CS 221 PROJECT 2 – CRAWLING

KARTHIK PRASAD	42686317
PHANI SHEKHAR MANTRIPRAGADA	85686586
RISHABH SHAH	79403075

1. How much time did it take to crawl the entire domain?

Crawling time was approximately **8** hours.

2. How many unique pages did you find in the entire domain?

Number of Unique pages in the entire set: **19779**

<

Removed following websites:

- Greater than 1Mb size
- Traps having URL size>150 characters, URL path containing over 10 forward slashes ('/')
- Subdomains – Archive.ics.uci.edu, wics.ics.uci.edu, ftp.ics.uci.edu, duttgroup.ics.uci.edu
- Dynamic websites having more than two variables

>

3. How many subdomains did you find? Submit the list of subdomains ordered alphabetically and the number of unique pages detected in each subdomain.

Number of unique Subdomains found in the entire list: **78**

archive.ics.uci.edu, 1
asterix.ics.uci.edu, 46
asterixdb.ics.uci.edu, 45
betapro.proteomics.ics.uci.edu, 1
calendar.ics.uci.edu, 10
ccsw.ics.uci.edu, 2
cdb.ics.uci.edu, 11
cert.ics.uci.edu, 13
cgvw.ics.uci.edu, 156
checkmate.ics.uci.edu, 1
chime.ics.uci.edu, 1
circadiomics.ics.uci.edu, 4
closeup.ics.uci.edu, 12
computableplant.ics.uci.edu, 69
contact.ics.uci.edu, 5
cradl.ics.uci.edu, 179
cybert.ics.uci.edu, 18
dataguard.ics.uci.edu, 11
dataprotector.ics.uci.edu, 1
dblp.ics.uci.edu, 1
dynamo.ics.uci.edu, 30
elms.ics.uci.edu, 1
esl.ics.uci.edu, 5
evoke.ics.uci.edu, 1155

fano.ics.uci.edu, 8973
flamingo.ics.uci.edu, 29
fr.ics.uci.edu, 9
frost.ics.uci.edu, 4
ftp.ics.uci.edu, 1
gonet.genomics.ics.uci.edu, 773
grape.ics.uci.edu, 173
graphics.ics.uci.edu, 6
graphmod.ics.uci.edu, 220
hci.ics.uci.edu, 1
hobbes.ics.uci.edu, 9
hombao.ics.uci.edu, 15
i-sensorium.ics.uci.edu, 5
ibook.ics.uci.edu, 6
informatics.ics.uci.edu, 1
ipubmed.ics.uci.edu, 4
isg.ics.uci.edu, 10
jujube.ics.uci.edu, 16
lineup.ics.uci.edu, 11
luci.ics.uci.edu, 3
mine10.ics.uci.edu, 3
mine5.ics.uci.edu, 3
mlearn.ics.uci.edu, 16
mondego.ics.uci.edu, 50
motifmap.ics.uci.edu, 1
mupro.proteomics.ics.uci.edu, 3
niagara.ics.uci.edu, 1
pepito.proteomics.ics.uci.edu, 5
pregelix.ics.uci.edu, 11
psearch.ics.uci.edu, 3
radicle.ics.uci.edu, 1
riscit.ics.uci.edu, 2
satware.ics.uci.edu, 6
sconce.ics.uci.edu, 3
scratch.proteomics.ics.uci.edu, 4
sherlock.ics.uci.edu, 7
sli.ics.uci.edu, 1944
soc.ics.uci.edu, 52
sourcerer.ics.uci.edu, 1
sprout.ics.uci.edu, 51
students.ics.uci.edu, 1
tastier.ics.uci.edu, 1
test-cs.ics.uci.edu, 18
testlab.ics.uci.edu, 5
tmbpro.ics.uci.edu, 5
vid.ics.uci.edu, 1
vision.ics.uci.edu, 185
wildhog.ics.uci.edu, 1
woda15.ics.uci.edu, 1
www-db.ics.uci.edu, 11
www.graphics.ics.uci.edu, 6
www.ics.uci.edu, 5315
www.isg.ics.uci.edu, 9

4. What is the longest page in terms of number of words?

Longest page: **https://sli.ics.uci.edu/Classes/2009W-Comments**
Count of number of words: **88430**

5. What are the 500 most common words in this domain?

Here are the 500 most common words in ics.uci.edu domain:

level 70132
relatedness 67840
protein 32449
0 30971
1 30261
d 29359
computer 26650
data 26448
activity 26077
information 24107
eppstein 23687
2 22423
science 22387
irvine 19378
3 19121
school 18700
research 18498
2015 17070
publications 16096
can 15995
complex 14529
will 14096
function 14093
web 14057
reply 13487
ics 13344
unknown 13006
two 12904
uc 12702
uci 12435
page 12430
name 12393
4 12211
5 12105
subunit 12100
binding 11928
null 11909
university 11499
mutant 11352
database 11151
server 11070
time 10887

design	10620
c	10130
algorithms	9991
one	9977
experimental	9910
6	9843
factor	9714
bren	9538
news	9414
systems	9388
software	9383
dna	9269
http	9243
transcription	9225
affinity	9086
molecular	9047
m	9013
fano	8978
use	8927
cytoplasm	8913
e	8868
citation	8845
may	8620
edu	8606
precipitation	8603
new	8430
s	8325
10	8269
involved	8005
california	7921
rna	7841
required	7800
group	7719
computing	7710
j	7696
people	7549
nucleus	7511
also	7502
ii	7471
pdf	7338
viable	7268
2014	7263
7	7256
nuclear	7210
cell	7205
n	7190
number	7161
work	7130
hybrid	7105
using	7011
project	6996
based	6993
process	6989

author 6915
policies 6894
social 6848
graph 6789
go 6735
june 6681
values 6664
component 6633
contact 6543
site 6467
class 6396
11 6381
www 6368
good 6322
lab 6241
2013 6148
8 6144
students 6142
year 6135
acm 6083
synthetic 6069
p 6059
home 6048
pages 6015
academic 5959
like 5957
lethality 5912
kinase 5872
projects 5841
membrane 5817
post 5816
processing 5741
set 5645
code 5616
system 5535
com 5513
g 5494
file 5483
game 5464
sensu 5348
read 5345
type 5344
mrna 5331
proc 5312
polymerase 5304
see 5282
t 5221
15 5177
large 5078
user 5064
12 5033
cs 5029
gene 5026

paper 4994
make 4963
web site 4923
b 4921
pp 4866
student 4851
graduate 4848
first 4820
march 4745
technology 4736
best 4717
applications 4700
comments 4651
2012 4609
point 4607
blog 4603
support 4592
title 4587
engineering 4583
computational 4576
2011 4511
problems 4511
world 4509
r 4502
many 4500
version 4425
2008 4374
dependent 4373
example 4367
development 4341
conference 4337
problem 4314
application 4298
learning 4295
list 4291
k 4289
2010 4276
graphs 4256
analysis 4231
faculty 4221
22 4175
get 4119
different 4107
search 4055
algorithm 4040
int 4023
dynamic 4003
transport 3999
size 3991
25 3985
used 3980
points 3970
20 3950

pol 3917
last 3900
2009 3883
pmwiki 3877
really 3873
orf 3868
order 3859
01 3840
david 3839
value 3824
results 3815
classes 3810
course 3807
00 3803
x 3801
human 3790
evoke 3777
links 3766
program 3760
password 3755
network 3724
univ 3687
text 3674
find 3669
line 3665
events 3665
distributed 3651
networks 3648
1086 3643
19 3627
via 3624
geometry 3614
symp 3608
regulation 3581
9 3578
family 3536
lecture 3492
end 3492
theory 3489
o 3451
promoter 3451
l 3435
string 3427
just 3425
mitochondrial 3420
sensitive 3396
parallel 3386
model 3378
october 3377
biosynthesis 3357
ribosomal 3342
models 3320
july 3317

email 3317
small 3282
authors 3259
query 3258
message 3258
made 3255
professor 3249
expression 3202
editor 3196
community 3194
14 3193
day 3177
knowledge 3171
growth 3163
informatics 3159
acid 3150
specific 3137
center 3118
view 3111
modified 3107
ieee 3106
homolog 3103
change 3099
programming 3093
open 3064
1072 3060
notes 3043
h 3039
international 3037
abstract 3031
initiation 3022
f 3004
top 2993
need 2992
website 2991
chromatin 2981
organization 2974
personal 2968
directory 2968
fungi 2952
1077 2930
response 2926
2016 2922
structural 2918
cellular 2904
putative 2898
splicing 2895
given 2848
sets 2837
come 2821
back 2816
space 2815
high 2811

actin 2808
30 2802
similar 2793
mitotic 2782
note 2772
users 2770
talk 2764
show 2758
inviable 2748
golgi 2745
organizations 2744
input 2742
three 2739
24 2739
security 2723
profiles 2723
method 2721
planar 2706
journal 2705
august 2703
geometric 2694
spindle 2687
03 2684
part 2678
wiki 2675
quality 2670
date 2666
pm 2662
pre 2658
department 2653
login 2646
studio 2644
2007 2628
13 2627
area 2620
possible 2619
atpase 2619
html 2612
chromosome 2606
media 2604
source 2601
study 2596
include 2592
org 2586
17 2585
add 2579
undergraduate 2573
related 2560
files 2560
16 2558
us 2555
series 2554
assembly 2554

2004 2521
management 2520
comp 2514
1090 2511
description 2509
virtual 2502
machine 2502
dept 2498
architecture 2494
recent 2491
ubiquitin 2490
biogenesis 2485
bibtex 2481
dealer 2477
2003 2465
efficient 2462
methods 2461
content 2459
23 2458
histone 2456
vacuolar 2452
control 2452
available 2451
office 2449
next 2449
rrna 2444
education 2443
now 2441
including 2433
speed 2432
08 2423
help 2423
agree 2421
2002 2419
cycle 2417
learn 2413
image 2411
biological 2409
form 2408
replication 2401
hall 2400
following 2400
functions 2395
result 2394
reviewer 2388
article 2385
04 2383
general 2381
alpha 2375
v 2362
amino 2358
thanks 2357
minimum 2356

areas 2349
games 2348
prospective 2339
local 2337
non 2329
communication 2328
link 2327
tubulin 2325
privacy 2324
2005 2324
single 2324
translation 2322
product 2312
wall 2312
y 2307
award 2296
forms 2295
technique 2295
18 2279
documents 2270
better 2266
2006 2259
donald 2259
repair 2258
public 2255
index 2250
workshop 2248
beta 2244
even 2240
great 2239
want 2237
images 2236
29 2230
technologies 2227
2001 2220
capital 2217
associated 2212
visit 2209
multiple 2208
way 2204
mobile 2203
object 2202
april 2199
statistics 2199
double 2193
co 2190
case 2190
linear 2186
details 2183
02 2175
structure 2169
within 2155
staff 2155

05	2143
plan	2139
must	2139
1089	2137
member	2133
degrees	2123
policy	2123
07	2113
poster	2100
november	2097
tree	2096
ca	2095
arxiv	2085
added	2084
php	2080
place	2076
1080	2075
mutants	2074
09	2072
1995	2070
often	2068
phd	2066
ph	2060
report	2058
28	2049
foundation	2045
1997	2044
friends	2044
seminar	2039
log	2032
techniques	2026
sciences	2025
several	2017
2000	2009

6. What are the 20 most common 3-grams?

Here are the 20 most common 3-grams in ics.uci.edu domain:

relatedness level 0	22336
activity relatedness level	11785
relatedness level 1	10496
information computer science	9230
computer science uc	9073
experimental web server	8974
fano experimental web	8974
d eppstein school	8973
server d eppstein	8973
web server d	8973
molecular function unknown	8577
d eppstein publications	8513
eppstein publications citation	8513
publications citation database	8513

unknown relatedness level	6641
relatedness level 2	5125
relatedness level 3	5081
function unknown relatedness	4668
relatedness level 6	4522
cytoplasm relatedness level	4480