Big_data_generate_data

March 16, 2019

In [6]:
```
import numpy as np
from sklearn.datasets import fetch_20newsgroups
news_groups_dataset = fetch_20newsgroups(shuffle=True, remove=('headers','footers','qu
print("number of data entries: %s" % np.shape(news_groups_dataset.data))
print("average number of word per entry: %f"
      % np.mean([len(text.split(' ')) for text in news_groups_dataset.data]))
```

```
number of data entries: 11314
average number of word per entry: 206.159802
```

In [7]:
```
from sklearn.datasets import make_classification
X,y = make_classification(n_samples=10**5, n_features=5, n_informative=3, random_state=
D = np.c_[y, X]
np.savetxt('large_dataset_10_5.csv',D, delimiter=',')
del(D, X, y)
X,y = make_classification(n_samples=10**6, n_features=5, n_informative=3, random_state=
D = np.c_[y, X]
np.savetxt('large_dataset_10_6.csv',D, delimiter=',')
del(D, X, y)
X,y = make_classification(n_samples=10**7, n_features=5, n_informative=3, random_state=
D = np.c_[y, X]
np.savetxt('large_dataset_10_7.csv',D, delimiter=',')
del(D, X, y)
```

In [ ]: