Data Wrangle in MongoDB

- 1. Query data in MongoDB
- 2. Extract data array for building model
- 3. Visualize data using Matplotlib
- 4. Variable importance, correlation matrix or mutual information?
- 5. Extract into dataframe format? mongoDB pipe to Pandas/Graphlab?

Explore Mongo data, produced from extract.ipynb.

1. Query data in MongoDB

In [2]:
```
import pymongo

#
# Connect to MongoDB server
#
def get_db(host="localhost", port=27017, username=None, password=None,
 db="dataLogs"):
    """ A util for making a connection to mongo """
    from pymongo import MongoClient

    if username and password:
        mongo_uri = 'mongodb://%s:%s@%s:%s/%s' % (username, password,
host, port, db)
        conn = MongoClient(mongo_uri)
    else:
        conn = MongoClient(host, port)

    return conn[db]


def get_coll(host="localhost", port=27017, username=None, password=Non
e, db="dataLogs",collection="logs"):
    db=get_db(host, port, username, password, db)
    return db[collection]
'''
# database
db = get_db('localhost', 27017, None, None,"dataLogs")
# Collection
logs= db.logs
test= db.test
'''
logs = get_coll('localhost', 27017, None, None,"dataLogs", "logs")
test = get_coll('localhost', 27017, None, None,"dataLogs", "test")
```

In [3]:
```
#
# peek one record
#
logs.find_one()
```

Out[3]:
```
{u'Accept-Encoding': u'gzip,deflate,sdch',
 u'Accept-Language': u'en-US,en;q=0.8,ms;q=0.6,id;q=0.4',
 u'Cache-Control': u'max-age=0, private, must-revalidate',
 u'Connection': u'keep-alive',
 u'Content-Length': u'3726',
 u'Content-Type': u'text/html; charset=utf-8',
 u'Date': u'Tue, 14 Oct 2014 00',
 u'Etag': u'"ea927ca92e0c7a85b168d210be0a5df4"',
 u'Num-Cookie': 2,
 u'Server': u'WEBrick/1.3.1 (Ruby/2.1.2/2014-05-08)',
 u'Set-Cookie': [[{u'request_method': u'GET'}, {u' path': u'/'}],
  [{u'_sample_app_session': u'NlljR05mSytQNzJ4N0gxQzA1VkJBQ3lDOUpEOWpw
WFJnaGVOaUdFU0ZDek82RE5VK0I5V2NpY0RrRFcrSjZhT21oOTRZQW5UUzBhTDlIMUJ1Qj
NnZUZva3VCSEUzZ2FZeWsxMU5xcnE4S093aGpwaVBRbUUxeTFkSENINGFsdzMzUER6RVhH
MkVyVCtmUTdzeS91S2lVc3hPU2V4Y1VOKy9GMzg0QUNNZFFzMFBUdWVUY1FvNmdiTnVRUG
tQbzlULS1jWGEzRS9VcEtoWDc4bWtBTlNSajN3PT0%3D--a87e267085a6649f66c98a67
1fa8dac9629f7bd9'},
   {u' path': u'/'},
   {u' HttpOnly': None}]],
 u'X-Content-Type-Options': u'nosniff',
 u'X-Frame-Options': u'SAMEORIGIN',
 u'X-Request-Id': u'1c96e7f2-2bd3-4070-b2b9-73ad728c111f',
 u'X-Runtime': u'0.017533',
 u'X-Ua-Compatible': u'chrome=1',
 u'X-Xhr-Current-Location': u'/',
 u'X-Xss-Protection': u'1; mode=block',
 u'_id': ObjectId('54c98ee07765b4344c49f5e7'),
 u'address': u'50.59.22.130,5',
 u'cert': u'',
 u'code': u'200',
 u'content': u'0',
 u'error': u'0',
 u'headers': u'399',
 u'host': u'54.165.254.99',
 u'http': u'path',
 u'httpversion': u'1#1',
 u'ip': u'54.165.254.99',
 u'method': u'GET',
 u'msg': u'OK',
 u'path': u'/',
 u'port': u'80',
 u'request': u'789',
 u'requestcount': u'1',
 u'response': u'4930',
 u'scheme': u'http',
 u'timestamp_end': u'1413247021.055149',
 u'timestamp_start': u'1413247021.052226'}
```

In [117]:
```
#
# Extract all features in the first event

#
features=logs.find_one().keys()  # list of features
features.remove("_id")  # _id is useless for this study
import unicodedata
features = [x.encode("UTF8") for x in features]
features
```

Out[117]:
```
['Content-Length',
 'code',
 'Accept-Language',
 'X-Xhr-Current-Location',
 'X-Request-Id',
 'Etag',
 'requestcount',
 'X-Ua-Compatible',
 'X-Frame-Options',
 'port',
 'content',
 'X-Runtime',
 'ip',
 'Date',
 'Num-Cookie',
 'scheme',
 'method',
 'X-Xss-Protection',
 'http',
 'Accept-Encoding',
 'Set-Cookie',
 'Cache-Control',
 'Server',
 'msg',
 'host',
 'address',
 'path',
 'response',
 'timestamp_end',
 'X-Content-Type-Options',
 'request',
 'Connection',
 'headers',
 'cert',
 'error',
 'timestamp_start',
 'Content-Type',
 'httpversion']
```

In [118]:
```
# Simple query to peek Content-Type variable
#
query ={"Content-Type":{'$exists':True}}
iter=logs.find(query,{"_id":0,"Content-Type":1},limit=10)
for item in iter:
    print item
```

```
{u'Content-Type': u'text/html; charset=utf-8'}
{u'Content-Type': u'text/css'}
{u'Content-Type': u'application/javascript'}
{u'Content-Type': u'text/css'}
{u'Content-Type': u'text/css'}
{u'Content-Type': u'text/css'}
{u'Content-Type': u'text/css'}
{u'Content-Type': u'application/javascript'}
{u'Content-Type': u'application/javascript'}
{u'Content-Type': u'application/javascript'}
```

In [119]:
```
#find distinct values in field

features1=["error","response","headers","content","request",
           "X-Runtime","Date",
           "Content-Length""timestamp_end","timestamp_start", "Last-Mo
dified",
           "cert","code","requestcount"]

features2 =["X-Frame-Options","X-Xss-Protection","X-Content-Type-Optio
ns","X-Ua-Compatible",
            "X-Xhr-Current-Location",
            "Content-Type","Cache-Control",
            "Server","httpversion","port","scheme","http","path","host
", "Connection",
            "ip","Referer","code",
            "Accept-Encoding","Accept-Language",
            "method",
            "msg",
            "address"]

features3=["Etag","X-Request-Id"]

for q in features2:
    print q, "has the below distinct values:"
    print logs.distinct(q)
    print "\n\n"
```

```
X-Frame-Options has the below distinct values:
[u'SAMEORIGIN']



X-Xss-Protection has the below distinct values:
[u'1; mode=block']
```

X-Content-Type-Options has the below distinct values:
[u'nosniff']

X-Ua-Compatible has the below distinct values:
[u'chrome=1']

X-Xhr-Current-Location has the below distinct values:
[u'/', u'/signin', u'/sessions', u'/users/1', u'/users', u'/users/2',
u'/users/3', u'/users/3/followers', u'/about', u'/contact', u'/users/1
/edit']

Content-Type has the below distinct values:
[u'text/html; charset=utf-8', u'text/css', u'application/javascript',
u'image/vnd.microsoft.icon', u'application/x-www-form-urlencoded']

Cache-Control has the below distinct values:
[u'max-age=0, private, must-revalidate', u'public, must-revalidate', u
'max-age=0', u'no-cache']

Server has the below distinct values:
[u'WEBrick/1.3.1 (Ruby/2.1.2/2014-05-08)']

httpversion has the below distinct values:
[u'1#1']

port has the below distinct values:
[u'80']

scheme has the below distinct values:
[u'http']

http has the below distinct values:
[u'path']

path has the below distinct values:

[u'/', u'/assets/application.css?body=1', u'/assets/jquery_ujs.js?body
=1', u'/assets/static_pages.css?body=1', u'/assets/sessions.css?body=1
', u'/assets/users.css?body=1', u'/assets/custom.css?body=1', u'/asset
s/jquery.js?body=1', u'/assets/turbolinks.js?body=1', u'/assets/bootst
rap-affix.js?body=1', u'/assets/bootstrap-transition.js?body=1', u'/as
sets/bootstrap-alert.js?body=1', u'/assets/bootstrap-button.js?body=1'
, u'/assets/bootstrap-carousel.js?body=1', u'/assets/bootstrap-collaps
e.js?body=1', u'/assets/bootstrap-scrollspy.js?body=1', u'/assets/boot
strap-modal.js?body=1', u'/assets/bootstrap-dropdown.js?body=1', u'/as
sets/bootstrap-tooltip.js?body=1', u'/assets/bootstrap-tab.js?body=1',
 u'/assets/bootstrap.js?body=1', u'/assets/bootstrap-popover.js?body=1
', u'/assets/sessions.js?body=1', u'/assets/bootstrap-typeahead.js?bod
y=1', u'/assets/static_pages.js?body=1', u'/assets/users.js?body=1', u
'/assets/application.js?body=1', u'/favicon.ico', u'/signin', u'/sessi
ons', u'/users/1', u'/users', u'/users/2', u'/users/3', u'/users/3/fol
lowers', u'/about', u'/contact', u'/users/1/edit']

host has the below distinct values:
[u'54.165.254.99']

Connection has the below distinct values:
[u'keep-alive']

ip has the below distinct values:
[u'54.165.254.99']

Referer has the below distinct values:
[u'http']

code has the below distinct values:
[u'200', u'304', u'302']

Accept-Encoding has the below distinct values:
[u'gzip,deflate,sdch', u'gzip,deflate']

Accept-Language has the below distinct values:
[u'en-US,en;q=0.8,ms;q=0.6,id;q=0.4']

method has the below distinct values:
[u'GET', u'POST']

```
msg has the below distinct values:
[u'OK', u'Not Modified', u'Found']



address has the below distinct values:
[u'50.59.22.130,5']
```

In [120]:
```
n=1
import sys

features11=["error","response","headers","content","Content-Length","r
equest"]
features12=["X-Runtime","Date",
           "cert","code","request","requestcount"]

print features11
try:
    iter = logs.find({},limit=10, skip=n)  #.sort('Content-Length', di
rection=1)
    for item in iter:
        print [item.get(f) for f in features11] #,item.get("content")
#,  item["Content-Length"] == item["Content"]
except:
    print "Error trying to read collection:", sys.exc_info()[0]
```

```
['error', 'response', 'headers', 'content', 'Content-Length', 'request
']
[u'0', u'1108', u'778', u'0', u'513', u'1179']
[u'0', u'16091', u'762', u'0', u'15477', u'1162']
[u'0', u'591', u'778', u'0', u'0', u'1198']
[u'0', u'591', u'778', u'0', u'0', u'1195']
[u'0', u'591', u'778', u'0', u'0', u'1190']
[u'0', u'134336', u'778', u'0', u'133735', u'1192']
[u'0', u'273816', u'762', u'0', u'273200', u'1176']
[u'0', u'13268', u'762', u'0', u'12654', u'1162']
[u'0', u'4097', u'762', u'0', u'3485', u'1166']
[u'0', u'2370', u'762', u'0', u'1758', u'1171']
```

In [120]:

In [121]:
```
#
#  feature-dependence using mongodb's group operation
#    (min, max, average)
#
def variableDependenc(field):

    minMaxR=logs.aggregate([{"$group":
                           {"_id":"$"+field,
                            "minRes":{"$min":"$response"},
                            "maxRes":{"$max":"$response"},
                            "avgRes":{"$avg":"$response"}
                            }}]
                          )
    print "\n=== %s min max avg======"%field
    for item in minMaxR['result']:
        print item["_id"],item["minRes"],item["maxRes"], item["avgRes"
]
    return minMaxR

for f in features2:
    variableDependenc(f)

for f in ["method","path"]:
    variableDependenc(f)
```

```
=== X-Frame-Options min max avg======
None 10523 8933 0.0
SAMEORIGIN 1091 7101 0.0

=== X-Xss-Protection min max avg======
None 10523 8933 0.0
1; mode=block 1091 7101 0.0

=== X-Content-Type-Options min max avg======
None 10523 8933 0.0
nosniff 1091 7101 0.0

=== X-Ua-Compatible min max avg======
None 10523 8933 0.0
chrome=1 1091 7101 0.0

=== X-Xhr-Current-Location min max avg======
/users/1/edit 1099 6552 0.0
/contact 1094 5377 0.0
/ 4930 4930 0.0
/users/3/followers 1104 6719 0.0
None 10523 8933 0.0
/users/3 1094 6254 0.0
/users/2 1094 6248 0.0
/about 1091 5501 0.0
/users/1 1094 7101 0.0
/users 1091 6134 0.0
/sessions 1388 1388 0.0
```

/signin 1092 5430 0.0

=== Content-Type min max avg======
application/x-www-form-urlencoded 1388 1388 0.0
None 1091 426 0.0
image/vnd.microsoft.icon 424 424 0.0
text/css 1108 591 0.0
application/javascript 10523 8933 0.0
text/html; charset=utf-8 4930 7101 0.0

=== Cache-Control min max avg======
max-age=0 1091 7101 0.0
None 424 424 0.0
public, must-revalidate 10523 8933 0.0
no-cache 423 426 0.0
max-age=0, private, must-revalidate 1091 6719 0.0

=== Server min max avg======
WEBrick/1.3.1 (Ruby/2.1.2/2014-05-08) 10523 8933 0.0

=== httpversion min max avg======
1#1 10523 8933 0.0

=== port min max avg======
80 10523 8933 0.0

=== scheme min max avg======
http 10523 8933 0.0

=== http min max avg======
path 10523 8933 0.0

=== path min max avg======
/contact 1094 5377 0.0
/users/2 1094 6248 0.0
/users/1 1094 7101 0.0
/sessions 1388 1388 0.0
/signin 1092 5430 0.0
/users/1/edit 1099 6552 0.0
/favicon.ico 424 424 0.0
/assets/bootstrap-tab.js?body=1 4110 426 0.0
/assets/bootstrap-button.js?body=1 3454 426 0.0
/assets/bootstrap-dropdown.js?body=1 424 5026 0.0
/users/3 1094 6254 0.0
/assets/bootstrap-typeahead.js?body=1 425 8933 0.0
/assets/bootstrap.js?body=1 426 621 0.0
/users/3/followers 1104 6719 0.0
/assets/bootstrap-carousel.js?body=1 425 6671 0.0
/assets/turbolinks.js?body=1 13268 426 0.0
/assets/users.css?body=1 425 591 0.0
/assets/bootstrap-collapse.js?body=1 425 5349 0.0
/users 1091 6134 0.0
/assets/jquery.js?body=1 273816 426 0.0
/assets/bootstrap-scrollspy.js?body=1 425 5269 0.0
/assets/bootstrap-tooltip.js?body=1 10523 426 0.0
/assets/custom.css?body=1 134336 426 0.0

/assets/custom.css?body=1 154990 426 0.0
/assets/bootstrap-modal.js?body=1 421 7269 0.0
/assets/jquery_ujs.js?body=1 16091 426 0.0
/assets/bootstrap-transition.js?body=1 2370 426 0.0
/assets/sessions.js?body=1 424 638 0.0
/assets/bootstrap-affix.js?body=1 4097 426 0.0
/assets/bootstrap-alert.js?body=1 3138 426 0.0
/assets/users.js?body=1 424 639 0.0
/ 4930 4930 0.0
/assets/application.css?body=1 1108 426 0.0
/assets/bootstrap-popover.js?body=1 3728 426 0.0
/assets/sessions.css?body=1 424 591 0.0
/assets/static_pages.js?body=1 425 639 0.0
/assets/static_pages.css?body=1 425 591 0.0
/about 1091 5501 0.0
/assets/application.js?body=1 1192 426 0.0

=== host min max avg======
54.165.254.99 10523 8933 0.0

=== Connection min max avg======
keep-alive 10523 8933 0.0

=== ip min max avg======
54.165.254.99 10523 8933 0.0

=== Referer min max avg======
http 10523 8933 0.0
None 424 4930 0.0

=== code min max avg======
302 1388 1388 0.0
304 1091 426 0.0
200 10523 8933 0.0

=== Accept-Encoding min max avg======
gzip,deflate 1388 1388 0.0
gzip,deflate,sdch 10523 8933 0.0

=== Accept-Language min max avg======
en-US,en;q=0.8,ms;q=0.6,id;q=0.4 10523 8933 0.0

=== method min max avg======
POST 1388 1388 0.0
GET 10523 8933 0.0

=== msg min max avg======
Not Modified 1091 426 0.0
Found 1388 1388 0.0
OK 10523 8933 0.0

=== address min max avg======
50.59.22.130,5 10523 8933 0.0

=== method min max avg======
POST 1388 1388 0.0
GET 10523 8933 0.0

```
GET 10523 6935 0.0

=== path min max avg======
/contact 1094 5377 0.0
/users/2 1094 6248 0.0
/users/1 1094 7101 0.0
/sessions 1388 1388 0.0
/signin 1092 5430 0.0
/users/1/edit 1099 6552 0.0
/favicon.ico 424 424 0.0
/assets/bootstrap-tab.js?body=1 4110 426 0.0
/assets/bootstrap-button.js?body=1 3454 426 0.0
/assets/bootstrap-dropdown.js?body=1 424 5026 0.0
/users/3 1094 6254 0.0
/assets/bootstrap-typeahead.js?body=1 425 8933 0.0
/assets/bootstrap.js?body=1 426 621 0.0
/users/3/followers 1104 6719 0.0
/assets/bootstrap-carousel.js?body=1 425 6671 0.0
/assets/turbolinks.js?body=1 13268 426 0.0
/assets/users.css?body=1 425 591 0.0
/assets/bootstrap-collapse.js?body=1 425 5349 0.0
/users 1091 6134 0.0
/assets/jquery.js?body=1 273816 426 0.0
/assets/bootstrap-scrollspy.js?body=1 425 5269 0.0
/assets/bootstrap-tooltip.js?body=1 10523 426 0.0
/assets/custom.css?body=1 134336 426 0.0
/assets/bootstrap-modal.js?body=1 421 7269 0.0
/assets/jquery_ujs.js?body=1 16091 426 0.0
/assets/bootstrap-transition.js?body=1 2370 426 0.0
/assets/sessions.js?body=1 424 638 0.0
/assets/bootstrap-affix.js?body=1 4097 426 0.0
/assets/bootstrap-alert.js?body=1 3138 426 0.0
/assets/users.js?body=1 424 639 0.0
/ 4930 4930 0.0
/assets/application.css?body=1 1108 426 0.0
/assets/bootstrap-popover.js?body=1 3728 426 0.0
/assets/sessions.css?body=1 424 591 0.0
/assets/static_pages.js?body=1 425 639 0.0
/assets/static_pages.css?body=1 425 591 0.0
/about 1091 5501 0.0
/assets/application.js?body=1 1192 426 0.0
```

In [121]:

In [121]:

2 Check Content-Length, content, headers and response values

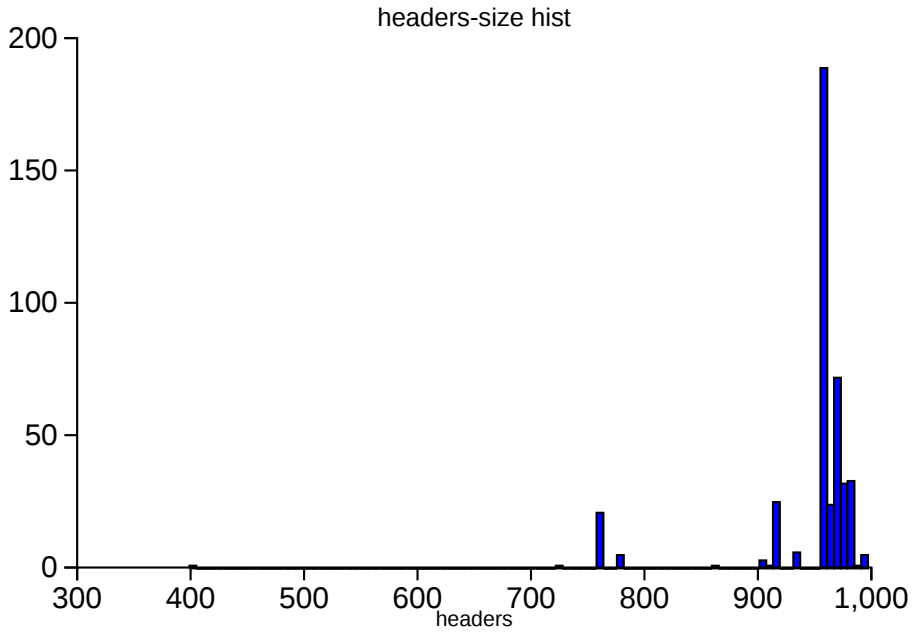
There are several features related to the request size: Content-Length, content, headers and response.

In [122]:
```
#
#  Check whether Content-Length == content is always true
#
iter = logs.find({})  #.sort('Content-Length', direction=1)
ntrue =0
nfalse =0
nprint =10
nrecord =0
for item in iter:
    nrecord +=1
    if item.get("Content-Length"):
        if int((item.get("Content-Length"))) != int(item.get("content"
)):
            nfalse +=1
            if nprint>0:
                print item.get("Content-Length"), item.get("content")
                nprint -=1
        else:
            ntrue +=1

print "Number of records : ",nrecord
print "Number of Content-Length == content : ", ntrue
print "Number of Content-Length != content : ", nfalse

#
# There is a None value in Content-Length,  and 0 for content.
# Neither seems to be a good choice
#
```

```
3726 0
513 0
15477 0
133735 0
273200 0
12654 0
3485 0
1758 0
2526 0
2843 0
Number of records :  420
Number of Content-Length == content :  5
Number of Content-Length != content :  33
```

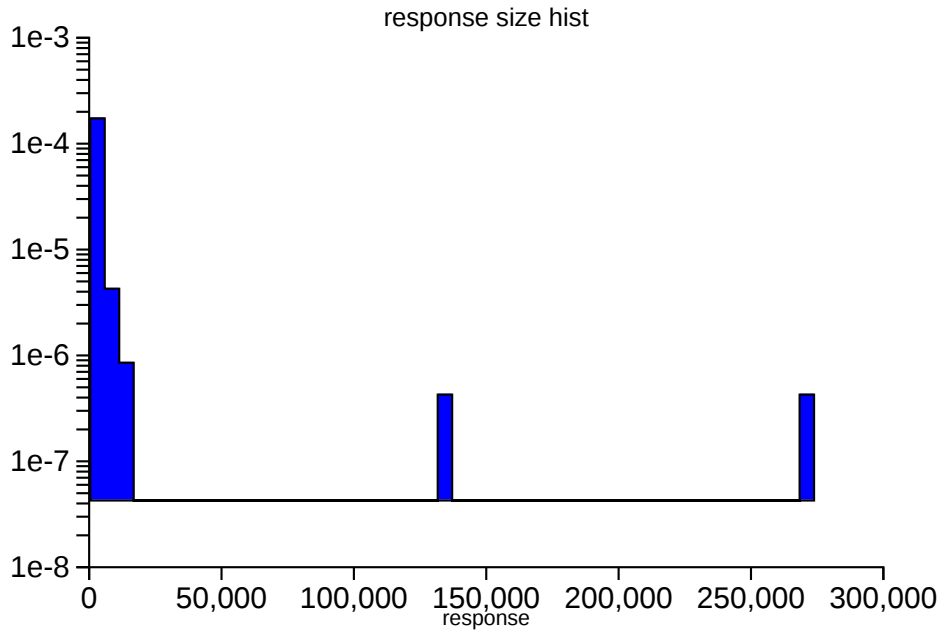
In [123]:
```
%matplotlib inline
import matplotlib.pyplot as plt
import mpld3
mpld3.enable_notebook()
```

In [124]:
```
iter=logs.find({"headers":{'$exists':True}},{"_id":0, "headers":1})
hsize =[int(item.values()[0]) for item in iter]

#plt.interactive(True)
plt.title("headers-size hist")
plt.xlabel("headers")
plt.hist(hsize,bins=100)
plt.show()
#mpld3.show
```



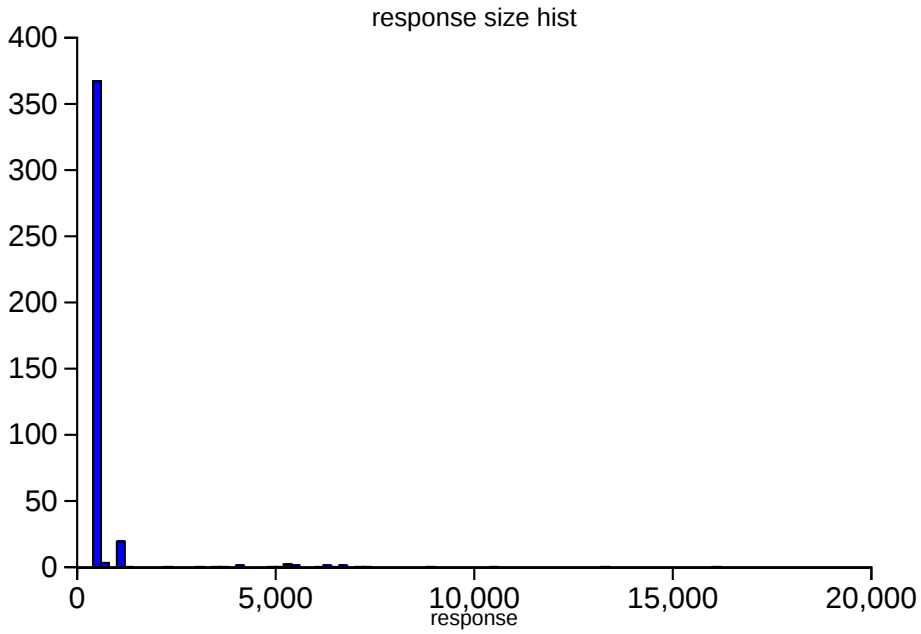
In [125]:
```
responseSize ="response"
iter=logs.find({"or":[{"response":{'$exists':False}}, {"response":0}]}
)
t= [item for item in iter]
print "%s \'response\' doenot exist or zero-value! \n "%len(t)


query ={responseSize:{'$exists':True},responseSize:{'$exists':True}}

iter=logs.find(query,{"_id":0, responseSize:1})
rsize =[int(item.values()[0]) for item in iter]


#plt.interactive(True)
plt.title("response size hist")
plt.xlabel("response")
plt.hist(rsize,bins=50, normed=1, histtype="stepfilled", log=True)
#plt.semilogy()
#plt.show()
#mpld3.show


#fig = plt.figure()
#ax = fig.add_subplot(2, 1,1)
```

```
          0 'response' doenot exist or zero-value!


Out[125]: (array([  1.76789383e-04,   4.35441830e-06,   8.70883660e-07,
                    0.00000000e+00,   0.00000000e+00,   0.00000000e+00,
                    0.00000000e+00,   0.00000000e+00,   0.00000000e+00,
                    0.00000000e+00,   0.00000000e+00,   0.00000000e+00,
                    0.00000000e+00,   0.00000000e+00,   0.00000000e+00,
                    0.00000000e+00,   0.00000000e+00,   0.00000000e+00,
                    0.00000000e+00,   0.00000000e+00,   0.00000000e+00,
                    4.35441830e-07,   0.00000000e+00,   0.00000000e+00,
                    0.00000000e+00,   0.00000000e+00,   0.00000000e+00,
                    0.00000000e+00,   0.00000000e+00,   0.00000000e+00,
                    0.00000000e+00,   0.00000000e+00,   0.00000000e+00,
                    0.00000000e+00,   0.00000000e+00,   0.00000000e+00,
                    0.00000000e+00,   0.00000000e+00,   0.00000000e+00,
                    0.00000000e+00,   0.00000000e+00,   0.00000000e+00,
                    0.00000000e+00,   4.35441830e-07]),
             array([    421. ,    5888.9,   11356.8,   16824.7,   22292.6,   27760
            .5,
                     33228.4,   38696.3,   44164.2,   49632.1,   55100. ,   60567
            .9,
                     66035.8,   71503.7,   76971.6,   82439.5,   87907.4,   93375
            .3,
                     98843.2,  104311.1,  109779. ,  115246.9,  120714.8,  126182
            .7,
                    131650.6,  137118.5,  142586.4,  148054.3,  153522.2,  158990
            .1,
                    164458. ,  169925.9,  175393.8,  180861.7,  186329.6,  191797
            .5,
                    197265.4,  202733.3,  208201.2,  213669.1,  219137. ,  224604
            .9,
                    230072.8,  235540.7,  241008.6,  246476.5,  251944.4,  257412
            .3,
                    262880.2,  268348.1,  273816. ]),
             <a list of 1 Patch objects>)
```



In [126]:
```
#
#
xlim=20000
#plt.set_xlim([0,xlim])
#plt.set_ylim([0,xlim])
plt.title("response size hist")
plt.xlabel("response")
#plt.hist(rsize,bins=100)
plt.hist(rsize,bins=100,range=[0,xlim])


# no
#
```

```
Out[126]: (array([   0.,    0.,  368.,    4.,    0.,   20.,    1.,    0.,    0.,
                     0.,    0.,    1.,    0.,    0.,    0.,    1.,    0.,    1.,
                     1.,    0.,    2.,    0.,    0.,    0.,    1.,    1.,    3.,
                     2.,    0.,    0.,    1.,    2.,    1.,    2.,    0.,    1.,
                     1.,    0.,    0.,    0.,    0.,    0.,    0.,    0.,    1.,
                     0.,    0.,    0.,    0.,    0.,    0.,    0.,    1.,    0.,
                     0.,    0.,    0.,    0.,    0.,    0.,    0.,    0.,    0.,
                     0.,    0.,    0.,    1.,    0.,    0.,    0.,    0.,    0.,
                     0.,    0.,    0.,    0.,    0.,    0.,    0.,    0.,    1.,
                     0.,    0.,    0.,    0.,    0.,    0.,    0.,    0.,    0.,
                     0.,    0.,    0.,    0.,    0.,    0.,    0.,    0.,    0.,
                0.]),
            array([    0.,    200.,    400.,    600.,    800.,   1000.,   1200.,
                    1400.,   1600.,   1800.,   2000.,   2200.,   2400.,   2600.,
                    2800.,   3000.,   3200.,   3400.,   3600.,   3800.,   4000.,
                    4200.,   4400.,   4600.,   4800.,   5000.,   5200.,   5400.,
                    5600.,   5800.,   6000.,   6200.,   6400.,   6600.,   6800.,
                    7000.,   7200.,   7400.,   7600.,   7800.,   8000.,   8200.,
                    8400.,   8600.,   8800.,   9000.,   9200.,   9400.,   9600.,
                    9800.,  10000.,  10200.,  10400.,  10600.,  10800.,  11000.,
                   11200.,  11400.,  11600.,  11800.,  12000.,  12200.,  12400.,

                   12600.,  12800.,  13000.,  13200.,  13400.,  13600.,  13800.,
                   14000.,  14200.,  14400.,  14600.,  14800.,  15000.,  15200.,
                   15400.,  15600.,  15800.,  16000.,  16200.,  16400.,  16600.,
                   16800.,  17000.,  17200.,  17400.,  17600.,  17800.,  18000.,
                   18200.,  18400.,  18600.,  18800.,  19000.,  19200.,  19400.,
                   19600.,  19800.,  20000.]),
            <a list of 100 Patch objects>)
```



In [127]:
```
# response = content + header ?
#
# response - content - header ~ 170, overhead
#
#
#
```

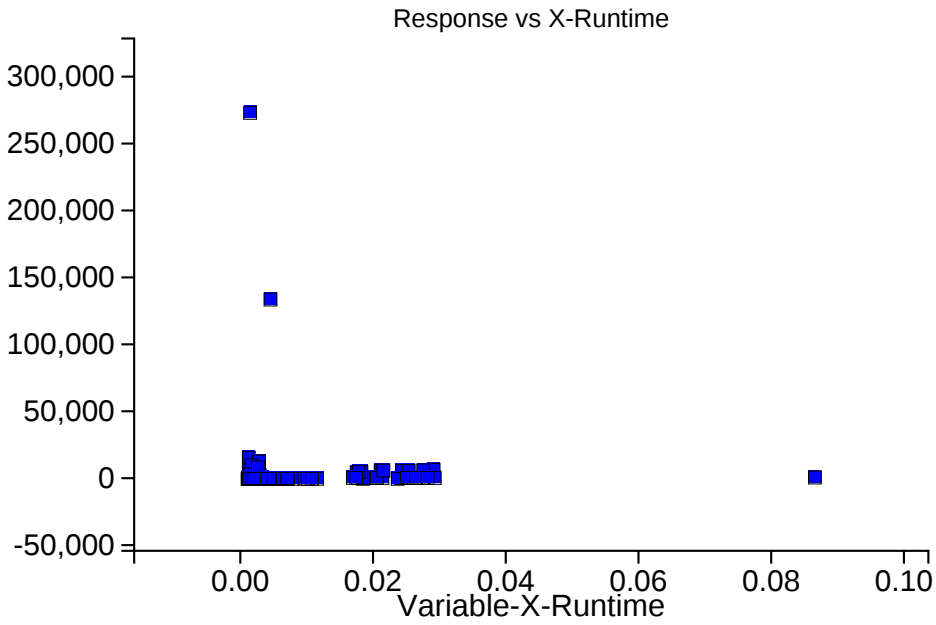
3. Feature dependence study

- "X-Runtime", "timestamp_end","timestamp_start",
- "X-Frame-Options","X-Xss-Protection","X-Content-Type-Options","X-Ua-Compatible","X-Xhr-Current-Location"
- "Content-Type","Etag","Cache-Control", "Server","Date",
- "cert","code","requestcount",
- "port","scheme","http","path","host","headers", "Connection",
- "ip","Referer"
- "Accept-Encoding","Accept-Language","method","msg"

In [128]:
```
f="X-Runtime"

def featureDependence(f):
    iter = logs.find({"response":{"$exists":True},f:{"$exists":True}},
{"_id":0,"response":1,f:1})

    x,y=[],[]
    for item in iter:
        y.append(int(item["response"]))
        x.append(float(item[f]))

    print len(x),len(y)
    title="Response vs %s"%f
    plt.figure(title)
    plt.title(title)

    plt.xlabel("Variable-%s"%f,fontsize="x-large")

    plt.plot(x,y,"bs")


    # Pad margins so that markers don't get clipped by the axes
    plt.margins(0.2)
    # Tweak spacing to prevent clipping of tick-labels
    plt.subplots_adjust(bottom=0.15)
    #
```

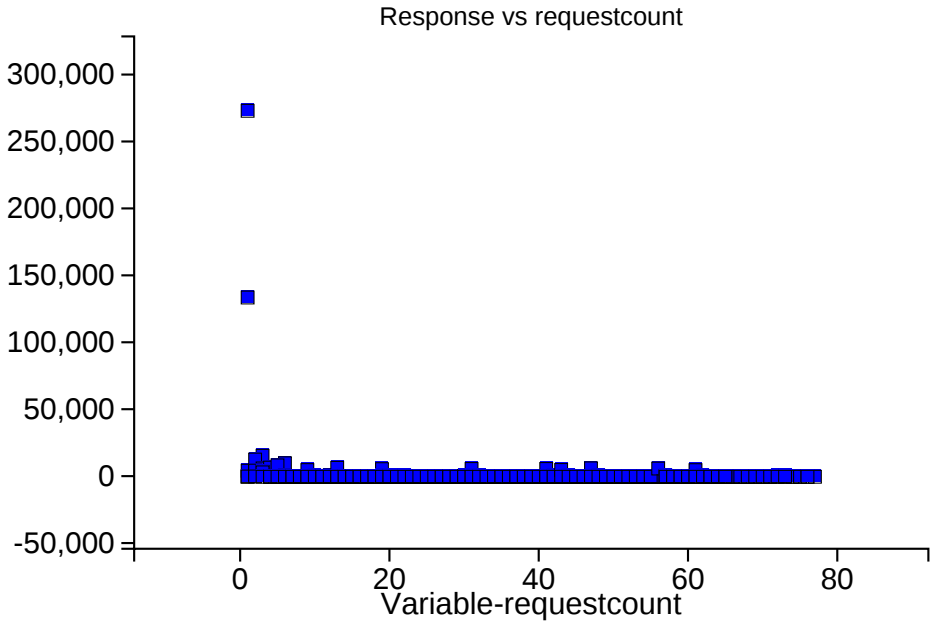
In [129]:
```
# X-Runtime

featureDependence(f="X-Runtime")
```

```
419 419
```



In [130]:
```
#requestCount
featureDependence(f="requestcount")
```

```
420 420
```

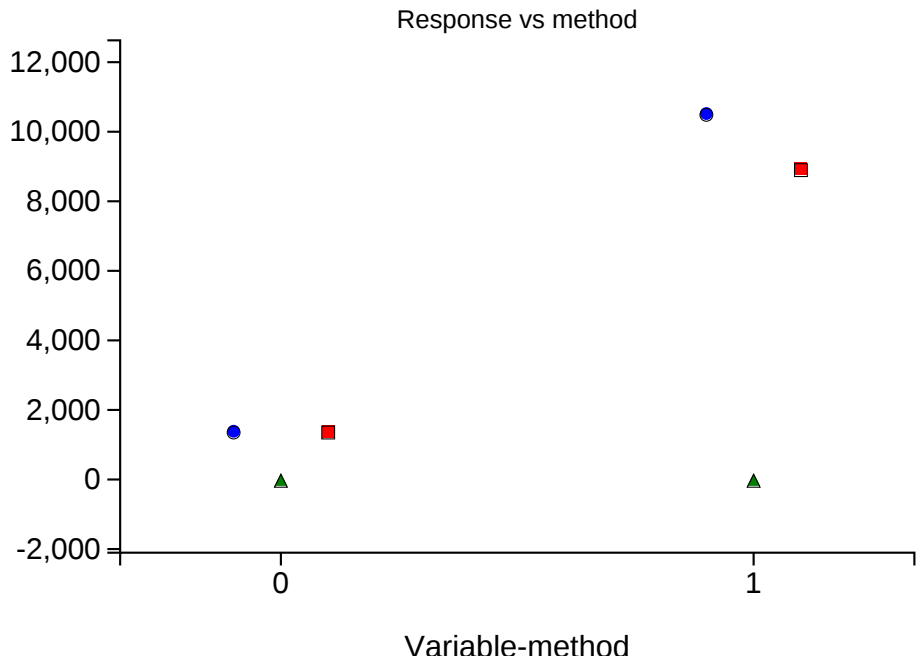


In [131]:

```
f="method"
def checkFeatureDependence(f, option=0):
    Rm = variableDependenc(f)

    ymin = [int(item.get("minRes") or 0) for item in Rm["result"]]
    ymax = [int(item.get("maxRes") or 0) for item in Rm["result"]]
    yavg = [int(item.get("avgRes") or 0) for item in Rm["result"]]

    xTicks=[str(item.get("_id")) for item in Rm["result"]]
    #    print item.get("minRes")  #.get("minRes")
    xTicks
    import numpy as np
    x= np.arange(len(xTicks))

    title="Response vs %s"%f
    plt.figure(title)
    plt.title(title)
    #ax = fig.add_subplot(111)
    #xtickNames = ax.set_xticklabels(xTicks)
    #plt.setp(xtickNames, rotation=45, fontsize=100)

    plt.xlabel("Variable-%s"%f,fontsize="x-large")
    #plt.ylabel("size")

    if option==0:
        plt.plot(x-0.1, ymin,"bo", x+0.1, ymax,"rs", x,yavg,"g^")
    elif option ==1:
        plt.plot(x,ymax,"g^")  # only show average

    plt.xticks(x, xTicks, rotation='vertical')

    # Pad margins so that markers don't get clipped by the axes
    plt.margins(0.2)
    # Tweak spacing to prevent clipping of tick-labels
    plt.subplots_adjust(bottom=0.15)
    #plt.show()
    print x,xTicks
```

In [132]:
```
# method  = "GET" or "post"
checkFeatureDependence(f="method")
```

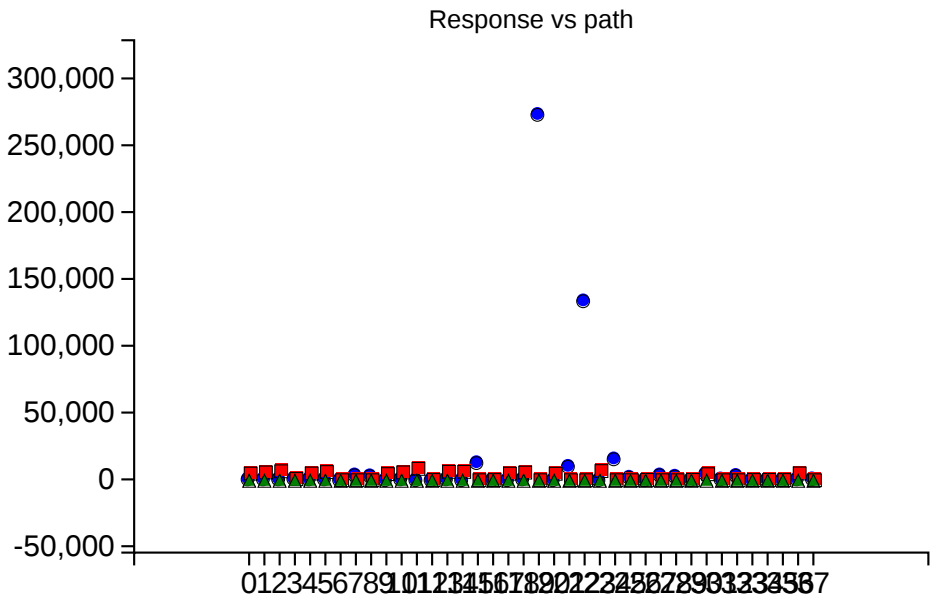
```
=== method min max avg======
POST 1388 1388 0.0
GET 10523 8933 0.0
[0 1] ['POST', 'GET']
```



In [133]:
```
# path
checkFeatureDependence(f="path", option=0)
```

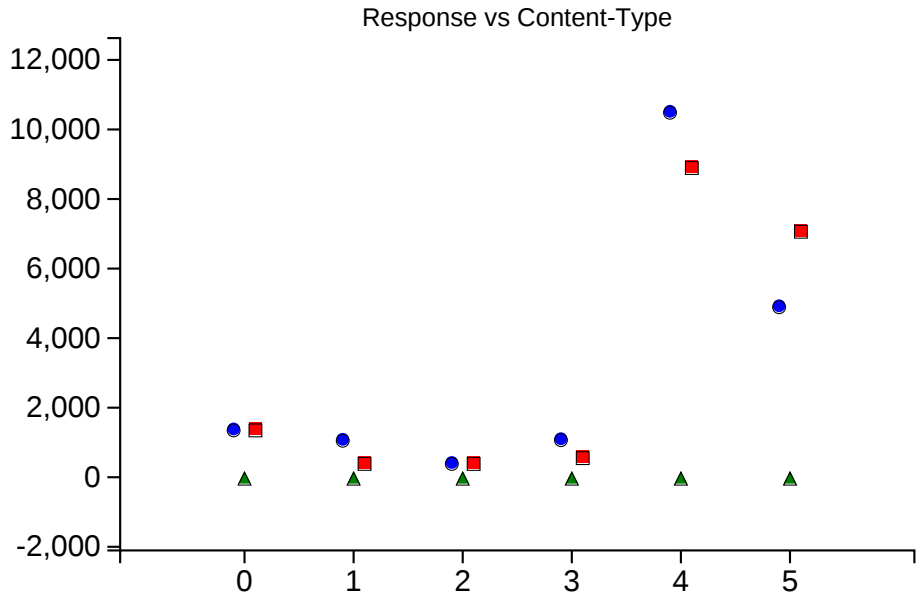
```
=== path min max avg======
/contact 1094 5377 0.0
/users/2 1094 6248 0.0
/users/1 1094 7101 0.0
/sessions 1388 1388 0.0
/signin 1092 5430 0.0
/users/1/edit 1099 6552 0.0
/favicon.ico 424 424 0.0
/assets/bootstrap-tab.js?body=1 4110 426 0.0
/assets/bootstrap-button.js?body=1 3454 426 0.0
/assets/bootstrap-dropdown.js?body=1 424 5026 0.0
/users/3 1094 6254 0.0
/assets/bootstrap-typeahead.js?body=1 425 8933 0.0
/assets/bootstrap.js?body=1 426 621 0.0
/users/3/followers 1104 6719 0.0
/assets/bootstrap-carousel.js?body=1 425 6671 0.0
/assets/turbolinks.js?body=1 13268 426 0.0
/assets/users.css?body=1 425 591 0.0
/assets/bootstrap-collapse.js?body=1 425 5349 0.0
/users 1091 6134 0.0
/assets/jquery.js?body=1 273816 426 0.0
/assets/bootstrap-scrollspy.js?body=1 425 5269 0.0
```

