

Úklid a čištění jako věda

A také zpráva o jednom vítězství

Pavel Pecina

Michal Marek, Miroslav Spousta

Ústav formální a aplikované lingvistiky
Matematicko-fyzikální fakulta Univerzity Karlovy



Mixer, 21.listopad, 2007

Motivace


WAC3
2007

WAC3 - 2007



Web as Corpus 2007, UCLouvain, Louvain-la-Neuve, September 15-16 2007 (Belgium)

WAC3
<ul style="list-style-type: none">◦ Call for papers◦ Submit a paper◦ Registration◦ Program◦ Scientific committee◦ Travel info. & venue◦ Local organisation team◦ Associated events◦ SIGWAC◦ Previous Workshops◦ Pictures
Cleaneval
<ul style="list-style-type: none">◦ Information◦ Scientific committee
Organisation
 UCL Université catholique de Louvain

SIGWAC

SIGWAC is the Special Interest Group of the Association for Computational Linguistics (ACL) on Web As Corpus. Its objectives are

- ◆ to promote interest in the use of the web as a source of linguistic data, and as an object of study in its own right;
- ◆ to provide members of the ACL with a special interest in the web-as-corpus with a means of exchanging news of recent research developments and other matters of interest;
- ◆ to sponsor meetings and workshops on the web as corpus that appear to be timely and worthwhile.

More info on [SIGWAC](#)

Last update : July, 2007



Motivace



WAC3 - 2007



Web as Corpus 2007, UCLouvain, Louvain-la-Neuve, September 15-16 2007 (Belgium)

WAC3	SIGWAC
<ul style="list-style-type: none">◦ Call for papers◦ Submit a paper◦ Registration◦ Program◦ Scientific committee◦ Travel info. & venue◦ Local organisation team◦ Associated events◦ SIGWAC◦ Previous Workshops◦ Pictures	<p>SIGWAC is the Special Interest Group of the Association for Computational Linguistics (ACL) on Web as Corpus Its objectives are</p> <ul style="list-style-type: none">◆ to promote interest in the use of the web as a source of linguistic data, and as an object of study in its own right;◆ to provide members of the ACL with a special interest in the web-as-corpus with a means of exchanging news of recent research developments and other matters of interest;◆ to sponsor meetings and workshops on the web as corpus that appear to be timely and worthwhile. <p>More info on SIGWAC</p>
Cleaneval	
<ul style="list-style-type: none">◦ Information◦ Scientific committee	
Organisation	
 <p>UCL Université catholique de Louvain</p>	

Last update : July, 2007

Motivace

WAC3 - 2007

Web as Corpus 2007, UCLouvain, Louvain-la-Neuve, September 15-16 2007 (Belgium)

WAC3

- » Call for papers
- » Submit a paper
- » Registration
- » Program
- » Scientific committee
- » Travel info. & venue
- » Local organisation team
- » Associated events
- » SIGWAC
- » Previous Workshops
- » Pictures

Clothesline

- » Information
- » Scientific committee

Organization


UCL
Université
catholique
de Louvain

SIGWAC

SIGWAC is the Special Interest Group of the Association for Computational Linguistics (ACL) on Web As Corpus. Its objectives are

- ◆ to promote interest in the use of the web as a source of linguistic data, and as an object of study in its own right;
- ◆ to provide members of the ACL with a special interest in the web-as-corpus with a means of exchanging news of recent research developments and other matters of interest;
- ◆ to sponsor meetings and workshops on the web as corpus that appear to be timely and worthwhile.

More info on **SIGWAC**

Last update : July, 2007

Navigation icons: back, forward, search, etc.

Aplikace

1. Indexování webových stránek pro vyhledávání

- ▶ text (slova), který nesouvisí s obsahem stránky, je pro indexaci nevhodný,
vyhledání na základě takových slov nevede k nalezení relevantní stránky
- ▶ vyčištěná stránka se dál neprezentuje,
při vyhledávání se odkazuje na původní stránku
- ▶ při čištění důraz kladen na recall,
nízkou precision "zachráni" indexační algoritmus

2. Vytváření webových korpusů pro lingvistické účely

- ▶ text, který nesouvisí s obsahem stránky, není pro takové využití vhodný
fragmentace, opakující se slova, chybějící segmentace, negramatičnost
- ▶ čištění se provádí jako základní předzpracování,
veškerá další zpracování probíhá až následně na vyčištěných datech
- ▶ při čištění důraz kladen na precision,
nízký recall lze "dohnat" zpracováním většího množství dat

Web jako korpus

Idea

- ▶ využití dat dostupných na internetu jako zdroje lingvistické evidence
- ▶ budť jako klasické korpusy nebo jako kolekce pro vyhledávání informací
- ▶ zaměření na textová data (zatím!)

Přístupy

- ▶ **vzdálený** – využití vyhledávacích služeb jako rozhraní k přístupu k datům,
pravděpodobnost řetezce slov odhadnuta na základě počtu "hitů"
- ▶ **lokální** – automatické stahování dat a vytváření klasického korpusu

Problémy

- ▶ stahování: *technologie, infrastruktura*
- ▶ detekce kódování a identifikace jazyka
- ▶ čištění: *odstranění nežádoucích částí*

Čištění webových stránek

Postup

1. odlišit zrno od plev
2. zrno ponechat
3. plevy vyhodit

Plevy

- ▶ menu, navigační panely
- ▶ seznamy (interních i externích) odkazů
- ▶ copyrightové informace
- ▶ hlavičky, patičky a další prvky šablon
- ▶ reklama

Zrno

- ▶ text, který tvoří "obsah" stránky
- ▶ strukturován pomocí následujících značek:
 - <h> nadpisy
 - <p> odstavce
 - <l> položky seznamů

Příklad

The screenshot shows the homepage of the WAC3 - 2007 conference website. At the top, there's a banner with the text "WAC3 - 2007" and a small logo. Below the banner, the main title is "Web as Corpus 2007, UCLouvain, Louvain-la-Neuve, September 15-16 2007 (Belgium)". On the left side, there's a sidebar with navigation links: "WAC3", "Call for papers", "Submit a paper", "Registration", "Program", "Scientific committee", "Travel info. & venue", "Local organisation team", "Associated events", "SIGWAC", "Previous Workshops", and "Pictures". The main content area has a section titled "SIGWAC" which includes a description of SIGWAC as the Special Interest Group of the Association for Computational Linguistics (ACL) on Web as Corpus, its objectives, and a bulleted list of its goals. It also links to "More info on SIGWAC". At the bottom, there's a footer with the UCL logo and text indicating the last update was in July 2007.

WAC3

- o Call for papers
- o Submit a paper
- o Registration
- o Program
- o Scientific committee
- o Travel info. & venue
- o Local organisation team
- o Associated events
- o SIGWAC
- o Previous Workshops
- o Pictures

Class news

- o Information
- o Scientific committee

Organisation

 **UCL**
Université catholique de Louvain

Last update: July, 2007

SIGWAC

SIGWAC is the Special Interest Group of the Association for Computational Linguistics (ACL) on Web as Corpus
Its objectives are

- ◆ to promote interest in the use of the web as a source of linguistic data, and as an object of study in its own right;
- ◆ to provide members of the ACL with a special interest in the web-as-corpus with a means of exchanging news of recent research developments and other matters of interest;
- ◆ to sponsor meetings and workshops on the web as corpus that appear to be timely and worthwhile.

More info on **SIGWAC**

<h>Web as Corpus 2007, UCLouvain, Louvain-la-Neuve, September 15-16 2007 (Belgium)

<h>SIGWAC

<p>SIGWAC is the Special Interest Group of the Association for Computational Linguistics (ACL) on Web as Corpus Its objectives are

to promote interest in the use of the web as a source of linguistic data, and as an object of study in its own right;

to provide members of the ACL with a special interest in the web-as-corpus with a means of exchanging news of recent research developments and other matters of interest;

to sponsor meetings and workshops on the web as corpus that appear to be timely and worthwhile.

<p>More info on SIGWAC

Příklad



WAC3 - 2007

Web as Corpus 2007, UCLouvain, Louvain-la-Neuve, September 15-16 2007 (Belgium)

WAC3

- Call for papers
- Submit a paper
- Registration
- Program
- Scientific committee
- Travel info. & venue
- Local organisation team
- Associated events
- SIGWAC
- Previous Workshops
- Pictures

Claire neval

- Information
- Scientific committee

Organisation

 **UCL**
Université catholique de Louvain

Last update : July, 2007

SIGWAC

SIGWAC is the Special Interest Group of the Association for Computational Linguistics (ACL) on Web as Corpus
Its objectives are

- ◆ to promote interest in the use of the web as a source of linguistic data, and as an object of study in its own right;
- ◆ to provide members of the ACL with a special interest in the web-as-corpus with a means of exchanging news of recent research developments and other matters of interest;
- ◆ to sponsor meetings and workshops on the web as corpus that appear to be timely and worthwhile.

More info on **SIGWAC**

<**h**>Web as Corpus 2007, UCLouvain, Louvain-la-Neuve, September 15-16 2007 (Belgium)

<**h**>SIGWAC

<**p**>SIGWAC is the Special Interest Group of the Association for Computational Linguistics (ACL) on Web as Corpus Its objectives are

<**l**>to promote interest in the use of the web as a source of linguistic data, and as an object of study in its own right;

<**l**>to provide members of the ACL with a special interest in the web-as-corpus with a means of exchanging news of recent research developments and other matters of interest;

<**l**>to sponsor meetings and workshops on the web as corpus that appear to be timely and worthwhile.

<**p**>More info on SIGWAC

Příklad

The screenshot shows the WAC3 - 2007 website. The main header features the logo for "WAC3 - 2007". Below it, a red box highlights the text: "Web as Corpus 2007, UCLouvain, Louvain-la-Neuve, September 15-16 2007 (Belgium)". On the left, there's a sidebar with links for "Call for papers", "Submit a paper", "Registration", "Program", "Scientific committee", "Travel info. & venue", "Local organisation team", "Associated events", "SIGWAC", "Previous Workshops", and "Pictures". The main content area has a red box around the "SIGWAC" section. This section includes a heading "SIGWAC", a description of SIGWAC as the Special Interest Group of the Association for Computational Linguistics (ACL) on Web as Corpus, and its objectives. The objectives are listed in a bulleted list:

- ◆ to promote interest in the use of the web as a source of linguistic data, and as an object of study in its own right;
- ◆ to provide members of the ACL with a special interest in the web-as-corpus with a means of exchanging news of recent research developments and other matters of interest;
- ◆ to sponsor meetings and workshops on the web as corpus that appear to be timely and worthwhile.

Below this, a link reads "More info on SIGWAC". At the bottom, there's a footer with the UCL logo and the text "Last update: July, 2007".

<h>Web as Corpus 2007, UCLouvain, Louvain-la-Neuve, September 15-16 2007 (Belgium)

SIGWAC

<p>SIGWAC is the Special Interest Group of the Association for Computational Linguistics (ACL) on Web as Corpus Its objectives are

to promote interest in the use of the web as a source of linguistic data, and as an object of study in its own right;

to provide members of the ACL with a special interest in the web-as-corpus with a means of exchanging news of recent research developments and other matters of interest;

to sponsor meetings and workshops on the web as corpus that appear to be timely and worthwhile.

<p>More info on SIGWAC

Příklad



WAC3 - 2007

Web as Corpus 2007, UCLouvain, Louvain-la-Neuve, September 15-16 2007 (Belgium)

WAC3

- Call for papers
- Submit a paper
- Registration
- Program
- Scientific committee
- Travel info. & venue
- Local organisation team
- Associated events
- SIGWAC
- Previous Workshops
- Pictures

Clauses

- Information
- Scientific committee

Organisation

 **UCL**
Université catholique de Louvain

Last update : July, 2007

SIGWAC

SIGWAC is the Special Interest Group of the Association for Computational Linguistics (ACL) on Web as Corpus
Its objectives are

- ◆ to promote interest in the use of the web as a source of linguistic data, and as an object of study in its own right;
- ◆ to provide members of the ACL with a special interest in the web-as-corpus with a means of exchanging news of recent research developments and other matters of interest;
- ◆ to sponsor meetings and workshops on the web as corpus that appear to be timely and worthwhile.

More info on **SIGWAC**

<h>Web as Corpus 2007, UCLouvain, Louvain-la-Neuve, September 15-16 2007 (Belgium)

<h>SIGWAC

<p>SIGWAC is the Special Interest Group of the Association for Computational Linguistics (ACL) on Web as Corpus Its objectives are

<!>to promote interest in the use of the web as a source of linguistic data, and as an object of study in its own right;

<!>to provide members of the ACL with a special interest in the web-as-corpus with a means of exchanging news of recent research developments and other matters of interest;

<!>to sponsor meetings and workshops on the web as corpus that appear to be timely and worthwhile.

<p>More info on SIGWAC

Příklad

The screenshot shows the homepage of the WAC3 - 2007 conference website. At the top, there are navigation links for Motivace, Cíle, Prostředky, Soutěž, Výsledky, and Závěr. Below these are two logo boxes: one for WAC3 and another for the Association for Computational Linguistics (ACL). The main content area features a large red box containing the text: "Web as Corpus 2007, UCLouvain, Louvain-la-Neuve, September 15-16 2007 (Belgium)". To the left is a sidebar with links for Call for papers, Submit a paper, Registration, Program, Scientific committee, Travel info. & venue, Local organisation team, Associated events, SIGWAC, Previous Workshops, and Pictures. Below this is a "Class notes" section with links for Information and Scientific committee. At the bottom is an "Organisation" section featuring the UCLouvain logo.

Web as Corpus 2007, UCLouvain, Louvain-la-Neuve, September 15-16 2007 (Belgium)

SIGWAC

SIGWAC is the Special Interest Group of the Association for Computational Linguistics (ACL) on Web as Corpus
Its objectives are

- ◆ to promote interest in the use of the web as a source of linguistic data, and as an object of study in its own right;
- ◆ to provide members of the ACL with a special interest in the web-as-corpus with a means of exchanging news of recent research developments and other matters of interest;
- ◆ to sponsor meetings and workshops on the web as corpus that appear to be timely and worthwhile.

More info on **SIGWAC**

Last update : July, 2007

UCL
Université catholique de Louvain

<**h**>Web as Corpus 2007, UCLouvain, Louvain-la-Neuve, September 15-16 2007 (Belgium)

SIGWAC

<**p**>SIGWAC is the Special Interest Group of the Association for Computational Linguistics (ACL) on Web as Corpus Its objectives are

<**!>**to promote interest in the use of the web as a source of linguistic data, and as an object of study in its own right;

<**!>**to provide members of the ACL with a special interest in the web-as-corpus with a means of exchanging news of recent research developments and other matters of interest;

<**!>**to sponsor meetings and workshops on the web as corpus that appear to be timely and worthwhile.

<**p**>More info on SIGWAC

Příklad



WAC3 - 2007

Web as Corpus 2007, UCLouvain, Louvain-la-Neuve, September 15-16 2007 (Belgium)

WAC3

- Call for papers
- Submit a paper
- Registration
- Program
- Scientific committee
- Travel info. & venue
- Local organisation team
- Associated events
- SIGWAC
- Previous Workshops
- Pictures

Clauses

- Information
- Scientific committee

Organisation

 **UCL**
Université catholique de Louvain

Last update : July, 2007

SIGWAC

SIGWAC is the Special Interest Group of the Association for Computational Linguistics (ACL) on Web as Corpus
Its objectives are

- ◆ to promote interest in the use of the web as a source of linguistic data, and as an object of study in its own right;
- ◆ to provide members of the ACL with a special interest in the web-as-corpus with a means of exchanging news of recent research developments and other matters of interest;
- ◆ to sponsor meetings and workshops on the web as corpus that appear to be timely and worthwhile.

More info on **SIGWAC**

<**h**>Web as Corpus 2007, UCLouvain, Louvain-la-Neuve, September 15-16 2007 (Belgium)

<**h**>SIGWAC

<**p**>SIGWAC is the Special Interest Group of the Association for Computational Linguistics (ACL) on Web as Corpus Its objectives are

<**l**>to promote interest in the use of the web as a source of linguistic data, and as an object of study in its own right;

<**l**>to provide members of the ACL with a special interest in the web-as-corpus with a means of exchanging news of recent research developments and other matters of interest;

<**l**>to sponsor meetings and workshops on the web as corpus that appear to be timely and worthwhile.

<**p**>More info on SIGWAC

Příklad

The screenshot shows the homepage of the WAC3 - 2007 conference website. At the top, there are navigation links for Motivace, Cíle, Prostředky, Soutěž, Výsledky, and Závěr. Below these are two large images: one of a computer monitor showing a presentation slide and another of a globe. A red box highlights the title "Web as Corpus 2007, UCLouvain, Louvain-la-Neuve, September 15-16 2007 (Belgium)". On the left, there's a sidebar with sections for WAC3, Call for papers, Submit a paper, Registration, Program, Scientific committee, Travel info. & venue, Local organisation team, Associated events, SIGWAC, Previous Workshops, and Pictures. Another red box highlights the "SIGWAC" section. The SIGWAC section contains a box with the text: "SIGWAC is the Special Interest Group of the Association for Computational Linguistics (ACL) on Web as Corpus. Its objectives are". Below this are three bullet points: 1. To promote interest in the use of the web as a source of linguistic data, and as an object of study in its own right; 2. To provide members of the ACL with a special interest in the web-as-corpus with a means of exchanging news of recent research developments and other matters of interest; 3. To sponsor meetings and workshops on the web as corpus that appear to be timely and worthwhile. At the bottom, there's a "More info on SIGWAC" link, a copyright notice for Université catholique de Louvain, and a footer with "Last update: July, 2007".

Web as Corpus 2007, UCLouvain, Louvain-la-Neuve, September 15-16 2007 (Belgium)

SIGWAC

SIGWAC is the Special Interest Group of the Association for Computational Linguistics (ACL) on Web as Corpus. Its objectives are

- ◆ to promote interest in the use of the web as a source of linguistic data, and as an object of study in its own right;
- ◆ to provide members of the ACL with a special interest in the web-as-corpus with a means of exchanging news of recent research developments and other matters of interest;
- ◆ to sponsor meetings and workshops on the web as corpus that appear to be timely and worthwhile.

More info on **SIGWAC**

Last update: July, 2007

UCL
Université
catholique
de Louvain

<h>Web as Corpus 2007, UCLouvain, Louvain-la-Neuve, September 15-16 2007 (Belgium)

<h>SIGWAC

<p>SIGWAC is the Special Interest Group of the Association for Computational Linguistics (ACL) on Web as Corpus Its objectives are

<i>to promote interest in the use of the web as a source of linguistic data, and as an object of study in its own right;

<i>to provide members of the ACL with a special interest in the web-as-corpus with a means of exchanging news of recent research developments and other matters of interest;

<i>to sponsor meetings and workshops on the web as corpus that appear to be timely and worthwhile.

<p>More info on **SIGWAC**

Příklad

The screenshot shows the WAC3 - 2007 website. At the top, there's a banner with the text "WAC3 - 2007". Below it, a red box highlights the title "Web as Corpus 2007, UCLouvain, Louvain-la-Neuve, September 15-16 2007 (Belgium)". On the left, there's a sidebar with links for "Call for papers", "Submit a paper", "Registration", "Program", "Scientific committee", "Travel info. & venue", "Local organisation team", "Associated events", "SIGWAC", "Previous Workshops", and "Pictures". The main content area has a section titled "SIGWAC" with a red border. It contains text about SIGWAC being the Special Interest Group of the Association for Computational Linguistics (ACL) on Web as Corpus, its objectives, and a "More info on SIGWAC" link. The footer includes the UCL logo and a "Last update: July, 2007" note.

Web as Corpus 2007, UCLouvain, Louvain-la-Neuve, September 15-16 2007 (Belgium)

SIGWAC

SIGWAC is the Special Interest Group of the Association for Computational Linguistics (ACL) on Web as Corpus Its objectives are

- ◆ to promote interest in the use of the web as a source of linguistic data, and as an object of study in its own right;
- ◆ to provide members of the ACL with a special interest in the web-as-corpus with a means of exchanging news of recent research developments and other matters of interest;
- ◆ to sponsor meetings and workshops on the web as corpus that appear to be timely and worthwhile.

More info on **SIGWAC**

Last update: July, 2007

UCL
Université catholique de Louvain

<h>Web as Corpus 2007, UCLouvain, Louvain-la-Neuve, September 15-16 2007 (Belgium)

SIGWAC

<p>SIGWAC is the Special Interest Group of the Association for Computational Linguistics (ACL) on Web as Corpus Its objectives are

to promote interest in the use of the web as a source of linguistic data, and as an object of study in its own right;

to provide members of the ACL with a special interest in the web-as-corpus with a means of exchanging news of recent research developments and other matters of interest;

to sponsor meetings and workshops on the web as corpus that appear to be timely and worthwhile.

<p>More info on SIGWAC

Příklad



WAC3 - 2007



Web as Corpus 2007, UCLouvain, Louvain-la-Neuve, September 15-16 2007 (Belgium)

WAC3

- Call for papers
- Submit a paper
- Registration
- Program
- Scientific committee
- Travel info. & venue
- Local organisation team
- Associated events
- SIGWAC
- Previous Workshops
- Pictures

SIGWAC

SIGWAC is the Special Interest Group of the Association for Computational Linguistics (ACL) on Web as Corpus. Its objectives are

- ◆ to promote interest in the use of the web as a source of linguistic data, and as an object of study in its own right;
- ◆ to provide members of the ACL with a special interest in the web-as-corpus with a means of exchanging news of recent research developments and other matters of interest;
- ◆ to sponsor meetings and workshops on the web as corpus that appear to be timely and worthwhile.

More info on [SIGWAC](#)

CleanEval

- Information
- Scientific committee

Organisation



UCL
Université catholique de Louvain

Last update : July, 2007

<**h**>Web as Corpus 2007, UCLouvain, Louvain-la-Neuve, September 15-16 2007 (Belgium)

<**h**>SIGWAC

<**p**>SIGWAC is the Special Interest Group of the Association for Computational Linguistics (ACL) on Web as Corpus Its objectives are

<**I**>to promote interest in the use of the web as a source of linguistic data, and as an object of study in its own right;

<**I**>to provide members of the ACL with a special interest in the web-as-corpus with a means of exchanging news of recent research developments and other matters of interest;

<**I**>to sponsor meetings and workshops on the web as corpus that appear to be timely and worthwhile.

<**p**>More info on SIGWAC

Jak na to ...

1. vstupní HTML soubor je pomocí utilitky Tidy převeden do validního XHTML
2. validní dokument je automaticky zbaven skriptů a stylů
3. dokument je parsován a rozdělen do textových bloků oddělených HTML značkami
4. pro každý blok jsou extrahovány hodnoty rysů, které popisují jeho "kvalitu"
5. na základě hodnot rysů svých a okolních bloků je každý blok klasifikován (označen) jednou z následujících značek (*sequence labeling*):
 - ▶ **<h>**, **<p>**, **<l>** - *nadpis, odstavec, položka seznamu*
 - ▶ **<c>** - *pokračování nadpisu, odstavce, nebo položky seznamu*
 - ▶ **<o>** - *ostatní*
6. bloky označené **<h>**, **<p>**, **<l>** zůstávají v textu a jsou jim předřazeny jejich značky
7. bloky označené **<c>** zůstávají v textu tak, jak jsou
8. bloky označené **<o>** jsou odstraněny

Příklad

Krok 0: Vstupní HTML soubor

```
<html>
<head>
<title>Sample Web Page</title>
<style> body color: green; </style>
</head>
<body>
<h1>Hello World!</h1>
<p>This is a simple webpage made of a paragraph and a list.</p>
<ul>
<li>It has <b>bold</b> fonts.
<li>And <i>italic</i>, too.
</ul>
<p><a href="mailto:mail@example.org">contact</a>
</body>
</html>
```

Příklad

Krok 1: Normalizace pomocí Tidy

```
<html>
<head>
<title>Sample Web Page</title>
<style> body color: green; </style>
</head>
<body>
<h1>Hello World!</h1>
<p>This is a simple webpage made of a paragraph and a list.</p>
<ul>
<li>It has <b>bold</b> fonts.</li>
<li>And <i>italic</i>, too.</li>
</ul>
<p><a href="mailto:mail@example.org">contact</a></p>
</body>
</html>
```

Příklad

Krok 2: Odstranění skriptu

```
<html>
<head>
<title>Sample Web Page</title>
<style> body color: green; </style>
</head>
<body>
<h1>Hello World!</h1>
<p>This is a simple webpage made of a paragraph and a list.</p>
<ul>
<li>It has <b>bold</b> fonts.</li>
<li>And <i>italic</i>, too.</li>
</ul>
<p><a href="mailto:mail@example.org">contact</a></p>
</body>
</html>
```

Příklad

Krok 3: Identifikace textových bloků

```
<html>
<head>
<title>Sample Web Page</title>
<style> body { color: green; } </style>
</head>
<body>
<h1>Hello World!</h1>
<p>This is a simple webpage made of a paragraph and a list.</p>
<ul>
<li>It has <b>bold</b> fonts.</li>
<li>And <i>italic</i>, too.</li>
</ul>
<p><a href="mailto:mail@example.org">contact</a></p>
</body>
</html>
```

Příklad

Krok 4: Získání hodnot rysů

```
<html>
<head>
<title>Sample Web Page</title>
<style> body color: green; </style>
</head>
<body>
<h1>Hello World!</h1>
<p>This is a simple webpage made of a paragraph and a list.</p>
<ul>
<li>It has <b>bold</b> fonts.</li>
<li>And <i>italic</i>, too.</li>
</ul>
<p><a href="mailto:mail@example.org">contact</a></p>
</body>
</html>
```

Příklad

Krok 5: Značkování

```
<html>
<head>
<title>Sample Web Page</title>
<style> body color: green; </style>
</head>
<body>
<h1>Hello World!</h1>
<p>This is a simple webpage made of a paragraph and a list.</p>
<ul>
<li>It has <b>bold</b> fonts.</li>
<li>And <i>italic</i>, too.</li>
</ul>
<p><a href="mailto:mail@example.org">contact</a></p>
</body>
</html>
```

Příklad

Krok 5: Značkování

```
<html>
<head>
<title><h>Sample Web Page</title>
<style> body color: gReen; </style>
</head>
<body>
<h1><h>Hello World!</h1>
<p><p>This is a simple webpage made of a paragraph and a list.</p>
<ul>
<li><l>It has <b><c>bold</b> <c>fonts.</l></li>
<li><l>And <i><c>italic</i><c>, too.</l></li>
</ul>
<p><a href="mailto:mail@example.org"><o>contact</a></p>
</body>
</html>
```

Příklad

Krok 6: bloky <h>, <p>, <l>

```
<html>
<head>
<title><h>Sample Web Page</title>
<style> body color: green; </style>
</head>
<body>
<h1><h>Hello World!</h1>
<p>This is a simple webpage made of a paragraph and a list.</p>
<ul>
<li><1>It has <b><c>bold</b> <c>fonts.</c></li>
<li><1>And <i><c>italic</c></i><c>, too.</c></li>
</ul>
<p><a href="mailto:mail@example.org"><o>contact</o></a></p>
</body>
</html>
```

Příklad

Krok 7: bloky <c>

```
<html>
<head>
<title><h>Sample Web Page</title>
<style> body color: green; </style>
</head>
<body>
<h1><h>Hello World!</h></h1>
<p><p>This is a simple webpage made of a paragraph and a list.</p>
<ul>
<li><l>It has <b> bold </b> fonts.</li>
<li><l>And <i> italic, too.</i></li>
</ul>
<p><a href="mailto:mail@example.org"><o>contact</o></a></p>
</body>
</html>
```

Příklad

Krok 8: bloky <o>

```
<html>
<head>
<title><h>Sample Web Page</title>
<style> body color: green; </style>
</head>
<body>
<h1><h>Hello World!</h></h1>
<p><p>This is a simple webpage made of a paragraph and a list.</p>
<ul>
<li><l>It has <b> bold </b> fonts.</l></li>
<li><l>And <i> italic, too.</i></l></li>
</ul>
<p><a href="mailto:mail@example.org">contact</a></p>
</body>
</html>
```

Příklad

Krok 9: Hotovo

```
<h>Sample Web Page
```

```
<h>Hello World!
```

```
<p>This is a simple webpage made of a paragraph and a list.
```

```
<l>It has bold fonts.
```

```
<l>And italic, too.
```

Příklad

Krok 9: Hotovo

```
<h>Sample Web Page
<h>Hello World!
<p>This is a simple webpage made of a paragraph and a list.
<l>It has bold fonts.
<l>And italic, too.
```

Typy rysů a jejich příklady

“Content-based”

- ▶ délka textu v bloku – počty znaků, slov, vět, odstavců
- ▶ rozdělení výskytu vybraných znaků (alfabetické, numerické, zvláštní ...)
- ▶ rozdělení výskytu vybraných řetězců (slova, neslova, URL, čísla ...)

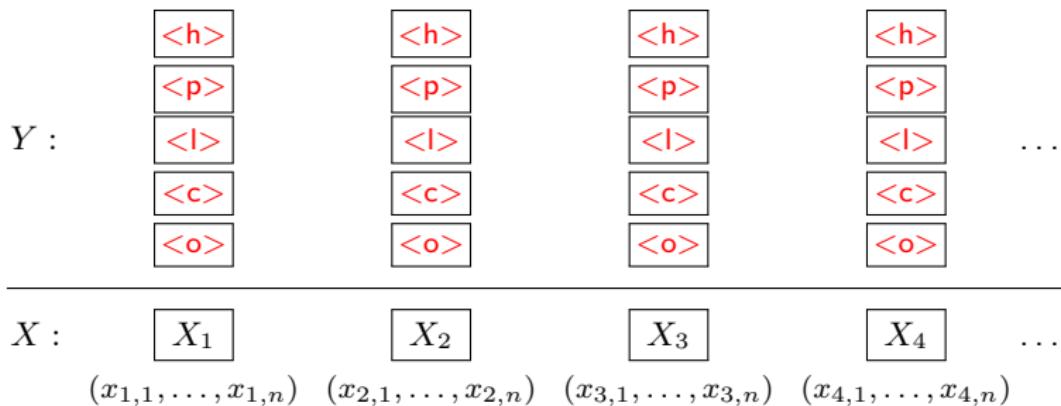
“Markup-based”

- ▶ typy a počty HTML značek, do kterých je blok uzavřen
- ▶ typy a počty HTML značek, které oddělují blok od *předchozího* bloku
- ▶ typy a počty HTML značek, které oddělují blok od *následujícího* bloku

“Document-based”

- ▶ relativní a absolutní poloha bloku v dokumentu (povrchová, v HTML struktuře)
- ▶ počet výskytu stejných bloků v dokumentu
- ▶ délka dokumentu – počty znaků, slov, vět, odstavců
- ▶ parametry rozdělení délek bloků dokumentu

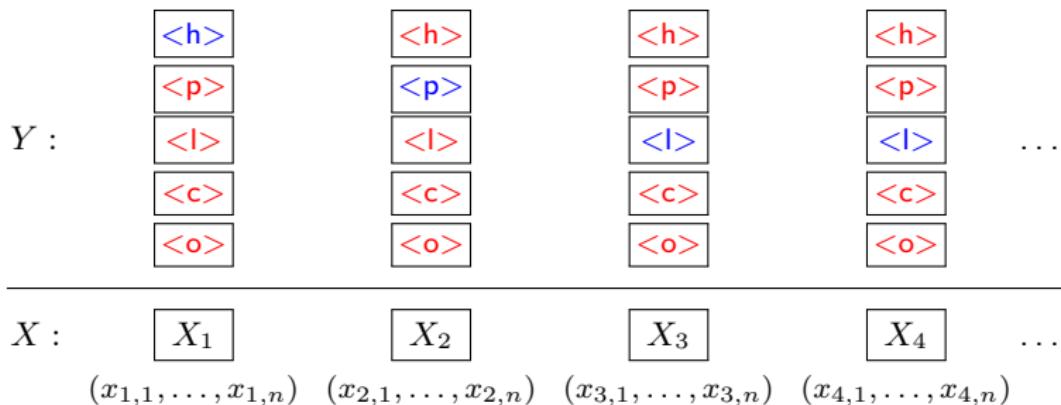
Značkování sekvencí (*sequence labeling*)



X : sekvence textových bloků X_1, \dots, X_N

Y : sekvence značek Y_1, \dots, Y_N

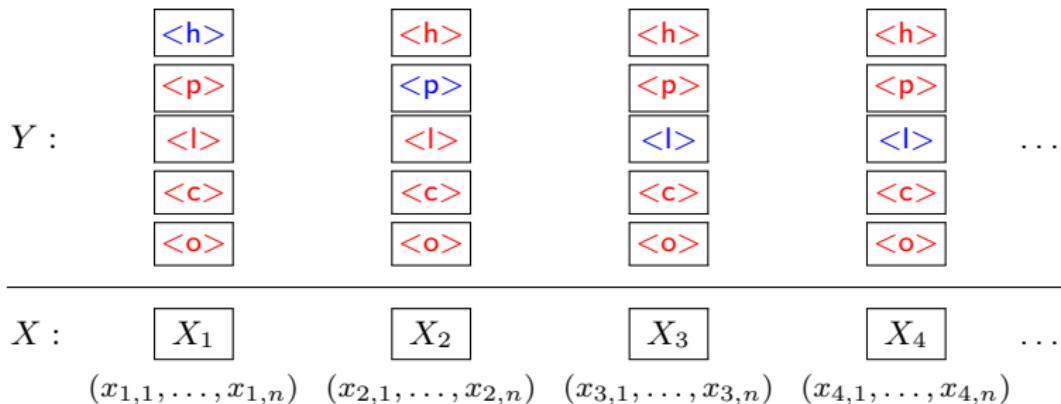
Značkování sekvencí (*sequence labeling*)



X : sekvence textových bloků X_1, \dots, X_N

Y : sekvence značek Y_1, \dots, Y_N

Značkování sekvencí (*sequence labeling*)



X : sekvence textových bloků X_1, \dots, X_N

Y : sekvence značek Y_1, \dots, Y_N

- ▶ Klasifikace: $p(Y_i|X_i)$
- ▶ Hidden Markov Models: $p(Y_i|X_i, X_{i-1})$
- ▶ Conditional Random Fields: $p(Y_i|X)$

Experimenty

Data

- ▶ 104 náhodně vybraných anglických www stránek segmentováno na textové bloky a anotováno
- ▶ o každém bloku bylo rozhodnuto, zda má být odstraněn, případně jakou značku mu přiřadit
- ▶ soubory rozděleny do 6 stejně velkých podmnožin

label	count
header	1 996
paragraph	3 419
list item	1 149
continuation	3 380
total (<i>content</i>)	9 944
other (<i>noise</i>)	12 557

Výsledky

- ▶ pomocí krosvalidace odhadnut podíl úspěšně označených bloků textu
- ▶ trénováno vždy na 5 dílech, testováno na 1, výsledky zprůměrovány

	full label set	content-noise
Exp-1	74.45	82.60
Exp-2	75.09	83.09
Exp-3	75.01	82.88

Cleaneval 2007

Soutěž v čištění webových stránek

"A shared task and competitive evaluation on the topic of cleaning arbitrary web pages, with the goal of preparing web data for use as a corpus, for linguistic and language technology research and development."

Organizátoři

Marco Baroni (*University of Trento*), Serge Sharoff (*Leeds University*)
Francis Chantree, Adam Kilgarriff (*Lexical Computing Ltd*)

Průběh

- ▶ zveřejněna data pro vývojové účely včetně **evaluačního skriptu** (*březen*)
- ▶ evaluační data připravena pro účastníky, výsledky (vyčištěné dokumenty) musí být odeslány zpět **do 24 hodin** po jejich přijetí (*červen*)
- ▶ vyhodnocování výsledků organizátory (*červenec*)
- ▶ příprava příspěvků na workshop (*srpen*)
- ▶ workshop, prezentace příspěvků, vyhlášení vítězů a analýza výsledků (*září*)

Cleaneval 2007: detaily

Příprava dat

- ▶ náhodný výběr z již existujícího anglického webového korpusu
- ▶ manuální čištění provádělo 23 anotátorů
- ▶ 57 stránek, včetně vyčištěné podoby určeno pro vývoj a ladění metod
- ▶ 653 stránek pro testování (vyčištěnou podobu soutěžící neviděli)

Evaluacní skript

- ▶ vstup:- *soubor vyčištěný automaticky (od soutěžícího týmu)*
 - *soubor vyčištěný ručně (anotovaný)*
- ▶ normalizace:
 - *převod na malá písmena*
 - *odstranění interpunkce*
 - *vertikalizace (jedno slovo na řádek)*
- ▶ evaluacní míra:
 - *založená na editační vzdálenosti (Levenshtein)*
 - *substituce se penalizuje dvojnásobně*
- ▶ dva režimy:
 - *značky se berou v úvahu (Markup+Text)*
 - *značky se ignorují (Text-only)*

Účastníci

- ▶ 9 týmů
- ▶ 7 zemí
- ▶ 4 světadíly

Team

Charles University, *Czech Republic*

Elhuyar Foundation, *Spain*

University of Amsterdam, *Netherlands*

IRST Center, *Italy*

GenieNows *Canada*

Osnabrück University, *Germany*

Chaudhury University, *India*

Osnabrück University, *Germany*

North West University, *South Africa*

Účastníci a výsledky

- ▶ 9 týmů
- ▶ 7 zemí
- ▶ 4 světadíly

Team	Text+Markup	Text-Only	Average
Charles University, Czech Republic	65.3	84.1	74.7
Elhuyar Foundation, Spain	65.3	83.4	74.3
University of Amsterdam, Netherlands	65.5	83.0	74.2
IRST Center, Italy	65.6	82.5	74.0
GenieNow Canada	63.9	83.4	73.6
Osnabrück University, Germany	60.3	82.9	71.6
Chaudhury University, India	59.5	80.9	70.2
Osnabrück University, Germany	53.5	73.5	63.5
North West University, South Africa	45.5	60.2	52.9

Další kroky

- ▶ vylepšit systém použitím jiných/lepších rysů
- ▶ natrénovat systém pro čištění stránek v češtine
- ▶ využít systém pro vytvoření velkého českého webového korpusu

Otázky a odpovědi

Děkuji za pozornost.