Final Project Step 1

Course: DS 5001
Module: Final
Date: 8 May 2022
Author: Thomas McIntyre gem5cm@virginia.edu
Purpose: This notebook will take in the raw text files and create corpus and lib tables and save them to the data folder

Creating LIB, CORPUS from Raw txt files

In [1]:
```
#https://www.kaggle.com/datasets/ashishsinhaiitr/lord-of-the-rings-text
#https://archive.org/stream/3.AStormOfSwords/3.%20A%20Storm%20of%20Swords_djvu.txt
#https://archive.org/stream/II.AClashOfKings_201803/II.%20A%20Clash%20of%20Kings_djvu.t
#https://archive.org/stream/1AGameOfThrones/1%20A%20Game%20of%20Thrones_djvu.txt

#gutenberg for the rest of the corpus
data_home = "corpus"
local_lib = "code"
OHCO = ['book_id', 'chap_num', 'para_num', 'sent_num', 'token_num']
SENTS = OHCO[:4]
PARAS = OHCO[:3]
CHAPS = OHCO[:2]
BOOKS = OHCO[:1]
```

In [2]:
```
import pandas as pd
import numpy as np
from glob import glob
import re
import nltk
import sys
import re
exec(open("code/methods.py").read())
```

In [3]:
```
epub_list = sorted(glob("corpus/*.*"))
epub_list
```

Out[3]:
```
['corpus/BARRIE_PETER_AND_WENDY-pg26654.txt',
 'corpus/BARRIE_PETER_PAN-pg16.txt',
 'corpus/BRONTE_JANE_EYRE-pg1260.txt',
 'corpus/BRONTE_WUTHERING_HEIGHTS-pg768.txt',
 'corpus/DICKENS_GREAT_EXPECTATIONS-pg1400.txt',
 'corpus/DICKENS_OLIVER_TWIST-pg730.txt',
 'corpus/HOMER-THE-ILIAD-pg6130.txt',
 'corpus/HOMER_THE_ODYSSEY-pg1727.txt',
 'corpus/MARTIN_A_CLASH_OF_KINGS-pg2.txt',
 'corpus/MARTIN_A_GAME_OF_THRONES-pg1.txt',
 'corpus/MARTIN_A_STORM_OF_SWORDS-pg3.txt',
 'corpus/TOLKIEN_THE_FELLOWSHIP_OF_THE_RING-pg4.txt']
```

In [4]:
```
chap_pats ={
    1260:{
        'start_line':57,
        'end_line':21031,
        'chapter': r"((CHAPTER)\s+\D+)|PREFACE"
```

```
    },
    768:{
        'start_line':34,
        'end_line':12370,
        'chapter': r"((CHAPTER)\s+\D+)"
    },
    1400:{
        'start_line':104,
        'end_line':20423,
        'chapter': r"((Chapter)\s+\D+)"
    },
    730:{
        'start_line':141,
        'end_line':18842,
        'chapter': r"((CHAPTER)\s+\D+)"
    },
    16:{
        'start_line':65,
        'end_line':6287,
        'chapter': r"((Chapter)\s+\D+)"
    },
    26654:{
        'start_line':140,
        'end_line':6413,
        'chapter': r"((CHAPTER)\s+\D+)"
    },
    1727:{
        'start_line':366,
        'end_line':10832,
        'chapter': r"((BOOK)\s+\D+)"
    },
    6130:{
        'start_line':2032,
        'end_line':23285,
        'chapter': r"((BOOK)\s+\D+)"
    },
    1:{
        'start_line':1674,
        'end_line':29533,
        'chapter': r"([A-Z]+[A-Z]+[A-Z]+)"
    },
    2:{
        'start_line':1700,
        'end_line':33952,
        'chapter': r"([A-Z]+[A-Z]+[A-Z]+)"
    },
    3:{
        'start_line':1684,
        'end_line':41418,
        'chapter': r"([A-Z]+[A-Z]+[A-Z]+)"
    },
    4:{
        'start_line':105,
        'end_line':4263,
        'chapter': r"_Chapter"
    }



}
```

In [5]:
```
LIB, DOCS = acquire_epubs(epub_list = epub_list, chap_pats = chap_pats, OHCO = OHCO)
LIB.at[1, "book_title"] = "A Game of Thrones, by RR Martin"
LIB.at[2, "book_title"] = "A Clash of Kings, by RR Martin"
LIB.at[3, "book_title"] = "A Storm of Swords, by RR Martin"
LIB.at[4, "book_title"] = "The Fellowship of the Ring, by JRR Tolkien"
LIB
```

```
BOOK ID 26654
BOOK ID 16
BOOK ID 1260
BOOK ID 768
/opt/conda/lib/python3.7/site-packages/ipykernel/__main__.py:50: FutureWarning: The defa
ult value of regex will change from True to False in a future version.
BOOK ID 1400
BOOK ID 730
BOOK ID 6130
BOOK ID 1727
BOOK ID 2
BOOK ID 1
BOOK ID 3
BOOK ID 4
```

Out[5]:

book_id	book_title	book_file
26654	Peter and Wendy, by James Matthew Barrie	corpus/BARRIE_PETER_AND_WENDY-pg26654.txt
16	Peter Pan, by James M. Barrie	corpus/BARRIE_PETER_PAN-pg16.txt
1260	Jane Eyre, by Charlotte Brontë	corpus/BRONTE_JANE_EYRE-pg1260.txt
768	Wuthering Heights, by Emily Brontë	corpus/BRONTE_WUTHERING_HEIGHTS-pg768.txt
1400	Great Expectations, by Charles Dickens	corpus/DICKENS_GREAT_EXPECTATIONS-pg1400.txt
730	Oliver Twist, by Charles Dickens	corpus/DICKENS_OLIVER_TWIST-pg730.txt
6130	The Iliad, by Homer	corpus/HOMER-THE-ILIAD-pg6130.txt
1727	The Odyssey, by Homer	corpus/HOMER_THE_ODYSSEY-pg1727.txt
2	A Clash of Kings, by RR Martin	corpus/MARTIN_A_CLASH_OF_KINGS-pg2.txt
1	A Game of Thrones, by RR Martin	corpus/MARTIN_A_GAME_OF_THRONES-pg1.txt
3	A Storm of Swords, by RR Martin	corpus/MARTIN_A_STORM_OF_SWORDS-pg3.txt
4	The Fellowship of the Ring, by JRR Tolkien	corpus/TOLKIEN_THE_FELLOWSHIP_OF_THE_RING-pg4.txt

In [6]:
```
regex_chap_list = [
    (26654,  r"((CHAPTER)\s+\D+)"),
    (16,    r"((Chapter)\s+\D+)"),
    (1260,  r"((CHAPTER)\s+\D+)|PREFACE"),
    (768,  r"((CHAPTER)\s+\D+)"),
    (1400,   r"((Chapter)\s+\D+)"),
    (730,  r"((CHAPTER)\s+\D+)"),
    (6130,  r"((BOOK)\s+\D+)"),
    (1727,   r"((BOOK)\s+\D+)"),
    (1,  r"[A-Z]+[A-Z]+[A-Z]+"),
    (2,  r"[A-Z]+[A-Z]+[A-Z]+"),
    (3,  r"[A-Z]+[A-Z]+[A-Z]+"),
```

```
    (4,  r"_Chapter")
]

LIB['chap_regex'] = LIB.index.map(pd.Series({x[0]:x[1] for x in regex_chap_list}))
```

In [7]:
```
DOCS
```

Out[7]:

			para_str
book_id	**chap_num**	**para_num**	
26654	**1**	**0**	PETER BREAKS THROUGH
		1	All children, except one, grow up. They soon k...
		2	Of course they lived at 14, and until Wendy ca...
		3	The way Mr. Darling won her was this: the many...
		4	Mr. Darling used to boast to Wendy that her mo...
...	**...**	**...**	...
4	**21**	**3**	The eighth night of their journey came. It was...
	22	**0**	The Breaking of the Fellowship
		1	Aragorn led them to the right arm of the River...
		2	Aragorn sprang swiftly away and went in pursui...
		3	So Frodo and Sam set off on the last stage of ...

43952 rows × 1 columns

In [8]:
```
CORPUS = tokenize(doc_df = DOCS, OHCO =OHCO, ws = True)
```

In [9]:
```
CORPUS['term_str'] = CORPUS.token_str.replace(r'[\W_]+', '', regex=True).str.lower()
CORPUS.head(10)
```

Out[9]:

					pos_tuple	pos	token_str	term_str
book_id	**chap_num**	**para_num**	**sent_num**	**token_num**				
26654	**1**	**0**	**0**	**0**	(PETER, NN)	NN	PETER	peter
				1	(BREAKS, NNP)	NNP	BREAKS	breaks
				2	(THROUGH, NNP)	NNP	THROUGH	through
			1	**0**	(All, DT)	DT	All	all
				1	(children,, NN)	NN	children,	children
				2	(except, IN)	IN	except	except
				3	(one,, JJ)	JJ	one,	one
				4	(grow, NN)	NN	grow	grow
				5	(up., NN)	NN	up.	up
			1	**0**	(They, PRP)	PRP	They	they

In [10]:
```
DF = CORPUS.reset_index()
DF26654 = DF[DF.book_id == 26654]
DF16 = DF[DF.book_id == 16]
DF1260 = DF[DF.book_id == 1260]
DF768 = DF[DF.book_id == 768]
DF1400 = DF[DF.book_id == 1400]
DF730 = DF[DF.book_id == 730]
DF6130 = DF[DF.book_id == 6130]
DF1727 = DF[DF.book_id == 1727]
DF1 = DF[DF.book_id == 1]
DF2 = DF[DF.book_id == 2]
DF3 = DF[DF.book_id == 3]
DF4 = DF[DF.book_id == 4]

tokens_lengths = [
    (26654,  DF26654.shape[0]),
    (16,    DF16.shape[0]),
    (1260,  DF1260.shape[0]),
    (768,  DF768.shape[0]),
    (1400,   DF1400.shape[0]),
    (730,  DF730.shape[0]),
    (6130,  DF6130.shape[0]),
    (1727,   DF1727.shape[0]),
    (1,  DF1.shape[0]),
    (2,  DF2.shape[0]),
    (3,  DF3.shape[0]),
    (4,  DF4.shape[0])
]

chapters_total = [
    (26654,  DF26654.chap_num.max()),
    (16,    DF16.chap_num.max()),
    (1260,  DF1260.chap_num.max()),
    (768,  DF768.chap_num.max()),
    (1400,   DF1400.chap_num.max()),
    (730, DF730.chap_num.max()),
    (6130,  DF6130.chap_num.max()),
    (1727,  DF1727.chap_num.max()),
    (1,  DF1.chap_num.max()),
    (2,   DF2.chap_num.max()),
    (3,  DF3.chap_num.max()),
    (4,   DF4.chap_num.max())

]
LIB["book_length"] = LIB.index.map(pd.Series({x[0]:x[1] for x in tokens_lengths}))
LIB["n_chaps"] = LIB.index.map(pd.Series({x[0]:x[1] for x in chapters_total}))
```

In [11]:
```
nltk_resources = [
    'tokenizers/punkt',
    'taggers/averaged_perceptron_tagger',
    'corpora/stopwords',
    'help/tagsets'
]

for rsc in nltk_resources:
    try:
        nltk.data.find(rsc)
    except IndexError:
        nltk.download(rsc)
```

In [12]:
```
CORPUS.head()
```

Out[12]:

book_id	chap_num	para_num	sent_num	token_num	pos_tuple	pos	token_str	term_str
26654	**1**	**0**	**0**	**0**	(PETER, NN)	NN	PETER	peter
				1	(BREAKS, NNP)	NNP	BREAKS	breaks
				2	(THROUGH, NNP)	NNP	THROUGH	through
		1	**0**	**0**	(All, DT)	DT	All	all
				1	(children,, NN)	NN	children,	children

In [13]:
```
meta_csv = """
26654, childrensfantasy, adventure, James M. Barrie
16, childrensfantasy, adventure, James M. Barrie
1260, gothic, romance, Charlotte Brontë
768, gothic, tragic, Emily Brontë
1400, gothic, regretful, Charles Dickens
730, crime, satire, Charles Dickens
6130, historical, serious, Homer
1727, historical, serious, Homer
1, fantasy, adventure, RR Martin
2, fantasy, adventure, RR Martin
3, fantasy, adventure, RR Martin
4, fantasy, adventure, JRR Tolkien
""".split('\n')[1:-1]
meta = pd.DataFrame([line.split(', ') for line in meta_csv], columns=['book_id','genre'
meta.book_id = meta.book_id.astype('int')
meta = meta.set_index('book_id')

### Appending genre, mood, author, label to LIB
LIB = pd.concat([LIB, meta], axis=1)
LIB['label'] = LIB.apply(lambda x: f"{x.book_title}-{x.genre[0].upper()}", 1)
LIB
```

Out[13]:

book_id	book_title	book_file	chap_regex	book_le
1	A Game of Thrones, by RR Martin	corpus/MARTIN_A_GAME_OF_THRONES-pg1.txt	[A-Z]+[A-Z]+[A-Z]+	29
2	A Clash of Kings, by RR Martin	corpus/MARTIN_A_CLASH_OF_KINGS-pg2.txt	[A-Z]+[A-Z]+[A-Z]+	32
3	A Storm of Swords, by RR Martin	corpus/MARTIN_A_STORM_OF_SWORDS-pg3.txt	[A-Z]+[A-Z]+[A-Z]+	41
4	The Fellowship of the Ring, by JRR Tolkien	corpus/TOLKIEN_THE_FELLOWSHIP_OF_THE_RING-pg4.txt	_Chapter	18

book_id	book_title	book_file	chap_regex	book_le
16	Peter Pan, by James M. Barrie	corpus/BARRIE_PETER_PAN-pg16.txt	((Chapter)\s+\D+)	4
730	Oliver Twist, by Charles Dickens	corpus/DICKENS_OLIVER_TWIST-pg730.txt	((CHAPTER)\s+\D+)	16
768	Wuthering Heights, by Emily Brontë	corpus/BRONTE_WUTHERING_HEIGHTS-pg768.txt	((CHAPTER)\s+\D+)	11
1260	Jane Eyre, by Charlotte Brontë	corpus/BRONTE_JANE_EYRE-pg1260.txt	((CHAPTER)\s+\D+)|PREFACE	19
1400	Great Expectations, by Charles Dickens	corpus/DICKENS_GREAT_EXPECTATIONS-pg1400.txt	((Chapter)\s+\D+)	18
1727	The Odyssey, by Homer	corpus/HOMER_THE_ODYSSEY-pg1727.txt	((BOOK)\s+\D+)	11
6130	The Iliad, by Homer	corpus/HOMER-THE-ILIAD-pg6130.txt	((BOOK)\s+\D+)	15
26654	Peter and Wendy, by James Matthew Barrie	corpus/BARRIE_PETER_AND_WENDY-pg26654.txt	((CHAPTER)\s+\D+)	4

In [14]:
```
CORPUS.to_csv("data/CORPUS.csv")
LIB.to_csv("data/LIB.csv")
```

In [ ]: