Final Project Step 7 sentiment analysis

```
Course:   DS 5001
Module:   Final
Date:     8 May 2022
Author:   Thomas McIntyre gem5cm@virginia.edu
Purpose:  This notebook will utlize the data created in step 2 to get
sentiment analysis.
```

In [1]:

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import plotly_express as px
from IPython.display import display, HTML
sns.set()
%matplotlib inline
```

In [2]:

```
data_home = "data"
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 [3]:

```
salex_csv = f'{data_home}/salex_nrc.csv'
```

In [4]:

```
SALEX = pd.read_csv(salex_csv).set_index('term_str')
SALEX.columns = [col.replace('nrc_','') for col in SALEX.columns]
VOCAB = pd.read_csv(f"{data_home}/VOCAB.csv").set_index("term_str")
BOW = pd.read_csv(f"{data_home}/BOW.csv").rename(columns = {"Unnamed: 2": "term_str"}).
TOKENS = pd.read_csv(f'{data_home}/CORPUS.csv').set_index(OHCO).sort_index()
LIB = pd.read_csv(f"{data_home}/LIB.csv").set_index('book_id').sort_index()
```

In [5]:

```
COMBO = TOKENS.join(LIB).join(SALEX, on='term_str').join(BOW, on=OHCO[:2] + ['term_str'
COMBO = COMBO.dropna()
COMBO = COMBO.sort_index()
COMBO
```

Out[5]:

					pos_tuple	pos	token_str	term_str	book_title
book_id	**chap_num**	**para_num**	**sent_num**	**token_num**					
1	**1**	**0**	**0**	**11**	('grow', 'VB')	VB	grow	grow	A Game of Thrones, by RR Martin
		1	**0**	**3**	('frighten', 'NN')	NN	frighten	frighten	A Game of Thrones, by RR Martin

					pos_tuple	pos	token_str	term_str	book_title
book_id	**chap_num**	**para_num**	**sent_num**	**token_num**					
			1	**10**	('smile.', 'NN')	NN	smile.	smile	A Game of Thrones, by RR Martin
		3	**2**	**9**	('quarrel', 'NN')	NN	quarrel	quarrel	A Game of Thrones, by RR Martin
			4	**1**	('mother', 'NN')	NN	mother	mother	A Game of Thrones, by RR Martin
...	**...**	**...**	**...**	**...**	...	...	...	...	...
26654	**17**	**168**	**2**	**30**	('innocent', 'JJ')	JJ	innocent	innocent	Peter and Wendy, by James Matthew Barrie
		175	**0**	**4**	('public', 'JJ')	JJ	public	public	Peter and Wendy, by James Matthew Barrie
			2	**0**	('Special', 'JJ')	JJ	Special	special	Peter and Wendy, by James Matthew Barrie
				6	('General', 'NNP')	NNP	General	general	Peter and Wendy, by James Matthew Barrie
			6	**17**	('public', 'JJ')	JJ	public	public	Peter and Wendy, by James Matthew Barrie

115022 rows × 26 columns

In [6]:
```
emo_cols = "anger anticipation disgust fear joy sadness surprise trust polarity".split(
```

In [14]:
```
BOOKS_SA = COMBO.groupby(OHCO[:1])[emo_cols+['tfidf']].mean().join(LIB.label)
BOOKS_SA.sort_values('polarity').style.background_gradient()
```

Out[14]:

	anger	anticipation	disgust	fear	joy	sadness	surprise	trust	polarity

book_id	anger	anticipation	disgust	fear	joy	sadness	surprise	trust	polarity	
book_id										
6130	0.380709	0.280774	0.175413	0.453515	0.280774	0.285471	0.120019	0.296890	-0.184888	0
3	0.267112	0.221901	0.265488	0.317015	0.265352	0.303253	0.157785	0.397735	-0.140640	0
2	0.251574	0.224347	0.262069	0.303696	0.260494	0.284223	0.147621	0.414995	-0.122726	0
1	0.228867	0.215536	0.251542	0.290538	0.248925	0.278951	0.150190	0.421977	-0.118732	0
768	0.294534	0.270243	0.229251	0.361167	0.322706	0.368590	0.184717	0.307524	-0.106950	0
16	0.232191	0.276826	0.216411	0.261948	0.397205	0.306132	0.156447	0.385482	-0.034265	0
26654	0.230665	0.276529	0.215827	0.261241	0.396133	0.305306	0.155126	0.385791	-0.030126	0
4	0.263484	0.302123	0.156534	0.333773	0.364500	0.308849	0.211658	0.357115	-0.019649	0
1727	0.270355	0.341573	0.161860	0.330194	0.369433	0.291740	0.194232	0.387090	-0.010006	0
1400	0.225302	0.301128	0.193912	0.283870	0.347330	0.292295	0.193369	0.397608	0.055714	0
730	0.227946	0.301236	0.196745	0.287043	0.333293	0.278111	0.200049	0.390921	0.074147	0
1260	0.198074	0.317286	0.160706	0.283929	0.375516	0.294131	0.195323	0.418157	0.097662	0

In [8]:

```
CHAPS_SA = COMBO.groupby(OHCO[:2])[emo_cols+['tfidf']].mean().join(LIB.label)
CHAPS_SA.sort_values('polarity')
```

Out[8]:

	anger	anticipation	disgust	fear	joy	sadness	surprise	trust

book_id	chap_num	anger	anticipation	disgust	fear	joy	sadness	surprise	trust
book_id	**chap_num**								
3	**18**	0.500000	0.500000	0.500000	1.000000	0.000000	1.000000	0.500000	0.000000
2	**3**	0.258929	0.107143	0.392857	0.375000	0.107143	0.482143	0.107143	0.205357
4	**17**	0.300699	0.188811	0.125874	0.524476	0.122378	0.370629	0.150350	0.202797
2	**141**	0.366197	0.119718	0.330986	0.330986	0.084507	0.345070	0.098592	0.316901
730	**50**	0.360656	0.174863	0.377049	0.469945	0.158470	0.398907	0.169399	0.191257
...	**...**	...	...	...	...	...	...	...	...
1400	**22**	0.146341	0.380488	0.112195	0.190244	0.473171	0.195122	0.243902	0.468293
730	**53**	0.133333	0.433333	0.100000	0.166667	0.644444	0.188889	0.188889	0.466667
3	**13**	0.100000	0.300000	0.100000	0.300000	0.600000	0.200000	0.200000	0.400000
	16	0.333333	0.000000	0.000000	0.333333	0.666667	0.000000	0.333333	0.333333
	12	0.000000	0.333333	0.166667	0.000000	0.666667	0.000000	0.166667	1.000000

533 rows × 11 columns

```
In [9]: SENTENCES_SA = COMBO.groupby(OHCO[:-1])[emo_cols].mean().join(LIB.label)
        SENTENCES_SA.sort_values('polarity')
```

Out[9]:

				anger	anticipation	disgust	fear	joy	sadness	surprise	tr
book_id	**chap_num**	**para_num**	**sent_num**								

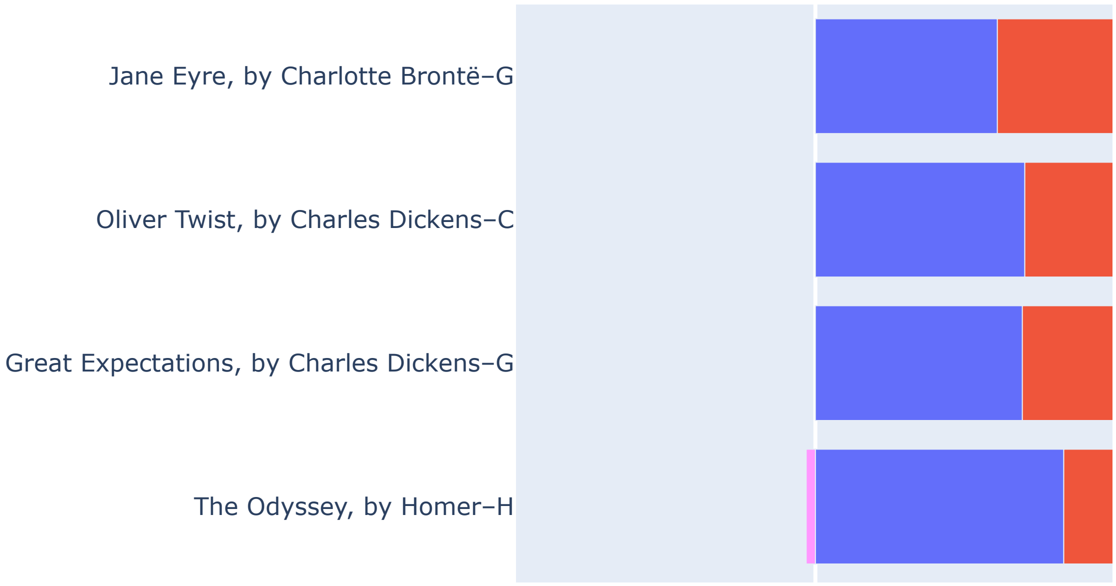
				anger	anticipation	disgust	fear	joy	sadness	surprise	tr
book_id	**chap_num**	**para_num**	**sent_num**								
3	**35**	**4**	**1**	1.0	0.0	0.0	1.0	0.0	1.0	0.0	
2	**112**	**127**	**0**	1.0	0.0	1.0	1.0	0.0	1.0	0.0	
6130	**2**	**17**	**1**	0.5	0.0	1.0	0.5	0.0	1.0	0.5	
2	**112**	**125**	**1**	1.0	0.0	0.0	0.0	0.0	0.0	0.0	
4	**8**	**1**	**87**	1.0	0.0	0.0	0.0	0.0	0.0	0.0	
	...	**...**	**...**	...	...	...	...	...	...	...	
	2	**3**	**55**	0.0	0.0	0.0	0.0	1.0	0.0	0.0	
			54	0.0	1.0	0.0	0.0	1.0	0.0	1.0	
			51	0.0	1.0	0.0	0.0	1.0	0.0	1.0	
			27	0.0	0.0	0.0	0.0	1.0	0.0	0.0	

				anger	anticipation	disgust	fear	joy	sadness	surprise	tr
book_id	chap_num	para_num	sent_num								
26654	17	175	6	0.0	1.0	0.0	0.0	0.0	0.0	0.0	

71179 rows × 10 columns

In [15]:

```
px.bar(BOOKS_SA.reset_index().sort_values('polarity'), emo_cols, 'label', orientation='
```



In [10]:

```
BOOKS_SA .to_csv("data/BOOKS_SA.csv")
CHAPS_SA.to_csv("data/CHAPS_SA.csv")
SENTENCES_SA.to_csv("data/SENTENCES_SA.csv")
```