CS221 Project 2

- Varad Meru (26648958)
- Nishaanth H Reddy (14765903)

1: Crawling time:

```
Time taken ~= 6 hours
```

2: Unique Pages:

```
C.no_print_urls()
50245
```

3: Sub-domains:

```
C.no_sub_domains1()
(u'http://alderis.ics.uci.edu/', 7)
(u'http://alumni.ics.uci.edu/', 5)
(u'http://ambassadors.ics.uci.edu/', 1)
(u'http://archive.ics.uci.edu/', 7719)
(u'https://asterix.ics.uci.edu/', 13)
(u'http://asterix.ics.uci.edu/', 15)
(u'http://asterixdb.ics.uci.edu/', 25)
(u'https://asterixdb.ics.uci.edu/', 17)
(u'http://auge.ics.uci.edu/', 1)
(u'http://calendar.ics.uci.edu/', 395)
(u'https://cbcl.ics.uci.edu/', 643)
(u'http://cert.ics.uci.edu/', 28)
(u'http://cgvw.ics.uci.edu/', 76)
(u'http://cleo.ics.uci.edu/', 4)
(u'http://cml.ics.uci.edu/', 64)
(u'http://computableplant.ics.uci.edu/', 97)
(u'http://cradl.ics.uci.edu/', 33)
(u'http://dblp.ics.uci.edu/', 2)
(u'https://docs.google.com/', 1)
(u'https://drive.google.com/', 1)
(u'http://drzaius.ics.uci.edu/', 3)
(u'http://dsm.ics.uci.edu/', 2)
(u'http://duttgroup-test.ics.uci.edu/', 1)
(u'https://duttgroup.ics.uci.edu/', 15760)
(u'http://dynamo.ics.uci.edu/', 24)
(u'https://elms.ics.uci.edu/', 1)
(u'http://emj.ics.uci.edu/', 3)
(u'http://emme.ics.uci.edu/', 14)
(u'http://esl.ics.uci.edu/', 4)
(u'http://evoke.ics.uci.edu/', 97)
(u'http://fano.ics.uci.edu/', 8498)
(u'http://flamingo.ics.uci.edu/', 725)
```

```
(u'http://frost.ics.uci.edu/', 10)
(u'http://ftp.ics.uci.edu/', 11)
(u'https://gats.ics.uci.edu/', 2)
(u'https://grape.ics.uci.edu/', 456)
(u'http://graphics.ics.uci.edu/', 52)
(u'http://graphmod.ics.uci.edu/', 248)
(u'http://hana.ics.uci.edu/', 30)
(u'http://hcc.ics.uci.edu/', 3)
(u'http://hci.ics.uci.edu/', 3)
(u'http://hobbes.ics.uci.edu/', 8)
(u'http://hombao.ics.uci.edu/', 4)
(u'http://honors.ics.uci.edu/', 13)
(u'http://i-sensorium.ics.uci.edu/', 5)
(u'http://informatics-study.ics.uci.edu/', 4)
(u'https://intranet.ics.uci.edu/', 5)
(u'http://ipubmed.ics.uci.edu/', 3)
(u'http://joplin.ics.uci.edu/', 1)
(u'http://jujube.ics.uci.edu/', 4)
(u'http://kdd.ics.uci.edu/', 44)
(u'http://lolth.ics.uci.edu/', 1)
(u'http://luci.ics.uci.edu/', 12)
(u'https://mailboss.ics.uci.edu/', 1)
(u'https://mailman.ics.uci.edu/', 650)
(u'http://metaviz.ics.uci.edu/', 5)
(u'http://mlearn.ics.uci.edu/', 485)
(u'http://mondego.ics.uci.edu/', 8)
(u'http://motifmap.ics.uci.edu/', 1)
(u'https://netreg.ics.uci.edu/', 2)
(u'http://pertea.ics.uci.edu/', 1)
(u'http://ppopp2013.ics.uci.edu/', 14)
(u'https://psearch.ics.uci.edu/', 3)
(u'http://riscit.ics.uci.edu/', 3)
(u'http://sdcl.ics.uci.edu/', 84)
(u'http://se.ics.uci.edu/', 2)
(u'https://sli.ics.uci.edu/', 203)
(u'http://sli.ics.uci.edu/', 464)
(u'http://soc.ics.uci.edu/', 261)
(u'https://student-council.ics.uci.edu/', 93)
(u'https://support.ics.uci.edu/', 5)
(u'https://timesheet.ics.uci.edu/', 2)
(u'https://transformativeplay.ics.uci.edu/', 19)
(u'http://vcp.ics.uci.edu/', 1)
(u'http://vip.ics.uci.edu/', 7)
(u'http://vision.ics.uci.edu/', 156)
(u'http://w3.ics.uci.edu/', 1)
(u'https://wearablegames.ics.uci.edu/', 4)
(u'https://webmail.ics.uci.edu/', 1)
(u'http://wics.ics.uci.edu/', 1494)
(u'http://www-db.ics.uci.edu/', 21)
(u'http://www.esl.ics.uci.edu/', 1)
(u'http://www.genomics.uci.edu/', 1)
(u'http://www.ics.uci.edu/', 7692)
(u'https://www.ics.uci.edu/', 280)
```

```
(u'http://www.informatics.uci.edu/', 31)
(u'http://www.physics.uci.edu/', 35)
```

4: Page with most words:

```
C.max_page()
http://www.ics.uci.edu/~xhx/project/MotifMap/sites/M00436.html
# of words:  48629
```

5: 500 most common words:

```
C.five_hundred()
0 204912
1 141845
3 101490
03 98955
2 97971
classification 81912
data 70142
10 69512
4 64736
01 63765
wics 61356
mrna 60781
2014 58950
real 58026
14 56378
5 53622
7 52127
password 50682
tools 49279
media 48104
11 47375
new 44069
s 43705
group 43000
17 42850
uci 42635
projects 41542
15 41254
8 40499
home 39809
9 39618
time 39525
12 39032
integer 38826
publications 37706
set 37180
files 36934
20 36054
content 35668
18 34792
```

categorical 34036
login 33486
6 33302
log 33192
information 33073
23 32894
04 32888
research 32507
2013 31207
protein 30780
16 30643
d 30599
19 30418
computer 30376
ics 29401
posted 29108
22 28958
100 28638
members 28581
project 28580
file 28339
page 26870
series 26806
web 26654
26 26180
photos 25752
06 25512
students 24389
user 24183
type 24061
regression 23894
13 23031
i 22800
t 22544
clustering 22362
lab 22198
sciences 22186
site 21844
date 21699
science 21565
citation 21541
09 21062
need 21001
28 20462
root 20384
following 20305
policy 20299
eppstein 20187
07 20180
domain 20111
contact 19994
school 19976
00 19953

21 19751
dutt 19104
view 18909
like 18908
married 18870
edu 18810
wiki 18692
recent 18686
learning 18359
http 18207
2011 18163
photo 18059
weekly 17861
seminar 17851
skip 17646
multivariate 17268
modified 17227
choose 17130
meeting 17118
08 17112
enter 17060
gallery 16932
database 16923
apr 16825
cc 16825
text 16770
systems 16719
currently 16679
class 16493
cecs 16432
remember 16394
internal 16308
2012 16292
share 16287
manager 16284
authentication 16234
enabled 16206
cookies 16199
license 16135
noted 16125
username 16123
social 16072
txt 16055
logged 15881
trace 15869
licensed 15844
alike 15831
forgotten 15831
krakow 15819
credentials 15806
attribution 15799
noncommercial 15776
changesmedia 15771

managersitemap 15771
unported 15771
namespaces 15679
lunch 15677
namespace 15646
filesuploadsearch 15539
c 15358
women 15275
05 15232
irvine 15020
family 14969
using 14968
x 14855
events 14814
leave 14812
comment 14711
list 14661
farewell 14562
sequential 14236
9002013 14203
year 14037
liu 13951
development 13833
angela 13800
m 13566
theory 13466
02 13455
2002013 13355
www 13225
e 13215
use 13071
server 13007
dataset 12873
software 12839
rows 12759
undo 12613
instances 12598
course 12524
work 12519
matrix 12473
search 12425
thumbnails 12377
week 12198
1000 12100
machine 12089
event 12073
pdf 11986
j 11958
jpg3264 11857
n 11830
workshop 11741
greater 11736
0004 11664

quarter 11629
don 11578
jpg1200 11532
help 11442
uc 11427
30 11419
start 11291
jan 11158
cs 11158
people 10972
game 10955
mar 10919
jpg200 10919
session 10812
prev 10706
25 10693
com 10629
oct 10540
non 10517
pm 10365
computing 10320
29 10290
dec 10235
nov 10185
feb 10177
student 10104
engineering 10102
held 10083
g 10062
university 10055
presentations 10045
based 9975
b 9831
documentsphotospresentations 9796
2015 9594
questions 9503
jun 9397
code 9393
27 9345
24 9324
sets 9289
2008 9217
4x 9216
experimental 9160
bren 9159
center 9112
info 9069
5x 8878
chr1 8867
2010 8848
july14 8846
4o 8838
make 8801

jpg 8726

physical 8724

jul 8710

chr2 8710

r 8648

network 8638

attributes 8590

fano 8528

sep 8512

1999 8482

task 8467

factor 8445

posts 8440

area 8437

aug 8419

provided 8305

april 8209

rights 8156

news 8132

24482014 8013

5o 8009

univariate 7974

version 7956

attribute 7922

available 7921

index 7886

used 7861

hypothetical 7853

author 7841

number 7743

collaboration 7735

classes 7708

digital 7706

xie 7692

fun 7692

open 7686

upload 7640

python 7636

future 7569

types 7568

kb 7560

algorithms 7545

format 7498

design 7498

repository 7473

program 7431

2009 7422

came 7410

general 7409

sorry 7400

50 7375

3x 7324

receptor 7303

life 7217
org 7209
anamakis 7149
paper 7104
supported 7064
default 7057
western 6966
3o 6935
read 6927
chr3 6921
android 6863
53 6837
programming 6824
y 6819
february 6781
member 6749
problem 6748
free 6739
thursday 6712
business 6667
java 6666
proceedings 6643
tour 6623
method 6612
bathen 6574
intelligent 6565
p 6552
example 6523
1990 6515
working 6492
k 6454
day 6453
study 6450
app 6434
chr5 6343
transcript 6296
different 6261
homolog 6234
google 6197
pages 6185
image 6184
recognition 6163
attendees 6163
cml 6153
numerical 6135
donation 6135
really 6129
company 6125
circuits 6120
got 6119
donate 6104
binding 6085
food 6047

hosted 6046
technology 6039
mixed 6030
containing 6022
directory 5995
tech 5952
contains 5938
monday 5928
academic 5906
panel 5891
conference 5890
f 5876
113 5849
l 5843
email 5777
witi 5770
experience 5766
methods 5718
browse 5710
large 5692
end 5687
reading 5671
related 5661
galleryuser 5629
documentsphotos 5629
interview 5619
setstable 5616
international 5580
jan28 5569
2006 5541
1998 5541
test 5533
model 5518
small 5502
chr6 5479
california 5406
talk 5392
listing 5388
jpg221 5342
2402013 5342
png120 5342
acm 5318
analysis 5307
just 5268
given 5247
written 5221
variant 5141
1988 5126
line 5123
meetings 5104
attributesyear 5052
taskattribute 5052
viewnamedata 5052

```
typesdefault 5052
chr11 5036
winter 5034
proc 5018
2002 5001
computational 4987
lecture 4978
field 4955
interviews 4948
o 4942
join 4908
language 4890
chr7 4851
chr12 4839
1994 4835
algorithm 4833
h 4828
littlebits 4827
6x 4823
documents 4812
chr4 4780
groups 4778
2007 4765
applications 4733
chrx 4729
activity 4729
chr10 4718
order 4711
80 4692
value 4669
gene 4667
networks 4658
speakers 4648
int 4646
spring 4627
associated 4617
ny 4613
slides 4599
jpg2592 4593
19362013 4593
knowledge 4580
games 4563
human 4549
problems 4534
added 4532
best 4528
40 4508
2000 4498
way 4464
1995 4447
6o 4432
31 4431
kinase 4428
```

```
informatics 4413
brownie 4391
function 4373
sensor 4368
troops 4362
chr9 4349
box 4347
000 4327
chenyinh 4323
title 4306
access 4305
values 4294
context 4286
department 4280
drosophila 4269
chr8 4242
change 4236
public 4231
opportunity 4228
1940 4227
end_lunch_outing 4216
54 4212
teams 4196
world 4193
1992 4188
good 4182
night 4174
ago 4170
notes 4163
features 4116
1996 4113
multiple 4103
david 4095
viewedithistory 4094
practice 4092
download 4079
```

6: 500 most common 2-grams:

```
C.twenty()
03 01 40777
01 23 24544
time series 21947
classification categorical 20060
classification real 19073
d eppstein 18316
integer real 17783
computer science 17323
home group 16572
01 22 16348
3 0 16263
group members 15953
research group 15952
```

new password 15941

cecs uci 15934

username password 15920

cookies enabled 15892

password remember 15885

set new 15868

projects publications 15847