|  |  |  |  |
| --- | --- | --- | --- |
| Dataset | Content | Size | Used in tasks |
| Py150 | Python source code,  AST | 8423 repos, 149993 files | Fine-tuning CUBERT models for Variable-Misuse Classification, Wrong Binary Operator, Swapped Operand, Function-Docstring Mismatch, Exception Type, Variable-Misuse Localization and Repair, |
| Bigquery github repos | GitHub metadata | 250M commits,  Metadata for 2 billion files, metadata for 3 million projects | Pre-training CUBERT model for Variable-Misuse Classification, Wrong Binary Operator,Swapped Operand, Function-Docstring Mismatch, Exception Type, Variable-Misuse Localization and Repair |
| https://libraries.io | Metadata for software packages | 306k packages | Empirical Comparison of Dependency Network, |
| GHTorrent | GitHub metadata |  |  |
|  |  |  |  |
| Labeled datasets |  |  |  |
|  |  |  |  |
| SEIP Lab Software Defect Prediction Data | Complexity metrics | 5 subsequent releases of 3 projects from the Java Eclipse community | Data collection and linking |
| PROMISE Software Engineering Repository | Numeric metrics;  module has reported defects (false/true) | 15 000 modules | Defect prediction |
| NASA Defect Dataset | Numeric metrics;  module has reported defects (false/true) | 51 000 modules | Defect prediction |
| GPHR | Java code and metrics | 3526 pairs of fragments, buggy and fixed, code metrics | Defect prediction |
| BugHunter Dataset | Hand-crafted metrics; fix-inducing commit;  number of reported bugs | 159k pairs for 3 granularity levels (file/class/method)  15 projects | Analyzing the importance of complexity metrics |
| Neural Code Translator  Dataset | Pairs of buggy and fixed abstracted method-level fragments | 46k pairs of small fragments (under 50 tokens)  50k pairs of medium fragments  (under 100 tokens) | Code refinement |
|  |  |  |  |