Homework 1

Part I.

Q1.

I think it's not enough to summarize all the information of the word in the document. The term-document matrix provides a basic, interpretable representation of words in a corpus but fails to capture deep semantic meaning. More advanced methods like word embeddings or transformers are needed to represent words in a way that reflects meaning, context, and relationships.

Q2.

For the word "juliet" top 10 similar word based on:

Term document frequency matrix

```
The 10 most similar words to "juliet" using cosine-similarity on term-document frequency matrix are:
1: juliet; 0.9999999999999999
2: capulet; 0.9899494936611666
3: falconer; 0.9899494936611665
4: fantasticoes; 0.9899494936611665
5: skains; 0.9899494936611665
6: versal; 0.9899494936611665
7: capulets; 0.9899494936611665
8: suspecting; 0.9899494936611665
9: needly; 0.9899494936611665
10: affray; 0.9899494936611665
```

Term context frequency matrix

```
The 10 most similar words to "juliet" using cosine-similarity on term-context frequency matrix are:
1: juliet; 1.0000000000000002
2: lucius; 0.7877964614494173
3: gloucester; 0.7818061482037953
4: servants; 0.7717159769405332
5: warwick; 0.7677308387215129
6: nurse; 0.759037389193082
7: paris; 0.7531720811811454
8: antonio; 0.7527362684737068
9: buckingham; 0.7489883918644319
10: brutus; 0.7485064965476168
```

For the word "ursula" top 10 similar word based on:

Term document frequency matrix

```
The 10 most similar words to "ursula" using cosine-similarity on term-document frequency matrix are:
1: ursula; 0.9999999999999998
2: hero; 0.9975445743429945
3: enigmatical; 0.9970544855015815
4: drovier; 0.9970544855015815
5: recheat; 0.9970544855015815
6: palabras; 0.9970544855015815
7: comprehended; 0.9970544855015815
8: crackers; 0.9970544855015815
9: endowed; 0.9970544855015815
10: crossness; 0.9970544855015815
```

Term context frequency matrix

```
The 10 most similar words to "ursula" using cosine-similarity on term-context frequency matrix are:
1: ursula; 1.0
2: attendants; 0.5843408972540847
3: vernon; 0.5690855010289954
4: gratiano; 0.5664542825449589
5: conrade; 0.5526818861547884
6: clown; 0.5515293588270307
7: sebastian; 0.5479364141766215
8: ferdinand; 0.5451485469582902
9: parolles; 0.5378337160789882
10: bardolph; 0.5365321611007914
```

For the word "sicily" top 10 similar word based on:

Term document frequency matrix

```
The 10 most similar words to "sicily" using cosine-similarity on term-document frequency matrix are:
1: sicily; 1.0000000000000002
2: nilus; 0.9707253433941511
3: laden; 0.9486832980505138
4: blemishes; 0.9486832980505138
5: enobarbus; 0.9428090415820636
6: mecaenas; 0.9428090415820635
7: surfeiter; 0.9428090415820635
8: outroar; 0.9428090415820635
9: crested; 0.9428090415820635
10: briefest; 0.9428090415820635
```

Term context frequency matrix

```
The 10 most similar words to "sicily" using cosine-similarity on term-context frequency matrix are:
1: sicily; 1.0
2: their; 0.6382434688422391
3: sleep; 0.6287005614198569
4: health; 0.6256295842887581
5: his; 0.6204280960595437
6: both; 0.6162318751932173
7: regard; 0.6106898569395511
8: conclusion; 0.6101521802011937
9: arms; 0.6063285349177694
10: forgery; 0.6029278668121224
```

Q3.

The term-context matrix produces more meaningful word similarities compared to the term-document matrix when finding words similar to "Juliet". Words that frequently appear in the same context windows as *"Juliet"* are more likely to be semantically related. For the term context frequency matrix, "nurse" and "paris" are closely related to Juliet in *Romeo and Juliet*, which makes sense, "Lucius" and "Gloucester" are characters from *Shakespearean plays*, meaning they likely appear in similar contexts, "servants" suggests that Juliet is often mentioned in the context of household roles.

Q4.

I didn't remove the target word in the result as from a homework perspective is can somehow help me to validate whether my implementation is correct as it should always be 1.0. I kept window size to 4 as it's not too large to include the useless information and not too small to only check the adjacent words to lose the sentence semantics.

Q5.

For the word "juliet" top 10 similar word based on:

Tf-idf matrix

```
The 10 most similar words to "juliet" using cosine-similarity on tf-idf matrix are:
1: juliet; 0.9999999999999998
2: benedicite; 0.9845619150110774
3: procures; 0.9845619150110774
4: ghostly; 0.8975502147028126
5: lucio; 0.8468242699368735
6: capulets; 0.8199600222390604
7: tybalt; 0.8199600222390604
8: romeo; 0.8199600222390604
9: montagues; 0.8199600222390604
10: stirreth; 0.8199600222390603
```

PPMI matrix

```
The 10 most similar words to "juliet" using cosine-similarity on PPMI matrix are:
1: juliet; 1.0
2: capulet; 0.1919614018467302
3: vauntingly; 0.14946850506091333
4: barnardine; 0.14096892942519404
5: provost; 0.1370683855294029
6: tybalt; 0.13685330053902423
7: montague; 0.13219117094886065
8: mercutio; 0.12983590962299896
9: romeo; 0.12548822860469921
10: stricken; 0.12338733674236771
```

For the word "ursula" top 10 similar word based on:

Tf-idf matrix

```
The 10 most similar words to "ursula" using cosine-similarity on tf-idf matrix are:
1: ursula; 1.0
2: dully; 0.9415674072270229
3: rhyming; 0.9415674072270229
4: alley; 0.9039596641294985
5: thirdly; 0.9039596641294985
6: benedick; 0.9039596641294985
7: beatrice; 0.9039596641294985
8: pedro; 0.9039596641294985
9: borachio; 0.9039596641294985
10: slandered; 0.9039596641294985
```

PPMI matrix

```
The 10 most similar words to "ursula" using cosine-similarity on PPMI matrix are:
1: ursula; 1.0000000000000002
2: leonato; 0.3165480067242844
3: benedick; 0.28301874413281863
4: conrade; 0.26780341691839504
5: margaret; 0.25691470263619753
6: beatrice; 0.25623177806740766
7: borachio; 0.2491165168193321
8: don; 0.2468159649424851
9: accusing; 0.2446761631827713
10: pedro; 0.23729425051182632
```

For the word “sicily” top 10 similar word based on:

Tf-idf matrix

```
The 10 most similar words to "sicily" using cosine-similarity on tf-idf matrix are:
1: sicily; 1.0000000000000002
2: nilus; 0.883453089087945
3: laden; 0.8795731614100674
4: blemishes; 0.8795731614100674
5: urgent; 0.8610056494914924
6: daintily; 0.8610056494914924
7: piteously; 0.8610056494914924
8: prognostication; 0.8610056494914924
9: castaway; 0.8610056494914924
10: nests; 0.8610056494914924
```

PPMI matrix

```
The 10 most similar words to "sicily" using cosine-similarity on PPMI matrix are:
1: sicily; 0.9999999999999998
2: aetna; 0.24574780537166444
3: goer; 0.21202793972201636
4: mackerel; 0.20436831300066793
5: anvil; 0.19847660763443473
6: coals; 0.17843253643671825
7: manhoods; 0.17707830386330262
8: possitable; 0.17459337252769694
9: personage; 0.16496516650129187
10: predecessors; 0.1563705560406094
```

Q6.

TF-IDF provides more relevant and meaningful word similarities than the raw term-document matrix. TF-IDF effectively ranks words based on importance and relevance, rather than just frequency. The term-document matrix is noisy and includes irrelevant words simply because they co-occur in documents. For TF-IDF results, Words like "ghostly" and "benedicite" may refer to the religious or spiritual themes in *Romeo and Juliet*. Lucio appears, which is a Shakespearean character but not from *Romeo and Juliet*—this suggests some overlap in Shakespearean themes rather than strict plot connections.

Q7.

PPMI provides better word similarities than the raw term-context matrix because it adjusts for word frequency biases and highlights meaningful associations. PPMI correctly ranks character names and important thematic words higher, whereas

the term-context matrix mixes in characters from other plays due to general co-occurrence patterns. From the PPMI result, stronger connections to *Romeo and Juliet* emerge, such as "Capulet" (Juliet's family name), "Tybalt" (Juliet's cousin), "Montague" (Romeo's family), "Mercutio" (Romeo's close friend), Romeo" (Juliet's love interest).

Q8.
TF-IDF (on the term-document matrix) produces more relevant words compared to the raw term-document frequency matrix. PPMI (on the term-context matrix) captures deeper semantic relationships than the raw term-context matrix. Both TF-IDF and PPMI prioritize important words by down weighting frequent but less informative words.
TF-IDF improves this by prioritizing unique associations as word "Juliet" was ranked similar to "Capulet", "Tybalt", "Romeo", "Montague", and "Mercutio", which are meaningful character connections.

PPMI Improves this by filtering out random co-occurrences as word "Juliet" was ranked similar to "Capulet", "Tybalt", "Montague", "Mercutio", and "Romeo", which are all characters in the play.

Part II.

Q1.
For the word "man" top 10 similar word based on PPMI:

```
The 10 most similar words to "man" using cosine-similarity on PPMI matrix are:
1: man; 0.9999999999999998
2: is; 0.3708996249698417
3: woman; 0.3436583176038O1
4: a; 0.3229108114174089
5: his; 0.29759648525737825
6: person; 0.2510788360624903
7: guy; 0.22578986855221014
8: young; 0.2250049719490408
9: boy; 0.22266370569010382
10: he; 0.2210719576309506
```

For the word "woman" top 10 similar word based on PPMI:

```
The 10 most similar words to "woman" using cosine-similarity on PPMI matrix are:
1: woman; 1.0
2: man; 0.3436583176038O1
3: her; 0.31291259073675687
4: is; 0.31194466849128444
5: a; 0.2563226372527975
6: lady; 0.24791175439333735
7: young; 0.24474668796263407
8: girl; 0.23183313989764404
9: and; 0.22365077503088426
10: while; 0.21341498422609875
```

For the word "boy" top 10 similar word based on PPMI:

```
The 10 most similar words to "boy" using cosine-similarity on PPMI matrix are:
1: boy; 0.9999999999999999
2: girl; 0.3174386661904272
3: little; 0.2771319661809781
4: child; 0.2704497730501622
5: young; 0.2669861016950528
6: is; 0.2231495952243061
7: man; 0.22266370569010382
8: his; 0.21028070492643333
9: woman; 0.20232662766655393
10: kid; 0.18958897518006596
```

For the word "girl" top 10 similar word based on PPMI:

```
The 10 most similar words to "girl" using cosine-similarity on ppmi matrix are:
1: girl; 1.0
2: little; 0.32011874551530095
3: boy; 0.3174386661904272
4: young; 0.27148696376018805
5: her; 0.2603310217945017
6: child; 0.23787134277618197
7: woman; 0.23183313989764404
8: is; 0.22500428572453998
9: baby; 0.1906917585655798
10: man; 0.19004813623665345
```

Q2.

"Man" is associated with "his", while "woman" is associated with "her". This reflects gendered language norms, reinforcing possessive pronoun stereotypes (e.g., "his" career vs. "her" beauty). This isn't inherently harmful but shows traditional gender role reinforcement in linguistic data.

"Man" is linked to "guy" while "woman" is linked to "lady". The male term "guy" is casual and neutral, while the female term "lady" carries politeness, formality, or old-fashioned undertones. This reflects gendered connotations in language where women are more often described with terms implying etiquette or social behaviour.

"Boy" and "Girl" are strongly linked to "young", "child", and "little". This makes sense since both terms are age-related.

"Boy" is associated with "man" (0.2227), while "girl" is associated with "baby" (0.1907). The boy and man progression reflects masculine maturity expectations. The girl and baby association reflects infantilization, which can reinforce perceptions of women as dependent or childlike.

Q3.

Associated words:

Man: guy

Woman: lady

First order similarity:

A man worth any woman, overbuys me.

A woman peering over the edge of a boat while a man walks behind her.

Second order similarity:

A guy in a green shirt is trying to tag out a man in a red shirt while playing baseball.

A young woman gives a lady a hug after graduation.

Associated words:

Boy: man

Girl: woman

First order similarity:

A boy skateboards on a cement wall.

A man and a boy go hiking through the forest.

Second order similarity:

A birds eye view of a young girl on a playground.

A man, woman, girl, and boy sit on a concrete bench by a beach.

Associated Words:

Doctor: he

Nurse: she

First order similarity:

Two doctors perform an operation.

A doctor in blue scrubs is monitoring a screen as he works along side other medical personnel.

Second order similarity:

Two nurses posing for a picture.

A nurse has gotten her clothes wet as she bathes a baby.

Associated Words:

Engineer: his

Teacher: her

First order similarity:

A group of railroad engineers are repairing the tracks in the tunnel.

A civil engineer with a black hat is drawing in the sand with his son.

Second order similarity:

Children gather around their teacher

A teacher is weaving a tale of adventure and delight for her students.

The association of "doctor" with "he" and "nurse" with "she" reflects traditional gender roles in healthcare professions. Such associations can reinforce stereotypes that doctors are male and nurses are female, which may contribute to gender bias in professional settings. The linkage of "engineer" with "he" and "teacher" with "she" mirrors societal stereotypes associating men with technical professions and women with educational roles. These associations can perpetuate gender biases in career expectations and opportunities.