In [1]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.cluster import MiniBatchKMeans
from sklearn.feature_extraction.text import TfidfVectorizer
```

In [2]:

```
data = pd.read_json('C:/Users/Hp/Downloads/combined.json/combined.json', lines=True)
data.head()
```

Out[2]:

	components	contents	date	id	title	topics
0	[National Security Division (NSD)]	PORTLAND, Oregon. – Mohamed Osman Mohamud, 23,...	2014-10-01T00:00:00-04:00	None	Convicted Bomb Plotter Sentenced to 30 Years	[]
1	[Environment and Natural Resources Division]	WASHINGTON – North Carolina’s Waccamaw River...	2012-07-25T00:00:00-04:00	12-919	$1 Million in Restitution Payments Announced t...	[]
2	[Environment and Natural Resources Division]	BOSTON– A $1-million settlement has been...	2011-08-03T00:00:00-04:00	11-1002	$1 Million Settlement Reached for Natural Reso...	[]
3	[Environment and Natural Resources Division]	WASHINGTON—A federal grand jury in Las Vegas...	2010-01-08T00:00:00-05:00	10-015	10 Las Vegas Men Indicted \r\nfor Falsifying V...	[]
4	[Environment and Natural Resources Division]	The U.S. Department of Justice, the U.S. Envir...	2018-07-09T00:00:00-04:00	18-898	$100 Million Settlement Will Speed Cleanup Wor...	[Environment]

In [4]:

```
tfidf = TfidfVectorizer(
    min_df = 5,
    max_df = 0.95,
    max_features = 8000,
    stop_words = 'english'
)
tfidf.fit(data.contents)
text = tfidf.transform(data.contents)
```

In [5]:

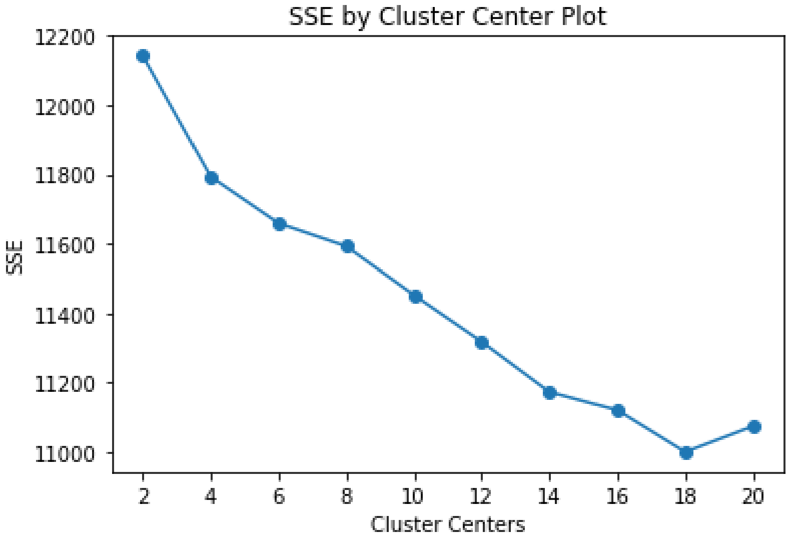
```
def find_optimal_clusters(data, max_k):
    iters = range(2, max_k+1, 2)

    sse = []
    for k in iters:
        sse.append(MiniBatchKMeans(n_clusters=k, init_size=1024, batch_size=2048, rando
m_state=20).fit(data).inertia_)
        print('Fit {} clusters'.format(k))

    f, ax = plt.subplots(1, 1)
    ax.plot(iters, sse, marker='o')
    ax.set_xlabel('Cluster Centers')
    ax.set_xticks(iters)
    ax.set_xticklabels(iters)
    ax.set_ylabel('SSE')
    ax.set_title('SSE by Cluster Center Plot')

find_optimal_clusters(text, 20)
```

```
Fit 2 clusters
Fit 4 clusters
Fit 6 clusters
Fit 8 clusters
Fit 10 clusters
Fit 12 clusters
Fit 14 clusters
Fit 16 clusters
Fit 18 clusters
Fit 20 clusters
```



In [6]:

```
clusters = MiniBatchKMeans(n_clusters=14, init_size=1024, batch_size=2048, random_state
=20).fit_predict(text)
```

In [7]:

```
def get_top_keywords(data, clusters, labels, n_terms):
    df = pd.DataFrame(data.todense()).groupby(clusters).mean()

    for i,r in df.iterrows():
        print('\nCluster {}'.format(i))
        print(','.join([labels[t] for t in np.argsort(r)[-n_terms:]]))

get_top_keywords(text, clusters, tfidf.get_feature_names(), 10)
```

```
Cluster 0
sexually,ceos,safe,children,project,exploitation,minor,sexual,childhood,ch
ild

Cluster 1
division,justice,united,office,indictment,fraud,department,district,crimin
al,attorney

Cluster 2
division,accounts,bank,taxes,attorney,false,returns,income,irs,tax

Cluster 3
hotline,provision,status,800,citizenship,immigration,ina,employment,discri
mination,osc

Cluster 4
allegations,act,services,government,medicare,false,settlement,care,claims,
health

Cluster 5
department,civil,lawsuit,rights,act,disabilities,hud,discrimination,fair,h
ousing

Cluster 6
beneficiaries,home,strike,oig,services,care,hhs,fraud,health,medicare

Cluster 7
prices,division,rigging,department,bid,competition,auctions,fraud,financia
l,antitrust

Cluster 8
county,police,ms,texas,abt,aka,racketeering,murder,members,gang

Cluster 9
act,epa,agreement,civil,settlement,department,voting,disabilities,rights,a
da

Cluster 10
sexual,children,safe,ceos,images,project,childhood,exploitation,pornograph
y,child

Cluster 11
terrorism,al,attorney,law,national,tribal,isil,support,violence,terrorist

Cluster 12
division,department,fbi,indictment,police,officers,attorney,victim,civil,r
ights

Cluster 13
preparing,prepared,return,irs,income,injunction,complaint,customers,return
s,tax
```

In [9]:

```
tfidf.get_feature_names()
```

Out[9]:

```
['00',
 '000',
 '0301',
 '0383',
 '05',
 '09',
 '10',
 '100',
 '1040',
 '107',
 '108',
 '1099',
 '11',
 '110',
 '112',
 '114',
 '115',
 '11th',
 '12',
 '120',
 '123',
 '125',
 '13',
 '130',
 '132',
 '135',
 '14',
 '140',
 '144',
 '145',
 '15',
 '150',
 '151',
 '16',
 '160',
 '168',
 '17',
 '170',
 '175',
 '18',
 '180',
 '188',
 '19',
 '190',
 '1960s',
 '1964',
 '1965',
 '1973',
 '1974',
 '1976',
 '1979',
 '1980',
 '1980s',
 '1981',
 '1982',
 '1983',
 '1984',
 '1985',
 '1986',
 '1987',
 '1988',
 '1989',
 '1990',
 '1990s',
 '1991',
 '1992',
 '1993',
 '1994',
 '1995',
 '1996',
 '1997',
 '1998',
 '1999',
 '20',
 '200',
 '2000',
 '2001',
 '2002',
 '2003',
 '2004',
 '2005',
 '2006',
 '2007',
 '2008',
 '2009',
 '2010',
 '2011',
 '2012',
 '2013',
 '2014',
 '2015',
 '2016',
 '2017',
 '2018',
 '2019',
 '202',
 '20530',
 '20th',
 '21',
 '210',
 '212',
 '215',
 '21st',
 '22',
 '220',
 '225',
 '23',
 '235',
 '237',
 '24',
 '240',
 '25',
 '250',
 '2515',
 '2525',
 '253',
 '255',
 '26',
 '260',
 '262',
 '264',
 '27',
 '270',
 '2735',
 '275',
 '28',
 '280',
 '29',
 '292',
 '30',
 '300',
 '307',
 '31',
 '312',
 '313',
 '32',
 '325',
 '3258',
 '326',
 '33',
 '330',
 '331',
 '335',
 '34',
 '35',
 '350',
 '36',
 '360',
 '362',
 '37',
 '375',
 '38',
 '380',
 '384',
 '39',
 '3931',
 '40',
 '400',
 '404',
 '41',
 '415',
 '42',
 '424',
 '43',
 '436',
 '44',
 '45',
 '450',
 '46',
 '47',
 '48',
 '480',
 '49',
 '50',
 '500',
 '504',
 '51',
 '514',
 '52',
 '53',
 '54',
 '55',
 '550',
 '553',
 '56',
 '57',
 '58',
 '59',
 '60',
 '600',
 '61',
 '616',
 '62',
 '63',
 '64',
 '647',
 '65',
 '657',
 '66',
 '6660',
 '669',
 '6694',
 '67',
 '68',
 '69',
 '70',
 '700',
 '71',
 '72',
 '720',
 '73',
 '74',
 '7400',
 '75',
 '750',
 '76',
 '7688',
 '77',
 '7743',
 '78',
 '79',
 '80',
 '800',
 '8000',
 '81',
 '8155',
 '82',
 '83',
 '84',
 '841',
 '85',
 '850',
 '86',
 '866',
 '87',
 '877',
 '88',
 '888',
 '896',
 '90',
 '900',
 '91',
 '911',
 '92',
 '93',
 '94',
 '95',
 '950',
 '96',
 '965',
 '97',
 '9777',
 '98',
 '99',
 'aaron',
 'abandoned',
 'abatement',
 'abbate',
 'abbott',
 'abc',
 'abdul',
 'abdur',
 'abetted',
 'abetting',
 'abide',
 'abiding',
 'ability',
 'able',
 'aboard',
 'abramoff',
 'abroad',
 'absence',
 'absent',
 'absentee',
 'abt',
 'abu',
 'abuse',
 'abused',
 'abuses',
 'abusing',
 'abusive',
 'academic',
 'academy',
 'accept',
 'acceptable',
 'acceptance',
 'accepted',
 'accepting',
 'access',
 'accessed',
 'accessibility',
 'accessible',
 'accessing',
 'accessories',
 'accessory',
 'accident',
 'accommodate',
 'accommodation',
 'accommodations',
 'accompanied',
 'accompanying',
 'accomplish',
 'accomplished',
 'accomplishments',
 'accordance',
 'according',
 'accordingly',
 'account',
 'accountability',
 'accountable',
 'accountant',
 'accountants',
 'accounted',
 'accountholder',
 'accountholders',
 'accounting',
 'accounts',
 'accuracy',
 'accurate',
 'accurately',
 'accusation',
 'accusations',
 'accused',
 'acevedo',
 'achieve',
 'achieved',
 'achievement',
 'achievements',
 'achieving',
 'acid',
 'acknowledge',
 'acknowledged',
 'acknowledges',
 'acosta',
 'acquire',
 'acquired',
 'acquiring',
 'acquisition',
 'acquisitions',
 'acquitted',
 'acre',
 'acres',
 'act',
 'acteam',
 'acted',
 'acting',
 'action',
 'actions',
 'active',
 'actively',
 'activities',
 'activity',
 'actors',
 'actress',
 'acts',
 'actual',
 'actually',
 'acute',
 'ada',
 'adam',
 'adams',
 'add',
 'added',
 'addiction',
 'adding',
 'addition',
 'additional',
 'additionally',
 'address',
 'addressed',
 'addresses',
 'addressing',
 'adequate',
 'adequately',
 'adhere',
 'adherence',
 'adjacent',
 'adkins',
 'administer',
 'administered',
 'administering',
 'administers',
 'administration',
 'administrative',
 'administrator',
 'administrators',
 'admiral',
 'admission',
 'admissions',
 'admit',
 'admits',
 'admitted',
 'admitting',
 'adopt',
 'adopted',
 'adoption',
 'adult',
 'adulterated',
 'adults',
 'advance',
 'advanced',
 'advances',
 'advancing',
 'advantage',
 'advantages',
 'adverse',
 'adversely',
 'advertise',
 'advertised',
 'advertisement',
 'advertisements',
 'advertisers',
 'advertising',
 'advice',
 'advise',
 'advised',
 'advising',
 'advisor',
 'advisors',
 'advisory',
 'advocacy',
 'advocate',
 'advocates',
 'aerospace',
 'affairs',
 'affect',
 'affected',
 'affecting',
 'affects',
 'affidavit',
 'affiliate',
 'affiliated',
 'affiliates',
 'affiliation',
 'affirmative',
 'afford',
 'affordable',
 'afforded',
 'afghan',
 'afghanistan',
 'afmls',
 'africa',
 'african',
 'aftermarket',
 'aftermath',
 'afternoon',
 'ag',
 'agac',
 'agbu',
 'age',
 'agencies',
 'agency',
 'agenda',
 'agent',
 'agents',
 'ages',
 'aggravated',
 'aggregate',
 'aggressive',
 'aggressively',
 'aggrieved',
 'aging',
 'ago',
 'agree',
 'agreed',
 'agreeing',
 'agreement',
 'agreements',
 'agrees',
 'agricultural',
 'agriculture',
 'aguilar',
 'ahead',
 'ahmad',
 'ahmed',
 'ahmedzay',
 'aid',
 'aided',
 'aiding',
 'aids',
 'aig',
 'aimed',
 'aims',
 'air',
 'aircraft',
 'airfield',
 'airline',
 'airlines',
 'airport',
 'airways',
 'ak',
 'aka',
 'al',
 'ala',
 'alabama',
 'alam',
 'alameda',
 'alan',
 'alaska',
 'albany',
 'albert',
 'alberto',
 'albuquerque',
 'alcohol',
 'alejandro',
 'alert',
 'alex',
 'alexander',
 'alexandria',
 'alexis',
 'ali',
 'aliases',
 'alien',
 'aliens',
 'alike',
 'alison',
 'allah',
 'allan',
 'allegation',
 'allegations',
 'allege',
 'alleged',
 'allegedly',
 'alleges',
 'allegiance',
 'alleging',
 'allen',
 'alliance',
 'allied',
 'allies',
 'alligator',
 'allocate',
 'allocated',
 'allocating',
 'allocation',
 'allow',
 'allowed',
 'allowing',
 'allows',
 'ally',
 'almighty',
 'alongside',
 'alonso',
 'alpha',
 'alphabetical',
 'alstom',
 'alter',
 'altered',
 'alternative',
 'alternatives',
 'altogether',
 'altonaga',
 'aluminum',
 'alvarado',
 'alvarez',
 'alvin',
 'alwan',
 'amanda',
 'ambassador',
 'amber',
 'ambulance',
 'amc',
 'amended',
 'amendment',
 'amendments',
 'america',
 'american',
 'americans',
 'americorps',
 'amgen',
 'amin',
 'aml',
 'ammonia',
 'ammunition',
 'amounts',
 'ams',
 'amy',
 'ana',
 'analyses',
 'analysis',
 'analyst',
 'analysts',
 'analyzing',
 'anchorage',
 'anderson',
 'andre',
 'andrea',
 'andres',
 'andrew',
 'andrews',
 'andré',
 'angel',
 'angela',
 'angeles',
 'animal',
 'animals',
 'ann',
 'anna',
 'anne',
 'anniversary',
 'announce',
 'announced',
 'announcement',
 'announcements',
 'announcing',
 'annual',
 'annually',
 'anonymity',
 'anonymous',
 'answer',
 'answers',
 'anthony',
 'anti',
 'anticipated',
 'anticipation',
 'anticompetitive',
 'antidumping',
 'antiques',
 'antitrust',
 'antoine',
 'antonio',
 'apartment',
 'apartments',
 'apd',
 'apex',
 'appeal',
 'appeals',
 'appear',
 'appearance',
 'appearances',
 'appeared',
 'appearing',
 'appellate',
 'appendix',
 'applaud',
 'apple',
 'applicable',
 'applicant',
 'applicants',
 'application',
 'applications',
 'applied',
 'applies',
 'apply',
 'applying',
 'appoint',
 'appointed',
 'appointment',
 'appreciate',
 'appreciates',
 'appreciation',
 'apprehend',
 'apprehended',
 'apprehending',
 'apprehension',
 'approach',
 'approached',
 'approaches',
 'appropriate',
 'appropriately',
 'approval',
 'approve',
 'approved',
 'approving',
 'approximately',
 'apps',
 'apr',
 'april',
 'aqap',
 'aquatic',
 'arab',
 'arabia',
 'arabian',
 'architects',
 'area',
 'areas',
 'arena',
 'argentina',
 'argueta',
 'arias',
 'arifjan',
 'arise',
 'arises',
 'arising',
 'ariz',
 'arizona',
