Replicability Challenge co-located with 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, OSIRRC@SIGIR 2019, Paris, France, July 25, 2019. 50–56. http://ceur-ws.org/Vol-2409/docker08.pdf

[62] Jiaxin Mao, Tetsuya Sakai, Cheng Luo, Peng Xiao, Yiqun Liu, and Zhicheng Dou. 2018. Overview of the NTCIR-14 We Want Web Task. In *NTCIR*.

[63] Iadh Ounis, Gianni Amati, Vassilis Plachouras, Ben He, Craig Macdonald, and Douglas Johnson. 2005. Terrier information retrieval platform. In *ECIR*. Springer, 517–519.

[64] Joao Palotti, Guido Zuccon, Jimmy, Pavel Pecina, Mihai Lupu, Lorraine Goeuriot, Liadh Kelly, and Allan Hanbury. 2017. CLEF 2017 Task Overview: The IR Task at the eHealth Evaluation Lab - Evaluating Retrieval Methods for Consumer Health Search. In *CLEF*.

[65] The pandas development team. 2020. *pandas-dev/pandas: Pandas*. https://doi.org/10.5281/zenodo.3509134

[66] Nick Craswell Li Deng Jianfeng Gao Xiaodong Liu Rangan Majumder Andrew McNamara Bhaskar Mitra Tri Nguyen Mir Rosenberg Xia Song Alina Stoica Saurabh Tiwary Tong Wang Payal Bajaj, Daniel Campos. 2016. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. In *InCoCo@NIPS*.

[67] Navid Rekabsaz, Oleg Lesota, Markus Schedl, Jon Brassey, and Carsten Eickhoff. 2021. TripClick: The Log Files of a Large Health Web Search Engine. In *SIGIR*.

[68] Kirk Roberts, Dina Demner-Fushman, Ellen Voorhees, William R. Hersh, Steven Bedrick, and Alexander J. Lazar. 2018. Overview of the TREC 2018 Precision Medicine Track. In *TREC*.

[69] Kirk Roberts, Dina Demner-Fushman, Ellen Voorhees, William R. Hersh, Steven Bedrick, Alexander J. Lazar, and Shubham Pant. 2017. Overview of the TREC 2017 Precision Medicine Track. In *TREC*.

[70] Kirk Roberts, Dina Demner-Fushman, Ellen Voorhees, William R. Hersh, Steven Bedrick, Alexander J. Lazar, Shubham Pant, and Funda Meric-Bernstam. 2019. Overview of the TREC 2019 Precision Medicine Track. In *TREC*.

[71] Kirk Roberts, Dina Demner-Fushman, Ellen M. Voorhees, and William R. Hersh. 2016. Overview of the TREC 2016 Clinical Decision Support Track. In *TREC*.

[72] Kirk Roberts, Matthew S. Simpson, Ellen Voorhees, and William R. Hersh. 2015. Overview of the TREC 2015 Clinical Decision Support Track. In *TREC*.

[73] Willie Rogers. 2000. TREC Mandarin LDC2000T52. https://catalog.ldc.upenn.edu/LDC2000T52

[74] Willie Rogers. 2000. TREC Spanish LDC2000T51. https://catalog.ldc.upenn.edu/LDC2000T51

[75] Evan Sandhaus. 2008. The new york times annotated corpus. *Linguistic Data Consortium, Philadelphia* 6, 12 (2008), e26752.

[76] Royal Sequiera and Jimmy Lin. 2017. Finally, a Downloadable Test Collection of Tweets. In *SIGIR*.

[77] Matthew S. Simpson, Ellen M. Voorhees, and William Hersh. 2014. Overview of the TREC 2014 Clinical Decision Support Track. In *TREC*.

[78] Alan Smeaton and Ross Wilkinson. 1996. Spanish and Chinese Document Retrieval in TREC-5. In *TREC*.

[79] Ian Soboroff, Shudong Huang, and Donna Harman. 2018. TREC 2018 News Track Overview. In *TREC*.

[80] Ian Soboroff, Shudong Huang, and Donna Harman. 2019. TREC 2019 News Track Overview. In *TREC*.

[81] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models. *arXiv preprint arXiv:2104.08663* (4 2021).

[82] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 809–819
.

[83] Ellen Voorhees. 2004. Overview of the TREC 2004 Robust Retrieval Track. In *TREC*.

[84] E. Voorhees, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, W. Hersh, Kyle Lo, Kirk Roberts, I. Soboroff, and Lucy Lu Wang. 2020. TREC-COVID: Constructing a Pandemic Information Retrieval Test Collection. *ArXiv* abs/2005.04474 (2020).

[85] Ellen M. Voorhees. 2005. Overview of the TREC 2005 Robust Retrieval Track. In *TREC*.

[86] Henning Wachsmuth, Shahbaz Syed, and Benno Stein. 2018. Retrieval of the Best Counterargument without Prior Topic Knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Melbourne, Australia). Association for Computational Linguistics, 241–251. http://aclweb.org/anthology/P18-1023

[87] David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or Fiction: Verifying Scientific Claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 7534–7550. https://doi.org/10.18653/v1/2020.emnlp-main.609

[88] Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, K. Funk, Rodney Michael Kinney, Ziyang Liu, W. Merrill, P. Mooney, D. Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, B. Stilson, A. Wade, K. Wang, Christopher Wilhelm, Boya Xie, D. Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020. CORD-19: The Covid-19 Open Research Dataset. *ArXiv* (2020).

[89] Ross Wilkinson. 1997. Chinese Document Retrieval at TREC-6. In *TREC*.

[90] Thomas Wolf, Quentin Lhoest, Patrick von Platen, Yacine Jernite, Mariama Drame, Julien Plu, Julien Chaumond, Clement Delangue, Clara Ma, Abhishek Thakur, Suraj Patil, Joe Davison, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angie McMillan-Major, Simon Brandeis, Sylvain Gugger, François Lagunas, Lysandre Debut, Morgan Funtowicz, Anthony Moi, Sasha Rush, Philipp Schmidd, Pierric Cistac, Victor Muštar, Jeff Boudier, and Anna Tordjmann. 2020. Datasets. *GitHub. Note: https://github.com/huggingface/datasets* 1 (2020).

[91] Peilin Yang, Hui Fang, and Jimmy Lin. 2017. Anserini: Enabling the Use of Lucene for Information Retrieval Research. *SIGIR* (2017).

[92] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 2369–2380. https://doi.org/10.18653/v1/D18-1259

[93] Andrew Yates, Kevin Martin Jose, Xinyu Zhang, and Jimmy Lin. 2020. Flexible IR pipelines with Capreolus. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 3181–3188.

[94] Guido Zuccon, Joao Palotti, Lorraine Goeuriot, Liadh Kelly, Mihai Lupu, Pavel Pecina, Henning Müller, Julie Budaher, and Anthony Deacon. 2016. The IR Task at the CLEF eHealth Evaluation Lab 2016: User-centred Health Information Retrieval. In *CLEF*.