NLP

Find your favorite news source and grab the article text.

1. Show the most common words in the article.
2. Show the most common words under a part of speech. (i.e. NOUN: {'Bob':12, 'Alice':4,})
3. Find a subject/object relationship through the dependency parser in any sentence.
4. Show the most common Entities and their types.
5. Find Entites and their dependency (hint: entity.root.head)
6. Find the most similar words in the article

Note: Yes, the notebook from the video is not provided, I leave it to you to make your own :) it's your final assignment for the semester. Enjoy!

In [2]:
```
import spacy
```

```
2023-05-01 20:31:51.995738: I tensorflow/core/platform/cpu_feature_gu
ard.cc:182] This TensorFlow binary is optimized to use available CPU
instructions in performance-critical operations.
To enable the following instructions: AVX2 FMA, in other operations,
rebuild TensorFlow with the appropriate compiler flags.
```

I selected "*Ukraine's Spring Offensive Comes With Immense Stakes for Future of the War*" by Michael Schwirtz, Julian E. Barnes, Eric Schmitt, and Thomas Gibbons-Neff.

This article was published in the New York Times on 24 APR 23. [https://www.nytimes.com/2023/04/24/us/politics/ukraine-russia-war-spring-offensive.html (https://www.nytimes.com/2023/04/24/us/politics/ukraine-russia-war-spring-offensive.html)](https://www.nytimes.com/2023/04/24/us/politics/ukraine-russia-war-spring-offensive.html)

In [4]:
```
processor = spacy.load('en_core_web_sm')
```

In [5]:
```
text = """ WASHINGTON — Ukraine is preparing to launch a counteroffens
American and NATO allies have supplied Ukraine with extensive artiller
At the same time, 12 Ukrainian combat brigades of about 4,000 soldiers
Although Ukraine shares few details of its operational plan with Ameri
“Everything hinges on this counteroffensive,” said Alexander Vershbow,
While Ukrainian officials have said their goal is to break through dug
Ukraine’s military faces many challenges — one reason that a stalemate
And yet American military officials say it is possible that Ukraine’s
“I’m optimistic that between this year and next year, I think Ukraine
Although Ukraine has deviated from the usual secrecy surrounding milit
Still, big gains are not guaranteed, or even necessarily likely. The b
Ukraine built the new combat brigades by pairing raw recruits with a s
Training on those tactics has gone well, according to multiple U.S. of
Soldiers fighting in Ukraine have said that, so far, sophisticated man
If the Ukrainians succeed in using these new tactics, even to a small
“If they can break through, then I think they can change the dynamic o
Major questions about Ukraine’s artillery and other ammunition supplie
The Ukrainian military has been firing thousands of artillery shells a
While Ukrainian forces can use drones to strike behind Russian front l
The Russians have challenges of their own.
Since the beginning of the invasion, there have been major doubts abou
But Russia is working to address those gaps. Russian troops have honed
In private meetings, Sergei K. Shoigu, the Russian defense minister, h
American intelligence officials have repeatedly warned that President
U.S. and European officials say Russia is preparing new rounds of mobi
American officials say that Mr. Putin faces a political cost for any m
Wagner’s prison recruits, quickly became cannon fodder.
Still, Russia’s capacity — and willingness — to absorb losses remains
```

```
A key focus of the United States and the West has been trying to stop

China appears to have been deterred, at least for the moment, from pro

Another apparent success has been Egypt. While U.S. officials were qui

After a diplomatic push by the United States and Britain, the Egyptian

U.S. officials said a production contract has been agreed with Egyptia

Some European countries, including France, are pushing for negotiation

For the Ukrainians to force a real negotiation, they must make sure “V

The Ukrainians have said they would not agree to any peace talks until

The chances that Mr. Putin will back down or cut his losses in respons

Celeste A. Wallander, the U.S. assistant secretary of defense for inte

"""
```

In [6]:
```
processed_text = processor(text)
```

Out[6]:
```
 WASHINGTON — Ukraine is preparing to launch a counteroffensive again
st Russian forces as early as next month, American officials say, in
the face of immense risks: Without a decisive victory, Western suppor
t for Ukraine could weaken, and Kyiv could come under increasing pres
sure to enter serious negotiations to end or freeze the conflict.

American and NATO allies have supplied Ukraine with extensive artille
ry and ammunition for the upcoming battle, and officials now say they
are hopeful the supplies will last — a change from two months ago whe
n weapons were only trickling in and U.S. officials were worried that
the supplies might run out.

At the same time, 12 Ukrainian combat brigades of about 4,000 soldier
s each are expected to be ready at the end of April, according to lea
ked Pentagon documents that offer a hint of Kyiv’s timetable. The Uni
ted States and NATO allies are training and supplying nine of those b
rigades, the documents said.

Although Ukraine shares few details of its operational plan with Amer
```

1. Most Common Words

In [7]:
```
import re
from collections import Counter
```

In [8]:
```
words = [token.text for token in processed_text if token.is_stop != Tr
         and token.is_punct !=True]
lower_words = [word.lower() for word in words]
```

In [9]:
```
Counter(lower_words).most_common(10)
```

Out[9]:
```
[('\n\n', 37),
 ('ukraine', 26),
 ('officials', 21),
 ('russia', 19),
 ('russian', 14),
 ('american', 13),
 ('said', 13),
 ('u.s.', 11),
 ('forces', 10),
 ('artillery', 10)]
```

2. Most Common Under Part of Speech

2.1 Nouns

In [10]:
```
nouns = [token.text for token in processed_text if token.is_stop != Tr
         and token.is_punct !=True and token.pos_ =='NOUN']
print(Counter(nouns).most_common(15))
```

```
[('officials', 21), ('artillery', 10), ('forces', 9), ('ammunition',
7), ('soldiers', 7), ('supplies', 6), ('battlefield', 6), ('year',
5), ('intelligence', 5), ('equipment', 5), ('allies', 4), ('time',
4), ('official', 4), ('army', 4), ('units', 4)]
```

2.2 Verbs

In [11]:
```
verbs = [token.text for token in processed_text if token.is_stop != Tr
         and token.is_punct !=True and token.pos_ =='VERB']
print(Counter(verbs).most_common(15))
```

```
[('said', 13), ('use', 5), ('according', 3), ('going', 3), ('remains
', 3), ('think', 3), ('coming', 3), ('send', 3), ('preparing', 2), ('
run', 2), ('leaked', 2), ('supplying', 2), ('including', 2), ('creati
ng', 2), ('break', 2)]
```

2.3 Adjectives

In [12]:
```
adj = [token.text for token in processed_text if token.is_stop != True
         and token.is_punct !=True and token.pos_ =='ADJ']
print(Counter(adj).most_common(15))
```

```
[('Russian', 13), ('American', 12), ('Ukrainian', 10), ('new', 7), ('
European', 6), ('likely', 3), ('senior', 3), ('significant', 3), ('mi
litary', 3), ('little', 3), ('domestic', 3), ('Western', 2), ('hopefu
l', 2), ('ready', 2), ('optimistic', 2)]
```

2.4 Adverbs

In [13]:
```
adv = [token.text for token in processed_text if token.is_stop != True
       and token.is_punct !=True and token.pos_ =='ADV']
print(Counter(adv).most_common(15))
```

```
[('effectively', 2), ('far', 2), ('early', 1), ('ago', 1), ('maybe',
1), ('openly', 1), ('necessarily', 1), ('heavily', 1), ('especially',
1), ('extremely', 1), ('numerically', 1), ('dangerously', 1), ('faste
r', 1), ('recently', 1), ('indefinitely', 1)]
```

3. Subject/ Object Relationship

In [14]:
```
def pr_tree(word, level):
    if word.is_punct:
        return
    for child in word.lefts:
        pr_tree(child, level+1)
    print('\t'* level + word.text + ' - ' + word.dep_)
    for child in word.rights:
        pr_tree(child, level+1)
```

In [15]:
```
for sentence in processed_text.sents:
    pr_tree(sentence.root, 0)
    print('')
```

```
                - dep
        WASHINGTON - dep
        Ukraine - nsubj
        is - aux
preparing - ROOT
                to - aux
        launch - xcomp
                        a - det
                counteroffensive - dobj
                        against - prep
                                        Russian - amod
                                forces - pobj
                        as - advmod
                early - advmod
                        as - prep
                                        next - amod
                                month - pobj
                        American - amod
                officials - nsubj
        say - parataxis
```

4. Most Common Entities

In [16]:
```
counts = Counter()
for entity in processed_text.ents:
    counts[entity.text.upper(), entity.label_] += 1
counts.most_common(15)
```

Out[16]:
```
[(('UKRAINE', 'GPE'), 26),
 (('RUSSIA', 'GPE'), 19),
 (('RUSSIAN', 'NORP'), 14),
 (('AMERICAN', 'NORP'), 13),
 (('U.S.', 'GPE'), 11),
 (('UKRAINIAN', 'NORP'), 9),
 (('PUTIN', 'PERSON'), 7),
 (('NATO', 'ORG'), 6),
 (('EUROPEAN', 'NORP'), 6),
 (('RUSSIANS', 'NORP'), 6),
 (('THE UNITED STATES', 'GPE'), 5),
 (('UKRAINIANS', 'NORP'), 5),
 (('WEST', 'LOC'), 5),
 (('KYIV', 'PERSON'), 3),
 (('PENTAGON', 'ORG'), 3)]
```

5. Entities and Dependencies

In [17]:
```
for entity in processed_text.ents:
    print(entity, entity.root.head)
```

```
WASHINGTON preparing
Ukraine preparing
Russian forces
next month as
American officials
Western support
Ukraine for
Kyiv come
American supplied
NATO allies
Ukraine supplied
two months ago from
U.S. officials
12 brigades
Ukrainian brigades
about 4,000 soldiers
the end of April at
Pentagon documents
Kyiv timetable
The United States training
```

6. Most Similar Words

In [18]:
```
n=0
for sentence in processed_text.sents:
    for noun_chunk in processed_text.noun_chunks:
        print(n, noun_chunk, processed_text.similarity(noun_chunk))
        n+=1
```

```
/var/folders/tq/y257w3hd2p59ywppnr10vgq40000gq/T/ipykernel_66626/942087018.py:4: UserWarning: [W007] The model you're using has no word vectors loaded, so the result of the Doc.similarity method will be based on the tagger, parser and NER, which may not give useful similarity judgements. This may happen if you're using one of the small models, e.g. `en_core_web_sm`, which don't ship with word vectors and only use context-sensitive tensors. You can always add your own word vectors, or use one of the larger models instead if available.
  print(n, noun_chunk, processed_text.similarity(noun_chunk))

0 Ukraine 0.43645340389226117
1 a counteroffensive 0.3972006628498517
2 Russian forces 0.475198998277585
3 next month 0.31365995793841134
4 American officials 0.40758906430368297
5 the face 0.3905851448945312
6 immense risks 0.3969868696808351
7 a decisive victory 0.4340192596790677
8 Western support 0.44655795511676094
9 Ukraine 0.45177978208133496
```

In [ ]: