An Empirical Comparison of Web Page Segmentation Algorithms

ECIR 2021





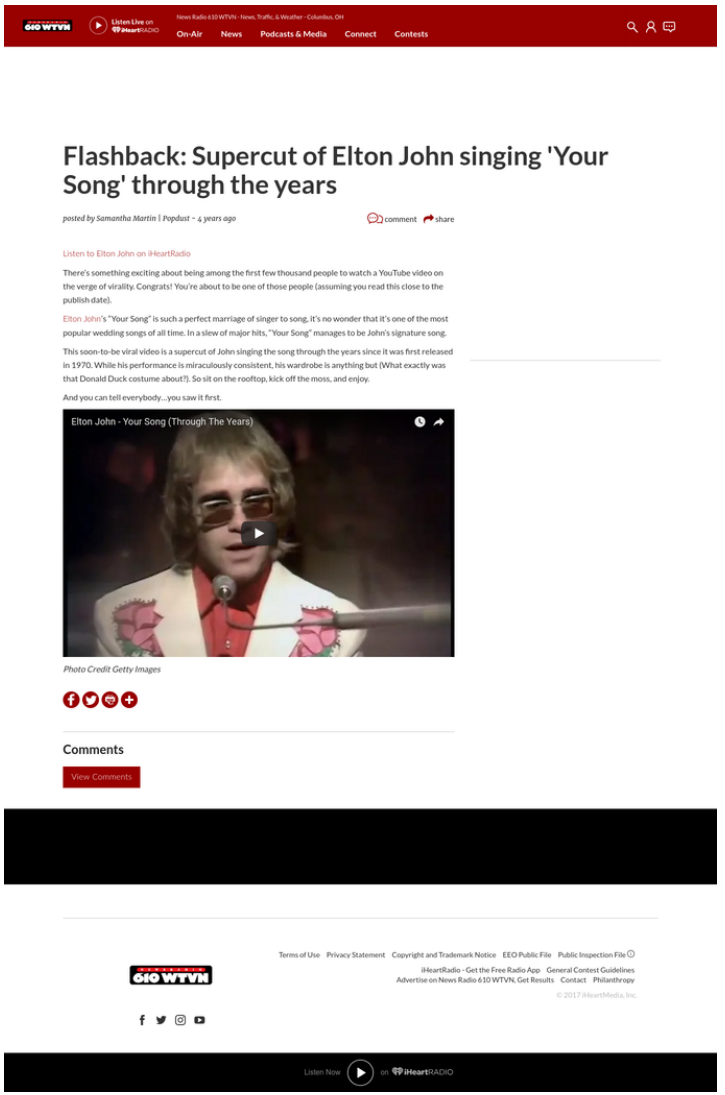
Johannes Kiesel[1] Lars Meyer[1] Florian Kneist[1] Benno Stein[1] Martin Potthast[2]

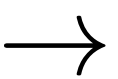
[1] Bauhaus-Universität Weimar

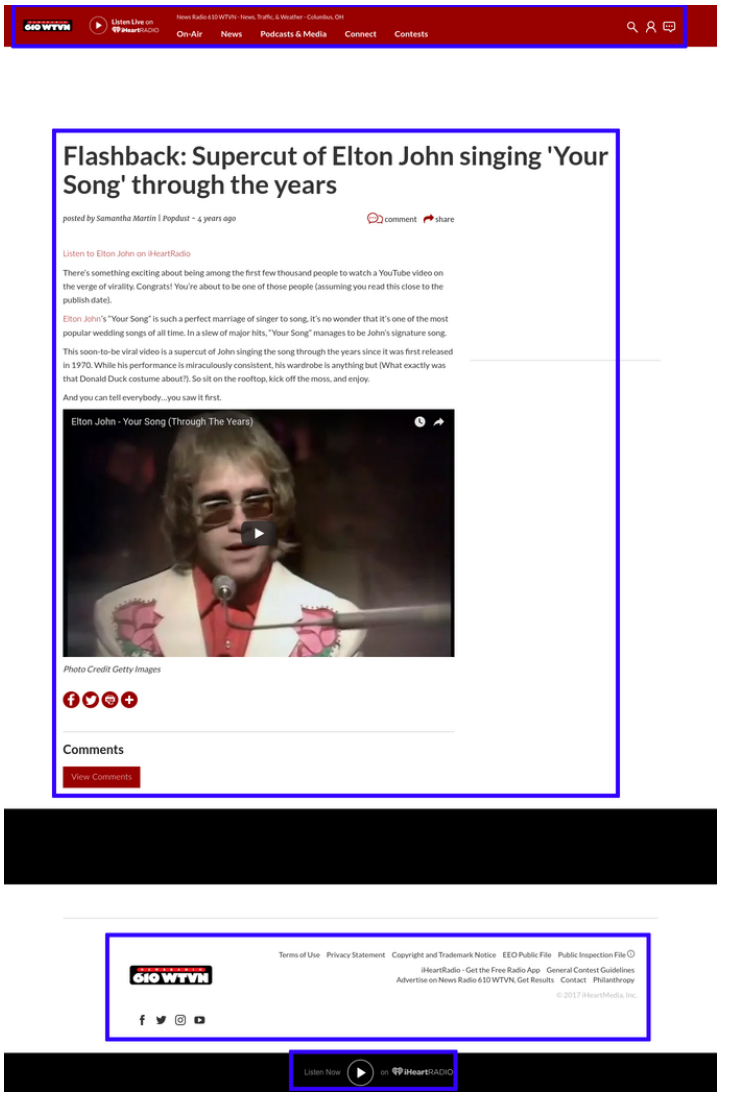
[2] Universität Leipzig

[1,2] Webis

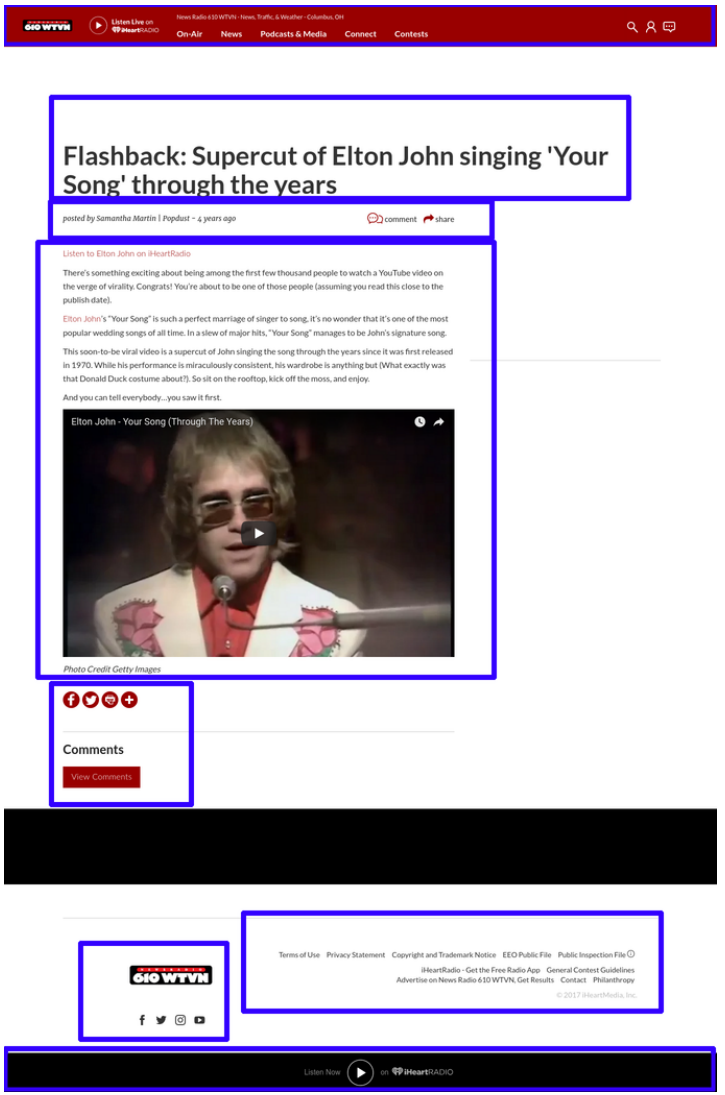
Web Page Segmentation

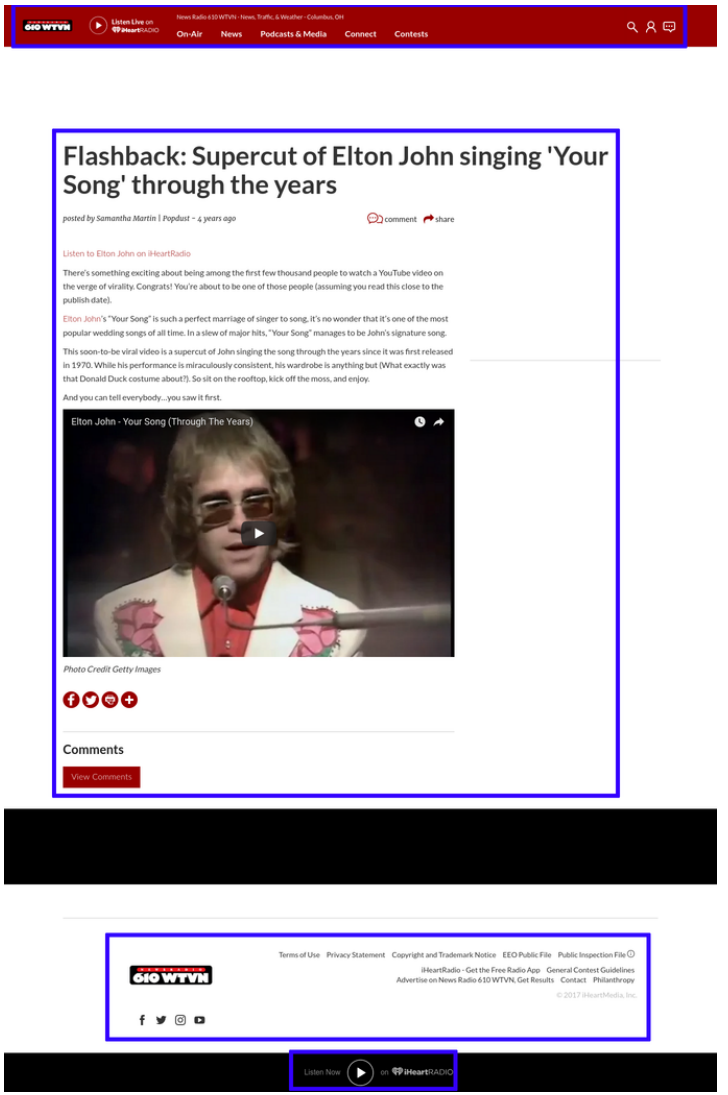




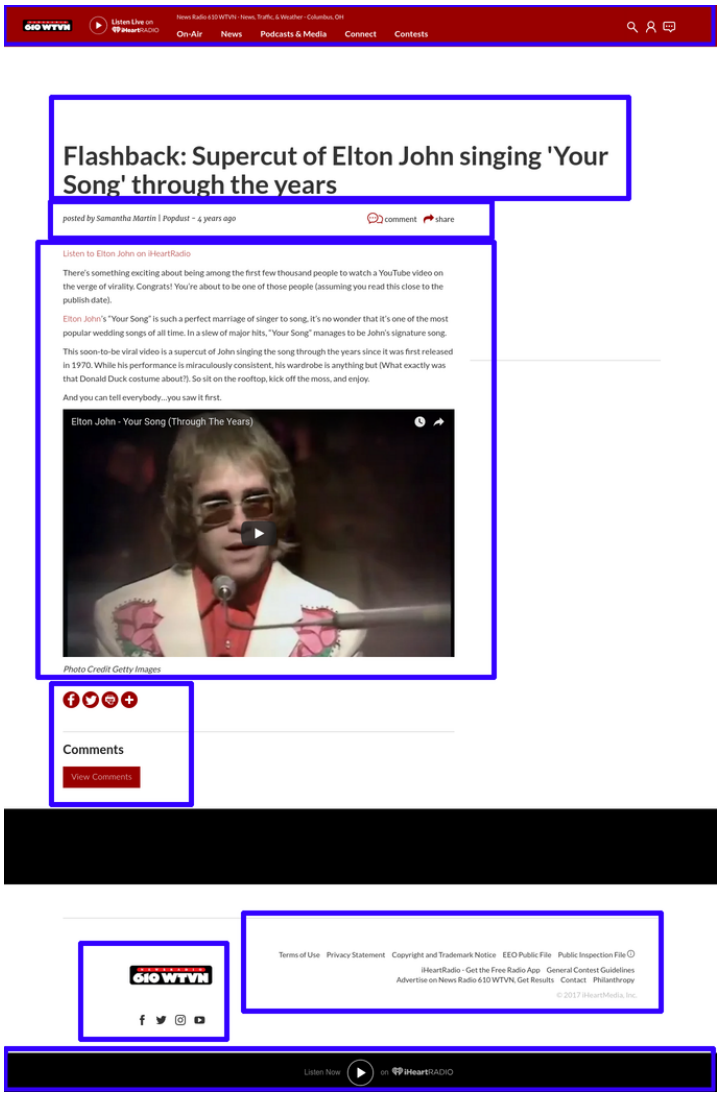


Web Page Segmentation

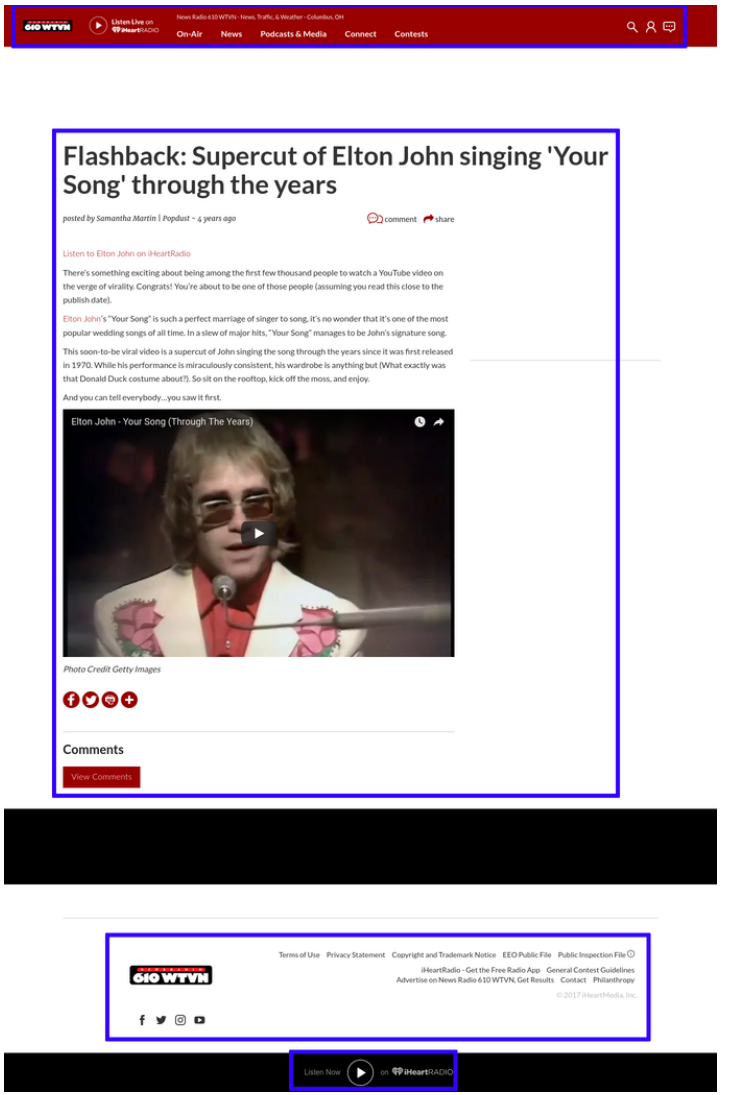




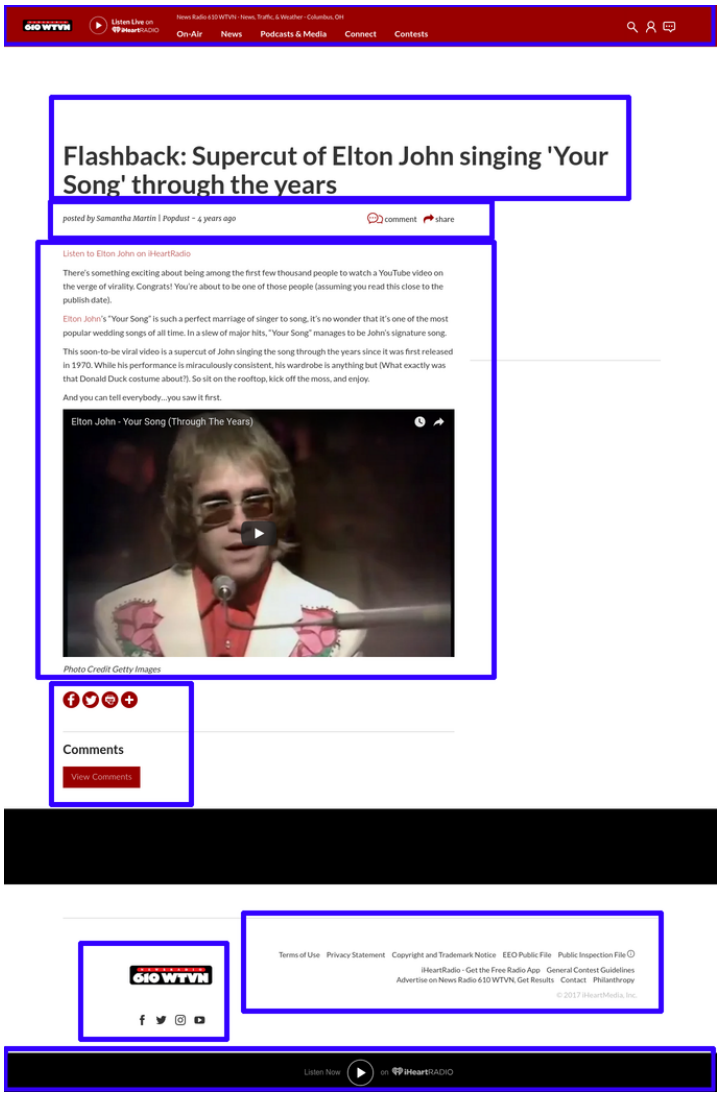
Web Page Segmentation



?

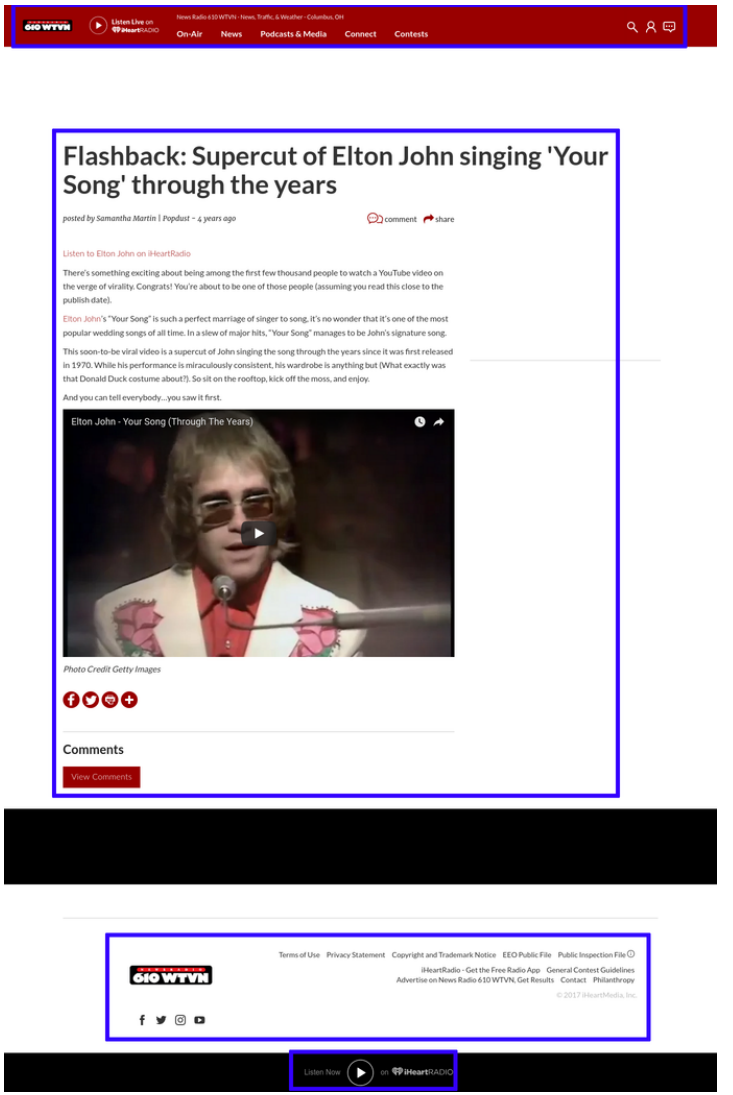


Web Page Segmentation

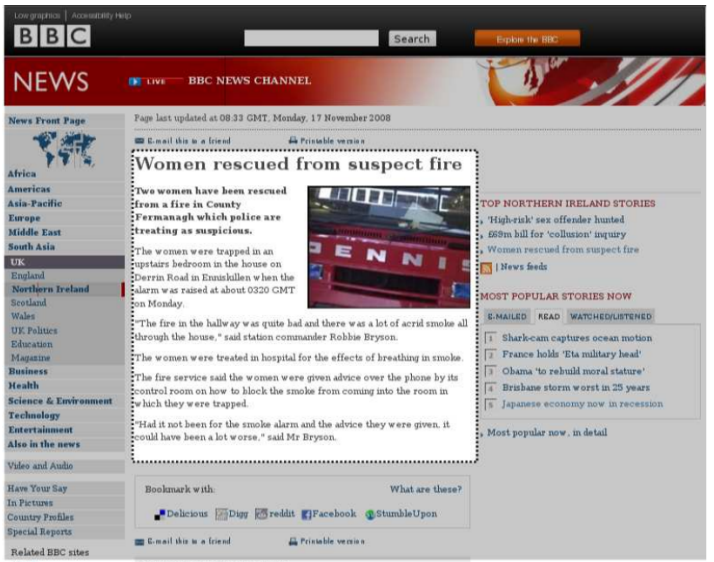


?

