To Mine, or Not to Mine

Data Analysis of the Works of William Shakespeare

[redacted] Text Mining

[redacted]

[redacted]

[redacted] Brian Hogan

William Shakespeare

- Lived 1564 – 1616, England
- Wrote or co-wrote 39 plays, 154 sonnets
- “Widely regarded as the greatest writer in the English language and the world's greatest dramatist” – encyclopedia entry

Shakespearean Musings…

- Text Mining Pursuits:
  - How does vocabulary of comedies differ from tragedies?
  - How does vocabulary differ by character? Hero, heroine, villain…
  - What characters are similar or different?
  - Can algorithms predict if a sentence belongs to a comedy or tragedy?

COMEDIES

[All's Well That Ends Well]()
[As You Like It]()
[Comedy of Errors]()
[Love's Labour's Lost]()
[Measure for Measure]()
[Merchant of Venice]()
[Merry Wives of Windsor]()
[Midsummer Night's Dream]()
[Much Ado about Nothing]()
[Taming of the Shrew]()
[Tempest]()
[Twelfth Night]()
[Two Gentlemen of Verona]()
[Winter's Tale]()

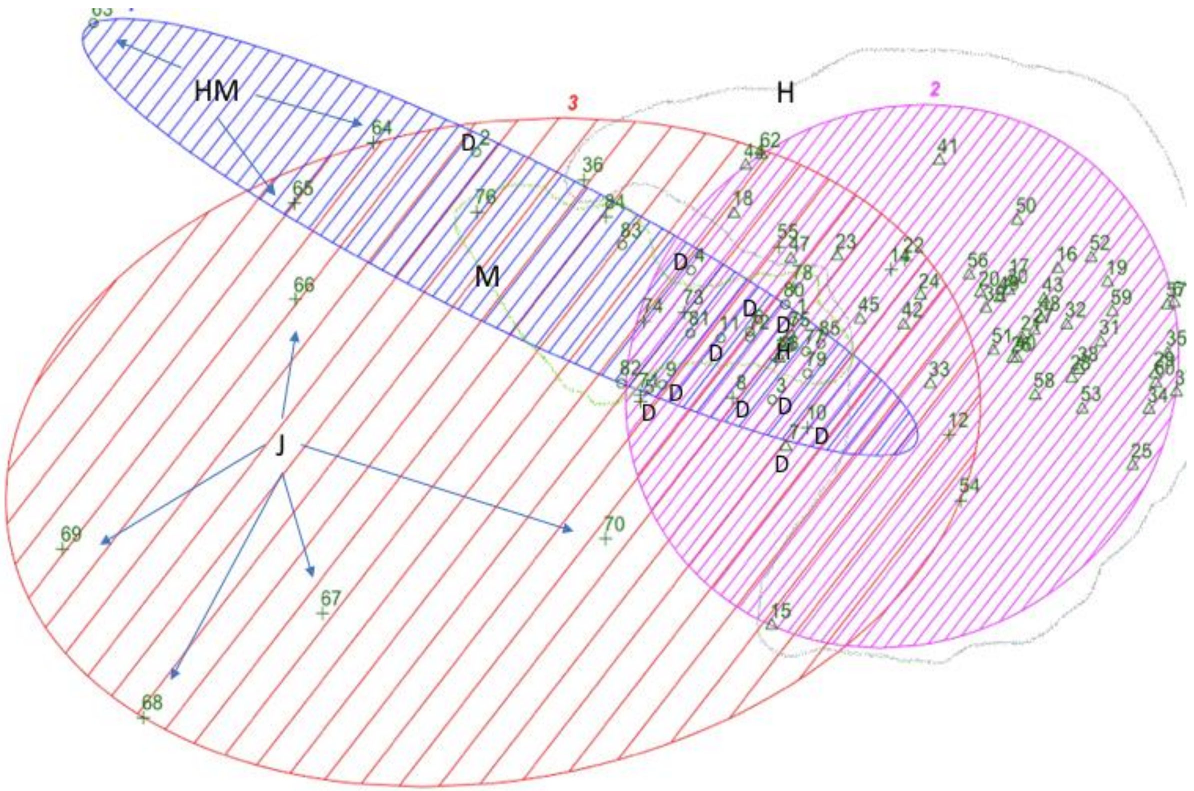
HISTORIES

[Henry IV, Part I]()
[Henry IV, Part II]()
[Henry V]()
[Henry VI, Part I]()
[Henry VI, Part II]()
[Henry VI, Part III]()
[Henry VIII]()
[King John]()
[Pericles]()
[Richard II]()
[Richard III]()

TRAGEDIES

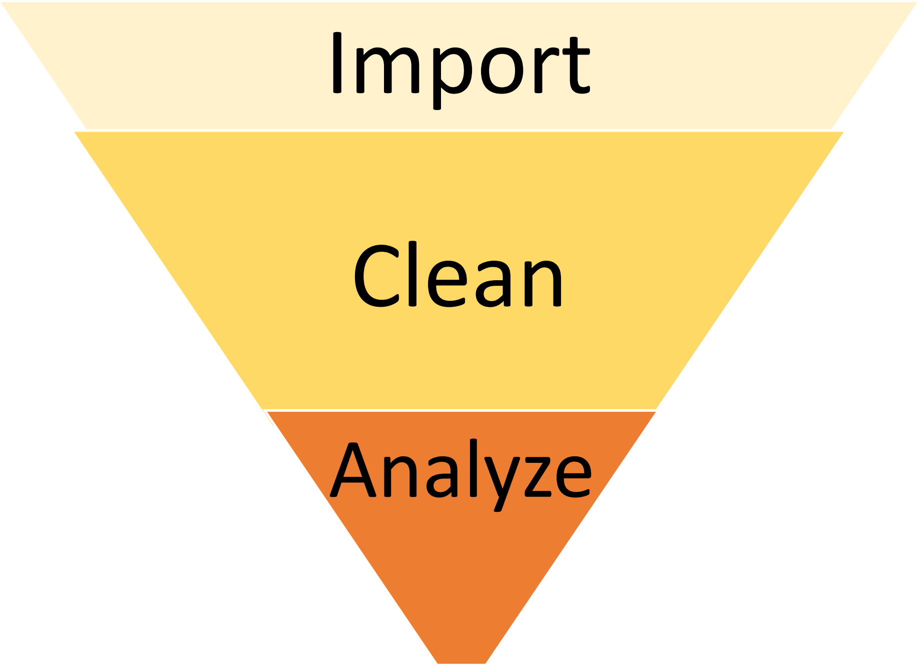
[Antony and Cleopatra]()
[Coriolanus]()
[Cymbeline]()
[Hamlet]()
[Julius Caesar]()
[King Lear]()
[Macbeth]()
[Othello]()
[Romeo and Juliet]()
[Timon of Athens]()
[Titus Andronicus]()
[Troilus and Cressida]()

Corpus



Data Preparation & Distributed Team Tasks

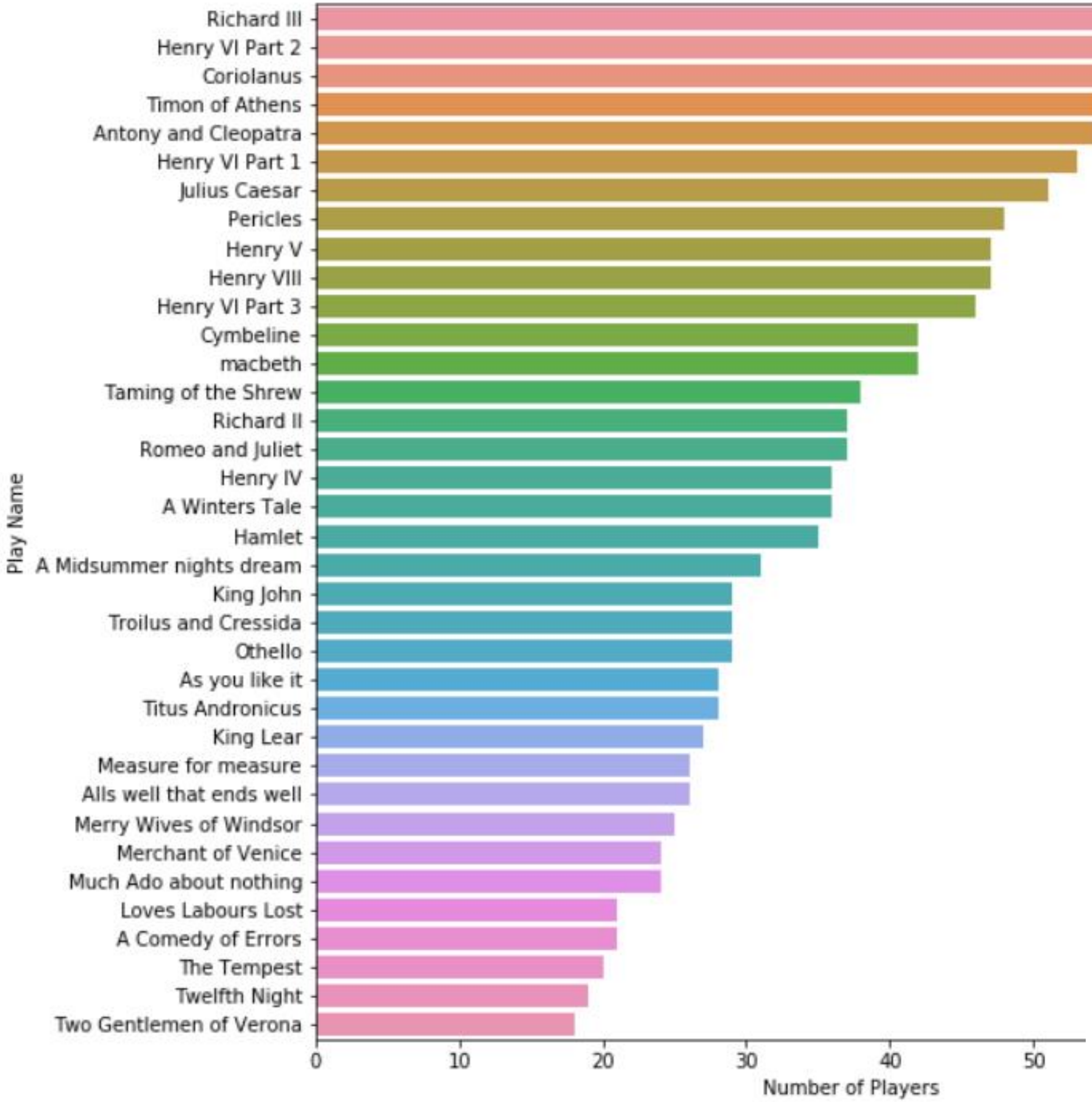
- Scripts ("data") available from MIT*
- Wrote a custom HTML parser which labeled each line of dialogue
- Team members independently answered their question, shared experience & findings
- Each team-member directed and performed analysis by algorithm assignment and interest

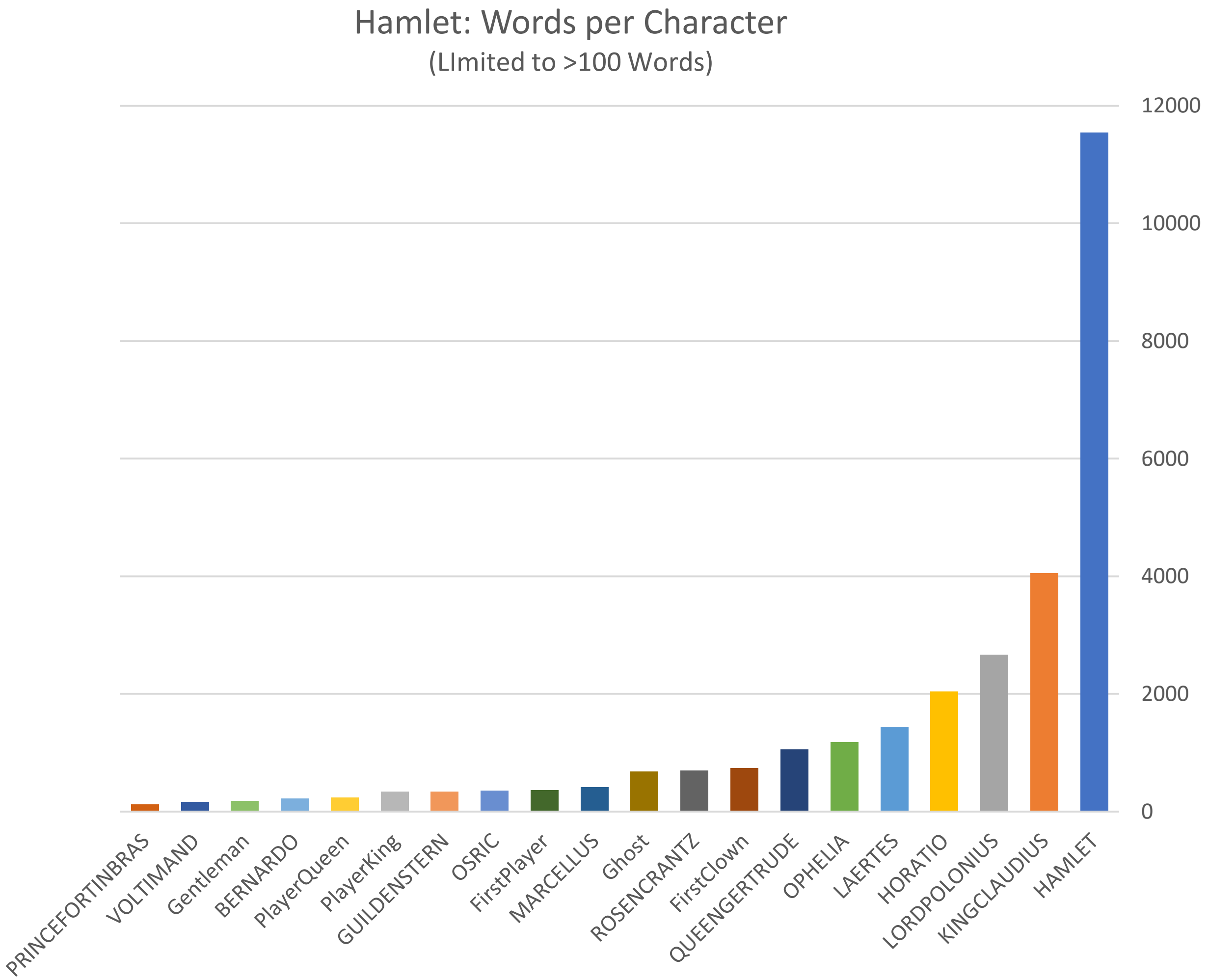


* http://mit.shakespeare.edu

Player and Dialogue Visualizations

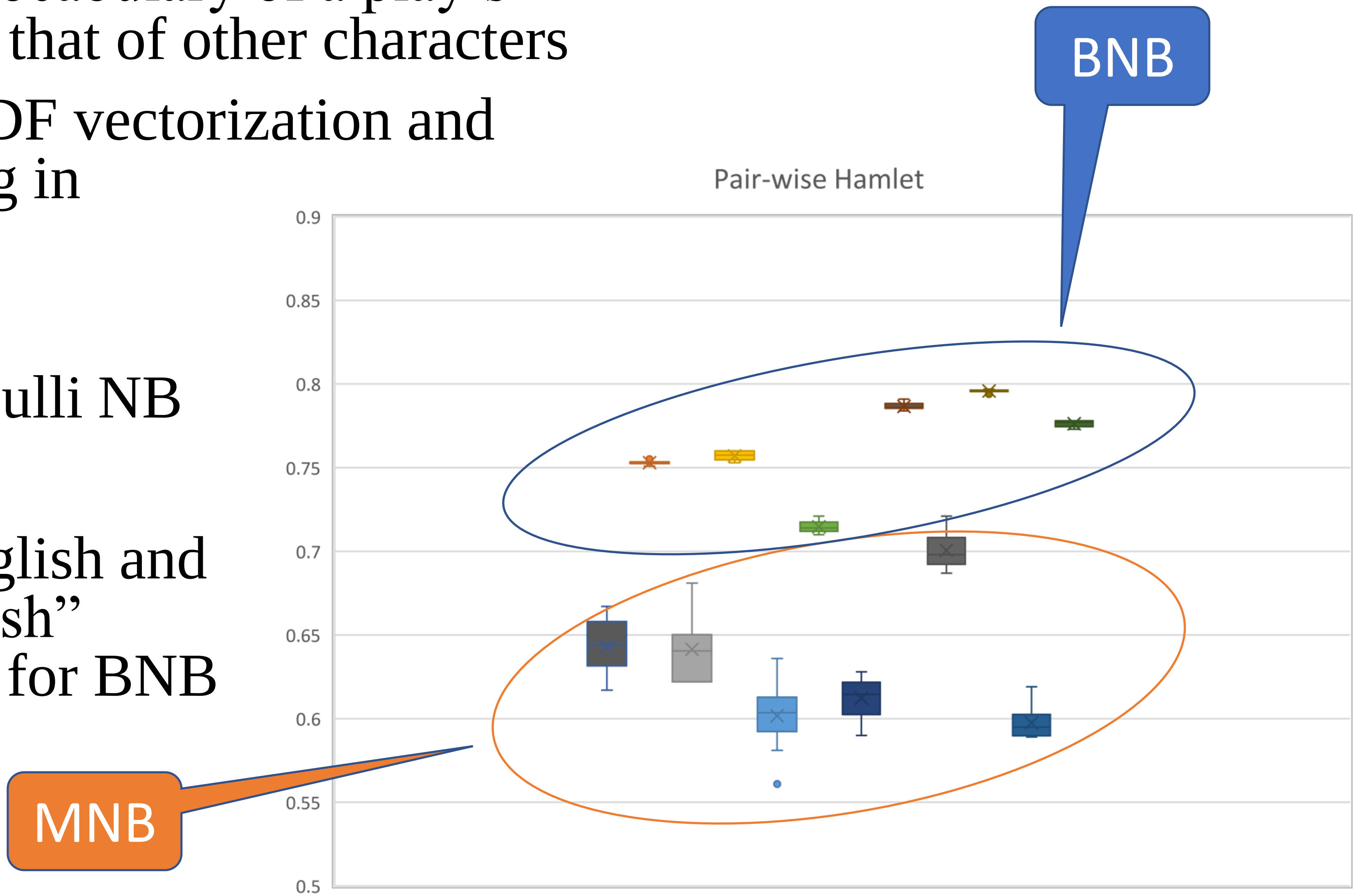
Frequency Distribution: Number of Players by Play



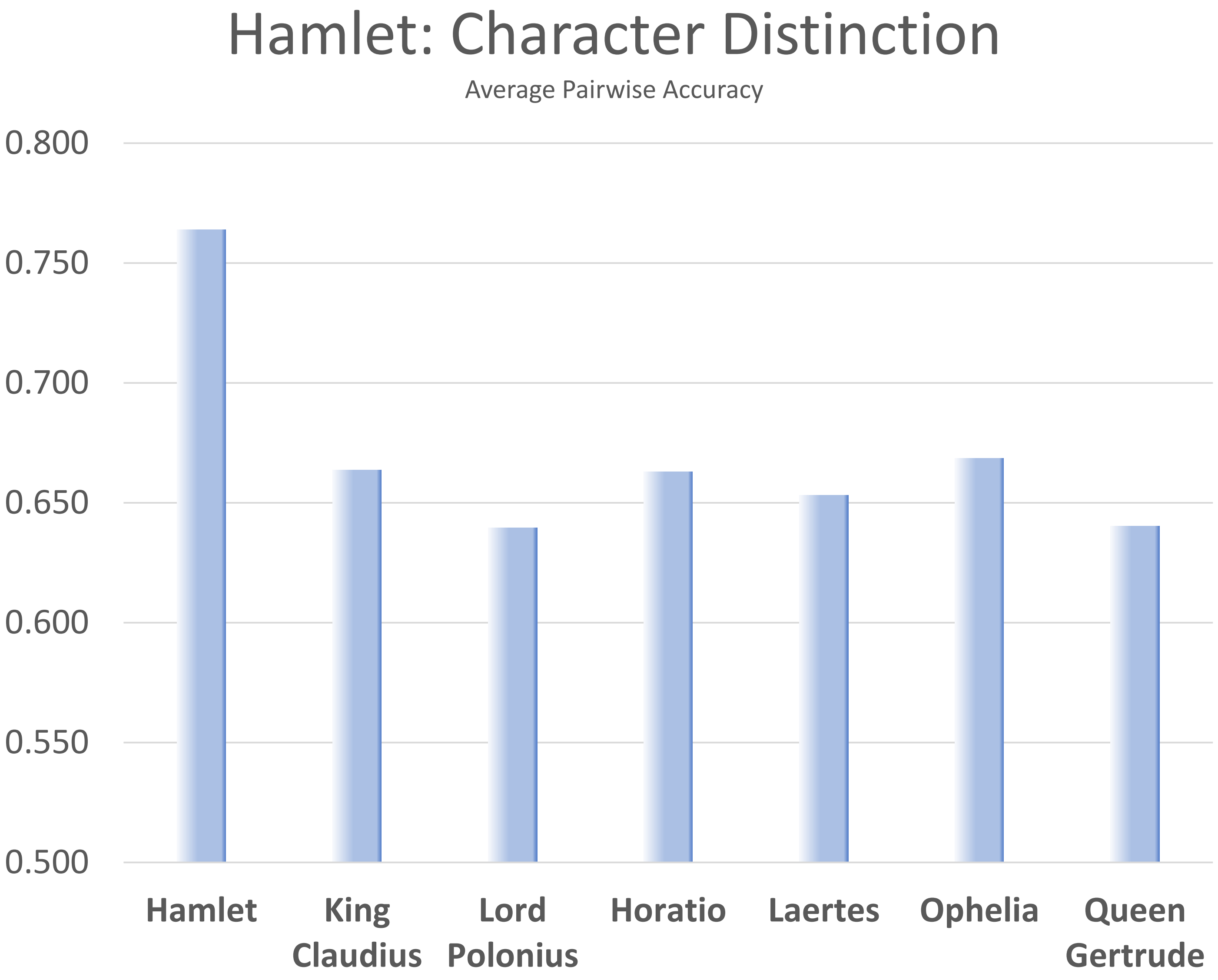


Character Distinction

- The amount that the vocabulary of a play's character differs from that of other characters
- Determined with TFIDF vectorization and Naïve Bayes modeling in pairwise comparisons
- Multinomial vs Bernoulli NB
- Tested removal of English and "Shakespearean English" stopwords: no benefit for BNB

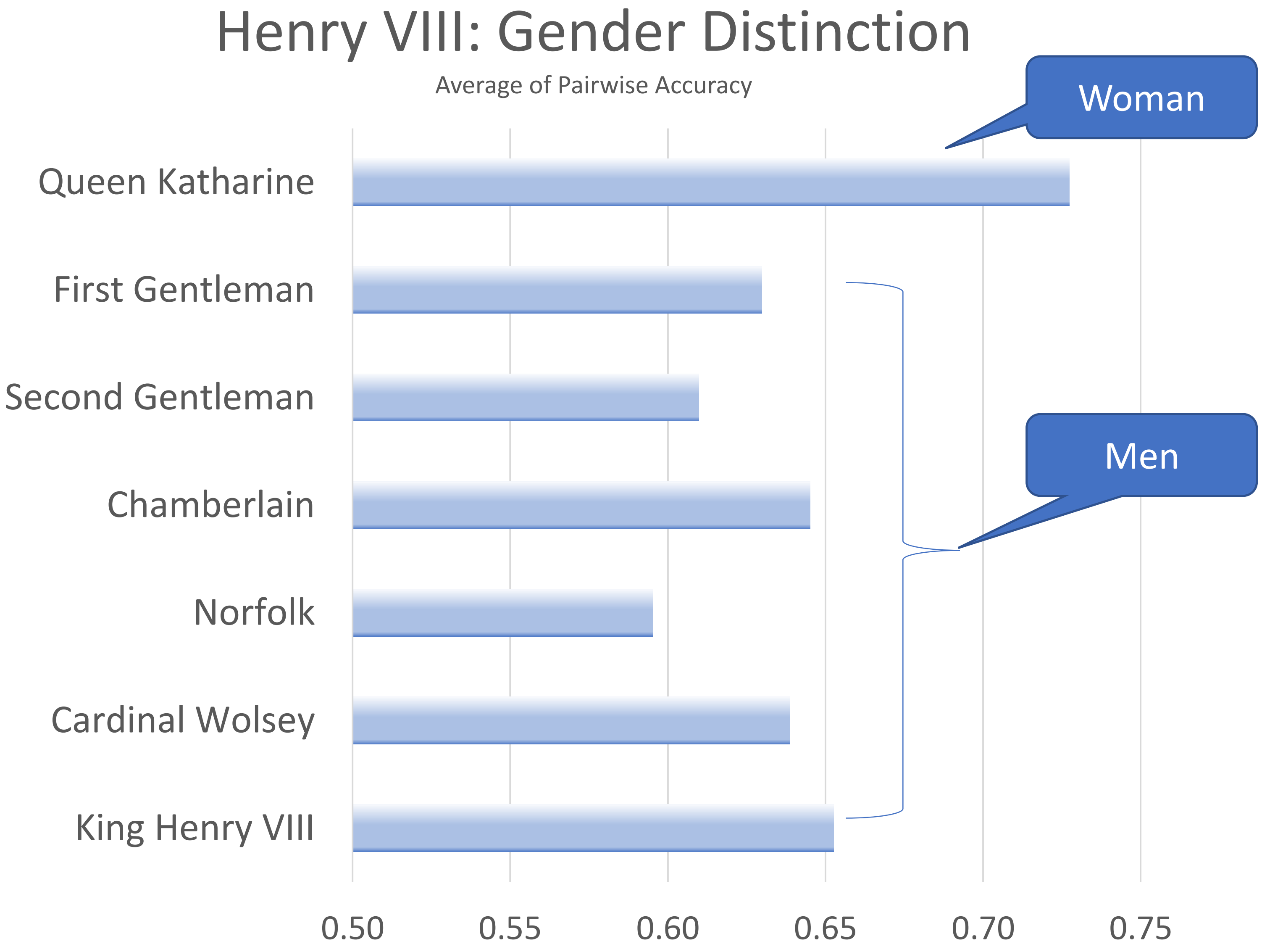


Only Hamlet’s word choices are distinctive

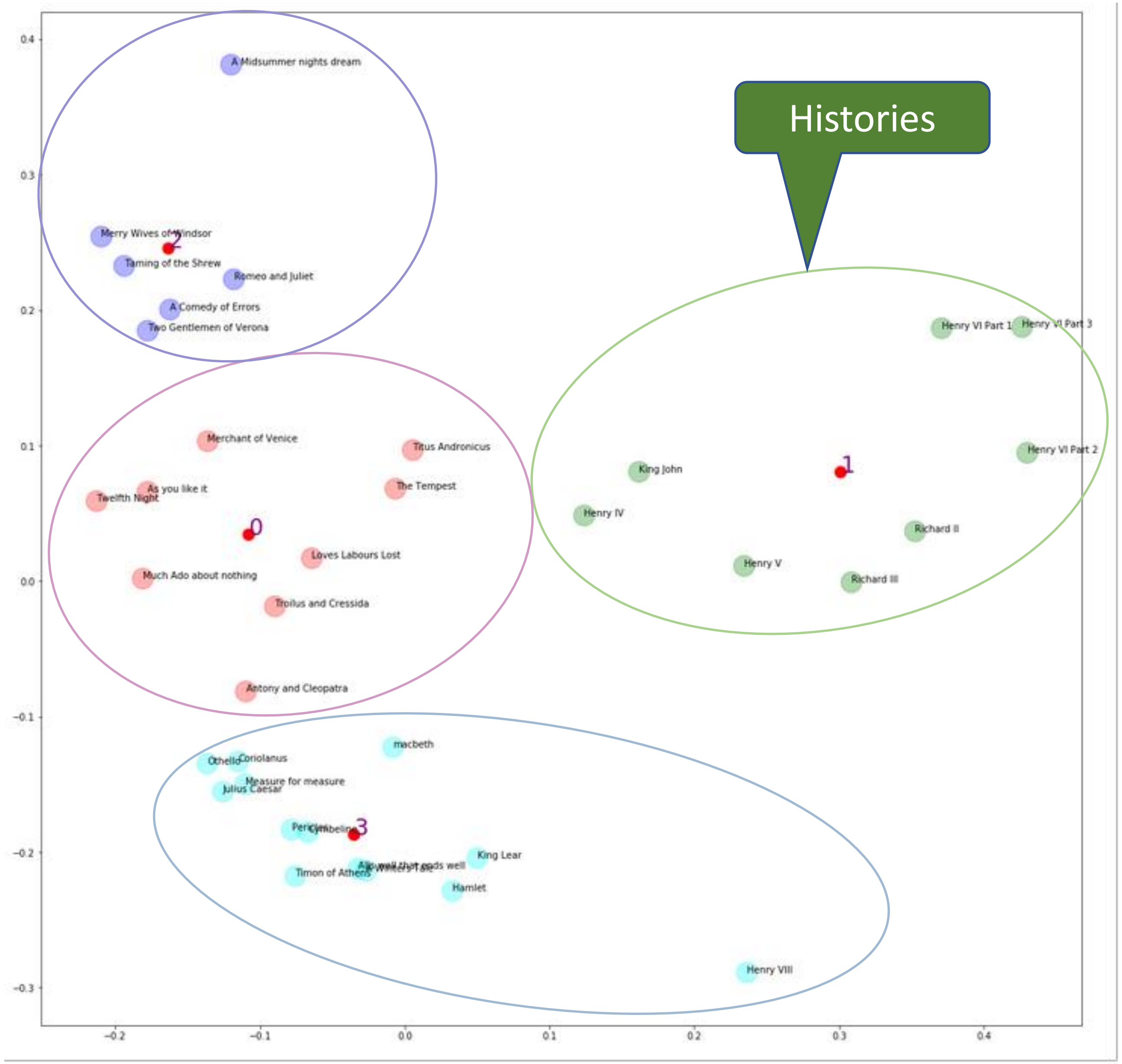


Henry VIII

Only the female character stood out



K-Means Clustering



Histories

Styles of Play

Number of Clusters (K)= 4

Cluster 0: As you like it, The Tempest, Merchant of Venice, Anthony and Cleopatra, Julius Caesar - Comedies, Tragedies, Romances

Cluster 1: King John, Henry VI Part 1-3, Henry IV, Henry V, Richard II, Richard III – The Histories

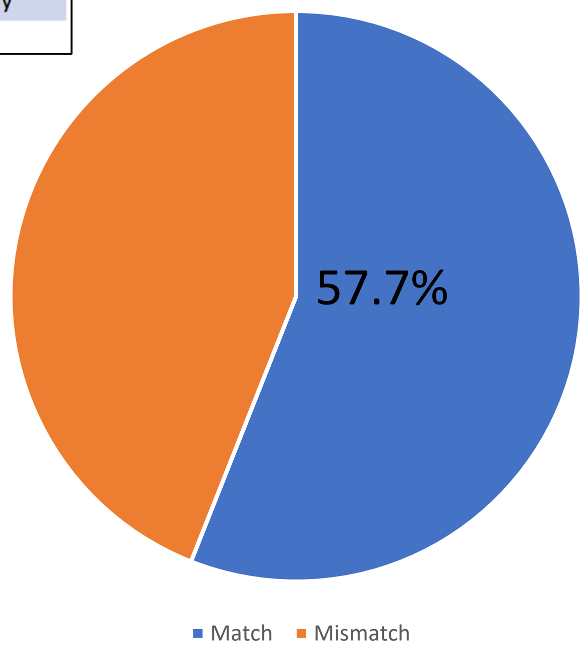
Cluster 2: Hamlet, King Lear, Othello, Winters Tale, Timon of Athens – Tragedies and a Comedy

Cluster 3: Romeo and Juliet, Merry Wives of Windsor, A Midsummer Night’s Dream, Titus Andronicus, Taming of the Shrew, Two gentleman of Verona - Comedies and Tragedies

Comedy vs Tragedy : Topic Modeling (LDA)

Play name	Category	LDA
All's Well That Ends Well	C	Tragedy
Antony and Cleopatra	T	Comedy
As You Like It	C	Tragedy
Comedy of Errors	C	Comedy
Coriolanus	T	Tragedy
Cymbeline	C	Tragedy
Hamlet	T	Tragedy
Julius Caesar	T	Tragedy
King Lear	T	Comedy
Love's Labour's Lost	C	Tragedy
Macbeth	T	Comedy
Measure for Measure	C	Tragedy
Merchant of Venice	C	Comedy
Midsummer Night's Dream	C	Comedy
Much Ado About Nothing	C	Comedy
Othello	T	Tragedy
Pericles	C	Comedy
Romeo and Juliet	T	Tragedy
Taming of the Shrew	C	Comedy
The Tempest	C	Tragedy
Timon of Athens	T	Comedy
Titus Andronicus	T	Tragedy
Troiles and Cressida	C	Tragedy
Twelfth Night	C	Comedy
Two Gentlemen of Verona	C	Comedy
Midsummer Night's Dream	C	Comedy

	LDA-Comedy	LDA-Tragedy
Comedy	9	7
Tragedy	4	6



Topic modeling Accuracy

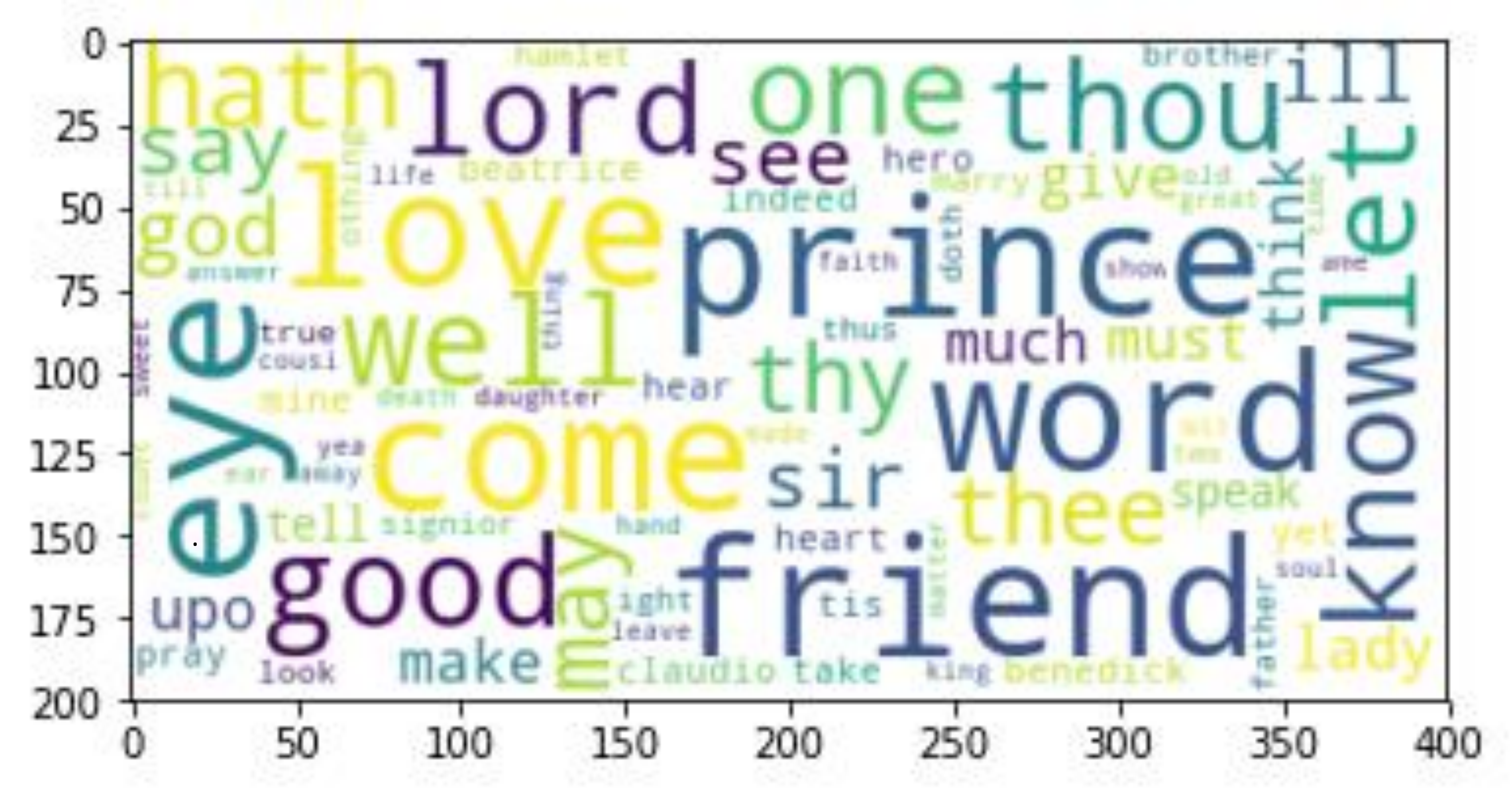
COMEDIES

All's Well That Ends Well
As You Like It
Comedy of Errors
Love's Labour's Lost
Measure for Measure
Merchant of Venice
Merry Wives of Windsor
Midsummer Night's Dream
Much Ado about Nothing
Taming of the Shrew
Tempest
Twelfth Night
Two Gentlemen of Verona
Winter's Tale

TRAGEDIES

Antony and Cleopatra
Coriolanus
Cymbeline
Hamlet
Julius Caesar
King Lear
Macbeth
Othello
Romeo and Juliet
Timon of Athens
Titus Andronicus
Troilus and Cressida

Much Ado… w “Lord, good and shall” top 3 frequency



Much Ado About Stopwords with a k-means Assessment

- Research has shown the importance of stopwords in literary text classification.
- According to Dr. Bei Yu, extremely common words can influence machine learning algorithms. Sentimental novels and plays create complex storylines and words like “my” might be a common word in one collection but not another so they should not be readily excluded (Yu, p. 330).*
- Shakespearean corpuses possess a small inventory of stopwords ~3%.
- TF-IDF & cosine similarity were assessed.
- Excluding stopwords did not impact k-means and SVM algorithm classifications.

Character Grouping with k-means by Vocabulary Usage – 55 Total Characters with 5 Clusters

With Stopwords

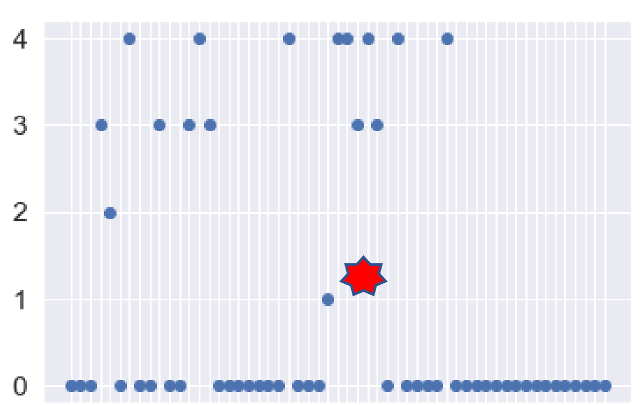
[8] Significant 2nd level cast (4)

[6] main family (3)

[1] Benedict – main character (2)

[1] Hamlet! – curious mistake distance math identified! (1)

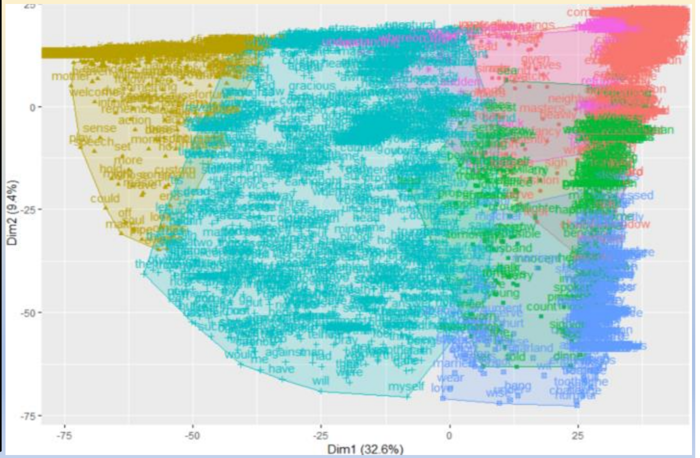
[39] Ensemble cast (0)



Without Stopwords

No change in cast grouping based on words.

K-means visualization by words instead of characters.



*Bei, Y. (2008). An evaluation of text classification methods for literary study. Literary and Linguistic Computing 23(3): 327-34

Conclusions

- Shakespeare *only occasionally* used different vocabularies between characters
- Word choice distinguished the Histories from the other genres, but not between the Comedies and Tragedies
- Multiple modeling techniques showed the vocabulary used in the Comedies and Tragedies do not differ significantly
- Reducing extremely common words with different techniques did not influence prediction outcomes