Andrew’s C/C++ Token Count Dataset 2016 (ACTCD16)

Project: Programming Language C++
Author: Andrew Tomazos <andrewtomazos@gmail.com>
Date: 2016-01-26

Abstract

We parsed 4,689,316,529 C/C++ tokens from 2,566,989 C/C++ source files taken from 11,423 open source packages of a popular Linux distribution. For each of the 50,325,647 distinct token spellings, we counted the number of occurrences, and output these tokens and counts into a single data file. We make that data file available for download as the ACTCD16 dataset.

Data File Location

http://www.tomazos.com/actcd16.txt.gz (392 MB)

Data File Format

The first line contains a single decimal integer N, the number of distinct tokens.

There are then N records in the text file, one for each distinct token.

The ith record has the following components, separated by space characters, and followed by a newline character:

1. A positive decimal integer C, the number of occurrences of the ith distinct token.
2. A positive decimal integer L, the byte length of the spelling of the ith distinct token.
3. A character K, denoting the kind of the ith token. One of ‘H’, ‘I’, ‘N’, ‘C, ‘S’, ‘O’. (See **Token Kind** below.)
4. A string S consisting of L bytes, containing the spelling of the ith distinct token.

The tokens are in a major order of K and then a minor order of S. Please note that S can contain embedded space and newline characters, so to parse correctly consume exactly L bytes to read S.

Token Kind

The character K has the following meaning:

K	Kind	Description	Examples

‘H’	Header	The token that appears after a #include.	`<iostream>`
			`“myheader.h”`

‘I’	Identifier	Includes identifiers and keywords.	`i`
			`class`
			`my_widget42`

‘N’	Numeric	Numeric literals including user-defined ones (formally ppnumbers)	0 123.245 1ull 42foo
‘C’	Characters	Character literals including user-defined ones.	‘a’ ‘c’foo
‘S’	Strings	String literals including user-defined and raw string literals.	“foo” L”bar”foo R”(baz)”
‘O’	Operators	Operators and punctuation tokens.	(++ *=

Example Usage

Download and decompress the file:

```
$ wget http://www.tomazos.com/actcd16.txt.gz
$ gunzip actcd16.txt.gz
```

List all identifiers starting with ‘a’ that occur more than 10,000 times:

```
$ cat > parser.cc
```

```
#include <iostream>
#include <string>

int main() {
  size_t N; // num distinct tokens
  std::cin >> N;
  for (size_t i = 0; i < N; ++i) {
    size_t C; // occurences
    size_t L; // token length
    char K; // token kind
    std::string S; // token spelling

    // parse ith record
    std::cin >> C >> L;
    if (std::cin.get() != ' ') return -1;
    K = std::cin.get();
    if (std::cin.get() != ' ') return -1;
    S = std::string(L,'\0');
    std::cin.read(&S[0], L);
    if (std::cin.get() != '\n') return -1;

    // filter the ith record
    if (K == 'I' // if is identifier
        && C > 100000 // and occurs more than 100000 times
        && S[0] == 'a') // and starts with 'a'
      std::cout << C << " " << S << std::endl;
  }
}
^D

$ g++ parser.cc
$ ./a.out < actcd16.txt
3381676 a
181014 a1
145980 a2
114112 abort
100355 abs
111368 ac
325634 action
111201 active
330159 adapter
227061 add
1008148 addr
```

```
303081 address
120441 ah
230454 alpha
104547 always_inline
342402 ap
117945 app
381772 append
126844 ar
120310 area
1025576 arg
145666 arg0
1011374 arg1
595394 arg2
264110 arg3
146184 arg4
947445 argc
586598 argp1
101011 argp2
1757950 args
1382158 argv
339457 array
101775 as_fusion_element
1292572 assert
370139 at
145216 atoi
148411 ats_ptr_type
505333 attr
105076 attribute
114224 attributes
120702 attrs
127015 auto
134324 avctx
```

Steps to Reproduce

To create ACTCD16 we took the following steps:

1. Mirror Ubuntu Repository

We used `apt-mirror` to download all the source packages of Ubuntu 15.10 Wily release in early January 2016 using the following apt line: `deb-src`

```
http://archive.ubuntu.com/ubuntu wily main restricted universe
multiverse
```

2. Unpack

We unpackaged all the packages source tarballs using `dpkg-source -x` on the .dsc files (debian source control files).

3. Mark C/C++ source files

We categorized all the source files based on file extension and marked those with one of the following file extensions as C/C++ files:

```
 ext | num_files
-----+---------
 cc  |  143375
 hxx |   38721
 cxx |   70075
 cpp |  417624
 hpp |   82556
 h   | 1017085
 c   |  803244
 C   |   20156
```

4. Tokenize C/C++ source files

We executed standard C++ translation phase 1 through 3 on the source files assuming a UTF-8 encoding. We found that 99.0% of the source files tokenized successfully. Of the remaining 1.0% the majority of the errors were decoding problems (most likely from ISO-8859 / Latin1 encoding), and we simply discarded these files. This resulted in 4,689,316,529 tokens.

Note that formally these are “preprocessing tokens” - they are the token sequence before macro replacement, source file inclusion or conditional compliation.

5. Count Tokens

We discovered of the 4,689,316,529 tokens there were 50,325,647 distinct spellings. We counted the number of each of these spellings forming a histogram. We then output this histogram as ACTCD16 in the data file format described earlier.

Largest Packages

The 300 largest packages used by the number of C/C++ Source Files they contain are as follows:

```
              package              | nfiles
-----------------------------------+--------
 oxide-qt                          |   76053
 chromium-browser                  |   75903
 android                           |   43955
 gcc-arm-none-eabi                 |   41327
 linux-raspi2                      |   39294
 linux                             |   38934
 linux-flo                         |   34500
 linux-mako                        |   34383
 linux-hammerhead                  |   34047
 linux-manta                       |   32028
 linux-chromebook                  |   31431
 linux-goldfish                    |   31366
 thunderbird                       |   23767
 firefox                           |   23292
 oce                               |   21612
 boost-mpi-source1.58              |   21434
 boost1.58                         |   21434
 libreoffice                       |   19821
 libreoffice-l10n                  |   19821
 qt4-x11                           |   19319
 ps3-kboot                         |   19064
 wine-gecko2.21                    |   17133
 wine-gecko-2.21                   |   16866
 aster                             |   14792
 llvm-toolchain-3.7                |   14732
 paraview                          |   14008
 llvm-toolchain-3.6                |   12934
 llvm-toolchain-3.5                |   12157
 qtwebkit-source                   |   12150
 vxl                               |   11973
 glibc                             |   11966
 qtwebkit-opensource-src           |   11014
 calligra                          |   10941
 virtualbox                        |   10774
```

```
phantomjs                          | 10682
llvm-toolchain-3.4                 | 10673
mame                               | 10295
openjfx                            |  9501
webkit2gtk                         |  9252
qtbase-opensource-src              |  8818
qtbase-opensource-src-gles         |  8816
root-system                        |  8203
webkitgtk                          |  7919
vtk                                |  7403
edk2                               |  7287
cbmc                               |  7052
vnc4                               |  6983
vtk6                               |  6869
u-boot                             |  6172
insighttoolkit                     |  5807
vice                               |  5737
blender                            |  5556
qtcreator                          |  5459
u-boot-linaro                      |  5354
kde4libs                           |  5134
mongodb                            |  5044
qtmobility                         |  5036
mariadb-10.0                       |  4936
kdepim                             |  4911
libc++                             |  4907
libexplain                         |  4755
scummvm                            |  4518
gdb                                |  4477
mythtv                             |  4342
juju-mongodb                       |  4311
digikam                            |  4193
redboot-imx                        |  4119
wxwidgets3.0                       |  4021
kodi                               |  3984
cgal                               |  3981
scilab                             |  3886
samba                              |  3886
povray                             |  3880
wine-development                   |  3862
konclude                           |  3854
wine1.6                            |  3761
newlib                             |  3707
```

```
flint                             |   3668
wxpython3.0                       |   3656
libboost-geometry-utils-perl      |   3560
hypre                             |   3526
mesa                              |   3488
percona-xtrabackup                |   3391
gnulib                            |   3256
wireshark                         |   3227
texlive-bin                       |   3197
emscripten                        |   3196
percona-xtradb-cluster-5.6        |   3165
grass                             |   3160
percona-server-5.6                |   3143
ugene                             |   3114
wxwidgets2.8                      |   3099
gromacs                           |   3098
mysql-5.6                         |   3080
openmpi                           |   3080
gcc-h8300-hms                     |   3029
kamailio                          |   3003
gimp                              |   2967
mgltools-utpackages               |   2951
ace                               |   2911
mpich                             |   2856
ncbi-blast+                       |   2853
codeblocks                        |   2847
gcc-arm-linux-androideabi         |   2840
gst-libav1.0                      |   2807
qemu                              |   2793
fis-gtm                           |   2756
sdcc                              |   2754
lammps                            |   2749
freefoam                          |   2742
nim                               |   2704
mingw-w64                         |   2675
freemat                           |   2618
kopete                            |   2614
ffmpeg                            |   2610
tendra                            |   2585
kadu                              |   2532
mldemos                           |   2501
codelite                          |   2498
cmake                             |   2498
```

```
lyx                               |   2447
alliance                          |   2441
condor                            |   2434
aegis                             |   2430
dovecot                           |   2387
quantlib                          |   2360
mrpt                              |   2337
freecad                           |   2336
ns3                               |   2276
binutils                          |   2271
spring                            |   2242
lapack                            |   2239
praat                             |   2228
0ad                               |   2210
abiword                           |   2204
kde4pimlibs                       |   2195
coin3                             |   2180
posixtestsuite                    |   2178
qutecom                           |   2170
genometools                       |   2164
ruby-passenger                    |   2133
openscenegraph                    |   2124
qgis                              |   2117
atlas                             |   2113
syslinux                          |   2110
pcl                               |   2087
doomsday                          |   2040
tulip                             |   2036
postbooks                         |   2033
marble                            |   2030
valgrind                          |   2024
ngspice                           |   2021
ardour3                           |   1995
alsa-driver                       |   1994
qtdeclarative-opensource-src-gles |   1988
qtdeclarative-opensource-src      |   1988
petsc                             |   1973
sflphone                          |   1965
opencv                            |   1952
gdal                              |   1948
psi4                              |   1937
gnuradio                          |   1935
dx                                |   1933
```

```
ogre-1.9                           |   1916
amarok                             |   1914
inkscape                           |   1913
opencollada                        |   1904
freemedforms-project               |   1900
ncl                                |   1897
strongswan                         |   1897
gmsh                               |   1886
ceph                               |   1881
motif                              |   1878
sfftobmp                           |   1864
ardour                             |   1848
musl                               |   1832
supertuxkart                       |   1825
grub2                              |   1823
gtk+3.0                            |   1823
php5                               |   1820
ossim                              |   1820
connectome-workbench               |   1815
openjdk-7-jre-dcevm                |   1814
cc1111                             |   1810
poco                               |   1802
aspectc++                          |   1793
vlc                                |   1786
mozjs24                            |   1764
mir                                |   1763
flightgear                         |   1763
psicode                            |   1761
openturns                          |   1759
postgresql-9.4                     |   1755
bombono-dvd                        |   1737
metview                            |   1725
openafs                            |   1714
icu                                |   1698
ogre-1.8                           |   1696
gettext                            |   1694
caret                              |   1682
saga                               |   1678
libwildmagic                       |   1672
xen                                |   1671
shogun                             |   1668
clamav                             |   1667
gridengine                         |   1661
```

```
mozc                              |   1653
ruby-lapack                       |   1653
lam                               |   1652
mysql-workbench                   |   1643
magics++                          |   1620
audacity                          |   1618
clam                              |   1595
krb5                              |   1581
qpid-cpp                          |   1578
efl                               |   1571
scorched3d                        |   1557
fftw3                             |   1544
fftw3-mpi                         |   1542
avidemux                          |   1540
openwalnut                        |   1535
arb                               |   1527
smlnj                             |   1521
openhpi                           |   1518
darkradiant                       |   1509
gst-plugins-bad1.0                |   1505
ncbi-tools6                       |   1500
resiprocate                       |   1494
eso-midas                         |   1493
trafficserver                     |   1492
plee-the-bear                     |   1483
gdcm                              |   1479
netsurf                           |   1478
gcl                               |   1477
openmw                            |   1471
evolution                         |   1468
qttools-opensource-src            |   1455
gccxml                            |   1455
libvirt                           |   1442
kdevplatform                      |   1436
octave                            |   1432
squid3                            |   1426
ghostscript                       |   1420
golang-race-detector-runtime      |   1408
votca-tools                       |   1404
seqan                             |   1402
z3                                |   1401
sipxtapi                          |   1393
lucene++                          |   1377
```

```
berkeley-abc                      |   1375
manaplus                          |   1368
zeroc-ice                         |   1368
rosegarden                        |   1357
polymake                          |   1356
net-snmp                          |   1354
nwchem                            |   1354
binutils-h8300-hms                |   1346
gsl-doc                           |   1344
madness                           |   1339
gem                               |   1335
xview                             |   1324
libunistring                      |   1320
babel                             |   1319
gazebo                            |   1318
heimdal                           |   1312
xorg-server                       |   1308
bind9                             |   1300
pjproject                         |   1292
psi-plus                          |   1291
ivtools                           |   1290
netw-ib-ox-ag                     |   1287
herwig++                          |   1285
gsl                               |   1278
netbeans                          |   1275
okteta                            |   1265
mednafen                          |   1265
qtxmlpatterns-opensource-src      |   1251
sra-sdk                           |   1251
rheolef                           |   1250
ipxe                              |   1250
openssl                           |   1243
sofa-framework                    |   1231
qtmultimedia-opensource-src       |   1230
dcmtk                             |   1229
qtmultimedia-opensource-src-gles  |   1229
meshlab                           |   1224
gtk+2.0                           |   1222
tomahawk                          |   1211
gambas3                           |   1210
gst-plugins-bad0.10               |   1209
cegui-mk2                         |   1199
mcrl2                             |   1192
```

```
 asterisk                                      |   1189
 mailutils                                     |   1185
 aria2                                         |   1185
 scribus                                       |   1180
 wesnoth-1.12                                  |   1174
 cln                                           |   1173
 nss                                           |   1172
 mygui                                         |   1171
```

Most Frequent Tokens

The 300 most frequent tokens in ACTCD16 are:

C	**S**
543765950	,
358949156	)
358944875	(
307344831	;
116381126	=
101736204	{
101720059	}
92373740	*
85549149	->
69422997	0
57860291	.
55601899	#
48131328	if
38863533	[
38863467	]
38339502	::
36655384	&
32581019	1
29941162	-
29516614	int
28511353	return
24262801	const
22558096	void
22202099	define
21003934	<
20506305	:
20224291	i

```
18201918 ==
17092004 struct
16088373 >
15251494 +
14494597 NULL
13651290 include
12701944 static
12626672 char
12093447 0x00
11893742 2
11759435 !
11170593 else
9912350 <<
9416127 case
9193167 ++
8626299 !=
8188949 60
7991273 &&
7808654 endif
7344347 unsigned
7200646 break
6276795 p
6249275 |
6192522 0x0000
6190089 3
5794389 for
5643239 x
5605464 ||
5591148 class
5573651 bool
5436634 4
5044400 s
4861142 0x00000000
4665835 data
4157034 c
4078325 sizeof
4067776 8
4056690 std
4027017 typename
3926738 n
3884974 false
3868877 name
3777040 j
```

```
3747758 value
3725777 typedef
3702276 +=
3685338 ret
3683256 type
3588750 this
3522381 result
3495003 goto
3477698 size
3438600 long
3381676 a
3347681 double
3128443 dev
2999241 5
2983296 true
2862864 ifdef
2841964 y
2831199 buf
2821790 len
2739842 /
2664784 d
2644349 ?
2605544 b
2570387 r
2546215 ifndef
2468462 6
2449369 10
2403453 16
2395027 PyObject
2356728 0xff
2320167 flags
2310696 while
2292010 string
2278111 >=
2258467 size_t
2255463 friend
2243511 7
2241835 public
2172131 v
2118112 virtual
2108599 t
2102019 0.000
2094626 extern
```

2089033	0x0
2047792	self
2043297	val
2024662	err
2014377	32
1977807	error
1977052	64
1970403	new
1964389	defined
1954184	info
1927267	line
1914805	state
1910129	m
1846307	status
1797245	140
1794111	T
1780043	count
1757950	args
1743074	f
1735120	k
1718225	<=
1706749	offset
1697796	float
1694322	inline
1680082	index
1678991	rc
1672705	~
1668517	0x20
1651209	template
1642755	9
1638159	0xFF
1625265	>>
1603683	priv
1598329	end
1593199	e
1567142	id
1559775	FALSE
1559606	QString
1530157	255
1503568	str
1499679	ctx
1482512	12
1440979	length

1430472	switch
1421490	tmp
1414027	fprintf
1403995	ptr
1390603	|=
1389356	next
1384192	""
1383697	24
1383243	base
1382158	argv
1356511	res
1356392	w
1354582	out
1348158	printf
1346531	uint32_t
1336340	buffer
1323521	pos
1321733	l
1315658	u32
1308885	obj
1306195	tu_int
1292572	assert
1280770	node
1271425	13
1270570	h
1264380	width
1219315	key
1213242	namespace
1201475	15
1200001	TRUE
1187825	file
1180122	4.000
1172384	src
1140363	11
1137478	N
1124026	48
1115399	--
1105409	0x01
1104083	static_cast
1061555	u
1043310	start
1041707	0.0
1029077	path

```
1025576 arg
1024079 z
1015272 it
1011374 arg1
1008148 addr
996064 u8
992648 default
982712 list
973294 height
965588 20
949986 enum
949321 0x001a
947445 argc
945012 continue
941055 14
934858 vector
933492 _
920835 msg
918226 short
909916 parent
897085 128
893646 uint8_t
890810 o
887565 fd
885461 cmd
883418 event
870300 q
856840 private
843474 item
837898 context
828844 mode
828672 dst
823783 tag
823418 0x02
821399 free
810485 0x80
810184 A
808866 stderr
804909 op
801663 res1
794025 int32_t
777969 strcmp
766707 operator
```

```
765424 gchar
764738 get
758795 0x65
753215 idx
753107 entry
751937 port
749446 -=
748511 0x10
746934 100
744014 g
738039 17
714820 '\x00'
714672 begin
711000 0x0000000000000000uLL
707804 fp
705127 boost
704540 reg
703055 strlen
699530 undef
690351 NI
685753 ##
685350 text
684170 19
683802 ch
681340 This
680684 stream
678756 iter
678044 0x04
675061 0xffffffff
663917 delete
662114 in
651879 device
649375 rv
642890 filename
633172 18
633038 target
624321 params
623555 %
621315 resultobj
618990 mask
616756 0x74
614886 1.0
613537 254
```

608187 dest
607738 L
605963 st
603484 50
599785 temp
599263 '\0'
596268 code
595394 arg2
593959 object
591661 25
590495 skb
588597 lock
586598 argp1
586275 endl
585752 memcpy

Attribution

To attribute a result to this dataset you can use the following:

```
Short Name: ACTCD16
Long Name: Andrew’s C/C++ Token Count Dataset 2016 (ACTCD16)
URL: http://www.tomazos.com/actcd16
```