ASSIGNMENT 5 - Named Entity Recognition

Matthew Dunne

You have been provided with a pickle file, containing the 100 news articles about Caterpillar. Identify what companies are mentioned most frequently in the news along with Caterpillar.

Load the data and discard non-English results

In [1]:

```
import pandas as pd
import nltk as nltk
import nltk.corpus
from nltk.text import Text
import pandas as pd
import re
import sys
```

In [2]:

```
directory=directory = 'C://Users//mjdun//Desktop//NLP//Assignments//'
data=pd.read_pickle(directory+'news_cat.pkl')
#it appears that all of the articles are listed as English
data[data['language']=='english']
data.describe(include='all')
```

Out[2]:

	crawled	language	text	title
count	100	100	100	100
unique	100	1	100	94
top	2018-01-31T13:18:56.020+02:00	english	model name / number: Areomax QR Code Link to T...	Caterpillar Gets a Leg or Two Up With a Winnin...
freq	1	100	1	4

In [3]:

```
data.head()
```

Out[3]:

	crawled	language	text	title
0	2018-01-30T23:03:51.004+02:00	english	by Abhishek K Global Telehandler Market 2023 D...	Global Telehandler Market 2023 Demand by Segme...
1	2018-01-30T23:06:46.024+02:00	english	favorite this post 2014 Caterpillar 314E LCR h...	2014 Caterpillar 314E LCR
2	2018-01-30T23:18:35.023+02:00	english	By: MAX NISEN The Amazon health care threat ha...	Amazon, Berkshire, JPMorgan health announcemen...
3	2018-01-30T23:20:54.012+02:00	english	QR Code Link to This Post MONTHLY PUBLIC AUCTI...	2005 Caterpillar CB534D Tandem Vibratory Rolle...
4	2018-01-30T23:28:30.000+02:00	english	QR Code Link to This Post 2007 CATERPILLAR D4G...	2007 CATERPILLAR D4G LGP CAB SCREEN/SWEEPS - O...

Identify what companies are most frequently mentioned along with Caterpillar (in both title and

Identify what companies are mentioned most frequently along with Caterpillar (in both title and the body of the article)

In [4]:

```
companies_mentioned=[]
for i in range(len(data)):
    entities = []
    labels = []
    #for each record take the title and the text and join them into a single string
    text=data['text'].iloc[i]
    title=data['title'].iloc[i]
    new='. '.join([title, text])
    for sent in nltk.sent_tokenize(new):
        for chunk in nltk.ne_chunk(nltk.pos_tag(nltk.word_tokenize(sent)), binary = False):
            if hasattr(chunk, 'label'):
                entities.append(' '.join(c[0] for c in chunk)) #Add space as between multi-token
entities
                labels.append(chunk.label())
    #because we are counting we do not want unique occurences
    entities_labels = list(zip(entities, labels))
    #take only those that are classified as organization
    companies = [company for company in entities_labels if company[1]=='ORGANIZATION']
    #append those to larger list element by element (not whole list)
    for j in companies:
        companies_mentioned.append(j)
```

In [5]:

```
company_df=pd.DataFrame(companies_mentioned)
company_df.columns = ["Entities", "Labels"]
company_df.describe(include='all')
```

Out[5]:

	Entities	Labels
count	1896	1896
unique	772	1
top	Caterpillar Inc.	ORGANIZATION
freq	96	1896

So we have identified 1896 tokens (772 unique) identified as an ORGANIZATION across the Text or Title of all articles.

Show a table or chart with your top-20 companies (sorted in the descending order)

In [6]:

```
#company_df['count']=company_df.groupby(['Entities']).agg(['count'])
company_count=company_df.groupby(['Entities']).agg(['count'])
comp_count=pd.DataFrame(company_count['Labels']['count'])
comp_count=pd.DataFrame({'Company':comp_count.index, 'Mentions':company_count['Labels']['count'] }
)
comp_count=comp_count.reset_index(drop=True)
top_20=comp_count.sort_values(by=['Mentions'], ascending=False)
top_20=top_20.reset_index(drop=True)
top_20.head(10)
```

Out[6]:

	Company	Mentions
0	Caterpillar Inc.	96
1	Caterpillar	87
2	NYSE	63
3	CAT	55

	Company	Mentions
4	Cat	36
5	Company	27
6	SEC	23
7	JPMorgan	22
8	Exchange Commission	20
9	Securities	20

Clearly there is a problem. The top results are mentions of Caterpillar itself, which we do not want. Some are also clearly not companies, even though they were identified as such by NLTK. We will filter these out.

In [21]:

```
#a list of "Company" names that are actually Caterpillar or are almost certainly not company names
cat=['Caterpillar Inc.', 'Caterpillar', 'CAT', 'Cat', 'Company', 'SEC', 'Exchange Commission', 'Sec
urities', 'Transportation', 'Construction Industries', 'Resource Industries', 'Financial Products'
, 'Energy', 'Investment', 'NOT', 'News', 'Ratings', 'VIOLATION', 'CFO Bradley', 'EPS', 'DIESEL', 'C
ountries', 'GENERATORS', 'FREE', 'NOS Events Center']
top_20=top_20[~top_20.Company.isin(cat)]
top_20.head(20)
```

Out[21]:

	Company	Mentions
2	NYSE	63
7	JPMorgan	22
17	LLC	14
18	Partners	13
19	Lincolnian Online	12
22	Vista Partners	11
23	Motley Fool	10
24	Dow	9
25	Bank	8
26	Capital Group	8
30	FMR	8
33	NASDAQ	8
34	WFG	7
35	Berkshire Hathaway	7
36	Vetr	7
37	AAPL	7
38	Credit Suisse Group	6
40	JPM	6
43	Dealer Mustang	6
44	Wonderland	6

The remaining are clearly company names, stock tickers, plausibly company names, or possibly parts of company names where NLTK did not get the whole name (e.g. LLC or Partners). These appear to be mostly financial related companies and probably just happend to be mentioned in the same article as Caterpillar.