

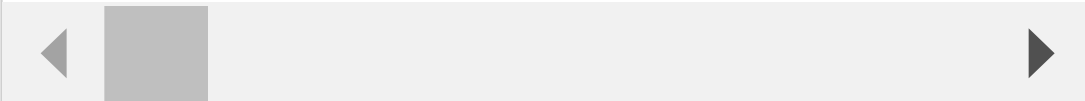
NLP

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 [1]: ▶ # Saving article text
         article = """American superstar Katie Ledecky equalled the record for
```



```
In [2]: ▶ # Printing article  
print(article)
```

American superstar Katie Ledecky equalled the record for the most gold medals by a female Olympian as she won the 800m freestyle title at the Paris Games. Ledecky clocked eight minutes 11.04 seconds to become the only woman - and only swimmer other than the great Michael Phelps - with four Olympic golds in the same event. It was Ledecky's ninth Olympic gold, moving her level with former Soviet gymnast Larisa Latynina, and taking her overall tally to 14 medals. Phelps has the most medals of any Olympian with 28, including 23 golds. "The four-times record is the one that means the most to me," Ledecky, 27, said afterwards. "3 August is the day I won in 2012, and I didn't want 3 August to be a day I didn't like moving forwards. "I put a lot of pressure on myself, so I'm happy I got the job done." Earlier on Saturday, Summer McIntosh's astonishing debut Games continued, with the Canadian 17-year-old securing her third gold with victory in the women's 200m individual medley. But Great Britain's 4x100m medley relay defence ended in disappointment, with the quartet finishing seventh. Ledecky's dominance over distance continues. Ledecky has won four medals in Paris alone - two golds, a silver and a bronze. She became the United States' most decorated female Olympian with silver in the women's 4x200m relay on Thursday. Such is her dominance in the 800m freestyle that she has lost just once over the distance in 13 years - and that was to rising star McIntosh at a regional meet earlier in 2024. McIntosh opted not to swim the 800m in Paris, meaning Ledecky's biggest rival was old foe Ariarne Titmus. Australia's Titmus beat Ledecky to 400m freestyle gold earlier in the week but she could not stay with the American in the closing stages of her favourite distance. The two shared a warm moment at the end of the race, with Ledecky raising both their arms in the air before Titmus applauded her opponent as she left the arena. "We have just seen a little bit of history there," Steve Parry, Olympic bronze medallist for Britain in 2004, said on BBC 5 Live. "Ledecky is the absolute queen of the pool. To be able to see someone dominate a distance event for 13 years is absolutely brilliant." Titmus took silver in 8:12.29, with Ledecky's American team-mate Paige Madden (8:13.00) completing the podium.

```

In [3]: ► # Installing packages
!pip install spacy
!python -m spacy download en_core_web_sm
!python -m spacy download en_core_web_md

# Importing libraries
import spacy
from collections import Counter

# Load the English model
nlp = spacy.load("en_core_web_sm")

c:/download/en_core_web_md-3.7.1/en_core_web_md-3.7.1-py3-none-
-any.whl (https://github.com/explosion/spacy-models/releases/d
ownload/en_core_web_md-3.7.1/en_core_web_md-3.7.1-py3-none-an
y.whl) (42.8 MB)
----- 42.8/42.8 MB 32.7
MB/s eta 0:00:00
Requirement already satisfied: spacy<3.8.0,>=3.7.2 in c:\users
\kaoui\anaconda3\lib\site-packages (from en-core-web-md==3.7.
1) (3.7.5)
Requirement already satisfied: packaging>=20.0 in c:\users\kao
ui\anaconda3\lib\site-packages (from spacy<3.8.0,>=3.7.2->en-c
ore-web-md==3.7.1) (22.0)
Requirement already satisfied: tqdm<5.0.0,>=4.38.0 in c:\users
\kaoui\anaconda3\lib\site-packages (from spacy<3.8.0,>=3.7.2->
en-core-web-md==3.7.1) (4.64.1)
Requirement already satisfied: setuptools in c:\users\kaoui\an
aconda3\lib\site-packages (from spacy<3.8.0,>=3.7.2->en-core-w
eb-md==3.7.1) (65.6.3)
Requirement already satisfied: wasabi<1.2.0,>=0.9.1 in c:\user
s\kaoui\anaconda3\lib\site-packages (from spacy<3.8.0,>=3.7.2-

```

```
In [6]: # Processing the text
doc = nlp(article)

# Load the medium English model for word vectors
nlp = spacy.load("en_core_web_md")

# Verify model name and word vectors
print(nlp.meta['name'])
print("Vectors loaded:", nlp.vocab.vectors_length > 0)

# Most common words
words = [token.text for token in doc if token.is_alpha]
word_freq = Counter(words)
most_common_words = word_freq.most_common(10)
print("Most common words:", most_common_words)

# Most common nouns
nouns = [token.text for token in doc if token.pos_ == "NOUN"]
noun_freq = Counter(nouns)
most_common_nouns = noun_freq.most_common(10)
print("Most common nouns:", most_common_nouns)

# Subject/Object relationships
print("\nSubject/Object relationships:")
for token in doc:
    if token.dep_ in ("nsubj", "dobj"):
        print(f"Token: {token.text}, Dep: {token.dep_}, Head: {token.h

# Most common entities and their types
entities = [(ent.text, ent.label_) for ent in doc.ents]
entity_freq = Counter(entities)
most_common_entities = entity_freq.most_common(10)
print("\nMost common entities:", most_common_entities)

# Entities and their dependency
print("\nEntities and their dependency:")
for ent in doc.ents:
    print(f"Entity: {ent.text}, Root Head: {ent.root.head.text}")
```

Token: mauden, Dep: nsubj, Head: completing

Token: podium, Dep: dobj, Head: completing

Most common entities: [(('Ledecky', 'PERSON'), 8), (('American', 'NORP'), 3), (('Olympian', 'NORP'), 3), (('800', 'CARDINAL'), 3), (('four', 'CARDINAL'), 3), (('Titmus', 'GPE'), 3), (('Paris', 'GPE'), 2), (('two', 'CARDINAL'), 2), (('13 years', 'DATE'), 2), (('McIntosh', 'PERSON'), 2)]

Entities and their dependency:

Entity: American, Root Head: superstar

Entity: Katie Ledecky, Root Head: equalled

Entity: Olympian, Root Head: by

Entity: 800, Root Head: m

Entity: the Paris Games, Root Head: at

Entity: Ledecky, Root Head: clocked

Entity: eight minutes 11.04 seconds, Root Head: clocked

Entity: Michael Phelps, Root Head: than

Entity: four, Root Head: golds

Entity: Ledecky, Root Head: gold

In [17]:  !python -m spacy download en_core_web_lg

Collecting en-core-web-lg==3.7.1

Downloading https://github.com/explosion/spacy-models/releases/download/en_core_web_lg-3.7.1/en_core_web_lg-3.7.1-py3-none-any.whl (https://github.com/explosion/spacy-models/releases/download/en_core_web_lg-3.7.1/en_core_web_lg-3.7.1-py3-none-any.whl) (587.7 MB)

----- 587.7/587.7 MB 2.6 MB/s e

ta 0:00:00

Requirement already satisfied: spacy<3.8.0,>=3.7.2 in c:\users\kaoui\anaconda3\lib\site-packages (from en-core-web-lg==3.7.1) (3.7.5)

Requirement already satisfied: packaging>=20.0 in c:\users\kaoui\anaconda3\lib\site-packages (from spacy<3.8.0,>=3.7.2->en-core-web-lg==3.7.1) (22.0)

Requirement already satisfied: pydantic!=1.8,!1.8.1,<3.0.0,>=1.7.4 in c:\users\kaoui\anaconda3\lib\site-packages (from spacy<3.8.0,>=3.7.2->en-core-web-lg==3.7.1) (1.10.12)

Requirement already satisfied: spacy-loggers<2.0.0,>=1.0.0 in c:\users\kaoui\anaconda3\lib\site-packages (from spacy<3.8.0,>=3.7.2->en-core-web-lg==3.7.1) (1.0.5)

Requirement already satisfied: murmurhash<1.1.0,>=0.28.0 in c:\users\kaoui\anaconda3\lib\site-packages (from spacy<3.8.0,>=3.7.2->en-core-web-lg==3.7.1) (1.0.10)

Requirement already satisfied: cymem<2.1.0,>=2.0.2 in c:\users\kaoui\anaconda3\lib\site-packages (from spacy<3.8.0,>=3.7.2->en-core-web-lg==3.7.1) (2.0.8)

Requirement already satisfied: wasabi<1.2.0,>=0.9.1 in c:\users\kaoui\anaconda3\lib\site-packages (from spacy<3.8.0,>=3.7.2->en-core-web-lg==3.7.1) (1.1.3)

Requirement already satisfied: setuptools in c:\users\kaoui\anaconda3\lib\site-packages (from spacy<3.8.0,>=3.7.2->en-core-web-lg==3.7.1) (65.6.3)

Requirement already satisfied: tqdm<5.0.0,>=4.38.0 in c:\users\kaoui\anaconda3\lib\site-packages (from spacy<3.8.0,>=3.7.2->en-core-web-lg==3.7.1) (4.64.1)

Requirement already satisfied: typer<1.0.0,>=0.3.0 in c:\users\kaoui\anaconda3\lib\site-packages (from spacy<3.8.0,>=3.7.2->en-core-web-lg==3.7.1) (0.12.3)

Requirement already satisfied: Jinja2 in c:\users\kaoui\anaconda3\lib\site-packages (from spacy<3.8.0,>=3.7.2->en-core-web-lg==3.7.1) (3.1.2)

Requirement already satisfied: preshed<3.1.0,>=3.0.2 in c:\users\kaoui\anaconda3\lib\site-packages (from spacy<3.8.0,>=3.7.2->en-core-web-lg==3.7.1) (3.0.9)

Requirement already satisfied: langcodes<4.0.0,>=3.2.0 in c:\users\kaoui\anaconda3\lib\site-packages (from spacy<3.8.0,>=3.7.2->en-core-web-lg==3.7.1) (3.4.0)

Requirement already satisfied: srsly<3.0.0,>=2.4.3 in c:\users\kaoui\anaconda3\lib\site-packages (from spacy<3.8.0,>=3.7.2->en-core-web-lg==3.7.1) (2.4.8)

Requirement already satisfied: requests<3.0.0,>=2.13.0 in c:\users\kaoui\anaconda3\lib\site-packages (from spacy<3.8.0,>=3.7.2->en-core-web-lg==3.7.1) (2.28.1)

Requirement already satisfied: numpy>=1.19.0 in c:\users\kaoui\anaconda3\lib\site-packages (from spacy<3.8.0,>=3.7.2->en-core-web-lg==3.7.1) (1.23.5)

Requirement already satisfied: catalogue<2.1.0,>=2.0.6 in c:\users\kaoui\anaconda3\lib\site-packages (from spacy<3.8.0,>=3.7.2->en-core-web-lg==3.7.1) (2.0.10)

Requirement already satisfied: thinc<8.3.0,>=8.2.2 in c:\users\kaoui\anaconda3\lib\site-packages (from spacy<3.8.0,>=3.7.2->en-core-web-lg==3.7.1) (8.2.5)

Requirement already satisfied: spacy-legacy<3.1.0,>=3.0.11 in c:\users\kaoui\anaconda3\lib\site-packages (from spacy<3.8.0,>=3.7.2->en-core-web-lg==3.7.1) (3.0.12)

Requirement already satisfied: weasel<0.5.0,>=0.1.0 in c:\users\kaoui\anaconda3\lib\site-packages (from spacy<3.8.0,>=3.7.2->en-core-web-lg==3.7.1) (0.4.1)

Requirement already satisfied: language-data>=1.2 in c:\users\kaoui\anaconda3\lib\site-packages (from langcodes<4.0.0,>=3.2.0->spacy<3.8.0,>=3.7.2->en-core-web-lg==3.7.1) (1.2.0)

Requirement already satisfied: typing-extensions>=4.2.0 in c:\users\kaoui\anaconda3\lib\site-packages (from pydantic!=1.8,!1.8.1,<3.0.0,>=1.7.4->spacy<3.8.0,>=3.7.2->en-core-web-lg==3.7.1) (4.12.2)

Requirement already satisfied: certifi>=2017.4.17 in c:\users\kaoui\anaconda3\lib\site-packages (from requests<3.0.0,>=2.13.0->spacy<3.8.0,>=3.7.2->en-core-web-lg==3.7.1) (2024.6.2)

Requirement already satisfied: urllib3<1.27,>=1.21.1 in c:\users\kaoui\anaconda3\lib\site-packages (from requests<3.0.0,>=2.13.0->spacy<3.8.0,>=3.7.2->en-core-web-lg==3.7.1) (1.26.14)

Requirement already satisfied: idna<4,>=2.5 in c:\users\kaoui\anaconda3\lib\site-packages (from requests<3.0.0,>=2.13.0->spacy<3.8.0,>=3.7.2->en-core-web-lg==3.7.1) (3.4)

Requirement already satisfied: charset-normalizer<3,>=2 in c:\users\kaoui\anaconda3\lib\site-packages (from requests<3.0.0,>=2.13.0->spacy<3.8.0,>=3.7.2->en-core-web-lg==3.7.1) (2.0.4)

Requirement already satisfied: confection<1.0.0,>=0.0.1 in c:\users\kaoui\anaconda3\lib\site-packages (from thinc<8.3.0,>=8.2.2->spacy<3.8.0,>=3.7.2->en-core-web-lg==3.7.1) (0.1.5)

Requirement already satisfied: blis<0.8.0,>=0.7.8 in c:\users\kaoui\anaconda3\lib\site-packages (from thinc<8.3.0,>=8.2.2->spacy<3.8.0,>=3.7.2->en-core-web-lg==3.7.1) (0.7.11)

Requirement already satisfied: colorama in c:\users\kaoui\anaconda3\lib\site-packages (from tqdm<5.0.0,>=4.38.0->spacy<3.8.0,>=3.7.2->en-core-web-lg==3.7.1) (0.4.6)

Requirement already satisfied: shellingham>=1.3.0 in c:\users\kaoui\anaconda3\lib\site-packages (from typer<1.0.0,>=0.3.0->spacy<3.8.0,>=3.7.2->en-core-web-lg==3.7.1) (1.5.4)

Requirement already satisfied: rich>=10.11.0 in c:\users\kaoui\anaconda3\lib\site-packages (from typer<1.0.0,>=0.3.0->spacy<3.8.0,>=3.7.2->en-core-web-lg==3.7.1) (13.7.1)

Requirement already satisfied: click>=8.0.0 in c:\users\kaoui\anaconda3\lib\site-packages (from typer<1.0.0,>=0.3.0->spacy<3.8.0,>=3.7.2->en-core-web-lg==3.7.1) (8.0.4)


Requirement already satisfied: cloudpathlib<1.0.0,>=0.7.0 in c:\users\kaoui\anaconda3\lib\site-packages (from weasel<0.5.0,>=0.1.0->spacy<3.8.0,>=3.7.2->en-core-web-lg==3.7.1) (0.18.1)

Requirement already satisfied: smart-open<8.0.0,>=5.2.1 in c:\users\kaoui\anaconda3\lib\site-packages (from weasel<0.5.0,>=0.1.0->spacy<3.8.0,>=3.7.2->en-core-web-lg==3.7.1) (5.2.1)

Requirement already satisfied: MarkupSafe>=2.0 in c:\users\kaoui\anaconda3\lib\site-packages (from jinja2->spacy<3.8.0,>=3.7.2->en-core-web-lg==3.7.1) (2.1.1)

Requirement already satisfied: marisa-trie>=0.7.7 in c:\users\kaoui\anaconda3\lib\site-packages (from language-data>=1.2->langcodes<4.0.0,>=3.2.0->spacy<3.8.0,>=3.7.2->en-core-web-lg==3.7.1) (1.2.0)

Requirement already satisfied: pygments<3.0.0,>=2.13.0 in c:\users\kaoui\anaconda3\lib\site-packages (from rich>=10.11.0->typer<1.0.0,>=0.3.0->spacy<3.8.0,>=3.7.2->en-core-web-lg==3.7.1) (2.18.0)
Requirement already satisfied: markdown-it-py>=2.2.0 in c:\users\kaoui\anaconda3\lib\site-packages (from rich>=10.11.0->typer<1.0.0,>=0.3.0->spacy<3.8.0,>=3.7.2->en-core-web-lg==3.7.1) (3.0.0)
Requirement already satisfied: mdurl~=0.1 in c:\users\kaoui\anaconda3\lib\site-packages (from markdown-it-py>=2.2.0->rich>=10.11.0->typer<1.0.0,>=0.3.0->spacy<3.8.0,>=3.7.2->en-core-web-lg==3.7.1) (0.1.2)
Installing collected packages: en-core-web-lg
Successfully installed en-core-web-lg-3.7.1
[+] Download and installation successful
You can now load the package via spacy.load('en_core_web_lg')

```
In [20]:  # Loading Larger model
nlp = spacy.load('en_core_web_lg')

doc = nlp(article)
```

```
In [21]: # Extracting unique words
unique_words = list(set([token.text for token in doc if token.is_alpha

# Creating function to find most similar words
def get_most_similar(word, topn=5):
    # Create a single token Doc object for the input word
    token_word = nlp(word)[0]
    if not token_word.has_vector:
        return [] # Skip words without vectors

    # Calculate similarity for each token in the document
    similarities = [(token.text, token.similarity(token_word))
                    for token in doc if token.has_vector]
    similarities = sorted(similarities, key=lambda item: -item[1])
    return similarities[:topn]

# Finding similar words for each unique word
similar_words = {}
for word in unique_words:
    try:
        similar_words[word] = get_most_similar(word)
    except KeyError:
        # Handle case where word may not be in the vocabulary
        similar_words[word] = []

# Printing the most similar words
print("\nMost similar words:")
for word, similarities in similar_words.items():
    print(f"{word}: {similarities}")
```

```
059)]
someone: [('someone', 1.0), ('just', 0.641502320766449), ('jus
t', 0.641502320766449), ('myself', 0.6153632998466492), ('me',
0.6127251982688904)]
moving: [('moving', 1.0), ('moving', 1.0), ('closing', 0.61347
47266769409), ('taking', 0.6060069799423218), ('rising', 0.580
3539156913757)]
day: [('day', 1.0), ('day', 1.0), ('week', 0.756441593170166),
('year', 0.6056016683578491), ('Saturday', 0.522563159465789
8)]
gymnast: [('gymnast', 1.0), ('swimmer', 0.7152723073959351),
('Olympian', 0.5664869546890259), ('Olympian', 0.5664869546890
259), ('Olympian', 0.5664869546890259)]
both: [('both', 1.0), ('and', 0.6645598411560059), ('and', 0.6
645598411560059), ('and', 0.6645598411560059), ('and', 0.66455
98411560059)]
dominance: [('dominance', 1.0), ('dominance', 1.0), ('dominat
e', 0.753794252872467), ('the', 0.601233720779419), ('the', 0.
601233720779419)]
year: [('year', 1.0), ('week', 0.7517561912536621), ('years',
```

