

Preface

People search on the web to find relevant information on topics they wish to know more about. Accordingly, companies analyze big data to discover new information that is relevant for their business. Today's search engines and big data analytics seek to fulfill such information needs ad-hoc i.e., immediately in response to a search query or similar. Often, the relevant information is hidden in large numbers of natural language texts from web pages and other documents. Instead of returning potentially relevant texts only, leading search and analytics applications have recently started to return relevant information directly. To obtain the information sought for from the texts, they perform *text mining*.

Text mining deals with tasks that target the inference of structured information from collections and streams of unstructured input texts. It covers all techniques needed to identify relevant texts, to extract relevant spans from these texts, and to convert the spans into high-quality information that can be stored in databases and analyzed statistically. Text mining requires task-specific text analysis processes that may consist of several interdependent steps. Usually, these processes are realized with *text analysis pipelines*. A text analysis pipeline employs a sequence of natural language processing algorithms where each algorithm infers specific types of information from the input texts. Although effective algorithms exist for various types, the use of text analysis pipelines is still restricted to a few predefined information needs. We argue that this is due to three problems:

First, text analysis pipelines are mostly constructed manually for the tasks to be addressed, because their design requires expert knowledge about the algorithms to be employed. When information needs have to be fulfilled that are unknown beforehand, text mining hence cannot be performed ad-hoc. Second, text analysis pipelines tend to be inefficient in terms of run-time, because their execution often includes analyzing texts with computationally expensive algorithms. When information needs have to be fulfilled ad-hoc, text mining hence cannot be performed in the large. And third, text analysis pipelines tend not to robustly achieve high effectiveness on all input texts (in terms of the correctness of the inferred information), because they often include algorithms that rely on domain-dependent features of texts. Generally, text mining hence cannot guarantee to infer high-quality information at present.

In this book, we tackle the outlined problems by investigating how to fulfill information needs from text mining ad-hoc in a run-time efficient and domain-robust manner. Text mining is studied within the broad field of computational linguistics, bringing together research from natural language processing, information retrieval, and data mining. On the basis of a concise introduction to the foundations and the state of the art of text mining, we observe that knowledge about a text analysis process as well as information obtained within the process can be exploited in order to improve the design, the execution, and the output of the text analysis pipeline that realizes the process. To do this fully automatically, we apply different techniques from classic and statistical artificial intelligence.

In particular, we first develop knowledge-based artificial intelligence approaches for an ad-hoc pipeline construction and for the optimal execution of a pipeline on its input. Then, we show how to theoretically and practically optimize and adapt the schedule of the algorithms in a pipeline based on information in the analyzed input texts in order to maximize the pipeline's run-time efficiency. Finally, we learn novel patterns in the overall structures of input texts statistically that remain strongly invariant across the domains of the texts and that, thereby, allow for more robust analysis results in a specific set of text analysis tasks.

We analyze all the developed approaches formally and we sketch how to implement them in software applications. On the basis of respective Java open-source applications that we provide online, we empirically evaluate the approaches on established and on newly created collections of texts. In our experiments, we address scientifically and industrially important text analysis tasks, such as the extraction of financial events from news articles or the fine-grained sentiment analysis of reviews.

Our findings presented in this book show that text analysis pipelines can be designed automatically, which process only portions of text that are relevant for the information need to be fulfilled. Through an informed scheduling, we improve the run-time efficiency of pipelines by up to more than one order of magnitude without compromising their effectiveness. Even on heterogeneous input texts, efficiency can be maintained by learning to predict the fastest pipeline for each text individually. Moreover, we provide evidence that the domain robustness of a pipeline's effectiveness substantially benefits from focusing on overall structure in argumentation-related tasks such as sentiment analysis.

We conclude that the developed approaches denote essential building blocks of enabling ad-hoc large-scale text mining in web search and big data analytics applications. In this regard, the book at hand serves as a guide for practitioners and interested readers that desire to know what to pay attention to in the context of text analysis pipelines. At the same time, we are confident that our scientific results prove valuable for other researchers who work on the automatic understanding of natural language and on the future of information search.

Text Analysis Pipelines

Towards Ad-hoc Large-Scale Text Mining

Wachsmuth, H.

2015, XX, 302 p. 74 illus. in color., Softcover

ISBN: 978-3-319-25740-2