The number of documents in ***Recreational Activity*** are ***3979***.

The number of documents in ***Computer Technology*** are ***3903***.

Should I strip the metadata in the mail ?

When evaluating text classifiers on the 20 Newsgroups data, you should strip newsgroup-related metadata. In scikit-learn, you can do this by setting remove=('headers', 'footers', 'quotes'). The F-score will be lower because it is more realistic.

Top 10 words for category: comp.sys.ibm.pc.hardware

scsi 1000

edu 1423

drive 1403

line 1101

com 1080

ide 573

subject 1024

use 1010

organ 972

card 769

Top 10 words for category: comp.sys.mac.hardware

edu 1899

line 1073

mac 1020

subject 997

organ 934

use 803

quadra 270

appl 664

problem 611

centri 223

Top 10 words for category: misc.forsale

edu 1751

00 1215

line 1044

subject 1008

sale 955

organ 981

univers 564

com 548

new 542

10 509

Top 10 words for category: soc.religion.christian

god 2577

christian 1760

edu 1638

church 937

subject 1176

jesus 904

homosexu 653

peopl 1073

sin 795

line 1052