In [8]:
```python
# Check path during tests
import sys
import os
module_path = os.path.abspath(os.path.join('..'))
if module_path not in sys.path:
    sys.path.append(module_path+"/")

# Plib imports
import Plib.News as news
import Plib.DataFarm.IEXdata as datafarm

# Other libraries
import pandas as pd
pd.set_option('display.max_columns', None)
pd.set_option('display.expand_frame_repr', False)
pd.set_option('max_colwidth', -1)
import datetime
from datetime import date, timedelta
import matplotlib.pyplot as plt
import numpy as np
```

In [9]:
```python
import spacy
from spacy import displacy
nlp = spacy.load('en')
```

In [53]:
```python
pd.read_pickle(PkFile)
```

In [18]:
```python
print('Downloaded Records:', len(df))
df.head(2)
```

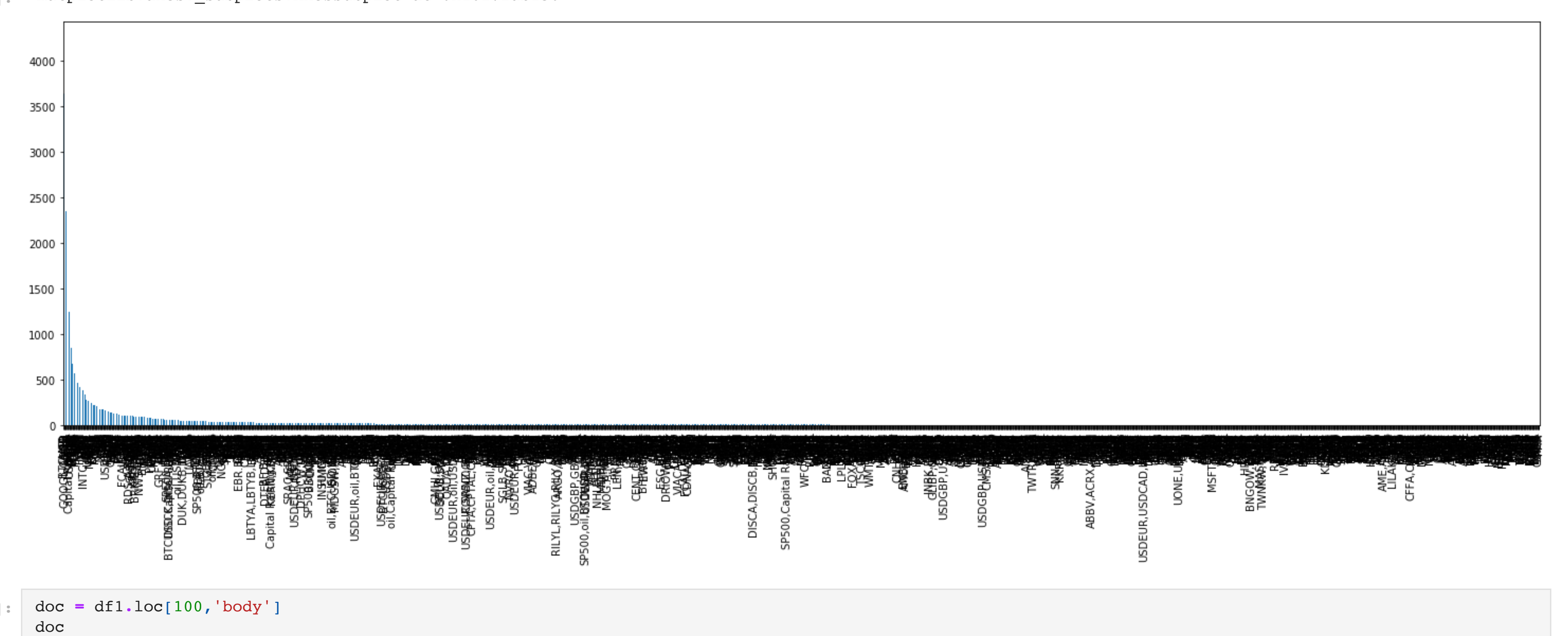
Downloaded Records: 73702

Out[18]:

	Date	Subject	Source	body	relevant
0	2020-10-01	You might need to erase and restore your iPhone if you see these bugs	AAPL	Apple published a new support document on its website informing iPhone owners that they need to reset and restore their phones if they run into certain bugs.	1
1	2020-10-01	FERC Carbon Pricing Conference Highlights State-Federal Divide on Clean Power Policy	EXC	Pricing carbon in federally regulated electricity markets could be more efficient at driving down emissions than clean energy subsidies and mandates in the states that belong to them. But that doesnt mean that states leery of federal intervention in their carbon reduction goals think it can replace what theyre already doing. Thats a key point emerging from Wednesdays carbon pricing conference held by the Federal Energy Regulatory Commission, which regulates the independent system operators (ISOs) and regional transmission organizations (RTOs) that manage transmission networks delivering electricity to about two-thirds of the country. The all-day conference, held at the request of power generators, industry groups and clean energy advocates, yielded broad consensus that pricing carbon is a cost-effective way to drive down emissions while fostering grid reliability. But it also underscored tensions between FERC and those states that say its actions have undermined their clean energy mandates and incentives.	1

In [14]:
```python
plt.figure(figsize=(25,7))
df.Source.value_counts().plot(kind='bar')
```

Out[14]: <matplotlib.axes._subplots.AxesSubplot at 0x10272ab150>



In [32]:
```python
doc = df1.loc[100,'body']
doc
```

Out[32]: 'Bed Bath & Beyond (NASDAQ: BBBY) shares are trading higher on Thursday after the company reported better-than-expected second-quarter EPS and sales results. The company also reported net sales from digital channels grew approximately 88%. Bed Bath & Beyond reported quarterly earnings of $0.50 per share, which beat the Full story available on Benzinga.com'

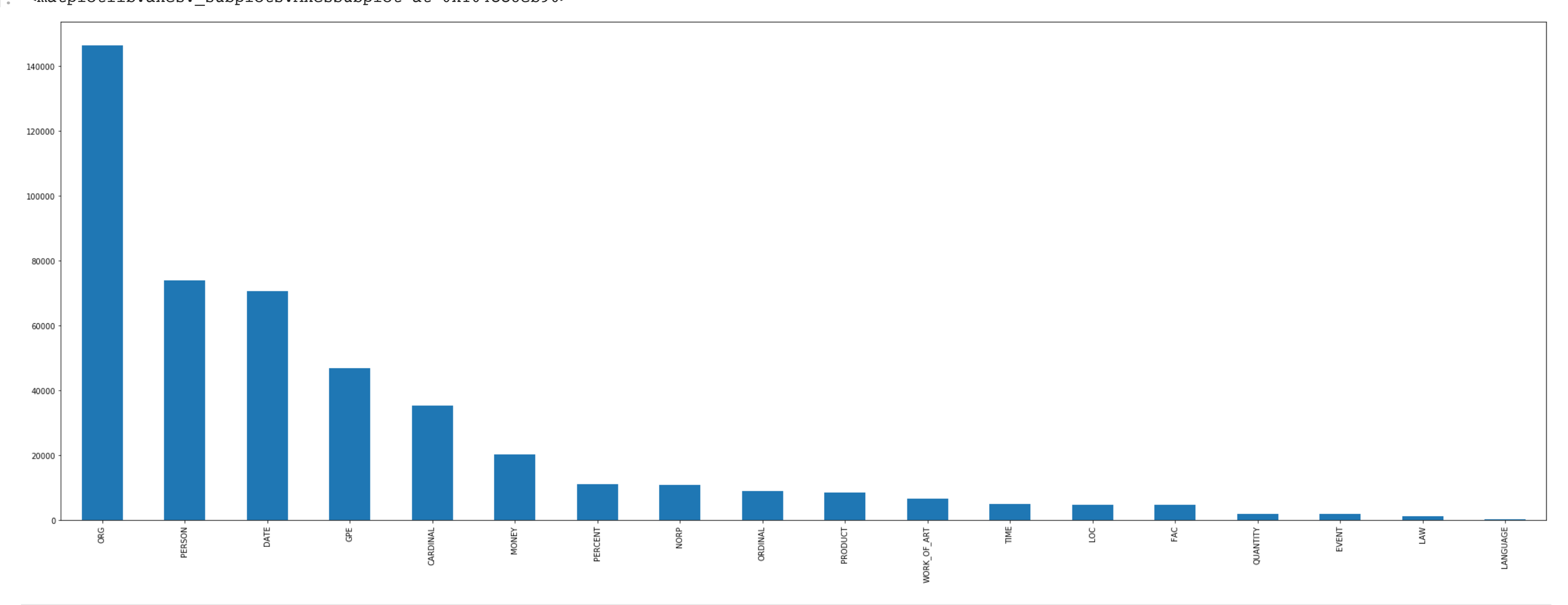
In [33]:
```python
text = nlp(doc)
displacy.render(text,style='ent',jupyter=True)
```

Bed Bath & Beyond ORG (NASDAQ: BBBY) shares are trading higher on Thursday DATE after the company reported better-than-expected second-quarter DATE EPS ORG and sales results. The company also reported net sales from digital channels grew approximately 88% PERCENT . Bed Bath & Beyond ORG reported quarterly DATE earnings of $ 0.50 MONEY per share, which beat the Full story available on Benzinga.com ORG

In [35]:
```python
nlp = spacy.load('en',
                 disable=['parser',
                          'tagger',
                          'textcat'])
from tqdm import tqdm_notebook
frames = []
for i in tqdm_notebook(range(len(df))):
    doc = df.loc[i,'body']
    text_id = df.loc[i,'Source']
    doc = nlp(doc)
    ents = [(e.text, e.start_char, e.end_char, e.label_)
        for e in doc.ents
        if len(e.text.strip(' --')) > 0]
    frame = pd.DataFrame(ents)
    frame['Source'] = text_id
    frames.append(frame)
npf = pd.concat(frames)
npf.columns = ['Text','Start','Stop','Type','Source']
```

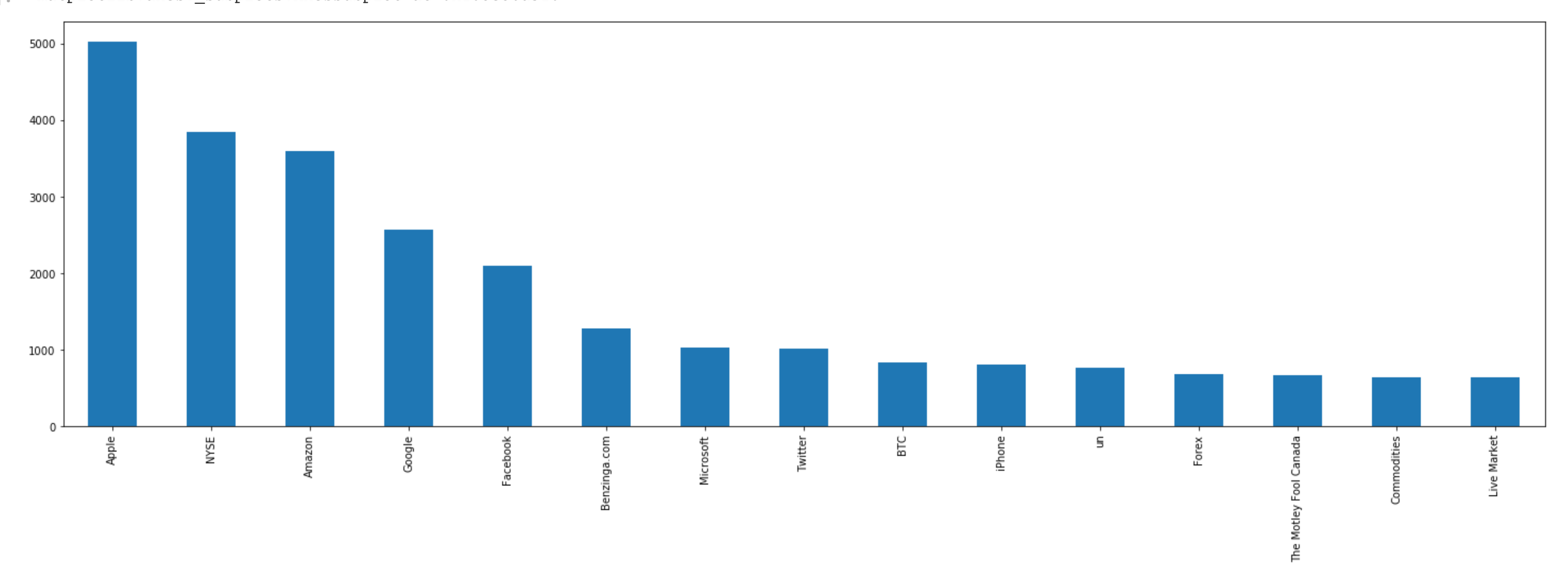
In [36]:
```python
plt.figure(figsize=(35,12))
npf.Type.value_counts().plot(kind='bar')
```

Out[36]: <matplotlib.axes._subplots.AxesSubplot at 0x104cc8eb90>



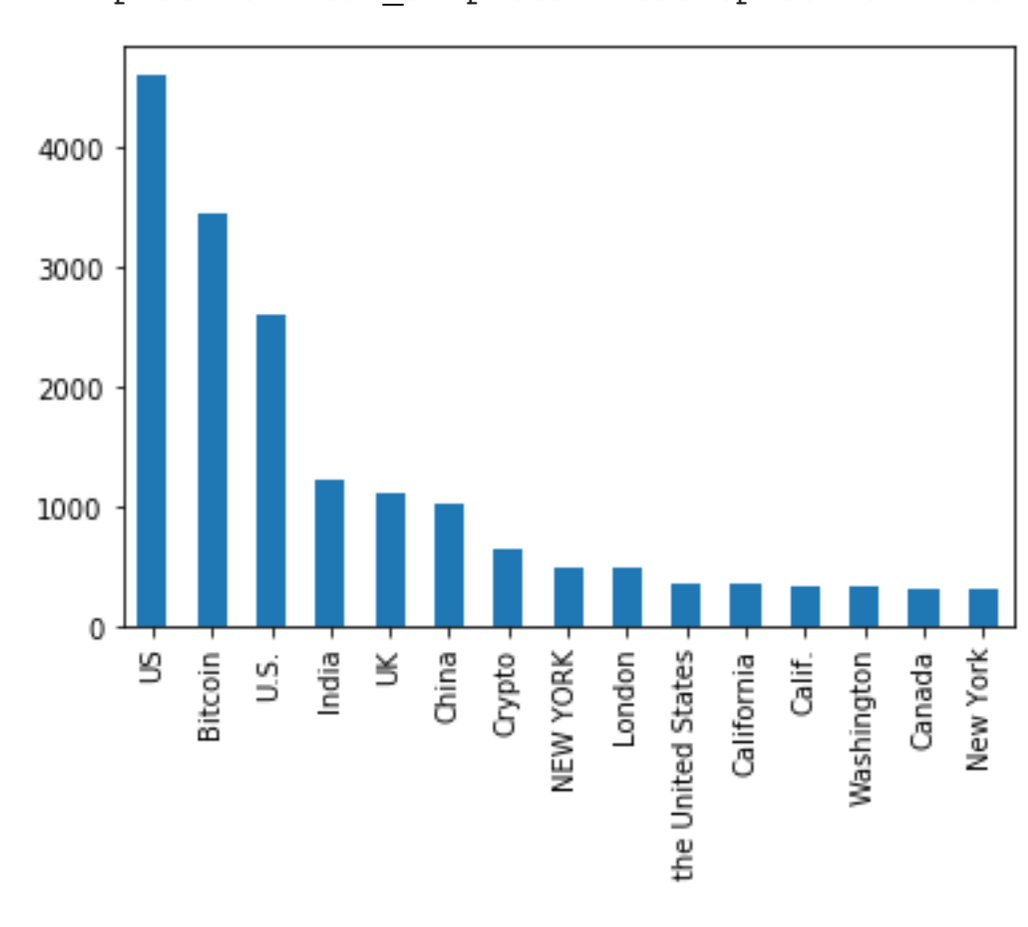
In [37]:
```python
plt.figure(figsize=(25,7))
orgs = npf[npf.Type == 'ORG']
orgs.Text.value_counts()[:15].plot(kind='bar')
```

Out[37]: <matplotlib.axes._subplots.AxesSubplot at 0x103e36d510>



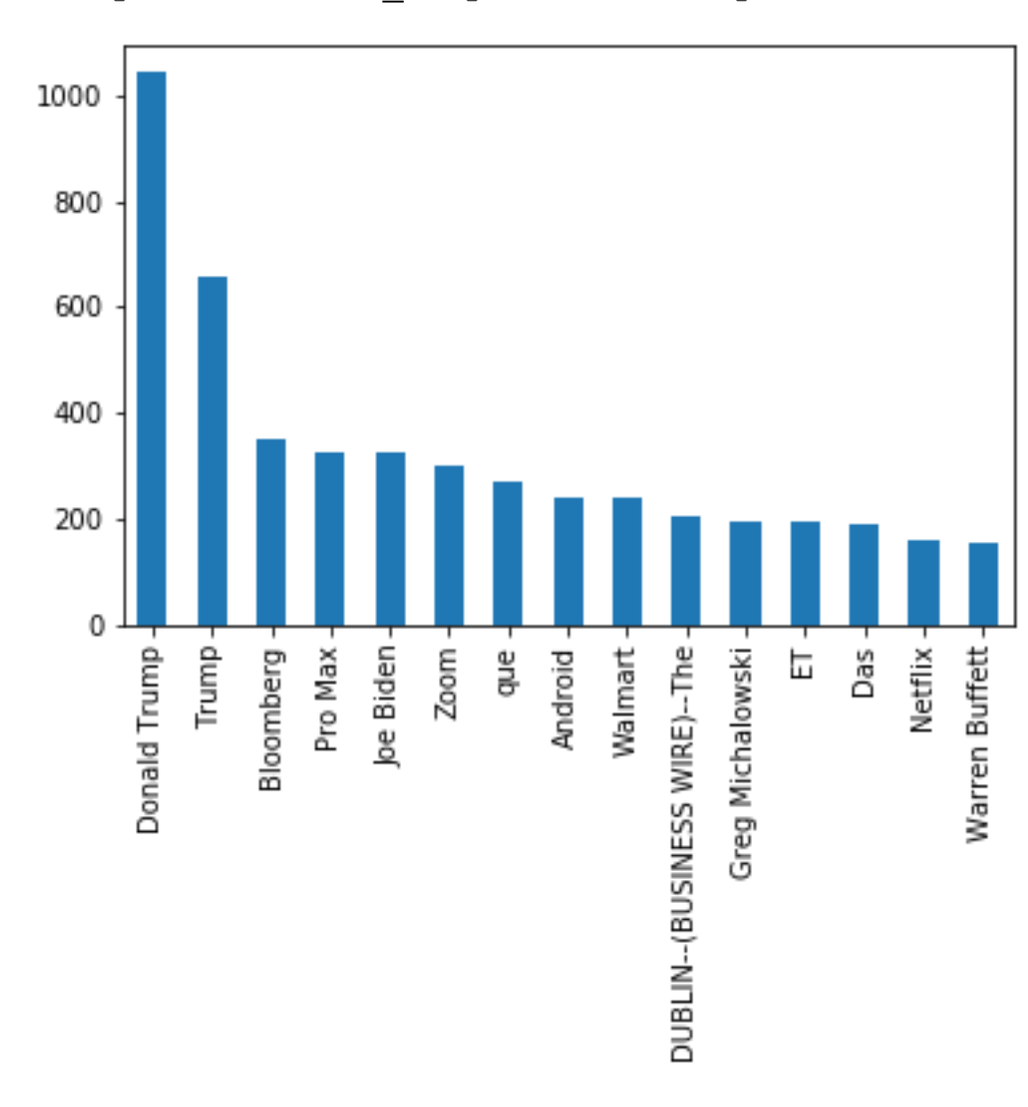
In [38]:
```python
orgs = npf[npf.Type == 'GPE']
orgs.Text.value_counts()[:15].plot(kind='bar')
```

Out[38]: <matplotlib.axes._subplots.AxesSubplot at 0x1060f59890>



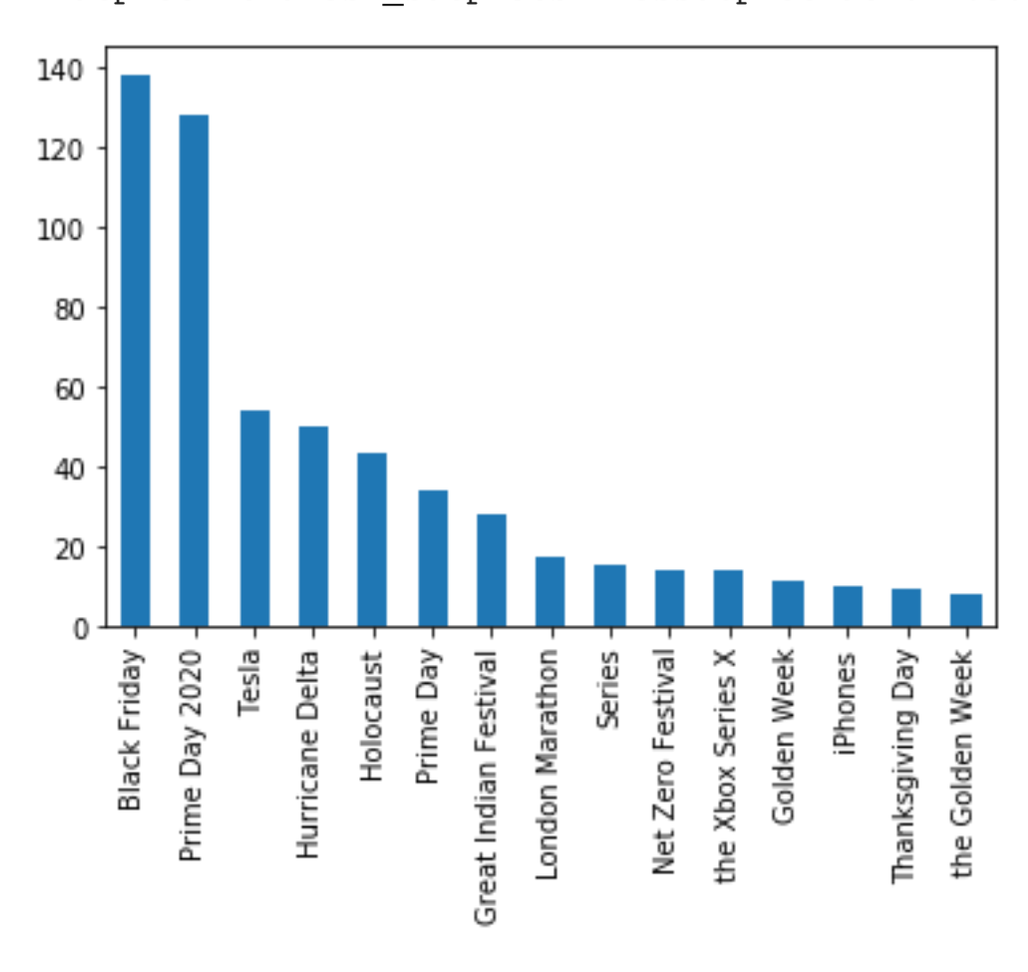
In [39]:
```python
orgs = npf[npf.Type == 'PERSON']
orgs.Text.value_counts()[:15].plot(kind='bar')
```

Out[39]: <matplotlib.axes._subplots.AxesSubplot at 0x103b6b6e90>



In [40]:
```python
orgs = npf[npf.Type == 'EVENT']
orgs.Text.value_counts()[:15].plot(kind='bar')
```

Out[40]: <matplotlib.axes._subplots.AxesSubplot at 0x10501c7f90>



In [43]:
```python
import nltk
nltk.download('vader_lexicon')
```

```
[nltk_data] Downloading package vader_lexicon to
[nltk_data]     /Users/rob/nltk_data...
[nltk_data]   Package vader_lexicon is already up-to-date!
```

Out[43]: True

In [44]:
```python
df1=df
df1.head(2)
```

Out[44]:

	Date	Subject	Source	body	relevant	Score
0	2020-10-01	You might need to erase and restore your iPhone if you see these bugs	AAPL	Apple published a new support document on its website informing iPhone owners that they need to reset and restore their phones if they run into certain bugs.	1	0.2960
1	2020-10-01	FERC Carbon Pricing Conference Highlights State-Federal Divide on Clean Power Policy	EXC	Pricing carbon in federally regulated electricity markets could be more efficient at driving down emissions than clean energy subsidies and mandates in the states that belong to them. But that doesnt mean that states leery of federal intervention in their carbon reduction goals think it can replace what theyre already doing. Thats a key point emerging from Wednesdays carbon pricing conference held by the Federal Energy Regulatory Commission, which regulates the independent system operators (ISOs) and regional transmission organizations (RTOs) that manage transmission networks delivering electricity to about two-thirds of the country. The all-day conference, held at the request of power generators, industry groups and clean energy advocates, yielded broad consensus that pricing carbon is a cost-effective way to drive down emissions while fostering grid reliability. But it also underscored tensions between FERC and those states that say its actions have undermined their clean energy mandates and incentives.	1	0.4019

In [45]:
```python
from nltk.sentiment.vader import SentimentIntensityAnalyzer as SIA

results = []

for mtext in df1['Subject']:
    pol_score = SIA().polarity_scores(mtext) # run analysis
    pol_score['headline'] = mtext # add headlines for viewing
    results.append(pol_score)
```

In [46]:
```python
df1['Score'] = pd.DataFrame(results)['compound']
df1.head(5)
```

Out[46]:

	Date	Subject	Source	body	relevant	Score
0	2020-10-01	You might need to erase and restore your iPhone if you see these bugs	AAPL	Apple published a new support document on its website informing iPhone owners that they need to reset and restore their phones if they run into certain bugs.	1	0.2960
1	2020-10-01	FERC Carbon Pricing Conference Highlights State-Federal Divide on Clean Power Policy	EXC	Pricing carbon in federally regulated electricity markets could be more efficient at driving down emissions than clean energy subsidies and mandates in the states that belong to them. But that doesnt mean that states leery of federal intervention in their carbon reduction goals think it can replace what theyre already doing. Thats a key point emerging from Wednesdays carbon pricing conference held by the Federal Energy Regulatory Commission, which regulates the independent system operators (ISOs) and regional transmission organizations (RTOs) that manage transmission networks delivering electricity to about two-thirds of the country. The all-day conference, held at the request of power generators, industry groups and clean energy advocates, yielded broad consensus that pricing carbon is a cost-effective way to drive down emissions while fostering grid reliability. But it also underscored tensions between FERC and those states that say its actions have undermined their clean energy mandates and incentives.	1	0.4019
2	2020-10-01	Macerich Schedules Third Quarter 2020 Earnings Release And Conference Call	MAC	SANTA MONICA, Calif., Oct. 1, 2020 /PRNewswire/ -- WHAT: Macerich (NYSE: MAC) Schedules Third Quarter 2020 Earnings Release WHEN: Earnings Results will be released before market open on Thursday, November 5, 2020. Management will hold a conference call at 10:00 am Pacific Time (1:00 pm	1	0.0000
3	2020-10-01	Extracorporeal Membrane Oxygenation (ECMO) Market 2020-2027 study and future prospects including key players LivaNova PLC, XENIOS AG, Terumo Corporation, MicroPort Scientific Corporation, etc.	LIVN	Reports and Data has released a new market report titled Global Extracorporeal Membrane Oxygenation (ECMO) Market which is an extensive study of the Extracorporeal Membrane Oxygenation (ECMO) market spanning over 100+ pages with pictorial representations of key statistical data. The	1	0.2960
4	2020-10-01	Nordstrom Launches BEAUTYCYCLE Nationwide	JWN	SEATTLE, Oct. 1, 2020 /PRNewswire/ -- BEAUTYCYCLE, the first beauty take-back and recycling program accepting all brands of beauty packaging at a major retailer, launches today at Nordstrom. Each year, more than 120 billion units of plastic packaging are produced by the beauty industry,	1	0.0000

In [47]:
```python
import Plib.DataFarm.Quandl as q
data1=q.getQDLData(symbol='CHRIS/CME_SP1', start_date='2020-09-01', end_date='2020-10-24')
# calculate daily returns
data1['Returns'] = 100*(np.log(data1.Adjusted_close) - np.log(data1.Adjusted_close.shift(1)))
data1=data1.reset_index()
data1['Date2'] = pd.to_datetime(data1['Date']).dt.date
data1.drop('Date', axis=1, inplace=True)
data1.rename(columns={'Date2':'Date'}, inplace=True)
data1=data1.set_index('Date')
data1.tail(7)
```

Out[47]:

Date	Open	High	Low	Close	Adjusted_close	openInterest	Volume	Returns
2020-10-15	3445.3	3483.9	3434.6	3475.4	3475.4	16050.0	955.0	-0.158130
2020-10-16	3483.0	3490.7	3463.6	3462.2	3462.2	16214.0	4723.0	-0.380536
2020-10-19	NaN	3495.7	3428.8	3422.7	3422.7	15465.0	1754.0	-1.147451
2020-10-20	NaN	3450.9	3419.3	3432.2	3432.2	15911.0	500.0	0.277174
2020-10-21	3428.3	3456.7	3421.3	3432.5	3432.5	16011.0	504.0	0.008740
2020-10-22	3428.3	3450.4	3405.3	3449.2	3449.2	16212.0	1748.0	0.485346
2020-10-23	NaN	3461.7	3437.8	3451.8	3451.8	15880.0	1300.0	0.075351

In [48]:
```python
# creates a daily score by summing the scores of the individual articles in each day
df2 = df1.groupby(['Date']).sum()
df2=df2[['Score']]
df2.head(15)
```

Out[48]:

Date	Score
2020-10-01	183.8323
2020-10-02	122.7037
2020-10-03	44.2328
2020-10-04	76.9506
2020-10-05	740.2076
2020-10-06	612.7881
2020-10-07	435.1747
2020-10-08	354.6099
2020-10-09	281.9905
2020-10-10	50.3463
2020-10-11	69.8771
2020-10-12	671.1880
2020-10-13	824.7670
2020-10-14	461.9290
2020-10-15	242.6412

In [49]:
```python
#Lag the sentiment score
df2['Lagged_Score'] = df2.shift(2)
df2.head(15)
```

Out[49]:

Date	Score	Lagged_Score
2020-10-01	183.8323	NaN
2020-10-02	122.7037	NaN
2020-10-03	44.2328	183.8323
2020-10-04	76.9506	122.7037
2020-10-05	740.2076	44.2328
2020-10-06	612.7881	76.9506
2020-10-07	435.1747	740.2076
2020-10-08	354.6099	612.7881
2020-10-09	281.9905	435.1747
2020-10-10	50.3463	354.6099
2020-10-11	69.8771	281.9905
2020-10-12	671.1880	50.3463
2020-10-13	824.7670	69.8771
2020-10-14	461.9290	671.1880
2020-10-15	242.6412	824.7670

In [50]:
```python
df3 = pd.merge(data1[['Returns']],df2[['Lagged_Score']], left_index=True, right_index=True, how='left')
df3.fillna(0, inplace=True)
# replace NaN with 0 permanently
df3.tail(24)
```

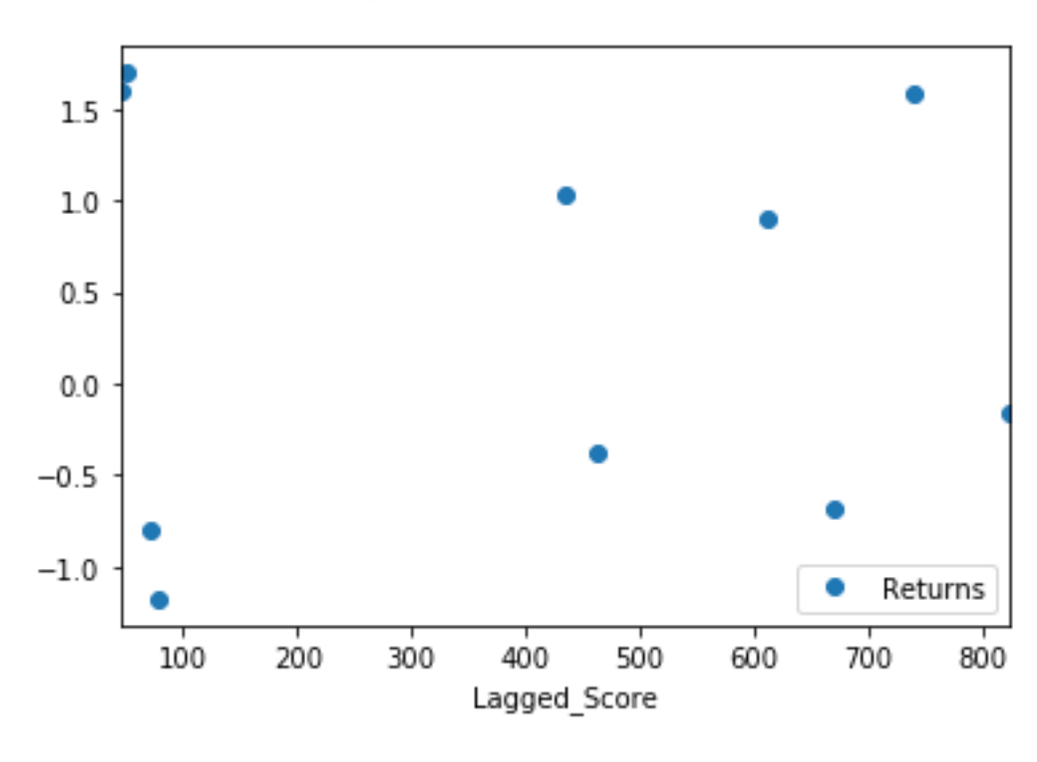
Out[50]:

Date	Returns	Lagged_Score
2020-09-22	0.736192	0.0000
2020-09-23	-2.085674	0.0000
2020-09-24	0.210227	0.0000
2020-09-25	1.511070	0.0000
2020-09-28	1.769904	0.0000
2020-09-29	-0.368280	0.0000
2020-09-30	0.547438	0.0000
2020-10-01	0.467284	0.0000
2020-10-02	-0.849876	0.0000
2020-10-05	1.595376	44.2328
2020-10-06	-1.176990	76.9506
2020-10-07	1.585832	740.2076
2020-10-08	0.897103	612.7881
2020-10-09	1.036069	435.1747
2020-10-12	1.698561	50.3463
2020-10-13	-0.795730	69.8771
2020-10-14	-0.684258	671.1880
2020-10-15	-0.158130	824.7670
2020-10-16	-0.380536	461.9290
2020-10-19	-1.147451	0.0000
2020-10-20	0.277174	0.0000
2020-10-21	0.008740	0.0000
2020-10-22	0.485346	0.0000
2020-10-23	0.075351	0.0000

In [51]:
```python
#remove scores between -0.5 amd 0.5
df4 = df3[(df3['Lagged_Score'] > 0.5) | (df3['Lagged_Score'] < -0.5)]
```

In [52]:
```python
df4.plot(x="Lagged_Score", y="Returns",style="o")
df4['Returns'].corr(df4['Lagged_Score'])
```

Out[52]: -0.0074820201270332344



In [ ]: