References

[1] R. Kosala and H. Blockeel, “Web mining research: a survey,” *ACM SIGKDD Explorations Newsletter*, vol. 2, no. 1, pp. 1–15, Jun. 2000. doi: 10.1145/360402.360406. [Online]. Available: https://dl.acm.org/doi/10.1145/360402.360406 [Page 1.]

[2] J. Wang and F. H. Lochovsky, “Data extraction and label assignment for web databases,” in *Proceedings of the twelfth international conference on World Wide Web - WWW ’03*. Budapest, Hungary: ACM Press, 2003. doi: 10.1145/775152.775179. ISBN 978-1-58113-680-7 p. 187. [Online]. Available: http://portal.acm.org/citation.cfm?doid=775152.775179 [Page 1.]

[3] A. Troestler and H. P. Lee, “The adaptation and standardization on websites of international companies : Analysis and comparison from websites of United States, Germany and Taiwan,” Ph.D. dissertation, Linköping University, 2007. [Online]. Available: http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-9801 [Page 1.]

[4] S. Flesca, G. Manco, E. Masciari, E. Rende, and A. Tagarelli, “Web wrapper induction: a brief survey,” *AI Communications*, vol. 17, no. 2, pp. 57–61, Apr. 2004. [Online]. Available: https://dl.acm.org/doi/10.5555/1218702.1218707 [Page 2.]

[5] “Document Object Model (DOM),” Dec. 2021. [Online]. Available: https://developer.mozilla.org/en-US/docs/Web/API/Document_Object_Model [Pages 2 and 5.]

[6] K. Lerman, S. N. Minton, and C. A. Knoblock, “Wrapper Maintenance: A Machine Learning Approach,” *Journal of Artificial Intelligence Research*, vol. 18, pp. 149–181, Feb. 2003. doi: 10.1613/jair.1145.

[Online]. Available: https://jair.org/index.php/jair/article/view/10325 [Page 2.]

[7] Hevner, March, Park, and Ram, “Design Science in Information Systems Research,” *MIS Quarterly*, vol. 28, no. 1, p. 75, 2004. doi: 10.2307/25148625. [Online]. Available: https://www.jstor.org/stable/10.2307/25148625 [Page 3.]

[8] R. Baumgartner, W. Gatterbauer, and G. Gottlob, “Web Data Extraction System,” in *Encyclopedia of Database Systems*, L. Liu and M. T. Özsu, Eds. Boston, MA: Springer US, 2009, pp. 3465–3471. ISBN 978-0-387-39940-9. [Online]. Available: http://link.springer.com/10.1007/978-0-387-39940-9_1154 [Page 5.]

[9] “XPath,” Jan. 2022. [Online]. Available: https://developer.mozilla.org/en-US/docs/Web/XPath [Page 6.]

[10] “Introducing JSON.” [Online]. Available: https://json.org/json-en.html [Page 6.]

[11] “What is Deep Learning?” May 2020. [Online]. Available: https://www.ibm.com/cloud/learn/deep-learning [Page 7.]

[12] J. Patterson and A. Gibson, *Deep learning: a practitioner’s approach*, 1st ed. Sebastopol, CA: O’Reilly, 2017. ISBN 978-1-4919-1425-0 [Pages 7, 8, and 9.]

[13] D. Hendrycks and K. Gimpel, “Gaussian Error Linear Units (GELUs),” *arXiv:1606.08415 [cs]*, Jul. 2020. [Online]. Available: http://arxiv.org/abs/1606.08415 [Page 9.]

[14] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997. doi: 10.1162/neco.1997.9.8.1735. [Online]. Available: https://direct.mit.edu/neco/article/9/8/1735-1780/6109 [Page 9.]

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, . Kaiser, and I. Polosukhin, “Attention is All you Need,” in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf [Page 10.]

[16] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, and X. Huang, “Pre-trained models for natural language processing: A survey,” *Science China Technological Sciences*, vol. 63, no. 10, pp. 1872–1897, Oct. 2020. doi: 10.1007/s11431-020-1647-3. [Online]. Available: https://link.springer.com/10.1007/s11431-020-1647-3 [Page 10.]

[17] K. Weiss, T. M. Khoshgoftaar, and D. Wang, “A survey of transfer learning,” *Journal of Big Data*, vol. 3, no. 1, p. 9, Dec. 2016. doi: 10.1186/s40537-016-0043-6. [Online]. Available: http://journalofbigdata.springeropen.com/articles/10.1186/s40537-016-0043-6 [Page 10.]

[18] X. Han, Z. Zhang, N. Ding, Y. Gu, X. Liu, Y. Huo, J. Qiu, Y. Yao, A. Zhang, L. Zhang, W. Han, M. Huang, Q. Jin, Y. Lan, Y. Liu, Z. Liu, Z. Lu, X. Qiu, R. Song, J. Tang, J.-R. Wen, J. Yuan, W. X. Zhao, and J. Zhu, “Pre-trained models: Past, present and future,” *AI Open*, vol. 2, pp. 225–250, 2021. doi: 10.1016/j.aiopen.2021.08.002. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2666651021000231 [Page 10.]

[19] G. Bonaccorso, *Machine Learning Algorithms*, 2nd ed. Packt, 2018. ISBN 978-1-78934-799-9 [Page 11.]

[20] M. Buckland and F. Gey, “The relationship between Recall and Precision,” *Journal of the American Society for Information Science*, vol. 45, no. 1, pp. 12–19, 1994. [Page 11.]

[21] E. Ferrara, P. De Meo, G. Fiumara, and R. Baumgartner, “Web data extraction, applications and techniques: A survey,” *Knowledge-Based Systems*, vol. 70, pp. 301–323, Nov. 2014. doi: 10.1016/j.knosys.2014.07.007. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0950705114002640 [Page 12.]

[22] S. M. Selkow, “The tree-to-tree editing problem,” *Information Processing Letters*, vol. 6, no. 6, pp. 184–186, Dec. 1977. doi: 10.1016/0020-0190(77)90064-3. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/0020019077900643 [Page 12.]

[23] P. Kilpeläinen, “Tree matching problems with applications to structured text databases,” Ph.D dissertation, University of Helsinki, Department of Computer Science, Helsinki, Finland, Nov. 1992. [Page 12.]

[24] N. Kushmerick, D. S. Weld, and R. B. Doorenbos, “Wrapper Induction for Information Extraction,” in *IJCAI*, 1997. [Page 12.]

[25] R. Mooney, "Relational learning of pattern-match rules for information extraction," in *Proceedings of the sixteenth national conference on artificial intelligence*, vol. 328, 1999, p. 334. [Page 12.]

[26] S. Soderland, "Learning information extraction rules for semi-structured and free text," *Machine learning*, vol. 34, no. 1, pp. 233–272, 1999, publisher: Springer. [Page 12.]

[27] I. Muslea and others, "Extraction patterns for information extraction tasks: A survey," in *The AAAI-99 workshop on machine learning for information extraction*, vol. 2. Orlando Florida, 1999, issue: 2. [Page 13.]

[28] S. Zhang and K. Balog, "Web Table Extraction, Retrieval, and Augmentation: A Survey," *ACM Transactions on Intelligent Systems and Technology*, vol. 11, no. 2, pp. 1–35, Apr. 2020. doi: 10.1145/3372117. [Online]. Available: https://dl.acm.org/doi/10.1145/3372117 [Page 13.]

[29] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv:1810.04805 [cs]*, May 2019. [Online]. Available: http://arxiv.org/abs/1810.04805 [Pages 15, 17, and 36.]

[30] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 19–27. [Page 16.]

[31] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding," *arXiv:1804.07461 [cs]*, Feb. 2019. [Online]. Available: http://arxiv.org/abs/1804.07461 [Page 16.]

[32] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ Questions for Machine Comprehension of Text," *arXiv:1606.05250 [cs]*, Oct. 2016. [Online]. Available: http://arxiv.org/abs/1606.05250 [Page 16.]

[33] R. Zellers, Y. Bisk, R. Schwartz, and Y. Choi, "SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference," *arXiv:1808.05326 [cs]*, Aug. 2018. [Online]. Available: http://arxiv.org/abs/1808.05326 [Page 16.]

[34] Y. Zhou, Y. Sheng, N. Vo, N. Edmonds, and S. Tata, "Simplified DOM Trees for Transferable Attribute Extraction from the Web," *arXiv:2101.02415 [cs]*, Jan. 2021. [Online]. Available: http://arxiv.org/abs/2101.02415 [Pages 16, 18, 19, 22, and 41.]

[35] Q. Hao, R. Cai, Y. Pang, and L. Zhang, "From one tree to a forest: a unified solution for structured web data extraction," in *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information - SIGIR '11*. Beijing, China: ACM Press, 2011. doi: 10.1145/2009916.2010020. ISBN 978-1-4503-0757-4 p. 775. [Online]. Available: http://portal.acm.org/citation.cfm?doid=2009916.2010020 [Pages 17, 18, 22, and 40.]

[36] J. Pennington, R. Socher, and C. Manning, "Glove: Global Vectors for Word Representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, 2014. doi: 10.3115/v1/D14-1162 pp. 1532–1543. [Online]. Available: http://aclweb.org/anthology/D14-1162 [Page 17.]

[37] J. Li, Y. Xu, L. Cui, and F. Wei, "MarkupLM: Pre-training of Text and Markup Language for Visually-rich Document Understanding," *arXiv:2110.08518 [cs]*, Oct. 2021. [Online]. Available: http://arxiv.org/abs/2110.08518 [Pages 17, 22, 27, and 29.]

[38] B. Y. Lin, Y. Sheng, N. Vo, and S. Tata, "FreeDOM: A Transferable Neural Architecture for Structured Information Extraction on Web Documents," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. Virtual Event CA USA: ACM, Aug. 2020. doi: 10.1145/3394486.3403153. ISBN 978-1-4503-7998-4 pp. 1092–1102. [Online]. Available: https://dl.acm.org/doi/10.1145/3394486.3403153 [Page 18.]

[39] X. Deng, P. Shiralkar, C. Lockard, B. Huang, and H. Sun, "DOM-LM: Learning Generalizable Representations for HTML Documents," *arXiv:2201.10608 [cs]*, Jan. 2022. [Online]. Available: http://arxiv.org/abs/2201.10608 [Page 18.]

[40] K. Peffers, T. Tuunanen, C. E. Gengler, M. Rossi, W. Hui, V. Virtanen, and J. Bragge, "Design Science Research Process: A Model for Producing and Presenting Information Systems

Research," *arXiv:2006.02763 [cs]*, Jun. 2020. [Online]. Available: http://arxiv.org/abs/2006.02763 [Page 21.]

[41] "AWS Deep Learning AMIs." [Online]. Available: https://aws.amazon.com/machine-learning/amis/ [Page 23.]

[42] "Amazon EC2 G4 Instances." [Online]. Available: https://aws.amazon.com/ec2/instance-types/g4/ [Page 23.]

[43] "Pytorch." [Online]. Available: https://pytorch.org/ [Page 27.]

[44] "Transformers." [Online]. Available: https://huggingface.co/transformers [Page 27.]

[45] "pickle — Python object serialization," Apr. 2022. [Online]. Available: https://docs.python.org/3/library/pickle.html [Page 27.]

[46] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *arXiv:1907.11692 [cs]*, Jul. 2019. [Online]. Available: http://arxiv.org/abs/1907.11692 [Page 29.]

[47] M. Malmsten, L. Börjeson, and C. Haffenden, "Playing with Words at the National Library of Sweden – Making a Swedish BERT," *arXiv:2007.01658 [cs]*, Jul. 2020. [Online]. Available: http://arxiv.org/abs/2007.01658 [Page 29.]

[48] "The AI community building the future." [Online]. Available: https://huggingface.co/ [Page 29.]

[49] "Cloud TPU." [Online]. Available: https://cloud.google.com/tpu/ [Page 36.]

[50] D. M. W. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," *arXiv:2010.16061 [cs, stat]*, Oct. 2020. [Online]. Available: http://arxiv.org/abs/2010.16061 [Page 36.]

TRITA-EECS-EX- 2022:192

For DIVA

```
{
”Author1”: { ”Last name”: ”Hodzic”,
”First name”: ”Amar”,
”Local User Id”: ”u15805os”,
”E-mail”: ”amarh@kth.se”,
”organisation”: {”L1”: ”School of Electrical Engineering and Computer Science”,
}
},
”Cycle”: ”2”,
”Course code”: ”DA231X”,
”Credits”: ”30”,
”Degree1”: {”Educational program”: ”Master’s Programme, Computer Science, 120 credits”
,”programcode”: ”TCSCM”
,”Degree”: ”Degree of Master (120 credits)”
,”subjectArea”: ”Computer Science and Engineering”
},
”Title”: {
”Main title”: ”Automated Extraction of Data from Insurance Websites”,
”Language”: ”eng” },
”Alternative title”: {
”Main title”: ”Automatiserad Datautvinning från Försäkringssidor”,
”Language”: ”swe”
},
”Supervisor1”: { ”Last name”: ”Dwivedi”,
”First name”: ”Ashish Kumar”,
”Local User Id”: ”u13recon”,
”E-mail”: ”dwvedi@kth.se”,
”organisation”: {”L1”: ”School of Electrical Engineering and Computer Science”,
”L2”: ”Computer Science” }
},
”Supervisor2”: { ”Last name”: ”Bränn”,
”First name”: ”Jesper”,
”E-mail”: ”jesper@insurely.com”,
”Other organisation”: ”Insurely”
},
”Supervisor3”: { ”Last name”: ”Sevelius”,
”First name”: ”Eric”,
”E-mail”: ”eric@insurely.se”,
”Other organisation”: ”Insurely”
},
”Examiner1”: { ”Last name”: ”Maguire Jr”,
”First name”: ”Gerald Quentin”,
”Local User Id”: ”u1d13i2c”,
”E-mail”: ”maguire@kth.se”,
”organisation”: {”L1”: ”School of Electrical Engineering and Computer Science”,
”L2”: ”Computer Science” }
},
”Cooperation”: { ”Partner_name”: ”Insurely”},
”National Subject Categories”: ”10201”,
”Other information”: {”Year”: ”2022”, ”Number of pages”: ”xiv,48”},
”Series”: { ”Title of series”: ”TRITA-EECS-EX” , ”No. in series”: ”2022:00” },
”Opponents”: { ”Name”: ”Natan Teferi Asegehegn”},
”Presentation”: { ”Date”: ”2022-05-24 13:00”

,”Language”:”eng”

,”Room”: ”via Zoom https://kth-se.zoom.us/j/62953003243”

,”Address”: ”Isafjordsgatan 22 (Kistagången 16)”

,”City”: ”Stockholm” },

”Number of lang instances”: ”2”,
”Abstract[eng]”: €€€€
Websites have become a critical source of information for many organizations in today's digital era.
However, extracting and organizing semi-structured data from web pages from multiple websites poses
challenges. This is especially true when a high level of automation is desired while maintaining
generality.

A natural progression in the quest for automation is to extend the methods for web data extraction
from only being able to handle a single website to handling multiple ones, usually within the same
domain. Although these websites share the same domain, the structure of the data can vary greatly. A
key question becomes how generalized such a system can be to encompass a large number of websites
while maintaining adequate accuracy.

The thesis examined the efficiency of automated web data extraction on multiple Swedish insurance
company websites. Previous work showed that good results can be achieved with a known English data
```