A Permissive Licenses

For the experiments presented in the paper, we use the following SPDX identifiers for our permissive license dataset:

- MIT-0
- MIT
- MIT-feh
- Apache-2.0
- BSD-3-Clause
- BSD-3-Clause-Clear
- BSD-3-Clause-No-Nuclear-License-2014
- BSD-2-Clause
- CC0-1.0
- EPL-1.0
- MPL-2.0
- Unlicense
- ISC
- Artistic-2.0
- deprecated_LGPL-3.0+
- deprecated_LGPL-2.1+
- ECL-2.0
- SHL-0.51
- MPL-2.0-no-copyleft-exception

After running all experiments, it was brought to our attention that licenses such as MPL, LGPL, and EPL were erroneously labeled as permissive when they are in fact weak copyleft licenses. We have removed these weak copyleft license files and will release the updated version of The Stack. The weak copyleft-licensed data is only a small part of the overall dataset (below 0.5% for the Python subset), hence we expect the experimental findings of the paper to remain unchanged. In the next section, we describe how we updated The Stack.

B The Stack v1.1

For the Stack v1.1, we rely on the Blue Oak Council[15] to classify the licenses. The new classification process results in 193 permissive licenses (which we list, for completeness, at the end of this section). The Stack v1.1 also includes more programming languages. We used the following list of programming language extensions `https://gist.github.com/ppisarczyk/43962d06686722d26d176fad46879d41` and obtained data for 370 programming languages. We show the updated statistics for the 30 popular programming languages in Table 7. In general, we find that the amount of data increased for most programming languages. Specifically, we see a large increase in data for C# (128.37 GB vs 215.07 GB), Tex (4.65 GB vs 8.19 GB), and Markdown (164.61 GB vs 245.26 GB). On the other hand, we see a minor decrease in data for Go (118.37 GB vs 112.86 GB) and Rust (40.35 GB vs 39.85 GB). We release this updated version to the research community.

[15] `https://blueoakcouncil.org/list`

	All-licenses		**Permissive**		**Perm. + near-dedup**	
Language	Size (GB)	Files (M)	Size (GB)	Files (M)	Size (GB)	Files (M)
Assembly	36.04	1.34	2.57	0.36	1.65	0.26
Batchfile	31.05	2.82	1.06	0.44	0.33	0.28
C	1461.23	95.57	255.29	21.38	75.93	11.21
C++	644.28	105.96	215.07	14.82	65.97	7.60
C#	1106.54	62.72	133.55	21.7	57.98	13.28
CMake	11.25	3.59	2.1	0.59	0.68	0.25
CSS	1040.53	50.47	150.2	5.89	34.86	3.59
Dockerfile	3.89	3.74	1.9	1.27	0.52	0.64
FORTRAN	26.67	1.21	3.81	0.29	2.08	0.19
GO	271.92	23.34	112.86	11.65	32.01	5.89
Haskell	15.79	2.06	5.85	0.8	2.75	0.58
HTML	9491.23	267.81	812.73	35.6	291.46	16.60
Java	1311.99	279.16	266.41	42.43	112.82	25.12
Javascript	5820.23	209.51	496.22	40.11	166.24	25.43
Julia	21.75	0.88	3.4	0.48	1.75	0.33
Lua	88.39	5.18	7.09	0.93	3.77	0.64
Makefile	39.36	6.34	6.88	1.48	2.14	0.80
Markdown	706.8	135.69	245.26	40.75	95.84	25.66
Perl	49.21	2.74	7.08	0.83	2.99	0.48
PHP	779.66	115.53	185.79	34.85	89.46	22.63
PowerShell	13.26	1.39	3.42	0.53	1.65	0.33
Python	737.89	106.91	200.93	24.21	80.13	15.15
Ruby	78.63	30.74	25.95	7.21	9.78	4.46
Rust	78.97	6.3	39.85	3.06	12.92	1.68
Scala	28.37	6.06	15.56	2.79	6.06	1.61
Shell	71.56	14.01	9.07	3.77	4.07	2.54
SQL	1438.73	10.2	19.94	1.39	12.68	1.07
TeX	69.4	4.01	8.19	0.71	5.86	0.59
Typescript	4145.01	75.8	131.01	19.59	36.61	12.82
Visual Basic	28.57	1.97	3.81	0.4	1.71	0.19
Total	29648.2	1633.05	3372.85	340.31	1212.7	201.90

Table 7: An overview of the amount of data we collected for The Stack v1.1. We show the size and number of files for different splits of the data: the all-license, permissive license, and permissive license with near-deduplication.

- MIT
- Apache-2.0
- BSD-3-Clause
- Unlicense
- CC0-1.0
- BSD-2-Clause
- CC-BY-4.0
- CC-BY-3.0
- 0BSD

- RSA-MD
- WTFPL
- MIT-0
- ISC
- ADSL
- BSL-1.0
- Zlib
- Artistic-2.0
- FTL
- MS-PL
- BSD-2-Clause-FreeBSD
- FSFAP
- BSD-Source-Code
- Apache-1.1
- BSD-4-Clause
- Ruby
- Artistic-1.0
- MulanPSL-1.0
- BSD-1-Clause
- X11
- CNRI-Python
- Beerware
- Condor-1.1
- PostgreSQL
- CECILL-B
- Intel
- Vim
- Naumen
- OML
- BSD-3-Clause-Clear
- AML
- PHP-3.01
- OpenSSL

- PSF-2.0
- Xnet
- Linux-OpenIB
- BSD-3-Clause-LBNL
- UPL-1.0
- AFL-3.0
- BlueOak-1.0.0
- Info-ZIP
- BSD-4-Clause-UC
- AAL
- LPPL-1.3c
- bzip2-1.0.6
- W3C
- W3C-20150513
- AFL-1.1
- DOC
- ICU
- CC-BY-2.0
- curl
- MTLL
- OLDAP-2.2.1
- ECL-2.0
- Adobe-Glyph
- CNRI-Python-GPL-Compatible
- BSD-2-Clause-Patent
- IJG
- PHP-3.0
- ZPL-2.1
- MIT-advertising
- NCSA
- Fair
- BSD-3-Clause-Attribution
- OLDAP-2.3

- NLPL
- BSD-3-Clause-Open-MPI
- ClArtistic
- Python-2.0
- NASA-1.3
- TCL
- Artistic-1.0-Perl
- blessing
- BSD-3-Clause-No-Nuclear-Warranty
- ImageMagick
- Net-SNMP
- Artistic-1.0-cl8
- OLDAP-2.5
- MIT-feh
- OLDAP-2.4
- MITNFA
- AFL-2.1
- libpng-2.0
- EFL-2.0
- OLDAP-2.7
- IBM-pibs
- libtiff
- OLDAP-2.8
- Cube
- Adobe-2006
- BSD-2-Clause-NetBSD
- zlib-acknowledgement
- OLDAP-2.6
- BSD-3-Clause-No-Nuclear-License-2014
- OLDAP-1.4
- Libpng
- MIT-CMU
- AFL-2.0

- JasPer-2.0
- LPL-1.02
- Zend-2.0
- TCP-wrappers
- XFree86-1.1
- FSFUL
- OLDAP-1.3
- SGI-B-2.0
- NetCDF
- CNRI-Jython
- Zed
- ZPL-2.0
- AFL-1.2
- Apache-1.0
- CC-BY-1.0
- OLDAP-2.1
- OLDAP-1.2
- OLDAP-2.0
- NTP
- LPL-1.0
- AMPAS
- Barr
- mpich2
- ANTLR-PD
- Xerox
- Spencer-94
- AMDPLPA
- BSD-3-Clause-No-Nuclear-License
- HPND
- ECL-1.0
- MirOS
- Qhull
- ZPL-1.1

- TU-Berlin-2.0
- Spencer-86
- SMLNJ
- xinetd
- OLDAP-2.2.2
- OGTSL
- MIT-enna
- Font-exception-2.0
- FSFULLR
- TU-Berlin-1.0
- xpp
- NRL
- W3C-19980720
- EFL-1.0
- eGenix
- Unicode-DFS-2016
- SWL
- Spencer-99
- Plexus
- VSL-1.0
- Leptonica
- Unicode-DFS-2015
- Mup
- Giftware
- OLDAP-2.2
- APAFML
- NBPL-1.0
- OLDAP-1.1
- Entessa
- Multics
- Newsletr
- psutils
- bzip2-1.0.5

- Afmparse
- diffmark
- BSD-2-Clause-Views
- DSDP
- MIT-Modern-Variant
- ANTLR-PD-fallback
- Bahyph
- BSD-3-Clause-Modification
- BSD-4-Clause-Shortened
- HTMLTIDY
- MIT-open-group
- MulanPSL-2.0
- OLDAP-2.0.1
- Saxpath
- Borceux
- Crossword
- CrystalStacker
- Rdisc
- Wsuipa

C Excluded file extensions

Part of this list was taken from `https://github.com/EleutherAI/github-downloader/blob/345e7c4cbb9e0dc8a0615fd995a08bf9d73b3fe6/download_repo_text.py`

'apk', 'app', 'bin', 'bmp', 'bz2', 'class', 'csv', 'dat', 'db', 'deb', 'dll', 'dylib', 'egg', 'eot', 'exe', 'gif', 'gitignore', 'glif', 'gradle', 'gz', 'ico', 'jar', 'jpeg', 'jpg', 'lib', 'lo', 'lock', 'log', 'mp3', 'mp4', 'nar', 'o', 'ogg', 'otf', 'p', 'pdb', 'pdf', 'png', 'pickle', 'pkl', 'ppt', 'pptx', 'pyc', 'pyd', 'pyo', 'rar', 'rkt', 'so', 'ss', 'svg', 'tar', 'tif', 'tiff', 'tsv', 'ttf', 'war', 'wav', 'webm', 'woff', 'woff2', 'xz', 'zip', 'zst'

D Included programming language extensions

This list of programming language extensions is taken from `https://gist.github.com/ppisarczyk/43962d06686722d26d176fad46879d41`.

.abap .asc .ash .ampl .mod .g4 .apib .apl .dyalog .asp .asax .ascx .ashx .asmx .aspx .axd .dats .hats .sats .as .adb .ada .ads .agda .als .apacheconf .vhost .cls .applescript .scpt .arc .ino .asciidoc .adoc .asc .aj .asm .a51 .inc .nasm .aug .ahk .ahkl .au3 .awk .auk .gawk .mawk .nawk .bat .cmd .befunge .bison .bb .bb .decls .bmx .bsv .boo .b .bf .brs .bro .c .cats .h .idc .w .cs .cake .cshtml .csx .cpp .c++ .cc .cp .cxx .h .h++ .hh .hpp .hxx .inc .inl .ipp .tcc .tpp .c-objdump .chs .clp .cmake .cmake.in .cob .cbl .ccp .cobol .cpy .css .csv .capnp .mss .ceylon .chpl .ch .ck .cirru .clw .icl .dcl .click .clj .boot .cl2 .cljc .cljs .cljs.hl .cljscm .cljx .hic .coffee ._coffee .cake .cjsx .cson .iced .cfm .cfml .cfc .lisp .asd .cl .l .lsp .ny .podsl .sexp .cp .cps .cl .coq .v .cppobjdump .c++-objdump .c++objdump .cpp-objdump .cxx-objdump .creole .cr .feature .cu .cuh .cy .pyx .pxd .pxi .d .di .d-objdump

.com .dm .zone .arpa .d .darcspatch .dpatch .dart .diff .patch .dockerfile .djs .dylan .dyl .intr .lid .E .ecl .eclxml .ecl .sch .brd .epj .e .ex .exs .elm .el .emacs .emacs.desktop .em .emberscript .erl .es .escript .hrl .xrl .yrl .fs .fsi .fsx .fx .flux .f90 .f .f03 .f08 .f77 .f95 .for .fpp .factor .fy .fancypack .fan .fs .for .eam.fs .fth .4th .f .for .forth .fr .frt .fs .ftl .fr .g .gco .gcode .gms .g .gap .gd .gi .tst .s .ms .gd .glsl .fp .frag .frg .fs .fsh .fshader .geo .geom .glslv .gshader .shader .vert .vrx .vsh .vshader .gml .kid .ebuild .eclass .po .pot .glf .gp .gnu .gnuplot .plot .plt .go .golo .gs .gst .gsx .vark .grace .gradle .gf .gml .graphql .dot .gv .man .1 .1in .1m .1x .2 .3 .3in .3m .3qt .3x .4 .5 .6 .7 .8 .9 .l .me .ms .n .rno .roff .groovy .grt .gtpl .gvy .gsp .hcl .tf .hlsl .fx .fxh .hlsli .html .htm .html.hl .inc .st .xht .xhtml .mustache .jinja .eex .erb .erb.deface .phtml .http .hh .php .haml .haml.deface .handlebars .hbs .hb .hs .hsc .hx .hxsl .hy .bf .pro .dlm .ipf .ini .cfg .prefs .pro .properties .irclog .weechatlog .idr .lidr .ni .i7x .iss .io .ik .thy .ijs .flex .jflex .json .geojson .lock .topojson .json5 .jsonld .jq .jsx .jade .j .java .jsp .js .__js .bones .es .es6 .frag .gs .jake .jsb .jscad .jsfl .jsm .jss .njs .pac .sjs .ssjs .sublime-build .sublime-commands .sublime-completions .sublime-keymap .sublime-macro .sublime-menu .sublime-mousemap .sublime-project .sublime-settings .sublime-theme .sublime-workspace .sublime_metrics .sublime_session .xsjs .xsjslib .jl .ipynb .krl .sch .brd .kicad_pcb .kit .kt .ktm .kts .lfe .ll .lol .lsl .lslp .lvproj .lasso .las .lasso8 .lasso9 .ldml .latte .lean .hlean .less .l .lex .ly .ily .b .m .ld .lds .mod .liquid .lagda .litcoffee .lhs .ls .__ls .xm .x .xi .lgt .logtalk .lookml .ls .lua .fcgi .nse .pd_lua .rbxs .wlua .mumps .m .m4 .m4 .ms .mcr .mtml .muf .m .mak .d .mk .mkfile .mako .mao .md .markdown .mkd .mkdn .mkdown .ron .mask .mathematica .cdf .m .ma .mt .nb .nbp .wl .wlt .matlab .m .maxpat .maxhelp .maxproj .mxt .pat .mediawiki .wiki .m .moo .metal .minid .druby .duby .mir .mirah .mo .mod .mms .mmk .monkey .moo .moon .myt .ncl .nl .nsi .nsh .n .axs .axi .axs.erb .axi.erb .nlogo .nl .lisp .lsp .nginxconf .vhost .nim .nimrod .ninja .nit .nix .nu .numpy .numpyw .numsc .ml .eliom .eliomi .ml4 .mli .mll .mly .objdump .m .h .mm .j .sj .omgrofl .opa .opal .cl .opencl .p .cls .scad .org .ox .oxh .oxo .oxygene .oz .pwn .inc .php .aw .ctp .fcgi .inc .php3 .php4 .php5 .phps .phpt .pls .pck .pkb .pks .plb .plsql .sql .sql .pov .inc .pan .psc .parrot .pasm .pir .pas .dfm .dpr .inc .lpr .pp .pl .al .cgi .fcgi .perl .ph .plx .pm .pod .psgi .t .6pl .6pm .nqp .p6 .p6l .p6m .pl .pl6 .pm .pm6 .t .pkl .l .pig .pike .pmod .pod .pogo .pony .ps .eps .ps1 .psd1 .psm1 .pde .pl .pro .prolog .yap .spin .proto .asc .pub .pp .pd .pb .pbi .purs .py .bzl .cgi .fcgi .gyp .lmi .pyde .pyp .pyt .pyw .rpy .tac .wsgi .xpy .pytb .qml .qbs .pro .pri .r .rd .rsx .raml .rdoc .rbbas .rbfrm .rbmnu .rbres .rbtbar .rbuistate .rhtml .rmd .rkt .rktd .rktl .scrbl .rl .raw .reb .r .r2 .r3 .rebol .red .reds .cw .rpy .rs .rsh .robot .rg .rb .builder .fcgi .gemspec .god .irbrc .jbuilder .mspec .pluginspec .podspec .rabl .rake .rbuild .rbw .rbx .ru .ruby .thor .watchr .rs .rs.in .sas .scss .smt2 .smt .sparql .rq .sqf .hqf .sql .cql .ddl .inc .prc .tab .udf .viw .sql .db2 .ston .svg .sage .sagews .sls .sass .scala .sbt .sc .scaml .scm .sld .sls .sps .ss .sci .sce .tst .self .sh .bash .bats .cgi .command .fcgi .ksh .sh.in .tmux .tool .zsh .sh-session .shen .sl .slim .smali .st .cs .tpl .sp .inc .sma .nut .stan .ML .fun .sig .sml .do .ado .doh .ihlp .mata .matah .sthlp .styl .sc .scd .swift .sv .svh .vh .toml .txl .tcl .adp .tm .tcsh .csh .tex .aux .bbx .bib .cbx .cls .dtx .ins .lbx .ltx .mkii .mkiv .mkvi .sty .toc .tea .t .txt .fr .nb .ncl .no .textile .thrift .t .tu .ttl .twig .ts .tsx .upc .anim .asset .mat .meta .prefab .unity .uno .uc .ur .urs .vcl .vhdl .vhd .vhf .vhi .vho .vhs .vht .vhw .vala .vapi .v .veo .vim .vb .bas .cls .frm .frx .vba .vbhtml .vbs .volt .vue .owl .webidl .x10 .xc .xml .ant .axml .ccxml .clixml .cproject .csl .csproj .ct .dita .ditamap .ditaval .dll.config .dotsettings .filters .fsproj .fxml .glade .gml .grxml .iml .ivy .jelly .jsproj .kml .launch .mdpolicy .mm .mod .mxml .nproj .nuspec .odd .osm .plist .pluginspec .props .ps1xml .psc1 .pt .rdf .rss .scxml .srdf .storyboard .stTheme .sublime-snippet .targets .tmCommand .tml .tmLanguage .tmPreferences .tmSnippet .tmTheme .ts .tsx .ui .urdf .ux .vbproj .vcxproj .vssettings .vxml .wsdl .wsf .wxi .wxl .wxs .x3d .xacro .xaml .xib .xlf .xliff .xmi .xml.dist .xproj .xsd .xul .zcml .xsp-config .xsp.metadata .xpl .xproc .xquery .xq .xql .xqm .xqy .xs .xslt .xsl .xojo_code .xojo_menu .xojo_report .xojo_script .xojo_toolbar .xojo_window .xtend .yml .reek .rviz .sublime-syntax .syntax .yaml .yaml-tmlanguage .yang .y .yacc .yy .zep .zimpl .zmpl .zpl .desktop .desktop.in .ec .eh .edn .fish .mu .nc .ooc .rst .rest .rest.txt .rst.txt .wisp .prg .ch .prw