125 Problems in Text Algorithms

String matching is one of the oldest algorithmic techniques, yet still one of the most pervasive in computer science. The past 20 years have seen technological leaps in applications as diverse as information retrieval and compression. This copiously illustrated collection of puzzles and exercises in key areas of text algorithms and combinatorics on words offers graduate students and researchers a pleasant and direct way to learn and practice with advanced concepts.

The problems are drawn from a large range of scientific publications, both classic and new. Building up from the basics, the book goes on to showcase problems in combinatorics on words (including Fibonacci or Thue–Morse words), pattern matching (including Knuth–Morris–Pratt and Boyer–Moore–like algorithms), efficient text data structures (including suffix trees and suffix arrays), regularities in words (including periods and runs) and text compression (including Huffman, Lempel–Ziv and Burrows–Wheeler–based methods).

MAXIME CROCHEMORE is Emeritus Professor at Université Gustave Eiffel and of King's College London. He holds an honorary doctorate from the University of Helsinki. He is the author of more than 200 articles on algorithms on strings and their applications, and co-author of several books on the subject.

THIERRY LECROQ is a professor in the Department of Computer Science at the University of Rouen Normandy (France). He is currently head of the research team Information Processing in Biology and Health of the Laboratory of Computer Science, Information Processing and System. He has been one of the coordinators of the working group in stringology of the French National Centre for Scientific Research for more than 10 years.

WOJCIECH RYTTER is a professor at the Faculty of Mathematics, Informatics and Mechanics, University of Warsaw. He is the author of a large number of publications on automata, formal languages, parallel algorithms and algorithms on texts. He is a co-author of several books on these subjects, including *Efficient Parallel Algorithms, Text Algorithms* and *Analysis of Algorithms and Data Structures*. He is a member of Academia Europaea.

125 Problems in Text Algorithms

With Solutions

MAXIME CROCHEMORE
Gustave Eiffel University

THIERRY LECROQ
University of Rouen Normandy

WOJCIECH RYTTER
University of Warsaw

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre,
New Delhi – 110025, India

79 Anson Road, #06–04/06, Singapore 079906

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning, and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781108835831
DOI: 10.1017/9781108869317

Illustrations designed by Hélène Crochemore

First published 2021

Printed in the United Kingdom by TJ Books Ltd, Padstow Cornwall

A catalogue record for this publication is available from the British Library.

Library of Congress Cataloging-in-Publication Data
Names: Crochemore, Maxime, 1947– author. | Lecroq, Thierry, author. |
Rytter, Wojciech, author.
Title: One twenty five problems in text algorithms / Maxime Crochemore,
Thierry Lecroq, Wojciech Rytter.
Other titles: 125 problems in text algorithms
Description: New York : Cambridge University Press, 2021. |
The numerals 125 are superimposed over "One twenty five" on the title page. |
Includes bibliographical references and index.
Identifiers: LCCN 2021002037 (print) | LCCN 2021002038 (ebook) |
ISBN 9781108835831 (hardback) | ISBN 9781108798853 (paperback) |
ISBN 9781108869317 (epub)
Subjects: LCSH: Text processing (Computer science)–Problems, exercises, etc. |
Computer algorithms–Problems, exercises, etc.
Classification: LCC QA76.9.T48 C758 2021 (print) |
LCC QA76.9.T48 (ebook) | DDC 005.13–dc23
LC record available at https://lccn.loc.gov/2021002037
LC ebook record available at https://lccn.loc.gov/2021002038

ISBN 978-1-108-83583-1 Hardback
ISBN 978-1-108-79885-3 Paperback

Contents

Preface

This book is about algorithms on texts, also called algorithmic stringology. Text (word, string, sequence) is one of the main unstructured data types and the subject is of vital importance in computer science.

The subject is versatile because it is a basic requirement in many sciences, especially in computer science and engineering. The treatment of unstructured data is a very lively area and demands efficient methods owing both to their presence in highly repetitive instructions of operating systems and to the vast amount of data that needs to be analysed on digital networks and equipments. The latter is clear for information technology companies that manage massive data in their data centres but also holds for most scientific areas beyond Computer science.

The book presents a collection of the most interesting representative problems in stringology. They are introduced in a short and pleasant way and open doors to more advanced topics. They were extracted from hundreds of serious scientific publications, some of which are more than a hundred years old and some are very fresh and up to date. Most of the problems are related to applications while others are more abstract. The core part of most of them is an ingenious short algorithmic solution except for a few introductory combinatorial problems.

This is not just yet another monograph on the subject but a series of problems (puzzles and exercises). It is a complement to books dedicated to the subject in which topics are introduced in a more academic and comprehensive way. Nevertheless, most concepts in the field are included in the book, which fills a missing gap and is very expected and needed, especially for students and teachers, as the first problem-solving textbook of the domain.

The book is organised into seven chapters:

'The Very Basics of Stringology' is a preliminary chapter introducing the terminology, basic concepts and tools for the next chapters and that reflects six main streams in the area.

'Combinatorial Puzzles' is about combinatorics on words, an important topic because many algorithms are based on combinatorial properties of their input.

'Pattern Matching' deals with the most classical subject, text searching and string matching.

'Efficient Data Structures' is about data structures for text indexing. They are used as fundamental tools in a large number of algorithms, such as special arrays and trees associated with texts.

'Regularities in Words' concerns regularities that occur in texts, in particular repetitions and symmetries, that have a strong influence on the efficiency of algorithms.

'Text Compression' is devoted to several methods of the practically important area of conservative text compression.

'Miscellaneous' contains various problems that do not fit in earlier chapters but certainly deserve presentation.

Problems listed in the book have been accumulated and developed over several years of teaching on string algorithms in our own different institutions in France, Poland, UK and USA. They have been taught mostly to master's students and are given with solutions as well as with references for further readings. The content also profits from the experience authors gained in writing previous textbooks.

Anyone teaching graduate courses on data structures and algorithms can select whatever they like from our book for their students. However, the overall book is not elementary and is intended as a reference for researchers, PhD and master's students, as well as for academics teaching courses on algorithms even if they are not directly related to text algorithms. It should be viewed as a companion to standard textbooks on the domain. The self-contained presentation of problems provides a rapid access to their understanding and to their solutions without requiring a deep background on the subject.

The book is useful for specialised courses on text algorithms, as well as for more general courses on algorithms and data structures. It introduces all required concepts and notions to solve problems but some prerequisites in bachelor- or sophomore-level academic courses on algorithms, data structures and discrete mathematics certainly help in grasping the material more easily.