PROJECT 2:   81646751(XI SUN) / 36088298(HONG CHANG)

Our Filter Rule of URL:

```
if(!mystring.matches("http://.*\\.ics\\.uci\\.edu.*"))
                    return false;
if(mystring.matches(".*\\.java"))
                    return false;
if(mystring.matches(".*ftp\\.ics\\.uci\\.edu.*"))
                    return false;
if(mystring.matches(".*seraja\\.ics\\.uci\\.edu.*"))
                    return false;
if(mystring.matches(".*fano\\.ics\\.uci\\.edu.*"))
                    return false;
if(mystring.contains("?"))
                    return false;
if(mystring.matches(".*edu/.*\\.(html|htm|php|jsp)"))
                    return true;
if(mystring.matches(".*edu/[^\\.]*"))
                    return true;
```

Q1. How much time did it take to crawl the entire domain?

It took 2h43min19sec (9799 seconds) to crawl the entire domain after applying above filter rules.

Q2. How many unique pages did you find in the entire domain? (Uniqueness is established by the URL)

NOTE: This does not contain some pages that we have used our filter scheme to eliminate.\

There are unique 27912 pages in entire domain.

Q3. How many subdomains did you find?

NOTE: This does not contain some subdomains that we have used our filter scheme to eliminate. For example, we do not consider about ftp.ics.uci.edu. We find www.ics.uci.edu is the largest domain, which contains 23273 different pages.

There are 66 subdomains in entire domain. (URL, # unique pages)

alderis.ics.uci.edu,7
archive.ics.uci.edu,1182
asterix.ics.uci.edu,9
auge.ics.uci.edu,16
awareness.ics.uci.edu,500
calendar.ics.uci.edu,2
cert.ics.uci.edu,30
cgvw.ics.uci.edu,2
chime.ics.uci.edu,1
cleo.ics.uci.edu,5
cml.ics.uci.edu,1
computableplant.ics.uci.edu,19
cradl.ics.uci.edu,4
dblp.ics.uci.edu,2
deepthought.ics.uci.edu,6
drzaius.ics.uci.edu,6
duttgroup.ics.uci.edu,1
emme.ics.uci.edu,15
esl.ics.uci.edu,6
evoke.ics.uci.edu,50
flamingo.ics.uci.edu,11
fr.ics.uci.edu,10
frost.ics.uci.edu,59
galen.ics.uci.edu,100
graphics.ics.uci.edu,63
graphmod.ics.uci.edu,162
hana.ics.uci.edu,15
hcc.ics.uci.edu,3
hci.ics.uci.edu,1
hobbes.ics.uci.edu,7
hombao.ics.uci.edu,4
honors.ics.uci.edu,21
i-sensorium.ics.uci.edu,6
ipubmed.ics.uci.edu,4
isg.ics.uci.edu,11
jujube.ics.uci.edu,7
kdd.ics.uci.edu,101
luci.ics.uci.edu,166
metaviz.ics.uci.edu,8
mlearn.ics.uci.edu,258
mondego.ics.uci.edu,5
motifmap.ics.uci.edu,1
ngs.ics.uci.edu,14
phoenix.ics.uci.edu,19
ppopp2013.ics.uci.edu,14
psearch.ics.uci.edu,4
sami.ics.uci.edu,9
sconce.ics.uci.edu,16
sdcl.ics.uci.edu,118
sherlock.ics.uci.edu,5
sli.ics.uci.edu,91
snekker.ics.uci.edu,1
soc.ics.uci.edu,11
sourcerer.ics.uci.edu,10

sprout.ics.uci.edu,37
student-council.ics.uci.edu,55
tastier.ics.uci.edu,1
testlab.ics.uci.edu,9
vcp.ics.uci.edu,1190
vip.ics.uci.edu,9
vision.ics.uci.edu,127
w3.ics.uci.edu,1
wics.ics.uci.edu,2
www-db.ics.uci.edu,3
www.ics.uci.edu,23273
xtune.ics.uci.edu,6

Q4. What is the longest page?

We detect that http://www.ics.uci.edu/~xhx/project/MotifMap/SNP/motif_sites_overlap_db_snp.list.html is longest page.

Q5. What are the 500 most common words in this domain?

NOTE:

1) According to the paper *word length, sentence length and frequency- ZIPF revised*, 99% of lexical words has length <15 character. We believe in practical search that input has very long length can be ignored, so in our tokenized file, we omit the word that has length > 15.
2) We have implemented our method that can choose including / not including numerical words for our statistics work. Because we find if we do not filter numbers, most of top words are biological terms or DNA/RNA sequence that containing numbers. Since this question asked us to output words, so we assume we only need to output alphabetical words, however, you can switch the method for further search engine projects in GenerateTokenFile.java
3) We use stopwordlist from http://www.ranks.nl/resources/stopwords.html (long version)
4) According to our observation, top words are related to biology field, student services and classes.

Top 500 frequent words list:

NOTE: The frequency rank list from high to low as following sequence:
left column> right column, then top row> bottom row, for example, mrna has frequency> protein, protein has frequency > data, which in second column.

mrna
protein
hypothetical
domain
factor
receptor
family
transcript
homolog
member
variant
binding
drosophila
kinase
open
chromosome
reading
frame
chrx
type
finger
und
zinc
transcription
alpha
class
beta
channel
box
gene
nuclear
carrier
data
solute
subfamily
growth
ru
will
associated
repeat
polypeptide
group
phosphatase
potassium
uci
subunit
cell
transmembrane
married
antigen
ics
oncogene
homeobox
interacting

krakow
computer
membrane
rna
inhibitor
site
rich
image
syndrome
region
software
ii
leucine
set
size
ras
lim
java
technical
cadherin
dehydrogenase
tyrosine
calcium
method
original
details
gamma
time
mitochondrial
enhancer
ribosomal
sequence
iso
acid
dna
systems
viral
glutamate
superfamily
voltage-gated
activating
forkhead
integration
sex
tumor
course
program
science
work
regulatory
neuronal
jan
b-cell
motif
leukemia
web
dec
determining
system
eos
canon
multiple
equiv
serine
specific
sry
-box
suppressor
university
public
http
mouse
jun
file
code
nov
oct
yeast
number
mar
protein-coupled
transporter
homeo
project
cerevisiae
feb
aug
complex
apr
domains
elegans
school
regulator
cancer
bren
element
design
large
molecule
students
hormone
server
sep
user
irvine
ankyrin
list
precursor
version
view
guanine
problem
lymphoma
immunoglobulin
nucleotide
org
sodium
early
rho
jul
ring
fibroblast
protease
gtpase
string
interleukin
activator
synthase
protocadherin
student
adhesion
ligand
activated
david
actin
collagen
sulfate
response
algorithms
oxidase
matrix
computing
delta
atpase
basic
listing
methods
polymerase
sema
transporting
tree
object
deleted
t-cell
crw
learning
classes
small
help
source
lab
enzyme

pou
cll
f-box
paper
activin
cytochrome
ionotropic
homology
frames
example
leucine-rich
myeloid
search
eppstein
similarity
contact
int
split
field
component
lang
sara
iii
expressed
pleckstrin
virus
engineering
synthetase
support
neural
phospholipase
graph
signal
induced
yancees
myosin
btb
values
kruppel-like
ecotropic
conference
lecture
wingless-type
mmtv
process
development
june
assignment
acm
protein-like
light
model
sorting
metalloprotease
programming
people
heparan
static
candidate
-like
pdf
isr
resource
application
processing
california
paired
cycle
subject
single
find
cytoplasmic
necrosis
three
ubiquitin
bone
histone
problems
phd
saturday
access
including
reductase
product
fibronectin
phosphoprotein
email
algorithm
division
international
differentiation
slide
network
cation
general
case
homeodomain
based
contactin
usm
returns
interface
final
endothelial
applications
points
latrophilin
write
ataxin
morphogenetic
package
short
point
retinoic
transforming
test
high
order
week
neurexin
de
slit
well
cs
machine
interactive
dual
glypican
analysis
pdz
collection
brain
state
t-box
synaptotagmin
cysteine
translation
department
technology
prev
g-protein
context
html
orphan
muscle
previous
message
human
acidic
distributed
informatics
july
transducin-like
banns
function
double
wd
translocation
management
john
personal
text

october
required
read
networks
questions
requirements
exchanger
carcinoma
link
create
sub-family
eukaryotic
files
thyroid
vision
threonine
rescue
disease
description
iroquois
note
tz
graduate
low
long
start
ets
ieee
breast
plasma
form
click
bruno-like
projects
specificity
abraham
resources
itr
copy
uc
anion
langsam
ps
dataguard
br
directory
repeats
integrin
glucosamine
zipper
document
theory
password
exchange
special
lead
change
cassette
cyclin
initiation
current
signalling
dominant
chain
models
return
chemokine
putative
collaboration
fragile
kh
control
programs
atp-binding
reference
apoptosis
embryonic
cyclase
papers
death
office
inducible
tm
good
space
request
proteasome
son
abstract
riken
sets
working
smad
april
protocol
cdna
dead
internet
march
max
security
variable
nexin
co-repressor
avian
second
policy
heavy
gef
glycoprotein
include
it's
database
rar-related
elongation
olfactory
interferon
pr
ny
insulin-like
notes

Q6. What are the 20 most common 2-grams?

Top 20 2-gram list

NOTE: The frequency rank list from high to low as following sequence: left column> right column, then top row> bottom row.

protein mrna
hypothetical protein
transcript variant
mrna hypothetical
reading frame
open reading
chromosome open
variant mrna
binding protein
frame mrna
mrna protein
mrna chrx
homolog drosophila
zinc finger
transcription factor

domain mrna	finger protein	mrna	chromosome
family member	drosophila mrna		

Q7. (extra problem) How many unique pages excluding similar contents?

There are 9387 unique pages(~ 59.7% of original) excluding similar contents in the whole domain.

NOTE: We used shingle algorithm to compute the similarity between different pages. Because of scale of the executing time, we chose 0.5 as our resemblance threshold and containment threshold. We did the computing and compressing at the same time, so we can do accomplish this task in several hours. (details could be discussed during face-to-face talk)