Named Entity Recognition & Topic Modeling

In [ ]:

```
# installs

!pip install sentence-transformers
```

In [ ]:

```
# for Bert NER
!pip install transformers
!pip install simpletransformers
```

In [ ]:

```
!pip install umap-learn
```

In [ ]:

```
!pip install hdbscan
```

In [ ]:

```
# pytorch for GPU

!pip3 install torch==1.9.0+cu111 torchvision==0.10.0+cu111 -f https://download.pytorch.org/whl/torch_stab
le.html
```

In [5]:

```
# check if GPU available
import torch
torch.cuda.is_available()
```

Out[5]:

```
True
```

In [6]:

```python
# imports

import pandas as pd
import numpy as np

import gensim

from sentence_transformers import SentenceTransformer
import hdbscan
import umap

# import umap.umap_ as umap # ALTERNATIVE UMAP IMPORT

import matplotlib.pyplot as plt

from sklearn.feature_extraction.text import CountVectorizer
```

In [ ]:

```python
# Read data into movies

movies = pd.read_csv('./mpst_full_data.csv')

# Print head
movies.head()
```

Out[ ]:

	imdb_id	title	plot_synopsis	tags	split	synopsis_source
0	tt0057603	I tre volti della paura	Note: this synopsis is for the orginal Italian...	cult, horror, gothic, murder, atmospheric	train	imdb
1	tt1733125	Dungeons & Dragons: The Book of Vile Darkness	Two thousand years ago, Nhagruul the Foul, a s...	violence	train	imdb
2	tt0033045	The Shop Around the Corner	Matuschek's, a gift store in Budapest, is the ...	romantic	test	imdb
3	tt0113862	Mr. Holland's Opus	Glenn Holland, not a morning person by anyone'...	inspiring, romantic, stupid, feel-good	train	imdb
4	tt0086250	Scarface	In May 1980, a Cuban man named Tony Montana (A...	cruelty, murder, dramatic, cult, violence, atm...	val	imdb

In [ ]:

```
# drop coulmns that won't be used
movies = movies.drop(columns=['split', 'synopsis_source'], axis=1)
```

In [ ]:

```
movies.head()
```

Out[ ]:

	imdb_id	title	plot_synopsis	tags
0	tt0057603	I tre volti della paura	Note: this synopsis is for the orginal Italian...	cult, horror, gothic, murder, atmospheric
1	tt1733125	Dungeons & Dragons: The Book of Vile Darkness	Two thousand years ago, Nhagruul the Foul, a s...	violence
2	tt0033045	The Shop Around the Corner	Matuschek's, a gift store in Budapest, is the ...	romantic
3	tt0113862	Mr. Holland's Opus	Glenn Holland, not a morning person by anyone'...	inspiring, romantic, stupid, feel-good
4	tt0086250	Scarface	In May 1980, a Cuban man named Tony Montana (A...	cruelty, murder, dramatic, cult, violence, atm...

PREPROCESSING

Preprocessing of plot_synopsis:

- remove punctuation
- remove stopwords
- remove Named Entities (Person)

In [ ]:

```
# remove punctuation and stopwords from text
# DO NOT convert to lowercase; Bert NER case-sensitive
def preprocess(text):
    no_punct = gensim.parsing.preprocessing.strip_punctuation(text)
    no_stops = gensim.parsing.preprocessing.remove_stopwords(no_punct)

    return no_stops

movies['plot_text_processed'] = movies['plot_synopsis'].map(lambda x: preprocess(x))

movies['plot_text_processed'].head()
```

Out[ ]:

```
0    Note synopsis orginal Italian release segments...
1    Two thousand years ago Nhagruul Foul sorcerer ...
2    Matuschek s gift store Budapest workplace Alfr...
3    Glenn Holland morning person s standards woken...
4    In May 1980 Cuban man named Tony Montana Al Pa...
Name: plot_text_processed, dtype: object
```

NAMED ENTITY REMOVAL

In [ ]:

```
# BERT named entitiy recognition model

from simpletransformers.ner import NERModel, NERArgs

model_args = NERArgs()
model_args.silent = True # no progress bar when running model (on multiple movies)

englishmodel = NERModel(
        model_type="bert",
        model_name="dslim/bert-base-NER",
        args=model_args,
        use_cuda=True
)
```

In [ ]:

```
def remove_named_entities_bert(text):
    result = []

    # get named entities in text
    # sensitive to case: input should not be all-lowercase
    prediction, raw_output = englishmodel.predict([text])

    # pred: [[{'Matuschek': 'B-PER'}, {'gift': 'O'}, ...]]
    for tag_dict in prediction[0]:
      for token in tag_dict: # iterate dict keys

        # keep token only if not a person entity
        # person entitities are 'B-PER' and 'I-PER'
        if 'PER' not in tag_dict[token]:
          result.append(token)
    return ' '.join(result) # text with entities removed


# TEST
test_movie = movies['plot_text_processed'][2]
print(test_movie)
print('RESULT:', remove_named_entities_bert(test_movie))
```

```
Matuschek s gift store Budapest workplace Alfred Kralik James Stewart newly hi Ed Klara Novak Margaret Su
llavan At work constantly irritate daily aggravation tempered fact secret pen pal trade long soul searchi
ng letters Romantic correspondence sent forth Alfred Klara trade barbs work dream someday meeting sensiti
ve caring unknown pen pal Christmas fast approaching store busy Alfred store time treated Mr Matuschek Fr
ank Morgan lately attitude changed Alfred loss Matuschek avoids explanation finally telling Alfred best l
eft Stunned Alfred accepts paycheck says goodbye including Klara For civil A long awaited meeting secret 
pen pals planned night Alfred having lost job desire Finding t fight curiosity wanders restaurant d agree
d meet peeks window fellow employee Of course Klara waiting chosen book wearing red carnation d agreed us
e signal Realizing d wrong irritation actually masking attraction finally enters goes table reveal true r
eason aware hurt pen pal t Alfred hurt rudeness finally leaves knowing wait night longer coming Meanwhile
store Mr Matuschek late night meeting private detective He knows wife having affair employees convinced t
rusted friend Alfred The detective tells Matuschek fact employee heart broken wife s infidelity retires o
ffice The delivery boy returning late enters prevents Matuschek shooting pistol Collapsing grief shame Ma
tuschekis rushed hospital The day Alfred visits Mr Matuschek sick bed asks Alfred s forgiveness puts work
manager store The delivery boy rewarded raise store clerk Klara arrives work late obviously heartbroken f
ailure correspondent materialize night When finds Alfred manager s office t believe discovers true faints
middle office Later resting home Alfred pays visit aunt brings letter secret pen pal explains meeting saw
Alfred Relieved misunderstanding swears Alfred ll work morning Alfred obviously working plan reveal Klara
Christmas Eve works day Mr Matuschek nearly recovered sickness stops things going final tally store best 
sales day 1928 Delighted hands bonuses takes new stock boy Christmas dinner Alfred Klara getting ready le
ave date mystery pen pal Alfred delays questions She s seen t know convince end engaged comes work He tel
ls mysterious pen pal stopped earlier fact fat bald older unemployed willing live Klara s income Alfred r
eveals puts red carnation lapel suddenly eveything clear
RESULT: s gift store Budapest workplace newly hi At work constantly irritate daily aggravation tempered f
act secret pen pal trade long soul searching letters Romantic correspondence sent forth trade barbs work 
dream someday meeting sensitive caring unknown pen pal Christmas fast approaching store busy Alfred store
time treated Mr lately attitude changed loss avoids explanation finally telling best left Stunned accepts
paycheck says goodbye including For civil A long awaited meeting secret pen pals
```

In [ ]:

```
# removed BERT-recognized named entities
movies['plot_text_processed'] = movies['plot_text_processed'].map(lambda x: remove_named_entities_bert
(x))

# Synopses after all preprocessing
movies['plot_text_processed'].head()
```

Out[ ]:

```
0    Note synopsis orginal Italian release segments...
1    Two thousand years ago Nhagruul Foul sorcerer ...
2    s gift store Budapest workplace newly hi At wo...
3    morning person s standards woken wife early br...
4    In May 1980 Cuban man named claims asylum Flor...
Name: plot_text_processed, dtype: object
```

In [ ]:

```
# more detailed preview on single synopsis
movies.plot_text_processed[0]
```

Out[ ]:

```
'Note synopsis orginal Italian release segments certain order introduces horror tales macabre supernatura
l known Three Faces Fear THE TELEPHONERosy attractive high priced Parisian girl returns spacious basement
apartment evening immediately gets beset series strange phone calls The caller soon identified ex pimp re
cently escaped prison Rosy terrified testimony landed man jail Looking solace phones lesbian lover The wo
men estranged time certain help agrees come night Seconds later calls promising matter calls protection r
evenge Unknown Rosy Mary caller impersonating arrives'
```

In [ ]:

```
# save preprocessed movies to csv
# maintain all other useful columns e.g. imdb_id, title
# use new csv for code below

movies.to_csv('bert_preprocessed_movies.csv', index=False)
```

BERT TOPIC MODELING

Adapted from Maarten Grootendorst's work on topic modeling with BERT:

https://towardsdatascience.com/topic-modeling-with-bert-779f7db187e6 (https://towardsdatascience.com/topic-modeling-with-bert-779f7db187e6)

In [7]:

```
# import csv with preprocessed movies
movies = pd.read_csv('./bert_preprocessed_movies.csv')

# Preview movies
movies.head()
```

Out[7]:

	imdb_id	title	plot_synopsis	tags	plot_text_processed
0	tt0057603	I tre volti della paura	Note: this synopsis is for the orginal Italian...	cult, horror, gothic, murder, atmospheric	Note synopsis orginal Italian release segments...
1	tt1733125	Dungeons & Dragons: The Book of Vile Darkness	Two thousand years ago, Nhagruul the Foul, a s...	violence	Two thousand years ago Nhagruul Foul sorcerer ...
2	tt0033045	The Shop Around the Corner	Matuschek's, a gift store in Budapest, is the ...	romantic	s gift store Budapest workplace newly hi At wo...
3	tt0113862	Mr. Holland's Opus	Glenn Holland, not a morning person by anyone'...	inspiring, romantic, stupid, feel-good	morning person s standards woken wife early br...
4	tt0086250	Scarface	In May 1980, a Cuban man named Tony Montana (A...	cruelty, murder, dramatic, cult, violence, atm...	In May 1980 Cuban man named claims asylum Flor...

In [8]:

```
# convert synopses data to list for clustering

movies_list = movies.plot_text_processed.tolist()

print(movies_list[4])
```

```
In May 1980 Cuban man named claims asylum Florida USA search American Dream departing Cuba Mariel boatlif
t 1980 When questioned tough talking INS officials notice tattoo s left arm black heart pitchfork identif
ies hitman detain camp called Freedomtown Cubans including s best friend Cuban Army buddy Ray local I 95
expressway government evaluates visa petitions After 30 days governmental dithering camp rumors receives
offer Cuban Mafia quickly relays If kill Roberto aide detained Freedomtown receive green cards agrees kil
ls riot
```

In [9]:

```
# EMBEDDING model

# without GPU
# model = SentenceTransformer('distilbert-base-nli-mean-tokens')

# with GPU
model = SentenceTransformer('distilbert-base-nli-mean-tokens', device='cuda')
```

In [10]:

```
# transform movie synopses into 768-dimensional vector embeddings:

embeddings = model.encode(movies_list, show_progress_bar=True)

# PREVIEW embeddings
for text, embedding in zip(movies_list, embeddings):
    print("Synopsis:", text)
    print("Embedding:", embedding[:10])
    print(len(embeddings[0]))
    print("")
    break
```

```
Synopsis: Note synopsis orginal Italian release segments certain order introduces horror tales macabre su
pernatural known Three Faces Fear THE TELEPHONERosy attractive high priced Parisian girl returns spacious
basement apartment evening immediately gets beset series strange phone calls The caller soon identified e
x pimp recently escaped prison Rosy terrified testimony landed man jail Looking solace phones lesbian lov
er The women estranged time certain help agrees come night Seconds later calls promising matter calls pro
tection revenge Unknown Rosy Mary caller impersonating arrives
Embedding: [-0.28233027 -0.5168891   0.7910191  -0.9591836  -0.3102247  -0.05852975
 -0.43264672 -0.6628994   0.4485602   0.3444074 ]
768
```

In [ ]:

```
# reduce dimensionality of embeddings for clustering
# a too low dimensionality results in a loss of information while a too high dimensionality results
# in poorer clustering results

# tune n_neighbors & n_components to get optimal results
umap_embeddings = umap.UMAP(n_neighbors=25,
                            n_components=5,
                            metric='cosine').fit_transform(embeddings)
```

In [72]:

```
# K-means clustering
from sklearn.cluster import KMeans

cluster = KMeans(n_clusters=20, random_state=0).fit(umap_embeddings)
```

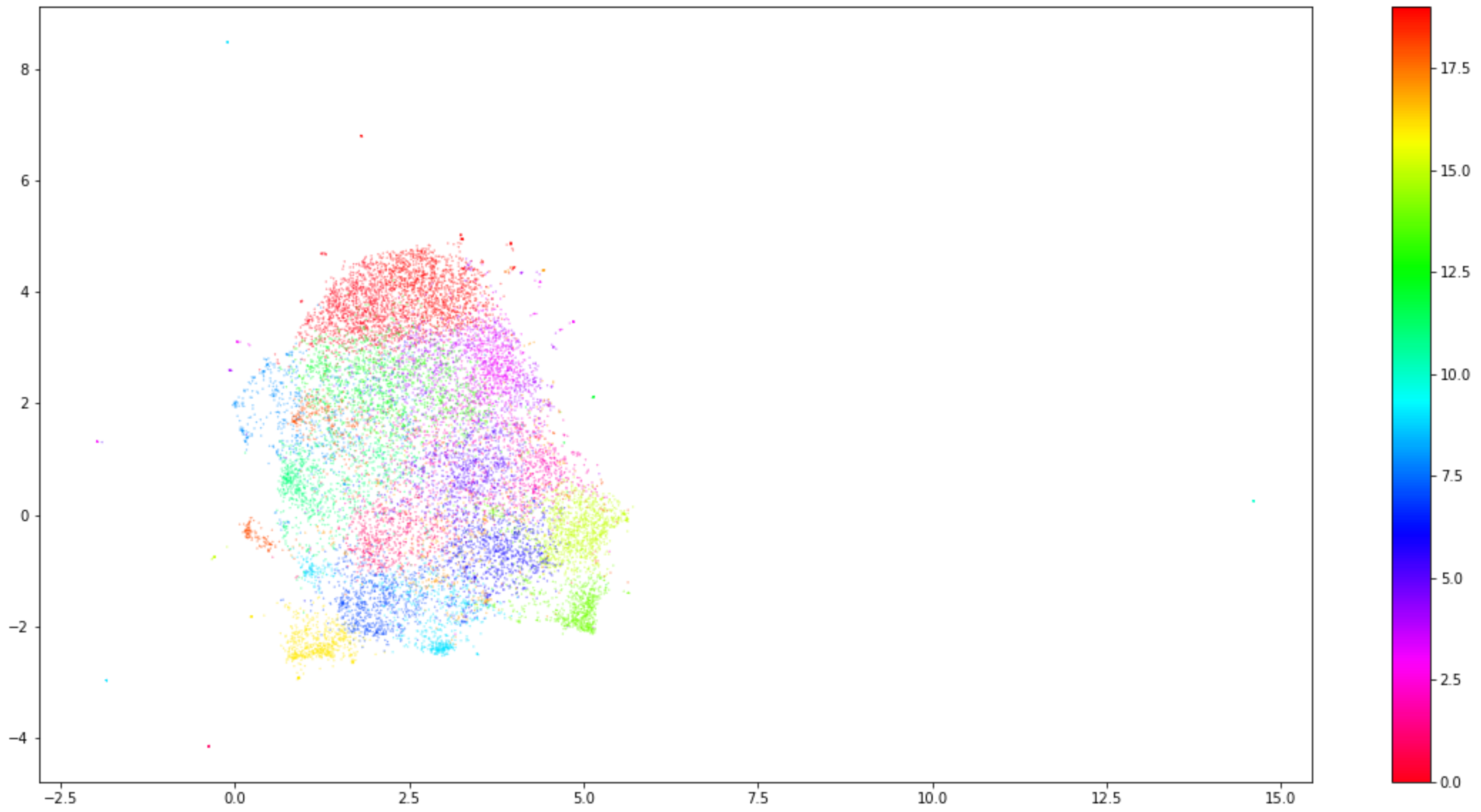
In [73]:

```
# VISUALIZE movie clusters

# make embeddings 2-dimensional to plot in 2d space
umap_data = umap.UMAP(n_neighbors=25, n_components=2, min_dist=0.0, metric='cosine').fit_transform(embeddings)
result = pd.DataFrame(umap_data, columns=['x', 'y'])
result['labels'] = cluster.labels_

# plot clusters
fig, ax = plt.subplots(figsize=(20, 10))
outliers = result.loc[result.labels == -1, :]
clustered = result.loc[result.labels != -1, :]
plt.scatter(outliers.x, outliers.y, color='#BDBDBD', s=0.05)
plt.scatter(clustered.x, clustered.y, c=clustered.labels, s=0.05, cmap='hsv_r')
plt.colorbar()
```

```
Out[73]: <matplotlib.colorbar.Colorbar at 0x7f0170600890>
```



CLUSTER TF_IDF

In [74]:
```
# group movies by cluster to get tf-idf across clusters

movies_df = pd.DataFrame(movies, columns=['plot_text_processed'])
movies_df['Topic'] = cluster.labels_
movies_df['Movie_ID'] = movies.imdb_id
movies_df['Title'] = movies.title
print(movies_df)

movies_per_topic = movies_df.groupby(['Topic'], as_index = False).agg({'plot_text_processed': ' '.join})
```

```
                                     plot_text_processed  Topic   Movie_ID  \
0      Note synopsis orginal Italian release segments...      5  tt0057603   
1      Two thousand years ago Nhagruul Foul sorcerer ...      7  tt1733125   
2      s gift store Budapest workplace newly hi At wo...     12  tt0033045   
3      morning person s standards woken wife early br...      8  tt0113862   
4      In May 1980 Cuban man named claims asylum Flor...     18  tt0086250   
...                                                  ...    ...        ...   
14823  In 1988 weatherman Harrisburg Pennsylvania tel...     12  tt0219952   
14824  In Russia media covers s disclosure identity I...      7  tt1371159   
14825  During North African Campaign World War II Cap...     14  tt0063443   
14826  catches unfaithful wife apartment boss apparen...      5  tt0039464   
14827  Sometime 1950s Chicago man returns home work f...     13  tt0235166   

                                               Title  
0                            I tre volti della paura  
1      Dungeons & Dragons: The Book of Vile Darkness  
2                         The Shop Around the Corner  
3                                 Mr. Holland's Opus  
4                                           Scarface  
...                                              ...  
14823                                  Lucky Numbers  
14824                                     Iron Man 2  
14825                                     Play Dirty  
14826                                      High Wall  
14827                               Against All Hope  

[14828 rows x 4 columns]
```