/assets/bootstrap-tooltip.js?body=1 10523 426 0.0
/assets/custom.css?body=1 134336 426 0.0
/assets/bootstrap-modal.js?body=1 421 7269 0.0
/assets/jquery_ujs.js?body=1 16091 426 0.0
/assets/bootstrap-transition.js?body=1 2370 426 0.0
/assets/sessions.js?body=1 424 638 0.0
/assets/bootstrap-affix.js?body=1 4097 426 0.0
/assets/bootstrap-alert.js?body=1 3138 426 0.0
/assets/users.js?body=1 424 639 0.0
/ 4930 4930 0.0
/assets/application.css?body=1 1108 426 0.0
/assets/bootstrap-popover.js?body=1 3728 426 0.0
/assets/sessions.css?body=1 424 591 0.0
/assets/static_pages.js?body=1 425 639 0.0
/assets/static_pages.css?body=1 425 591 0.0
/about 1091 5501 0.0
/assets/application.js?body=1 1192 426 0.0
[0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
23 24
 25 26 27 28 29 30 31 32 33 34 35 36 37] ['/contact', '/users/2', '/us
ers/1', '/sessions', '/signin', '/users/1/edit', '/favicon.ico', '/ass
ets/bootstrap-tab.js?body=1', '/assets/bootstrap-button.js?body=1', '/
assets/bootstrap-dropdown.js?body=1', '/users/3', '/assets/bootstrap-t
ypeahead.js?body=1', '/assets/bootstrap.js?body=1', '/users/3/follower
s', '/assets/bootstrap-carousel.js?body=1', '/assets/turbolinks.js?bod
y=1', '/assets/users.css?body=1', '/assets/bootstrap-collapse.js?body=
1', '/users', '/assets/jquery.js?body=1', '/assets/bootstrap-scrollspy
.js?body=1', '/assets/bootstrap-tooltip.js?body=1', '/assets/custom.cs
s?body=1', '/assets/bootstrap-modal.js?body=1', '/assets/jquery_ujs.js
?body=1', '/assets/bootstrap-transition.js?body=1', '/assets/sessions.
js?body=1', '/assets/bootstrap-affix.js?body=1', '/assets/bootstrap-al
ert.js?body=1', '/assets/users.js?body=1', '/', '/assets/application.c
ss?body=1', '/assets/bootstrap-popover.js?body=1', '/assets/sessions.c
ss?body=1', '/assets/static_pages.js?body=1', '/assets/static_pages.cs
s?body=1', '/about', '/assets/application.js?body=1']



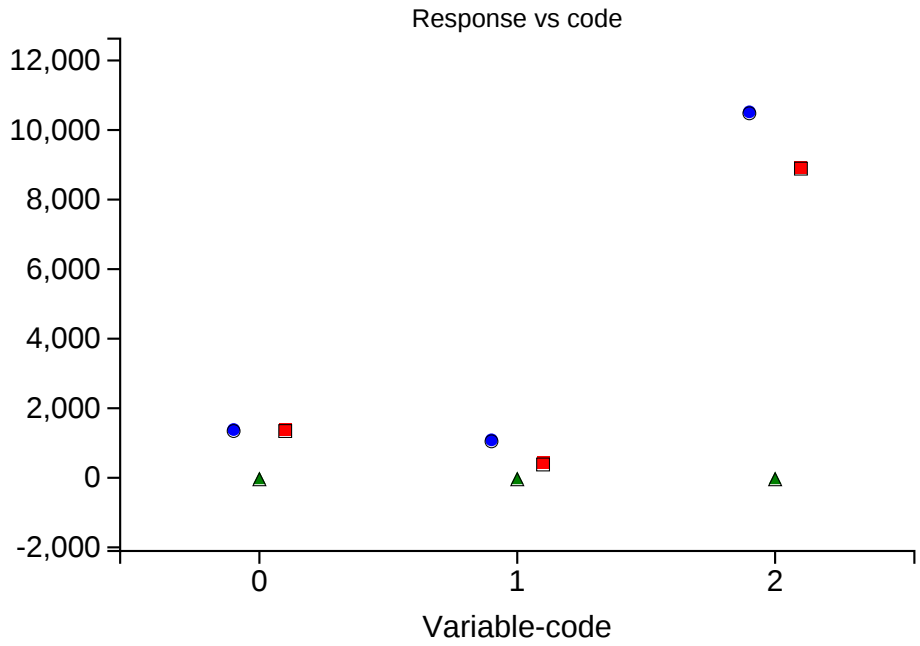
In [134]:
```
# path
checkFeatureDependence(f="Content-Type", option=0)
```

```
=== Content-Type min max avg======
application/x-www-form-urlencoded 1388 1388 0.0
None 1091 426 0.0
image/vnd.microsoft.icon 424 424 0.0
text/css 1108 591 0.0
application/javascript 10523 8933 0.0
text/html; charset=utf-8 4930 7101 0.0
[0 1 2 3 4 5] ['application/x-www-form-urlencoded', 'None', 'image/vnd
.microsoft.icon', 'text/css', 'application/javascript', 'text/html; ch
arset=utf-8']
```



In [135]:
```
# code = 200, 302, 304
checkFeatureDependence(f="code", option=0)
```

```
=== code min max avg======
302 1388 1388 0.0
304 1091 426 0.0
200 10523 8933 0.0
[0 1 2] ['302', '304', '200']
```



In [135]:

In [135]:

4. Data Reshape to Pandas dataframe

In [136]:
```
import pandas as pd


#def read_mongo(db, collection, query={}, host='localhost', port=27017
, username=None, password=None, no_id=True):
def read_mongo(collection, query={}, host='localhost',no_id=True):
    """ Read from Mongo and Store into DataFrame """

    # Make a query to the specific DB and Collection
    cursor = collection.find(query)

    # Expand the cursor and construct the DataFrame
    df =  pd.DataFrame(list(cursor))

    # Delete the _id
    if no_id:
        del df['_id']

    return df

df=read_mongo(logs)
df_test = read_mongo(test)
```

In [137]:
```
df
```

Out[137]:

	Accept-Encoding	Accept-Language	Cache-Control	Connection	Content-Length	C
0	gzip,deflate,sdch	en-US,en;q=0.8,ms;q=0.6,id;q=0.4	max-age=0, private, must-revalidate	keep-alive	3726	t
1	gzip,deflate,sdch	en-US,en;q=0.8,ms;q=0.6,id;q=0.4	public, must-revalidate	keep-alive	513	t
2	gzip,deflate,sdch	en-US,en;q=0.8,ms;q=0.6,id;q=0.4	public, must-revalidate	keep-alive	15477	a
3	gzip,deflate,sdch	en-US,en;q=0.8,ms;q=0.6,id;q=0.4	public, must-revalidate	keep-alive	0	t

4	gzip,deflate,sdch	en-US,en;q=0.8,ms;q=0.6,id;q=0.4	public, must-revalidate	keep-alive	0	t
5	gzip,deflate,sdch	en-US,en;q=0.8,ms;q=0.6,id;q=0.4	public, must-revalidate	keep-alive	0	t
6	gzip,deflate,sdch	en-US,en;q=0.8,ms;q=0.6,id;q=0.4	public, must-revalidate	keep-alive	133735	t
7	gzip,deflate,sdch	en-US,en;q=0.8,ms;q=0.6,id;q=0.4	public, must-revalidate	keep-alive	273200	a
8	gzip,deflate,sdch	en-	public, must-	keep-alive	12654	a

In [138]:
```
#df.dropna()
df["timeStampStartEnd"]=df['timestamp_end'].astype(float) - df['timest
amp_start'].astype(float)
df_test["timeStampStartEnd"]=df_test['timestamp_end'].astype(float) -
df_test['timestamp_start'].astype(float)
```

In [139]:
```
# Writing Pandas dataframe to CSV file
df.to_csv("output3_df.txt",sep=",")
df_test.to_csv("test_df.txt",sep=",")
```

In [140]:
```
# Correlation matrix between
#
#
df.ix[:,['Content-Length','content','response', "headers","X-Runtime"]
].astype(float).corr()
```

Out[140]:

	Content-Length	content	response	headers	X-Runtime
Content-Length	1.000000	-0.049260	0.999985	-0.059063	-0.108725
content	-0.049260	1.000000	-0.001312	0.037921	0.587040
response	0.999985	-0.001312	1.000000	-0.263237	0.014125
headers	-0.059063	0.037921	-0.263237	1.000000	0.001492
X-Runtime	-0.108725	0.587040	0.014125	0.001492	1.000000

In [141]:
```
df.ix[:,["X-Runtime","headers"]].astype(float).corr()
```

Out[141]:

	X-Runtime	headers
X-Runtime	1.000000	0.001492
headers	0.001492	1.000000

In [142]:
```
df.ix[:10,['content',"headers",'response', ]]
```

Out[142]:

	content	headers	response
0	0	399	4930
1	0	778	1108
2	0	762	16091
3	0	778	591
4	0	778	591
5	0	778	591
6	0	778	134336
7	0	762	273816
8	0	762	13268
9	0	762	4097
10	0	762	2370

In [143]:
```
print df.loc[df['response'].astype(int).isin([134336, 273816, 4110, 37
28])]
```

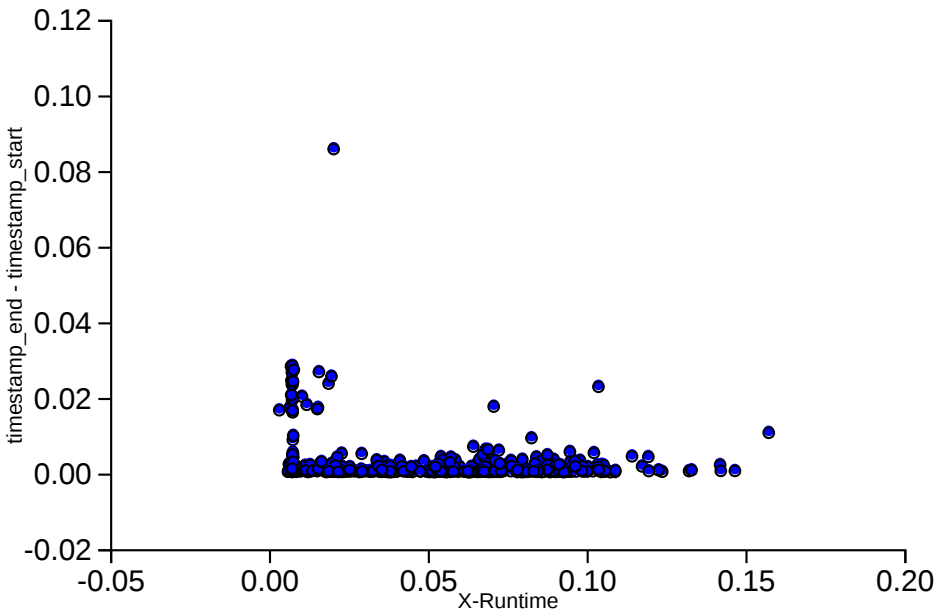
```
     Accept-Encoding                  Accept-Language  \
6   gzip,deflate,sdch  en-US,en;q=0.8,ms;q=0.6,id;q=0.4
7   gzip,deflate,sdch  en-US,en;q=0.8,ms;q=0.6,id;q=0.4
19  gzip,deflate,sdch  en-US,en;q=0.8,ms;q=0.6,id;q=0.4
```

```
21  gzip,deflate,sdch  en-US,en;q=0.8,ms;q=0.6,id;q=0.4

             Cache-Control  Connection Content-Length  \
6   public, must-revalidate  keep-alive         133735
7   public, must-revalidate  keep-alive         273200
19  public, must-revalidate  keep-alive           3498
21  public, must-revalidate  keep-alive           3116

             Content-Type                 Date  \
6                text/css  Tue, 14 Oct 2014 00
7   application/javascript  Tue, 14 Oct 2014 00
19  application/javascript  Tue, 14 Oct 2014 00
21  application/javascript  Tue, 14 Oct 2014 00

                                  Etag        Last-Modified  Num-Cooki
e  \
6   "2d2f6e06e330622fdc0afff0c86f0f54"  Tue, 24 Jun 2014 14
0
7   "6e3920d87a0f9be18aff05235d9c84f8"  Wed, 01 Oct 2014 05
0
19  "87194a63ad27787408db264c5deb3872"  Wed, 01 Oct 2014 05
0
21  "ecbdf22134362275e2e40c4913ff2734"  Wed, 01 Oct 2014 05
0

       ...     msg                                path port request
\
6      ...      OK            /assets/custom.css?body=1   80    1192

7      ...      OK             /assets/jquery.js?body=1   80    1176

19     ...      OK      /assets/bootstrap-tab.js?body=1   80    1165

21     ...      OK  /assets/bootstrap-popover.js?body=1   80    1169


   requestcount response scheme      timestamp_end    timestamp_start
 \
6             1   134336   http   1413247021.71032  1413247021.653419

7             1   273816   http  1413247021.694431  1413247021.656471

19            3     4110   http  1413247022.531667  1413247022.513663

21            3     3728   http  1413247022.758954  1413247022.753162


   timeStampStartEnd
6           0.056901
7           0.037960
19          0.018004
21          0.005792

[4 rows x 41 columns]
```

In [144]:
```
df["response"][df['response']==13433]
```

Out[144]:
```
Series([], name: response, dtype: object)
```

In [148]:
```
import matplotlib.pyplot as plt
plt.figure("X-Runtime vs timeStamp_diff")
plt.xlabel("X-Runtime")
plt.ylabel("timestamp_end - timestamp_start")


#plt.scatter(training_sframe['content'], prediction,)

x1=df['timestamp_end'].astype(float) - df['timestamp_start'].astype(fl
oat)
x2 = df['X-Runtime'].astype(float)
plt.scatter(x1,x2)
#plt.show()
plt.savefig("runtime.png")
```



In [146]:
```
df.ix[:,["request","response","content","headers"]]
```

Out[146]:

	request	response	content	headers
0	789	4930	0	399
1	1179	1108	0	778
2	1162	16091	0	762
3	1198	591	0	778
4	1195	591	0	778
5	1190	591	0	778
6	1192	134336	0	778
7	1176	273816	0	762

7	1170	273810	0	762
8	1162	13268	0	762
9	1166	4097	0	762
10	1171	2370	0	762
11	1166	3138	0	762
12	1168	3454	0	762
13	1170	6671	0	762
14	1170	5349	0	762
15	1170	5269	0	762
16	1166	7269	0	762
17	1170	5026	0	762
18	1168	10523	0	762
19	1165	4110	0	762
20	1161	621	0	762
21	1169	3728	0	762
22	1159	638	0	762
23	1171	8933	0	762
24	1164	639	0	762
25	1157	639	0	762
26	1163	1192	0	762
27	1106	424	0	722
28	1240	5430	0	862
29	1311	1092	0	932
...	...	...	...	...
390	1358	426	0	963
391	1365	426	0	963
392	1353	1094	0	972

In [146]:

In []: