Jovany Gonzalez
CISB 63

Text Summarization

For this project, I wanted to see what is the most common words in one topic and try to see if they appear in other various topics from BBC News.

Link to my dataset my dataset. (https://www.kaggle.com/datasets/pariza/bbc-news-summary)

Personal Github link (https://github.com/ojdgonzo/CISB62_Midterm---JGonzalez)

In [105]:

```
# importing our necessary libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import nltk
from nltk.corpus import stopwords
from nltk.tokenize import sent_tokenize, word_tokenize
from nltk.probability import FreqDist
import spacy
from spacy import displacy
from wordcloud import WordCloud
from PIL import Image


# ignoring warnings
import warnings
warnings.filterwarnings('ignore')
```

In [37]:

```
# I will only use a small portion of the dataset
# file path
# data\BBC News Summary\BBC News Summary\News Articles\business
text1 = open("data/BBC News Summary/BBC News Summary/News Articles/business
text2 = open("data/BBC News Summary/BBC News Summary/News Articles/enterta
text3 = open("data/BBC News Summary/BBC News Summary/News Articles/politic


# reading our text files
text1 = text1.read()
text2 = text2.read()
text3 = text3.read()

print(f"This is the length of the first text document: {len(text1)}")
print("--- \nText: \n")
print(text1)
```

```
This is the length of the first text document: 2560
---
Text: 

Ad sales boost Time Warner profit

Quarterly profits at US media giant TimeWarner jumped 76% to $1.13bn (Â£6
00m) for the three months to December, from $639m year-earlier.

The firm, which is now one of the biggest investors in Google, benefited
from sales of high-speed internet connections and higher advert sales. Ti
meWarner said fourth quarter sales rose 2% to $11.1bn from $10.9bn. Its p
rofits were buoyed by one-off gains which offset a profit dip at Warner B
ros, and less users for AOL.

Time Warner said on Friday that it now owns 8% of search-engine Google. B
ut its own internet business, AOL, had has mixed fortunes. It lost 464,00
0 subscribers in the fourth quarter profits were lower than in the preced
ing three quarters. However, the company said AOL's underlying profit bef
ore exceptional items rose 8% on the back of stronger internet advertisin
g revenues. It hopes to increase subscribers by offering the online servi
ce free to TimeWarner internet customers and will try to sign up AOL's ex
isting customers for high-speed broadband. TimeWarner also has to restate
2000 and 2003 results following a probe by the US Securities Exchange Com
mission (SEC), which is close to concluding.

Time Warner's fourth quarter profits were slightly better than analysts'
expectations. But its film division saw profits slump 27% to $284m, helpe
d by box-office flops Alexander and Catwoman, a sharp contrast to year-ea
rlier, when the third and final film in the Lord of the Rings trilogy boo
sted results. For the full-year, TimeWarner posted a profit of $3.36bn, u
p 27% from its 2003 performance, while revenues grew 6.4% to $42.09bn. "O
ur financial performance was strong, meeting or exceeding all of our full
-year objectives and greatly enhancing our flexibility," chairman and chi
ef executive Richard Parsons said. For 2005, TimeWarner is projecting ope
rating earnings growth of around 5%, and also expects higher revenue and
wider profit margins.

TimeWarner is to restate its accounts as part of efforts to resolve an in
quiry into AOL by US market regulators. It has already offered to pay $30
```

```
0m to settle charges, in a deal that is under review by the SEC. The comp
any said it was unable to estimate the amount it needed to set aside for
legal reserves, which it previously set at $500m. It intends to adjust th
e way it accounts for a deal with German music publisher Bertelsmann's pu
rchase of a stake in AOL Europe, which it had reported as advertising rev
enue. It will now book the sale of its stake in AOL Europe as a loss on t
he value of that stake.
```

In [38]:

```python
print(f"This is the length of the second text document: {len(text2)}")
print("--- \nText: \n")
```

```
This is the length of the second text document: 1582
---
Text: 

Jarre joins fairytale celebration

French musician Jean-Michel Jarre is to perform at a concert in Copenhage
n to mark the bicentennial of the birth of writer Hans Christian Anderse
n.

Denmark is holding a three-day celebration of the life of the fairy-tale
author, with a concert at Parken stadium on 2 April. Other stars are expe
cted to join the line-up in the coming months, and the Danish royal famil
y will attend. "Christian Andersen's fairy tales are timeless and univers
al," said Jarre. "For all of us, at any age there is always - beyond the
pure enjoyment of the tale - a message to learn." There are year-long cel
ebrations planned across the world to celebrate Andersen and his work, wh
ich includes The Emperor's New Clothes and The Little Mermaid. Denmark's
Crown Prince Frederik and Crown Princess Mary visited New York on Monday
to help promote the festivities. The pair were at a Manhattan library to
honour US literary critic Harold Bloom "the international icon we thought
we knew so well".

"Bloom recognizes the darker aspects of Andersen's authorship," Prince Fr
ederik said. Bloom is to be formally presented with the Hans Christian An
dersen Award this spring in Anderson's hometown of Odense. The royal coup
le also visited the Hans Christian Anderson School complex, where Queen M
ary read The Ugly Duckling to the young audience. Later at a gala dinner,
Danish supermodel Helena Christensen was named a Hans Christian Andersen
ambassador. Other ambassadors include actors Harvey Keitel and Sir Roger
Moore, athlete Cathy Freeman and Brazilian soccer legend Pele.
```

In [39]:

```python
print(f"This is the length of the first text document: {len(text3)}")
print("--- \nText: \n")
print(text3)
```

```
This is the length of the first text document: 3109
---
Text: 

Hewitt decries 'career sexism'

Plans to extend paid maternity leave beyond six months should be prominen
t in Labour's election manifesto, the Trade and Industry Secretary has sa
id.

Patricia Hewitt said the cost of the proposals was being evaluated, but i
t was an "increasingly high priority" and a "shared goal across governmen
t". Ms Hewitt was speaking at a gender and productivity seminar organised
by the Equal Opportunities Commission (EOC). Mothers can currently take u
p to six months' paid leave - and six unpaid. Ms Hewitt told the seminar:
"Clearly, one of the things we need to do in the future is to extend the
period of payment for maternity leave beyond the first six months into th
e second six months. "We are looking at how quickly we can do that, becau
se obviously there are cost implications because the taxpayer reimburses
the employers for the cost of that."

Ms Hewitt also announced a new drive to help women who want to work in ma
le dominated sectors, saying sexism at work was still preventing women re
aching their full potential. Plans include funding for universities to he
lp female science and engineering graduates find jobs and "taster course
s" for men and women in non-traditional jobs. Women in full-time work ear
n 19% less than men, according to the Equal Opportunities Commission (EO
C).

The minister told delegates that getting rid of "career sexism" was vital
to closing the gender pay gap.

"Career sexism limits opportunities for women of all ages and prevents th
em from achieving their full potential. "It is simply wrong to assume som
eone cannot do a job on the grounds of their sex," she said. Earlier, she
told BBC Radio 4's Today programme: "What we are talking about here is th
e fact that about six out of 20 women work in jobs that are low-paid and
typically dominated by women, so we have got very segregated employment.
"Unfortunately, in some cases, this reflects very old-fashioned and stere
otypical ideas about the appropriate jobs for women, or indeed for men. "
Career sexism is about saying that engineering, for instance, where only
10% of employees are women, is really a male-dominated industry. Construc
tion is even worse. "But it is also about saying childcare jobs are reall
y there for women and not suitable for men. Career sexism goes both way
s."

She added that while progress had been made, there was still a gap in pay
figures. "The average woman working full-time is being paid about 80p for
every pound a man is earning. For women working part-time it is 60p." The
Department for Trade and Industry will also provide funding to help a new
pay experts panel run by the TUC.

It has been set up to advise hundreds of companies on equal wage policie
```

```
s. Research conducted by the EOC last year revealed that many Britons bel
ieve the pay gap between men and women is the result of "natural differen
ces" between the sexes. Women hold less than 10% of the top positions in
FTSE 100 companies, the police, the judiciary and trade unions, according
to their figures. And retired women have just over half the income of the
ir male counterparts on average.
```

In [43]:

```
# tokenizing our text
stoken1 = sent_tokenize(text1)
stoken2 = sent_tokenize(text2)
stoken3 = sent_tokenize(text3)
```

In [46]:

```
# checking the other sentences length
print(len(stoken1))
print(len(stoken2))
print(len(stoken3))
```

```
20
13
26
```

In [62]:

```
# a small sample of the senteces from each token
stoken1[4:8]
```

Out[62]:

```
['Time Warner said on Friday that it now owns 8% of search-engine Googl
e.',
 'But its own internet business, AOL, had has mixed fortunes.',
 'It lost 464,000 subscribers in the fourth quarter profits were lower th
an in the preceding three quarters.',
 "However, the company said AOL's underlying profit before exceptional it
ems rose 8% on the back of stronger internet advertising revenues."]
```

In [58]:

```
stoken2[8:13]
```

Out[58]:

```
['"Bloom recognizes the darker aspects of Andersen\'s authorship," Prince
Frederik said.',
 "Bloom is to be formally presented with the Hans Christian Andersen Awar
d this spring in Anderson's hometown of Odense.",
 'The royal couple also visited the Hans Christian Anderson School comple
x, where Queen Mary read The Ugly Duckling to the young audience.',
 'Later at a gala dinner, Danish supermodel Helena Christensen was named
a Hans Christian Andersen ambassador.',
 'Other ambassadors include actors Harvey Keitel and Sir Roger Moore, ath
lete Cathy Freeman and Brazilian soccer legend Pele.']
```

In [61]:
```
tokens3[16:22]
```

Out[61]:
```
['"But it is also about saying childcare jobs are really there for women
and not suitable for men.',
 'Career sexism goes both ways."',
 'She added that while progress had been made, there was still a gap in p
ay figures.',
 '"The average woman working full-time is being paid about 80p for every
pound a man is earning.',
 'For women working part-time it is 60p."',
 'The Department for Trade and Industry will also provide funding to help
a new pay experts panel run by the TUC.']
```

In [64]:
```
# word tokenizing
words1 = word_tokenize(text1)
words2 = word_tokenize(text2)
words3 = word_tokenize(text3)
```

In [67]:
```
# total amount of words in each document
print(len(words1))
print(len(words2))
print(len(words3))
```

```
490
288
604
```

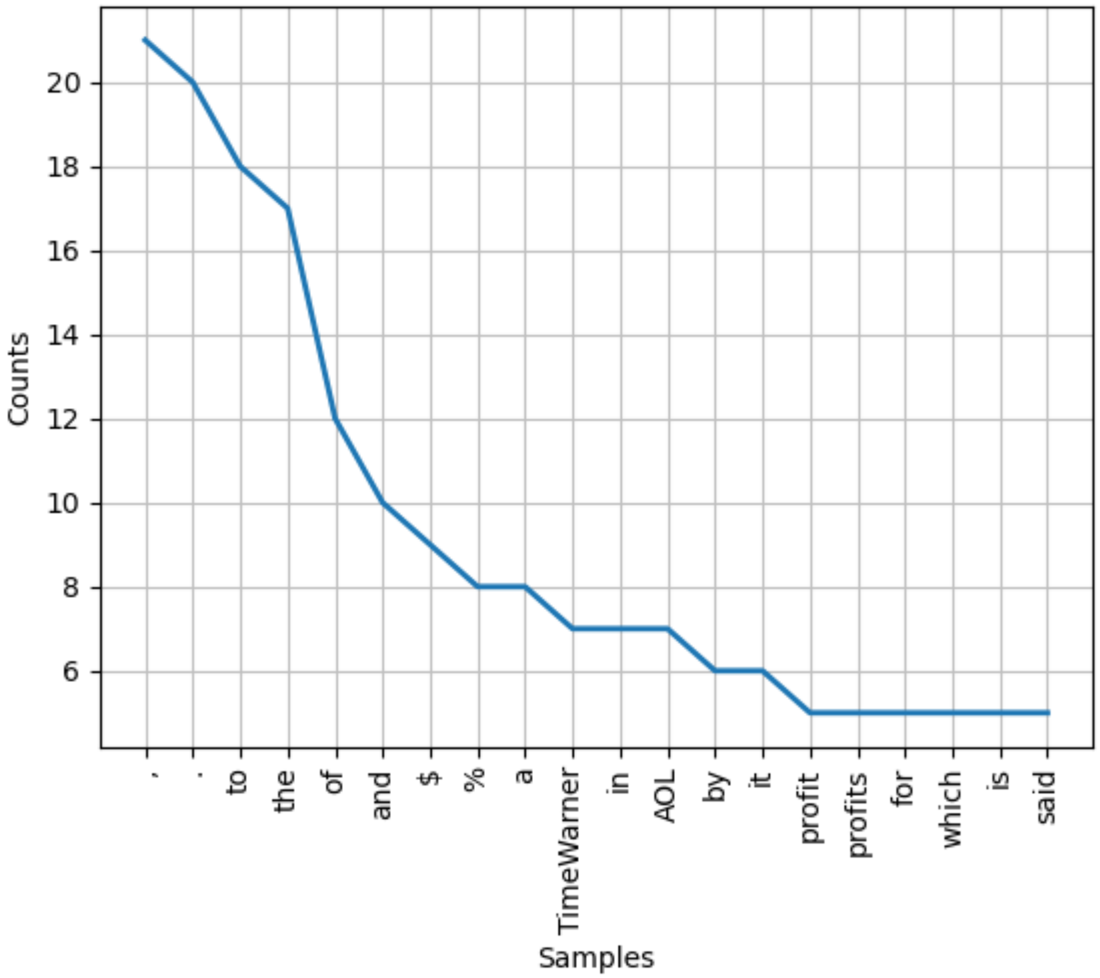
In [69]:
```
fdist1 = FreqDist(words1)
fdist2 = FreqDist(words2)
fdist3 = FreqDist(words3)
```

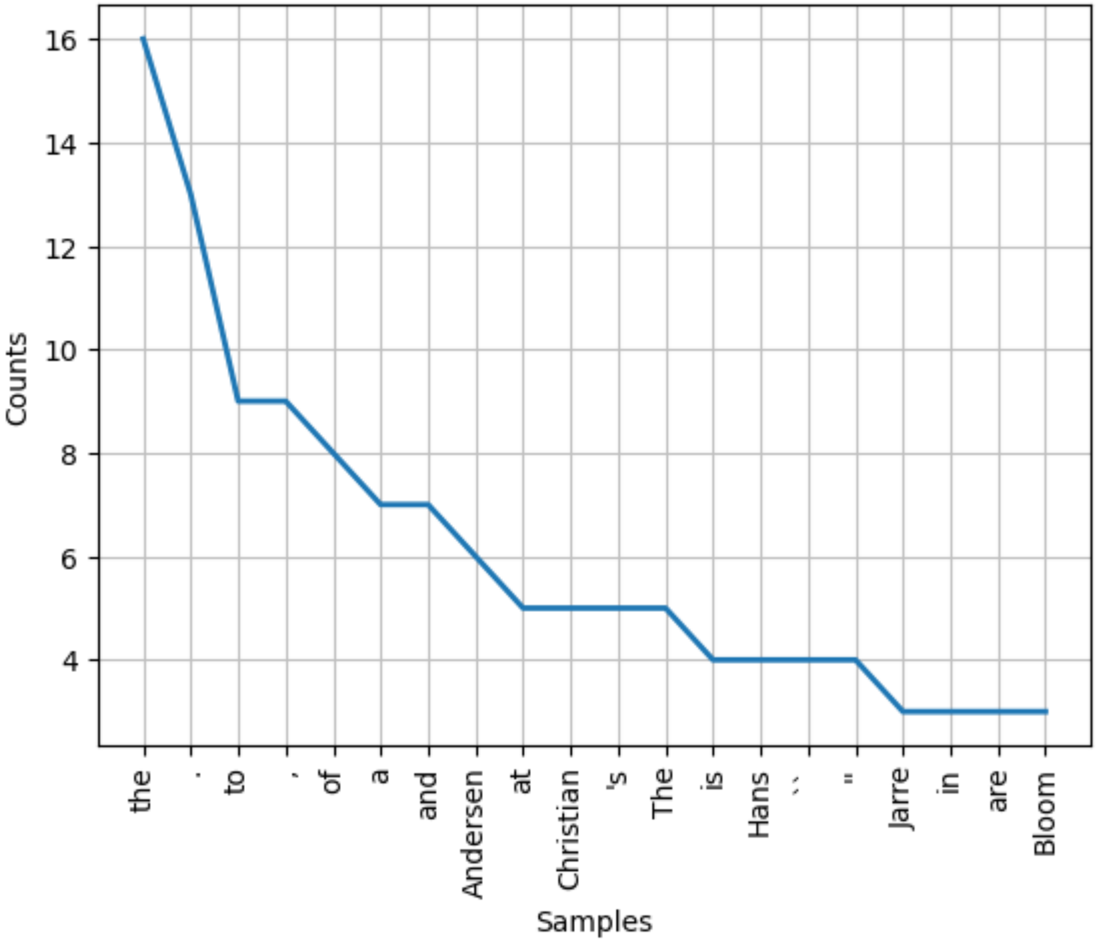
In [73]:
```
# finding the most common words between the three different articles
print(f"These are the 20 most commons words in the Business article: \n {f
print(f"These are the 20 most commons words in the Entertainment article: \
print(f"These are the 20 most commons words in the Politics article: \n {f
```

```
These are the 20 most commons words in the Business article:
 [(',', 21), ('.', 20), ('to', 18), ('the', 17), ('of', 12), ('and', 10),
('$', 9), ('%', 8), ('a', 8), ('TimeWarner', 7), ('in', 7), ('AOL', 7),
('by', 6), ('it', 6), ('profit', 5), ('profits', 5), ('for', 5), ('which
', 5), ('is', 5), ('said', 5)]
These are the 20 most commons words in the Entertainment article:
 [('the', 16), ('.', 13), ('to', 9), (',', 9), ('of', 8), ('a', 7), ('and
', 7), ('Andersen', 6), ('at', 5), ('Christian', 5), ("'s", 5), ('The',
5), ('is', 4), ('Hans', 4), ('``', 4), ("''", 4), ('Jarre', 3), ('in',
3), ('are', 3), ('Bloom', 3)]
These are the 20 most commons words in the Politics article:
 [('the', 27), ('.', 26), (',', 19), ('and', 14), ('``', 14), ('to', 13),
('of', 13), ('for', 12), ('women', 12), ('is', 11), ('in', 9), ("''", 9),
('a', 8), ('that', 8), ('six', 6), ('was', 6), ('are', 6), ('about', 6),
('Hewitt', 5), ('sexism', 5)]
```

In [72]:

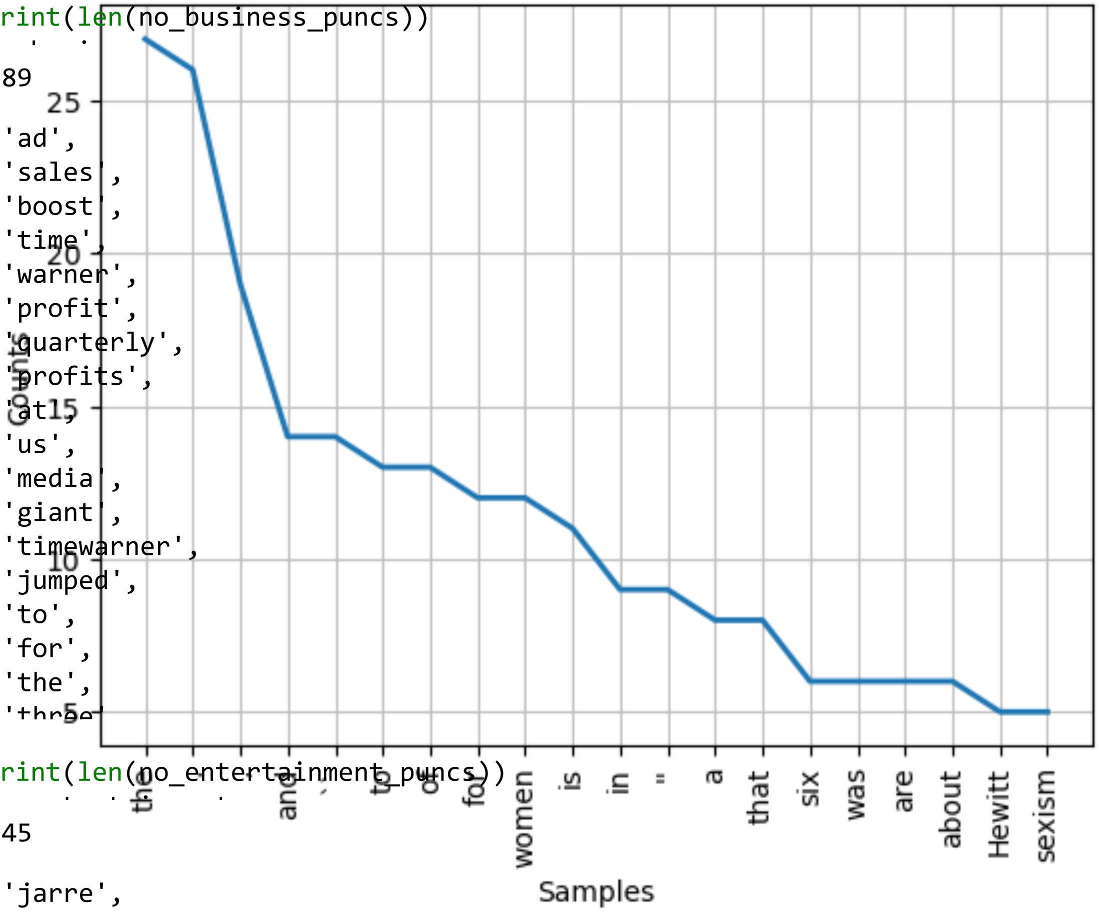
```
# plotting them out
fdist1.plot(20)
fdist2.plot(20)
fdist3.plot(20)
```





In [77]:

```
# taking out punctuations
no_business_puncs = []
no_entertainment_puncs = []
no_politics_puncs = []

for w in words1:
    if w.isalpha():
        no_business_puncs.append(w.lower())

for w in words2:
    if w.isalpha():
        no_entertainment_puncs.append(w.lower())

for w in words3:
    if w.isalpha():
        no_politics_puncs.append(w.lower())
```



In [81]:
```
print(len(no_business_puncs))
```

```
389
```

Out[81]:
```
['ad',
 'sales',
 'boost',
 'time',
 'warner',
 'profit',
 'quarterly',
 'profits',
 'at',
 'us',
 'media',
 'giant',
 'timewarner',
 'jumped',
 'to',
 'for',
 'the',
 'thre
```

In [82]:
```
print(len(no_entertainment_puncs))
```

```
245
```

Out[82]:
```
['jarre',
 'joins',
 'fairytale',
 'celebration',
 'french',
 'musician',
 'jarre',
 'is',
 'to',
 'perform',
 'at',
 'a',
 'concert',
 'in',
 'copenhagen',
 'to',
 'mark',
 'the'
```

Out[72]:
```
<Axes: xlabel='Samples', ylabel='Counts'>
```

In [83]:
```
print(len(no_politics_puncs))
```

```
506
```

Out[83]:
```
['hewitt',
 'decries',
 'plans',
 'to',
 'extend',
 'paid',
 'maternity',
 'leave',
 'beyond',
 'six',
 'months',
 'should',
 'be',
 'prominent',
 'in',
 'labour',
 'election',
 'manifesto'
```

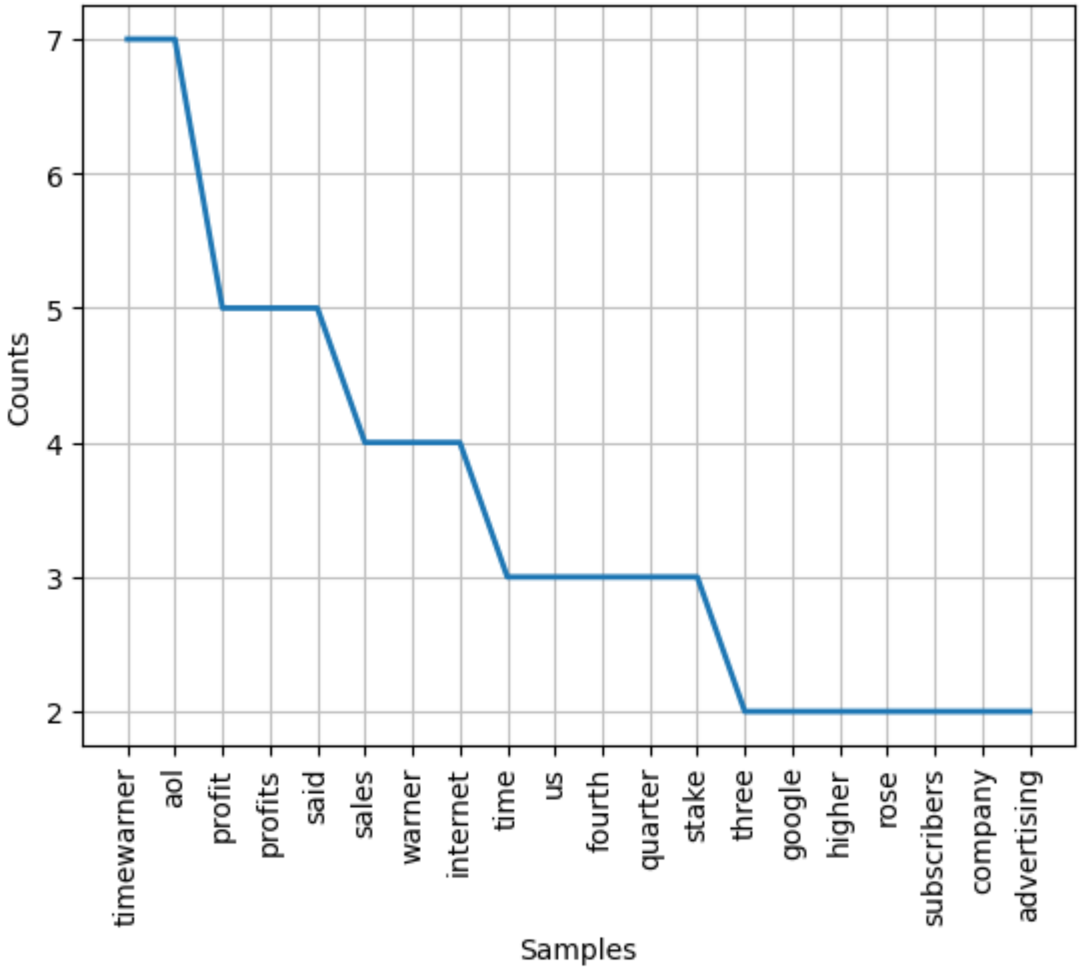
In [100]:
```
# [illegible]
```

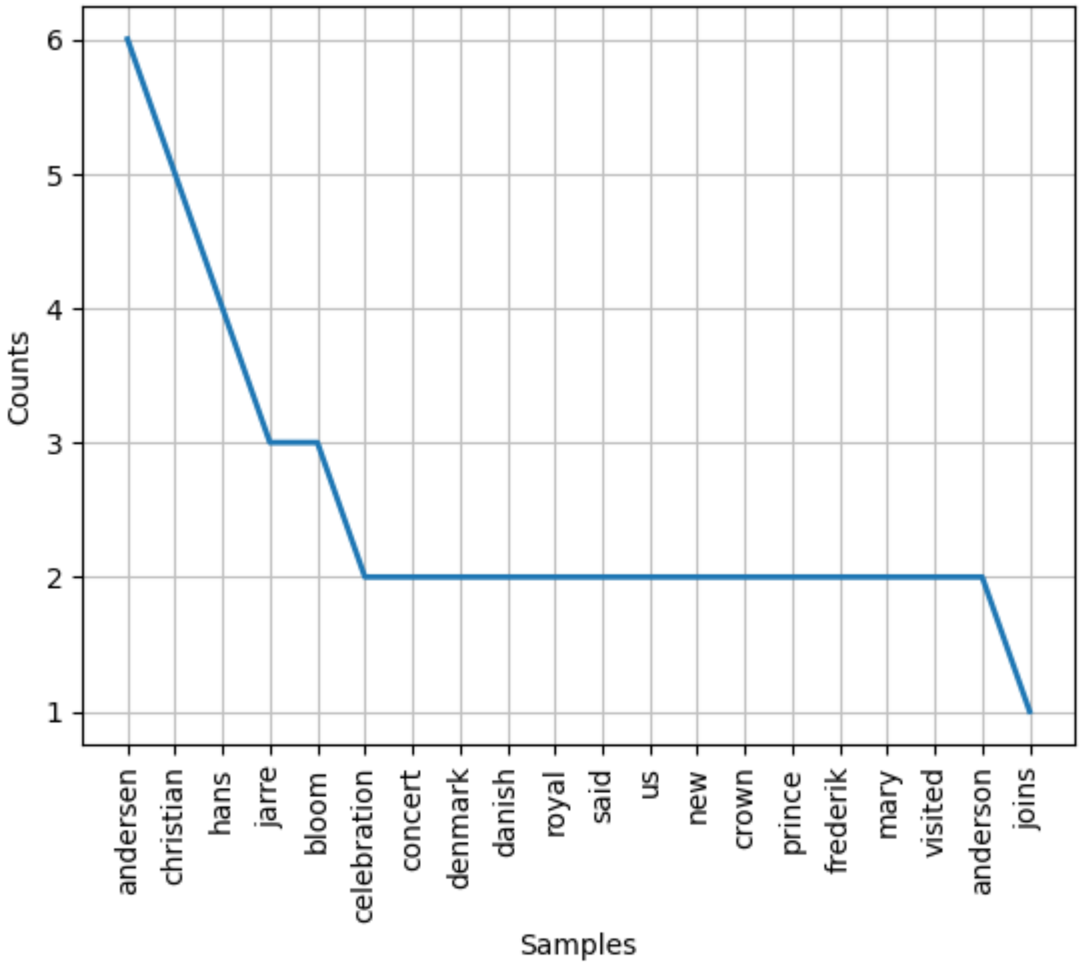
In [101]:
```
clean1 = []
clean2 = []
clean3 = []

for p in no_business_puncs:
    if p not in stopwords:
        clean1.append(p)

for p in no_entertainment_puncs:
    if p not in stopwords:
        clean2.append(p)

for p in no_politics_puncs:
    if p not in stopwords:
        clean3.append(p)
```

In [102]:
```
# time to plot them again to see how much has changed
fdist1 = FreqDist(clean1)
fdist2 = FreqDist(clean2)
fdist3 = FreqDist(clean3)
```

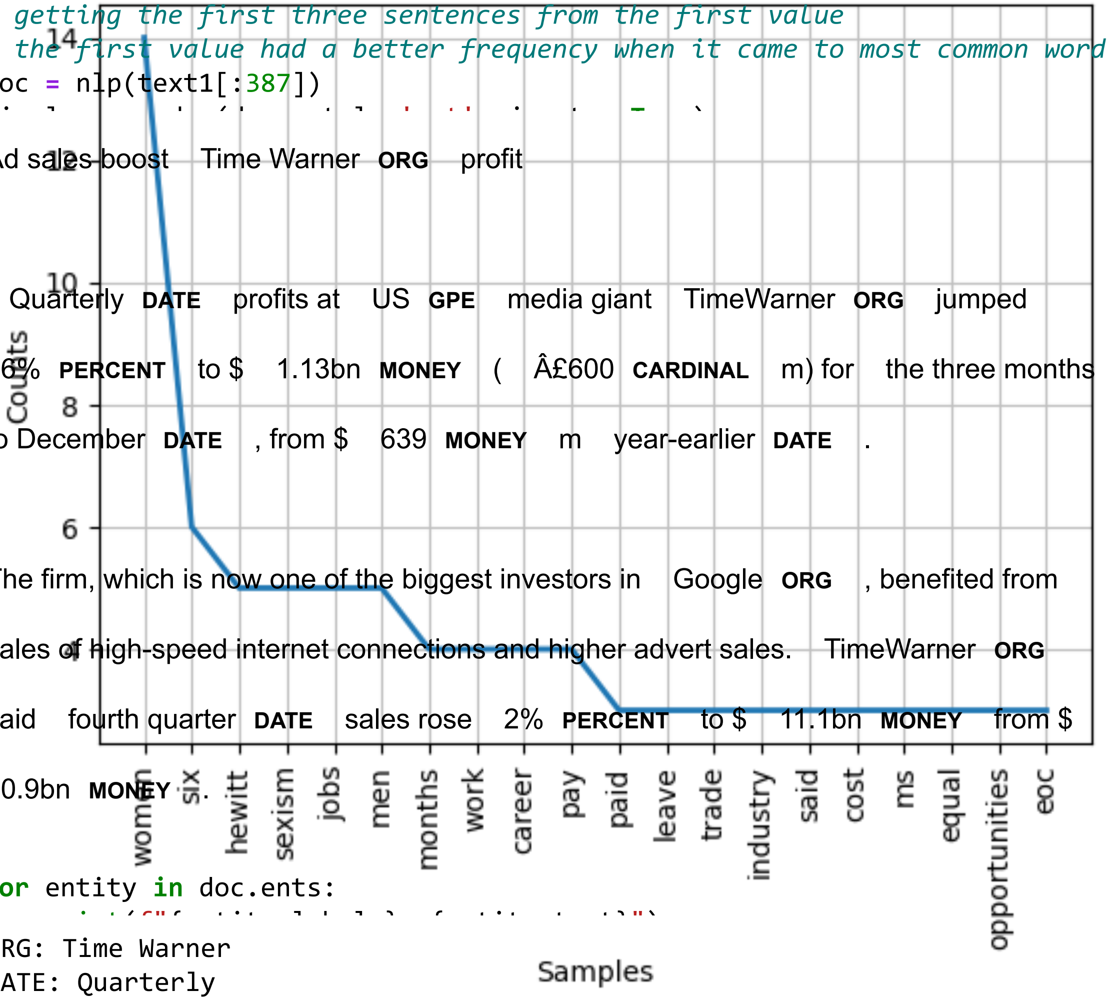
In [103]:

```
fdist1.plot(20)
fdist2.plot(20)
fdist3.plot(20)
```




6
5
4
3
2
1
Counts
andersen
christian
hans
jarre
bloom
celebration
concert
denmark
danish
royal
said
us
new
crown
prince
frederik
mary
visited
anderson
joins
Samples

In [106]:



Out[103]:
```
<AxesSubplot:xlabel='Samples', ylabel='Counts'>
```

In [119]:
```python
# getting the first three sentences from the first value
# the first value had a better frequency when it came to most common words
doc = nlp(text1[:387])
```

Ad sales boost Time Warner **ORG** profit

Quarterly **DATE** profits at US **GPE** media giant TimeWarner **ORG** jumped 76% **PERCENT** to $ 1.13bn **MONEY** (Â£600 **CARDINAL** m) for the three months to December **DATE** , from $ 639 **MONEY** m year-earlier **DATE** .

The firm, which is now one of the biggest investors in Google **ORG** , benefited from sales of high-speed internet connections and higher advert sales. TimeWarner **ORG** said fourth quarter **DATE** sales rose 2% **PERCENT** to $ 11.1bn **MONEY** from $ 10.9bn **MONEY** .

In [121]:
```python
for entity in doc.ents:
```

```
ORG: Time Warner
DATE: Quarterly
GPE: US
ORG: TimeWarner
PERCENT: 76%
MONEY: 1.13bn
CARDINAL: Â£600
DATE: the three months to December
MONEY: 639
DATE: year-earlier
ORG: Google
ORG: TimeWarner
DATE: fourth quarter
PERCENT: 2%
MONEY: 11.1bn
MONEY: 10.9bn
```

In [129]:

```
color = {'ORG': 'orange', 'DATE': 'light', 'PERCENT': 'aqua', 'MONEY': 'gr
options = {'ents': ['ORG', 'DATE', 'PERCENT', 'MONEY', 'CARDINAL'], 'color
```

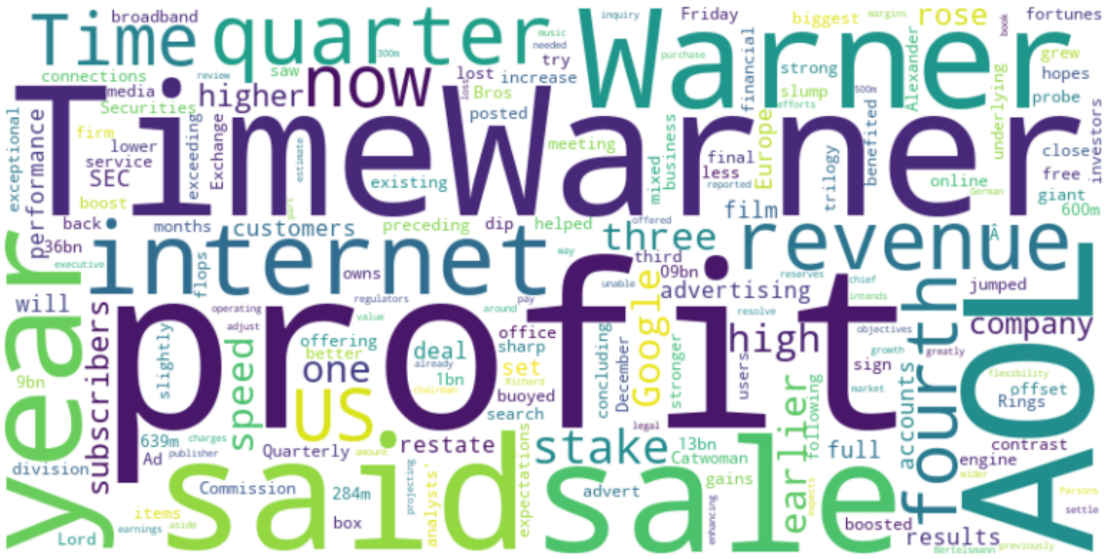
Ad sales boost Time Warner **ORG** profit

Quarterly **DATE** profits at US media giant TimeWarner **ORG** jumped 76% **PERCENT** to $ 1.13bn **MONEY** (Â£600 **CARDINAL** m) for the three months to December **DATE** , from $ 639 **MONEY** m year-earlier **DATE** .

The firm, which is now one of the biggest investors in Google **ORG** , benefited from sales of high-speed internet connections and higher advert sales. TimeWarner **ORG** said fourth quarter **DATE** sales rose 2% **PERCENT** to $ 11.1bn **MONEY** from $ 10.9bn **MONEY** .

In [130]:

```
# Generate a word cloud
wordcloud = WordCloud(width=800, height=400, background_color="white").gen

# Display the word cloud using matplotlib
plt.figure(figsize=(10, 5))
plt.imshow(wordcloud, interpolation="bilinear")
plt.axis("off")
```



While these might've been bad examples, I was hoping to see advertisements, BBC affliates, and/or BBC's name itself to appear quite often between the three articles. Maybe it would've been better to have web scraped BBC's website to find these parameters that I am looking for.

In [ ]: