Introduction to Common Crawl

Malte Ostendorff (Common Crawl Foundation)

Abstract:
Common Crawl is a free, open corpus of over 300 billion web pages collected across nearly two decades - a foundational resource cited in more than 12,000 research papers and widely used to train today's large language models. This talk will give an introduction to what Common Crawl is, how the crawl works, what datasets it provides, and how the changing state of the open web affects crawling.