In [79]:

```
# save movies with cluster_id (Topic) to csv

movies_df.to_csv('movies_with_topics.csv', index=False)
```

In [75]:

```
# cluster-based TF-IDF:
# treat all documents in a cluster as a single document and then apply TF-IDF

def cluster_tf_idf(movies, m, ngram_range=(1, 1)):
    count = CountVectorizer(ngram_range=ngram_range, stop_words="english").fit(movies)
    t = count.transform(movies).toarray() # term frequency
    w = t.sum(axis=1) # total number of words
    tf = np.divide(t.T, w) # regularization of frequent words in the cluster
    sum_t = t.sum(axis=0)
    idf = np.log(np.divide(m, sum_t)).reshape(-1, 1) # the total, unjoined, number of documents m is divided by the total frequency of word t across all classes n
    tf_idf = np.multiply(tf, idf)

    return tf_idf, count

tf_idf, count = cluster_tf_idf(movies_per_topic.plot_text_processed.values, m=len(movies_list))
```

GET WORDS PER TOPIC

In [ ]:

```
# get top n words per topic
def top_n_words_per_topic(tf_idf, count, movies_per_topic, n=20):
    words = count.get_feature_names()
    labels = list(movies_per_topic.Topic)
    tf_idf_transposed = tf_idf.T
    indices = tf_idf_transposed.argsort()[:, -n:]
    top_n_words = {label: [(words[j], tf_idf_transposed[i][j]) for j in indices[i]][::-1] for i, label in enumerate(labels)}
    return top_n_words

# get number of movies per topic
def extract_topic_sizes(df):
    topic_sizes = (df.groupby(['Topic'])
                     .plot_text_processed
                     .count()
                     .reset_index()
                     .rename({"Topic": "Topic", "plot_text_processed": "Size"}, axis='columns')
                     .sort_values("Size", ascending=False))
    return topic_sizes

top_n_words = top_n_words_per_topic(tf_idf, count, movies_per_topic, n=20)
topic_sizes = extract_topic_sizes(movies_df)
```

In [77]:

```
# get topic count and sizes
print("TOPIC COUNT:", len(topic_sizes))
print()

topic_sizes.head(20)
```

```
TOPIC COUNT: 20
```

Out[77]:

	Topic	Size
19	19	1244
0	0	1199
15	15	1039
11	11	891
4	4	849
6	6	841
2	2	840
5	5	831
7	7	784
3	3	774
14	14	747
12	12	744
9	9	678
8	8	617
1	1	606
13	13	606
16	16	588
18	18	489

In [78]:

```
# inspect the most important words per topic (cluster)

# get top 20 topics
topics = topic_sizes.Topic[:20]

# get top 20 words per topic for top 20 topics
for t in topics:
    print(top_n_words[t][:20])
    print()
```

```
[('mother', 0.012133563077918286), ('love', 0.011353456008737519), ('daughter', 0.010543057270909807), ('
family', 0.00995396319251385), ('married', 0.00842139003602417), ('wife', 0.00840262642843264), ('sister
', 0.00834604103580555), ('marriage', 0.007975795230714695), ('old', 0.007739022260598821), ('husband',
0.00751531002615941), ('father', 0.00744473544375262), ('year', 0.007296286352102224), ('woman', 0.007137
343923220253), ('house', 0.007116836578277079), ('son', 0.006984767454992409), ('home', 0.006946783260260
494), ('parents', 0.006832858256912007), ('years', 0.0066210478279094), ('life', 0.006611041120071244),
('children', 0.0065468929672527956)]

[('love', 0.009432395631349563), ('party', 0.008576832781289888), ('friends', 0.008222770383220048), ('sc
hool', 0.008105682715363098), ('friend', 0.007654225014366395), ('new', 0.00731843005473818), ('girl', 0.
00711419104707805), ('york', 0.006722684024222509), ('sex', 0.006361224106424399), ('day', 0.006306411320
186986), ('best', 0.006300383998628229), ('relationship', 0.006240373472348164), ('meets', 0.006208565268
844355), ('night', 0.005978848399057214), ('woman', 0.005939417712602014), ('married', 0.0058278478042071
4), ('girls', 0.005731302299965957), ('wedding', 0.005705558813263167), ('tells', 0.00559863075576886),
('young', 0.005479927828660325)]

[('police', 0.013689848255796519), ('gang', 0.012777266361687602), ('money', 0.011199711576659289), ('dru
g', 0.011130729994369546), ('prison', 0.010973712682612777), ('bank', 0.009502670114211754), ('crime', 0.
008783876609639004), ('robbery', 0.007988350856693293), ('car', 0.007319783628136014), ('man', 0.00724659
6008330496), ('boss', 0.006678355899460461), ('men', 0.00660492602281305), ('gangster', 0.006558444944068
9195), ('criminal', 0.006349557900324284), ('partner', 0.0059820785420556435), ('gun', 0.0059635082235288
96), ('officer', 0.005639057298634401), ('city', 0.005556313406640583), ('agent', 0.00548598260812328),
('detective', 0.005236974414127327)]

[('cat', 0.008361652454803362), ('dog', 0.006436502413668185), ('house', 0.006366833307277293), ('rabbit
', 0.00604903893328862), ('mouse', 0.005623827088021853), ('goes', 0.005577332059287839), ('door', 0.0055
66484646175416), ('opens', 0.005472827652578092), ('tries', 0.005438521456214372), ('head', 0.00538201246
6451296), ('cartoon', 0.0053165349588949525), ('away', 0.005183316653172766), ('water', 0.004954176031230
864), ('man', 0.0049450837822652026), ('room', 0.004929653951539343), ('inside', 0.004785141416324228),
('like', 0.0047822256106173225), ('named', 0.0046664213441223075), ('film', 0.004642034474045884), ('begi
ns', 0.004544563308617495)]

[('father', 0.01151473536676355), ('family', 0.010447453379011089), ('son', 0.008761712282234761), ('scho
ol', 0.008093543763185098), ('old', 0.006478332640459421), ('boy', 0.006334619053598468), ('life', 0.0062
325294317667965), ('year', 0.006222549845390651), ('parents', 0.006178459271779234), ('home', 0.005939789
75550456), ('brother', 0.00589924582426146), ('wife', 0.005872610307235395), ('work', 0.00570456843232524
2), ('new', 0.00568918153966361), ('mother', 0.005615045338918921), ('young', 0.005401700102357534), ('li
ves', 0.0053189577085109565), ('day', 0.005218842749916545), ('job', 0.0051945037673516116), ('time', 0.0
05105523116737533)]
```

```
[('police', 0.011917433732661324), ('killer', 0.011324185282273285), ('murder', 0.00920018676184696), ('k
illed', 0.00732271765412303), ('prison', 0.007253148905855938), ('kill', 0.007161752611533143), ('killing
', 0.007079498441013022), ('death', 0.006964012567717302), ('man', 0.006945693362801547), ('agent', 0.006
80420960530134), ('crime', 0.006771733620528644), ('case', 0.006285323158298694), ('detective', 0.0060809
14588034993), ('murders', 0.005953222686946647), ('dead', 0.005908456745903682), ('kills', 0.005777381811
632993), ('people', 0.005437905952135127), ('body', 0.005327260739405704), ('later', 0.00525109006967312
05), ('serial', 0.005202030937201458)]

[('car', 0.011610593410339072), ('police', 0.00910527533295875), ('night', 0.006434183326387309), ('man',
0.005816382411670087), ('new', 0.005814275061450149), ('drug', 0.005794032296341005), ('money', 0.0055214
08552471208), ('home', 0.005498986189902294), ('friend', 0.0054780532277730505), ('store', 0.005385041351
434347), ('wife', 0.005251637768280532), ('detective', 0.005002452151291742), ('tells', 0.004956546909274
09), ('goes', 0.004911799337955069), ('film', 0.004876893616874309), ('job', 0.004838779671897598), ('get
s', 0.004745907770005199), ('murder', 0.004737304609239666), ('city', 0.004702926717193595), ('takes', 0.
004696987703893356)]

[('woman', 0.011773865321904753), ('girl', 0.008547619039294357), ('murder', 0.008515067106589096), ('pol
ice', 0.008140576156739642), ('mother', 0.00730562020889787), ('house', 0.007138313811444179), ('killed',
0.0068981973594026925), ('young', 0.006809639824497741), ('body', 0.006367746026475298), ('death', 0.0062
66541010109057), ('car', 0.006243558619195987), ('night', 0.00619027585341748), ('murdered', 0.0060782593
5850716), ('later', 0.006023659119872515), ('film', 0.005990018742248258), ('dead', 0.005988901579677627
5), ('women', 0.0059436777998874216), ('killer', 0.005792966522377623), ('wife', 0.005782659962276625),
('murders', 0.005711287765908986)]

[('zombies', 0.00760461474345105), ('zombie', 0.00715333499823143), ('world', 0.006937090655951926), ('dr
', 0.006739127938516152), ('city', 0.006388186963738419), ('group', 0.005919309477153331), ('virus', 0.00
5818051159532781), ('earth', 0.005481691139899968), ('agent', 0.00547152194894732), ('nuclear', 0.0052620
94605876737), ('plane', 0.005256386564201292), ('power', 0.005186521990067796), ('human', 0.0051576509326
245095), ('bomb', 0.005137830848006803), ('team', 0.005114922096815029), ('death', 0.004951486479120973),
('known', 0.004898692076807437), ('people', 0.004767228053114456), ('village', 0.0046849303618772915), ('
killed', 0.004678271887249315)]

[('mother', 0.012353430279391618), ('wife', 0.00914988992119594), ('daughter', 0.008154034893876201), ('h
ome', 0.007846220962862884), ('woman', 0.00781313495844484), ('children', 0.007712526063238076), ('husban
d', 0.007704583399030875), ('house', 0.007405714331508128), ('father', 0.007096566249993493), ('family',
0.007031448712617368), ('sister', 0.006986727486912901), ('young', 0.0069767334140928595), ('old', 0.0066
20462888325893), ('hospital', 0.006537882625098194), ('parents', 0.006432644704701024), ('death', 0.00607
7759849640156), ('child', 0.00600852270627643), ('life', 0.005979599740302128), ('years', 0.0059215783960
328665), ('later', 0.005907097487499541)]

[('town', 0.015367218970187868), ('sheriff', 0.012115102412075663), ('ranch', 0.011702735902492883), ('me
n', 0.010898734691930773), ('mexico', 0.009007921045018772), ('gang', 0.008953265107426235), ('mexican',
0.008589237127654383), ('man', 0.008553634826123235), ('texas', 0.008532808868318866), ('gold', 0.0085154
2533763545), ('saloon', 0.008397953713626852), ('horse', 0.007909433986352547), ('cattle', 0.007874319262
509751), ('train', 0.007174478906671642), ('army', 0.0069585723674784345), ('truck', 0.00695158463423052
1), ('west', 0.0063315379964308935), ('desert', 0.006329707712681671), ('local', 0.006328134698476029),
('wagon', 0.006262451111969562)]

[('band', 0.009930448224894606), ('school', 0.008049582916761862), ('new', 0.007463133539065067), ('music
', 0.006775776621309871), ('party', 0.006379288879317427), ('friends', 0.0063259380035406996), ('film',
0.006052148959020411), ('york', 0.005942887289452823), ('friend', 0.0057209553081081594), ('money', 0.005
573571746618397), ('day', 0.005496909732671065), ('night', 0.005401429056510542), ('gets', 0.005384955710
625182), ('father', 0.005178536682404603), ('man', 0.005150279827945659), ('club', 0.005129118722651276),
('time', 0.005107599405269211), ('high', 0.005037286071661318), ('singer', 0.005000783110138659), ('song
', 0.004925002437084606)]

[('german', 0.020118180285494145), ('war', 0.01566999485058748), ('british', 0.01064182706525789), ('capt
ain', 0.010210315743291745), ('germany', 0.010118505584331094), ('king', 0.009582706499436296), ('army',
0.009129529094928712), ('soldiers', 0.009012812474768313), ('japanese', 0.008382762593173014), ('ship',
0.00836445580727963), ('nazi', 0.008193819160466853), ('american', 0.008059517493298692), ('men', 0.00743
30628003295715), ('germans', 0.007248996177776566), ('world', 0.0070942635170323665), ('battle', 0.006969
205517080662), ('general', 0.006825845901660539), ('mission', 0.006716607859907387), ('colonel', 0.006553
589661782705), ('island', 0.006299983111485906)]

[('christmas', 0.010854843907281416), ('dog', 0.008074666325263243), ('school', 0.007911782758528947), ('
fairy', 0.0073470261380174755), ('children', 0.007225274226532552), ('home', 0.007076041180663321), ('nam
ed', 0.0067947056630505084), ('day', 0.006758365872902555), ('boy', 0.006461296137579402), ('water', 0.00
6319870406766768), ('house', 0.0061625847774589735), ('old', 0.006135487048536592), ('family', 0.00584585
8424830349), ('night', 0.005838069226283256), ('young', 0.005683641154922042), ('boat', 0.005576354962929
246), ('boys', 0.005410590315286135), ('friends', 0.005392499400605375), ('year', 0.005389037537398288),
('mother', 0.005321550006060882)]

[('vampire', 0.01710799619005152), ('vampires', 0.011937439393038917), ('dr', 0.009923877398527374), ('bl
ood', 0.008493080207283829), ('castle', 0.007152677000236392), ('human', 0.007152177841419281), ('witches
', 0.006357252690047727), ('body', 0.006052098420608662), ('house', 0.0056446784990490054), ('dead', 0.00
5567788287156153), ('woman', 0.005433168636497983), ('years', 0.005382037994362874), ('doctor', 0.0053748
01195608667), ('night', 0.00525946890517127), ('witch', 0.005227065751912488), ('demon', 0.00515954391054
3656), ('death', 0.005135043744263047), ('wolf', 0.005114463522260685), ('evil', 0.0051087365678208095),
('young', 0.004928032454751872)]

[('woman', 0.011084503170973486), ('girl', 0.009785726657040026), ('sex', 0.008219187711531016), ('girls
', 0.007676488417922436), ('tells', 0.007628388736118668), ('young', 0.007402814425743817), ('home', 0.00
```

```
7158759852436653), ('gets', 0.006634688148761827), ('house', 0.006563766957342223), ('women', 0.006403728
867039018), ('room', 0.006294157889057803), ('night', 0.006166331046272642), ('says', 0.00615621996400605
04), ('car', 0.006094341053855404), ('apartment', 0.005930181860345025), ('mother', 0.00589031570661754
8), ('goes', 0.005712826223689817), ('asks', 0.005656518350331793), ('bus', 0.005625426610687027), ('late
r', 0.005597172643341462)]

[('earth', 0.02965008170218715), ('planet', 0.023785727812364287), ('space', 0.022681133281033876), ('ali
en', 0.014856713359143906), ('ship', 0.01167587811050586), ('crew', 0.010603281080899113), ('humans', 0.0
10450640959781333), ('mission', 0.009481799662547063), ('world', 0.009250270468423387), ('human', 0.00878
6484748947507), ('mars', 0.008278196862619172), ('moon', 0.00807512735163687), ('robot', 0.00806419200418
9324), ('aliens', 0.007958820275532071), ('called', 0.00754575593339074), ('known', 0.0072161173209306),
('war', 0.006916010127268063), ('base', 0.006893523765632677), ('robots', 0.006828421593961588), ('years
', 0.006763953573232793)]

[('team', 0.012207423712295561), ('fight', 0.010800291533332689), ('race', 0.009486665387598927), ('game
', 0.008914456785443958), ('football', 0.008781546575469843), ('boxing', 0.00804389544952374), ('win', 0.
006980333722321846), ('martial', 0.006941265894883308), ('father', 0.006793667849340874), ('boy', 0.00678
0496877367114), ('man', 0.0066727974154046224), ('match', 0.006572649828153049), ('coach', 0.006427242474
308313), ('tournament', 0.006090526463852468), ('boxer', 0.006090526463852468), ('mr', 0.0057432345369845
98), ('boys', 0.0056532202434488145), ('school', 0.005604066722743995), ('car', 0.005444122080754431), ('
brother', 0.005441714114195664)]

[('french', 0.009572975740587514), ('father', 0.007758333887325886), ('king', 0.0069619525629088545), ('w
ar', 0.006880401842018503), ('pirate', 0.006366813224717813), ('story', 0.00634743797475261), ('brother',
0.005968802814171913), ('captain', 0.005819506179366585), ('ship', 0.005333451286584721), ('life', 0.0053
14657828387505), ('president', 0.005284550774356312), ('british', 0.005234019131363051), ('england', 0.00
5204949879693833), ('count', 0.005183222873631379), ('man', 0.005165558693164276), ('son', 0.005148554205
844435), ('years', 0.005075670280447128), ('film', 0.004989840299134072), ('american', 0.0049417534380117
21), ('book', 0.004830311153033089)]

[('que', 0.2631084587997156), ('en', 0.1674595727896377), ('una', 0.16701245787001143), ('la', 0.16281492
2339793), ('el', 0.14860061641351127), ('se', 0.11806576972607997), ('es', 0.08171685314703747), ('lo',
0.07265360667936893), ('coche', 0.054513951596356434), ('tem', 0.054513951596356434), ('está', 0.05182429
913792748), ('um', 0.04991596034886094), ('le', 0.04864513222211654), ('para', 0.04843573778624595), ('pe
lícula', 0.04231671778906723), ('ela', 0.04231671778906723), ('ele', 0.04231671778906723), ('niña', 0.042
31671778906723), ('él', 0.04231671778906723), ('algo', 0.04231671778906723)]
```

```
In [ ]:
```