NCleaner
A lightweight and efficient tool for cleaning Web pages

Stefan Evert
University of Osnabrück

stefan.evert@uos.de | purl.org/stefan.evert
The Web as Corpus

- Almost unlimited amounts of data
- Broad range of genres, speakers, etc.
- Always up-to-date
- Freely accessible
- More reasons at WAC-4 on Sunday!

- But it's a little bit messy ...
WaCky problems

- Different languages and encodings
- WaC spam (not quite the same as Web spam)
- Duplicate and derivative Web pages
- Boilerplate and advertising
- Lots of typos, spelling errors, 1337 5P34K, ...
- Non-native speakers (esp. for English)
- Lack of metadata (speaker, genre, ...)
WaCky problems

- Different languages and encodings
- WaC spam (not quite the same as Web spam)
- Duplicate and derivative Web pages
- Boilerplate and advertising
- Lots of typos, spelling errors, 1337 5P34K, ...
- Non-native speakers (esp. for English)
- Lack of metadata (speaker, genre, ...)

3
Lesson Of The Month
Basic Studio Portraiture

Ben Clay/Web Photo School, September, 2001

The basics of portrait photography could fill many large books. We have decided to concentrate on one application with a few variations on the theme for this lesson.

For our backdrop, we draped a black muslin drop cloth on a Boom attached to a Litsstand. Next, we set up a medium Photoflex MultiDome softbox as the main light source to the right of our model (#1 below). We attached the softbox to a Quantum Qflash strobe powered by a Quantum Turbo.

Because the softbox blocks the Qflash’s sensor, we set the flash to manual and dialed in the power, f/stop, and film speed settings by using the Mode, Set, and up/down buttons. We wanted the background to be slightly soft (out of focus), so we determined that the camera’s aperture should be set to f/8. To ensure that there would be no motion blur, we set the shutter speed to 1/250 of a sec. This first exposure shows the main light position and exposure. A one light portrait can be dramatic in effect because of the contrast between light and shadow (#2).

A longer lens does not distort a model’s face the way a normal or wide angle lens can, so we used the 140mm lens on our Contax 645. One of the great things about the Contax is that it comes with 90Å° prismfinder. The prismfinder allows you to look directly at your subject while shooting. This is especially advantageous for shooting portraits as the image is right side up, and the composition of the photo is easy to see.

In the next photo (#3), we used a 120mm macro lens on our Contax 645 for a close-up of our model.

---

Boilerplate example
Lesson Of The Month
Basic Studio Portraiture

Ben Clay/Web Photo School, September, 2001

The basics of portrait photography could fill many large books. We have decided to concentrate on one application with a few variations on the theme for this lesson.

For our backdrop, we draped a black muslin drop cloth on a Boom* attached to a Litesand. Next, we set up a medium Photoflex MultiDome softbox as the main light source to the right of our model (#1 below). We attached the softbox to a Quantum Qflash strobe powered by a Quantum Turbo.

Because the softbox blocks the Qflash’s sensor, we set the flash to manual and dialed in the power, fstop, and film speed settings by using the Mode, Set, and up/down buttons. We wanted the background to be slightly soft (out of focus), so we determined that the camera’s aperture should be set to f/8. To ensure that there would be no motion blur, we set the shutter speed to 1/250 of a sec. This first exposure shows the main light position and exposure. A one light portrait can be dramatic in effect because of the contrast between light and shadow (#2).

A longer lens does not distort a model’s face the way a normal or wide angle lens can, so we used the 140mm lens on our Contax 645. One of the great things about the Contax is that it comes with 90A* prismfinder. The prismfinder allows you to look directly at your subject while shooting. This is especially advantageous for shooting portraits as the image is right side up, and the composition of the photo is easy to see.
Lesson Of The Month
Basic Studio Portraiture

Ben Clay/Web Photo School, September, 2001
Lesson Of The Month
Basic Studio Portraiture

Ben Clay/Web Photo School, September, 2001
The basics of portrait photography could fill many large books. We have decided to concentrate on one application with a few variations on the theme for this lesson.

For our backdrop, we draped a black muslin drop cloth on a Boom attached to a Litestand. Next, we set up a medium Photoflex MultiDome softbox as the main light source to the right of our model (#1 below). We attached the softbox to a Quantum Qflash strobe powered by a Quantum Turbo.

Because the softbox blocks the Qflash's sensor, we set the flash to manual and dialed in the power, f/stop, and film speed settings by using the Mode, Set, and up/down buttons. We wanted the background to be slightly soft (out of focus), so we determined that the camera's aperture should be set to f/8. To ensure that there would be no motion blur, we set the shutter speed to 1/250 of a sec. This first exposure shows the main light position and exposure. A one light portrait can be dramatic in effect because of the contrast between light and shadow (#2).

A longer lens does not distort a model's face the same way a normal or wide angle lens can, so we used the 140mm lens on our Contax 645. One of the great things about the Contax is that it comes with 90° prismfinder. The prismfinder allows you to look directly at your subject while shooting. This is especially advantageous for shooting portraits as the image is right side up, and the composition of the photo is easy to see.

In order to fill in the shadow on the left side of the face, we attached a Litedisc reflector to a Litedisc holder to reflect light into the shadowed areas of our model. We used a soft gold reflector surface, which "warmed up" the model's face (#3).

...
... we added texture to the image. We then eye up and across the image (#8).

Understanding and experimenting with the different elements of your shot enables you to find the shot you're after.

This lesson will be posted in the free public section of the Web Photo School at: www.webphotoschool.com You will be able to enlarge the photos from thumbnails. If you would like to continue your digital step by step education lessons on editing, printing, and e-mailing your photos it will be on the private section of the Web Photo School.

[0901lesson20i1.jpg]

1

[0901lesson20i3.jpg]

3

...

Subscribe to Shutterbug now and receive 12 issues for ONLY $17.95 – and save 62% off the cover price!
If you're serious about photography you need to subscribe to Shutterbug.
Outside the US? Canada or International GIVE A GIFT [s.gif]
[mag_cover.jpg]
Email: _________________________
First Name: _________________________
Last Name: _________________________
Boilerplate removal HowTo
Boilerplate removal HowTo

- HTML tag density (BTE)
- Formatting (lists, colour, CSS classes, etc.)
- Keywords (e.g. Disclaimer, Google Ad)
- Average sentence length, ...
- Grammaticality, POS distribution, ...
- Supervised machine learning
- Sequence models (e.g. CRF)
Boilerplate removal HowTo

- HTML tag density (BTE)
- Formatting (lists, colour, CSS classes, etc.)
- Keywords (e.g. Disclaimer, Google Ad)
- Average sentence length, ...
- Grammaticality, POS distribution, ...
- Supervised machine learning
- Sequence models (e.g. CRF)
- *Or you could do something totally naïve...*
Naïve boilerplate removal

- Extract plain text from Web page, then apply standard n-gram classifier
- Makes no use of ...
  - HTML structure & typographical markup
  - Tag density information
  - Sequential patterns (stretches of clean or dirty text)
  - Linguistic features (grammaticality, POS, ...)
- An interesting baseline experiment
  - if you happen to have training data available
CleanEval results (2007)

<table>
<thead>
<tr>
<th>Team</th>
<th>Text</th>
<th>Seg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bauer et al. (Osnabrück)</td>
<td>73.5</td>
<td>53.5</td>
</tr>
<tr>
<td>Marek, Pecina &amp; Sprousta (Prague)</td>
<td>84.1</td>
<td>65.3</td>
</tr>
<tr>
<td>Hofmann &amp; Weerkamp (Amsterdam)</td>
<td>83.0</td>
<td>65.5</td>
</tr>
<tr>
<td>Chaudhury (India)</td>
<td>80.9</td>
<td>59.5</td>
</tr>
<tr>
<td>Conradie (South Africa)</td>
<td>60.2</td>
<td>45.5</td>
</tr>
<tr>
<td>Gao &amp; Abou-Assaleh (GenieKnows)</td>
<td>83.4</td>
<td>63.9</td>
</tr>
<tr>
<td>Girardi (IRST)</td>
<td>82.5</td>
<td>65.6</td>
</tr>
<tr>
<td>Saralegi &amp; Leturia (Elhuyar Foundation)</td>
<td>83.4</td>
<td>65.3</td>
</tr>
<tr>
<td><strong>Evert (Osnabrück)</strong></td>
<td><strong>82.9</strong></td>
<td><strong>60.3</strong></td>
</tr>
</tbody>
</table>

from Baroni, Chantree, Kilgarriff & Sharoff (2008)
(see there for details of scoring algorithm)
# CleanEval results (2007)

<table>
<thead>
<tr>
<th>Team</th>
<th>Text</th>
<th>Seg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bauer et al. (Osnabrück)</td>
<td>73.5</td>
<td>53.5</td>
</tr>
<tr>
<td>Marek, Pecina &amp; Sprousta (Prague)</td>
<td>84.1</td>
<td>65.3</td>
</tr>
<tr>
<td>Hofmann &amp; Weerkamp (Amsterdam)</td>
<td>83.0</td>
<td>65.5</td>
</tr>
<tr>
<td>Chaudhury (India)</td>
<td>80.9</td>
<td>59.5</td>
</tr>
<tr>
<td>Conradie (South Africa)</td>
<td>60.2</td>
<td>45.5</td>
</tr>
<tr>
<td>Gao &amp; Abou-Assaleh (GenieKnows)</td>
<td>83.4</td>
<td>63.9</td>
</tr>
<tr>
<td>Girardi (IRST)</td>
<td>82.5</td>
<td>65.6</td>
</tr>
<tr>
<td>Saralegi &amp; Leturia (Elhuyar Foundation)</td>
<td>83.4</td>
<td>65.3</td>
</tr>
<tr>
<td><strong>Evert (Osnabrück)</strong></td>
<td><strong>82.9</strong></td>
<td><strong>60.3</strong></td>
</tr>
</tbody>
</table>

from Baroni, Chantree, Kilgarriff & Sharoff (2008) (see there for details of scoring algorithm)
**NCleaner architecture**

- character-level n-gram models (clean vs. dirty)
- default: \( n = 3 \) (has little influence)
- geometric interpolation
- heuristics only do not perform well
- n-gram models can be applied to non-HTML data (or existing text dumps of Web pages)
NCleaner implementation

◆ Portable & easy to use
  • platform-independent Perl implementation
  • optional: efficient C code for n-gram models
◆ Lightweight
  • standard parameter file: 2.3 MB (uncompressed)
◆ Fast
  • 20 million words / hour (Perl)
  • 120 million words / hour (Perl + C)
◆ Open source @ webascorpus.sf.net
The basics of portrait photography could fill many large books. We have decided to concentrate on one application with a few variations on the theme for this lesson.

For our backdrop, we draped a black muslin drop cloth on a Boom stand. Next, we set up a medium Photoflex MultiDome softbox as the main light source to the right of our model (Fig. 1 below). We attached the softbox to a Quantum Qflash strobe powered by a Quantum Turbo.

Because the softbox blocks the Qflash’s sensor, we set the flash to manual and dialed in the power, f/stop, and film speed settings by using the Mode, Set, and up/down buttons. We wanted the background to be slightly soft (out of focus), so we determined that the camera’s aperture should be set to f/8. To ensure that there would be no motion blur, we set the shutter speed to 1/250 of a sec. This first exposure shows the main light position and exposure. A one light portrait can be dramatic in effect because of the contrast between light and shadow (Fig. 2).

A longer lens does not distort a model’s face the way a normal or wide angle lens can, so we used the 140mm lens on our Contax 645. One of the great things about the Contax is that it comes with 90A* prismfinder. The prismfinder allows you to look directly at your subject while shooting. This is especially advantageous for shooting portraits as the image is right side up, and the composition of the photo is easy to see.
The basics of portrait photography could fill many large books. We have decided to concentrate on one application with a few variations on the theme for this lesson.

For our backdrop, we draped a black muslin drop cloth on a Boom attached to a Litestand. Next, we set up a medium Photoflex MultiDome softbox as the main light source to the right of our model (#1 below). We attached the softbox to a Quantum Qflash strobe powered by a Quantum Turbo.

Because the softbox blocks the Qflash's sensor, we set the flash to manual and dialed in the power, fstop, and film speed settings by using the Mode, Set, and up/down buttons. We wanted the background to be slightly soft (out of focus), so we determined that the camera’s aperture should be set to f/8. To ensure that there would be no motion blur, we set the shutter speed to 1/250 of a sec. This first exposure shows the main light position and exposure. A one light portrait can be dramatic in effect because of the contrast between light and shadow (#2).

A longer lens does not distort a model's face the way a normal or wide angle lens can, so we used the 140mm lens on our Contax 645. One of the great things about the Contax is that it comes with 90A° prismfinder. The prismfinder allows you to look directly at your subject while shooting. This is especially advantageous for shooting portraits as the image is right side up, and the composition of the photo is easy to see.
Evaluation

cross-validation

CleanEval test set

(percentage of words, micro-averaged, using cleaneval.py script)
Language-independent?

- Statistical methods are language-independent, but require training data for each new language
  - NCleaner standard parameter file was trained on 168 manually cleaned English Web pages

- Can NCleaner be used for other languages?
  1. re-train NCleaner on as little data as possible
  2. apply standard parameter file (trained on English) to other European languages
Learning curve

NCleaner learning curve

Accuracy

F-score  precision  recall

Training size (tokens)

0  50000  100000  150000  200000  250000  300000  350000

88  90  92  94  96  98
A case study for German

- Downloaded 10 random German Web pages
- Manually cleaned
- Evaluation of standard NCleaner parameter file
- Some pages work very well, others poorly
Juristen

Karrierekick Großkanzlei

Von Corinna Budras

05. Oktober 2005

Der Auftrag klingt simpel: „Liefern Sie einen für beide Seiten akzeptablen Vertrag ab - vermeiden Sie Extrempositionen.“ Doch die Verantwortung ist riesig.


Bedingungen wie in der Wirklichkeit. Nur daß hier im achten Stock eines Hochhauses mitten im Frankfurter Bankenviertel keine Topanwälte sitzen, sondern junge Hochschulabsolventen, die das noch werden wollen: sechzehn hochqualifizierte Juristen, die meisten mit einem Prädikatsexamen und diversen Auslandsaufenthalten.


Den drei besten Kandidaten winkt ein Stipendium für einen dreimonatigen Aufenthalt in einem Auslandsbüro von Linklaters. Diese Station werden sie während ihres Referendariats vor dem zweiten Staatsexamen absolvieren.
Karrierekick Großkanzlei

Von Corinna Budras

05. Oktober 2005
Der Auftrag klingt simpel: „Liefern Sie einen für beide Seiten akzeptablen Vertrag ab - vermeiden Sie Extrempositionen.“ Doch die Verantwortung ist riesig.


Bedingungen wie in der Wirklichkeit. Nur daß hier im achten Stock eines Hochhauses mitten im Frankfurter Bankenviertel keine Topanwälte sitzen, sondern junge Hochschulabsolventen, die das noch werden wollen: sechszehn hochqualifizierte Juristen, die meisten mit einem Prädikatsexamen und diversen Auslandsaufenthalten.

NCleaner highlights

- State-of-the-art accuracy (almost :-)
- Lightweight
- Fast
- Portable & easy to use
- Open source

[Software description]

NCleaner-1.0

NCleaner is a simple tool for automatic boilerplate removal, using character-level n-gram models as classifiers. Since it does not make use of HTML structure, it can also be applied to existing text dumps of Web pages. NCleaner participated in CleanEval-1 (under the working title StupidOS), where it achieved competitive results for text cleanup (though not for segmentation accuracy). The first official release includes both a fully portable Perl implementation, as well as C code for faster processing on supported platforms (more than 20 million words per hour on a standard desktop computer).

http://webascorpus.sf.net/
Next steps

- Get better training data
- Improve parameter tuning
- Add sequencing model (HMM)
- Include HTML tags in n-gram models
Thank you!