HW 5 - NLP - Kenwan Cheung

You have been provided with a pickle file, containing the 100 news articles about Caterpillar.

Discard non-English results.

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

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

In [17]:
```
from IPython.core.display import display, HTML
display(HTML("<style>.container { width:100% !important; }</style>"))
```

In [42]:
```
import nltk as nltk
# nltk.internals.config_java("C:/Program Files (x86)/Java/jdk1.8.0_162/bin/java.exe")
import nltk.corpus
from nltk.text import Text
import pandas as pd
import re
import sys
import numpy as np
```

In [2]:
```
# nltk.download('punkt')
# nltk.download('averaged_perceptron_tagger')
# nltk.download('maxent_ne_chunker')
nltk.download('words')
```

```
[nltk_data] Downloading package words to
[nltk_data]     C:\Users\Boog\AppData\Roaming\nltk_data...
[nltk_data]   Package words is already up-to-date!
```

Out[2]: True

In [3]:
```
import os
from graphviz import Source
java_path = "C:/Program Files (x86)/Java/jdk1.8.0_162/bin/java.exe"
# java_path = "C:\ProgramData\Oracle\Java\javapath\java.exe"
os.environ['JAVAHOME'] = java_path

# Change the path according to your system
stanford_classifier = 'D:\datascience\StanfordParser\stanford-ner-2017-06-09\classifiers\english.all.3class.distsim.crf.ser.gz'
stanford_ner_path = 'D:\datascience\StanfordParser\stanford-ner-2017-06-09\stanford-ner.jar'

stanford_parser_path = 'D:\datascience\StanfordParser\stanford-parser-full-2017-06-09\stanford-parser.jar'
stanford_parser_models_path = 'D:\datascience\StanfordParser\stanford-parser-full-2017-06-09\stanford-parser-3.8.0-models.jar'

from nltk.parse.stanford import StanfordDependencyParser
from nltk.tag.stanford import StanfordNERTagger

# Creating Tagger Object
st = StanfordNERTagger(stanford_classifier, stanford_ner_path, encoding='utf-8')
sdp = StanfordDependencyParser(path_to_jar=stanford_parser_path, path_to_models_jar=stanford_parser_models_path)
```

```
c:\users\boog\dev\cfehome\lib\site-packages\nltk\tag\stanford.py:183: DeprecationWarning:
The StanfordTokenizer will be deprecated in version 3.2.5.
Please use nltk.tag.corenlp.CoreNLPPOSTagger or nltk.tag.corenlp.CoreNLPNERTagger instead.
  super(StanfordNERTagger, self).__init__(*args, **kwargs)
```

In [4]:
```
# import

cat = pd.read_pickle('../wk5/news_cat.pkl')
```

In [5]:
```
cat.head()
```

Out[5]:

	crawled	language	text	title

	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...

In [6]:
```
cat_english = cat[(cat.language == 'english')]
cat_english.head()
```

Out[6]:

	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...

In [7]:
```
cat_text = cat_english.text.astype(str)
```

```
cat_title = cat_english.title.astype(str)
```

In [8]:
```
cat_total = cat_text.append(cat_title)
```

In [16]:
```
print(cat_total)
```

```
0     by Abhishek K Global Telehandler Market 2023 D...
1     favorite this post 2014 Caterpillar 314E LCR h...
2     By: MAX NISEN The Amazon health care threat ha...
3     QR Code Link to This Post MONTHLY PUBLIC AUCTI...
4     QR Code Link to This Post 2007 CATERPILLAR D4G...
5     Elite Wealth Management Inc. Acquires Shares o...
6     Entertainment: Cast announced for 'Alice's Adv...
7     Daniel Leist/Getty Images General view of atmo...
8     The Dow Market Happenings For Tuesday, January...
9     DJIA Market Happenings For Monday, January 29,...
10    Fresh Engine Oil and Filter This machine is re...
11    Top Stock Picks for the Week of Jan 29, 2018 (...
12    QR Code Link to This Post Ag Parts Supply in H...
13    model name / number: D8 QR Code Link to This P...
14    Caterpillar Marine announces open order board ...
15    Republic NOLA New Orleans, LA - 11:00 PM Ages:...
16    0 SHARE \nDoes the name Land Rover ring any be...
17    Horizon’s Nine Stars for Sale By Kelley Sanfor...
18    Tweet \nSecurity National Trust Co. grew its h...
19    One year later, Peoria progresses after Caterp...
20    QR Code Link to This Post 2004 938G Series 2 C...
21    Jan 30, 2018 at 8:30 AM By John Irwin \nMonday...
22    DUBLIN- The "Diesel Power Engine Market by Ope...
23    Jan. 31 -- PEORIA -- Peoria natives and transp...
24    DIGGER MAN BLOG Earthworks for Dozers \nGPS an...
25    favorite this post Caterpillar c9 engine - $85...
26    Successful Cases[sornyang]Xuzhou Bonovo Machin...
27    State: Kentucky \nCompany Description \nProgre...
28    State: Alabama \nCompany Description \nProgres...
29    State: Alabama \nCompany Description \nProgres...
                          ...
70    Caterpillar Inc. (CAT) Position Cut by British...
71    Analysts’ Recent Ratings Updates for Caterpill...
```

```
72    Cat Engines For Adelphia Gateway Project Compr...
73           Caterpillar (CAT) Upgraded to Hold by Vetr
74    Boeing beats, says it'll deliver more planes i...
75    Caterpillar : Free Research Report as Caterpil...
76    IFG Advisory LLC Takes $623,000 Position in Ca...
77    FineMark National Bank & Trust Has $1.36 Milli...
78    Oakmont Partners LLC Invests $387,000 in Cater...
79                    World Book Fair, 6-14 January 2018
80       Keep Asking: Is this a need or is this a want?
81    Samsung essentially informs the US government ...
82    Regentatlantic Capital LLC Sells 913 Shares of...
83    Louisiana State Employees Retirement System Ha...
84    Convergence Investment Partners LLC Purchases ...
85    440 Investment Group LLC Invests $303,000 in C...
86    World: 3 Giants Form Health Alliance, Rocking ...
87                             "Déjà vu All Over Again?"
88    Stock Futures Slide: Market Shudders As Amazon...
89    Atossa Genetics Inc : "Déjà vu All Over Again?...
90      3 Giants Form Health Alliance, Rocking Insurers
91    Caterpillar Inc. (NYSE:CAT) Shares Sold by Par...
92         What to Expect From Caterpillar Inc. in 2018
93                             "Déjà vu All Over Again?"
94     Electric Forklift & Charger (West Chester) $5500
95         What to Expect From Caterpillar Inc. in 2018
96    Like to trade my Mitsubishi Excavator for a Mo...
97    One year after Caterpillar's headquarters anno...
98    1,613 Shares in Caterpillar Inc. (NYSE:CAT) Pu...
99    Caterpillar Inc. (CAT) Holdings Cut by Weather...
Length: 200, dtype: object
```

Now we've loaded just english results.

We need to identify which words are organizations only.

In [10]:
```
entities = []
labels = []
for chunk in nltk.ne_chunk(nltk.pos_tag(nltk.word_tokenize(str(cat_total))), 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())

#entities_labels = list(zip(entities, labels))
entities_labels = list(set(zip(entities, labels))) #unique entities
```

In [12]:
```
entities_df = pd.DataFrame(entities_labels)
entities_df.columns = ["Entities", "Labels"]
entities_df.head()
```

Out[12]:

	Entities	Labels
0	Parts Supply	PERSON
1	SHARE	ORGANIZATION
2	Abhishek K Global Telehandler	PERSON
3	Republic	ORGANIZATION
4	Picks	PERSON

Let's parse out which rows have "caterpillar in their name. We will repeat the exercise again.

In [19]:
```
cat_rows = cat_total[cat_total.str.contains("Caterpillar")]
```

In [69]:
```
entities = []
labels = []
for chunk in nltk.ne_chunk(nltk.pos_tag(nltk.word_tokenize(str(cat_rows))), binary = False):
    if hasattr(chunk, 'label'):
#         print(chunk)
        entities.append(' '.join(c[0] for c in chunk)) #Add space as between multi-token entities
        labels.append(chunk.label())

# entities_labels = list(zip(entities, labels))
entities_labels = list(set(zip(entities, labels))) #unique entities
```

In [70]:
```
entities_df = pd.DataFrame(entities_labels)
entities_df.columns = ["Entities", "Labels"]
# entities_df = entities_df.sort_values(by="Entities")
```

In [71]:
```
entities_df.head(20)
```

Out[71]:

	Entities	Labels
0	Louisiana State Employees Retirement System Ha	PERSON
1	H	GPE
2	Alabama	PERSON
3	Holdings Cut	ORGANIZATION
4	Investment Group	ORGANIZATION
5	DIGGER	ORGANIZATION
6	Marine	PERSON
7	LLC	ORGANIZATION
8	Peoria	GSP
9	British	GPE
10	Oakmont Partners	ORGANIZATION
11	Winnin	GPE
12	Picks	PERSON
13	Ca	GPE
14	Diesel Power Engine Market	PERSON
15	Capital	ORGANIZATION
16	Peoria	GPE
17	John Irwin	PERSON

	Entities	Labels
18	Weather	PERSON
19	Shares	PERSON

Let's get just the rows with organization in them.

In [72]:
```
org_token = entities_df[entities_df.Labels == "ORGANIZATION"].reset_index(drop=True)
```

In [73]:
```
print(org_token)
```

```
                          Entities        Labels
0                     Holdings Cut  ORGANIZATION
1                  Investment Group  ORGANIZATION
2                            DIGGER  ORGANIZATION
3                               LLC  ORGANIZATION
4                  Oakmont Partners  ORGANIZATION
5                           Capital  ORGANIZATION
6                              Week  ORGANIZATION
7                          Partners  ORGANIZATION
8                  NOLA New Orleans  ORGANIZATION
9                                LA  ORGANIZATION
10                              Par  ORGANIZATION
11                          Caterpil  ORGANIZATION
12                 Caterpillar Inc.  ORGANIZATION
13  Very Hungry Caterpillar Rocker   ORGANIZATION
14           FineMark National Bank  ORGANIZATION
15                              NYSE  ORGANIZATION
16       DJIA Market Happenings For  ORGANIZATION
17           Recent Ratings Updates  ORGANIZATION
18                              GSND  ORGANIZATION
19                            PUBLIC  ORGANIZATION
20                            Dozers  ORGANIZATION
21                              Vetr  ORGANIZATION
22                             SHARE  ORGANIZATION
23                           Republic  ORGANIZATION
```

```
24                   LLC Purchases  ORGANIZATION
25                             CAT  ORGANIZATION
26                     IFG Advisory  ORGANIZATION
27       Dow Market Happenings For  ORGANIZATION
```

Discussion

We have now pulled the most commonly mentioned companies that exist with Caterpillar in the news.