Ian MacDonald

CIS 530

Homework 4 Writeup

Comparing Plays

In this section, plays from each of Shakespeare's main genres – Comedy (*All's Well That Ends Well*), Tragedy (*Cymbeline*), and History (*Henry IV*) - were chosen and compared to all the other Shakespeare plays given. The top 10 most similar plays to each were calculated by each of the similarity functions – Cosine Similarity, Jaccard Similarity, and Dice Similarity – and reported below. The expectation is that the most similar plays for each genre will be plays in the same genre, based on the assumption that plays in the same genre should have a lot more words in common. To determine the efficacy of the similarity functions, the chosen play was included during the function testing, as it should show up as the most similar play for each metric, with a similarity score of 1, which it did. However, after checking to make sure the functions were working well, the chosen play was removed, and is removed in all of the result tables below as well, as it is assumed that each play would be most similar to itself.

Comedy (*All's Well That Ends Well*)

Most Similar Plays By Similarity Function

Cosine Similarity	Jaccard Similarity	Dice Similarity
Cymbeline	*Cymbeline*	*Cymbeline*
Othello	*Othello*	*Othello*
Measure for Measure	*Measure for Measure*	*Measure for Measure*
King Lear	*King Lear*	*King Lear*
Timon of Athens	*Timon of Athens*	*Timon of Athens*
Henry VIII	*Henry VIII*	*Henry VIII*
A Winters Tale	*A Winters Tale*	*A Winters Tale*
Much Ado About Nothing	*Merchant of Venice*	*Merchant of Venice*
Merchant of Venice	*Much Ado About Nothing*	*Much Ado About Nothing*
Hamlet	*Antony and Cleopatra*	*Antony and Cleopatra*

Analysis

Oddly enough, the most similar play to this comedy was a tragedy, *Cymbeline*, which coincidentally was the tragedy chosen for closer analysis for this report. Given that it shows up at the top of each of these lists, it will be interesting to see if/where it shows up on the list for *Cymbeline*. In general, the results for this play were unexpected, as there are only a few comedies here (*Measure for Measure, A Winter's Tale*, *Much Ado About Nothing, Merchant of Venice*). However, after a bit of research, it turns out that *All's Well That Ends Well* is categorized as a "Shakespearean Problem Play", along with some other famous Shakespeare plays which can be found on Wikipedia (https://en.wikipedia.org/wiki/Shakespearean_problem_play). Fascinatingly though, all of the plays commonly denoted as these "problem plays" are listed in the above table, save for one. Given that, it seems as if the model did pretty well in finding similarities between these plays, even if the similarities were not the ones that were expected.

Tragedy (*Cymbeline*)

Most Similar Plays By Similarity Function

Cosine Similarity	Jaccard Similarity	Dice Similarity
King Lear	*King Lear*	*King Lear*
Antony and Cleopatra	*Antony and Cleopatra*	*Antony and Cleopatra*
A Winters Tale	*A Winters Tale*	*A Winters Tale*
Timon of Athens	*Coriolanus*	*Coriolanus*
Coriolanus	*Timon of Athens*	*Timon of Athens*
Richard III	*Richard III*	*Richard III*
Othello	*Othello*	*Othello*
Alls Well That Ends Well	*Alls Well That Ends Well*	*Alls Well That Ends Well*
Henry VIII	*Henry VIII*	*Henry VIII*
Richard II	*Richard II*	*Richard II*

Analysis

The similar plays in this cohort were closer to the plays that were expected. Most of them are tragedies, with a few historical plays thrown in as well. The historical plays are not unexpected though, as *Cymbeline*, otherwise known as *Cymbeline, King of Britain*, likely has a lot of similar undertones to the historical plays, since it is based on legends of a real historical British King (according to Wikipedia, https://en.wikipedia.org/wiki/Cymbeline). Another couple interesting things here are that all of the similarity functions are in total agreement – same plays, same order, and that *All's Well That Ends Well* does indeed show up in *Cymbeline's* list of similar plays, although at a lower rank than when looked at the other way around.

History (*Henry VI*)

Most Similar Plays By Similarity Function

Cosine Similarity	Jaccard Similarity	Dice Similarity
King John	*King John*	*King John*
Richard II	*Richard II*	*Richard II*
Richard III	*Richard III*	*Richard III*
Henry VI Part 3	*Henry VI Part 3*	*Henry VI Part 3*
Merchant of Venice	*King Lear*	*King Lear*
Cymbeline	*Cymbeline*	*Cymbeline*
King Lear	*Romeo and Juliet*	*Romeo and Juliet*
Romeo and Juliet	*Merchant of Venice*	*Merchant of Venice*
Henry VI Part 2	*Henry VI Part 2*	*Henry VI Part 2*
Much Ado About Nothing	*Othello*	*Othello*

Analysis

This set of plays was both expected and unexpected. What was unexpected was how low the other plays about Henry VI were, coming in at numbers 4 and 9 in the rankings. Given that they are both about Henry VI (I assume), it is interesting that they aren't ranked numbers 1 and 2 as expected. However, they are both still on the rankings list, which was expected, and are joined by a number of

other historical plays, including *Richard II*, *Richard III*, and *King John*, as well as *Cymbeline*, which seems to be pretty popular. However, given the information about *Cymbeline* above, it is not surprising to see it here.

