Distributions of Words & Sentences

This assignment is comprised of three tasks for ITCS 4111 and an additional task for ITCS 5111:

1. The first task is to compute the frequency vs. rank distribution of the words in Moby Dick. For this, you will need to tokenize the document and create a vocabulary mapping word types to their document frequency.
2. The second task is to segment the document into sentences and compute the sentence length distribution. Here you will experiment with spaCy's default sentence segmenter as well as the simple rule-based Sentencizer.
3. The third task is the same as the first except that we use subword tokenization.
4. Use spacy's NE recognizer to find all named entities in the first 2,500 paragraphs. Count how many times they appear in the document and consolidate them based on their most frequent type.

Manish Kumar Govind:

Submission Instructions

1. Click the Save button at the top of the Jupyter Notebook.
2. Please make sure to have entered your name above.
3. Select Cell -> All Output -> Clear. This will clear all the outputs from all cells (but will keep the content of ll cells).
4. Select Cell -> Run All. This will run all the cells in order, and will take several minutes.
5. Once you've rerun everything, select File -> Download as -> PDF via LaTeX and download a PDF version showing the code and the output of all cells, and save it in the same folder that contains the notebook file.
6. Look at the PDF file and make sure all your solutions are there, displayed correctly. The PDF is the only thing we will see when grading!
7. Submit **both** your PDF and notebook on Canvas.
8. Make sure your your Canvas submission contains the correct files by downloading it after posting it on Canvas.

Word distributions using the SpaCy tokenizer (40 + 10 points)

First, create the spaCy tokenizer.

In [203…
```
from spacy.lang.en import English
nlp = English()

tokenizer = nlp.tokenizer
```

Create a *vocab* dictionary. This dictionary will map tokens to their counts in the input text file.

In [204…
```
vocab = {}
```

Read the input file line by line.

1. Tokenize each line.
2. For each token in the line that contains only letters, convert it to lower case and increment the corresponding count in the dictionary.
    - If the token does not exist in the dictionary yet, insert it with a count of 1. For example, the first time the token 'water' is encountered, the code should evaluate *vocab['water'] = 1*.

At the end of this code segment, *vocab* should map each word type to the number of times it appeared in the entire document. There should be 16830 word types and 214287 words in Moby Dick.

In [205…
```
with open('../data/melville-moby_dick.txt', 'r') as f:
    for line in f:
        tokens = tokenizer(line)

        for token in tokens:
            # Check if the token is a word (contains only letters)
            if token.is_alpha:
                # Convert the token to lowercase
                l_token = token.text.lower()

                # Check if the token is already in the dictionary
                if l_token in vocab:
                    # Increment the count in the dictionary
                    vocab[l_token] += 1
                else:
                    # Initialize the count to 1 if not present in the dictionary
                    vocab[l_token] = 1
print('There are', len(vocab), 'word types in Moby Dick.')
print('There are', sum(vocab.values()), 'words in Moby Dick.')
```

```
There are 16830 word types in Moby Dick.
There are 214287 words in Moby Dick.
```

Create a list *ranked* of tuples *(word, freq)* that contains all the words in the vocabulary *vocab* sorted by frequency. For example, if *vocab = {'duck':2, 'goose':5, 'turkey':3}*, then *ranked = [('goose', 5), ('turkey', 3), ('duck', 2)]*.

In [206…
```
sorted_vocab = dict(sorted(vocab.items(), key=lambda item: item[1], reverse=True))

#  top 10 words with the highest frequencies

ranked = list(sorted_vocab.items())
```

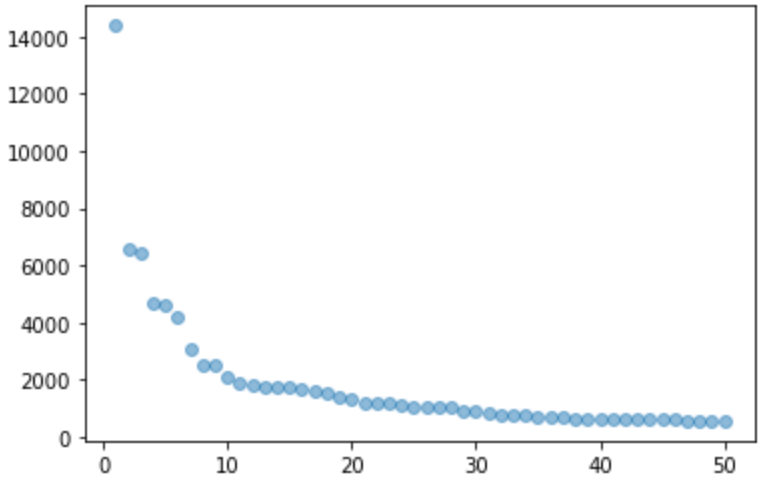
Print the top 10 words in the sorted list.

In [207…
```
print('Size of vocabulary:', len(ranked))
for word, freq in ranked[:10]:
    print(word, freq)
```

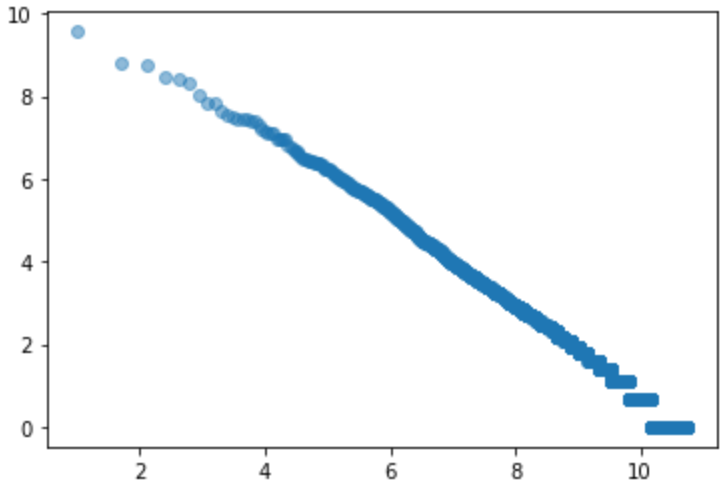
```
Size of vocabulary: 16830
the 14388
of 6606
and 6414
a 4698
to 4618
in 4164
that 3061
his 2527
it 2489
i 2068
```

Plot the frequency vs. rank of the top ranked words in Moby Dick.

In [208…
```
import matplotlib.pyplot as plt
ranks = range(1, 50 + 1)
freqs = [t[1] for t in ranked[:50]]
plt.scatter(ranks, freqs, c='#1f77b4', alpha=0.5)
plt.show()
```



In [209…
```
import math
ranks = [1 + math.log(r) for r in range(1, len(ranked) + 1)]
freqs = [math.log(t[1]) for t in ranked]
plt.scatter(ranks, freqs, c='#1f77b4', alpha=0.5)
plt.show()
```



Sentence distributions (40 + 10 points)

First, try to create the spaCy nlp object from the entire text of Moby Dick. This will likely not work, it is not a good idea to read all the text.

In [210…
```
import spacy

nlp = spacy.load("en_core_web_sm")
text = open('../data/melville-moby_dick.txt', 'r').read()
doc = nlp(text)
```

```
---------------------------------------------------------------------------
ValueError                                Traceback (most recent call last)
```

```
Input In [210], in <cell line: 5>()
      3 nlp = spacy.load("en_core_web_sm")
      4 text = open('../data/melville-moby_dick.txt', 'r').read()
----> 5 doc = nlp(text)

File ~\AppData\Roaming\Python\Python39\site-packages\spacy\language.py:1030, in Languag
e.__call__(self, text, disable, component_cfg)
   1009 def __call__(
   1010     self,
   1011     text: Union[str, Doc],
   (...)
   1014     component_cfg: Optional[Dict[str, Dict[str, Any]]] = None,
   1015 ) -> Doc:
   1016     """Apply the pipeline to some text. The text can span multiple sentences,
   1017     and can contain arbitrary whitespace. Alignment into the original string
   1018     is preserved.
   (...)
   1028     DOCS: https://spacy.io/api/language#call
   1029     """
-> 1030     doc = self._ensure_doc(text)
   1031     if component_cfg is None:
   1032         component_cfg = {}

File ~\AppData\Roaming\Python\Python39\site-packages\spacy\language.py:1121, in Languag
e._ensure_doc(self, doc_like)
   1119     return doc_like
   1120 if isinstance(doc_like, str):
-> 1121     return self.make_doc(doc_like)
   1122 if isinstance(doc_like, bytes):
   1123     return Doc(self.vocab).from_bytes(doc_like)

File ~\AppData\Roaming\Python\Python39\site-packages\spacy\language.py:1110, in Languag
e.make_doc(self, text)
   1104 """Turn a text into a Doc object.
   1105
   1106 text (str): The text to process.
   1107 RETURNS (Doc): The processed doc.
   1108 """
   1109 if len(text) > self.max_length:
-> 1110     raise ValueError(
   1111         Errors.E088.format(length=len(text), max_length=self.max_length)
   1112     )
   1113 return self.tokenizer(text)

ValueError: [E088] Text of length 1220066 exceeds maximum of 1000000. The parser and NER
 models require roughly 1GB of temporary memory per 100,000 characters in the input. Thi
s means long texts may cause memory allocation errors. If you're not using the parser or
 NER, it's probably safe to increase the `nlp.max_length` limit. The limit is in number
 of characters, so you can check whether your inputs are too long by checking `len(text)
`.
```

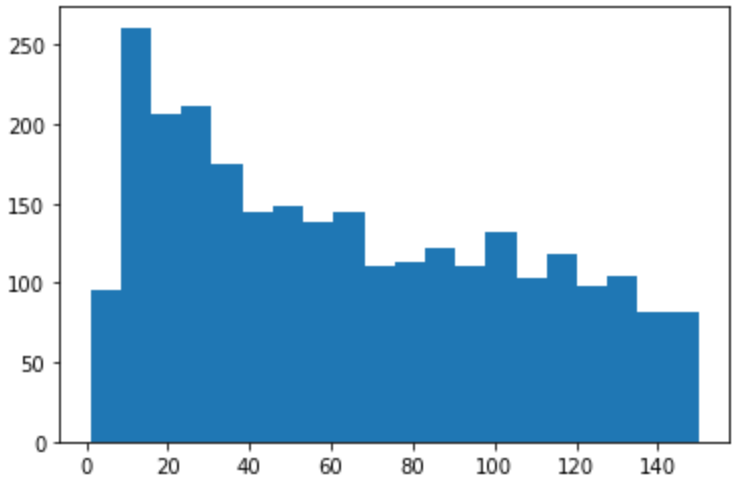
Instead, read the document paragraph by paragraph, i.e. in chunks of text separated by empty lines. Before using spaCy to segment a paragraph into sentences, replace each end of line character with a whitespace, to allow a sentence to span multiple lines. After sentence segmentation, for each sentence in the paragraph append its length (in tokens) to *lengths*. Use the default *nlp* class to process each paragraph and split it into sentences. Stop after processing 1000 paragraphs. This will be slow, so be patient.

In [211…

```
import spacy
import time

nlp = spacy.load("en_core_web_sm")

# the number of paragraphs read so far.
count = 0
```

```
# stores the length of each sentence processed so far.
lengths = []
sens=[]
paragraph = ""
flag = False
# make sure the file is read line by line.
start_time = time.time()
with open('../data/melville-moby_dick.txt', 'r') as f:
    # YOUR CODE GOES HERE
    for line in f :
        if line.strip():
            paragraph += line.strip() + ' '
            flag = True
        else:
            if flag :
                doc = nlp(paragraph)
                # Split the paragraph into sentences and append their lengths.
                for sentence in doc.sents:
                    #print(sentence.text  + "----line nene--------------------")
                    lengths.append(len(sentence.text))

                count += 1
                #print(count)
                paragraph = ""
                flag = False

        if count >= 1000:
            break

end_time = time.time()
tot_time  = end_time - start_time
print(tot_time)
len150 = [l for l in lengths if l <= 150]
plt.hist(len150, bins = 20)
plt.show()
```

```
17.19482159614563
```



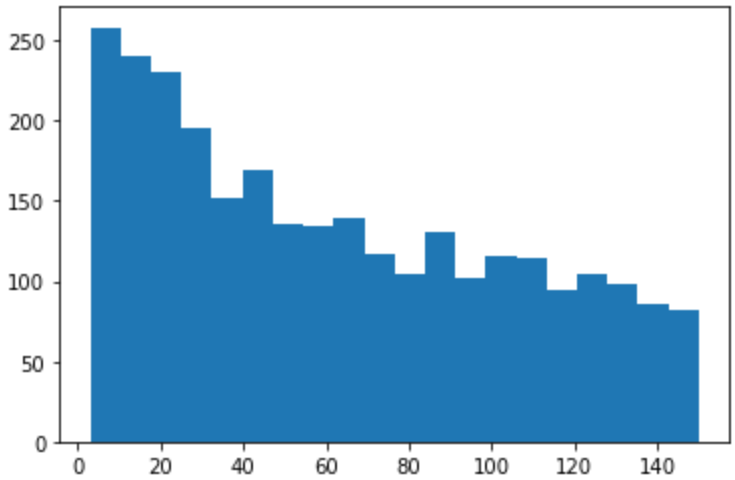
Next, do the same processing as above, but use the more robust Sentencizer to split paragraphs into sentences. Note the speedup.

In [212…

```
from spacy.lang.en import English
import time

nlp = English()
nlp.add_pipe("sentencizer")

# the number of paragraphs read so far.
count = 0
```

```
# stores the length of each sentence processed so far.
lengths = []
sens2 = []
paragraph = ""
flag = False
# make sure the file is read line by line.
start_time = time.time()
with open('../data/melville-moby_dick.txt', 'r') as f:
    # YOUR CODE GOES HERE
    for line in f :
        if line.strip():
            paragraph += line.strip() + ' '
            flag = True
        else:
            if flag :
                doc = nlp(paragraph)
                # Split the paragraph into sentences and append their lengths.
                for sentence in doc.sents:
                    #print(sentence.text  + "----line nene--------------------")
                    lengths.append(len(sentence.text))
                count += 1
                #print(count)
                paragraph = ""
                flag = False

        if count >= 1000:
            break

end_time = time.time()
tot_time_sentencizer  = end_time - start_time
print(tot_time_sentencizer)
len150 = [l for l in lengths if l <= 150]
plt.hist(len150, bins = 20)
plt.show()
```

```
0.8420326709747314
```



In [213…

```
#speed up between twi approaches

speed_up = tot_time / tot_time_sentencizer
print(speed_up)
```

```
20.420610967791806
```

Note the difference between the two histograms. Identify at least 5 examples of sentences in Moby Dick that are segmented differently by the two approaches. Copy them below and explain the differences. Which method seems to be more accurate?

Difference in a segmenting sentences and Speedup

appraoch1 -- without sentencizer pipeline approach2 -- with sentencizer pipeline

sentence in approach 1 :
1) Sw.
2) and Dan.
3) HVAL.

sentence in approach 2 :
1) Sw. and Dan. HVAL.

in approach 1 it sentencized into 3 sentences and in approach 2 only one because in first approach after period it is considering as new sentence

---

sentence in approach 1 :

1) It is more immediately from the Dut.
2) and Ger.

sentence in approach 2 :

1) It is more immediately from the Dut. and Ger.

in approach one it sentencized into 2 sentences and in approach 2 only one because in first approach after period it is considering a new sentence

---

sentence in approach 1 :
1) Among the former, one was of a most monstrous size.
2) ...

sentence in approach 2 :
1) Among the former, one was of a most monstrous size. ...

In this sentence after the period and space approach 1 splitted into 2 sentences but approach 2 considered as one sentence

---

sentence in approach 1 :
1) He visited this country also with a view of catching horse-whales, which had bones of very great value for their teeth, of which he brought some to the king.
2) ...

sentence in approach 2 :
1) He visited this country also with a view of catching horse-whales, which had bones of very great value for their teeth, of which he brought some to the king. ...

In this sentence after the period and space approach 1 splitted into 2 sentences but approach 2 considered as one sentence

---

sentence in approach 1 :
1) APOLOGY
2) FOR RAIMOND SEBOND.

sentence in approach 2 :
1) APOLOGY FOR RAIMOND SEBOND

---

sentence in approach 1 :
1) IBID.
2) "HISTORY OF LIFE AND DEATH."

sentence in approach 2 :
1) IBID. "HISTORY OF LIFE AND DEATH."

In approach one it sentencized into 2 sentences and in approach 2 only one because in first approach after period it is considering a new sentence

Approach 2 is more accurate than Appraoch 1

Word distribution using OpenAI's subword tokenization (30 points)

In this part, we will compute the frequency vs. rank based on the the BPE subword tokenization created by the tiktoken module from OpenAI.

Read the input file line by line.

1. Tokenize each line using `tiktoken` encoder and decoder for GPT-3.5.
2. For each token in the line that contains only letters, convert it to lower case and increment the corresponding count in the dictionary.
   - If the token does not exist in the dictionary yet, insert it with a count of 1. For example, the first time the token 'water' is encountered, the code should evaluate *vocab['water'] = 1*.

At the end of this code segment, *vocab* should map each word type to the number of times it appeared in the entire document. There should be 14619 unique types and 248615 total tokens in Moby Dick.

In [214…

```
import tiktoken

# To get the tokeniser corresponding to a specific model in the OpenAI API:
enc = tiktoken.encoding_for_model("gpt-4")

vocab = {}
with open('../data/melville-moby_dick.txt', 'r') as f:
    for line in f:
        # YOUR CODE HERE
        line =  line.strip()

        if line :
            tokens = enc.encode(line)
            #print(tokens)
```

```
            for token in tokens:
                token = enc.decode_single_token_bytes(token)
                token = token.decode('utf-8').strip()
                if token.isalpha():
                  # Convert to lowercase
                    #token = token.lower()
                    vocab[token] = vocab.get(token, 0) + 1


print('There are', len(vocab), 'unique tokens in Moby Dick.')
print('There are', sum(vocab.values()), 'tokens in Moby Dick.')
```

```
There are 14619 unique tokens in Moby Dick.
There are 248615 tokens in Moby Dick.
```

Rank the tokens based on their frequency, then plot frequency vs. rank.

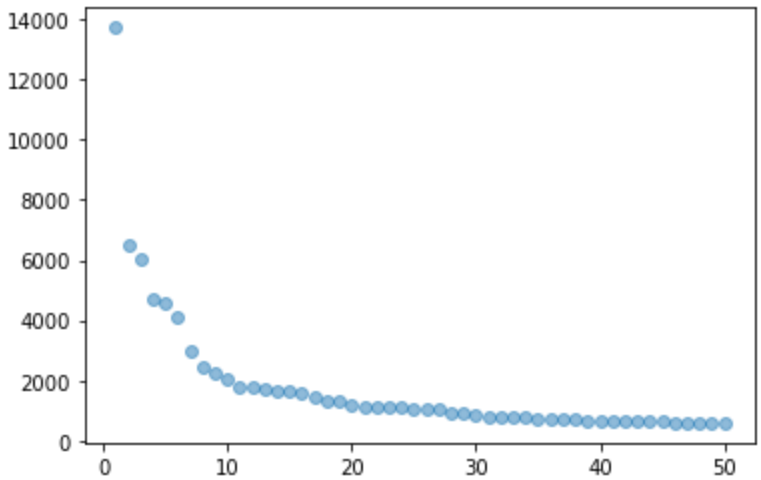
In [215…

```
sorted_vocab = dict(sorted(vocab.items(), key=lambda item: item[1], reverse=True))

#  top 10 words with the highest frequencies

ranked = list(sorted_vocab.items())

print('Size of vocabulary:', len(ranked))
for word, freq in ranked[:10]:
    print(word, freq)
```

```
Size of vocabulary: 14619
the 13739
of 6527
and 6024
a 4711
to 4582
in 4137
that 2982
his 2473
it 2260
I 2071
```

In [216…

```
import matplotlib.pyplot as plt
ranks = range(1, 50 + 1)
freqs = [t[1] for t in ranked[:50]]
plt.scatter(ranks, freqs, c='#1f77b4', alpha=0.5)
plt.show()
```



[5111] Named Entities (10 + 10 + 10 + 10 + 10 points)

Useful documentation is at:

- https://spacy.io/usage/linguistic-features#named-entities
- https://spacy.io/api/entityrecognizer

In [217…
```
import spacy

nlp = spacy.load("en_core_web_sm")

# These are all the entity types covered by spaCy's NE recognizer.
nlp.pipe_labels['ner']
```

Out[217]:
```
['CARDINAL',
 'DATE',
 'EVENT',
 'FAC',
 'GPE',
 'LANGUAGE',
 'LAW',
 'LOC',
 'MONEY',
 'NORP',
 'ORDINAL',
 'ORG',
 'PERCENT',
 'PERSON',
 'PRODUCT',
 'QUANTITY',
 'TIME',
 'WORK_OF_ART']
```

Read the first 2,500 paragraphs in Moby Dick and extract all named entities into a dictionary `ne_counts` that maps each *named entity* to its frequency. By *named entity* we mean a tuple *(name, type)* where *name* is the entity name as a string, and *type* is its entity type. For example, if the name 'Ahab' appears with the NE type 'PERSON' 50 times, then the dictionary should map the key *('Ahab', 'PERSON')* to the value *50*.

In [218…
```
# The number of paragraphs read so far.
count = 0
# Stores the dictionary of named entites and their counts.
ne_counts = {}

count = 0
# stores the length of each sentence processed so far.
lengths = []
paragraph = ""
flag = False

# make sure the file is read line by line.


with open('../data/melville-moby_dick.txt', 'r') as f:
    # YOUR CODE GOES HERE
    for line in f :

        if line.strip() :
            paragraph += line
            flag = True
        else:
            if flag :

                doc = nlp(paragraph.strip())
                for ent in doc.ents:
```

```
                        ne_tuple = (ent.text, ent.label_)
                        if ne_tuple in ne_counts:
                            ne_counts[ne_tuple] += 1
                        else:
                            ne_counts[ne_tuple] = 1
                count += 1
                #print(count)
                paragraph = ""
                lines =0
                flag = False

        if count >= 2500:
            break

for ne_tuple, frequency in ne_counts.items():
    print(ne_tuple, frequency)
```

```
('Moby Dick', 'PERSON') 57
('Herman Melville', 'PERSON') 1
('Grammar School', 'ORG') 1
('Sw', 'ORG') 1
('Dan', 'PERSON') 3
('HVALT', 'ORG') 2
('WHALE', 'ORG') 9
('Ger', 'GPE') 1
('WALLEN', 'PERSON') 1
('A.S. WALW-IAN', 'ORG') 1
("--RICHARDSON'S", 'PERSON') 1
('KETOS', 'ORG') 1
('GREEK', 'NORP') 1
('CETUS', 'ORG') 1
('LATIN', 'PRODUCT') 1
('WHOEL', 'ORG') 1
('WAL', 'ORG') 1
('ENGLISH', 'FAC') 1
('FRENCH', 'NORP') 5
('PEKEE-NUEE-NUEE', 'PRODUCT') 2
('FEGEE', 'ORG') 1
('Vaticans', 'NORP') 1
('Leviathan', 'GPE') 44
('a Sub-Sub', 'ORG') 1
('Thou', 'PERSON') 27
('one', 'CARDINAL') 350
('Hampton Court', 'ORG') 1
('seven', 'CARDINAL') 4
('Gabriel', 'PERSON') 19
('Michael', 'PERSON') 1
('Raphael', 'ORG') 1
('--GENESIS', 'GPE') 1
('Leviathan maketh', 'WORK_OF_ART') 1
('Jonah', 'GPE') 38
('--PSALMS', 'ORG') 1
('that day', 'DATE') 7
('Leviathan', 'PERSON') 10
("--HOLLAND'S", 'WORK_OF_ART') 2
('Indian Sea', 'LOC') 1
('Whales', 'PERSON') 6
('Whirlpooles', 'PERSON') 1
('Balaene', 'PERSON') 1
('four acres', 'QUANTITY') 1
('two days', 'DATE') 3
('THE TRUE', 'WORK_OF_ART') 1
('forty-eight', 'CARDINAL') 1
('some fifty yards', 'QUANTITY') 1
('six', 'CARDINAL') 7
('sixty', 'CARDINAL') 2
```

('--OTHER', 'PERSON') 1
('890', 'CARDINAL') 1
('Nick', 'PERSON') 1
('Job', 'PERSON') 1
('two', 'CARDINAL') 205
('ANNALS', 'ORG') 1
('PSALMS', 'ORG') 1
('--IBID', 'ORG') 1
('thro', 'PERSON') 1
('FAERIE QUEEN', 'ORG') 1
('WILLIAM', 'ORG') 2
('thirty years', 'DATE') 2
('Nescio', 'PERSON') 1
('HIS V. E.', 'PERSON') 1
('Commonwealth', 'ORG') 1
('Latin', 'NORP') 4
('Civitas', 'ORG') 1
('Created', 'ORG') 1
('Hugest', 'LOC') 1
('--DRYDEN', 'PERSON') 1
('twelve or thirteen feet', 'QUANTITY') 1
('EDGE', 'ORG') 1
('PURCHAS', 'ORG') 1
("T. HERBERT'S", 'PERSON') 1
('VOYAGES', 'PRODUCT') 1
('ASIA', 'LOC') 1
('AFRICA', 'LOC') 1
('SIXTH', 'ORDINAL') 1
('N.E.', 'PERSON') 1
('Jonas', 'GPE') 1
('first', 'ORDINAL') 171
('Shetland', 'GPE') 1
('One', 'CARDINAL') 19
('Spitzbergen', 'LOC') 2
('1652', 'DATE') 1
('eighty feet', 'QUANTITY') 2
('500', 'CARDINAL') 1
('Pitferren', 'GPE') 1
('KINROSS', 'GPE') 1
('Sperma', 'ORG') 1
("--RICHARD STRAFFORD'S", 'PERSON') 1
('PHIL', 'ORG') 1
('A.D.', 'GPE') 2
('1668', 'DATE') 1
('--N. E. PRIMER', 'PERSON') 1
('a hundred to one', 'CARDINAL') 1
('GLOBE', 'ORG') 3
('1729', 'DATE') 1
("--ULLOA'S SOUTH AMERICA", 'WORK_OF_ART') 1
('fifty', 'CARDINAL') 12
('seven-fold', 'CARDINAL') 1
('--GOLDSMITH', 'PRODUCT') 1
('NAT', 'ORG') 1
('HIST', 'ORG') 1
('--GOLDSMITH', 'NORP') 1
('the afternoon', 'TIME') 2
('Asiatics', 'NORP') 2
("--COOK'S\nVOYAGES", 'WORK_OF_ART') 1
("VON TROIL'S", 'ORG') 1
('1772', 'DATE') 1
('Nantuckois', 'PERSON') 1
('1778', 'DATE') 2
('THE NANTUCKET WHALE-FISHERY', 'ORG') 1
('Spain', 'GPE') 5
('Europe', 'LOC') 4
('BURKE', 'ORG') 1

('tenth', 'ORDINAL') 3
('--COWPER', 'GPE') 1
("ON THE QUEEN'S", 'ORG') 1
('LONDON', 'GPE') 1
('Ten or fifteen gallons', 'QUANTITY') 1
('London Bridge', 'FAC') 1
("--PALEY'S", 'PERSON') 1
('THEOLOGY', 'ORG') 1
('40 degrees', 'QUANTITY') 1
('Spermacetti Whales', 'PERSON') 1
('May', 'DATE') 5
('millions', 'CARDINAL') 5
("--MONTGOMERY'S WORLD BEFORE THE FLOOD", 'WORK_OF_ART') 1
('Paean', 'NORP') 1
('Io', 'LOC') 1
('Atlantic', 'LOC') 16
('the Polar Sea', 'LOC') 1
('LAMB', 'NORP') 1
('the year 1690', 'DATE') 1
('NANTUCKET', 'ORG') 5
('Susan', 'PERSON') 1
('a Gothic Arch', 'NORP') 1
('the Pacific ocean', 'LOC') 1
('no less than forty years\nago', 'DATE') 1
('Tom', 'PERSON') 2
('Christian', 'NORP') 17
("--COOPER'S PILOT", 'WORK_OF_ART') 1
('the Berlin Gazette', 'GPE') 1
('--ECKERMANN', 'PERSON') 1
('Chace', 'PERSON') 1
('THE PACIFIC OCEAN', 'LOC') 1
('BY OWEN CHACE OF NANTUCKET', 'ORG') 1
('FIRST', 'ORDINAL') 1
('NEW YORK', 'GPE') 1
('1821', 'DATE') 1
('one night', 'TIME') 4
('10,440 yards', 'QUANTITY') 1
('nearly six', 'CARDINAL') 1
('English', 'LANGUAGE') 39
('three or', 'CARDINAL') 1
('"Mad with the agonies', 'WORK_OF_ART') 1
('Sperm Whale', 'PERSON') 13
('the Sperm Whale', 'GPE') 9
('late years', 'DATE') 2
('1839', 'DATE') 2
('The Cachalot', 'WORK_OF_ART') 1
('Greenland or Right Whale', 'WORK_OF_ART') 1
('1840', 'DATE') 1
('October 13', 'DATE') 1
('Three', 'CARDINAL') 8
('lee', 'PERSON') 6
('Raise up your wheel', 'WORK_OF_ART') 1
('Sperm Whales', 'PERSON') 4
('THAR', 'ORG') 1
('Two miles and a half', 'QUANTITY') 1
('1846', 'DATE') 1
('Globe', 'ORG') 1
('Nantucket', 'GPE') 58
('HUSSEY', 'NORP') 1
('1828', 'DATE') 2
('Webster', 'PERSON') 1
('National', 'ORG') 1
('eight or nine thousand', 'CARDINAL') 1
('WEBSTER', 'PERSON') 1
('SENATE', 'ORG') 1
('ON THE APPLICATION FOR THE ERECTION OF A', 'ORG') 1

('BREAKWATER', 'ORG') 1
('WHALEMAN', 'PERSON') 1
('ADVENTURES', 'PRODUCT') 1
('REV', 'ORG') 1
('HENRY T. CHEEVER', 'PERSON') 1
('Samuel', 'PERSON') 3
('BROTHER', 'ORG') 1
('Dutch', 'NORP') 22
('the Northern Ocean', 'LOC') 1
('India', 'GPE') 5
("--MCCULLOCH'S COMMERCIAL", 'FAC') 1
('North-West Passage', 'ORG') 1
('U.S. EX', 'ORG') 1
('Pedestrians', 'NORP') 1
('London', 'GPE') 11
('THE ARCTIC OCEAN', 'LOC') 1
('American', 'NORP') 28
('The Whale', 'WORK_OF_ART') 1
('Terra Del Fuego', 'WORK_OF_ART') 1
("--DARWIN'S VOYAGE OF A NATURALIST", 'WORK_OF_ART') 1
('--WHARTON THE WHALE KILLER', 'WORK_OF_ART') 1
('harpooneer', 'ORG') 1
('--NANTUCKET SONG', 'PERSON') 1
('Whale', 'GPE') 5
('--WHALE SONG', 'PERSON') 1
('CHAPTER 1', 'LAW') 2
('Some years ago', 'DATE') 1
('November', 'DATE') 1
('Cato', 'ORG') 1
('Indian', 'NORP') 48
('a few hours', 'TIME') 1
('Sabbath afternoon', 'PERSON') 1
('Corlears Hook', 'PERSON') 1
('Whitehall', 'ORG') 1
('thousands upon thousands', 'CARDINAL') 1
('China', 'GPE') 8
('week days', 'DATE') 1
('ten', 'CARDINAL') 16
('Saco', 'ORG') 1
('Prairies', 'ORG') 2
('June', 'DATE') 1
('Tiger', 'PERSON') 1
('Niagara', 'PERSON') 2
('your thousand miles', 'QUANTITY') 1
('Tennessee', 'GPE') 1
('Rockaway', 'ORG') 1
('Persians', 'NORP') 1
('Greeks', 'PERSON') 2
('Jove', 'PERSON') 7
('Narcissus', 'PERSON') 1
('nights', 'TIME') 2
('Cook', 'PERSON') 12
('barques', 'ORG') 1
('brigs', 'PERSON') 1
('Egyptians', 'NORP') 4
('the Van Rensselaers', 'FAC') 1
('Randolphs', 'NORP') 1
('Hardicanutes', 'LOC') 1
('Seneca', 'PRODUCT') 1
('Stoics', 'ORG') 1
('the New Testament', 'ORG') 1
('a single\npenny', 'MONEY') 1
('Pythagorean', 'NORP') 1
('second', 'ORDINAL') 45
('Fates', 'NORP') 2
('Providence', 'GPE') 2

('THE UNITED STATES', 'GPE') 1
('ONE', 'CARDINAL') 2
('AFFGHANISTAN', 'ORG') 1
('thousand', 'CARDINAL') 14
('Patagonian', 'NORP') 3
('sail forbidden\nseas', 'PERSON') 1
('CHAPTER 2', 'LAW') 1
('Carpet-Bag', 'ORG') 1
('Cape Horn', 'LOC') 12
('Pacific', 'LOC') 28
('Manhatto', 'PERSON') 1
('New Bedford', 'GPE') 18
('Saturday', 'DATE') 5
('night', 'TIME') 22
('December', 'DATE') 3
('Nantucket', 'ORG') 6
('Monday', 'DATE') 2
('Nantucket', 'PERSON') 3
('Tyre', 'ORG') 1
('the Red-Men', 'ORG') 1
('a night', 'TIME') 2
('another night', 'TIME') 1
('Ishmael', 'GPE') 12
('the night', 'TIME') 11
('Ishmael', 'PERSON') 4
('The\nCrossed', 'WORK_OF_ART') 1
('Sword-Fish Inn', 'WORK_OF_ART') 1
('ten inches', 'QUANTITY') 4
('this hour', 'TIME') 3
('the last day of the week', 'DATE') 1
('Gomorrah', 'PERSON') 2
('The Crossed Harpoons', 'WORK_OF_ART') 1
('The\nSword-Fish?"--this', 'WORK_OF_ART') 1
('"The Trap', 'WORK_OF_ART') 1
('Black Parliament', 'ORG') 1
('Tophet', 'GPE') 2
('A hundred', 'CARDINAL') 1
('Angel of Doom', 'PRODUCT') 1
("'The Trap", 'WORK_OF_ART') 1
('Spouter Inn:--Peter Coffin', 'LOC') 1
('I.', 'ORG') 12
('Peter', 'PERSON') 1
('Euroclydon', 'GPE') 3
('Paul', 'PERSON') 1
('Euroclydon', 'PERSON') 2
('Death', 'PERSON') 3
('a million years ago', 'DATE') 1
('summer', 'DATE') 9
('Lazarus', 'PERSON') 2
('Sumatra', 'NORP') 2
('Lazarus', 'ORG') 2
('Moluccas', 'GPE') 2
('Czar', 'PERSON') 3
('CHAPTER 3', 'LAW') 1
('The Spouter-Inn', 'ORG') 1
('Spouter-Inn', 'ORG') 1
('New England', 'LOC') 6
('over three', 'CARDINAL') 1
('half', 'CARDINAL') 90
('the Black Sea', 'LOC') 1
('midnight', 'TIME') 18
('four', 'CARDINAL') 41
('elements.--It', 'PERSON') 1
('Hyperborean', 'ORG') 2
('Time', 'ORG') 3
('three', 'CARDINAL') 134

```
('fifty years ago', 'DATE') 1
('Nathan Swain', 'PERSON') 1
('fifteen', 'CARDINAL') 1
('Javan', 'GPE') 1
('years', 'DATE') 11
('the Cape of Blanco', 'LOC') 1
('forty feet', 'QUANTITY') 3
('Jonah', 'PERSON') 33
('a penny', 'MONEY') 2
('Battery', 'LOC') 1
('Iceland', 'GPE') 1
('Landlord', 'PERSON') 7
('the\nharpooneer', 'PRODUCT') 1
('I.  "', 'PERSON') 1
('Grampus', 'PERSON') 1
('this morning', 'TIME') 2
("three years'", 'DATE') 3
('Hurrah', 'PERSON') 7
('Feegees', 'ORG') 1
('Labrador', 'GPE') 3
('six feet', 'QUANTITY') 2
('Southerner', 'PRODUCT') 1
('Alleghanian Ridge', 'GPE') 1
('Virginia', 'GPE') 3
('a few\nminutes', 'TIME') 1
('Bulkington', 'PERSON') 8
("about nine o'clock", 'TIME') 1
('Suppose', 'PERSON') 1
('about four inches', 'QUANTITY') 1
('the next morning', 'TIME') 1
('late hours', 'TIME') 1
("twelve o'clock", 'TIME') 1
('airley', 'PERSON') 1
('Sunday', 'DATE') 5
('morning', 'TIME') 9
('BROWN', 'PERSON') 1
('Broke', 'WORK_OF_ART') 1
('Mt. Hecla', 'LOC') 1
('New\nZealand', 'GPE') 1
('last Sunday', 'DATE') 1
('Sabbath', 'PERSON') 3
('Sal', 'PERSON') 3
('Sam', 'PERSON') 2
('Johnny', 'PERSON') 1
('about one night', 'TIME') 1
('South American', 'NORP') 1
('a rainy day', 'DATE') 1
('now', 'DATE') 1
('Nod', 'PERSON') 1
('New Zealand', 'GPE') 5
("Thirty\nYears' War", 'DATE') 1
('the South Seas', 'LOC') 2
("three days' old", 'DATE') 1
('Congo', 'GPE') 2
('First', 'ORDINAL') 16
('you?"--he', 'PRODUCT') 1
('Peter Coffin', 'PERSON') 2
('Coffin', 'GPE') 3
('Angels', 'ORG') 1
('Queequeg', 'NORP') 98
('peddlin', 'PERSON') 1
('Queequeg', 'GPE') 114
('CHAPTER 4', 'LAW') 1
('Counterpane', 'ORG') 1
('next morning', 'TIME') 3
('Queequeg', 'PERSON') 4
```

('Cretan', 'GPE') 2
('a few days', 'DATE') 1
('supperless,--my', 'NORP') 1
("only two o'clock", 'TIME') 1
('afternoon', 'TIME') 3
('the 21st June', 'DATE') 1
('the longest day', 'DATE') 1
('the year', 'DATE') 5
('third', 'ORDINAL') 12
('sixteen entire hours', 'TIME') 1
('Sixteen hours', 'TIME') 1
('several hours', 'TIME') 2
('one single inch', 'QUANTITY') 1
('the morning', 'TIME') 3
('days', 'DATE') 10
('weeks', 'DATE') 4
('months', 'DATE') 5
('this very hour', 'TIME') 1
("the past night's", 'TIME') 1
('the broad day', 'DATE') 1
('Newfoundland', 'GPE') 1
('caterpillar', 'ORG') 1
('Rogers', 'ORG') 1
('CHAPTER 5', 'LAW') 1
('three\ndays', 'DATE') 1
('satin', 'NORP') 1
('Andes', 'LOC') 6
('Mungo Park', 'FAC') 1
('Scotch', 'NORP') 1
('Siberia', 'LOC') 2
('Africa', 'LOC') 7
('Mungo', 'ORG') 1
('the Green Mountains', 'LOC') 2
('CHAPTER 6', 'LAW') 1
('Broadway', 'FAC') 2
('Chestnut', 'PERSON') 1
('Mediterranean', 'LOC') 8
('Regent Street', 'LOC') 1
('Lascars', 'GPE') 1
('Malays', 'DATE') 1
('Bombay', 'GPE') 1
('Apollo', 'ORG') 1
('Yankees', 'ORG') 3
('Wapping', 'GPE') 2
('Feegeeans', 'NORP') 1
('Tongatobooarrs', 'ORG') 1
('Erromanggoans', 'NORP') 1
('Pannangians', 'NORP') 1
('Brighggians', 'NORP') 1
('weekly', 'DATE') 1
('a few hours old', 'TIME') 1
('two acres', 'QUANTITY') 1
('Canaan', 'PERSON') 2
('America', 'GPE') 7
('Herr Alexander', 'PERSON') 1
('every night', 'TIME') 2
('August', 'DATE') 3
('seventh', 'ORDINAL') 2
('Salem', 'GPE') 1
('Puritanic', 'ORG') 1
('CHAPTER 7', 'LAW') 1
('Whaleman', 'PERSON') 1
('the Indian Ocean', 'LOC') 1
('quote:--', 'ORG') 1
('JOHN TALBOT', 'PERSON') 1
('the age of eighteen', 'DATE') 1

('the Isle of Desolation', 'GPE') 1
('Patagonia', 'GPE') 2
('November 1st, 1836', 'DATE') 1
('NATHAN COLEMAN', 'PERSON') 1
('WALTER CANNY', 'PERSON') 1
('SETH MACY', 'PERSON') 1
('PACIFIC', 'LOC') 1
('December 31st, 1839', 'DATE') 1
('Japan', 'GPE') 10
('AUGUST 3d, 1833', 'ORG') 1
('Faith', 'ORG') 1
('Elephanta', 'GPE') 2
('yesterday', 'DATE') 5
('the Life Insurance Companies', 'ORG') 1
('Adam', 'PERSON') 5
('sixty round centuries\nago', 'DATE') 1
('Methinks', 'ORG') 4
('Life and Death', 'ORG') 1
('CHAPTER 8', 'LAW') 1
('Mapple', 'PERSON') 9
('winter', 'DATE') 7
('February', 'DATE') 1
('Quebec', 'GPE') 2
('Ehrenbreitstein', 'GPE') 1
('Nelson', 'PERSON') 3
('the Holy Bible', 'ORG') 1
('earth', 'LOC') 7
('CHAPTER 9', 'LAW') 1
('Sermon', 'PERSON') 1
('Midships', 'ORG') 1
('tell--', 'PERSON') 1
('Deliverer', 'PRODUCT') 1
('Bible', 'WORK_OF_ART') 8
('only four', 'CARDINAL') 2
('Scriptures', 'ORG') 1
('the\nkelpy', 'ORG') 1
('Shipmates', 'ORG') 2
('Amittai', 'PERSON') 1
('Joppa', 'GPE') 6
('Tarshish', 'NORP') 7
('Cadiz', 'GPE') 1
('Cadiz', 'PERSON') 1
('Cadiz', 'ORG') 1
('Joppa', 'PERSON') 2
('Jaffa', 'PERSON') 2
('Syrian', 'NORP') 3
('Cadiz', 'NORP') 1
('two thousand miles', 'QUANTITY') 2
('Straits of Gibraltar', 'LOC') 1
('those days', 'DATE') 3
('Joe', 'PERSON') 1
('Harry\nlad', 'PERSON') 1
('Sodom', 'ORG') 1
('five hundred', 'CARDINAL') 1
('Frighted Jonah', 'PERSON') 1
("Now Jonah's Captain", 'PERSON') 1
('Jonah\nnow', 'PERSON') 1
('that stifling hour', 'TIME') 1
('Roman', 'NORP') 12
('Aye', 'ORG') 16
('Startled', 'PERSON') 1
('white moon', 'FAC') 1
('Hebrew', 'LANGUAGE') 3
('Straightway', 'GPE') 2
('Sin', 'ORG') 1
('Nineveh', 'GPE') 3

('ten thousand', 'CARDINAL') 4
("hell'--when", 'PRODUCT') 1
('Truth', 'ORG') 3
('Falsehood', 'PERSON') 1
('Gospel', 'GPE') 1
('Pilot Paul', 'PERSON') 1
('kelson', 'GPE') 3
('Delight', 'PERSON') 4
('no quarter', 'DATE') 1
('Keel of the Ages', 'WORK_OF_ART') 1
('Thy', 'PERSON') 2
('Thine', 'ORG') 1
('CHAPTER 10', 'LAW') 1
('Bosom Friend', 'ORG') 1
('the Spouter-Inn', 'ORG') 2
('fiftieth', 'ORDINAL') 2
('the next fifty', 'DATE') 1
('number one', 'CARDINAL') 1
('more than fifty', 'CARDINAL') 1
('Pagan', 'GPE') 5
('Washington', 'GPE') 4
('George\nWashington', 'PERSON') 1
('the\nnight', 'TIME') 1
('Socratic', 'NORP') 1
('some twenty thousand miles', 'QUANTITY') 1
('Jupiter', 'LOC') 3
('evening', 'TIME') 5
("last night's", 'TIME') 2
('thirty dollars', 'MONEY') 1
('Presbyterian Church', 'ORG') 1
('Presbyterian', 'NORP') 1
('nearly morning', 'TIME') 1
('CHAPTER 11', 'LAW') 1
('kneepans', 'NORP') 1
('arctic', 'LOC') 1
('day', 'DATE') 5
("twelve-o'clock-at-night", 'TIME') 1
('Tomahawk', 'PRODUCT') 2
('CHAPTER 12', 'LAW') 1
('Biographical', 'ORG') 1
('Rokovoko', 'FAC') 2
('West', 'LOC') 6
('South', 'LOC') 4
('Christendom', 'ORG') 4
('Sag Harbor', 'LOC') 1
('Struck', 'PERSON') 1
('Prince of Wales', 'GPE') 1
('Czar Peter', 'PERSON') 1
('Christians', 'NORP') 6
('Sag\nHarbor', 'PERSON') 1
('pagan', 'GPE') 1
('Christianity', 'NORP') 2
('thirty', 'CARDINAL') 5
('Potluck', 'PERSON') 1
('CHAPTER 13', 'LAW') 1
('Next morning', 'TIME') 4
("Peter Coffin's", 'PERSON') 1
('Nantucket', 'NORP') 2
('Sag Harbor', 'GPE') 1
('Rokovoko', 'ORG') 1
('Grace', 'PERSON') 1
('Grace', 'ORG') 3
('Moss', 'PERSON') 2
('Tartar', 'NORP') 1
('Sultan', 'LOC') 1
('hour', 'TIME') 2

('Capting, Capting', 'WORK_OF_ART') 1
('Hallo', 'WORK_OF_ART') 1
('three minutes', 'TIME') 2
('Queequeg', 'LANGUAGE') 8
('A few minutes', 'TIME') 1
('that hour', 'TIME') 3
('CHAPTER 14', 'LAW') 1
('Eddystone', 'PERSON') 1
('twenty years', 'DATE') 1
('Canada', 'GPE') 2
('Rome', 'GPE') 4
('Laplander', 'PRODUCT') 1
('Illinois', 'GPE') 3
('Nantucketers', 'PERSON') 9
("Behring's Straits", 'FAC') 1
('Himmalehan', 'GPE') 1
('Mastodon', 'GPE') 1
('ant-hill', 'GPE') 1
('Alexanders', 'ORG') 1
('Poland', 'GPE') 2
('Mexico', 'GPE') 3
('Texas', 'GPE') 2
('Cuba', 'GPE') 2
('two thirds', 'CARDINAL') 3
('Nantucketer', 'ORG') 18
('Merchant', 'ORG') 1
('Noah', 'PERSON') 6
('Alps', 'LOC') 2
('Earthsman', 'PRODUCT') 1
('CHAPTER 15', 'LAW') 1
('late in the evening', 'TIME') 1
('Hosea\nHussey', 'PERSON') 1
('the Try Pots', 'ORG') 2
('Cousin Hosea', 'PERSON') 1
('the Try Pots', 'FAC') 1
('Two', 'CARDINAL') 3
('TWO', 'CARDINAL') 5
('Hussey', 'PERSON') 13
('Hosea Hussey', 'PERSON') 2
('said--"Clam', 'ORG') 1
('Cod', 'PERSON') 1
('Shirt', 'PERSON') 1
('codfish', 'ORG') 1
('one morning', 'TIME') 5
('Hosea', 'PERSON') 1
('four years', 'DATE') 3
('a half', 'CARDINAL') 1
('only three barrels', 'QUANTITY') 1
('ILE', 'ORG') 1
('CHAPTER 16', 'LAW') 1
('Yojo', 'PERSON') 17
('two or three', 'CARDINAL') 3
('Lent or Ramadan', 'FAC') 1
('XXXIX Articles', 'PERSON') 1
("three-years'", 'DATE') 2
('Tit', 'PERSON') 2
('Pequod', 'GPE') 152
('TIT-BIT', 'ORG') 1
('PEQUOD', 'ORG') 1
('Massachusetts', 'GPE') 3
('Indians', 'NORP') 2
('Medes', 'ORG') 1
('Japanese', 'NORP') 9
('French', 'NORP') 16
('Egypt', 'GPE') 3
('Cologne', 'GPE') 4

('Canterbury', 'GPE') 1
('Becket', 'ORG') 1
('more than half', 'CARDINAL') 2
('Peleg', 'PERSON') 55
("Thorkill-Hake's", 'ORG') 1
('Ethiopian', 'NORP') 2
('Tartar', 'PRODUCT') 1
('some one', 'CARDINAL') 4
('some ten feet', 'QUANTITY') 3
('the apex united', 'FAC') 1
("Pottowottamie Sachem's", 'PERSON') 1
('noon', 'TIME') 10
('Quaker', 'ORG') 10
('Thou', 'GPE') 4
('Quakerish Nantucketer', 'PERSON') 1
('Cape Cod', 'PERSON') 2
('Ahab', 'PERSON') 344
('Bildad', 'PERSON') 10
('Young', 'PERSON') 1
('merchant--', 'PRODUCT') 1
('Answer', 'ORG') 1
('Bildad', 'GPE') 65
('this day', 'DATE') 8
('Quakers', 'PERSON') 2
('Scandinavian', 'NORP') 2
('Pagan Roman', 'PERSON') 1
('whaleman', 'PERSON') 16
('Nantucket\nQuakerism', 'ORG') 1
('Horn', 'LOC') 2
('gore', 'PERSON') 2
('Categut', 'ORG') 1
('Indolence', 'ORG') 1
('Scriptures', 'PERSON') 1
('the last thirty years', 'DATE') 1
('275th', 'ORDINAL') 3
('275th', 'CARDINAL') 1
('the Thunder Cloud', 'FAC') 1
('200th', 'ORDINAL') 1
('Captain Peleg', 'PERSON') 2
('Thou knowest', 'PERSON') 1
('seven hundred', 'CARDINAL') 3
('LAY--', 'PERSON') 1
('LAYS', 'ORG') 1
('seven\nhundred and seventy-seven', 'CARDINAL') 1
('seven hundred and seventy-seven', 'CARDINAL') 1
('Seven hundred and seventy-seventh', 'CARDINAL') 1
('Captain Peleg', 'WORK_OF_ART') 1
('thou', 'CARDINAL') 7
('seven hundred and seventy-seventh', 'CARDINAL') 2
('Thou Bildad', 'PERSON') 1
('hundredth', 'CARDINAL') 1
('Cape', 'LOC') 4
('Mark', 'PERSON') 4
('AHAB', 'ORG') 3
('Ahab of old', 'PERSON') 1
('thou knowest', 'PERSON') 1
('Tistig', 'PERSON') 1
('Gayhead', 'DATE') 1
('years ago', 'DATE') 3
('CHAPTER 17', 'LAW') 1
('all day', 'DATE') 9
('night-fall', 'DATE') 2
('Presbyterian Christians', 'NORP') 1
('Ramadan;--but', 'ORG') 1
('Presbyterians', 'NORP') 1
('Pagans', 'NORP') 1

('La! la', 'WORK_OF_ART') 1
('La', 'GPE') 1
('Betty', 'PERSON') 1
('Snarles the Painter', 'ORG') 1
('one about a mile', 'QUANTITY') 1
('eight or ten hours', 'TIME') 1
('Ramadan', 'DATE') 3
('the Atlantic\nOcean', 'LOC') 1
("nearly\neleven o'clock", 'TIME') 1
('bearskin jacket', 'PERSON') 1
('four feet', 'QUANTITY') 1
('all night', 'TIME') 2
('Lents', 'PERSON') 1
('Ramadans', 'NORP') 2
('Hygiene', 'PERSON') 1
("about two\no'clock", 'CARDINAL') 1
('CHAPTER 18', 'LAW') 1
('Congregational Church', 'ORG') 2
('First Congregational Church', 'ORG') 1
("Deacon Deuteronomy Coleman's", 'PERSON') 1
("Deacon Deuteronomy's", 'PERSON') 1
('Deacon Deuteronomy', 'PERSON') 1
('Catholic Church', 'ORG') 1
('First Congregation', 'ORG') 1
('Deacon', 'PERSON') 1
('Quohog', 'PERSON') 9
('Hedgehog', 'PERSON') 1
('ninetieth', 'ORDINAL') 1
('this:--', 'PERSON') 1
('The Latter Day Coming', 'WORK_OF_ART') 1
('Lose', 'PERSON') 1
('Belial', 'ORG') 1
('Spurn', 'PRODUCT') 1
('Bell', 'ORG') 1
('Scriptural', 'NORP') 1
('Nat Swaine', 'PERSON') 1
('Vineyard', 'ORG') 2
('thou knowest', 'GPE') 1
('Death and the Judgment', 'WORK_OF_ART') 2
('CHAPTER 19', 'LAW') 1
('fifth', 'ORDINAL') 4
('three days', 'DATE') 6
('Spaniard', 'NORP') 2
('Anyhow', 'PERSON') 1
('Morning', 'TIME') 3
('Morning', 'WORK_OF_ART') 1
('a hundred yards', 'QUANTITY') 1
('Elijah', 'PERSON') 10
('the Cape Horn', 'LOC') 2
('Captain\nPeleg', 'PERSON') 1
('the day', 'DATE') 10
('Tistig', 'ORG') 1
('a hundred', 'CARDINAL') 4
('CHAPTER 20', 'LAW') 1
('several days', 'DATE') 3
('Island', 'GPE') 1
('SHE', 'ORG') 1
('Aunt Charity', 'ORG') 1
('Quakeress', 'ORG') 2
('the last day', 'DATE') 2
('these days', 'DATE') 1
('every day', 'DATE') 5
('next day', 'DATE') 2
('CHAPTER 21', 'LAW') 1
('Going Aboard', 'PERSON') 1
("nearly six o'clock", 'TIME') 1

```
('Lookee', 'PERSON') 1
('Pacific Oceans', 'NORP') 1
('said--"Did', 'PERSON') 1
('five', 'CARDINAL') 11
('ottomans', 'NORP') 1
('eight', 'CARDINAL') 8
('last night', 'TIME') 1
('Captain?--Ahab', 'PERSON') 1
('Starbuck', 'PERSON') 146
('CHAPTER 22', 'LAW') 1
('Christmas', 'DATE') 4
('Muster', 'ORG') 1
('Booble Alley', 'ORG') 1
('Watts', 'LOC') 1
('Spring', 'WORK_OF_ART') 1
('spring', 'DATE') 4
('Scotch-cap', 'PERSON') 1
('Spring', 'DATE') 6
('heard,--', 'ORG') 1
('Stand', 'ORG') 1
('Jews', 'NORP') 1
('Jordan', 'PERSON') 1
('winter night', 'TIME') 1
('the spring', 'DATE') 1
('midsummer', 'DATE') 1
('Capes', 'ORG') 1
('thousands', 'CARDINAL') 5
('Eastern', 'ORG') 2
('Stubb', 'PERSON') 19
('Flask', 'PERSON') 9
('this day three years', 'DATE') 1
("ye'll", 'PERSON') 3
('three per cent', 'MONEY') 1
('twenty cents', 'MONEY') 1
('CHAPTER 23', 'LAW') 1
('mid-winter', 'DATE') 1
("four years'", 'DATE') 2
('Wonderfullest', 'PERSON') 1
('six-inch', 'QUANTITY') 1
('CHAPTER 24', 'LAW') 1
('Advocate', 'ORG') 1
('Martial Commanders', 'ORG') 1
("De Witt's", 'GPE') 1
('Louis XVI', 'PERSON') 1
('France', 'GPE') 5
('Dunkirk', 'ORG') 1
('Britain', 'GPE') 1
('the years 1750', 'DATE') 1
('1788', 'DATE') 1
('L1,000,000', 'PERSON') 1
('eighteen thousand', 'CARDINAL') 1
('4,000,000 of dollars', 'MONEY') 1
('20,000,000', 'MONEY') 1
('every year', 'DATE') 1
('7,000,000', 'MONEY') 1
('sixty years', 'DATE') 1
('Egyptian', 'NORP') 7
('many years', 'DATE') 2
('Vancouver', 'GPE') 2
('European', 'NORP') 2
('Exploring Expeditions', 'ORG') 1
('Krusensterns', 'NORP') 1
('Krusenstern', 'NORP') 1
('South Sea Voyages', 'LOC') 1
('Spanish', 'NORP') 8
('Peru', 'GPE') 4
```

```
('Chili', 'PERSON') 1
('Bolivia', 'GPE') 2
('Australia', 'GPE') 2
('Dutchman', 'GPE') 1
('Australian', 'NORP') 1
('Polynesia', 'GPE') 1
('WHALE NO FAMOUS AUTHOR', 'ORG') 1
('Alfred the Great', 'WORK_OF_ART') 1
('Norwegian', 'NORP') 1
('Parliament', 'ORG') 1
('Edmund Burke', 'PERSON') 1
('Benjamin Franklin', 'PERSON') 1
('Mary Morrel', 'PERSON') 1
('Mary Folger', 'PERSON') 1
('Folgers', 'ORG') 1
('kin', 'ORG') 1
('Benjamin', 'PERSON') 1
('WHALING', 'ORG') 1
('Cetus', 'ORG') 2
('three hundred', 'CARDINAL') 3
('MSS', 'ORG') 1
('Yale College', 'ORG') 1
('Harvard', 'ORG') 2
('CHAPTER 25', 'LAW') 1
('Britons', 'GPE') 1
('CHAPTER 26', 'LAW') 1
('Knights and Squires', 'ORG') 2
('Indies', 'ORG') 1
('those thousand-fold', 'CARDINAL') 1
('Outward', 'PERSON') 1
('hundreds', 'CARDINAL') 5
('august', 'DATE') 3
('Bunyan', 'GPE') 1
('Cervantes', 'PERSON') 1
('Andrew Jackson', 'PERSON') 1
('CHAPTER 27', 'LAW') 1
('a Cape-Cod-man', 'PRODUCT') 1
('joiner', 'PERSON') 1
('flank', 'ORG') 1
('Flask', 'ORG') 38
('Tisbury', 'ORG') 1
('Martha', 'PERSON') 3
('King-Post', 'ORG') 10
('Arctic', 'LOC') 4
('three mates', 'QUANTITY') 1
('Flask', 'GPE') 38
('Gothic Knight', 'PERSON') 1
('Tashtego', 'ORG') 40
('Gay Head', 'PERSON') 1
('Gay-Headers', 'PERSON') 1
('Oriental', 'NORP') 6
('Antarctic', 'LOC') 5
('Puritans', 'NORP') 1
('the Prince of the Powers', 'GPE') 1
('Tashtego', 'PERSON') 2
('Third', 'ORDINAL') 3
('Ahasuerus', 'PRODUCT') 1
('Daggoo', 'GPE') 13
('Ahasuerus', 'PERSON') 1
('the present day', 'DATE') 6
('Americans', 'NORP') 6
('Herein', 'PERSON') 1
('the American Canals and Railroads', 'ORG') 1
('Azores', 'PERSON') 2
('Greenland', 'GPE') 25
('Hull', 'PERSON') 1
```

```
('the Shetland Islands', 'GPE') 1
('Islanders', 'NORP') 4
('ISOLATOES', 'ORG') 1
('ISOLATO', 'ORG') 1
('Clootz', 'ORG') 1
('Black Little Pip', 'EVENT') 1
('Alabama', 'GPE') 6
('CHAPTER 28', 'LAW') 1
('Elijah', 'ORG') 1
('Polar', 'PERSON') 1
('mornings', 'TIME') 1
('Reality', 'PRODUCT') 1
('Cellini', 'GPE') 1
('Perseus', 'PERSON') 4
('Gay-Head\n', 'PERSON') 1
('Manxman', 'PERSON') 10
('Gay-Head Indian', 'PERSON') 1
('quarter', 'DATE') 5
('about half an inch', 'QUANTITY') 1
('that morning', 'TIME') 3
('April', 'DATE') 2
('CHAPTER 29', 'LAW') 1
('Some days', 'DATE') 1
('Quito', 'GPE') 2
('Tropic', 'ORG') 1
('Persian', 'NORP') 4
('Earls', 'PERSON') 1
('hours', 'TIME') 5
('six inches', 'QUANTITY') 1
('Captain\nAhab', 'PERSON') 1
('less than half', 'CARDINAL') 1
('more than three hours', 'TIME') 1
('twenty-four', 'CARDINAL') 1
('Dough-Boy', 'PERSON') 9
('a morning', 'TIME') 1
('Tic-Dolly-row', 'ORG') 1
('Dough-Boy', 'ORG') 1
('eleventh', 'ORDINAL') 1
('twelfth', 'ORDINAL') 1
('CHAPTER 30', 'LAW') 1
('Pipe', 'FAC') 1
('Norse times', 'ORG') 1
('Danish', 'NORP') 2
('Khan', 'PERSON') 1
('Leviathans', 'NORP') 7
('CHAPTER 31', 'LAW') 1
('Queen Mab', 'PERSON') 1
('merman', 'PERSON') 1
('Slid', 'PERSON') 1
('Humpback', 'PERSON') 1
('Halloa', 'PERSON') 3
("I.  '", 'PERSON') 1
('England', 'GPE') 12
('CHAPTER 32', 'LAW') 1
('Zoology', 'ORG') 1
('Captain Scoresby', 'PERSON') 2
('1820', 'DATE') 1
('Surgeon Beale', 'PERSON') 1
('John Hunter', 'PERSON') 2
('Lesson', 'PERSON') 1
('Aristotle', 'ORG') 1
('Thomas Browne', 'PERSON') 1
('Gesner', 'ORG') 1
('Ray', 'PERSON') 1
('Rondeletius', 'ORG') 1
('Willoughby', 'PERSON') 1
```

```
('Green', 'PERSON') 1
('Artedi', 'GPE') 1
('Sibbald', 'GPE') 1
('Brisson', 'GPE') 1
('Lacepede', 'ORG') 1
('Bonneterre', 'PERSON') 1
('Desmarest', 'ORG') 1
('Baron Cuvier', 'PERSON') 1
('Frederick Cuvier', 'PERSON') 2
('Owen', 'PERSON') 4
('Scoresby', 'PERSON') 5
('Beale', 'GPE') 1
('Bennett', 'PERSON') 2
('J.  Ross Browne', 'PERSON') 1
('Miriam Coffin', 'GPE') 1
('Olmstead', 'PERSON') 1
('some seventy years', 'DATE') 1
('this present day', 'DATE') 1
('of past days', 'DATE') 1
('Charing\nCross', 'ORG') 1
('only two', 'CARDINAL') 2
('Beale', 'ORG') 4
('South-Sea', 'LOC') 2
('a minute', 'TIME') 4
('Post-Office', 'ORG') 1
('Job', 'GPE') 1
('Behold', 'PERSON') 1
('System of Nature', 'WORK_OF_ART') 1
('1776', 'DATE') 1
('the year 1850', 'DATE') 1
('Linnaeus', 'ORG') 2
('feminam mammis', 'GPE') 1
('Simeon Macey', 'PERSON') 1
('Charley Coffin', 'GPE') 2
('Charley', 'PERSON') 1
('Above,', 'ORG') 1
('Linnaeus', 'PERSON') 1
('Lamatins', 'LOC') 1
('Dugongs', 'PERSON') 1
('wet hay', 'FAC') 1
('the Kingdom of Cetology', 'GPE') 1
('III', 'ORG') 2
('OCTAVO', 'ORG') 1
('GRAMPUS', 'GPE') 2
('DUODECIMO', 'ORG') 1
('IV', 'GPE') 2
('HUMP', 'GPE') 2
('V. the RAZOR-BACK WHALE', 'PERSON') 1
('VI', 'ORG') 1
('SULPHUR', 'ORG') 1
('CHAPTER I.', 'ORG') 2
('Trumpa', 'GPE') 1
('Physeter', 'FAC') 1
('Anvil Headed', 'ORG') 1
('Cachalot', 'LOC') 1
('French', 'LANGUAGE') 2
('Germans', 'NORP') 3
('the Long Words', 'ORG') 1
('Some centuries\nago', 'DATE') 1
('Sperm', 'NORP') 1
('the Greenland Whale', 'LOC') 3
('CHAPTER II', 'LAW') 3
('the Black Whale', 'LOC') 1
('the Great Whale', 'FAC') 1
('the True Whale', 'ORG') 1
('Folios', 'PRODUCT') 1
```

```
('the Baliene Ordinaire', 'FAC') 1
('Swedes', 'NORP') 1
('two centuries', 'DATE') 2
('the Brazil Banks', 'ORG') 1
('Fin-Back', 'ORG') 1
('New York', 'GPE') 4
('The Fin-Back', 'ORG') 1
('Cain', 'PERSON') 2
('the Fin-Back', 'FAC') 1
('WHALEBONE WHALES', 'PERSON') 1
('Castle', 'ORG') 1
('CHAPTER VI', 'LAW') 1
('Tartarian', 'NORP') 1
('Prodigies', 'ORG') 1
('Adieu', 'PERSON') 1
('Sulphur Bottom', 'PERSON') 1
('BOOK I. (', 'PERSON') 1
('BOOK II', 'LAW') 1
('NARWHALE', 'ORG') 2
('THRASHER', 'ORG') 1
('V.', 'ORG') 1
('Folio', 'PERSON') 3
('Octavo', 'PERSON') 1
('fifteen to twenty-five', 'CARDINAL') 1
('the Black Fish', 'LOC') 1
('the Hyena Whale', 'ORG') 1
('Mephistophelean', 'NORP') 1
('some sixteen or', 'CARDINAL') 1
('eighteen feet', 'QUANTITY') 1
('Hyena', 'GPE') 2
('thirty gallons', 'QUANTITY') 1
('NOSTRIL', 'ORG') 1
('some sixteen feet', 'QUANTITY') 1
('five feet', 'QUANTITY') 5
('fifteen feet', 'QUANTITY') 2
('Narwhale', 'ORG') 5
('the Polar\nSea', 'GPE') 1
('Unicorn', 'ORG') 1
('Unicornism', 'NORP') 1
('Black Letter', 'ORG') 2
('Martin\nFrobisher', 'PERSON') 1
('Queen Bess', 'ORG') 1
('Greenwich\nPalace', 'FAC') 1
('Thames', 'ORG') 1
('Martin', 'PERSON') 1
('a long period', 'DATE') 1
('Irish', 'NORP') 1
('the Earl of Leicester', 'FAC') 1
('Feegee', 'ORG') 1
('Killer', 'PERSON') 2
('Exception', 'ORG') 1
('Bonapartes', 'GPE') 1
('CHAPTER V.', 'PERSON') 1
('Thrasher', 'PERSON') 1
('BOOK II', 'WORK_OF_ART') 1
('DUODECIMOES.--These', 'NORP') 1
('The Algerine Porpoise', 'LOC') 1
('Mealy', 'GPE') 1
('WHALES', 'PERSON') 1
('HUZZA', 'ORG') 1
('more than one', 'CARDINAL') 4
('Fourth', 'ORDINAL') 2
('July', 'DATE') 2
('Huzza\nPorpoise', 'PERSON') 1
('one good gallon', 'CARDINAL') 1
('ALGERINE PORPOISE).--A', 'ORG') 1
```

```
('the Huzza Porpoise', 'LOC') 2
('Porpoise', 'PERSON') 1
('Folio, Octavo', 'WORK_OF_ART') 1
('Duodecimo magnitude:--The Bottle-Nose Whale', 'PERSON') 1
('the Pudding-Headed Whale', 'ORG') 1
('the Cape Whale', 'LOC') 1
('the Cannon Whale', 'ORG') 1
('the Coppered Whale', 'FAC') 1
('Elephant Whale', 'WORK_OF_ART') 1
('the Iceberg Whale', 'ORG') 1
('the Quog Whale', 'FAC') 1
('the Blue Whale', 'ORG') 1
('Icelandic', 'GPE') 1
('Leviathanism', 'NORP') 1
('Strength', 'PERSON') 1
('Patience', 'ORG') 1
('CHAPTER 33', 'LAW') 1
('harpooneer', 'ORDINAL') 3
('Dutch Fishery', 'PERSON') 1
('Specksynder', 'ORG') 1
('supreme', 'ORG') 1
('the British Greenland Fishery', 'ORG') 1
('Specksioneer', 'PERSON') 1
('Harpooneer', 'PRODUCT') 1
('Southern', 'NORP') 5
('Mesopotamian', 'NORP') 1
('Divine Inert', 'PERSON') 1
('Nicholas the Czar', 'ORG') 1
('plebeian', 'NORP') 1
('Captain', 'ORG') 6
('Kings', 'PERSON') 1
('CHAPTER 34', 'LAW') 1
('The Cabin-Table', 'ORG') 1
('the lee quarter-boat', 'DATE') 1
('daily', 'DATE') 5
('Flask', 'PRODUCT') 2
("King Ahab's", 'PERSON') 1
('Abjectus', 'ORG') 1
('Belshazzar', 'PERSON') 5
('Caesar', 'ORG') 1
('Coronation', 'NORP') 1
('Frankfort', 'ORG') 1
('German', 'NORP') 15
('Imperial\nElectors', 'ORG') 1
("Dough-Boy's", 'PERSON') 1
("Dough-Boy's", 'ORG') 1
('African', 'NORP') 3
('Moorish', 'NORP') 1
('Missouri', 'GPE') 2
('Summer', 'DATE') 1
('CHAPTER 35', 'LAW') 1
('fifteen thousand miles', 'QUANTITY') 1
("five years'", 'DATE') 1
('Babel', 'GPE') 1
('Asia', 'LOC') 4
('Babel', 'PRODUCT') 1
('Saint Stylites', 'GPE') 1
('Napoleon', 'ORG') 3
('Vendome', 'PERSON') 1
('some one hundred', 'CARDINAL') 1
('fifty feet', 'QUANTITY') 4
('Louis Philippe', 'PERSON') 1
('Louis Blanc', 'PERSON') 1
('Louis the Devil', 'ORG') 1
('Baltimore', 'GPE') 1
('Hercules', 'GPE') 6
```

```
('Trafalgar Square', 'PERSON') 1
('Obed Macy', 'PERSON') 1
('Obed', 'PERSON') 1
('A few years ago', 'DATE') 1
('Bay', 'FAC') 1
('every two hours', 'TIME') 1
('Colossus', 'PERSON') 1
('Rhodes', 'PERSON') 1
('three years', 'DATE') 6
("a long three or four years'", 'DATE') 1
('the various hours', 'TIME') 1
('several entire months', 'DATE') 1
('Alps', 'GPE') 1
('Sleet', 'PERSON') 4
('A Voyage', 'WORK_OF_ART') 2
('Icebergs', 'LOC') 1
('the Lost Icelandic Colonies', 'ORG') 1
('Old Greenland', 'GPE') 1
('Glacier', 'LOC') 2
('Sleet', 'LOC') 1
('three or four', 'CARDINAL') 2
('Phaedon', 'ORG') 1
('Bowditch', 'GPE') 2
('Platonist', 'NORP') 1
('Childe Harold', 'PERSON') 1
('Ten thousand', 'CARDINAL') 1
('Platonists', 'NORP') 1
('Crammer', 'ORG') 1
('Pantheistic', 'ORG') 1
('Descartian', 'NORP') 1
('mid-day', 'DATE') 4
('Heed', 'PERSON') 1
('Pantheists', 'NORP') 1
('CHAPTER 36', 'LAW') 1
('one\nmorning', 'TIME') 1
("D'ye mark him", 'ORG') 1
('Twill', 'ORG') 1
('The hours', 'TIME') 1
('cried:--', 'PERSON') 1
('sixteen dollar', 'MONEY') 1
('Skin', 'PERSON') 1
('Tash', 'PERSON') 2
('annual', 'DATE') 3
('Moby Dick--Moby Dick', 'PERSON') 1
('Stubb and Flask', 'ORG') 1
('Captain Ahab, I', 'WORK_OF_ART') 1
('Aye', 'WORK_OF_ART') 1
('Good Hope', 'PERSON') 1
('Norway Maelstrom', 'GPE') 1
('Steward', 'ORG') 2
('thou requirest', 'ORG') 1
('Madness', 'GPE') 1
('pasteboard', 'ORG') 1
('Turkish', 'NORP') 2
('Pagan', 'PERSON') 1
('Chilian', 'NORP') 2
('Reckon', 'ORG') 1
('Speak', 'LOC') 1
('speak!--Aye', 'ORG') 1
('capstan', 'GPE') 7
('Satan', 'LOC') 1
('the\nyears', 'DATE') 1
('Leyden', 'PERSON') 1
('Perchance', 'ORG') 1
('some three feet', 'QUANTITY') 2
('Forthwith', 'PERSON') 1
```

('three to three', 'CARDINAL') 1
('Bestow', 'PERSON') 1
('CHAPTER 37', 'LAW') 1
('Sunset', 'GPE') 1
('AHAB SITTING ALONE', 'PERSON') 1
('Yonder', 'GPE') 2
('Tis', 'PERSON') 2
('Paradise', 'LOC') 2
('Burkes', 'PERSON') 1
('Naught', 'PRODUCT') 1
('CHAPTER 38', 'LAW') 1
('MAINMAST', 'NORP') 1
('democrat', 'NORP') 1
('FORECASTLE', 'ORG') 1
('an hour', 'TIME') 3
('CHAPTER 39', 'LAW') 1
('First Night Watch', 'ORG') 1
('Mogul', 'ORG') 5
('la!', 'GPE') 1
('skirra!', 'GPE') 1
('sir--(ASIDE', 'GPE') 1
('CHAPTER 40', 'LAW') 1
('Midnight', 'TIME') 1
('Forecastle', 'PERSON') 1
('SAILORS', 'ORG') 1
('CHORUS', 'ORG') 1
('1ST', 'CARDINAL') 1
('SINGS', 'ORG') 1
("MATE'S VOICE", 'PERSON') 2
('Eight', 'CARDINAL') 3
('2ND', 'CARDINAL') 1
('SCUTTLE', 'ORG') 2
('DUTCH', 'NORP') 1
('Amsterdam', 'GPE') 3
('Hist', 'ORG') 2
('Blanket Bay', 'LOC') 1
('Beat', 'PERSON') 1
('Jig', 'PERSON') 1
('merry', 'PERSON') 2
('gallop', 'ORG') 1
('SICILIAN', 'NORP') 2
('TAMBOURINE', 'ORG') 1
('Rig', 'PERSON') 1
('Jinglers', 'PERSON') 1
('Merry-mad', 'PERSON') 1
('Pip', 'PERSON') 10
('Split', 'PERSON') 1
('Spell oh!--whew', 'ORG') 1
('SKY', 'ORG') 1
('Ganges', 'PERSON') 1
('Seeva', 'PERSON') 1
('NUDGING', 'ORG') 1
('TAHITAN', 'ORG') 1
('MAT', 'ORG') 1
('Hail', 'PERSON') 1
('girls!--the Heeva-Heeva', 'FAC') 1
('Tahiti', 'GPE') 6
('the first\nday', 'DATE') 1
('Pirohitee', 'ORG') 1
('villages?--The', 'PERSON') 1
('PORTUGUESE', 'NORP') 1
('pell-mell', 'PERSON') 2
('Crack', 'ORG') 1
('Cattegat', 'ORG') 1
('Baltic', 'NORP') 1
('4TH', 'CARDINAL') 1

```
('ENGLISH', 'NORP') 2
('JAGO', 'PERSON') 1
('5TH', 'CARDINAL') 1
('row!', 'DATE') 1
('WHIFF', 'ORG') 1
('Humph', 'NORP') 1
('BELFAST', 'GPE') 1
('arrah a row!', 'PERSON') 1
('Fair', 'ORG') 2
('Abel', 'PERSON') 1
('Sweet', 'PERSON') 1
('Jollies', 'LOC') 1
('Crish', 'NORP') 1
('Blang', 'GPE') 1
('Jimmini', 'PERSON') 1
('anaconda', 'GPE') 2
('CHAPTER 41', 'LAW') 1
('White Whale', 'ORG') 5
('Sperm Whale', 'LOC') 4
('twelvemonth', 'DATE') 2
('Sperm Whale', 'WORK_OF_ART') 3
('the White Whale', 'ORG') 29
('a\nthousand miles', 'QUANTITY') 1
('the Sperm Whale', 'ORG') 4
('the Sperm Whale', 'LOC') 17
('North', 'LOC') 3
('Cuvier', 'ORG') 1
('Baron', 'PERSON') 2
('Povelson', 'PERSON') 1
('the earlier\ndays', 'DATE') 1
('Strello', 'PERSON') 1
('Portugal', 'GPE') 1
('Arethusa', 'GPE') 1
('Syracuse', 'GPE') 1
('the Holy Land', 'WORK_OF_ART') 1
("the White\nWhale's", 'ORG') 1
('Arkansas', 'GPE') 1
('six inch', 'QUANTITY') 1
('Venetian', 'NORP') 4
('Malay', 'LANGUAGE') 1
('one-half', 'CARDINAL') 1
('Ophites of the east', 'LOC') 1
('months of days', 'DATE') 1
('mid winter', 'DATE') 1
('Patagonian Cape', 'LOC') 1
('Hudson', 'PERSON') 1
('Northman', 'NORP') 1
('Highland', 'FAC') 3
('a\nthousand fold', 'CARDINAL') 1
('Hotel de Cluny', 'ORG') 1
('Caryatid', 'ORG') 1
('State', 'ORG') 2
('the very day', 'DATE') 1
('Gnawed', 'PERSON') 1
('Job', 'LOC') 1
('White Whale', 'WORK_OF_ART') 2
('CHAPTER 42', 'LAW') 1
('The Whiteness of The Whale', 'ORG') 1
('Pegu', 'PERSON') 1
('Siam', 'NORP') 1
('Hanoverian', 'NORP') 1
('Caesarian', 'NORP') 1
('Romans', 'NORP') 3
('the Red Men of America', 'ORG') 1
('Justice', 'ORG') 1
('Greek', 'NORP') 8
```

('Great Jove', 'PERSON') 1
('Iroquois', 'GPE') 2
('White Dog', 'FAC') 1
('Romish', 'NORP') 2
('Vision of St. John', 'ORG') 1
('four-and-twenty', 'DATE') 1
('the white bear of the poles', 'ORG') 1
('Polar', 'NORP') 5
('Polar', 'ORG') 1
('REQUIN', 'ORG') 1
('Coleridge', 'PERSON') 1
('Wondrous', 'ORG') 1
('Abraham', 'PERSON') 3
('Goney', 'PERSON') 2
('Coleridge', 'ORG') 1
('Rhyme', 'ORG') 2
('grey\nalbatrosses', 'ORG') 1
('Xerxes', 'PERSON') 1
('the Rocky Mountains', 'LOC') 1
('Alleghanies', 'ORG') 1
('Ohio', 'GPE') 3
('the White Steed', 'ORG') 1
('Albatross', 'LOC') 2
('Albino', 'ORG') 1
('Albino', 'PERSON') 1
('the Southern Seas', 'PRODUCT') 1
('Froissart', 'GPE') 1
('White Hoods of Ghent', 'ORG') 1
('snowy mantle', 'GPE') 1
('the Middle American States', 'GPE') 1
('the White\nTower', 'FAC') 1
('the Byward Tower', 'FAC') 1
('Bloody', 'ORG') 1
('the White Mountains', 'LOC') 1
('New Hampshire', 'GPE') 2
('Blue Ridge', 'GPE') 1
('the White Sea', 'LOC') 1
('the Yellow Sea', 'LOC') 1
('Central Europe', 'LOC') 1
('Hartz', 'PERSON') 1
('Blocksburg', 'GPE') 1
('Lima', 'GPE') 9
('Pizarro', 'PERSON') 1
('Second', 'ORDINAL') 4
('Vermont', 'GPE') 1
('the sunniest day', 'DATE') 1
('Oregon', 'GPE') 2
('thousands of miles', 'QUANTITY') 1
('bison herds', 'PERSON') 1
('Lapland', 'GPE') 1
('Albino', 'NORP') 1
('CHAPTER 43', 'LAW') 1
('Cabaco', 'ORG') 6
('Archy', 'PERSON') 5
('Cholo', 'ORG') 1
('Caramba', 'WORK_OF_ART') 1
('fifty miles', 'QUANTITY') 1
('Grin', 'PERSON') 1
('CHAPTER 44', 'LAW') 1
('Chart', 'ORG') 1
('this night', 'TIME') 1
('the timeliest day', 'DATE') 1
('Lieutenant Maury', 'PERSON') 1
('the National\nObservatory', 'ORG') 1
('April 16th, 1851', 'DATE') 1
('five degrees', 'QUANTITY') 2

```
('twelve', 'CARDINAL') 6
('the twelve months', 'DATE') 1
('each month', 'DATE') 1
('VEIN', 'ORG') 1
('some few miles', 'QUANTITY') 1
('this year', 'DATE') 1
('the preceding season', 'DATE') 1
('a former year', 'DATE') 1
('Seychelle', 'GPE') 1
('Volcano Bay', 'LOC') 1
('the Japanese Coast', 'LOC') 1
('season', 'DATE') 2
('several consecutive years', 'DATE') 1
('Zodiac', 'ORG') 2
('the\nmonomaniac old', 'ORG') 1
('sixty degrees', 'QUANTITY') 1
('the premature hour', 'TIME') 1
('sixty-five days', 'DATE') 1
('the Persian Gulf', 'LOC') 2
('the Bengal Bay', 'LOC') 1
('Pampas', 'PERSON') 1
('Harmattans', 'NORP') 1
('Trades', 'ORG') 1
('Mufti', 'ORG') 1
('Nay', 'ORG') 1
('CHAPTER 45', 'LAW') 1
('Affidavit', 'GPE') 1
('nearly two years', 'DATE') 1
('three-year', 'DATE') 1
('three years previous', 'DATE') 1
('Secondly', 'ORDINAL') 3
('Rinaldo\nRinaldini', 'PERSON') 1
('Cambyses or Caesar', 'ORG') 1
('Ombay', 'ORG') 1
('thou terror', 'PERSON') 1
('the Tattoo Land', 'ORG') 1
('Don Miguel', 'PERSON') 2
('Marius', 'PERSON') 1
('Sylla', 'NORP') 1
('the Narragansett Woods', 'ORG') 1
('Captain Butler', 'PERSON') 1
('Annawon', 'FAC') 1
('New Guinea', 'GPE') 3
('gallon', 'CARDINAL') 1
('at least one', 'CARDINAL') 2
('two-fold', 'CARDINAL') 1
('the Sperm\nWhale', 'GPE') 4
('the year 1820', 'DATE') 1
('Essex', 'NORP') 1
('Pollard', 'PERSON') 3
('the Pacific Ocean', 'LOC') 1
('One day', 'DATE') 2
('ten minutes', 'TIME') 2
('Owen Chace', 'PERSON') 1
('a few miles', 'QUANTITY') 1
('Chace', 'ORG') 1
('45,--he', 'GPE') 1
('ANIMAL', 'ORG') 1
('the year 1807', 'DATE') 1
('Thirdly', 'ORDINAL') 1
('Some eighteen', 'CARDINAL') 1
('twenty years ago', 'DATE') 1
('Oahu', 'GPE') 1
('Sandwich Islands', 'GPE') 1
('Some weeks', 'DATE') 1
('Commodore', 'ORG') 1
```

('Valparaiso', 'PERSON') 1
('Russian', 'NORP') 4
('Krusenstern', 'PERSON') 1
('Discovery Expedition', 'PRODUCT') 1
('the present century', 'DATE') 1
('Langsdorff', 'PERSON') 2
('seventeenth', 'ORDINAL') 1
('thirteenth', 'ORDINAL') 1
('the next\nday', 'DATE') 1
('Ochotsh', 'GPE') 1
('some days', 'DATE') 2
('three feet', 'QUANTITY') 2
("D'Wolf", 'PERSON') 2
('New Englander', 'NORP') 1
('Dorchester near', 'ORG') 1
('Boston', 'GPE') 2
('Siberian', 'NORP') 1
('Dampier', 'LOC') 1
('John Ferdinando', 'PERSON') 1
('Juan Fernandes', 'PERSON') 1
("four o'clock", 'TIME') 1
('about one hundred and fifty', 'CARDINAL') 1
('the Main of America', 'ORG') 1
('Davis', 'PERSON') 2
('early hour of the morning', 'TIME') 1
('Pusie Hall', 'PERSON') 1
('several consecutive\nminutes', 'TIME') 1
('only one', 'CARDINAL') 4
('millionth', 'ORDINAL') 1
('Solomon', 'PERSON') 8
('the sixth Christian century', 'DATE') 1
('Procopius', 'PERSON') 3
('the days', 'DATE') 4
('Justinian', 'PERSON') 1
('Belisarius', 'ORG') 1
('fifty years', 'DATE') 1
('British', 'NORP') 2
('Dardanelles', 'ORG') 2
('Propontis', 'LOC') 1
('Propontis', 'ORG') 1
('BRIT', 'ORG') 1
('half a century', 'DATE') 2
('CHAPTER 46', 'LAW') 1
('Moby', 'PERSON') 2
('Dick', 'PERSON') 2
('the White Whale', 'FAC') 5
('Crusaders', 'PRODUCT') 1
('some months', 'DATE') 1
('every minute', 'TIME') 1
('CHAPTER 47', 'LAW') 1
('Fates', 'GPE') 1
('Gay-Header', 'PERSON') 1
('Fate', 'ORG') 2
('about two miles', 'QUANTITY') 1
('one foot', 'QUANTITY') 2
('CHAPTER 48', 'LAW') 1
('The First Lowering', 'ORG') 1
('the starboard quarter', 'DATE') 1
('Chinese', 'NORP') 4
('Manillas;--a', 'PRODUCT') 1
('Fedallah', 'PERSON') 17
('fourth', 'ORDINAL') 4
('Captain Ahab?--', 'WORK_OF_ART') 1
('only five', 'CARDINAL') 1
('Pull', 'PERSON') 2
('a thousand pounds', 'QUANTITY') 1

('pull?--pull', 'ORG') 1
('a single inch', 'QUANTITY') 1
("Stubb's", 'ORG') 2
("The White Whale's", 'ORG') 1
('Elijah', 'GPE') 1
('Mississippi', 'GPE') 2
('some two feet', 'QUANTITY') 2
('Roar', 'PERSON') 1
('know;--merry', 'PERSON') 1
('Squall', 'PERSON') 1
('nearer and nearer', 'CARDINAL') 1
('CHAPTER 49', 'LAW') 1
('all the days', 'DATE') 1
('so many months or weeks', 'DATE') 1
('CHAPTER 50', 'LAW') 1
('Crew', 'PERSON') 1
('Beelzebub', 'ORG') 2
('Fedallah', 'NORP') 1
('mannerly', 'ORDINAL') 1
('Asiatic', 'NORP') 3
('these modern days', 'DATE') 1
('the moon', 'LOC') 1
('Genesis', 'ORG') 1
('Rabbins', 'PERSON') 1
('CHAPTER 51', 'LAW') 1
('Days, weeks', 'DATE') 1
('the ivory', 'GPE') 1
('Azores', 'ORG') 1
('the Cape de Verdes', 'LOC') 1
('the Carrol Ground', 'ORG') 1
('St. Helena', 'GPE') 1
('the\nmoon', 'LOC') 2
('such unusual hours', 'TIME') 1
('some\ndays', 'DATE') 1
('night after night', 'TIME') 1
('one whole day', 'DATE') 1
('every morning', 'TIME') 1
('Cape Tormentoto', 'PERSON') 1
('hours and hours', 'TIME') 2
('CHAPTER 52', 'LAW') 1
('Crozetts', 'GPE') 1
('nearly four years', 'DATE') 1
("the White Whale's", 'ORG') 3
('hailed--"Ahoy', 'ORG') 1
('voice,--"Up helm', 'PERSON') 1
('Cyclades', 'ORG') 1
('CHAPTER 53', 'LAW') 1
('five minutes', 'TIME') 2
('the Pine Barrens', 'ORG') 1
('New York State', 'GPE') 1
('Salisbury Plain', 'PERSON') 1
('Pine Barrens', 'PERSON') 1
('Salisbury Plains', 'PERSON') 1
("King's Mills", 'ORG') 1
('a date a year', 'DATE') 1
('two later', 'DATE') 1
('some third', 'CARDINAL') 1
('one language', 'TIME') 1
('Yankee', 'ORG') 1
('one day', 'DATE') 3
('ten years', 'DATE') 2
('mid-Atlantic', 'DATE') 1
('Pirates', 'GPE') 1
('Pirates', 'PRODUCT') 1
('Johnson', 'PERSON') 3
("Noah Webster's", 'PERSON') 1

```
('some fifteen thousand', 'CARDINAL') 1
('Lexicon', 'GPE') 1
('GENERALLY', 'GPE') 1
('ON A CRUISING-GROUND', 'ORG') 1
('TIME', 'ORG') 1
('ON\nBOARD', 'ORG') 1
('OTHER', 'ORG') 1
('CHAPTER 54', 'LAW') 1
("The Town-Ho's Story", 'ORG') 1
('THE GOLDEN INN', 'FAC') 1
('The Cape of Good Hope', 'WORK_OF_ART') 1
('Goney', 'PRODUCT') 1
('the Town-Ho', 'ORG') 3
('Polynesians', 'ORG') 1
('Interweaving', 'GPE') 1
('Gallipagos', 'LOC') 1
('the Golden\nInn', 'FAC') 1
('Pedro', 'PERSON') 1
('Sebastian', 'PERSON') 3
('Some two years', 'DATE') 1
('the Town-Ho, Sperm Whaler of\n', 'ORG') 1
('Golden Inn', 'WORK_OF_ART') 1
('Line', 'LOC') 1
('One morning', 'TIME') 1
('six-and-thirty', 'DATE') 1
('Radney', 'ORG') 9
('Steelkilt', 'ORG') 30
('Lakeman', 'PERSON') 22
('Buffalo', 'GPE') 3
('Lakeman', 'GPE') 1
('Don Sebastian', 'PERSON') 6
('Lake Erie', 'LOC') 1
('Don', 'PERSON') 7
('Callao', 'ORG') 1
('Manilla', 'PERSON') 5
('Ontario', 'GPE') 1
('Huron', 'GPE') 1
('Superior', 'LOC') 1
('Polynesian', 'NORP') 2
('East', 'LOC') 7
('Mackinaw', 'PRODUCT') 1
('Gothic', 'ORG') 2
('Tartar\nEmperors', 'ORG') 1
('Cleveland', 'GPE') 1
('Winnebago', 'GPE') 1
('Borean', 'NORP') 1
('Radney', 'PERSON') 7
('Nantucket beach', 'GPE') 2
('more than a day', 'DATE') 1
("the Town-Ho's", 'ORG') 1
('the lee\nscupper-holes', 'FAC') 1
('Steelkilt Charlemagne', 'ORG') 1
('Charlemagne', 'PERSON') 1
('Rad', 'PERSON') 3
('every evening', 'TIME') 1
('the Town-Ho', 'FAC') 2
('Radney', 'LOC') 3
('a few inches', 'QUANTITY') 2
('Canallers', 'ORG') 4
('Don Pedro', 'PERSON') 4
('Canallers', 'PRODUCT') 2
('Erie Canal', 'PERSON') 1
('Senor', 'PERSON') 3
('North', 'PERSON') 3
('three hundred and sixty miles', 'QUANTITY') 1
('Roman', 'ORG') 1
```

('Mohawk', 'ORG') 1
('Venetianly', 'PRODUCT') 1
("'Well", 'WORK_OF_ART') 1
('Pardon', 'GPE') 1
('Limeese', 'NORP') 1
('Venice', 'GPE') 2
('Corrupt as Lima', 'WORK_OF_ART') 1
('Dominic', 'GPE') 1
('Canaller', 'PERSON') 2
('Mark Antony', 'PERSON') 1
('Nile', 'LOC') 2
('Cleopatra', 'PERSON') 2
('Sydney', 'GPE') 1
('the Grand Canal', 'ORG') 1
('Hardly', 'ORG') 1
('about three or', 'CARDINAL') 1
('Parisians', 'NORP') 1
("'Come", 'WORK_OF_ART') 1
("'Turn", 'WORK_OF_ART') 3
('Captain', 'PERSON') 2
('Shall', 'PERSON') 3
('the hours', 'TIME') 1
('Only three', 'CARDINAL') 1
('Shut', 'PERSON') 1
('the last night', 'TIME') 1
('a few minutes', 'TIME') 3
('three quarters', 'CARDINAL') 1
("'Say", 'WORK_OF_ART') 1
("'You are a coward!'", 'WORK_OF_ART') 1
("'So I am", 'WORK_OF_ART') 1
('the Town-Ho', 'PRODUCT') 2
("two o'clock", 'TIME') 1
('the\nmorning', 'TIME') 1
('the third day', 'DATE') 1
('Shipmate', 'PERSON') 1
('HIM', 'ORG') 1
('next night', 'TIME') 1
('Twenty-four hours', 'TIME') 1
('Teneriffe', 'PERSON') 2
('Jesu', 'PERSON') 1
('St. Dominic', 'GPE') 1
('Spaniards', 'NORP') 2
('Sirs', 'ORG') 1
('fifty yards', 'QUANTITY') 1
('The White Whale', 'WORK_OF_ART') 2
('harpooneer', 'FAC') 2
('five or six', 'CARDINAL') 1
('five hundred\nmiles', 'QUANTITY') 1
('the fourth day', 'DATE') 1
('six days', 'DATE') 1
('Adios', 'NORP') 1
('Some ten days', 'DATE') 1
('Tahitians', 'NORP') 1
('Where Steelkilt', 'WORK_OF_ART') 1
('the Holy Evangelists', 'ORG') 1
('the Golden Inn', 'FAC') 1
('Nay', 'PERSON') 1
('Evangelists', 'NORP') 1
('the Holy Book before me', 'WORK_OF_ART') 1
('CHAPTER 55', 'LAW') 1
('the Monstrous Pictures of Whales', 'ORG') 1
('Hindoo', 'PERSON') 5
('Grecian', 'NORP') 2
('Saladin', 'PERSON') 1
("St.\nGeorge's", 'GPE') 1
('Brahmins', 'PERSON') 1

('the Matse Avatar', 'ORG') 1
('Galleries', 'ORG') 1
('Guido', 'PERSON') 2
('Andromeda', 'ORG') 2
('Hogarth', 'PERSON') 1
('Perseus Descending', 'WORK_OF_ART') 1
('Hogarthian', 'NORP') 1
('one inch', 'QUANTITY') 1
('Prodromus', 'NORP') 1
('Scotch Sibbald', 'PERSON') 1
('Bibles', 'PERSON') 1
('Italian', 'NORP') 4
('about the 15th century', 'DATE') 1
('Saratoga', 'NORP') 2
('Baden-Baden', 'GPE') 1
('the "Advancement of Learning', 'ORG') 1
('Harris', 'PERSON') 1
('A Whaling Voyage to Spitzbergen', 'WORK_OF_ART') 1
('Jonas', 'PERSON') 1
('Peter Peterson', 'PERSON') 1
('Friesland', 'GPE') 1
('English', 'NORP') 5
('Spermaceti Whale Fisheries', 'ORG') 1
('Picture of a Physeter or Spermaceti whale', 'WORK_OF_ART') 1
('1793', 'DATE') 1
('some five feet', 'QUANTITY') 1
("Goldsmith's Animated Nature", 'WORK_OF_ART') 1
('1807', 'DATE') 1
('nineteenth century', 'DATE') 1
('1825', 'DATE') 1
('Bernard Germain', 'PERSON') 1
('Count de Lacepede', 'PERSON') 1
('1836', 'DATE') 1
("Frederick Cuvier's", 'PERSON') 2
('Desmarest', 'PERSON') 1
('Richard III', 'PERSON') 1
('Platonian', 'NORP') 2
("Jeremy Bentham's", 'PERSON') 1
('Jeremy', 'PERSON') 1
('Hunter', 'PERSON') 1
('CHAPTER 56', 'LAW') 1
('the Less Erroneous Pictures', 'ORG') 1
('the True Pictures', 'ORG') 1
('Pliny', 'GPE') 2
('Colnett', 'PERSON') 2
('Huggins', 'PERSON') 2
('Sperm Whale', 'ORG') 4
('J. Ross Browne', 'ORG') 1
('the Right Whale', 'ORG') 3
('Scoresby', 'GPE') 1
('Garnery', 'GPE') 3
('Right Whale', 'PERSON') 1
('Versailles', 'PRODUCT') 1
('the Northern Lights', 'LOC') 1
('one tenth', 'CARDINAL') 1
('Leuwenhoeck', 'NORP') 1
('ninety-six', 'DATE') 1
('Peace', 'FAC') 1
('Leviathanic', 'GPE') 3
('Right Whale', 'ORG') 1
('CHAPTER 57', 'LAW') 1
('Teeth', 'GPE') 1
('Wood', 'GPE') 1
('Sheet-Iron', 'ORG') 1
('Stone', 'ORG') 1
('Mountains', 'ORG') 1

```
('Tower-hill', 'FAC') 1
('KEDGER', 'ORG') 1
('Sag Harbor', 'PERSON') 1
('the King of the\nCannibals', 'WORK_OF_ART') 1
('Hawaiian', 'NORP') 2
('Achilles', 'PERSON') 1
('Albert Durer', 'PERSON') 1
('Wooden', 'PERSON') 1
('South Sea', 'LOC') 5
('the Soloma Islands', 'LOC') 1
('Mendanna', 'PERSON') 1
('Figuera', 'PERSON') 1
('the Argo-Navis', 'ORG') 1
('Hydrus', 'ORG') 1
('CHAPTER 58', 'LAW') 1
('Crozetts', 'PERSON') 1
('the minute', 'TIME') 1
('the second day', 'DATE') 1
('a Sperm Whaler', 'ORG') 1
('the "Brazil Banks"', 'ORG') 1
('the Banks of Newfoundland', 'ORG') 1
('Columbus', 'GPE') 2
('tens and hundreds of thousands', 'CARDINAL') 1
('Portuguese', 'NORP') 2
('last year', 'DATE') 1
('Hebrews', 'LOC') 1
('Korah', 'PERSON') 1
('Push', 'PERSON') 1
('CHAPTER 59', 'LAW') 1
('Java', 'PERSON') 4
('The White Whale', 'ORG') 1
('twenty and thirty\nfeet', 'QUANTITY') 1
('Kraken of Bishop', 'ORG') 1
('Pontoppodan', 'PERSON') 1
('Squid', 'GPE') 1
('Bishop', 'ORG') 3
('CHAPTER 60', 'LAW') 1
('Hemp', 'ORG') 1
('Circassian', 'NORP') 1
('only two-thirds', 'CARDINAL') 1
('one\nhundred', 'CARDINAL') 1
('twenty pounds', 'QUANTITY') 1
('three tons', 'QUANTITY') 1
('two hundred', 'CARDINAL') 2
('almost an entire morning', 'TIME') 1
('nearly three feet', 'QUANTITY') 1
('one half-inch', 'QUANTITY') 1
('twenty', 'CARDINAL') 3
('half-inch', 'QUANTITY') 1
('Calais', 'PERSON') 1
('Edward', 'PERSON') 1
('Mazeppa', 'PERSON') 1
('CHAPTER 61', 'LAW') 1
('Stubb Kills', 'PERSON') 1
('Squid', 'PERSON') 1
('The next day', 'DATE') 1
('Indian Ocean', 'LOC') 1
('Rio de la Plata', 'GPE') 1
('Luff', 'PERSON') 1
('Ontario Indians', 'NORP') 1
('Paddles', 'ORG') 1
('galliot', 'PERSON') 1
('the Gay-Header', 'ORG') 1
('Koo-loo', 'PERSON') 1
('Grenadier', 'ORG') 1
('Tashtego!--give', 'GPE') 1
```

```
('Whole Atlantics and Pacifics', 'ORG') 1
('Pull up!--close to!', 'WORK_OF_ART') 1
('the\nday', 'DATE') 1
('gush', 'PERSON') 1
('gush', 'ORG') 1
('red gore', 'PERSON') 1
('CHAPTER 62', 'LAW') 1
('twenty or thirty feet', 'QUANTITY') 1
('four barrels', 'QUANTITY') 1
('CHAPTER 63', 'LAW') 1
('hurler', 'ORG') 1
('CHAPTER 64', 'LAW') 1
('eighteen', 'CARDINAL') 1
('thirty-six', 'CARDINAL') 1
('one hundred and eighty', 'CARDINAL') 1
('hour after hour', 'TIME') 1
('Hang-Ho', 'ORG') 1
('a mile', 'QUANTITY') 2
('Stubb', 'GPE') 1
('flavorish', 'NORP') 1
('About midnight', 'TIME') 1
('a few\ninches', 'TIME') 1
('Fleece', 'PERSON') 10
('Ebony', 'ORG') 1
('Belubed', 'PERSON') 1
('dan de shark', 'PERSON') 1
("bred'ren", 'PERSON') 1
('Gor', 'ORG') 1
('Massa Stubb', 'PERSON') 1
('cried--', 'ORG') 1
('Cussed', 'ORG') 1
('ninety', 'CARDINAL') 1
('Hind de hatchway', 'PERSON') 1
('de Roanoke', 'PERSON') 1
('berry joosy', 'PERSON') 1
('Cape-Down', 'LOC') 1
('Cape-Town', 'GPE') 1
('Drop', 'PERSON') 1
('Aloft', 'LOC') 1
("D'ye", 'ORG') 1
('dan Massa Shark', 'PERSON') 1
('CHAPTER 65', 'LAW') 1
('three centuries ago', 'DATE') 1
('the Right\nWhale', 'ORG') 1
('Henry', 'GPE') 1
('Dunfermline', 'LOC') 1
('one hundred feet', 'QUANTITY') 1
('Esquimaux', 'PERSON') 2
('Zogranda', 'GPE') 1
('Englishmen', 'PRODUCT') 1
('several months', 'DATE') 1
('the third month', 'DATE') 1
('Brute', 'GPE') 1
('Fejee', 'ORG') 3
('geese', 'NORP') 1
('the Society for the Suppression of\nCruelty', 'ORG') 1
('the last month', 'DATE') 1
('CHAPTER 66', 'LAW') 1
('The Shark Massacre', 'ORG') 1
('the Southern Fishery', 'LOC') 1
('late at night', 'TIME') 1
('two for an hour', 'TIME') 1
('six hours', 'TIME') 1
('twenty to thirty feet', 'QUANTITY') 1
('Ingin', 'ORG') 1
('CHAPTER 67', 'LAW') 1
```

('some one hundred pounds', 'QUANTITY') 1
('the end of the', 'DATE') 1
('Whereupon', 'ORG') 1
('CHAPTER 68', 'LAW') 1
('satin', 'PERSON') 1
('one hundred barrels', 'QUANTITY') 1
('ten barrels', 'QUANTITY') 1
('ten tons', 'QUANTITY') 1
('Upper Mississippi', 'FAC') 1
('Agassiz', 'ORG') 1
('Freeze', 'PERSON') 1
('Borneo', 'PERSON') 1
("St. Peter's", 'GPE') 1
("St. Peter's", 'PERSON') 2
('CHAPTER 69', 'LAW') 1
('SHOALS', 'ORG') 1
('ROCKS', 'ORG') 1
('CHAPTER 70', 'LAW') 1
('Sphynx', 'GPE') 1
('some eight', 'CARDINAL') 1
('ten feet', 'QUANTITY') 1
('nearly one third', 'CARDINAL') 2
('about half', 'CARDINAL') 2
('Holofernes', 'ORG') 1
('Judith', 'ORG') 1
('lotus', 'ORG') 1
('Decapitation', 'WORK_OF_ART') 1
('Speak', 'GPE') 2
('thou hast', 'PERSON') 1
('Sail ho', 'PERSON') 1
('St. Paul', 'GPE') 2
('CHAPTER 71', 'LAW') 1
('the American Whale Fleet', 'ORG') 1
('Jeroboam', 'PERSON') 8
('Mayhew', 'PERSON') 2
('Neskyeuna', 'GPE') 2
('Oceanica', 'PERSON') 1
('Captain Mayhew', 'PERSON') 3
('Shaker', 'NORP') 1
('Shakers', 'ORG') 1
('some year', 'DATE') 1
('Macey', 'GPE') 4
('Macey', 'PERSON') 1
('about fifty', 'CARDINAL') 1
('Ahab answered--"Aye', 'PERSON') 1
("blasphemer's end", 'DATE') 1
('two or three years', 'DATE') 1
('Soon Starbuck', 'PERSON') 1
('Har', 'PERSON') 1
('Harry Macey', 'PERSON') 1
('Ship Jeroboam;--why', 'PERSON') 1
('Captain Mayhew', 'WORK_OF_ART') 1
('CHAPTER 72', 'LAW') 1
('The Monkey-Rope', 'ORG') 1
('about, half', 'CARDINAL') 1
('Siamese', 'NORP') 3
('Temperance Society', 'ORG') 1
("Aunt Charity's", 'ORG') 1
('CHAPTER 73', 'LAW') 1
("a Sperm Whale's", 'ORG') 1
('the past night', 'TIME') 1
('Right Whales', 'PERSON') 1
('Crozetts', 'ORG') 1
('a Sperm Whale', 'LOC') 1
('Cut', 'WORK_OF_ART') 2
('only a few feet', 'QUANTITY') 1

("the Sperm Whale's", 'ORG') 12
('Israelites', 'NORP') 1
('a dark night', 'TIME') 1
('hands--"Aye', 'GPE') 1
('John', 'PERSON') 2
("mad,--'I", 'PERSON') 1
("'Take him,'", 'WORK_OF_ART') 1
('Adventures', 'PERSON') 1
('Locke', 'PERSON') 1
('Kant', 'PERSON') 1
('Parsee', 'GPE') 10
('CHAPTER 74', 'LAW') 1
("The Sperm Whale's", 'ORG') 1
("the Sperm\nWhale's", 'GPE') 2
('some thirty degrees', 'QUANTITY') 1
('about thirty', 'CARDINAL') 1
('Euclid', 'GPE') 1
('Herschel', 'ORG') 1
('Kentucky', 'GPE') 1
('some fifteen feet', 'QUANTITY') 1
('some few days', 'DATE') 2
('Michigan', 'GPE') 1
('forty-two', 'CARDINAL') 1
('CHAPTER 75', 'LAW') 1
("the Right\nWhale's", 'FAC') 1
("Sperm Whale's", 'ORG') 2
('Two hundred\nyears ago', 'DATE') 1
('F', 'ORG') 1
('carpenter', 'PERSON') 1
('about twenty feet', 'QUANTITY') 1
('some 500 gallons', 'QUANTITY') 1
('Peruvian', 'NORP') 3
('Mackinaw', 'ORG') 1
('twelve feet', 'QUANTITY') 1
('Purchas', 'ORG') 1
('Hackluyt', 'GPE') 1
('about two hundred and fifty', 'CARDINAL') 1
("Queen Anne's", 'GPE') 1
('Haarlem', 'PERSON') 1
('Turkey', 'GPE') 1
('Stoic', 'NORP') 1
('Spinoza', 'GPE') 1
('CHAPTER 76', 'LAW') 1
('The Battering-Ram', 'ORG') 1
("the Sperm Whale's", 'GPE') 1
('twenty feet', 'QUANTITY') 3
('the Isthmus of Darien', 'FAC') 1
('Truth', 'GPE') 1
('Lais', 'GPE') 1
('CHAPTER 77', 'LAW') 1
('The Great Heidelburgh Tun', 'ORG') 1
('the Baling of the Case', 'WORK_OF_ART') 1
('Euclidean', 'PERSON') 1
('Heidelburgh Tun of the Sperm Whale', 'WORK_OF_ART') 1
('Heidelburgh', 'GPE') 1
('Rhenish', 'NORP') 1
('about five hundred gallons', 'QUANTITY') 1
('the Heidelburgh Tun', 'FAC') 2
('the Heidelburgh Tun of the Sperm Whale', 'FAC') 1
('one third', 'CARDINAL') 1
('more than twenty-six', 'CARDINAL') 1
('the\nspermaceti', 'ORG') 1
('that quarter', 'DATE') 1
('Heidelburgh Tun', 'PERSON') 1
('CHAPTER 78', 'LAW') 1
('Turkish Muezzin', 'ORG') 1

('Tun', 'PRODUCT') 3
('oozy', 'GPE') 1
('eightieth', 'ORDINAL') 1
('Tun of Heidelburgh', 'WORK_OF_ART') 1
('Avast', 'PERSON') 1
('Table-Rock', 'LOC') 1
('both!--it', 'PERSON') 1
("the Gay-Header's", 'ORG') 1
('Only one', 'CARDINAL') 1
('Plato', 'ORG') 2
('CHAPTER 79', 'LAW') 1
('Physiognomist', 'PERSON') 1
('Gibraltar', 'GPE') 1
('Gall', 'PERSON') 2
('Dome', 'CARDINAL') 1
('Pantheon', 'ORG') 1
('Lavater', 'PERSON') 2
('Spurzheim', 'GPE') 1
('Phidias', 'ORG') 1
('Shakespeare', 'PERSON') 1
('Melancthon', 'GPE') 1
('Orient World', 'ORG') 1
('May-day', 'DATE') 1
('William Jones', 'PERSON') 1
('Ishmael', 'NORP') 1
('Chaldee', 'NORP') 2
('CHAPTER 80', 'LAW') 1
('Sphinx', 'FAC') 1
('at least twenty', 'CARDINAL') 2
('CHAPTER 81', 'LAW') 1
('The Pequod Meets The Virgin', 'ORG') 1
('Jungfrau', 'PERSON') 4
('Derick De Deer', 'PERSON') 2
('Bremen', 'GPE') 1
('Yarman', 'ORG') 1
('Yarman', 'PRODUCT') 1
('Newcastle', 'GPE') 1
('Jungfrau', 'GPE') 1
('Derick', 'PERSON') 10
('Aware', 'ORG') 1
('half an acre', 'QUANTITY') 1
('yaw\n', 'PERSON') 1
('Hindostan', 'GPE') 1
('Yarman', 'PERSON') 3
('Halloo', 'PERSON') 1
('DO', 'ORG') 1
('spring,--he', 'ORG') 1
('Yarman', 'GPE') 1
('three thousand dollars', 'MONEY') 1
('fifty thousand', 'CARDINAL') 1
('two-and-twenty', 'DATE') 1
('Gayhead', 'PERSON') 1
('Yarman', 'NORP') 1
('Sail', 'PERSON') 2
('portcullis\njaw', 'PERSON') 1
('Blinding', 'GPE') 1
("St. Bernard's", 'GPE') 1
('Davy Jones', 'PERSON') 1
('less than 2000', 'CARDINAL') 1
('eight inches', 'QUANTITY') 1
('eight day', 'DATE') 1
('Suspended', 'PERSON') 1
('Xerxes', 'NORP') 1
('West Indian', 'NORP') 1
('Sperm\nWhale', 'WORK_OF_ART') 1
('Right Whale', 'WORK_OF_ART') 1

("the Fin-Back's", 'FAC') 1
('Virgin', 'PERSON') 1
('keels', 'PERSON') 1
('the Fin-Backs', 'ORG') 1
('CHAPTER 82', 'LAW') 1
('Glory of Whaling', 'WORK_OF_ART') 1
('the knightly days', 'DATE') 1
('Perseus', 'GPE') 2
('the\nsea-coast', 'LOC') 1
('Arkite', 'PERSON') 1
('Italy', 'GPE') 1
('St. George', 'GPE') 6
('Ezekiel', 'ORG') 1
('George', 'PERSON') 1
('Philistines', 'GPE') 1
('Dagon', 'PERSON') 1
('Israel', 'GPE') 1
("St. George's", 'GPE') 1
('Crockett', 'PERSON') 1
('Kit Carson', 'PERSON') 1
('Shaster', 'PERSON') 2
('Hindoos', 'GPE') 1
('Vishnoo', 'PERSON') 5
('Vedas', 'PERSON') 2
('CHAPTER 83', 'LAW') 1
('Jonah Historically Regarded', 'PERSON') 1
('Sag-Harbor', 'LOC') 1
('this:--He', 'WORK_OF_ART') 1
("Bishop Jebb's", 'PERSON') 1
("the Right Whale's", 'ORG') 1
('Sag-Harbor', 'ORG') 1
('The\nWhale', 'WORK_OF_ART') 1
('Eagle', 'ORG') 1
('the Mediterranean Sea', 'LOC') 1
("three days'", 'DATE') 1
('Tigris', 'ORG') 2
('more than three', 'CARDINAL') 1
('Nineveh', 'PERSON') 3
('the Cape of Good Hope', 'LOC') 4
('Red Sea', 'LOC') 1
('Bartholomew Diaz', 'PERSON') 1
('Sag-Harbor', 'PERSON') 1
('Catholic', 'NORP') 1
('Turks', 'NORP') 3
('some three centuries ago', 'DATE') 1
("Harris's Voyages", 'ORG') 1
('Turkish Mosque', 'ORG') 1
('Mosque', 'GPE') 1
('CHAPTER 84', 'LAW') 1
('Actium', 'ORG') 1
('jerking boat', 'PERSON') 1
('ten or twelve feet', 'QUANTITY') 1
('waistband', 'GPE') 1
('today', 'DATE') 1
('Orleans', 'LOC') 1
('Monongahela', 'GPE') 1
('CHAPTER 85', 'LAW') 1
('Fountain', 'GPE') 1
('six thousand years', 'DATE') 1
('some centuries', 'DATE') 1
('fifteen and a\n', 'DATE') 1
("quarter minutes past one o'clock P.M.", 'TIME') 1
('this sixteenth day', 'DATE') 1
('1851', 'DATE') 1
('at least eight feet', 'QUANTITY') 1
('full hour', 'TIME') 1

```
('eleven minutes', 'TIME') 1
('Remark', 'ORG') 1
('about one seventh', 'CARDINAL') 1
('Erie Canal', 'ORG') 1
('Pyrrho', 'GPE') 1
('Devil', 'PERSON') 1
('Dante', 'PERSON') 1
('CHAPTER 86', 'LAW') 1
('Tail', 'ORG') 1
('at least fifty', 'CARDINAL') 1
('Roman', 'PERSON') 1
('Eckerman', 'LANGUAGE') 1
('Angelo', 'PERSON') 1
('Son', 'PERSON') 1
('Five', 'CARDINAL') 1
('Fourth', 'GPE') 1
('Fifth', 'ORDINAL') 2
('Darmonodes', 'ORG') 1
('at least thirty feet', 'QUANTITY') 1
('BREACH', 'ORG') 1
('Satan', 'GPE') 1
('Baltic of Hell', 'LOC') 1
('Dantean', 'PRODUCT') 1
('Isaiah', 'PERSON') 1
('Persia', 'GPE') 1
('Ptolemy Philopater', 'PERSON') 1
('King Juba', 'PERSON') 1
('Free-Mason', 'ORG') 1
('CHAPTER 87', 'LAW') 1
('Malacca', 'GPE') 1
('Birmah', 'ORG') 1
('Sumatra, Java,', 'ORG') 1
('Bally', 'ORG') 1
('Timor', 'GPE') 1
('sally', 'PERSON') 1
('Sunda', 'GPE') 6
('Malacca', 'PERSON') 1
('Sumatra', 'PERSON') 2
('Java Head', 'PERSON') 2
('Straits', 'LOC') 3
('Baltic', 'LOC') 1
('Danes', 'NORP') 1
('centuries', 'DATE') 2
('Sumatra', 'ORG') 1
('Javan', 'LOC') 1
('Philippine', 'NORP') 1
('three years afloat', 'DATE') 1
('answer--"Well', 'ORG') 1
('the Sperm\nWhales', 'GPE') 1
('thousands on thousands', 'CARDINAL') 1
('some two or three', 'CARDINAL') 1
('one half', 'CARDINAL') 1
('noon-day', 'DATE') 1
('bush', 'PERSON') 1
('gaunt', 'PERSON') 1
('Cockatoo Point', 'PRODUCT') 1
('Porus', 'PERSON') 2
('Alexander', 'ORG') 2
('tens of thousands', 'CARDINAL') 2
('Witness', 'PERSON') 1
("about\nthree minutes' time", 'TIME') 1
('Nantucket Indians', 'NORP') 1
('satin-like', 'PERSON') 1
('Titanic', 'ORG') 1
('at least two', 'CARDINAL') 1
('three square miles', 'QUANTITY') 1
```

('Gulfweed', 'PERSON') 1
('some fourteen feet', 'QUANTITY') 1
('some six feet', 'QUANTITY') 1
('the final spring', 'DATE') 1
('Tartar', 'ORG') 1
('Madame Leviathan', 'GPE') 1
('all seasons', 'DATE') 1
('nine months', 'DATE') 1
('Esau', 'LOC') 1
('Jacob:--a', 'ORG') 1
('desperado Arnold', 'PRODUCT') 1
('Hudson', 'LOC') 1
('between two', 'CARDINAL') 2
('CHAPTER 88', 'LAW') 1
('Sperm\nWhales', 'GPE') 1
('twenty to fifty', 'CARDINAL') 1
('Ottoman', 'NORP') 4
('more than one-third', 'CARDINAL') 1
('half a dozen yards', 'QUANTITY') 1
('EMBONPOINT', 'ORG') 1
('the summer', 'DATE') 1
('Equator', 'ORG') 1
('Bashaw', 'PERSON') 2
('Lothario', 'GPE') 1
('Grand Turks', 'PERSON') 1
('Turk', 'PERSON') 1
('Vidocq', 'GPE') 1
('Frenchman', 'NORP') 6
('Sperm\nWhales', 'PERSON') 1
('Almost', 'WORK_OF_ART') 1
('Daniel\nBoone', 'ORG') 1
('forty-barrel', 'QUANTITY') 1
('Forty-barrel', 'QUANTITY') 1
('Yale', 'ORG') 1
('about three-fourths', 'CARDINAL') 1
('Forty-barrel-bull', 'QUANTITY') 1
('CHAPTER 89', 'LAW') 1
('Loose-Fish', 'ORG') 4
('Holland', 'GPE') 4
('States', 'GPE') 1
('A.D. 1695', 'DATE') 1
("Justinian's Pandects", 'ORG') 1
('the Chinese Society for the Suppression of Meddling', 'ORG') 1
("People's Business", 'ORG') 1
("Queen Anne's", 'PERSON') 1
('nine-inch', 'QUANTITY') 1
('Coke', 'ORG') 1
('Some fifty years ago', 'DATE') 1
('Northern', 'LOC') 1
('Erskine', 'PERSON') 3
('Erskine', 'NORP') 1
('con', 'ORG') 1
('the Temple of the Law', 'ORG') 1
('Temple of the Philistines', 'WORK_OF_ART') 1
('Possession', 'PRODUCT') 1
('Republican', 'NORP') 1
('Mordecai', 'ORG') 1
('Woebegone', 'PERSON') 1
('Savesoul', 'PERSON') 1
('hundreds of thousands', 'CARDINAL') 1
('Savesoul', 'GPE') 1
('John Bull', 'PERSON') 1
('Ireland', 'GPE') 1
('Jonathan', 'PERSON') 1
('1492', 'DATE') 1
('Greece', 'GPE') 2

('the United\nStates', 'GPE') 1
('CHAPTER 90', 'LAW') 1
('De balena', 'ORG') 1
('rex habeat caput', 'ORG') 1
('BRACTON', 'ORG') 1
('the Laws of England', 'WORK_OF_ART') 1
('Honourary Grand Harpooneer', 'PERSON') 1
('the last two years', 'DATE') 1
('Dover', 'GPE') 1
('Sandwich', 'GPE') 1
('the Cinque Ports', 'ORG') 2
('Warden', 'GPE') 2
('the Cinque Port', 'ORG') 1
('Blackstone', 'ORG') 2
('Duke', 'ORG') 2
('a quarter or a half', 'DATE') 1
('Wellington', 'GPE') 1
('Sovereign', 'PRODUCT') 1
('Sovereign', 'LOC') 1
('Plowdon', 'PERSON') 2
('the King and Queen', 'ORG') 1
('Queen', 'PERSON') 3
('Queen-Gold', 'WORK_OF_ART') 1
("King's", 'ORG') 1
('Bench', 'FAC') 1
('William Prynne', 'PERSON') 1
('Prynne', 'PERSON') 1
('CHAPTER 91', 'LAW') 1
('The Pequod Meets The Rose-Bud', 'ORG') 1
('Ambergriese', 'NORP') 1
('T. BROWNE', 'PERSON') 1
('V.E.', 'PERSON') 1
('week', 'DATE') 1
('the other day', 'DATE') 1
('Assyrian', 'NORP') 1
('Frenchmen', 'GPE') 1
('Bouton\nde Rose,"--Rose-button', 'PERSON') 1
('BOUTON', 'ORG') 1
('ROSE', 'ORG') 1
('bawled--"Bouton-de-Rose', 'ORG') 1
('Bouton-de-Roses', 'PERSON') 1
('Guernsey', 'PERSON') 11
('Bouton-de-Rose-bud', 'ORG') 1
('The WHITE Whale', 'WORK_OF_ART') 1
('Cachalot Blanche', 'ORG') 1
('Bouton-de-Rose', 'PERSON') 1
('Guernseyman', 'PRODUCT') 1
('CABINET', 'ORG') 1
('Captain', 'PRODUCT') 2
('Monsieur', 'PERSON') 4
('only yesterday', 'DATE') 1
('St. Jago', 'GPE') 1
('Monsieur', 'ORG') 3
('Bordeaux', 'GPE') 1
('Some six', 'CARDINAL') 1
('CHAPTER 92', 'LAW') 1
('Ambergris', 'ORG') 1
('1791', 'DATE') 1
('the English', 'ORG') 1
('House of Commons', 'ORG') 1
('late day', 'DATE') 1
('grey amber', 'ORG') 1
('Mecca', 'GPE') 1
('Brandreth', 'PERSON') 1
('Corinthians', 'GPE') 1
('Paracelsus', 'PERSON') 1

('more than two centuries ago', 'DATE') 1
('Schmerenburgh', 'FAC') 1
('Smeerenberg', 'PERSON') 1
('Fogo Von Slack', 'PERSON') 1
('Smells', 'ORG') 1
('berg', 'PERSON') 1
('fifty days', 'DATE') 1
('Jew', 'NORP') 1
('CHAPTER 93', 'LAW') 1
('Castaway', 'ORG') 1
('Pippin', 'PERSON') 1
('nick-name', 'PERSON') 1
('Pip and Dough-Boy', 'ORG') 1
('year', 'DATE') 1
('three hundred and sixty-five', 'CARDINAL') 1
('Julys', 'GPE') 1
("New Year's", 'EVENT') 1
('Tolland County', 'GPE') 1
('Connecticut', 'GPE') 1
('crystal', 'LOC') 1
('the King of Hell', 'WORK_OF_ART') 1
('less\nthan half', 'DATE') 1
('BOAT', 'ORG') 1
('Stick', 'FAC') 1
('Wisdom', 'ORG') 1
('CHAPTER 94', 'LAW') 1
('Hand', 'LOC') 1
('Constantine', 'ORG') 1
('only a few minutes', 'TIME') 1
('the hour', 'TIME') 2
('Paracelsan', 'ORG') 1
('Plum', 'ORG') 1
('citron', 'GPE') 1
('Louis', 'GPE') 1
('Gros', 'PERSON') 1
('first day', 'DATE') 1
('vineyards', 'LOC') 1
('Champagne', 'ORG') 1
('stringy affair', 'PERSON') 1
('Gurry', 'PERSON') 1
('Nippers', 'ORG') 1
('pairs,--a', 'GPE') 1
('CHAPTER 95', 'LAW') 1
('Cassock', 'ORG') 1
('Kentuckian', 'NORP') 2
('Queen Maachah', 'PERSON') 1
('Judea', 'ORG') 1
('King Asa', 'ORG') 1
('Kedron', 'ORG') 1
('15th', 'ORDINAL') 1
('the First Book of\nKings', 'LAW') 1
('CHAPTER 96', 'LAW') 1
('The Try-Works', 'ORG') 1
('eight square', 'QUANTITY') 1
("about nine o'clock at night", 'TIME') 1
('Hydriote', 'ORG') 1
('Canaris', 'ORG') 1
('Tartarean', 'PERSON') 1
('long hours', 'TIME') 1
('Uppermost', 'WORK_OF_ART') 1
('I. Lo', 'PERSON') 1
('Dismal Swamp', 'ORG') 1
('Campagna', 'PERSON') 1
('Sahara', 'LOC') 1
('millions of miles', 'QUANTITY') 1
('Sorrows', 'PERSON') 1

('Cowper, Young', 'ORG') 1
('Pascal', 'PERSON') 1
('Rousseau', 'PERSON') 1
('Rabelais', 'ORG') 1
('I.E.', 'ORG') 1
('Catskill', 'LOC') 1
('CHAPTER 97', 'LAW') 1
('Aladdin', 'ORG') 1
('CHAPTER 98', 'LAW') 1
('Stowing Down', 'PERSON') 1
('Shadrach', 'DATE') 1
('Abednego', 'PERSON') 1
('six-barrel', 'QUANTITY') 1
('no night', 'TIME') 1
('ninety-six hours', 'TIME') 1
('Pythagoras', 'NORP') 1
('thousand years ago', 'DATE') 1
('CHAPTER 99', 'LAW') 1
('every hour', 'TIME') 2
('South America', 'LOC') 1
('Spanishly', 'GPE') 1
('QUITO', 'ORG') 1
('Libra', 'ORG') 2
('Lucifer', 'PERSON') 1
('six months', 'DATE') 2
('Aries', 'ORG') 2
('Trinity', 'ORG') 1
('Righteousness', 'ORG') 1
('nine', 'CARDINAL') 1
('Negro Hill', 'FAC') 1
("Corlaer's Hook", 'FAC') 1
('Humph', 'PRODUCT') 1
('Popayan', 'NORP') 1
('Golconda', 'ORG') 1
('Bowditch', 'PERSON') 1
('Daboll', 'ORG') 2
('Hem', 'LOC') 1
('Taurus', 'PRODUCT') 2
('Bull', 'PERSON') 2
('Jimimi', 'GPE') 1
('Gemini', 'PERSON') 1
('Twins', 'ORG') 2
('Gemini', 'GPE') 1
('Crab', 'PRODUCT') 1
('Virtue', 'GPE') 1
('Lion', 'PERSON') 1
('dabs', 'NORP') 1
('Virgo', 'ORG') 1
('Virgin', 'GPE') 1
('Scales', 'EVENT') 1
('Scorpio', 'PRODUCT') 1
('Scorpion', 'LOC') 1
('Sagittarius', 'ORG') 2
('Archer', 'PERSON') 2
('Capricornus', 'LOC') 1
('Aquarius', 'PERSON') 1
('Fishes', 'PRODUCT') 1
('Adieu', 'PRODUCT') 1
('sixteen dollars', 'MONEY') 1
('two cents', 'MONEY') 1
('nine hundred and sixty', 'CARDINAL') 2
('Prick', 'ORG') 1
('a month', 'DATE') 1
('two score years ago', 'DATE') 1
('Copenhagen', 'ORG') 1
('Dodge', 'ORG') 1

```
('Cannibal', 'ORG') 1
('Surgeon', 'PERSON') 1
('Ho!', 'GPE') 1
('Murray', 'PERSON') 1
('Grammar', 'PERSON') 1
('caw', 'PERSON') 4
('Tolland county', 'GPE') 1
('Hish', 'NORP') 1
('hish!  ', 'PERSON') 1
('Jenny', 'PERSON') 3
('CHAPTER 100', 'LAW') 1
('Meets', 'NORP') 1
('Samuel Enderby', 'PERSON') 4
('Trumpet', 'LOC') 1
('less than a minute', 'TIME') 1
('Jump', 'PERSON') 1
('capstan', 'PERSON') 1
('Englishman', 'PERSON') 3
('last season', 'DATE') 1
('Englishman', 'NORP') 1
('four or', 'CARDINAL') 1
('Mounttop', 'PERSON') 3
('--Mounttop', 'PERSON') 1
('midday', 'TIME') 1
('Bunger', 'PERSON') 2
('Bunger', 'PRODUCT') 8
('Boomer', 'PERSON') 3
('Sammy', 'PERSON') 1
("about three o'clock", 'TIME') 1
('Jack Bunger', 'PERSON') 1
('more than two feet', 'QUANTITY') 1
('Englishmen', 'FAC') 1
('Twice', 'WORK_OF_ART') 1
('Divine Providence', 'PERSON') 1
('Ceylon', 'GPE') 1
('lancet', 'ORG') 1
('CHAPTER 101', 'LAW') 1
('Enderby & Sons', 'ORG') 1
('that year (1775', 'DATE') 1
('some score of years previous', 'DATE') 1
('1726', 'DATE') 1
('Coffins', 'PERSON') 1
('Maceys of Nantucket', 'ORG') 1
('Amelia', 'GPE') 2
('Enderbys', 'GPE') 1
('Cape\nHorn', 'LOC') 1
('the South Sea', 'LOC') 1
('Commanded', 'ORG') 1
('Post-Captain', 'ORG') 1
('Rattler', 'PERSON') 1
('1819', 'DATE') 1
('Enderbies', 'ORG') 1
('trumps', 'PERSON') 1
('Saxon', 'PERSON') 1
('Flip', 'PERSON') 1
('ten gallons', 'QUANTITY') 1
('the Samuel Enderby', 'PERSON') 1
('Hollanders', 'ORG') 1
('Zealanders', 'NORP') 1
('Danes', 'ORG') 1
('Dan Coopman', 'PERSON') 1
('Fitz Swackhammer', 'ORG') 1
('Snodhead', 'PERSON') 3
('Santa Claus', 'ORG') 1
("St. Pott's", 'GPE') 1
('The Cooper', 'WORK_OF_ART') 1
```

('The Merchant', 'WORK_OF_ART') 1
('Smeer', 'PRODUCT') 1
('Fat', 'WORK_OF_ART') 1
('180', 'CARDINAL') 2
('400,000', 'CARDINAL') 1
('60,000', 'CARDINAL') 1
('150,000', 'CARDINAL') 1
('550,000', 'CARDINAL') 1
('72,000', 'CARDINAL') 1
('2,800', 'CARDINAL') 1
('20,000', 'CARDINAL') 1
('Texel & Leyden', 'ORG') 1
('144,000', 'CARDINAL') 1
('550', 'CARDINAL') 2
('Geneva', 'GPE') 1
('10,800 barrels', 'QUANTITY') 2
('Platonic', 'ORG') 1
('Spitzbergen', 'ORG') 1
('Texel', 'PERSON') 1
('Leyden', 'GPE') 1
('Polar Seas', 'PERSON') 1
('the short summer', 'DATE') 1
('three months', 'DATE') 1
('30', 'CARDINAL') 1
('5,400', 'CARDINAL') 1
('two barrels', 'QUANTITY') 2
("twelve weeks'", 'DATE') 1
('Equator', 'PERSON') 1
('two or three centuries ago', 'DATE') 1
('CHAPTER 102', 'LAW') 1
('Hitherto', 'PERSON') 1
('Cetacea', 'ORG') 1
('Tranquo', 'PERSON') 3
('Tranque', 'ORG') 5
('Dey of Algiers', 'ORG') 1
('Arsacidean', 'NORP') 4
('Pupella', 'ORG') 3
('Bamboo-Town', 'ORG') 1
('every month', 'DATE') 1
("Dar'st", 'CARDINAL') 1
('a Leviathanic Museum', 'ORG') 1
('Hull', 'GPE') 1
('Manchester', 'GPE') 1
('River Whale', 'LOC') 1
('the United States', 'GPE') 1
('Yorkshire', 'GPE') 1
('Burton Constable', 'PERSON') 1
('Clifford Constable', 'PERSON') 1
('Clifford', 'PERSON') 2
('Clifford', 'ORG') 1
('CHAPTER', 'ORG') 3
('Skeleton', 'GPE') 1
('seventy tons', 'QUANTITY') 1
('sixty feet', 'QUANTITY') 1
('between eighty-five', 'CARDINAL') 1
('less than forty feet', 'QUANTITY') 1
('at least ninety tons', 'QUANTITY') 1
('thirteen', 'CARDINAL') 1
('one thousand one hundred', 'CARDINAL') 1
('seventy-two', 'CARDINAL') 2
('ninety feet', 'QUANTITY') 1
('about one\nfifth', 'CARDINAL') 1
('some twenty feet', 'QUANTITY') 1
('less than a third', 'CARDINAL') 1
('nearly six feet', 'QUANTITY') 1
('eight feet', 'QUANTITY') 1

('at least sixteen feet', 'QUANTITY') 1
('Pompey', 'PERSON') 1
('Pillar', 'PRODUCT') 1
('forty and odd', 'CARDINAL') 1
('less than three feet', 'QUANTITY') 1
('more than four', 'CARDINAL') 1
('only two inches', 'QUANTITY') 1
('CHAPTER 104', 'LAW') 1
('The Fossil Whale', 'PERSON') 1
('Vesuvius', 'PERSON') 1
('Friends', 'ORG') 1
('panoramas', 'NORP') 1
('Fossil Whales', 'PERSON') 2
('thirty years past', 'DATE') 1
('Lombardy', 'GPE') 1
('Scotland', 'GPE') 1
('the States of Louisiana', 'GPE') 1
('the year 1779', 'DATE') 1
('the Rue\nDauphine', 'FAC') 1
('Paris', 'GPE') 1
('Tuileries', 'ORG') 2
('Antwerp', 'GPE') 1
('Cetacean', 'GPE') 1
('the year 1842', 'DATE') 1
('Creagh', 'PERSON') 1
('Basilosaurus', 'GPE') 1
('the English Anatomist', 'ORG') 1
('Zeuglodon', 'GPE') 1
('London Geological Society', 'ORG') 1
('Saturn', 'PERSON') 1
('grey chaos', 'ORG') 1
('Tropics', 'ORG') 1
('the 25,000 miles', 'QUANTITY') 1
('Himmalehs', 'NORP') 2
('Pharaoh', 'PERSON') 2
('Methuselah', 'PERSON') 1
('marl', 'PERSON') 1
('Denderah', 'ORG') 1
('some fifty years ago', 'DATE') 1
('John Leo', 'PERSON') 2
('Barbary', 'PERSON') 1
('Rafters', 'ORG') 1
('Beams', 'PRODUCT') 1
('Whale-Bones', 'ORG') 1
('Rocks', 'ORG') 1
('Miles', 'PERSON') 1
('Sea', 'LOC') 1
('Ground', 'ORG') 1
('Arch', 'PERSON') 1
('Camel', 'ORG') 1
('a hundred Years', 'DATE') 1
('this Afric Temple of the Whale', 'FAC') 1
('CHAPTER 105', 'LAW') 1
('Eternities', 'ORG') 1
('Tertiary', 'GPE') 2
('seventy feet', 'QUANTITY') 1
('seventy-two feet', 'QUANTITY') 1
('a hundred feet', 'QUANTITY') 1
('the present hour', 'TIME') 1
('Whales', 'GPE') 1
('Aldrovandus', 'PERSON') 1
('eight hundred feet', 'QUANTITY') 1
('Rope Walks', 'ORG') 1
('the\ndays', 'DATE') 1
('Banks', 'GPE') 1
('the Academy of Sciences', 'ORG') 1

('Iceland Whales', 'PERSON') 1
('one hundred and twenty yards', 'CARDINAL') 1
('three hundred and sixty feet', 'QUANTITY') 1
('Lacepede', 'PRODUCT') 1
('3', 'CARDINAL') 1
('one hundred\nmetres', 'CARDINAL') 1
('three hundred and twenty-eight feet', 'CARDINAL') 1
('thousands of years', 'DATE') 2
('Smithfield', 'ORG') 1
('Behring', 'GPE') 1
('Leviathan\ncan', 'PERSON') 1
('forty years ago', 'DATE') 1
('a\nperiod ago', 'DATE') 1
('the\npresent day', 'DATE') 1
('Forty', 'DATE') 1
('the Sperm Whales', 'PERSON') 1
('forty-eight months', 'DATE') 1
('forty', 'DATE') 1
('Canadian', 'NORP') 1
('forty\nthousand', 'CARDINAL') 1
('the last century', 'DATE') 1
('Swiss', 'NORP') 1
('less than 13,000', 'CARDINAL') 1
('annually', 'DATE') 1
('Harto', 'GPE') 1
('Goa', 'PERSON') 1
('the King of Siam', 'WORK_OF_ART') 1
('4,000', 'CARDINAL') 1
('Semiramis', 'ORG') 1
('Hannibal', 'ORG') 1
('Americas', 'LOC') 1
('New Holland', 'GPE') 1
('the age of a century', 'DATE') 1
('seventy-five years ago', 'DATE') 1
('Windsor Castle', 'PERSON') 1
('Kremlin', 'ORG') 1
('Netherlands', 'GPE') 1
('CHAPTER 106', 'LAW') 1
("Ahab's Leg", 'PERSON') 1
('Grief', 'NORP') 1
('Joy', 'PERSON') 1
('jaw-ivory', 'PERSON') 1
('CHAPTER 107', 'LAW') 1
('Carpenter', 'PERSON') 4
('Saturn', 'PRODUCT') 1
('bull', 'ORG') 1
('MULTUM', 'PERSON') 1
('PARVO', 'GPE') 1
('automaton', 'GPE') 1
('some sixty years', 'DATE') 1
('CHAPTER 108', 'LAW') 1
('BUSILY', 'GPE') 1
('IVORY', 'NORP') 1
('LEG', 'ORG') 1
('IVORY', 'GPE') 1
('LEATHER STRAPS', 'ORG') 1
('PADS', 'ORG') 1
('BENCH', 'ORG') 1
('FORWARD', 'ORG') 1
('BLACKSMITH', 'ORG') 1
('Drat', 'NORP') 1
('drat', 'NORP') 1
('SNEEZES', 'ORG') 2
('SNEEZES)--why', 'PERSON') 1
('SNEEZES)--yes', 'PERSON') 1
('Smut', 'PERSON') 1

('Mogulship', 'PERSON') 1
('TIMES', 'ORG') 1
('Prometheus', 'GPE') 1
('Africans', 'NORP') 1
('Prometheus', 'PERSON') 1
('Imprimis', 'ORG') 1
('about a quarter', 'CARDINAL') 1
('ho!', 'GPE') 1
('no;--a', 'CARDINAL') 1
('thou', 'ORDINAL') 4
('most solitary hours', 'TIME') 1
('Bungle', 'ORG') 1
('Praetorians', 'NORP') 1
('about one', 'CARDINAL') 1
('CHAPTER 109', 'LAW') 1
('Cabin', 'PRODUCT') 1
('Formosa', 'GPE') 1
('the Bashee Isles', 'FAC') 1
('Matsmai', 'PERSON') 1
('Sikoke', 'PERSON') 1
('new ivory', 'GPE') 1
('Up Burtons', 'WORK_OF_ART') 1
('a week', 'DATE') 1
('tinker', 'ORG') 1
('a year', 'DATE') 1
('twenty thousand miles', 'QUANTITY') 1
('Typhoons', 'ORG') 2
('keel.--On', 'ORG') 1
('me?--On', 'PERSON') 1
('Pequod.--On', 'ORG') 1
('Ahab\nbeware', 'PERSON') 1
('Burton', 'PERSON') 1
('Tierce', 'ORG') 1
('Aristotle', 'PERSON') 1
('Eternity', 'ORG') 1
('Zoroaster', 'PERSON') 1
('Death', 'GPE') 1
('Long Island', 'LOC') 2
('Rarmai', 'WORK_OF_ART') 1
('Antilles', 'PRODUCT') 1
('Antilles', 'PERSON') 2
('behind;--I', 'PERSON') 1
('Rig-a-dig', 'PERSON') 1
('there?--Hark', 'PERSON') 1
('almost half', 'CARDINAL') 1
('a day', 'DATE') 1
('a few indolent days', 'DATE') 1
('Many spare hours', 'TIME') 1
('CHAPTER 111', 'LAW') 1
('Bashee', 'ORG') 2
('Ephesian', 'NORP') 1
('Potters', 'ORG') 1
('Magian', 'NORP') 1
('Californian', 'NORP') 1
('Japans', 'NORP') 1
('Pan', 'PERSON') 2
('White Whale', 'FAC') 1
('Delta', 'LOC') 1
('brooks', 'ORG') 1
('CHAPTER 112', 'LAW') 1
('Blacksmith', 'ORG') 1
('Perth', 'GPE') 2
('an early period', 'DATE') 1
('the age of nearly sixty', 'DATE') 1
('every Sunday', 'DATE') 1
('Labor', 'ORG') 1

('Remote', 'ORG') 1
('Wild', 'LOC') 1
('Unshored', 'ORG') 1
('Perth', 'PERSON') 12
('CHAPTER 113', 'LAW') 1
("Mother Carey's", 'PERSON') 1
('Perth', 'ORG') 2
('Said', 'PERSON') 1
('Quick', 'ORG') 1
('Parsee', 'ORG') 1
('the Icy Sea', 'LOC') 1
('Ahoy', 'PERSON') 1
('CHAPTER 114', 'LAW') 1
('Gilder', 'ORG') 1
('twelve, fifteen, eighteen', 'DATE') 1
('twenty hours', 'TIME') 1
('sixty or seventy minutes', 'QUANTITY') 1
('blue hill', 'LOC') 1
('May-time', 'DATE') 1
('ye,--though', 'GPE') 1
('murmured:--', 'PERSON') 1
('I am Stubb', 'WORK_OF_ART') 1
('CHAPTER 115', 'LAW') 1
('The Pequod Meets The Bachelor', 'ORG') 1
('some few weeks', 'DATE') 1
('Bachelor', 'ORG') 6
('POKE', 'ORG') 1
('Bastille', 'GPE') 1
('Hast', 'ORG') 1
('CHAPTER 116', 'LAW') 1
('Niger', 'ORG') 1
('Typhoon', 'DATE') 4
('thou', 'GPE') 1
('CHAPTER 117', 'LAW') 1
('leeward', 'ORG') 1
('Asphaltites', 'NORP') 1
('Gomorrah', 'ORG') 1
('Parsee:--a', 'ORG') 1
('Quadrant', 'PERSON') 1
('The season', 'DATE') 1
('Thou sea-mark', 'PERSON') 1
('Pilot', 'WORK_OF_ART') 1
('Foolish toy!', 'WORK_OF_ART') 1
('Commodores', 'PRODUCT') 1
('Captains', 'PERSON') 1
('Curse', 'PERSON') 2
('Horatii', 'ORG') 1
('CHAPTER 119', 'LAW') 1
('Bengal', 'GPE') 1
('Typhoon', 'GPE') 3
('the windward quarter', 'DATE') 1
('Avast Stubb', 'PERSON') 1
('Madman', 'PERSON') 1
('thou hast none', 'PERSON') 1
('Old Thunder', 'WORK_OF_ART') 1
('God', 'PERSON') 1
('Mene, Mene,', 'WORK_OF_ART') 1
('Satanic', 'PRODUCT') 1
('Herculaneum', 'GPE') 1
('thou madest', 'LOC') 1
('NINE', 'CARDINAL') 1
('REST', 'ORG') 1
('thou knowest', 'LOC') 1
('thou hermit', 'GPE') 1
('CHAPTER 120', 'LAW') 1
('AHAB STANDING BY', 'PERSON') 1

```
('HELM', 'ORG') 1
('Loftiest', 'PERSON') 1
('CHAPTER 121', 'LAW') 1
('Midnight.--The Forecastle Bulwarks', 'PERSON') 1
('Marine Insurance', 'ORG') 1
('Ahab,--aye', 'PERSON') 1
('us,--were', 'ORG') 1
('CHAPTER 122', 'LAW') 1
('Lightning', 'GPE') 1
('CHAPTER 123', 'LAW') 1
('Some hours', 'TIME') 1
('HO', 'GPE') 1
('FAIR', 'ORG') 1
('twenty-four hours', 'TIME') 1
('Ahabs', 'NORP') 1
('law.--Aye', 'ORG') 1
('Mary', 'PERSON') 1
('this day\nweek', 'DATE') 1
("Thou know'st", 'PERSON') 1
('CHAPTER 124', 'LAW') 1
('Babylonian', 'NORP') 1
('Ho', 'GPE') 1
('ho', 'GPE') 1
('Yoke', 'PERSON') 1
('Heading East', 'WORK_OF_ART') 1
('sun', 'PERSON') 1
('CHAPTER 125', 'LAW') 1
('The Log and Line', 'ORG') 1
('Rains', 'PERSON') 1
('many hours', 'TIME') 1
('Tahitian', 'NORP') 3
('some thirty', 'CARDINAL') 1
('Twill', 'PERSON') 1
("Queen Nature's", 'PERSON') 1
('Isle of Man', 'LOC') 1
('the Isle of Man', 'GPE') 1
('Man', 'GPE') 1
('Snap', 'PERSON') 1
('one long festoon', 'TIME') 1
('Jerk', 'PERSON') 2
('Lo', 'PERSON') 2
('dong', 'PERSON') 2
('Ding', 'PERSON') 1
('Mend', 'LOC') 1
('CHAPTER 126', 'LAW') 1
('The Life-Buoy', 'ORG') 1
('Herod', 'PERSON') 1
('the pagan\nharpooneers', 'ORG') 1
```

Create a list `ranked_ne` containing all the items in the `ne_counts` dictionary that is sorted in descending order by their frequency.

In [219…

```python
sorted_items = dict(sorted(ne_counts.items(), key=lambda item: item[1], reverse=True))

ranked_ne = list(sorted_items.items()) # YOUR CODE GOES HERE

# This should display 2974 unique named entities, with the top two being
# ('Ahab', 'PERSON') 347 and ('one', 'CARDINAL') 335
print('Unique named entities:', len(ranked_ne))
for ne, count in ranked_ne[:50]:
    print(ne, count)
```

```
Unique named entities: 3001
('one', 'CARDINAL') 350
('Ahab', 'PERSON') 344
```

```
('two', 'CARDINAL') 205
('first', 'ORDINAL') 171
('Pequod', 'GPE') 152
('Starbuck', 'PERSON') 146
('three', 'CARDINAL') 134
('Queequeg', 'GPE') 114
('Queequeg', 'NORP') 98
('half', 'CARDINAL') 90
('Bildad', 'GPE') 65
('Nantucket', 'GPE') 58
('Moby Dick', 'PERSON') 57
('Peleg', 'PERSON') 55
('Indian', 'NORP') 48
('second', 'ORDINAL') 45
('Leviathan', 'GPE') 44
('four', 'CARDINAL') 41
('Tashtego', 'ORG') 40
('English', 'LANGUAGE') 39
('Jonah', 'GPE') 38
('Flask', 'ORG') 38
('Flask', 'GPE') 38
('Jonah', 'PERSON') 33
('Steelkilt', 'ORG') 30
('the White Whale', 'ORG') 29
('American', 'NORP') 28
('Pacific', 'LOC') 28
('Thou', 'PERSON') 27
('Greenland', 'GPE') 25
('Dutch', 'NORP') 22
('night', 'TIME') 22
('Lakeman', 'PERSON') 22
('Gabriel', 'PERSON') 19
('One', 'CARDINAL') 19
('Stubb', 'PERSON') 19
('New Bedford', 'GPE') 18
('midnight', 'TIME') 18
('Nantucketer', 'ORG') 18
('Christian', 'NORP') 17
('Yojo', 'PERSON') 17
('the Sperm Whale', 'LOC') 17
('Fedallah', 'PERSON') 17
('Atlantic', 'LOC') 16
('ten', 'CARDINAL') 16
('First', 'ORDINAL') 16
('Aye', 'ORG') 16
('French', 'NORP') 16
('whaleman', 'PERSON') 16
('German', 'NORP') 15
```

Consolidate named entities

Some names appear with more than one type, most often due to errors in named entity recognition. One way to fix such errors is to use the fact that typically a name occurs with just one meaning in a document, as such it has just one type. In this part of the assignment, we will consolidate the extracted names such that the counts for the same name appearing with multiple types are added together, and by associating the name with the type that it appears with most often.

Create a dictionary `ne_types` that maps each name to a dictionary that contains all the types the name appears with, where each type is mapped to the corresponding count. Use information from the dictionary `ne_counts` above.

In [220…
```
# Create a dictionary to store the consolidated named entities
ne_types = {}

# Iterate through the ne_counts dictionary
for (name, ent_type), count in ne_counts.items():
    if name in ne_types:
        if ent_type in ne_types[name]:
            ne_types[name][ent_type] += count
        else:
            ne_types[name][ent_type] = count
    else:
        ne_types[name] = {ent_type: count}


print(ne_types['Queequeg']) # this should print {'GPE': 109, 'NORP': 98, 'PERSON': 4, 'L

print(ne_types['Gabriel']) # this should print {'PERSON': 18, 'ORG': 1}
```

```
{'NORP': 98, 'GPE': 114, 'PERSON': 4, 'LANGUAGE': 8}
{'PERSON': 19}
```

Create the consolidated dictionary `ne_cons` that maps each name to a tuple that contains its most frequent type and the total count over all types. Use information from the dictionary `ne_types` above.

In [221…
```
# Create the consolidated dictionary ne_cons
ne_cons = {}

final_ne_types = {}

for name, type_counts in ne_types.items():
    most_frequent_type = max(type_counts, key=type_counts.get)
    final_ne_types[name] = most_frequent_type

# Iterate through the final_ne_types dictionary (from the previous step)
for name, most_frequent_type in final_ne_types.items():
    # Get the total count for the current name by summing all type counts
    total_count = sum(ne_types[name].values())
    # Create a tuple (most frequent type, total count) for the current name
    ne_cons[name] = (most_frequent_type, total_count)


print(ne_cons['Queequeg']) # this should print ('GPE', 219)

print(ne_cons['Gabriel']) # this should print ('PERSON', 19)
```

```
('GPE', 224)
('PERSON', 19)
```

Create a list `ranked_nec` that contains only the consolidated entries from `ne_cons` whose type is among the types listed in the list `types` below, sorted in descending order based on their total counts.

In [222…
```
types = ['PERSON', 'GPE', 'ORG', 'LOC', 'FAC']

# YOUR CODE HERE

# Filter and sort the consolidated named entities
sort_ne = sorted(
    ((name, (ent_type, count)) for name, (ent_type, count) in ne_cons.items() if ent_typ
    key=lambda x: x[1][1],
    reverse=True
)

# Create the ranked_nec list containing the filtered and sorted entries
ranked_nec = []
```

```
for name, (entity_type, count) in sort_ne:
        data = {}
        ne = (name , entity_type)
        data[ne]= count
        ranked_nec.append(data)



# This should display 1632 consolidated named entities, with the top two entries being
# Ahab ('PERSON', 347) and Queequeg ('GPE', 219)
# Display the number of consolidated named entities
print('Consolidated named entities:', len(ranked_nec))

# Print the top 30 entries
for entry in ranked_nec[:30]:
    for ne , count in entry.items():
        print(ne , count)
```

```
Consolidated named entities: 1668
('Ahab', 'PERSON') 344
('Queequeg', 'GPE') 224
('Pequod', 'GPE') 152
('Starbuck', 'PERSON') 146
('Flask', 'ORG') 87
('Bildad', 'GPE') 75
('Jonah', 'GPE') 71
('Nantucket', 'GPE') 69
('Moby Dick', 'PERSON') 57
('Peleg', 'PERSON') 55
('Leviathan', 'GPE') 54
('Tashtego', 'ORG') 42
('the White Whale', 'ORG') 34
('Thou', 'PERSON') 31
('the Sperm Whale', 'LOC') 30
('Steelkilt', 'ORG') 30
('Pacific', 'LOC') 28
('Greenland', 'GPE') 25
('Sperm Whale', 'PERSON') 24
('Lakeman', 'PERSON') 23
('Stubb', 'PERSON') 20
('Gabriel', 'PERSON') 19
('Radney', 'ORG') 19
('New Bedford', 'GPE') 18
('Nantucketer', 'ORG') 18
('Fedallah', 'PERSON') 18
('Ishmael', 'GPE') 17
('Aye', 'ORG') 17
('Yojo', 'PERSON') 17
('Atlantic', 'LOC') 16
```

[Bonus points 1] (10 points) Select one name from the dictionary `ne_counts` that appears frequently with 2 types and explain why you think spaCy's named entity recognizer associated the name with those 2 types.

Ans: {'NORP': 98, 'GPE': 114, 'PERSON': 4, 'LANGUAGE': 8}

for the name 'Queequeg' the NER of spacy recognizes into 4 different types . I think its because of the sentences syntactic dependency between object ,verb and subject

[Bonus points 2] (20 points) Find all the syntactic dependency paths connecting the subject Ahab with a direct object, e.g. 'Ahab' ---> nsubj ---> <verb> ---> dobj ---> <object>. Rank all the object words based

on how frequently they appear connected to 'Ahab' through this syntactic pattern, and for the top 10 objects display the list of verbs that are used with each object.

Useful documentation is at:

- https://spacy.io/usage/linguistic-features#dependency-parse

In [223…

```
# YOUR CODE HERE
import spacy

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

# Dictionary to store object words and their associated verbs
obj_verbs = {}


count = 0
paragraph = ""
flag = False
# Read the text file line by line
with open('../data/melville-moby_dick.txt', 'r') as f:


    for line in f:
        if line.strip():
            paragraph += line.strip()
            flag = True
        else:
            if flag:
                doc = nlp(paragraph)
                for sentence in doc.sents:
                    for token in sentence:
                        if token.text == 'Ahab' and token.dep_ == 'nsubj' and token.head
                                for child in token.head.children:
                                    if child.dep_ == 'dobj':
                                        word = child.text.lower()
                                        verb = token.head.text
                                        # Store the object word and associated verb
                                        if word in obj_verbs:
                                            obj_verbs[word].append(verb)
                                        else:
                                            obj_verbs[word] = [verb]
                count += 1
                paragraph = ""
                flag = False

            if count >= 2500:
                break

# Rank object words based on their frequency
sorted_obj_verbs = sorted(obj_verbs.items(), key=lambda x: len(x[1]), reverse=True)

# Display the top 10 object words and associated verbs
for object_word, verbs in sorted_obj_verbs[:10]:
    print(f"Object Word: {object_word}")
    print(f"Associated Verbs: {', '.join(verbs)}")
```

```
Object Word: humanities
Associated Verbs: has
Object Word: planks
Associated Verbs: paced
Object Word: ye
Associated Verbs: kicked
```

```
Object Word: he
Associated Verbs: kicked
Object Word: glimpse
Associated Verbs: had
Object Word: maze
Associated Verbs: threading
Object Word: prey
Associated Verbs: leaped
Object Word: that
Associated Verbs: knew
Object Word: sails
Associated Verbs: commanded
Object Word: reserve
Associated Verbs: manifested
```

Bonus points

Anything extra goes here. For example:

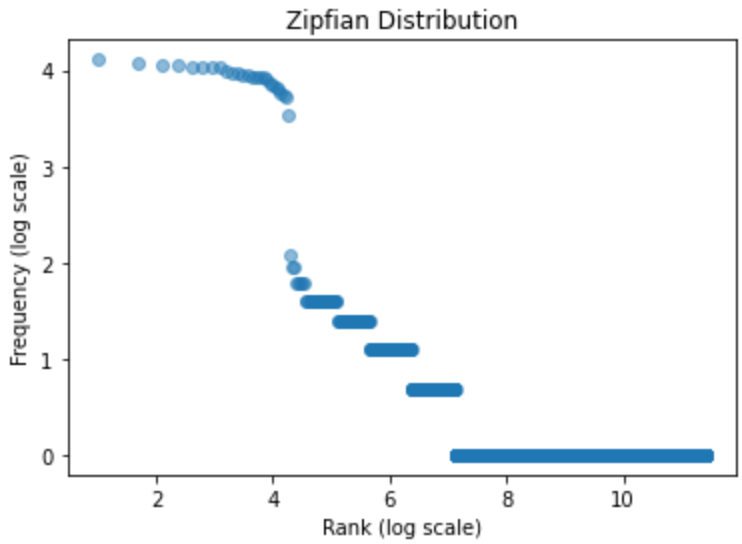
- Write code Li (1992) showing that just random typing of letters including a space will generate “words” with a Zipfian distribution. Generate at least 1 million characters before your compute word frequencies.
    - Show mathematically that random typing results in a Zipf's distribution by computing probabilities for all words that contain just 1 letter, 2 letters, ...
- Implement the BPE algorithm, where you break ties by selecting to merge in lexicographic order. Train the BPE algorithm on a large corpus and then use it to do subword tokenization on the Mobby Dick corpus. What are the top 10 most frequenct tokens and how does it compare with what you got from `tiktokenizer`.

In [224…

```python
import random


# Set the seed for reproducibility
random.seed(50)

# Generate random text with spaces (1 million characters)
length = 1000000
random_text = ''.join(random.choice('abcdefghijklmnopqrstuvwxyz ') for _ in range(length

# Split the random text into words based on spaces
words = random_text.split()

#print(words)

freq = {}

for word in words :
    freq[word] = freq.get(word, 0) + 1


sorted_freq = sorted(freq.items(), key=lambda x: x[1], reverse=True)
# top 10 words and their frequencies
print("Top 10 Words and Frequencies:")
for word, frequency in sorted_freq[:10]:
    print(f"{word}: {frequency}")

# Plot the word frequencies to visualize the Zipfian distribution
import matplotlib.pyplot as plt
```

```
# Get the frequencies and ranks
frequencies = [freq for _, freq in sorted_freq]
ranks = list(range(1, len(frequencies) + 1))


import math
ranks = [1 + math.log(r) for r in ranks]
freqs = [math.log(freq) for freq in frequencies]
plt.scatter(ranks, freqs, c='#1f77b4', alpha=0.5)
plt.xlabel('Rank (log scale)')
plt.ylabel('Frequency (log scale)')
plt.title('Zipfian Distribution')
plt.show()
```

```
Top 10 Words and Frequencies:
x: 61
i: 59
f: 58
k: 58
w: 57
q: 57
h: 57
j: 57
u: 54
a: 53
```



In [ ]: