CS591
HW2 Writeup
Ziran Min
U59274427
minziran@bu.edu

Note: Source code is not included in this report. Some useful output will be shown.

1. Most similar characters

In will_play_text.csv, there is a column containing the name of the character who spoke in each line. In order to find the most similar pair of characters, we need to create a new term-document matrix like the one in previous part of the homework. Instead of putting play names as column names, we need the name of each character in every play.

Therefore, I make some change in the read_in_shakespeare() function that can give me tuples with character name and the list of words that character says, and the list of all unique character names. There are 934 different characters in Shakespeare's collection. I also create a dictionary with keys of each character and values of the play where the character is from.

Then I use the same create_tf_idf_matrix function to weight the matrix. Since the goal is to find the most similar pair of character, I want to change my rank function to a most_similar_characters function which uses one of the similarity functions to find the **character** who has the **highest similarity score** with the **target character**.

In this way, I write a for loop to compute a 2-D list containing the most similar character and the corresponding score of each target character. By sorting the list by similarity scores, I can find the most similar pair that has the highest score.

The following partial output is the top 6 most similar pair by using Cosine Similarity score. In each line, the first name is the target character, the second is his/her most similar character, the score is their similarity scores.

```
[0.9999999999999999, 'Outlaws', 'Second Pirate']],
[0.9999999999999998, 'Second Pirate', 'Outlaws'],
[0.9675301011144135, 'PHILIP', 'JOSEPH'],
[0.9675301011144135, 'JOSEPH', 'PHILIP'],
[0.9141380916379601, 'NICHOLAS', 'JOSEPH'],
[0.8072107528240503, 'Ostler', 'FRANCIS'],
……
```

The following partial output is the top 6 most similar pair by using Jaccard Similarity score.

```
[0.6930698044441733, 'PHILIP', 'JOSEPH']],
[0.6930698044441733, 'JOSEPH', 'PHILIP'],
[0.6666666666666667, 'Second Pirate', 'Outlaws'],
[0.6666666666666667, 'Outlaws', 'Second Pirate'],
[0.6491641781438824, 'NICHOLAS', 'JOSEPH'],
[0.41340030394467553, 'Second Herald', 'First Herald'],
[0.41340030394467553, 'First Herald', 'Second Herald'],
……
```

The following partial output is the top 6 most similar pair by using Dice Similarity score.

```
[0.8187137974168819, 'PHILIP', 'JOSEPH']],
[0.8187137974168819, 'JOSEPH', 'PHILIP'],
[0.8, 'Second Pirate', 'Outlaws'],
[0.8, 'Outlaws', 'Second Pirate'],
[0.7872644661424918, 'NICHOLAS', 'JOSEPH'],
[0.5849727112565517, 'Second Herald', 'First Herald'],
……
```

Now we can see that (`'PHILIP', 'JOSEPH'`), (`'NICHOLAS', 'JOSEPH'`) (`'Outlaws', 'Second Pirate'`), and (`'First Herald', 'Second Herald'`) are top 4 most similar pairs of character.

By using the dictionary I created before, I can check which play these characters come from. `'PHILIP', 'JOSEPH'`, and `'JOSEPH'` come from “Taming of the Shrew”, so the first two pairs make sense. Both `'First Herald'` and `'Second Herald'` are from “Richard II”, the fourth pair make sense two. Two character coming from same play implies they may have many conversations and may say many similar words.

However, `'Outlaws'` is from “Two Gentlemen of Verona” and `'Second Pirate'` is from “Pericles”, but both plays are comedies.

2. Most Central Play

In previous part of the homework, we can find the top 10 most similar play of a target play. We can compare the genre of the target play with the genres of the top similar plays to figure out how "central" the target play is.

In provided 36 Shakespeare's work, 15 are Comedies (labeled as "C"), 12 are Tragedies (labeled as "T") and 9 are Histories (labeled as "H").

My method of finding most central play is: For a target play, we can get its top 9 most similar play. Then we count how many are C, how many are T, and how many are H. **The closer the three numbers are, the more central the target play is**. The best case is a play with numbers 3, 3, 3. In order to measure how close three numbers are, I will compute their variance. The play that has the lowest "three number variance" will be the most central play.

The follow table the is the three number and variance of each play by using Cosine Similarity score.

	C	T	H	Var
Henry IV	2	3	4	0.667
Alls well that ends well	6	3	0	6
Loves Labours Lost	3	5	1	2.666
Taming of the Shrew	8	1	0	12.667
Antony and Cleopatra	3	6	0	6
Coriolanus	3	5	1	2.667
Hamlet	2	4	3	0.667
A Midsummer nights dream	4	5	0	4.667
Merry Wives of Windsor	8	1	0	12.667
Romeo and Juliet	5	4	0	4.667
Richard II	0	2	7	8.667
King John	0	3	6	6
macbeth	0	5	4	4.667
Timon of Athens	3	4	2	0.667
A Winters Tale	4	4	1	2
The Tempest	3	4	2	0.667
Henry VI Part 2	0	2	7	8.667
As you like it	6	3	0	6
Julius Caesar	4	5	0	4.667
A Comedy of Errors	7	2	0	8.667
Henry VIII	4	4	1	2
Measure for measure	6	3	0	6
Richard III	0	3	6	6
Two Gentlemen of Verona	6	3	0	6

Henry VI Part 1	0	2	7	8.667
Much Ado about nothing	8	1	0	12.667
Henry V	1	2	6	4.667
Troilus and Cressida	5	3	1	2.667
Twelfth Night	7	2	0	8.667
Merchant of Venice	5	4	0	4.667
Henry VI Part 3	0	3	6	6
Othello	7	2	0	8.667
Cymbeline	3	5	1	2.667
King Lear	5	2	2	2
Pericles	3	5	1	2.667
Titus Andronicus	2	1	6	4.667

The follow table the is the three number and variance of each play by using Jaccard Similarity score and Dice Similarity score.

	C	T	H	Var
Henry IV	3	4	2	0.667
Alls well that ends well	5	4	0	4.667
Loves Labours Lost	8	1	0	12.667
Taming of the Shrew	8	1	0	12.667
Antony and Cleopatra	2	6	1	4.667
Coriolanus	2	6	1	4.667
Hamlet	2	5	2	2
A Midsummer nights dream	7	2	0	8.667
Merry Wives of Windsor	8	1	0	12.667
Romeo and Juliet	4	4	1	2
Richard II	0	2	7	8.667
King John	2	2	5	2
macbeth	4	3	2	0.667
Timon of Athens	6	3	0	6
A Winters Tale	3	5	1	2.667
The Tempest	5	3	1	2.667
Henry VI Part 2	0	2	7	8.667
As you like it	7	2	0	8.667
Julius Caesar	6	3	0	6
A Comedy of Errors	7	2	0	8.667
Henry VIII	3	5	1	2.667
Measure for measure	6	2	1	4.667
Richard III	1	3	5	2.667
Two Gentlemen of Verona	8	1	0	12.667
Henry VI Part 1	1	1	7	8
Much Ado about nothing	8	1	0	12.667
Henry V	0	3	6	6
Troilus and Cressida	2	6	1	4.667
Twelfth Night	8	1	0	12.667

Merchant of Venice	7	1	1	8
Henry VI Part 3	0	3	6	6
Othello	4	5	0	4.667
Cymbeline	2	5	2	2
King Lear	3	5	1	2.667
Pericles	6	2	1	4.667
Titus Andronicus	1	3	5	2.667

Jaccard Similarity score and Dice Similarity score will give me exactly the same result because Dice = 2*Jaccard/(1+ Jaccard).

Therefore in all we can find that by using Cosine Similarity Score, “Henry IV”, “Hamlet”, “Timon of Athens” and “The Tempest” are the most central plays. By using Jaccard Similarity score and Dice Similarity score, “Henry IV” and “macbeth” are the most central plays.

One guess of why famous tragedy like “Hamlet” will be central could be that many words can be used for both positive and negative meaning or for irony. In other worlds, many worlds can be widely used in both comedies or tragedies.