1.1
The term-document matrix dimensions are not sufficient to fully represent the meaning of words. While this matrix captures word frequencies across documents, it lacks the ability to encode fine-grained semantic relationships such as synonymy, antonymy, or other lexical relations. Additionally:

- **Sparsity**: Given the size of the corpus, many words do not appear in all documents, leading to a sparse matrix. This sparsity makes it difficult to capture meaningful relationships between words.
- **Loss of word order**: The term-document matrix operates under a bag-of-words model, which completely ignores word order and syntactic structure, crucial components for understanding meaning.
- **Polysemy**: Words with multiple meanings (polysemy) are conflated into a single vector, which could confuse the distinct meanings of the same word.

In conclusion, while the term-document matrix can provide some insight into word frequencies and document associations, it is not sufficient for deeper semantic analysis. Methods like word embeddings or contextual embeddings are more effective for capturing nuanced word meanings.

1.2:

Here are the top 10 most similar words to "Juliet", "Player", and "Attain" using cosine similarity on both the term-document and term-context matrices.

For "Juliet":

- **Term-Document Matrix**:
  1. Capulet (0.9899), 2. Mercutio (0.9899), 3. Romeo (0.9899)
     Notable associations: These characters co-occur frequently with Juliet in the same play, which is why they score highly.
- **Term-Context Matrix**:
  1. Lucius (0.7877), 2. Gloucester (0.7818), 3. Servants (0.7717)
     Notable associations: Here, the words reflect broader relationships within Shakespeare's works, indicating a stronger semantic and contextual similarity.

For "Player":

- **Term-Document Matrix**:
  1. Alarm (0.9058), 2. Incestuous (0.8744), 3. Summit (0.8682)
     Notable associations: These words might co-occur in documents with "Player," but do not seem semantically relevant.
- **Term-Context Matrix**:
  1. Like (0.5587), 2. Word (0.5475), 3. Maid (0.5441)
     Notable associations: These reflect more meaningful, context-based relationships related to language and social roles.

For "Attain":

- **Term-Document Matrix**:
  1. Common (0.7103), 2. Worthier (0.6888), 3. Rent (0.6885)
     Notable associations: These words are related to broad concepts, but some associations like "flatterers" seem irrelevant.

- **Term-Context Matrix**:
    1. Compell (0.6647), 2. Perform (0.6498), 3. Condemn (0.6440)
    Notable associations: These are more closely tied to action-based or semantic meanings related to "attain."

```
The 10 most similar words to "juliet" using cosine-similarity on term-document frequency matrix are:
1: capulet; 0.9899494936611666
2: mercutio; 0.9899494936611666
3: romeo; 0.9899494936611666
4: pump; 0.9899494936611665
5: laura; 0.9899494936611665
6: pitcher; 0.9899494936611665
7: behoveful; 0.9899494936611665
8: hurdle; 0.9899494936611665
9: capulets; 0.9899494936611665
10: petrucio; 0.9899494936611665
```

```
The 10 most similar words to "juliet" using cosine-similarity on term-context frequency matrix are:
1: lucius; 0.7877964614494171
2: gloucester; 0.7818061482037952
3: servants; 0.7717159769405332
4: warwick; 0.7677308387215129
5: nurse; 0.7590373891930818
6: paris; 0.7531720811811452
7: antonio; 0.7527362684737068
8: buckingham; 0.7489883918644318
9: brutus; 0.7485064965476168
10: soldiers; 0.747553738371912
```

```
The 10 most similar words to "player" using cosine-similarity on term-document frequency matrix are:
1: alarm; 0.9058216273156765
2: incestuous; 0.8744746321952062
3: summit; 0.8682431421244593
4: origin; 0.8682431421244593
5: norway; 0.8663254299163511
6: pyrrhus; 0.8555183532243665
7: players; 0.8360171835451684
8: flints; 0.8320502943378437
9: umbrage; 0.8320502943378437
10: uncharge; 0.8320502943378437
```

```
The 10 most similar words to "player" using cosine-similarity on term-context frequency matrix are:
1: like; 0.558705530411695
2: word; 0.5475484148719328
3: maid; 0.5441246972696346
4: gentlewoman; 0.5426510523430812
5: roman; 0.5347033094034588
6: glass; 0.5284293109365638
7: beggar; 0.5265442421247153
8: fool; 0.5261284617417139
9: slave; 0.5261055748301167
10: cock; 0.5225520536761179

The 10 most similar words to "attain" using cosine-similarity on term-document frequency matrix are:
1: common; 0.7103939035548414
2: worthier; 0.6888467201936644
3: rent; 0.6885303726590963
4: press; 0.683536555146996
5: flatterers; 0.674199862463242
6: fawn; 0.6713171133426189
7: refused; 0.669438681395203
8: low; 0.6684515873086949
9: redress; 0.6648376737370494
10: ambitious; 0.6647097715994328

The 10 most similar words to "attain" using cosine-similarity on term-context frequency matrix are:
1: compell; 0.6646856131598861
2: perform; 0.6498131986875082
3: condemn; 0.6440049685174982
4: abhorr; 0.640630896458971
5: conquer; 0.6172401694534742
6: deliver; 0.6015850344396534
7: constrain; 0.5916249558537158
8: betray; 0.587595038795117
9: enrich; 0.5824740684220558 10:
slaughter; 0.5804598291845563
```

1.3:

Based on the results above, the term-context frequency matrix tends to produce similar words that are more contextually relevant and semantically coherent compared to the term-document frequency matrix. This observation can be backed up by examining the top associated terms for "juliet" and "player," as well as the other examples provided.

Term-Document Frequency Matrix Observations

- The term-document matrix captures how frequently words co-occur with documents, which might lead to high similarity scores between words that frequently appear in the same documents but don't necessarily have similar meanings.
- For "juliet," words like "capulet," "mercutio," "romeo" are expected as they are central characters in "Romeo and Juliet." However, the presence of words like "pump," "laura," and "pitcher" among the top similar words indicates that the similarity might be driven more by their joint presence in specific documents rather than semantic similarity. - For "player," the term-document matrix associates words like "alarm," "incestuous," and "norway," which seem unrelated in a semantic sense but might appear in similar contexts or specific plays where the word "player" is also used.

Term-Context Frequency Matrix Observations

- The term-context matrix focuses on the immediate linguistic context of words, capturing more about their functional and semantic relationships. This typically yields associations that are more reflective of a word's meaning and usage in language.
- For "juliet," the similar words include characters ("lucius," "gloucester") and roles ("nurse," "servants"), which are more indicative of the social and relational context of Juliet within the narrative structures of Shakespeare's works.
- For "player," the associated words ("like," "word," "maid") reflect a broader range of usage and are more indicative of the word's function and thematic associations in text, showing a diversity of context that is more semantically meaningful.

Conclusion

The term-context matrix is more effective in capturing semantically meaningful associations between words due to its focus on the words' immediate linguistic contexts. This approach better reflects how words are used in relation to one another within the text, leading to associations that are more indicative of the words' meanings and functions.

The term-document matrix, while useful for identifying topical and document-level associations, may conflate frequency of co-occurrence across documents with semantic similarity, leading to less meaningful associations from a linguistic perspective.

The results for "juliet" and "player" in the term-context matrix are more semantically coherent, demonstrating that understanding word similarity based on immediate context rather than document-level co-occurrence provides richer insights into word meanings and their relationships.

1.4

Several key decisions were made during the implementation of the term-context matrix and related functions. These decisions directly impacted the results:

1. Excluding Target Words from Their Own Context

Decision: In constructing the term-context matrix, I chose to exclude the target word from its own context. This means that when calculating the context for a word, the word itself is not included as part of its context.

Reason: Allowing a word to co-occur with itself would artificially inflate its frequency in the matrix, which could result in an overly high similarity score between the word and itself, thus distorting the final similarity ranking.

Impact: By excluding self-co-occurrence, the matrix captures external relationships between words rather than reflecting a word's self-association. This leads to a more accurate measurement of the semantic relationships between the target word and other words. Excluding self-co-occurrence emphasizes the relationships with surrounding words, resulting in more meaningful contextual associations.

Example: In the case of "Juliet," if self-co-occurrence were allowed, it would significantly increase the similarity score between "Juliet" and itself, resulting in overly strong self-association. However, the analysis focuses more on Juliet's relationships with other characters (e.g., "Romeo" or "Capulet"), so excluding self-co-occurrence helps better capture these external relationships.

2. Window Size for the Term-Context Matrix

Decision: I selected a window size of 4 for the term-context matrix, meaning that each word's context consists of the 4 words before and after it.

Reason: The window size needs to balance capturing local semantic relationships and maintaining semantic precision. A smaller window size (e.g., 2) could capture tighter grammatical structures but might miss broader thematic and contextual relationships. Conversely, a larger window size (e.g., 8 or 10) could include more distant words, which may capture more semantic information but also introduce noise and unrelated words, reducing the accuracy of the similarity results. Choosing a window size of 4 strikes a good balance between capturing local context and maintaining semantic relevance.

Impact: A smaller window generally better captures the immediate grammatical and semantic relationships between words, which is useful for identifying fine-grained associations. A larger window helps capture broader thematic relationships but may introduce more noise. By choosing a window

size of 4, I ensured that the model captures enough meaningful contextual information while minimizing unrelated noise.

Example: In "Juliet's" context, a smaller window (e.g., 2) might only capture the closest words around "Juliet," missing out on broader contextual relationships, such as her interactions with other characters. A larger window (e.g., 8 or 10) might introduce irrelevant words that do not directly relate to "Juliet." Therefore, a window size of 4 strikes the right balance by capturing both local and wider context effectively.

3. Handling Stopwords and Low-Frequency Words

Decision: I decided to remove stopwords (e.g., "the," "and") and low-frequency words from the corpus. These stopwords frequently appear in the corpus but do not carry meaningful semantic information. For low-frequency words, I set a frequency threshold to filter out words that appear too infrequently, in order to reduce the sparsity of the matrix and improve computational efficiency.

Reason: Removing stopwords helps reduce noise, as these words frequently appear in most documents but do not differentiate between them. Low-frequency words, on the other hand, may appear too infrequently to provide useful contextual information. Filtering these words improves computational performance and ensures that the context matrix is more compact and representative of meaningful word associations.

Impact: By removing stopwords, the results focus more on contextually important words rather than on common words that do not contribute much to the semantic meaning. Filtering low-frequency words reduces the sparsity of the matrix, making computations more efficient while also improving the semantic relevance of word relationships. Retaining higher-frequency words ensures that the model captures representative contexts, leading to more reliable results.

Example: Without removing stopwords, words like "the" or "is" would frequently appear in "Juliet's" context, diluting the similarity scores with irrelevant words. By removing these stopwords, more meaningful associations, like "Romeo" or "Capulet," take precedence in the similarity calculations.

4. Using Frequency Threshold to Avoid Sparse Matrices

Decision: To reduce matrix sparsity, I chose to retain only words that appear above a certain frequency threshold in the entire corpus. This decision was made to avoid a large number of zero values in the matrix, which would improve computational efficiency.

Reason: In large corpora, many words have very low frequencies, which could lead to a sparse matrix and increased computational complexity. At the same time, these low-frequency words often do not provide useful contextual information. Setting a frequency threshold filters out these rarely occurring words, making the matrix more compact and efficient for analysis.

Impact: By retaining only higher-frequency words, the sparsity of the matrix is significantly reduced, improving the speed and efficiency of computations. This also ensures that the focus remains on words that occur frequently enough to provide meaningful context, rather than being influenced by sparsely occurring words.

Example: For instance, in Shakespeare's complete works, some very rare words (e.g., words that appear only once or twice) might not carry much semantic value. If we included these words, the matrix would contain many zero values, increasing computational difficulty. By setting a frequency threshold, we can more efficiently analyze the associations between commonly occurring words without the interference of noise from rare words.

Conclusion

By excluding self-co-occurrence, selecting an appropriate window size, removing stopwords, and filtering low-frequency words, I was able to more accurately capture the semantic relationships between words. These decisions helped reduce noise and sparsity, improved the efficiency and accuracy of the matrices, and ensured that the results were more semantically meaningful. These methods not only improved computational efficiency but also provided stronger semantic interpretations of word relationships.

1.5

Here are the top 10 most similar words to "Juliet", "Player", and "Attain" using tf-idf and PPMI-weighted matrices:

For "Juliet":

- **tf-idf Matrix**:
    1. Capulets (0.9899), 2. Mab (0.9899), 3. Jule (0.9899)
       Notable associations: These words are more thematically focused, highlighting relevant characters and events.
- **PPMI Matrix**:
    1. Capulet (0.1940), 2. Barnardine (0.1402), 3. Provost (0.1384)
       Notable associations: PPMI captures stronger semantic relationships, emphasizing family and character connections.

For "Player":

- **tf-idf Matrix**:
    1. Alarm (0.9058), 2. Incestuous (0.8744), 3. Summit (0.8682)
       Notable associations: Similar to the term-document matrix, these associations appear thematically but are not semantically strong.
- **PPMI Matrix**:

    1. Tripped (0.2406), 2. Describes (0.2205), 3. Culled (0.2004)
       Notable associations: PPMI provides more action-related terms, which are semantically relevant to the concept of "Player."

For "Attain":

- **tf-idf Matrix**:
    1. Common (0.7103), 2. Worthier (0.6888), 3. Rent (0.6885)
       Notable associations: These words are related to broad concepts, but some associations like "flatterers" seem irrelevant.
- **PPMI Matrix**:
    1. Easeful (0.2471), 2. Delicious (0.1641), 3. Curlish (0.1582)
       Notable associations: PPMI captures stronger semantic relationships focused on action-oriented words.

```
The 10 most similar words to "juliet" using cosine-similarity on tf-idf matrix are:
1: capulets; 0.9899494936611666
2: mab; 0.9899494936611666
3: jule; 0.9899494936611666
4: lammas; 0.9899494936611666
5: susan; 0.9899494936611666
6: capulet; 0.9899494936611665
7: pump; 0.9899494936611665
8: laura; 0.9899494936611665
9: pitcher; 0.9899494936611665
10: behoveful; 0.9899494936611665

The 10 most similar words to "juliet" using cosine-similarity on PPMI matrix are:
1: capulet; 0.19400627099532375
2: barnardine; 0.14025071040829773
3: provost; 0.13841558109464036
4: tybalt; 0.1379093837639379
5: montague; 0.1341426209994393
6: mercutio; 0.1332444904955642
7: romeo; 0.12769146827499245
8: vauntingly; 0.1214175553181901
9: stricken; 0.12006293302129989 10:
banditti; 0.11859571208140018

The 10 most similar words to "player" using cosine-similarity on tf-idf matrix are:
1: alarm; 0.9058216273156767
2: incestuous; 0.8744746321952062
3: summit; 0.8682431421244592
4: origin; 0.8682431421244592
```

```
5: norway; 0.866325429916351
6: pyrrhus; 0.8555183532243664
7: players; 0.8360171835451684
8: flints; 0.8320502943378438
9: umbrage; 0.8320502943378438
10: uncharge; 0.8320502943378438

The 10 most similar words to "player" using cosine-similarity on PPMI matrix are:
1: tripped; 0.24061513929634293
2: describes; 0.2205497898147668
3: culled; 0.200475208041475
4: strutting; 0.19409457453521184
5: dennis; 0.1879801978653447
6: chanticleer; 0.18538271063135758
7: quadrangle; 0.1846913830088136
8: football; 0.1837278414906618
9: dismantle; 0.17680655015855573 10:
swans; 0.17380591054098038

The 10 most similar words to "attain" using cosine-similarity on tf-idf matrix are:
1: common; 0.7103939035548413
2: worthier; 0.6888467201936644
3: rent; 0.6885303726590964
4: press; 0.6835365551469959
5: flatterers; 0.6741998624632421
6: fawn; 0.6713171133426189
7: refused; 0.669438681395203
8: low; 0.6684515873086949
9: redress; 0.6648376737370493
10: ambitious; 0.6647097715994328

The 10 most similar words to "attain" using cosine-similarity on PPMI matrix are:
1: easeful; 0.24709681058546018
2: delicious; 0.16419175699426702
3: curlish; 0.15824783711956636
4: outlook; 0.15038785952154243
5: clement; 0.14652779677258476
6: oracles; 0.13931930457793196
7: congruing; 0.13604821798586397
8: recomforture; 0.13510363264019787
```

9: feebly; 0.1344374752941664 10:
uncouple; 0.13271859762960725

1.6:

Weighting with TF-IDF (Term Frequency-Inverse Document Frequency) compared to using the unweighted term-document matrix brings several key differences and improvements in capturing the importance and relevance of words within documents and across a corpus. The date illustrate these differences effectively:

TF-IDF Weighting

- Focus on Relevance: TF-IDF diminishes the weight of terms that occur very frequently across the corpus (thus likely to be less informative) and increases the weight of terms that occur frequently in a small number of documents. This helps to highlight words that are particularly relevant to specific documents.
- Discrimination of Important Terms: For "juliet," TF-IDF brings up closely related characters and terms like "capulets," "mab," and "capulet," which are directly relevant to Juliet's context within "Romeo and Juliet." This suggests that TF-IDF effectively identifies terms with specific significance to the target word's context.
- Reduction of Common Words: TF-IDF effectively filters out common words that might not contribute to understanding the unique context or meaning associated with a target word, focusing instead on terms that offer more distinctive insights.

Unweighted Term-Document Matrix

- Equal Weighting: In the unweighted term-document matrix, all occurrences are treated equally, leading to a representation that might emphasize common but less informative words.
- Less Discrimination: Without the discriminatory power of TF-IDF, the unweighted matrix might not as effectively highlight the words that are uniquely important to a document or set of documents, potentially obscuring more meaningful relationships.

Observations from Your Results

- The results from the TF-IDF matrix for "juliet" and "player" show a mix of closely related terms and some that seem less directly related. This indicates that while TF-IDF improves relevance, the interpretation of "similarity" still depends on the context and the specific relationships captured by the document corpus.
- For the PPMI matrix, which emphasizes word co-occurrence beyond mere presence within the same document, you see a different set of associations that might capture more about the relationships or roles that words play within the same contexts.

Conclusion

When comparing tf-idf to the unweighted term-document matrix, tf-idf clearly enhances the focus on contextually important words by reducing the influence of common but less informative terms. For example, in the tf-idf results for "Juliet," words like "Capulets" and "Mab" are highly relevant and directly tied to the character's context within the play, whereas the unweighted matrix includes irrelevant words like "Pump."

Tf-idf is particularly effective in reducing noise and highlighting words that carry specific significance to the document, making it more useful for document-focused analyses.

1.7:

Weighting with Positive Pointwise Mutual Information (PPMI) compared to using the unweighted term-context matrix represents a significant shift in how word relationships are quantified, moving from raw co-occurrence frequencies to a measure that emphasizes the significance of those co-occurrences. Let's explore how PPMI weighting compares to the unweighted approach, especially in light of the results you provided:

PPMI Weighting

- Emphasizes Meaningful Associations: PPMI focuses on the strength of association between words beyond their mere co-occurrence, by comparing observed co-occurrences to what would be expected by chance. This highlights words that are more meaningfully related.
- Reduces the Impact of High-Frequency Words: Similar to TF-IDF, PPMI can mitigate the dominance of high-frequency words that occur in many contexts but may not have a strong association with specific words. It prioritizes contextually unique relationships. - Captures Semantic Similarity: The PPMI weighting scheme is particularly adept at revealing semantic similarities and related concepts, as seen in the associations with characters and thematic elements closely related to "juliet" and "player."

Unweighted Term-Context Matrix

- Raw Co-occurrence Counts: The unweighted matrix simply tallies how often words appear near each other, which can surface frequent but not necessarily meaningful associations.
- Potential Noise from Common Words: Without adjusting for the overall frequency of words, common terms that appear in many contexts could unduly influence the representation, potentially obscuring more interesting or unique word relationships.

Observations from Your Results

- PPMI for "juliet": The PPMI results show closely related characters and terms (e.g., "capulet," "tybalt," "montague"), which reflects a nuanced understanding of Juliet's narrative context. This suggests PPMI's effectiveness in identifying words with a significant contextual relationship.
- PPMI for "player": The similar words identified through PPMI are more varied and seem to capture a range of associations that might reflect different aspects or contexts of "player" within the corpus, demonstrating PPMI's ability to identify a broader range of significant relationships.

Conclusion

PPMI emphasizes meaningful word associations by focusing on how often words co-occur beyond chance. This produces semantically coherent results, as seen in the case of "Juliet" being associated with "Capulet," "Tybalt," and "Romeo" — characters directly related to her in the narrative.

In contrast, the unweighted term-context matrix may capture frequent but less significant co-occurrences. For example, "Juliet" might be associated with general terms that co-occur often but lack direct relevance.

PPMI is more effective for identifying deeper semantic relationships, especially when context matters more than simple co-occurrence.

1.8

Based on the results comparing the term-document, tf-idf, term-context, and PPMI matrices, several interesting patterns emerge, highlighting the strengths and weaknesses of each approach.

Term-Document Matrix vs. Term-Context Matrix

- **Term-Document Matrix**:
  The term-document matrix is effective for identifying word co-occurrence at the document level. It tends to group words together based on their appearance in the same documents. However, this method often results in high similarity scores for words that share thematic relevance, even if they are not semantically related. For example, in the case of "Juliet," words like "Capulet" and "Mercutio" are highly associated, which makes sense given the narrative context. However, other words like "Pump" and "Pitcher" appear in the top results, likely because they frequently co-occur in the same document, but they lack any real semantic relationship to "Juliet."

  - **Strengths**: This matrix is useful for thematic analysis and identifying words that frequently appear together in the same document or play. It is effective for tasks such as document classification or topic modeling.

  - **Weaknesses**: The term-document matrix fails to capture deeper semantic relationships between words. It is prone to highlighting irrelevant words that co-occur frequently in documents without reflecting any real linguistic or semantic connections.

- **Term-Context Matrix**:
  The term-context matrix, on the other hand, focuses on the linguistic context in which words appear. This method is more effective at capturing the relationships between words that appear close to one another in sentences or phrases. For example, "Juliet" is more likely to be associated with "Nurse" or "Servants" in the term-context matrix, as these words appear in close proximity to "Juliet" in the text. This makes the term-context matrix a better choice for capturing syntactic and semantic relationships.

    - **Strengths**: It captures more meaningful semantic associations between words by focusing on their immediate context. This is useful for tasks such as synonym detection or semantic similarity analysis.
    - **Weaknesses**: The term-context matrix may struggle with longer-range dependencies, where words are semantically related but do not appear in the same immediate context. Additionally, the results can be sensitive to the chosen window size for context.

tf-idf Matrix vs. Unweighted Term-Document Matrix

- **tf-idf Matrix**:
  The tf-idf matrix improves upon the unweighted term-document matrix by down-weighting common words that appear frequently across many documents, such as function words, and up-weighting words that are particularly important to individual documents. This provides a more nuanced understanding of word importance in a corpus. For example, in the tf-idf matrix for "Juliet," terms like "Capulets" and "Mab" are more prominent, which are more closely tied to the character and her context in the play.
    - **Strengths**: The tf-idf matrix is particularly useful for identifying document-specific terms, making it ideal for information retrieval, text categorization, and topic detection.
    - **Weaknesses**: While it improves upon the raw term-document matrix, tf-idf is still based on document-level associations and may not capture deeper semantic relationships between words.
- **Unweighted Term-Document Matrix**:
  This matrix, while simple and effective for certain tasks, treats all word occurrences equally, regardless of their importance. As seen in the unweighted term-document matrix for "Juliet," it brings up many irrelevant words like "Pump" and "Pitcher," which have little semantic relevance but co-occur frequently in the same documents.
    - **Strengths**: The method is simple and computationally efficient, suitable for basic document analysis.
    - **Weaknesses**: It fails to account for word importance or the relevance of terms, leading to the inclusion of common but irrelevant words in the results.

PPMI Matrix vs. Unweighted Term-Context Matrix

- **PPMI Matrix**:
  The PPMI matrix offers significant improvements over the raw term-context matrix by focusing on the strength of association between words, adjusted for how often they would be expected

to co-occur by chance. This makes PPMI particularly effective at capturing meaningful semantic relationships. For example, in the PPMI results for "Juliet," characters like "Capulet" and "Romeo" appear at the top, which are central to Juliet's narrative, reflecting the strong contextual and relational ties between these words.

  - **Strengths**: PPMI excels at revealing deep semantic relationships between words by emphasizing rare but meaningful co-occurrences. It is especially useful for tasks that require a fine-grained understanding of word meaning and context.

  - **Weaknesses**: One limitation of PPMI is that it may not handle very rare words well, especially if there is insufficient data to calculate meaningful associations. Additionally, it can be more computationally expensive than other methods due to the additional complexity involved in computing pointwise mutual information.

- **Unweighted Term-Context Matrix**:
  The unweighted term-context matrix captures word co-occurrence in a local context, but it may give too much weight to frequently occurring words that appear in many contexts, without considering their importance. For example, words like "Lucius" and "Gloucester" might appear near "Juliet" simply due to their frequent presence in the text, but their relationship is not as strong as those captured by PPMI.

  - **Strengths**: This matrix is simple and can quickly capture local word relationships, making it useful for fast computations.

  - **Weaknesses**: It tends to overestimate the importance of frequent words and does not adjust for the significance of word associations, leading to less meaningful results compared to PPMI.

Emerging Patterns

- **Semantic Relationships**: The PPMI matrix consistently outperforms other methods when it comes to capturing semantic relationships. For example, "Juliet" is closely tied to characters and roles that have strong narrative connections to her story. The tf-idf and term-context matrices also perform well but tend to emphasize thematic or local co-occurrences rather than deep semantic links.

- **Contextual Sensitivity**: The term-context and PPMI matrices excel at identifying words that share similar contexts, such as "Juliet" and "Nurse," which frequently co-occur within the same conversations or scenes. This makes them ideal for tasks that require understanding word usage in context.

- **Document-Level Focus**: The term-document and tf-idf matrices are better suited for document-level analyses. They are more effective at identifying words that frequently appear together across entire documents or themes but may struggle with capturing fine-grained semantic relationships.

Conclusion

In summary, the choice of method depends on the specific goals of the task. The **PPMI matrix** is superior for capturing meaningful semantic relationships, making it the best choice for tasks involving word meaning and contextual analysis. The **tf-idf matrix** is effective for document-level relevance, while the **term-context matrix** provides valuable insights into word usage in context. The **unweighted term-document matrix**, while simple, is the weakest method for semantic tasks but can still be useful for broader thematic analyses.

Each method has its strengths and weaknesses, and selecting the right approach depends on the nature of the problem at hand. For synonym detection and semantic analysis, PPMI and term-context matrices are the most effective. For document classification and topic modeling, tf-idf and term-document matrices offer better results.

2.1:

Top 10 Words for Identity Labels

- For "Gay": Retired (0.6366), Old (0.5457), Bisexual (0.3592)
- For "Feminine": Christians (0.5980), Teenagers (0.3885), Male (0.3060)
- For "Black": Polish (0.3500), Irish (0.2810), Female (0.2638)
- For "Man": Men (0.3890), Woman (0.3338), Elderly (0.2778

```
The 10 most similar words to "gay" using cosine-similarity on PPMI matrix are:
1: retired; 0.6366321775399617
2: old; 0.5457130438275298
3: israeli; 0.39327783923021287
4: he; 0.3737333454119325
5: buddhists; 0.3691219872162691
6: christian; 0.3610506886832052
7: bisexual; 0.3592571448235483
8: mongolian; 0.3316653260315132
9: irish; 0.32003787473284095
10: men; 0.31834750666389566
```

```
The 10 most similar words to "feminine" using cosine-similarity on PPMI matrix are:
1: christians; 0.5980943331136044
2: italian; 0.42701586948934556
3: teenagers; 0.388550687416894
4: teenager; 0.36900691809598074
5: teenage; 0.34112080680062973
6: caucasian; 0.31955036814871285
7: senior; 0.3145544608883388
8: male; 0.30603201237654654
9: young; 0.2768592787238018
10: french; 0.26819581708811346

The 10 most similar words to "black" using cosine-similarity on PPMI matrix are:
1: polish; 0.35009180567933385
2: christian; 0.33059937270372086
3: italian; 0.28812705611934575
4: mexican; 0.2874820853822535
5: irish; 0.28105164009581274
6: masculine; 0.28104490318324893
7: egyptian; 0.2667068972334261
8: brazilian; 0.26612919997889894
9: female; 0.2638523661731573
10: senior; 0.26327384133316545

The 10 most similar words to "man" using cosine-similarity on PPMI matrix are:
1: men; 0.389046802716966
2: woman; 0.3338111953766085
3: women; 0.2917678383497404
4: elderly; 0.277821275677564
5: handicapped; 0.2306126479268863
6: old; 0.22824808945107922
7: girls; 0.2110843532724488
8: he; 0.1929320405876781
9: hindu; 0.19140438905836143
10: jewish; 0.1889527956605055

The 10 most similar words to "american" using cosine-similarity on PPMI matrix are:
1: male; 0.49152366954233107
2: caucasian; 0.4884615125641355
3: women; 0.30607903092726474
```

```
4: latino; 0.30488947866356275
5: african; 0.27868941901791655
6: female; 0.2650020113811229
7: hispanic; 0.2582803408409433
8: elderly; 0.23485175860767538
9: woman; 0.22837491641718666 10:
pakistani; 0.2081085875561799
```

2.1.2:

The results from the cosine similarity on the PPMI matrix for words like "gay," "feminine," "black," "man," and "american" reveal interesting and, in some cases, potentially stereotypical associations between terms. Let's discuss the associations and their potential implications, especially in terms of reflecting social stereotypes:

Gay

- The association of "gay" with terms like "retired," "old," and "bisexual" seems to mix unrelated concepts (age and sexual orientation) with more relevant terms. The presence of "he" and "men" could reflect stereotypical gender associations with sexual orientation. However, the connections with "retired" and "old" do not immediately seem relevant and may be artifacts of the dataset's biases or the limited context window used for analysis.

Feminine

- The term "feminine" is associated with "christians," "italian," "teenagers," and "male," among others. The mix of religious, ethnic, age-related, and gender terms suggests a broad and potentially stereotypical set of associations with the concept of femininity. For example, linking "feminine" with specific religions or nationalities may reflect cultural stereotypes present in the dataset.

Black

- The associations of "black" with various nationalities ("polish," "italian," "mexican") and "masculine" might reflect stereotypical or biased correlations within the dataset. The linkage with "senior" and "female" also suggests a diverse range of contexts in which "black" appears, potentially indicating the dataset's complexity or bias.

Man

- The word "man" is most closely associated with "men," "woman," "women," and "elderly," showing expected connections with gender and somewhat with age. The presence of "handicapped" and "old" in the list might reflect societal stereotypes or biases about gender roles or attributes.

American

- "American" is associated with "male," "caucasian," "women," "latino," and "african," suggesting a nuanced representation of American identity that includes race and gender. The strong association with "male" and "caucasian" might reflect stereotypical views of the "average American" or predominant representations in the dataset.

Interpretation and Discussion

The associations observed can reflect the biases inherent in the dataset used to generate the PPMI matrix and by extension, the societal stereotypes and biases that exist in the broader culture. Some associations seem to reinforce stereotypes (e.g., linking "feminine" with "teenagers" or "man" with "elderly" and "handicapped"), while others may result from the co-occurrence patterns in the data without necessarily reflecting direct stereotypes.

It's important to note that the presence of these associations in a mathematical model does not validate them but rather reflects the biases present in the dataset. Such findings underscore the need for careful consideration and mitigation of biases in NLP applications, especially those related to identity and human attributes. When interpreting these results, one must consider the context, the construction of the PPMI matrix, and the broader societal implications of reinforcing or challenging these stereotypes through technology.

In summary, the observed associations highlight the complexity of language and its reflection of societal norms and biases. They serve as a reminder of the critical need for diversity and inclusivity in data collection and the importance of ethical considerations in NLP research and applications.

Representational harms in NLP models, such as those trained on the SNLI corpus, occur when the model's outputs reinforce harmful stereotypes or biases, thereby perpetuating negative representations of certain groups or individuals. These harms can be broadly categorized into two types:

1. Stereotyping Harm: When a model reinforces unfounded or harmful generalizations about a group. This can lead to oversimplified, misleading, or demeaning portrayals. 2. Denigration Harm: When a model's representations demean or degrade groups or individuals based on their identity attributes (e.g., race, gender, sexual orientation).

Examination and Examples from the Provided Associations

1. Stereotyping Harm:

- "Gay" associated with "retired" and "old": This association could suggest a stereotype that being gay is linked with older age or a particular lifestyle, which does not represent the diversity and breadth of the LGBTQ+ community. It's a form of stereotyping harm because it might reinforce incorrect assumptions about gay individuals' life stages or experiences.
- "Feminine" linked with "teenagers", "teenager", and "young": These associations might suggest that femininity is primarily a characteristic of the young or that it diminishes with age, reinforcing age-related stereotypes about femininity and potentially marginalizing older individuals' identities.

2. Denigration Harm:

- While the provided associations do not directly suggest denigration harm through explicit negative portrayals, the implication of certain stereotypes can indirectly lead to denigration. For example, if the model systematically associates positive attributes with certain groups (e.g., "american" with "male" and "caucasian") and less with others, it might contribute to a form of representational harm by elevating some groups over others in a manner that could be seen as diminishing the value or identity of those not similarly highlighted.

Interpretation and Conclusion

Upon examining the associations, it's evident that there could be potential for stereotyping harm within the SNLI corpus when using a bag-of-words model. The associations suggest a risk of reinforcing stereotypes related to age, gender, and ethnicity. For instance, the association of identity terms with seemingly unrelated attributes (e.g., "gay" with "old" or "feminine" with "teenagers") can inadvertently perpetuate stereotypes about these groups.

The representational harms identified primarily fall under the category of stereotyping. These harms arise from the model's potential to reinforce narrow, possibly misleading conceptions of social groups, which could influence perceptions and treatment of individuals based on those biased representations.

While direct denigration harm was not explicitly identified in the provided examples, the subtler implications of the associations—such as reinforcing societal hierarchies or preferences for certain identities over others—could contribute to a broader context of representational harm.

In conclusion, while the specific examples from the cosine similarity analysis on the PPMI matrix of the SNLI corpus provide a limited view, they highlight the importance of critically evaluating and mitigating biases in language models. Such evaluations are crucial for developing more inclusive, fair, and representative NLP systems.

2.2.1:
The retired gay couple are dressed. an old gay
couple walks through the woods

Not find any context word occur together of separately for feminine

A young girl's hand with blue nail polish and a tattoo that says' no regrets' in cursive lettering is clenched in a loose fist in front of a black and white zippered handbag.

“Man men in white jackets with the number ""3"" on the back and black pants face away from the camera on a raised platform while a crowd gathers behind them." Man woman and a napping child traveling.
A man woman and little boy sitting in a meadow.

An African American male is a Electronic background worker.
0,two african american males looked away from the crowd
0,two african american males were driving down the road 0,two
aftrican american males were dancing in a crowd An African
American male is at a skatepark.
An African American male is with another person.
An African American male is sitting on a couch.
"An African American, Hispanic and Caucasian American are talking.”

The examples give us a glimpse into how certain identity terms, such as "gay," "feminine," "man," and "African American," appear in the SNLI corpus, both in contexts where they occur directly and in contexts suggesting second-order similarity (where similar context words are used). Let's examine these in the context of first-order and second-order similarities and discuss the findings.

Gay

First-Order Similarity:

- "The retired gay couple are dressed."
- "an old gay couple walks through the woods."

These sentences directly mention "gay couples" alongside descriptors like "retired" and "old," indicating a direct association between the identity term "gay" and aging or life stages. This can be reflective of stereotyping harm if such associations implicitly suggest that being gay is predominantly an attribute of older individuals, potentially overlooking the diversity within the LGBTQ+ community across all ages.

Feminine

No Direct Context Provided:

- Not finding any context word occur together or separately for "feminine" indicates a potential gap in the dataset or a lack of diverse contexts in which femininity is discussed. This absence itself can be telling, suggesting representational harm through omission, where the complexity and diversity of feminine identities are not adequately represented.

Man

First-Order Similarity:

- “Man men in white jackets with the number ""3"" on the back and black pants face away from the camera on a raised platform while a crowd gathers behind them."
- "Man woman and a napping child traveling."
- "A man woman and little boy sitting in a meadow."

These sentences show men in various social roles and settings, from sports or events to family contexts. The presence of men in diverse scenarios can offer a more nuanced representation, though the specific activities and roles highlighted may still reflect societal norms or stereotypes (e.g., men associated with sports or as part of a nuclear family).

American male

First-Order Similarity:

- Descriptions of American males in various activities and settings, from skateparks to social gatherings.

These examples show American male individuals in a range of everyday activities, which can help counteract stereotypical representations by showcasing normalcy and diversity within African American experiences. However, the repeated mention of race in contexts where it might not be directly relevant could also risk othering or unnecessarily highlighting race in situations where it isn't the focal point.

Discussion and Findings

The examples provided illustrate both first-order similarity, where identity terms directly appear with specific contexts, and hint at second-order similarities in how certain activities or settings might be associated with specific identities, even if not explicitly mentioned.

- Representation Diversity: The diversity in representation for terms like "man" and "African American" in various contexts is crucial for mitigating representational harms. It showcases individuals in a range of roles and activities, challenging monolithic stereotypes.
- Absence and Omission: The lack of diverse contexts for "feminine" raises concerns about omission and the importance of inclusive representation. Representation through omission can be as impactful as direct representation, as it shapes perceptions of what is considered notable or normal.
- Stereotyping through Context: The association of "gay" with "old" or "retired" couples might inadvertently reinforce stereotypes or narrow perceptions about the LGBTQ+ community. Such associations suggest the importance of diverse and nuanced representations that reflect the full spectrum of experiences within any identity group.

In conclusion, examining first-order and second-order similarities in dataset contexts reveals much about the representational dynamics at play. It underscores the need for careful, inclusive, and diverse dataset curation to ensure that NLP models built on these datasets do not perpetuate representational harms or stereotypes.