 'ark',
 'arkansas',
 'arl',
 'arlington',
 'arm',
 'armando',
 'armed',
 'armenian',
 'armor',
 'arms',
 'armstrong',
 'army',
 'arnold',
 'arose',
 'arraigned',
 'arraignment',
 'arrange',
 'arranged',
 'arrangement',
 'arrangements',
 'arranging',
 'array',
 'arrest',
 'arrested',
 'arrestee',
 'arrests',
 'arrival',
 'arrived',
 'arriving',
 'arsenal',
 'arson',
 'art',
 'arthritis',
 'arthrocare',
 'arthur',
 'article',
 'articles',
 'artificially',
 'arturo',
 'aryan',
 'asbestos',
 'ashe',
 'asi',
 'asia',
 'asian',
 'aside',
 'ask',
 'asked',
 'asking',
 'asks',
 'aslanyan',
 'aspect',
 'aspects',
 'asphalt',
 'assault',
 'assaulted',
 'assaulting',
 'assaults',
 'assembled',
 'assemblies',
 'assembly',
 'asserted',
 'asserts',
 'assess',
 'assessed',
 'assessing',
 'assessment',
 'assessments',
 'asset',
 'assets',
 'assigned',
 'assignment',
 'assignments',
 'assist',
 'assistance',
 'assistant',
 'assistants',
 'assisted',
 'assisting',
 'associate',
 'associated',
 'associates',
 'association',
 'associations',
 'assume',
 'assumed',
 'assurance',
 'assure',
 'assured',
 'asthma',
 'atascosa',
 'atc',
 'atf',
 'atlanta',
 'atlantic',
 'atlas',
 'atm',
 'atmospheric',
 'atms',
 'atp',
 'atr',
 'attached',
 'attaché',
 'attack',
 'attacked',
 'attacking',
 'attacks',
 'attempt',
 'attempted',
 'attempting',
 'attempts',
 'attend',
 'attendance',
 'attended',
 'attending',
 'attention',
 'attorney',
 'attorneys',
 'attract',
 'au',
 'auction',
 'auctions',
 'audio',
 'audit',
 'audited',
 'auditor',
 'auditors',
 'audits',
 'aug',
 'august',
 'aurora',
 'austin',
 'australia',
 'australian',
 'authorities',
 'authority',
 'authorization',
 'authorize',
 'authorized',
 'authorizes',
 'authorizing',
 'auto',
 'automated',
 'automatic',
 'automobile',
 'automobiles',
 'automotive',
 'auxiliary',
 'availability',
 'available',
 'avenue',
 'average',
 'aves',
 'aviation',
 'avila',
 'avoid',
 'avoided',
 'avoiding',
 'await',
 'awaiting',
 'awaits',
 'award',
 'awarded',
 'awardees',
 'awarding',
 'awards',
 'aware',
 'awareness',
 'away',
 'awlaki',
 'ayala',
 'aziz',
 'azteca',
 'ba',
 'backed',
 'background',
 'backpage',
 'bacteria',
 'bad',
 'badge',
 'bae',
 'baer',
 'bag',
 'baghdad',
 'bagram',
 'bags',
 'bahamas',
 'bail',
 'bailey',
 'bailout',
 'baker',
 'balance',
 'balances',
 'bales',
 'ball',
 'ballast',
 'ballistic',
 'ballot',
 'ballots',
 'baltimore',
 'ban',
 'bancgroup',
 'band',
 'bandes',
 'bangladesh',
 'bank',
 'banker',
 'bankers',
 'banking',
 'bankruptcy',
 'banks',
 'banned',
 'banque',
 'bar',
 'barack',
 'barbara',
 'barclays',
 'barlow',
 'barnes',
 'barred',
 'barrels',
 'barrett',
 'barrier',
 'barriers',
 'barring',
 'barrio',
 'barry',
 'bars',
 'base',
 'based',
 'bases',
 'basic',
 'basis',
 'bass',
 'bates',
 'baton',
 'battery',
 'battle',
 'baum',
 'baumann',
 'bay',
 'bayou',
 'bazaarvoice',
 'bba',
 'beach',
 'bear',
 'bearing',
 'beat',
 'beaten',
 'beating',
 'beatings',
 'beaumont',
 'beck',
 'becker',
 'bed',
 'bedford',
 'bedroom',
 'beef',
 'beemsterboer',
 'beer',
 'began',
 'begin',
 'beginning',
 'begins',
 'begun',
 'behalf',
 'behavior',
 'behavioral',
 'belgium',
 'belief',
 'beliefs',
 'believe',
 'believed',
 'believes',
 'believing',
 'beliveau',
 'belize',
 'bell',
 'bello',
 'belonged',
 'belonging',
 'beltran',
 'beltre',
 'ben',
 'bench',
 'benchmark',
 'bend',
 'bender',
 'beneficial',
 'beneficiaries',
 'beneficiary',
 'benefit',
 'benefited',
 'benefits',
 'benefitted',
 'benitez',
 'benjamin',
 'bennett',
 'bentley',
 'benton',
 'benzene',
 'berg',
 'berger',
 'berkeley',
 'berlin',
 'berman',
 'bermuda',
 'bernard',
 'bernstein',
 'berry',
 'best',
 'beth',
 'betray',
 'betrayed',
 ...]
```

```
In [ ]:
```