

Jakub Radoszewski

Project title: “New Trends in Text Algorithms”

Project Description for the General Public

With the development of digitization and its applications in almost every area of technology and everyday life, massive amounts of data in a digital format are currently being produced and stored. Terabytes of new generally unstructured sequential data—which we further refer to for simplicity as texts—appear daily on the Internet. In order to take advantage of this resource, very efficient algorithms (methods) for processing textual data are in need. Design of such algorithms is the main goal of this project.

Problems considered in this project include searching for patterns and various kinds of regularities in texts. We aim both at algorithms that are efficient in practise and at providing detailed theoretical explanations for the presented algorithms and their behaviour which will increase the understanding of those methods. We will consider several scenarios that arise from practical applications for which already known algorithms often turn out to be unsatisfactory. As an example, we will consider a scenario in which a data text arrives as a stream and there is not sufficient memory to store the whole text or even its significant parts. We will also combine known methods with cutting-edge machine learning techniques in order to obtain algorithms that achieve high efficiency on real-world data. In particular, an expected outcome of the project will be specific implementations of selected algorithms.