In [39]:
```
import pandas as pd
```

In [40]:
```
data = pd.read_csv('raw_partner_headlines.csv')
```

In [41]:
```
data.head(5)
```

Out[41]:

	Unnamed: 0	headline	url	publisher	date	stock
0	2	Agilent Technologies Announces Pricing of $5….....	http://www.gurufocus.com/news/1153187/agilent-...	GuruFocus	2020-06-01 00:00:00	A
1	3	Agilent (A) Gears Up for Q2 Earnings: What's i...	http://www.zacks.com/stock/news/931205/agilent...	Zacks	2020-05-18 00:00:00	A
2	4	J.P. Morgan Asset Management Announces Liquida...	http://www.gurufocus.com/news/1138923/jp-morga...	GuruFocus	2020-05-15 00:00:00	A
3	5	Pershing Square Capital Management, L.P. Buys ...	http://www.gurufocus.com/news/1138704/pershing...	GuruFocus	2020-05-15 00:00:00	A
4	6	Agilent Awards Trilogy Sciences with a Golden ...	http://www.gurufocus.com/news/1134012/agilent-...	GuruFocus	2020-05-12 00:00:00	A

In [42]:
```
# for loop to filter dataset by keywords

keyword_list = ['Supply Chain','China']

#create an empty dataframe called result
result = pd.DataFrame()

for index, row in data.iterrows():
    text = row['headline']
    #for each keyword in my keyword list, if keyword is in text
    # then we append this row to result
    for keyword in keyword_list:
        if keyword in text:
            result = result.append(row)
        else:
            continue
```

In [43]:
```
import re
from sklearn import feature_extraction
stop_words = feature_extraction.text.ENGLISH_STOP_WORDS
from nltk.stem import PorterStemmer
from nltk.stem import WordNetLemmatizer

def preprocess(text):
  text = text.lower() #lowercase
  text = re.sub(r'[^\w\s]', '', text) #remove punctuations
  text = re.sub(r'\d+', '', text) #remove numbers
  text = " ".join(text.split()) #stripWhitespace
  text = text.split()
  text = [x for x in text if x not in stop_words] #remove stopwords
  text = [x for x in text if x not in ["star", "starwars","jedi"]] #remove task specific stopwords
  text = " ".join(text)
  # stemmer_ps = PorterStemmer()
  # text = [stemmer_ps.stem(word) for word in text.split()] #stemming
  # text = " ".join(text)
  # lemmatizer = WordNetLemmatizer()
  # text = [lemmatizer.lemmatize(word) for word in text.split()]  #lemmatization
  # text = " ".join(text)
  return(text)
```

In [44]:
```
result
```

Out[44]:

	Unnamed: 0	headline	url	publisher	date	stock
103	105.0	KraneShares Joins with Nasdaq Dorsey Wright to...	https://www.benzinga.com/node/14643902	GuruFocus	2019-10-23 00:00:00	A
358	360.0	Agilent (A) Opens Logistics Center, Expands Ch...	http://www.zacks.com/stock/news/309535/agilent...	Zacks	2018-06-28 00:00:00	A
490	492.0	Stocks Turn Up; Alphabet At Record High, China...	http://www.investors.com/market-trend/stock-ma...	Investor's Business Daily	2017-05-23 00:00:00	A
619	621.0	Wall Street Breakfast: China Stock Plunge Weig...	http://seekingalpha.com/article/3447576-wall-s...	Seeking Alpha	2015-08-18 00:00:00	A
647	649.0	Pressures from China, Greek Markets - Ahead of...	http://www.zacks.com/stock/news/176398/pressur...	Zacks	2015-05-28 00:00:00	A
...	...	...	...	...	...	...
1845541	1849861.0	China Zenix Auto reports Q4 results	http://seekingalpha.com/news/2432526-china-zen...	Seeking Alpha	2015-04-17 00:00:00	ZX
1845542	1849862.0	China Zenix: This Probably Won't End Well	http://seekingalpha.com/article/2777505-china-...	Seeking Alpha	2014-12-24 00:00:00	ZX
1845545	1849865.0	Can The Uptrend Continue for China Zenix (ZX)?...	http://www.zacks.com/stock/news/125942/can-the...	Zacks	2014-03-11 00:00:00	ZX
1845548	1849868.0	China Zenix Auto International (ZX) Worth Watc...	http://www.zacks.com/stock/news/115709/china-z...	Zacks	2013-12-02 00:00:00	ZX
1845549	1849869.0	China Zenix Auto International Ltd (ZX) Enters...	http://www.zacks.com/stock/news/105808/china-z...	Zacks	2013-08-06 00:00:00	ZX

34951 rows × 6 columns

In [7]:
```
result['text_processed']=result['headline'].apply(lambda x:preprocess(str(x)))
result['text_processed']=result['text_processed'].apply(lambda x:x.split())
```

In [8]:
```
result
```

Out[8]:

	Unnamed: 0	headline	url	publisher	date	stock	text_processed
103	105.0	KraneShares Joins with Nasdaq Dorsey Wright to...	https://www.benzinga.com/node/14643902	GuruFocus	2019-10-23 00:00:00	A	[kraneshares, joins, nasdaq, dorsey, wright, l...
358	360.0	Agilent (A) Opens Logistics Center, Expands Ch...	http://www.zacks.com/stock/news/309535/agilent...	Zacks	2018-06-28 00:00:00	A	[agilent, opens, logistics, center, expands, c...
490	492.0	Stocks Turn Up; Alphabet At Record High, China...	http://www.investors.com/market-trend/stock-ma...	Investor's Business Daily	2017-05-23 00:00:00	A	[stocks, turn, alphabet, record, high, china, ...
619	621.0	Wall Street Breakfast: China Stock Plunge Weig...	http://seekingalpha.com/article/3447576-wall-s...	Seeking Alpha	2015-08-18 00:00:00	A	[wall, street, breakfast, china, stock, plunge...
647	649.0	Pressures from China, Greek Markets - Ahead of...	http://www.zacks.com/stock/news/176398/pressur...	Zacks	2015-05-28 00:00:00	A	[pressures, china, greek, markets, ahead, wall...
...	...	...	...	...	...	...	...
1845541	1849861.0	China Zenix Auto reports Q4 results	http://seekingalpha.com/news/2432526-china-zen...	Seeking Alpha	2015-04-17 00:00:00	ZX	[china, zenix, auto, reports, q, results]
1845542	1849862.0	China Zenix: This Probably Won't End Well	http://seekingalpha.com/article/2777505-china-...	Seeking Alpha	2014-12-24 00:00:00	ZX	[china, zenix, probably, wont, end]
1845545	1849865.0	Can The Uptrend Continue for China Zenix (ZX)?...	http://www.zacks.com/stock/news/125942/can-the...	Zacks	2014-03-11 00:00:00	ZX	[uptrend, continue, china, zenix, zx, tale, tape]
1845548	1849868.0	China Zenix Auto International (ZX) Worth Watc...	http://www.zacks.com/stock/news/115709/china-z...	Zacks	2013-12-02 00:00:00	ZX	[china, zenix, auto, international, zx, worth,...

	Unnamed: 0	headline	url	publisher	date	stock	text_processed
1845549	1849869.0	China Zenix Auto International Ltd (ZX) Enters...	http://www.zacks.com/stock/news/105808/china-z...	Zacks	2013-08-06 00:00:00	ZX	[china, zenix, auto, international, zx, enters...

34951 rows × 7 columns

In [9]:
```
from gensim import corpora
dictionary = corpora.Dictionary(result['text_processed'])
dictionaryDF = pd.DataFrame()
dictionaryDF['id']=dictionary.keys()
dictionaryDF['word']=dictionary.values()
dictionaryDF
```

Out[9]:

	id	word
0	0	china
1	1	dorsey
2	2	joins
3	3	kraneshares
4	4	launch
...	...	...
8949	8949	scoops
8950	8950	premiere
8951	8951	junqiu
8952	8952	zenixs
8953	8953	probably

8954 rows × 2 columns

In [10]:
```
dictionaryDF = pd.DataFrame()
dictionaryDF['id']=dictionary.keys()
dictionaryDF['word']=dictionary.values()
dictionaryDF
```

Out[10]:

	id	word
0	0	china
1	1	dorsey
2	2	joins
3	3	kraneshares
4	4	launch
...	...	...
8949	8949	scoops
8950	8950	premiere
8951	8951	junqiu
8952	8952	zenixs
8953	8953	probably

8954 rows × 2 columns

In [11]:
```
dictionary.filter_extremes(no_below=100,no_above=10000)
# no_below (int, optional) — Keep tokens which are contained in at least no_below documents.
# no_above (float, optional) — Keep tokens which are contained in no more than no_above documents (fract
dictionaryDF = pd.DataFrame()
dictionaryDF['id']=dictionary.keys()
dictionaryDF['word']=dictionary.values()
dictionaryDF
```

Out[11]:

	id	word
0	0	china
1	1	nasdaq
2	2	expands
3	3	high
4	4	record
...	...	...
381	381	sinopec
382	382	profit
383	383	transcript
384	384	recession
385	385	brexit

386 rows × 2 columns

In [12]:
```
dictionary = corpora.Dictionary(result['text_processed'])
dictionary.filter_extremes(no_below=100,keep_tokens=['windshield'])
# keep_tokens (iterable of str) – Iterable of tokens that must stay in dictionary after filtering.
dictionaryDF = pd.DataFrame()
dictionaryDF['id']=dictionary.keys()
dictionaryDF['word']=dictionary.values()
dictionaryDF
```

Out[12]:

	id	word
0	0	nasdaq
1	1	expands
2	2	high
3	3	record
4	4	stocks
...	...	...
380	380	sinopec
381	381	profit
382	382	transcript
383	383	recession
384	384	brexit

385 rows × 2 columns

In [13]:
```
dictionary.filter_extremes(keep_n=5)
# keep_n (int, optional) – Keep only the first keep_n most frequent tokens.
dictionaryDF = pd.DataFrame()
dictionaryDF['id']=dictionary.keys()
dictionaryDF['word']=dictionary.values()
dictionaryDF
```

Out[13]:

	id	word
0	0	stocks
1	1	trade
2	2	uschina
3	3	chinas
4	4	growth

In [14]:
```
dictionary = corpora.Dictionary(result['text_processed'])
result['text_ids']=result['text_processed'].apply(lambda x:dictionary.doc2bow(x))
```

In [15]:
```
from gensim import models
num_topics=5
ldamodel = models.ldamodel.LdaModel(result['text_ids'], num_topics = num_topics, id2word=dictionary, pas
topics = ldamodel.print_topics(num_words=4)
for topic in topics:
    print(topic)
```

```
(0, '0.115*"china" + 0.048*"analyst" + 0.045*"blog" + 0.018*"solar"')
(1, '0.120*"china" + 0.051*"chinas" + 0.045*"growth" + 0.015*"trade"')
(2, '0.113*"china" + 0.041*"chinas" + 0.037*"trade" + 0.018*"market"')
(3, '0.116*"china" + 0.029*"stocks" + 0.026*"chinas" + 0.026*"trade"')
(4, '0.108*"china" + 0.017*"q" + 0.016*"earnings" + 0.016*"chinas"')
```

In [16]:
```
ldamodel = models.ldamodel.LdaModel(result['text_ids'], num_topics = 5, id2word=dictionary, passes=1, ra
```

In [17]:
```
ldamodel = models.ldamodel.LdaModel(result['text_ids'], num_topics = 10, id2word=dictionary, passes=5, r
```

In [18]:
```
topics = ldamodel.print_topics(num_words=7)
for i in range(num_topics):
    print(topics[i])
```

```
(0, '0.114*"china" + 0.081*"analyst" + 0.076*"blog" + 0.027*"roundup" + 0.023*"petrochina" + 0.022*"st
ock" + 0.019*"solar"')
(1, '0.135*"china" + 0.077*"growth" + 0.037*"chinas" + 0.018*"factory" + 0.014*"sector" + 0.013*"activ
ity" + 0.012*"stocks"')
(2, '0.126*"china" + 0.033*"trade" + 0.031*"chinas" + 0.021*"wall" + 0.019*"street" + 0.019*"breakfas
t" + 0.018*"market"')
(3, '0.121*"china" + 0.023*"chinas" + 0.023*"oil" + 0.018*"stocks" + 0.012*"supply" + 0.010*"rally" +
0.009*"crude"')
(4, '0.099*"china" + 0.030*"chinas" + 0.025*"pmi" + 0.017*"buys" + 0.015*"stocks" + 0.014*"capital" +
0.013*"road"')
```

In [19]:
```
result['year'] = pd.to_datetime(result['date'], errors='coerce').dt.year
result['year'] = result['year'].astype(str)
result.head()
```

Out[19]:

	Unnamed: 0	headline	url	publisher	date	stock	text_processed	text_ids	year
103	105.0	KraneShares Joins with Nasdaq Dorsey Wright to...	https://www.benzinga.com/node/14643902	GuruFocus	2019-10-23 00:00:00	A	[kraneshares, joins, nasdaq, dorsey, wright, l...	[(0, 1), (1, 1), (2, 1), (3, 2), (4, 1), (5, 1...	2019
358	360.0	Agilent (A) Opens Logistics Center, Expands Ch...	http://www.zacks.com/stock/news/309535/agilent...	Zacks	2018-06-28 00:00:00	A	[agilent, opens, logistics, center, expands, c...	[(0, 1), (9, 1), (10, 1), (11, 1), (12, 1), (1...	2018
490	492.0	Stocks Turn Up; Alphabet At Record High, China...	http://www.investors.com/market-trend/stock-ma...	Investor's Business Daily	2017-05-23 00:00:00	A	[stocks, turn, alphabet, record, high, china, ...	[(0, 1), (15, 1), (16, 1), (17, 1), (18, 2), (...	2017
619	621.0	Wall Street Breakfast: China Stock Plunge Weig...	http://seekingalpha.com/article/3447576-wall-s...	Seeking Alpha	2015-08-18 00:00:00	A	[wall, street, breakfast, china, stock, plunge...	[(0, 1), (21, 1), (22, 1), (23, 1), (24, 1), (...	2015
647	649.0	Pressures from China, Greek Markets - Ahead of...	http://www.zacks.com/stock/news/176398/pressur...	Zacks	2015-05-28 00:00:00	A	[pressures, china, greek, markets, ahead, wall...	[(0, 1), (22, 1), (25, 1), (26, 1), (29, 1), (...	2015

In [20]:
```
# create sub-dataframes with specific year
# eg. year_2020 contains all AAL data for the year 2020
year_2018 = result[result['year']=='2018']
year_2019 = result[result['year']=='2019']
year_2020 = result[result['year']=='2020']
year_2018.head()
```

Out[20]:

	Unnamed: 0	headline	url	publisher	date	stock	text_processed	text_ids	year
358	360.0	Agilent (A) Opens Logistics Center, Expands Ch...	http://www.zacks.com/stock/news/309535/agilent...	Zacks	2018-06-28 00:00:00	A	[agilent, opens, logistics, center, expands, c...	[(0, 1), (9, 1), (10, 1), (11, 1), (12, 1), (1...	2018
1128	1131.0	Copper drops amid weak China economic data; Fr...	https://seekingalpha.com/news/3417207-copper-d...	Seeking Alpha	2018-12-14 00:00:00	AA	[copper, drops, amid, weak, china, economic, d...	[(0, 1), (39, 1), (50, 1), (51, 1), (52, 1), (...	2018
1145	1148.0	Base metals slip on U.S.-China trade tension; ...	https://seekingalpha.com/news/3402718-base-met...	Seeking Alpha	2018-10-30 00:00:00	AA	[base, metals, slip, uschina, trade, tension, ...	[(33, 1), (41, 1), (43, 1), (57, 1), (58, 1),	2018

In [21]:
```
# for loop to filter dataset by keywords

keyword_list = ['Supply Chain','China']

#create an empty dataframe called result
result_2018 = pd.DataFrame()

for index, row in year_2018.iterrows():
    text = row['headline']
    #for each keyword in my keyword list, if keyword is in text
    # then we append this row to result
    for keyword in keyword_list:
        if keyword in text:
            result_2018 = result_2018.append(row)
        else:
            continue

result_2018.head()
```

Out[21]:

	Unnamed: 0	headline	url	publisher	date	stock	text_processed	text_ids	year
358	360.0	Agilent (A) Opens Logistics Center, Expands Ch...	http://www.zacks.com/stock/news/309535/agilent...	Zacks	2018-06-28 00:00:00	A	[agilent, opens, logistics, center, expands, c...	[(0, 1), (9, 1), (10, 1), (11, 1), (12, 1), (1...	2018
1128	1131.0	Copper drops amid weak China economic data; Fr...	https://seekingalpha.com/news/3417207-copper-d...	Seeking Alpha	2018-12-14 00:00:00	AA	[copper, drops, amid, weak, china, economic, d...	[(0, 1), (39, 1), (50, 1), (51, 1), (52, 1), (...	2018
1145	1148.0	Base metals slip on U.S.-China trade tension; ...	https://seekingalpha.com/news/3402718-base-met...	Seeking Alpha	2018-10-30 00:00:00	AA	[base, metals, slip, uschina, trade, tension, ...	[(33, 1), (41, 1), (43, 1), (57, 1), (58, 1), ...	2018

	Unnamed: 0	headline	url	publisher	date	stock	text_processed	text_ids	year
2943	2967.0	Dorman Products: Strong Q3, But China Tariffs ...	https://seekingalpha.com/article/4215770-dorma...	Seeking Alpha	2018-10-30 00:00:00	AAP	[dorman, products, strong, q, china, tariffs, ...	[(0, 1), (85, 1), (95, 1), (96, 1), (97, 1), (...	2018
3041	3065.0	Auto Stock Roundup: TSLA Hikes Prices in China...	http://www.zacks.com/stock/news/311190/auto-st...	Zacks	2018-07-12 00:00:00	AAP	[auto, stock, roundup, tsla, hikes, prices, ch...	[(0, 1), (24, 1), (81, 1), (87, 1), (94, 1), (...	2018

In [22]:
```
result_2018
```

Out[22]:

	Unnamed: 0	headline	url	publisher	date	stock	text_processed	text_ids	yea
358	360.0	Agilent (A) Opens Logistics Center, Expands Ch...	http://www.zacks.com/stock/news/309535/agilent...	Zacks	2018-06-28 00:00:00	A	[agilent, opens, logistics, center, expands, c...	[(0, 1), (9, 1), (10, 1), (11, 1), (12, 1), (1...	201
1128	1131.0	Copper drops amid weak China economic data; Fr...	https://seekingalpha.com/news/3417207-copper-d...	Seeking Alpha	2018-12-14 00:00:00	AA	[copper, drops, amid, weak, china, economic, d...	[(0, 1), (39, 1), (50, 1), (51, 1), (52, 1), (...	201
1145	1148.0	Base metals slip on U.S.-China trade tension; ...	https://seekingalpha.com/news/3402718-base-met...	Seeking Alpha	2018-10-30 00:00:00	AA	[base, metals, slip, uschina, trade, tension, ...	[(33, 1), (41, 1), (43, 1), (57, 1), (58, 1), ...	201
2943	2967.0	Dorman Products: Strong Q3, But China Tariffs ...	https://seekingalpha.com/article/4215770-dorma...	Seeking Alpha	2018-10-30 00:00:00	AAP	[dorman, products, strong, q, china, tariffs, ...	[(0, 1), (85, 1), (95, 1), (96, 1), (97, 1), (...	201
3041	3065.0	Auto Stock Roundup: TSLA Hikes Prices in China...	http://www.zacks.com/stock/news/311190/auto-st...	Zacks	2018-07-12 00:00:00	AAP	[auto, stock, roundup, tsla, hikes, prices, ch...	[(0, 1), (24, 1), (81, 1), (87, 1), (94, 1), (...	201
...	...	...	...	...	...	...	...	...	.
1845501	1849821.0	China Zenix Auto International's (ZX) CEO Junq...	https://seekingalpha.com/article/4175145-china...	Seeking Alpha	2018-05-17 00:00:00	ZX	[china, zenix, auto, internationals, zx, ceo, ...	[(0, 1), (81, 1), (85, 1), (86, 1), (121, 1), ...	201

	Unnamed: 0	headline	url	publisher	date	stock	text_processed	text_ids	yea
1845502	1849822.0	China Zenix Auto reports Q1 results	https://seekingalpha.com/news/3357513-china-ze...	Seeking Alpha	2018-05-17 00:00:00	ZX	[china, zenix, auto, reports, q, results]	[(0, 1), (81, 1), (85, 1), (86, 1), (533, 1), ...	201
1845503	1849823.0	China could be the winner from auto industry r...	https://seekingalpha.com/news/3346095-china-wi...	Seeking Alpha	2018-04-17 00:00:00	ZX	[china, winner, auto, industry, reset]	[(0, 1), (81, 1), (662, 1), (1315, 1), (1316, 1)]	201
1845506	1849826.0	China Zenix Auto International's (ZX) CEO Junq...	https://seekingalpha.com/article/4156860-china...	Seeking Alpha	2018-03-15 00:00:00	ZX	[china, zenix, auto, internationals, zx, ceo, ...	[(0, 1), (81, 1), (85, 1), (86, 1), (121, 1), ...	201
1845508	1849828.0	China Zenix Auto reports Q4 results	https://seekingalpha.com/news/3339363-china-ze...	Seeking Alpha	2018-03-15 00:00:00	ZX	[china, zenix, auto, reports, q, results]	[(0, 1), (81, 1), (85, 1), (86, 1), (533, 1), ...	201

7130 rows × 9 columns

In [23]:
```
import re
from sklearn import feature_extraction
stop_words = feature_extraction.text.ENGLISH_STOP_WORDS
from nltk.stem import PorterStemmer
from nltk.stem import WordNetLemmatizer

def preprocess(text):
  text = text.lower() #lowercase
  text = re.sub(r'[^\w\s]', '', text) #remove punctuations
  text = re.sub(r'\d+', '', text) #remove numbers
  text = " ".join(text.split()) #stripWhitespace
  text = text.split()
  text = [x for x in text if x not in stop_words] #remove stopwords
  text = [x for x in text if x not in ["star", "starwars","jedi"]] #remove task specific stopwords
  text = " ".join(text)
  # stemmer_ps = PorterStemmer()
  # text = [stemmer_ps.stem(word) for word in text.split()] #stemming
  # text = " ".join(text)
  # lemmatizer = WordNetLemmatizer()
  # text = [lemmatizer.lemmatize(word) for word in text.split()]  #lemmatization
  # text = " ".join(text)
  return(text)
```

In [24]:
```
result_2018['text_processed']=result_2018['headline'].apply(lambda x:preprocess(str(x)))
result_2018['text_processed']=result_2018['text_processed'].apply(lambda x:x.split())
result_2018.head(5)
result_2018.shape
```

Out[24]: (7130, 9)

In [25]:
```
from gensim import corpora
dictionary = corpora.Dictionary(result_2018['text_processed'])
dictionaryDF = pd.DataFrame()
dictionaryDF['id']=dictionary.keys()
dictionaryDF['word']=dictionary.values()
dictionaryDF
```

Out[25]:

	id	word
0	0	agilent
1	1	center
2	2	china
3	3	expands
4	4	foothold
...	...	...
3162	3162	esports
3163	3163	junqiu
3164	3164	zenixs
3165	3165	zx
3166	3166	zenix

3167 rows × 2 columns

In [26]:
```
dictionary = corpora.Dictionary(result_2018['text_processed'])
result_2018['text_ids']=result_2018['text_processed'].apply(lambda x:dictionary.doc2bow(x))

from gensim import models
num_topics=5
ldamodel = models.ldamodel.LdaModel(result_2018['text_ids'], num_topics = num_topics, id2word=dictionary
topics = ldamodel.print_topics(num_words=4)
for topic in topics:
    print(topic)
```

```
(0, '0.144*"china" + 0.041*"tariffs" + 0.025*"trade" + 0.013*"chinas"')
(1, '0.069*"china" + 0.055*"chinas" + 0.045*"trade" + 0.018*"uschina"')
(2, '0.127*"china" + 0.021*"chinas" + 0.015*"new" + 0.012*"tariffs"')
(3, '0.104*"trade" + 0.097*"china" + 0.036*"uschina" + 0.028*"war"')
(4, '0.132*"china" + 0.024*"growth" + 0.024*"trade" + 0.020*"chinas"')
```

In [27]:
```
ldamodel = models.ldamodel.LdaModel(result_2018['text_ids'], num_topics = 5, id2word=dictionary, passes=
ldamodel = models.ldamodel.LdaModel(result_2018['text_ids'], num_topics = 10, id2word=dictionary, passes


topics = ldamodel.print_topics(num_words=8)
for i in range(num_topics):
    print(topics[i])
```

```
(0, '0.136*"china" + 0.036*"trade" + 0.024*"tariffs" + 0.011*"amid" + 0.011*"weighs" + 0.011*"uschina"
+ 0.011*"data" + 0.010*"wto"')
(1, '0.080*"china" + 0.063*"trade" + 0.030*"chinas" + 0.029*"uschina" + 0.024*"war" + 0.022*"data" +
0.011*"lower" + 0.011*"oil"')
(2, '0.138*"china" + 0.014*"earnings" + 0.014*"stocks" + 0.014*"trump" + 0.012*"tariffs" + 0.010*"grow
th" + 0.010*"q" + 0.010*"market"')
(3, '0.105*"trade" + 0.093*"china" + 0.044*"talks" + 0.032*"uschina" + 0.028*"wall" + 0.026*"breakfas
t" + 0.026*"street" + 0.016*"deal"')
(4, '0.144*"china" + 0.042*"trade" + 0.021*"war" + 0.014*"chinas" + 0.013*"growth" + 0.013*"markets" +
0.011*"economic" + 0.008*"way"')
```

In [28]:
```
# 2019
```

In [29]:
```
keyword_list = ['Supply Chain','China']

#create an empty dataframe called result
result_2019 = pd.DataFrame()

for index, row in year_2019.iterrows():
    text = row['headline']
    #for each keyword in my keyword list, if keyword is in text
    # then we append this row to result
    for keyword in keyword_list:
        if keyword in text:
            result_2019 = result_2019.append(row)
        else:
            continue

result_2019.head()
```

Out[29]:

	Unnamed: 0	headline	url	publisher	date	stock	text_processed	text_ids	year
103	105.0	KraneShares Joins with Nasdaq Dorsey Wright to...	https://www.benzinga.com/node/14643902	GuruFocus	2019-10-23 00:00:00	A	[kraneshares, joins, nasdaq, dorsey, wright, l...	[(0, 1), (1, 1), (2, 1), (3, 2), (4, 1), (5, 1...	2019
990	993.0	Alcoa cut at Gabelli as aluminum may be challe...	https://seekingalpha.com/news/3507410-alcoa-cu...	Seeking Alpha	2019-10-21 00:00:00	AA	[alcoa, cut, gabelli, aluminum, challenged, in...	[(0, 1), (32, 1), (33, 1), (34, 1), (35, 1), (...	2019
1049	1052.0	Copper tumbles as U.S.-China trade dispute weighs	https://seekingalpha.com/news/3469297-copper-t...	Seeking Alpha	2019-06-05 00:00:00	AA	[copper, tumbles, uschina, trade, dispute, wei...	[(27, 1), (39, 1), (40, 1), (41, 1), (42, 1), ...	2019
1123	1126.0	Copper bounces as China eases reserve requirem...	https://seekingalpha.com/news/3420767-copper-b...	Seeking Alpha	2019-01-04 00:00:00	AA	[copper, bounces, china, eases, reserve, requi...	[(0, 1), (39, 1), (41, 1), (44, 1), (45, 1), (...	2019

	Unnamed: 0	headline	url	publisher	date	stock	text_processed	text_ids	year
2886	2910.0	Tesla (TSLA) to Break Ground on Gigafactory Pl...	http://www.zacks.com/stock/news/346025/tesla-t...	Zacks	2019-01-07 00:00:00	AAP	[tesla, tsla, break, ground, gigafactory, plan...	[(0, 1), (89, 1), (90, 1), (91, 1), (92, 1), (...	2019

In [30]:
```
import re
from sklearn import feature_extraction
stop_words = feature_extraction.text.ENGLISH_STOP_WORDS
from nltk.stem import PorterStemmer
from nltk.stem import WordNetLemmatizer

def preprocess(text):
  text = text.lower() #lowercase
  text = re.sub(r'[^\w\s]', '', text) #remove punctuations
  text = re.sub(r'\d+', '', text) #remove numbers
  text = " ".join(text.split()) #stripWhitespace
  text = text.split()
  text = [x for x in text if x not in stop_words] #remove stopwords
  text = [x for x in text if x not in ["star", "starwars","jedi"]] #remove task specific stopwords
  text = " ".join(text)
  # stemmer_ps = PorterStemmer()
  # text = [stemmer_ps.stem(word) for word in text.split()] #stemming
  # text = " ".join(text)
  # lemmatizer = WordNetLemmatizer()
  # text = [lemmatizer.lemmatize(word) for word in text.split()]  #lemmatization
  # text = " ".join(text)
  return(text)
```

In [31]:
```
result_2019['text_processed']=result_2019['headline'].apply(lambda x:preprocess(str(x)))
result_2019['text_processed']=result_2019['text_processed'].apply(lambda x:x.split())
result_2019.head(5)
result_2019.shape
```

Out[31]: (7056, 9)

In [32]:
```
from gensim import corpora
dictionary = corpora.Dictionary(result_2019['text_processed'])
dictionaryDF = pd.DataFrame()
dictionaryDF['id']=dictionary.keys()
dictionaryDF['word']=dictionary.values()
dictionaryDF
```

Out[32]:

	id	word
0	0	china
1	1	dorsey
2	2	joins
3	3	kraneshares
4	4	launch
...	...	...
3265	3265	chains
3266	3266	dhl
3267	3267	ltl
3268	3268	reddaway
3269	3269	youtube

3270 rows × 2 columns

In [33]:
```
dictionary = corpora.Dictionary(result_2019['text_processed'])
result_2019['text_ids']=result_2019['text_processed'].apply(lambda x:dictionary.doc2bow(x))

from gensim import models
num_topics=5
ldamodel = models.ldamodel.LdaModel(result_2019['text_ids'], num_topics = num_topics, id2word=dictionary
topics = ldamodel.print_topics(num_words=4)
for topic in topics:
    print(topic)
```

```
(0, '0.069*"china" + 0.056*"trade" + 0.053*"uschina" + 0.027*"war"')
(1, '0.120*"china" + 0.045*"trade" + 0.020*"war" + 0.014*"chinas"')
(2, '0.095*"china" + 0.036*"trade" + 0.020*"uschina" + 0.016*"wall"')
(3, '0.090*"china" + 0.055*"trade" + 0.037*"chinas" + 0.028*"uschina"')
(4, '0.112*"china" + 0.027*"stocks" + 0.015*"trade" + 0.015*"watch"')
```

In [34]:
```
# 2020
keyword_list = ['Supply Chain','China']

#create an empty dataframe called result
result_2020 = pd.DataFrame()

for index, row in year_2020.iterrows():
    text = row['headline']
    #for each keyword in my keyword list, if keyword is in text
    # then we append this row to result
    for keyword in keyword_list:
        if keyword in text:
            result_2020 = result_2020.append(row)
        else:
            continue

result_2020.head()
```

Out[34]:

	Unnamed: 0	headline	url	publisher	date	stock	text_processed	text_ids	year
2710	2734.0	Auto Stock Roundup: TSLA's China-Made Model 3 ...	http://www.zacks.com/stock/news/700424/auto-st...	Zacks	2020-01-03 00:00:00	AAP	[auto, stock, roundup, tslas, chinamade, model...	[(5, 1), (24, 1), (81, 1), (82, 1), (83, 1), (...	2020
6287	6372.0	UBP Investment Advisors SA Buys iShares Short ...	http://www.gurufocus.com/news/1137997/ubp-inve...	GuruFocus	2020-05-15 00:00:00	ABBV	[ubp, investment, advisors, sa, buys, ishares,...	[(0, 1), (204, 1), (205, 1), (206, 1), (207, 1...	2020
15731	15829.0	CWM Advisors, LLC Buys Accenture PLC, China Te...	http://www.gurufocus.com/news/1131421/cwm-advi...	GuruFocus	2020-05-08 00:00:00	ACN	[cwm, advisors, llc, buys, accenture, plc, chi...	[(0, 1), (204, 1), (206, 1), (364, 1), (479, 1...	2020

	Unnamed: 0	headline	url	publisher	date	stock	text_processed	text_ids	year
20248	20359.0	US-China Escalation Sinks Hong Kong and Hits R...	https://talkmarkets.com/content/us-china-escal...	TalkMarkets	2020-05-22 00:00:00	ACWI	[uschina, escalation, sinks, hong, kong, hits,...	[(43, 1), (58, 1), (243, 1), (524, 1), (525, 1...	2020
20254	20365.0	Risking An Investment War With China	https://talkmarkets.com/content/risking-an-inv...	TalkMarkets	2020-05-15 00:00:00	ACWI	[risking, investment, war, china]	[(0, 1), (209, 1), (529, 1), (530, 1)]	2020

In [35]:
```
import re
from sklearn import feature_extraction
stop_words = feature_extraction.text.ENGLISH_STOP_WORDS
from nltk.stem import PorterStemmer
from nltk.stem import WordNetLemmatizer

def preprocess(text):
  text = text.lower() #lowercase
  text = re.sub(r'[^\w\s]', '', text) #remove punctuations
  text = re.sub(r'\d+', '', text) #remove numbers
  text = " ".join(text.split()) #stripWhitespace
  text = text.split()
  text = [x for x in text if x not in stop_words] #remove stopwords
  text = [x for x in text if x not in ["star", "starwars","jedi"]] #remove task specific stopwords
  text = " ".join(text)
  # stemmer_ps = PorterStemmer()
  # text = [stemmer_ps.stem(word) for word in text.split()] #stemming
  # text = " ".join(text)
  # lemmatizer = WordNetLemmatizer()
  # text = [lemmatizer.lemmatize(word) for word in text.split()]  #lemmatization
  # text = " ".join(text)
  return(text)
```

In [36]:
```
result_2020['text_processed']=result_2020['headline'].apply(lambda x:preprocess(str(x)))
result_2020['text_processed']=result_2020['text_processed'].apply(lambda x:x.split())
result_2020.head(5)
result_2020.shape
```

Out[36]: (680, 9)

In [37]:
```
from gensim import corpora
dictionary = corpora.Dictionary(result_2019['text_processed'])
dictionaryDF = pd.DataFrame()
dictionaryDF['id']=dictionary.keys()
dictionaryDF['word']=dictionary.values()
dictionaryDF
```

Out[37]:

	id	word
0	0	china
1	1	dorsey
2	2	joins
3	3	kraneshares
4	4	launch
...	...	...
3265	3265	chains
3266	3266	dhl
3267	3267	ltl
3268	3268	reddaway
3269	3269	youtube

3270 rows × 2 columns

In [38]:
```
dictionary = corpora.Dictionary(result_2020['text_processed'])
result_2020['text_ids']=result_2020['text_processed'].apply(lambda x:dictionary.doc2bow(x))

from gensim import models
num_topics=5
ldamodel = models.ldamodel.LdaModel(result_2020['text_ids'], num_topics = num_topics, id2word=dictionary
topics = ldamodel.print_topics(num_words=4)
for topic in topics:
    print(topic)
```

```
(0, '0.086*"china" + 0.023*"buys" + 0.016*"ishares" + 0.014*"market"')
(1, '0.069*"china" + 0.031*"coronavirus" + 0.023*"stocks" + 0.011*"buys"')
(2, '0.045*"china" + 0.021*"coronavirus" + 0.020*"market" + 0.019*"uschina"')
(3, '0.082*"china" + 0.015*"stock" + 0.014*"market" + 0.014*"jones"')
(4, '0.048*"china" + 0.026*"ishares" + 0.016*"chinas" + 0.016*"chain"')
```

In [ ]:

In [ ]: