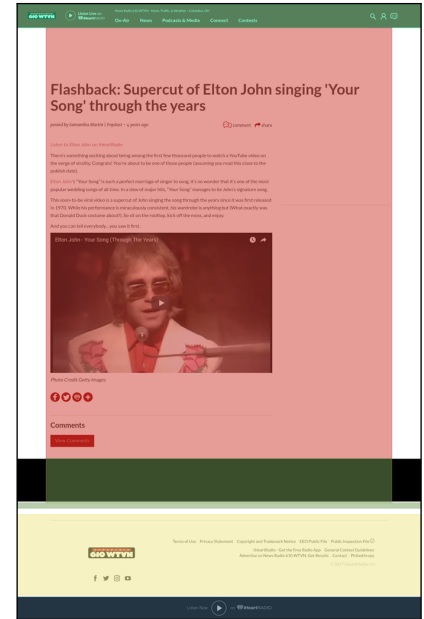
$$t_l = 256; s_{\min} = 45$$



$$t_l = 512; s_{\min} = 45$$

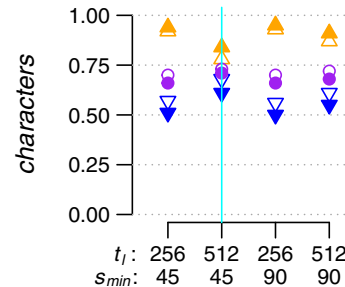
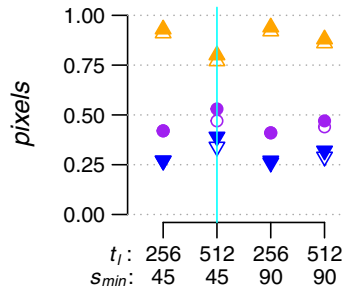
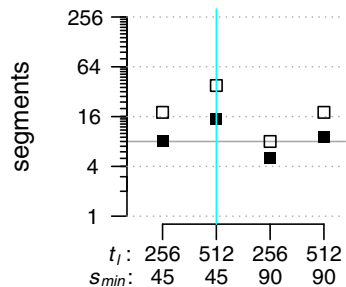


$$t_l = 256; s_{\min} = 90$$



$$t_l = 512; s_{\min} = 90$$

Number of segments and comparison with ground-truth



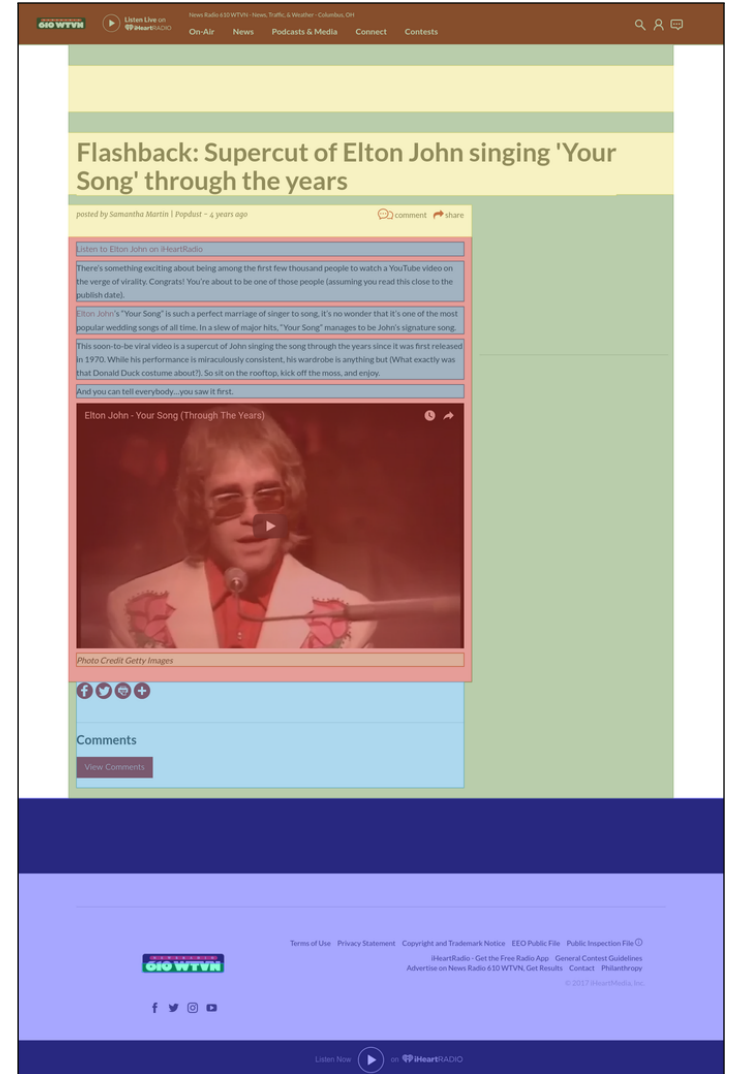
Legend: \square segments \circ F_B^* ∇ P_B^3 \triangle R_B^3 \blacksquare \bullet \blacktriangledown \blacktriangle fitted

MMDetection

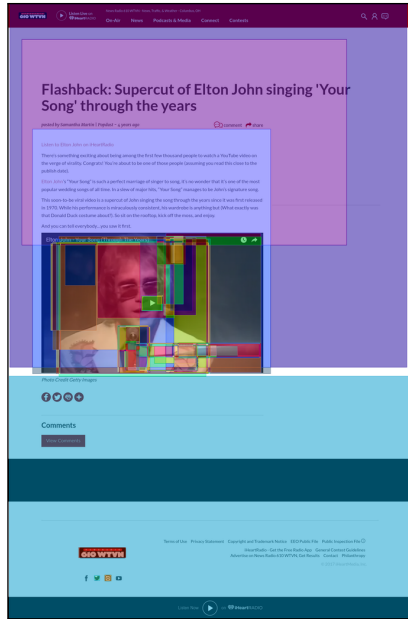
One Hybrid Task Cascade model from the MMDetection toolbox by Chen et al., 2019.

Model was state-of-the-art in 2020 as per the MSCOCO object detection task leaderboard

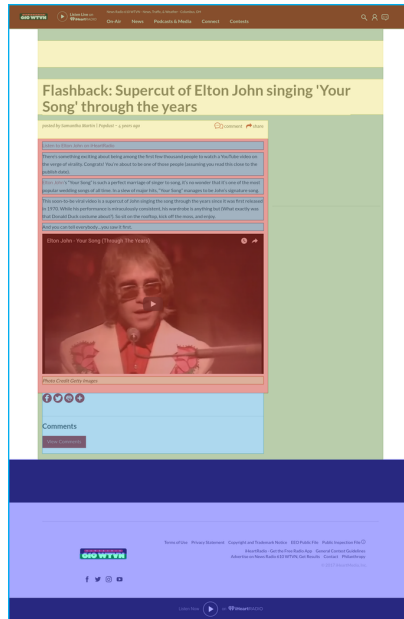
- ❑ Uses the web page screenshot as sole input
- ❑ Neural network
- ❑ Trained on object detection in real-world images (photos)



MMDetection: Optimization

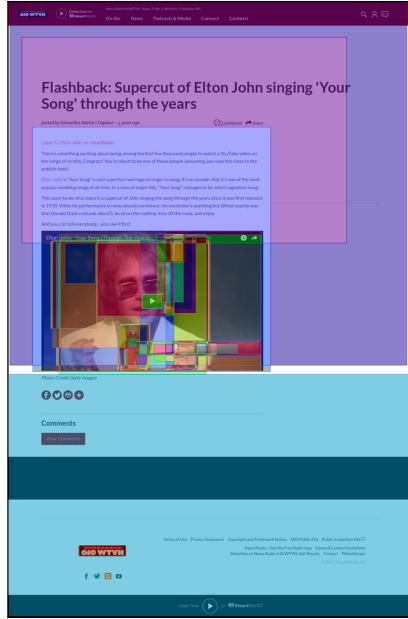


Original

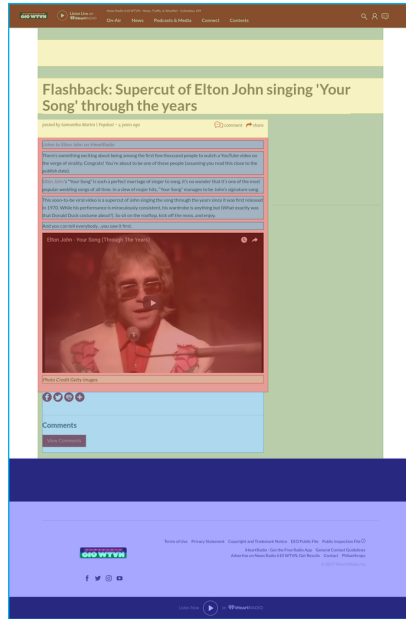


Fitted

MMDetection: Optimization

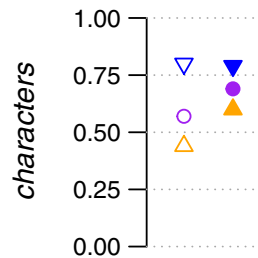
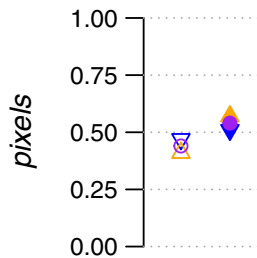
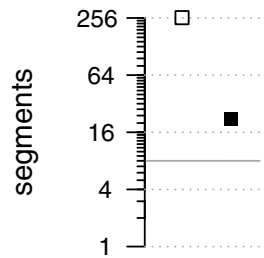


Original



Fitted

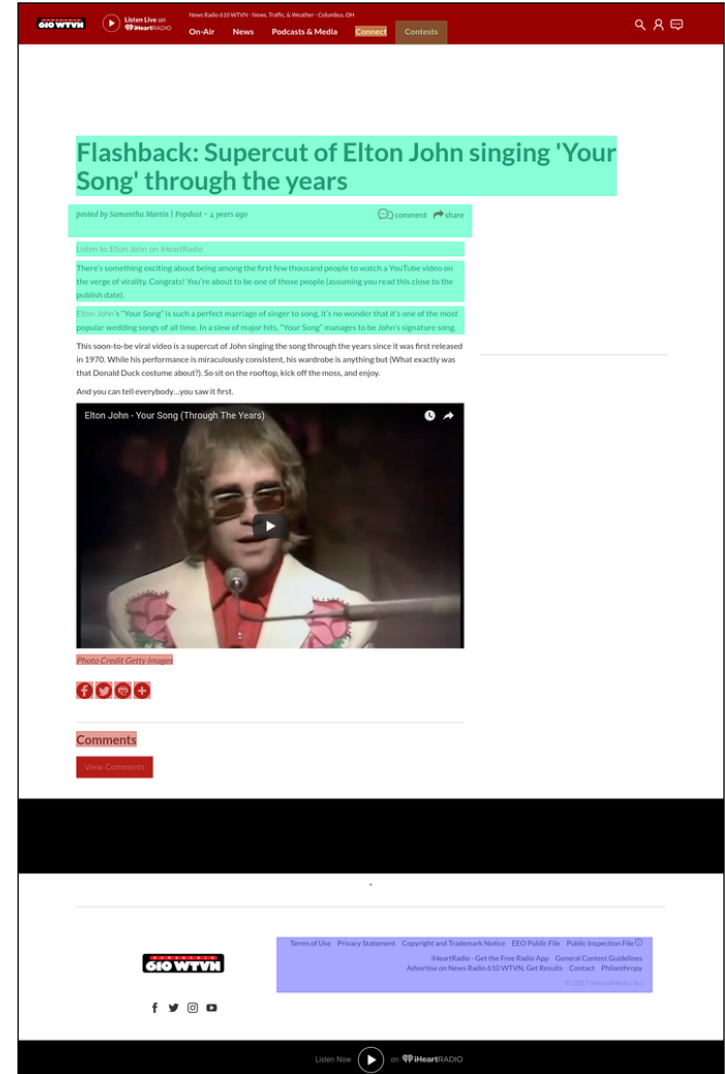
Number of segments and comparison with ground-truth



Legend: □ segments ○ F_B^* ▼ P_B^3 △ R_B^3 ■ / ● / ▼ / ▲ fitted

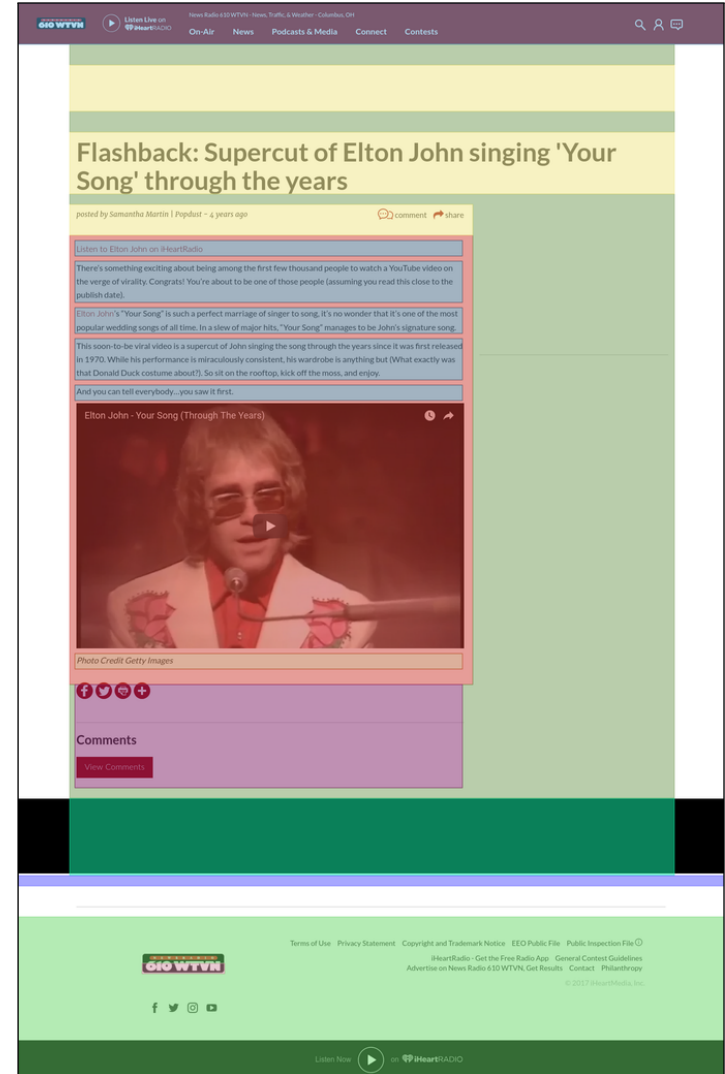
Meier et al., 2017

- ❑ Uses the web page screenshot and the location of text nodes as input
- ❑ Convolutional neural network
- ❑ Requires fixed-size input images: cropping to 4096 pixels height
- ❑ Originally developed/trained for newspaper segmentation
- ❑ 10-fold cross-evaluation on the Webis-WebSeg-20
- ❑ No detailed comparison to other algorithms due to differences in the setup

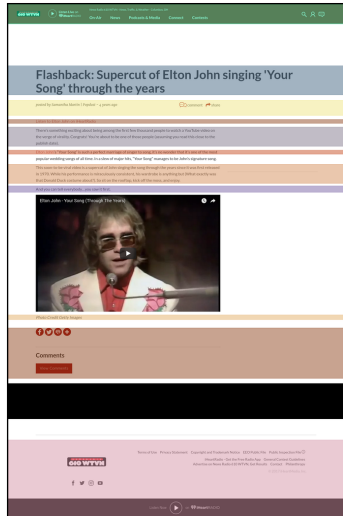


Min-vote Ensemble

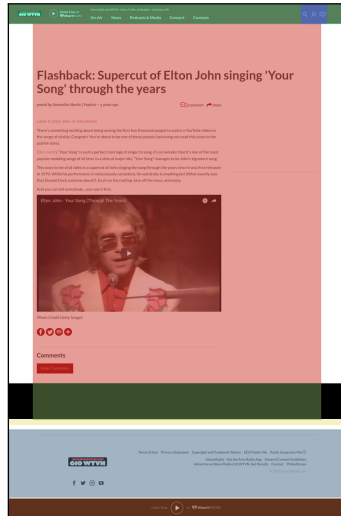
- ❑ Ensemble of VIPS, HEPS, Cormier et al., and MMDetection
- ❑ Parameter $n \in [1, 4]$
- ❑ Ignores elements which less than n algorithms placed into segments
- ❑ Standard hierarchical agglomerative clustering
- ❑ Similarity of two elements is the ratio of algorithms that place these elements in the same segment
- ❑ Similarity thresholds is $\frac{n-0.5}{4}$
Roughly: group elements together if at least n algorithms did so



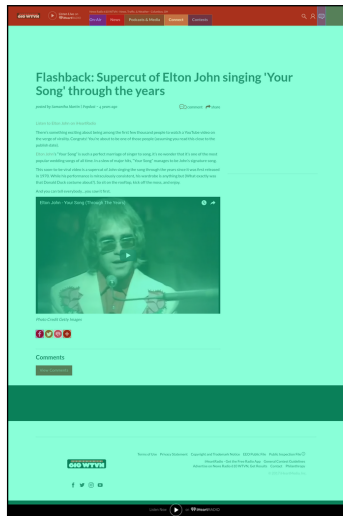
Min-vote Ensemble



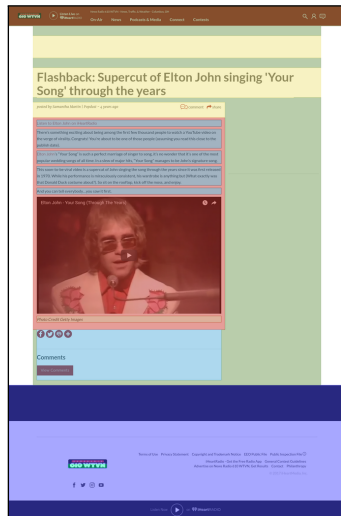
VIPS



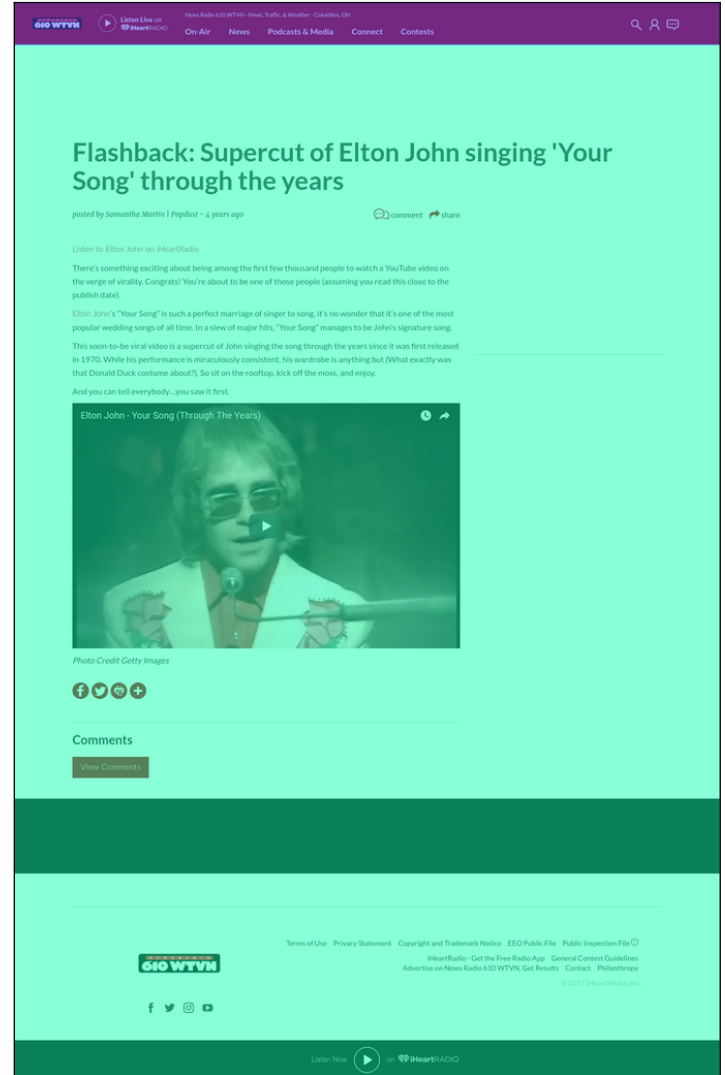
Cormier et al.



HEPS

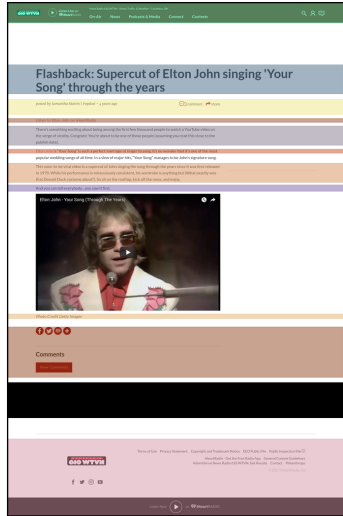


MMDetection

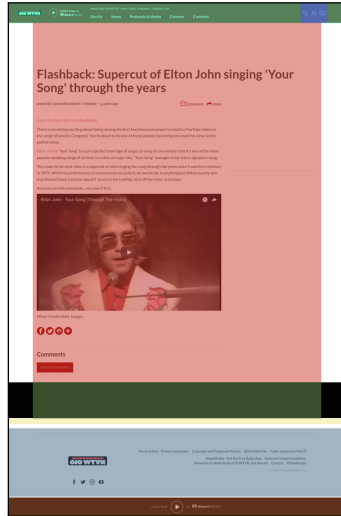


Min-vote@1

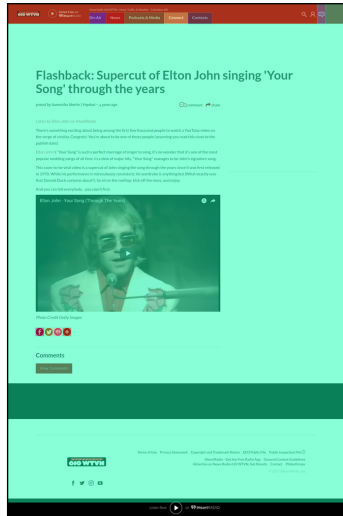
Min-vote Ensemble



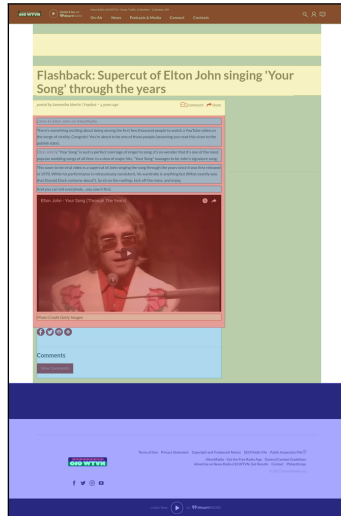
VIPS



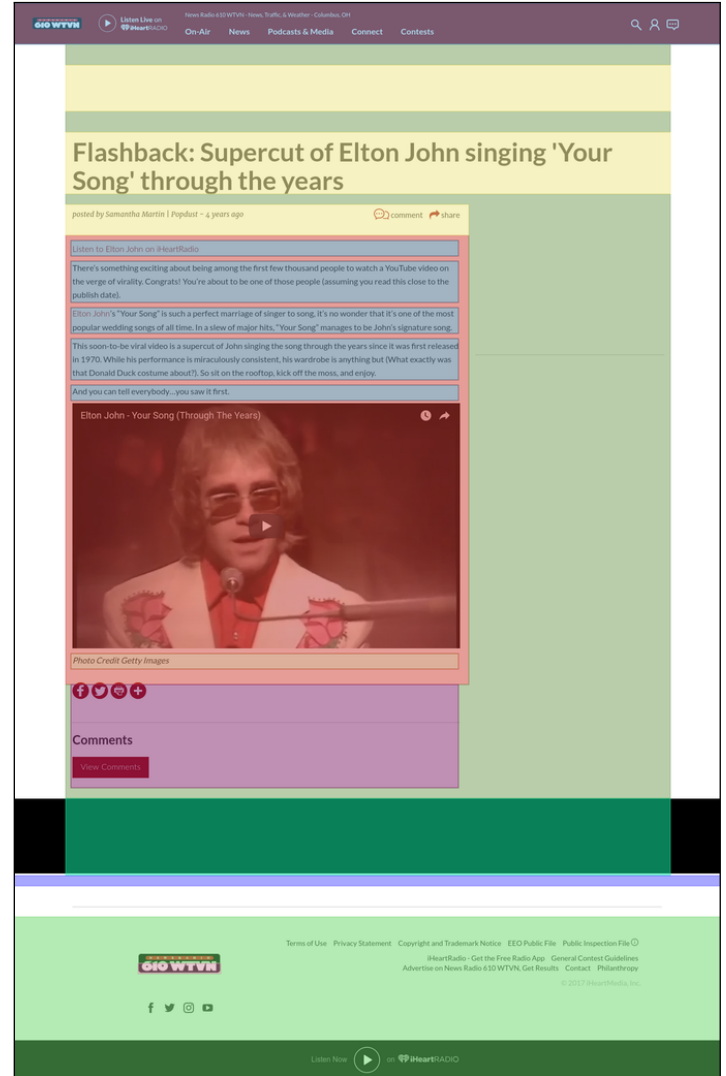
Cormier et al.



HEPS

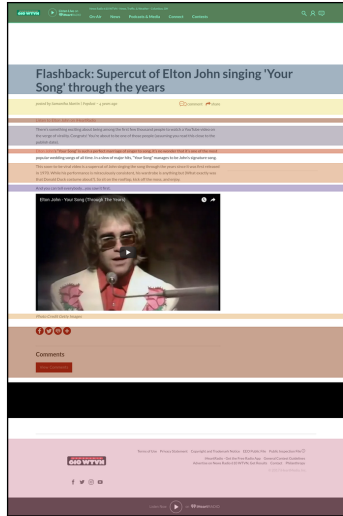


MMDetection

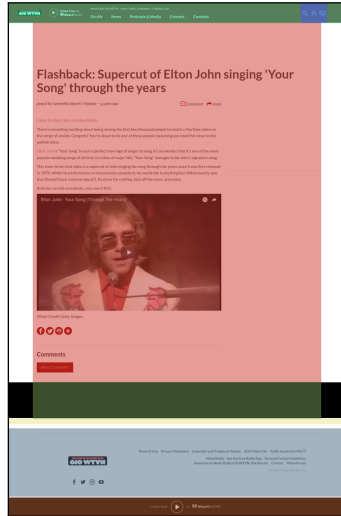


Min-vote@2

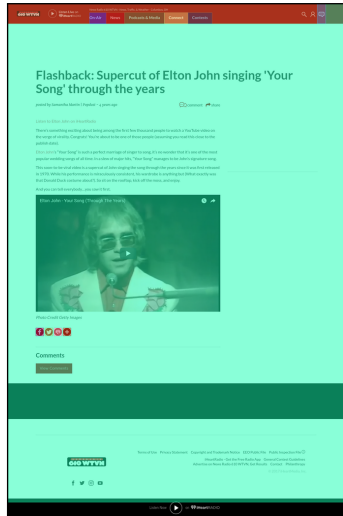
Min-vote Ensemble



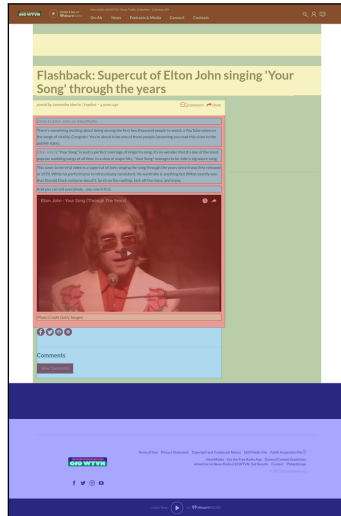
VIPS



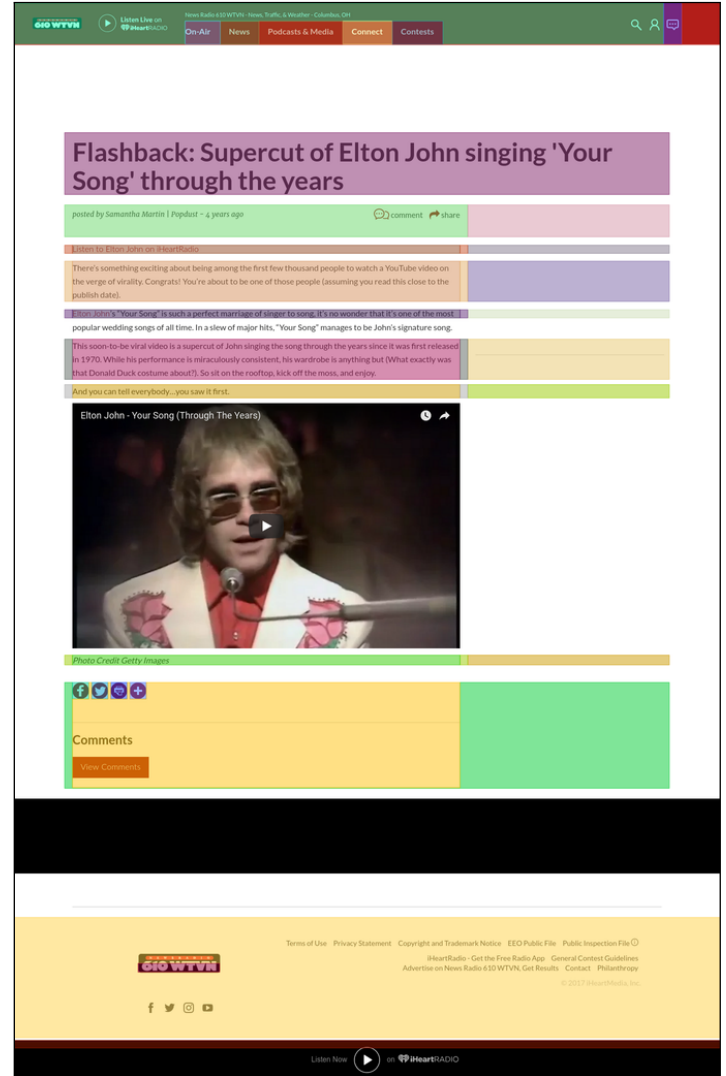
Cormier et al.



HEPS



MMDetection



Min-vote@4

Results

Measure	Baseline	VIPS	HEPS	Corm.	MMD.	Meier	MV@1	MV@2	MV@3	MV@4	
Segments	1.0	16.1	36.1	15.3	23.0	4.6	6.5	18.7	36.5	69.5	
<i>pixels</i>	F_{B^3}	0.24	0.38	0.33	0.36	0.42	0.32	0.30	0.39	0.30	0.28
	$F_{B^3}^*$	0.28	0.47	0.44	0.53	0.54	0.50	0.35	0.50	0.45	0.42
	P_{B^3}	0.16	0.36	0.36	0.39	0.51	0.48	0.22	0.38	0.60	0.68
	R_{B^3}	1.00	0.67	0.56	0.80	0.57	0.52	0.96	0.72	0.36	0.30
<i>edges_F</i>	F_{B^3}	0.44	0.59	0.48	0.51	0.53	0.41	0.50	0.56	0.39	0.34
	$F_{B^3}^*$	0.49	0.68	0.58	0.65	0.61	0.55	0.56	0.66	0.49	0.45
	P_{B^3}	0.32	0.66	0.61	0.55	0.73	0.55	0.40	0.61	0.81	0.87
	R_{B^3}	1.00	0.69	0.55	0.80	0.53	0.55	0.96	0.71	0.36	0.30
<i>edges_C</i>	F_{B^3}	0.45	0.61	0.49	0.53	0.54	0.42	0.51	0.57	0.39	0.35
	$F_{B^3}^*$	0.49	0.68	0.59	0.66	0.62	0.56	0.56	0.67	0.50	0.46
	P_{B^3}	0.32	0.67	0.62	0.56	0.74	0.55	0.40	0.63	0.82	0.88
	R_{B^3}	1.00	0.70	0.56	0.80	0.53	0.57	0.96	0.72	0.36	0.31
<i>nodes</i>	F_{B^3}	0.42	0.63	0.43	0.52	0.52	0.44	0.49	0.54	0.34	0.31
	$F_{B^3}^*$	0.46	0.70	0.54	0.65	0.61	0.56	0.55	0.65	0.44	0.42
	P_{B^3}	0.30	0.69	0.63	0.53	0.74	0.52	0.38	0.64	0.85	0.88
	R_{B^3}	1.00	0.71	0.46	0.82	0.51	0.61	0.96	0.65	0.29	0.27
<i>chars</i>	F_{B^3}	0.52	0.67	0.50	0.61	0.61	0.50	0.59	0.62	0.40	0.39
	$F_{B^3}^*$	0.57	0.75	0.60	0.71	0.69	0.61	0.64	0.71	0.50	0.49
	P_{B^3}	0.39	0.77	0.73	0.61	0.79	0.59	0.48	0.72	0.90	0.92
	R_{B^3}	1.00	0.72	0.51	0.84	0.60	0.63	0.96	0.71	0.35	0.33

Results

Measure		Baseline	VIPS	HEPS	Corm.	MMD.	Meier	MV@1	MV@2	MV@3	MV@4
Segments		1.0	16.1	36.1	15.3	23.0	4.6	6.5	18.7	36.5	69.5
<i>pixels</i>	F_{B^3}	0.24	0.38	0.33	0.36	0.42	0.32	0.30	0.39	0.30	0.28
	$F_{B^3}^*$	0.28	0.47	0.44	0.53	0.54	0.50	0.35	0.50	0.45	0.42
	P_{B^3}	0.16	0.36	0.36	0.39	0.51	0.48	0.22	0.38	0.60	0.68
	R_{B^3}	1.00	0.67	0.56	0.80	0.57	0.52	0.96	0.72	0.36	0.30
<i>edges_F</i>	F_{B^3}	0.44	0.59	0.48	0.51	0.53	0.41	0.50	0.56	0.39	0.34
	$F_{B^3}^*$	0.49	0.68	0.58	0.65	0.61	0.55	0.56	0.66	0.49	0.45
	P_{B^3}	0.32	0.66	0.61	0.55	0.73	0.55	0.40	0.61	0.81	0.87
	R_{B^3}	1.00	0.69	0.55	0.80	0.53	0.55	0.96	0.71	0.36	0.30
<i>edges_C</i>	F_{B^3}	0.45	0.61	0.49	0.53	0.54	0.42	0.51	0.57	0.39	0.35
	$F_{B^3}^*$	0.49	0.68	0.59	0.66	0.62	0.56	0.56	0.67	0.50	0.46
	P_{B^3}	0.32	0.67	0.62	0.56	0.74	0.55	0.40	0.63	0.82	0.88
	R_{B^3}	1.00	0.70	0.56	0.80	0.53	0.57	0.96	0.72	0.36	0.31
<i>nodes</i>	F_{B^3}	0.42	0.63	0.43	0.52	0.52	0.44	0.49	0.54	0.34	0.31
	$F_{B^3}^*$	0.46	0.70	0.54	0.65	0.61	0.56	0.55	0.65	0.44	0.42
	P_{B^3}	0.30	0.69	0.63	0.53	0.74	0.52	0.38	0.64	0.85	0.88
	R_{B^3}	1.00	0.71	0.46	0.82	0.51	0.61	0.96	0.65	0.29	0.27
<i>chars</i>	F_{B^3}	0.52	0.67	0.50	0.61	0.61	0.50	0.59	0.62	0.40	0.39
	$F_{B^3}^*$	0.57	0.75	0.60	0.71	0.69	0.61	0.64	0.71	0.50	0.49
	P_{B^3}	0.39	0.77	0.73	0.61	0.79	0.59	0.48	0.72	0.90	0.92
	R_{B^3}	1.00	0.72	0.51	0.84	0.60	0.63	0.96	0.71	0.35	0.33

Results

Measure	Baseline	VIPS	HEPS	Corm.	MMD.	Meier	MV@1	MV@2	MV@3	MV@4	
Segments	1.0	16.1	36.1	15.3	23.0	4.6	6.5	18.7	36.5	69.5	
<i>pixels</i>	F_{B^3}	0.24	0.38	0.33	0.36	0.42	0.32	0.30	0.39	0.30	0.28
	$F_{B^3}^*$	0.28	0.47	0.44	0.53	0.54	0.50	0.35	0.50	0.45	0.42
	P_{B^3}	0.16	0.36	0.36	0.39	0.51	0.48	0.22	0.38	0.60	0.68
	R_{B^3}	1.00	0.67	0.56	0.80	0.57	0.52	0.96	0.72	0.36	0.30
<i>chars</i>	F_{B^3}	0.52	0.67	0.50	0.61	0.61	0.50	0.59	0.62	0.40	0.39
	$F_{B^3}^*$	0.57	0.75	0.60	0.71	0.69	0.61	0.64	0.71	0.50	0.49
	P_{B^3}	0.39	0.77	0.73	0.61	0.79	0.59	0.48	0.72	0.90	0.92
	R_{B^3}	1.00	0.72	0.51	0.84	0.60	0.63	0.96	0.71	0.35	0.33

Results

Measure	Baseline	VIPS	HEPS	Corm.	MMD.	Meier	MV@1	MV@2	MV@3	MV@4	
Segments	1.0	16.1	36.1	15.3	23.0	4.6	6.5	18.7	36.5	69.5	
<i>pixels</i>	F_{B^3}	0.24	0.38	0.33	0.36	0.42	0.32	0.30	0.39	0.30	0.28
	$F_{B^3}^*$	0.28	0.47	0.44	0.53	0.54	0.50	0.35	0.50	0.45	0.42
	P_{B^3}	0.16	0.36	0.36	0.39	0.51	0.48	0.22	0.38	0.60	0.68
	R_{B^3}	1.00	0.67	0.56	0.80	0.57	0.52	0.96	0.72	0.36	0.30
<i>chars</i>	F_{B^3}	0.52	0.67	0.50	0.61	0.61	0.50	0.59	0.62	0.40	0.39
	$F_{B^3}^*$	0.57	0.75	0.60	0.71	0.69	0.61	0.64	0.71	0.50	0.49
	P_{B^3}	0.39	0.77	0.73	0.61	0.79	0.59	0.48	0.72	0.90	0.92
	R_{B^3}	1.00	0.72	0.51	0.84	0.60	0.63	0.96	0.71	0.35	0.33

Results

Measure	Baseline	VIPS	HEPS	Corm.	MMD.	Meier	MV@1	MV@2	MV@3	MV@4	
Segments	1.0	16.1	36.1	15.3	23.0	4.6	6.5	18.7	36.5	69.5	
<i>pixels</i>	F_{B^3}	0.24	0.38	0.33	0.36	0.42	0.32	0.30	0.39	0.30	0.28
	$F_{B^3}^*$	0.28	0.47	0.44	0.53	0.54	0.50	0.35	0.50	0.45	0.42
	P_{B^3}	0.16	0.36	0.36	0.39	0.51	0.48	0.22	0.38	0.60	0.68
	R_{B^3}	1.00	0.67	0.56	0.80	0.57	0.52	0.96	0.72	0.36	0.30
<i>chars</i>	F_{B^3}	0.52	0.67	0.50	0.61	0.61	0.50	0.59	0.62	0.40	0.39
	$F_{B^3}^*$	0.57	0.75	0.60	0.71	0.69	0.61	0.64	0.71	0.50	0.49
	P_{B^3}	0.39	0.77	0.73	0.61	0.79	0.59	0.48	0.72	0.90	0.92
	R_{B^3}	1.00	0.72	0.51	0.84	0.60	0.63	0.96	0.71	0.35	0.33

Results

Measure	Baseline	VIPS	HEPS	Corm.	MMD.	Meier	MV@1	MV@2	MV@3	MV@4	
Segments	1.0	16.1	36.1	15.3	23.0	4.6	6.5	18.7	36.5	69.5	
<i>pixels</i>	F_{B^3}	0.24	0.38	0.33	0.36	0.42	0.32	0.30	0.39	0.30	0.28
	$F_{B^3}^*$	0.28	0.47	0.44	0.53	0.54	0.50	0.35	0.50	0.45	0.42
	P_{B^3}	0.16	0.36	0.36	0.39	0.51	0.48	0.22	0.38	0.60	0.68
	R_{B^3}	1.00	0.67	0.56	0.80	0.57	0.52	0.96	0.72	0.36	0.30
<i>chars</i>	F_{B^3}	0.52	0.67	0.50	0.61	0.61	0.50	0.59	0.62	0.40	0.39
	$F_{B^3}^*$	0.57	0.75	0.60	0.71	0.69	0.61	0.64	0.71	0.50	0.49
	P_{B^3}	0.39	0.77	0.73	0.61	0.79	0.59	0.48	0.72	0.90	0.92
	R_{B^3}	1.00	0.72	0.51	0.84	0.60	0.63	0.96	0.71	0.35	0.33

Results

Measure	Baseline	VIPS	HEPS	Corm.	MMD.	Meier	MV@1	MV@2	MV@3	MV@4	
Segments	1.0	16.1	36.1	15.3	23.0	4.6	6.5	18.7	36.5	69.5	
<i>pixels</i>	F_{B^3}	0.24	0.38	0.33	0.36	0.42	0.32	0.30	0.39	0.30	0.28
	$F_{B^3}^*$	0.28	0.47	0.44	0.53	0.54	0.50	0.35	0.50	0.45	0.42
	P_{B^3}	0.16	0.36	0.36	0.39	0.51	0.48	0.22	0.38	0.60	0.68
	R_{B^3}	1.00	0.67	0.56	0.80	0.57	0.52	0.96	0.72	0.36	0.30
<i>chars</i>	F_{B^3}	0.52	0.67	0.50	0.61	0.61	0.50	0.59	0.62	0.40	0.39
	$F_{B^3}^*$	0.57	0.75	0.60	0.71	0.69	0.61	0.64	0.71	0.50	0.49
	P_{B^3}	0.39	0.77	0.73	0.61	0.79	0.59	0.48	0.72	0.90	0.92
	R_{B^3}	1.00	0.72	0.51	0.84	0.60	0.63	0.96	0.71	0.35	0.33

Results

Measure	Baseline	VIPS	HEPS	Corm.	MMD.	Meier	MV@1	MV@2	MV@3	MV@4	
Segments	1.0	16.1	36.1	15.3	23.0	4.6	6.5	18.7	36.5	69.5	
<i>pixels</i>	F_{B^3}	0.24	0.38	0.33	0.36	0.42	0.32	0.30	0.39	0.30	0.28
	$F_{B^3}^*$	0.28	0.47	0.44	0.53	0.54	0.50	0.35	0.50	0.45	0.42
	P_{B^3}	0.16	0.36	0.36	0.39	0.51	0.48	0.22	0.38	0.60	0.68
	R_{B^3}	1.00	0.67	0.56	0.80	0.57	0.52	0.96	0.72	0.36	0.30
<i>chars</i>	F_{B^3}	0.52	0.67	0.50	0.61	0.61	0.50	0.59	0.62	0.40	0.39
	$F_{B^3}^*$	0.57	0.75	0.60	0.71	0.69	0.61	0.64	0.71	0.50	0.49
	P_{B^3}	0.39	0.77	0.73	0.61	0.79	0.59	0.48	0.72	0.90	0.92
	R_{B^3}	1.00	0.72	0.51	0.84	0.60	0.63	0.96	0.71	0.35	0.33

Results

Measure	Baseline	VIPS	HEPS	Corm.	MMD.	Meier	MV@1	MV@2	MV@3	MV@4	
Segments	1.0	16.1	36.1	15.3	23.0	4.6	6.5	18.7	36.5	69.5	
<i>pixels</i>	F_{B^3}	0.24	0.38	0.33	0.36	0.42	0.32	0.30	0.39	0.30	0.28
	$F_{B^3}^*$	0.28	0.47	0.44	0.53	0.54	0.50	0.35	0.50	0.45	0.42
	P_{B^3}	0.16	0.36	0.36	0.39	0.51	0.48	0.22	0.38	0.60	0.68
	R_{B^3}	1.00	0.67	0.56	0.80	0.57	0.52	0.96	0.72	0.36	0.30
<i>chars</i>	F_{B^3}	0.52	0.67	0.50	0.61	0.61	0.50	0.59	0.62	0.40	0.39
	$F_{B^3}^*$	0.57	0.75	0.60	0.71	0.69	0.61	0.64	0.71	0.50	0.49
	P_{B^3}	0.39	0.77	0.73	0.61	0.79	0.59	0.48	0.72	0.90	0.92
	R_{B^3}	1.00	0.72	0.51	0.84	0.60	0.63	0.96	0.71	0.35	0.33

Results

Measure	Baseline	VIPS	HEPS	Corm.	MMD.	Meier	MV@1	MV@2	MV@3	MV@4	
Segments	1.0	16.1	36.1	15.3	23.0	4.6	6.5	18.7	36.5	69.5	
<i>pixels</i>	F_{B^3}	0.24	0.38	0.33	0.36	0.42	0.32	0.30	0.39	0.30	0.28
	$F_{B^3}^*$	0.28	0.47	0.44	0.53	0.54	0.50	0.35	0.50	0.45	0.42
	P_{B^3}	0.16	0.36	0.36	0.39	0.51	0.48	0.22	0.38	0.60	0.68
	R_{B^3}	1.00	0.67	0.56	0.80	0.57	0.52	0.96	0.72	0.36	0.30
<i>chars</i>	F_{B^3}	0.52	0.67	0.50	0.61	0.61	0.50	0.59	0.62	0.40	0.39
	$F_{B^3}^*$	0.57	0.75	0.60	0.71	0.69	0.61	0.64	0.71	0.50	0.49
	P_{B^3}	0.39	0.77	0.73	0.61	0.79	0.59	0.48	0.72	0.90	0.92
	R_{B^3}	1.00	0.72	0.51	0.84	0.60	0.63	0.96	0.71	0.35	0.33

Results

Measure	Baseline	VIPS	HEPS	Corm.	MMD.	Meier	MV@1	MV@2	MV@3	MV@4	
Segments	1.0	16.1	36.1	15.3	23.0	4.6	6.5	18.7	36.5	69.5	
<i>pixels</i>	F_{B^3}	0.24	0.38	0.33	0.36	0.42	0.32	0.30	0.39	0.30	0.28
	$F_{B^3}^*$	0.28	0.47	0.44	0.53	0.54	0.50	0.35	0.50	0.45	0.42
	P_{B^3}	0.16	0.36	0.36	0.39	0.51	0.48	0.22	0.38	0.60	0.68
	R_{B^3}	1.00	0.67	0.56	0.80	0.57	0.52	0.96	0.72	0.36	0.30
<i>chars</i>	F_{B^3}	0.52	0.67	0.50	0.61	0.61	0.50	0.59	0.62	0.40	0.39
	$F_{B^3}^*$	0.57	0.75	0.60	0.71	0.69	0.61	0.64	0.71	0.50	0.49
	P_{B^3}	0.39	0.77	0.73	0.61	0.79	0.59	0.48	0.72	0.90	0.92
	R_{B^3}	1.00	0.72	0.51	0.84	0.60	0.63	0.96	0.71	0.35	0.33

Results

Measure	Baseline	VIPS	HEPS	Corm.	MMD.	Meier	MV@1	MV@2	MV@3	MV@4	
Segments	1.0	16.1	36.1	15.3	23.0	4.6	6.5	18.7	36.5	69.5	
<i>pixels</i>	F_{B^3}	0.24	0.38	0.33	0.36	0.42	0.32	0.30	0.39	0.30	0.28
	$F_{B^3}^*$	0.28	0.47	0.44	0.53	0.54	0.50	0.35	0.50	0.45	0.42
	P_{B^3}	0.16	0.36	0.36	0.39	0.51	0.48	0.22	0.38	0.60	0.68
	R_{B^3}	1.00	0.67	0.56	0.80	0.57	0.52	0.96	0.72	0.36	0.30
<i>chars</i>	F_{B^3}	0.52	0.67	0.50	0.61	0.61	0.50	0.59	0.62	0.40	0.39
	$F_{B^3}^*$	0.57	0.75	0.60	0.71	0.69	0.61	0.64	0.71	0.50	0.49
	P_{B^3}	0.39	0.77	0.73	0.61	0.79	0.59	0.48	0.72	0.90	0.92
	R_{B^3}	1.00	0.72	0.51	0.84	0.60	0.63	0.96	0.71	0.35	0.33

Results

Measure	Baseline	VIPS	HEPS	Corm.	MMD.	Meier	MV@1	MV@2	MV@3	MV@4	
Segments	1.0	16.1	36.1	15.3	23.0	4.6	6.5	18.7	36.5	69.5	
<i>pixels</i>	F_{B^3}	0.24	0.38	0.33	0.36	0.42	0.32	0.30	0.39	0.30	0.28
	$F_{B^3}^*$	0.28	0.47	0.44	0.53	0.54	0.50	0.35	0.50	0.45	0.42
	P_{B^3}	0.16	0.36	0.36	0.39	0.51	0.48	0.22	0.38	0.60	0.68
	R_{B^3}	1.00	0.67	0.56	0.80	0.57	0.52	0.96	0.72	0.36	0.30
<i>chars</i>	F_{B^3}	0.52	0.67	0.50	0.61	0.61	0.50	0.59	0.62	0.40	0.39
	$F_{B^3}^*$	0.57	0.75	0.60	0.71	0.69	0.61	0.64	0.71	0.50	0.49
	P_{B^3}	0.39	0.77	0.73	0.61	0.79	0.59	0.48	0.72	0.90	0.92
	R_{B^3}	1.00	0.72	0.51	0.84	0.60	0.63	0.96	0.71	0.35	0.33

Results

Measure	Baseline	VIPS	HEPS	Corm.	MMD.	Meier	MV@1	MV@2	MV@3	MV@4	
Segments	1.0	16.1	36.1	15.3	23.0	4.6	6.5	18.7	36.5	69.5	
<i>pixels</i>	F_{B^3}	0.24	0.38	0.33	0.36	0.42	0.32	0.30	0.39	0.30	0.28
	$F_{B^3}^*$	0.28	0.47	0.44	0.53	0.54	0.50	0.35	0.50	0.45	0.42
	P_{B^3}	0.16	0.36	0.36	0.39	0.51	0.48	0.22	0.38	0.60	0.68
	R_{B^3}	1.00	0.67	0.56	0.80	0.57	0.52	0.96	0.72	0.36	0.30
<i>chars</i>	F_{B^3}	0.52	0.67	0.50	0.61	0.61	0.50	0.59	0.62	0.40	0.39
	$F_{B^3}^*$	0.57	0.75	0.60	0.71	0.69	0.61	0.64	0.71	0.50	0.49
	P_{B^3}	0.39	0.77	0.73	0.61	0.79	0.59	0.48	0.72	0.90	0.92
	R_{B^3}	1.00	0.72	0.51	0.84	0.60	0.63	0.96	0.71	0.35	0.33

Conclusion

- ❑ Empirical evaluation of
 - 5 web page segmentation algorithms on
 - 8490 web pages
- ❑ Usage of web archiving technology for reproducibility
- ❑ VIPS performs best overall, but not for *pixel* segments
- ❑ Competitive performance for purely visual approaches
- ❑ When fitted to DOM nodes, also a generic object detection algorithm trained on photos performs competitively