CIS530 HW3

Ignacio Arranz, Jishnu Renugopal

January 30, 2018

1 How do I know if my rankings are good

Rank	Cosine	Jaccard	Dice
1	All's well...	All's well...	All's well...
2	A Winter's Tale	A Winter's Tale	A Winter's Tale
3	As you like it	Measure for measure	Measure for measure
4	Cymbeline	Cymbeline	Cymbeline
5	Othello	Othello	Othello
6	Merchant of Venice	As you like it	As you like it
7	Twelfth Night	King Lear	King Lear
8	King Lear	Merchant of Venice	Merchant of Venice
9	Measure for measure	Much Ado about nothing	Much Ado about nothing
10	Much Ado about nothing	Antony and Cleopatra	Antony and Cleopatra

Table 1: Similarity to 'All's well that ends well'.

For all methods we can see a very similar ranking. For starters, identifying the same play as the first one is a positive sign.

Second, the fact that "A Winter's Tale", "As You Like it" and "Measure for Measure" rank highly, is also indicative of a good algorithm as these are all comedies.

Further online research helps validate this, telling us about similarities between "All's well that ends well" and "Measure for measure":

- All's Well That Ends Well, written about 1598, or six years previous to Measure for Measure, turns on the same dramatic device, the substitution of one bed partner for another. Critics point out that while this works well as a part of the plot in All's Well, in Measure for Measure it seems tacked on.

2 Segmenting Shakespeare's plays

2.1 Segmentation of term document matrix

One way of analyzing the methods was producing a segmentation of the plays based on the vector representation of every play, taken from the term document matrix. The table and graph below show how the plays have been segmented. You will see subtle differences between the two, as the segment produced in the table was used applying K-means to a transposed term-document matrix (which become a document-term matrix), whereas the segments of the graph were produced with 2 principal components after performing PCA over all words in each play.

The interesting conclusions is that there is a clear segment of the "Henrys". There's another segment of "King" plays, which as seen in the graph is placed very closely to the one of Henry's. This indicates close similarity too.

1	2	3	4	5	6
Macbeth	Henry VIII	Hamlet	King John	Merchant of Venice	Henry VI P2
2 Gentl. of Verona	A Winters Tale	Richard III	Henry VI P1	Twelfth Night	Henry VI P3
A Comedy of Er..	Troilus & Cressida		Richard II	As you like it	Henry V
Julius Cesar	Romeo and Juliet		Titus Andr.	Much Ado ...	Henry IV
Pericles	Othello			Measure for me..	
The Tempest	Coriolanus			Merry Wives of ...	
A Midsummer..	Antony & Cleo..			Taming of the...	
Timon of Athens	Cymbeline			Loves Labours Lost	
				All's well...	

Table 2: Segments of plays.

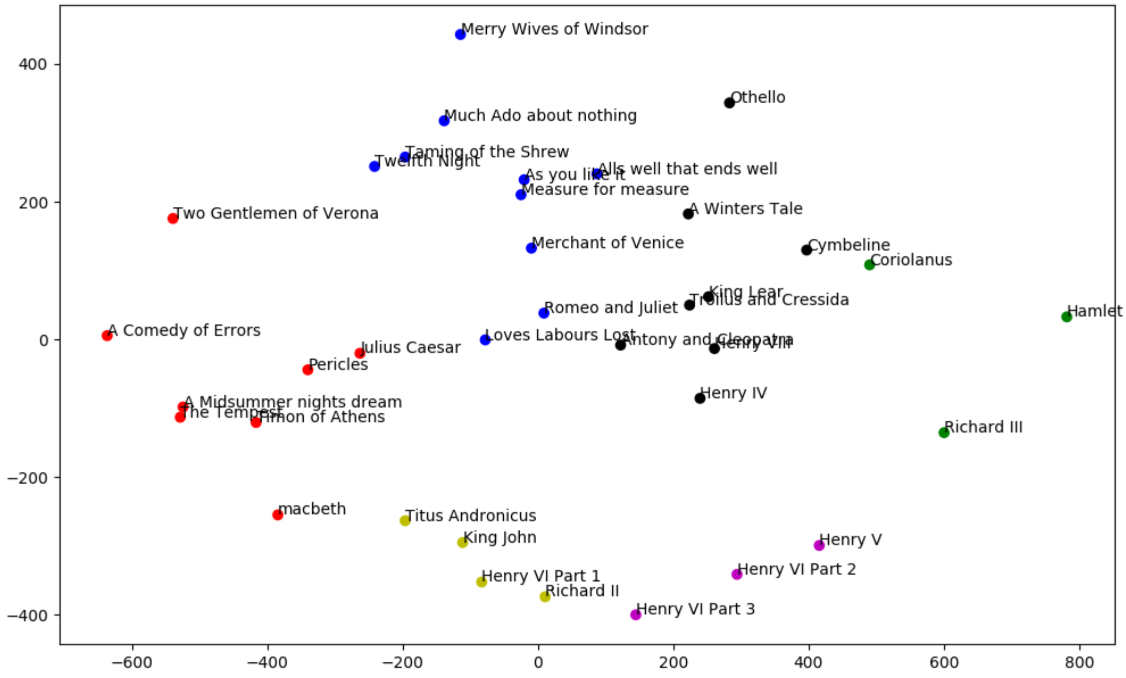

Figure 1: Shakespeare Plays segmented by their similarity. Axes are principal components.

When we decided to look at the Principal Components to assess which words where most heavily influencing each PC, we realized that the top words were essentially stopwords.

	PC1	PC2
1	the	you
2	and	i
3	of	a
4	to	her
5	my	sir
6	i	she
7	in	it
8	a	he
9	you	is
10	his	not

Table 3: Principal Components.

This suggests that the results of applying a segmentation with the term-document matrix may not be ideal.
Note: The lecture on Monday Jan 28th confirmed that applying euclidean distance as a measure of distance on the term-document matrix was not good practice, but we wanted to keep the conclusion we had arrived to, to show how our analysis evolved.

2.2 Segmentation of term document matrix after normalizing document vectors

When we normalize the vectors, we get the following segmentation, which is the equivalent of making a segmentation with cosine similarity distances:

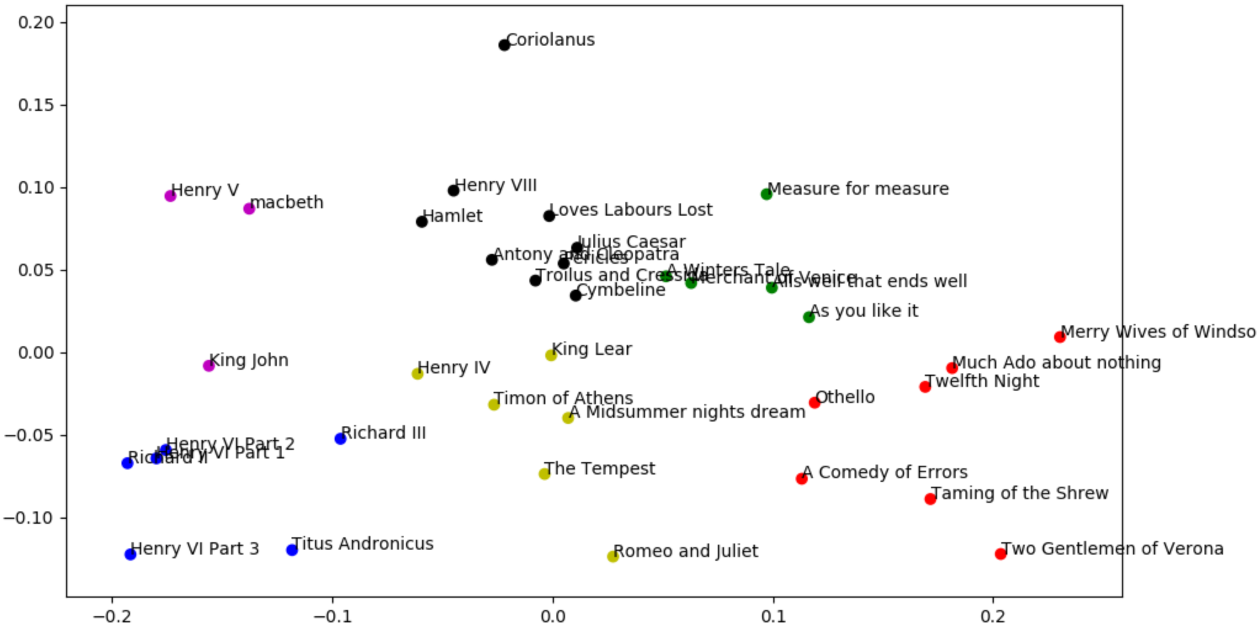

Figure 2: Shakespeare Plays segmented by their similarity, with normalized vectors. Axes are principal components.

	PC1	PC2
1	you 0.41	the 0.43
2	i 0.41	you 0.31
3	her 0.18	he 0.17
-3	of -0.23	and -0.26
-2	and -0.27	my -0.27
-1	the -0.32	thou -0.28

Table 4: Principal Components without stopwords.

The main words of the principal components are still stopwords, so this probably indicates that when doing cosine similarity, stopwords have a very strong weighting in the similarity of plays, when they should not.

2.3 Segmentation of term document matrix without stopwords

Given the results for the principal components in the previous segmentation, we created a different term-document matrix, without stopwords. The results have a noticeable change, as we see the plays change their segments and similar plays. There is a much stronger segmentation of Henry VI Parts I, II and III together with Richard III, for example.

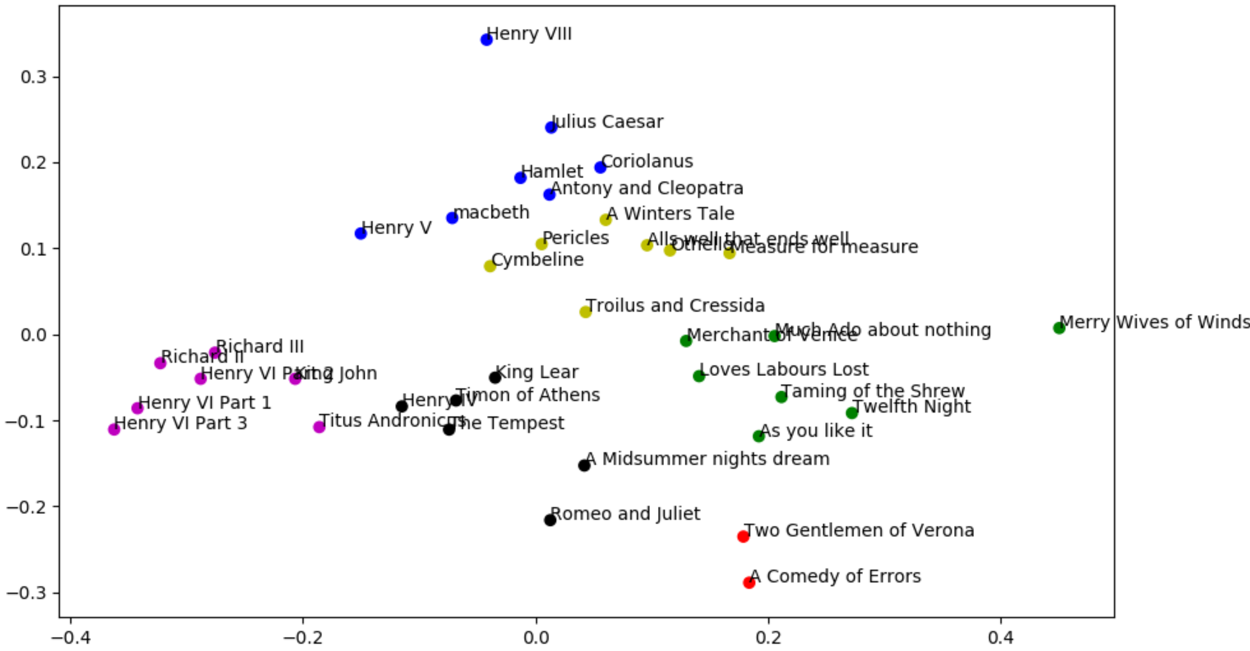

Figure 3: Shakespeare Plays segmented by their similarity, but after normalizing and removing stopwords. Axes are principal components.

	PC1	PC2
1	observant 0.36	glorious 0.21
2	questant 0.18	candle 0.19
3	garrison 0.16	approacheth 0.14
-3	fust -0.25	questant -0.24
-2	approacheth -0.29	fust -0.27
-1	unloading -0.33	portal -0.42

Table 5: Principal Components without stopwords.

When computing the frequency with which the words appeared in each document, we saw that "approacheth" occured in Henry VI Part I, Henry VI Part III and The Two Gentlemen of Verona. Given it is the word with highest coefficient in PC1, it seems logical that Henry VI Parts I and III are at the far left on that axis uses the word. Applying principal components may be oversimplifying the segmentation, and may also have lower coefficients for words with a very high normalized frequency.

2.4 Segmentation of tf-idf matrix

When computing a segmentation with the TF-IDF matrix, the visual results with PCA were surprising, as two plays were very different from the rest.

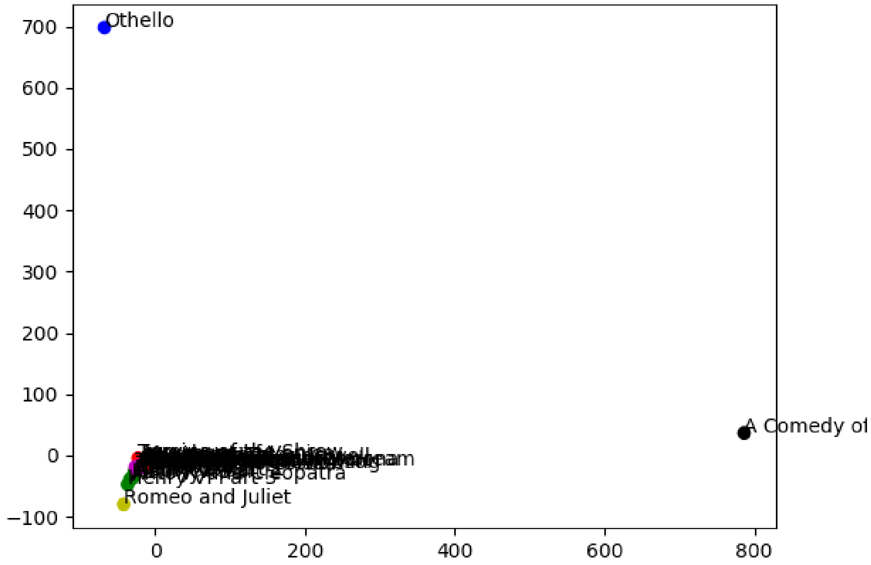

Figure 4: Shakespeare Plays segmented by their similarity using TF-IDF matrix. Axes are principal components.

When we look at the Principal Components, we can see that the names are the key factors for each principal component.

	PC1	PC2
1	antipholus 0.92	cassio 0.69
2	dromio 0.25	iago 0.43
3	syracuse 0.11	desdemona 0.31

Table 6: Principal Components with TF-IDF.

Antipholus of Syracuse and his servant Dromio of Syracuse are the main characters of Comedy of Errors. The reason for this type of segmentation likely is that Antipholus is a name that does not appear on any other play, but appears with great frequency on Comedy of Errors (it appears 211 times in Comedy of Errors and does not appear on any other play). Romeo appears 146 times in Romeo and Juliet, Juliet appears 9 times in Measure for Measure and 63 times in Romeo and Juliet, Similarly for the other principal component, Cassio, Iago and Desdemona are main characters of Othello.
TF-IDF seems to have a magnifying effect, and is logically heavily routed in character names. This then becomes not a great measure of similarity, as the strongest variables will be names of characters and the strongest association will be drawn between plays with characters with the same name.
An interesting analysis would be to draw similarities of texts after removing character names.

2.5 Segmentation of tf-idf matrix built from term-document matrix that excludes character names

Seeing the relevance that character names had on the tf-idf matrix, we decided to exclude these from the term-document matrix to then perform tf-idf. The results are very interesting:

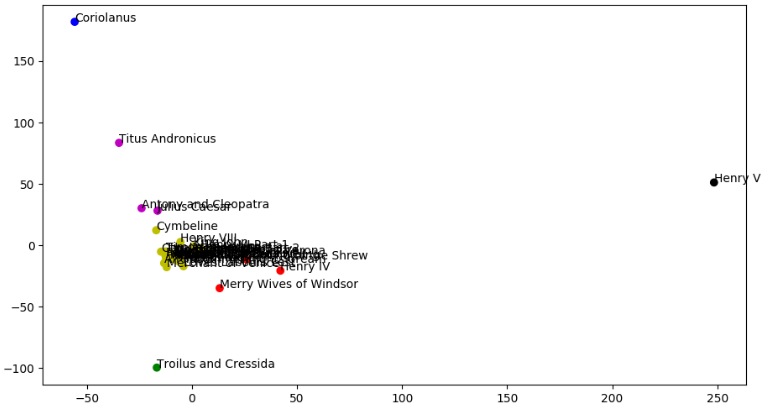

Figure 5: Shakespeare Plays segmented by their similarity using TF-IDF matrix, but excluding character names from the term-document matrix. Axes are principal components.

When we look now at the principal components, we see that actual words are highly represented in each principal component:

	PC1	PC2
1	je 0.92	rome 0.69
2	vous 0.25	consul 0.43
3	kate 0.25	corioli 0.43
4	les 0.11	volsces 0.31

Table 7: Principal Components with TF-IDF, after excluding character names.

Three out of the first four words in PC1 are in french. This is clear evidence that Henry V is the only Shakespeare play where french is spoken. Similarly, it seems evident that both Coriolanus and Titus Andronicus are both set in Rome. Evidently so, the following is the ranking of plays by its use of the word 'rome':

	Play	Frequency
1	Titus Andronicus	110
2	Coriolanus	102
3	Julius Caesar	42
4	Antony and Cleopatra	34
5	Cymbeline	13
5	King John	10
5	Henry VII	10

Table 8: Number of times the word Rome is repeated in a play.

3 Understanding Shakespearean vocabulary

Whenever we had to read Shakespeare in High School, the main challenge was understanding what the word really was being used for. As much of the english was nothing like other things we read, this posed a challenge. For this reason, we decided to select specific words that people may use differently nowadays to see what their most similar words are in Shakesperean english, **while testing our similarity functions**.

3.1 What dost this verb mean?

"O Romeo, Romeo, wherefore **art** thou Romeo?"

For the word **art** , the most similar words are:

- Jaccard or Dice on PPMI: **art** , **am** , **was** , **tis** , **being** , **been** , hast...
- Jaccard or Dice on term-context: art, hast, dost, wilt, shalt, tis...

Seems interesting that PPMI seems to find similarity to other verbs, while term-context seems to find it to other tenses of the same verb. So we can continue the analysis with other verbs.

For the word **dost** , the most similar words are:

- Jaccard or Dice on PPMI: **dost** , **didst** , **does** , should, wilt, **doth** ...
- Jaccard or Dice on term-context: dost, wilt, hast, shalt, art...

Again, we see the same trend (not as clear this time, though). But we can conclude that PPMI matrices will enable us to define similarity by the meaning of the action, while term-context will do it by the tense used.

This seems logical, as "art","hast","dost","wilt" are usually preceded by "thou" (or followed by it in questions).

3.2 'tis but an unknown noun

"Your face, my **thane** , is as a book where men may read strange matters". For the word **thane** , the most similar words are:

- Jaccard or Dice on term-context: thane, image, bishop, cawdor...
- Cosine on PPMI: thane, cawdor, governor, macduff...
- Jaccard or Dice on PPMI: thane, wolsey, supposed, ashamed, discharge...

According to Oxford English Dictionary:
thane - (in Scotland) a man, often the chief of a clan, who held land from a Scottish king and ranked with an earl's son. Example: "the Thane of Cawdor".

It shouldn't come as a surprise that for the term-context matrix, thane and cawdor are not similar, as we used a distance of one word and "Thane of Cawdor" is the usual format of the expression. For PPMI, on the other hand, cawdor is its most similar word, meaning they appear frequently together. Here the term similarity is probably not the correct one, as "thane" and "cawdor" are more **complementary** than they are similar.

4 Character Sentiment Analysis

4.1 How do popular characters feel?

One analysis we thought would be interesting would be to measure the average polarity of sentences said by different characters (we used a Python library called TextBlob to do this). We thought best to select the most prominent characters and compare their behaviors, as for characters that don't speak as much it may be harder to assess how they feel as they really didn't have that big a chance to express themselves.

- **Queen Margaret** seems to be very upset or pessimistic: "No sleep close up that deadly eye of thine, Unless it be while some tormenting dream Affrights thee with a hell of ugly devils."
- **Macbeth** may have also had some rough days: "Out, out, brief candle! Life's but a walking shadow, a poor player that struts and frets his hour upon the stage and then is heard no more: it is a tale told by an idiot, full of sound and fury, signifying nothing."

Character	Sum of polarity	Average Polarity	Number of lines
gloucester	68.7512	0.0358	1920
hamlet	100.408	0.0634	1582
iago	80.1745	0.0691	1161
falstaff	62.9565	0.0564	1117
king henry v	66.3063	0.0611	1086
brutus	44.7923	0.0426	1051
othello	65.6497	0.0707	928
mark antony	61.9507	0.0668	927
king henry vi	56.5477	0.0617	917
duke vincentio	66.6639	0.0733	909
timon	59.8252	0.0684	875
queen margaret	13.2562	0.0157	847
clown	54.7841	0.0681	804
king lear	24.6141	0.0307	801
king richard ii	32.3479	0.0407	794
macbeth	17.5582	0.0224	783
titus andronicus	26.5705	0.0346	768
prospero	53.1673	0.0714	745
...	...	...	...
clarence	-3.54417	-0.0137	258

Table 9: Number of lines said by each character, average polarity and sum of all polarity.

A character called **Clarence** struck us for his low average polarity. Upon some research, we found out he has a monologue in Richard III which starts with "O, I have passed a miserable night, So full of fearful dreams, of ugly sights" and continues in that same tone, which explains his low polarity.

5 Comparison with SimLex-999

Co-ocurrence matrix	Cosine similarity	Dice similarity	Jaccard similarity
Term-document	-0.057	-0.074	-0.074
Term-context	-0.041	-0.043	-0.043
TF-IDF	-0.059	-0.051	-0.051
PPMI	0.0015	-0.035	-0.035

Table 10: Correlation with human judgements.

It was observed that there was almost no correlation for all settings with the human similarity ratings given by the SimLex-999 dataset. This can be attributed to the changes in language over time. Diachronic studies have shown that the usage of words and their meanings have considerably evolved over time.