(http://www.pieriandata.com)

Parts of Speech Project

For this assessment we'll be using the short story The Tale of Peter Rabbit (https://en.wikipedia.org/wiki/The_Tale_of_Peter_Rabbit) by Beatrix Potter (1902). The story is in the public domain; the text file was obtained from Project Gutenberg (https://www.gutenberg.org/ebooks/14838.txt.utf-8).

In [2]:
```
# RUN THIS CELL to perform standard imports:
import spacy
nlp = spacy.load('en_core_web_sm')
from spacy import displacy
```

1. Create a Doc object from the file `peterrabbit.txt`

> HINT: Use `with open('../TextFiles/peterrabbit.txt') as f:`

In [9]:
```
with open('../TextFiles/peterrabbit.txt') as f:
    doc = nlp(f.read())
```

2. For every token in the third sentence, print the token text, the POS tag, the fine-grained TAG tag, and the description of the fine-grained tag.

In [20]:
```
for token in list(doc.sents)[2]:
    print(f'{token.text:{12}} {token.pos_:{10}} {token.tag_:{8}} {spacy.
```

```
They         PRON       PRP      pronoun, personal
lived        VERB       VBD      verb, past tense
with         ADP        IN       conjunction, subordinating or preposi
tion
their        ADJ        PRP$     pronoun, possessive
Mother       PROPN      NNP      noun, proper singular
in           ADP        IN       conjunction, subordinating or preposi
tion
a            DET        DT       determiner
sand         NOUN       NN       noun, singular or mass
-            PUNCT      HYPH     punctuation mark, hyphen
bank         NOUN       NN       noun, singular or mass
,            PUNCT      ,        punctuation mark, comma
underneath   ADP        IN       conjunction, subordinating or preposi
tion
the          DET        DT       determiner
root         NOUN       NN       noun, singular or mass
of           ADP        IN       conjunction, subordinating or preposi
tion
a            DET        DT       determiner

             SPACE              None
very         ADV        RB       adverb
big          ADJ        JJ       adjective
fir          NOUN       NN       noun, singular or mass
-            PUNCT      HYPH     punctuation mark, hyphen
tree         NOUN       NN       noun, singular or mass
.            PUNCT      .        punctuation mark, sentence closer


          SPACE      _SP      None
```

3. Provide a frequency list of POS tags from the entire document

In [23]:
```
POS_count = doc.count_by(spacy.attrs.POS)
for k,v in sorted(POS_count.items()):
    print(f'{k}.  {doc.vocab[k].text} : {v}')
```

```
83.  ADJ : 83
84.  ADP : 127
85.  ADV : 75
88.  CCONJ : 61
89.  DET : 90
91.  NOUN : 176
92.  NUM : 8
93.  PART : 36
94.  PRON : 72
95.  PROPN : 75
96.  PUNCT : 174
99.  VERB : 182
102.  SPACE : 99
```

4. CHALLENGE: What percentage of tokens are nouns?
HINT: the attribute ID for 'NOUN' is 91

In [48]:
```
# len(doc) gives the number of tokens in doc

100 * POS_count[91]/len(doc)
```

Out[48]:
```
13.990461049284578
```

5. Display the Dependency Parse for the third sentence

In [6]:
```

```

They PRON lived VERB with ADP their ADJ Mother PROPN in ADP a DET sand- NOUN bank, NOUN underneath ADP the DET root NOUN of ADP a DET very ADV big ADJ fir- NOUN tree. NOUN SPACE nsubj prep poss pobj prep det compound pobj prep det pobj prep det advmod amod compound punct

In [34]:
```
displacy.render(list(doc.sents)[2], style = 'dep', jupyter = True, optio
```

They PRON lived VERB with ADP their ADJ Mother PROPN in ADP a DET sand- NOUN bank, NOUN underneath ADP the DET root NOUN of ADP a DET very ADV big ADJ fir- NOUN tree. NOUN SPACE nsubj prep poss pobj prep det compound pobj prep det pobj prep det advmod amod compound punct

*6. Show the first two named entities from Beatrix Potter's *The Tale of Peter Rabbit* **

In [38]:
```
for ent in list(doc.ents)[0:2]:
    print(ent.text + ' - ' + ent.label_ + ' - ' + spacy.explain(ent.labe
```

```
The Tale of Peter Rabbit - WORK_OF_ART - Titles of books, songs, etc.
Beatrix Potter - PERSON - People, including fictional
```

7. How many sentences are contained in *The Tale of Peter Rabbit*?

In [39]:
```
len(list(doc.sents))
```

Out[39]: 56

8. CHALLENGE: How many sentences contain named entities?

In [41]:
```
count = 0
for sent in doc.sents:
    if sent.ents:
        count += 1
print(count)
```

```
51
```

In [50]:
```
# Instructor's solution

list_of_sents = [nlp(sent.text) for sent in doc.sents]
list_of_ners = [doc for doc in list_of_sents if doc.ents]
len(list_of_ners)
```

Out[50]: 49

9. CHALLENGE: Display the named entity visualization for `list_of_sents[0]` from the previous problem

In [44]:
```
list_of_sents = list(doc.sents)
displacy.render(list_of_sents[0], style = 'ent', jupyter = True)
```

The Tale of Peter Rabbit **WORK_OF_ART** , by Beatrix Potter **PERSON** (1902 **DATE**).