Notes for Paper

Matt Warner

Last updated: December 6, 2025

Contents

※ Section completeness

There are no limitations on the format or content of README files, however GitHub recommends to include some common sections to make it easier for users to find information quickly. [?]

- Description
- Installation
- Usage
- License
- Contribution
- Table of Contents
- Credits

The problem with determining whether or not a readme contains these sections is that devlopers often use different terms to refer to the same concept. For example, They may include a "build" section that contains *installation* information, or a "quickstart" section that describes usage information. This makes it difficult to

extract sections from a readme with simple string matching techniques like regex. Instead, we opted for a keyword based implementation where each section name maps to a set of common keywords that developers use to refer to the section. In a prior study, two authors manually inspected different samples of 383 READMES to determine the set of keywords for each section. After inspection, they found the following keywords:

Section	**Keywords**
Description	describe, description, overview, about, summary, introduction, "what is"
Usage	use, usage, quickstart, run, start document, docs example, demo, sample troubleshoot
Installation	install, build, setup, download, compile
License	license, licence, copyright
Credits	credit, acknowledge, author
Table of Contents	content
Contribution	contribute, contribution, contributing

Table 1: Section names and keyword sets used detection.

1.1 Method

GitHub recommends that developers write their README files using GitHub Flavored Markdown (GFM), an extended version of standard Markdown. However, there is no requirement that developers follow this convention. Developers can instead use HTML, reStructuredText, or even plain text to write their READMEs. This makes section detection challenging since a README written in HTML would express a section heading as `<h1>Section Heading</h1>`, while a Markdown README would represent the same structure as # Section Heading. To parse Markdown-based READMEs, we use the Mistune library, which converts Markdown into an abstract syntax tree (AST). This representation makes it easier to extract section headings and analyze document structure.

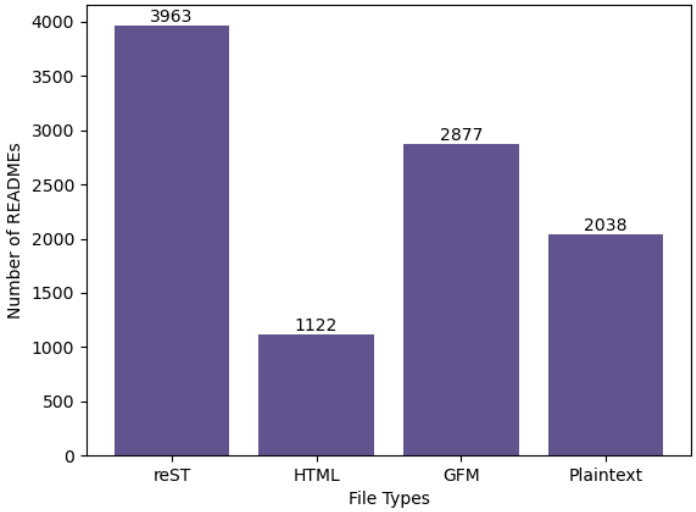

Figure 1: Number of README files observed in each file format