JODATUT2019 Datan kuvaileminen

Nimi: Petri Salminen
Opnum: 243840
Email: petri.salminen@tuni.fi

Asentaakseni holoviews:n, syötin seuraavan komennon komentoriviltä: `pip install "holoviews[recommended]"` . Importin pitäisi nyt onnistua.

In [1]:
```
import pandas as pd
import matplotlib
import holoviews as hv
hv.extension('matplotlib')
```

Aloitetaan datan kuvaileminen lataamalla data sisään ja katsomalla sen 5 ensimmäistä alkiota.

In [2]:
```
df = pd.read_csv('data/weblog_clean.csv', sep=",", index_col=0)
df.head()
```

Out[2]:

	IP	Time	URL	Status	RequestType	HttpVersion
0	176161281	2017-11-29 06:58:55	/login.php	200	GET	1.1
1	176161281	2017-11-29 06:59:02	/process.php	302	POST	1.1
2	176161281	2017-11-29 06:59:03	/home.php	200	GET	1.1
3	176357889	2017-11-29 06:59:04	/js/vendor/moment.min.js	200	GET	1.1
4	176292353	2017-11-29 06:59:06	/bootstrap-3.3.7/js/bootstrap.js	200	GET	1.1

In [3]:
```
%matplotlib inline
```

Seuraavaksi katsokaamme perustietoja datasta. Tästä esimerkiksi näkee, miten edellisessä tehtävässä numeerisiksi muutetut muuttujat luetaan nyt oikein int64 tai float64 -muuttujina.

In [4]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 15789 entries, 0 to 16006
Data columns (total 6 columns):
IP             15789 non-null int64
Time           15789 non-null object
URL            15789 non-null object
Status         15789 non-null int64
RequestType    15789 non-null object
HttpVersion    15789 non-null float64
dtypes: float64(1), int64(2), object(3)
memory usage: 863.5+ KB
```

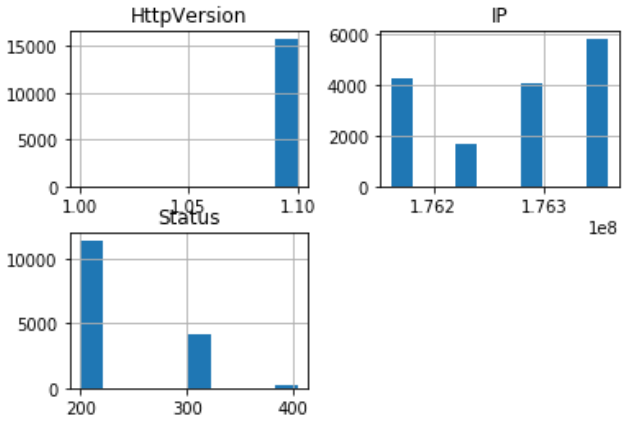
Time luettiin näköjään objectina. Muutetaan se datetime-muuttujaksi.

In [5]:
```
df['Time'] = pd.to_datetime(df['Time'], format="%Y-%m-%d %H:%M:%S")# 2017-11-29 06:58:55
```

Eräs näppärä tapa tutkia muuttujien sisältöä on laittaa ne histogramiin. Tämä käy näppärästi alla olevalla komennolla, mikäli data on numeerista. Valitettavasti tämä kyseinen lokidata ei ollut hirveän nättiä.

In [6]:
```
df.hist()
```

Out[6]:
```
array([[<matplotlib.axes._subplots.AxesSubplot object at 0x7f9704afc828>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f9704ac8cf8>],
       [<matplotlib.axes._subplots.AxesSubplot object at 0x7f9704a7a160>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f9704aa26d8>]],
      dtype=object)
```

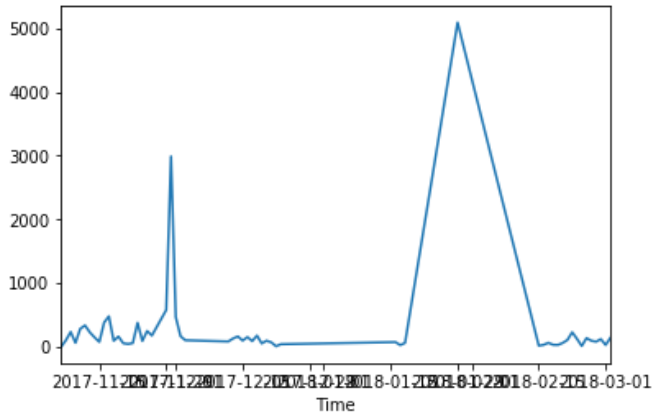


Histogrammeista näkee, että HttpVersion-muuttuja on oikeastaan täysin turha, koska se sisältää vain arvoa 1.1. Toivoin alun perin, että IP-muuttujasta olisi ollut paljon apua, mutta ilmeisesti kyseessä onkin 4 käyttäjää koko datasetissä. Hieman pettynyt olen. Status-muuttujassakaan ei ole mitään oikeasti erikoista. 200 dominoi ja sen lisäksi muutamat 302 (redirect) ja 404 (not found).

Seuraavaksi on hyvä tutkia alkioiden jakautumista ajan suhteen:

In [7]:
```
df.groupby(df.Time.dt.date).size().plot()
```

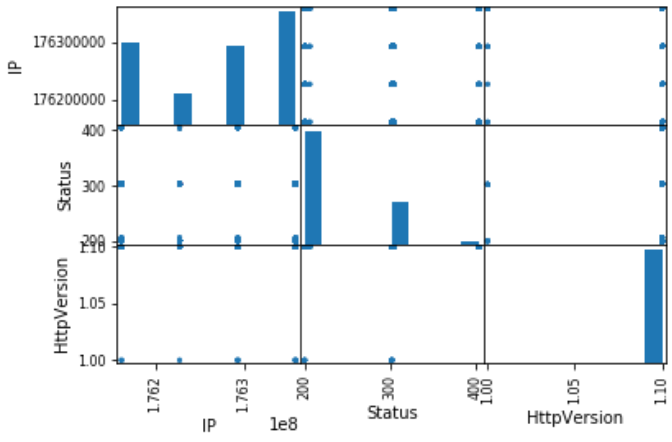
Out[7]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f970299ccc0>
```



Ylläolevasta kuvaajasta lienee aika selvää, että kyseiseen datasettiin liittyvät HTTP-pyynnöt on tehty pääosin kahtena päivänä. Lisäksi on pieniä määriä satunnaisilta päiviltä ajalla Marraskuu 2017- Maaliskuu 2018.

In [8]:
```
pd.plotting.scatter_matrix(df) # Tästä ei paljoa hurskastuttu
```

Out[8]:
```
array([[<matplotlib.axes._subplots.AxesSubplot object at 0x7f97049cc7f0>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f97028919e8>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f9702835f60>],
       [<matplotlib.axes._subplots.AxesSubplot object at 0x7f9702864518>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f970280ca90>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f97027ba048>],
       [<matplotlib.axes._subplots.AxesSubplot object at 0x7f97027e15c0>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f9702789b70>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f9702789ba8>]],
      dtype=object)
```



Tutkitaan vielä URL-muuttujaa. Sitä on hieman hankalampi tutkia, koska se on selkeästi tekstipohjainen muuttuja. Mitä tässä tapahtuu on se, että groupby:lla ryhmitellään URL-muuttujan mukaan. Sen jälkeen otetaan size(), jolloin size lasketaan jokaisesta erikseen.

In [9]:
```
df.groupby(df.URL).size()
```

Out[9]:

URL	
/	862
//home.php	2
/action.php	83
/adminpanel.php	1
/allsubmission.php	140
/allsubmission.php?name=abc145	1
/allsubmission.php?name=ham05	2
/allsubmission.php?name=mahadi	1
/allsubmission.php?name=moshiur_cse15	1
/allsubmission.php?name=shawon	2
/allsubmission.php?name=zerocool	1
/allsubmission.php?page=2	11
/allsubmission.php?page=3	7
/allsubmission.php?page=4	2
/allsubmission.php?page=5	2
/announcement.php	8
/archive.php	246
/archive.php?page=1	8
/archive.php?page=2	55
/bootstrap-3.3.7/js/bootstrap.js	191
/bootstrap-3.3.7/js/bootstrap.min.js	382
/compile.php	96
/compiler.php	98
/contest.php	249
/contestproblem.php	12
/contestproblem.php?name=	3
/contestproblem.php?name=ROJ%20Testing%20Contest%201	1
/contestproblem.php?name=RUET%20CSE%20Contest%203	4
/contestproblem.php?name=RUET%20OJ%20Final%20Test	4
/contestproblem.php?name=RUET%20OJ%20Server%20Testing%20Contest	467
	...
/showcode.php?id=294&nm=Rakib_1603065	1
/showcode.php?id=296&nm=bruce	1
/showcode.php?id=300&nm=x	1
/showcode.php?id=301&nm=Shawon14012	2
/showcode.php?id=304&nm=vinoth	1
/showcode.php?id=308&nm=ham05	2
/showcode.php?id=309&nm=ham05	2
/showcode.php?id=313&nm=abc145	1
/sign.php	127
/sign.php?value=fail	9
/standings.php	1
/standings.php?id=12	2
/standings.php?id=13	7
/standings.php?id=14	1
/standings.php?id=16	153
/standings.php?id=3	1
/standings.php?id=4	2
/submit.php	1
/submit.php?id=55	12
/submit.php?id=58	8
/submit.php?id=63	1
/submit.php?id=64	11
/submit.php?id=67	2
/submit.php?id=68	2
/submit.php?id=71	1
/submit.php?id=73	3
/submit.php?id=76	2
/submit.php?id=77	8
/submit.php?id=78	3
/update.php	7

Length: 265, dtype: int64

Ylläolevissa muuttujissa on selkeästi hieman hassusti se, että siellä on GET-parametrit mukana. Ne voi riisua, koska ne eivät suoranaisesti kuulu URLiin.

In [10]:
```
# Riisutaan GET-parametrit URLeista
df["URL2"] = pd.Series([url.split("?")[0] for url in df.URL])
df.groupby(df.URL2).size()
```

Out[10]:

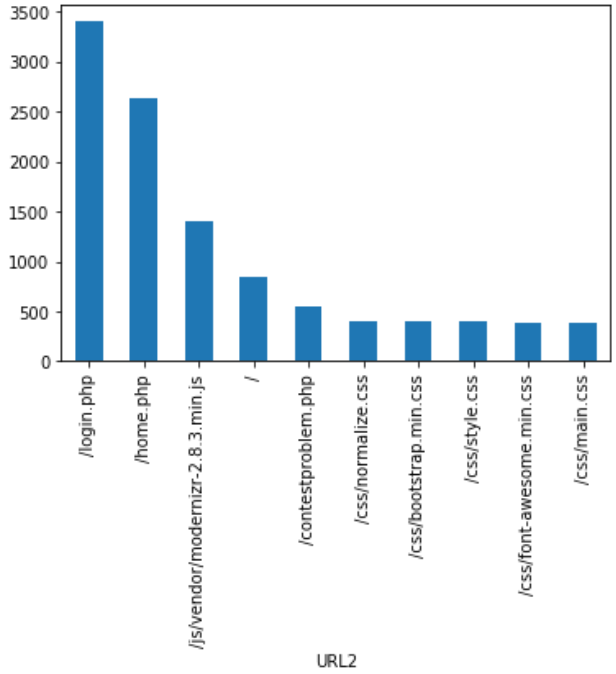
URL2	
/	848
//home.php	2
/action.php	82
/adminpanel.php	1
/allsubmission.php	167
/announcement.php	8
/archive.php	307
/bootstrap-3.3.7/js/bootstrap.js	188
/bootstrap-3.3.7/js/bootstrap.min.js	374
/compile.php	94
/compiler.php	94
/contest.php	244
/contestproblem.php	549
/contestshowcode.php	6
/contestsubmission.php	224
/contestsubmit.php	52
/countdown.php	72
/createadmin.php	4
/css/bootstrap.min.css	395
/css/bootstrap.min.css.map	1
/css/font-awesome.min.css	391
/css/main.css	390
/css/normalize.css	402
/css/style.css	393
/dboot/js/bootstrap.min.js	5
/dcss/bootstrap-datetimepicker.min.css	5
/description.php	124
/details.php	290
/djquery/jquery-1.8.3.min.js	5
/djs/bootstrap-datetimepicker.js	5
	...
/editcontest.php	3
/editcontestproblem.php	12
/favicon.ico	18
/fonts/fontawesome-webfont.eot	6
/fonts/fontawesome-webfont.woff	22
/fonts/fontawesome-webfont.woff2	238
/fonts/glyphicons-halflings-regular.woff	2
/fonts/glyphicons-halflings-regular.woff2	3
/home.php	2640
/img/ruet.png	209
/index.php	4
/js/chart.min.js	56
/js/jquery.min.js	55
/js/vendor/jquery-1.12.0.min.js	382
/js/vendor/modernizr-2.8.3.min.js	1407
/js/vendor/moment.min.js	170
/login.php	3400
/logout.php	44
/pcompile.php	76
/process.php	313
/profile.php	142
/robots.txt	214
/setcontest.php	6
/setcontestproblem.php	3
/setproblem.php	10
/showcode.php	53
/sign.php	129
/standings.php	166
/submit.php	54
/update.php	7

Length: 62, dtype: int64

Jes! Vihdoinkin joku muuttuja, jossa on myös hieman enemmän hajontaa. Harmin paikka kylläkin, että kyseessä on kategorinen muuttuja, eikä sitä voi hyödyntää niin näppärästi matematiikassa. Jos päädytään käyttämään OneHotEncodingia jossain vaiheessa, on erittäin hyvä, että URLien määrä saatiin pienennettyä 265:sta 62:een kappaleeseen.

In [20]:
```
df.groupby(df.URL2).size().nlargest(10).plot.bar()
```

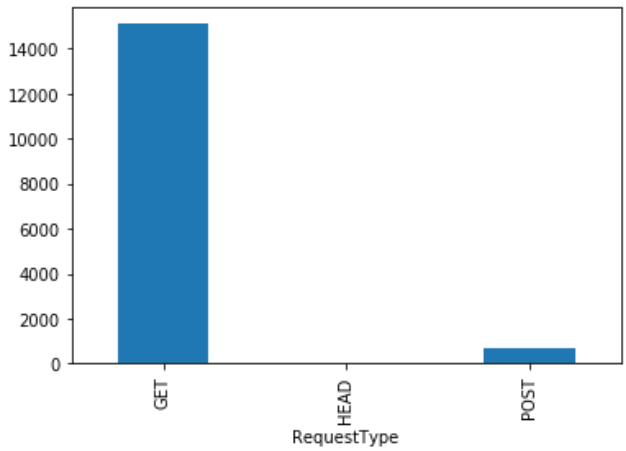
Out[20]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f96fddbccf8>
```



RequestTypessä ei ole paljoakaan hajontaa. Harmi :(

In [12]:
```
df.groupby(df.RequestType).size().plot.bar()
```

Out[12]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f97025b7eb8>
```



Tallennetaan tämänhetkinen Dataframe levylle, jotta sitä voidaan hyödyntää näppärästi seuraavassa tehtävässä.

In [21]:
```
from pickle import dump, HIGHEST_PROTOCOL
with open('df.pickle', 'wb') as f:
    dump(df, f, HIGHEST_PROTOCOL)
```

Hyödylliset linkit

- Python Pickle Module for saving Objects by serialization (https://pythonprogramming.net/python-pickle-module-save-objects-serialization/)
- Monissa koodipätkissä hyödynnetty etenkin Pandasin dokumentaatiota (https://pandas.pydata.org/pandas-docs/stable/reference/index.html)

Helpot ja hankalat asiat

- Lähtödatahan tässä taitaa se hankalin olla. Kyllä tässä pitäisi jokin pokaali saada, jos saan mitään järkevää ulos tästä datasta (mikä ei ole itsestään selvää)
- Histogrammit ovat pandasilla erittäin iisejä. Myös groupby, kunhan sen hiffailee :) SQL-taidoista on hyötyä.

In [ ]:
```
```