Dickens 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("dickens_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(("dickens_pre_LIB.csv"), index_col = ['book_id'])
```

In [6]:

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

In [7]:

```
VOCAB = pd.read_csv("dickens_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	st

term_str	n	n_chars	p	i	max_pos	n_pos	cat_pos	stop	stem_porter	st
term_str										
0	60	1	1.207251e-05	16.337915	CD	4	{'RB', 'CD', 'NN', 'JJ'}	0	0	
1	38	1	7.645923e-06	16.996878	CD	5	{'NNP', 'CD', 'VB', 'NN', 'JJ'}	0	1	
10	8	2	1.609668e-06	19.244805	CD	4	{'NNP', 'IN', 'CD', 'NN'}	0	10	
100	4	3	8.048340e-07	20.244805	CD	4	{'JJ', 'IN', 'CD', 'NN'}	0	100	
1000	1	4	2.012085e-07	22.244805	JJ	1	{'JJ'}	0	1000	

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

Out[10]:

			n	tf	tfidf
book_id	**chap_id**	**term_str**			
98	**1**	**a**	23	0.291139	0.000000
		about	2	0.025316	0.002626
		achievements	1	0.012658	0.082362
		adjacent	1	0.012658	0.059284
		after	2	0.025316	0.002792

In [11]:
```
LIB.head()
```

Out[11]:

	source_file_path	title	chap_regex	author
book_id				
98	Dickens/98-a_tale_of_two_cities.txt	a tale of two cities	^\s*CHAPTER\s*[IVXLCM]+\.$	dickens
564	Dickens/564-the_mystery_of_edwin_drood.txt	the mystery of edwin drood	^CHAPTER\s[IVXLCM]+\.$	dickens
580	Dickens/580-the_pickwick_papers.txt	the pickwick papers	^CHAPTER\s[IVXLCM]+\.\s[A-Z]+	dickens

	source_file_path	title	chap_regex	author	
book_id					
588	Dickens/588-master_humphreys_clock.txt	master humphreys clock	^(?:[IVXLCM]+$|TO THE READERS OF)	dickens	s
644	Dickens/644-the_haunted_man_and_the_ghosts_bar...	the haunted man and the ghosts bargain	^CHAPTER\s[IVXLCM]+$	dickens	s

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]:

	book_id	chap_id	n	tf	tfidf	n_vocab	n_chars	p	i
term_str									
0	588	7	2	0.040000	0.304882	60	1	1.207251e-05	16.337915
0	786	16	1	0.012987	0.098988	60	1	1.207251e-05	16.337915
0	882	47	1	0.001244	0.009480	60	1	1.207251e-05	16.337915
0	912	3	3	0.005714	0.043555	60	1	1.207251e-05	16.337915
0	1414	1	49	0.182836	1.393584	60	1	1.207251e-05	16.337915
...	...	...	...	...	...	...	...	...	...
æolian	699	4	1	0.003333	0.030690	2	6	4.024170e-07	21.244805

	book_id	chap_id	n	tf	tfidf	n_vocab	n_chars	p	i
term_str									
æolian	872	10	1	0.003731	0.034355	2	6	4.024170e-07	21.244805
æsop	35536	2	1	0.007576	0.077326	1	4	2.012085e-07	22.244805
éclat	918	3	1	0.014493	0.147928	1	5	2.012085e-07	22.244805
élite	882	28	1	0.004545	0.046396	1	5	2.012085e-07	22.244805

1336044 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]:

			n	tf	tfidf
book_id	**chap_id**	**term_str**			
98	**6**	**lock**	1	0.004608	0.014574
	4	**watchtower**	1	0.004926	0.050281
	39	**watchmen**	1	0.003704	0.024991
	20	**watchmen**	1	0.004695	0.031679
	38	**fit**	1	0.004405	0.009153
...	**...**	**...**	...	...	...
35536	**11**	**refers**	1	0.014493	0.086364
	4	**sturdy**	1	0.008547	0.035578
	8	**referred**	1	0.027027	0.074259
	2	**court**	1	0.007576	0.012577
	13	**hundreds**	1	0.045455	0.168146

1336044 rows × 3 columns

In [14]:

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

Out[14]:

```
0.03497452084379164
```

In [15]:

```
n_topics = 40
n_terms = 2000
```

In [16]:

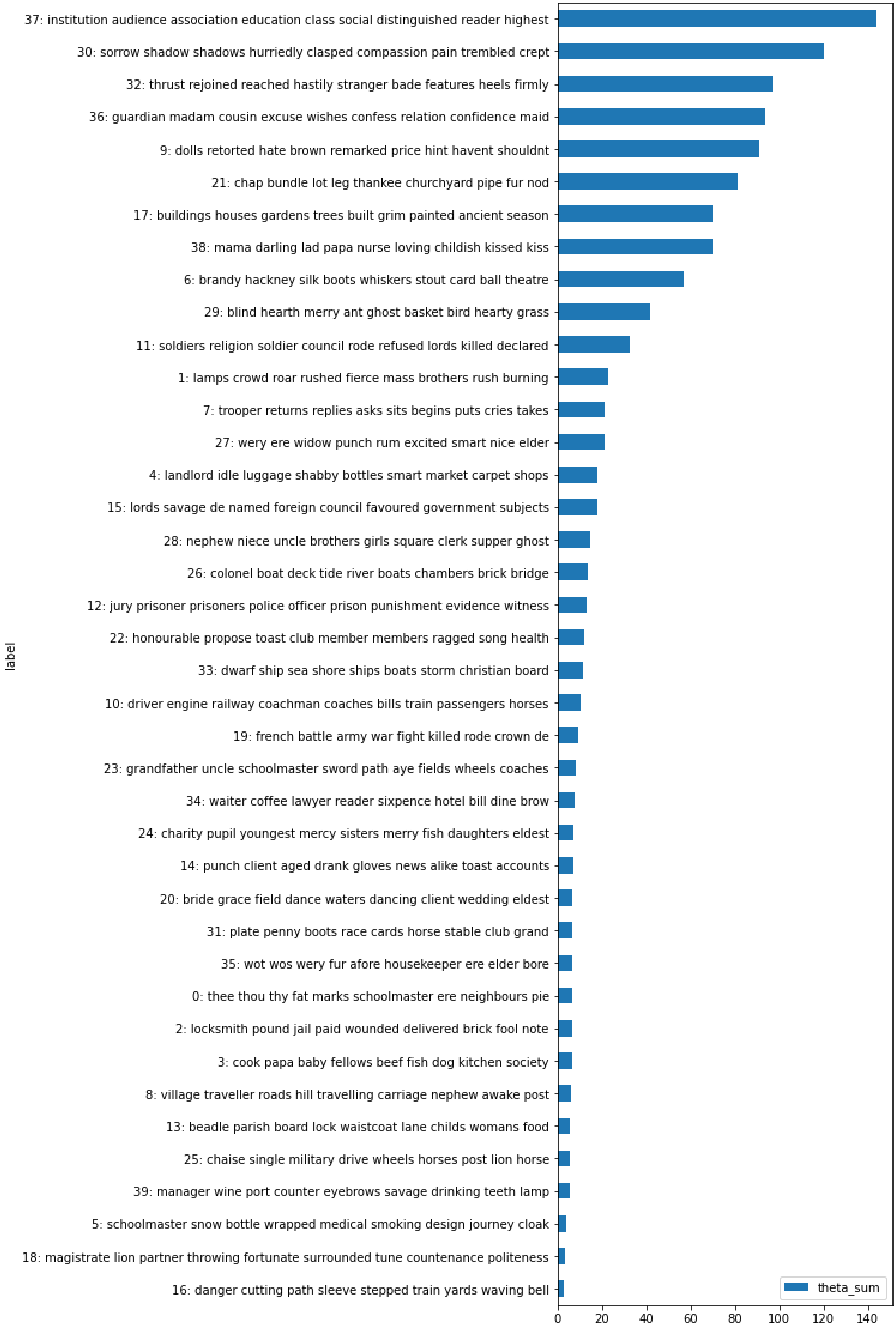
```
tm = TopicModel(filtered_BOW)
tm.n_topics = n_topics
tm.n_terms = n_terms
```

In [17]:

```
tm.create_X()
tm.get_model()
tm.describe_topics()
tm.get_model_stats()
```

In [18]:

```
tm.plot_topics()
```


37: institution audience association education class social distinguished reader highest
30: sorrow shadow shadows hurriedly clasped compassion pain trembled crept
32: thrust rejoined reached hastily stranger bade features heels firmly
36: guardian madam cousin excuse wishes confess relation confidence maid
9: dolls retorted hate brown remarked price hint havent shouldnt
21: chap bundle lot leg thankee churchyard pipe fur nod
17: buildings houses gardens trees built grim painted ancient season
38: mama darling lad papa nurse loving childish kissed kiss
6: brandy hackney silk boots whiskers stout card ball theatre
29: blind hearth merry ant ghost basket bird hearty grass
11: soldiers religion soldier council rode refused lords killed declared
1: lamps crowd roar rushed fierce mass brothers rush burning
7: trooper returns replies asks sits begins puts cries takes
27: wery ere widow punch rum excited smart nice elder
4: landlord idle luggage shabby bottles smart market carpet shops
15: lords savage de named foreign council favoured government subjects
28: nephew niece uncle brothers girls square clerk supper ghost
26: colonel boat deck tide river boats chambers brick bridge
12: jury prisoner prisoners police officer prison punishment evidence witness
22: honourable propose toast club member members ragged song health
33: dwarf ship sea shore ships boats storm christian board
10: driver engine railway coachman coaches bills train passengers horses
19: french battle army war fight killed rode crown de
23: grandfather uncle schoolmaster sword path aye fields wheels coaches
34: waiter coffee lawyer reader sixpence hotel bill dine brow
24: charity pupil youngest mercy sisters merry fish daughters eldest
14: punch client aged drank gloves news alike toast accounts
20: bride grace field dance waters dancing client wedding eldest
31: plate penny boots race cards horse stable club grand
35: wot wos wery fur afore housekeeper ere elder bore
0: thee thou thy fat marks schoolmaster ere neighbours pie
2: locksmith pound jail paid wounded delivered brick fool note
3: cook papa baby fellows beef fish dog kitchen society
8: village traveller roads hill travelling carriage nephew awake post
13: beadle parish board lock waistcoat lane childs womans food
25: chaise single military drive wheels horses post lion horse
39: manager wine port counter eyebrows savage drinking teeth lamp
5: schoolmaster snow bottle wrapped medical smoking design journey cloak
18: magistrate lion partner throwing fortunate surrounded tune countenance politeness
16: danger cutting path sleeve stepped train yards waving bell
label
theta_sum
0
20
40
60
80
100
120
140

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**								
98	**1**	0.000188	0.000188	0.000188	0.000188	0.000188	0.000188	0.000188	0.000
	2	0.000088	0.000088	0.000088	0.000088	0.000088	0.000088	0.000088	0.000
	3	0.000110	0.000110	0.000110	0.000110	0.000110	0.000110	0.000110	0.00
	4	0.000044	0.000044	0.000044	0.000044	0.029046	0.000044	0.119610	0.000
	5	0.004181	0.147800	0.000043	0.000043	0.000043	0.000043	0.000043	0.000
...	**...**	...	...	...	...	...	...	...	
35536	**9**	0.063746	0.000171	0.000171	0.000171	0.000171	0.000171	0.000171	0.00
	10	0.103054	0.216522	0.000455	0.000455	0.000455	0.000455	0.000455	0.000
	11	0.000212	0.000212	0.000212	0.000212	0.000212	0.000212	0.000212	0.00
	12	0.000333	0.000333	0.000333	0.000333	0.000333	0.000333	0.000333	0.000
	13	0.983475	0.000424	0.000424	0.000424	0.000424	0.000424	0.000424	0.000

1182 rows × 40 columns

In [20]:

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

Out[20]:

term_str	lie	understood	youth	third	quickly	difficulty	weak	
topic_id								
0	0.025000	0.025000	11.751554	0.025000	0.025000	3.262398	0.025000	1.
1	12.812068	0.025000	0.025000	9.663417	0.025000	11.992605	0.025000	2
2	0.025000	4.280178	0.025000	5.442221	3.280720	0.025000	0.025000	0.
3	0.025000	4.667417	1.360998	0.025000	0.025000	1.254139	1.199507	22.
4	0.025000	1.138922	0.025000	34.606390	0.027057	5.846419	17.142279	0.
5	0.025000	0.025000	1.743291	2.077884	0.025000	7.946871	0.025000	9.
6	0.025000	15.103489	0.025000	55.330021	0.025000	31.046006	0.025000	70.
7	15.329177	9.860270	7.410073	0.025000	4.568890	6.134426	10.463160	0.
8	1.476239	0.025000	0.025000	4.418940	0.034412	0.025000	0.025000	0.
9	17.413441	47.980245	47.235042	24.647709	32.139769	18.499233	55.647671	40.
10	0.025000	7.102742	0.025000	6.487948	8.816408	0.025000	0.025000	0.
11	0.025000	4.368480	0.025000	28.657035	18.196925	15.769301	29.126880	4
12	5.345035	0.025000	0.025000	3.588518	0.025000	0.025000	0.085434	0.
13	0.025000	0.025000	2.969477	0.025000	0.025000	1.955967	0.025000	0.

term_str	lie	understood	youth	third	quickly	difficulty	weak	
topic_id								
14	0.025000	3.209910	0.025000	0.025000	0.025000	3.202380	0.025000	0.
15	2.483488	7.096558	10.364135	22.040881	0.025000	6.711489	14.041432	0.
16	0.025000	2.316688	1.039195	4.086619	4.202465	0.025000	0.025000	0.
17	98.959734	6.335168	10.413182	31.841678	8.840451	10.584972	0.025000	24.
18	1.614629	0.025000	0.025000	0.025000	1.279513	0.025000	2.859411	0.
19	1.412702	0.025000	7.157730	20.303078	0.025000	0.025000	0.025000	0.
20	0.025000	9.854981	20.248930	0.025000	3.502576	0.025000	0.025000	12.
21	41.048301	40.170254	19.307265	22.998346	3.833785	47.101910	39.504454	115
22	0.025000	1.040363	0.025000	0.025000	0.025000	0.025000	0.025000	0.
23	0.025000	0.025000	0.025000	0.025000	4.000744	0.025000	8.025413	0.
24	0.025000	7.422136	49.838859	0.025000	0.025000	0.025000	0.025000	0.
25	0.025000	0.025000	1.431668	1.669838	0.025000	0.025000	1.349551	5
26	2.887782	15.308292	0.025000	9.385745	0.025000	0.206176	0.025000	4.
27	1.385641	4.261198	0.207395	18.903448	13.199940	2.743130	2.677996	22
28	0.025000	0.025000	19.575153	0.025000	6.220376	0.025000	0.025000	0.
29	27.047825	0.025000	8.362203	0.025000	15.968917	14.570688	5.169756	2.
30	136.305300	17.793508	48.049766	48.845857	116.075363	9.859348	80.865008	0.
31	0.025000	0.025000	0.025000	0.025000	0.025000	9.962279	0.025000	0.
32	23.866075	0.025000	6.755022	24.154232	108.327798	28.696075	33.288211	48.
33	0.025000	7.790811	0.025000	0.025000	7.007532	0.025000	5.211737	0.
34	0.025000	0.025000	5.137558	3.162239	0.025004	1.467164	0.025000	7.
35	0.025000	8.280918	0.025000	0.025000	2.040621	1.202033	0.025000	0.
36	1.909779	77.488227	44.040666	9.478441	26.925608	56.061650	24.737023	10.
37	19.086118	72.326728	59.306196	44.784513	2.956403	95.567785	35.640348	0.
38	25.066668	59.402517	50.869642	0.025000	40.929058	42.032812	66.439728	28.
39	0.025000	0.025000	0.025000	0.025000	4.224665	2.972745	0.025000	0.

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
37	61499.551297	143.579508	10.12	institution audience association education cla...	society respect human institution class knowle...	37: institution audience association education...
30	62807.348840	119.986306	9.99	sorrow shadow shadows hurriedly clasped compas...	breast shadow raised die grave broken earth se...	30: sorrow shadow shadows hurriedly clasped co...
32	53922.415636	96.720517	10.07	thrust rejoined reached hastily stranger bade ...	rejoined stranger reached reply bill hastily c...	32: thrust rejoined reached hastily stranger b...
36	55886.255698	93.561189	10.15	guardian madam cousin excuse wishes confess re...	guardian cousin confidence beg breakfast excus...	36: guardian madam cousin excuse wishes confes...
9	51910.885690	90.708829	10.16	dolls retorted hate brown remarked price hint ...	retorted brown shaking exclaimed rejoined laug...	9: dolls retorted hate brown remarked price hi...
21	46285.859104	81.136282	10.07	chap bundle lot leg thankee churchyard pipe fu...	bottle pipe leg whats piece bit property wante...	21: chap bundle lot leg thankee churchyard pip...
17	43945.628765	69.992820	9.90	buildings houses gardens trees built grim pain...	houses city green windows trees sea sun yard s...	17: buildings houses gardens trees built grim ...
38	41487.755610	69.599950	10.06	mama darling lad papa nurse loving childish ki...	mama parlour loved hardly darling baby lad swe...	38: mama darling lad papa nurse loving childis...
6	32009.362092	56.846382	9.65	brandy hackney silk boots whiskers stout card ...	boots everybody oclock wine blue green pair pl...	6: brandy hackney silk boots whiskers stout ca...
29	23761.898690	41.652583	9.66	blind hearth merry ant ghost basket bird heart...	blind merry cheerful beside comfort ghost bles...	29: blind hearth merry ant ghost basket bird h...
11	17767.267836	32.324709	9.57	soldiers religion soldier council rode refused...	soldiers prison died sent thousand tried relig...	11: soldiers religion soldier council rode ref...
1	11037.564733	22.715944	8.97	lamps crowd roar rushed fierce mass brothers r...	crowd lamps windows brothers doors noise ran d...	1: lamps crowd roar rushed fierce mass brother...

	phi_sum	theta_sum	h	top_terms_rel	top_terms	label
topic_id						
7	10328.489388	21.056573	8.62	trooper returns replies asks sits begins puts ...	returns goes takes cries replies makes trooper...	7: trooper returns replies asks sits begins pu...
27	13008.275634	20.915748	9.56	wery ere widow punch rum excited smart nice elder	wery exclaimed ere countenance servant feeling...	27: wery ere widow punch rum excited smart nic...
4	10984.966741	17.975642	9.17	landlord idle luggage shabby bottles smart mar...	landlord idle rain market smart dirty walking ...	4: landlord idle luggage shabby bottles smart ...
15	8799.698888	17.715621	9.19	lords savage de named foreign council favoured...	thousand lords ran merry named sent died seven...	15: lords savage de named foreign council favo...
28	7192.831248	14.580303	8.85	nephew niece uncle brothers girls square clerk...	uncle nephew brothers niece spirit rejoined gi...	28: nephew niece uncle brothers girls square c...
26	7005.464802	13.339811	8.42	colonel boat deck tide river boats chambers br...	boat river colonel tide board deck bridge boat...	26: colonel boat deck tide river boats chamber...
12	7121.281232	12.985642	8.33	jury prisoner prisoners police officer prison ...	prisoner prison officer police prisoners jury ...	12: jury prisoner prisoners police officer pri...
22	3814.899205	11.799269	8.38	honourable propose toast club member members r...	honourable member members toast propose health...	22: honourable propose toast club member membe...
33	5566.574329	11.501317	8.34	dwarf ship sea shore ships boats storm christi...	sea ship dwarf board shore ships boat wild boats	33: dwarf ship sea shore ships boats storm chr...
10	5522.554745	10.533876	8.40	driver engine railway coachman coaches bills t...	driver horses horse engine railway train stati...	10: driver engine railway coachman coaches bil...
19	4884.785353	9.194422	8.44	french battle army war fight killed rode crown de	french army battle war thousand crown fight ho...	19: french battle army war fight killed rode c...
23	4493.800657	7.960685	8.40	grandfather uncle schoolmaster sword path aye ...	uncle grandfather schoolmaster horses sword di...	23: grandfather uncle schoolmaster sword path ...

