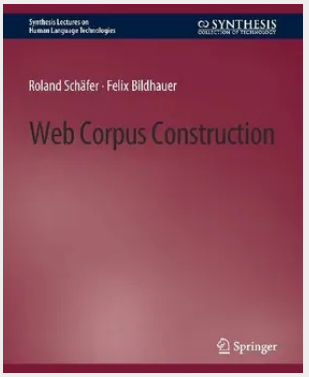
Schäfer / Bildhauer

Web Corpus Construction



The World Wide Web constitutes the largest existing source of texts written in a great variety of languages. A feasible and sound way of exploiting this data for linguistic research is to compile a static corpus for a given language. There are several adavantages of this approach: (i) Working with such corpora obviates the problems encountered when using Internet search engines in quantitative linguistic research (such as non-transparent ranking algorithms). (ii) Creating a corpus from web data is virtually free. (iii) The size of corpora compiled from the WWW may exceed by several orders of magnitudes the size of language resources offered elsewhere. (iv) The data is locally available to the user, and it can be linguistically post-processed and queried with the tools preferred by her/him. This book addresses the main practical tasks in the creation of web corpora up to giga-token size. Among these tasks are the sampling process (i.e., web crawling) and the usual cleanups including boilerplate removal and removal of duplicated content. Linguistic processing and problems with linguistic processing coming from the different kinds of noise in web corpora are also covered. Finally, the authors show how web corpora can be evaluated and compared to other corpora (such as traditionally compiled corpora). For additional material please visit the companion website: sites.morganclaypool.com/wcc Table of Contents: Preface / Acknowledgments / Web Corpora / Data Collection / Post-Processing / Linguistic Processing / Corpus Evaluation and Comparison / Bibliography / Authors' Biographies

The World Wide Web constitutes the largest existing source of texts written in a great variety of languages. A feasible and sound way of exploiting this data for linguistic research is to compile a static corpus for a given language. There are several adavantages of this approach: (i) Working with such corpora obviates the problems encountered when using Internet search engines in quantitative linguistic research (such as non-transparent ranking algorithms). (ii) Creating a corpus from web data is virtually free. (iii) The size of corpora compiled from the WWW may exceed by several orders of magnitudes the size of language resources offered elsewhere. (iv) The data is locally available to the user, and it can be linguistically post-processed and queried with the tools preferred by her/him. This book addresses the main practical tasks in the creation of web corpora up to giga-token size. Among these tasks are the sampling process (i.e., web crawling) and the usual cleanups including boilerplate removal and removal of duplicated content. Linguistic processing and problems with linguistic processing coming from the different kinds of noise in web corpora are also covered. Finally, the authors show how web corpora can be evaluated and compared to other corpora (such as traditionally compiled corpora). For additional material please visit the companion website: sites.morganclaypool.com/wcc Table of Contents: Preface / Acknowledgments / Web Corpora / Data Collection / Post-Processing / Linguistic Processing / Corpus Evaluation and Comparison / Bibliography / Authors' Biographies

35,30 €
32,99 € (zzgl. MwSt.)

Lieferfrist: bis zu 10 Tage

Artikelnummer: 9783031010248
Medium: Buch
ISBN: 978-3-031-01024-8
Verlag: Springer
Erscheinungstermin: 22.07.2013
Sprache(n): Englisch
Auflage: Erscheinungsjahr 2013
Serie: Synthesis Lectures on Human Language Technologies
Produktform: Kartoniert
Gewicht: 291 g
Seiten: 129
Format (B x H): 191 x 235 mm

Kundenservice Verlag Dr. Otto Schmidt KG
Neumannstraße 10, 40235 Düsseldorf | kundenservice@otto-schmidt.de | 0221-93738-997 (Inland)