CIS 530 - HW3 - Spring 2019

Authors:

Ilayda Ipek Onur

Srinivas Suri

1. How do I know if my rankings are good for the plays?

a. Using Term Document Matrix

We have compared ‘Merry Wives of Windsor’; a comedy novel written by Shakespeare with the top 10 novels it is similar to. Our Similarities are reported below:

S.no	Cosine	Cosine similarity	Jaccard	Jaccard Similarity	Dice	Dice Similarity
1	Twelfth Night	0.962169	Much Ado about nothing	0.547375	Much Ado about nothing	0.707489
2	Much Ado about nothing	0.959597	Twelfth Night	0.539325	Twelfth Night	0.700729
3	As you like it	0.958945	As you like it	0.533452	As you like it	0.695753
4	Alls well that ends well	0.957455	Taming of the Shrew	0.529853	Taming of the Shrew	0.692685
5	Taming of the Shrew	0.956370	Alls well that ends well	0.529193	Alls well that ends well	0.692120
6	Merchant of Venice	0.950144	Merchant of Venice	0.520350	Merchant of Venice	0.684513
7	Othello	0.949623	Measure for measure	0.517154	Measure for measure	0.681742
8	Measure for measure	0.941885	Othello	0.507620	Othello	0.673406

9	A Winters Tale	0.940986	A Winters Tale	0.493742	A Winters Tale	0.661080
10	Two Gentlemen of Verona	0.936886	King Lear	0.487117	King Lear	0.655116

Our Observations:

- We observe that 9 out of 10 novels reportes are indeed in the category of comedies of Shakespeare according to (https://en.wikipedia.org/wiki/Shakespeare%27s_plays#Canonical_plays) .
- We observed that Jaccard and Dice similarity have the same top 10 novels for ‘Merry Wives of Windsor’;This can be accounted to the similarity between Dice and Jaccard scores.
- We have observed that the only novel that does not belong to Comedies section is Othella. Although we are not sure why, we found it interesting to report this observation.

b. Using Tf-Idf Matrix

S.no	Cosine	Cosine similarity	Jaccard	Jaccard Similarity	Dice	Dice Similarity
1	Henry V	0.107132	Henry V	0.083738	Henry V	0.154536
2	Henry IV	0.094073	Henry IV	0.082986	Henry IV	0.153253
3	Much Ado about nothing	0.080636	Much Ado about nothing	0.082070	Much Ado about nothing	0.151691
4	King Lear	0.079816	King Lear	0.080429	King Lear	0.148883
5	Hamlet	0.078445	Hamlet	0.079664	Hamlet	0.147573
6	Loves Labours Lost	0.078211	Othello	0.078945	Othello	0.146337
7	A Winters Tale	0.077877	As you like it	0.078149	As you like it	0.144969
8	Measure for measure	0.075633	A Winters Tale	0.078110	A Winters Tale	0.144902

9	Troilus and Cressida	0.075375	Romeo and Juliet	0.076734	Romeo and Juliet	0.142531
10	Romeo and Juliet	0.075319	Alls well that ends well	0.075871	Alls well that ends well	0.141040

Our Observations:

- We observe again that Jaccard and Dice Similarity have given the same top 10 results for ‘Merry Wives of Windsor’.
- We observe that ‘Merry Wives of Windsor’ is more closer to Henry V and Henry IV than ‘Twelfth night’ or ‘Much Ado About nothing’. It’s interesting to see this when we changed the similarity to tf-idf matrix. On further Googling (https://www.shmoop.com/merry-wives-of-windsor/ , https://www.tandfonline.com/doi/abs/10.1080/00144940.1989.9933946), we found that both these novels are indeed related with each other and tf-idf results reported better similarities than just the term-document matrix.

Word Rankings

We looked at two characters

For ‘juliet’ top 10 similar words are

TERM-CONTEXT MATRIX window size = 4

Cosine	**Jaccard**	**Dice**
lucius	nurse	nurse
warwick	silvia	silvia
buckingham	proteus	proteus
brutus	marcus	marcus
others	valentine	valentine
thy	paris	paris
cold	othello	othello
montague	beatrice	beatrice
fortune	montague	montague
officers	leonato	leonato

PPMI MATRIX   k=1 smoothing

Cosine	**Jaccard**	**Dice**
lady	nurse	nurse
mistress	romeo	romeo
romeo	silvia	silvia
nurse	paris	paris
capulet	proteus	proteus
s	hamlet	hamlet
and	claudio	claudio

page	leonato	leonato
come	anne	anne
here	marcus	marcus

For ‘hamlet’ top 10 similar words are

TERM-CONTEXT MATRIX window size = 4

Cosine	**Jaccard**	**Dice**
timon	nurse	nurse
talbot	timon	timon
fortune	angelo	angelo
good	talbot	talbot
angelo	meat	meat
sweet	clifford	clifford
death	prisoner	prisoner
warwick	romeo	romeo
and	harry	harry
dead	chance	chance

PPMI MATRIX window size = 4

Cosine	**Jaccard**	**Dice**
lord	nurse	nurse
king	gracious	gracious
queen	timon	timon
come	romeo	romeo
my	sister	sister

but	clarence	clarence
and	laertes	laertes
now	cousin	cousin
son	sight	sight
what	cassio	cassio

Observations:

For TC_matrix, Jaccard and Dice performed better than Cosine similarity. However, for PPMI, both extracted meaningful words in different ways.

For example, for both ‘juliet’ and ‘hamlet’, cosine similarity on PPMI resulted in words describing the character rather than bringing other characters -such as nurse- who are associated with the character.

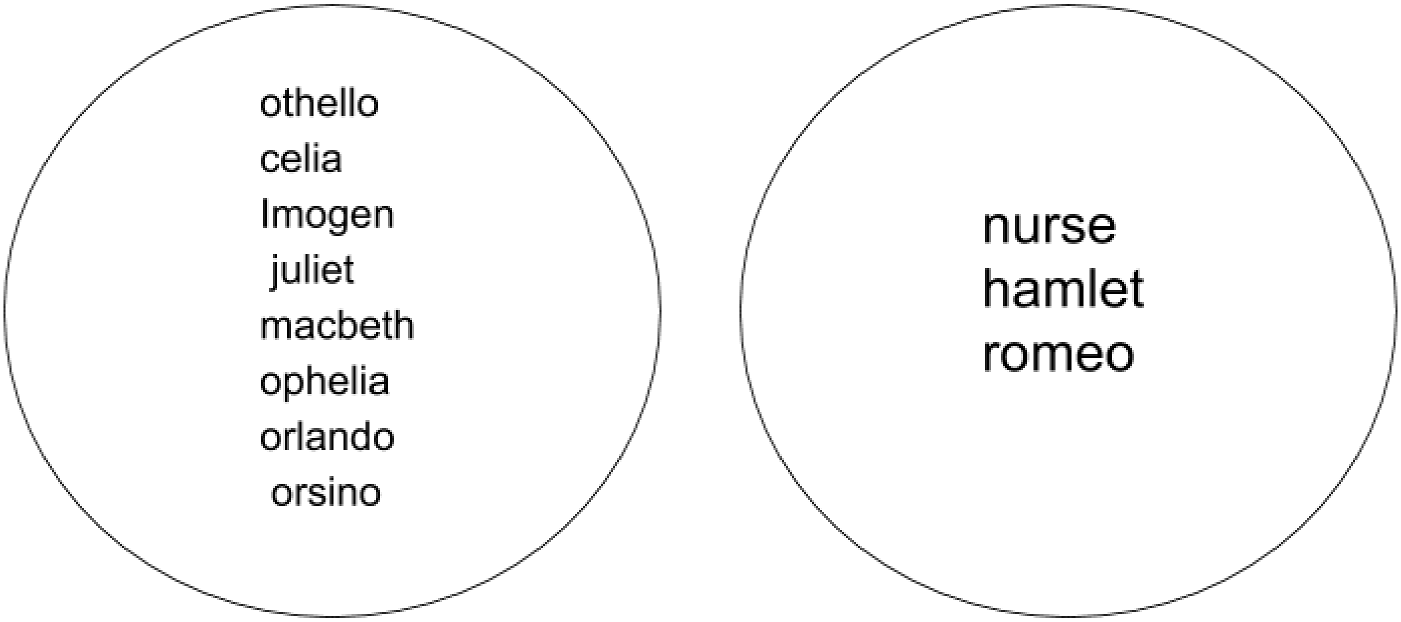
If we didn’t know who ‘Hamlet’ is, cosine similarity on PPMI would be very useful because ‘hamlet’ is a lord, son of a king and a queen.

Character Clustering

We used a seed character list =
['othello', 'celia', 'imogen', 'juliet', 'macbeth', 'ophelia','nurse', 'hamlet', 'orlando', 'orsino', 'romeo']

And used all the matrices we have created to cluster these characters into two groups
We used parameters: *n_clusters=2, init='random', max_iter=300, random_state=0, n_init=30*

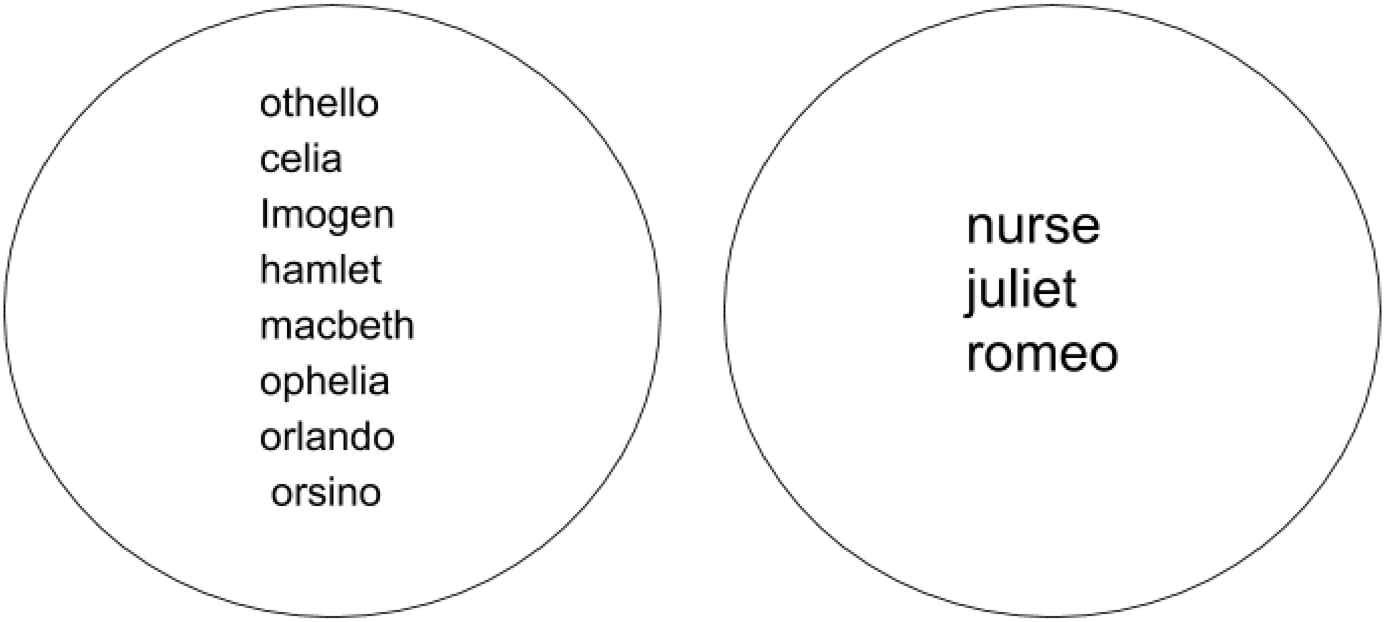
For TC_Matrix and PPMI_matrix, groupings were identical



othello
celia
Imogen
juliet
macbeth
ophelia
orlando
orsino

nurse
hamlet
romeo

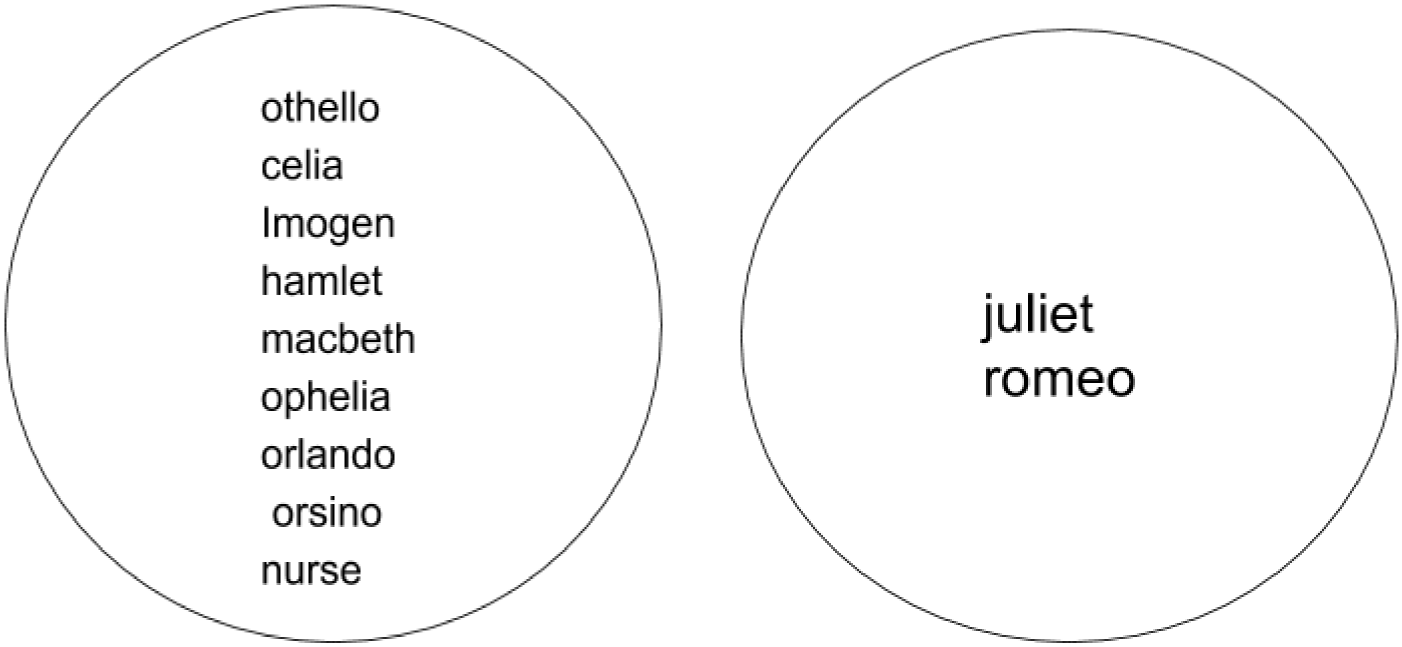
For TD_Matrix:



othello
celia
Imogen
hamlet
macbeth
ophelia
orlando
orsino

nurse
juliet
romeo

For TF_IDF_Matrix:



Character Clustering Observations:

Term-Context and PPMI were better at grouping characters based on their genders. Except for ‘orlando’ and ‘orsino’ - male characters- and ‘nurse’ - a female character-, groupings were accurate. This indicates that females were represented similarly in term-context matrices. If we had more male characters, the results might have been different.

TD_Matrix on the other hand grouped the characters of the play “Romeo and Juliet”. This is not surprising because word vectors have features related to plays and ‘romeo’,’juliet’ and ‘nurse’ will have similar vectors as they appear in same plays.

Effect of Smoothing on PPMI

Running similarities on the word juliet but this time no smoothing gave the following cosine similarity:

Cosine
disliken
tribe
success
mood
pipes
friday
scaring
lovers
hue

When we compare these words to the words that we got in the previous section for ‘juliet’, we can observe that these words are less correlated with the character. Smoothing seems to be helping to make more meaningful associations.