Assignment 1: Data Wrangling and Vectorization

Name: Yunzhen WU

MSDS 453 Natural Language Processing - Section 56

Date: April 20, 2025

1. Introduction and Problem Statement

This assignment explores document-level vectorization techniques for textual data in the context of natural language processing. The primary goal is to refine a corpus-wide vocabulary that effectively captures content across a large set of movie reviews. The dataset consists of film reviews submitted by the class, with each student contributing a set of documents. My subset focuses on ten reviews of the movie *Oblivion*. The objective is twofold: to first qualitatively select terms based on subjective interpretation (Part 1), and then quantitatively evaluate different vectorization methods (Part 2) using the class-wide corpus.

Through data wrangling and various vectorization techniques such as TF-IDF, Word2Vec, Doc2Vec, and ELMo, this report aims to evaluate how well selected terms and embeddings capture document similarity and token relationships. The goal is to inform a refined vector design suitable for future unsupervised tasks such as clustering or classification.

2. Data and Preprocessing

An initial exploratory data analysis (EDA) was performed to understand the distribution of genres, review lengths, and vocabulary richness. The review texts varied significantly in length and tone, which suggested the need for normalization before any modeling. Word clouds and frequency plots helped identify frequently occurring terms, but also revealed the presence of noise such as stopwords and non-informative phrases.

Preprocessing involved a multi-step pipeline: all documents were lowercased, tokenized using NLTK, stripped of punctuation and special characters, lemmatized with WordNetLemmatizer, and finally stemmed using PorterStemmer. Stopwords were removed to

focus on content-bearing terms. These preprocessing steps ensured that the downstream models focused on semantic content rather than grammatical or stylistic variations.

3. Research Design and Modeling Methods

For the qualitative part, 10 reviews of the film *Oblivion* were analyzed. The following ten terms were initially selected: *jack harper, earth, space station, alien invasion, memory, science fiction, mission, identity, terraforming expedition,* and *tom cruise*. These terms were chosen because they either captured central characters *(jack harper, tom cruise*), core settings (*earth, space station*), major themes (*memory, identity, mission*), or domain-specific sci-fi elements (*terraforming expedition, alien invasion, science fiction*). These selections reflect both frequency and conceptual relevance.

The list was refined to a subset of four most representative terms: *jack harper, alien invasion, memory*, and *space station*. These terms were ultimately selected based on their significance in distinguishing *Oblivion* thematically (refer to Table 1 in the Appendix). For instance, jack harper appears prominently in reviews as the film's protagonist; alien invasion and space station are critical to the film's plot structure and setting; and memory is a recurring motif that ties into character identity and narrative twist.

For the quantitative part, 4 vectorization strategies were evaluated using the class corpus:

- TF-IDF vectorization using unigrams and bigrams, to capture sparse frequency-based signals.
- Word2Vec embeddings at three different dimensions (100, 200, and 300), to observe how embedding size affects token similarity and clustering behavior.

- Doc2Vec document embeddings, also evaluated at 100, 200, and 300 dimensions, to represent reviews as dense vectors suitable for clustering.
- ELMo contextual embeddings, using the TensorFlow Hub model, to generate deep semantic representations at the document level.

The tokenized output was used differently depending on the embedding method (joined strings for TF-IDF, token lists for Word2Vec/Doc2Vec, and raw strings for ELMo). Each vectorization method produced either document-level or token-level embeddings that were analyzed using cosine similarity, with results visualized through heatmaps and t-SNE plots.

4. Results

4.1 TF-IDF Results

After applying TF-IDF vectorization to the full class corpus, two tables were generated: one showing the top 10 terms with the highest average TF-IDF scores across all reviews (Table 2), and another showing the TF-IDF scores of the 10 terms selected in Part 1 (Table 3).

The global top terms—such as “film” (3.68), “movi” (2.81), and “barbi” (2.32)—are frequent across many documents and mostly reflect generic or stemmed forms common in movie reviews. While high-scoring, these terms may offer limited semantic depth for clustering purposes.

By contrast, most of the Part 1 terms showed lower TF-IDF scores. Notably, “earth” and “mission”had moderate values, while “jack harper” scored 0.17, and others like “alien invasion”, “memory”, and “science fiction” did not appear in the vocabulary at all. These absences suggest that while such terms are important for describing *Oblivion*, they are not prevalent enough across the broader corpus to score highly.

The histogram of mean TF-IDF scores shows that most terms have very low importance, with the vast majority concentrated near zero (Figure 1). This confirms the expected sparsity of TF-IDF features in large text corpora, where only a few terms meaningfully distinguish one document from another.

Finally, a cosine similarity heatmap based on TF-IDF vectors for all Sci-Fi movie documents was generated (Figure 2). Strong diagonal bands indicate consistent document identity, while visible clusters (e.g., among *Oblivion* reviews) suggest that TF-IDF embeddings capture enough semantic overlap to group similar documents together.

4.2 Word2Vec Results

From the cosine similarity heatmaps (Figure 3-5), we observed clear clustering patterns across all three dimensions. However, the 300-dimension heatmap appeared to show slightly more refined and coherent clusters compared to 100 and 200 dimensions. The similarity between related terms, such as film-related vocabulary or character names, was more consistent in the higher-dimensional setting.

The t-SNE visualizations (Figure 6-8) further support this finding. In the 100-dimension t-SNE plot, word placements were generally scattered, with fewer distinct groupings. The 200-dimension projection showed some structure, while the 300-dimension formed more visually distinct clusters, suggesting better separation and organization of semantically related words.

4.3 Doc2Vec and ELMo Document Similarity

From the Doc2Vec heatmaps (Figure 9-11), the similarity matrices generally appeared dense and uniform across all three embedding sizes. This suggests that many reviews share

common vocabulary structures. However, increasing the embedding size led to slightly clearer separation among some documents, especially in the 300-dimension version (Figure 11), where small clusters began to form more distinctly.

The Doc2Vec t-SNE plots (Figure 12-14) also showed clearer separations as the dimension increased. At 100 dimensions, most points were scattered with weak groupings. At 200, some genre-based clusters started to emerge. The 300-dimension t-SNE provided the most readable structure, showing tight clusters for certain groups of movies (e.g., action and sci-fi).

In comparison, the ELMo results differed significantly. The t-SNE (Figure 16) showed a sparse distribution of documents with little clear clustering. The ELMo cosine similarity heatmap (Figure 15) showed moderate-to-strong similarity values among many documents, but the structure lacked the genre-based grouping seen in Doc2Vec.

5. Analysis and Interpretations

The experiments demonstrate that no single vectorization method fully captures the nuanced semantics of all movie reviews. Each technique offers distinct strengths. TF-IDF, while sparse and frequency-driven, showed clear genre-specific clustering in Sci-Fi documents, indicating its usefulness in identifying similar reviews based on term overlap. However, many narratively important terms—like “memory” and “alien invasion”—did not score highly, reflecting TF-IDF’s limitation in capturing context or latent meaning.

Word2Vec embeddings improved upon this by capturing deeper token-level semantics. As the embedding size increased from 100 to 300, there are stronger clustering and more coherent term groupings in both heatmaps and t-SNE visualizations. This supports the

conclusion that higher-dimensional embeddings better preserve semantic structure and inter-term relationships, making them more suitable for tasks requiring token similarity.

Doc2Vec, in contrast, was more effective for document-level understanding. The 300-dimensional setting produced the most structured clusters (Figure 11 and 14), indicating that it captured genre and thematic similarity across reviews. The clear genre-based clustering seen in the 300-dimensional t-SNE plot suggests that Doc2Vec embeddings can be helpful for grouping documents in classification or retrieval settings.

ELMo embeddings offered deep contextual understanding but showed less pronounced structure in visualizations (Figures 15–16). While cosine similarities were high, they lacked strong genre separation. This could be due to ELMo's smaller sample size or its tendency to emphasize subtle contextual shifts over broader topic distinctions.

These findings highlight the trade-offs between frequency-based, distributional, and contextual vectorization strategies. For token-level tasks, Word2Vec (300D) performed best; for document-level clustering, Doc2Vec (300D) was most interpretable; and ELMo, though powerful, was less suited for coarse-grained grouping in this setup.

6. Conclusions

This assignment successfully applied multiple vectorization techniques to a corpus of movie reviews, combining qualitative judgment with quantitative evaluation. In Part 1, ten domain-relevant terms were manually selected from *Oblivion* reviews and later refined to four based on frequency and narrative relevance. However, the TF-IDF results showed that not all meaningful terms achieve high statistical significance across the broader corpus, reinforcing the importance of combining human insight with algorithmic tools.

In Part 2, TF-IDF revealed semantic groupings among documents but was limited in nuance. Word2Vec embeddings captured more refined token similarities, especially at higher dimensions. Doc2Vec embeddings were most effective for document-level clustering, particularly with 300-dimensional vectors. ELMo embeddings provided rich contextual information but lacked the visual separability seen in Doc2Vec outputs.

Overall, this study recommends using Doc2Vec (300D) for document clustering and Word2Vec (300D) for token-level similarity. Combining these methods, supported by strong preprocessing and exploratory analysis, can build a robust foundation for downstream NLP tasks such as genre classification, review recommendation, or semantic search.

Appendix

Table 1: Mean frequencies of 10 selected terms

	Oblivion	All Sci-Fi	All Non-Sci-Fi
earth	2.5	0.52	0.105556
science fiction	1.7	0.66	0.011111
tom cruise	1.5	0.48	0.016667
jack harper	0.9	0.18	0.000000
mission	0.5	0.12	0.072222
memory	0.4	0.12	0.016667
space station	0.3	0.06	0.000000
alien invasion	0.3	0.10	0.000000
identity	0.1	0.04	0.027778
terraforming expedition	0.1	0.02	0.000000

Table 2: Top 10 highest mean TF-IDF terms across the corpus

	Mean TF-IDF
film	3.68
movi	2.81
ha	2.34
barbi	2.32
wa	2.28
like	2.10
one	2.10
time	1.80
charact	1.80
make	1.63

Table 3: Mean TF-IDF scores for selected terms

	All Movies
jack harper	0.17
earth	0.61
space station	0.07
alien invasion	NaN
memory	NaN
science fiction	NaN
mission	0.35
identity	NaN
terraforming expedition	NaN
tom cruise	NaN

Figure 1: Histogram of mean TF-IDF scores

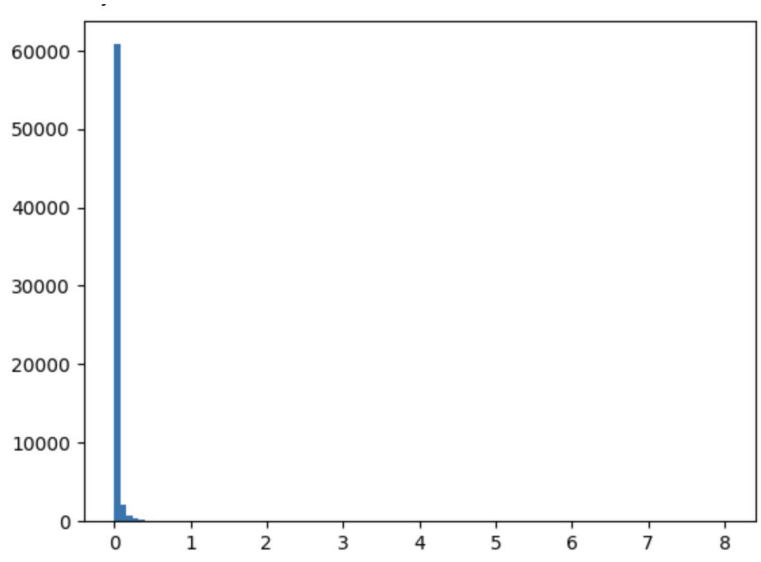


Figure 2: TF-IDF cosine similarity heatmap (Sci-Fi documents)

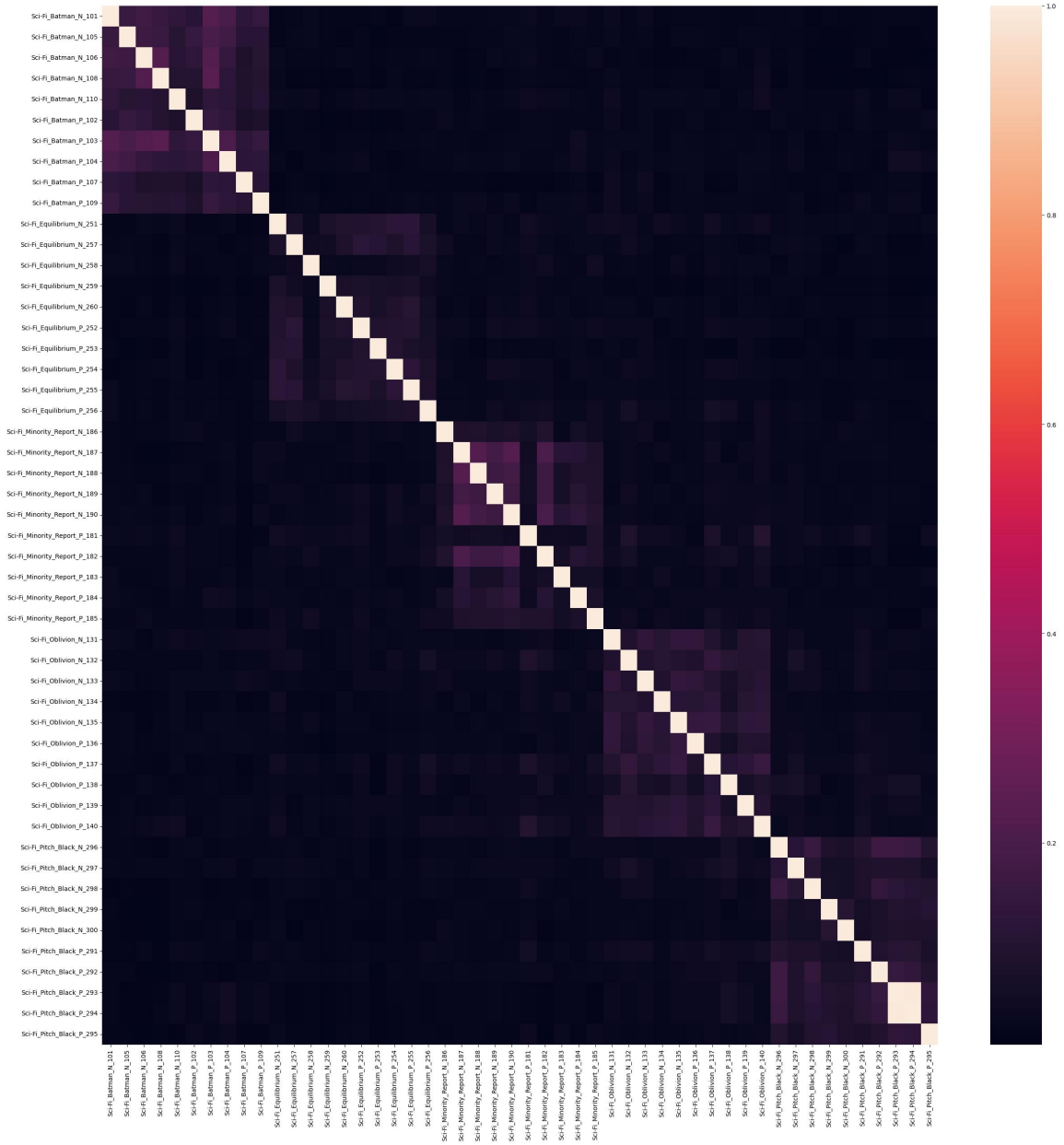
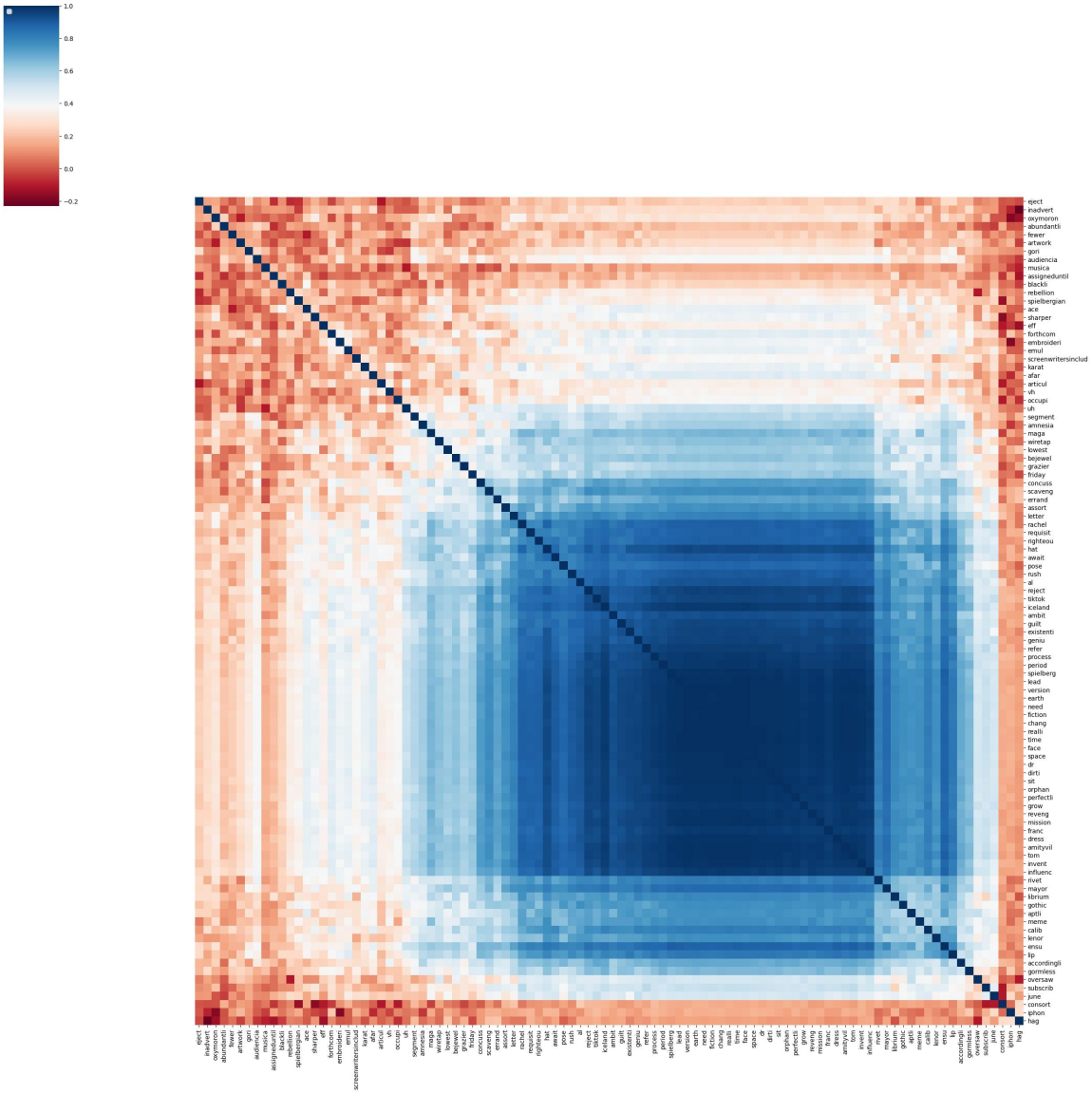
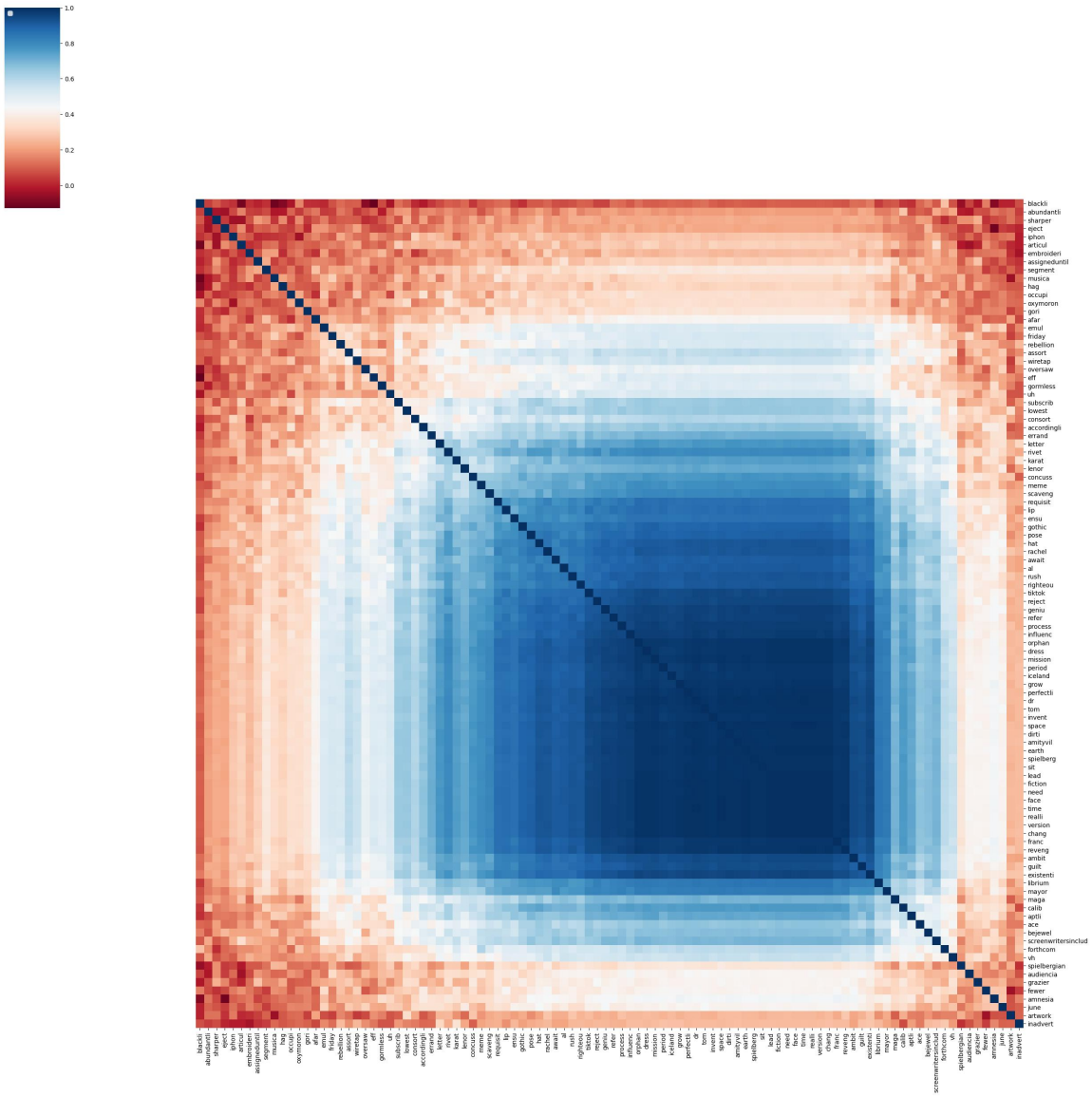


Figure 3–5: Word2Vec heatmaps (100/200/300 dims)




1.0
0.8
0.6
0.4
0.2
0.0
blackli
abundantli
sharper
eject
iphon
articul
embroideri
assigneduntil
segment
musica
haq
occupi
oxymoron
gori
afar
emul
friday
rebellion
assort
wiretap
oversaw
eff
gormless
uh
subscrib
lowest
consort
accordingli
errand
ketter
rivet
karat
kenor
concuss
meme
scaveng
requisit
kp
ensu
gothic
pose
hat
rachel
await
al
rush
righteou
tiktok
reject
geniu
refer
process
influenc
orphan
dress
mission
period
iceland
grow
perfectli
dr
tom
invent
space
dirti
amityvil
earth
spielberg
sit
lead
fiction
need
face
time
realli
version
chang
franc
reveng
ambit
guilt
existenti
librium
mayor
maga
calib
aptli
ace
bejewel
screenwritersinclud
forthcom
vh
spielbergian
audiencia
grazier
fewer
amnesia
june
artwork
inadvert

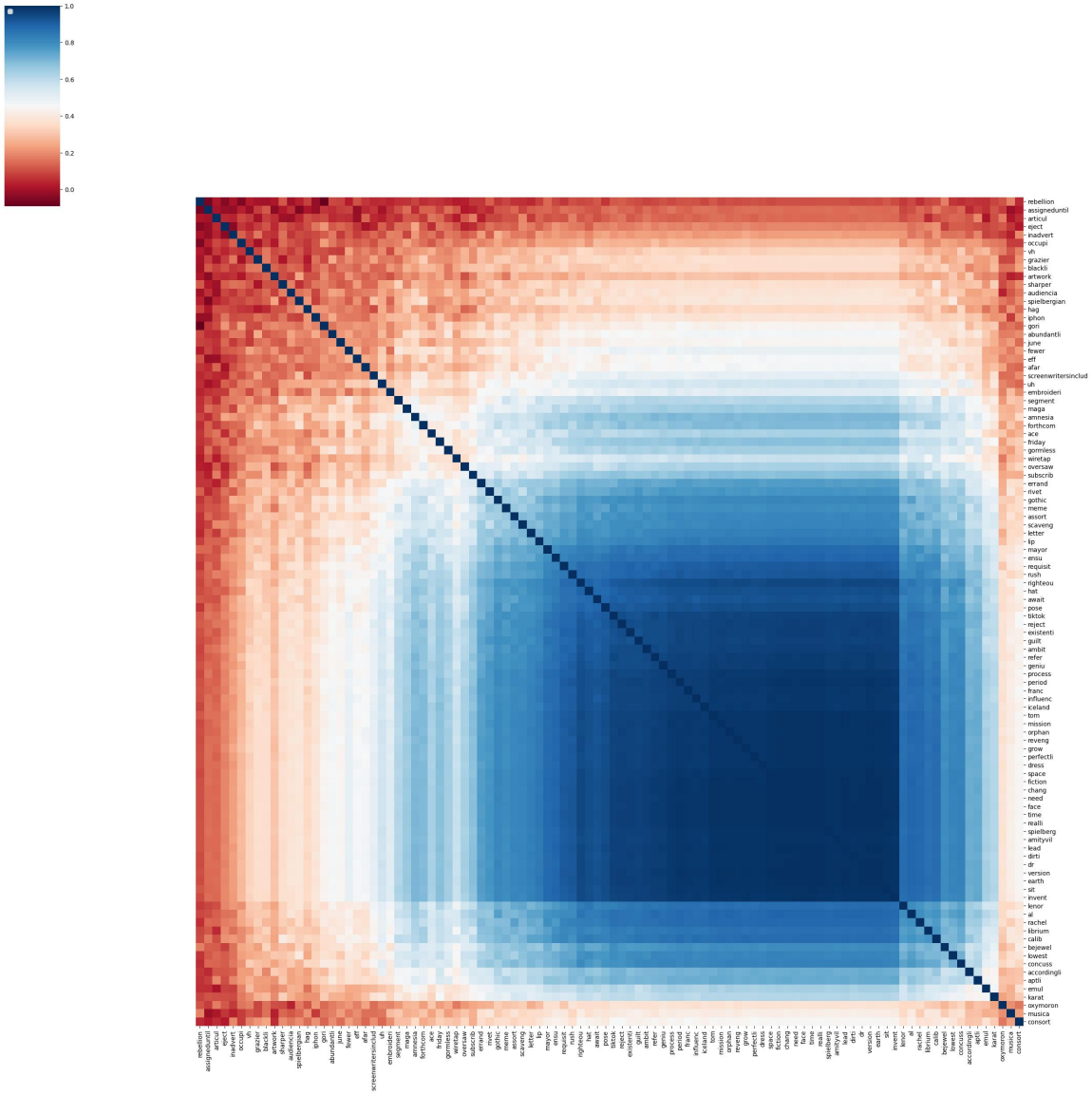
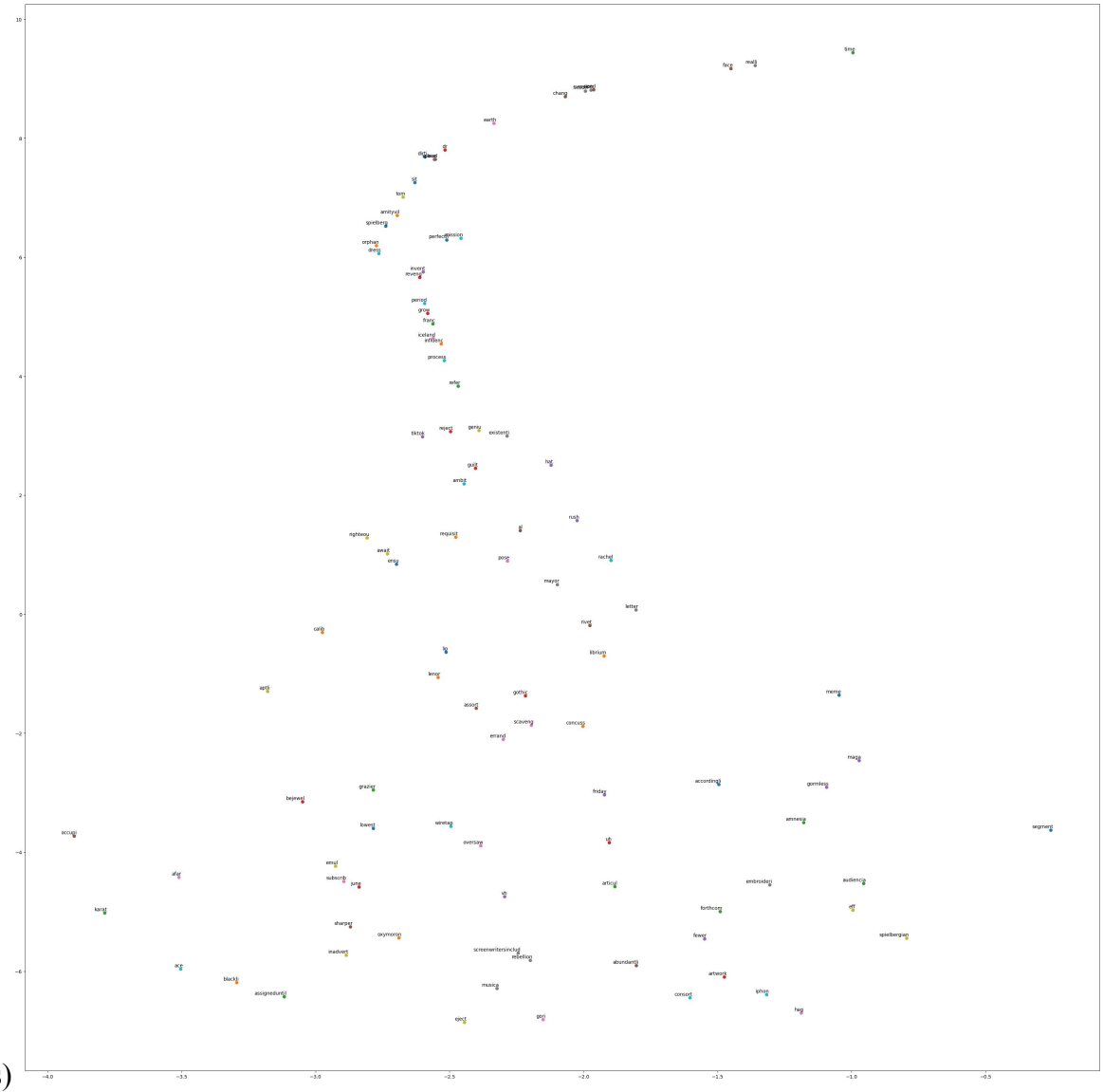
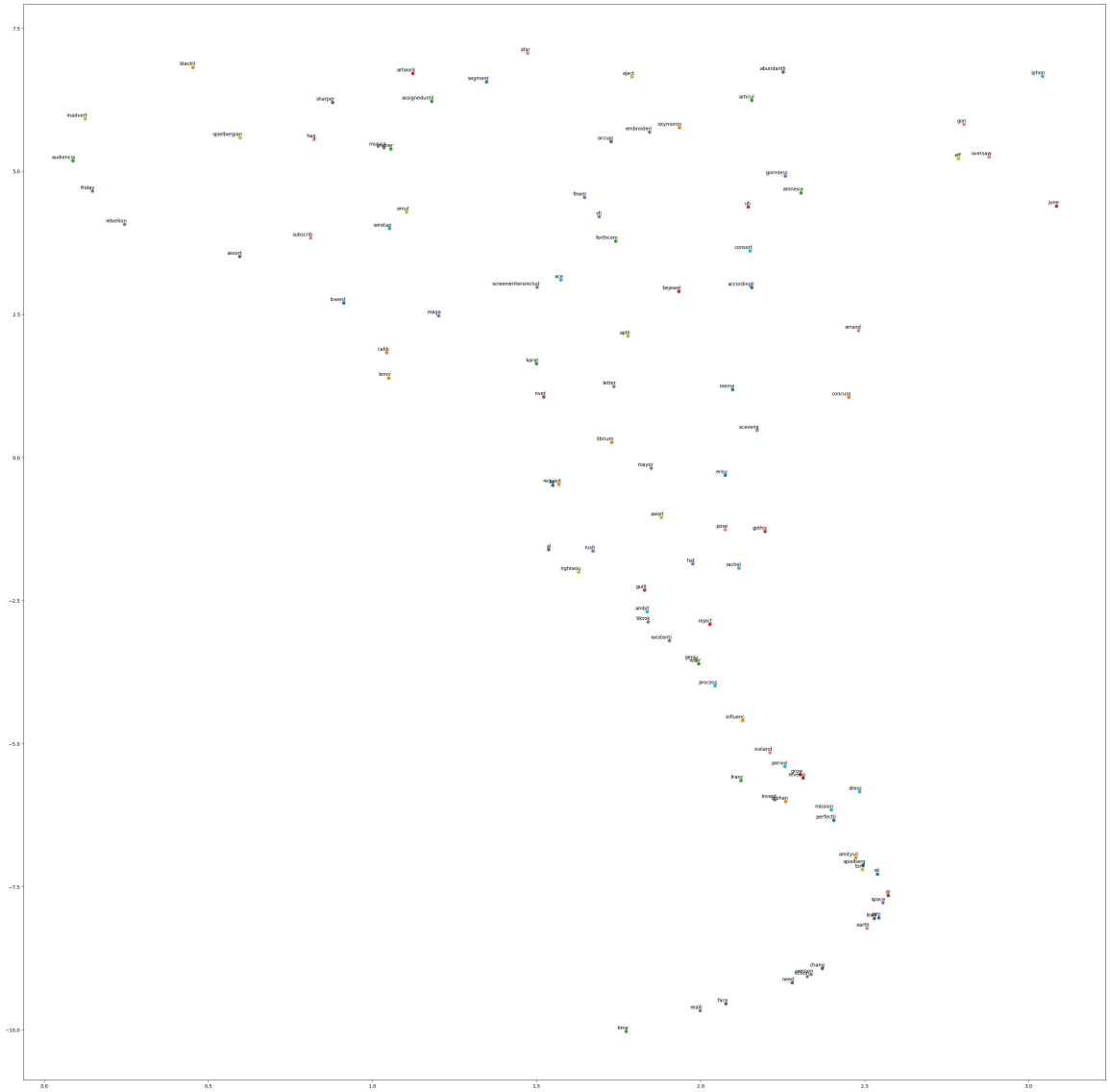

1.0
0.8
0.6
0.4
0.2
0.0
rebellion
assigneduntil
articul
eject
inadvert
occupi
vh
grazier
blackli
artwork
sharper
audiencia
spielbergian
hag
iphon
gon
abundantli
june
fewer
eff
afar
screenwritersinclud
uh
embroideri
segment
maga
amnesia
forthcom
ace
friday
gormless
wiretap
oversaw
subscrib
errand
rivet
gothic
meme
assort
scaveng
letter
lip
mayor
ensu
requisit
rush
righteou
hat
await
pose
tiktok
reject
existenti
guilt
ambit
refer
geniu
process
period
franc
influenc
iceland
tom
mission
orphan
reveng
grow
perfecti
dress
space
fiction
chang
need
face
time
realli
spielberg
amityvil
lead
dirti
dr
version
earth
sit
invent
lenor
al
rachel
librium
calib
bejewel
lowest
concuss
accordingli
aptli
emul
karat
oxymoron
musica
consort

Figure 6–8: Word2Vec t-SNE plots (100/200/300 dims)




7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
after
blacklist
artwork
segment
aspect
abundantli
iphon
sharpen
assignedunitl
articul
inadvert
embroidered
oxymoron
occupi
oversaw
audienc
gormless
amnesia
friday
fewer
june
rebellion
emul
subscrib
wristwatch
forthcom
assert
consult
screenwritersincludi
accordingli
lowest
beyond
mango
errand
apart
calib
lesser
lower
letter
meme
concuss
river
scaveng
fibrium
mayor
enrol
reclaim
await
pose
gothic
rachel
eighteou
pupil
ambul
reject
existentli
process
influenc
iceland
period
franc
invest
mission
perform
amityvill
spielberg
space
earth
chang
need
face
exalt
time

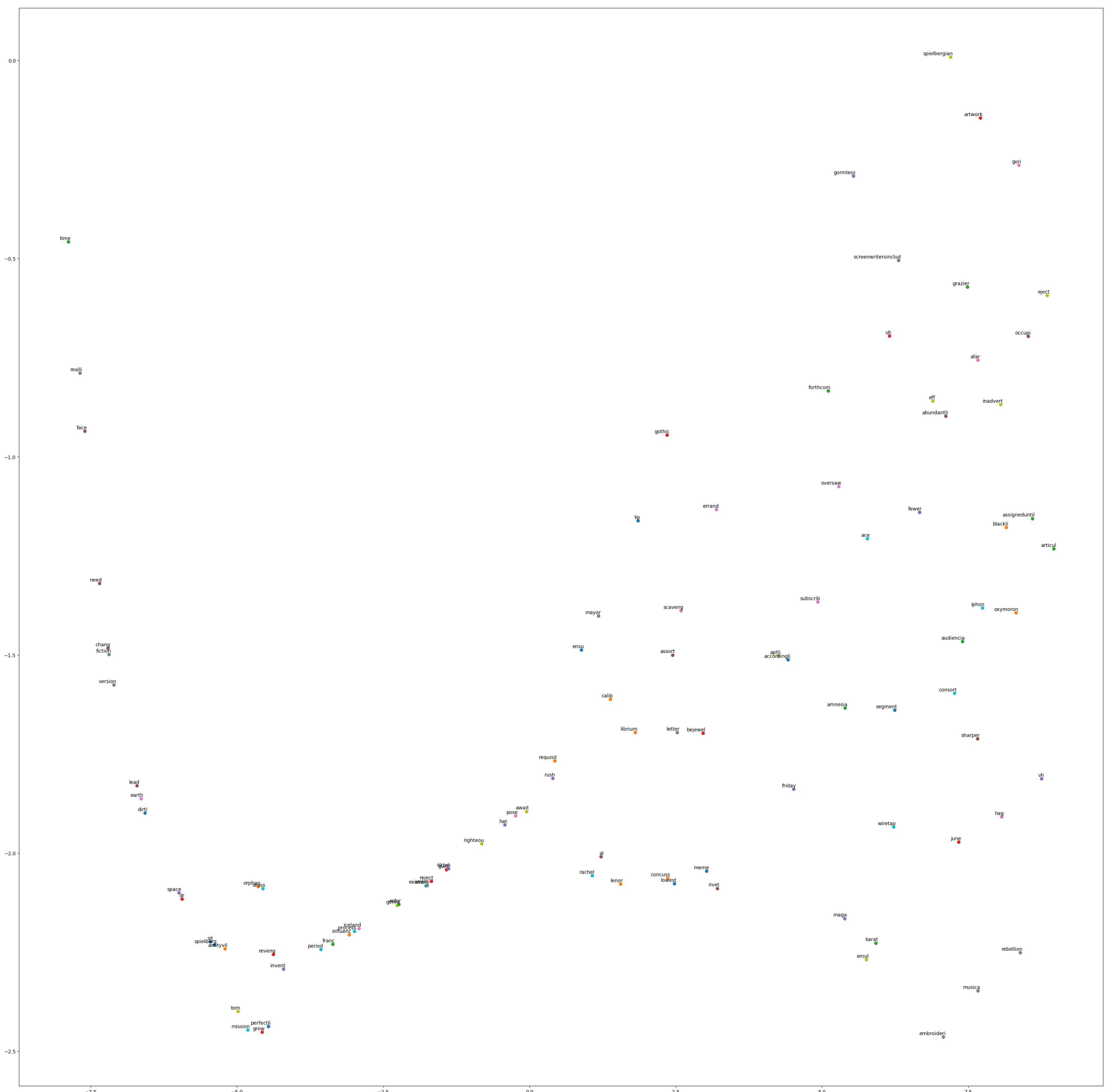
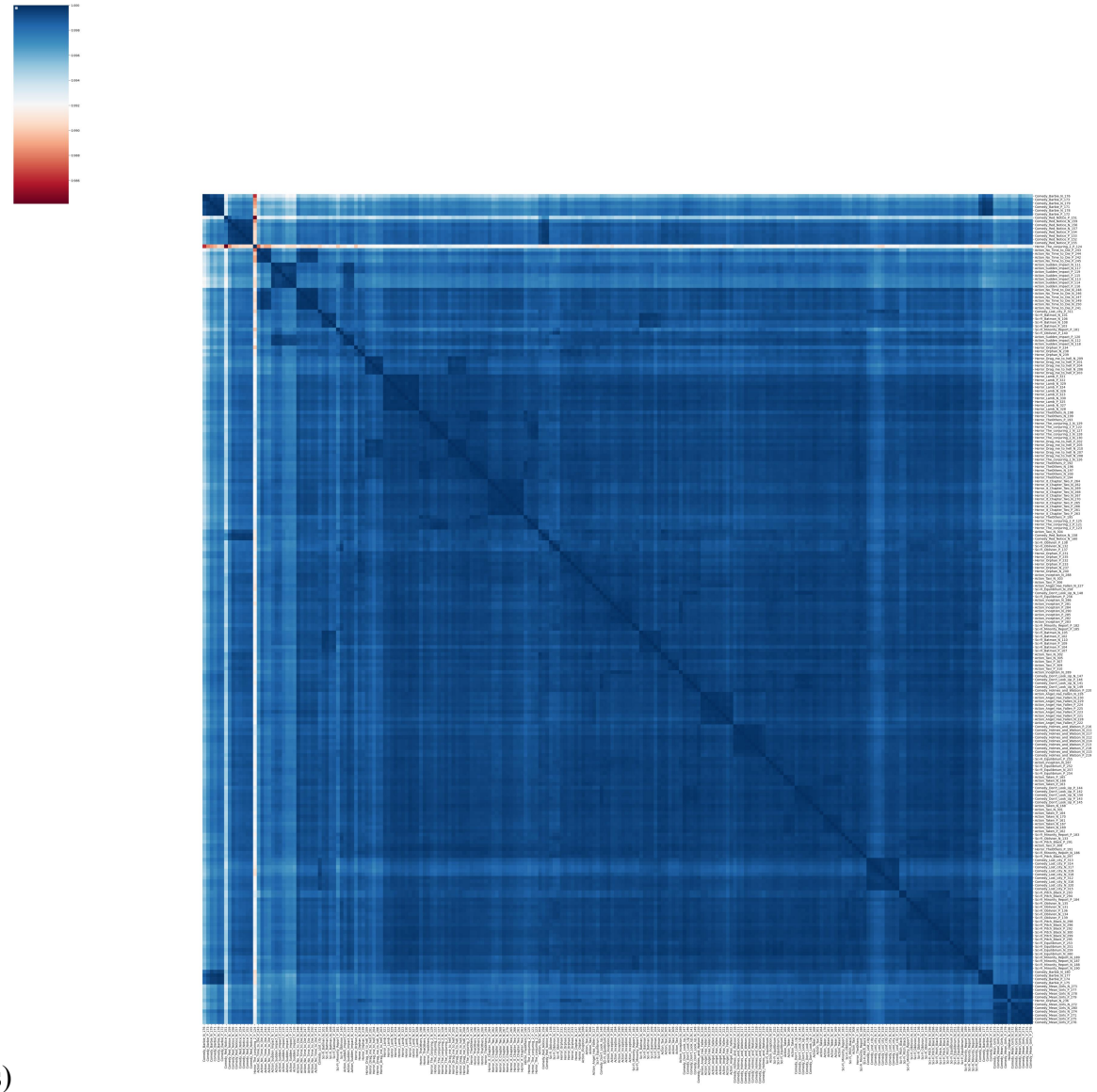
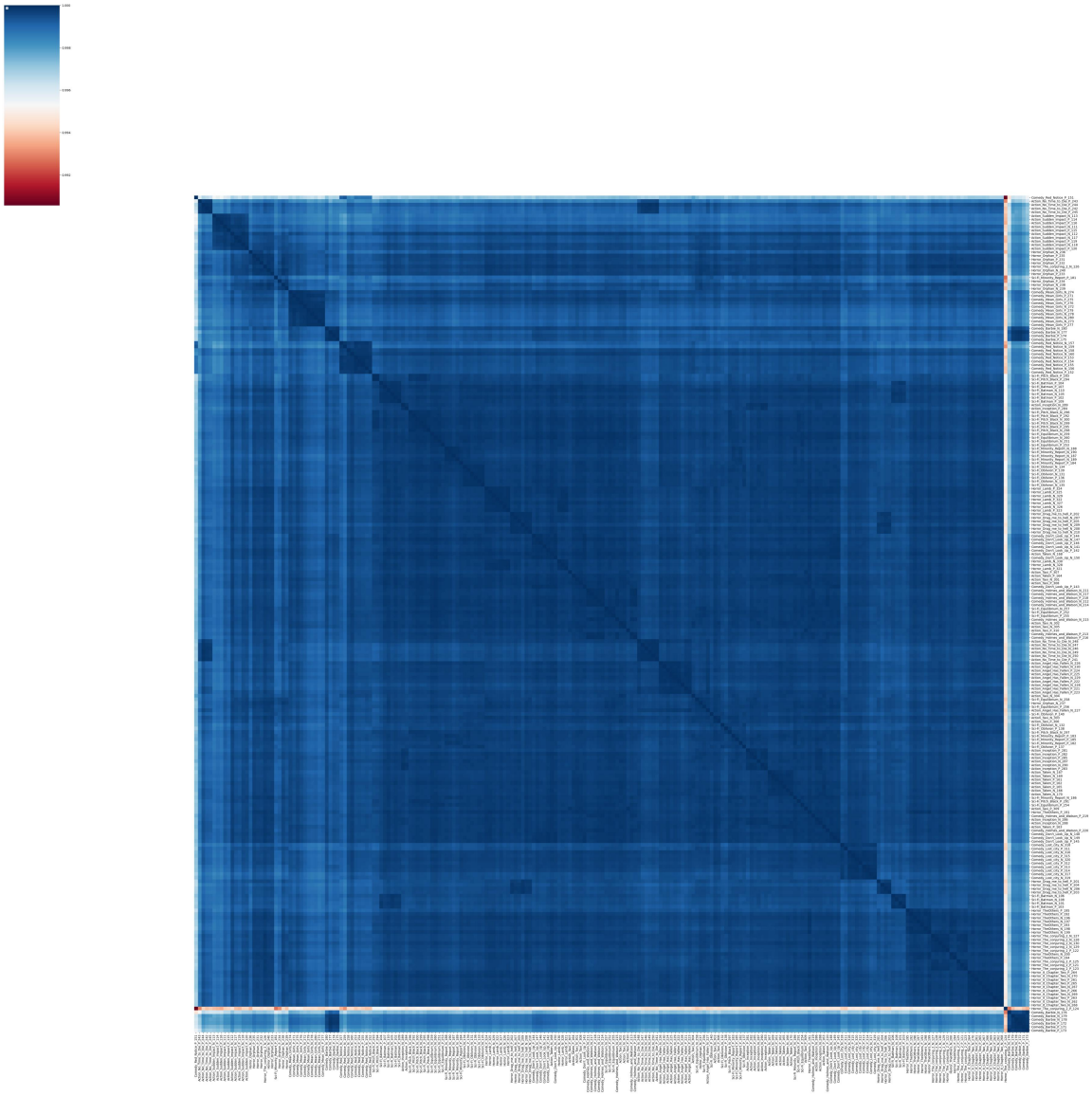

0.0
−0.5
−1.0
−1.5
−2.0
−2.5
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
spielbergian
artwork
gori
gormless
time
screenwritersinclud
grazier
eject
occupi
afar
realli
forthcom
abundantli
inadvert
gothic
face
oversaw
errand
fewer
assignedunti
blackli
ace
articul
need
subscrib
iphon
oxymoron
scaveng
mayor
audiencia
chang
fiction
assort
version
consort
amnesia
segment
librium
letter
bejewel
sharper
requisit
rush
lead
earth
dirti
friday
await
pose
righteou
wiretap
rachel
lenor
concuss
lowest
meme
rivet
space
orphan
maga
karat
spielberg
franc
influenc
iceland
period
reveng
invent
emul
rebellion
musica
tom
perfecti
mission
embroideri

Figure 9–11: Doc2Vec heatmaps (100/200/300 dims)





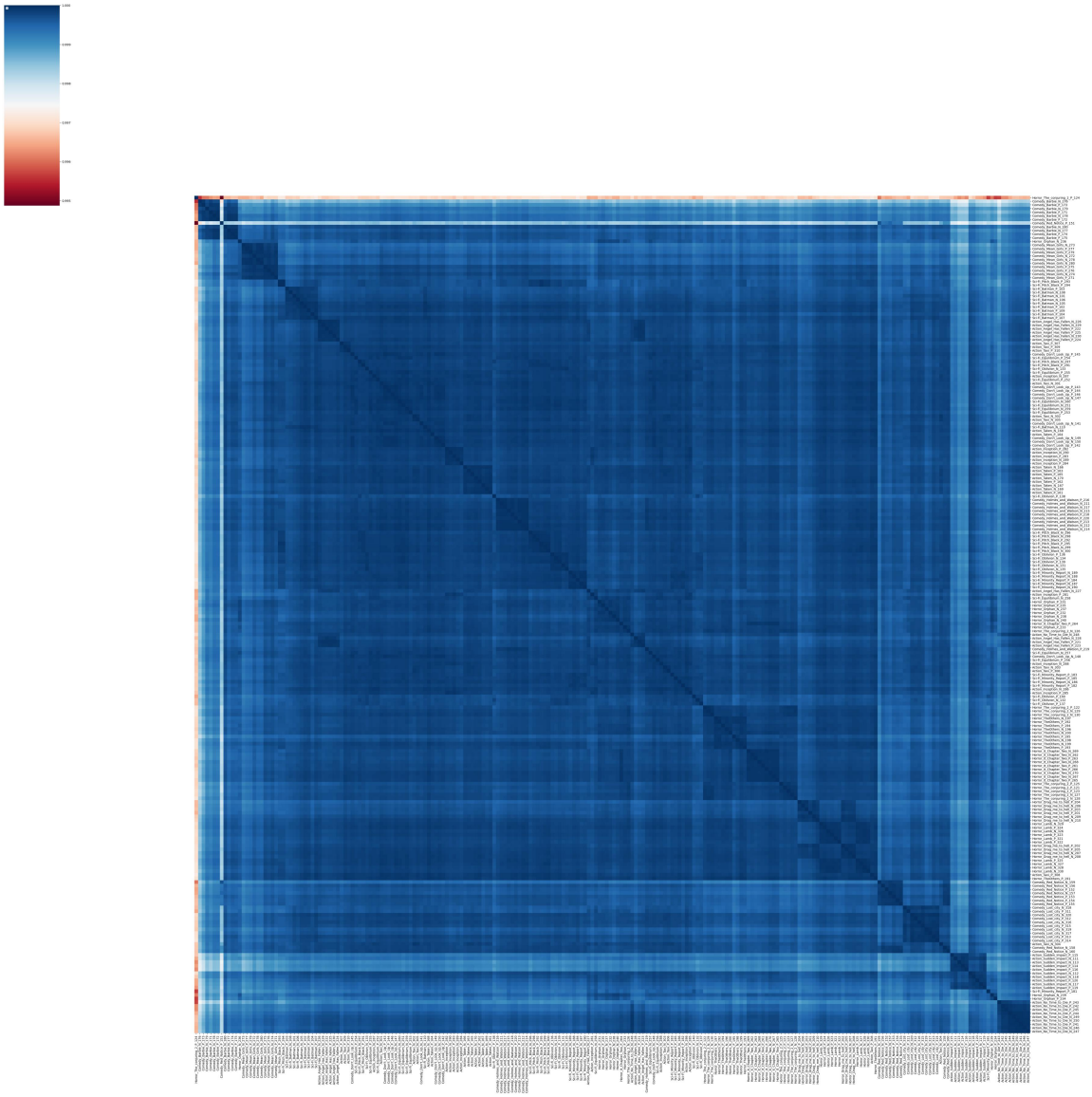
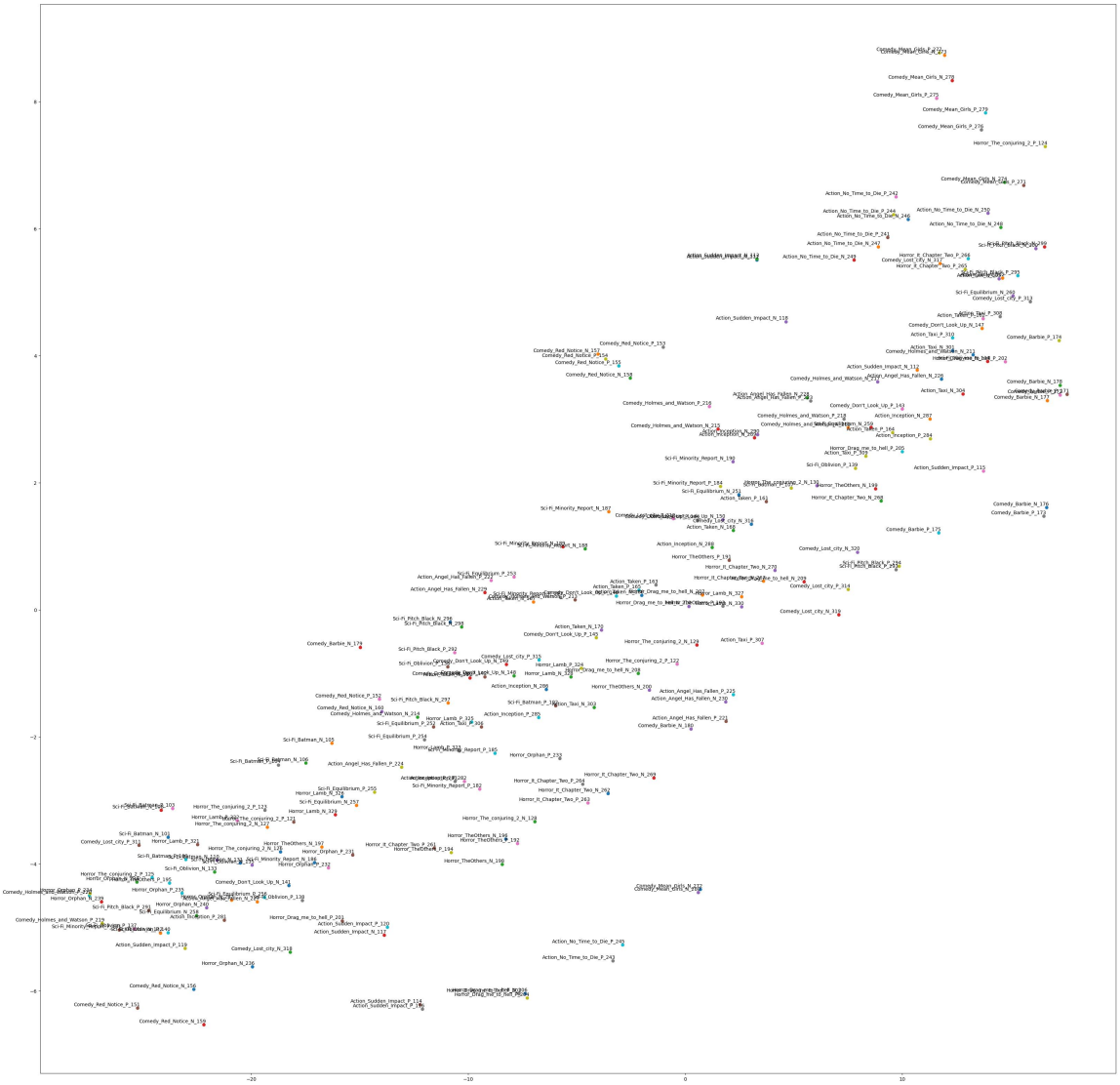
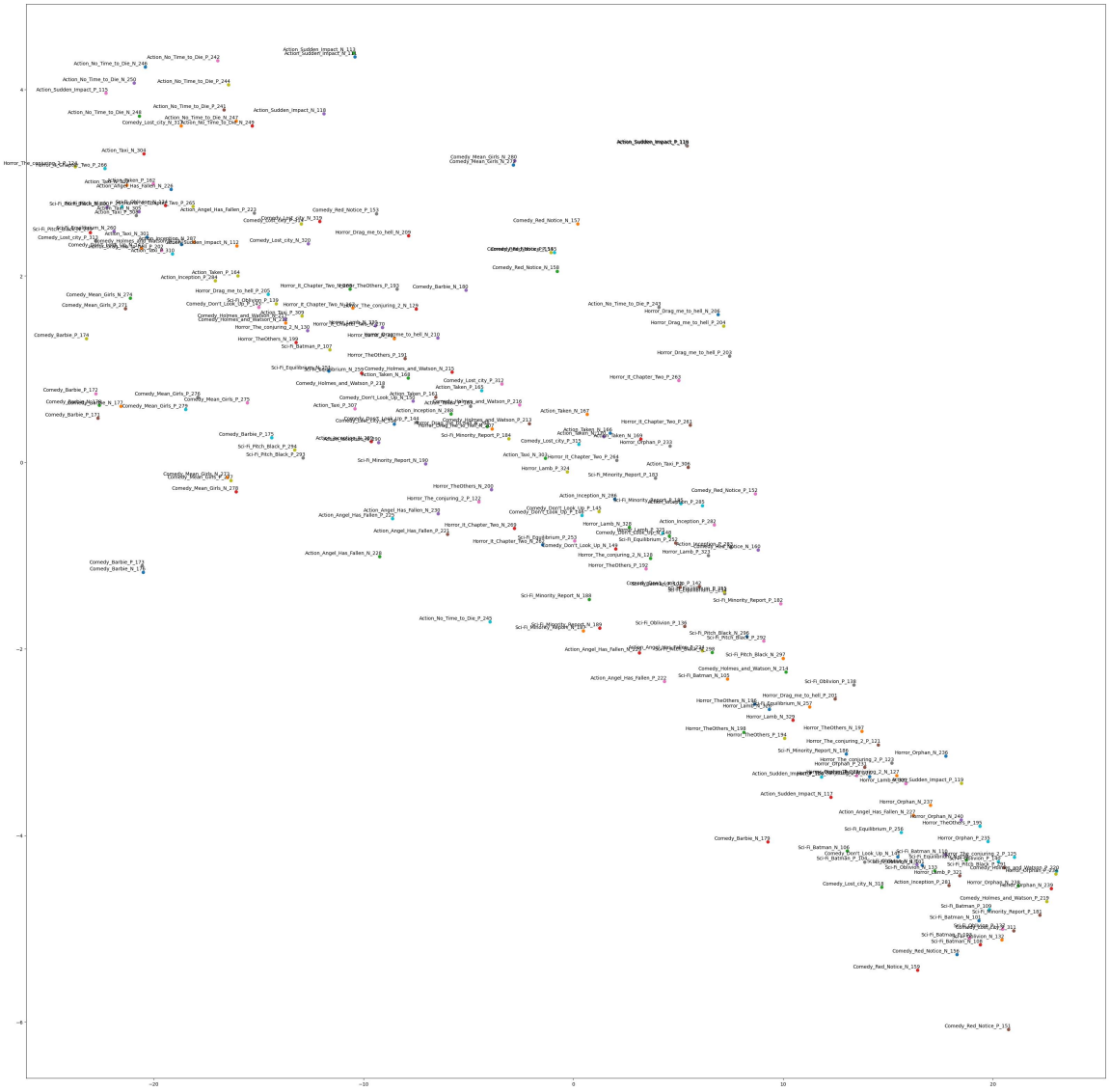


Figure 12–14: Doc2Vec t-SNE plots (100/200/300 dims)





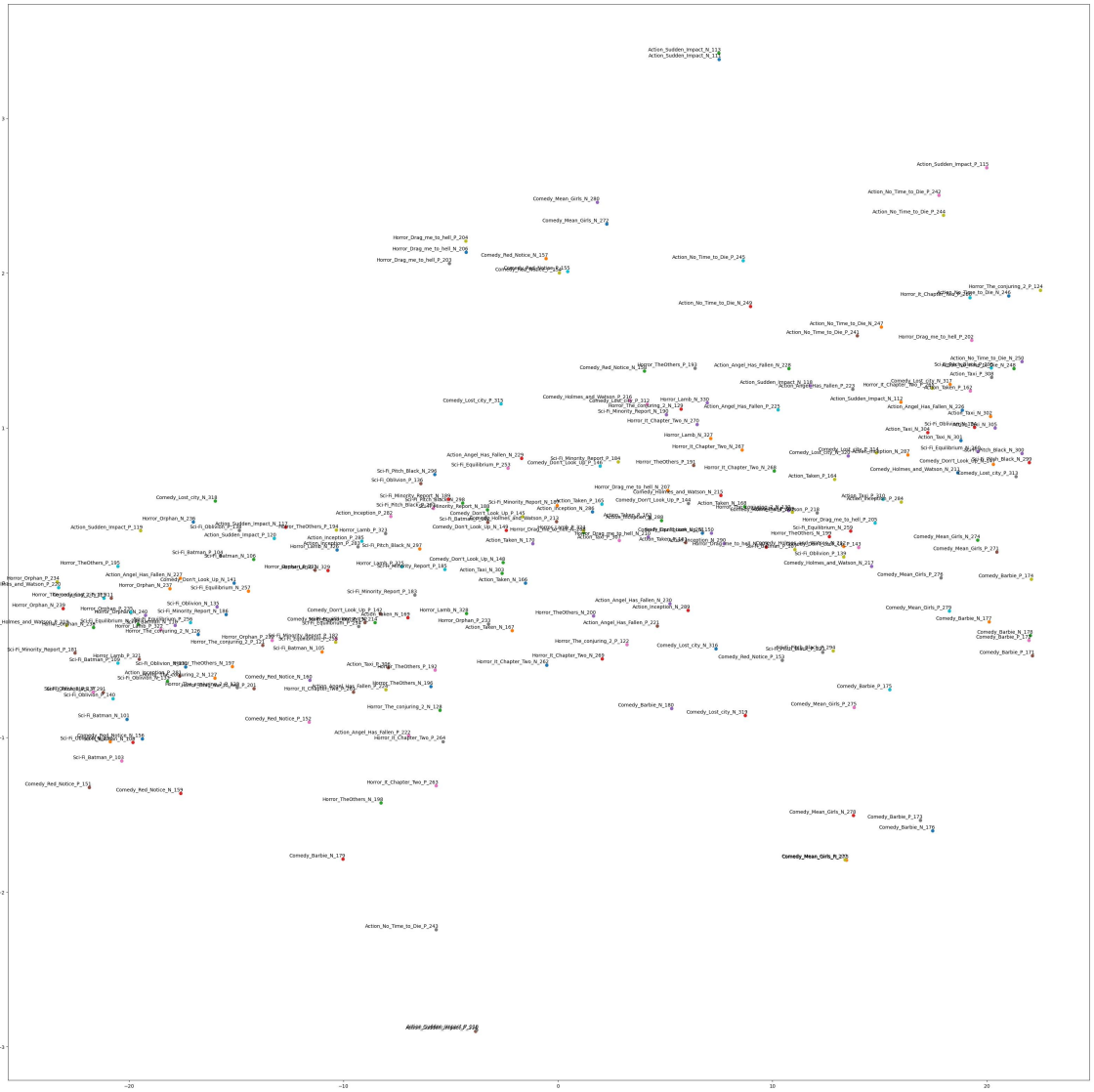

Action_Sudden_Impact_N_114
Action_Sudden_Impact_P_115
Action_No_Time_to_Die_P_242
Action_No_Time_to_Die_P_244
Comedy_Mean_Girls_N_280
Comedy_Mean_Girls_N_277
Horror_Drag_me_to_hell_P_204
Horror_Drag_me_to_hell_N_206
Horror_Drag_me_to_hell_P_203
Comedy_Red_Notice_N_157
Action_No_Time_to_Die_P_245
Horror_The_conjuring_2_P_124
Action_No_Time_to_Die_N_246
Action_No_Time_to_Die_N_247
Action_No_Time_to_Die_P_243
Horror_Drag_me_to_hell_P_202
Action_No_Time_to_Die_N_250
Action_Taxi_P_309
Comedy_Lost_city_N_317
Action_Sudden_Impact_N_116
Action_Angel_Has_Fallen_P_222
Horror_It_Chapter_Two_P_264
Action_Sudden_Impact_N_112
Action_Angel_Has_Fallen_N_226
Action_Taxi_N_302
Action_Taxi_N_306
Action_Tool_N_301
Comedy_Don't_Look_Up_P_143
Comedy_Holmes_and_Watson_N_271
Comedy_Lost_city_P_313
Horror_Drag_me_to_hell_N_207
Comedy_Holmes_and_Watson_N_272
Action_Taken_P_164
Comedy_Lost_city_N_318
Horror_Orphan_N_236
Action_Sudden_Impact_P_119
Horror_Orphan_N_237
Sci-Fi_Batman_N_104
Sci-Fi_Batman_P_105
Horror_TheOthers_P_195
Action_Angel_Has_Fallen_N_227
Comedy_Don't_Look_Up_N_141
Horror_Orphan_P_234
Comedy_Holmes_and_Watson_P_220
Horror_Orphan_N_237
Sci-Fi_Equilibrium_N_257
Sci-Fi_Minority_Report_P_182
Sci-Fi_Oblivion_N_135
Horror_Orphan_N_239
Sci-Fi_Minority_Report_N_186
Horror_Lamb_N_321
Horror_TheOthers_N_200
Action_Angel_Has_Fallen_N_230
Action_Inception_N_299
Comedy_Mean_Girls_P_279
Comedy_Barbie_N_177
Action_Angel_Has_Fallen_P_227
Horror_Orphan_P_233
Action_Taken_N_167
Horror_The_conjuring_2_P_122
Comedy_Lost_city_N_316
Comedy_Red_Notice_P_152
Horror_It_Chapter_Two_N_262
Horror_It_Chapter_Two_P_263
Comedy_Barbie_N_178
Comedy_Barbie_P_171
Comedy_Barbie_P_175
Comedy_Mean_Girls_P_275
Comedy_Barbie_N_180
Comedy_Lost_city_N_312
Horror_The_conjuring_2_N_128
Action_Angel_Has_Fallen_P_222
Horror_It_Chapter_Two_P_264
Sci-Fi_Batman_P_103
Comedy_Red_Notice_P_151
Comedy_Red_Notice_N_150
Horror_TheOthers_N_198
Horror_It_Chapter_Two_P_262
Comedy_Mean_Girls_N_278
Comedy_Barbie_P_173
Comedy_Barbie_N_176
Comedy_Barbie_N_172
Comedy_Mean_Girls_P_276
Action_No_Time_to_Die_P_241
Action_Sudden_Impact_P_116

Figure 15&16: ELMo cosine similarity heatmap and t-SNE visualization