A Dataset for Content Error Detection in Web Archives

Johannes Kiesel     Fabienne Hubricht     Benno Stein     Martin Potthast

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.

<table>
<thead>
<tr>
<th>Content error</th>
<th>Agreement</th>
<th>Corrections</th>
<th>Distribution</th>
<th>%Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ad page</td>
<td>0.65</td>
<td>329</td>
<td>9865 105</td>
<td>1.1</td>
</tr>
<tr>
<td>Loading indicators</td>
<td>0.89</td>
<td>48</td>
<td>9950 50</td>
<td>0.5</td>
</tr>
<tr>
<td>Pop-ups</td>
<td>0.82</td>
<td>394</td>
<td>9297 315</td>
<td>3.9</td>
</tr>
<tr>
<td>CAPTCHAs</td>
<td>0.91</td>
<td>124</td>
<td>9865 60</td>
<td>0.8</td>
</tr>
<tr>
<td>Error messages</td>
<td>0.89</td>
<td>331</td>
<td>9554 83</td>
<td>4.5</td>
</tr>
</tbody>
</table>