Overall Play Comparison Discussion

Overall, the play comparisons gave results that were mostly expected – tragedies were similar to tragedies, historical plays were similar to historical plays. Even the unexpected results, when looked into more closely, actually made some sense – most notably the other "Shakespeare Problem Plays" associated with *All's Well That Ends Well*. In terms of comparing the similarity functions, they all gave relatively similar results. Jaccard and Dice similarity were identical, which isn't surprising given that Dice uses the Jaccard Index in it's calculations. Additionally, since there's only around 30 plays to choose from, and each play consists of lots of words, there probably isn't a lot of variance. Therefore, it doesn't make too much of a difference what similarity function is used here, and they all yield pretty much the same results.

Comparing Words

This section chooses a few different words, and computes the top 10 most similar words to each word, using each similarity function as well as both a TF-IDF matrix and a PPMI matrix. The words chosen were "Juliet", "Romeo", "Hippolyta", and "Beseech". The first two ("Romeo", "Juliet") were chosen since both characters are from the same play, and each meet similar fates, which should make the comparisons easier by eliminating as much variance as possible.

"Hippolyta" was chosen as an opposite to "Romeo" and "Juliet". As the main character in *A Midsummer's Nights Dream* (a comedy), and a Queen whose marriage celebration is one of the central points of the play, the words associated with her are expected to be much happier than those associated with "Romeo" and "Juliet", due to the tragedies that happen to them in their respective play.

Finally, "Beseech" was chosen because it is an outdated word, one that has since fallen out of use since Shakespearean times. However, it translates to "ask urgently", so it would be interesting to see if the words similar to "beseech" are words similar to "ask" or words associated with "ask".

For each word, the word itself is removed from the rankings for the same reasons as listed above for the plays. Similarly to above, it was left in for testing to check the efficacy of the functions, then removed afterward, and it can be assumed here as well that the word itself would be the most similar word in all rankings.

Juliet

Most Similar Words By Similarity Function, TF-IDF Matrix

Cosine Similarity (TF-IDF)	Jaccard Similarity (TF-IDF)	Dice Similarity (TF-IDF)
warwick	orlando	orlando
lucius	silvia	silvia
gloucester	proteus	proteus
antonio	cassio	cassio
helena	kent	kent
othello	hermia	hermia
servants	nurse	nurse

brutus	demetrius	demetrius
claudio	iago	iago
clifford	marcus	marcus

Most Similar Words By Similarity Function, PPMI Matrix

Cosine Similarity (PPMI)	Jaccard Similarity (PPMI)	Dice Similarity (PPMI)
capulet	silvia	silvia
lucio	tybalt	tybalt
cleomenes	capulet	capulet
tybalt	lucio	lucio
katarina	leonato	leonato
silvia	montague	montague
olivia	olivia	olivia
rosencrantz	proteus	proteus
banditti	senators	senators
executioners	anne	anne

Analysis

Most of the words associated with “Juliet” are other names, which makes sense since names are likely used in similar fashion across all the plays. There are a few characters from *Romeo and Juliet* listed, such as “Capulet”, “Tybalt”, and “Montague”, which makes sense, and there seems to be a fair amount of female names listed, so when comparing to “Romeo” it will be interesting to see if the female to male ratio of names stays somewhat constant, or if it is very different for the two. Additionally, it is interesting that “Romeo” does not pop up here at all. Other than the names though, there isn’t much here.

Romeo

Most Similar Words By Similarity Function, TF-IDF Matrix

Cosine Similarity (TF-IDF)	Jaccard Similarity (TF-IDF)	Dice Similarity (TF-IDF)
caesar	hector	hector
hector	talbot	talbot
she	cassio	cassio
there	troilus	troilus
he	hamlet	hamlet
antonio	antonio	antonio
dead	friar	friar
god	soldier	soldier
here	living	living
tybalt	clarence	clarence

Most Similar Words By Similarity Function, PPMI Matrix

Cosine Similarity (PPMI)	Jaccard Similarity (PPMI)	Dice Similarity (PPMI)
mercutio	mercutio	mercutio
tybalt	tybalt	tybalt
hist	clifford	clifford
booted	juliet	juliet
kinsman	hamlet	hamlet
pined	marcius	marcius
juliet	friar	friar
Clifford	lysander	lysander
vintner	wounded	wounded
doff	claudio	claudio

Analysis

Similarly to “Juliet”, “Romeo” has mostly names as it’s similar words. However, it’s interesting to note that almost every name listed here is a male name, so there is definitely a difference in the male to female ratio of names. Another interesting thing to note is that the only female name that does show up here is “Juliet”, even though for “Juliet” we did not see the name “Romeo”. Also interesting is that there’s a significantly higher amount of non-person words here then when compared to “Juliet”, and some of those words actually make a lot of sense when paired with “Romeo”, such as “booted”, “soldier”, “living”, and (tragically), “dead”.

Hippolyta

Most Similar Words By Similarity Function, TF-IDF

Cosine Similarity (TF-IDF)	Jaccard Similarity (TF-IDF)	Dice Similarity (TF-IDF)
clown	theseus	theseus
lucius	voltimand	voltimand
egeus	egeus	egeus
maria	musicians	musicians
attendants	montano	montano
Diomedes	oberon	oberon
Bertram	dogberry	dogberry
gratiano	exton	exton
lafeu	oswald	oswald
helena	sheriff	sheriff

Most Similar Words By Similarity Function, PPMI Matrix

Cosine Similarity (PPMI)	Jaccard Similarity (PPMI)	Dice Similarity (PPMI)
banditti	chatillon	chatillon
executioners	executioners	executioners
theseus	theseus	theseus
blackamoors	abhorson	abhorson
philostrate	philostrate	philostrate

abhorson	blackamoors	blackamoors
salanio	voltimand	voltimand
egeus	escalus	escalus
frederick	dogberry	dogberry
dogberry	egeus	egeus

Analysis

"Hippolyta" again runs into many of the same issues as "Romeo" and "Juliet", with many of the similar words being other characters. "Theseus" does pop up as a top word in most of the rankings, which is expected, along with words such as "musicians" and "attendants", which are likely to be associated with royalty, especially at a wedding ceremony. A couple of words do seem to be a bit odd even though they have relatively high rankings, such as "clown" and "executioners". A quick Google search mentions there are executioners in the play, so maybe they do have something to do with "Hippolyta", but having never read the play it's hard to say for sure.

Beseech

Most Similar Words By Similarity Function, TF-IDF Matrix

Cosine Similarity (TF-IDF)	Jaccard Similarity (TF-IDF)	Dice Similarity (TF-IDF)
thank	thank	thank
peseech	entreat	entreat
implored	please	please
entreat	pray	pray
please	attend	attend
requesting	pleasure	pleasure
bonjour	understand	understand
allege	offend	offend
christenings	accept	accept
cabins	humbly	humbly

Most Similar Words By Similarity Function, PPMI Matrix

Cosine Similarity (PPMI)	Jaccard Similarity (PPMI)	Dice Similarity (PPMI)
thank	thank	thank
peseech	entreat	entreat
implored	please	please
entreat	pray	pray
please	attend	attend
requesting	pleasure	pleasure
bonjour	understand	understand
allege	offend	offend
christenings	accept	accept
cabins	humbly	humbly

Analysis

The list of words here is most in line with what was expected for any of the words in this section. As mentioned above, beseech is an outdated term that translates to “ask urgently”, and the words above are all either very similar in definition to the word ask (“implored”, “requesting”) or words very commonly seen with somebody asking for something (“please”, “thanks”, “humbly”, “attend”, etc.). Additionally, it seems cosine similarity was able to catch “peseech”, which is assumed to be a typo of “beseech”.

Overall Word Comparison Discussion

Overall, the word comparison discussion was a mixed bag. The analysis of the character words was made difficult due to the fact most of the similar words were just other characters, but even through that a lot of the words that popped up as similar made a lot of sense. The words in the “beseech” section though were pretty close to the kinds of words that were expected beforehand. For every comparison word, the Jaccard and Dice Similarities performed exactly the same, which is the same behavior seen for the plays. Again, this may be because the Jaccard and Dice formulas are very similar. For most of the comparison words, the matrix/similarity function combinations didn’t seem to yield drastically better or worse similar words. The only exception would be for the word “beseech”, where cosine similarity was able to catch the word “peseech” for both matrices. This indicates that cosine similarity performed well, because it was able to catch a nonsensical, typo word (which in this case can almost be thought of as a word that was randomly substituted in for “beseech”, and thus should have very similar context around it) and determine that it was used similarly to “beseech”. Because cosine similarity was able to catch that, I would say it performed better than Jaccard or Dice for that particular word, otherwise they performed pretty similarly across the board. Both matrices I would say also performed equally well, there wasn’t a huge difference between the two in terms of being better or worse.