

Web content mining

IKT 407 – Web mining

Sigbjørn Tvedt
Christian Kroken

Why?

- Improve accessibility of webpages
- Classify images
 - Formulas
 - Text
 - Images
 - Logos

How?

- Ruby
 - Camellia
 - Blob detection
- Python
 - Pil (Python image library)
 - Cropping

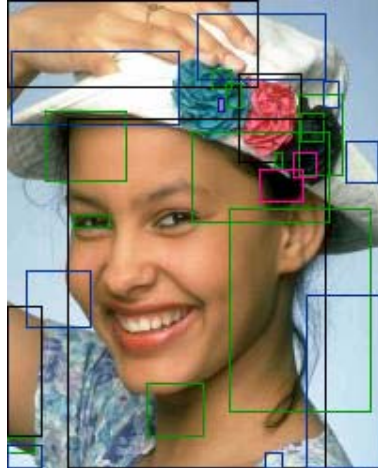
Methods

- Color detection
- Text pattern detection
- Intersection

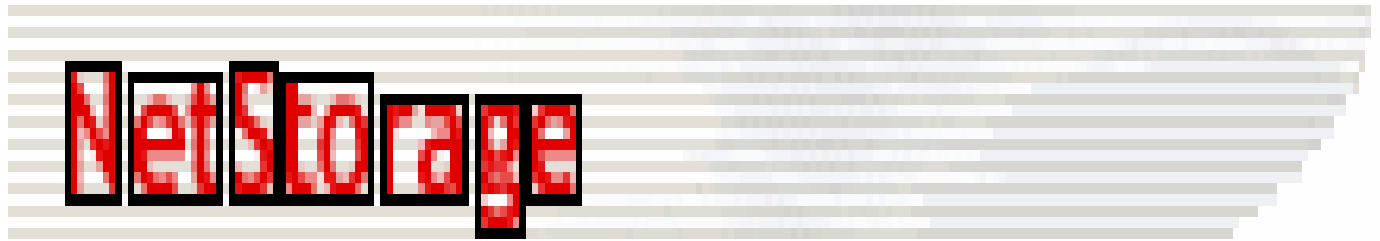
	<i>Image</i>	<i>Text</i>	<i>Text logo</i>	<i>Formula</i>
<i>Intersections</i>	<i>Yes</i>	<i>No</i>	<i>Yes</i>	<i>No</i>
<i>Line detection</i>	<i>No</i>	<i>Yes</i>	<i>Yes</i>	<i>No</i>

Classification

- Image



- Logo



Classification (cont.)

- Text



- Formula

$$F(x) = \int_5^{\vartheta} \frac{Az \delta y}{X \delta x}$$

Optimization

<i>44 files</i>	<i>Time used with DoubleScan Total/average</i>	<i>Errors with DoubleScan</i>	<i>Time used without DoubleScan Total/average</i>	<i>Errors without DoubleScan</i>
<i>Without maximum intersections</i>	<i>23.469s 0.574s</i>	<i>3 Errors 3 not classified</i>	<i>14.125s 0.353s</i>	<i>4 Errors 4 not classified</i>
<i>With maximum intersections set to 100</i>	<i>12.469s 0.304s</i>	<i>2 errors 3 not classified</i>	<i>5.344s 0.134s</i>	<i>3 Errors 4 not classified</i>