requirements:

numpy, pandas, nltk, scikit-learn, matplotlib, seaborn

Download 20news groups

In [1]:
```
from sklearn.datasets import fetch_20newsgroups
```

In [2]:
```
groups = fetch_20newsgroups()
```

In [3]:
```
groups.keys()
```

Out[3]:
```
dict_keys(['data', 'filenames', 'target_names', 'target', 'DESCR'])
```

In [4]:
```
import numpy as np
np.unique(groups.target)
```

Out[4]:
```
array([ 0,  1,  2,  3,  4,  5,  6,  7,  8,  9, 10, 11, 12, 13, 14, 15, 16,
       17, 18, 19])
```

In [5]:
```
# visualize target
import seaborn as sns
sns.distplot(groups.target)
import matplotlib.pyplot as plt
plt.show()
```

```
<Figure size 640x480 with 1 Axes>
```

Preprocessing Data

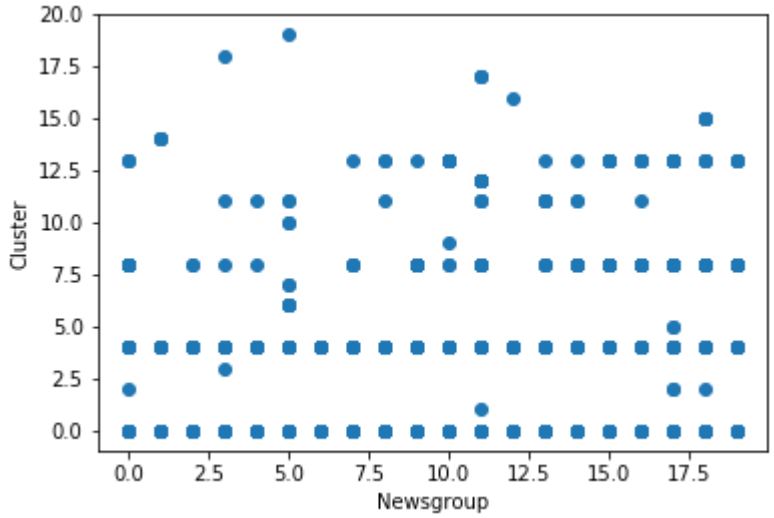
In [6]:
```
from sklearn.feature_extraction.text import CountVectorizer
from nltk.corpus import names
from nltk.stem import WordNetLemmatizer
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt


def letters_only(astr):
    for c in astr:
        if not c.isalpha():
            return False

    return True


cv = CountVectorizer(stop_words="english", max_features=500)
groups = fetch_20newsgroups()
cleaned = []
all_names = set(names.words())
lemmatizer = WordNetLemmatizer()


for post in groups.data:
    cleaned.append(' '.join([lemmatizer.lemmatize(word.lower())
                             for word in post.split()
                             if letters_only(word)
                             and word not in all_names]))


transformed = cv.fit_transform(cleaned)
print(cv.get_feature_names())
```

```
['able', 'accept', 'access', 'according', 'act', 'action', 'actually', 'add',
'address', 'ago', 'agree', 'algorithm', 'allow', 'american', 'anonymous', 'an
swer', 'anybody', 'apple', 'application', 'apr', 'arab', 'area', 'argument',
'armenian', 'article', 'ask', 'asked', 'assume', 'atheist', 'attack', 'attemp
t', 'available', 'away', 'bad', 'based', 'basic', 'belief', 'believe', 'bes
t', 'better', 'bible', 'big', 'bike', 'bit', 'black', 'board', 'body', 'boo
k', 'box', 'build', 'bus', 'business', 'buy', 'ca', 'california', 'called',
'came', 'car', 'card', 'care', 'carry', 'case', 'cause', 'center', 'certain',
'certainly', 'chance', 'change', 'check', 'child', 'chip', 'christian', 'chur
ch', 'city', 'claim', 'clear', 'clipper', 'code', 'college', 'color', 'come',
'coming', 'command', 'comment', 'common', 'communication', 'company', 'comput
er', 'computing', 'consider', 'considered', 'contact', 'control', 'controlle
r', 'copy', 'correct', 'cost', 'country', 'couple', 'course', 'cover', 'creat
e', 'crime', 'current', 'cut', 'data', 'day', 'db', 'deal', 'death', 'departm
ent', 'design', 'device', 'did', 'difference', 'different', 'discussion', 'di
sk', 'display', 'division', 'dod', 'doe', 'doing', 'drive', 'driver', 'drug',
'early', 'earth', 'easy', 'effect', 'email', 'encryption', 'end', 'engineerin
g', 'entry', 'error', 'especially', 'event', 'evidence', 'exactly', 'exampl
e', 'expect', 'experience', 'explain', 'face', 'fact', 'faq', 'far', 'fast',
'federal', 'feel', 'figure', 'file', 'final', 'following', 'food', 'force',
'form', 'free', 'friend', 'ftp', 'function', 'game', 'general', 'getting', 'g
iven', 'gmt', 'goal', 'god', 'going', 'good', 'got', 'government', 'graphic',
'great', 'greek', 'ground', 'group', 'guess', 'gun', 'guy', 'ha', 'hand', 'ha
rd', 'hardware', 'having', 'head', 'health', 'hear', 'heard', 'hell', 'help',
'high', 'history', 'hit', 'hockey', 'hold', 'home', 'hope', 'house', 'human',
'ibm', 'idea', 'image', 'important', 'include', 'includes', 'including', 'ind
ividual', 'info', 'information', 'instead', 'institute', 'interested', 'inter
esting', 'international', 'internet', 'israeli', 'issue', 'jew', 'jewish', 'j
ob', 'just', 'key', 'kill', 'killed', 'kind', 'know', 'known', 'la', 'large',
'later', 'law', 'le', 'lead', 'league', 'left', 'let', 'level', 'life', 'ligh
t', 'like', 'likely', 'line', 'list', 'little', 'live', 'local', 'long', 'lon
ger', 'look', 'looking', 'lost', 'lot', 'love', 'low', 'machine', 'mail', 'ma
in', 'major', 'make', 'making', 'man', 'manager', 'matter', 'maybe', 'mean',
'medical', 'member', 'memory', 'men', 'message', 'method', 'military', 'milli
on', 'mind', 'mode', 'model', 'money', 'monitor', 'month', 'moral', 'mouse',
'muslim', 'na', 'nasa', 'national', 'near', 'need', 'needed', 'network', 'ne
w', 'news', 'nice', 'north', 'note', 'number', 'offer', 'office', 'old', 'ope
n', 'opinion', 'order', 'original', 'output', 'package', 'particular', 'pas
t', 'pay', 'pc', 'people', 'period', 'person', 'personal', 'phone', 'place',
'play', 'player', 'point', 'police', 'policy', 'political', 'position', 'poss
ible', 'post', 'posted', 'posting', 'power', 'president', 'press', 'pretty',
'previous', 'price', 'private', 'probably', 'problem', 'product', 'program',
'project', 'provide', 'public', 'purpose', 'question', 'quite', 'radio', 'rat
e', 'read', 'reading', 'real', 'really', 'reason', 'recently', 'reference',
'religion', 'religious', 'remember', 'reply', 'report', 'research', 'respons
e', 'rest', 'result', 'return', 'right', 'road', 'rule', 'run', 'running', 'r
ussian', 'said', 'sale', 'san', 'save', 'saw', 'say', 'saying', 'school', 'sc
ience', 'screen', 'scsi', 'second', 'section', 'security', 'seen', 'sell', 's
end', 'sense', 'sent', 'serial', 'server', 'service', 'set', 'shall', 'shor
t', 'shot', 'similar', 'simple', 'simply', 'single', 'site', 'situation', 'si
ze', 'small', 'software', 'sort', 'sound', 'source', 'space', 'special', 'spe
cific', 'speed', 'standard', 'start', 'started', 'state', 'statement', 'sto
p', 'strong', 'study', 'stuff', 'subject', 'sun', 'support', 'sure', 'taken',
'taking', 'talk', 'talking', 'tape', 'tax', 'team', 'technical', 'technolog
y', 'tell', 'term', 'test', 'texas', 'text', 'thanks', 'thing', 'think', 'thi
nking', 'thought', 'time', 'tin', 'today', 'told', 'took', 'total', 'tried',
'true', 'truth', 'try', 'trying', 'turkish', 'turn', 'type', 'understand', 'u
```

```
nit', 'united', 'university', 'unix', 'unless', 'usa', 'use', 'used', 'user',
'using', 'usually', 'value', 'various', 'version', 'video', 'view', 'wa', 'wa
nt', 'wanted', 'war', 'water', 'way', 'weapon', 'week', 'went', 'western', 'w
hite', 'widget', 'willing', 'win', 'window', 'woman', 'word', 'work', 'workin
g', 'world', 'write', 'written', 'wrong', 'year', 'york', 'young']
```

Data Clustering

In [7]:
```
km = KMeans(n_clusters=20)
km.fit(transformed)
labels = groups.target
plt.scatter(labels, km.labels_)
plt.xlabel('Newsgroup')
plt.ylabel('Cluster')
plt.show()
```



Topic modeling

In [8]:
```
from sklearn.decomposition import NMF
nmf = NMF(n_components=100, random_state=43).fit(transformed)

for topic_idx, topic in enumerate(nmf.components_):
    label = '{}: '.format(topic_idx)
    print(label, " ".join([cv.get_feature_names()[i]
                           for i in topic.argsort()[:-9:-1]]))
```

```
 0:  wa thought later took left order seen taken
 1:  db bit data place stuff add time line
 2:  server using display screen support code mouse application
 3:  file section information write source change entry number
 4:  disk drive hard controller support card board head
 5:  entry rule program source number info email build
 6:  new york sale change service result study early
 7:  image software user package using display include support
 8:  window manager application using offer user information course
 9:  gun united control house american second national issue
10:  hockey league team game division player list san
11:  turkish government sent war study came american world
12:  program change technology display information version application rate
13:  space nasa technology service national international small communication
14:  government political federal sure free private local country
15:  output line open write read return build section
16:  people country doing tell live killed lot saying
17:  widget application value set type return function list
18:  child case rate le report area research group
19:  jew jewish world war history help research arab
20:  armenian russian muslim turkish world city road today
21:  president said group tax press working package job
22:  ground box usually power code current house white
23:  russian president american support food money important private
24:  ibm color week memory hardware monitor software standard
25:  anonymous posting service server user group message post
26:  la win san went list year radio near
27:  work job young school lot private create business
28:  encryption technology access device policy security government data
29:  tape driver work memory using cause note following
30:  war military world attack way united russian force
31:  god bible shall man come life hell love
32:  atheist religious religion belief god sort feel idea
33:  data available information user research set model based
34:  center research medical institute national study test north
35:  think lot try trying talk kind agree certainly
36:  water city division list public similar north high
37:  section military shall weapon person division application mean
38:  good cover great pretty probably bad issue life
39:  drive head single mode set using model type
40:  israeli arab attack policy true apr fact stop
41:  use note using usually similar available standard work
42:  know tell way come sure understand let saw
43:  car speed driver change high buy different design
44:  internet email address information anonymous user network mail
45:  like look sound long little guy pretty having
46:  going come way mean kind sure working got
47:  state united public national political federal member local
48:  dod bike member computer list started live email
49:  greek killed act word western muslim turkish talk
50:  computer information public internet list issue network communication
51:  law act federal specific issue clear order moral
52:  book read reference list copy second study offer
53:  argument form true evidence event truth particular known
54:  make sense difference little sure making end tell
55:  scsi hard pc drive device bus different data
56:  time long having able lot order light response
```

```
57:  gun rate crime city death study control difference
58:  right second free shall security mean left american
59:  went came said told started saw took woman
60:  power period second san special le play goal
61:  used using product way function version note single
62:  problem work having using help apple running error
63:  available version widget server includes sun set support
64:  question answer ask asked science reason claim post
65:  san information police said group league political including
66:  number serial large men report following million le
67:  year ago old best sale hit long project
68:  want help let life reason trying copy tell
69:  point way different line algorithm exactly idea view
70:  run running home version start hit win speed
71:  got shot play took goal went hit lead
72:  thing saw sure got trying kind seen asked
73:  graphic send mail message package server various computer
74:  university science department general computer thanks engineering texas
75:  just maybe start thought big probably look getting
76:  key message public security algorithm standard method attack
77:  doe mean anybody actually different ask reading difference
78:  game win sound play left second lead great
79:  ha able called taken given past exactly looking
80:  believe belief christian truth evidence claim mean different
81:  drug study information war group reason usa evidence
82:  need help phone able needed kind thanks bike
83:  did death let money fact man wanted body
84:  chip clipper serial algorithm phone communication encryption key
85:  card driver video support mode mouse board bus
86:  church christian member group true bible different view
87:  ftp available anonymous general nasa package source version
88:  better player best play probably hit maybe big
89:  human life person moral kill claim reason world
90:  bit using let change mode attack size quite
91:  say mean word act clear said read simply
92:  health medical public national care study service user
93:  article post usa read world discussion opinion gmt
94:  team player win play city look bad great
95:  day come word christian said tell little way
96:  really lot sure look fact idea actually feel
97:  unit disk size serial total national got return
98:  image color version free available display current better
99:  woman men muslim religion way man great world
```

In [ ]: