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]:
```
!pip install -U pip setuptools wheel
```

```
Requirement already satisfied: pip in /opt/anaconda3/lib/python3.9/si
te-packages (22.2.2)
Collecting pip
  Downloading pip-23.1.1-py3-none-any.whl (2.1 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 2.1/2.1 MB 7.4 MB/s eta
0:00:0000:0100:01
Requirement already satisfied: setuptools in /opt/anaconda3/lib/pytho
n3.9/site-packages (63.4.1)
Collecting setuptools
  Downloading setuptools-67.7.2-py3-none-any.whl (1.1 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 1.1/1.1 MB 9.4 MB/s eta
0:00:00:00:0100:01
Requirement already satisfied: wheel in /opt/anaconda3/lib/python3.9/
site-packages (0.37.1)
Collecting wheel
  Downloading wheel-0.40.0-py3-none-any.whl (64 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 64.5/64.5 kB 2.6 MB/s e
ta 0:00:00
Installing collected packages: wheel, setuptools, pip
  Attempting uninstall: wheel
    Found existing installation: wheel 0.37.1
    Uninstalling wheel-0.37.1:
      Successfully uninstalled wheel-0.37.1
  Attempting uninstall: setuptools
    Found existing installation: setuptools 63.4.1
    Uninstalling setuptools-63.4.1:
      Successfully uninstalled setuptools-63.4.1
  Attempting uninstall: pip
    Found existing installation: pip 22.2.2
    Uninstalling pip-22.2.2:
      Successfully uninstalled pip-22.2.2
ERROR: pip's dependency resolver does not currently take into account
all the packages that are installed. This behaviour is the source of
the following dependency conflicts.
spyder 5.3.3 requires pyqt5<5.16, which is not installed.
spyder 5.3.3 requires pyqtwebengine<5.16, which is not installed.
anaconda-project 0.11.1 requires ruamel-yaml, which is not installed.
conda-repo-cli 1.0.20 requires clyent==1.2.1, but you have clyent 1.
2.2 which is incompatible.
conda-repo-cli 1.0.20 requires nbformat==5.4.0, but you have nbformat
5.5.0 which is incompatible.
Successfully installed pip-23.1.1 setuptools-67.7.2 wheel-0.40.0
```

In [2]:
```
!pip install -U spacy
```

```
Collecting spacy
  Downloading spacy-3.5.2-cp39-cp39-macosx_10_9_x86_64.whl (6.9 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 6.9/6.9 MB 7.0 MB/s eta
0:00:0000:0100:01
Collecting spacy-legacy<3.1.0,>=3.0.11 (from spacy)
  Downloading spacy_legacy-3.0.12-py2.py3-none-any.whl (29 kB)
Collecting spacy-loggers<2.0.0,>=1.0.0 (from spacy)
  Downloading spacy_loggers-1.0.4-py3-none-any.whl (11 kB)
Collecting murmurhash<1.1.0,>=0.28.0 (from spacy)
  Downloading murmurhash-1.0.9-cp39-cp39-macosx_10_9_x86_64.whl (18 k
B)
Collecting cymem<2.1.0,>=2.0.2 (from spacy)
  Downloading cymem-2.0.7-cp39-cp39-macosx_10_9_x86_64.whl (32 kB)
Collecting preshed<3.1.0,>=3.0.2 (from spacy)
  Downloading preshed-3.0.8-cp39-cp39-macosx_10_9_x86_64.whl (107 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 107.9/107.9 kB 5.7 MB/s
eta 0:00:00
Collecting thinc<8.2.0,>=8.1.8 (from spacy)
  Downloading thinc-8.1.9-cp39-cp39-macosx_10_9_x86_64.whl (865 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 865.2/865.2 kB 7.0 MB/s
eta 0:00:0000:0100:01
Collecting wasabi<1.2.0,>=0.9.1 (from spacy)
  Downloading wasabi-1.1.1-py3-none-any.whl (27 kB)
Collecting srsly<3.0.0,>=2.4.3 (from spacy)
  Downloading srsly-2.4.6-cp39-cp39-macosx_10_9_x86_64.whl (492 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 492.2/492.2 kB 4.2 MB/s
eta 0:00:0000:0100:01
Collecting catalogue<2.1.0,>=2.0.6 (from spacy)
  Downloading catalogue-2.0.8-py3-none-any.whl (17 kB)
Collecting typer<0.8.0,>=0.3.0 (from spacy)
  Downloading typer-0.7.0-py3-none-any.whl (38 kB)
Collecting pathy>=0.10.0 (from spacy)
  Downloading pathy-0.10.1-py3-none-any.whl (48 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 48.9/48.9 kB 2.3 MB/s e
ta 0:00:00
Requirement already satisfied: smart-open<7.0.0,>=5.2.1 in /opt/anaco
nda3/lib/python3.9/site-packages (from spacy) (5.2.1)
Requirement already satisfied: tqdm<5.0.0,>=4.38.0 in /opt/anaconda3/
lib/python3.9/site-packages (from spacy) (4.64.1)
Requirement already satisfied: numpy>=1.15.0 in /opt/anaconda3/lib/py
thon3.9/site-packages (from spacy) (1.19.5)
Requirement already satisfied: requests<3.0.0,>=2.13.0 in /opt/anacon
da3/lib/python3.9/site-packages (from spacy) (2.28.1)
Collecting pydantic!=1.8,!=1.8.1,<1.11.0,>=1.7.4 (from spacy)
  Downloading pydantic-1.10.7-cp39-cp39-macosx_10_9_x86_64.whl (2.9 M
B)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 2.9/2.9 MB 8.2 MB/s eta
0:00:0000:01:00:01
Requirement already satisfied: jinja2 in /opt/anaconda3/lib/python3.9
/site-packages (from spacy) (2.11.3)
Requirement already satisfied: setuptools in /opt/anaconda3/lib/pytho
n3.9/site-packages (from spacy) (67.7.2)
Requirement already satisfied: packaging>=20.0 in /opt/anaconda3/lib/
python3.9/site-packages (from spacy) (21.3)
```

```
Collecting langcodes<4.0.0,>=3.2.0 (from spacy)
  Downloading langcodes-3.3.0-py3-none-any.whl (181 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 181.6/181.6 kB 7.6 MB/s
eta 0:00:00
Requirement already satisfied: pyparsing!=3.0.5,>=2.0.2 in /opt/anaco
nda3/lib/python3.9/site-packages (from packaging>=20.0->spacy) (3.0.
9)
Collecting typing-extensions>=4.2.0 (from pydantic!=1.8,!=1.8.1,<1.1
1.0,>=1.7.4->spacy)
  Downloading typing_extensions-4.5.0-py3-none-any.whl (27 kB)
Requirement already satisfied: charset-normalizer<3,>=2 in /opt/anaco
nda3/lib/python3.9/site-packages (from requests<3.0.0,>=2.13.0->spac
y) (2.0.4)
Requirement already satisfied: idna<4,>=2.5 in /opt/anaconda3/lib/pyt
hon3.9/site-packages (from requests<3.0.0,>=2.13.0->spacy) (3.3)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in /opt/anaconda
3/lib/python3.9/site-packages (from requests<3.0.0,>=2.13.0->spacy)
(1.26.11)
Requirement already satisfied: certifi>=2017.4.17 in /opt/anaconda3/l
ib/python3.9/site-packages (from requests<3.0.0,>=2.13.0->spacy) (202
2.9.24)
Collecting blis<0.8.0,>=0.7.8 (from thinc<8.2.0,>=8.1.8->spacy)
  Downloading blis-0.7.9-cp39-cp39-macosx_10_9_x86_64.whl (6.1 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 6.1/6.1 MB 9.7 MB/s eta
0:00:00:00:0100:01
Collecting confection<1.0.0,>=0.0.1 (from thinc<8.2.0,>=8.1.8->spacy)
  Downloading confection-0.0.4-py3-none-any.whl (32 kB)
Requirement already satisfied: click<9.0.0,>=7.1.1 in /opt/anaconda3/
lib/python3.9/site-packages (from typer<0.8.0,>=0.3.0->spacy) (8.0.4)
Requirement already satisfied: MarkupSafe>=0.23 in /opt/anaconda3/lib
/python3.9/site-packages (from jinja2->spacy) (2.0.1)
Installing collected packages: cymem, wasabi, typing-extensions, type
r, spacy-loggers, spacy-legacy, murmurhash, langcodes, catalogue, bli
s, srsly, pydantic, preshed, pathy, confection, thinc, spacy
  Attempting uninstall: typing-extensions
    Found existing installation: typing-extensions 3.10.0.2
    Uninstalling typing-extensions-3.10.0.2:
      Successfully uninstalled typing-extensions-3.10.0.2
ERROR: pip's dependency resolver does not currently take into account
all the packages that are installed. This behaviour is the source of
the following dependency conflicts.
tensorflow 2.6.5 requires typing-extensions<3.11,>=3.7, but you have
typing-extensions 4.5.0 which is incompatible.
Successfully installed blis-0.7.9 catalogue-2.0.8 confection-0.0.4 cy
mem-2.0.7 langcodes-3.3.0 murmurhash-1.0.9 pathy-0.10.1 preshed-3.0.8
pydantic-1.10.7 spacy-3.5.2 spacy-legacy-3.0.12 spacy-loggers-1.0.4 s
rsly-2.4.6 thinc-8.1.9 typer-0.7.0 typing-extensions-4.5.0 wasabi-1.
1.1
```

In [5]:
```
!python -m spacy download en_core_web_sm
```

```
Collecting en-core-web-sm==3.5.0
  Downloading https://github.com/explosion/spacy-models/releases/download/en_core_web_sm-3.5.0/en_core_web_sm-3.5.0-py3-none-any.whl (https://github.com/explosion/spacy-models/releases/download/en_core_web_sm-3.5.0/en_core_web_sm-3.5.0-py3-none-any.whl) (12.8 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 12.8/12.8 MB 10.8 MB/s
eta 0:00:0000:0100:01
Requirement already satisfied: spacy<3.6.0,>=3.5.0 in /opt/anaconda3/lib/python3.9/site-packages (from en-core-web-sm==3.5.0) (3.5.2)
Requirement already satisfied: spacy-legacy<3.1.0,>=3.0.11 in /opt/anaconda3/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (3.0.12)
Requirement already satisfied: spacy-loggers<2.0.0,>=1.0.0 in /opt/anaconda3/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (1.0.4)
Requirement already satisfied: murmurhash<1.1.0,>=0.28.0 in /opt/anaconda3/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (1.0.9)
Requirement already satisfied: cymem<2.1.0,>=2.0.2 in /opt/anaconda3/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (2.0.7)
Requirement already satisfied: preshed<3.1.0,>=3.0.2 in /opt/anaconda3/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (3.0.8)
Requirement already satisfied: thinc<8.2.0,>=8.1.8 in /opt/anaconda3/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (8.1.9)
Requirement already satisfied: wasabi<1.2.0,>=0.9.1 in /opt/anaconda3/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (1.1.1)
Requirement already satisfied: srsly<3.0.0,>=2.4.3 in /opt/anaconda3/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (2.4.6)
Requirement already satisfied: catalogue<2.1.0,>=2.0.6 in /opt/anaconda3/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (2.0.8)
Requirement already satisfied: typer<0.8.0,>=0.3.0 in /opt/anaconda3/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (0.7.0)
Requirement already satisfied: pathy>=0.10.0 in /opt/anaconda3/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (0.10.1)
Requirement already satisfied: smart-open<7.0.0,>=5.2.1 in /opt/anaconda3/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (5.2.1)
Requirement already satisfied: tqdm<5.0.0,>=4.38.0 in /opt/anaconda3/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (4.64.1)
Requirement already satisfied: numpy>=1.15.0 in /opt/anaconda3/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (1.19.5)
Requirement already satisfied: requests<3.0.0,>=2.13.0 in /opt/anaconda3/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (2.28.1)
```

```
Requirement already satisfied: pydantic!=1.8,!=1.8.1,<1.11.0,>=1.7.4
in /opt/anaconda3/lib/python3.9/site-packages (from spacy<3.6.0,>=3.
5.0->en-core-web-sm==3.5.0) (1.10.7)
Requirement already satisfied: jinja2 in /opt/anaconda3/lib/python3.9
/site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (2.1
1.3)
Requirement already satisfied: setuptools in /opt/anaconda3/lib/pytho
n3.9/site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0)
(67.7.2)
Requirement already satisfied: packaging>=20.0 in /opt/anaconda3/lib/
python3.9/site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-sm==3.
5.0) (21.3)
Requirement already satisfied: langcodes<4.0.0,>=3.2.0 in /opt/anacon
da3/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->en-core-we
b-sm==3.5.0) (3.3.0)
Requirement already satisfied: pyparsing!=3.0.5,>=2.0.2 in /opt/anaco
nda3/lib/python3.9/site-packages (from packaging>=20.0->spacy<3.6.0,>
=3.5.0->en-core-web-sm==3.5.0) (3.0.9)
Requirement already satisfied: typing-extensions>=4.2.0 in /opt/anaco
nda3/lib/python3.9/site-packages (from pydantic!=1.8,!=1.8.1,<1.11.0,
>=1.7.4->spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (4.5.0)
Requirement already satisfied: charset-normalizer<3,>=2 in /opt/anaco
nda3/lib/python3.9/site-packages (from requests<3.0.0,>=2.13.0->spacy
<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (2.0.4)
Requirement already satisfied: idna<4,>=2.5 in /opt/anaconda3/lib/pyt
hon3.9/site-packages (from requests<3.0.0,>=2.13.0->spacy<3.6.0,>=3.
5.0->en-core-web-sm==3.5.0) (3.3)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in /opt/anaconda
3/lib/python3.9/site-packages (from requests<3.0.0,>=2.13.0->spacy<3.
6.0,>=3.5.0->en-core-web-sm==3.5.0) (1.26.11)
Requirement already satisfied: certifi>=2017.4.17 in /opt/anaconda3/l
ib/python3.9/site-packages (from requests<3.0.0,>=2.13.0->spacy<3.6.
0,>=3.5.0->en-core-web-sm==3.5.0) (2022.9.24)
Requirement already satisfied: blis<0.8.0,>=0.7.8 in /opt/anaconda3/l
ib/python3.9/site-packages (from thinc<8.2.0,>=8.1.8->spacy<3.6.0,>=
3.5.0->en-core-web-sm==3.5.0) (0.7.9)
Requirement already satisfied: confection<1.0.0,>=0.0.1 in /opt/anaco
nda3/lib/python3.9/site-packages (from thinc<8.2.0,>=8.1.8->spacy<3.
6.0,>=3.5.0->en-core-web-sm==3.5.0) (0.0.4)
Requirement already satisfied: click<9.0.0,>=7.1.1 in /opt/anaconda3/
lib/python3.9/site-packages (from typer<0.8.0,>=0.3.0->spacy<3.6.0,>=
3.5.0->en-core-web-sm==3.5.0) (8.0.4)
Requirement already satisfied: MarkupSafe>=0.23 in /opt/anaconda3/lib
/python3.9/site-packages (from jinja2->spacy<3.6.0,>=3.5.0->en-core-w
eb-sm==3.5.0) (2.0.1)
Installing collected packages: en-core-web-sm
Successfully installed en-core-web-sm-3.5.0
✔ Download and installation successful
You can now load the package via spacy.load('en_core_web_sm')
```

In [1]:
```
import spacy
```

In [2]:
```
nlp = spacy.load("en_core_web_sm")
import en_core_web_sm
nlp = en_core_web_sm.load()
doc = nlp("This is a sentence.")
print([(w.text, w.pos_) for w in doc])
```

```
[('This', 'PRON'), ('is', 'AUX'), ('a', 'DET'), ('sentence', 'NOUN'),
('.', 'PUNCT')]
```

src: https://www.theonion.com/pedestrian-thankfully-just-dented-1850358958 (https://www.theonion.com/pedestrian-thankfully-just-dented-1850358958)

In [3]:
```
short_text = ("MINNEAPOLIS—Breathing a deep sigh of relief, local driv
short_text
```

Out[3]:
```
'MINNEAPOLIS—Breathing a deep sigh of relief, local driver Rob Glasse
r was reportedly thankful Friday after confirming the pedestrian he h
ad struck with his car was just dented. “Well, thank God it’s nothing
serious,” said Glasser, bending down to examine the small dent on the
pedestrian’s forehead, which he noted could not be larger than 3 inch
es across. “I was worried when I heard that crunch, but honestly, it’
s really no damage at all. Who knows, maybe he was already like that
before I bumped him. My car looks worse than he does. You could proba
bly take a plunger to that and have it sorted right in a minute. No h
arm, no foul, I guess.” At press time, sources reported Glasser had l
eft a note pinned to the man and sped off. '
```

In [4]:
```
short = nlp(short_text)
```

src: https://arstechnica.com/science/2021/12/mary-queen-of-scots-sealed-her-final-missive-with-an-intricate-spiral-letterlock/ (https://arstechnica.com/science/2021/12/mary-queen-of-scots-sealed-her-final-missive-with-an-intricate-spiral-letterlock/)

In [5]:
```
long_text = ("On the eve of her execution for treason in February 1587
             "The authors are an interdisciplinary team of researchers
             "As we reported previously, co-author Jana Dambrogio, a c
             "Dambrogio has been studying the practice of letterlockin
             "Queen Elizabeth I, Machiavelli, Galileo Galilei, and Mar
             "Earlier this year, Dambrogio's team was able to use X-ra
             "The unopened letters in the Brienne Collection meant tha
             "A high percentage of the material evidence for the lette
             "Certain common repairs might have been made that can als
             "Fortunately, the field of conservation is now shifting t
             "The raw materials for creating a spiral letterlock are s
             "The almost guaranteed destruction of the lock is precise
             "Mary, Queen of Scots, led a colorful life filled with po
             "The letter is now housed in the British Library. There i
             "Mary's final letter, dated February 8, 1587, was clearly
             "Now housed at the National Library of Scotland, this let
             "The team admits to being stumped by this particular spir
             "Dambrogio's group also acquired a 1570 spiral-locked let
             "Although it had been opened, breaking the lock, almost a
             "Elizabeth I used a spiral lock for a letter written to H
             "According to the authors, this makes it an excellent exa
             "A 1574 letter from Mary also used a variant of the spira
             "In addition, there were two spiral-locked letters by uni
             "The other letter is from 1633. The lock is missing, like
             "One goal for the Unlocking History Research Group is to
             "According to Smith, studying letterlocking can provide a
             "As additional letters with such locking mechanisms come
             "DOI: Electronic British Library Journal, 2021. 10.23636/
```

In [6]:
```
article = nlp(long_text)
```

1. Show the most common words in the article.

In [7]:
```
freq = {}
for token in short:
    if not token.is_stop and not token.is_punct and token.is_alpha:
        if token.text not in freq:
            freq[token.text] = 1
        else:
            freq[token.text] += 1
```

In [8]:
```
sorted_freq = sorted(freq.items(), key=lambda x: x[1], reverse=True)

for word, freq in sorted_freq[:5]:
    print(f"{word}: {freq}")
```

```
Glasser: 3
pedestrian: 2
car: 2
MINNEAPOLIS: 1
Breathing: 1
```

In [9]:
```
freq = {}
for token in article:
    if not token.is_stop and not token.is_punct and token.is_alpha:
        if token.text not in freq:
            freq[token.text] = 1
        else:
            freq[token.text] += 1
```

In [10]:
```
sorted_freq = sorted(freq.items(), key=lambda x: x[1], reverse=True)

for word, count in sorted_freq[:5]:
    print(f"{word}: {count}")
```

```
lock: 41
letter: 31
spiral: 21
paper: 17
letters: 16
```

2. Show the most common words under a part of speech. (i.e. NOUN: {'Bob':12, 'Alice':4,})

In [11]:
```
for token in article:
    if not token.is_stop and not token.is_punct and token.is_alpha:
        print(token.pos_)
```

```
NOUN
NOUN
NOUN
PROPN
PROPN
PROPN
PROPN
VERB
NOUN
PROPN
PROPN
PROPN
PROPN
VERB
NOUN
NOUN
VERB
ADJ
ADJ
NOUN
```

In [12]:
```
vcount = {}
for token in article:
    if not token.is_stop and not token.is_punct and token.is_alpha and
        if token.text not in vcount:
            vcount[token.text] = 1
        else:
            vcount[token.text] += 1
```

In [13]:
```
sorted_verbs = sorted(vcount.items(), key=lambda x: x[1], reverse=True
print("VERBS:", sorted_verbs[:5])
```

```
VERBS: [('letterlocking', 6), ('written', 6), ('locked', 5), ('opened
', 5), ('preserved', 5)]
```

In [14]:
```
ncount = {}
for token in article:
    if not token.is_stop and not token.is_punct and token.is_alpha and
        if token.text not in ncount:
            ncount[token.text] = 1
        else:
            ncount[token.text] += 1
```

In [15]:
```
sorted_nouns = sorted(ncount.items(), key=lambda x: x[1], reverse=True
print("NOUNS:", sorted_nouns[:5])
```

```
NOUNS: [('lock', 41), ('letter', 31), ('paper', 17), ('letters', 16),
('evidence', 14)]
```

In [16]:
```
pncount = {}
for token in article:
    if not token.is_stop and not token.is_punct and token.is_alpha and
        if token.text not in pncount:
            pncount[token.text] = 1
        else:
            pncount[token.text] += 1
```

In [17]:
```
sorted_pnouns = sorted(pncount.items(), key=lambda x: x[1], reverse=Tr
print("PROPER NOUNS:", sorted_pnouns[:5])
```

```
PROPER NOUNS: [('Mary', 12), ('Dambrogio', 12), ('Elizabeth', 8), ('Q
ueen', 5), ('Scots', 4)]
```

In [18]:
```
acount = {}
for token in article:
    if not token.is_stop and not token.is_punct and token.is_alpha and
        if token.text not in acount:
            acount[token.text] = 1
        else:
            acount[token.text] += 1
```

In [19]:
```
sorted_adj = sorted(acount.items(), key=lambda x: x[1], reverse=True)
print("ADJECTIVES:", sorted_adj[:5])
```

```
ADJECTIVES: [('spiral', 21), ('locked', 4), ('usual', 4), ('intricate
', 3), ('ingenious', 3)]
```

3. Find a subject/object relationship through the dependency parser in any sentence.

In [20]:
```
# working with shorter text block
for sent in short.sents:
    print(sent.text)
```

```
MINNEAPOLIS—Breathing a deep sigh of relief, local driver Rob Glasser
was reportedly thankful Friday after confirming the pedestrian he had
struck with his car was just dented.
“Well, thank God it’s nothing serious,” said Glasser, bending down to
examine the small dent on the pedestrian’s forehead, which he noted c
ould not be larger than 3 inches across.
“I was worried when I heard that crunch, but honestly, it’s really no
damage at all.
Who knows, maybe he was already like that before I bumped him.
My car looks worse than he does.
You could probably take a plunger to that and have it sorted right in
a minute.
No harm, no foul, I guess.”
At press time, sources reported Glasser had left a note pinned to the
man and sped off.
```

In [21]:
```
from spacy import displacy
sentence_spans = list(short.sents)
```

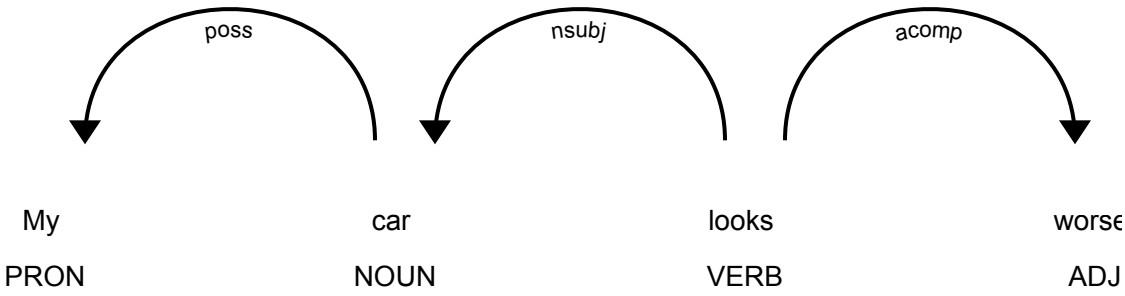
In [22]:
```
sentence_spans[4]
```

Out[22]:
```
My car looks worse than he does.
```

In [23]:
```
for token in sentence_spans[4]:
    print("text:", token.text, "\tdependency:", token.dep_, "\troot he
            "\tdependents:",[child for child in token.children])
```

```
text: My        dependency: poss        root head: car NOUN     depen
dents: []
text: car       dependency: nsubj       root head: looks VERB   depen
dents: [My]
text: looks     dependency: ROOT        root head: looks VERB   depen
dents: [car, worse, .]
text: worse     dependency: acomp       root head: looks VERB   depen
dents: [does]
text: than      dependency: mark        root head: does VERB    depen
dents: []
text: he        dependency: nsubj       root head: does VERB    depen
dents: []
text: does      dependency: advcl       root head: worse ADJ    depen
dents: [than, he]
text: .         dependency: punct       root head: looks VERB   depen
dents: []
```

In [24]:
```
displacy.render(sentence_spans[4], style="dep")
```



In [25]:
```
# working with longer article
for sent in article.sents:
    print(sent.text)
```

```
On the eve of her execution for treason in February 1587, Mary, Queen
of Scots, penned a letter to King Henri III of France and secured it
with a paper lock that featured an intricate spiral mechanism.
So-called "letterlocking" was a common practice to protect private le
tters from prying eyes, but this spiral lock is particularly ingeniou
s and delicate because it incorporates a built-in self-destruct featu
re, according to a new paper published in the Electronic British Libr
ary Journal.
The authors are an interdisciplinary team of researchers working unde
r the umbrella of the Unlocking History Research Group.
In this paper, they describe a dozen examples of a spiral lock in let
ters dated between 1568 and 1638, including one from Mary's former mo
ther-in-law, Catherine de Medici, as well as her arch-rival, Elizabet
h I, who signed Mary's death warrant.
As we reported previously, co-author Jana Dambrogio, a conservator at
MIT Libraries, coined the term "letterlocking" after discovering such
letters while a fellow at the Vatican Secret Archives in 2000.
The Vatican letters dated back to the 15th and 16th centuries, and th
ey featured strange slits and corners that had been sliced off.
```

In [26]:
```
sentence_spans = list(article.sents)
```

In [27]:
```
sentence_spans[4]
```

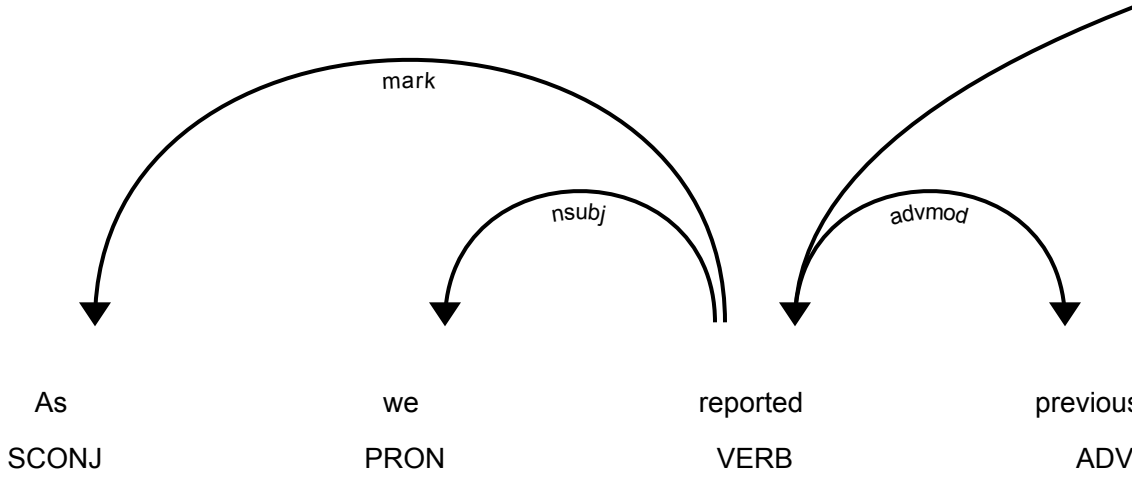
Out[27]:
```
As we reported previously, co-author Jana Dambrogio, a conservator at
MIT Libraries, coined the term "letterlocking" after discovering such
letters while a fellow at the Vatican Secret Archives in 2000.
```

In [28]:
```
for token in sentence_spans[4]:
    print("text:", token.text, "\tdependency:", token.dep_, "\troot he
            "\tdependents:",[child for child in token.children])
```

```
text: As        dependency: mark        root head: reported VERB
        dependents: []
text: we        dependency: nsubj       root head: reported VERB
        dependents: []
text: reported  dependency: advcl       root head: coined VERB  depen
dents: [As, we, previously, ,]
text: previously        dependency: advmod      root head: reported V
ERB     dependents: []
text: ,         dependency: punct       root head: reported VERB
        dependents: []
text: co        dependency: compound    root head: Dambrogio PROPN
        dependents: []
text: -         dependency: compound    root head: Dambrogio PROPN
        dependents: []
text: author    dependency: compound    root head: Dambrogio PROPN
        dependents: []
text: Jana      dependency: compound    root head: Dambrogio PROPN
        dependents: []
text: Dambrogio         dependency: nsubj       root head: coined VER
B       dependents: [co, -, author, Jana, ,, conservator, ,]
text: ,         dependency: punct       root head: Dambrogio PROPN
        dependents: []
text: a         dependency: det         root head: conservator NOUN
        dependents: []
text: conservator       dependency: appos       root head: Dambrogio
PROPN   dependents: [a, at]
text: at        dependency: prep        root head: conservator NOUN
        dependents: [Libraries]
text: MIT       dependency: compound    root head: Libraries PROPN
        dependents: []
text: Libraries         dependency: pobj        root head: at ADP
        dependents: [MIT]
text: ,         dependency: punct       root head: Dambrogio PROPN
        dependents: []
text: coined    dependency: ROOT        root head: coined VERB  depen
dents: [reported, Dambrogio, letterlocking, .]
text: the       dependency: det         root head: term NOUN    depen
dents: []
text: term      dependency: nsubj       root head: letterlocking VERB
        dependents: [the]
text: "         dependency: punct       root head: letterlocking VERB
        dependents: []
text: letterlocking     dependency: ccomp       root head: coined VER
B       dependents: [term, ", ", after]
text: "         dependency: punct       root head: letterlocking VERB
        dependents: []
text: after     dependency: prep        root head: letterlocking VERB
        dependents: [discovering]
text: discovering       dependency: pcomp       root head: after ADP
        dependents: [letters, fellow]
text: such      dependency: amod        root head: letters NOUN
        dependents: []
```

```
text: letters   dependency: dobj        root head: discovering VERB
        dependents: [such]
text: while     dependency: mark        root head: fellow NOUN  depen
dents: []
text: a         dependency: det         root head: fellow NOUN  depen
dents: []
text: fellow    dependency: advcl       root head: discovering VERB
        dependents: [while, a, at, in]
text: at        dependency: prep        root head: fellow NOUN  depen
dents: [Archives]
text: the       dependency: det         root head: Archives PROPN
        dependents: []
text: Vatican   dependency: compound    root head: Archives PROPN
        dependents: []
text: Secret    dependency: compound    root head: Archives PROPN
        dependents: []
text: Archives  dependency: pobj        root head: at ADP       depen
dents: [the, Vatican, Secret]
text: in        dependency: prep        root head: fellow NOUN  depen
dents: [2000]
text: 2000      dependency: pobj        root head: in ADP       depen
dents: []
text: .         dependency: punct       root head: coined VERB  depen
dents: []
```

In [29]:
```
displacy.render(sentence_spans[4], style="dep")
```



4. Show the most common Entities and their types.

In [30]:
```
import pandas as pd
import numpy as np
```

In [31]:
```
# working with shorter text block
short_ents = {}
for token in short:
    if not token.is_stop and not token.is_punct and token.is_alpha:
        if not token.ent_iob_ == 'O':
            print(token.text, token.ent_type_, token.ent_iob_)
            short_ents[token.text] = token.ent_type_
```

```
Rob PERSON B
Glasser PERSON I
Friday DATE B
Glasser ORG B
larger QUANTITY B
inches QUANTITY I
Glasser ORG B
```

In [32]:
```
short_df_ents = pd.DataFrame(short_ents.items(), columns =['word', 'en
```

In [33]:
```
short_df_ents.nunique()
```

Out[33]:
```
word      5
entity    4
dtype: int64
```

In [34]:
```
# looking for entities in the long article
article_ents = {}
for token in article:
    if not token.is_stop and not token.is_punct and token.is_alpha:
        if not token.ent_iob_ == 'O':
            print(token.text, token.ent_type_, token.ent_iob_)
            article_ents[token.text] = token.ent_type_
```

```
February DATE B
Mary PERSON B
King PERSON B
Henri PERSON I
III PERSON I
France GPE B
Electronic ORG I
British ORG I
Library ORG I
Journal ORG I
Unlocking ORG I
History ORG I
Research ORG I
Group ORG I
dozen CARDINAL B
Mary PERSON B
Catherine PERSON B
de PERSON I
Medici PERSON I
```

5. Find Entites and their dependency (hint: entity.root.head)

In [36]:
```
article.ents
```

Out[36]:
```
(February 1587,
 Mary,
 King Henri III,
 France,
 the Electronic British Library Journal,
 the Unlocking History Research Group,
 dozen,
 between 1568 and 1638,
 one,
 Mary,
 Catherine de Medici,
 Elizabeth,
 Mary,
 Jana Dambrogio,
 MIT Libraries,
 2000,
 Vatican,
 the 15th and 16th centuries,
 Dambrogio,
```

In [37]:
```
for ents in article.ents:
    print("text:", ents.text, "\troot text", ents.root.text, "\tdepend
```

```
text: February 1587     root text February      dependency: in
text: Mary      root text Mary  dependency: penned
text: King Henri III    root text III   dependency: to
text: France    root text France        dependency: of
text: the Electronic British Library Journal    root text Journal
        dependency: in
text: the Unlocking History Research Group      root text Group
        dependency: of
text: dozen     root text dozen         dependency: examples
text: between 1568 and 1638     root text between       dependency: d
ated
text: one       root text one   dependency: including
text: Mary      root text Mary  dependency: mother
text: Catherine de Medici       root text Medici        dependency: d
escribe
text: Elizabeth         root text Elizabeth     dependency: I
text: Mary      root text Mary  dependency: warrant
text: Jana Dambrogio    root text Dambrogio     dependency: coined
text: MIT Libraries     root text Libraries     dependency: at
```

6. Find the most similar words in the article

In [38]:
```
!python -m spacy download en_core_web_md
nlp = spacy.load("en_core_web_md")
```

```
Collecting en-core-web-md==3.5.0
  Downloading https://github.com/explosion/spacy-models/releases/download/en_core_web_md-3.5.0/en_core_web_md-3.5.0-py3-none-any.whl (https://github.com/explosion/spacy-models/releases/download/en_core_web_md-3.5.0/en_core_web_md-3.5.0-py3-none-any.whl) (42.8 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 42.8/42.8 MB 10.1 MB/s
eta 0:00:0000:0100:01
Requirement already satisfied: spacy<3.6.0,>=3.5.0 in /opt/anaconda3/lib/python3.9/site-packages (from en-core-web-md==3.5.0) (3.5.2)
Requirement already satisfied: spacy-legacy<3.1.0,>=3.0.11 in /opt/anaconda3/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-md==3.5.0) (3.0.12)
Requirement already satisfied: spacy-loggers<2.0.0,>=1.0.0 in /opt/anaconda3/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-md==3.5.0) (1.0.4)
Requirement already satisfied: murmurhash<1.1.0,>=0.28.0 in /opt/anaconda3/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-md==3.5.0) (1.0.9)
Requirement already satisfied: cymem<2.1.0,>=2.0.2 in /opt/anaconda3/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-md==3.5.0) (2.0.7)
Requirement already satisfied: preshed<3.1.0,>=3.0.2 in /opt/anaconda3/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-md==3.5.0) (3.0.8)
Requirement already satisfied: thinc<8.2.0,>=8.1.8 in /opt/anaconda3/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-md==3.5.0) (8.1.9)
Requirement already satisfied: wasabi<1.2.0,>=0.9.1 in /opt/anaconda3/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-md==3.5.0) (1.1.1)
Requirement already satisfied: srsly<3.0.0,>=2.4.3 in /opt/anaconda3/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-md==3.5.0) (2.4.6)
Requirement already satisfied: catalogue<2.1.0,>=2.0.6 in /opt/anaconda3/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-md==3.5.0) (2.0.8)
Requirement already satisfied: typer<0.8.0,>=0.3.0 in /opt/anaconda3/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-md==3.5.0) (0.7.0)
Requirement already satisfied: pathy>=0.10.0 in /opt/anaconda3/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-md==3.5.0) (0.10.1)
Requirement already satisfied: smart-open<7.0.0,>=5.2.1 in /opt/anaconda3/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-md==3.5.0) (5.2.1)
Requirement already satisfied: tqdm<5.0.0,>=4.38.0 in /opt/anaconda3/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-md==3.5.0) (4.64.1)
Requirement already satisfied: numpy>=1.15.0 in /opt/anaconda3/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-md==3.5.0) (1.19.5)
Requirement already satisfied: requests<3.0.0,>=2.13.0 in /opt/anaconda3/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->en-core-we
```

```
b-md==3.5.0) (2.28.1)
Requirement already satisfied: pydantic!=1.8,!=1.8.1,<1.11.0,>=1.7.4
in /opt/anaconda3/lib/python3.9/site-packages (from spacy<3.6.0,>=3.
5.0->en-core-web-md==3.5.0) (1.10.7)
Requirement already satisfied: jinja2 in /opt/anaconda3/lib/python3.9
/site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-md==3.5.0) (2.1
1.3)
Requirement already satisfied: setuptools in /opt/anaconda3/lib/pytho
n3.9/site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-md==3.5.0)
(67.7.2)
Requirement already satisfied: packaging>=20.0 in /opt/anaconda3/lib/
python3.9/site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-md==3.
5.0) (21.3)
Requirement already satisfied: langcodes<4.0.0,>=3.2.0 in /opt/anacon
da3/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->en-core-we
b-md==3.5.0) (3.3.0)
Requirement already satisfied: pyparsing!=3.0.5,>=2.0.2 in /opt/anaco
nda3/lib/python3.9/site-packages (from packaging>=20.0->spacy<3.6.0,>
=3.5.0->en-core-web-md==3.5.0) (3.0.9)
Requirement already satisfied: typing-extensions>=4.2.0 in /opt/anaco
nda3/lib/python3.9/site-packages (from pydantic!=1.8,!=1.8.1,<1.11.0,
>=1.7.4->spacy<3.6.0,>=3.5.0->en-core-web-md==3.5.0) (4.5.0)
Requirement already satisfied: charset-normalizer<3,>=2 in /opt/anaco
nda3/lib/python3.9/site-packages (from requests<3.0.0,>=2.13.0->spacy
<3.6.0,>=3.5.0->en-core-web-md==3.5.0) (2.0.4)
Requirement already satisfied: idna<4,>=2.5 in /opt/anaconda3/lib/pyt
hon3.9/site-packages (from requests<3.0.0,>=2.13.0->spacy<3.6.0,>=3.
5.0->en-core-web-md==3.5.0) (3.3)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in /opt/anaconda
3/lib/python3.9/site-packages (from requests<3.0.0,>=2.13.0->spacy<3.
6.0,>=3.5.0->en-core-web-md==3.5.0) (1.26.11)
Requirement already satisfied: certifi>=2017.4.17 in /opt/anaconda3/l
ib/python3.9/site-packages (from requests<3.0.0,>=2.13.0->spacy<3.6.
0,>=3.5.0->en-core-web-md==3.5.0) (2022.9.24)
Requirement already satisfied: blis<0.8.0,>=0.7.8 in /opt/anaconda3/l
ib/python3.9/site-packages (from thinc<8.2.0,>=8.1.8->spacy<3.6.0,>=
3.5.0->en-core-web-md==3.5.0) (0.7.9)
Requirement already satisfied: confection<1.0.0,>=0.0.1 in /opt/anaco
nda3/lib/python3.9/site-packages (from thinc<8.2.0,>=8.1.8->spacy<3.
6.0,>=3.5.0->en-core-web-md==3.5.0) (0.0.4)
Requirement already satisfied: click<9.0.0,>=7.1.1 in /opt/anaconda3/
lib/python3.9/site-packages (from typer<0.8.0,>=0.3.0->spacy<3.6.0,>=
3.5.0->en-core-web-md==3.5.0) (8.0.4)
Requirement already satisfied: MarkupSafe>=0.23 in /opt/anaconda3/lib
/python3.9/site-packages (from jinja2->spacy<3.6.0,>=3.5.0->en-core-w
eb-md==3.5.0) (2.0.1)

[notice] A new release of pip is available: 23.1.1 -> 23.1.2
[notice] To update, run: pip install --upgrade pip
✔ Download and installation successful
You can now load the package via spacy.load('en_core_web_md')
```

Playing with the short article...

In [39]:
```
for token in short:
    print(token.text, token.has_vector, token.vector_norm, token.is_oo
```

```
MINNEAPOLIS True 7.70223 True
— True 7.9372907 True
Breathing True 7.6862473 True
a True 7.46388 True
deep True 7.8781786 True
sigh True 5.191081 True
of True 8.503868 True
relief True 7.36309 True
, True 8.605738 True
local True 8.424441 True
driver True 6.6867614 True
Rob True 9.381631 True
Glasser True 8.183234 True
was True 7.4873757 True
reportedly True 9.882214 True
thankful True 8.498257 True
Friday True 9.483415 True
after True 8.673904 True
confirming True 10.041585 True
```

In [40]:
```
wordmatch = []
for token1 in short:
    if not token1.is_stop and not token1.is_punct and token1.is_alpha:
        for token2 in short:
            if not token2.is_stop and not token2.is_punct and toke
                print(token1.text, token2.text, token1.similarity(
                wordmatch.append([token1.text, token2.text, token1
```

```
MINNEAPOLIS MINNEAPOLIS 1.0
MINNEAPOLIS Breathing 0.14857754111289978
MINNEAPOLIS deep 0.04189905896782875
MINNEAPOLIS sigh 0.19953212141990662
MINNEAPOLIS relief 0.3221192955970764
MINNEAPOLIS local 0.07103551179170609
MINNEAPOLIS driver 0.2511027455329895
MINNEAPOLIS Rob 0.31025511026382446
MINNEAPOLIS Glasser 0.4344996213912964
MINNEAPOLIS reportedly -0.09513086080551147
MINNEAPOLIS thankful 0.11801502108573914
MINNEAPOLIS Friday 0.4273734986782074
MINNEAPOLIS confirming 0.05401059240102768
MINNEAPOLIS pedestrian 0.248304083943367
MINNEAPOLIS struck -0.017169537022709846
MINNEAPOLIS car 0.19550490379333496
MINNEAPOLIS dented 0.07231322675943375
MINNEAPOLIS thank 0.05907164141535759
MINNEAPOLIS God 0.3531319200992584
```

In [41]:
```
sims = pd.DataFrame(wordmatch, columns =['word1', 'word2', 'similar'])
sims
```

Out[41]:

	word1	word2	similar
0	MINNEAPOLIS	MINNEAPOLIS	1.000000
1	MINNEAPOLIS	Breathing	0.148578
2	MINNEAPOLIS	deep	0.041899
3	MINNEAPOLIS	sigh	0.199532
4	MINNEAPOLIS	relief	0.322119
...	...	...	...
3595	sped	left	0.470161
3596	sped	note	-0.100636
3597	sped	pinned	0.392562
3598	sped	man	0.074111
3599	sped	sped	1.000000

3600 rows × 3 columns

In [42]:
```
sims.drop(sims[sims.word1 == sims.word2].index, inplace=True)
sims
```

Out[42]:

	word1	word2	similar
1	MINNEAPOLIS	Breathing	0.148578
2	MINNEAPOLIS	deep	0.041899
3	MINNEAPOLIS	sigh	0.199532
4	MINNEAPOLIS	relief	0.322119
5	MINNEAPOLIS	local	0.071036
...	...	...	...
3594	sped	Glasser	-0.020840
3595	sped	left	0.470161
3596	sped	note	-0.100636
3597	sped	pinned	0.392562
3598	sped	man	0.074111

3530 rows × 3 columns

In [43]:
```
sims = sims[pd.DataFrame(np.sort(sims[['word1','word2']].values,1)).du
```

In [44]:
```
sims.sort_values(by=['similar'], ascending=False)
```

Out[44]:

	word1	word2	similar
3314	left	struck	0.832439
2311	bumped	heard	0.824619
2488	worse	larger	0.821064
1382	small	deep	0.765311
1272	bending	confirming	0.754127
...	...	...	...
1240	Glasser	looks	-0.255124
3513	man	honestly	-0.264137
2423	looks	small	-0.264331
3516	man	maybe	-0.265780
3549	sped	reportedly	-0.306034

1990 rows × 3 columns

Now the long article...

In [45]:
```
for token in article:
    print(token.text, token.has_vector, token.vector_norm, token.is_oo
```

```
On True 9.251914 True
the True 9.057247 True
eve True 6.008178 True
of True 8.758128 True
her True 8.11128 True
execution True 5.418924 True
for True 7.724888 True
treason True 6.4567266 True
in True 8.664272 True
February True 10.999555 True
1587 True 9.338388 True
, True 8.479347 True
Mary True 9.384924 True
, True 8.763777 True
Queen True 7.7121825 True
of True 9.96618 True
Scots True 8.631713 True
, True 9.489109 True
penned True 8.428752 True
```

In [46]:
```
wordmatch = []
for token1 in article:
    if not token1.is_stop and not token1.is_punct and token1.is_alpha:
        for token2 in article:
            if not token2.is_stop and not token2.is_punct and toke
                print(token1.text, token2.text, token1.similarity(
                wordmatch.append([token1.text, token2.text, token1
```

```
eve eve 1.0
eve execution 0.5623785257339478
eve treason 0.2912912666797638
eve February 0.3517320454120636
eve Mary 0.24419082701206207
eve Queen 0.2718174457550049
eve Scots 0.2571503818035126
eve penned 0.07273944467306137
eve letter 0.5419021844863892
eve King 0.1850900650024414
eve Henri 0.10094331204891205
eve III 0.33087581396102905
eve France 0.41677331924438477
eve secured 0.01711602322757244
eve paper 0.2934509515762329
eve lock 0.49582287669181824
eve featured -0.034924812614917755
eve intricate 0.10223338007926941
eve spiral 0.2363453060388565
```

In [47]:
```
sims = pd.DataFrame(wordmatch, columns =['word1', 'word2', 'similar'])
sims
```

Out[47]:

	word1	word2	similar
0	eve	eve	1.000000
1	eve	execution	0.562379
2	eve	treason	0.291291
3	eve	February	0.351732
4	eve	Mary	0.244191
...	...	...	...
1223231	DOIs	British	0.123152
1223232	DOIs	Library	0.101893
1223233	DOIs	Journal	0.339970
1223234	DOIs	gyhc	0.059619
1223235	DOIs	DOIs	1.000000

1223236 rows × 3 columns

In [48]:
```
sims.drop(sims[sims.word1 == sims.word2].index, inplace=True)
sims
```

Out[48]:

	word1	word2	similar
1	eve	execution	0.562379
2	eve	treason	0.291291
3	eve	February	0.351732
4	eve	Mary	0.244191
5	eve	Queen	0.271817
...	...	...	...
1223230	DOIs	Electronic	0.043682
1223231	DOIs	British	0.123152
1223232	DOIs	Library	0.101893
1223233	DOIs	Journal	0.339970
1223234	DOIs	gyhc	0.059619

1216864 rows × 3 columns

In [49]:
```
sims = sims[pd.DataFrame(np.sort(sims[['word1','word2']].values,1)).du
sims
```

Out[49]:

	word1	word2	similar
29	eve	spiral	0.047444
30	eve	lock	0.411784
41	eve	paper	0.487388
57	eve	paper	0.495836
61	eve	spiral	0.144801
...	...	...	...
1223230	DOIs	Electronic	0.043682
1223231	DOIs	British	0.123152
1223232	DOIs	Library	0.101893
1223233	DOIs	Journal	0.339970
1223234	DOIs	gyhc	0.059619

1024974 rows × 3 columns

In [50]:
```
sims[~(pd.get_dummies(sims.word1).add(pd.get_dummies(sims.word2), fill
```

Out[50]:

	word1	word2	similar
29	eve	spiral	0.047444
30	eve	lock	0.411784
41	eve	paper	0.487388
63	eve	letters	0.325608
66	eve	Mary	0.297831
...	...	...	...
1223225	DOIs	archives	0.308498
1223226	DOIs	started	-0.113308
1223227	DOIs	looking	-0.109028
1223229	DOIs	iceberg	0.263524
1223234	DOIs	gyhc	0.059619

191890 rows × 3 columns

In [51]:
```
sims = sims.drop_duplicates(subset=['word1', 'word2'])
```

In [52]:
```
sims.sort_values(by=['similar'], ascending=False)
```

Out[52]:

	word1	word2	similar
505826	composed	identified	0.938752
636055	placed	sliced	0.915345
122311	folded	papered	0.910078
122250	folded	filled	0.909937
143325	tampered	papered	0.906364
...	...	...	...
411623	fill	simple	-0.357765
1064344	Frances	fill	-0.372835
633004	British	fill	-0.374001
161494	dates	spiral	-0.393375
411476	fill	British	-0.395347

288950 rows × 3 columns

In [ ]:
```

```