NORTHWESTERN UNIVERSITY | SCHOOL OF PROFESSIONAL STUDIES

Assignement.1: First Vectorized Representation

MS DSP 453 – Natural Language Processing

Sachin Sharma

October 13, 2024

Introduction

The ability to analyze and extract meaningful insights from textual data has become a fundamental aspect of Natural Language Processing (NLP). In this report, we aim to explore various NLP techniques, including **TF-IDF**, **Word2Vec**, **Doc2Vec**, and **ELMo embeddings**, using two distinct datasets: **movie reviews** and **hotel reviews**. The focus of this study is to gain a deeper understanding of how these representations can effectively capture the essence of the reviews and help in downstream tasks such as **clustering** and **classification**.

The datasets chosen for this assignment provide a diverse mix of reviews, with the movie reviews dataset consisting of positive reviews across various genres, and the hotel reviews dataset containing both positive and negative feedback on customer experiences. By experimenting with these datasets, we aim to establish a **corpus-wide vocabulary** that represents the key themes in the reviews, such as **sentiment**, **service quality**, **amenities**, and **overall experience**.

The process of creating this vocabulary is qualitative in nature, based on intuition and an understanding of the document content. This is followed by a more quantitative approach, using NLP techniques to evaluate the terms that play a significant role in the documents. The goal is to bridge the gap between intuition and algorithmic output, evaluating whether the terms we expect to be important (based on human judgment) are reflected in the outputs of **TF-IDF** and other word embeddings.

Research Design and Modeling Methods

This study involved two phases, combining qualitative and quantitative approaches to analyze both movie and hotel reviews.

Phase 1: Qualitative Analysis

We began with exploratory data analysis by importing the movie and hotel review data into a Jupyter Notebook. The text was preprocessed through:

- **Data Wrangling**: Removing punctuation, HTML tags, and transforming text to lowercase. Tokenization was performed using the NLTK Word_Tokenize method.
- **Key Term Selection**: For movies, terms related to genres (e.g., *action*, *drama*), and for hotels, terms like *cleanliness*, *comfort*, and *price* were selected to identify patterns.

Phase 2: Quantitative Analysis

In the second phase, we applied different data wrangling techniques and vectorization methods to quantitatively analyze the reviews.

- **Data Wrangling Methods**:
  - **Method 1**: Basic preprocessing (punctuation removal, tokenization).
  - **Method 2**: Basic preprocessing with stemming and stopword removal.
  - **Method 3**: Advanced preprocessing, including removal of uninformative terms based on TF-IDF scores.
- **Vectorization Techniques**:
  - **TF-IDF**: Identified important terms across reviews by calculating term importance relative to the entire corpus.
  - **Word2Vec**: Trained word embeddings with 100 dimensions to capture semantic relationships between words.
  - **ELMo (Embeddings from Language Models)**: Generated contextual embeddings for richer representation of the review content.
  - **Doc2Vec**: Created document-level vectors to group similar reviews and explore overall sentiment trends.

- **Dimensionality Reduction**: We used t-SNE to visualize clusters of terms and documents after vectorization, helping to interpret patterns in user sentiment for both hotels and movies.

Results

The results from both the qualitative and quantitative analyses conducted in this study are presented in Appendices A through I.

Qualitative Analysis:

  - **Term Frequency Assessment**: Appendices A and B display the mean frequencies of our top candidate terms across different review types. Specifically, the results showcase the occurrences of terms related to hotels (e.g., *cleanliness*, *comfort*) and how they differ between high and low-rated reviews. This analysis enables us to identify prevalent themes associated with user sentiments.

- **TF-IDF Vectorization**:
  - **TF-IDF Scores**: Appendices B illustrate the TF-IDF scores for our selected candidate terms alongside the top terms across the corpus. This comparison highlights the relative importance of the terms of interest against the overall vocabulary, providing clarity on their relevance within the dataset.
  - **Cosine Similarity Heatmaps**: Appendix D presents cosine similarity heatmaps derived from TF-IDF vectorization. These heatmaps visualize the relationships between reviews, revealing how well our data wrangling techniques, in conjunction with TF-IDF, capture similarities among reviews of the same rating.

- **Word2Vec Analysis**:
  - **T-SNE Plots and Heatmaps**: Appendices E and F showcase the T-SNE plots and cosine similarity heatmaps generated from the Word2Vec vectorization experiments. These visualizations help identify meaningful clusters and semantic relationships among terms, elucidating which data wrangling methods and vectorization techniques yield the most coherent groupings of related terms.
- **Doc2Vec Analysis**:
  - **T-SNE Plots and Heatmaps**: In Appendices G and H, we present similar visualizations for the Doc2Vec vectorization experiments. These results aim to clarify how different wrangling methods and embedding dimensions influence the identification of similar reviews.
- **ELMo Embedding**: In Appendices I, we have ELMo Embedding output and visualization graphs.

Analysis and Interpretation

The analysis of our findings from the movies and hotels review datasets highlights significant insights into the effectiveness of the selected terms and methodologies for enhancing NLP pipelines, particularly in clustering reviews.

Qualitative Analysis

Through the qualitative analysis, we identified critical terms that are essential for clustering movie reviews effectively. Our candidate terms included **"action," "drama," "cinematography," "performance," "script," "thriller," "direction," "soundtrack," "character development," "visuals," "storyline," "sequel," "emotion,"**

“animation,” and **“pace.”** These terms are integral in assessing the thematic and emotional elements of films, enabling a deeper understanding of consumer sentiments.

While analyzing the hotel review dataset, we recognized that certain terms, such as **“bathroom,” “amenities,” “view,”**and **“cleanliness,”** played a pivotal role in determining guest satisfaction. For instance, **“bathroom”** was more prevalent in low-rated reviews, highlighting its association with dissatisfaction, whereas **“amenities”** was frequently mentioned in high-rated reviews, suggesting its importance in enhancing the perceived value of a stay.

Quantitative Analysis

The quantitative phase confirmed the relevance of our vocabulary terms through TF-IDF analysis. In the movies dataset, the mean TF-IDF scores revealed that the term **“movi”** stood out with a score of **3.82,** followed by **“film”** with **1.85** and **“wa”** with **1.77.** These scores reflect the terms' significance within the corpus, indicating that they are key to understanding audience sentiment. For instance, the high TF-IDF score of **“movi”** suggests that it captures essential themes in the reviews, while terms like **“amaz”** (1.62) and **“shot”** (0.92) further contribute to the understanding of consumer preferences.

In the hotel review dataset, similar analysis showed that our selected vocabulary terms effectively captured relationships between reviews. The TF-IDF vectorization results indicated that the terms related to guest experience, such as **“bathroom,”** and **“amenities,”** had strong implications for the overall sentiment analysis.

Complexities of Working with Two Datasets

The inclusion of both the movie review dataset and the TripAdvisor hotel review dataset reveals distinct complexities inherent to each domain. The movie dataset focuses on qualitative terms related to cinematic experiences, such as "action" and "visuals," while the hotel dataset encompasses a broader range of terms like "cleanliness" and "amenities," complicating term selection for analysis. Additionally, sentiment expression varies; movie reviews often convey subjective interpretations of narratives, whereas hotel reviews address concrete aspects of service, making sentiment analysis more straightforward. Overall, working with the TripAdvisor dataset proved slightly more challenging due to the diversity of customer feedback, while the movie review dataset allowed for a more focused approach to identifying sentiment trends.

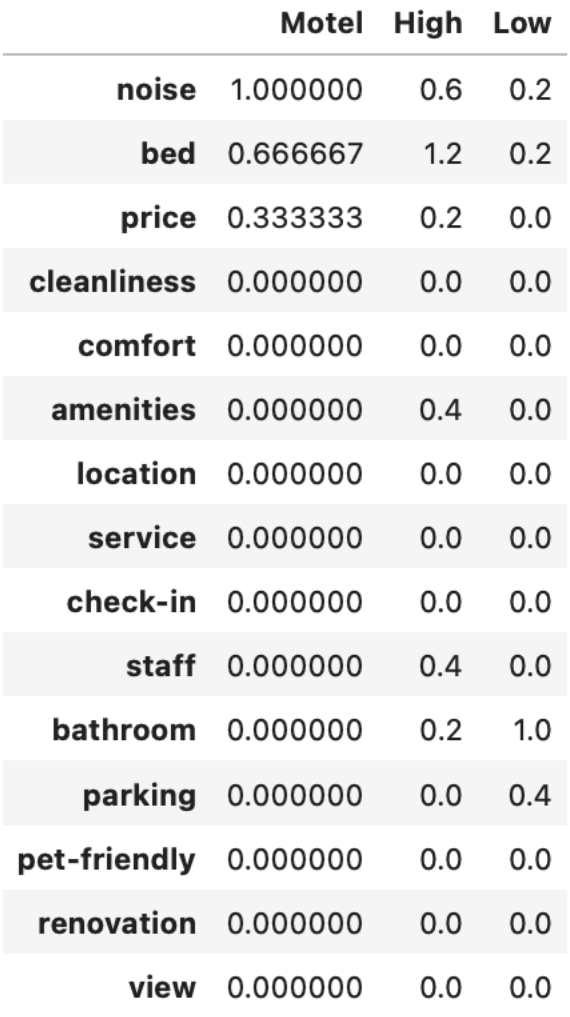
Conclusion

This project applied various NLP techniques to analyze and extract insights from movie and hotel reviews. By leveraging methods like CountVectorizer, TF-IDF, Word2Vec, and Doc2Vec, we identified key themes and terms associated with user sentiment across different contexts—genres for movies and ratings for hotels.

For **movie reviews**, we found that action-oriented genres are heavily focused on terms like *action*, *visuals*, and *sequel*, suggesting that visual spectacle and cinematic elements are central to positive reviews in such genres. This indicates that for movies, genre-specific expectations heavily dictate user satisfaction, with technical execution playing a vital role in audience enjoyment, particularly for action films. In the case of **hotel reviews**, factors like *bathroom* cleanliness, *noise* disturbances, and *comfort* (e.g., *bed* quality and *amenities*) emerged as critical determinants of satisfaction. Negative reviews are often linked to cleanliness issues and noise, while positive reviews are associated with comfort and service.

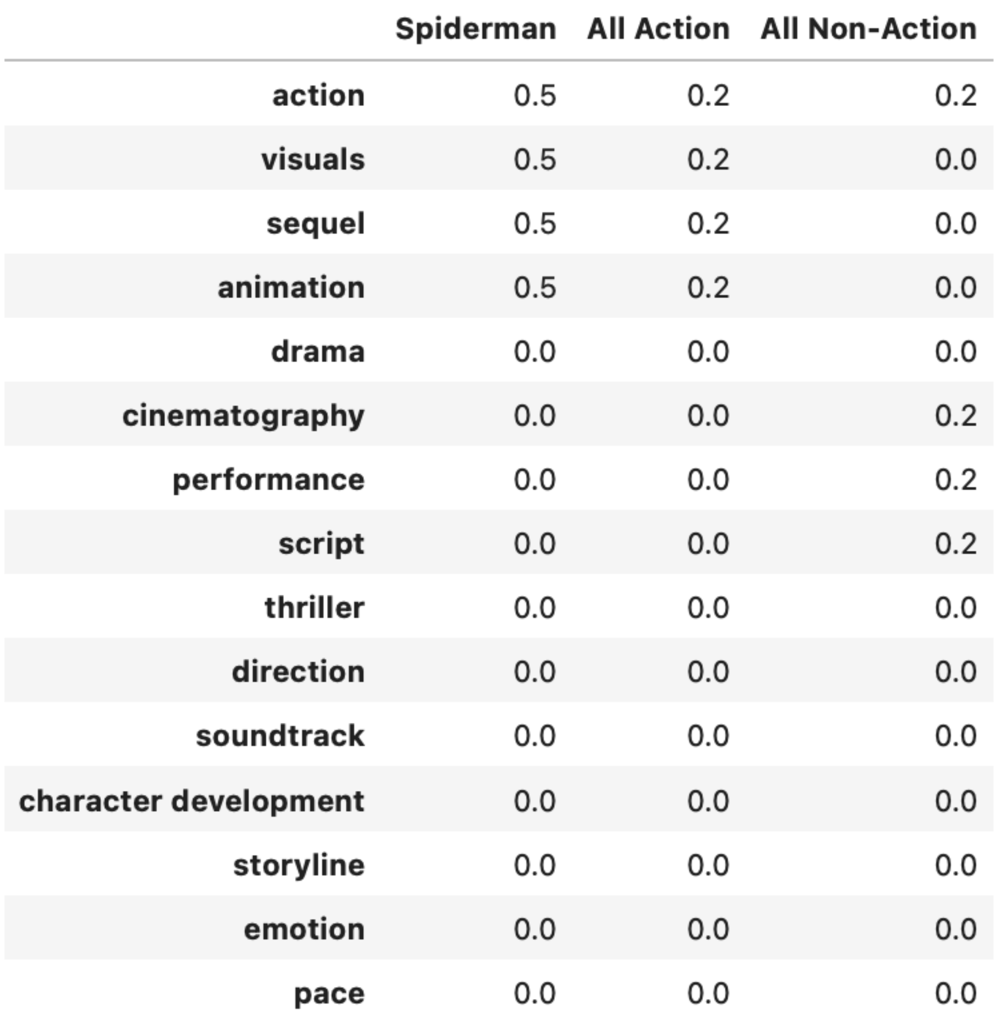
Appendix A –Mean Frequency of Terms of Interest in Subsets of Documents

Hotel Dataset



	Motel	High	Low
noise	1.000000	0.6	0.2
bed	0.666667	1.2	0.2
price	0.333333	0.2	0.0
cleanliness	0.000000	0.0	0.0
comfort	0.000000	0.0	0.0
amenities	0.000000	0.4	0.0
location	0.000000	0.0	0.0
service	0.000000	0.0	0.0
check-in	0.000000	0.0	0.0
staff	0.000000	0.4	0.0
bathroom	0.000000	0.2	1.0
parking	0.000000	0.0	0.4
pet-friendly	0.000000	0.0	0.0
renovation	0.000000	0.0	0.0
view	0.000000	0.0	0.0

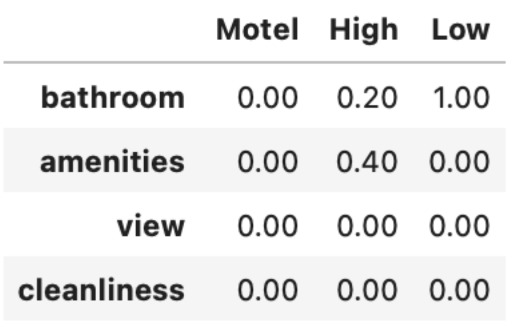
Movie Dataset



	Spiderman	All Action	All Non-Action
action	0.5	0.2	0.2
visuals	0.5	0.2	0.0
sequel	0.5	0.2	0.0
animation	0.5	0.2	0.0
drama	0.0	0.0	0.0
cinematography	0.0	0.0	0.2
performance	0.0	0.0	0.2
script	0.0	0.0	0.2
thriller	0.0	0.0	0.0
direction	0.0	0.0	0.0
soundtrack	0.0	0.0	0.0
character development	0.0	0.0	0.0
storyline	0.0	0.0	0.0
emotion	0.0	0.0	0.0
pace	0.0	0.0	0.0

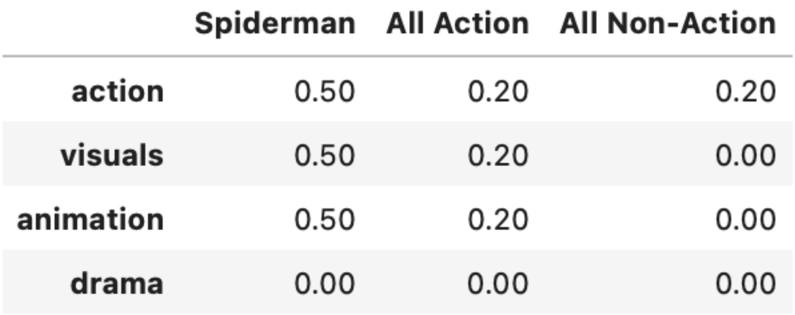
Appendix B –TF-IDF Scores for the Important Prevalent Vocabulary Terms

Hotel Dataset



	Motel	High	Low
bathroom	0.00	0.20	1.00
amenities	0.00	0.40	0.00
view	0.00	0.00	0.00
cleanliness	0.00	0.00	0.00

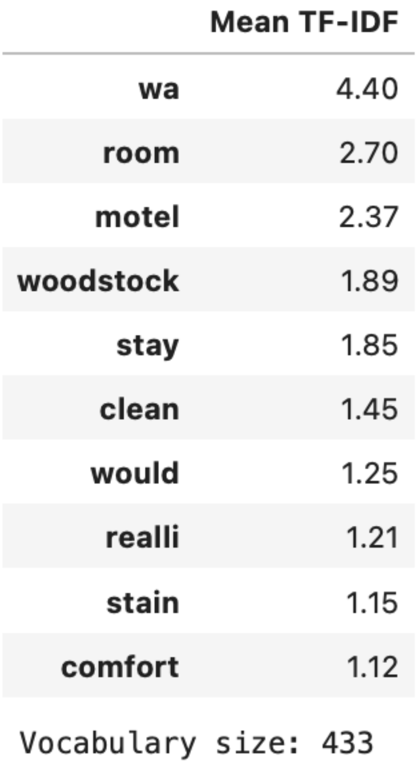
Movie Dataset



	Spiderman	All Action	All Non-Action
action	0.50	0.20	0.20
visuals	0.50	0.20	0.00
animation	0.50	0.20	0.00
drama	0.00	0.00	0.00

Appendix C – Terms with the Top Highest TF-IDF Mean Scores

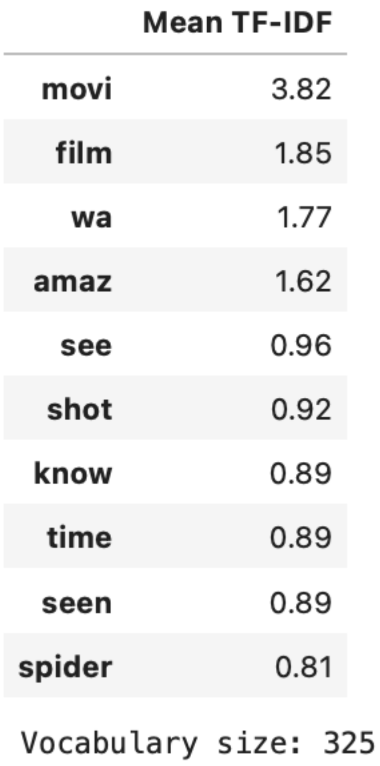
Hotel Dataset



	Mean TF-IDF
wa	4.40
room	2.70
motel	2.37
woodstock	1.89
stay	1.85
clean	1.45
would	1.25
realli	1.21
stain	1.15
comfort	1.12

```
Vocabulary size: 433
```

Movie Dataset

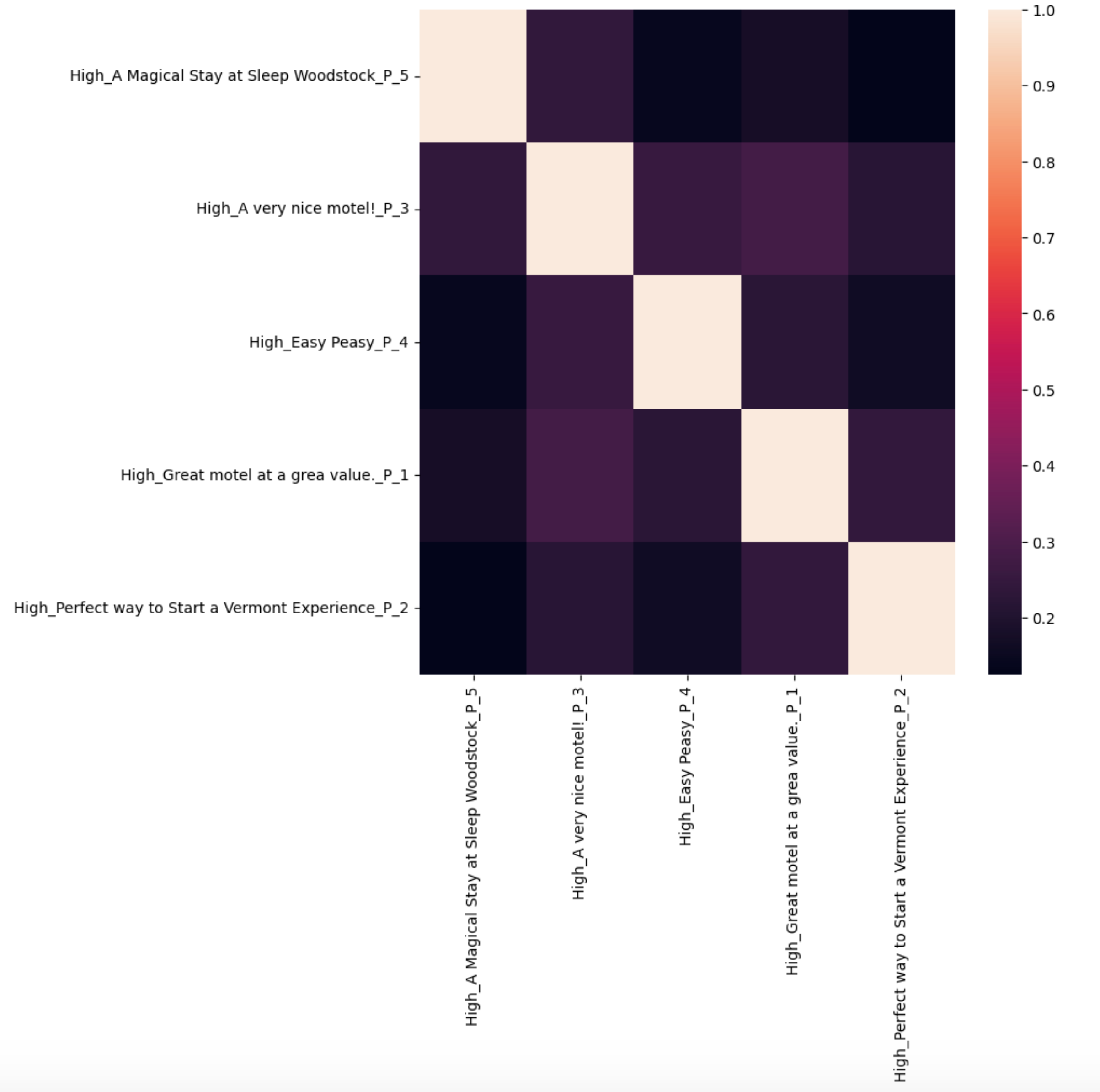


	Mean TF-IDF
movi	3.82
film	1.85
wa	1.77
amaz	1.62
see	0.96
shot	0.92
know	0.89
time	0.89
seen	0.89
spider	0.81

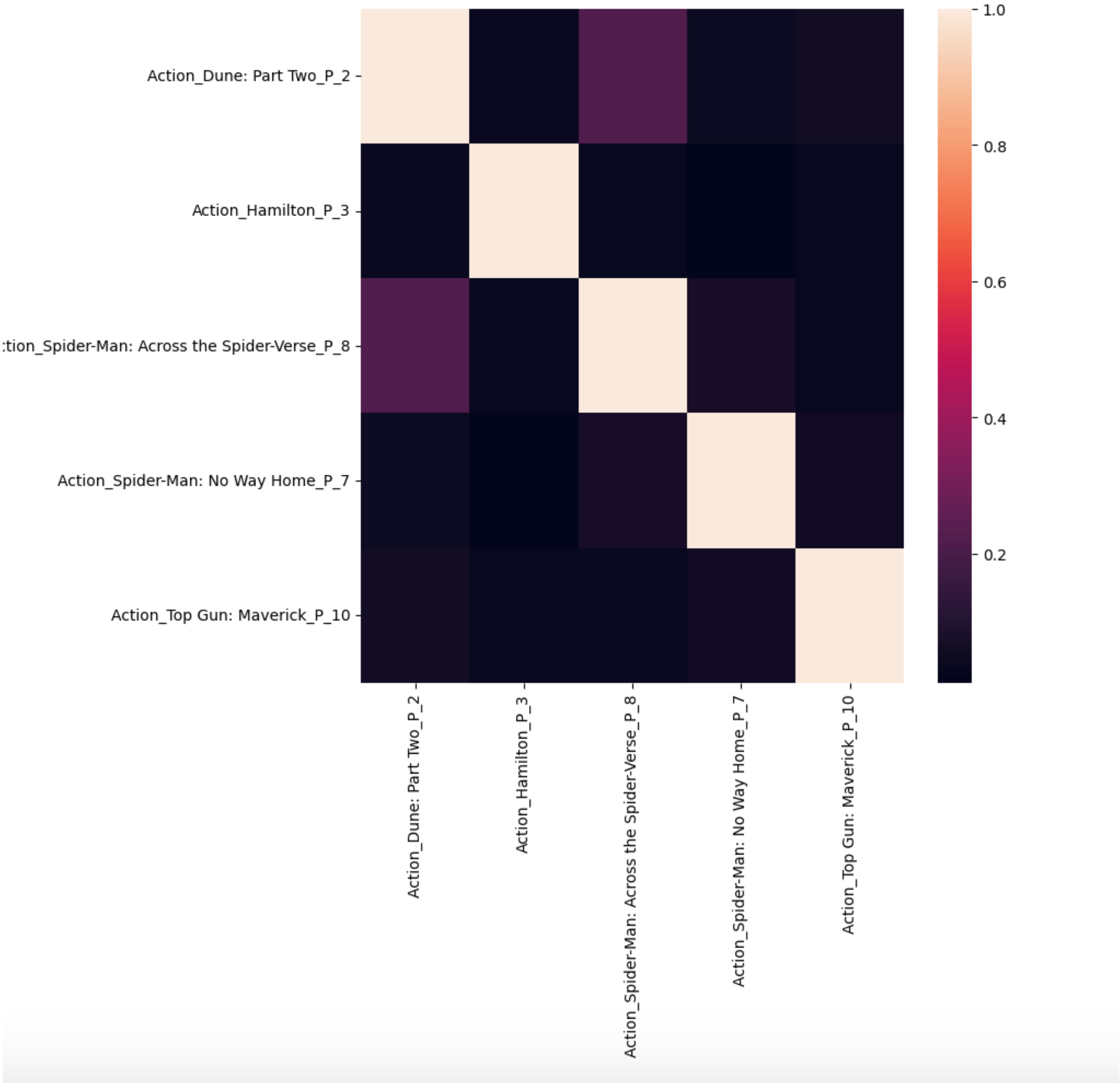
```
Vocabulary size: 325
```

Appendix D – Document Cosine Similarity Heatmaps Using TF-IDF Vectorization

Hotel Dataset

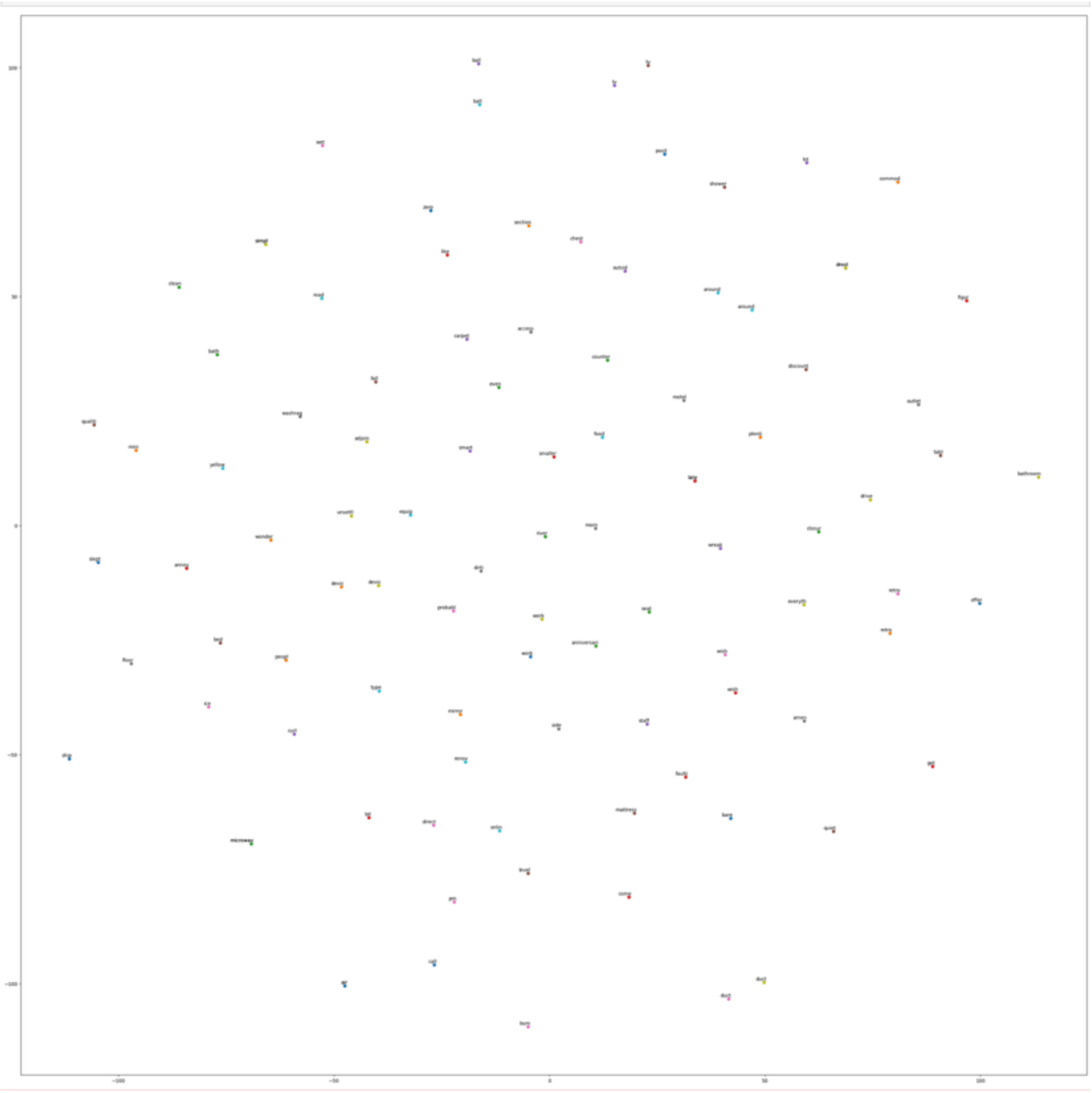


Movie Dataset

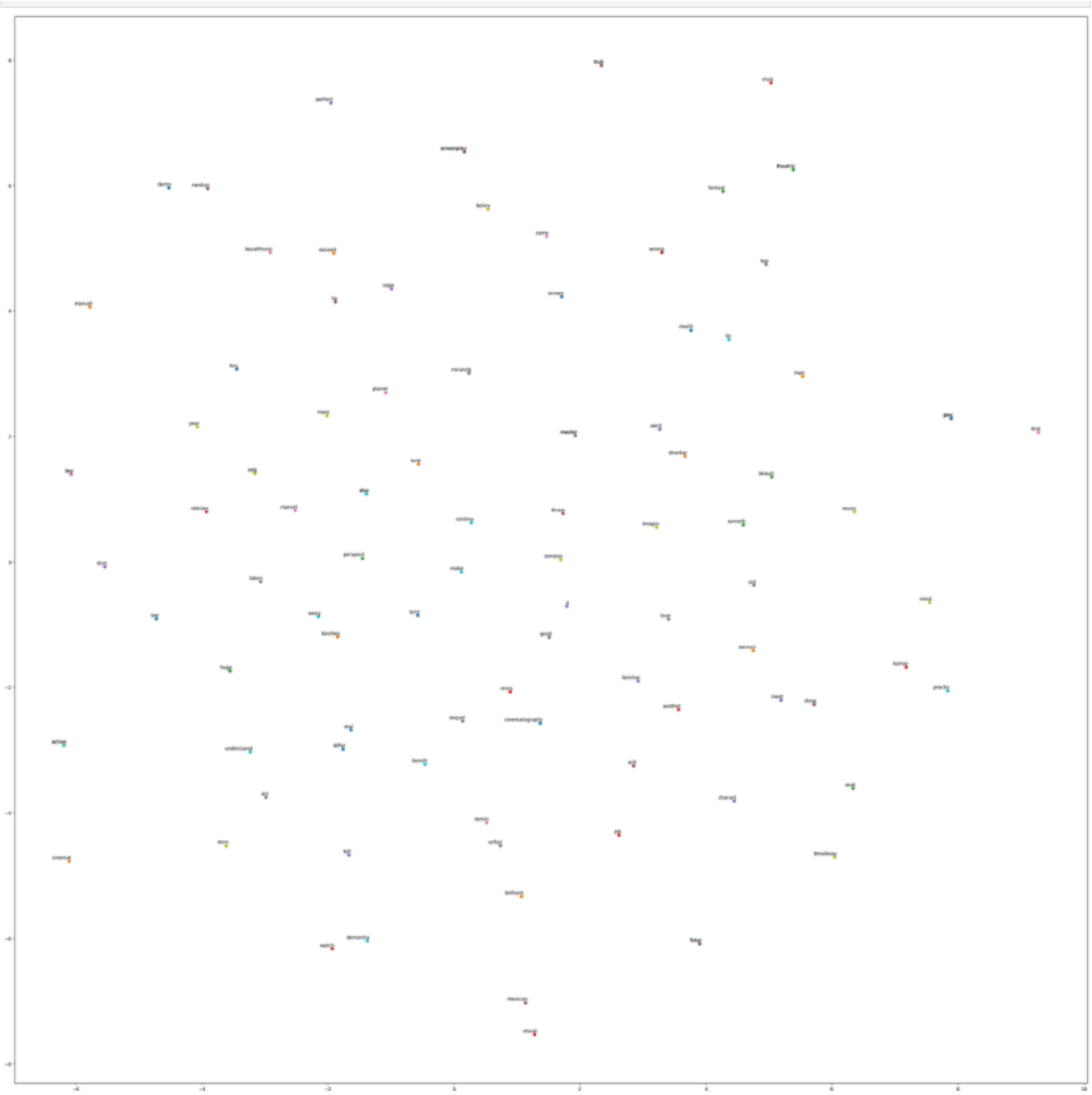


Appendix E - T-SNE Plot of Documents Using word2vec Vectorization

Hotel Dataset

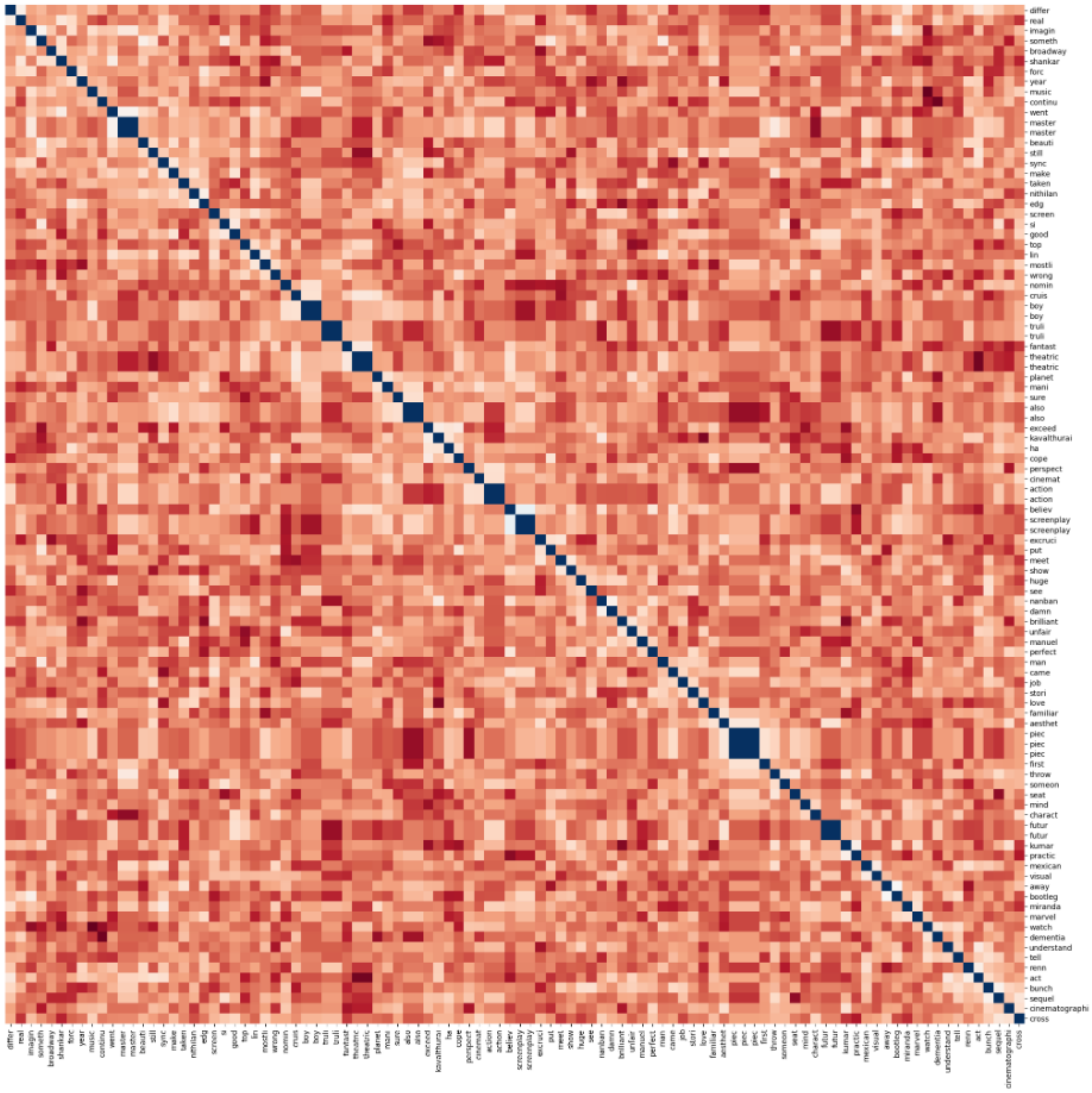


Movie Dataset

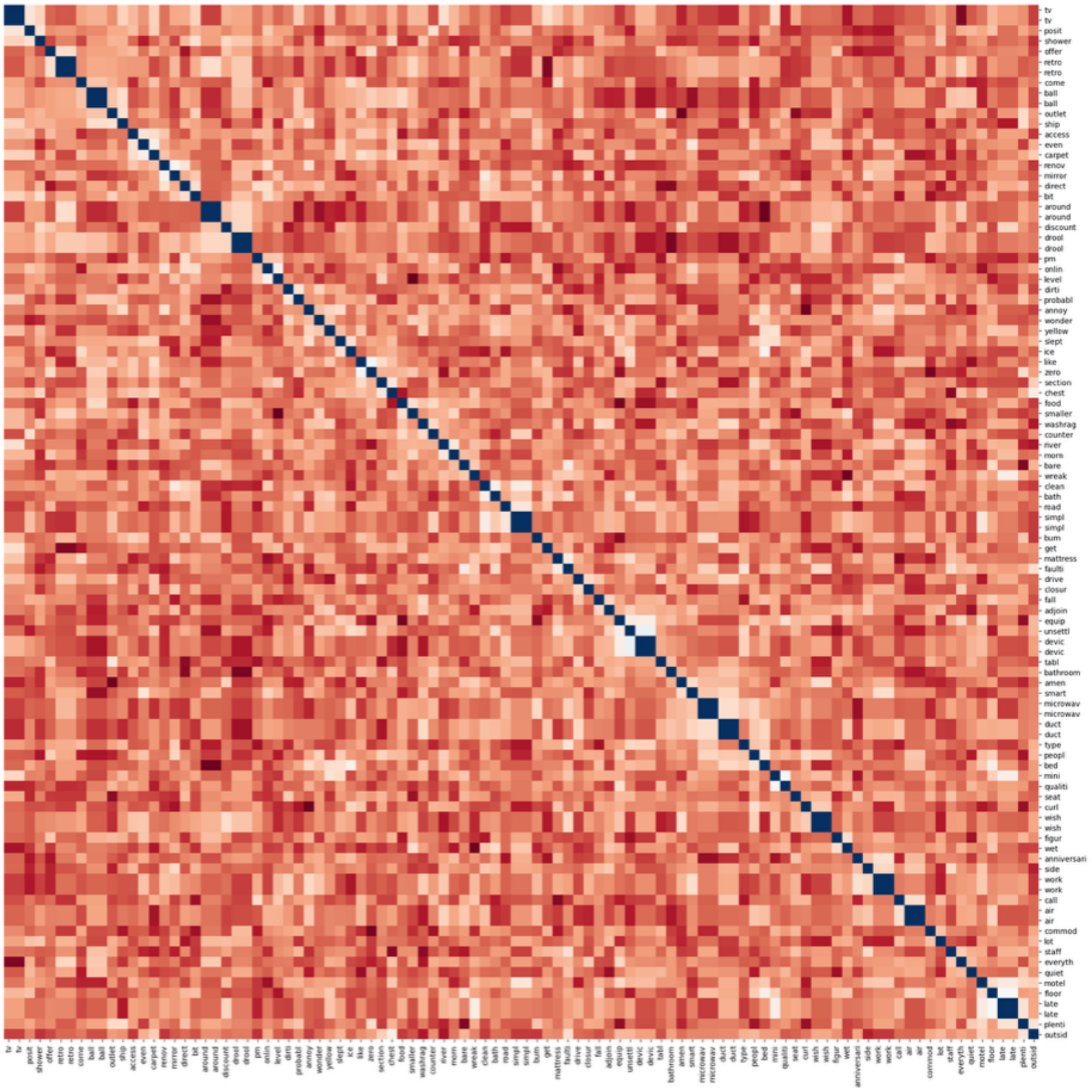


Appendix F – T-SNE Plot of the Top TF-IDF Terms Using Word2Vec Vectorization

Movie Dataset

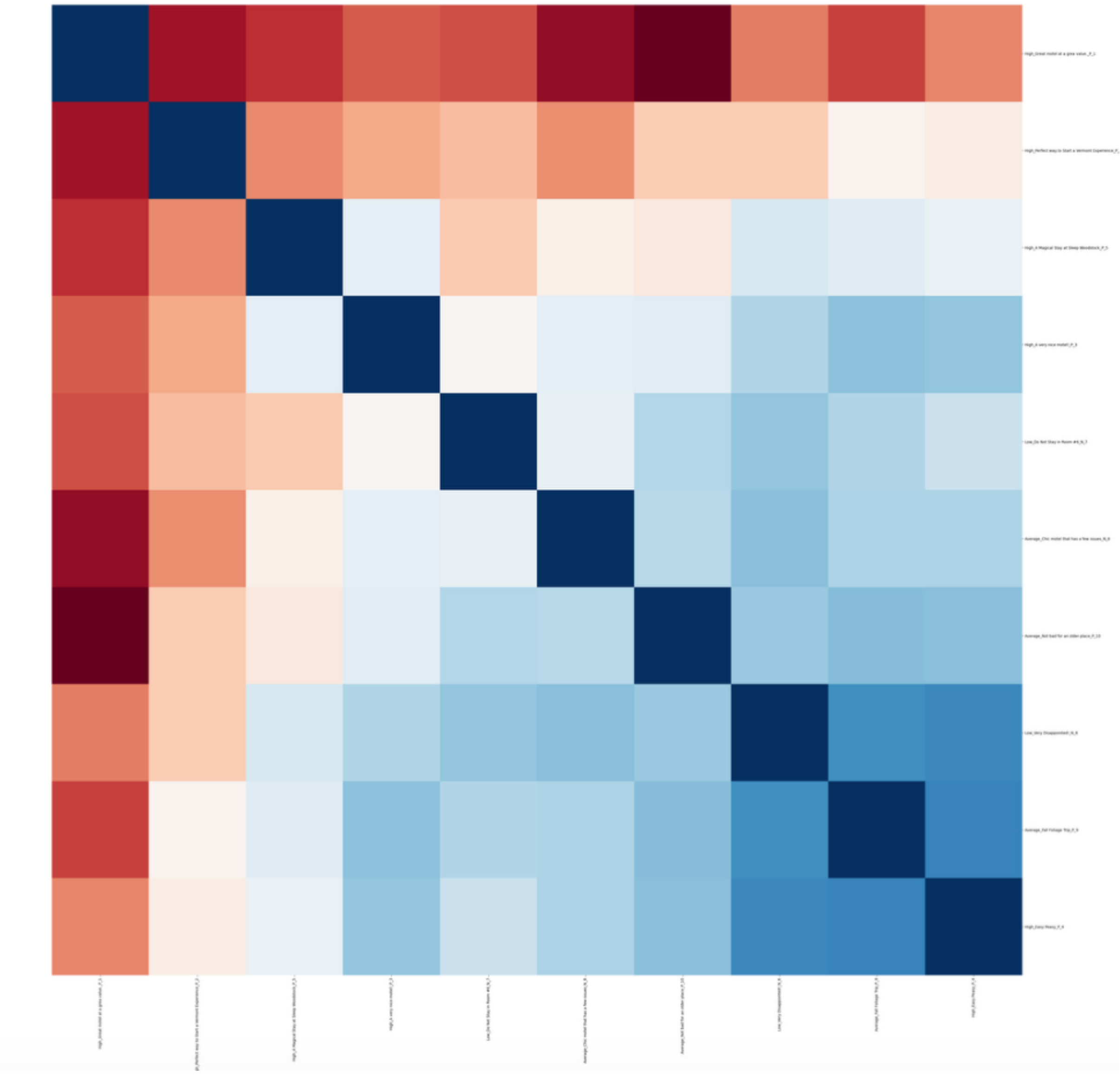


Hotel Dataset

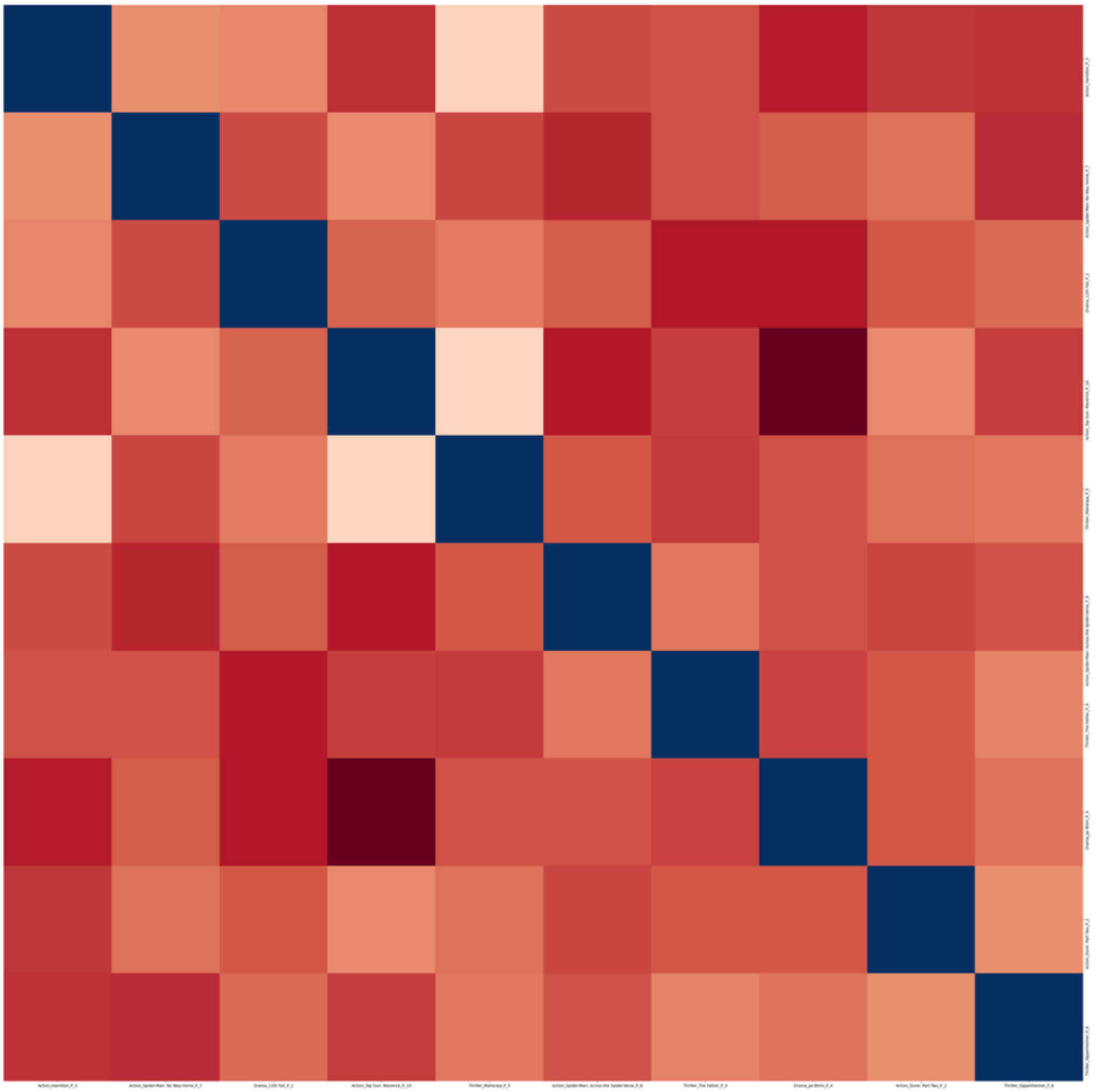


Appendix G – Document Cosine Similarity Heatmap Using Doc2Vec Vectorization

Hotel Dataset

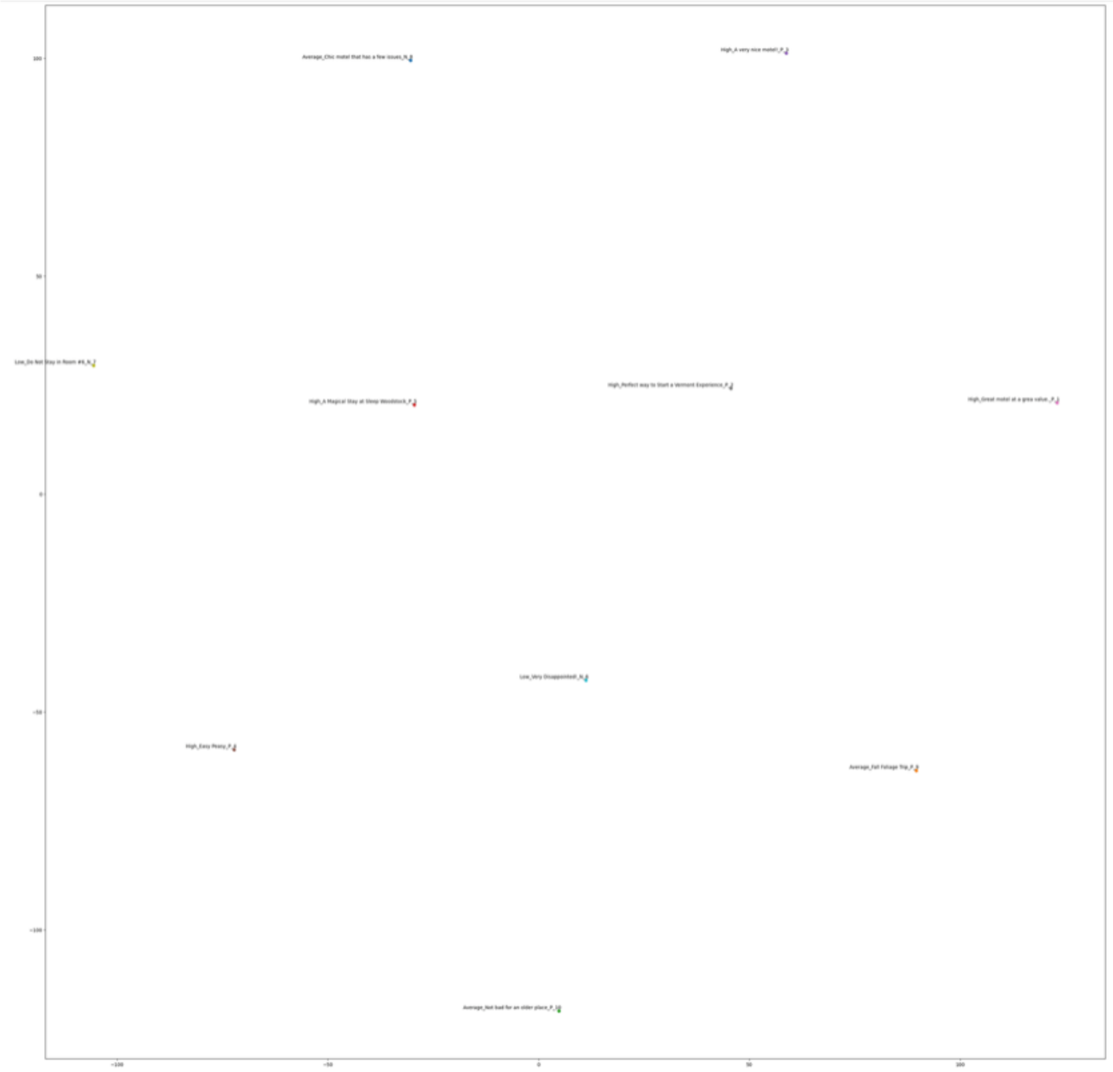


Movie Dataset

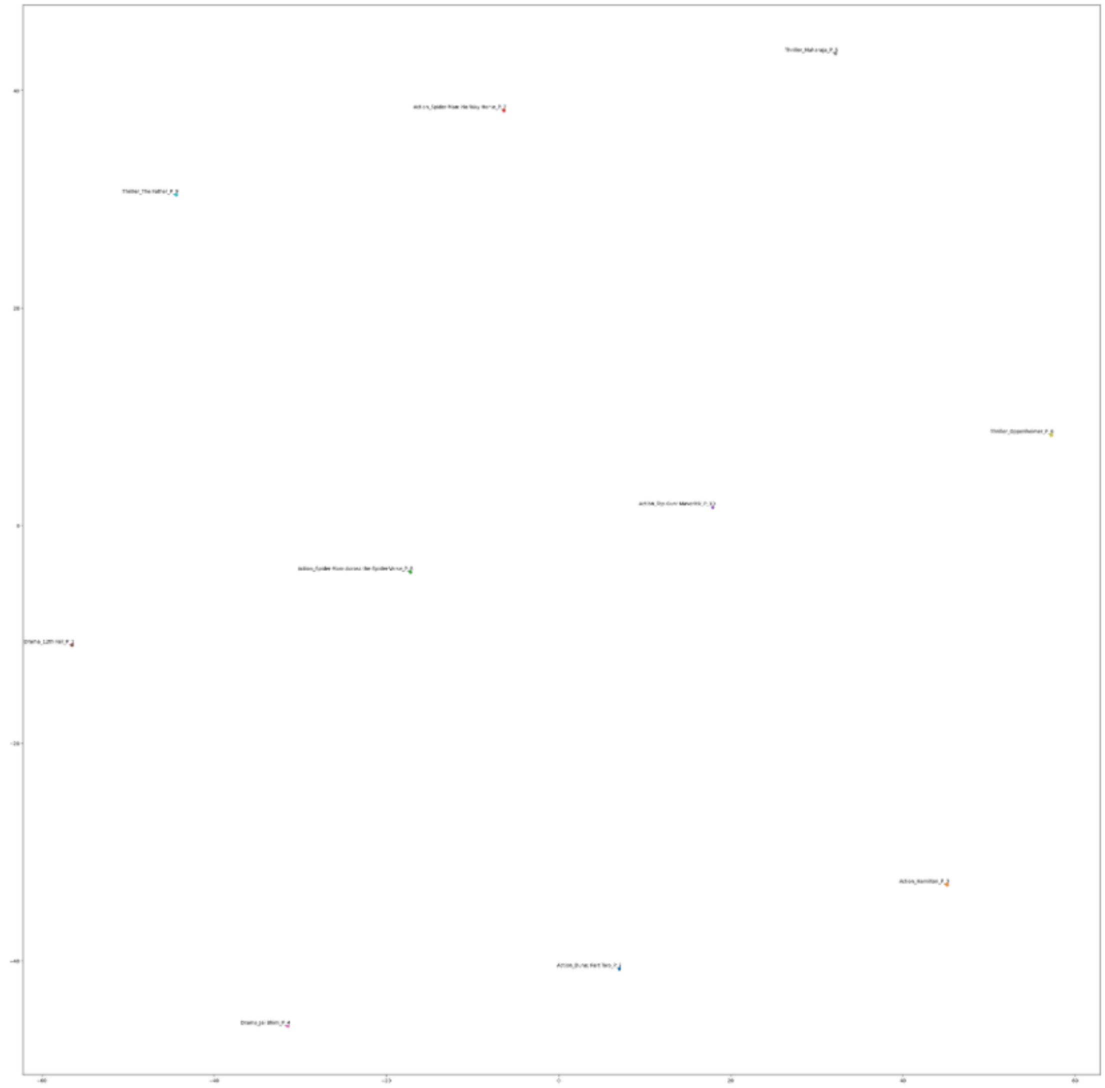


Appendix H – T-SNE Plot of Documents Using doc2vec Vectorization

Hotel Dataset

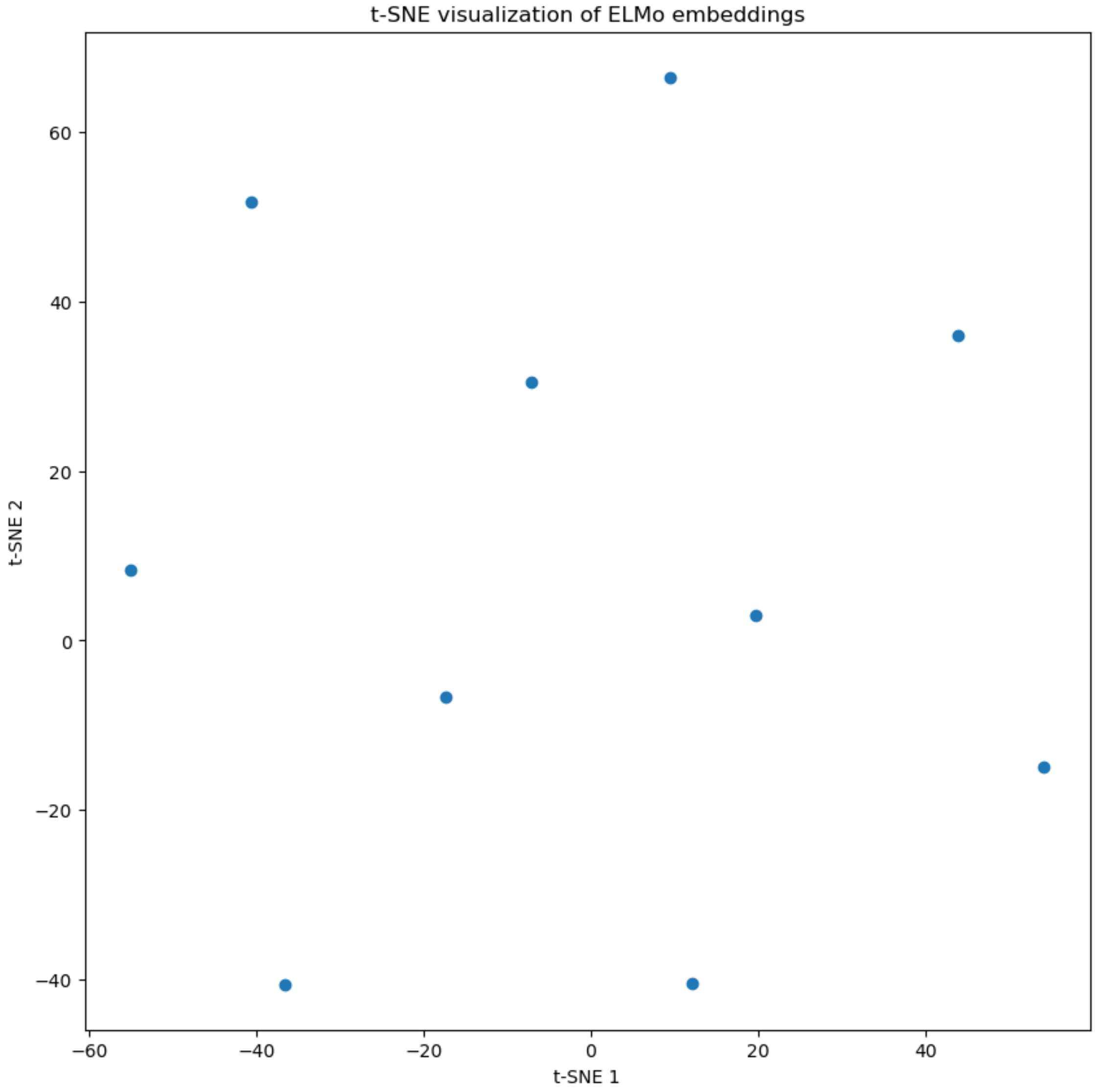


Movie Dataset

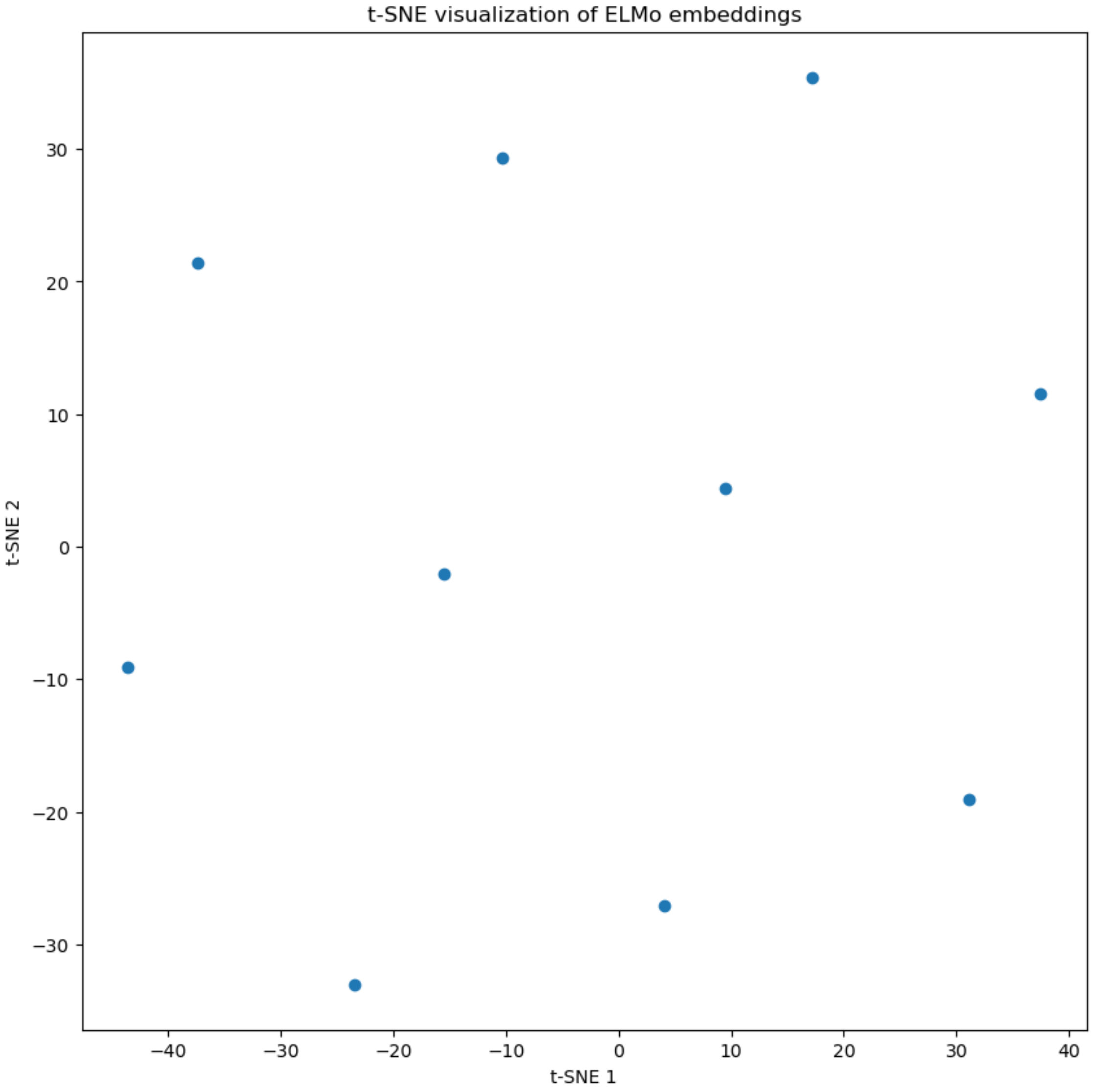


Appendix I – ELMo Embedding output Visualization

Hotel Dataset



Movie Dataset


t-SNE visualization of ELMo embeddings
t-SNE 2
30
20
10
0
−10
−20
−30
−40
−30
−20
−10
0
10
20
30
40
t-SNE 1