	phi_sum	theta_sum	h	top_terms_rel	top_terms	label
topic_id						
34	2601.979589	7.701786	8.22	waiter coffee lawyer reader sixpence hotel bil...	waiter coffee bill lawyer pen reader hotel shi...	34: waiter coffee lawyer reader sixpence hotel...
24	4033.945328	6.918712	7.97	charity pupil youngest mercy sisters merry fis...	charity sisters merry mercy pupil youngest dau...	24: charity pupil youngest mercy sisters merry...
14	1731.221631	6.882468	8.92	punch client aged drank gloves news alike toas...	punch aged client shoulder drank society news ...	14: punch client aged drank gloves news alike ...
20	4135.461960	6.734693	8.45	bride grace field dance waters dancing client ...	bride grace field dance dancing green tree wat...	20: bride grace field dance waters dancing cli...
31	2838.494836	6.616429	8.15	plate penny boots race cards horse stable club...	boots plate horse horses race week penny dust ...	31: plate penny boots race cards horse stable ...
35	4065.736007	6.508327	8.11	wot wos wery fur afore housekeeper ere elder bore	wot wos wery afore fur bore ere housekeeper pipe	35: wot wos wery fur afore housekeeper ere eld...
0	3283.886528	6.490354	8.12	thee thou thy fat marks schoolmaster ere neigh...	thee thy fat thou schoolmaster voices marks er...	0: thee thou thy fat marks schoolmaster ere ne...
2	1970.532319	6.455095	8.34	locksmith pound jail paid wounded delivered br...	locksmith pound paid parlour note honest women...	2: locksmith pound jail paid wounded delivered...
3	2835.357018	6.401074	9.14	cook papa baby fellows beef fish dog kitchen s...	baby papa cook fellows society dog kitchen hal...	3: cook papa baby fellows beef fish dog kitche...
8	3026.361312	5.852191	7.60	village traveller roads hill travelling carria...	village traveller roads post carriage hill sto...	8: village traveller roads hill travelling car...
13	2535.695729	5.711818	8.11	beadle parish board lock waistcoat lane childs...	beadle parish board waistcoat gate lock lane d...	13: beadle parish board lock waistcoat lane ch...
25	2308.756196	5.379319	7.92	chaise single military drive wheels horses pos...	chaise single horses post horse military stage...	25: chaise single military drive wheels horses...

	phi_sum	theta_sum	h	top_terms_rel	top_terms	label
topic_id						
39	2375.148475	5.329834	7.92	manager wine port counter eyebrows savage drin...	wine manager port drinking teeth drink shoulde...	39: manager wine port counter eyebrows savage ...
5	2026.085858	4.125338	8.49	schoolmaster snow bottle wrapped medical smoki...	schoolmaster bottle snow journey wine companio...	5: schoolmaster snow bottle wrapped medical sm...
18	1548.489716	3.473238	8.38	magistrate lion partner throwing fortunate sur...	magistrate lion partner throwing countenance t...	18: magistrate lion partner throwing fortunate...
16	1529.431283	3.035389	8.68	danger cutting path sleeve stepped train yards...	danger bell below ran train line mouth path ring	16: danger cutting path sleeve stepped train y...

Top terms associated with the most frequent topic

In [22]:

```
top_topic = tm.TOPIC.theta_sum.idxmax()

top_topic
```

Out[22]:

```
37
```

In [23]:

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

Out[23]:

```
'institution audience association education class social distinguished reader hi
ghest'
```

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]:

```
['institution', 'audience', 'association', 'education', 'class']
```

In [62]:

```
# 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[62]:

			0	1	2	3	4	5
book_id	**title**	**type**						
98	**a tale of two cities**	**novel**	0.003095	0.049294	0.021320	0.000289	0.008174	0.002549
564	**the mystery of edwin drood**	**novel**	0.000882	0.014619	0.000058	0.000058	0.005031	0.000058
580	**the pickwick papers**	**novel**	0.017492	0.006685	0.000039	0.001701	0.013594	0.002172
588	**master humphreys clock**	**stories**	0.016547	0.008285	0.000048	0.000048	0.000048	0.000048
644	**the haunted man and the ghosts bargain**	**stories**	0.001645	0.000020	0.000020	0.000020	0.000020	0.000020
650	**pictures from italy**	**non-fiction**	0.000054	0.014090	0.000054	0.000054	0.027197	0.000054
653	**the chimes**	**novel**	0.003679	0.012996	0.000026	0.000026	0.000026	0.000026
675	**american notes**	**non-fiction**	0.000161	0.007471	0.000035	0.000035	0.009271	0.000035
676	**the battle of life**	**novel**	0.002710	0.005891	0.000019	0.000019	0.000019	0.000019
699	**a childs history of england**	**non-fiction**	0.004349	0.012696	0.000068	0.001068	0.000068	0.000068
700	**the old curiosity shop**	**novel**	0.019703	0.008615	0.000066	0.000886	0.003344	0.003749
730	**oliver twist**	**novel**	0.002013	0.057279	0.000070	0.001779	0.006075	0.001888
766	**david copperfield**	**novel**	0.000740	0.006051	0.000061	0.001274	0.009263	0.003977
786	**hard times**	**novel**	0.025390	0.012005	0.000083	0.062298	0.000083	0.001818
807	**hunted down**	**stories**	0.000178	0.008649	0.000178	0.000178	0.003736	0.000178
809	**holiday romance**	**stories**	0.005316	0.000063	0.000063	0.322941	0.000063	0.002141
810	**george silvermans explanation**	**stories**	0.005418	0.076060	0.000560	0.000560	0.000560	0.000560
821	**dombey and sons**	**novel**	0.000219	0.011038	0.000039	0.003943	0.003668	0.000108
824	**speeches of charles dickens**	**non-fiction**	0.001904	0.000243	0.000243	0.006764	0.002184	0.000567
872	**reprinted pieces**	**stories**	0.005178	0.009277	0.043381	0.000675	0.065346	0.000420

			0	1	2	3	4	5	
book_id	title	type							
882	sketches by boz	stories	0.000113	0.013861	0.000094	0.000102	0.067170	0.003194	0
883	our mutual friend	novel	0.000803	0.012758	0.000047	0.003109	0.011917	0.020673	0
888	the lazy tour of two idle apprentices	stories	0.000023	0.009487	0.000023	0.000023	0.331001	0.000023	0
912	the mudfog and other sketches	stories	0.000069	0.024095	0.000313	0.019520	0.000069	0.000069	0
914	the uncommerical traveller	non-fiction	0.002214	0.019466	0.000064	0.001466	0.072747	0.013172	0
916	sketches of young couples	stories	0.000141	0.003650	0.000141	0.007156	0.005711	0.000141	0
917	barnaby rudge	stories	0.001239	0.071070	0.049184	0.000122	0.002032	0.000067	0
918	sketches of young gentlemen	stories	0.000149	0.000149	0.000149	0.000149	0.000149	0.000149	
922	sunday under three heads	non-fiction	0.003468	0.000050	0.000050	0.000050	0.019717	0.000050	0
927	the lamplighter	stories	0.039326	0.037789	0.000029	0.000029	0.000029	0.000029	0
967	nicholas nickleby	novel	0.007838	0.008518	0.000045	0.003139	0.003025	0.005003	0
968	martin chuzzlewit	novel	0.007203	0.003847	0.000038	0.000038	0.008150	0.001971	0
1023	bleak house	novel	0.001276	0.006708	0.000048	0.000106	0.009292	0.004088	0
1289	three ghost stories	stories	0.002306	0.000032	0.000032	0.000032	0.043389	0.000032	0
1394	the holly tree	stories	0.002180	0.039885	0.000077	0.000077	0.053799	0.042825	0
1400	great expectations	novel	0.000264	0.040057	0.000083	0.020751	0.007403	0.003673	0
1406	the perils of certain english prisoners	stories	0.000024	0.000024	0.000024	0.000024	0.000024	0.000024	0
1407	a message from the sea	stories	0.000050	0.000050	0.000050	0.000050	0.000050	0.000050	0
1413	tom tiddlers ground	stories	0.000107	0.000107	0.000107	0.002767	0.066765	0.000107	0
1414	somebodys luggage	stories	0.006034	0.019426	0.000045	0.000045	0.106043	0.000045	0

			0	1	2	3	4	5	
book_id	**title**	**type**							
1415	**doctor marigold**	**stories**	0.000280	0.000280	0.191444	0.000280	0.107787	0.000280	0
1416	**mrs lirripers lodgings**	**stories**	0.000070	0.013488	0.000070	0.000070	0.030797	0.000070	0
1421	**mrs lirripers legacy**	**stories**	0.003441	0.022894	0.000079	0.000079	0.000079	0.000079	0
1435	**miscellaneous papers**	**non-fiction**	0.008926	0.004221	0.000120	0.000120	0.000120	0.000894	0
1467	**some christmas stories**	**stories**	0.002766	0.000074	0.000074	0.166397	0.011583	0.000074	0
2324	**a house to let**	**stories**	0.005841	0.000043	0.000043	0.016862	0.070749	0.000043	0
19337	**a christmas carol**	**novel**	0.004572	0.007833	0.000043	0.000043	0.006771	0.002314	0
20795	**the cricket on the hearth**	**novel**	0.001631	0.000019	0.000019	0.000019	0.000019	0.000019	0
27924	**mugby junction**	**stories**	0.000031	0.043337	0.000031	0.000031	0.056755	0.000031	0
35536	**the poems and verses of charles dickens**	**stories**	0.091732	0.019923	0.000288	0.000288	0.006258	0.000288	0

In [66]:

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

Out[66]:

```
type
non-fiction    37
novel          30
stories        37
dtype: int64
```

In [80]:

```
# 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[80]:

	topic_id	title	type	phi_sum	theta_sum	h	top_terms_rel
book_id							

book_id	topic_id	title	type	phi_sum	theta_sum	h	top_terms_rel	
1421	38	mrs lirripers legacy	stories	41487.755610	69.599950	10.060000	mama darling lad papa nurse loving childish kissed kiss	pa da
821	38	dombey and sons	novel	41487.755610	69.599950	10.060000	mama darling lad papa nurse loving childish kissed kiss	pa da
766	38	david copperfield	novel	41487.755610	69.599950	10.060000	mama darling lad papa nurse loving childish kissed kiss	pa da
912	37	the mudfog and other sketches	stories	61499.551297	143.579508	10.120000	institution audience association education class social distinguished reader highest	sc
807	37	hunted down	stories	61499.551297	143.579508	10.120000	institution audience association education class social distinguished reader highest	sc
810	37	george silvermans explanation	stories	61499.551297	143.579508	10.120000	institution audience association education class social distinguished reader highest	sc
824	37	speeches of charles dickens	non-fiction	61499.551297	143.579508	10.120000	institution audience association education class social distinguished reader highest	sc

	topic_id	title	type	phi_sum	theta_sum	h	top_terms_rel	
book_id								
588	37	master humphreys clock	stories	61499.551297	143.579508	10.120000	institution audience association education class social distinguished reader highest	sc
916	37	sketches of young couples	stories	61499.551297	143.579508	10.120000	institution audience association education class social distinguished reader highest	sc
918	37	sketches of young gentlemen	stories	61499.551297	143.579508	10.120000	institution audience association education class social distinguished reader highest	sc
1435	37	miscellaneous papers	non-fiction	61499.551297	143.579508	10.120000	institution audience association education class social distinguished reader highest	sc
35536	37	the poems and verses of charles dickens	stories	61499.551297	143.579508	10.120000	institution audience association education class social distinguished reader highest	sc
1023	36	bleak house	novel	55886.255698	93.561189	10.150000	guardian madam cousin excuse wishes confess relation confidence maid	e m
564	36	the mystery of edwin drood	novel	55886.255698	93.561189	10.150000	guardian madam cousin excuse wishes confess relation confidence maid	e m

book_id	topic_id	title	type	phi_sum	theta_sum	h	top_terms_rel	t
922	34	sunday under three heads	non-fiction	2601.979589	7.701786	8.220000	waiter coffee lawyer reader sixpence hotel bill dine brow	wa i
1406	33	the perils of certain english prisoners	stories	5566.574329	11.501317	8.340000	dwarf ship sea shore ships boats storm christian board	d s
967	32	nicholas nickleby	novel	53922.415636	96.720517	10.070000	thrust rejoined reached hastily stranger bade features heels firmly	fu
917	32	barnaby rudge	stories	53922.415636	96.720517	10.070000	thrust rejoined reached hastily stranger bade features heels firmly	fu
730	32	oliver twist	novel	53922.415636	96.720517	10.070000	thrust rejoined reached hastily stranger bade features heels firmly	fu
700	32	the old curiosity shop	novel	53922.415636	96.720517	10.070000	thrust rejoined reached hastily stranger bade features heels firmly	fu
653	30	the chimes	novel	62807.348840	119.986306	9.990000	sorrow shadow shadows hurriedly clasped compassion pain trembled crept	gr e
2324	30	a house to let	stories	62807.348840	119.986306	9.990000	sorrow shadow shadows hurriedly clasped compassion pain trembled crept	gr e

	topic_id	title	type	phi_sum	theta_sum	h	top_terms_rel	
book_id								
1467	30	some christmas stories	stories	62807.348840	119.986306	9.990000	sorrow shadow shadows hurriedly clasped compassion pain trembled crept	gr e
786	30	hard times	novel	62807.348840	119.986306	9.990000	sorrow shadow shadows hurriedly clasped compassion pain trembled crept	gr e
644	30	the haunted man and the ghosts bargain	stories	62807.348840	119.986306	9.990000	sorrow shadow shadows hurriedly clasped compassion pain trembled crept	gr e
98	30	a tale of two cities	novel	62807.348840	119.986306	9.990000	sorrow shadow shadows hurriedly clasped compassion pain trembled crept	gr e
20795	29	the cricket on the hearth	novel	23761.898690	41.652583	9.660000	blind hearth merry ant ghost basket bird hearty grass	b g
19337	29	a christmas carol	novel	23761.898690	41.652583	9.660000	blind hearth merry ant ghost basket bird hearty grass	b g
580	27	the pickwick papers	novel	13008.275634	20.915748	9.560000	wery ere widow punch rum excited smart nice elder	co wh
1394	21	the holly tree	stories	46285.859104	81.136282	10.070000	chap bundle lot leg thankee churchyard pipe fur nod	w

	topic_id	title	type	phi_sum	theta_sum	h	top_terms_rel
book_id							
1416	21	mrs lirripers lodgings	stories	46285.859104	81.136282	10.070000	chap bundle lot leg thankee churchyard pipe fur nod
1415	21	doctor marigold	stories	46285.859104	81.136282	10.070000	chap bundle lot leg thankee churchyard pipe fur nod
1414	21	somebodys luggage	stories	46285.859104	81.136282	10.070000	chap bundle lot leg thankee churchyard pipe fur nod
1413	21	tom tiddlers ground	stories	46285.859104	81.136282	10.070000	chap bundle lot leg thankee churchyard pipe fur nod
1407	21	a message from the sea	stories	46285.859104	81.136282	10.070000	chap bundle lot leg thankee churchyard pipe fur nod
1400	21	great expectations	novel	46285.859104	81.136282	10.070000	chap bundle lot leg thankee churchyard pipe fur nod
676	20	the battle of life	novel	4135.461960	6.734693	8.450000	bride grace field dance waters dancing client wedding eldest
914	17	the uncommerical traveller	non-fiction	43945.628765	69.992820	9.900000	buildings houses gardens trees built grim painted ancient season
872	17	reprinted pieces	stories	43945.628765	69.992820	9.900000	buildings houses gardens trees built grim painted ancient season

	topic_id	title	type	phi_sum	theta_sum	h	top_terms_rel	t
book_id								
675	17	american notes	non-fiction	43945.628765	69.992820	9.900000	buildings houses gardens trees built grim painted ancient season	h
650	17	pictures from italy	non-fiction	43945.628765	69.992820	9.900000	buildings houses gardens trees built grim painted ancient season	h
1289	16	three ghost stories	stories	1529.431283	3.035389	8.680000	danger cutting path sleeve stepped train yards waving bell	c r
699	11	a childs history of england	non-fiction	17767.267836	32.324709	9.570000	soldiers religion soldier council rode refused lords killed declared	h tri
27924	10	mugby junction	stories	5522.554745	10.533876	8.400000	driver engine railway coachman coaches bills train passengers horses	dri ho ra
968	9	martin chuzzlewit	novel	51910.885690	90.708829	10.160000	dolls retorted hate brown remarked price hint havent shouldnt	la
927	9	the lamplighter	stories	51910.885690	90.708829	10.160000	dolls retorted hate brown remarked price hint havent shouldnt	la
883	9	our mutual friend	novel	51910.885690	90.708829	10.160000	dolls retorted hate brown remarked price hint havent shouldnt	la

	topic_id	title	type	phi_sum	theta_sum	h	top_terms_rel
book_id							
882	6	sketches by boz	stories	32009.362092	56.846382	9.650000	brandy hackney silk boots whiskers stout card ball theatre
888	4	the lazy tour of two idle apprentices	stories	10984.966741	17.975642	9.170000	landlord idle luggage shabby bottles smart market carpet shops
809	3	holiday romance	stories	2835.357018	6.401074	9.140000	cook papa baby fellows beef fish dog kitchen society

Works and Top Terms Associated with Each Topic

In [109…

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

In [110…

```
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[110…

	count	title	top_terms_rel
topic_id			
37	9	master humphreys clock, hunted down, george silvermans explanation, speeches of charles dickens, the mudfog and other sketches, sketches of young couples, sketches of young gentlemen, miscellaneous papers, the poems and verses of charles dickens	institution audience association education class social distinguished reader highest
21	7	the holly tree, great expectations, a message from the sea, tom tiddlers ground, somebodys luggage, doctor marigold, mrs lirripers lodgings	chap bundle lot leg thankee churchyard pipe fur nod
30	6	a tale of two cities, the haunted man and the ghosts bargain, the chimes, hard times, some christmas stories, a house to let	sorrow shadow shadows hurriedly clasped compassion pain trembled crept
17	4	pictures from italy, american notes, reprinted pieces, the uncommerical traveller	buildings houses gardens trees built grim painted ancient season

	count	title	top_terms_rel
topic_id			
32	4	the old curiosity shop, oliver twist, barnaby rudge, nicholas nickleby	thrust rejoined reached hastily stranger bade features heels firmly
38	3	david copperfield, dombey and sons, mrs lirripers legacy	mama darling lad papa nurse loving childish kissed kiss
9	3	our mutual friend, the lamplighter, martin chuzzlewit	dolls retorted hate brown remarked price hint havent shouldnt
29	2	a christmas carol, the cricket on the hearth	blind hearth merry ant ghost basket bird hearty grass
36	2	the mystery of edwin drood, bleak house	guardian madam cousin excuse wishes confess relation confidence maid
16	1	three ghost stories	danger cutting path sleeve stepped train yards waving bell
11	1	a childs history of england	soldiers religion soldier council rode refused lords killed declared
20	1	the battle of life	bride grace field dance waters dancing client wedding eldest
4	1	the lazy tour of two idle apprentices	landlord idle luggage shabby bottles smart market carpet shops
27	1	the pickwick papers	wery ere widow punch rum excited smart nice elder
10	1	mugby junction	driver engine railway coachman coaches bills train passengers horses
33	1	the perils of certain english prisoners	dwarf ship sea shore ships boats storm christian board
34	1	sunday under three heads	waiter coffee lawyer reader sixpence hotel bill dine brow

topic_id	count	title	top_terms_rel
6	1	sketches by boz	brandy hackney silk boots whiskers stout card ball theatre
3	1	holiday romance	cook papa baby fellows beef fish dog kitchen society

In [111…

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

M09: Word Embeddings

In [28]:

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

In [29]:

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

In [30]:

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

In [31]:

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

In [32]:

```
W2V.head()
```

Out[32]:

term_str	0	1	2	3	4	5	6	7
0	-0.086669	0.040908	0.093690	0.034416	0.038068	-0.176783	0.064615	0.325522
1	-0.108333	0.138857	0.135974	0.027027	0.064037	-0.160321	0.058961	0.240217
1841	-0.020596	0.155038	0.096934	0.055339	-0.058595	-0.225135	-0.012934	0.295945
1842	-0.042258	0.149591	0.054136	0.151247	0.021083	-0.241527	-0.052447	0.246394
1844	0.056815	0.099444	0.027985	0.069834	-0.027610	-0.272625	0.030844	0.289546

5 rows × 100 columns

In [33]:

```
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 [34]:

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

In [35]:

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

In [36]:

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

In [37]:

```
COORDS
```

Out[37]:

	x	y	n	dfidf	pos_group	log_n
term_str						
0	6.764849	-7.233729	60	45.732311	CD	4.094345
1	5.966598	-7.085509	38	134.929244	CD	3.637586
1841	-11.627644	-28.026684	11	63.333804	CD	2.397895
1842	4.192721	-2.623560	17	68.850862	CD	2.833213
1844	4.297374	-2.655729	12	51.797616	CD	2.484907
...	...	...	...	...	...	...
zealous	-50.287689	32.855015	51	221.858487	JJ	3.931826
zenith	-48.237213	-23.476273	12	79.464622	NN	2.484907
zest	-53.370853	-10.908454	18	108.667608	NN	2.890372
zoological	4.185986	-24.232597	10	63.333804	JJ	2.302585
à	-6.250517	76.625725	50	74.223410	NN	3.912023

16515 rows × 6 columns

In [112…

```
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 [113…

```
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 [67]:

```
noun_COORDS = COORDS.loc[COORDS.pos_group == 'NN']

noun_COORDS
```

Out[67]:

	x	y	n	dfidf	pos_group	log_n
term_str						
aaron	-14.241907	95.049797	16	32.828057	NN	2.772589
aback	-21.607054	-14.214270	19	99.312229	NN	2.944439
abandonment	-46.959370	-8.260768	14	84.585470	NN	2.639057
abbey	10.990131	84.507027	184	237.391975	NN	5.214936
abbeys	8.985194	86.130791	12	39.425431	NN	2.484907
...	...	...	...	...	...	...
yup	9.715828	47.414139	11	10.207014	NN	2.397895
zeal	-58.171692	-14.548573	43	202.217270	NN	3.761200
zenith	-48.237213	-23.476273	12	79.464622	NN	2.484907
zest	-53.370853	-10.908454	18	108.667608	NN	2.890372
à	-6.250517	76.625725	50	74.223410	NN	3.912023

9450 rows × 6 columns

Noun tSNE plot

In [114…

```
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

- ease, liberty, credit, comfort, use, reign → idea that comfort, ease related to money, reign... class differences?
- mistrust, warrant, venture, judge, play → trust and judgment??
- nurse, servant, housekeeper, lad, boy, fellow, physician → domestic occupations / roles (gender roles also...??)
- collision, dart, crack, fight, rolls, ooze, plough, whisking → action, trepidation??
- courtyard, chapel, prison, house, room → locations where many scenes occur

Analogies and Similarities (vector algebra)

In [40]:

```
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 [41]:

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

Out[41]:

	term	sim
0	girl	0.837573
1	baby	0.777271
2	child	0.754371

In [42]:

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

Out[42]:

	term	sim
0	son	0.798225
1	sister	0.772192
2	wife	0.757186

In [43]:

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

Out[43]:

	term	sim

	term	sim
0	niece	0.781419
1	daughter	0.780853
2	father	0.768486

In [44]:

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

Out[44]:

	term	sim
0	lady	0.795574
1	housekeeper	0.761616
2	widow	0.739700
3	girl	0.737553
4	priest	0.701211

In [45]:

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

Out[45]:

	term	sim
0	gentleman	0.786933
1	person	0.630580
2	clergyman	0.606372
3	housekeeper	0.582045
4	genlmn	0.570784

In [46]:

```
complete_analogy('day', 'sun', 'night', 5)
```

Out[46]:

	term	sim
0	moon	0.775283
1	wind	0.751996
2	rain	0.725211
3	sky	0.721906
4	clouds	0.717258

In [115…

```
complete_analogy('king', 'rich', 'servant', 5)
```

Out[115…

	term	sim
0	handsome	0.684049
1	shabby	0.681852

	term	sim
2	nice	0.668962
3	queer	0.638489
4	smart	0.619074

In [116…

```
complete_analogy('lord', 'rich', 'servant', 5)
```

Out[116…

	term	sim
0	shabby	0.690646
1	tall	0.620994
2	neat	0.619176
3	handsome	0.602545
4	dirty	0.597426

In [117…

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

Out[117…

	term	sim
0	voyage	0.703066
1	arrival	0.602386
2	trial	0.586183
3	departure	0.584255
4	eve	0.583341

In [118…

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

Out[118…

	term	sim
0	trial	0.686139
1	judgment	0.629092
2	success	0.626624
3	absence	0.619787
4	departure	0.615857

In [119…

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

Out[119…

	term	sim
0	sex	0.598334
1	existence	0.597544

	term	sim
2	history	0.596154
3	affairs	0.587139
4	misfortunes	0.586858

In [120…

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

Out[120…

	term	sim
0	wretch	0.687669
1	silly	0.675156
2	creetur	0.664741
3	brute	0.664144
4	villain	0.646810

In [121…

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

Out[121…

	term	sim
0	vagabond	0.628066
1	monster	0.606369
2	devil	0.605471
3	brute	0.586062
4	madman	0.570789

In [122…

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

Out[122…

	term	sim
0	devilish	0.627629
1	artful	0.615905
2	industrious	0.598349
3	handy	0.592551
4	thoughtless	0.591561

In [123…

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

Out[123…

	term	sim
0	reasonable	0.563518
1	useful	0.553319

	term	sim
2	sensible	0.551846
3	uncommon	0.531260
4	absurd	0.530238

Similarites

In [47]:

```
get_most_similar('joy')
```

Out[47]:

	term	sim
0	grief	0.766438
1	delight	0.747182
2	gratitude	0.740493
3	admiration	0.739902
4	compassion	0.734604
5	sympathy	0.719900
6	contempt	0.703811
7	affection	0.699071
8	firmness	0.689040
9	tenderness	0.684935

In [70]:

```
get_most_similar('servant')
```

Out[70]:

	term	sim
0	maid	0.814667
1	nurse	0.768200
2	lodger	0.754837
3	housekeeper	0.713236
4	wife	0.708198
5	clerk	0.707178
6	daughter	0.704906
7	priest	0.698037
8	niece	0.691128
9	relation	0.687183

In [75]:

```
get_most_similar('king')
```

Out[75]:

	term	sim
0	queen	0.784870
1	duke	0.745500
2	earl	0.684949
3	prince	0.681742
4	pope	0.674839
5	president	0.660339
6	henry	0.626848
7	army	0.620704
8	barons	0.611010
9	archbishop	0.594176

In [76]:

```
get_most_similar('knowledge')
```

Out[76]:

	term	sim
0	experience	0.781393
1	crime	0.739452
2	existence	0.739142
3	memory	0.731023
4	design	0.730296
5	belief	0.729291
6	weakness	0.726590
7	merits	0.724801
8	imagination	0.721957
9	wealth	0.713997

In [71]:

```
get_most_similar('church')
```

Out[71]:

	term	sim
0	cathedral	0.838048
1	hall	0.806354
2	inn	0.801049
3	tower	0.800412
4	gallery	0.791313
5	maypole	0.789417
6	village	0.786119

	term	sim
7	churchyard	0.784754
8	palace	0.784076
9	park	0.773788

In [72]:

```
get_most_similar('poor')
```

Out[72]:

	term	sim
0	wretched	0.647956
1	wicked	0.642902
2	silly	0.641307
3	dearest	0.636960
4	miserable	0.619008
5	brave	0.613389
6	foolish	0.604288
7	darling	0.593470
8	sick	0.591913
9	dear	0.589218

In [77]:

```
get_most_similar('rich')
```

Out[77]:

	term	sim
0	shabby	0.714806
1	rare	0.693201
2	hungry	0.684507
3	lazy	0.684202
4	funny	0.676628
5	clever	0.675720
6	healthy	0.668881
7	handsome	0.667154
8	clad	0.666034
9	thirsty	0.662678

In [73]:

```
get_most_similar('money')
```

Out[73]:

	term	sim
0	trouble	0.642048

	term	sim
1	debt	0.623403
2	em	0.584595
3	shelter	0.566257
4	security	0.564140
5	evidence	0.554088
6	comfort	0.541906
7	match	0.533711
8	employment	0.530339
9	luggage	0.529704

In [74]:

```
get_most_similar('duty')
```

Out[74]:

	term	sim
0	feelings	0.683477
1	weakness	0.646468
2	kindness	0.643070
3	conduct	0.642216
4	readers	0.632523
5	advice	0.632393
6	memory	0.630948
7	consent	0.628821
8	happiness	0.626369
9	belief	0.624754

In [79]:

```
get_most_similar('kindness')
```

Out[79]:

	term	sim
0	friendship	0.771272
1	affection	0.769321
2	happiness	0.761251
3	gratitude	0.754086
4	tenderness	0.752351
5	devotion	0.750932
6	gratification	0.732739
7	goodness	0.731585

	term	sim
8	praise	0.720555
9	vanity	0.716911

In [48]:

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

Out[48]:

	term	sim
0	gentleman	0.836856
1	woman	0.793284
2	person	0.756895
3	lady	0.663957
4	soldier	0.660217
5	dog	0.659054
6	clergyman	0.658894
7	boy	0.643173
8	chap	0.641801
9	priest	0.631909

In [49]:

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

Out[49]:

	term	sim
0	further	0.301181
1	high	0.262769
2	particular	0.242779
3	greater	0.242681
4	great	0.240039
5	special	0.238332
6	sooner	0.236409
7	moral	0.235225
8	favourable	0.233940
9	vast	0.232693

In [50]:

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

Out[50]:

	term	sim
0	girl	0.861149
1	man	0.793284

	term	sim
2	creature	0.791718
3	lady	0.773513
4	wretch	0.771722
5	housekeeper	0.761022
6	gentleman	0.760796
7	priest	0.744040
8	boy	0.735630
9	chap	0.719009

In [51]:

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

Out[51]:

	term	sim
0	jane	0.467104
1	miss	0.425777
2	screamed	0.413791
3	sobbed	0.393252
4	lucie	0.392523
5	tippins	0.391272
6	weeping	0.389153
7	maria	0.385204
8	sobbing	0.376695
9	girl	0.368548

In [52]:

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

Out[52]:

	term	sim
0	gentleman	0.379001
1	men	0.300544
2	himself	0.290749
3	outward	0.280235
4	an	0.268891
5	violent	0.259708
6	themselves	0.257858
7	suspicious	0.248906
8	change	0.248604

	term	sim
9	stronger	0.245816

Save

In [53]:

```
# 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
- Setting pandas df column width with `pd.set_option(display.max_colwidth', None)` to prevent truncating column values: https://pandas.pydata.org/docs/user_guide/options.html

In [ ]:

```
```