Homework 3: Vector Space Models

Author: Shashank Garg

Report

Co-occurrence Matrix considered: Term document matrix, term-context matrix

Weighting Measures considered: Frequency, Tf-Idf, PPMI

Similarity measures considered: Cosine similarity, Jaccard Similarity, Dice Similarity

I took both the frequency as well as Tf-idf as the weighting scheme in separate instances for the term document matrix and attempted to compare the similarities of the Shakespeare's plays based on the genres. 3 genres were considered:

- Comedies
- Historical
- Tragedies

For the comedy play, I took **Merry Wives of Windsor** as a candidate play title and found out the 10 most similar plays. Following are the results:

Matrix used: Term Document Matrix

Weighting: Frequency

The 10 most similar plays to "Merry Wives of Windsor" using compute_cosine_similarity is:

1: Merry Wives of Windsor

2: Twelfth Night

3: Much Ado about nothing

4: As you like it

5: Alls well that ends well

6: Taming of the Shrew

7: Merchant of Venice

8: **Othello (misclassification - original category: tragedy)**

9: Measure for measure

10: A Winters Tale

The 10 most similar plays to "Merry Wives of Windsor" using compute_jaccard_similarity are:

1: Merry Wives of Windsor

2: Much Ado about nothing

3: Twelfth Night

4: As you like it

5: Taming of the Shrew

6: Alls well that ends well

7: Merchant of Venice

8: Measure for measure

9: **Othello (misclassification - original category: tragedy)**

10: A Winters Tale

The 10 most similar plays to "Merry Wives of Windsor" using compute_dice_similarity are:

1: Merry Wives of Windsor

2: Much Ado about nothing

3: Twelfth Night

4: As you like it

5: Taming of the Shrew

6: Alls well that ends well

7: Merchant of Venice

8: Measure for measure

9: **Othello (misclassification - original category: tragedy)**

10: A Winters Tale

Result: With term frequency as weighting scheme all three measures of similarity classified all but one play correctly. However, results of jaccard similarity and dice similarity are identical.

Matrix used: Term Document Matrix

Weighting Scheme: Tf-idf

The 10 most similar plays to "Merry Wives of Windsor" using compute_cosine_similarity are:

1: Merry Wives of Windsor

2: **Henry IV (misclassification – original category: Historical)**

3**: Henry V (misclassification – original category: Historical)**

4: **Henry VIII (misclassification – original category: Historical)**

5: **King John (misclassification – original category: Historical)**

6: **Richard III (misclassification – original category: Historical)**

7: Loves Labours Lost

8: **King Lear (misclassification – original category: Tragedies)**

9: **Henry VI Part 2 (misclassification – original category: Historical)**

10: Alls well that ends well

The 10 most similar plays to "Merry Wives of Windsor" using compute_jaccard_similarity are:

1: Merry Wives of Windsor

2: **Henry IV (misclassification – original category: Historical)**

3: **Henry V (misclassification – original category: Historical)**

4: **King Lear (misclassification – original category: Tragedies)**

5: As you like it

6: Much Ado about nothing

7: **Hamlet (misclassification – original category: Tragedies)**

8: Alls well that ends well

9: A Winters Tale

10: **Othello (misclassification – original category: Tragedies)**

The 10 most similar plays to "Merry Wives of Windsor" using compute_dice_similarity are:

1: Merry Wives of Windsor

2: **Henry IV (misclassification – original category: Historical)**

3: **Henry V (misclassification – original category: Historical)**

4: **King Lear (misclassification – original category: Tragedies)**

5: As you like it

6: Much Ado about nothing

7: **Hamlet (misclassification – original category: Tragedies)**

8: Alls well that ends well

9: A Winters Tale

10: **Othello (misclassification – original category: Tragedies)**

Result: Cosine similarity performed worse than jaccard and dice similarity measures.

A similar behavior was seen in tragedy as well as historical plays.

For the historical play, the play Troilus and Cressida was chosen as candidate play.

Attempt by Tragedies play : Troilus and Cressida

Matrix used: Term Document Matrix

Weighting Scheme: Frequency

The 10 most similar plays to "Troilus and Cressida" using compute_cosine_similarity are:

1: Troilus and Cressida

2: Hamlet

3: Cymbeline

4: King Lear

5: Henry IV XXX Histories

6: As you like it XXX Comedies

7: A Winters Tale XXX Comedies

8: Alls well that ends well XXX Comedies

9: Pericles XXX Comedies

10: Merchant of Venice XXX Comedies

The 10 most similar plays to "Troilus and Cressida" using compute_jaccard_similarity are:

1: Troilus and Cressida

2: King Lear

3: Antony and Cleopatra

4: Othello

5: Cymbeline

6: A Winters Tale XXX Comedies

7: Hamlet

8: Henry IV XXX Histories

9: Alls well that ends well XXX Comedies

10: Romeo and Juliet

The 10 most similar plays to "Troilus and Cressida" using compute_dice_similarity are:

1: Troilus and Cressida

2: King Lear

3: Antony and Cleopatra

4: Othello

5: Cymbeline

6: A Winters Tale XXX Comedies

7: Hamlet

8: Henry IV XXX Histories

9: Alls well that ends well XXX Comedies

10: Romeo and Juliet

Attempt by Tragedies play : Troilus and Cressida

Matrix used: Term Document Matrix

Weighting Scheme: Tf-idf

The 10 most similar plays to "Troilus and Cressida" using compute_cosine_similarity are:

1: Troilus and Cressida

2: Loves Labours Lost XXX Comedies

3: Alls well that ends well XXX Comedies

4: Henry V XXX Historical

5: Henry VI Part 1 XXX Historical

6: Henry VI Part 2 XXX Historical

7: King John XXX Historical

8: Twelfth Night XXX Comedies

9: King Lear

10: Henry VI Part 3 XXX Historical

The 10 most similar plays to "Troilus and Cressida" using compute_jaccard_similarity are:

1: Troilus and Cressida

2: Hamlet

3: King Lear

4: Cymbeline

5: King John XXX Historical

6: Coriolanus

7: Othello

8: Antony and Cleopatra

9: Alls well that ends well XXX Comedies

10: Henry V XXX Historical

The 10 most similar plays to "Troilus and Cressida" using compute_dice_similarity are:

1: Troilus and Cressida

2: Hamlet

3: King Lear

4: Cymbeline

5: King John XXX Historical

6: Coriolanus

7: Othello

8: Antony and Cleopatra

9: Alls well that ends well XXX Comedies

10: Henry V XXX Historical

Result: When term frequency is used as weighting measure 6 plays were misclassified while 8 plays were misclassified with tf-idf when cosine similarity was used. However, with using Jaccard and Dice similarity measure both the weighting scheme misclassified only 3 plays.

Similarity of words:

Candidate Word: troubled

Following are the results:

The 10 most similar words to "troubled" using compute_cosine_similarity on term-context frequency matrix are:

1: troubled

2: prayerbook

3: restorative

4: snaffle

5: wanion

6: flail

7: warrener

8: popinjay

9: overmastered

10: trowel

The 10 most similar words to "troubled" using compute_jaccard_similarity on term-context frequency matrix are:

1: troubled

2: suitor

3: torch

4: drowned

5: sounded

6: talking

7: acquainted

8: companion

9: prophet

10: horn

The 10 most similar words to "troubled" using compute_dice_similarity on term-context frequency matrix are:

1: troubled

2: suitor

3: torch

4: drowned

5: sounded

6: talking

7: acquainted

8: companion

9: prophet

10: horn

The 10 most similar words to "troubled" using compute_cosine_similarity on PPMI matrix are:

1: troubled

2: fainted

3: chafing

4: magnanimity

5: armory

6: slab

7: bettered

8: admirer

9: rotundity

10: displant

The 10 most similar words to "troubled" using compute_jaccard_similarity on PPMI matrix are:

1: troubled

2: suitor

3: oppress

4: stirr

5: whipp

6: drowned

7: encounter

8: feasting

9: charged

10: accused

The 10 most similar words to "troubled" using compute_dice_similarity on PPMI matrix are:

1: troubled

2: suitor

3: oppress

4: stirr

5: whipp

6: drowned

7: encounter

8: feasting

9: charged

10: accused

Result: With term-context matrix the similarity performance was worse on all three measures than when using PPMI matrix.