≈

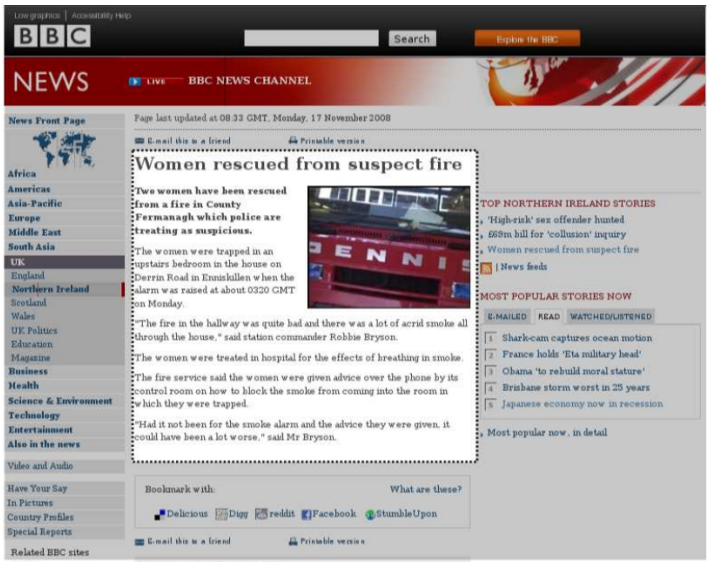


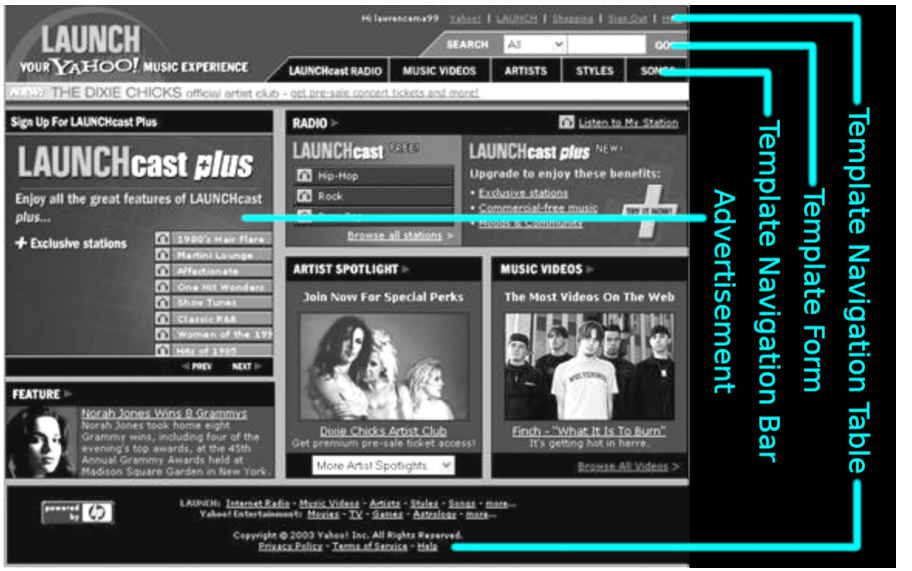
Web Page Segmentation: Downstream Tasks (Examples)



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

Web Page Segmentation: Downstream Tasks (Examples)

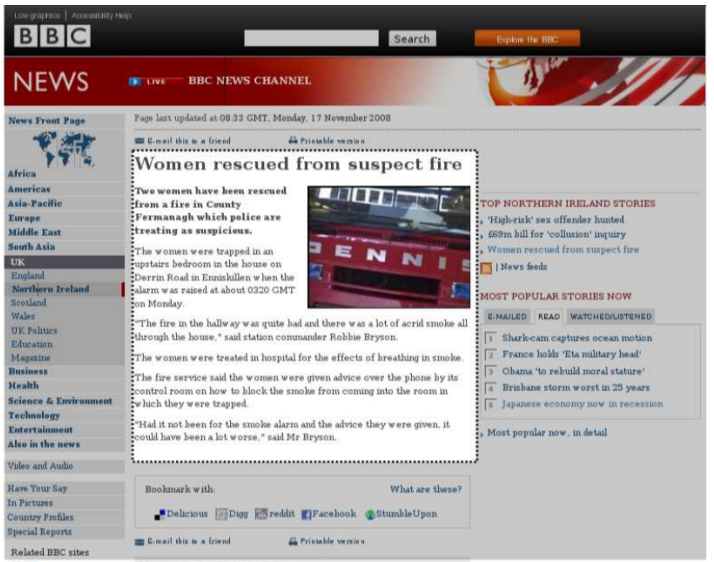


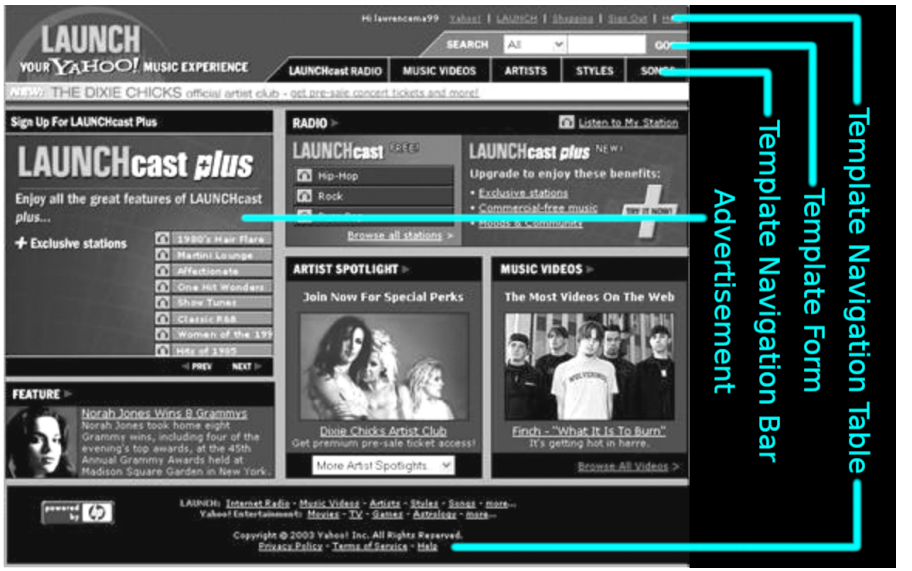


- **Content Extraction**
  Image: *Language Independent Content Extraction from Web Pages.* Javier et al., DIR'09.

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

Web Page Segmentation: Downstream Tasks (Examples)

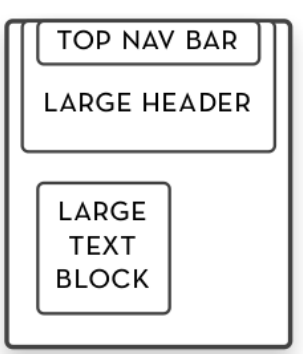


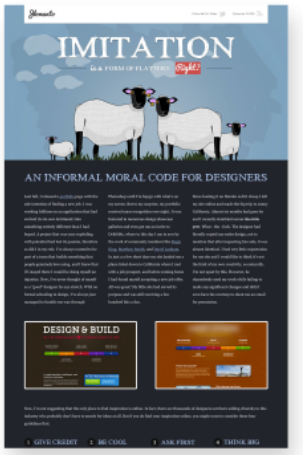


- **Content Extraction**
  Image: *Language Independent Content Extraction from Web Pages.* Javier et al., DIR'09.

- **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.



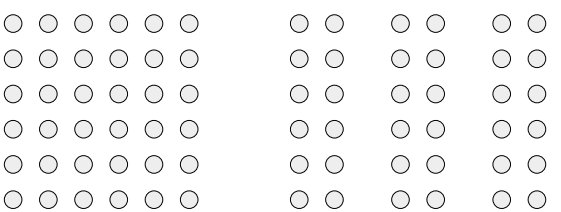


Concept Formation: Web Page Segment

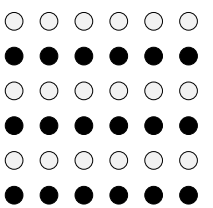
A web page segment is a part of a web page containing those elements that belong together as per agreement among a majority of viewers.

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

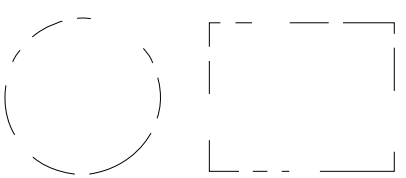
Gestalt Principles provide common ground



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

Segmentation $S = \{s_1, \ldots, s_m\}$ 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.

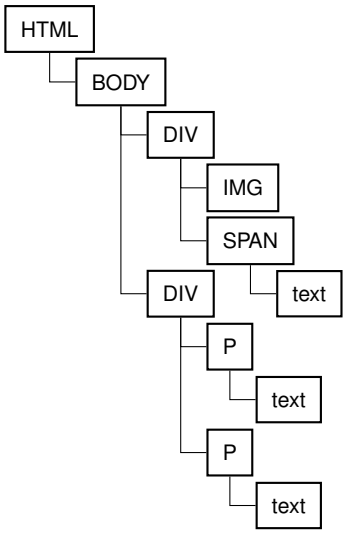
Elements $E = \{e_1, \ldots, e_n\}$

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

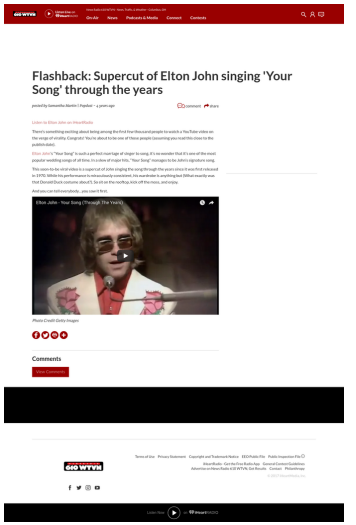
Suggested sets of elements:

Listen Live on iHeartRADIO News Radio 610WTVN-News, Traffic, Weather - Columbus, OH On-Air News Podcasts Media Connect Contests Flashback: 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 There's something exciting about being among the

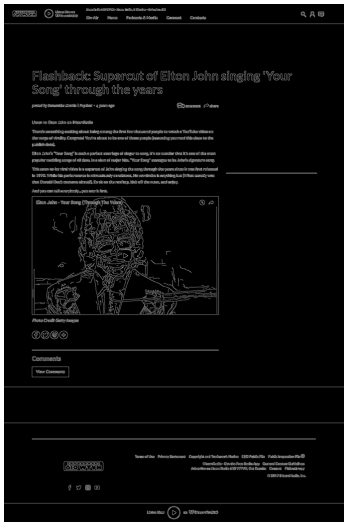
Characters



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

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

Precision

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

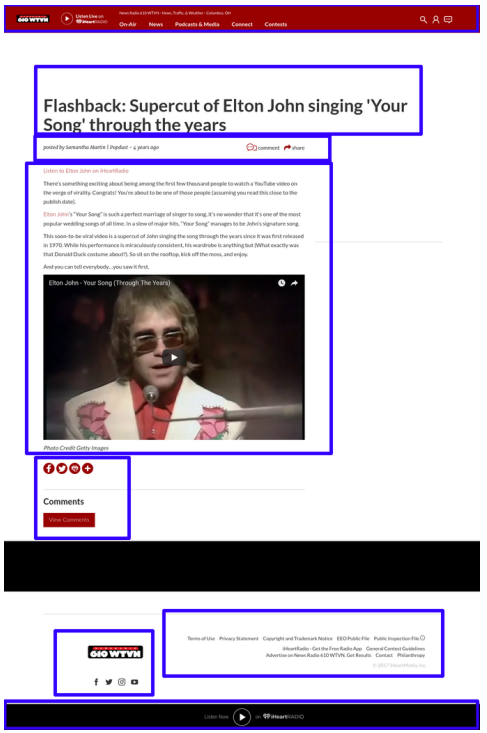
Recall

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

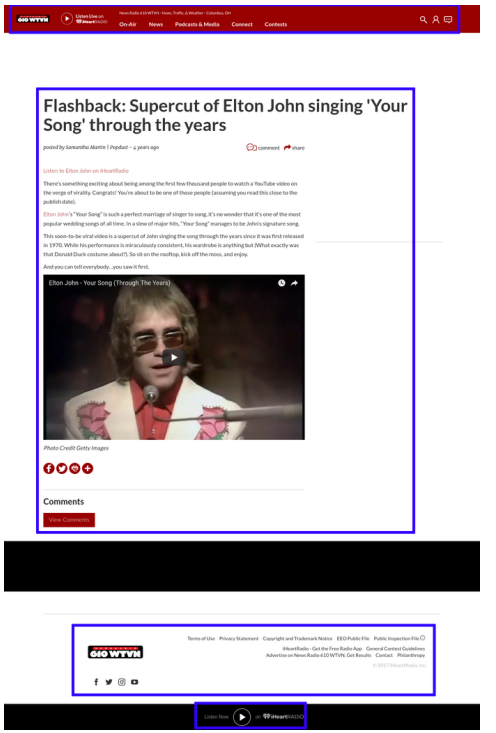
F-Measure, F_{B^3}, is defined as the harmonic mean of precision and recall as usual

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

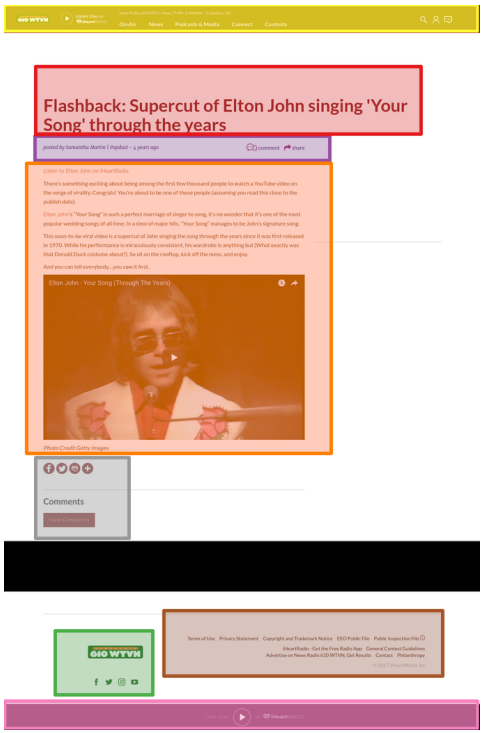
Evaluation Framework for Web Page Segmentation

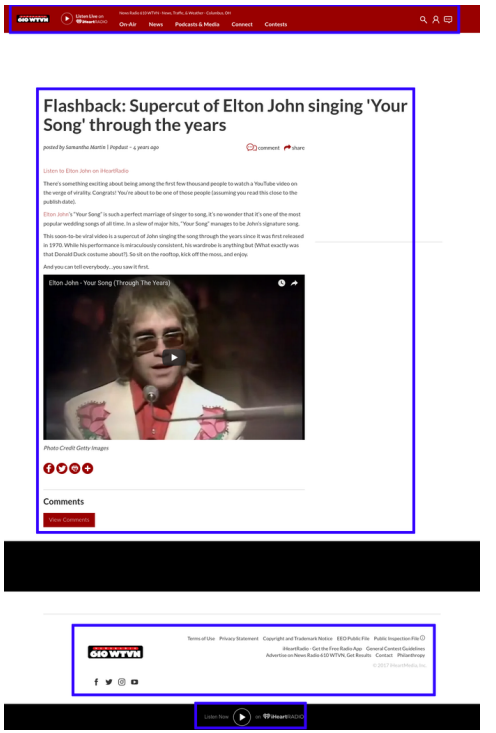


$\overset{?}{\approx}$



Evaluation Framework for Web Page Segmentation

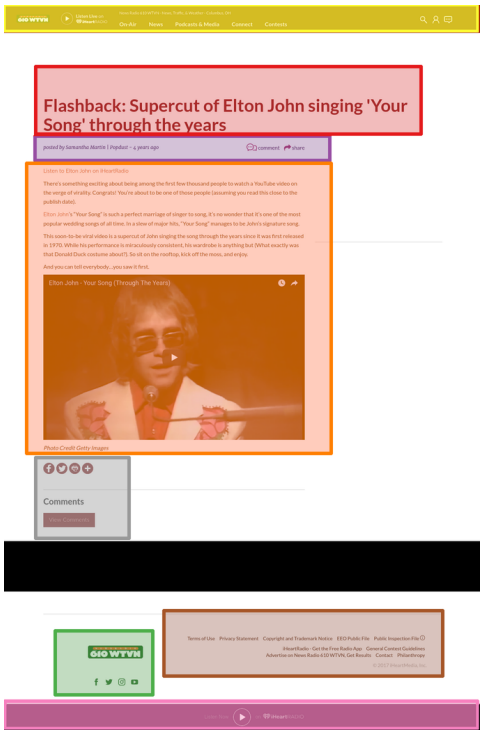




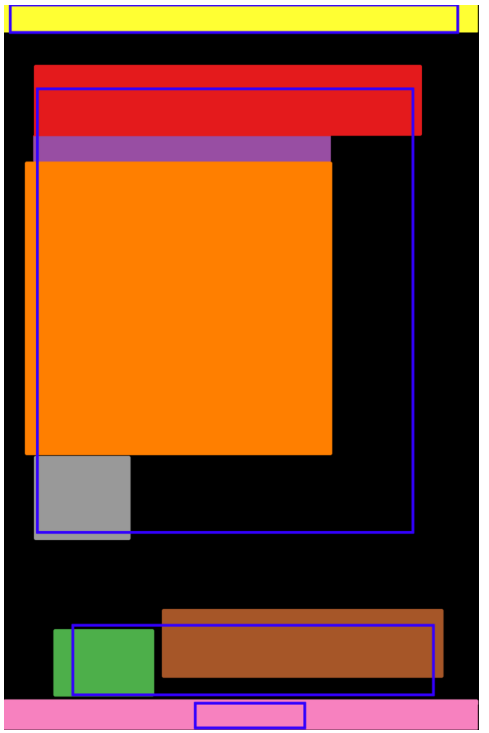
S Characters of S in S^* S^*

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

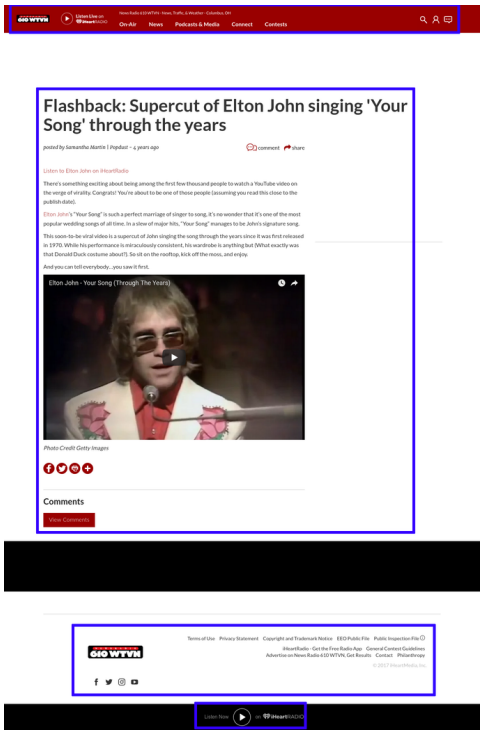
Evaluation Framework for Web Page Segmentation



S



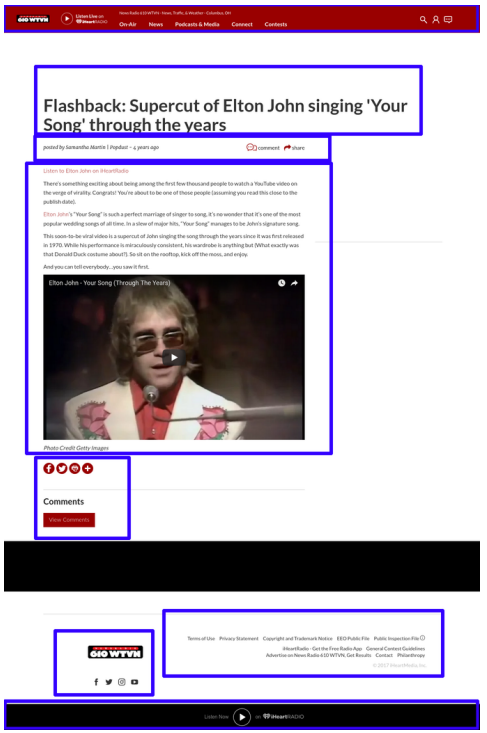
Pixels of S in S^*

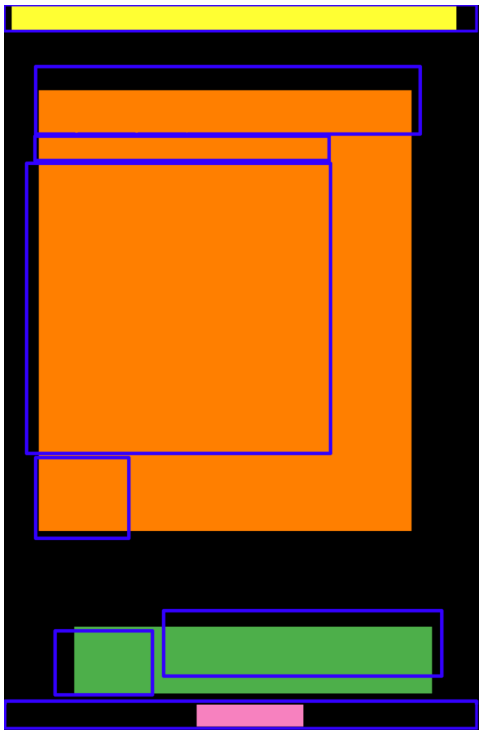


S^*

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

Evaluation Framework for Web Page Segmentation



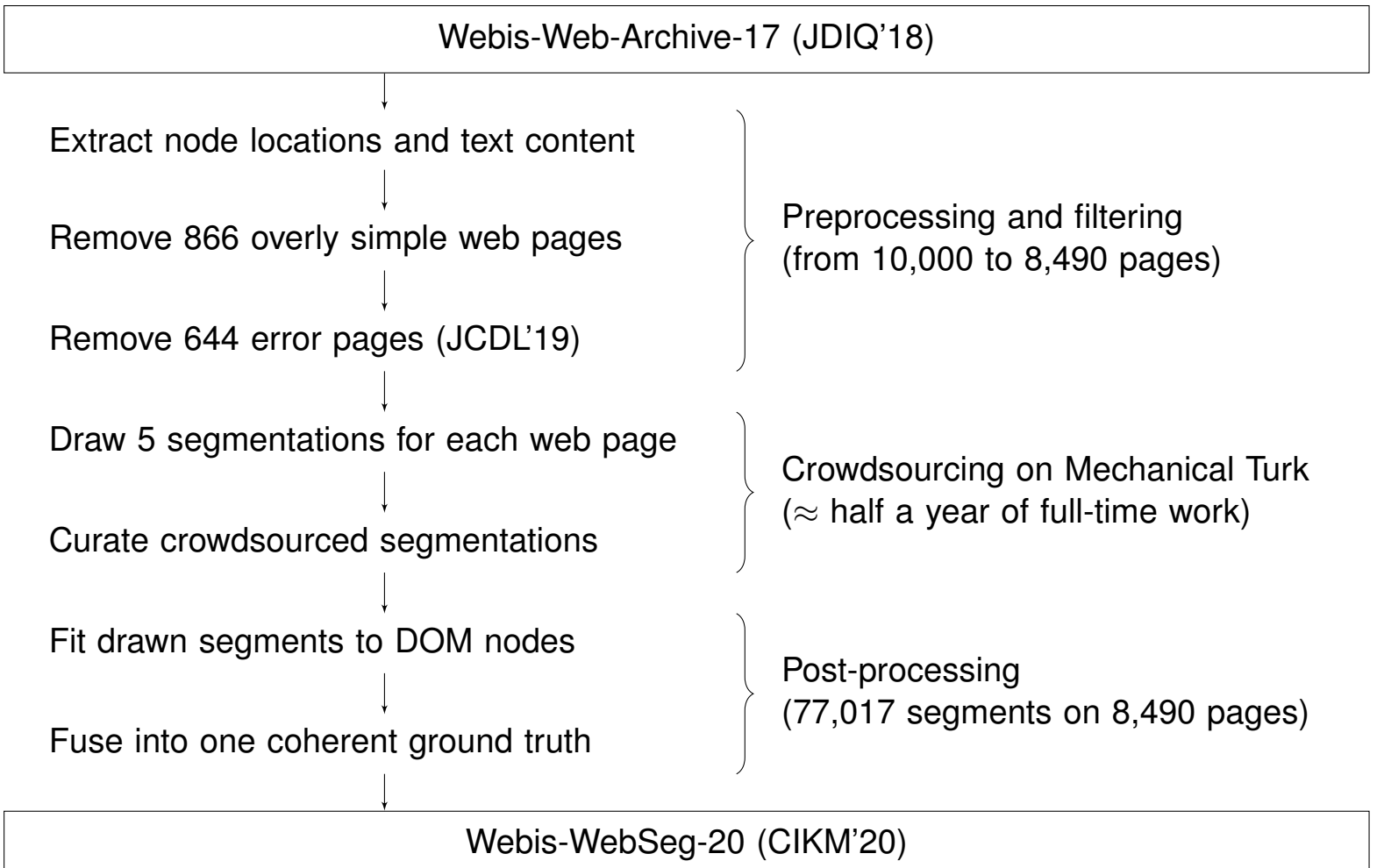


S Pixels of S^* in S S^*

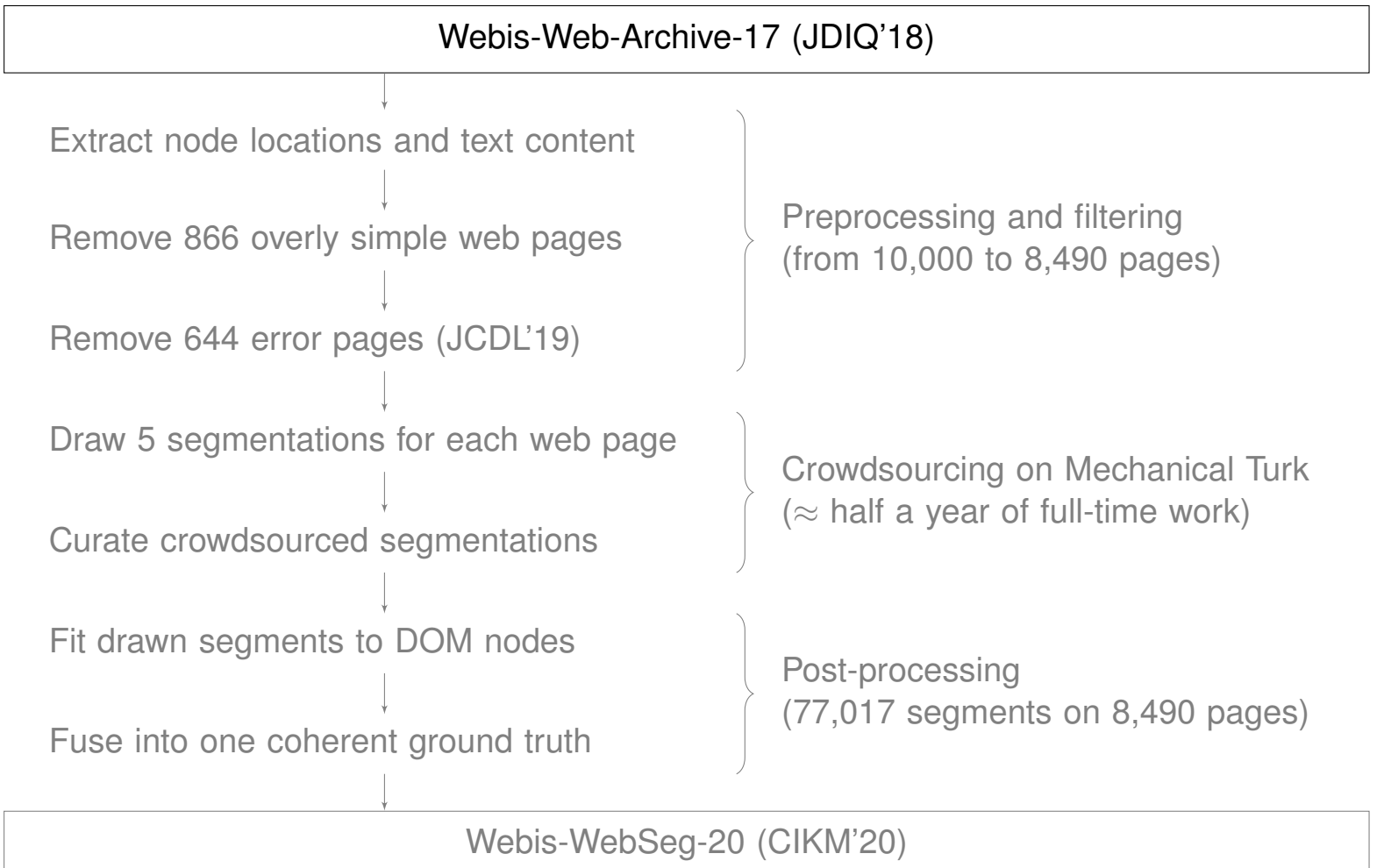
$$P_{B^3}(S, S^*) = \text{avg}_e \left(\frac{|\text{elements in same segment as } e \text{ in both } S \text{ and } S^*|}{|\text{elements in same segment as } e \text{ in } S|} \right)$$

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

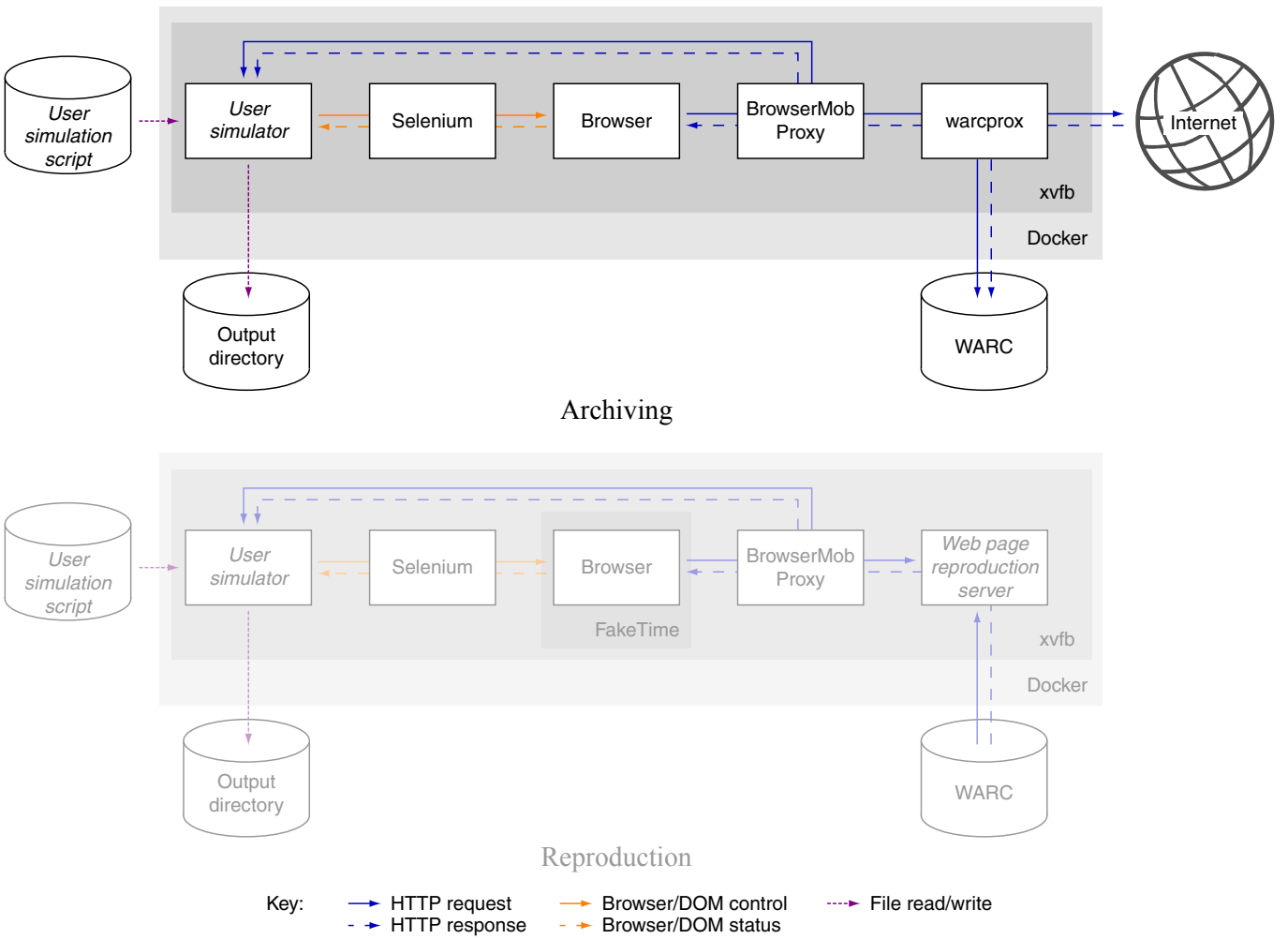
The Webis-WebSeg-20 Dataset



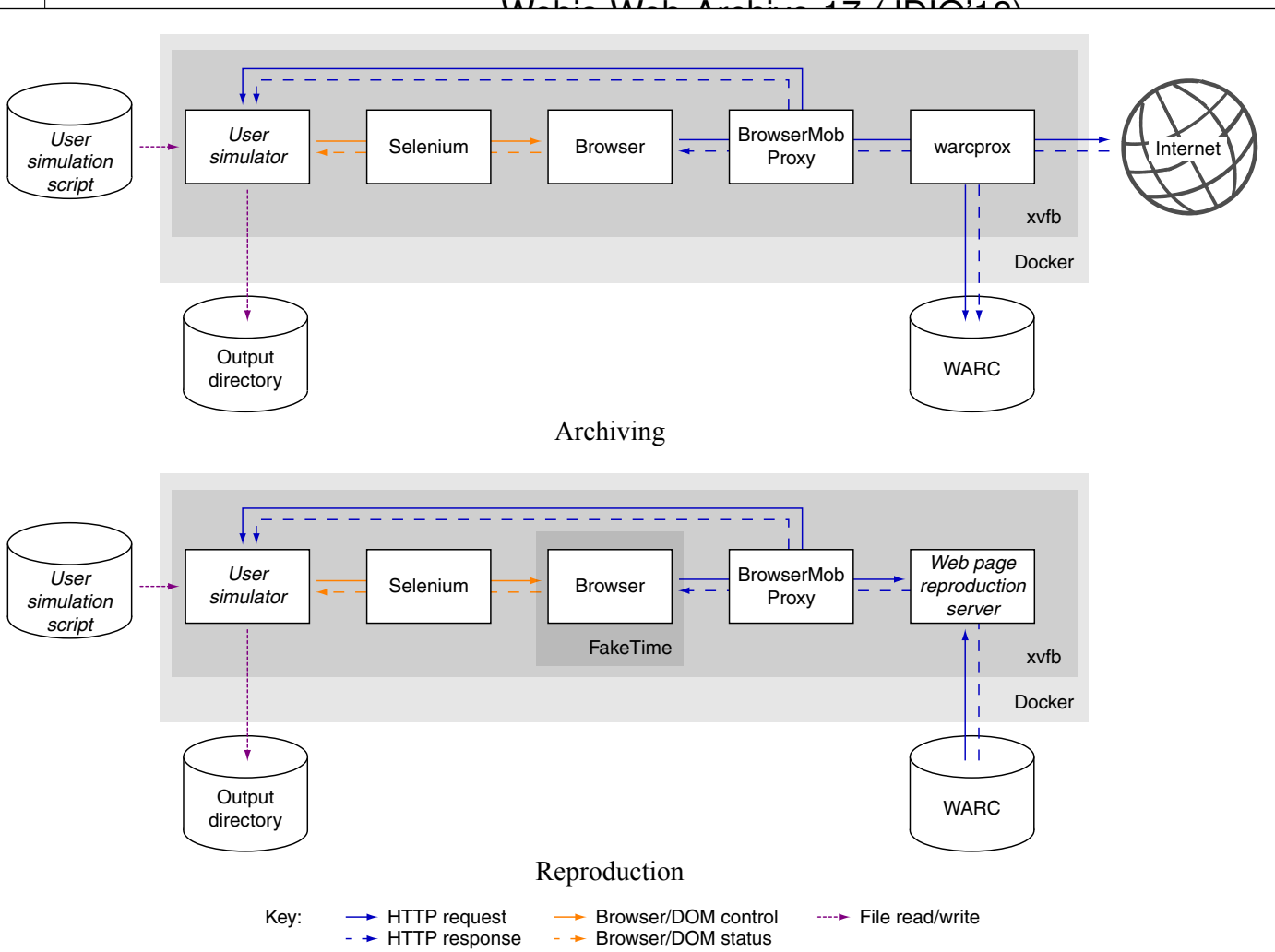
The Webis-WebSeg-20 Dataset



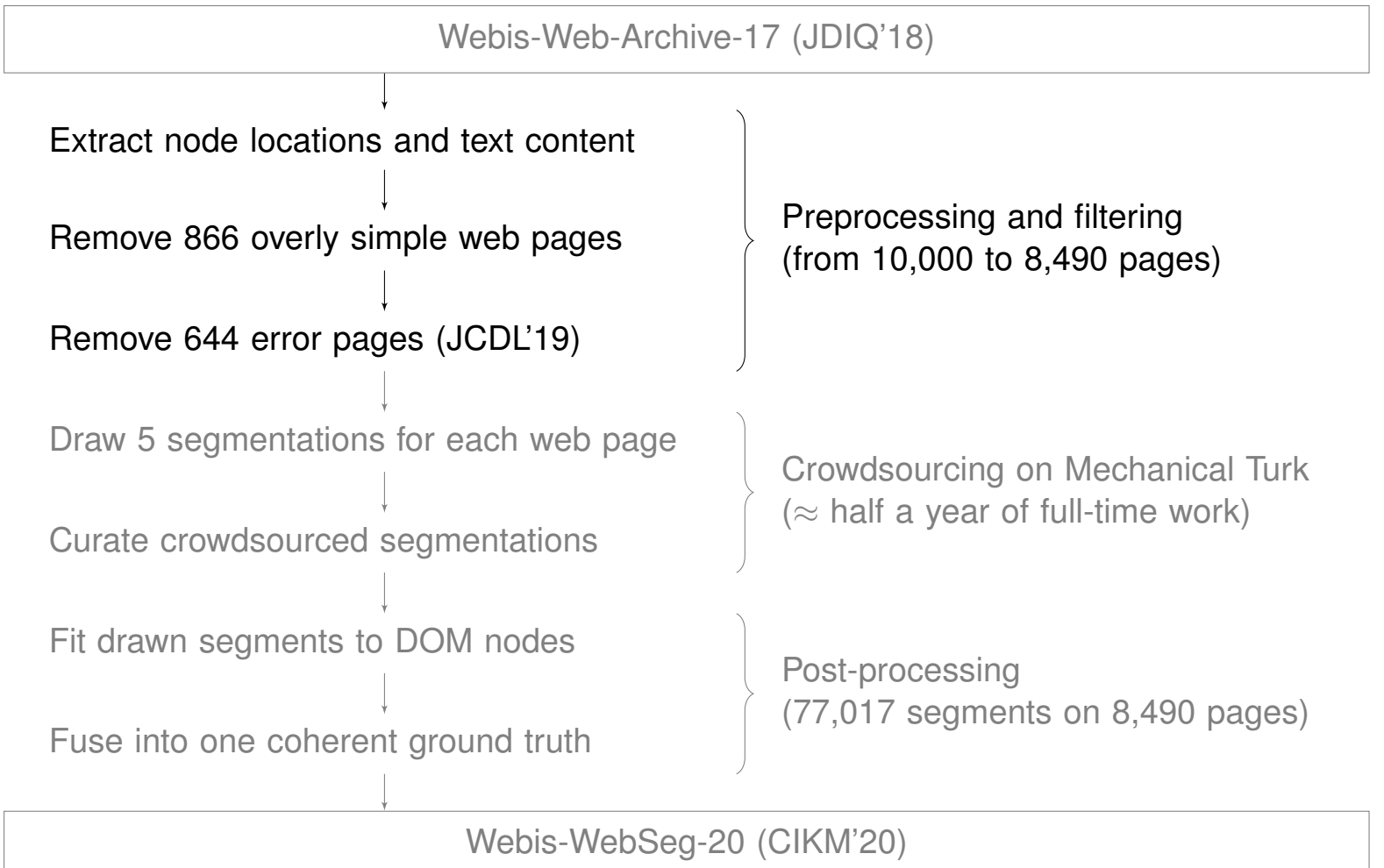
The Webis-WebSeg-20 Dataset



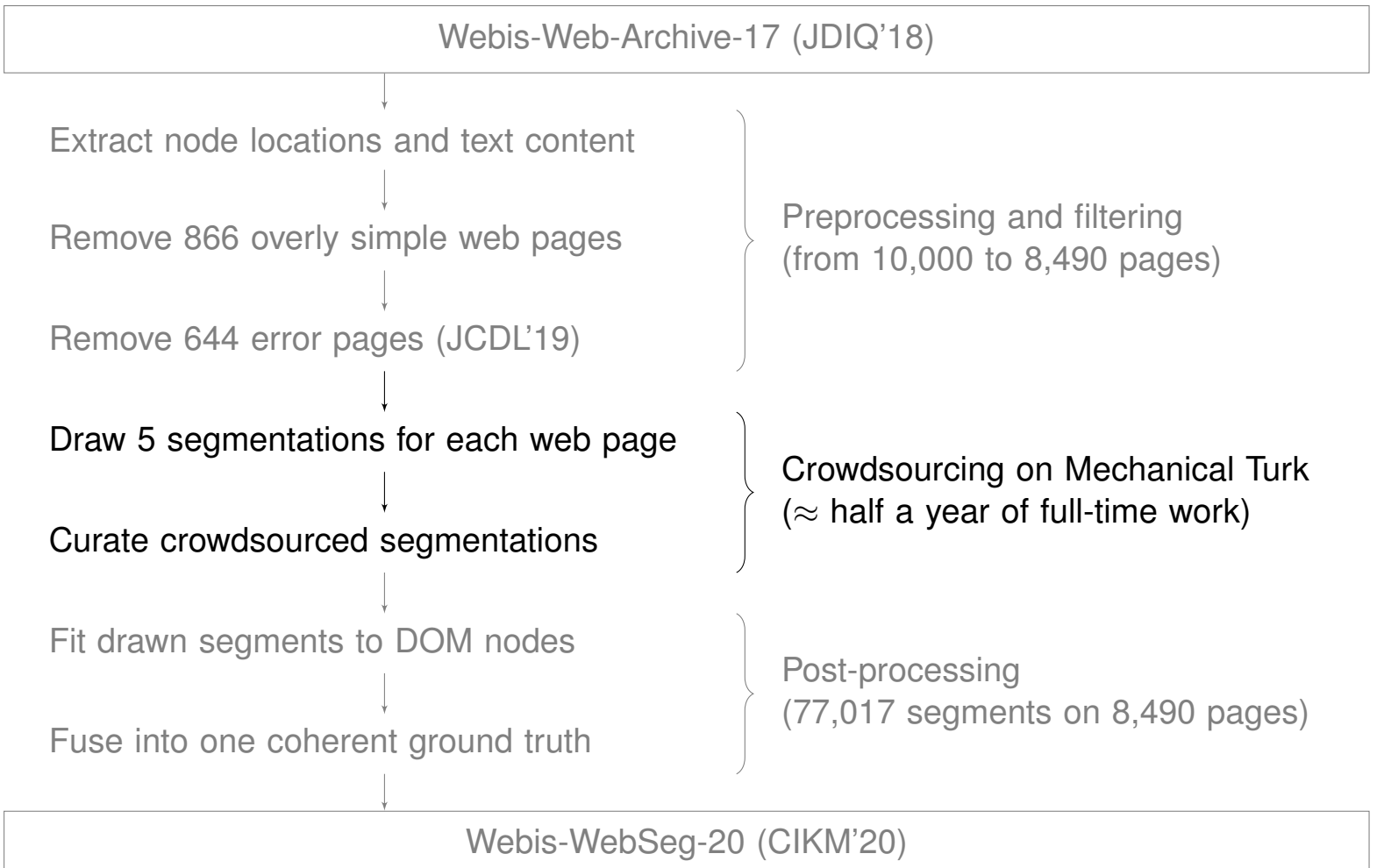
The Webis-WebSeg-20 Dataset



The Webis-WebSeg-20 Dataset



The Webis-WebSeg-20 Dataset

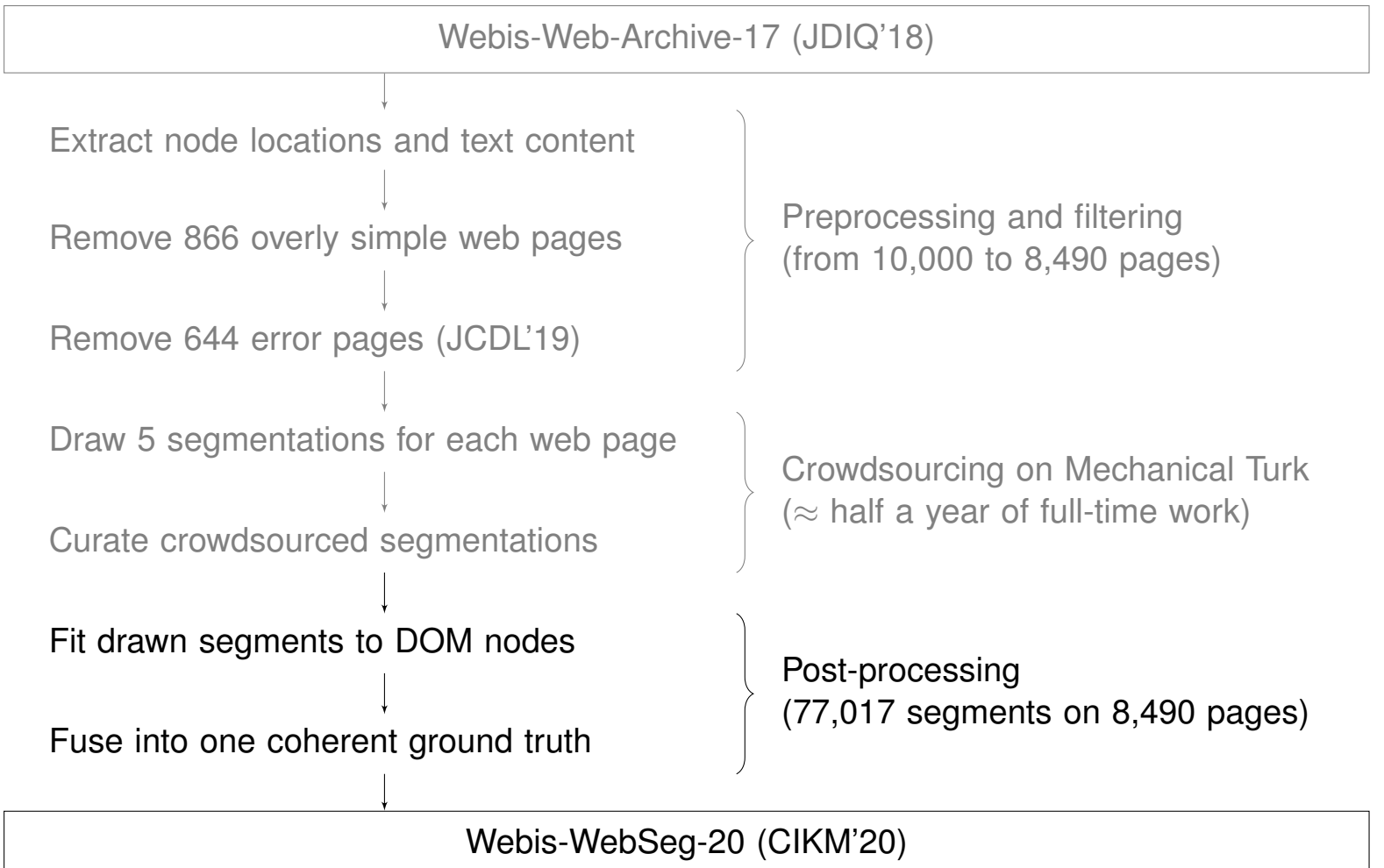


The Webis-WebSeg-20 Dataset

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

Extract node locations and text content

The Webis-WebSeg-20 Dataset

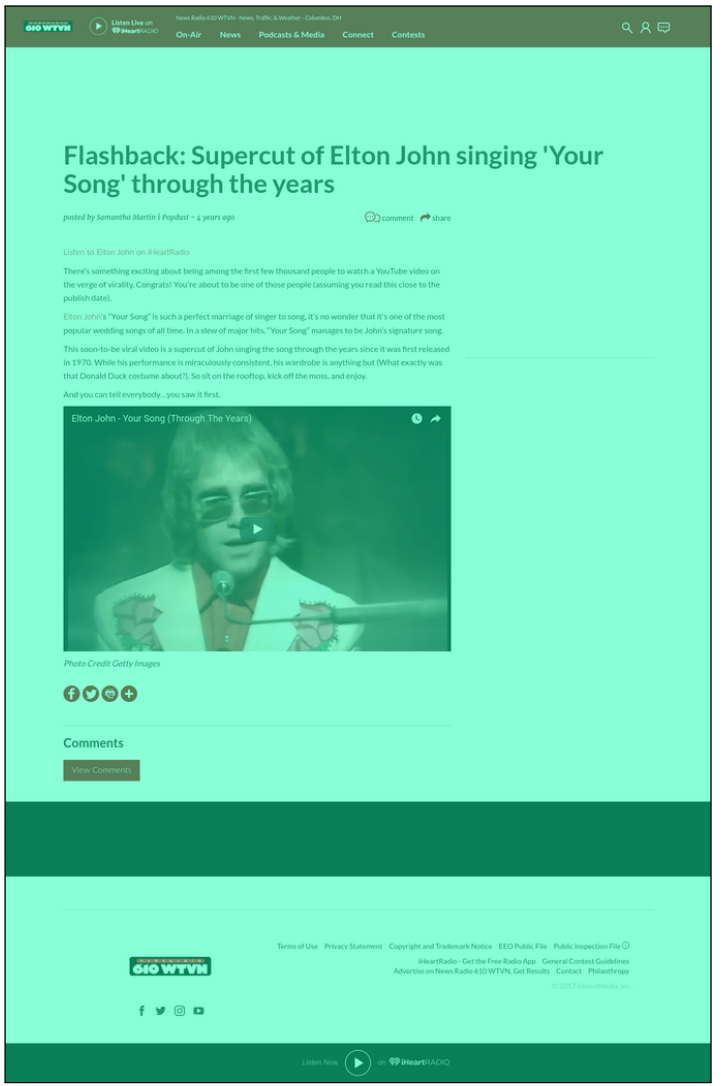


Algorithms

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

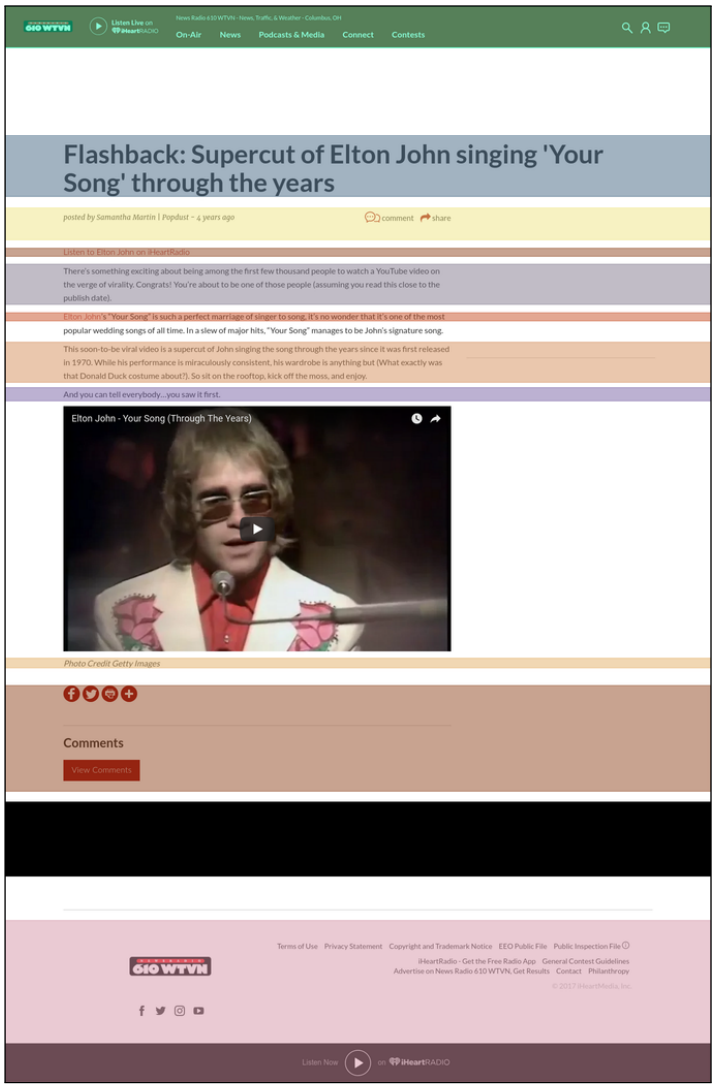


VIPS

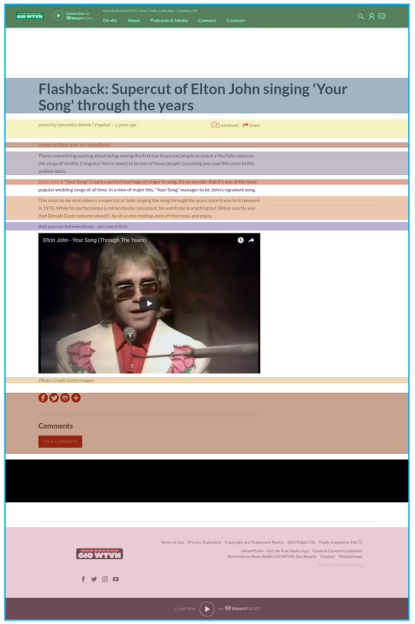
“VIsion-based Page Segmentation algorithm” by Cai et al., 2003

- 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

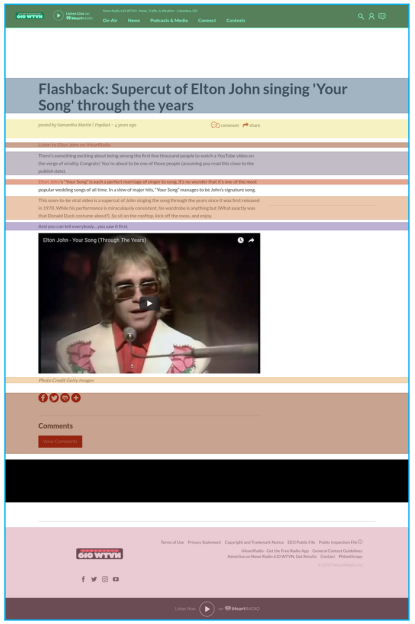


VIPS: Optimization for Permitted Degree of Coherence (PDoC)

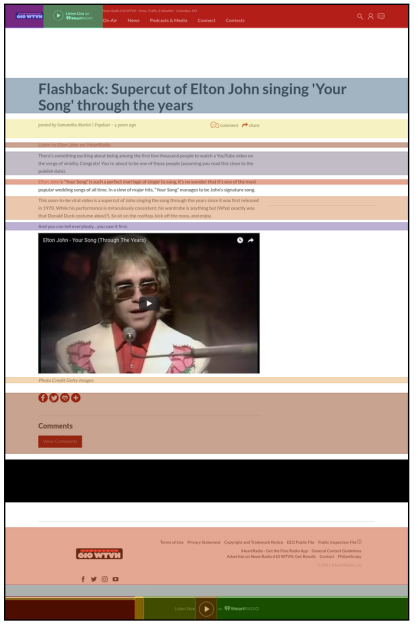


PDoC $\in [1, 6]$

VIPS: Optimization for Permitted Degree of Coherence (PDoC)

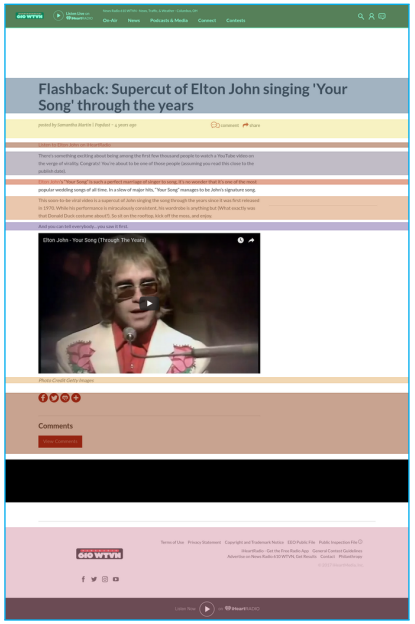


PDoC $\in [1, 6]$

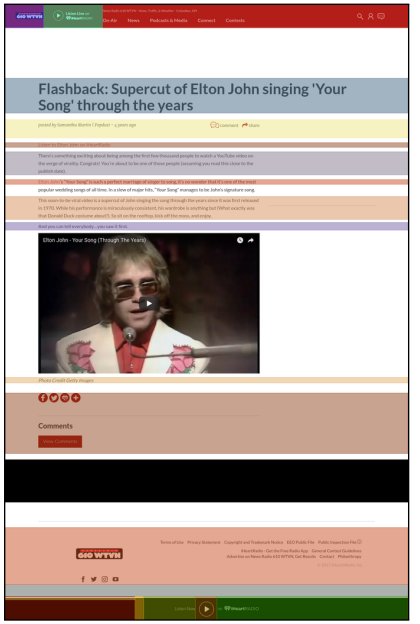


PDoC $= 7$

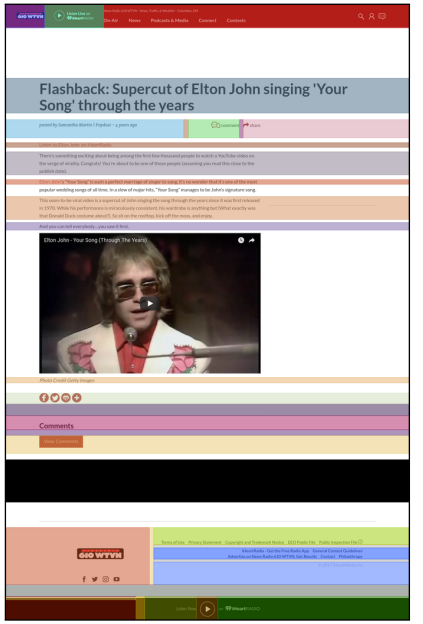
VIPS: Optimization for Permitted Degree of Coherence (PDoC)



PDoC $\in [1, 6]$

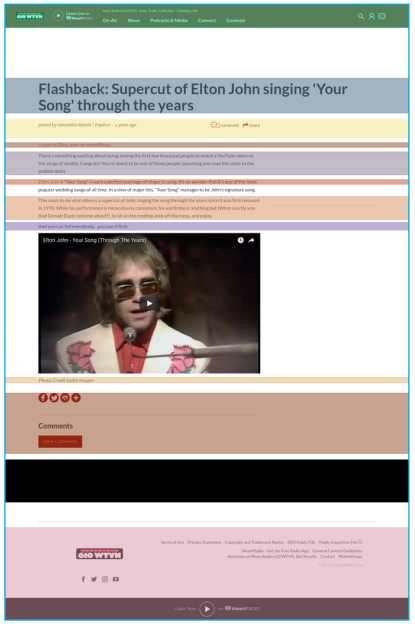


PDoC $= 7$

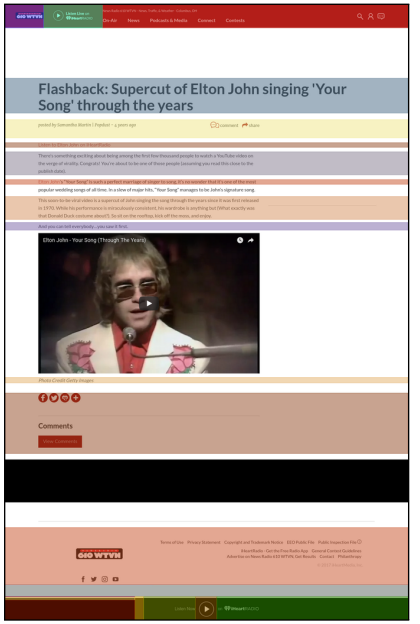


PDoC $\in [8, 9]$

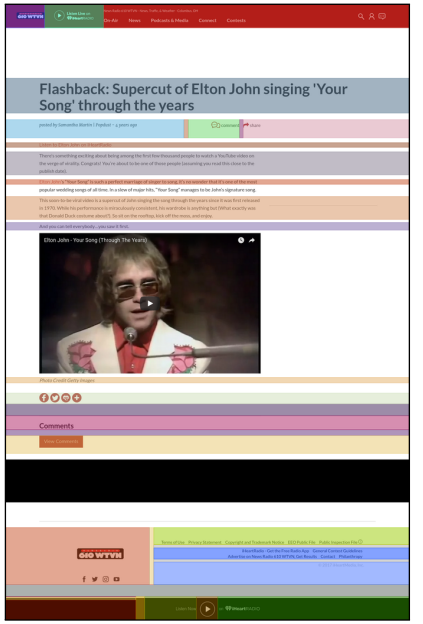
VIPS: Optimization for Permitted Degree of Coherence (PDoC)



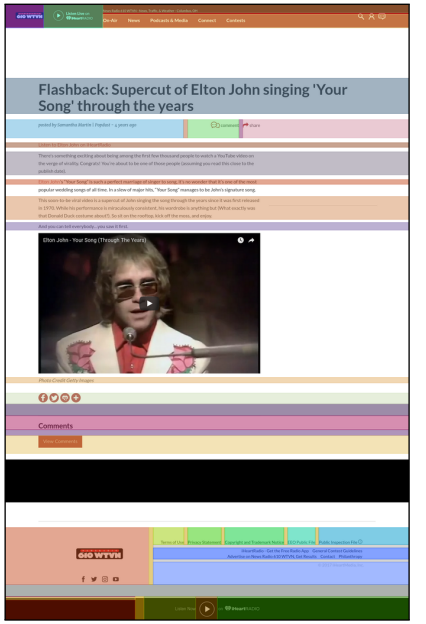
PDoC $\in [1, 6]$



PDoC $= 7$

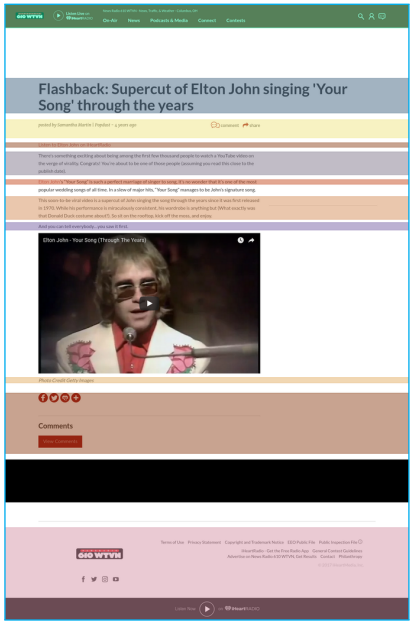


PDoC $\in [8, 9]$

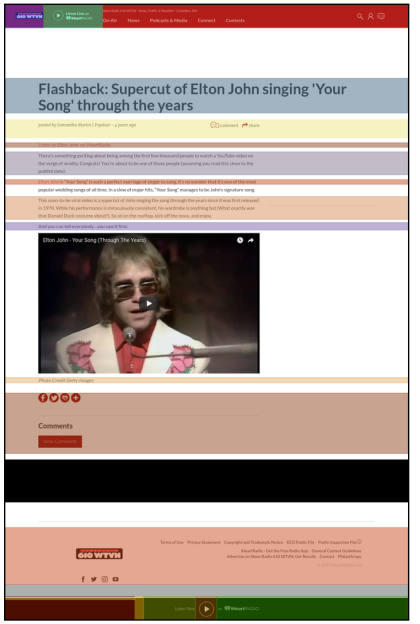


PDoC $\in [10, 11]$

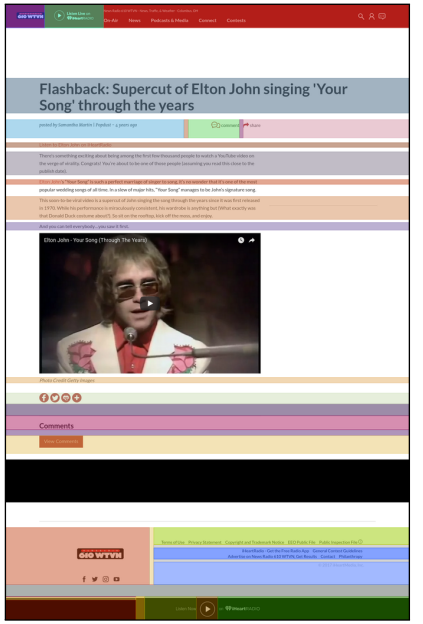
VIPS: Optimization for Permitted Degree of Coherence (PDoC)



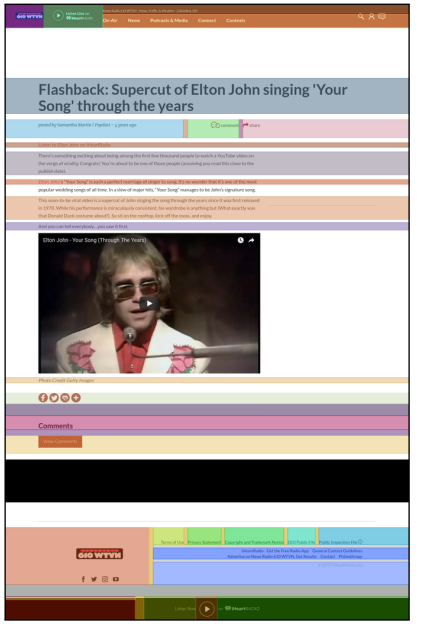
PDoC $\in [1, 6]$



PDoC $= 7$

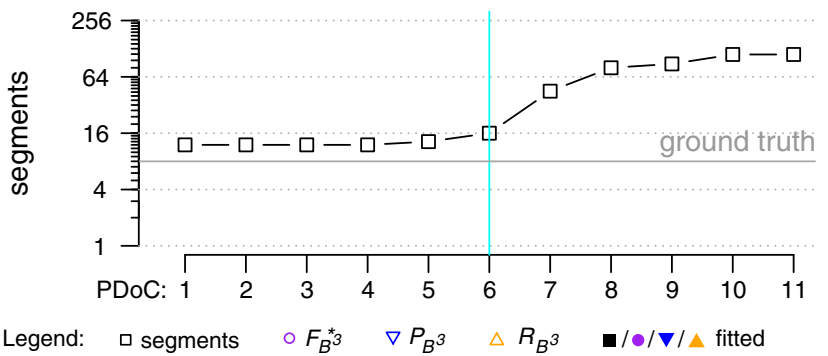


PDoC $\in [8, 9]$

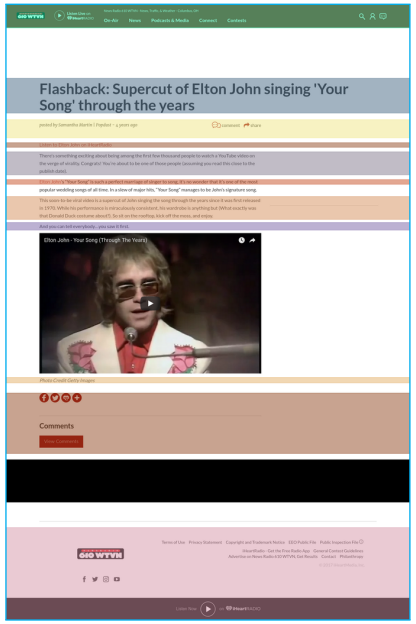


PDoC $\in [10, 11]$

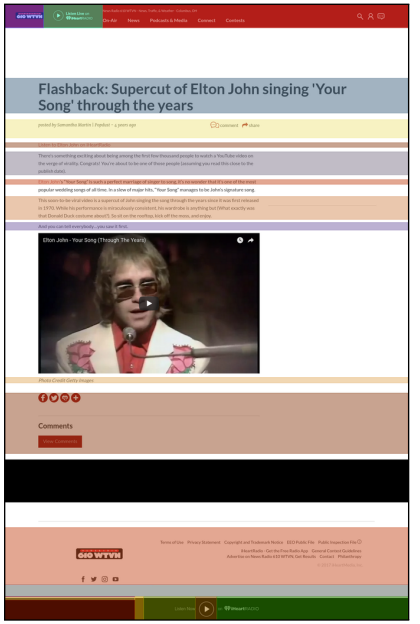
Number of segments



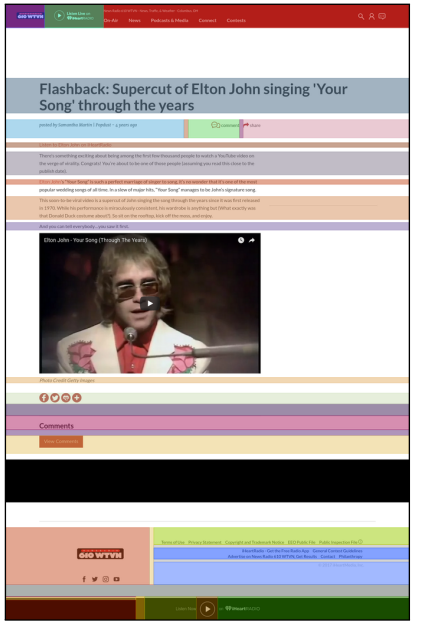
VIPS: Optimization for Permitted Degree of Coherence (PDoC)



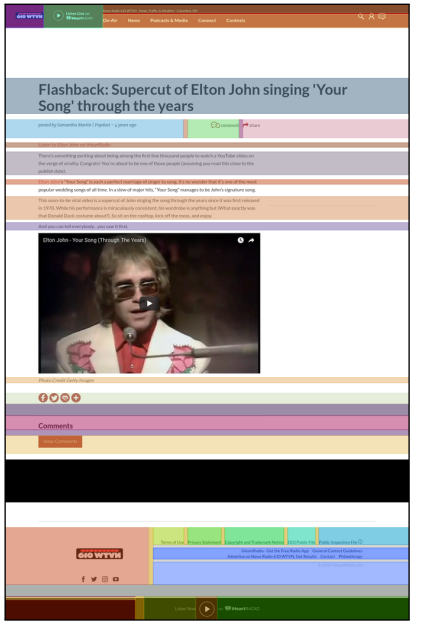
$\text{PDoC} \in [1, 6]$



$\text{PDoC} = 7$

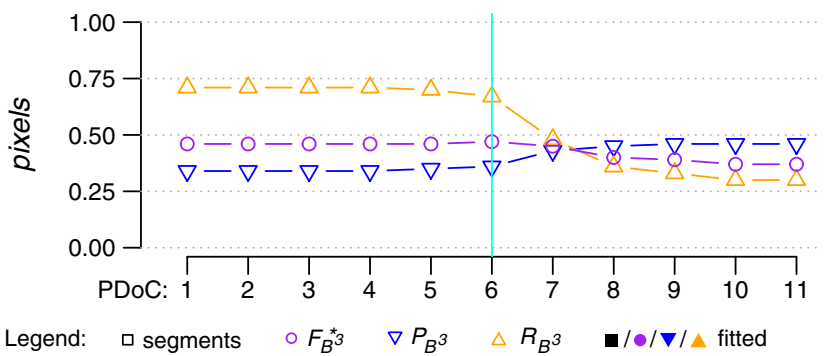


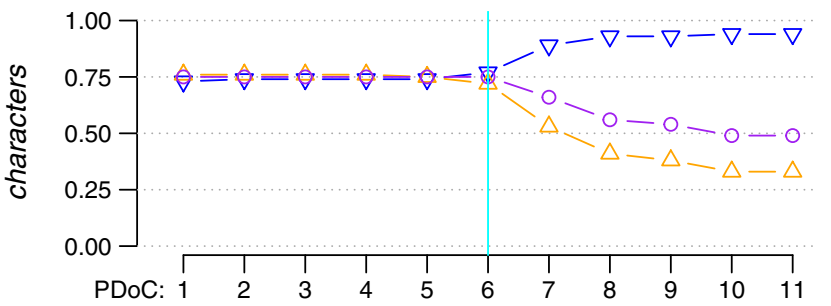
$\text{PDoC} \in [8, 9]$



$\text{PDoC} \in [10, 11]$

Comparison with ground-truth



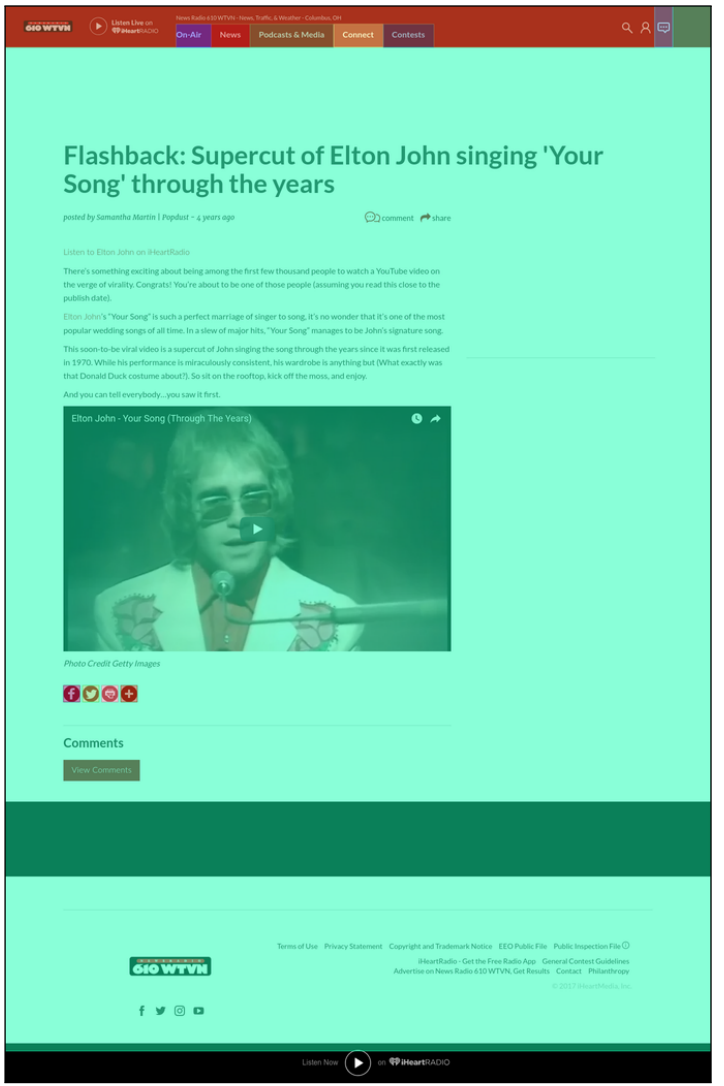


HEPS

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

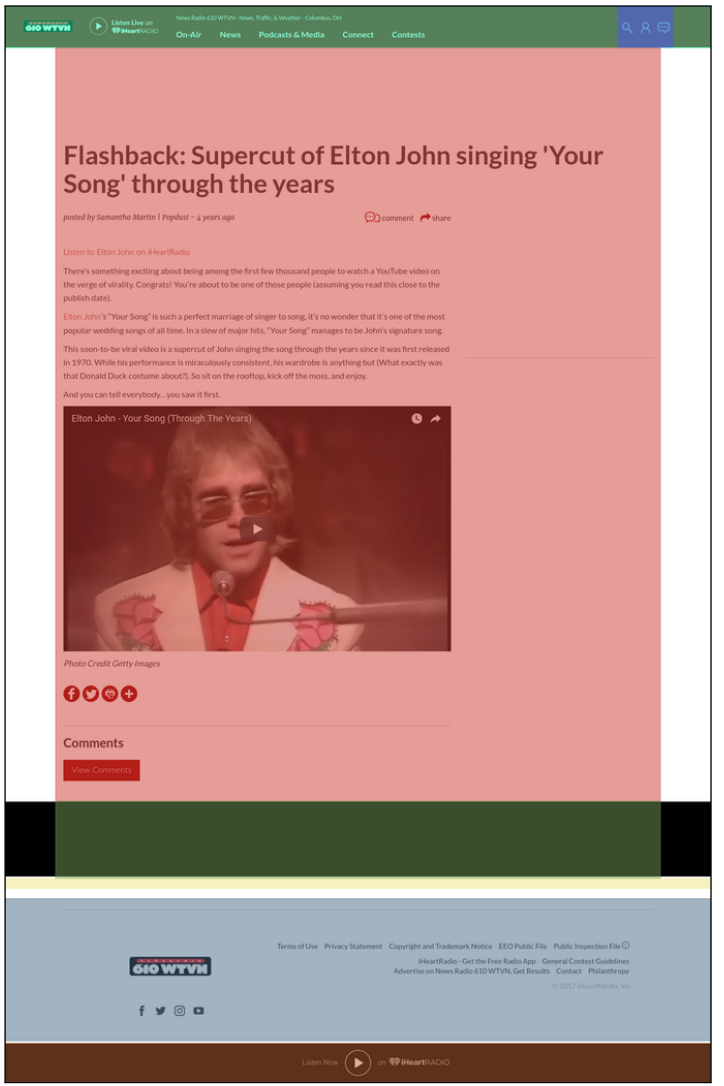
- Identifies headings and their segments by heuristic rules
- A heading is “both visually prominent and described the topic of a segment”

We slightly adopted the author’s original implementation

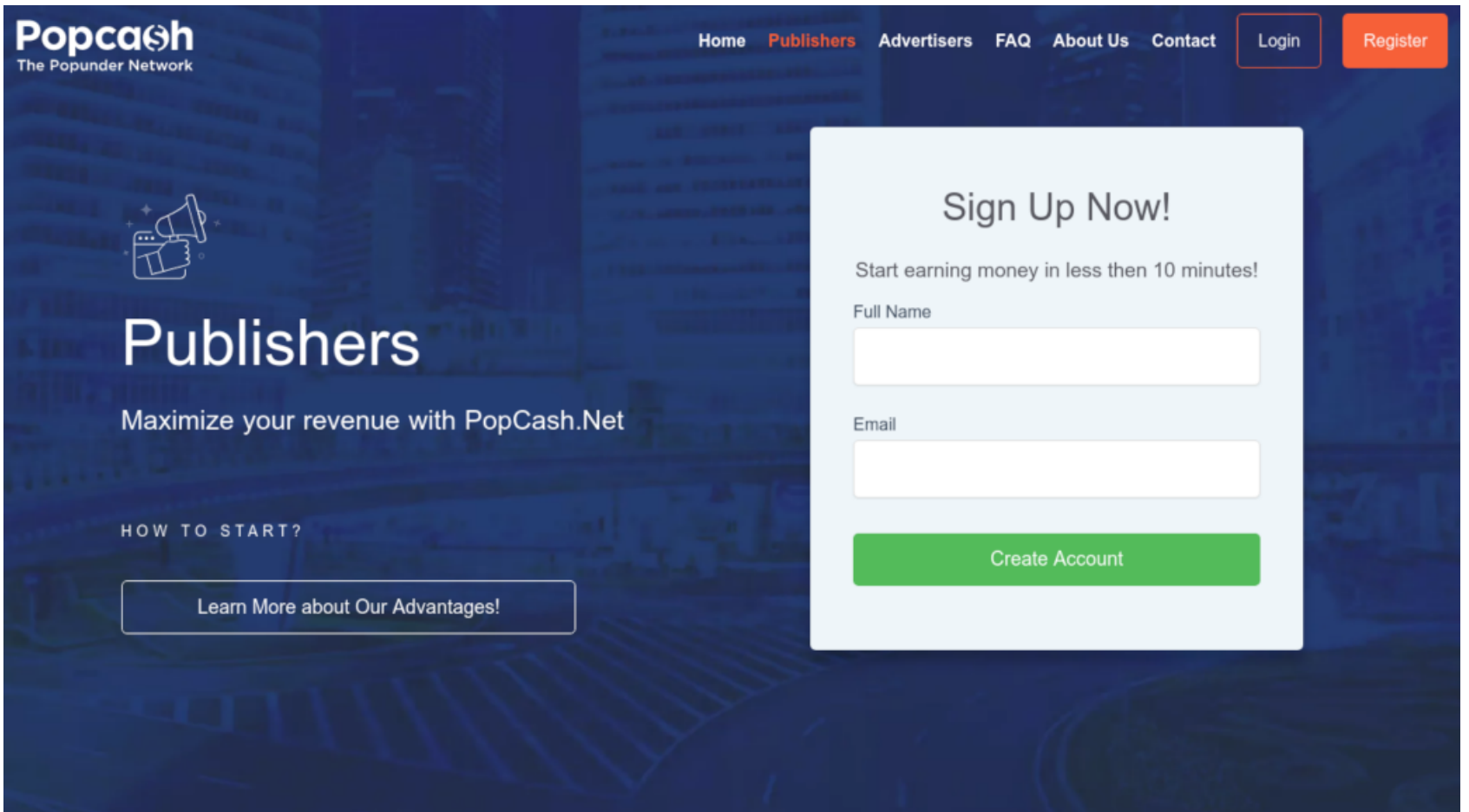


Cormier et al.

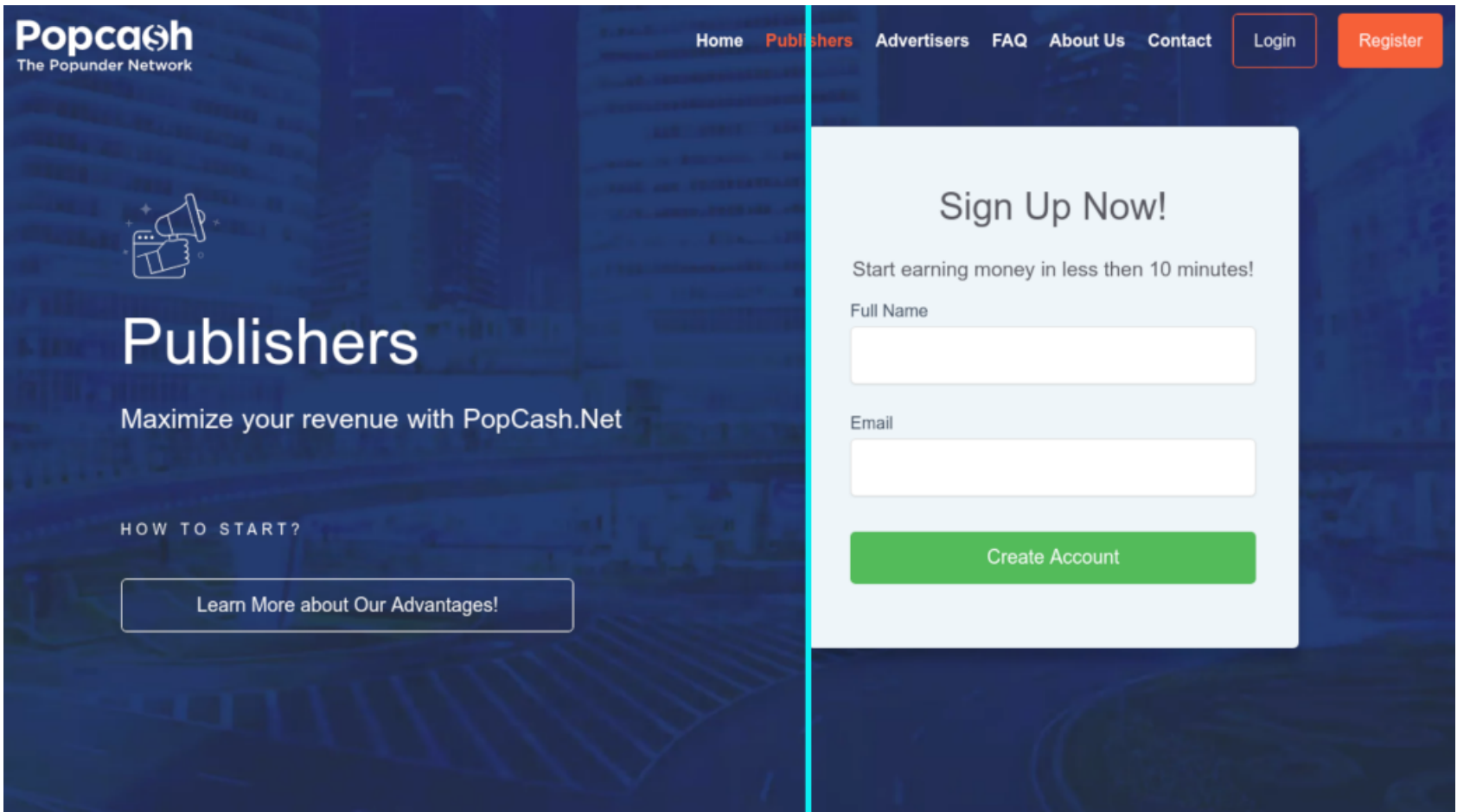
- Uses the web page screenshot as sole input
- Identifies locally significant horizontal and vertical edge pixels
- Identifies horizontal and vertical “semantically significant” lines of such pixels
- Recursively splits segments by most semantically significant line



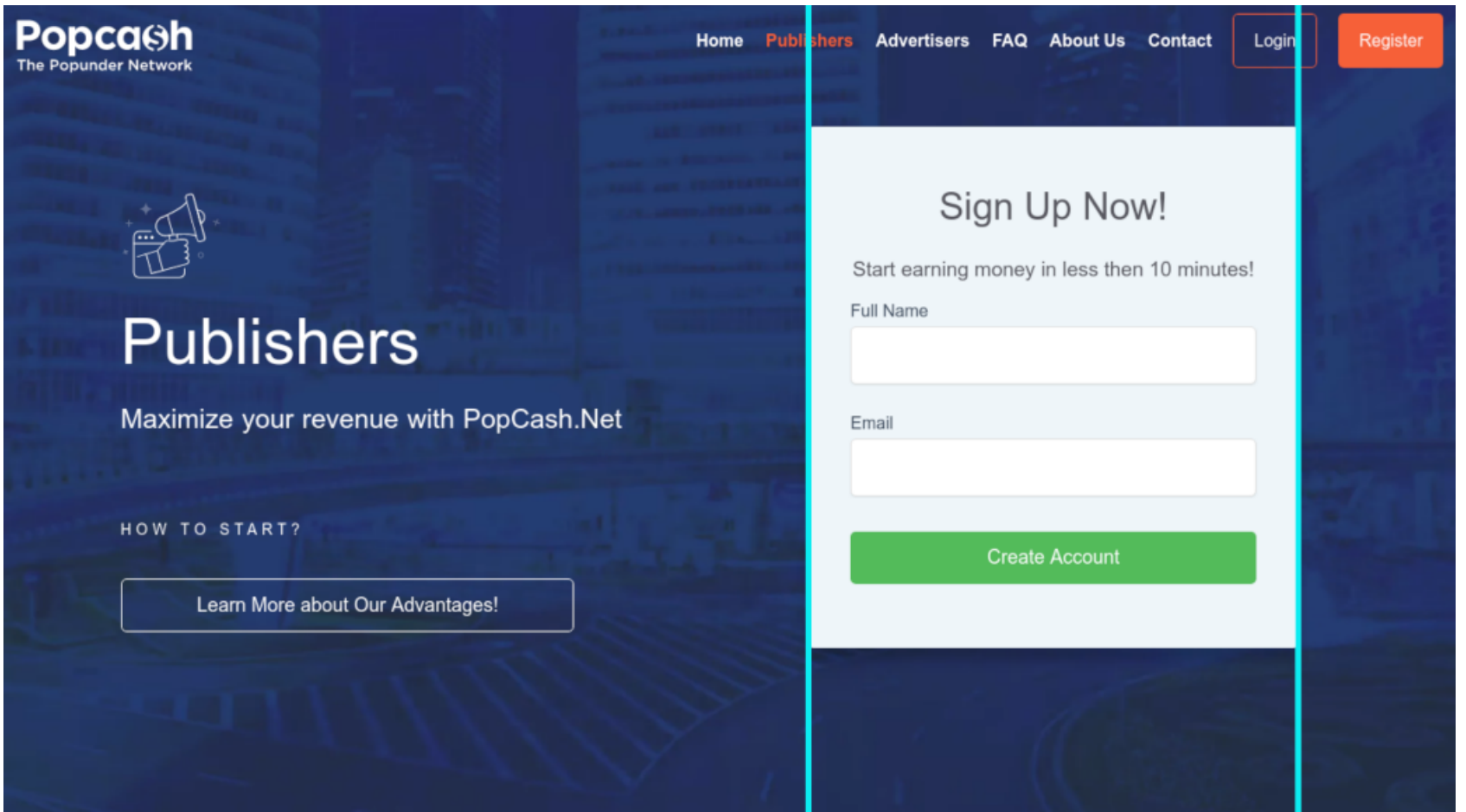
Cormier et al.



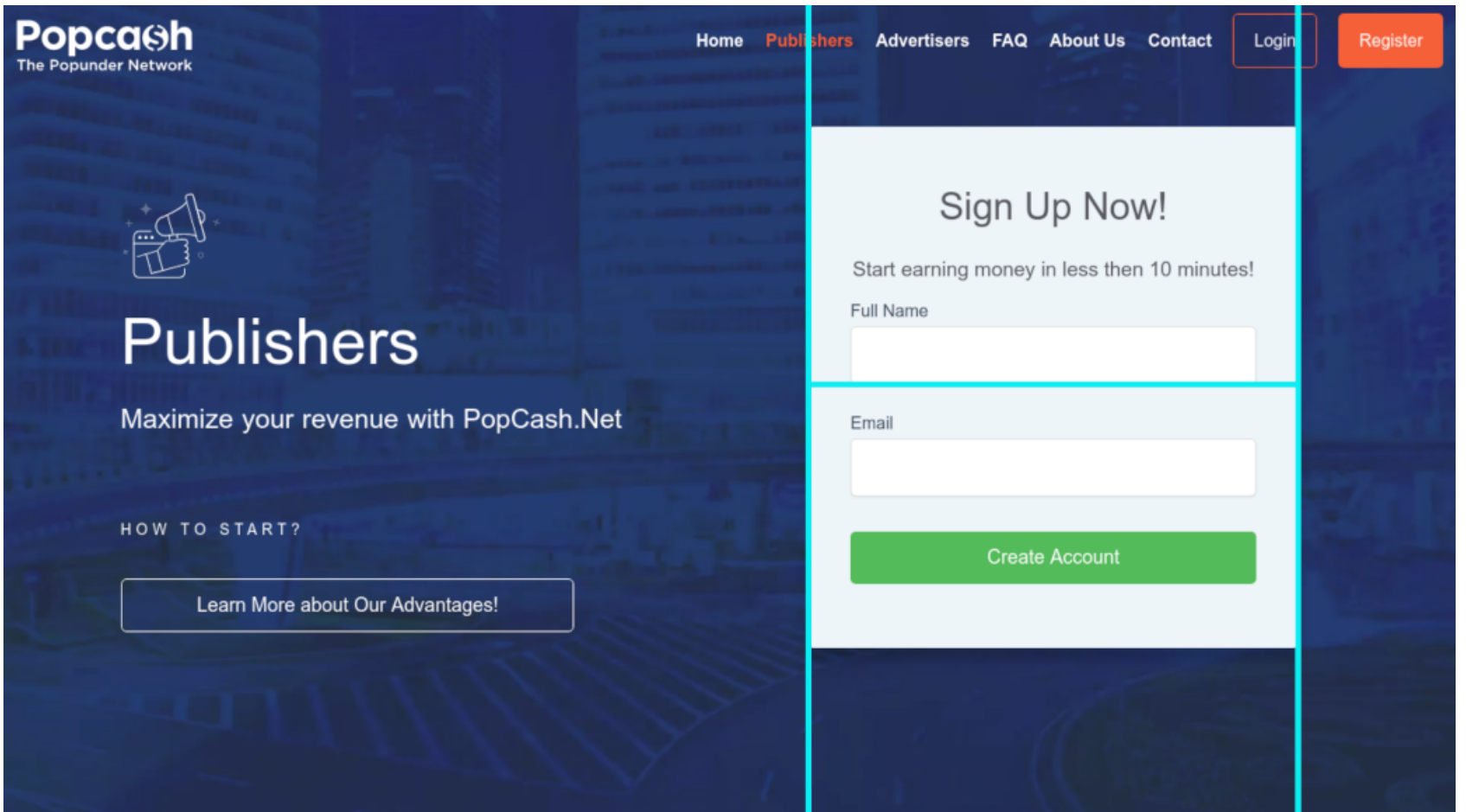
Cormier et al.



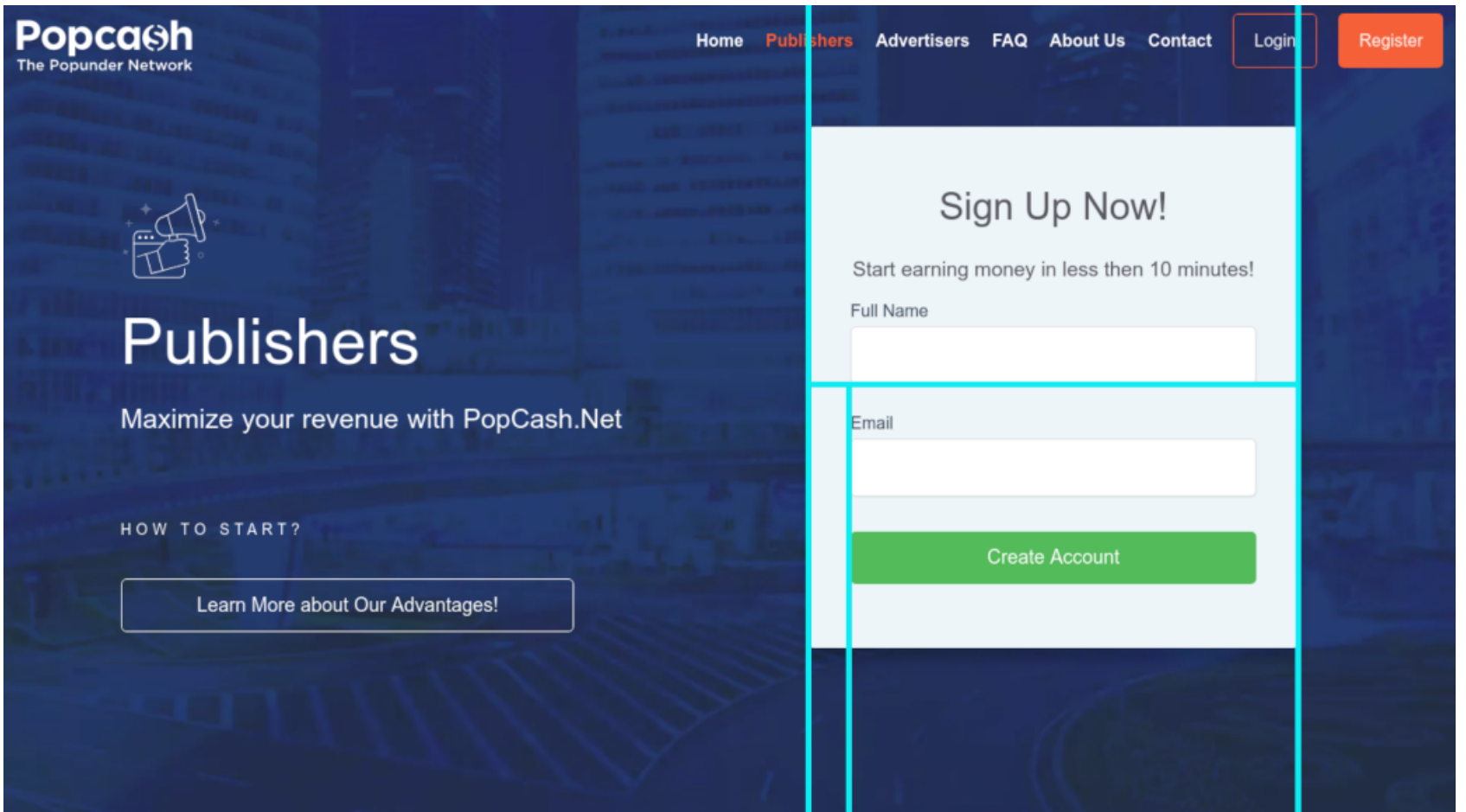
Cormier et al.



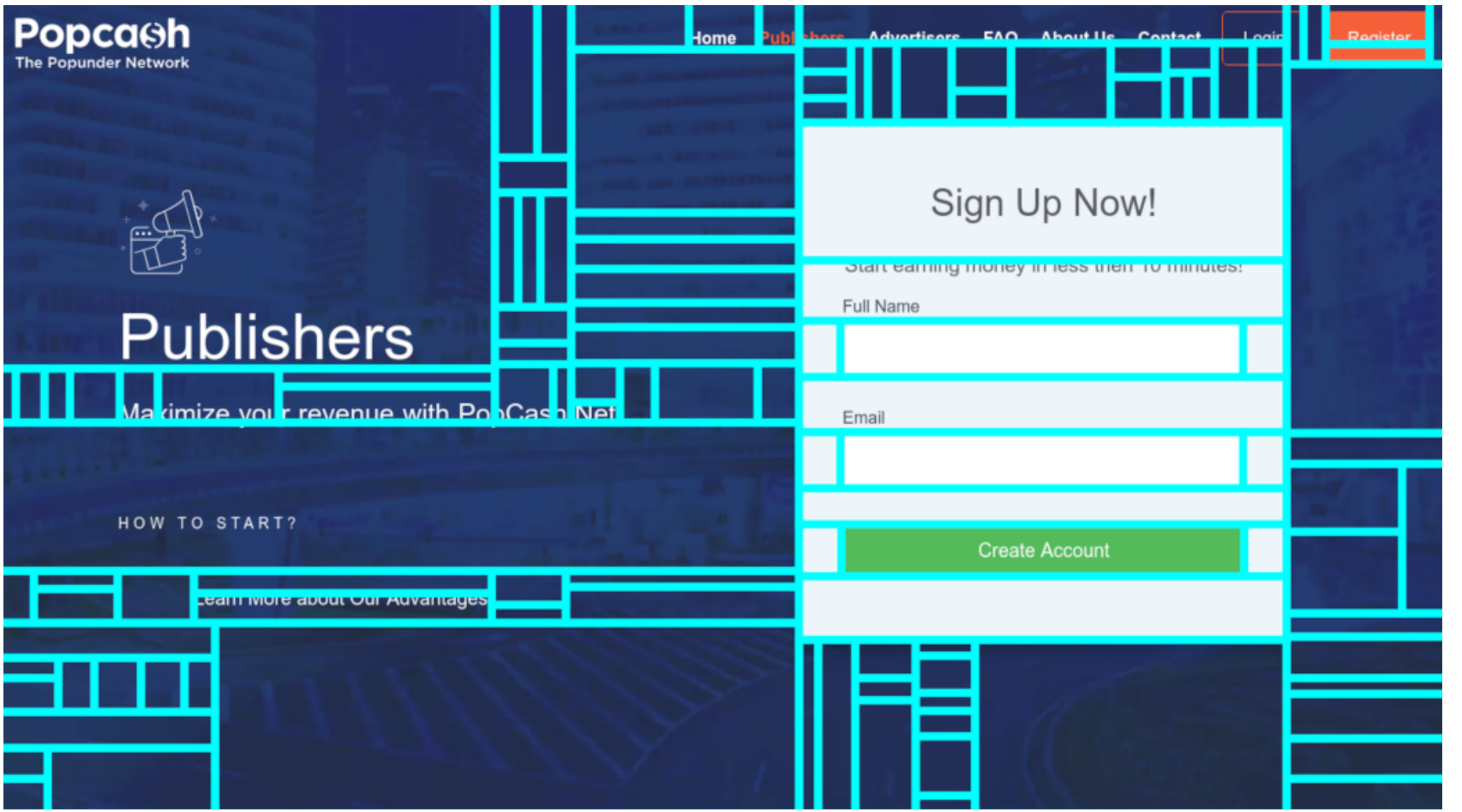
Cormier et al.



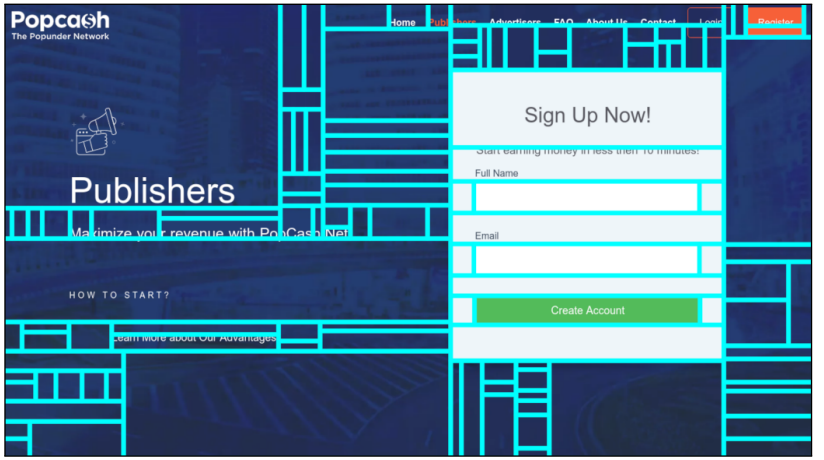
Cormier et al.



Cormier et al.

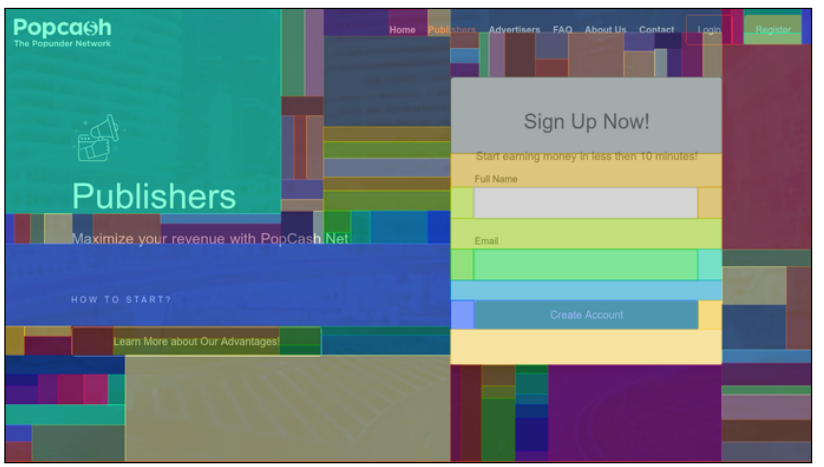


Cormier et al.: Fitting to DOM Nodes

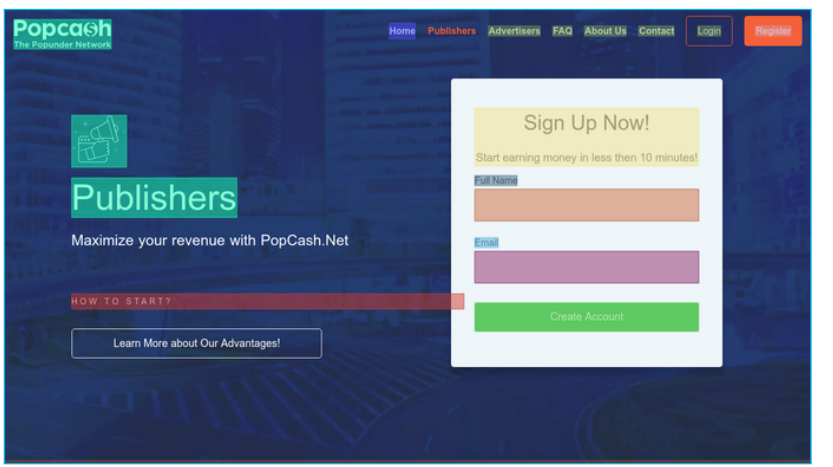


Original (borders)

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

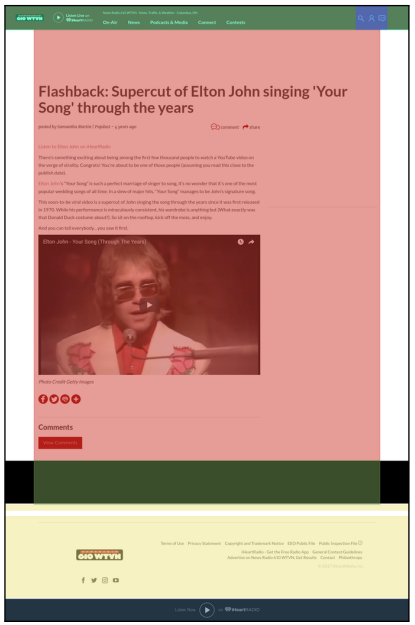


Original

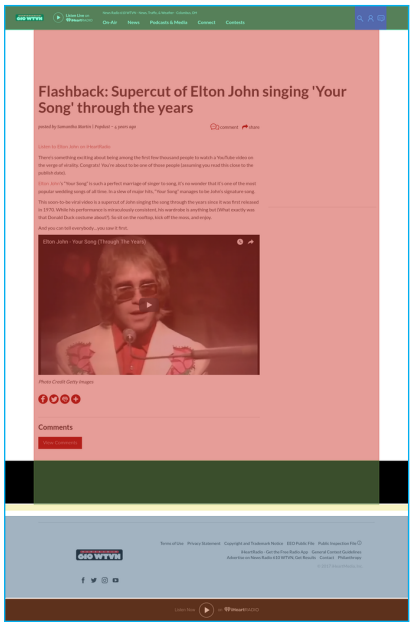


Fitted

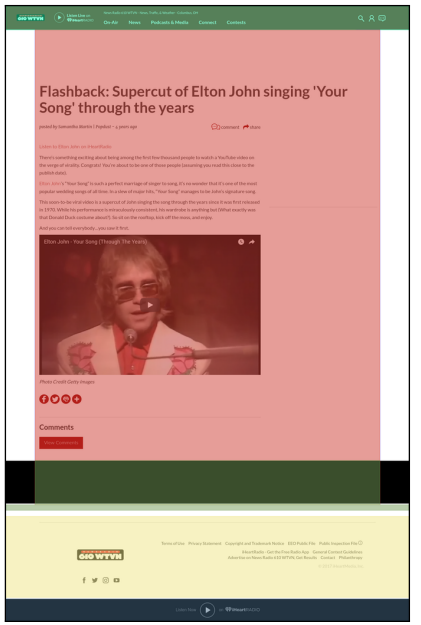
Cormier et al.: Optimization



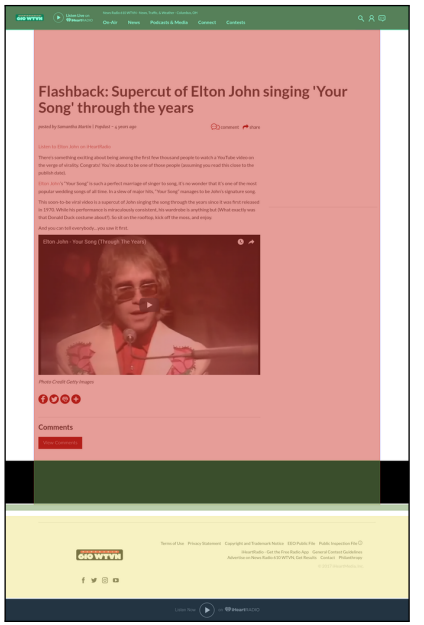
$t_l = 256; s_{\text{min}} = 45$



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

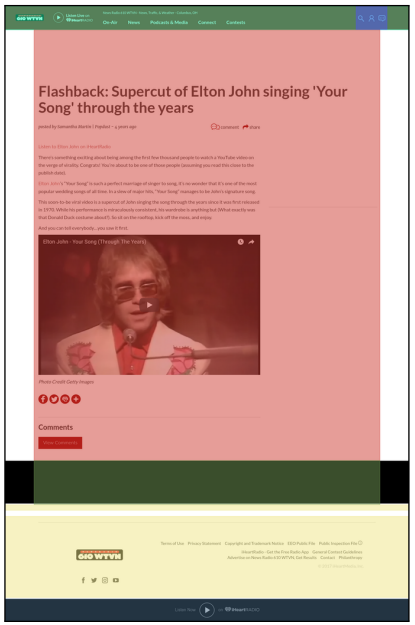


$t_l = 256; s_{\text{min}} = 90$

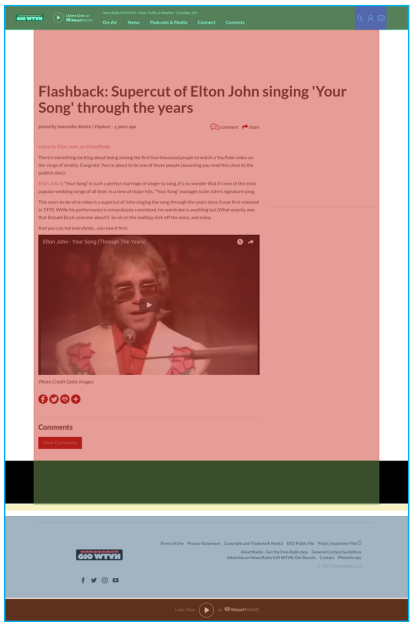


$t_l = 512; s_{\text{min}} = 90$

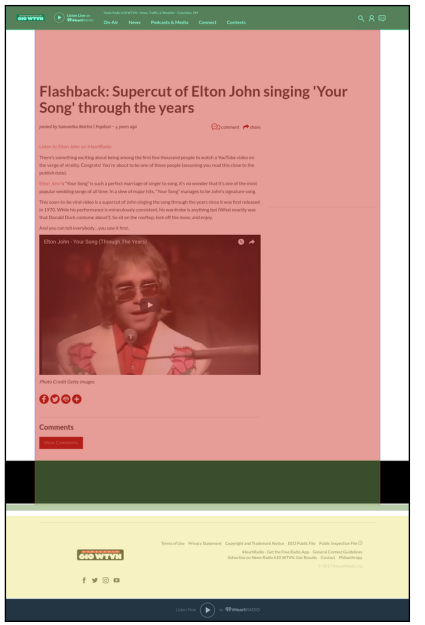
Cormier et al.: Optimization



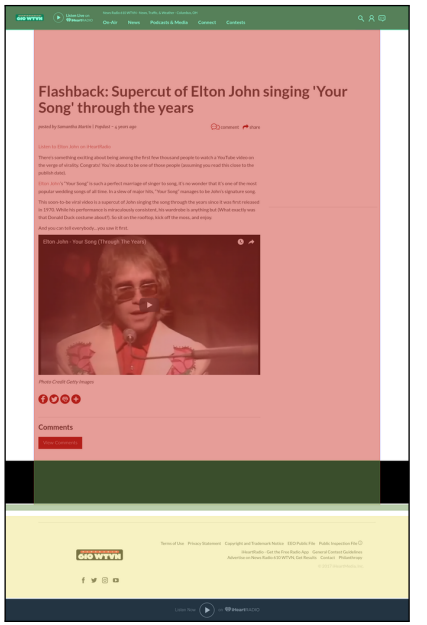
$t_l = 256; s_{\text{min}} = 45$



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

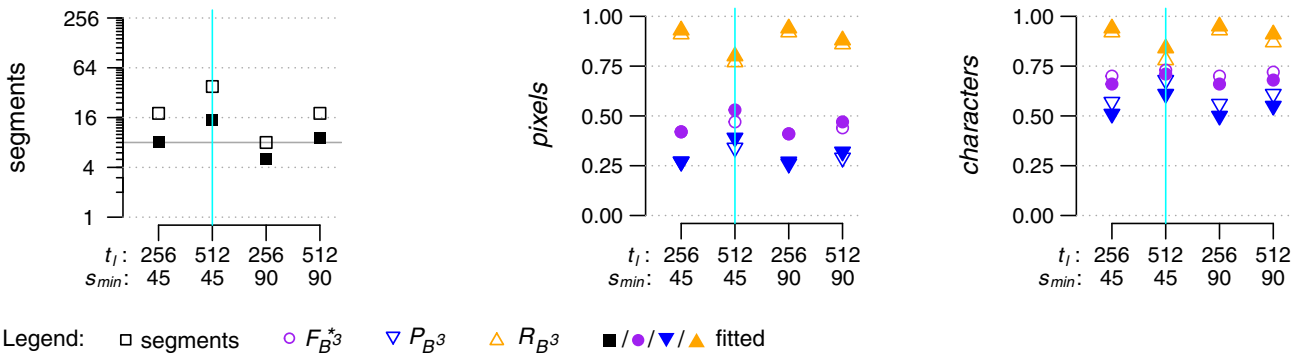


$t_l = 256; s_{\text{min}} = 90$



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

Number of segments and comparison with ground-truth

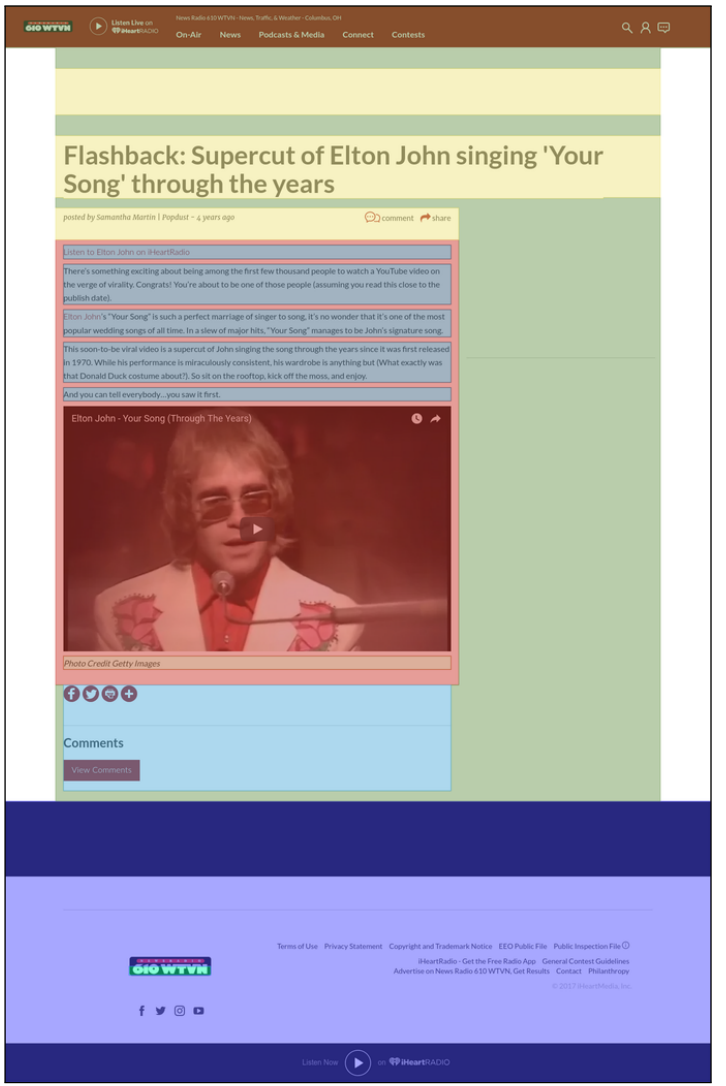


MMDetection

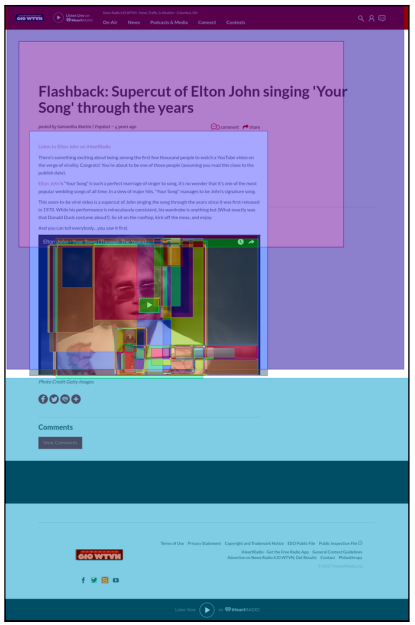
One Hybrid Task Cascade model from the MMDetection toolbox by Chen et al., 2019.

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

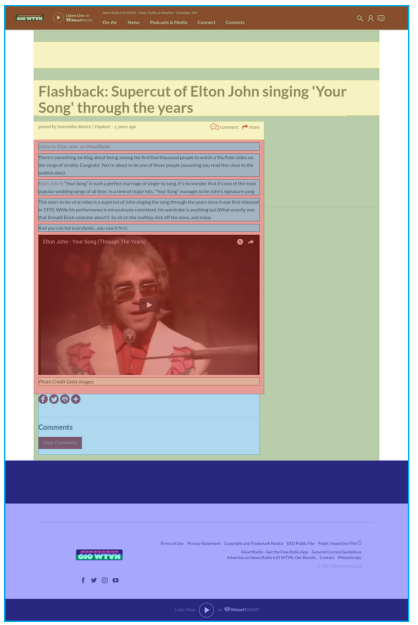
- Uses the web page screenshot as sole input
- Neural network
- Trained on object detection in real-world images (photos)



MMDetection: Optimization

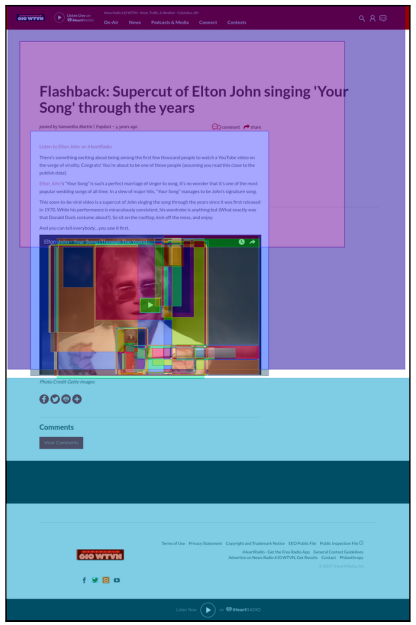


Original

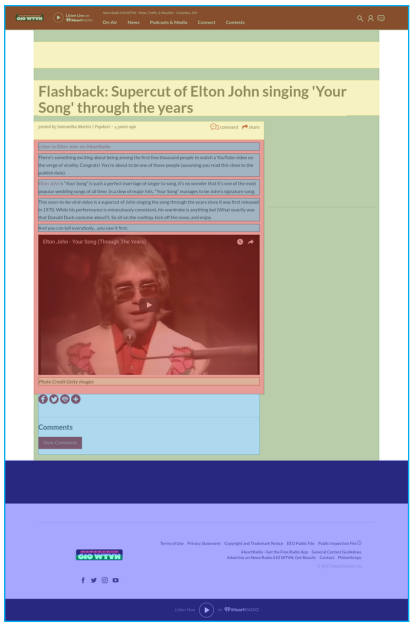


Fitted

MMDetection: Optimization

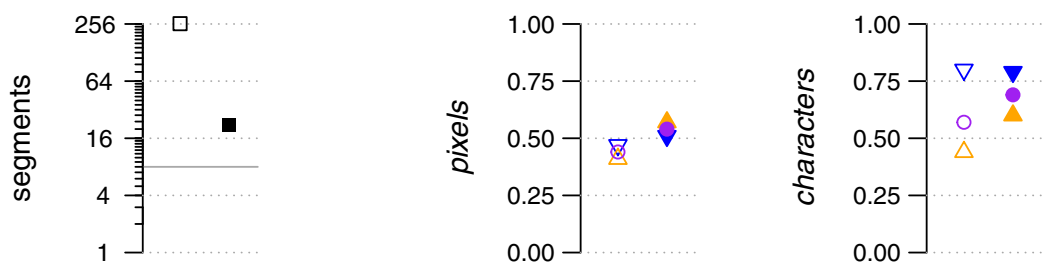


Original



Fitted

Number of segments and comparison with ground-truth

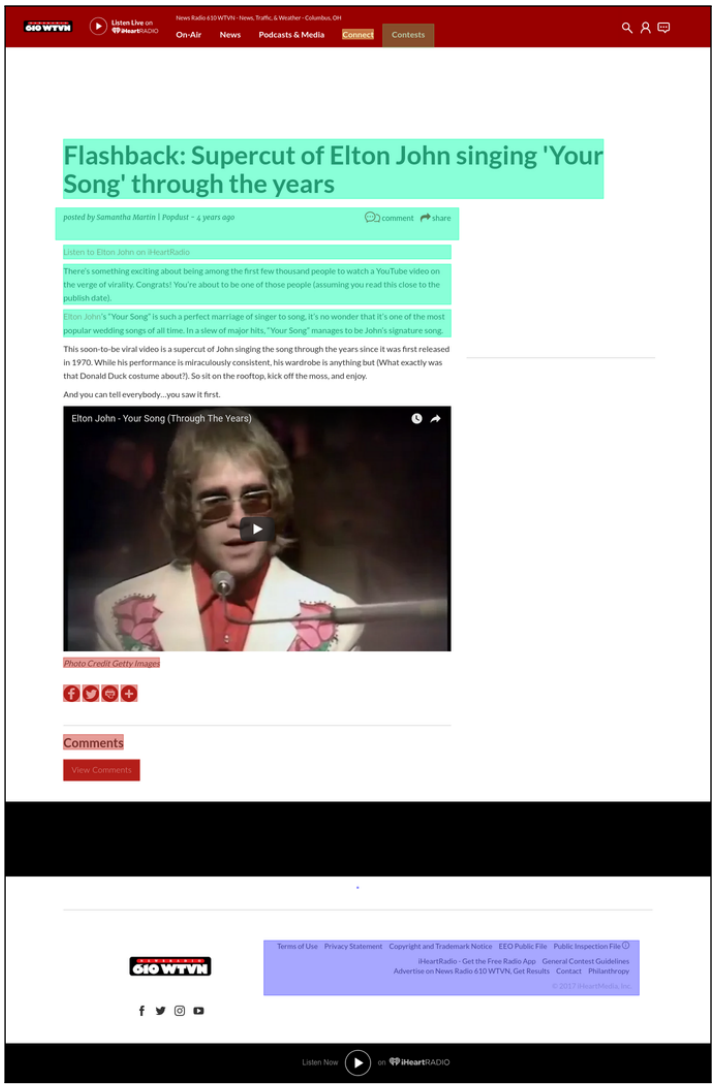


Legend: segments $F_{B^3}^*$ P_{B^3} R_{B^3} fitted

Meier et al.

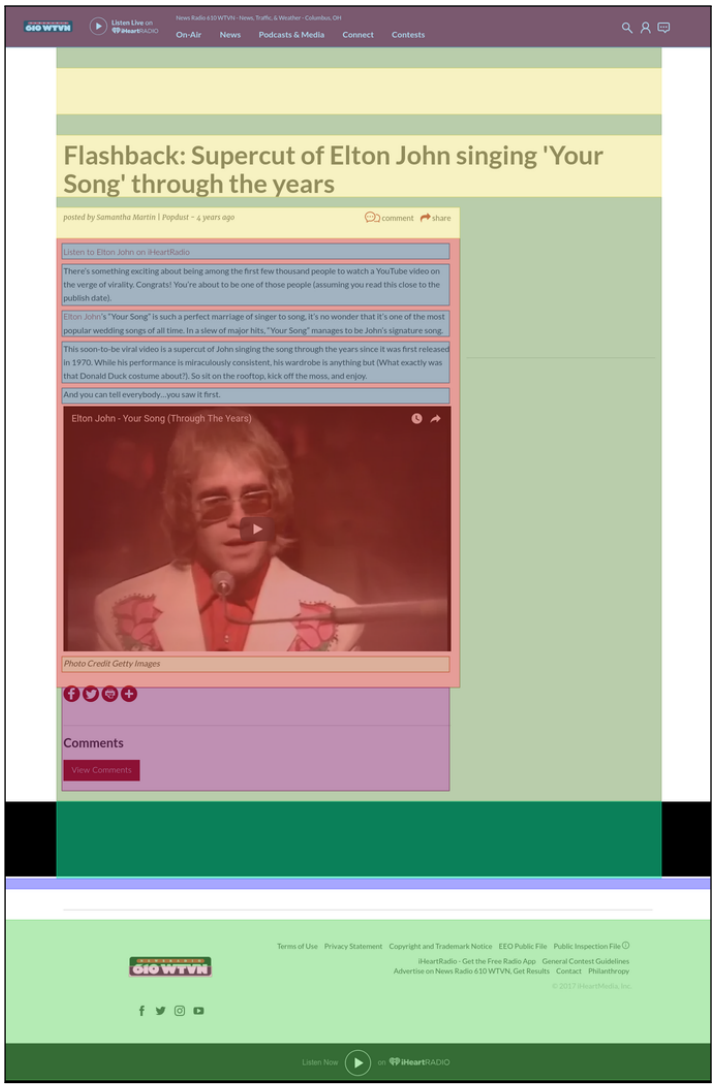
Meier et al., 2017

- 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

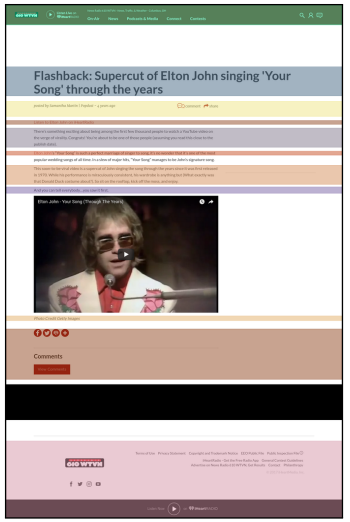


Min-vote Ensemble

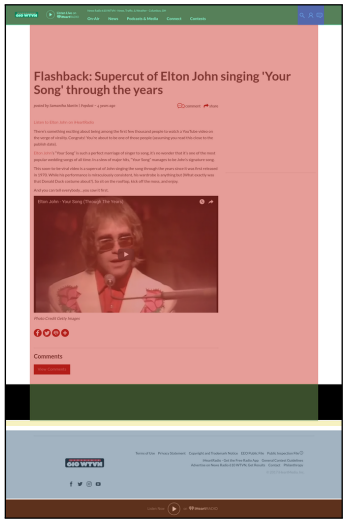
- Ensemble of VIPS, HEPS, Cormier et al., and MMDetection
- Parameter $n \in [1, 4]$
- Ignores elements which less than n algorithms placed into segments
- Standard hierarchical agglomerative clustering
- Similarity of two elements is the ratio of algorithms that place these elements in the same segment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



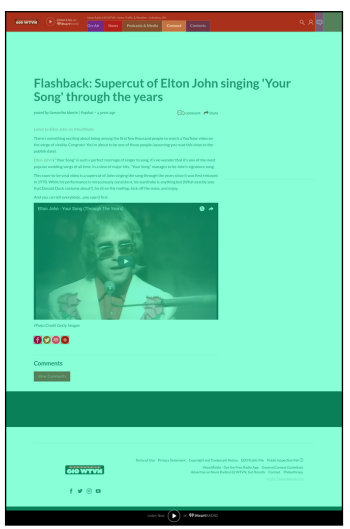
Min-vote Ensemble



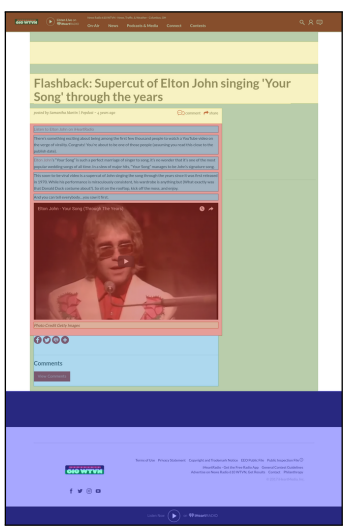
VIPS



Cormier et al.

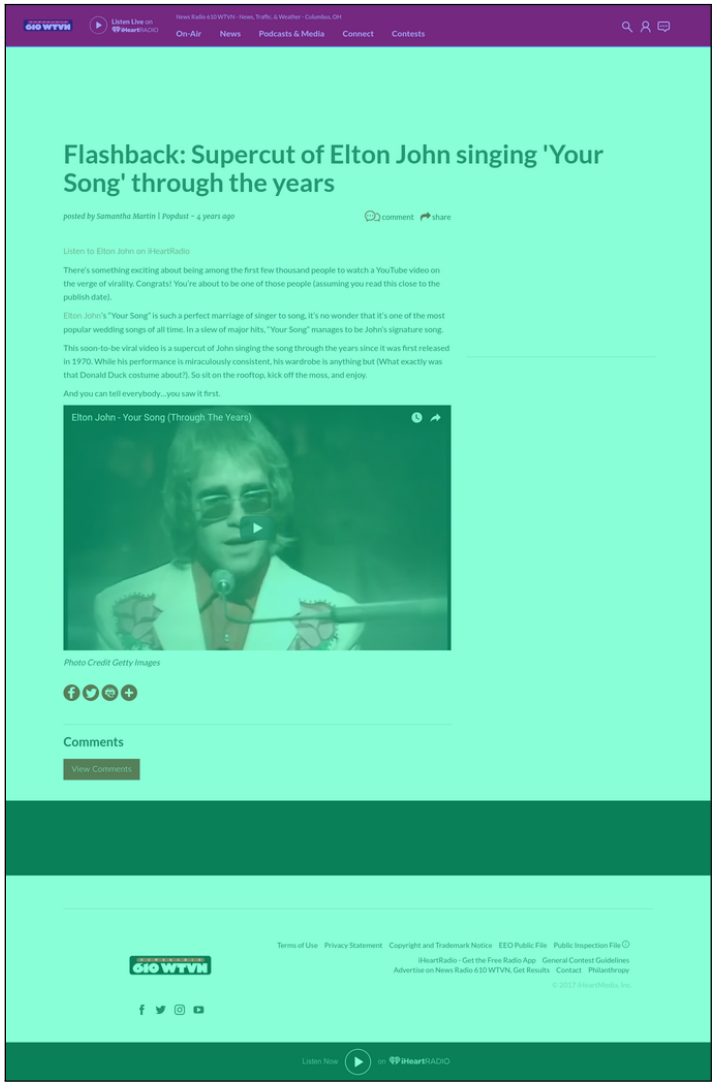


HEPS



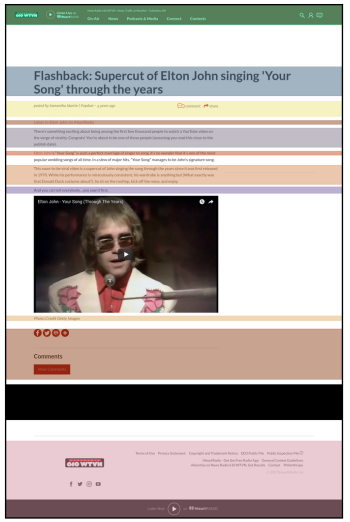
MMDetection

⇒

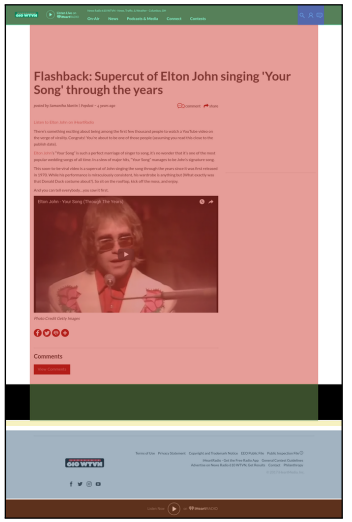


Min-vote@1

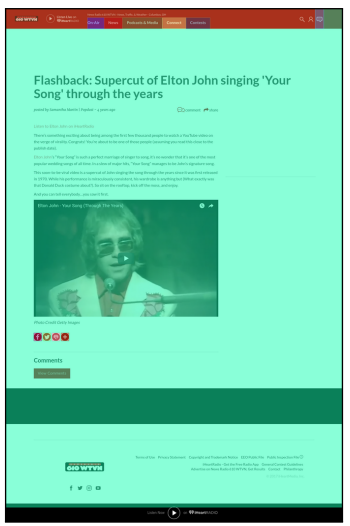
Min-vote Ensemble



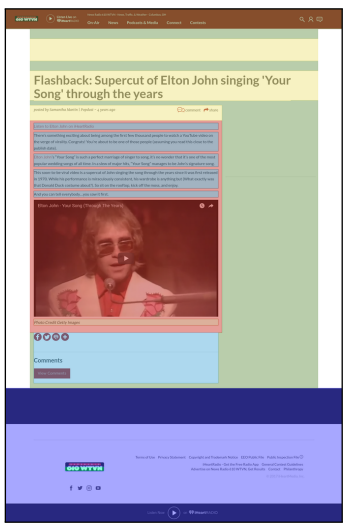
VIPS



Cormier et al.

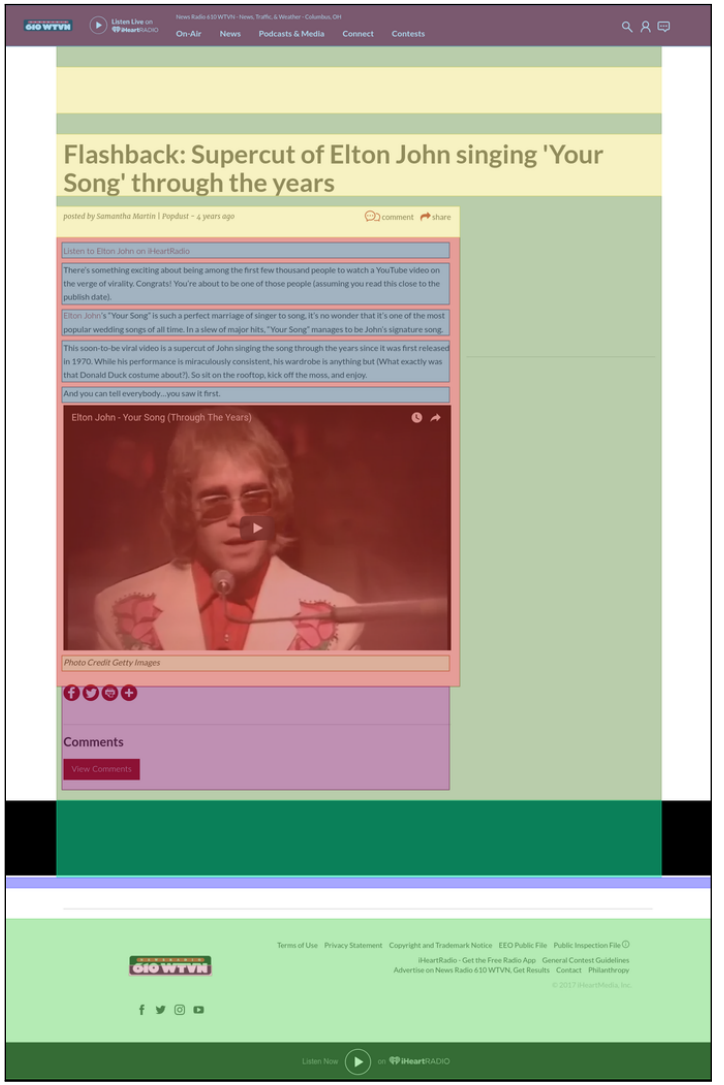


HEPS



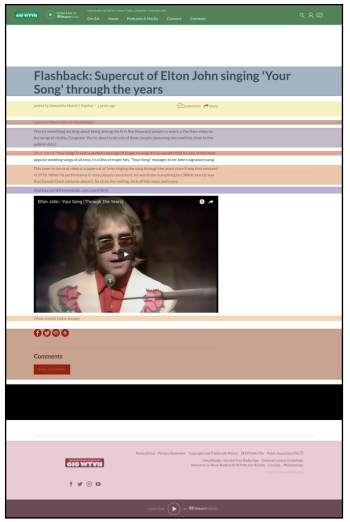
MMDetection

⇒

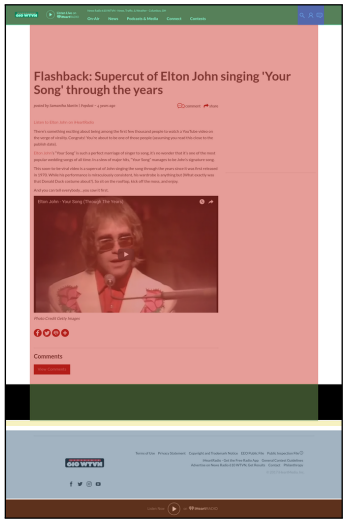


Min-vote@2

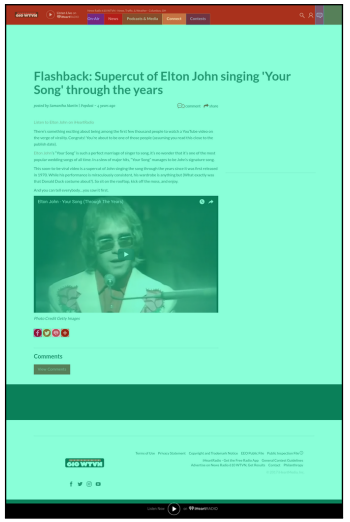
Min-vote Ensemble



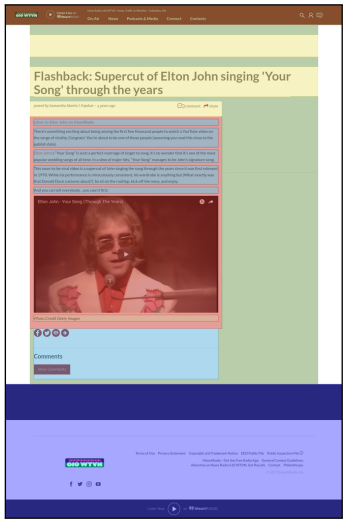
VIPS



Cormier et al.

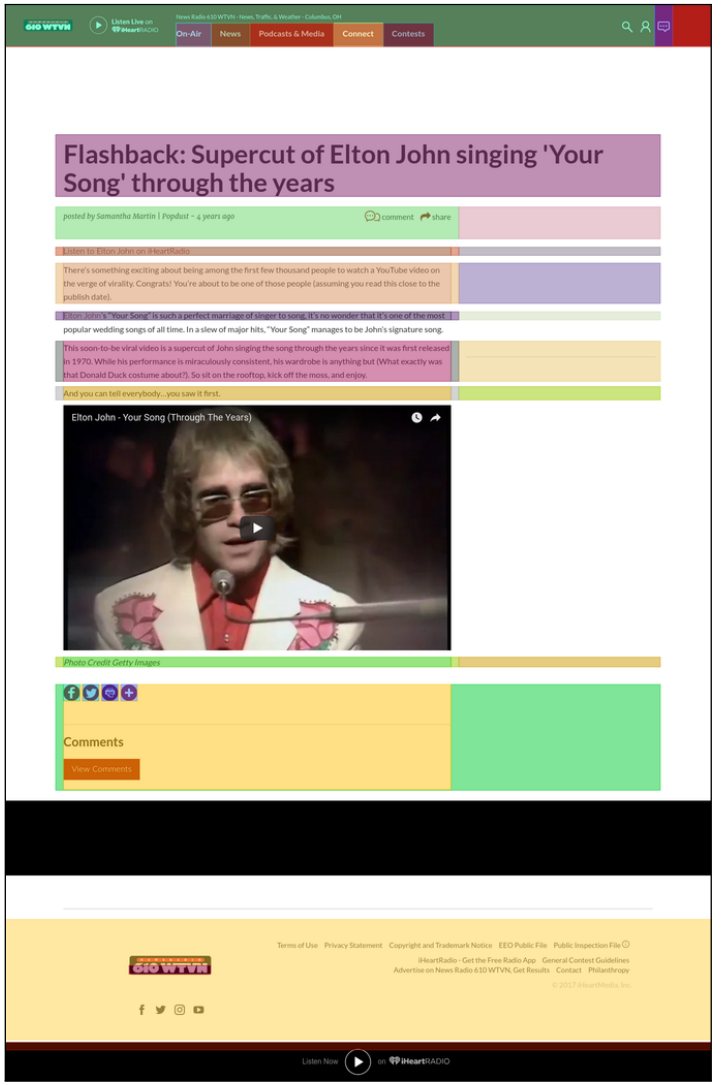


HEPS



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
pixels	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
edges$_F$	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
edges$_C$	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
nodes	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
chars	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
pixels	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
edges$_F$	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
edges$_C$	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
nodes	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
chars	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
pixels	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
chars	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
pixels	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
chars	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
pixels	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
chars	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
pixels	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
chars	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
pixels	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
chars	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
pixels	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
chars	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
pixels	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
chars	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
pixels	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
chars	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
pixels	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
chars	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
pixels	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
chars	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
pixels	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
chars	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
pixels	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
chars	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