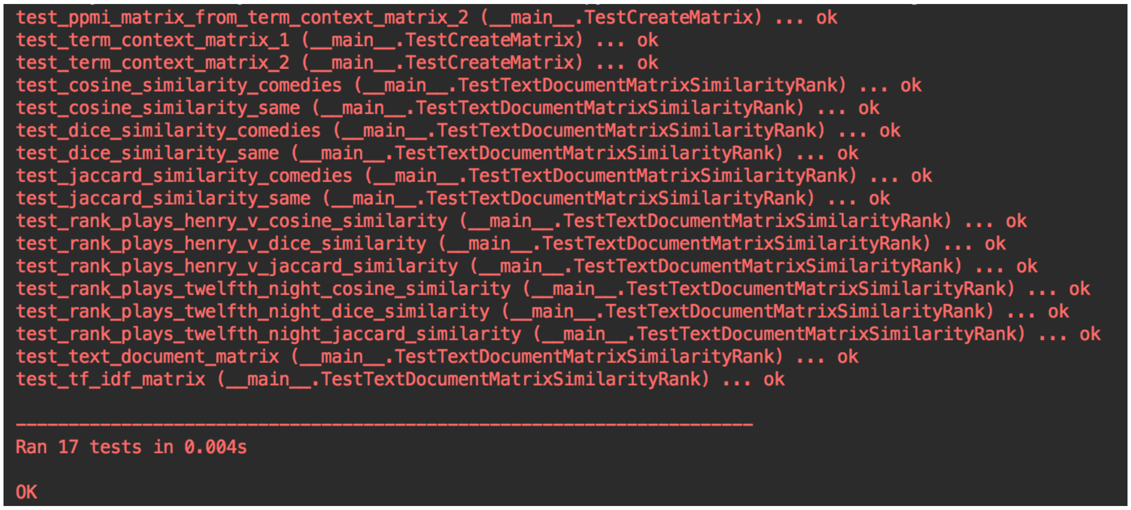
CIS530 Homework 3:
Vector Space Models

Maria Kustikova (mkust) and Devanshu Jain (devjain)

Due Date: January 31, 2018

1 Testing

In order to ensure that the implementation of functions (create_term_document_matrix, create_term_context_matrix, create_PPMI_matrix, compute_tf_idf_matrix, compute_cosine_similarity, compute_jaccard_similarity, compute_dice_similarity, rank_plays, rank_words) is valid we have written unit tests for each of the required functions that follow examples from Chapter 15 of the textbook [4]. Please note that the unit tests are included in the submission under the name of *tests.py*. Figure 1 displays the outcome of running 17 unit tests.



```
test_ppmi_matrix_from_term_context_matrix_2 (__main__.TestCreateMatrix) ... ok
test_term_context_matrix_1 (__main__.TestCreateMatrix) ... ok
test_term_context_matrix_2 (__main__.TestCreateMatrix) ... ok
test_cosine_similarity_comedies (__main__.TestTextDocumentMatrixSimilarityRank) ... ok
test_cosine_similarity_same (__main__.TestTextDocumentMatrixSimilarityRank) ... ok
test_dice_similarity_comedies (__main__.TestTextDocumentMatrixSimilarityRank) ... ok
test_dice_similarity_same (__main__.TestTextDocumentMatrixSimilarityRank) ... ok
test_jaccard_similarity_comedies (__main__.TestTextDocumentMatrixSimilarityRank) ... ok
test_jaccard_similarity_same (__main__.TestTextDocumentMatrixSimilarityRank) ... ok
test_rank_plays_henry_v_cosine_similarity (__main__.TestTextDocumentMatrixSimilarityRank) ... ok
test_rank_plays_henry_v_dice_similarity (__main__.TestTextDocumentMatrixSimilarityRank) ... ok
test_rank_plays_henry_v_jaccard_similarity (__main__.TestTextDocumentMatrixSimilarityRank) ... ok
test_rank_plays_twelfth_night_cosine_similarity (__main__.TestTextDocumentMatrixSimilarityRank) ... ok
test_rank_plays_twelfth_night_dice_similarity (__main__.TestTextDocumentMatrixSimilarityRank) ... ok
test_rank_plays_twelfth_night_jaccard_similarity (__main__.TestTextDocumentMatrixSimilarityRank) ... ok
test_text_document_matrix (__main__.TestTextDocumentMatrixSimilarityRank) ... ok
test_tf_idf_matrix (__main__.TestTextDocumentMatrixSimilarityRank) ... ok

----------------------------------------------------------------------
Ran 17 tests in 0.004s

OK
```

Figure 1: The result of running unit test for required functions

2 Rank Plays

For ranking of plays, we used term-document and tf-idf matrices and experimented with three similarity measure.

2.1 Term-Document Matrix

The 10 most similar plays to 'Hamlet' using three similarity measures can be seen in Table 1. As can be seen from Table 1 dice and jaccard similarity measures give the same ranking (at least in the top 10). This may be justified due to almost similar formulae used to compute the similarity. The Cosine similarity measure, on the other hand, is quite different to dice and jaccard similarities. For instance, 'Henry VII' is ranked as number 1 based on the cosine similarity, but is ranked as 8 for dice and jacard similarities. Moreover, ‘Othello’ is ranked as number 1 based on the dice and jacard similarities, but ‘Othello’ is not even present in the too 10 of the Cosine similarity However, we can see all the three measures have 6 plays that are the same although the order is a little different. Based on the canonical grouping plays [1] ‘Hamlet’ is categorized into tragedy and we can see 7 plays in the same category (Troilus and Cressida, Macbeth, Hamlet, King Lear, Othello, Antony and Cleopatra, Cymbeline) appear in top 10 most similar plays across the similarity metrics. However, ‘Alls well that ends well’ is categorized as a comedy, but appears in Table 1. Without analyzing each of the Shakespeare’s plays in thorough details and based on the data we collected, we believe our rankings are consistent with canonical representation seen in [1].

Rank	Cosine Similarity	Jaccard Similarity	Dice Similarity
1	Henry VIII (0.9738)	Othello (0.5769)	Othello (0.7317)
2	A Winters Tale (0.9726)	Cymbeline (0.569)	Cymbeline (0.7253)
3	Troilus and Cressida (0.9715)	A Winters Tale (0.5605)	A Winters Tale (0.7184)
4	Cymbeline (0.9713)	King Lear (0.5586)	King Lear (0.7168)
5	King Lear (0.9713)	Richard III (0.5555)	Richard III (0.7142)
6	Alls well that ends well (0.971)	Coriolanus (0.5458)	Coriolanus (0.7062)
7	Richard III (0.9696)	Troilus and Cressida (0.5407)	Troilus and Cressida (0.7019)
8	Pericles (0.9693)	Henry VIII (0.5379)	Henry VIII (0.6995)
9	macbeth (0.9687)	Alls well that ends well (0.5369)	Alls well that ends well (0.6987)
10	Loves Labours Lost (0.9683)	Antony and Cleopatra (0.5332)	Antony and Cleopatra (0.6955)

Table 1: The 10 most similar plays to ‘Hamlet’ using three similarity measures on term-document matrix

2.2 tf-idf Matrix

As can be seen from Table 2 dice and jaccard similarity measures give the same ranking (at least in the top 10) again. This may be justified again due to almost similar formulae used to compute the similarity.

Rank	Cosine Similarity	Jaccard Similarity	Dice Similarity
1	Henry V (0.0254)	King Lear (0.09)	King Lear (0.1651)
2	King John (0.0217)	Othello (0.0894)	Othello (0.1641)
3	Alls well that ends well (0.0197)	Cymbeline (0.0892)	Cymbeline (0.1638)
4	Henry VIII (0.0194)	Henry VIII (0.083)	Henry VIII (0.1533)
5	Richard II (0.0181)	A Winters Tale (0.0828)	A Winters Tale (0.153)
6	King Lear (0.0169)	Troilus and Cressida (0.0823)	Troilus and Cressida (0.152)
7	Richard III (0.0166)	Alls well that ends well (0.0816)	Alls well that ends well (0.1509)
8	Henry VI Part 2 (0.0159)	King John (0.0813)	King John (0.1504)
9	A Winters Tale (0.0156)	Richard III (0.0807)	Richard III (0.1493)
10	Henry IV (0.0153)	macbeth (0.0784)	macbeth (0.1454)

Table 2: The 10 most similar plays to ‘Hamlet’ using three similarity measures on tf-idf matrix

3 Rank Words

For ranking of words, we used term-context and PPMI matrices with three similarity measure.

3.1 Term-Context Matrix

The 10 most similar words to ‘death’ using three similarity measures for just term-context matrix can be seen in Table 3. As in Section 2, dice and jaccard similarities measures give the same rankings (at least in the top 10) and may arise due to almost similar formulae used to compute the similarity. Based on the initial observations of Table 3 dice and jaccard similarities give a more sensible results, since we expected the words, such as ‘blood’, ‘heart’, ‘heaven’, ‘life’ to occur. Unlike in Section 2, there are less words that occur across all similarity measures. Something we found interesting and was emphasized during the previous lecture is that distributional models like this one will sometimes identify antonyms to be quite similar. This can be seen in this Table as the word ‘life’ is ranked number 1 for jaccard and dice similarity.

Rank	Cosine Similarity	Jaccard Similarity	Dice Similarity
1	death (1.0)	death (1.0)	death (1.0) (0.5569)
2	fortune (0.912)	life (0.3859)	life (0.5569)
3	nature (0.8815)	honour (0.3583)	honour (0.5276)
4	virtue (0.8534)	name (0.347)	name (0.5152)
5	sorrow (0.8466)	blood (0.339)	blood (0.5063)
6	england (0.8455)	heart (0.3372)	heart (0.5043)
7	blood (0.8454)	father (0.3289)	father (0.495)
8	name (0.8367)	son (0.325)	son (0.4905)
9	wit (0.835)	time (0.318)	time (0.4826)
10	the (0.8283)	heaven (0.3135)	heaven (0.4774)

Table 3: The 10 most similar words to the ‘death’ using three similarity measures on term-context matrix

3.2 PPMI Matrix

Table 4 contains 10 most similar words to the ‘death’ using three similarity measures on PPMI matrix, which is one of the weighting schemes and can be a better way of measuring the the association between words. From the initial observations of this matrix, we believe there is a similar amount words that ‘make sense’ in comparison to just using raw frequencies in Table 3. The words we identify as making sense are ‘honour’, ‘die’, ‘life’, ‘timeless’, ‘fear’. However, we expected ‘blood’, ‘heart’, ‘heaven’ to also occur in top 10. Interesting thing to not is that we can see that some of the words ranked in top 10 are not even words and are result of default parsing method.

Rank	Cosine Similarity	Jaccard Similarity	Dice Similarity
1	death (1.0)	death (1.0)	death (1.0)
2	humphrey (0.0824)	die (0.0517)	die (0.0983)
3	die (0.0808)	life (0.0501)	life (0.0955)
4	timeless (0.0795)	till (0.0476)	till (0.0908)
5	dearth (0.0787)	whose (0.0464)	whose (0.0887)
6	to (0.076)	fear (0.0456)	fear (0.0872)
7	by (0.0744)	honour (0.0451)	honour (0.0863)
8	s (0.0743)	father (0.0435)	father (0.0834)
9	of (0.0741)	doth (0.0427)	doth (0.0819)
10	thy (0.0739)	any (0.0423)	any (0.0811)

Table 4: The 10 most similar words to the ‘death’ using three similarity measures on PPMI matrix

4 Extra Credit: Character Analysis

We have chosen to follow the specification guidelines of the character analysis with extra addition of a case study character comparison. Please note that:

- blue is used throughout this report to portray a female character
- Cosine Similarity is used to identify similarity between characters, although other similarity metrics could have been used
- In this analysis we used term-character matrix which is classified as an optional fun extra credit option

4.1 Case study: ‘Hamlet’

We first started by selecting a specific play and looking at the similarity of the characters within this particular play. Due to our familiarity of the plot of ‘Hamlet’ we decided to pick it as a case study. Table 5 displays number of lines each character has for ‘Hamlet’ based on the given data and a default parsing function. It can be seen from this table that the character frequency varies a lot, for instance the main character Hamlet has 1582 line, whereas a secondary character ‘Servant’ has only 1 line. Table 6 shows the most similar characters in the play and Table 7 shows the least similar characters. Without delving into the plot of a play it is hard to see a particular pattern in both tables other than that Table 7 has some values of cosine similarity equal nearly to 0, which clearly has to do with not enough lines for a particular character. For example, ‘Servant’ has one line and it is: “Sailors, sir: they say they have letters for you”. ‘CYMBELINE’ has three lines and it is: “ACT I SCENE I”, “ Elsinore”, “A platform before the castle. FRANCISCO at his post. Enter to him BERNARDO”. It makes sense that ‘Servant’ and ‘CYMBELINE’ have cosine similarity set to 0, as their lines are very specific and limited to a particular context.

If we consider the plot of the play and the overall theme, then it makes sense that the main characters of ‘Hamlet’ are similar in Table 7. The whole play is a tragedy and is about revenge, bitter and melancholy that is revolving around the main protagonist prince ‘Hamlet’. It is interesting to note that protagonist Hamlet and his antagonist ‘King Claudius’ are ranked quite similar to each other and are third in Table 7.

Character	Number of Lines
HAMLET	1582
KING CLAUDIUS	594
LORD POLONIUS	370
HORATIO	303
LAERTES	216
OPHELIA	187
QUEEN GERTRUDE	166
First Clown	99
Ghost	96
ROSENCRANTZ	96
MARCELLUS	69
GUILDENSTERN	55
First Player	52
OSRIC	48
Player King	45
BERNARDO	39
Player Queen	31
PRINCE FORTINBRAS	30
Gentleman	24
VOLTIMAND	23
Second Clown	19
REYNALDO	15
Captain	13
First Priest	13
FRANCISCO	12
LUCIANUS	7
Lord	7
First Ambassador	6
All	5
First Sailor	5
Messenger	5
Prologue	4
Danes	4
CYMBELINE	3
Servant	1

Table 5: Number of lines each character has 'Hamlet'

Character 1	Character 2	Cosine Similarity
HAMLET	LORD POLONIUS	0.9355
HAMLET	LAERTES	0.9329
HAMLET	KING CLAUDIUS	0.9324
HAMLET	HORATIO	0.9154
LORD POLONIUS	LAERTES	0.9138
KING CLAUDIUS	LORD POLONIUS	0.8985
LORD POLONIUS	HORATIO	0.8971
HORATIO	LAERTES	0.8936
KING CLAUDIUS	HORATIO	0.8826
HAMLET	QUEEN GERTRUDE	0.8823
LORD POLONIUS	OPHELIA	0.8811
KING CLAUDIUS	QUEEN GERTRUDE	0.8739

Table 6: Most similar characters in 'Hamlet'

Character 1	Character 2	Cosine Similarity
LUCIANUS	Servant	0.0
All	Danes	0.0
All	Servant	0.0
CYMBELINE	Servant	0.0
REYNALDO	Prologue	0.0124
REYNALDO	Servant	0.0185
REYNALDO	LUCIANUS	0.022
Ghost	Servant	0.0222
REYNALDO	All	0.028
LUCIANUS	CYMBELINE	0.0313
Gentleman	Servant	0.0355
VOLTIMAND	Servant	0.0392

Table 7: Least similar characters in ‘Hamlet’

4.2 Most and least ‘main’ characters

The total number of unique characters in the given data is 1328. In order to remove noise from the data seen in Section 4.1 we decided to introduce the concept of the ‘main’ and ‘secondary’ characters. Having experimented with various thresholds (5 - 20), we decided to select a static number of 10 characters from each of the Shakespeare’s play based on the number of lines a character has in the descending order. Table 8 displays the actual data we gathered for each character in each play. After the initial pre-filtering is applied, we were left with 360 characters (36 plays × 10 characters per play), which made it possible to perform a combination (360 choose 2) and compare each character to another.

Name of the play	List of (character names, number of lines) in desc order
Henry IV	{FALSTAFF,654}, {HOTSPUR,583}, {PRINCE HENRY,582}, {KING HENRY IV,355} ...
Alls well that ends well	{HELENA,498}, {KING,403}, {PAROLLES,387}, {COUNTESS,298}, {LAFEU,287} ...
Loves Labours Lost	{BIRON,647}, {FERDINAND,323}, {PRINCESS,297}, {ADRIANO DE ARMADO,281} ...
⋮	⋮
Pericles	{PERICLES,645}, {GOWER,298}, {MARINA,189}, {SIMONIDES,178} ...
Titus Andronicus	{TITUS ANDRONICUS,768}, {AARON,375}, {MARCUS ANDRONICUS,277} ...

Table 8: ‘Main’ character selection based on the number of lines

4.3 Most similar ‘main’ characters

Table 9 displays most similar ‘main’ characters across all plays. From what we can see the cosine similarity is very high in comparison to the similarities observed in Section 4.1. Without analyzing each play thoroughly, it can be seen that some of the character similarities make sense. For instance, ‘King Richard II’ and ‘King Henry V’, or ‘King Richard II’ and ‘King John’, or ‘King Henry VII’ and ‘Cardinal Wolsey’ have have high similarity scores. From the gathered data it seems that there is a difference between vocabulary used by a royal person as supposed to a non-royal person. This was our initial intuition/hypothesis which we believe will still hold if all the plays are analyzed in thorough details.

Character 1	Character 2	Cosine Similarity
HAMLET (Hamlet)	PORTIA (Merchant of Venice)	0.9593
KING RICHARD II (Richard II)	KING HENRY V (Henry V)	0.9591
BRUTUS (Julius Caesar)	CASSIUS (Julius Caesar)	0.9529
ROSALIND (As you like it)	IAGO (Othello)	0.9522
HAMLET (Hamlet)	MACBETH (macbeth)	0.9519
HAMLET (Hamlet)	IAGO (Othello)	0.9508
HOTSPUR (Henry IV)	HAMLET (Hamlet)	0.9504
HAMLET (Hamlet)	BASTARD (King John)	0.9501
PAROLLES (Alls well that ends well)	HAMLET (Hamlet)	0.9501
HAMLET (Hamlet)	GLOUCESTER (Richard III)	0.9497
MACBETH (macbeth)	KING HENRY V (Henry V)	0.9497
BIRON (Loves Labours Lost)	HAMLET (Hamlet)	0.9486
CORIOLANUS (Coriolanus)	POSTHUMUS LEONATUS (Cymbeline)	0.9485
CARDINAL WOLSEY (Henry VIII)	GLOUCESTER (Richard III)	0.9483
KING RICHARD II (Richard II)	KING JOHN (King John)	0.9483
PORTIA (Merchant of Venice)	BASSANIO (Merchant of Venice)	0.9481
HAMLET (Hamlet)	CARDINAL WOLSEY (Henry VIII)	0.9476
MARK ANTONY (Antony and Cleopatra)	MACBETH (macbeth)	0.9473
LEONTES (A Winters Tale)	OTHELLO (Othello)	0.9471
KING HENRY VIII (Henry VIII)	CARDINAL WOLSEY (Henry VIII)	0.9467
ROMEO (Romeo and Juliet)	JULIET (Romeo and Juliet)	0.9465

Table 9: ‘Main’ characters that are most similar

4.4 Least similar ‘main’ characters

Table 10 displays the least similar ‘main’ characters across all plays. It is quite hard to derive certain pattern in this table. We can see here is that plays *Timon of Athens* and *Measure for measure* are common occurrence in the table. According to the wikipedia [3], these plays are classified as *problem plays* - which are characterized by their complex tone. This may be a reason why the characters from these plays have such a low similarity to other characters from other plays. Write something more?

Character 1	Character 2	Cosine Similarity
AGRIPPA (Antony and Cleopatra)	LUCIUS (Julius Caesar)	0.4073
Second Senator (Timon of Athens)	LUCIUS (Julius Caesar)	0.4186
First Senator (Timon of Athens)	LUCIUS (Julius Caesar)	0.4193
LUCIUS (Julius Caesar)	KING OF FRANCE (Henry V)	0.4203
Second Senator (Timon of Athens)	MARIANA (Measure for measure)	0.4224
First Citizen (Coriolanus)	MARIANA (Measure for measure)	0.4251
MARIANA (Measure for measure)	KING OF FRANCE (Henry V)	0.4356
AGRIPPA (Antony and Cleopatra)	MARIANA (Measure for measure)	0.4377
First Senator (Timon of Athens)	MARIANA (Measure for measure)	0.4454
AGRIPPA (Antony and Cleopatra)	Porter (macbeth)	0.4471
Second Senator (Timon of Athens)	LUCETTA (Two Gentlemen of Verona)	0.449
MARIANA (Measure for measure)	PANTHINO (Two Gentlemen of Verona)	0.4572
MARIANA (Measure for measure)	Chorus (Henry V)	0.4582
SLENDER (Merry Wives of Windsor)	Second Senator (Timon of Athens)	0.4583
Second Senator (Timon of Athens)	DROMIO OF EPHESUS (A Comedy of Errors)	0.4605
First Senator (Coriolanus)	MARIANA (Measure for measure)	0.4609
MARIANA (Measure for measure)	CANTERBURY (Henry V)	0.4676
LADY PERCY (Henry IV)	AGRIPPA (Antony and Cleopatra)	0.4682
QUINCE (A Midsummer nights dream)	LUCIUS (Julius Caesar)	0.4684
PRINCE (Romeo and Juliet)	LUCIUS (Julius Caesar)	0.4685
AGRIPPA (Antony and Cleopatra)	SILVIUS (As you like it)	0.4699

Table 10: ‘Main’ characters that are least similar

4.5 Female vs Male character comparison

Table 11 displays a first female vs male comparison we did as a part of our analysis. For this task we have chosen a different approach to the one described in Section 4.2. Since the given data does not contain a label indicating the gender of a character, we had done a manual job of separating female and male characters based on the list of notable female characters in Shakespeare’s plays [2]. To be more specific, we hard-coded a list of 37 female characters and every character that is in this list are considered a female character, and the rest are assumed to be male by default. Hence, the female character make up to 2.7% of all the unique characters in each play. As can be seen from Table 11 the average cosine similarity between female characters is high with little variance. This indicates that notable female characters are quite similar to each other. However, the average cosine similarity of the male only characters is small in comparison with a higher standard deviation. This could be because we select only ‘main’ female characters but the rest are male characters, which makes such comparison unfair as we saw previously in Section 4.1. Some of the male characters will have a very low number of lines, which will bring down the average of the cosine similarity between male characters.

To overcome this problem we decided to match the ‘main’ male character for each female character on the list [2]. We thus restricted our data to 37 female characters and 37 male characters. Table 12 shows the result of running the same type of comparison as in Table 11, but with different male data. From Table 12 we can see that separating female and male characters does not make that much of a difference, and in fact comparing the mix of both characters gives a higher average than comparing female characters. This was not something we expected and it would be an interesting future work to see the reason behind such numbers. As a side note, throughout the report we highlighted female characters with blue and we could not derive an obvious pattern which is consistent with the results of the comparison we performed in this section.

Category	Average	Median	Standard Deviation
Female and Female	0.7837	0.826	0.1186
Male and Male	0.3643	0.3443	0.2291
(Female and Male) or (Male and Female)	0.5258	0.5517	0.2314

Table 11: Female vs Male Comparison 1

Category	Average	Median	Standard Deviation
Female and Female	0.7837	0.826	0.1186
Male and Male	0.8727	0.8778	0.0415
(Female and Male) or (Male and Female)	0.8229	0.8452	0.0834

Table 12: Female vs Male Comparison 2

5 Clustering Plays

We clustered the vector representations (Term-Document Matrix) of the plays using K-Means algorithm into 3 clusters (Tragedy, Comedy and History):

5.1 Cluster 1

1. Henry IV
2. Antony and Cleopatra
3. Coriolanus
4. Hamlet
5. A Winters Tale
6. Henry VI Part 2
7. Henry VIII
8. Richard III
9. Henry V
10. Troilus and Cressida
11. Henry VI Part 3
12. Othello
13. Cymbeline
14. King Lear

5.2 Cluster 2

1. A Midsummer nights dream
2. Richard II
3. King John
4. macbeth

5. Timon of Athens
6. The Tempest
7. Julius Caesar
8. A Comedy of Errors
9. Henry VI Part 1
10. Pericles
11. Titus Andronicus

5.3 Cluster 3

1. Alls well that ends well
2. Loves Labours Lost
3. Taming of the Shrew
4. Merry Wives of Windsor
5. Romeo and Juliet
6. As you like it
7. Measure for measure
8. Two Gentlemen of Verona
9. Much Ado about nothing
10. Twelfth Night
11. Merchant of Venice

We can see above that the clustering method makes a lot of confusion between plays belonging to the categories: *History* and *Tragedy.* This is also consistent with the data we have seen in Table 1 since many of the historical plays were similar to 'Hamlet', which is considered to be tragedy.

6 Extra Credit: Kendall's Tau

In this section, we used the Simlex-999 dataset to compute the kendall's tau coefficient between our measures of similarity and the human judgements. The dataset consisted of 999 pairs of words scored on a range of [1,10], with 10 denoting the highest order of similarity between the words. We filtered 687 pairs out of them by choosing only the pairs whose words were present in the Shakespeare vocabulary. We experimented with cosine similarity measure on all types of matrix and achieved the same Kendall's tau coefficient of 0.09913 across them.

References

[1] Canonical plays. https://en.wikipedia.org/wiki/Shakespeare%27s_plays#Canonical_plays.

[2] Notable female characters. https://en.wikipedia.org/wiki/Women_in_Shakespeare%27s_works.

[3] Shakespearean problem play. https://en.wikipedia.org/wiki/Shakespearean_problem_play.

[4] Dan Jurafsky and James H. Martin. Speech and language processing. https://web.stanford.edu/~jurafsky/slp3/15.pdf.