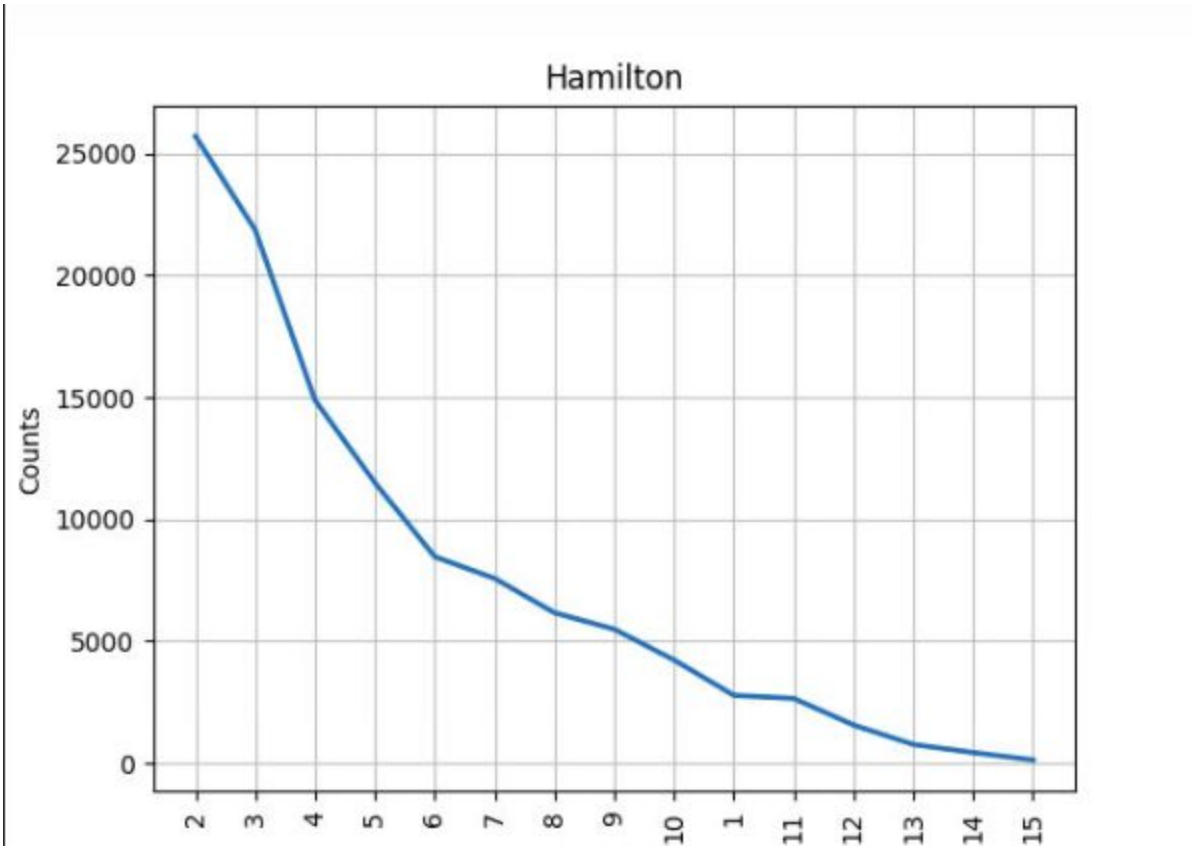
Natural Language Processing

```
import nltk
import matplotlib

papers = {'Madison':
[10,14,37,38,40,41,42,43,44,45,46,47,48],'Hamilton':[1,6,7,8,9,11,12,13,15,16,17,21,22,2
3,24,25,26,27,28,29,30,31,32,33,34,35,36,59,60,61,65,66,67,68,69,70,71,72,73,74,75,76,77
,78,79,80,81,82,83,84,85],'Jay': [2,3,4,5],'Shared': [18,19,20],'Disputed':
[49,50,51,52,53,54,55,56,57,58,62,63]}

def read_files(filename):
   strings = []
   for file in filename:
       with open(f'federalist_{file}.txt') as f:
           strings.append(f.read())
   return ('\n'.join(strings))

federalist_by_author = {}

for author,files in papers.items():
   federalist_by_author[author] = read_files(files)

authors = ('Hamilton', 'Madison','Disputed', 'Jay','Shared')

author_tokens = {}
length_distribution = {}

for author in authors:
   tokens = nltk.word_tokenize(federalist_by_author[author])

   author_tokens[author] = ([token for token in tokens if any(c.isalpha() for c in
token)])

   token_lengths = [len(token) for token in author_tokens[author]]

   length_distribution[authors] = nltk.FreqDist(token_lengths)

   length_distribution[authors].plot(15,title=author)
```

Output :


Hamilton
Counts
25000
20000
15000
10000
5000
0
2
3
4
5
6
7
8
9
10
1
11
12
13
14
15