wordcloud 이용하여 단어빈도 분석하기

통계적 처리

gutenberg 내 corpus들의 평균 단어 길이, 평균 문장 길이, 그리고 각 어휘 항목이 텍스트에 평균적으로 나타나는 횟수(어휘 다양성 점수)를 먼저 알아보겠습니다.

```
평균단어길이:4, 평균 문장 길이 : 24,
각 어휘 항목이 텍스트에 평균적으로 나타나는 횟수 : 26, austen-emma.txt

평균단어길이:4, 평균 문장 길이 : 26,
각 어휘 항목이 텍스트에 평균적으로 나타나는 횟수 : 16, austen-persuasion.txt

평균단어길이:4, 평균 문장 길이 : 28,
각 어휘 항목이 텍스트에 평균적으로 나타나는 횟수 : 22, austen-sense.txt

평균단어길이:4, 평균 문장 길이 : 33,
각 어휘 항목이 텍스트에 평균적으로 나타나는 횟수 : 79, bible-kjv.txt

평균단어길이:4, 평균 문장 길이 : 19,
각 어휘 항목이 텍스트에 평균적으로 나타나는 횟수 : 5, blake-poems.txt

평균단어길이:4, 평균 문장 길이 : 19,
각 어휘 항목이 텍스트에 평균적으로 나타나는 횟수 : 14, bryant-stories.txt

평균단어길이:4, 평균 문장 길이 : 17,
각 어휘 항목이 텍스트에 평균적으로 나타나는 횟수 : 12, burgess-busterbrown.txt

평균단어길이:4, 평균 문장 길이 : 20,
각 어휘 항목이 텍스트에 평균적으로 나타나는 횟수 : 12, carroll-alice.txt

평균단어길이:4, 평균 문장 길이 : 20,
```

```
각 어휘 항목이 텍스트에 평균적으로 나타나는 횟수 : 11, chesterton-ball.txt

평균단어길이:4, 평균 문장 길이 : 22,
각 어휘 항목이 텍스트에 평균적으로 나타나는 횟수 : 11, chesterton-brown.txt

평균단어길이:4, 평균 문장 길이 : 18,
각 어휘 항목이 텍스트에 평균적으로 나타나는 횟수 : 10, chesterton-thursday.txt

평균단어길이:4, 평균 문장 길이 : 20,
각 어휘 항목이 텍스트에 평균적으로 나타나는 횟수 : 24, edgeworth-parents.txt

평균단어길이:4, 평균 문장 길이 : 25,
각 어휘 항목이 텍스트에 평균적으로 나타나는 횟수 : 15, melville-moby_dick.txt

평균단어길이:4, 평균 문장 길이 : 52,
각 어휘 항목이 텍스트에 평균적으로 나타나는 횟수 : 10, milton-paradise.txt

평균단어길이:4, 평균 문장 길이 : 11,
각 어휘 항목이 텍스트에 평균적으로 나타나는 횟수 : 8, shakespeare-caesar.txt

평균단어길이:4, 평균 문장 길이 : 12,
각 어휘 항목이 텍스트에 평균적으로 나타나는 횟수 : 7, shakespeare-hamlet.txt

평균단어길이:4, 평균 문장 길이 : 12,
각 어휘 항목이 텍스트에 평균적으로 나타나는 횟수 : 6, shakespeare-macbeth.txt

평균단어길이:4, 평균 문장 길이 : 36,
각 어휘 항목이 텍스트에 평균적으로 나타나는 횟수 : 12, whitman-leaves.txt
```

제가 중점적으로 분석할 austen, cheserton, sakespeare 세 작가의 작품은

austen : austen-emma.txt

cheserton : chesterton-brown.txt

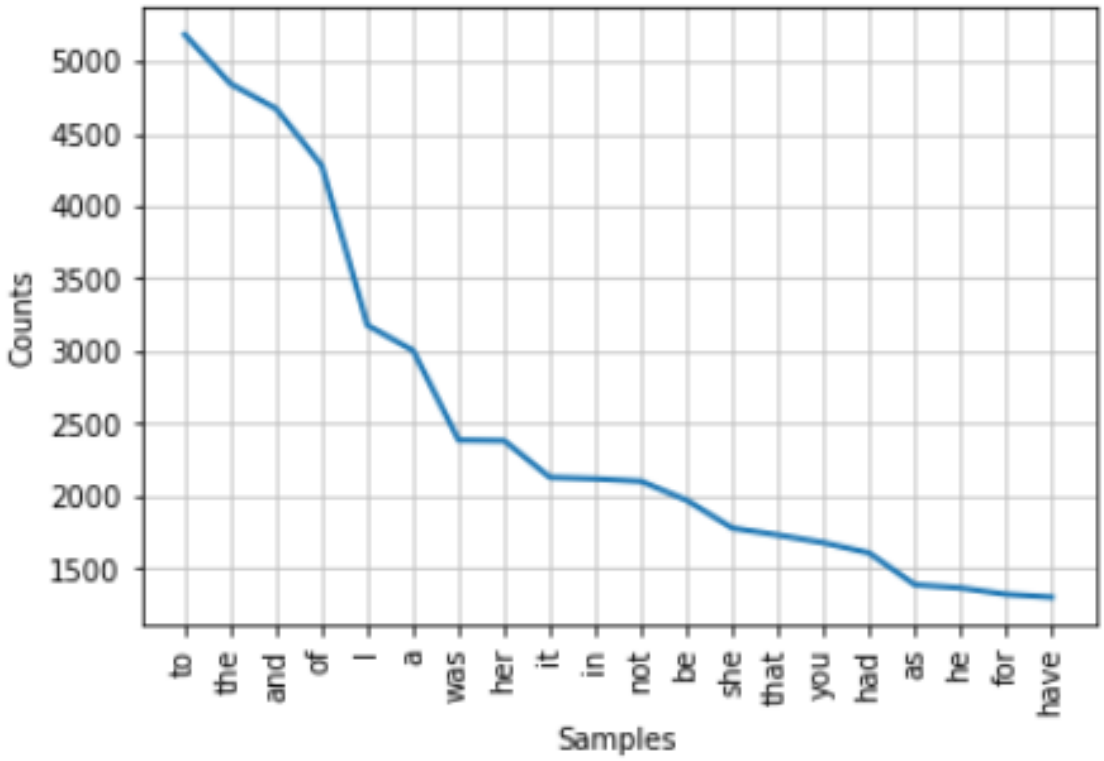
sakespeare : shakespeare-hamlet.txt

이렇게 3가지를 중심으로 해서 분석할 예정입니다.

소설 내 단어 출현 빈도

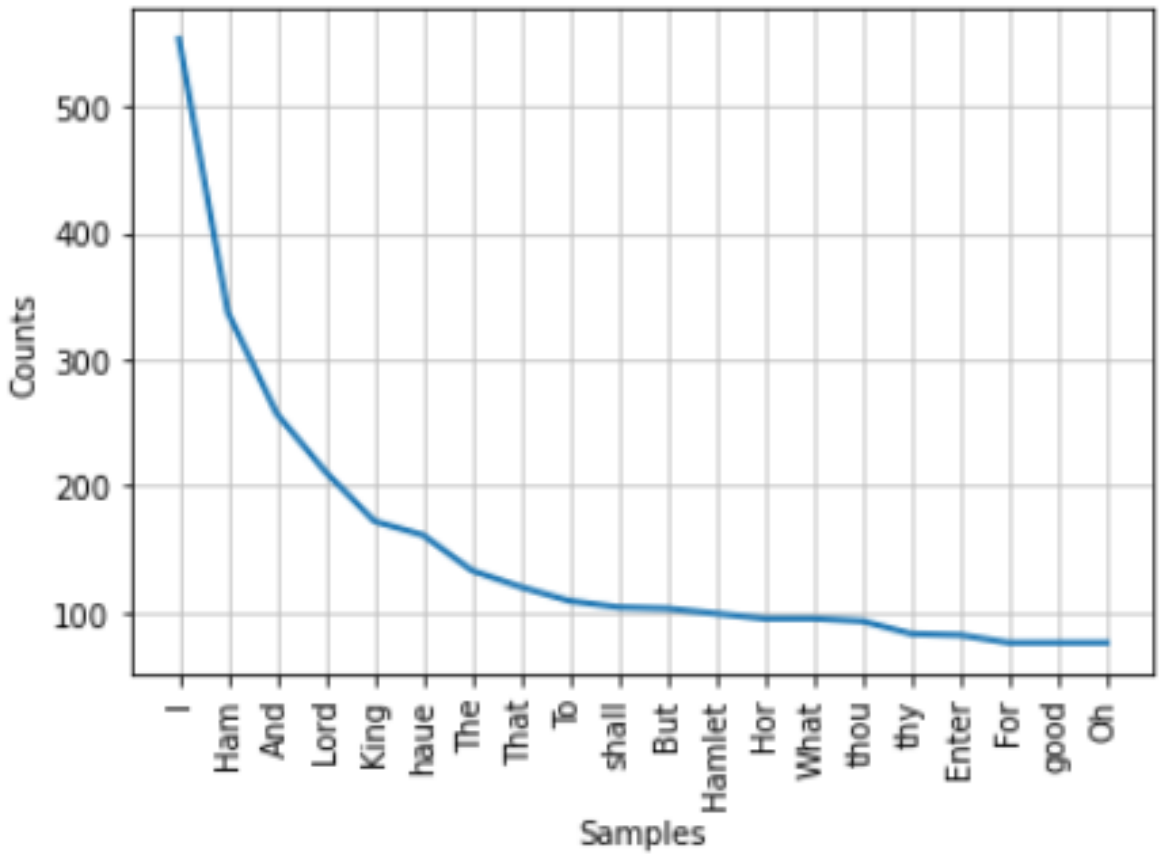
- 각 코퍼스를 단어 수준으로 tokenizing 후 단어 별 count를 시각화해보겠습니다.

1. **Emma**



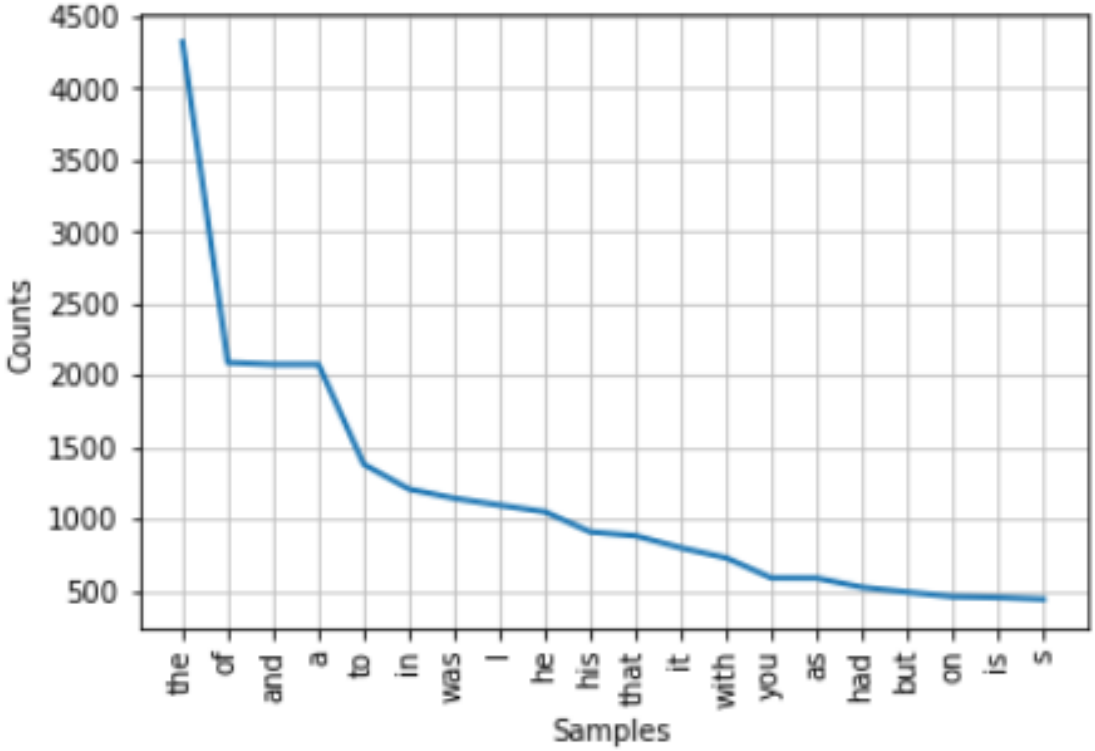
Emma의 경우 여성 중심의 소설이라 her이나 she의 빈도가 높은 것이 눈이 띕니다

2. **shakespeare - Hamlet**



Hamlet의 경우 주인공인 hamlet을 비롯해 Lord나 King의 빈도가 높은 것이 눈에 띕니다. 이는 햄릿의 배경이 작중 12세기 덴마크의 왕가인 것에서 비롯된 것으로 보입니다.

3. **chesterton**

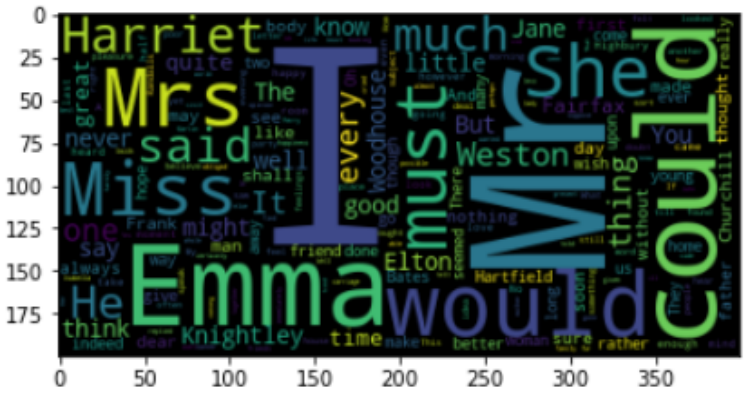


체스터턴의 브라운 신부 같은 경우 브라운 신부가 남자여서 he가 많이 언급되는 것 외에는 대부분이 조사나 지칭대명사라 유의미한 분석은 하기 어렵습니다.

- 이후 불용어(stopwords)를 제거한 후 단어의 빈도수를 시각화해보겠습니다 (데이터프레임 및 워드 클라우드 시각화). 위의 워드 클라우드는 불용어를 제거한 후 빈도수 중심의 워드 클라우드이고 아래의 워드 클라우드는 불용어를 제거한 후 토큰을 다시 문장으로 합친 후 워드클라우드 시킨 것입니다.

	word	freq
0	I	3178
1	Mr	1153
2	Emma	865
3	could	825
4	would	815
5	Mrs	699
6	Miss	592
7	must	564
8	She	562
9	Harriet	506
10	said	484
11	much	478
12	He	441
13	Weston	439
14	one	413
15	It	400
16	every	398
17	thing	398
18	Knightley	389
19	Elton	385

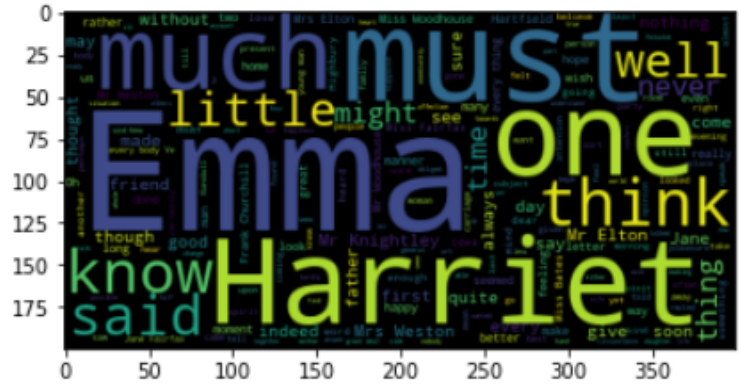
Emma



위의 워드클라우드는 불용어를 제거한 후 빈도수로 워드클라우드한 그림입니다. 불용어를 제거한 emma 소설의 가장 많이 등장하는 단어는 역시 Emma이고 mrs와 miss입니다. 또한 1인칭 대명사인 I가 많이 등장하는데 이는 주인공인 Emma 관점으로 소설이 진행되기 때문으로 보입니다.

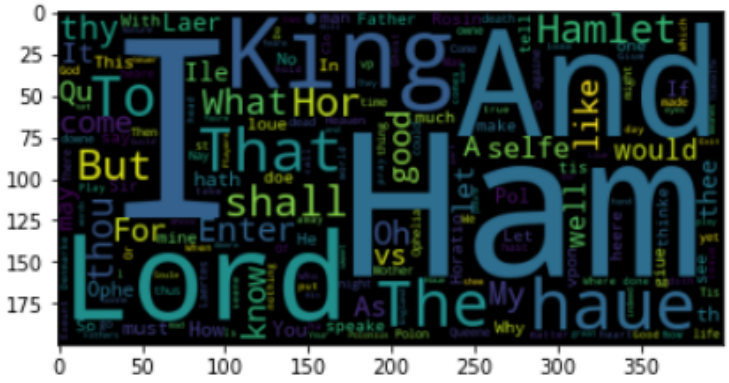
아래 그림은 불용어를 제거한 토큰을 문장으로 다시 합쳐 워드클라우드 시킨 그림입니다. 우리는 1인칭 주인공이 Emma이고 그 주변 인물이 Harriet이라는 것을 유추해볼 수 있습니다.

또한 과거형의 가능/의지 동사인 could와 would가 많이 등장하는 것으로 보아 주인공이 과거 회상 혹은 행동을 후회하거나 복기하는 장면이 많았을 것으로 유추해볼 수 있습니다.



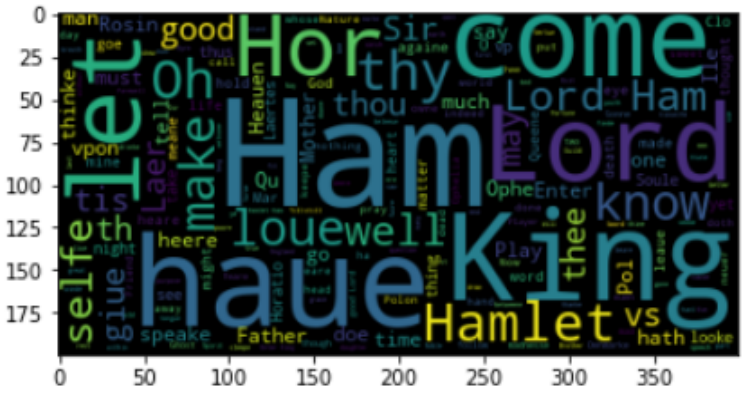
	word	freq
0	I	553
1	Ham	337
2	And	257
3	Lord	211
4	King	172
5	haue	161
6	The	133
7	That	120
8	To	109
9	shall	104
10	But	103
11	Hamlet	99
12	Hor	95
13	What	95
14	thou	93
15	thy	83
16	Enter	82
17	For	76
18	good	76
19	Oh	76

Hamlet



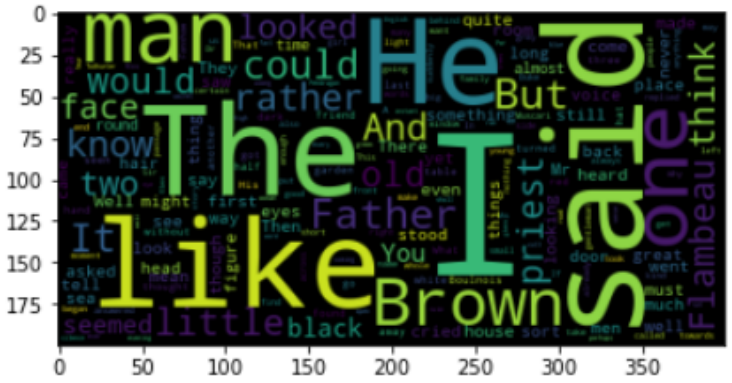
불용어를 제거한 후 가장 많이 등장한 단어는 Ham과 I 입니다. 이로 유추해볼 수 있는 것은 작 중 hamlet을 ham으로 줄여 부르는 경우가 많지 않았나 생각해볼 수 있습니다.

Hamlet의 경우 역접인 but 그리고 what과 같은 의문형 단어가 많이 등장하는 것으로 보아 소설 내 반전과 작 중 주인공의 심적 갈등이 많았음을 유추해볼 수 있을 것 같습니다.

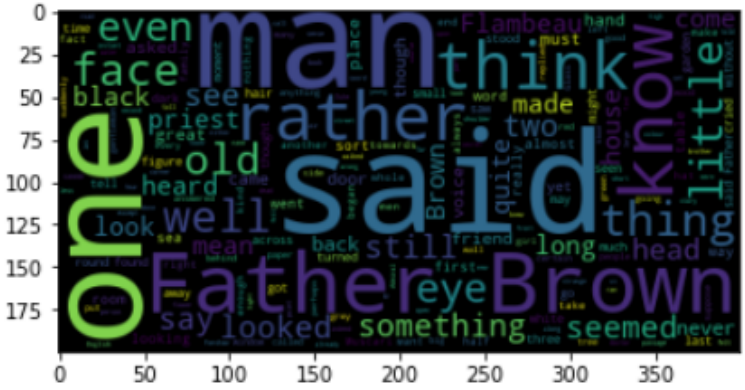


	word	freq
0	I	1093
1	said	415
2	The	346
3	like	328
4	He	310
5	man	303
6	one	264
7	Brown	261
8	Father	205
9	But	205
10	It	186
11	could	170
12	know	151
13	And	147
14	little	146
15	rather	142
16	priest	134
17	would	132
18	think	131
19	Flambeau	129

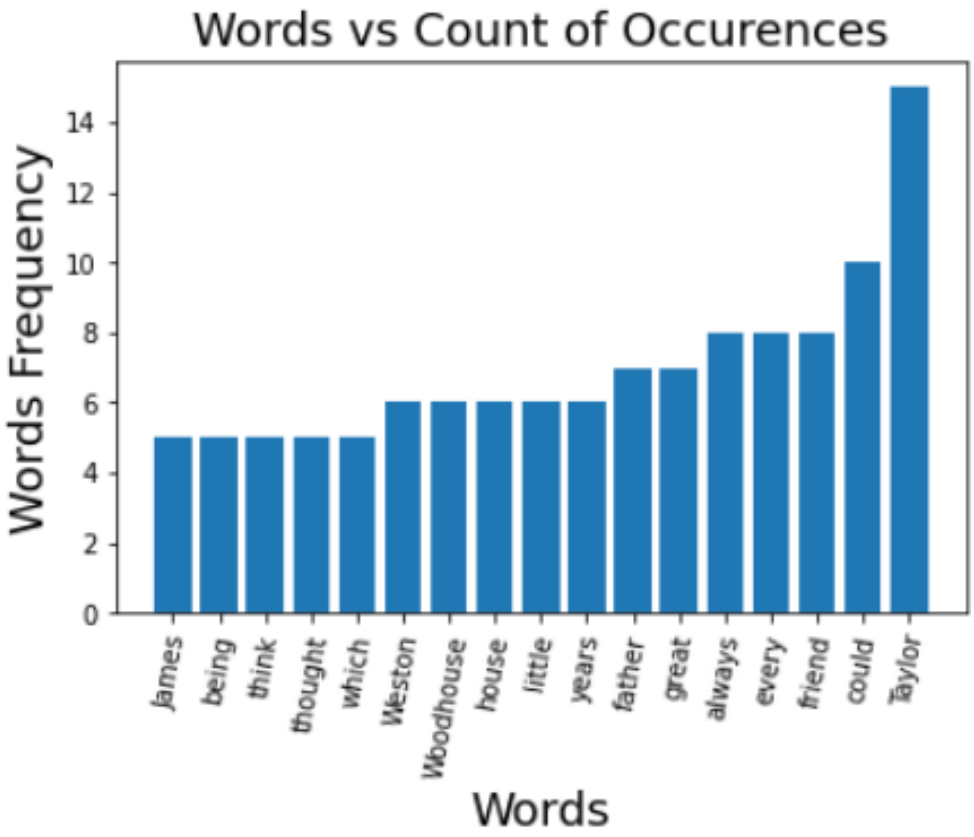
Father Borwn - **Chesterton**



Father Brown은 Brown 신부가 주인공으로 등장하는 추리소설입니다. 그에 따라 I 혹은 said 그리고 like과 같은 단어가 많이 등장하는데 이는 주인공이 추리를 함에 따라 등장하는 단어들일 것이라 추측됩니다. 또한 주인공의 신분에 따라 priest 그리고 Father과 같은 단어도 빈도가 높게 나옵니다.

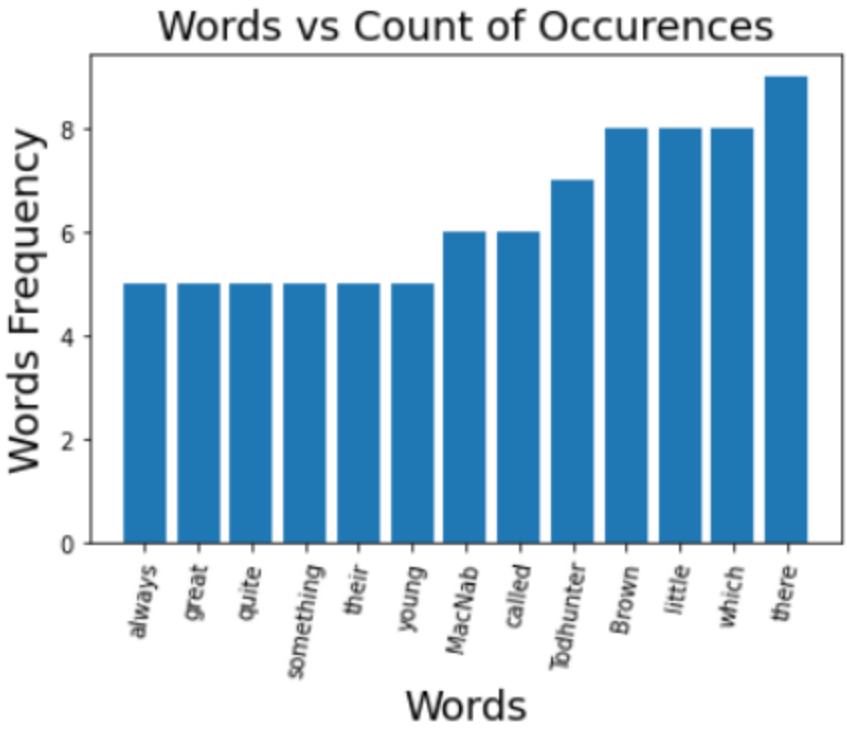


- 이번에는 **Emma와 인칭 대명사를 제외한 4글자 이상의 단어들의 빈도 수를 뽑아 시각화해 보겠습니다. 이를 통해 주인공 외 인물이나 다른 단어의 빈도 수를 확인해 볼 예정입니다.**



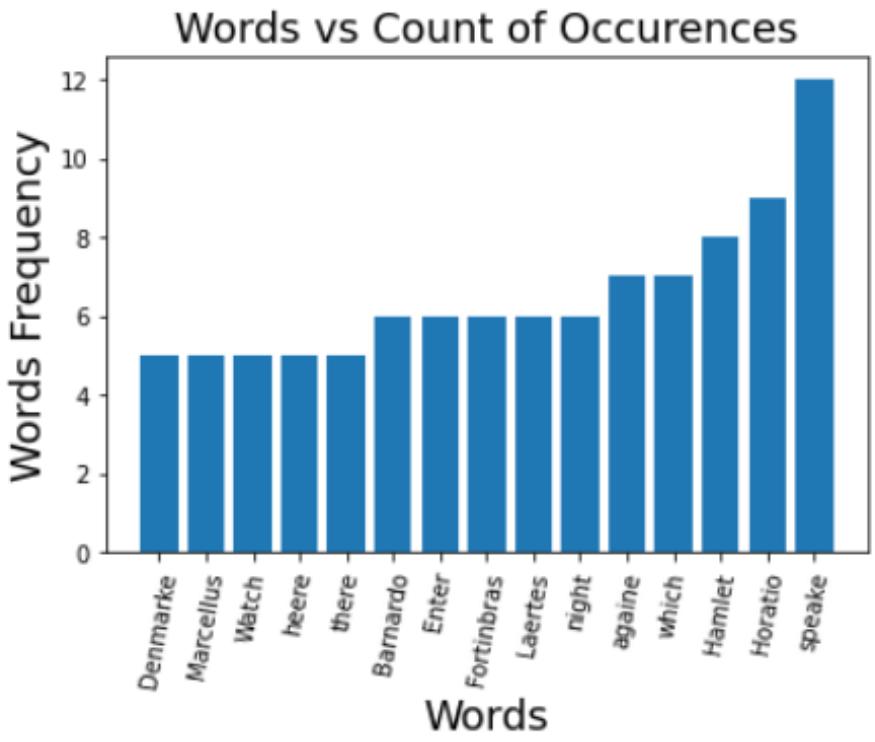
Emma

Taylor라는 단어가 많이 나온 것으로 보아 이 인물이 emma 다음으로 중요한 인물인 것을 유추해볼 수 있습니다. 또한 friend가 많이 나오는데 이는 작 중 emma가 친구를 주선해주는 일을 많이 해서 이로 인해 많이 등장한 것으로 유추해볼 수 있습니다.



Chesterton- Brown

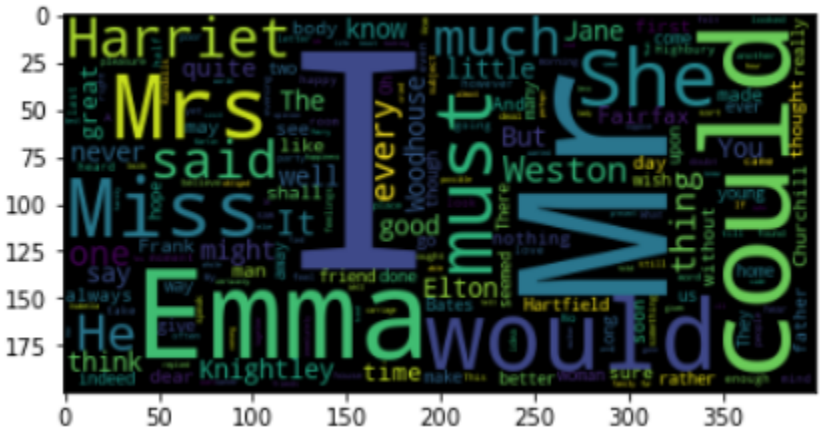
There과 which 그리고 Brown이 많이 등장하는 것으로 보아 추리하는 장면에 이러한 단어가 많이 쓰인 것으로 유추해볼 수 있습니다. 또한 눈에 띄는 단어(명사)로는 MacNab 그리고 TodHunter이 있습니다. MacNab이라는 단체와 TodHunter이라는 surname을 가진 인물이 빈번히 등장한다는 것을 알 수 있습니다.



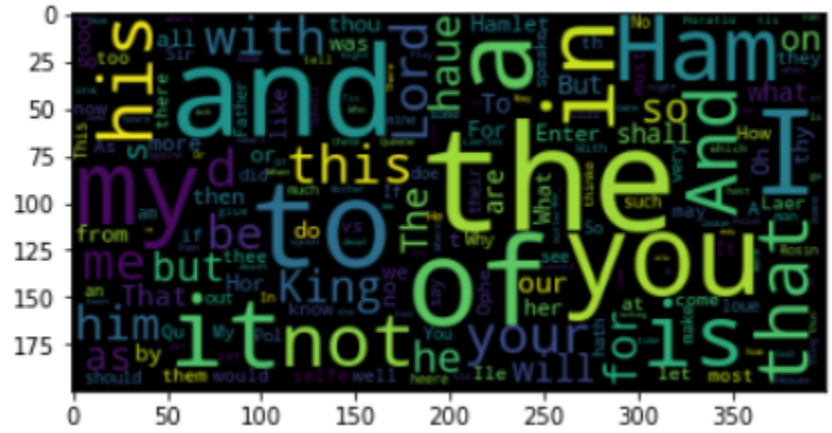
Hamlet

speake라는 단어가 제일 많이 등장하는 데 이는 speak의 구식 표현이다. 이를 통해 hamlet이 고전 소설임을 알 수 있다. 또한 주인공인 hamlet보다 horatio가 더 많이 등장하는 데 이 인물 또한 대단히 중요한 인물임을 유추해볼 수 있다. 또한 night의 등장 빈도로 보아 해당 시간대가 작중 중요한 시간대임을 유추해 볼 수 있습니다.

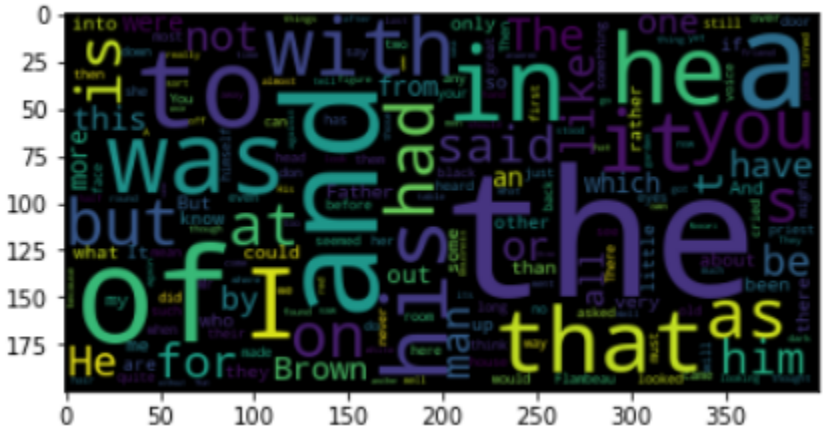
- **글자 빈도수에 따른 워드 클라우드**



Emma



Shakespeare-Hamlet

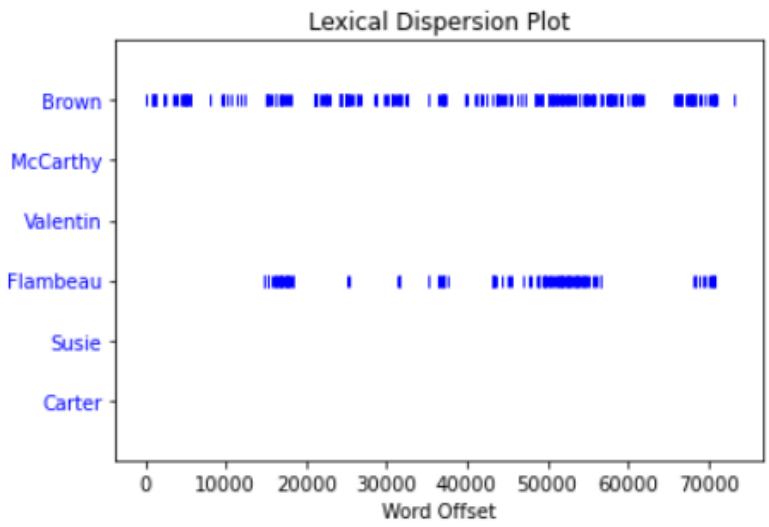


chesterton-Brown

소설 내 등장인물 출현 빈도

Emma - 제인 오스틴

Hamlet - 셰익스피어

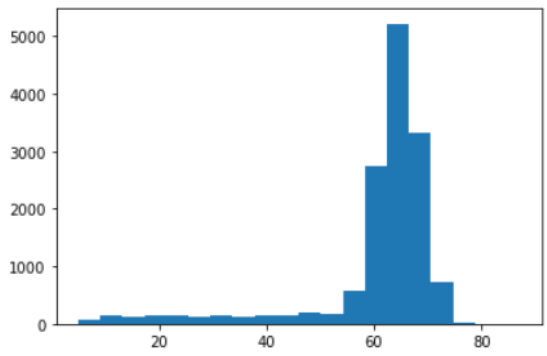


Father Brown - **Chesterton**

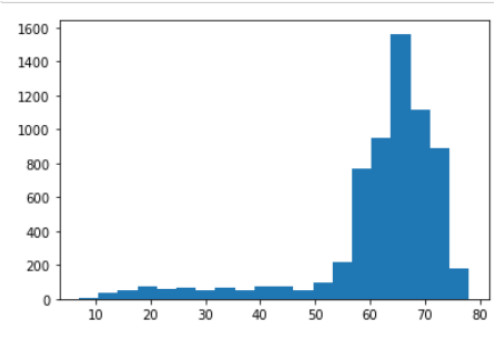
- Hamlet에 비해 Emma와 Brown은 주인공이 더 많이 등장하는 것으로 보아 작 중 시점이 주인공인 emma와 brown **1인칭**으로 전개되는 것으로 보입니다. 반대로 Hamlet은 hamlet의 등장이 주인공이라 많기는 하지만 위의 두 작품만큼 많지 않습니다. 이 둘의 차이는 **작중 시점의 차이**인 것으로 유추할 수 있습니다. (전지적 작가 시점과 1인칭 주인공 사이의 시점 차이인 것으로 유추됩니다.)

문장 단위 분석

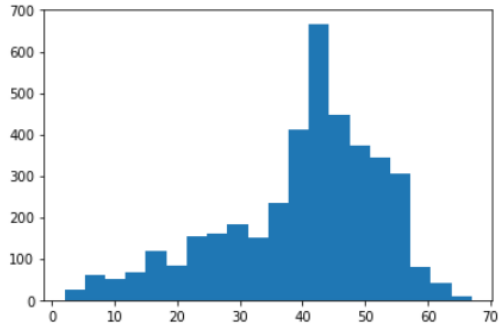
문장 길이 별 빈도



Emma



chesterton



Hamlet

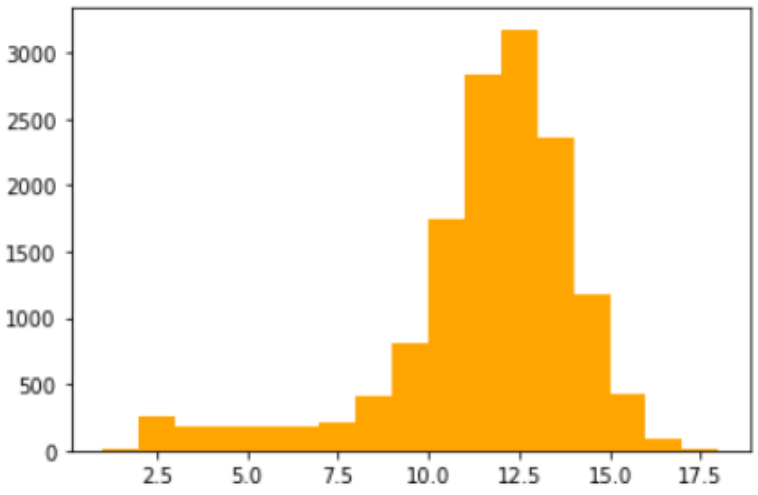
Emma와 chesterton의 소설의 문장 길이 빈도가 유사한 것하고 한쪽으로 skewed된 것에 비해 Hamlet은 그보다 덜 skewed 되어 있고 고르게 분포되어 있는 것을 알 수 있습니다.

skewed 정도 : Emma > chesterton > Hamlet

문장 당 단어의 개수/ 그에 따른 빈도

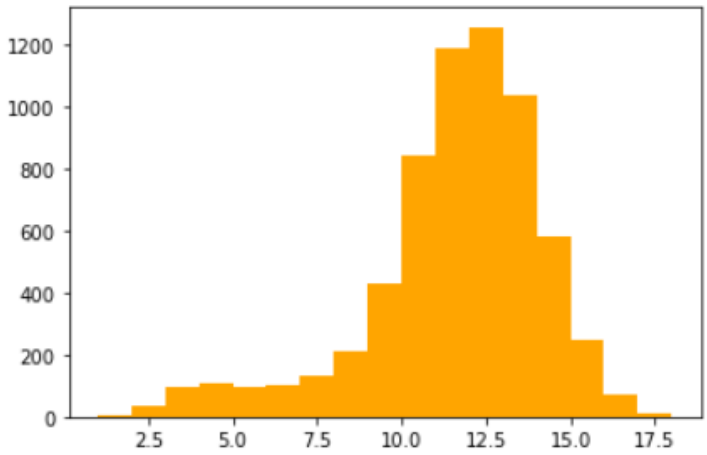
- **Emma**

	number
count	14283.000000
mean	11.073794
std	2.626604
min	1.000000
25%	10.000000
50%	12.000000
75%	13.000000
max	17.000000



- **Chesterton**

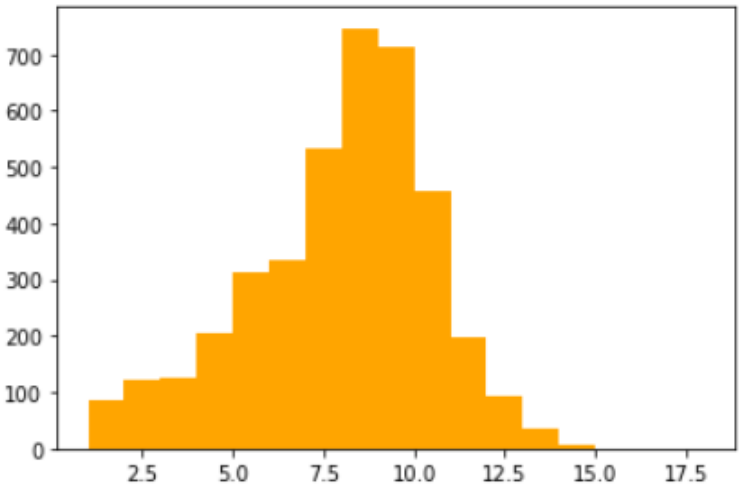
	number
count	6443.000000
mean	11.116871
std	2.606829
min	1.000000
25%	10.000000
50%	11.000000
75%	13.000000
max	18.000000



Emma와 Chesterton의 문장 내 단어의 수 분포는 유사하나 emma의 경우가 조금 더 빈도가 높음을 알 수 있습니다. 빈도 차이는 이는 count의 차이에서 나온 것임을 알 수 있습니다. 이는 해당 분포의 mean과 standard dev가 유사한 것을 통해서도 알 수 있습니다.

- **Hamlet**

	number
count	3966.000000
mean	7.464700
std	2.527756
min	1.000000
25%	6.000000
50%	8.000000
75%	9.000000
max	14.000000



hamlet의 경우 문장 길이당 단어의 개수 분포가 조금 더 왼쪽으로 치우쳐짐을 알 수 있습니다. 이는 mean의 차이를 통해서도 알 수 있습니다. (7.46 < 11)