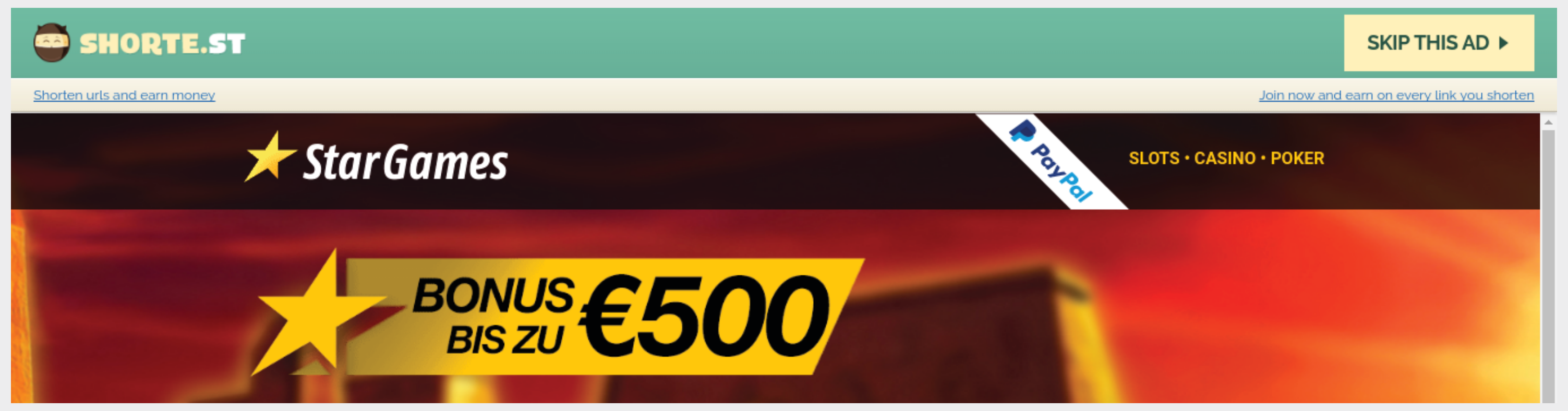
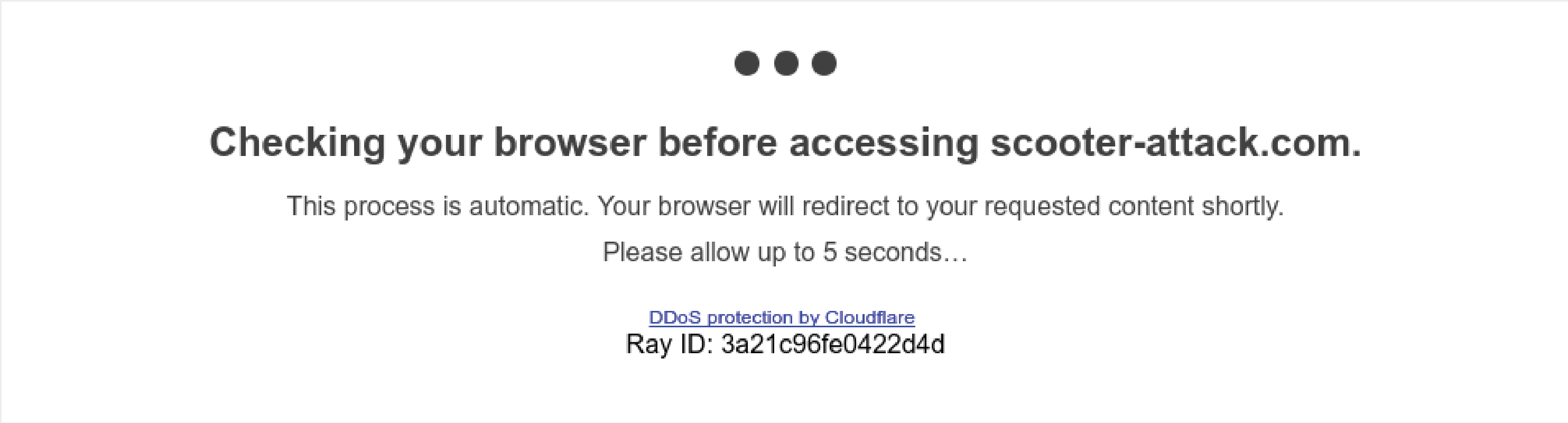
Johannes Kiesel   Fabienne Hubricht   Benno Stein   Martin Potthast

A Dataset for Content Error Detection in Web Archives





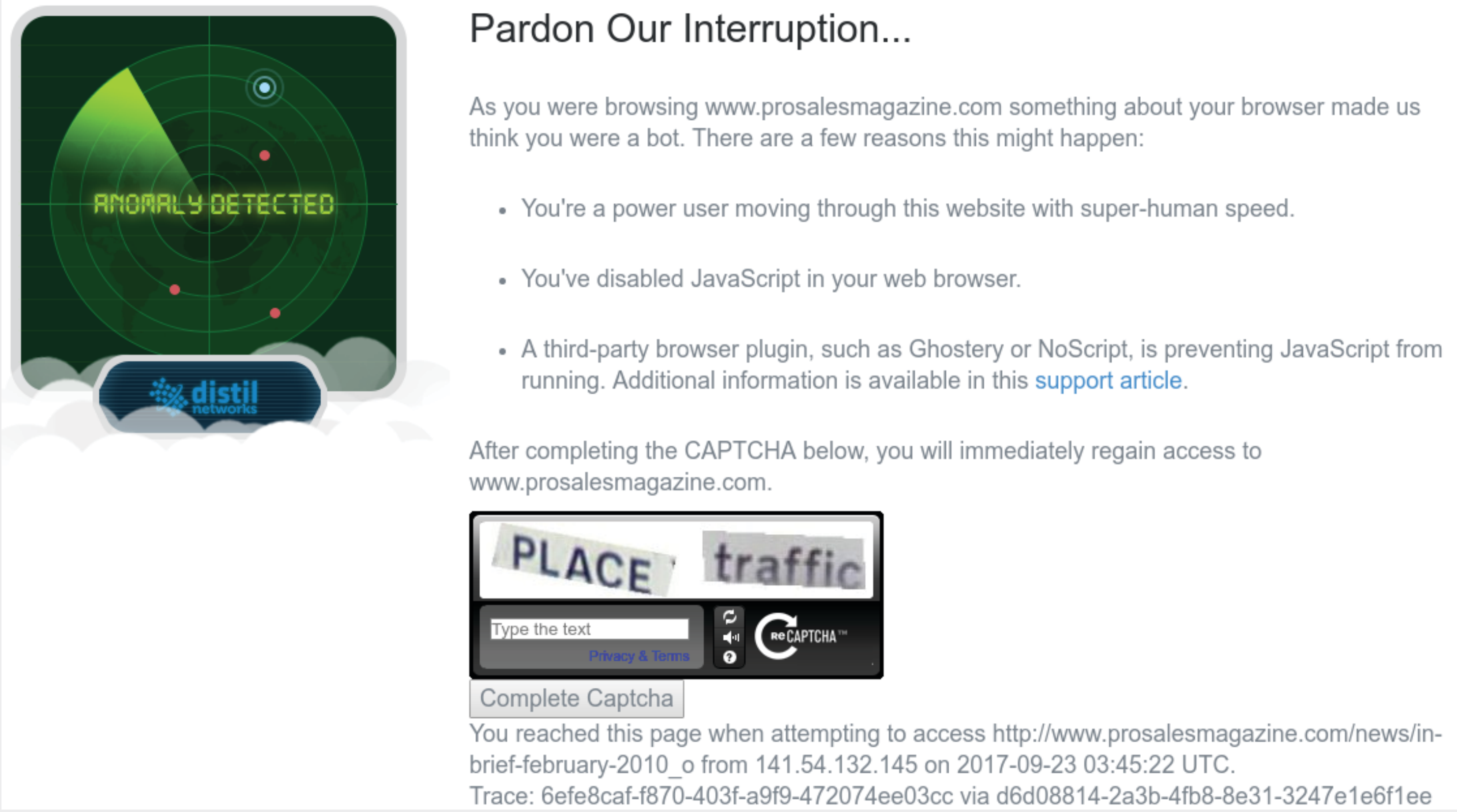
●●●

Checking your browser before accessing scooter-attack.com.

This process is automatic. Your browser will redirect to your requested content shortly.

Please allow up to 5 seconds…

DDoS protection by Cloudflare
Ray ID: 3a21c96fe0422d4d



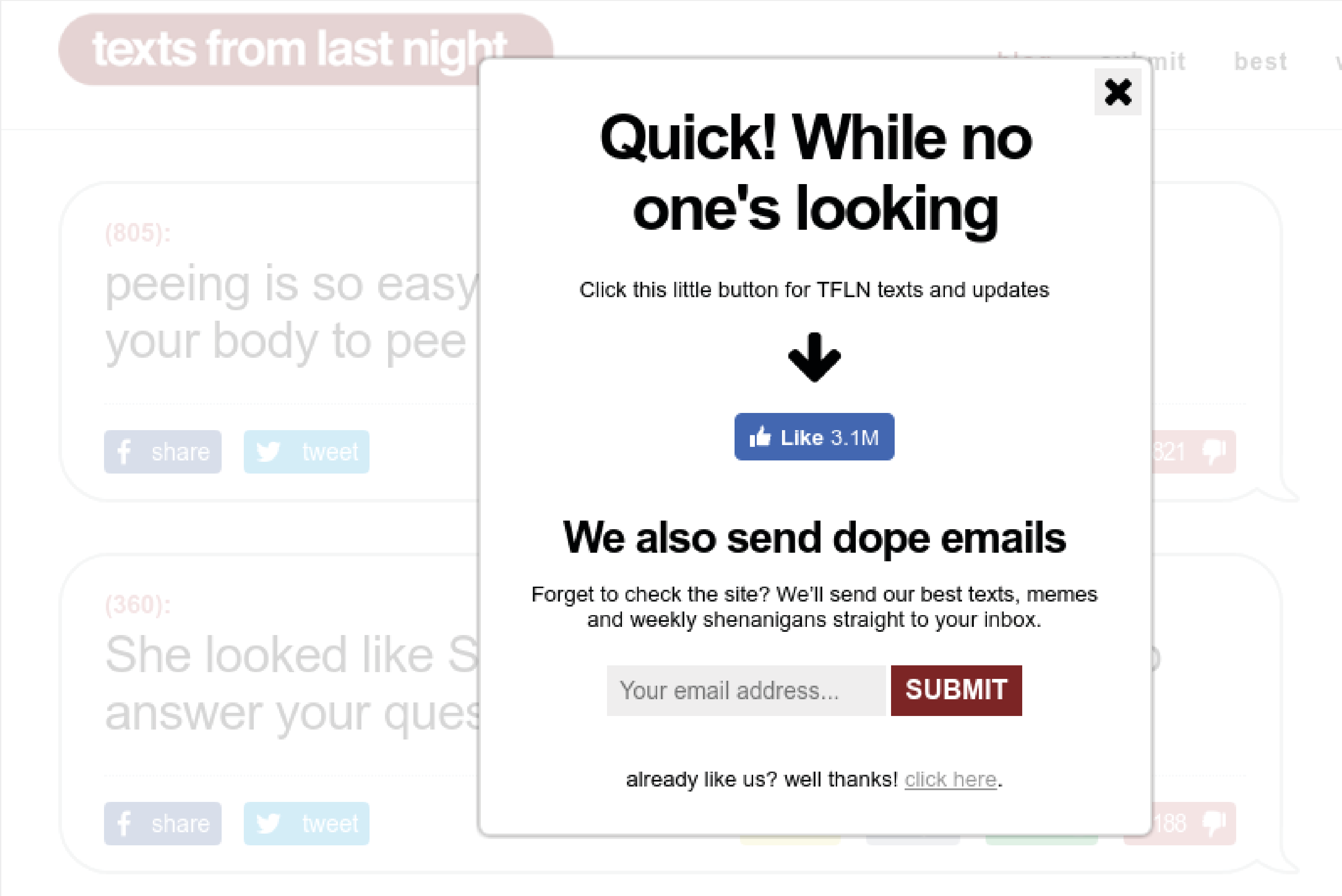
Pardon Our Interruption...

As you were browsing www.prosalesmagazine.com something about your browser made us think you were a bot. There are a few reasons this might happen:

- You're a power user moving through this website with super-human speed.
- You've disabled JavaScript in your web browser.
- A third-party browser plugin, such as Ghostery or NoScript, is preventing JavaScript from running. Additional information is available in this support article.

After completing the CAPTCHA below, you will immediately regain access to www.prosalesmagazine.com.

You reached this page when attempting to access http://www.prosalesmagazine.com/news/in-brief-february-2010_o from 141.54.132.145 on 2017-09-23 03:45:22 UTC.
Trace: 6efe8caf-f870-403f-a9f9-472074ee03cc via d6d08814-2a3b-4fb8-8e31-3247e1e6f1ee



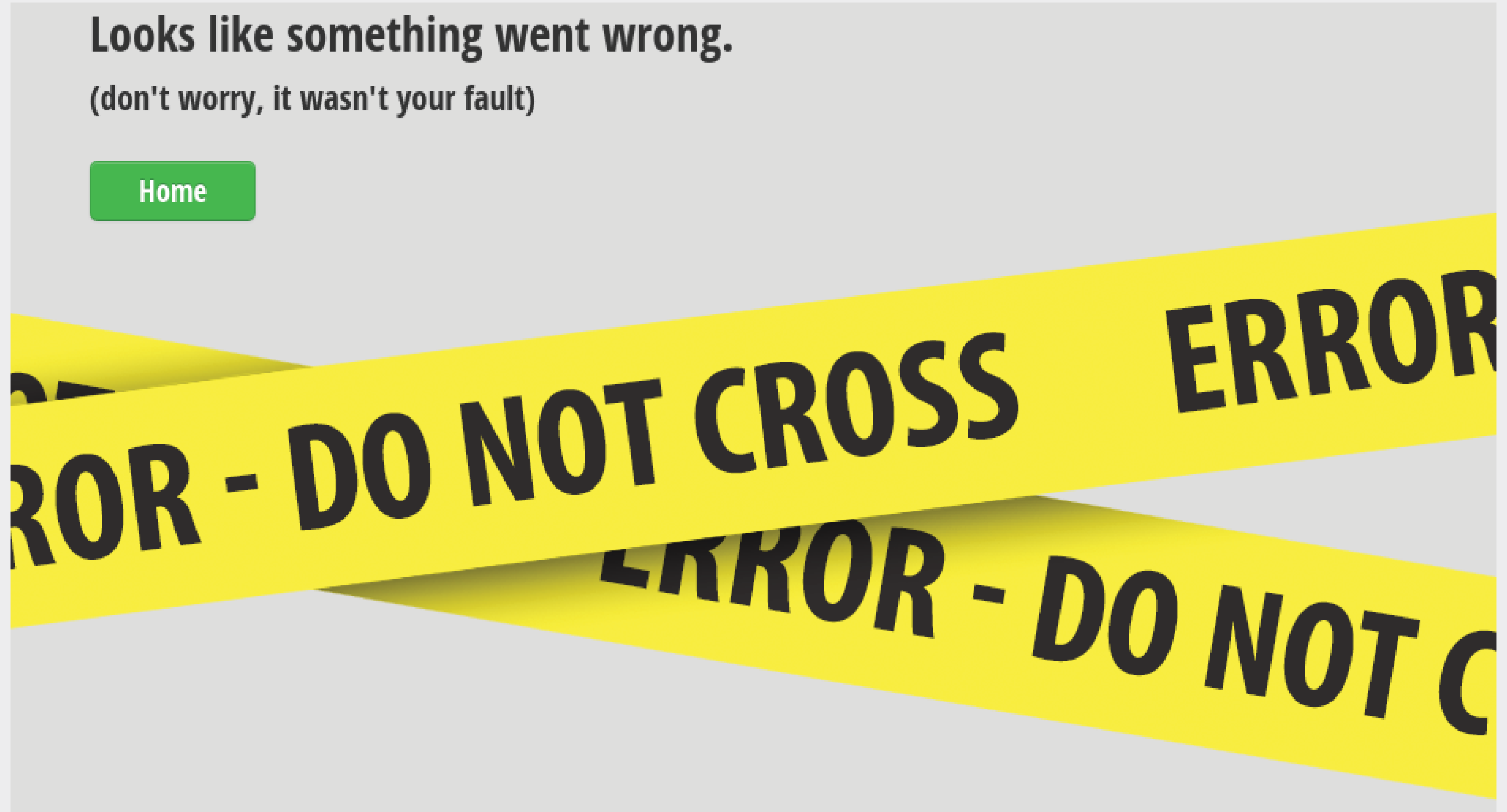
Quick! While no one's looking

Click this little button for TFLN texts and updates

We also send dope emails

Forget to check the site? We'll send our best texts, memes and weekly shenanigans straight to your inbox.

already like us? well thanks! click here.


Looks like something went wrong.
(don't worry, it wasn't your fault)
Home
ROR - DO NOT CROSS   ERROR
ERROR - DO NOT C

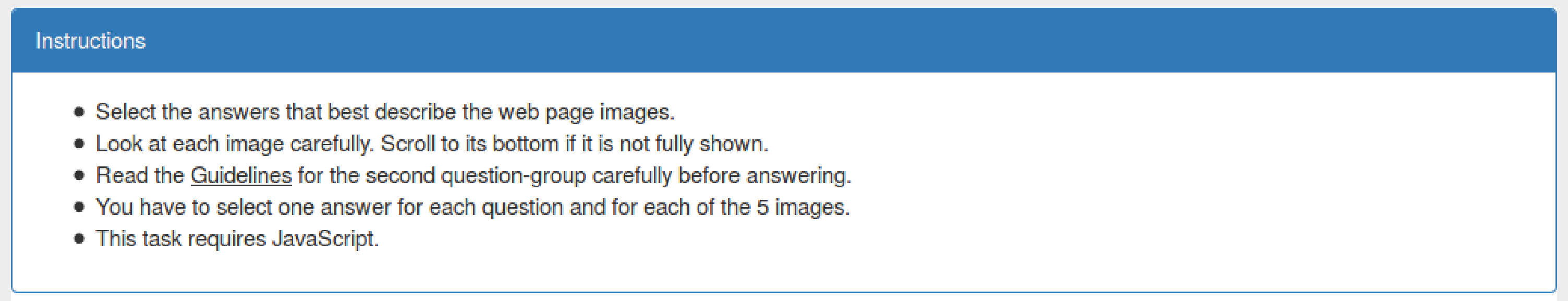
A Dataset of 10,000 Web Pages

Content errors if a web page is different than one expects.
We define 5 types of content errors.

Webis-Web-Archive-17 sampled from Common Crawl.
Annotated for reproduction errors in previous work

Annotation using Mechanical Turk with 5 annotators per page.
Distinction whether errors can be ignored when using the page.

Curation by two of the authors if annotators disagreed.
Assuring consistency in the labels.



Instructions

- Select the answers that best describe the web page images.
- Look at each image carefully. Scroll to its bottom if it is not fully shown.
- Read the Guidelines for the second question-group carefully before answering.
- You have to select one answer for each question and for each of the 5 images.
- This task requires JavaScript.

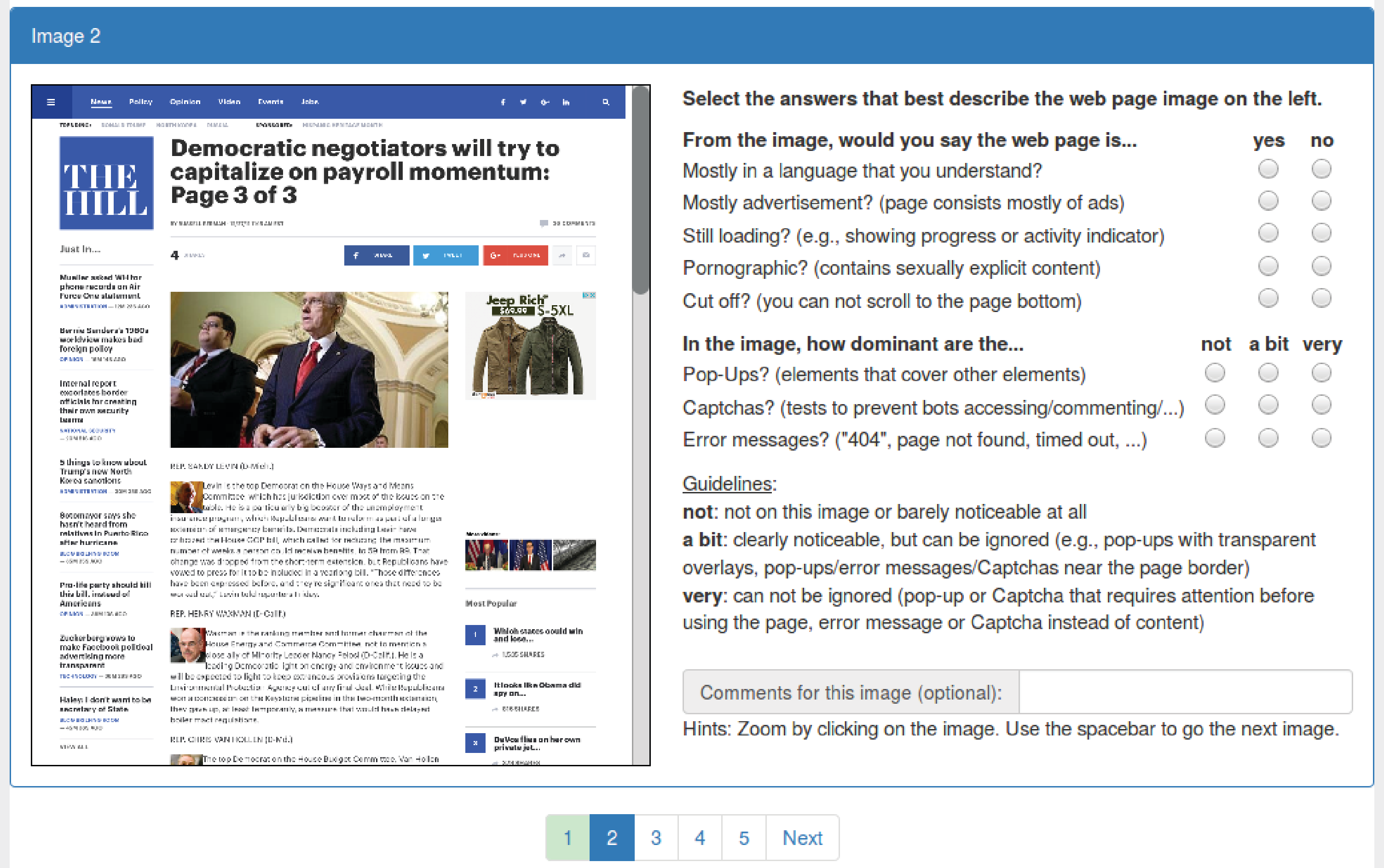


Image 2

News
Policy
Opinion
Video
Events
Jobs
THE HILL
Democratic negotiators will try to capitalize on payroll momentum: Page 3 of 3
Just In...
Jeep Rich S-5XL
Most Popular

Select the answers that best describe the web page image on the left.

From the image, would you say the web page is...	**yes**	**no**
Mostly in a language that you understand?	○	○
Mostly advertisement? (page consists mostly of ads)	○	○
Still loading? (e.g., showing progress or activity indicator)	○	○
Pornographic? (contains sexually explicit content)	○	○
Cut off? (you can not scroll to the page bottom)	○	○

In the image, how dominant are the...	**not**	**a bit**	**very**
Pop-Ups? (elements that cover other elements)	○	○	○
Captchas? (tests to prevent bots accessing/commenting/...)	○	○	○
Error messages? ("404", page not found, timed out, ...)	○	○	○

Guidelines:
not: not on this image or barely noticeable at all
a bit: clearly noticeable, but can be ignored (e.g., pop-ups with transparent overlays, pop-ups/error messages/Captchas near the page border)
very: can not be ignored (pop-up or Captcha that requires attention before using the page, error message or Captcha instead of content)

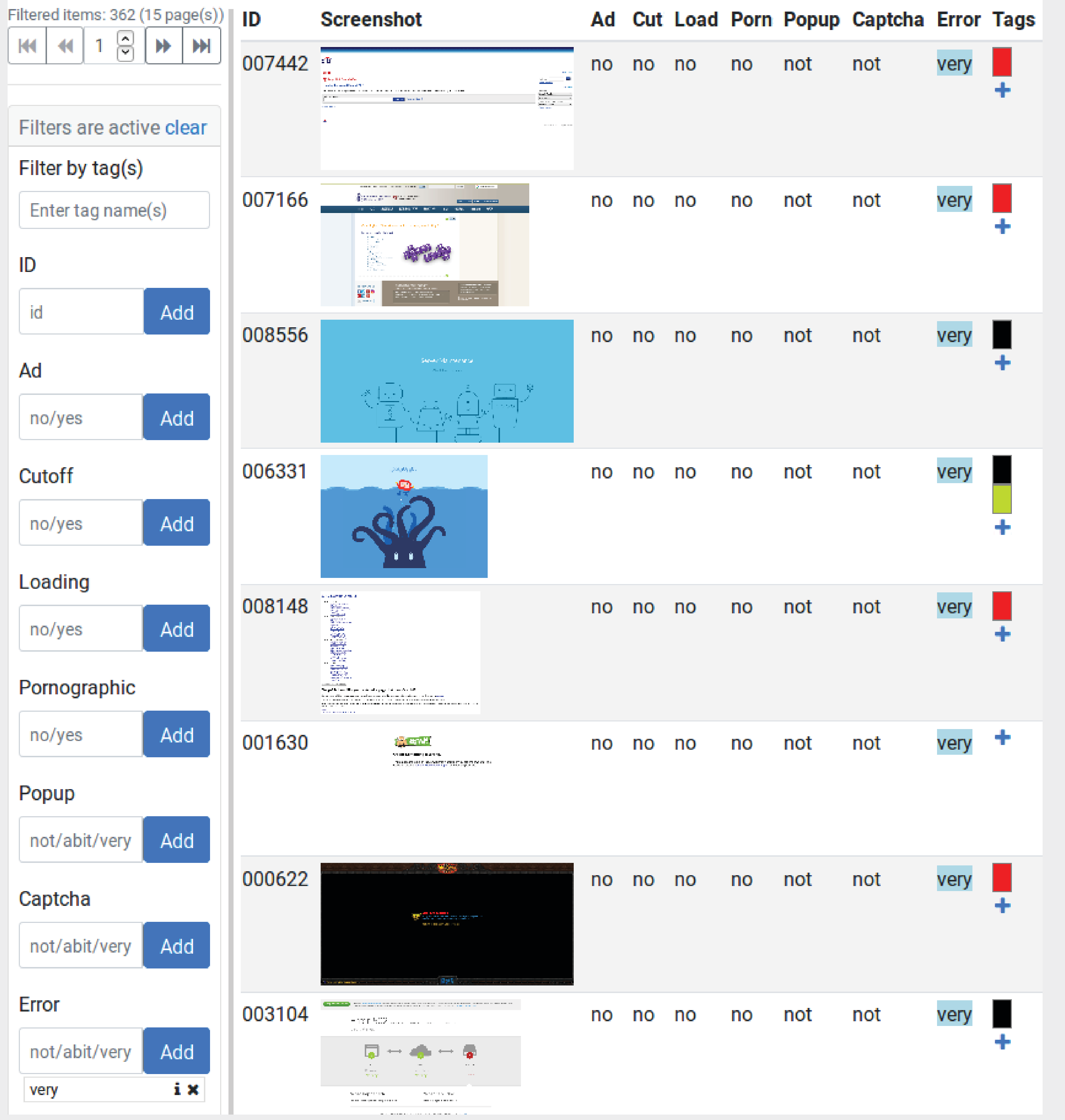
Comments for this image (optional):

Hints: Zoom by clicking on the image. Use the spacebar to go the next image.

1 2 3 4 5 Next

Dataset   **Viewer**

Content error	Agreement	Corrections	Distribution			%Error
			No	Yes		
Ad page	0.65	329	9865	105		1.1
Loading indicators	0.89	48	9950	50		0.5
			Not	A bit	Very	
Pop-ups	0.82	394	9297	315	388	3.9
CAPTCHAs	0.91	124	9865	60	75	0.8
Error messages	0.89	331	9554	83	363	4.5



ID	Screenshot	Ad	Cut	Load	Porn	Popup	Captcha	Error	Tags
007442		no	no	no	no	not	not	very	🟥 +
007166		no	no	no	no	not	not	very	🟥 +
008556		no	no	no	no	not	not	very	⬛ +
006331		no	no	no	no	not	not	very	⬛ 🟩 +
008148		no	no	no	no	not	not	very	🟥 +
001630		no	no	no	no	not	not	very	+
000622		no	no	no	no	not	not	very	🟥 +
003104		no	no	no	no	not	not	very	⬛ +