GitHubGo: A Distributed Search Engine for GitHub READMEs

Alex Ding
Brown University
alexander_ding@brown.edu

Megan Frisella
Brown University
megan_frisella@brown.edu

Sophia Liu
Brown University
sophia_liu@brown.edu

Tianran Zhang
Brown University
tianran_zhang@brown.edu

Abstract

This paper presents the design, implementation, and evaluation of GitHubGo, a distributed and scalable GitHub README search engine developed for CS1380 at Brown University. GitHubGo implements GitHub README crawling and indexing workflows on a custom distributed MapReduce pipeline. It supports arbitrary-length search strings and uses TF-IDF scores to index search terms. The distributed system runs on 16 Amazon Web Services EC2 servers. GitHubGo achieves a throughput of 1247 queries per second doing distributed search over 300k GitHub READMEs.