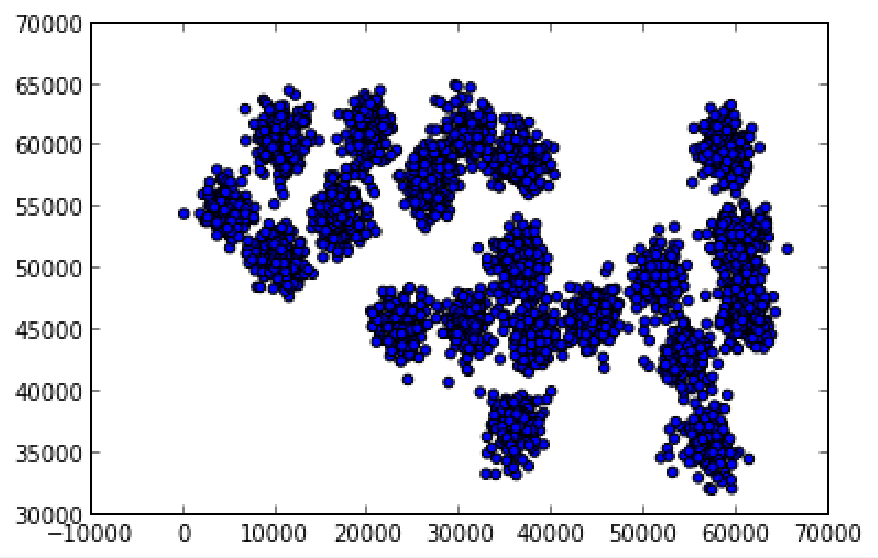
0.1 Hadoop ve KMeans

Some text goes here

```
from pandas import *
df = read_csv("synthetic.txt",names=['a','b'],sep=" ")
scatter(df['a'],df['b'])
```

```
<matplotlib.collections.PathCollection at 0xa6f320c>
```



```
!ssh localhost -l hduser /home/hduser/Downloads/hadoop*/bin/stop-all.sh
!ssh localhost -l hduser /home/hduser/Downloads/hadoop*/bin/start-all.sh
```

```
no jobtracker to stop

localhost: no tasktracker to stop

no namenode to stop

localhost: no datanode to stop

localhost: no secondarynamenode to stop

starting namenode, logging to /home/hduser/Downloads/hadoop-1.0.4/libexec/../logs/hadoop-hduser-namenode

localhost: starting datanode, logging to /home/hduser/Downloads/hadoop-1.0.4/libexec/../logs/hadoop-hdus

localhost: starting secondarynamenode, logging to /home/hduser/Downloads/hadoop-1.0.4/libexec/../logs/ha

starting jobtracker, logging to /home/hduser/Downloads/hadoop-1.0.4/libexec/../logs/hadoop-hduser-jobtra
```

```
localhost: starting tasktracker, logging to /home/hduser/Downloads/hadoop-1.0.4/libexec/../logs/hadoop-h
```

```
!ssh localhost -l hduser /home/hduser/Downloads/hadoop*/bin/hadoop dfs -mkdir /tmp/
```

```
!ssh localhost -l hduser /home/hduser/Downloads/hadoop*/bin/hadoop dfs -ls /
```

```
Found 3 items
drwxr-xr-x   - hduser supergroup          0 2013-02-25 17:23 /app
drwxr-xr-x   - hduser supergroup          0 2013-02-26 12:49 /tmp
drwxr-xr-x   - hduser supergroup          0 2013-02-26 11:45 /user
```

```
!ssh localhost -l hduser /home/hduser/Downloads/hadoop*/bin/hadoop dfs -copyFromLocal
    $HOME/Documents/classnotes/stat/stat_hadoop_kmeans/synthetic.txt /user/hduser
```

```
copyFromLocal: Target /user/hduser/synthetic.txt already exists
```

```
!cp mapper.py /tmp/
!chmod a+r /tmp/mapper.py
!chmod a+x /tmp/mapper.py
!ssh localhost -l hduser /home/hduser/Downloads/hadoop*/bin/hadoop dfs -rmr /user/hduser
    /output
```

```
Deleted hdfs://localhost:54310/user/hduser/output

packageJobJar: [/app/hadoop/tmp/hadoop-unjar4901234381683438453/] [] /tmp/streamjob8733192798849770485.j

13/02/27 01:00:48 INFO util.NativeCodeLoader: Loaded the native-hadoop library
13/02/27 01:00:48 WARN snappy.LoadSnappy: Snappy native library not loaded
13/02/27 01:00:48 INFO mapred.FileInputFormat: Total input paths to process : 1

13/02/27 01:00:48 INFO streaming.StreamJob: getLocalDirs(): [/app/hadoop/tmp/mapred/local]
13/02/27 01:00:48 INFO streaming.StreamJob: Running job: job_201302270038_0001
13/02/27 01:00:48 INFO streaming.StreamJob: To kill this job, run:
13/02/27 01:00:48 INFO streaming.StreamJob: /home/hduser/Downloads/hadoop-1.0.4/libexec/../bin/hadoop jo
13/02/27 01:00:48 INFO streaming.StreamJob: Tracking URL: http://localhost:50030/jobdetails.jsp?jobid=jo

13/02/27 01:00:49 INFO streaming.StreamJob:  map 0%  reduce 0%

13/02/27 01:01:37 INFO streaming.StreamJob:  map 100%  reduce 100%
13/02/27 01:01:37 INFO streaming.StreamJob: To kill this job, run:
13/02/27 01:01:37 INFO streaming.StreamJob: /home/hduser/Downloads/hadoop-1.0.4/libexec/../bin/hadoop jo
13/02/27 01:01:37 INFO streaming.StreamJob: Tracking URL: http://localhost:50030/jobdetails.jsp?jobid=jo
13/02/27 01:01:37 ERROR streaming.StreamJob: Job not successful. Error: # of failed Map Tasks exceeded a
13/02/27 01:01:37 INFO streaming.StreamJob: killJob...
Streaming Command Failed!
```

```
print open("mapper.py").read()
```

```
#!/usr/bin/python
import os,sys,itertools
import numpy as np
from numpy import linalg as la
os.environ[’MPLCONFIGDIR’]=’/tmp’
import pandas as pd

centers = pd.read_csv("/tmp/centers.csv",header=None,sep=",")
print centers[:4]

def dist(vect,x):
    return np.fromiter(itertools.imap(np.linalg.norm, vect-x),dtype=np.float)

def closest(x):
    d = dist(np.array(centers)[:,1:3],np.array(x))
    return np.argmin(d)

df = pd.read_csv(sys.stdin,header=None,sep="   ")
df[’closest’] = df.apply(closest,axis=1)
print df[:20]
```

```
import os,sys,itertools
from numpy import linalg as la
import pandas as pd
k = 10
df = read_csv("synthetic.txt",names=[’a’,’b’],sep=" ")
centers = df.take(np.random.permutation(len(df))[:k])
centers.to_csv("/tmp/centers.csv",header=None)

os.system("cp mapper.py /tmp/")
os.system("chmod a+r /tmp/mapper.py")
os.system("chmod a+x /tmp/mapper.py")

os.system("ssh localhost -l hduser /home/hduser/Downloads/hadoop*/bin/hadoop dfs -rm /user/hduser/centers.csv")
os.system("ssh localhost -l hduser /home/hduser/Downloads/hadoop*/bin/hadoop dfs -copyFromLocal /tmp/centers.csv /user/hduser")
os.system("ssh localhost -l hduser /home/hduser/Downloads/hadoop*/bin/hadoop dfs -cat /user/hduser/centers.csv")

centers = pd.read_csv("/tmp/centers.csv",header=None,sep=",")
print centers[:4]

os.system("ssh localhost -l hduser /home/hduser/Downloads/hadoop*/bin/hadoop dfs -rmr /user/hduser/output")
os.system("ssh localhost -l hduser /home/hduser/Downloads/hadoop*/bin/hadoop jar /home/hduser/Downloads/hadoop*/contrib/streaming/hadoop-*streaming*.jar -input patent/apat63_99_sampled.txt -output output -mapper /tmp/mapper.py -numReduceTasks 0 ")
```

```
0      1      2
0  2145  28810  55914
1  2687   9657  50553
2  2091  30265  59751
3  1830  37166  58232

1024
```

```
xfrom IPython.core.display import HTML
def css_styling():
    styles = open("../../custom.css", "r").read()
    return HTML(styles)
css_styling()
```

```
<IPython.core.display.HTML at 0xa34402c>
```