Full Corpus 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("full_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(("full_LIB.csv"), index_col = ['book_id'])
```

In [6]:
```
CORPUS = pd.read_csv(("full_CORPUS.csv"), index_col = OHCO)
```

In [7]:
```
VOCAB = pd.read_csv("full_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]:
```
VOCAB.head()
```

Out[8]:

term_str	term_rank	n	n_chars	p	i	max_pos	n_pos	cat_pos	stop	ste

term_str	term_rank	n	n_chars	p	i	max_pos	n_pos	cat_pos	stop	ste
the	1	418963	3	0.052764	4.244302	DT	22	{'PRP', 'FW', 'RB', 'NN', 'JJS', 'NNP', 'VBZ',...	1	
and	2	310105	3	0.039054	4.678368	CC	20	{'PRP', 'FW', 'RB', 'PDT', 'NN', 'NNP', 'VBZ',...	1	
of	3	218996	2	0.027580	5.180221	IN	19	{'PRP', 'FW', 'RB', 'PDT', 'NN', 'NNP', 'VBZ',...	1	
to	4	206700	2	0.026032	5.263587	TO	23	{'WDT', 'FW', 'RB', 'PDT', 'NN', 'NNP', 'VBZ',...	1	
a	5	189310	1	0.023842	5.390375	DT	21	{'RBR', 'PRP', 'FW', 'RB', 'NN', 'NNP', 'VBZ',...	1	

5 rows × 21 columns

In [9]:
```
BOW.head()
```

Out[9]:

book_id	chap_id	term_str	n	tf	tfidf
70	**1**	**1835**	1	0.142857	1.262743
		1910	1	0.142857	1.225167
		a	2	0.285714	0.001081
		alphabet	1	0.142857	0.873820
		as	2	0.285714	0.007080

In [10]:
```
LIB.head()
```

Out[10]:

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
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 [11]:
```
# 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[11]:

term_str	book_id	chap_id	n	tf	tfidf	term_rank	n_vocab	n_chars	
0	588	7	2	0.040000	0.326445	7549	65	1	8.186(
0	786	16	1	0.012987	0.105989	7549	65	1	8.186(
0	882	47	1	0.001244	0.010151	7549	65	1	8.186(
0	912	3	3	0.005714	0.046635	7549	65	1	8.186(

	book_id	chap_id	n	tf	tfidf	term_rank	n_vocab	n_chars	
term_str									
0	1414	1	49	0.182836	1.492147	7549	65	1	8.186(
...	...	...	...	...	...	...	...	...	
étouffante	60900	5	1	0.007752	0.086520	50882	1	10	1.259:
évitant	3189	3	1	0.004132	0.046120	50885	1	7	1.259:
êtes	3189	3	1	0.004132	0.046120	50890	1	4	1.259:
öffnen	60900	6	1	0.004608	0.051434	50891	1	6	1.259:
übergeschlagen	60900	6	1	0.004608	0.051434	50895	1	14	1.259:

2213701 rows × 26 columns

In [12]:

```
# 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[12]:

			n	tf	tfidf	term_rank
book_id	chap_id	term_str				
70	**17**	**theological**	1	0.000749	0.005123	14167
	3	**vague**	1	0.004739	0.015596	2509
	12	**article**	1	0.062500	0.181779	1280
	2	**miserable**	5	0.003658	0.008645	1053
	4	**miserable**	1	0.004065	0.009608	1053
...	**...**	**...**	...	...	...	...
62739	**2**	**buckets**	2	0.005556	0.035590	11706
	4	**beguiled**	1	0.003534	0.018828	8063
		number	1	0.003534	0.006499	638
		afterward	1	0.003534	0.011499	2069
	2	**subheadings**	1	0.002778	0.031003	58356

2213701 rows × 4 columns

In [13]:
```
# removed ~ 3.5% of data when taking out proper nouns (singular and plural)
(BOW.shape[0] - filtered_BOW.shape[0]) / BOW.shape[0]
```

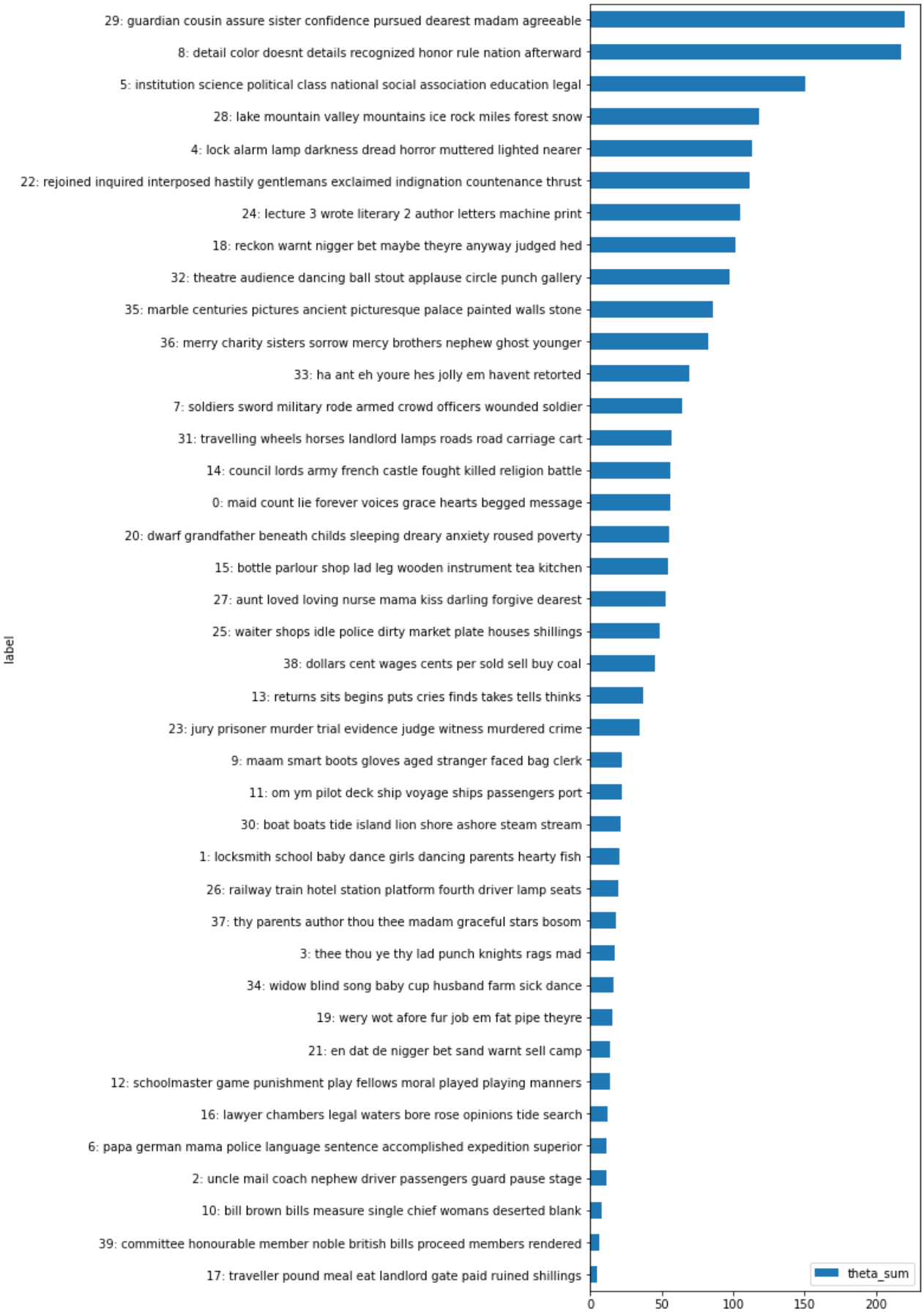
Out[13]:
```
0.04075437915339256
```

In [14]:
```
n_topics = 40
n_terms = 2000
```

In [15]:
```
tm = TopicModel(filtered_BOW)
tm.n_topics = n_topics
tm.n_terms = n_terms
```

In [16]:
```
tm.create_X()
tm.get_model()
tm.describe_topics()
tm.get_model_stats()
```

In [17]:
```
tm.plot_topics()
```



In [18]:
```
# table with distribution of topics for each doc
tm.THETA
```

Out[18]:

book_id	chap_id	topic_id 0	1	2	3	4	5	6	
70	1	0.002500	0.002500	0.002500	0.002500	0.002500	0.679131	0.002500	0.002
	2	0.024762	0.000005	0.000005	0.000005	0.000005	0.117217	0.000005	0.000
	3	0.130919	0.000040	0.000040	0.000040	0.000040	0.000040	0.029978	0.000
	4	0.000045	0.000045	0.000045	0.005706	0.000045	0.126276	0.000045	0.05
	5	0.000032	0.000032	0.000032	0.000032	0.000032	0.000032	0.000032	0.000
...	...	...	...	...	...	...	...	...	
62739	2	0.027685	0.000027	0.000027	0.000027	0.000027	0.167650	0.000027	0.20
	3	0.068494	0.000269	0.000269	0.088688	0.000269	0.109124	0.000269	0.000
	4	0.000063	0.000063	0.000063	0.000063	0.000063	0.496019	0.000063	0.05
	5	0.000179	0.000179	0.000179	0.000179	0.000179	0.654181	0.000179	0.066
	6	0.000714	0.000714	0.000714	0.000714	0.000714	0.000714	0.000714	0.00

2290 rows × 40 columns

In [19]:

```
# distrubution of words over topics
tm.PHI
```

Out[19]:

topic_id	term_str knowing	shaking	consider	twelve	closed	cry	shoulder	
0	29.803706	2.632424	33.944118	12.519358	7.809385	37.463780	2.403779	0
1	0.025000	0.025000	0.025000	0.025000	0.025000	35.215981	0.025000	9
2	0.025000	0.025000	0.025000	0.025000	1.708888	2.391782	1.659831	3
3	0.025000	0.025000	1.960878	0.025000	3.849825	3.454260	13.902374	18
4	75.480954	84.434048	20.012255	43.281975	196.011303	167.818798	164.086678	197
5	13.626903	0.025000	74.970403	40.375234	7.369458	0.025000	18.693192	0
6	0.025000	0.025000	0.025000	8.468638	2.351202	6.540831	0.025000	0
7	16.102627	1.654308	0.025000	9.536278	80.832588	90.249374	35.460406	50
8	32.223802	0.025000	71.012268	69.439364	44.537920	26.612234	44.513576	7
9	9.679891	0.025000	0.025000	7.111885	12.287194	0.874992	0.025000	1
10	0.025000	1.151489	0.025000	0.025000	14.875877	7.522422	3.942505	0
11	8.343296	4.381569	0.025000	17.679644	6.297713	0.032241	0.025000	1
12	0.025000	0.025000	7.743518	17.285492	7.963162	0.025000	0.025000	0
13	21.376924	0.025000	49.878052	0.025000	0.025000	0.025000	3.160418	0
14	31.847859	0.025000	23.676464	61.351506	0.025000	15.459161	0.025000	1
15	11.187464	74.811755	0.025000	0.025000	0.025000	12.286402	66.907514	49

term_str	knowing	shaking	consider	twelve	closed	cry	shoulder	
topic_id								
16	19.493094	0.025000	2.276214	1.726161	1.954341	0.025000	0.025000	0
17	1.322578	2.565197	0.025000	4.578351	1.402625	0.025000	0.025000	1
18	21.133021	13.256124	8.168496	23.891582	0.025000	34.942834	28.583829	93
19	0.025000	42.808918	1.464008	0.025000	3.513799	0.025000	8.640975	0
20	10.064987	0.025000	0.025000	0.081036	90.918706	13.785262	9.915553	13
21	1.142914	0.025000	0.025000	0.025000	0.025000	3.725885	10.471178	2
22	72.279834	96.609799	114.633582	32.197280	83.724876	85.895105	39.578045	3
23	5.362592	6.310348	3.166873	28.307653	13.007660	4.058319	0.025000	7
24	38.194408	0.025000	54.097946	21.780341	16.705373	10.180615	3.768525	26
25	31.775613	11.047148	0.025000	12.421110	0.025000	0.346572	1.846999	0
26	0.025000	0.025000	9.625074	9.392131	0.025000	0.025000	0.025000	2
27	42.020479	38.414494	12.402814	1.494123	32.853313	93.107579	21.174153	35
28	0.025000	0.025000	0.025000	85.718655	24.547001	0.025000	20.341226	29
29	217.276236	254.303413	280.261185	14.941531	29.955200	62.694751	167.677160	77
30	0.025000	0.025000	0.025000	38.662668	0.025000	30.517891	9.030942	34
31	4.249229	20.218358	0.485445	39.265459	27.336795	29.613348	37.690124	2
32	27.560903	7.941185	6.711003	69.540270	15.311309	10.087222	0.025666	10
33	50.759404	169.260632	12.142615	0.025000	40.692670	0.025000	79.046099	0
34	7.418504	0.025000	0.025000	0.025000	0.025000	6.886808	0.025000	25
35	0.328553	0.025000	0.025000	54.190975	25.983776	9.712992	21.011513	0
36	43.619225	12.648792	29.729033	22.771667	52.873042	34.567687	25.117739	7
37	0.025000	0.025000	6.849315	0.025000	0.025000	2.679873	0.025000	0
38	0.025000	0.025000	19.363443	95.155163	0.025000	0.025000	0.025000	16
39	0.025000	0.025000	0.025000	3.559470	0.025000	0.025000	0.025000	0

40 rows × 2000 columns

In [20]:
```
tm.TOPIC.sort_values('theta_sum', ascending = False)
```

Out[20]:

	phi_sum	theta_sum	h	top_terms_rel	top_terms	label
topic_id						
29	138476.137464	220.112446	10.22	guardian cousin assure sister confidence pursu...	sister understand observed guardian confidence...	29: guardian cousin assure sister confidence p...

topic_id	phi_sum	theta_sum	h	top_terms_rel	top_terms	label
8	89131.260172	217.495613	9.98	detail color doesnt details recognized honor r...	presently toward ones isnt everybody war able ...	8: detail color doesnt details recognized hono...
5	56343.692496	150.168832	9.78	institution science political class national s...	society human law character knowledge class ch...	5: institution science political class nationa...
28	48624.881840	118.187148	9.79	lake mountain valley mountains ice rock miles ...	miles mountain land lake rock mountains distan...	28: lake mountain valley mountains ice rock mi...
4	62850.031662	113.004098	9.81	lock alarm lamp darkness dread horror muttered...	figure slowly wind answered sound lips breast ...	4: lock alarm lamp darkness dread horror mutte...
22	72011.366796	111.941402	10.07	rejoined inquired interposed hastily gentleman...	inquired rejoined exclaimed countenance servan...	22: rejoined inquired interposed hastily gentl...
24	50656.218241	105.019908	9.73	lecture 3 wrote literary 2 author letters mach...	wrote letters write written story paper writin...	24: lecture 3 wrote literary 2 author letters ...
18	34475.634392	101.499568	9.25	reckon warnt nigger bet maybe theyre anyway ju...	warnt reckon hes minute hadnt everybody big ma...	18: reckon warnt nigger bet maybe theyre anywa...
32	50898.403590	97.576213	9.85	theatre audience dancing ball stout applause c...	party everybody wine glass stage appearance bl...	32: theatre audience dancing ball stout applau...
35	41938.713871	85.799296	9.61	marble centuries pictures ancient picturesque ...	stone picture walls ancient sea houses streets...	35: marble centuries pictures ancient pictures...
36	45127.533127	82.742349	9.75	merry charity sisters sorrow mercy brothers ne...	spirit merry bear brothers tears earth thought...	36: merry charity sisters sorrow mercy brother...
33	41170.236899	69.274817	9.54	ha ant eh youre hes jolly em havent retorted	ha youre hes em youll whats office eh pleasant	33: ha ant eh youre hes jolly em havent retorted
7	29074.794954	64.261290	9.31	soldiers sword military rode armed crowd offic...	crowd soldiers horse military sword force guar...	7: soldiers sword military rode armed crowd of...

topic_id	phi_sum	theta_sum	h	top_terms_rel	top_terms	label
31	34759.348239	56.914861	9.45	travelling wheels horses landlord lamps roads ...	road horses horse carriage wind weather trees ...	31: travelling wheels horses landlord lamps ro...
14	28395.221946	56.375642	9.46	council lords army french castle fought killed...	french sent army died war castle killed afterw...	14: council lords army french castle fought ki...
0	20027.842032	55.893150	8.94	maid count lie forever voices grace hearts beg...	lie ah none toward tears truth noble saying vo...	0: maid count lie forever voices grace hearts ...
20	29544.283556	55.121146	9.67	dwarf grandfather beneath childs sleeping drea...	dwarf grandfather strange sleep silence led no...	20: dwarf grandfather beneath childs sleeping ...
15	32163.255500	54.755932	9.43	bottle parlour shop lad leg wooden instrument ...	shop bottle parlour tea leg em wooden sister coat	15: bottle parlour shop lad leg wooden instrum...
27	29961.847416	53.136875	9.49	aunt loved loving nurse mama kiss darling forg...	aunt loved tears sweet sitting quiet loving da...	27: aunt loved loving nurse mama kiss darling ...
25	26769.288410	48.638219	9.40	waiter shops idle police dirty market plate ho...	houses streets shop idle waiter iron windows w...	25: waiter shops idle police dirty market plat...
38	17024.101091	45.659177	8.48	dollars cent wages cents per sold sell buy coal	dollars gold pay worth per sold silver cent go...	38: dollars cent wages cents per sold sell buy...
13	18311.256224	36.886177	8.63	returns sits begins puts cries finds takes tel...	comes looks goes takes returns makes knows cri...	13: returns sits begins puts cries finds takes...
23	15332.440038	34.452292	8.52	jury prisoner murder trial evidence judge witn...	court judge prisoner murder jury trial evidenc...	23: jury prisoner murder trial evidence judge ...
9	13835.357462	22.391066	8.68	maam smart boots gloves aged stranger faced ba...	maam stranger boots em inquired pocket clerk o...	9: maam smart boots gloves aged stranger faced...
11	12258.544953	21.998367	8.14	om ym pilot deck ship voyage ships passengers ...	ship sea om ym deck board pilot passengers ships	11: om ym pilot deck ship voyage ships passeng...
30	10004.087665	21.173252	8.08	boat boats tide island lion shore ashore steam...	boat boats island shore bank tide lion stream ...	30: boat boats tide island lion shore ashore s...

topic_id	phi_sum	theta_sum	h	top_terms_rel	top_terms	label
1	8929.895643	20.770954	8.58	locksmith school baby dance girls dancing pare...	school locksmith baby pocket girls laugh daugh...	1: locksmith school baby dance girls dancing p...
26	8317.639096	19.643036	8.38	railway train hotel station platform fourth dr...	train hotel station railway box line road dog ...	26: railway train hotel station platform fourt...
37	12154.044627	17.790633	9.47	thy parents author thou thee madam graceful st...	thy thee thou soul ye noble youth parents bear	37: thy parents author thou thee madam gracefu...
3	6787.624918	17.416021	7.43	thee thou ye thy lad punch knights rags mad	ye thee thou thy lad none art mad ah	3: thee thou ye thy lad punch knights rags mad
34	6427.723600	16.568204	7.98	widow blind song baby cup husband farm sick dance	husband blind widow baby song eat sick couple ...	34: widow blind song baby cup husband farm sic...
19	9985.989702	15.111265	8.68	wery wot afore fur job em fat pipe theyre	wery em wot afore job inquired youre fur hes	19: wery wot afore fur job em fat pipe theyre
21	5331.354009	14.043632	6.49	en dat de nigger bet sand warnt sell camp	de en dat nigger em knows warnt sell sand	21: en dat de nigger bet sand warnt sell camp
12	4499.035576	13.524702	8.06	schoolmaster game punishment play fellows mora...	schoolmaster game play school bad society fell...	12: schoolmaster game punishment play fellows ...
16	5192.973099	12.313669	8.45	lawyer chambers legal waters bore rose opinion...	rose lawyer chambers bore honour waters legal ...	16: lawyer chambers legal waters bore rose opi...
6	3704.572278	11.467062	7.34	papa german mama police language sentence acco...	papa german mama language police french talkin...	6: papa german mama police language sentence a...
2	5962.053206	11.438465	7.55	uncle mail coach nephew driver passengers guar...	uncle coach mail driver nephew guard pause pas...	2: uncle mail coach nephew driver passengers g...
10	3468.842872	8.112984	6.46	bill brown bills measure single chief womans d...	bill brown single bills measure chief speech d...	10: bill brown bills measure single chief woma...
39	2456.571850	6.332991	7.50	committee honourable member noble british bill...	honourable committee noble member british bank...	39: committee honourable member noble british ...

topic_id	phi_sum	theta_sum	h	top_terms_rel	top_terms	label
17	1823.899489	4.987247	8.42	traveller pound meal eat landlord gate paid ru...	traveller pound eat gate paid meal landlord of...	17: traveller pound meal eat landlord gate pai...

Top 5 terms associated with the most frequent topic

In [21]:
```
top_topic = tm.TOPIC.theta_sum.idxmax()

top_topic
```

Out[21]:
```
29
```

In [22]:
```
tm.TOPIC.sort_values('theta_sum', ascending = False).loc[top_topic, 'top_terms_r
```

Out[22]:
```
'guardian cousin assure sister confidence pursued dearest madam agreeable'
```

In [23]:
```
# 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 [24]:
```
top_five_terms
```

Out[24]:
```
['guardian', 'cousin', 'assure', 'sister', 'confidence']
```

In [95]:
```
# 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[95]:

book_id	title	type	0	1	2	3	4	5
70	what is man	non-fiction	0.009390	0.008849	0.000235	0.001757	0.007265	0.145150
74	the adventures of tom sawyer	novel	0.049413	0.025284	0.000110	0.001742	0.079709	0.005079
76	the adventures of huckleberry finn	novel	0.003389	0.008033	0.002257	0.002569	0.000090	0.000090

			0	1	2	3	4	5	
book_id	title	type							
86	a connecticut yankee in king arthurs court	novel	0.070338	0.000363	0.000599	0.049563	0.035567	0.046094	0
91	tom sawyer abroad	novel	0.005992	0.000081	0.010122	0.000081	0.001582	0.008627	0
93	tom sawyer detective	novel	0.018958	0.000113	0.017373	0.000113	0.000113	0.000113	
98	a tale of two cities	novel	0.011382	0.000795	0.021161	0.003003	0.257681	0.020277	0
102	the tragedy of puddnhead wilson	novel	0.048508	0.008584	0.008988	0.000097	0.029773	0.030765	0
119	a tramp abroad	non-fiction	0.012653	0.003515	0.000066	0.004471	0.030154	0.029588	0
142	the 30000 bequest and other stories	stories	0.047183	0.000579	0.022021	0.001838	0.001521	0.087963	0
245	life on the mississippi	non-fiction	0.006648	0.003169	0.004721	0.000427	0.026438	0.074778	0
564	the mystery of edwin drood	novel	0.003491	0.010043	0.003267	0.006959	0.147611	0.031131	0
580	the pickwick papers	novel	0.003281	0.009057	0.019654	0.000318	0.020148	0.024977	0
588	master humphreys clock	stories	0.000044	0.000044	0.001081	0.004862	0.016351	0.099404	0
644	the haunted man and the ghosts bargain	stories	0.013539	0.008731	0.000018	0.000755	0.338508	0.036795	0
650	pictures from italy	non-fiction	0.000049	0.000049	0.000049	0.000049	0.025876	0.048508	0
653	the chimes	novel	0.010035	0.012778	0.004614	0.003634	0.261995	0.035002	0
675	american notes	non-fiction	0.000032	0.004439	0.007907	0.000644	0.009644	0.144144	0
676	the battle of life	novel	0.032070	0.048985	0.002432	0.001528	0.027884	0.009867	
699	a childs history of england	non-fiction	0.007595	0.003193	0.004027	0.003908	0.015744	0.010165	0
700	the old curiosity shop	novel	0.003832	0.008643	0.002243	0.004485	0.047036	0.011589	0
730	oliver twist	novel	0.004533	0.001820	0.000843	0.000063	0.110861	0.015866	0
766	david copperfield	novel	0.001296	0.016388	0.004283	0.000894	0.063482	0.026414	0

			0	1	2	3	4	5	
book_id	title	type							
786	hard times	novel	0.017459	0.000072	0.000950	0.018537	0.174327	0.074307	0
807	hunted down	stories	0.000160	0.000160	0.006459	0.005958	0.143385	0.155136	0
809	holiday romance	stories	0.000057	0.305651	0.001860	0.001941	0.000057	0.000057	0
810	george silvermans explanation	stories	0.009921	0.047064	0.001910	0.006749	0.071651	0.143381	0
821	dombey and sons	novel	0.006023	0.000209	0.012255	0.000376	0.100640	0.013093	
824	speeches of charles dickens	non-fiction	0.002244	0.007369	0.000677	0.000307	0.002318	0.553153	0
872	reprinted pieces	stories	0.002692	0.022891	0.000107	0.002791	0.021129	0.067194	0
882	sketches by boz	stories	0.000087	0.006586	0.013932	0.000180	0.016988	0.035101	0
883	our mutual friend	novel	0.004088	0.006129	0.000191	0.001634	0.138487	0.029852	0
888	the lazy tour of two idle apprentices	stories	0.000021	0.000021	0.000359	0.000021	0.137140	0.051788	
912	the mudfog and other sketches	stories	0.003319	0.000290	0.037002	0.000062	0.000062	0.184076	0
914	the uncommerical traveller	non-fiction	0.002452	0.004144	0.004081	0.000912	0.044953	0.119505	0
916	sketches of young couples	stories	0.006602	0.075274	0.000124	0.000124	0.000124	0.080526	0
917	barnaby rudge	stories	0.001484	0.039080	0.002127	0.001687	0.124637	0.023546	0
918	sketches of young gentlemen	stories	0.004089	0.005804	0.000138	0.000138	0.013997	0.112845	0
922	sunday under three heads	non-fiction	0.000046	0.000046	0.000046	0.000046	0.000046	0.272929	0
927	the lamplighter	stories	0.000026	0.000026	0.009641	0.009697	0.025532	0.000026	0
967	nicholas nickleby	novel	0.002125	0.011280	0.010865	0.006124	0.048602	0.023867	0
968	martin chuzzlewit	novel	0.003355	0.001296	0.003595	0.004087	0.045424	0.053055	0
1023	bleak house	novel	0.004758	0.005565	0.001183	0.001195	0.063027	0.034179	

			0	1	2	3	4	5	
book_id	title	type							
1044	extract from captain stormfields visit to Heaven	stories	0.000026	0.000026	0.000026	0.003027	0.000026	0.038712	0
1086	a horses tale	novel	0.059202	0.006429	0.003931	0.000380	0.000380	0.014288	0
1289	three ghost stories	stories	0.000028	0.011209	0.000028	0.000028	0.356886	0.064719	0
1394	the holly tree	stories	0.000070	0.008868	0.001150	0.001172	0.005605	0.007670	0
1400	great expectations	novel	0.008051	0.010427	0.009295	0.000273	0.147545	0.016101	
1406	the perils of certain english prisoners	stories	0.014194	0.000021	0.000021	0.000021	0.027778	0.000021	0
1407	a message from the sea	stories	0.000043	0.000043	0.000043	0.000043	0.050603	0.000043	0
1413	tom tiddlers ground	stories	0.035451	0.000092	0.000092	0.000092	0.000092	0.014516	0
1414	somebodys luggage	stories	0.000039	0.000039	0.000039	0.002410	0.047335	0.065210	0
1415	doctor marigold	stories	0.000243	0.000243	0.000243	0.000243	0.090357	0.000243	0
1416	mrs lirripers lodgings	stories	0.010001	0.125790	0.000063	0.000063	0.014245	0.001928	0
1421	mrs lirripers legacy	stories	0.033660	0.000070	0.000070	0.002316	0.015992	0.000070	0
1435	miscellaneous papers	non-fiction	0.011929	0.005109	0.000878	0.005464	0.018538	0.437036	0
1467	some christmas stories	stories	0.014210	0.039305	0.009365	0.000068	0.037513	0.032199	0
1837	the prince and the pauper	novel	0.125743	0.014046	0.002184	0.225778	0.082276	0.012667	0
2324	a house to let	stories	0.014811	0.000038	0.002889	0.001239	0.039083	0.025552	0
2874	personal recollections of joan of arc vol 1	non-fiction	0.254703	0.003826	0.002709	0.008486	0.019684	0.004559	0
2875	personal recollections of joan of arc vol 2	non-fiction	0.297637	0.001720	0.001515	0.004186	0.031298	0.032773	0
2895	following the equator	non-fiction	0.011044	0.001414	0.000457	0.001887	0.006131	0.089574	

			0	1	2	3	4	5	
book_id	title	type							
3171	in defense of harriet shelley	non-fiction	0.000034	0.012185	0.000034	0.012122	0.000034	0.085039	0
3172	fenimore coopers literary offences	non-fiction	0.000033	0.000033	0.000033	0.000033	0.000033	0.124599	0
3173	essays on paul bourget	non-fiction	0.000034	0.000034	0.000034	0.000034	0.000034	0.333928	0
3176	the innocents abroad	non-fiction	0.029505	0.002304	0.000949	0.006506	0.008949	0.026396	
3177	roughing it	novel	0.012992	0.002765	0.014333	0.000394	0.019562	0.043561	0
3178	the gilded age	novel	0.025880	0.006614	0.005308	0.010801	0.022757	0.113402	0
3179	the american claimant	novel	0.026804	0.001621	0.000900	0.000354	0.025078	0.051622	0
3180	a double barrelled detective story	stories	0.060737	0.000187	0.011478	0.003985	0.087922	0.000187	0
3181	the stolen white elephant	stories	0.011084	0.000069	0.000069	0.000069	0.000069	0.095697	0
3182	some rambling notes of an idle excursion	non-fiction	0.010924	0.000043	0.000043	0.000595	0.029328	0.016605	0
3183	the facts concerning the recent carnival of crime in connecticut	stories	0.000027	0.000027	0.000027	0.000027	0.000027	0.000027	0
3184	alonzo fitz and other stories	stories	0.070844	0.035533	0.006579	0.076966	0.000100	0.055767	0
3185	those extraordinary twins	stories	0.051618	0.000121	0.000121	0.000121	0.006009	0.040372	
3186	the mysterious stranger and other stories	stories	0.088795	0.002064	0.007931	0.000100	0.046225	0.009184	0
3188	mark twain speeches	non-fiction	0.017354	0.015212	0.001637	0.000908	0.001994	0.176339	0
3189	sketches new and old	stories	0.030225	0.014409	0.002662	0.003720	0.024096	0.081282	0

			0	1	2	3	4	5	
book_id	title	type							
3190	1601 conversation as it was by the social fireside in the time of the tudors	stories	0.020241	0.000091	0.000091	0.168709	0.000091	0.182268	0
3191	goldsmiths friend abroad again	stories	0.034448	0.000290	0.000290	0.020891	0.000290	0.016789	0
3192	the curious republic of gondour and other whimsical sketches	stories	0.002787	0.000366	0.000366	0.000366	0.000366	0.207245	0
3199	the letters of mark twain	non-fiction	0.010030	0.000719	0.000595	0.000965	0.004062	0.039612	0
3250	how to tell a story and other essays	non-fiction	0.000154	0.002918	0.023603	0.000154	0.046718	0.000154	0
3251	the man that corrupted hadleyburg and other stories	stories	0.106797	0.003092	0.001099	0.000571	0.024848	0.105268	
19337	a christmas carol	novel	0.000037	0.025192	0.018158	0.004069	0.085491	0.000037	0
19484	editorial wild oats	stories	0.016739	0.021627	0.010532	0.000121	0.021929	0.033859	
19987	chapters from my autobiography	non-fiction	0.015152	0.005424	0.004267	0.001678	0.014649	0.047126	0
20795	the cricket on the hearth	novel	0.003587	0.098341	0.000017	0.001369	0.051946	0.004145	
27924	mugby junction	stories	0.007883	0.005528	0.015614	0.000027	0.186989	0.020384	0
33077	the treaty with china its provisions explained	non-fiction	0.000025	0.000025	0.000025	0.000025	0.000025	0.565564	0
35536	the poems and verses of charles dickens	stories	0.028606	0.016627	0.000251	0.033804	0.022100	0.078869	0
60900	merry tales	stories	0.041890	0.000041	0.001007	0.000921	0.060213	0.010121	0
61522	the 1000000 bank note	stories	0.000019	0.000019	0.000019	0.000344	0.006644	0.027626	

book_id	title	type	0	1	2	3	4	5	
62636	**to the person sitting in darkness**	**non-fiction**	0.059227	0.000036	0.000036	0.000036	0.000036	0.114464	(
62739	**king leopolds soliloquy**	**stories**	0.040921	0.000234	0.000234	0.014970	0.000234	0.277346	(

In [56]:
```
# most common topics by work type
book_mean_theta.groupby('type').mean().idxmax(axis = 1)
```

Out[56]:
```
type
non-fiction     8
novel          29
stories         8
dtype: int64
```

In [81]:
```
# 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[81]:

book_id	topic_id	title	type	phi_sum	theta_sum	h	top_terms_rel	
1415	38	doctor marigold	stories	17024.101091	45.659177	8.480000	dollars cent wages cents per sold sell buy coal	(g
61522	37	the 1000000 bank note	stories	12154.044627	17.790633	9.470000	thy parents author thou thee madam graceful stars bosom	t r p
19337	36	a christmas carol	novel	45127.533127	82.742349	9.750000	merry charity sisters sorrow mercy brothers nephew ghost younger	s t
676	36	the battle of life	novel	45127.533127	82.742349	9.750000	merry charity sisters sorrow mercy brothers nephew ghost younger	s t

book_id	topic_id	title	type	phi_sum	theta_sum	h	top_terms_rel	
3176	35	the innocents abroad	non-fiction	41938.713871	85.799296	9.610000	marble centuries pictures ancient picturesque palace painted walls stone	p
650	35	pictures from italy	non-fiction	41938.713871	85.799296	9.610000	marble centuries pictures ancient picturesque palace painted walls stone	p
20795	33	the cricket on the hearth	novel	41170.236899	69.274817	9.540000	ha ant eh youre hes jolly em havent retorted	ha w
927	33	the lamplighter	stories	41170.236899	69.274817	9.540000	ha ant eh youre hes jolly em havent retorted	ha w
917	33	barnaby rudge	stories	41170.236899	69.274817	9.540000	ha ant eh youre hes jolly em havent retorted	ha w
912	32	the mudfog and other sketches	stories	50898.403590	97.576213	9.850000	theatre audience dancing ball stout applause circle punch gallery	a b
882	32	sketches by boz	stories	50898.403590	97.576213	9.850000	theatre audience dancing ball stout applause circle punch gallery	a b
916	32	sketches of young couples	stories	50898.403590	97.576213	9.850000	theatre audience dancing ball stout applause circle punch gallery	a b
918	32	sketches of young gentlemen	stories	50898.403590	97.576213	9.850000	theatre audience dancing ball stout applause circle punch gallery	a b

book_id	topic_id	title	type	phi_sum	theta_sum	h	top_terms_rel	r
1394	31	the holly tree	stories	34759.348239	56.914861	9.450000	travelling wheels horses landlord lamps roads road carriage cart	
1406	30	the perils of certain english prisoners	stories	10004.087665	21.173252	8.080000	boat boats tide island lion shore ashore steam stream	is
883	29	our mutual friend	novel	138476.137464	220.112446	10.220000	guardian cousin assure sister confidence pursued dearest madam agreeable	th
564	29	the mystery of edwin drood	novel	138476.137464	220.112446	10.220000	guardian cousin assure sister confidence pursued dearest madam agreeable	th
766	29	david copperfield	novel	138476.137464	220.112446	10.220000	guardian cousin assure sister confidence pursued dearest madam agreeable	th
786	29	hard times	novel	138476.137464	220.112446	10.220000	guardian cousin assure sister confidence pursued dearest madam agreeable	th
807	29	hunted down	stories	138476.137464	220.112446	10.220000	guardian cousin assure sister confidence pursued dearest madam agreeable	th

book_id	topic_id	title	type	phi_sum	theta_sum	h	top_terms_rel	
810	29	george silvermans explanation	stories	138476.137464	220.112446	10.220000	guardian cousin assure sister confidence pursued dearest madam agreeable	th
821	29	dombey and sons	novel	138476.137464	220.112446	10.220000	guardian cousin assure sister confidence pursued dearest madam agreeable	th
968	29	martin chuzzlewit	novel	138476.137464	220.112446	10.220000	guardian cousin assure sister confidence pursued dearest madam agreeable	th
1400	29	great expectations	novel	138476.137464	220.112446	10.220000	guardian cousin assure sister confidence pursued dearest madam agreeable	th
1407	29	a message from the sea	stories	138476.137464	220.112446	10.220000	guardian cousin assure sister confidence pursued dearest madam agreeable	th
1416	29	mrs lirripers lodgings	stories	138476.137464	220.112446	10.220000	guardian cousin assure sister confidence pursued dearest madam agreeable	th
1421	29	mrs lirripers legacy	stories	138476.137464	220.112446	10.220000	guardian cousin assure sister confidence pursued dearest madam agreeable	th

book_id	topic_id	title	type	phi_sum	theta_sum	h	top_terms_rel	
1467	29	some christmas stories	stories	138476.137464	220.112446	10.220000	guardian cousin assure sister confidence pursued dearest madam agreeable	th
2324	29	a house to let	stories	138476.137464	220.112446	10.220000	guardian cousin assure sister confidence pursued dearest madam agreeable	th
1023	29	bleak house	novel	138476.137464	220.112446	10.220000	guardian cousin assure sister confidence pursued dearest madam agreeable	th
119	28	a tramp abroad	non-fiction	48624.881840	118.187148	9.790000	lake mountain valley mountains ice rock miles forest snow	
3182	28	some rambling notes of an idle excursion	non-fiction	48624.881840	118.187148	9.790000	lake mountain valley mountains ice rock miles forest snow	
3177	28	roughing it	novel	48624.881840	118.187148	9.790000	lake mountain valley mountains ice rock miles forest snow	
245	28	life on the mississippi	non-fiction	48624.881840	118.187148	9.790000	lake mountain valley mountains ice rock miles forest snow	

book_id	topic_id	title	type	phi_sum	theta_sum	h	top_terms_rel	
888	25	the lazy tour of two idle apprentices	stories	26769.288410	48.638219	9.400000	waiter shops idle police dirty market plate houses shillings	s irc w
872	25	reprinted pieces	stories	26769.288410	48.638219	9.400000	waiter shops idle police dirty market plate houses shillings	s irc w
914	25	the uncommerical traveller	non-fiction	26769.288410	48.638219	9.400000	waiter shops idle police dirty market plate houses shillings	s irc w
35536	24	the poems and verses of charles dickens	stories	50656.218241	105.019908	9.730000	lecture 3 wrote literary 2 author letters machine print	w w s w
19484	24	editorial wild oats	stories	50656.218241	105.019908	9.730000	lecture 3 wrote literary 2 author letters machine print	w w s w
3199	24	the letters of mark twain	non-fiction	50656.218241	105.019908	9.730000	lecture 3 wrote literary 2 author letters machine print	w w s w
3190	24	1601 conversation as it was by the social fireside in the time of the tudors	stories	50656.218241	105.019908	9.730000	lecture 3 wrote literary 2 author letters machine print	w w s w
3188	24	mark twain speeches	non-fiction	50656.218241	105.019908	9.730000	lecture 3 wrote literary 2 author letters machine print	w w s w
3171	24	in defense of harriet shelley	non-fiction	50656.218241	105.019908	9.730000	lecture 3 wrote literary 2 author letters machine print	w w s w
1414	24	somebodys luggage	stories	50656.218241	105.019908	9.730000	lecture 3 wrote literary 2 author letters machine print	w w s w

book_id	topic_id	title	type	phi_sum	theta_sum	h	top_terms_rel	
967	22	nicholas nickleby	novel	72011.366796	111.941402	10.070000	rejoined inquired interposed hastily gentlemans exclaimed indignation countenance thrust	co tu
580	22	the pickwick papers	novel	72011.366796	111.941402	10.070000	rejoined inquired interposed hastily gentlemans exclaimed indignation countenance thrust	co tu
730	22	oliver twist	novel	72011.366796	111.941402	10.070000	rejoined inquired interposed hastily gentlemans exclaimed indignation countenance thrust	co tu
700	20	the old curiosity shop	novel	29544.283556	55.121146	9.670000	dwarf grandfather beneath childs sleeping dreary anxiety roused poverty	g sl
588	20	master humphreys clock	stories	29544.283556	55.121146	9.670000	dwarf grandfather beneath childs sleeping dreary anxiety roused poverty	g sl
1044	18	extract from captain stormfields visit to Heaven	stories	34475.634392	101.499568	9.250000	reckon warnt nigger bet maybe theyre anyway judged hed	wa
93	18	tom sawyer detective	novel	34475.634392	101.499568	9.250000	reckon warnt nigger bet maybe theyre anyway judged hed	wa

book_id	topic_id	title	type	phi_sum	theta_sum	h	top_terms_rel	
91	18	tom sawyer abroad	novel	34475.634392	101.499568	9.250000	reckon warnt nigger bet maybe theyre anyway judged hed	w
76	18	the adventures of huckleberry finn	novel	34475.634392	101.499568	9.250000	reckon warnt nigger bet maybe theyre anyway judged hed	w
74	18	the adventures of tom sawyer	novel	34475.634392	101.499568	9.250000	reckon warnt nigger bet maybe theyre anyway judged hed	w
3181	17	the stolen white elephant	stories	1823.899489	4.987247	8.420000	traveller pound meal eat landlord gate paid ruined shillings	of
1413	17	tom tiddlers ground	stories	1823.899489	4.987247	8.420000	traveller pound meal eat landlord gate paid ruined shillings	of
699	14	a childs history of england	non-fiction	28395.221946	56.375642	9.460000	council lords army french castle fought killed religion battle	.
1086	8	a horses tale	novel	89131.260172	217.495613	9.980000	detail color doesnt details recognized honor rule nation afterward	t
3173	8	essays on paul bourget	non-fiction	89131.260172	217.495613	9.980000	detail color doesnt details recognized honor rule nation afterward	t

book_id	topic_id	title	type	phi_sum	theta_sum	h	top_terms_rel	t
3172	8	fenimore coopers literary offences	non-fiction	89131.260172	217.495613	9.980000	detail color doesnt details recognized honor rule nation afterward	
2895	8	following the equator	non-fiction	89131.260172	217.495613	9.980000	detail color doesnt details recognized honor rule nation afterward	
2874	8	personal recollections of joan of arc vol 1	non-fiction	89131.260172	217.495613	9.980000	detail color doesnt details recognized honor rule nation afterward	
3178	8	the gilded age	novel	89131.260172	217.495613	9.980000	detail color doesnt details recognized honor rule nation afterward	
142	8	the 30000 bequest and other stories	stories	89131.260172	217.495613	9.980000	detail color doesnt details recognized honor rule nation afterward	
102	8	the tragedy of puddnhead wilson	novel	89131.260172	217.495613	9.980000	detail color doesnt details recognized honor rule nation afterward	
86	8	a connecticut yankee in king arthurs court	novel	89131.260172	217.495613	9.980000	detail color doesnt details recognized honor rule nation afterward	

book_id	topic_id	title	type	phi_sum	theta_sum	h	top_terms_rel	t
3180	8	a double barrelled detective story	stories	89131.260172	217.495613	9.980000	detail color doesnt details recognized honor rule nation afterward	
70	8	what is man	non-fiction	89131.260172	217.495613	9.980000	detail color doesnt details recognized honor rule nation afterward	
3179	8	the american claimant	novel	89131.260172	217.495613	9.980000	detail color doesnt details recognized honor rule nation afterward	
3183	8	the facts concerning the recent carnival of crime in connecticut	stories	89131.260172	217.495613	9.980000	detail color doesnt details recognized honor rule nation afterward	
3184	8	alonzo fitz and other stories	stories	89131.260172	217.495613	9.980000	detail color doesnt details recognized honor rule nation afterward	
3185	8	those extraordinary twins	stories	89131.260172	217.495613	9.980000	detail color doesnt details recognized honor rule nation afterward	
3186	8	the mysterious stranger and other stories	stories	89131.260172	217.495613	9.980000	detail color doesnt details recognized honor rule nation afterward	

book_id	topic_id	title	type	phi_sum	theta_sum	h	top_terms_rel	
3189	8	sketches new and old	stories	89131.260172	217.495613	9.980000	detail color doesnt details recognized honor rule nation afterward	t
3191	8	goldsmiths friend abroad again	stories	89131.260172	217.495613	9.980000	detail color doesnt details recognized honor rule nation afterward	t
3250	8	how to tell a story and other essays	non-fiction	89131.260172	217.495613	9.980000	detail color doesnt details recognized honor rule nation afterward	t
3251	8	the man that corrupted hadleyburg and other stories	stories	89131.260172	217.495613	9.980000	detail color doesnt details recognized honor rule nation afterward	t
19987	8	chapters from my autobiography	non-fiction	89131.260172	217.495613	9.980000	detail color doesnt details recognized honor rule nation afterward	t
60900	8	merry tales	stories	89131.260172	217.495613	9.980000	detail color doesnt details recognized honor rule nation afterward	t
62636	8	to the person sitting in darkness	non-fiction	89131.260172	217.495613	9.980000	detail color doesnt details recognized honor rule nation afterward	t

book_id	topic_id	title	type	phi_sum	theta_sum	h	top_terms_rel
1435	5	miscellaneous papers	non-fiction	56343.692496	150.168832	9.780000	institution science political class national social association education legal
675	5	american notes	non-fiction	56343.692496	150.168832	9.780000	institution science political class national social association education legal
824	5	speeches of charles dickens	non-fiction	56343.692496	150.168832	9.780000	institution science political class national social association education legal
922	5	sunday under three heads	non-fiction	56343.692496	150.168832	9.780000	institution science political class national social association education legal
62739	5	king leopolds soliloquy	stories	56343.692496	150.168832	9.780000	institution science political class national social association education legal
3192	5	the curious republic of gondour and other whimsical sketches	stories	56343.692496	150.168832	9.780000	institution science political class national social association education legal
33077	5	the treaty with china its provisions explained	non-fiction	56343.692496	150.168832	9.780000	institution science political class national social association education legal

book_id	topic_id	title	type	phi_sum	theta_sum	h	top_terms_rel	
27924	4	mugby junction	stories	62850.031662	113.004098	9.810000	lock alarm lamp darkness dread horror muttered lighted nearer	fi br
1289	4	three ghost stories	stories	62850.031662	113.004098	9.810000	lock alarm lamp darkness dread horror muttered lighted nearer	fi br
653	4	the chimes	novel	62850.031662	113.004098	9.810000	lock alarm lamp darkness dread horror muttered lighted nearer	fi br
644	4	the haunted man and the ghosts bargain	stories	62850.031662	113.004098	9.810000	lock alarm lamp darkness dread horror muttered lighted nearer	fi br
98	4	a tale of two cities	novel	62850.031662	113.004098	9.810000	lock alarm lamp darkness dread horror muttered lighted nearer	fi br
1837	3	the prince and the pauper	novel	6787.624918	17.416021	7.430000	thee thou ye thy lad punch knights rags mad	y tl
809	1	holiday romance	stories	8929.895643	20.770954	8.580000	locksmith school baby dance girls dancing parents hearty fish	b
2875	0	personal recollections of joan of arc vol 2	non-fiction	20027.842032	55.893150	8.940000	maid count lie forever voices grace hearts begged message	tc

Works and Top Terms Associated with Each Topic

In [89]:

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

In [92]:
```
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.reset_index().style.background_gradient(cmap='YlGnBu', subset = ['topic
```

Out[92]:

	topic_id	count	title	top_terms_rel
0	8	23	what is man, a connecticut yankee in king arthurs court, the tragedy of puddnhead wilson, the 30000 bequest and other stories, a horses tale, personal recollections of joan of arc vol 1, following the equator, fenimore coopers literary offences, essays on paul bourget, the gilded age, the american claimant, a double barrelled detective story, the facts concerning the recent carnival of crime in connecticut, alonzo fitz and other stories, those extraordinary twins, the mysterious stranger and other stories, sketches new and old, goldsmiths friend abroad again, how to tell a story and other essays, the man that corrupted hadleyburg and other stories, chapters from my autobiography, merry tales, to the person sitting in darkness	detail color doesnt details recognized honor rule nation afterward
1	29	15	the mystery of edwin drood, david copperfield, hard times, hunted down, george silvermans explanation, dombey and sons, our mutual friend, martin chuzzlewit, bleak house, great expectations, a message from the sea, mrs lirripers lodgings, mrs lirripers legacy, some christmas stories, a house to let	guardian cousin assure sister confidence pursued dearest madam agreeable
2	24	7	somebodys luggage, in defense of harriet shelley, mark twain speeches, 1601 conversation as it was by the social fireside in the time of the tudors, the letters of mark twain, editorial wild oats, the poems and verses of charles dickens	lecture 3 wrote literary 2 author letters machine print
3	5	7	american notes, speeches of charles dickens, sunday under three heads, miscellaneous papers, the curious republic of gondour and other whimsical sketches, the treaty with china its provisions explained, king leopolds soliloquy	institution science political class national social association education legal
4	4	5	a tale of two cities, the haunted man and the ghosts bargain, the chimes, three ghost stories, mugby junction	lock alarm lamp darkness dread horror muttered lighted nearer
5	18	5	the adventures of tom sawyer, the adventures of huckleberry finn, tom sawyer abroad, tom sawyer detective, extract from captain stormfields visit to Heaven	reckon warnt nigger bet maybe theyre anyway judged hed
6	32	4	sketches by boz, the mudfog and other sketches, sketches of young couples, sketches of young gentlemen	theatre audience dancing ball stout applause circle punch gallery

	topic_id	count	title	top_terms_rel
7	28	4	a tramp abroad, life on the mississippi, roughing it, some rambling notes of an idle excursion	lake mountain valley mountains ice rock miles forest snow
8	22	3	the pickwick papers, oliver twist, nicholas nickleby	rejoined inquired interposed hastily gentlemans exclaimed indignation countenance thrust
9	25	3	reprinted pieces, the lazy tour of two idle apprentices, the uncommerical traveller	waiter shops idle police dirty market plate houses shillings
10	33	3	barnaby rudge, the lamplighter, the cricket on the hearth	ha ant eh youre hes jolly em havent retorted
11	36	2	the battle of life, a christmas carol	merry charity sisters sorrow mercy brothers nephew ghost younger
12	17	2	tom tiddlers ground, the stolen white elephant	traveller pound meal eat landlord gate paid ruined shillings
13	20	2	master humphreys clock, the old curiosity shop	dwarf grandfather beneath childs sleeping dreary anxiety roused poverty
14	35	2	pictures from italy, the innocents abroad	marble centuries pictures ancient picturesque palace painted walls stone
15	37	1	the 1000000 bank note	thy parents author thou thee madam graceful stars bosom

	topic_id	count	title	top_terms_rel
16	0	1	personal recollections of joan of arc vol 2	maid count lie forever voices grace hearts begged message
17	31	1	the holly tree	travelling wheels horses landlord lamps roads road carriage cart
18	30	1	the perils of certain english prisoners	boat boats tide island lion shore ashore steam stream
19	1	1	holiday romance	locksmith school baby dance girls dancing parents hearty fish
20	14	1	a childs history of england	council lords army french castle fought killed religion battle
21	3	1	the prince and the pauper	thee thou ye thy lad punch knights rags mad
22	38	1	doctor marigold	dollars cent wages cents per sold sell buy coal

In [91]:
```
# reset width to default: https://pandas.pydata.org/docs/user_guide/options.html
pd.set_option('display.max_colwidth', 50)
```

M09: Word Embeddings

In [27]:
```
w2v_params = dict(
    min_count = 10,
    workers = 1,
    # vector_size = 246,
    vector_size = 100,
    window = 2
)
```

In [28]:
```
SENTS = CORPUS.groupby(OHCO[:-1]).term_str.apply(lambda  x:  x.tolist())
```

In [29]:
```
model = word2vec.Word2Vec(SENTS.values, **w2v_params)
```

In [30]:
```
W2V = pd.DataFrame(model.wv.get_normed_vectors(), index=model.wv.index_to_key)
W2V.index.name = 'term_str'
W2V = W2V.sort_index()
```

In [31]:
```
W2V.head()
```

Out[31]:

term_str	0	1	2	3	4	5	6	7
0	-0.168197	0.018952	0.025413	0.005438	0.097372	-0.171935	0.085254	0.057859
04	-0.031947	0.000754	-0.015000	-0.020203	0.085838	-0.088574	0.087527	0.199322
08	-0.052396	0.006009	0.034031	0.008314	0.056115	-0.126131	0.092966	0.208461
1	-0.120617	-0.009931	-0.044832	0.057515	0.133221	-0.062641	0.090187	-0.013538
10	-0.064771	0.053249	-0.015686	0.066639	0.094703	-0.102791	0.043700	0.041097

5 rows × 100 columns

In [32]:
```
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 [33]:
```
tsne_engine = TSNE(**tsne_params)
tsne_model = tsne_engine.fit_transform(W2V)
```

In [34]:
```
COORDS = pd.DataFrame(tsne_model, columns=['x','y'], index=W2V.index).join(VOCAB
```

In [35]:
```
COORDS['log_n'] = np.log(COORDS['n'])
```

In [36]:
```
COORDS
```

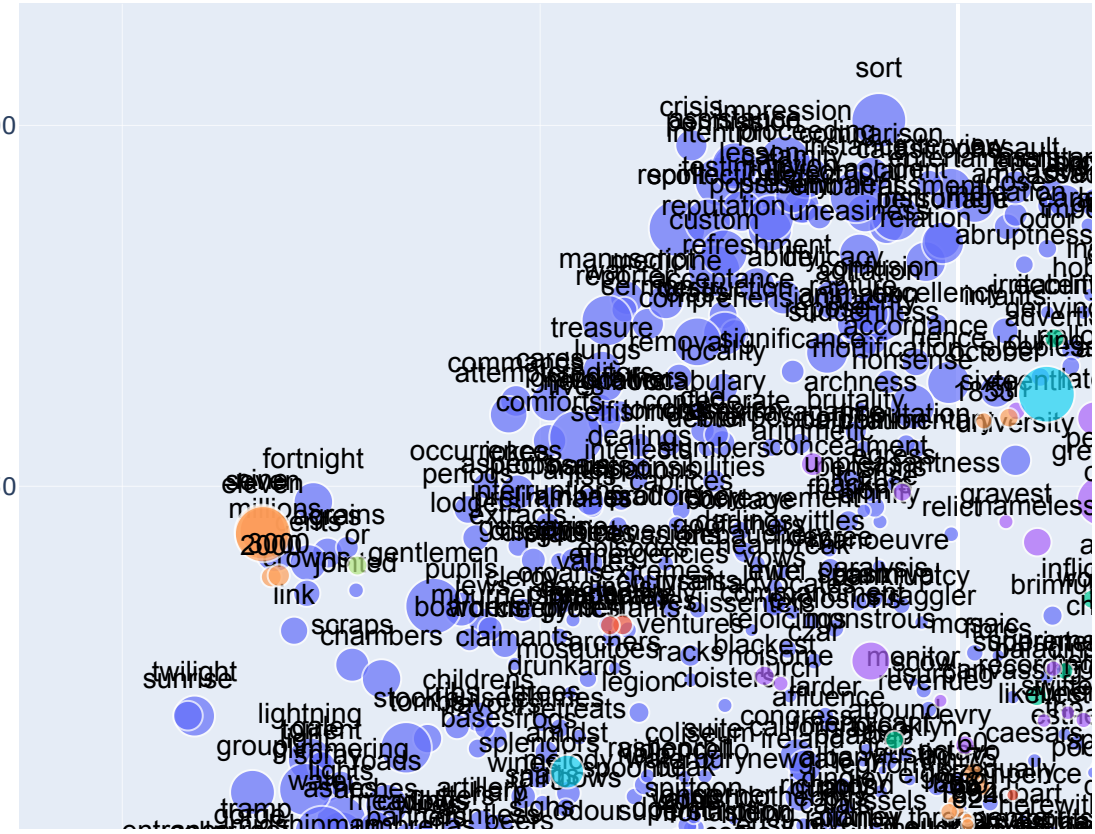
Out[36]:

term_str	x	y	n	dfidf	pos_group	log_n
0	-0.171599	61.323936	65	65.289055	CD	4.174387
04	0.542611	5.074944	10	20.322264	NN	2.302585
08	0.813468	5.348079	10	11.161132	NN	2.302585
1	0.145392	62.795574	369	640.100772	CD	5.910797

term_str	x	y	n	dfidf	pos_group	log_n
10	0.960314	64.455231	143	390.523832	CD	4.962845
...	...	...	...	...	...	...
zoological	-21.948164	17.195534	18	120.252374	JJ	2.890372
zu	60.205723	-7.450152	22	28.728508	NN	3.091042
zulu	-13.550667	38.664394	12	58.476439	NN	2.484907
à	74.929085	4.147956	94	125.841724	NN	4.543295
était	61.743019	-4.612803	13	11.161132	NN	2.564949

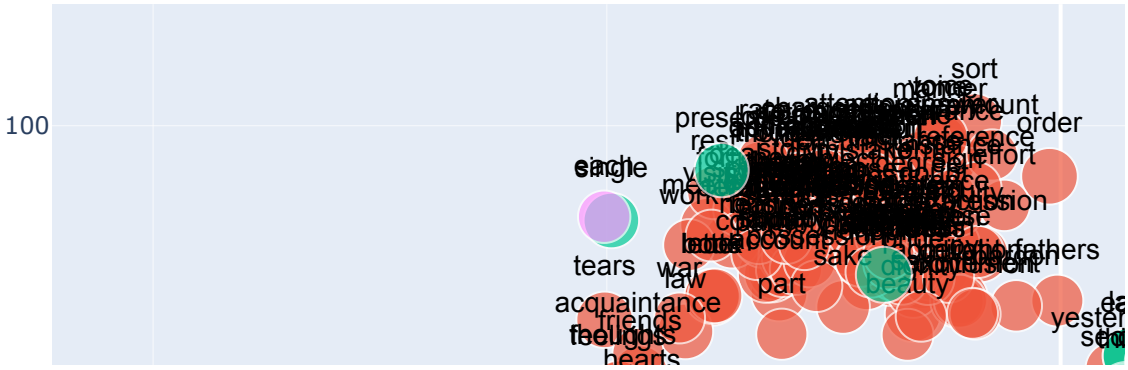
22373 rows × 6 columns

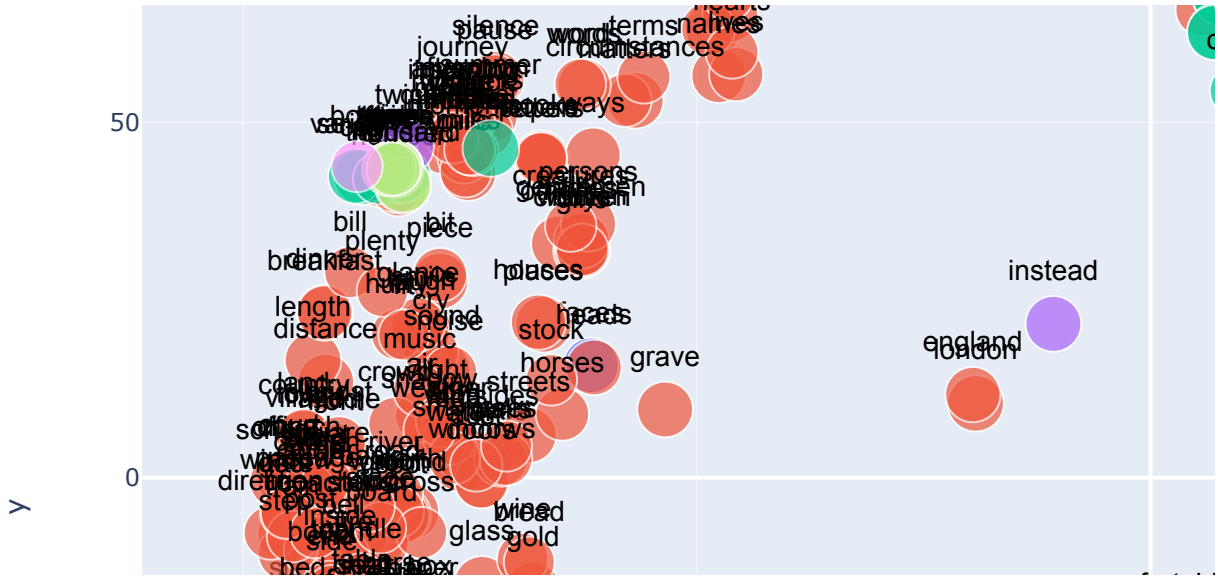
In [37]:
```
px.scatter(COORDS.reset_index().sample(1000),
           'x', 'y',
           text='term_str',
           color='pos_group',
           hover_name='term_str',
           size='dfidf',
           height=1000).update_traces(
                mode='markers+text',
                textfont=dict(color='black', size=14, family='Arial'),
                textposition='top center')
```



In [38]:

```
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')
```





With Nouns Only (not proper ones)

In [53]:
```
noun_COORDS = COORDS.loc[COORDS.pos_group == 'NN']

noun_COORDS
```

Out[53]:

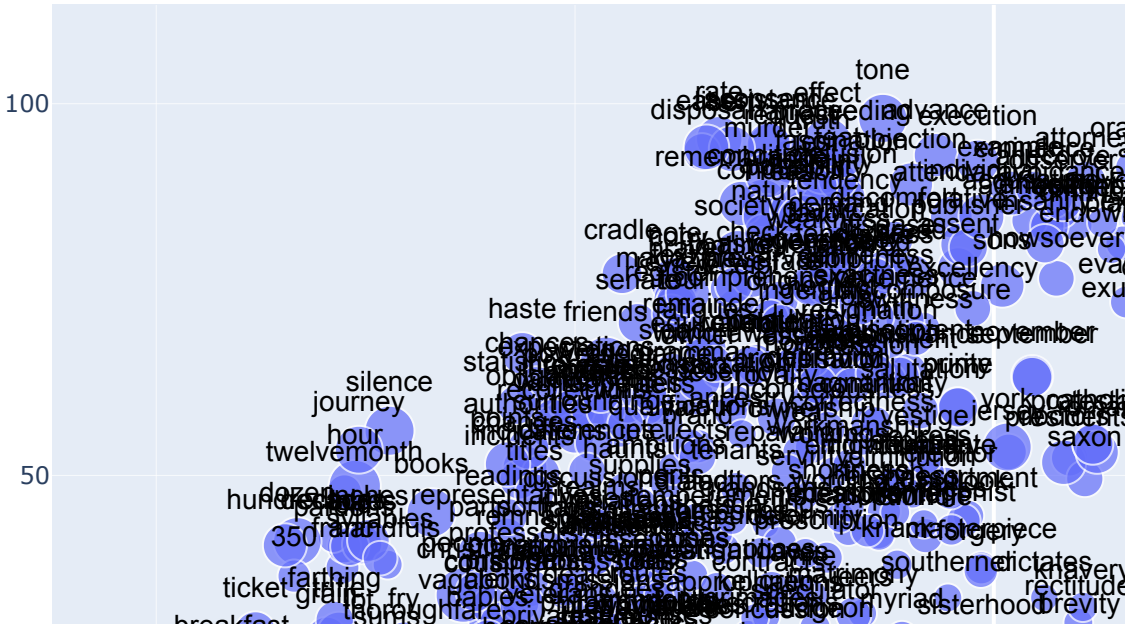
term_str	x	y	n	dfidf	pos_group	log_n
04	0.542611	5.074944	10	20.322264	NN	2.302585

term_str	x	y	n	dfidf	pos_group	log_n
08	0.813468	5.348079	10	11.161132	NN	2.302585
350	-83.536514	36.456417	24	78.392038	NN	3.178054
87	2.034327	5.669988	14	51.457016	NN	2.639057
89	1.161920	5.738420	15	44.196019	NN	2.708050
...	...	...	...	...	...	...
zone	-32.028049	2.642661	12	78.392038	NN	2.484907
zu	60.205723	-7.450152	22	28.728508	NN	3.091042
zulu	-13.550667	38.664394	12	58.476439	NN	2.484907
à	74.929085	4.147956	94	125.841724	NN	4.543295
était	61.743019	-4.612803	13	11.161132	NN	2.564949

13143 rows × 6 columns

Noun tSNE plot

In [54]:
```
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

- luggage, country, town, borough, inn, tent → travel
- nostrils, knuckles, wrist, sleeve, armor, grip → battle, struggle??
- scavengers, hostlers, philanthropists, presbyterians, architects, wayfarers → occupation, action, charity??
- dissensions, swindlers, aims, rivals, foes → enmity, opposition
- woods, ice, gravel, dust, deserts, valleys, plants → nature, outdoors

Analogies and Similarities (vector algebra)

In [39]:

```
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 [40]:
```
complete_analogy('man', 'boy', 'woman', 3)
```

Out[40]:

	term	sim
0	girl	0.851387
1	baby	0.768748
2	child	0.764450

In [41]:
```
complete_analogy('girl', 'daughter', 'boy', 3)
```

Out[41]:

	term	sim
0	son	0.806970
1	niece	0.796292
2	nephew	0.792255

In [42]:
```
complete_analogy('girl', 'sister', 'boy', 3)
```

Out[42]:

	term	sim
0	niece	0.803975
1	nephew	0.780578
2	brother	0.761206

In [43]:
```
complete_analogy('man', 'gentleman', 'woman', 5)
```

Out[43]:

	term	sim
0	lady	0.837602
1	girl	0.756685
2	housekeeper	0.730277
3	widow	0.726901
4	matron	0.669566

In [44]:
```
complete_analogy('woman', 'lady', 'man', 5)
```

Out[44]:

	term	sim

	term	sim
0	gentleman	0.824193
1	person	0.687966
2	student	0.618382
3	clergyman	0.597353
4	lawyer	0.588978

In [45]:
```
complete_analogy('day', 'sun', 'night', 5)
```

Out[45]:

	term	sim
0	moon	0.758980
1	rain	0.719536
2	sky	0.714280
3	sunlight	0.712312
4	clouds	0.711592

In [57]:
```
complete_analogy('king', 'money', 'servant', 5)
```

Out[57]:

	term	sim
0	purse	0.618348
1	lodgings	0.538427
2	meals	0.536840
3	medicine	0.535820
4	property	0.535807

In [58]:
```
complete_analogy('king', 'royal', 'servant', 5)
```

Out[58]:

	term	sim
0	keepers	0.603850
1	cabinet	0.594091
2	boarding	0.564780
3	private	0.564711
4	ladyships	0.561722

In [67]:
```
complete_analogy('king', 'rich', 'servant', 5)
```

Out[67]:

	term	sim

	term	sim
0	nice	0.608984
1	shabby	0.604531
2	handsome	0.594732
3	clever	0.577550
4	sturdy	0.547286

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

Out[68]:

	term	sim
0	shabby	0.655889
1	lazy	0.580029
2	tall	0.578269
3	clad	0.566695
4	sailor	0.561728

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

Out[71]:

	term	sim
0	voyage	0.678766
1	trip	0.639491
2	pilgrimage	0.582225
3	visit	0.535242
4	marriage	0.534133

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

Out[72]:

	term	sim
0	commission	0.621347
1	trial	0.615661
2	introduction	0.601732
3	petition	0.596050
4	request	0.594801

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

Out[73]:

	term	sim

	term	sim
0	affairs	0.570672
1	estate	0.568760
2	religion	0.567143
3	society	0.564692
4	rights	0.558945

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

Out[74]:

	term	sim
0	devil	0.663740
1	villain	0.663168
2	girl	0.662057
3	creetur	0.660398
4	beggar	0.653608

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

Out[75]:

	term	sim
0	vagabond	0.561235
1	thief	0.558269
2	foreigner	0.557487
3	devil	0.557404
4	villain	0.556923

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

Out[76]:

	term	sim
0	brave	0.642014
1	innocent	0.607824
2	clever	0.591082
3	foolish	0.579697
4	minded	0.575066

In [77]:
```
complete_analogy('woman', 'wise', 'man', 5)
```

Out[77]:

	term	sim

	term	sim
0	reasonable	0.554601
1	sane	0.533982
2	superior	0.519525
3	useful	0.519105
4	rational	0.509219

Similarites

In [46]:
```
get_most_similar('joy')
```

Out[46]:

	term	sim
0	delight	0.778237
1	grief	0.755719
2	terror	0.754897
3	gratitude	0.746084
4	admiration	0.743671
5	gladness	0.729121
6	bitterness	0.727012
7	horror	0.711439
8	earnestness	0.698146
9	rage	0.697935

In [47]:
```
get_most_similar('man')
```

Out[47]:

	term	sim
0	gentleman	0.824611
1	person	0.804918
2	woman	0.780459
3	student	0.718592
4	foreigner	0.698749
5	dog	0.664404
6	creature	0.659028
7	boy	0.656878
8	chap	0.649437
9	soldier	0.648450

In [48]:
```
get_most_similar(positive=['man'], negative=['woman'])
```

Out[48]:

	term	sim
0	diplomatic	0.262227
1	mark	0.260356
2	line	0.251799
3	men	0.246355
4	point	0.236271
5	express	0.231483
6	transact	0.229944
7	patent	0.225700
8	further	0.225159
9	record	0.224411

In [49]:
```
get_most_similar(positive='woman')
```

Out[49]:

	term	sim
0	girl	0.859669
1	man	0.780459
2	creature	0.777335
3	lady	0.759753
4	boy	0.735077
5	wretch	0.728275
6	gentleman	0.722586
7	rascal	0.721813
8	chap	0.712864
9	widow	0.702231

In [50]:
```
get_most_similar(positive=['woman'], negative=['man'])
```

Out[50]:

	term	sim
0	jane	0.476249
1	sweet	0.452006
2	peasant	0.427259
3	weeping	0.420330
4	mary	0.416504
5	baby	0.409885

	term	sim
6	girl	0.401163
7	eldest	0.397838
8	buxom	0.397097
9	sally	0.395690

In [51]:
```
get_most_similar(['man','woman'],['boy','girl'])
```

Out[51]:

	term	sim
0	gentleman	0.331652
1	outward	0.317353
2	moral	0.316811
3	person	0.288236
4	material	0.287824
5	himself	0.286915
6	crime	0.286068
7	sane	0.280468
8	indifference	0.270962
9	prosperous	0.270836

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

Out[59]:

	term	sim
0	experience	0.746146
1	theory	0.725863
2	ideas	0.721126
3	power	0.716801
4	wisdom	0.712457
5	imagination	0.711024
6	design	0.709023
7	recollection	0.708704
8	belief	0.706114
9	profession	0.706039

In [60]:
```
get_most_similar('kindness')
```

Out[60]:

	term	sim

	term	sim
0	gratitude	0.717582
1	devotion	0.710808
2	homage	0.709736
3	condescension	0.696773
4	fortitude	0.696501
5	generosity	0.696357
6	friendship	0.696182
7	fidelity	0.695630
8	forgiveness	0.695235
9	affection	0.692774

In [61]:
```
get_most_similar('adventure')
```

Out[61]:

	term	sim
0	event	0.794039
1	episode	0.781311
2	engagement	0.779032
3	interview	0.772411
4	incident	0.764948
5	anecdote	0.752455
6	exposition	0.752424
7	absurdity	0.750883
8	enterprise	0.747795
9	performance	0.743935

In [78]:
```
get_most_similar('poor')
```

Out[78]:

	term	sim
0	miserable	0.638907
1	wretched	0.638791
2	wicked	0.615074
3	sick	0.604765
4	foolish	0.601449
5	friendless	0.590747
6	silly	0.584538

	term	sim
7	peasant	0.579586
8	brave	0.578708
9	darling	0.575868

In [80]:
```
get_most_similar('money')
```

Out[80]:

	term	sim
0	trouble	0.681588
1	food	0.657160
2	debt	0.618269
3	purchase	0.603006
4	property	0.599279
5	wages	0.586387
6	reward	0.582792
7	security	0.576184
8	bill	0.565190
9	employment	0.563871

In [79]:
```
get_most_similar('rich')
```

Out[79]:

	term	sim
0	healthy	0.670826
1	clever	0.637012
2	picturesque	0.627724
3	thirsty	0.616304
4	tough	0.599806
5	pure	0.598837
6	colored	0.591493
7	hungry	0.588078
8	prim	0.587944
9	showy	0.585651

Save

In [52]:
```
# W2V.to_csv(f'{data_home}/{data_prefix}/{data_prefix}-W2V.csv')
# VOCAB.to_csv(f'{data_home}/{data_prefix}/{data_prefix}-VOCAB.csv')
```

```
# SENTS.to_csv(f'{data_home}/{data_prefix}/{data_prefix}-GENSIM_DOCS.csv')
```

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 [ ]: