Bibliography

[1] AIZERMAN, A.; BRAVERMAN, E. ; ROZONER, L. **Automation and remote control**. Theoretical foundations of the potential function method in pattern recognition learning, journal, v.25, p. 821–837, 1964.

[2] BOSER, B.; GUYON, I. ; VAPNIK, V. **A training algorithm for optimal margin classifiers**. In: PROCEEDINGS OF THE FIFTH ANNUAL WORKSHOP ON COMPUTATIONAL LEARNING THEORY, p. 144–152. ACM, 1992.

[3] **The C# Language**. http://msdn.microsoft.com/en-us/vcsharp/aa336809. [Online; accessed 01-August-2011].

[4] ÇELIK, T.; BOS, B.; HICKSON, I. ; LIE, H. W. **Cascading style sheets level 2 revision 1 (CSS 2.1) specification**. Candidate recommendation, W3C, Sept. 2009. http://www.w3.org/TR/2009/CR-CSS2-20090908.

[5] **Cleaneval home page**. http://cleaneval.sigwac.org.uk. [Online; accessed 18-January-2011].

[6] CORTES, C.; VAPNIK, V. **Machine Learning**. Support-vector networks, journal, v.20, p. 273–297, 1995. 10.1007/BF00994018.

[7] **CSS current work & how to participate**. http://www.w3.org/Style/CSS/current-work. [Online; accessed 01-August-2011].

[8] **DOM CSS Properties – MDC Doc Center**. https://developer.mozilla.org/en/DOM/CSS. [Online; accessed 13-April-2011].

[9] **Firebug**. http://getfirebug.com/. [Online; accessed 21-July-2011].

[10] **Mozilla firefox web browser**. http://getfirefox.com/. [Online; accessed 21-July-2011].

[11] **Gecko**. https://developer.mozilla.org/en/Gecko. [Online; accessed 01-September-2010].

[12] HALL, M.; FRANK, E.; HOLMES, G.; PFAHRINGER, B.; REUTEMANN, P. ; WITTEN, I. **ACM SIGKDD Explorations Newsletter**. The weka data mining software: an update, journal, v.11, n.1, p. 10–18, 2009.

[13] HICKSON, I. **HTML 5**. W3C working draft, W3C, May 2011. http://www.w3.org/TR/2011/WD-html5-20110525/.

[14] HU, Y.; XIN, G.; SONG, R.; HU, G.; SHI, S.; CAO, Y. ; LI, H. **Title extraction from bodies of html documents and its application to web page retrieval**. In: SIGIR '05: PROCEEDINGS OF THE 28TH ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, p. 250–257, New York, NY, USA, 2005. ACM.

[15] JACOBS, I.; RAGGETT, D. ; HORS, A. L. **HTML 4.01 specification**. W3C recommendation, W3C, Dec. 1999. http://www.w3.org/TR/1999/REC-html401-19991224.

[16] **java.com: Java + you**. http://www.java.com/en/. [Online; accessed 01-August-2011].

[17] **JPype – Java to Python integration**. http://jpype.sourceforge.net. [Online; accessed 01-August-2011].

[18] LABER, E. S.; DE SOUZA, C. P.; JABOUR, I. V.; DE AMORIM, E. C. F.; CARDOSO, E. T.; RENTERÍA, R. P.; TINOCO, L. C. ; VALENTIM, C. D. **A fast and simple method for extracting relevant content from news webpages**. In: Cheung, D. W.-L.; Song, I.-Y.; Chu, W. W.; Hu, X. ; Lin, J. J., editors, CIKM, p. 1685–1688. ACM, 2009.

[19] LEVENSHTEIN, V. I. **Doklady Akademii Nauk SSSR**. Binary codes with correction of deletions, insertions and substitution of symbols, journal, v.163, n.4, p. 845–848, 1965.

[20] **The XML C parser and toolkit of Gnome**. http://xmlsoft.org/. [Online; accessed 01-August-2011].

[21] **libxml2dom**. http://www.boddie.org.uk/python/libxml2dom.html. [Online; accessed 01-August-2011].

[22] LUO, P.; FAN, J.; LIU, S.; LIN, F.; XIONG, Y. ; LIU, J. **Web article extraction for web printing: a dom+visual based approach**.

In: PROCEEDINGS OF THE 9TH ACM SYMPOSIUM ON DOCUMENT ENGINEERING, p. 66–69. ACM, 2009.

[23] MAREK, M.; PECINA, P. ; SPOUSTA, M. **Cahiers du Cental**. Web page cleaning with conditional random fields, journal, v.5, p. 1, 2007.

[24] MCCARRON, S.; ISHIKAWA, M. **XHTML™ 1.1 - module-based XHTML - second edition**. W3C recommendation, W3C, Nov. 2010. http://www.w3.org/TR/2010/REC-xhtml11-20101123.

[25] **MSHTML**. http://msdn.microsoft.com/en-us/library/bb508516(v=VS.85).aspx. [Online; accessed 01-September-2010].

[26] NICOL, G.; CHAMPION, M.; HÉGARET, P. L.; ROBIE, J.; WOOD, L.; HORS, A. L. ; BYRNE, S. **Document object model (DOM) level 3 core specification**. W3C recommendation, W3C, Apr. 2004. http://www.w3.org/TR/2004/REC-DOM-Level-3-Core-20040407.

[27] **Opera Presto 2.1 – Web standards supported by Opera's core – Dev.Opera**. http://dev.opera.com/articles/view/presto-2-1-web-standards-supported-by/. [Online; accessed 13-April-2011].

[28] PASTERNACK, J.; ROTH, D. **Extracting article text from the web with maximum subsequence segmentation**. In: PROCEEDINGS OF THE 18TH INTERNATIONAL CONFERENCE ON WORLD WIDE WEB, p. 971–980. ACM, 2009.

[29] **Python programming language – official website**. http://www.python.org. [Online; accessed 01-August-2011].

[30] **Python Webkit DOM Bindings**. http://www.gnu.org/s/pythonwebkit/. [Online; accessed 01-August-2011].

[31] ROBIE, J.; CHAMBERLIN, D. ; SNELSON, M. D. J. **XML path language (XPath) 3.0**. W3C working draft, W3C, June 2011. http://www.w3.org/TR/2011/WD-xpath-30-20110614/.

[32] RUZZO, W.; TOMPA, M. **A linear time algorithm for finding all maximal scoring subsequences**. In: PROCEEDINGS OF THE SEVENTH INTERNATIONAL CONFERENCE ON INTELLIGENT SYSTEMS FOR MOLECULAR BIOLOGY, p. 234–241. AAAI Press, 1999.

[33] SPENGLER, A.; BORDES, A. ; GALLINARI, P. **A comparison of discriminative classifiers for web news content extraction**. In:

PROCEEDINGS OF RIAO 2010, 9TH INT. CONF. ON ADAPTIVITY, PERSONALIZATION AND FUSION OF HETEROGENEOUS INFORMATION, 2010.

[34] SPENGLER, A.; GALLINARI, P. **Learning to Extract Content from News Webpages**. In: PROCEEDINGS OF THE 2009 INTERNATIONAL CONFERENCE ON ADVANCED INFORMATION NETWORKING AND APPLICATIONS WORKSHOPS, p. 709–714. IEEE Computer Society, 2009.

[35] SPENGLER, A.; GALLINARI, P. **Document structure meets page layout: loopy random fields for web news content extraction**. In: PROCEEDINGS OF THE 10TH ACM SYMPOSIUM ON DOCUMENT ENGINEERING, p. 151–160. ACM, 2010.

[36] TAN, P.-N.; STEINBACH, M. ; KUMAR, V. **Introduction to Data Mining**. Addison-Wesley, 2005.

[37] **The WebKit Open Source Project – CSS (Cascading Style Sheets)**. http://www.webkit.org/projects/css/index.html. [Online; accessed 13-April-2011].

[38] VAN KESTEREN, A. **HTML 5 differences from HTML 4**. W3C working draft, W3C, Aug. 2009. http://www.w3.org/TR/2009/WD-html5-diff-20090825/.

[39] VIEIRA, K.; DA SILVA, A. S.; PINTO, N.; DE MOURA, E. S.; CAVALCANTI, J. M. B. ; FREIRE, J. **A fast and robust method for web page template detection and removal**. In: CIKM '06: PROCEEDINGS OF THE 15TH ACM INTERNATIONAL CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT, p. 258–267, New York, NY, USA, 2006. ACM.

[40] WANG, J.; CHEN, C.; WANG, C.; PEI, J.; BU, J.; GUAN, Z. ; ZHANG, W. V. **Can we learn a template-independent wrapper for news article extraction from a single training site?** In: PROCEEDINGS OF THE 15TH ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, p. 1345–1354. ACM, 2009.

[41] WANG, J.; HE, X.; WANG, C.; PEI, J.; BU, J.; CHEN, C.; GUAN, Z. ; LU, G. **News article extraction with template-independent wrapper**. In: PROCEEDINGS OF THE 18TH INTERNATIONAL CONFERENCE ON WORLD WIDE WEB, p. 1085–1086. ACM, 2009.

[42] **WebKit**. http://webkit.org/. [Online; accessed 01-September-2010].

[43] **Using Weka from Python**. http://kogs-www.informatik.uni-hamburg.de/ meine/weka-python/. [Online; accessed 01-August-2011].

[44] WIKIPEDIA. **Comparison of layout engines (cascading style sheets) — Wikipedia, the free encyclopedia**, 2010. [Online; accessed 22-September-2010].

[45] **World Wide Web Consortium (W3C)**. http://www.w3c.org/. [Online; accessed 14-September-2010].

[46] XUE, Y.; HU, Y.; XIN, G.; SONG, R.; SHI, S.; CAO, Y.; LIN, C.-Y. ; LI, H. **Information Processing and Management**. Web page title extraction and its application, journal, v.43, n.5, p. 1332–1347, 2007.

A
List of sites in the RCD4 corpus

Below is a list of domains in the RCD4 corpus, in alphabetical order:

- http://abclocal.go.com: 1 page
- http://abcnews.go.com: 6 pages
- http://ausiellofiles.ew.com: 2 pages
- http://blogs.wsj.com: 1 page
- http://g1.globo.com: 7 pages
- http://globoesporte.globo.com: 1 page
- http://hollywoodinsider.ew.com: 1 page
- http://jbonline.terra.com.br: 8 pages
- http://money.cnn.com: 2 pages
- http://music-mix.ew.com: 1 page
- http://news.bbc.co.uk: 4 pages
- http://noticias.uol.com.br: 1 page
- http://oglobo.globo.com: 8 pages
- http://online.wsj.com: 7 pages
- http://politicalticker.blogs.cnn.com: 1 page
- http://popwatch.ew.com: 1 page
- http://vestibular.uol.com.br: 1 page
- http://washington.bizjournals.com: 5 pages
- http://www.20minutos.es: 5 pages
- http://www.abc-7.com: 1 page
- http://www.aolnews.com: 1 page
- http://www.bbc.co.uk: 1 page
- http://www.boston.com: 4 pages
- http://www.cbc.ca: 5 pages
- http://www.cnn.com: 3 pages

- http://www.consortiumnews.com: 5 pages
- http://www.economist.com: 4 pages
- http://www.elmundo.es: 5 pages
- http://www.estadao.com.br: 8 pages
- http://www.foxnews.com: 4 pages
- http://www.guardian.co.uk: 5 pages
- http://www.health.com: 1 page
- http://www.ireport.com: 1 page
- http://www.marca.com: 5 pages
- http://www.mcclatchydc.com: 5 pages
- http://www.msnbc.msn.com: 5 pages
- http://www.noticias.com: 5 pages
- http://www.nysun.com: 5 pages
- http://www.nytimes.com: 10 pages
- http://www.pcworld.com: 4 pages
- http://www.politicsdaily.com: 1 page
- http://www.reuters.com: 8 pages
- http://www.sciencedaily.com: 5 pages
- http://www.slate.com: 5 pages
- http://www.sphere.com: 4 pages
- http://www.time.com: 7 pages
- http://www.usatoday.com: 5 pages
- http://www.valoronline.com.br: 4 pages
- http://www.washingtonpost.com: 1 page
- http://www1.folha.uol.com.br: 10 pages