In [2]:
```
import pandas as ps
import requests
import bs4
from bs4 import BeautifulSoup
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
%load_ext autoreload
%autoreload 2
import spacy
from spacy.displacy.render import EntityRenderer
from IPython.core.display import display, HTML
import scispacy
import en_core_sci_sm
from spacy import displacy
```

```
The autoreload extension is already loaded. To reload it, use:
  %reload_ext autoreload
```

In [6]:
```
url ="https://economictimes.indiatimes.com/news/sports"
```

In [7]:
```
headers = {
    'User-Agent':"Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/85.0.4183.102 Safari/537.36"
}
```

In [8]:
```
s = requests.get(url,{'headers':headers})
```

In [9]:
```
soup=bs4.BeautifulSoup(s.text,'html.parser')
```

In [10]:
```
Headline = soup.find_all('div',{'class':'eachStory'})[0].find_all('a')[-1].text
```

In [11]:
```
Time=soup.find_all('div',{'class':'eachStory'})[0].find_all('time')[-1].text
```

In [12]:
```
Content=soup.find_all('div',{'class':'eachStory'})[0].find_all('p')[-1].text
```

In [13]:
```
NewsType=soup.find_all('div',{'class':'timeEdition tac'})[0].find_all('a')[-1].text
```

In [14]:
```
data=[[url,NewsType,Headline,Time,Content]]
```

In [15]:
```
data
```

Out[15]:
```
[['https://economictimes.indiatimes.com/news/sports',
  'E-Paper',
  'What moving the Indian Premier League to U.A.E. means for cricket',
  'Sep 13, 2020, 10:10 AM IST',
  'The UAE isn't one of the top cricketing jurisdictions that play five-day games called "tests," typically limited to England and a clutch of her for
mer colonies. It does, however, excel as a transportation hub and a venue for international events. And IPL is a global business — valued at some $6.8
billion — headquartered in Mumbai.']]
```

In [16]:
```
df=ps.DataFrame(data, columns=['url','NewsType','Headline','Time','Content'])
```

In [17]:
```
df
```

Out[17]:

	url	NewsType	Headline	Time	Content
0	https://economictimes.indiatimes.com/news/sports	E-Paper	What moving the Indian Premier League to U.A.E...	Sep 13, 2020, 10:10 AM IST	The UAE isn't one of the top cricketing jurisd...

In [18]:
```
for i in range(1,5):
    url ="https://economictimes.indiatimes.com/news/sports"
    s = requests.get(url,{'headers':headers})
    if(s.status_code==200):
        print("Data is fetched successfuly",i)
        soup=bs4.BeautifulSoup(s.text,'html.parser')
        Headline = soup.find_all('div',{'class':'eachStory'})[i].find_all('a')[-1].text
        DatePosted=soup.find_all('div',{'class':'eachStory'})[i].find_all('time')[-1].text
        ArticleContent=soup.find_all('div',{'class':'eachStory'})[i].find_all('p')[-1].text
        NewsType=soup.find_all('div',{'class':'timeEdition tac'})[0].find_all('a')[-1].text
        data.insert(i,[url,NewsType,Headline,DatePosted,ArticleContent])
    else:
        print("Url not found",i)
df=ps.DataFrame(data, columns=['url','NewsType','Headline','DatePosted','ArticleContent'])
df.to_csv('newsportal_12sep.csv')
```

```
Data is fetched successfuly 1
Data is fetched successfuly 2
Data is fetched successfuly 3
Data is fetched successfuly 4
```

In [16]:
```
df = ps.read_csv('newsportal_12sep.csv')
df.head(5)
```

Out[16]:

	Unnamed: 0	url	NewsType	Headline	DatePosted	ArticleContent
0	0	https://economictimes.indiatimes.com/news/sports	E-Paper	What moving the Indian Premier League to U.A.E...	Sep 13, 2020, 10:10 AM IST	The UAE isn't one of the top cricketing jurisd...
1	1	https://economictimes.indiatimes.com/news/sports	E-Paper	Japan's Naomi Osaka beats Victoria Azarenka to...	Sep 13, 2020, 06:50 AM IST	Osaka, the fourth seed, overcame her unseeded ...
2	2	https://economictimes.indiatimes.com/news/sports	E-Paper	Star Neymar back in PSG squad for home game vs...	Sep 12, 2020, 09:44 PM IST	Captain Marquinhos and striker Mauro Icardi al...
3	3	https://economictimes.indiatimes.com/news/sports	E-Paper	Can't be held online, BCCI indefinitely postpo...	Sep 11, 2020, 05:24 PM IST	After taking a legal opinion on the subject, B...
4	4	https://economictimes.indiatimes.com/news/sports	E-Paper	Cricket takes back seat as lieutenant colonel ...	Sep 10, 2020, 02:02 PM IST	An excited Dhoni promptly expressed his deligh...

In [17]:
```
df.dtypes
```

Out[17]:
```
Unnamed: 0         int64
url               object
NewsType          object
Headline          object
DatePosted        object
ArticleContent    object
dtype: object
```

In [18]:
```
df.shape
```

Out[18]:
```
(17, 6)
```

In [19]:
```
duplicate_rows_df = df[df.duplicated()]
print('number of duplicate rows:', duplicate_rows_df.shape)
```

```
number of duplicate rows: (0, 6)
```

In [20]:
```
df = df.drop_duplicates()
df.head(5)
```

Out[20]:

	Unnamed: 0	url	NewsType	Headline	DatePosted	ArticleContent
0	0	https://economictimes.indiatimes.com/news/sports	E-Paper	What moving the Indian Premier League to U.A.E...	Sep 13, 2020, 10:10 AM IST	The UAE isn't one of the top cricketing jurisd...
1	1	https://economictimes.indiatimes.com/news/sports	E-Paper	Japan's Naomi Osaka beats Victoria Azarenka to...	Sep 13, 2020, 06:50 AM IST	Osaka, the fourth seed, overcame her unseeded ...
2	2	https://economictimes.indiatimes.com/news/sports	E-Paper	Star Neymar back in PSG squad for home game vs...	Sep 12, 2020, 09:44 PM IST	Captain Marquinhos and striker Mauro Icardi al...
3	3	https://economictimes.indiatimes.com/news/sports	E-Paper	Can't be held online, BCCI indefinitely postpo...	Sep 11, 2020, 05:24 PM IST	After taking a legal opinion on the subject, B...
4	4	https://economictimes.indiatimes.com/news/sports	E-Paper	Cricket takes back seat as lieutenant colonel ...	Sep 10, 2020, 02:02 PM IST	An excited Dhoni promptly expressed his deligh...

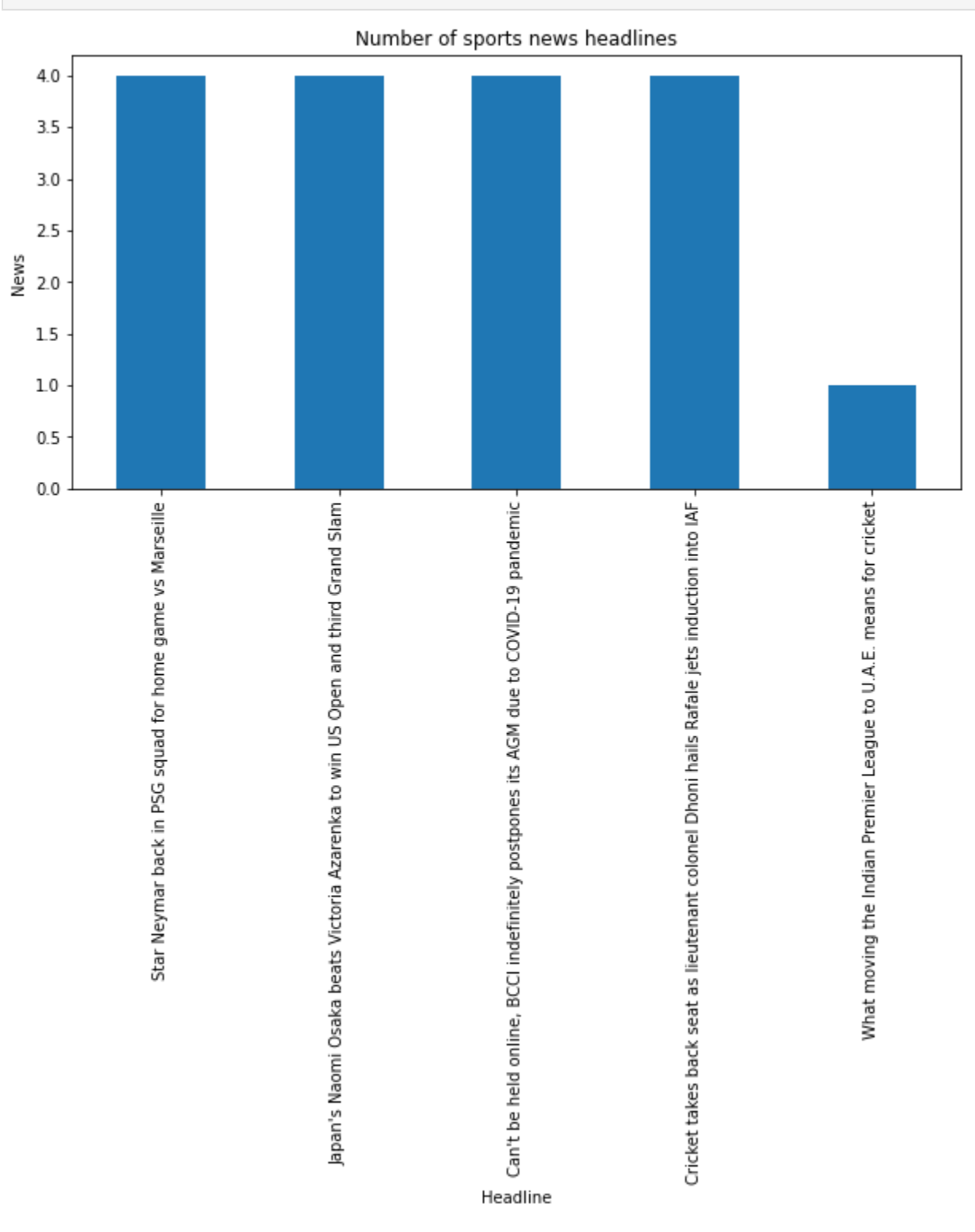
In [21]:
```
df.shape
```

Out[21]:
```
(17, 6)
```

In [22]:
```
df = df.dropna()
df.count()
```

Out[22]:
```
Unnamed: 0        17
url               17
NewsType          17
Headline          17
DatePosted        17
ArticleContent    17
dtype: int64
```

In [23]:
```
# Plotting a Histogram
df.Headline.value_counts().nlargest(40).plot(kind='bar', figsize=(10,5))
plt.title('Number of sports news headlines')
plt.ylabel('News ')
plt.xlabel('Headline');
```



In [24]:
```
df[df.duplicated()]
```

Out[24]:

Unnamed: 0	url	NewsType	Headline	DatePosted	ArticleContent

In [25]:
```
lower = lambda x: x.lower()
```

In [26]:
```
df = ps.DataFrame(df['ArticleContent'].apply(lower))
df.columns = ['text']
display(df)
```

	text
0	the uae isn't one of the top cricketing jurisd...
1	osaka, the fourth seed, overcame her unseeded ...
2	captain marquinhos and striker mauro icardi al...
3	after taking a legal opinion on the subject, b...
4	an excited dhoni promptly expressed his deligh...
...	...
12	an excited dhoni promptly expressed his deligh...
13	osaka, the fourth seed, overcame her unseeded ...
14	captain marquinhos and striker mauro icardi al...
15	after taking a legal opinion on the subject, b...
16	an excited dhoni promptly expressed his deligh...

17 rows × 1 columns

In [27]:
```
def custom_render(doc, df, column, options={}, page=False, minify=False, idx=0):
    """Overload the spaCy built-in rendering to allow custom part-of-speech (POS) tags.

    Keyword arguments:
    doc -- a spaCy nlp doc object
    df -- a pandas dataframe object
    column -- the name of of a column of interest in the dataframe
    options -- various options to feed into the spaCy renderer, including colors
    page -- rendering markup as full HTML page (default False)
    minify -- for compact HTML (default False)
    idx -- index for specific query or doc in dataframe (default 0)

    """
    renderer, converter = EntityRenderer, parse_custom_ents
    renderer = renderer(options=options)
    parsed = [converter(doc, df=df, idx=idx, column=column)]
    html = renderer.render(parsed, page=page, minify=minify).strip()
    return display(HTML(html))

def parse_custom_ents(doc, df, idx, column):
    """Parse custom entity types that aren't in the original spaCy module.

    Keyword arguments:
    doc -- a spaCy nlp doc object
    df -- a pandas dataframe object
    idx -- index for specific query or doc in dataframe
    column -- the name of of a column of interest in the dataframe

    """
    if column in df.columns:
        entities = df[column][idx]
        ents = [{'start': ent[1], 'end': ent[2], 'label': ent[3]}
                for ent in entities]
    else:
        ents = [{'start': ent.start_char, 'end': ent.end_char, 'label': ent.label_}
            for ent in doc.ents]
    return {'text': doc.text, 'ents': ents, 'title': None}

def render_entities(idx, df, options={}, column='named_ents'):
    """A wrapper function to get text from a dataframe and render it visually in jupyter notebooks

    Keyword arguments:
    idx -- index for specific query or doc in dataframe (default 0)
    df -- a pandas dataframe object
    options -- various options to feed into the spaCy renderer, including colors
    column -- the name of of a column of interest in the dataframe (default 'named_ents')

    """
    text = df['text'][idx]
    custom_render(nlp(text), df=df, column=column, options=options, idx=idx)
```

In [35]:
```
# colors for additional part of speech tags we want to visualize
options = {
    'colors': {'COMPOUND': '#FE6BFE', 'PROPN': '#18CFE6', 'NOUN': '#18CFE6', 'NP': '#1EECA6', 'ENTITY': '#FF8800'}
}
```

In [36]:
```
ps.set_option('display.max_rows', 10) # edit how jupyter will render our pandas dataframes
ps.options.mode.chained_assignment = None # prevent warning about working on a copy of a dataframe
```

In [41]:
```
nlp = spacy.load('en_core_sci_sm')
```

In [42]:
```
def extract_named_ents(text):
    """Extract named entities, and beginning, middle and end idx using spaCy's out-of-the-box model.

    Keyword arguments:
    text -- the actual text source from which to extract entities

    """
    return [(ent.text, ent.start_char, ent.end_char, ent.label_) for ent in nlp(text).ents]

def add_named_ents(df):
    """Create new column in data frame with named entity tuple extracted.

    Keyword arguments:
    df -- a dataframe object

    """
    df['named_ents'] = df['text'].apply(extract_named_ents)
```

In [43]:
```
add_named_ents(df)
display(df)
```

	text	named_ents
0	the uae isn't one of the top cricketing jurisd...	[(uae, 4, 7, ENTITY), (jurisdictions, 40, 53, ...
1	osaka, the fourth seed, overcame her unseeded ...	[(osaka, 0, 5, ENTITY), (seed, 18, 22, ENTITY)...
2	captain marquinhos and striker mauro icardi al...	[(captain, 0, 7, ENTITY), (marquinhos, 8, 18, ...
3	after taking a legal opinion on the subject, b...	[(legal opinion, 15, 28, ENTITY), (subject, 36...
4	an excited dhoni promptly expressed his deligh...	[(excited, 3, 10, ENTITY), (dhoni, 11, 16, ENT...
...	...	...
12	an excited dhoni promptly expressed his deligh...	[(excited, 3, 10, ENTITY), (dhoni, 11, 16, ENT...
13	osaka, the fourth seed, overcame her unseeded ...	[(osaka, 0, 5, ENTITY), (seed, 18, 22, ENTITY)...
14	captain marquinhos and striker mauro icardi al...	[(captain, 0, 7, ENTITY), (marquinhos, 8, 18, ...
15	after taking a legal opinion on the subject, b...	[(legal opinion, 15, 28, ENTITY), (subject, 36...
16	an excited dhoni promptly expressed his deligh...	[(excited, 3, 10, ENTITY), (dhoni, 11, 16, ENT...

17 rows × 2 columns

In [44]:
```
column = 'named_ents'
render_entities(9, df, options=options, column=column)
```

osaka ENTITY , the fourth seed ENTITY , overcame her unseeded opponent 1-6, 6-3, 6-3 in 1hr 53min inside a near-empty ENTITY arthur ENTITY ashe stadium ENTITY at flushing meadows.

In [46]:
```
def extract_nouns(text):
    """Extract a few types of nouns, and beginning, middle and end idx using spaCy's POS (part of speech) tagger.

    Keyword arguments:
    text -- the actual text source from which to extract entities

    """
    keep_pos = ['PROPN', 'NOUN']
    return [(tok.text, tok.idx, tok.idx+len(tok.text), tok.pos_) for tok in nlp(text) if tok.pos_ in keep_pos]

def add_nouns(df):
    """Create new column in data frame with nouns extracted.

    Keyword arguments:
    df -- a dataframe object

    """
    df['nouns'] = df['text'].apply(extract_nouns)
```

In [47]:
```
add_nouns(df)
display(df)
```

	text	named_ents	nouns
0	the uae isn't one of the top cricketing jurisd...	[(uae, 4, 7, ENTITY), (jurisdictions, 40, 53, ...	[(uae, 4, 7, NOUN), (jurisdictions, 40, 53, NO...
1	osaka, the fourth seed, overcame her unseeded ...	[(osaka, 0, 5, ENTITY), (seed, 18, 22, ENTITY)...	[(seed, 18, 22, NOUN), (opponent, 46, 54, NOUN...
2	captain marquinhos and striker mauro icardi al...	[(captain, 0, 7, ENTITY), (marquinhos, 8, 18, ...	[(striker, 23, 30, NOUN), (mauro, 31, 36, NOUN...
3	after taking a legal opinion on the subject, b...	[(legal opinion, 15, 28, ENTITY), (subject, 36...	[(opinion, 21, 28, NOUN), (subject, 36, 43, NO...
4	an excited dhoni promptly expressed his deligh...	[(excited, 3, 10, ENTITY), (dhoni, 11, 16, ENT...	[(dhoni, 11, 16, NOUN), (delight, 40, 47, NOUN...
...	...	...	...
12	an excited dhoni promptly expressed his deligh...	[(excited, 3, 10, ENTITY), (dhoni, 11, 16, ENT...	[(dhoni, 11, 16, NOUN), (delight, 40, 47, NOUN...
13	osaka, the fourth seed, overcame her unseeded ...	[(osaka, 0, 5, ENTITY), (seed, 18, 22, ENTITY)...	[(seed, 18, 22, NOUN), (opponent, 46, 54, NOUN...
14	captain marquinhos and striker mauro icardi al...	[(captain, 0, 7, ENTITY), (marquinhos, 8, 18, ...	[(striker, 23, 30, NOUN), (mauro, 31, 36, NOUN...
15	after taking a legal opinion on the subject, b...	[(legal opinion, 15, 28, ENTITY), (subject, 36...	[(opinion, 21, 28, NOUN), (subject, 36, 43, NO...
16	an excited dhoni promptly expressed his deligh...	[(excited, 3, 10, ENTITY), (dhoni, 11, 16, ENT...	[(dhoni, 11, 16, NOUN), (delight, 40, 47, NOUN...

17 rows × 3 columns

In [48]:
```
column = 'nouns'
render_entities(0, df, options=options, column=column)
```

the uae NOUN isn't one of the top cricketing jurisdictions NOUN that play five-day games NOUN called " tests NOUN ," typically limited to england and a clutch NOUN of her former colonies NOUN . it does, however, excel NOUN as a transportation NOUN hub NOUN and a venue NOUN for international events NOUN . and ipl NOUN is a global business NOUN — NOUN valued at some $6.8 billion — headquartered in mumbai NOUN .

In [49]:
```
def extract_named_nouns(row_series):
    """Combine nouns and non-numerical entities.

    Keyword arguments:
    row_series -- a Pandas Series object

    """
    ents = set()
    idxs = set()
    # remove duplicates and merge two lists together
    for noun_tuple in row_series['nouns']:
        for named_ents_tuple in row_series['named_ents']:
            if noun_tuple[1] == named_ents_tuple[1]:
                idxs.add(noun_tuple[1])
                ents.add(named_ents_tuple)
        if noun_tuple[1] not in idxs:
            ents.add(noun_tuple)

    return sorted(list(ents), key=lambda x: x[1])

def add_named_nouns(df):
    """Create new column in data frame with nouns and named ents.

    Keyword arguments:
    df -- a dataframe object

    """
    df['named_nouns'] = df.apply(extract_named_nouns, axis=1)
```

In [50]:
```
add_named_nouns(df)
display(df)
```

	text	named_ents	nouns	named_nouns
0	the uae isn't one of the top cricketing jurisd...	[(uae, 4, 7, ENTITY), (jurisdictions, 40, 53, ...	[(uae, 4, 7, NOUN), (jurisdictions, 40, 53, NO...	[(uae, 4, 7, ENTITY), (jurisdictions, 40, 53, ...
1	osaka, the fourth seed, overcame her unseeded ...	[(osaka, 0, 5, ENTITY), (seed, 18, 22, ENTITY)...	[(seed, 18, 22, NOUN), (opponent, 46, 54, NOUN...	[(seed, 18, 22, ENTITY), (opponent, 46, 54, NO...
2	captain marquinhos and striker mauro icardi al...	[(captain, 0, 7, ENTITY), (marquinhos, 8, 18, ...	[(striker, 23, 30, NOUN), (mauro, 31, 36, NOUN...	[(striker mauro, 23, 36, ENTITY), (mauro, 31, ...
3	after taking a legal opinion on the subject, b...	[(legal opinion, 15, 28, ENTITY), (subject, 36...	[(opinion, 21, 28, NOUN), (subject, 36, 43, NO...	[(opinion, 21, 28, NOUN), (subject, 36, 43, EN...
4	an excited dhoni promptly expressed his deligh...	[(excited, 3, 10, ENTITY), (dhoni, 11, 16, ENT...	[(dhoni, 11, 16, NOUN), (delight, 40, 47, NOUN...	[(dhoni, 11, 16, ENTITY), (delight, 40, 47, NO...
...	...	...	...	...
12	an excited dhoni promptly expressed his deligh...	[(excited, 3, 10, ENTITY), (dhoni, 11, 16, ENT...	[(dhoni, 11, 16, NOUN), (delight, 40, 47, NOUN...	[(dhoni, 11, 16, ENTITY), (delight, 40, 47, NO...
13	osaka, the fourth seed, overcame her unseeded ...	[(osaka, 0, 5, ENTITY), (seed, 18, 22, ENTITY)...	[(seed, 18, 22, NOUN), (opponent, 46, 54, NOUN...	[(seed, 18, 22, ENTITY), (opponent, 46, 54, NO...
14	captain marquinhos and striker mauro icardi al...	[(captain, 0, 7, ENTITY), (marquinhos, 8, 18, ...	[(striker, 23, 30, NOUN), (mauro, 31, 36, NOUN...	[(striker mauro, 23, 36, ENTITY), (mauro, 31, ...
15	after taking a legal opinion on the subject, b...	[(legal opinion, 15, 28, ENTITY), (subject, 36...	[(opinion, 21, 28, NOUN), (subject, 36, 43, NO...	[(opinion, 21, 28, NOUN), (subject, 36, 43, EN...
16	an excited dhoni promptly expressed his deligh...	[(excited, 3, 10, ENTITY), (dhoni, 11, 16, ENT...	[(dhoni, 11, 16, NOUN), (delight, 40, 47, NOUN...	[(dhoni, 11, 16, ENTITY), (delight, 40, 47, NO...

17 rows × 4 columns

In [51]:
```
column = 'named_nouns'
render_entities(1, df, options=options, column=column)
```

osaka, the fourth seed ENTITY , overcame her unseeded opponent NOUN 1-6, 6-3, 6-3 in 1hr 53min NOUN inside a near-empty arthur ENTITY ashe NOUN stadium ENTITY at flushing NOUN meadows NOUN .

In [52]:
```
add_named_ents(df)
display(df)
```

	text	named_ents	nouns	named_nouns
0	the uae isn't one of the top cricketing jurisd...	[(uae, 4, 7, ENTITY), (jurisdictions, 40, 53, ...	[(uae, 4, 7, NOUN), (jurisdictions, 40, 53, NO...	[(uae, 4, 7, ENTITY), (jurisdictions, 40, 53, ...
1	osaka, the fourth seed, overcame her unseeded ...	[(osaka, 0, 5, ENTITY), (seed, 18, 22, ENTITY)...	[(seed, 18, 22, NOUN), (opponent, 46, 54, NOUN...	[(seed, 18, 22, ENTITY), (opponent, 46, 54, NO...
2	captain marquinhos and striker mauro icardi al...	[(captain, 0, 7, ENTITY), (marquinhos, 8, 18, ...	[(striker, 23, 30, NOUN), (mauro, 31, 36, NOUN...	[(striker mauro, 23, 36, ENTITY), (mauro, 31, ...
3	after taking a legal opinion on the subject, b...	[(legal opinion, 15, 28, ENTITY), (subject, 36...	[(opinion, 21, 28, NOUN), (subject, 36, 43, NO...	[(opinion, 21, 28, NOUN), (subject, 36, 43, EN...
4	an excited dhoni promptly expressed his deligh...	[(excited, 3, 10, ENTITY), (dhoni, 11, 16, ENT...	[(dhoni, 11, 16, NOUN), (delight, 40, 47, NOUN...	[(dhoni, 11, 16, ENTITY), (delight, 40, 47, NO...
...	...	...	...	...
12	an excited dhoni promptly expressed his deligh...	[(excited, 3, 10, ENTITY), (dhoni, 11, 16, ENT...	[(dhoni, 11, 16, NOUN), (delight, 40, 47, NOUN...	[(dhoni, 11, 16, ENTITY), (delight, 40, 47, NO...
13	osaka, the fourth seed, overcame her unseeded ...	[(osaka, 0, 5, ENTITY), (seed, 18, 22, ENTITY)...	[(seed, 18, 22, NOUN), (opponent, 46, 54, NOUN...	[(seed, 18, 22, ENTITY), (opponent, 46, 54, NO...
14	captain marquinhos and striker mauro icardi al...	[(captain, 0, 7, ENTITY), (marquinhos, 8, 18, ...	[(striker, 23, 30, NOUN), (mauro, 31, 36, NOUN...	[(striker mauro, 23, 36, ENTITY), (mauro, 31, ...
15	after taking a legal opinion on the subject, b...	[(legal opinion, 15, 28, ENTITY), (subject, 36...	[(opinion, 21, 28, NOUN), (subject, 36, 43, NO...	[(opinion, 21, 28, NOUN), (subject, 36, 43, EN...
16	an excited dhoni promptly expressed his deligh...	[(excited, 3, 10, ENTITY), (dhoni, 11, 16, ENT...	[(dhoni, 11, 16, NOUN), (delight, 40, 47, NOUN...	[(dhoni, 11, 16, ENTITY), (delight, 40, 47, NO...

17 rows × 4 columns

In [53]:
```
df.to_csv('newsportal_nameentity.csv')
```

In [ ]: