Twain Topic Model and Word Embeddings

DS 5001: Exploratory Text Analytics

Cecily Wolfe (cew4pf)

Spring 2022

In [1]:

```
import pandas as pd
import numpy as np
from gensim.models import word2vec
from sklearn.manifold import TSNE
import plotly.express as px
```

In [2]:

```
from topicmodel import TopicModel
```

In [3]:

```
OHCO = ['book_id','chap_id','para_num','sent_num','token_num']
```

In [4]:

```
BOW = pd.read_csv("twain_BOW.csv")
BOW['term_str'] = BOW['term_str'].astype('str')
BOW = BOW.set_index(['book_id', 'chap_id', 'term_str'])
```

In [5]:

```
LIB = pd.read_csv(("twain_pre_LIB.csv"), index_col = ['book_id'])
```

In [6]:

```
CORPUS = pd.read_csv(("twain_pre_CORPUS.csv"), index_col = OHCO)
```

In [7]:

```
VOCAB = pd.read_csv("twain_pre_VOCAB.csv")

VOCAB['term_str'] = VOCAB['term_str'].astype('str')

VOCAB = VOCAB.set_index('term_str')

VOCAB['pos_group'] = VOCAB.max_pos.str.slice(0,2)
```

In [8]:

```
CHAPS = CORPUS.groupby(OHCO[:2]+['term_str']).term_str.count().unstack()
VOCAB['df'] = CHAPS.count()
VOCAB['dfidf'] = VOCAB.df * np.log2(len(CHAPS)/VOCAB.df)
```

In [9]:

```
VOCAB.head()
```

Out[9]:

	n	n_chars	p	i	max_pos	n_pos	cat_pos	stop	stem_porter	stem_

term_str	n	n_chars	p	i	max_pos	n_pos	cat_pos	stop	stem_porter	stem_
term_str										
0	5	1	0.000002	19.180285	CD	1	{'CD'}	0	0	
00	3	2	0.000001	19.917251	NN	2	{'NN', 'NNS'}	0	00	
01	3	2	0.000001	19.917251	NNS	2	{'NN', 'NNS'}	0	01	
02	4	2	0.000001	19.502213	NN	3	{'POS', 'NN', 'NNP'}	0	02	
03	6	2	0.000002	18.917251	NN	3	{'POS', 'NN', 'NNS'}	0	03	

In [10]:

```
BOW.head()
```

Out[10]:

book_id	chap_id	term_str	n	tf	tfidf
70	**1**	**1835**	1	0.142857	1.159106
		1910	1	0.142857	1.075540
		a	2	0.285714	0.002238
		alphabet	1	0.142857	0.991974
		as	2	0.285714	0.013615

In [11]:

```
LIB.head()
```

Out[11]:

book_id	source_file_path	title	chap_regex	author	type
70	Twain/70-what_is_man.txt	what is man	WHAT IS MAN?|THE DEATH OF JEAN|THE TURNING-POI...	twain	non-fiction
74	Twain/74-the_adventures_of_tom_sawyer.txt	the adventures of tom sawyer	^\s*CHAPTER\s*[IVXLCM]+$	twain	novel
76	Twain/76-the_adventures_of_huckleberry_finn.txt	the adventures of huckleberry finn	^\s*CHAPTER\s*(?:[IVXLCM]+\.|THE LAST)$	twain	novel

book_id	source_file_path	title	chap_regex	author	type
86	Twain/86-a_connecticut_yankee_in_king_arthurs_...	a connecticut yankee in king arthurs court	^\s*(?:PREFACE|A WORD OF EXPLANATION|THE STRAN...	twain	novel
91	Twain/91-tom_sawyer_abroad.txt	tom sawyer abroad	CHAPTER\s[IVXLCM]+\.	twain	novel

M08: Topic Models

In [12]:

```
# join BOW and VOCAB
joint_BOW = BOW.reset_index().set_index('term_str').join(VOCAB, rsuffix = "_voca

# remove nan
joint_BOW = joint_BOW.loc[~joint_BOW.isna().any(axis = 1)]

# remove proper nouns
joint_BOW = joint_BOW.loc[~joint_BOW.max_pos.isin(['NNP', 'NNPS'])]

joint_BOW
```

Out[12]:

term_str	book_id	chap_id	n	tf	tfidf	n_vocab	n_chars	p	
0	3199	1	2	0.008439	0.076909	5	1	1.683290e-06	19.180
0	3251	6	3	0.004587	0.041806	5	1	1.683290e-06	19.180
00	3199	24	3	0.012448	0.125897	3	2	1.009974e-06	19.91
01	3199	25	3	0.013699	0.138544	3	2	1.009974e-06	19.91
02	3186	14	1	0.005464	0.049802	4	2	1.346632e-06	19.50
...	...	...	...	...	...	...	...	...	
étouffante	60900	5	1	0.007752	0.078401	1	10	3.366579e-07	21.50
évitant	3189	3	1	0.004132	0.041792	1	7	3.366579e-07	21.50
êtes	3189	3	1	0.004132	0.041792	1	4	3.366579e-07	21.50
öffnen	60900	6	1	0.004608	0.046607	1	6	3.366579e-07	21.50
übergeschlagen	60900	6	1	0.004608	0.046607	1	14	3.366579e-07	21.50

877057 rows × 19 columns

In [13]:
```
# recover filtered BOW --> drop cols added by VOCAB and reset index to book_id,

filtered_BOW = joint_BOW.drop(joint_BOW.loc[:, 'n_vocab':].columns, axis = 1).re

# sort by book id
filtered_BOW = filtered_BOW.sort_values('book_id')

filtered_BOW
```

Out[13]:

book_id	chap_id	term_str	n	tf	tfidf
70	10	read	3	0.014423	0.019551
		stock	1	0.004808	0.013125
	16	stock	1	0.010989	0.030000
	17	stock	2	0.001498	0.004090
	2	inert	1	0.000732	0.005080
...	...	...	...	...	...
62739	4	two	5	0.017668	0.003841
	5	two	4	0.038095	0.008282
	4	most	3	0.010601	0.004905
	2	everything	2	0.005556	0.006925
		officials	1	0.002778	0.012337

877057 rows × 3 columns

In [14]:
```
# removed ~ 5% of data when taking out proper nouns (singular and plural)
(BOW.shape[0] - filtered_BOW.shape[0]) / BOW.shape[0]
```

Out[14]: 0.05007110465109982

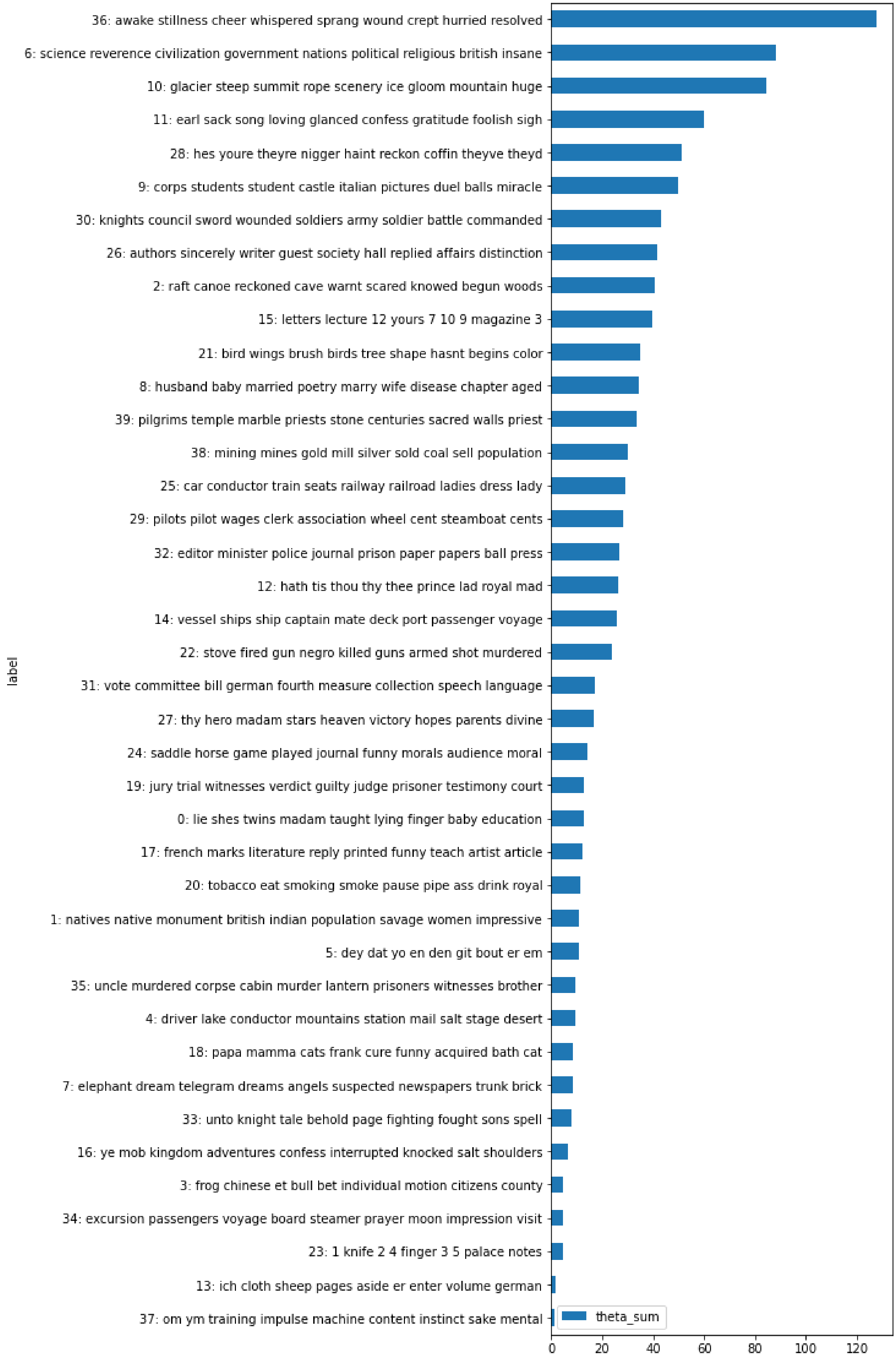
In [15]:
```
n_topics = 40
n_terms = 2000
```

In [16]:
```
tm = TopicModel(filtered_BOW)
tm.n_topics = n_topics
tm.n_terms = n_terms
```

In [17]:
```
tm.create_X()
tm.get_model()
tm.describe_topics()
tm.get_model_stats()
```

In [18]:

```
tm.plot_topics()
```


36: awake stillness cheer whispered sprang wound crept hurried resolved
6: science reverence civilization government nations political religious british insane
10: glacier steep summit rope scenery ice gloom mountain huge
11: earl sack song loving glanced confess gratitude foolish sigh
28: hes youre theyre nigger haint reckon coffin theyve theyd
9: corps students student castle italian pictures duel balls miracle
30: knights council sword wounded soldiers army soldier battle commanded
26: authors sincerely writer guest society hall replied affairs distinction
2: raft canoe reckoned cave warnt scared knowed begun woods
15: letters lecture 12 yours 7 10 9 magazine 3
21: bird wings brush birds tree shape hasnt begins color
8: husband baby married poetry marry wife disease chapter aged
39: pilgrims temple marble priests stone centuries sacred walls priest
38: mining mines gold mill silver sold coal sell population
25: car conductor train seats railway railroad ladies dress lady
29: pilots pilot wages clerk association wheel cent steamboat cents
32: editor minister police journal prison paper papers ball press
12: hath tis thou thy thee prince lad royal mad
14: vessel ships ship captain mate deck port passenger voyage
22: stove fired gun negro killed guns armed shot murdered
31: vote committee bill german fourth measure collection speech language
27: thy hero madam stars heaven victory hopes parents divine
24: saddle horse game played journal funny morals audience moral
19: jury trial witnesses verdict guilty judge prisoner testimony court
0: lie shes twins madam taught lying finger baby education
17: french marks literature reply printed funny teach artist article
20: tobacco eat smoking smoke pause pipe ass drink royal
1: natives native monument british indian population savage women impressive
5: dey dat yo en den git bout er em
35: uncle murdered corpse cabin murder lantern prisoners witnesses brother
4: driver lake conductor mountains station mail salt stage desert
18: papa mamma cats frank cure funny acquired bath cat
7: elephant dream telegram dreams angels suspected newspapers trunk brick
33: unto knight tale behold page fighting fought sons spell
16: ye mob kingdom adventures confess interrupted knocked salt shoulders
3: frog chinese et bull bet individual motion citizens county
34: excursion passengers voyage board steamer prayer moon impression visit
23: 1 knife 2 4 finger 3 5 palace notes
13: ich cloth sheep pages aside er enter volume german
37: om ym training impulse machine content instinct sake mental
label
theta_sum
0
20
40
60
80
100
120

In [19]:
```
# table with distribution of topics for each doc
tm.THETA
```

Out[19]:

	topic_id	0	1	2	3	4	5	6	
book_id	**chap_id**								
70	**1**	0.002273	0.002273	0.002273	0.002273	0.002273	0.002273	0.678106	0.002
	2	0.000005	0.000005	0.000005	0.000005	0.000005	0.000005	0.000005	0.000
	3	0.000036	0.000036	0.000036	0.000036	0.010489	0.000036	0.000036	0.000
	4	0.000038	0.000038	0.000038	0.000038	0.000038	0.000038	0.435486	0.000
	5	0.000028	0.000028	0.000028	0.000028	0.000028	0.000028	0.105762	0.000
...	**...**	...	...	...	...	...	...	...	
62739	**2**	0.000023	0.073183	0.036687	0.000023	0.005265	0.000023	0.379604	0.000
	3	0.000227	0.000227	0.000227	0.000227	0.000227	0.000227	0.411287	0.000
	4	0.000054	0.050824	0.000054	0.000054	0.025275	0.000054	0.653492	0.000
	5	0.000144	0.000144	0.000144	0.000144	0.000144	0.000144	0.692254	0.000
	6	0.000581	0.000581	0.000581	0.000581	0.000581	0.000581	0.513626	0.000

1108 rows × 40 columns

In [20]:
```
# distrubution of words over topics
tm.PHI
```

Out[20]:

term_str	german	ancient	allowed	art	thou	private	month	
topic_id								
0	8.025308	0.025000	0.025000	7.762330	0.025000	0.025000	0.025000	
1	0.025000	0.025000	0.025000	0.025000	0.025000	0.025000	0.025000	
2	0.025000	0.025000	27.847490	0.025000	0.025000	0.025000	0.025000	13
3	0.025000	0.025000	0.025000	3.552965	0.025000	0.025000	0.025000	
4	0.025000	0.025000	3.561333	0.859023	0.025000	0.025000	0.725470	
5	0.025000	0.025000	0.025000	0.025000	0.025000	0.025000	7.134957	
6	1.606724	58.063440	78.634398	65.521863	0.025000	60.510445	25.791504	
7	0.025000	1.403531	0.025000	0.025000	0.025000	2.251557	0.025000	
8	0.025000	15.812970	3.202335	5.097929	3.700189	12.573707	37.355050	
9	142.250248	52.863735	40.318434	109.878943	0.025000	41.759558	28.939902	
10	14.314383	23.783597	0.025000	0.025000	0.025000	0.025000	5.974309	16
11	0.025000	6.575866	13.534899	11.425903	2.618971	42.418284	10.252781	
12	0.025000	26.779278	0.025000	57.983577	325.409821	5.168668	0.025000	
13	39.439127	0.025000	0.025000	1.055050	1.027437	2.131427	0.025000	

term_str	german	ancient	allowed	art	thou	private	month	
topic_id								
14	0.025000	0.489465	4.510259	9.014678	0.025000	0.025000	23.411455	
15	23.423745	3.189324	25.701895	3.291962	0.025000	41.679256	107.879904	
16	0.025000	0.025000	0.037356	0.025000	0.025000	0.962268	0.025000	
17	0.025000	5.672838	13.243987	57.442185	0.025000	9.417616	0.025000	
18	10.479892	7.029229	8.349322	0.025000	0.025000	0.025000	0.025000	
19	0.025000	0.025000	3.775352	0.025000	0.025000	0.025000	0.025000	
20	4.107050	0.025000	0.025000	0.025000	0.025000	0.025000	0.025000	
21	0.025000	0.540559	0.025000	13.926766	0.025000	0.096237	0.025000	
22	10.543879	4.634565	1.686472	1.183793	0.025000	16.212240	0.557923	
23	0.025000	1.297191	0.025000	0.025000	0.025000	0.025000	0.025000	
24	11.474755	4.285487	0.025000	0.025000	0.025000	0.025000	0.025000	
25	6.360447	0.025000	1.401918	0.025000	0.025000	30.328421	0.025000	
26	0.025000	6.716472	19.627906	7.721242	0.025000	58.224471	0.025000	
27	0.025000	8.012943	13.513383	41.480125	65.362325	0.025000	0.025000	
28	0.025000	0.025000	47.382806	0.025000	0.025000	18.850184	0.025000	1
29	0.025000	8.749385	21.605058	0.025000	0.025000	8.264212	64.480949	1
30	0.025000	0.887764	12.717148	0.025000	0.025000	9.045312	1.674156	
31	132.133853	0.025000	0.025000	0.025000	0.025000	10.678900	0.899617	
32	0.025000	0.025000	0.025000	0.025000	0.025000	10.873960	12.503969	
33	0.025000	0.025000	0.025000	7.498789	0.025000	0.025000	7.400956	
34	0.025000	10.432333	0.025000	2.889008	0.025000	0.025000	0.025000	
35	0.025000	0.025000	0.025000	0.025000	0.025000	0.025000	10.402455	
36	0.025000	4.499162	61.774934	0.025000	0.025000	7.718947	19.768136	
37	0.025000	1.025000	0.025000	0.025000	0.025000	2.025000	0.025000	
38	3.165588	4.399512	0.025000	0.025000	0.025000	19.640129	46.296508	
39	0.025000	150.456353	4.098316	2.888871	12.031257	0.744201	0.025000	

40 rows × 2000 columns

In [21]:

```
tm.TOPIC.sort_values('theta_sum', ascending = False)
```

Out[21]:

	phi_sum	theta_sum	h	top_terms_rel	top_terms	label
topic_id						

topic_id	phi_sum	theta_sum	h	top_terms_rel	top_terms	label
36	56674.817260	127.577716	10.13	awake stillness cheer whispered sprang wound c...	boys sat voice answer followed fire tried brok...	36: awake stillness cheer whispered sprang wou...
6	44859.419532	88.110400	9.92	science reverence civilization government nati...	government history law race state nation human...	6: science reverence civilization government n...
10	40207.637360	84.184373	9.90	glacier steep summit rope scenery ice gloom mo...	distance mountain foot deep ground behind sun ...	10: glacier steep summit rope scenery ice gloo...
11	30476.677087	59.764951	9.89	earl sack song loving glanced confess gratitud...	father happy child herself sat voice wife stra...	11: earl sack song loving glanced confess grat...
28	19540.766202	51.483016	9.05	hes youre theyre nigger haint reckon coffin th...	hes reckon theres nigger youre wont duke youll...	28: hes youre theyre nigger haint reckon coffi...
9	23108.804196	49.600283	9.86	corps students student castle italian pictures...	pictures picture table german castle art fine ...	9: corps students student castle italian pictu...
30	16947.303924	43.172027	9.28	knights council sword wounded soldiers army so...	war battle army child march sent herself frenc...	30: knights council sword wounded soldiers arm...
26	19734.061371	41.714986	9.67	authors sincerely writer guest society hall re...	perhaps father wrote books society hall suppos...	26: authors sincerely writer guest society hal...
2	16895.707949	40.930780	9.07	raft canoe reckoned cave warnt scared knowed b...	warnt raft big boys mile begun run reckon minute	2: raft canoe reckoned cave warnt scared knowe...
15	30783.182917	39.717141	9.62	letters lecture 12 yours 7 10 9 magazine 3	letters write wrote written send story yours w...	15: letters lecture 12 yours 7 10 9 magazine 3
21	15361.746982	34.968000	9.23	bird wings brush birds tree shape hasnt begins...	tree black bird makes comes goes big heaven looks	21: bird wings brush birds tree shape hasnt be...
8	13636.500040	34.248539	9.43	husband baby married poetry marry wife disease...	wife child chapter husband friend married doct...	8: husband baby married poetry marry wife dise...
39	16583.089258	33.596436	9.14	pilgrims temple marble priests stone centuries...	stone church ancient marble walls pilgrims bui...	39: pilgrims temple marble priests stone centu...

topic_id	phi_sum	theta_sum	h	top_terms_rel	top_terms	label
38	12089.462258	30.269851	9.18	mining mines gold mill silver sold coal sell p...	gold silver rich worth sold mine mining mines ...	38: mining mines gold mill silver sold coal se...
25	11914.347299	29.169184	9.12	car conductor train seats railway railroad lad...	train car hotel lady ladies gentlemen public c...	25: car conductor train seats railway railroad...
29	11689.267902	28.193424	9.07	pilots pilot wages clerk association wheel cen...	pilot pay cent pilots boat wages bank clerk buy	29: pilots pilot wages clerk association wheel...
32	11099.074186	26.910419	9.24	editor minister police journal prison paper pa...	paper public editor school write office papers...	32: editor minister police journal prison pape...
12	11305.222550	26.289940	9.13	hath tis thou thy thee prince lad royal mad	thou thy thee prince hath none tis ye royal	12: hath tis thou thy thee prince lad royal mad
14	12234.126069	25.794035	8.62	vessel ships ship captain mate deck port passe...	ship captain sea boat island deck ships island...	14: vessel ships ship captain mate deck port p...
22	11200.302142	24.080291	9.41	stove fired gun negro killed guns armed shot m...	killed shot kill war horse stove officer box road	22: stove fired gun negro killed guns armed sh...
31	6793.966237	17.230533	8.87	vote committee bill german fourth measure coll...	bill german vote speech committee language nob...	31: vote committee bill german fourth measure ...
27	13266.099095	16.544906	9.58	thy hero madam stars heaven victory hopes pare...	thy heaven father thee voice soul alone woman ...	27: thy hero madam stars heaven victory hopes ...
24	4336.777031	14.208223	8.56	saddle horse game played journal funny morals ...	horse game played saddle memory dog stage reme...	24: saddle horse game played journal funny mor...
19	5515.768416	12.868162	7.65	jury trial witnesses verdict guilty judge pris...	judge court jury trial law evidence prisoner m...	19: jury trial witnesses verdict guilty judge ...
0	4497.493923	12.664865	8.79	lie shes twins madam taught lying finger baby ...	lie father child truth shes twins son school p...	0: lie shes twins madam taught lying finger ba...
17	4065.521650	12.361717	8.34	french marks literature reply printed funny te...	french american literature art article convers...	17: french marks literature reply printed funn...

topic_id	phi_sum	theta_sum	h	top_terms_rel	top_terms	label
20	2543.415053	11.481418	8.48	tobacco eat smoking smoke pause pipe ass drink...	eat smoke tobacco pause cat royal smoking pipe...	20: tobacco eat smoking smoke pause pipe ass d...
1	3976.042408	11.110674	8.35	natives native monument british indian populat...	native natives women indian monument british p...	1: natives native monument british indian popu...
5	4426.796219	10.822974	7.30	dey dat yo en den git bout er em	en dat dey den yo git nigger em bout	5: dey dat yo en den git bout er em
35	3342.756757	9.713668	8.34	uncle murdered corpse cabin murder lantern pri...	uncle brother cabin murder kill boys murdered ...	35: uncle murdered corpse cabin murder lantern...
4	3532.911084	9.562080	8.16	driver lake conductor mountains station mail s...	lake driver mountains stage station desert sno...	4: driver lake conductor mountains station mai...
18	3592.375037	8.454726	8.35	papa mamma cats frank cure funny acquired bath...	papa cats mamma remember lady cat prince table...	18: papa mamma cats frank cure funny acquired ...
7	1770.259201	8.391929	8.57	elephant dream telegram dreams angels suspecte...	elephant dream dreams office telegram arrived ...	7: elephant dream telegram dreams angels suspe...
33	2631.197347	8.253225	8.29	unto knight tale behold page fighting fought s...	unto tale knight story pass page women seven hair	33: unto knight tale behold page fighting foug...
16	2294.407743	6.795957	7.76	ye mob kingdom adventures confess interrupted ...	ye boys mob tree bad books master fair school	16: ye mob kingdom adventures confess interrup...
3	2569.318196	4.918893	8.09	frog chinese et bull bet individual motion cit...	frog chinese et bull bet citizens article floo...	3: frog chinese et bull bet individual motion ...
34	1544.790390	4.890485	8.43	excursion passengers voyage board steamer pray...	board excursion passengers reached visit voyag...	34: excursion passengers voyage board steamer ...
23	1665.885870	4.600938	7.57	1 knife 2 4 finger 3 5 palace notes	1 2 knife 4 3 girl grand finger letters	23: 1 knife 2 4 finger 3 5 palace notes
13	1621.287051	1.970443	8.23	ich cloth sheep pages aside er enter volume ge...	ich cloth die german aside pages sheep girls e...	13: ich cloth sheep pages aside er enter volum...

	phi_sum	theta_sum	h	top_terms_rel	top_terms	label
topic_id						
37	5405.414808	1.378391	8.87	om ym training impulse machine content instinc...	om ym outside training mans machine self spiri...	37: om ym training impulse machine content ins...

Top 5 terms associated with the most frequent topic

In [22]:
```
top_topic = tm.TOPIC.theta_sum.idxmax()

top_topic
```

Out[22]:
```
36
```

In [23]:
```
tm.TOPIC.sort_values('theta_sum', ascending = False).loc[top_topic, 'top_terms_r
```

Out[23]:
```
'awake stillness cheer whispered sprang wound crept hurried resolved'
```

In [24]:
```
# find topic (theta) that is most frequent (highest total prob across all docs)
top_five_terms = tm.TOPIC.sort_values('theta_sum', ascending = False).loc[top_to
```

In [25]:
```
top_five_terms
```

Out[25]:
```
['awake', 'stillness', 'cheer', 'whispered', 'sprang']
```

In [26]:
```
# join THETA and LIB tables
joint_theta = tm.THETA.join(LIB)

# add title column to index
joint_theta = joint_theta.set_index('title', append = True)

# drop other LIB cols and get mean topic distribution for each book
book_mean_theta = joint_theta.drop(joint_theta.loc[:, 'year':].columns, axis = 1

book_mean_theta.style.background_gradient(axis=None)
```

Out[26]:

			0	1	2	3	4	5	
book_id	**title**	**type**							
70	**what is man**	**non-fiction**	0.015837	0.000207	0.024724	0.000207	0.006011	0.000207	0
74	**the adventures of tom sawyer**	**novel**	0.014999	0.000651	0.130010	0.001421	0.000094	0.000426	0
76	**the adventures of huckleberry finn**	**novel**	0.007904	0.000073	0.370789	0.000357	0.000398	0.074444	0

book_id	title	type	0	1	2	3	4	5
86	a connecticut yankee in king arthurs court	novel	0.000081	0.000668	0.026693	0.000808	0.003958	0.000081
91	tom sawyer abroad	novel	0.000068	0.001430	0.546590	0.000437	0.014668	0.092809
93	tom sawyer detective	novel	0.003444	0.000096	0.272056	0.000096	0.000096	0.000096
102	the tragedy of puddnhead wilson	novel	0.033287	0.000502	0.000084	0.000084	0.000716	0.162329
119	a tramp abroad	non-fiction	0.025215	0.001831	0.007867	0.001248	0.005697	0.001029
142	the 30000 bequest and other stories	stories	0.047651	0.021243	0.003721	0.002725	0.001459	0.000107
245	life on the mississippi	non-fiction	0.004254	0.008460	0.041460	0.000363	0.023363	0.000891
1044	extract from captain stormfields visit to Heaven	stories	0.000021	0.000021	0.082854	0.000021	0.000021	0.000021
1086	a horses tale	novel	0.158913	0.002295	0.007775	0.016298	0.018987	0.001130
1837	the prince and the pauper	novel	0.000099	0.001300	0.000099	0.000814	0.000872	0.000099
2874	personal recollections of joan of arc vol 1	non-fiction	0.013243	0.002239	0.004701	0.000090	0.000090	0.000090
2875	personal recollections of joan of arc vol 2	non-fiction	0.002437	0.026268	0.000100	0.001516	0.000253	0.000281
2895	following the equator	non-fiction	0.007173	0.054910	0.003186	0.000760	0.012864	0.000062
3171	in defense of harriet shelley	non-fiction	0.000029	0.000029	0.000029	0.000029	0.001713	0.000029
3172	fenimore coopers literary offences	non-fiction	0.000028	0.000028	0.101781	0.000028	0.016329	0.000028
3173	essays on paul bourget	non-fiction	0.000029	0.020714	0.000029	0.000029	0.000029	0.000029
3176	the innocents abroad	non-fiction	0.002929	0.001412	0.001813	0.001334	0.007316	0.000137
3177	roughing it	novel	0.000554	0.010776	0.020091	0.003813	0.057298	0.000470

book_id	title	type	0	1	2	3	4	5
3178	the gilded age	novel	0.002170	0.024885	0.007543	0.000865	0.002244	0.005937
3179	the american claimant	novel	0.026801	0.002196	0.000763	0.000391	0.000068	0.003472
3180	a double barrelled detective story	stories	0.012571	0.000720	0.024443	0.000851	0.000168	0.000168
3181	the stolen white elephant	stories	0.000058	0.000058	0.000058	0.000058	0.013198	0.000058
3182	some rambling notes of an idle excursion	non-fiction	0.000035	0.000035	0.039449	0.000035	0.000035	0.000035
3183	the facts concerning the recent carnival of crime in connecticut	stories	0.000024	0.000024	0.000024	0.000024	0.011788	0.000024
3184	alonzo fitz and other stories	stories	0.038418	0.002598	0.006892	0.000085	0.000753	0.000174
3185	those extraordinary twins	stories	0.052447	0.000104	0.000104	0.001013	0.000104	0.000602
3186	the mysterious stranger and other stories	stories	0.000082	0.002489	0.017501	0.000082	0.000082	0.000082
3188	mark twain speeches	non-fiction	0.004950	0.014325	0.008165	0.001100	0.002569	0.000683
3189	sketches new and old	stories	0.020941	0.001853	0.008130	0.039235	0.003288	0.017860
3190	1601 conversation as it was by the social fireside in the time of the tudors	stories	0.000079	0.003155	0.000079	0.000079	0.000079	0.000079
3191	goldsmiths friend abroad again	stories	0.000255	0.022364	0.000255	0.014304	0.000255	0.000255
3192	the curious republic of gondour and other whimsical sketches	stories	0.000296	0.000296	0.009651	0.000296	0.002430	0.000296

book_id	title	type	0	1	2	3	4	5	
3199	the letters of mark twain	non-fiction	0.002300	0.000659	0.000636	0.002814	0.003943	0.000079	C
3250	how to tell a story and other essays	non-fiction	0.000135	0.000135	0.000135	0.000135	0.000135	0.198836	C
3251	the man that corrupted hadleyburg and other stories	stories	0.024478	0.003901	0.001486	0.008960	0.000970	0.000053	0
19484	editorial wild oats	stories	0.000104	0.000104	0.000104	0.198686	0.000104	0.010594	C
19987	chapters from my autobiography	non-fiction	0.005820	0.002969	0.003783	0.003217	0.002543	0.002551	C
33077	the treaty with china its provisions explained	non-fiction	0.000020	0.006284	0.000020	0.196933	0.000020	0.000020	0
60900	merry tales	stories	0.000035	0.000035	0.017146	0.000035	0.000035	0.000035	C
61522	the 1000000 bank note	stories	0.000016	0.000016	0.000016	0.000016	0.000016	0.000016	0
62636	to the person sitting in darkness	non-fiction	0.000031	0.000031	0.000031	0.012267	0.000031	0.000031	
62739	king leopolds soliloquy	stories	0.000193	0.020848	0.006304	0.000193	0.005271	0.000193	0

In [27]:

```
# most common topics by work type
book_mean_theta.groupby('type').mean().idxmax(axis = 1)
```

Out[27]:

```
type
non-fiction     6
novel          36
stories        36
dtype: int64
```

In [56]:

```
tm.TOPIC.loc[11]
```

Out[56]:

```
phi_sum                                                  30476.677087
theta_sum                                                   59.764951
h                                                                9.89
top_terms_rel     earl sack song loving glanced confess gratitud...
top_terms         father happy child herself sat voice wife stra...
label             11: earl sack song loving glanced confess grat...
Name: 11, dtype: object
```

In [28]:

```
# table with most popular topic for each book --> rename new col created to topi
max_topic = book_mean_theta.apply(lambda x: x.idxmax(), axis = 1).reset_index().
```

```
# join with tm.TOPIC for words for each topic
max_topic = max_topic.join(tm.TOPIC).reset_index().set_index('book_id')

max_topic['top_five_terms'] = max_topic.apply(lambda x: x.top_terms_rel.split()[

max_topic.sort_values('topic_id', ascending = False).drop('label', axis = 1).sty
```

Out[28]:

book_id	topic_id	title	type	phi_sum	theta_sum	h	top_terms_rel	t
3176	39	the innocents abroad	non-fiction	16583.089258	33.596436	9.140000	pilgrims temple marble priests stone centuries sacred walls priest	ma pilg
19987	36	chapters from my autobiography	non-fiction	56674.817260	127.577716	10.130000	awake stillness cheer whispered sprang wound crept hurried resolved	fol ti
3191	36	goldsmiths friend abroad again	stories	56674.817260	127.577716	10.130000	awake stillness cheer whispered sprang wound crept hurried resolved	fol ti
3185	36	those extraordinary twins	stories	56674.817260	127.577716	10.130000	awake stillness cheer whispered sprang wound crept hurried resolved	fol ti
3182	36	some rambling notes of an idle excursion	non-fiction	56674.817260	127.577716	10.130000	awake stillness cheer whispered sprang wound crept hurried resolved	fol ti
3180	36	a double barrelled detective story	stories	56674.817260	127.577716	10.130000	awake stillness cheer whispered sprang wound crept hurried resolved	fol ti
102	36	the tragedy of puddnhead wilson	novel	56674.817260	127.577716	10.130000	awake stillness cheer whispered sprang wound crept hurried resolved	fol ti

book_id	topic_id	title	type	phi_sum	theta_sum	h	top_terms_rel	t
86	36	a connecticut yankee in king arthurs court	novel	56674.817260	127.577716	10.130000	awake stillness cheer whispered sprang wound crept hurried resolved	fol t
74	36	the adventures of tom sawyer	novel	56674.817260	127.577716	10.130000	awake stillness cheer whispered sprang wound crept hurried resolved	fol t
3250	33	how to tell a story and other essays	non-fiction	2631.197347	8.253225	8.290000	unto knight tale behold page fighting fought sons spell	kn p s
3189	32	sketches new and old	stories	11099.074186	26.910419	9.240000	editor minister police journal prison paper papers ball press	pa sc
19484	32	editorial wild oats	stories	11099.074186	26.910419	9.240000	editor minister police journal prison paper papers ball press	pa sc
2875	30	personal recollections of joan of arc vol 2	non-fiction	16947.303924	43.172027	9.280000	knights council sword wounded soldiers army soldier battle commanded	v a m
2874	30	personal recollections of joan of arc vol 1	non-fiction	16947.303924	43.172027	9.280000	knights council sword wounded soldiers army soldier battle commanded	v a m
93	28	tom sawyer detective	novel	19540.766202	51.483016	9.050000	hes youre theyre nigger haint reckon coffin theyve theyd	h nig v y
76	28	the adventures of huckleberry finn	novel	19540.766202	51.483016	9.050000	hes youre theyre nigger haint reckon coffin theyve theyd	h nig v y

book_id	topic_id	title	type	phi_sum	theta_sum	h	top_terms_rel	to
61522	27	the 1000000 bank note	stories	13266.099095	16.544906	9.580000	thy hero madam stars heaven victory hopes parents divine	th fa v wo
3178	26	the gilded age	novel	19734.061371	41.714986	9.670000	authors sincerely writer guest society hall replied affairs distinction	fat so qui
60900	22	merry tales	stories	11200.302142	24.080291	9.410000	stove fired gun negro killed guns armed shot murdered	k ho o
1086	21	a horses tale	novel	15361.746982	34.968000	9.230000	bird wings brush birds tree shape hasnt begins color	b co b
1044	21	extract from captain stormfields visit to Heaven	stories	15361.746982	34.968000	9.230000	bird wings brush birds tree shape hasnt begins color	b co b
3190	17	1601 conversation as it was by the social fireside in the time of the tudors	stories	4065.521650	12.361717	8.340000	french marks literature reply printed funny teach artist article	lite cor m
3199	15	the letters of mark twain	non-fiction	30783.182917	39.717141	9.620000	letters lecture 12 yours 7 10 9 magazine 3	let wri st
1837	12	the prince and the pauper	novel	11305.222550	26.289940	9.130000	hath tis thou thy thee prince lad royal mad	th l ti
3186	11	the mysterious stranger and other stories	stories	30476.677087	59.764951	9.890000	earl sack song loving glanced confess gratitude foolish sigh	fatl ch stra
3184	11	alonzo fitz and other stories	stories	30476.677087	59.764951	9.890000	earl sack song loving glanced confess gratitude foolish sigh	fatl ch stra

book_id	topic_id	title	type	phi_sum	theta_sum	h	top_terms_rel	tc
3183	11	the facts concerning the recent carnival of crime in connecticut	stories	30476.677087	59.764951	9.890000	earl sack song loving glanced confess gratitude foolish sigh	fatl ch stra
3179	11	the american claimant	novel	30476.677087	59.764951	9.890000	earl sack song loving glanced confess gratitude foolish sigh	fatl ch stra
3177	10	roughing it	novel	40207.637360	84.184373	9.900000	glacier steep summit rope scenery ice gloom mountain huge	b
245	10	life on the mississippi	non-fiction	40207.637360	84.184373	9.900000	glacier steep summit rope scenery ice gloom mountain huge	b
119	10	a tramp abroad	non-fiction	40207.637360	84.184373	9.900000	glacier steep summit rope scenery ice gloom mountain huge	b
3171	8	in defense of harriet shelley	non-fiction	13636.500040	34.248539	9.430000	husband baby married poetry marry wife disease chapter aged	c
3181	7	the stolen white elephant	stories	1770.259201	8.391929	8.570000	elephant dream telegram dreams angels suspected newspapers trunk brick	arr cc
70	6	what is man	non-fiction	44859.419532	88.110400	9.920000	science reverence civilization government nations political religious british insane	gc h ı pc

book_id	topic_id	title	type	phi_sum	theta_sum	h	top_terms_rel	tc
3188	6	mark twain speeches	non-fiction	44859.419532	88.110400	9.920000	science reverence civilization government nations political religious british insane	gc h pc
142	6	the 30000 bequest and other stories	stories	44859.419532	88.110400	9.920000	science reverence civilization government nations political religious british insane	gc h pc
2895	6	following the equator	non-fiction	44859.419532	88.110400	9.920000	science reverence civilization government nations political religious british insane	gc h pc
3172	6	fenimore coopers literary offences	non-fiction	44859.419532	88.110400	9.920000	science reverence civilization government nations political religious british insane	gc h pc
3173	6	essays on paul bourget	non-fiction	44859.419532	88.110400	9.920000	science reverence civilization government nations political religious british insane	gc h pc
62739	6	king leopolds soliloquy	stories	44859.419532	88.110400	9.920000	science reverence civilization government nations political religious british insane	gc h pc
3192	6	the curious republic of gondour and other whimsical sketches	stories	44859.419532	88.110400	9.920000	science reverence civilization government nations political religious british insane	gc h pc

	topic_id	title	type	phi_sum	theta_sum	h	top_terms_rel	t
book_id								
3251	6	the man that corrupted hadleyburg and other stories	stories	44859.419532	88.110400	9.920000	science reverence civilization government nations political religious british insane	gc h i pc
33077	6	the treaty with china its provisions explained	non-fiction	44859.419532	88.110400	9.920000	science reverence civilization government nations political religious british insane	gc h i pc
62636	6	to the person sitting in darkness	non-fiction	44859.419532	88.110400	9.920000	science reverence civilization government nations political religious british insane	gc h i pc
91	2	tom sawyer abroad	novel	16895.707949	40.930780	9.070000	raft canoe reckoned cave warnt scared knowed begun woods	m r

Works and Top Terms Associated with Each Topic

In [64]:

```
# set option so that columns not truncated
pd.set_option('display.max_colwidth', None)
```

In [65]:

```
works_df = max_topic.groupby('topic_id').agg({'topic_id': 'size', 'title': lambd
                    .rename({'topic_id': 'count'}, axis = 1) \
                    .sort_values('count', ascending = False)

works_df['top_terms_rel'] = tm.TOPIC.top_terms_rel

works_df
```

Out[65]:

	count	title	top_terms_rel
topic_id			

topic_id	count	title	top_terms_rel
6	11	what is man, the 30000 bequest and other stories, following the equator, fenimore coopers literary offences, essays on paul bourget, mark twain speeches, the curious republic of gondour and other whimsical sketches, the man that corrupted hadleyburg and other stories, the treaty with china its provisions explained, to the person sitting in darkness, king leopolds soliloquy	science reverence civilization government nations political religious british insane
36	8	the adventures of tom sawyer, a connecticut yankee in king arthurs court, the tragedy of puddnhead wilson, a double barrelled detective story, some rambling notes of an idle excursion, those extraordinary twins, goldsmiths friend abroad again, chapters from my autobiography	awake stillness cheer whispered sprang wound crept hurried resolved
11	4	the american claimant, the facts concerning the recent carnival of crime in connecticut, alonzo fitz and other stories, the mysterious stranger and other stories	earl sack song loving glanced confess gratitude foolish sigh
10	3	a tramp abroad, life on the mississippi, roughing it	glacier steep summit rope scenery ice gloom mountain huge
21	2	extract from captain stormfields visit to Heaven, a horses tale	bird wings brush birds tree shape hasnt begins color
32	2	sketches new and old, editorial wild oats	editor minister police journal prison paper papers ball press
30	2	personal recollections of joan of arc vol 1, personal recollections of joan of arc vol 2	knights council sword wounded soldiers army soldier battle commanded
28	2	the adventures of huckleberry finn, tom sawyer detective	hes youre theyre nigger haint reckon coffin theyve theyd
26	1	the gilded age	authors sincerely writer guest society hall replied affairs distinction

topic_id	count	title	top_terms_rel
33	1	how to tell a story and other essays	unto knight tale behold page fighting fought sons spell
27	1	the 1000000 bank note	thy hero madam stars heaven victory hopes parents divine
2	1	tom sawyer abroad	raft canoe reckoned cave warnt scared knowed begun woods
22	1	merry tales	stove fired gun negro killed guns armed shot murdered
17	1	1601 conversation as it was by the social fireside in the time of the tudors	french marks literature reply printed funny teach artist article
15	1	the letters of mark twain	letters lecture 12 yours 7 10 9 magazine 3
12	1	the prince and the pauper	hath tis thou thy thee prince lad royal mad
8	1	in defense of harriet shelley	husband baby married poetry marry wife disease chapter aged
7	1	the stolen white elephant	elephant dream telegram dreams angels suspected newspapers trunk brick
39	1	the innocents abroad	pilgrims temple marble priests stone centuries sacred walls priest

In [66]:

```
# reset width to default: https://pandas.pydata.org/docs/user_guide/options.html
pd.set_option('display.max_colwidth', 50)
```

M09: Word Embeddings

In [29]:

```
w2v_params = dict(
    min_count = 10,
    workers = 1,
    # vector_size = 246,
    vector_size = 100,
    window = 2
)
```

In [30]:

```
SENTS = CORPUS.groupby(OHCO[:-1]).term_str.apply(lambda  x:  x.tolist())
```

In [31]:

```
model = word2vec.Word2Vec(SENTS.values, **w2v_params)
```

In [32]:

```
W2V = pd.DataFrame(model.wv.get_normed_vectors(), index=model.wv.index_to_key)
W2V.index.name = 'term_str'
W2V = W2V.sort_index()
```

In [33]:

```
W2V.head()
```

Out[33]:

term_str	0	1	2	3	4	5	6	7
04	-0.114225	0.095736	0.051546	0.055306	0.074904	-0.105503	0.057929	0.241827
08	-0.101873	0.127570	0.037040	0.017632	0.039355	-0.123453	0.033405	0.277762
1	-0.106604	0.047069	0.027273	0.010269	0.007156	-0.112879	0.013426	0.232790
10	-0.083764	-0.035320	0.101737	-0.096142	0.090019	-0.128170	-0.072589	0.229303
100	-0.125445	0.081001	0.076093	-0.188683	0.015055	-0.188129	-0.080429	0.243127

5 rows × 100 columns

In [34]:

```
tsne_params = dict(
    learning_rate = 200., #'auto' or [10.0, 1000.0]
    perplexity = 40,
    n_components = 2,
    init = 'random', # 'pca'
    n_iter = 2500,
    random_state = 23
)
```

In [35]:

```
tsne_engine = TSNE(**tsne_params)
tsne_model = tsne_engine.fit_transform(W2V)
```

In [36]:

```
COORDS = pd.DataFrame(tsne_model, columns=['x','y'], index=W2V.index).join(VOCAB
```

In [37]:

```
COORDS['log_n'] = np.log(COORDS['n'])
```

In [38]:

```
COORDS
```

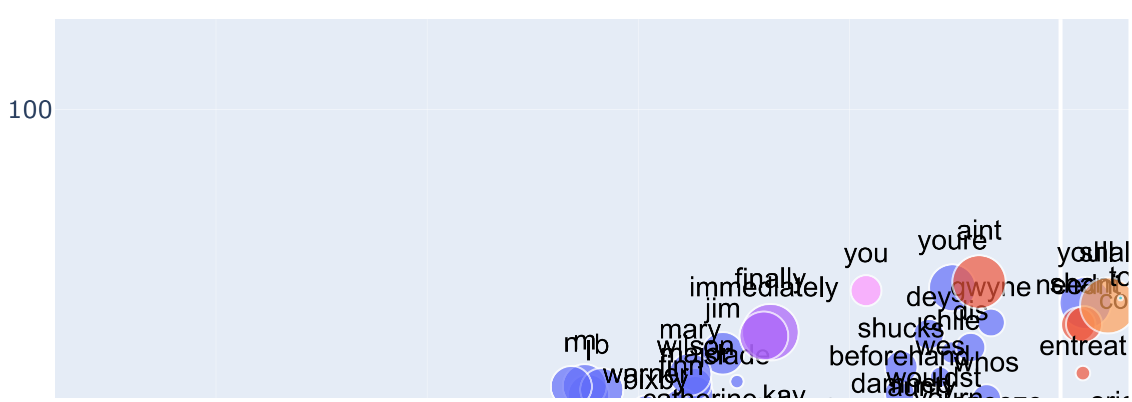
Out[38]:

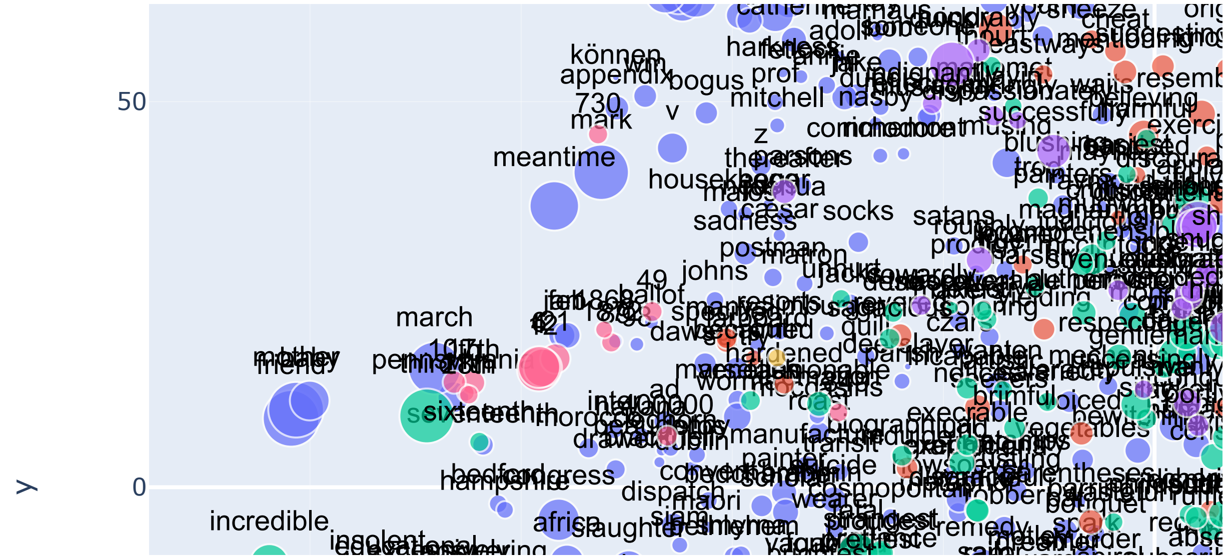
	x	y	n	dfidf	pos_group	log_n
term_str						
04	-49.950764	19.800415	10.0	18.227484	NN	2.302585
08	-50.145111	19.059072	10.0	10.113742	NN	2.302585
1	-57.705616	17.749544	331.0	428.368264	CD	5.802118
10	-58.390575	15.621562	135.0	288.917371	CD	4.905275
100	-55.285233	11.433803	62.0	181.458686	CD	4.127134
...	...	...	...	...	...	...
zest	-12.579935	5.808639	12.0	67.918141	NN	2.484907
zu	-51.100506	52.666790	22.0	25.586339	NN	3.091042
à	-56.147686	51.326492	44.0	51.144711	NN	3.784190
était	-53.847881	50.613708	13.0	10.113742	NN	2.564949
NaN	-14.547632	61.285461	NaN	NaN	NaN	NaN

13676 rows × 6 columns

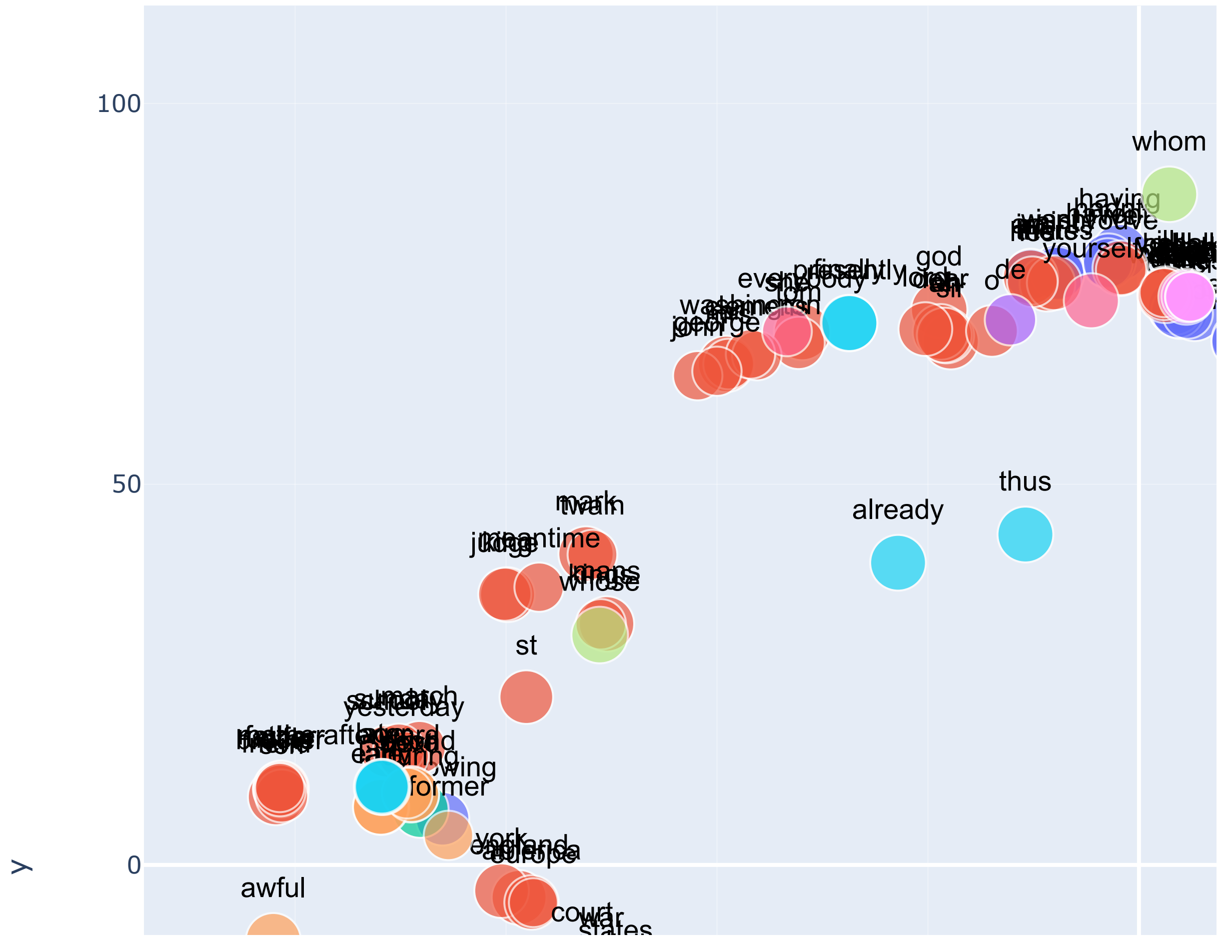
In [39]:

```
px.scatter(COORDS.reset_index().sample(1000),
           'x', 'y',
           text='term_str',
           color='pos_group',
           hover_name='term_str',
           size='dfidf',
           height=1000).update_traces(
                mode='markers+text',
                textfont=dict(color='black', size=14, family='Arial'),
                textposition='top center')
```





In [40]:
```
px.scatter(COORDS.reset_index().sort_values('dfidf', ascending=False).head(1000)
           'x', 'y',
           text='term_str',
           color='pos_group',
           hover_name='term_str',
           size='dfidf',
           height=1000).update_traces(
                mode='markers+text',
                textfont=dict(color='black', size=14, family='Arial'),
                textposition='top center')
```


100
50
0
y
whom
already
thus
mark
twain
st
god
de
o
former
awful

With Nouns Only (not proper ones)

In [41]:
```
noun_COORDS = COORDS.loc[COORDS.pos_group == 'NN']

noun_COORDS
```

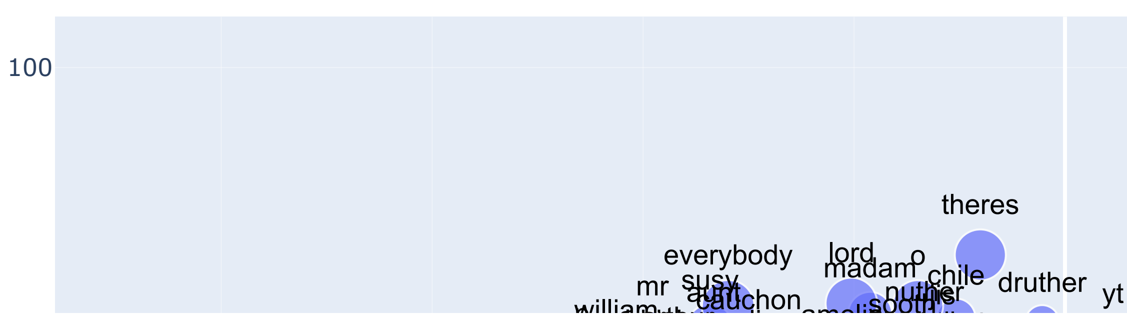
Out[41]:

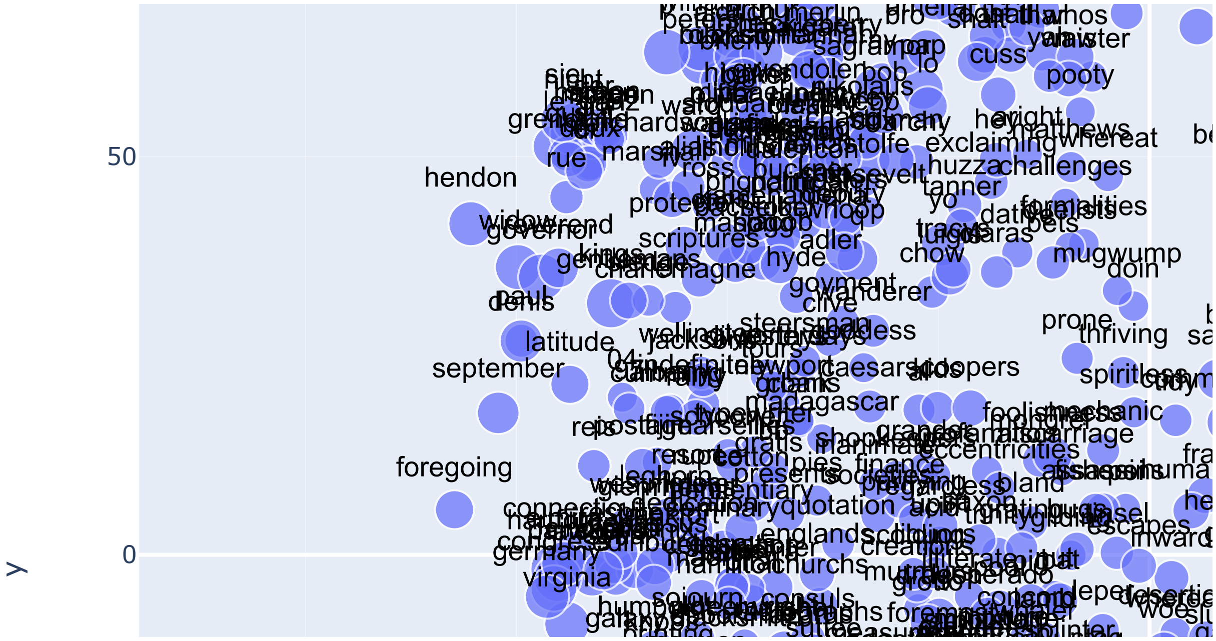
term_str	x	y	n	dfidf	pos_group	log_n
04	-49.950764	19.800415	10.0	18.227484	NN	2.302585
08	-50.145111	19.059072	10.0	10.113742	NN	2.302585
350	-53.410347	9.169445	24.0	67.918141	NN	3.178054
87	-49.367786	18.569710	13.0	38.959070	NN	2.564949
89	-51.060143	18.927137	15.0	38.959070	NN	2.708050
...	...	...	...	...	...	...
zermatt	34.265057	-67.250267	46.0	67.918141	NN	3.828641
zest	-12.579935	5.808639	12.0	67.918141	NN	2.484907
zu	-51.100506	52.666790	22.0	25.586339	NN	3.091042
à	-56.147686	51.326492	44.0	51.144711	NN	3.784190
était	-53.847881	50.613708	13.0	10.113742	NN	2.564949

7916 rows × 6 columns

Noun tSNE plot

In [67]:
```
px.scatter(noun_COORDS.reset_index().sample(1000),
           'x', 'y',
           text='term_str',
           color='pos_group',
           hover_name='term_str',
           size = 'log_n',
           height=1000).update_traces(
                mode='markers+text',
                textfont=dict(color='black', size=14, family='Arial'),
                textposition='top center')
```





Clusters in Nouns Plot

- Morning, summer, hour, seconds, times, ages → time (of day, year)
- Care, excuse, play, blow, cheer → carefree, mischievious, antics
- Honesty, ability, affection, powers, protection, worship → reverence, ability
- Accord, stupidity, piety, criticisms, devotions, genuineness, prosperity → conflicting views on religion, agreement

- Plunder, ordeal, crusades, conflagrations, pilgrimage, caution, tranquility → conflicting faces of religious activities throughout history

Analogies and Similarities (vector algebra)

In [43]:

```
def complete_analogy(A, B, C, n=2):
    try:
        cols = ['term', 'sim']
        return pd.DataFrame(model.wv.most_similar(positive=[B, C], negative=[A])
    except KeyError as e:
        print('Error:', e)
        return None

def get_most_similar(positive, negative=None):
    return pd.DataFrame(model.wv.most_similar(positive, negative), columns=['ter
```

In [44]:

```
complete_analogy('man', 'boy', 'woman', 3)
```

Out[44]:

	term	sim
0	child	0.776828
1	girl	0.767200
2	lady	0.722114

In [45]:

```
complete_analogy('girl', 'daughter', 'boy', 3)
```

Out[45]:

	term	sim
0	brother	0.823369
1	sister	0.801899
2	darling	0.779816

In [46]:

```
complete_analogy('girl', 'sister', 'boy', 3)
```

Out[46]:

	term	sim
0	darling	0.804392
1	brother	0.759961
2	liege	0.728171

In [47]:

```
complete_analogy('man', 'gentleman', 'woman', 5)
```

Out[47]:

	term	sim
0	lady	0.878961
1	girl	0.817366

	term	sim
2	fellow	0.735923
3	soldier	0.704647
4	farmer	0.695843

In [48]:
```
complete_analogy('woman', 'lady', 'man', 5)
```

Out[48]:

	term	sim
0	gentleman	0.824686
1	master	0.699578
2	citizen	0.692256
3	person	0.685369
4	stranger	0.683805

In [49]:
```
complete_analogy('day', 'sun', 'night', 5)
```

Out[49]:

	term	sim
0	rain	0.783103
1	wind	0.757743
2	darkness	0.744765
3	storm	0.722417
4	curtain	0.719862

In [68]:
```
complete_analogy('king', 'rich', 'servant', 5)
```

Out[68]:

	term	sim
0	slender	0.711932
1	graceful	0.702313
2	handsome	0.695950
3	splendid	0.687642
4	fat	0.687409

In [69]:
```
complete_analogy('lord', 'rich', 'servant', 5)
```

Out[69]:

	term	sim
0	handsome	0.720002
1	graceful	0.715430

	term	sim
2	slender	0.702366
3	coarse	0.686488
4	dumb	0.669372

In [70]:
```
complete_analogy('man', 'journey', 'woman', 5)
```

Out[70]:

	term	sim
0	voyage	0.706367
1	trip	0.658042
2	stretch	0.630050
3	spring	0.614480
4	flight	0.605543

In [71]:
```
complete_analogy('woman', 'marriage', 'man', 5)
```

Out[71]:

	term	sim
0	commission	0.753536
1	services	0.752548
2	birth	0.733643
3	powers	0.730161
4	departure	0.726789

In [72]:
```
complete_analogy('man', 'property', 'woman', 5)
```

Out[72]:

	term	sim
0	affairs	0.755897
1	rights	0.741495
2	society	0.733355
3	sorrow	0.725145
4	religion	0.721457

In [73]:
```
complete_analogy('man', 'fool', 'woman', 5)
```

Out[73]:

	term	sim
0	devil	0.696721
1	child	0.651625

	term	sim
2	lad	0.635922
3	girl	0.624696
4	beggar	0.623952

In [74]:
```
complete_analogy('woman', 'fool', 'man', 5)
```

Out[74]:

	term	sim
0	person	0.644702
1	hurry	0.603647
2	dog	0.591042
3	stranger	0.585422
4	chance	0.574519

In [75]:
```
complete_analogy('man', 'wise', 'woman', 5)
```

Out[75]:

	term	sim
0	innocent	0.697347
1	foolish	0.680637
2	brave	0.677902
3	simple	0.635330
4	ignorant	0.630998

In [76]:
```
complete_analogy('woman', 'wise', 'man', 5)
```

Out[76]:

	term	sim
0	worthy	0.670954
1	reasonable	0.644145
2	useful	0.641109
3	correct	0.634468
4	likely	0.620507

Similarites

In [50]:
```
get_most_similar('joy')
```

Out[50]:

	term	sim

	term	sim
0	delight	0.801221
1	admiration	0.784409
2	gratitude	0.747991
3	sorrow	0.747686
4	astonishment	0.731498
5	fright	0.726452
6	blessing	0.722288
7	spirit	0.719269
8	glory	0.714442
9	excitement	0.705790

In [51]:

```
get_most_similar('man')
```

Out[51]:

	term	sim
0	person	0.850920
1	gentleman	0.794088
2	woman	0.766792
3	stranger	0.740261
4	dog	0.718034
5	fellow	0.690010
6	fool	0.664921
7	citizen	0.647081
8	girl	0.645880
9	slave	0.635982

In [52]:

```
get_most_similar(positive=['man'], negative=['woman'])
```

Out[52]:

	term	sim
0	money	0.344735
1	business	0.268716
2	necessary	0.265689
3	government	0.258725
4	chance	0.255060
5	yourself	0.251223
6	public	0.250115

	term	sim
7	wrong	0.246364
8	going	0.243778
9	further	0.243301

In [53]:

```
get_most_similar(positive='woman')
```

Out[53]:

	term	sim
0	girl	0.866291
1	gentleman	0.831650
2	lady	0.820941
3	fellow	0.811571
4	man	0.766792
5	soldier	0.765872
6	person	0.760113
7	creature	0.756210
8	slave	0.755404
9	child	0.738329

In [54]:

```
get_most_similar(positive=['woman'], negative=['man'])
```

Out[54]:

	term	sim
0	young	0.452947
1	sweet	0.428815
2	friendless	0.419484
3	sister	0.403715
4	gray	0.398251
5	jane	0.391606
6	colored	0.371316
7	husband	0.370074
8	peasant	0.368533
9	old	0.367854

In [55]:

```
get_most_similar(['man','woman'],['boy','girl'])
```

Out[55]:

	term	sim
0	free	0.305291

	term	sim
1	human	0.292779
2	neither	0.269210
3	nor	0.239546
4	honorable	0.234591
5	an	0.230809
6	lack	0.229228
7	reasonable	0.228167
8	utter	0.226937
9	independent	0.226737

In [57]:
```
get_most_similar('knowledge')
```

Out[57]:

	term	sim
0	quality	0.846008
1	method	0.830475
2	genius	0.826863
3	system	0.826707
4	statement	0.825331
5	invention	0.824492
6	importance	0.820199
7	language	0.815986
8	crime	0.814865
9	wisdom	0.811844

In [58]:
```
get_most_similar('rich')
```

Out[58]:

	term	sim
0	handsome	0.765206
1	graceful	0.745025
2	pure	0.732829
3	charming	0.727296
4	nice	0.724187
5	picturesque	0.721800
6	neat	0.716759
7	comely	0.710631

	term	sim
8	fine	0.710626
9	beautiful	0.706341

In [59]:
```
get_most_similar('poor')
```

Out[59]:

	term	sim
0	brave	0.669315
1	young	0.642734
2	friendless	0.611581
3	devil	0.585920
4	sick	0.568551
5	weak	0.567297
6	gentle	0.563504
7	child	0.557940
8	girl	0.547192
9	innocent	0.546639

In [77]:
```
get_most_similar('money')
```

Out[77]:

	term	sim
0	trouble	0.772665
1	food	0.682620
2	stock	0.677449
3	orders	0.652256
4	delay	0.649710
5	use	0.641453
6	chance	0.635294
7	wages	0.626738
8	profit	0.626591
9	purpose	0.626240

Sources

- Dropping multiple columns by name starting with `drop` and `loc` : https://www.geeksforgeeks.org/how-to-drop-one-or-multiple-columns-in-pandas-dataframe/

- Adding a new index level from the columns of a dataframe: https://stackoverflow.com/questions/14744068/prepend-a-level-to-a-pandas-multiindex

In [ ]:

```

```

In [ ]:

```

```