



To Mine, or Not to Mine

Data Analysis of the Works of
William Shakespeare

IST-736 Text Mining

Summer 2019

Dr. Gates

Jeff Victor, Arthur Fu, Olga Perera, 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](#)

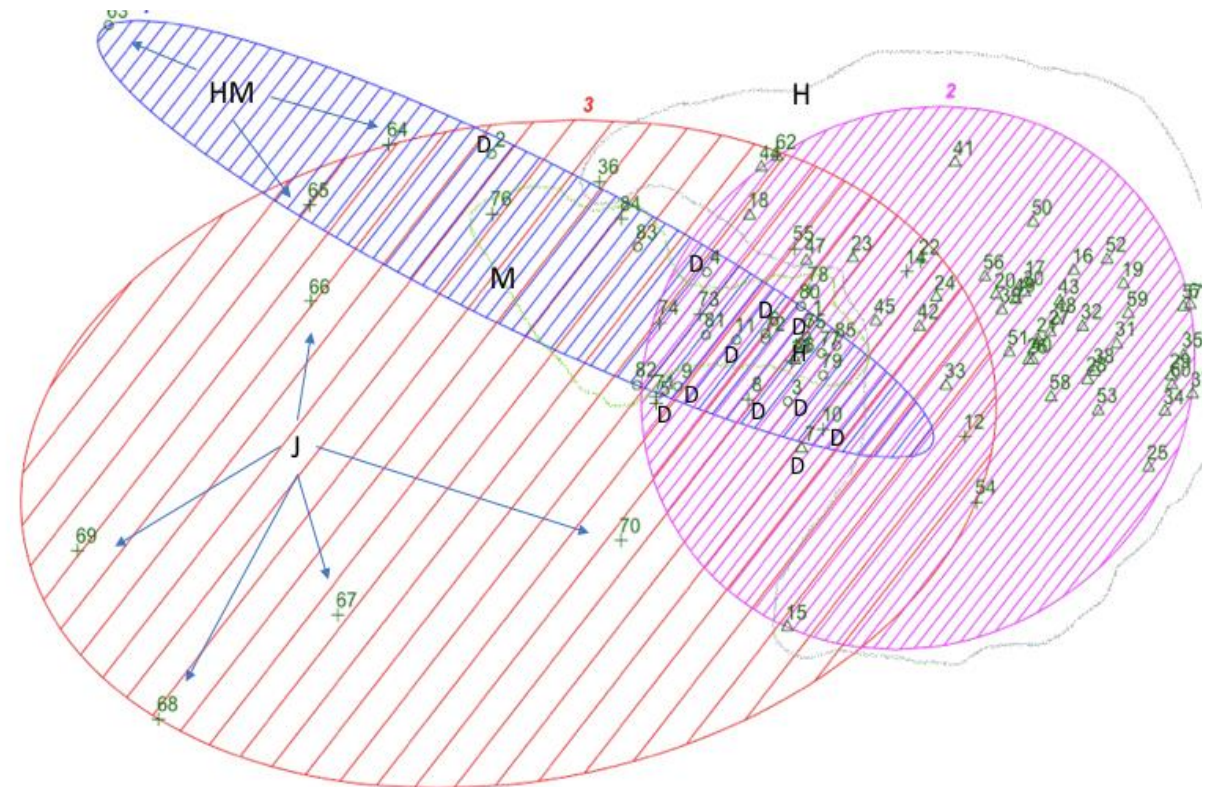
HISTORIES

[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

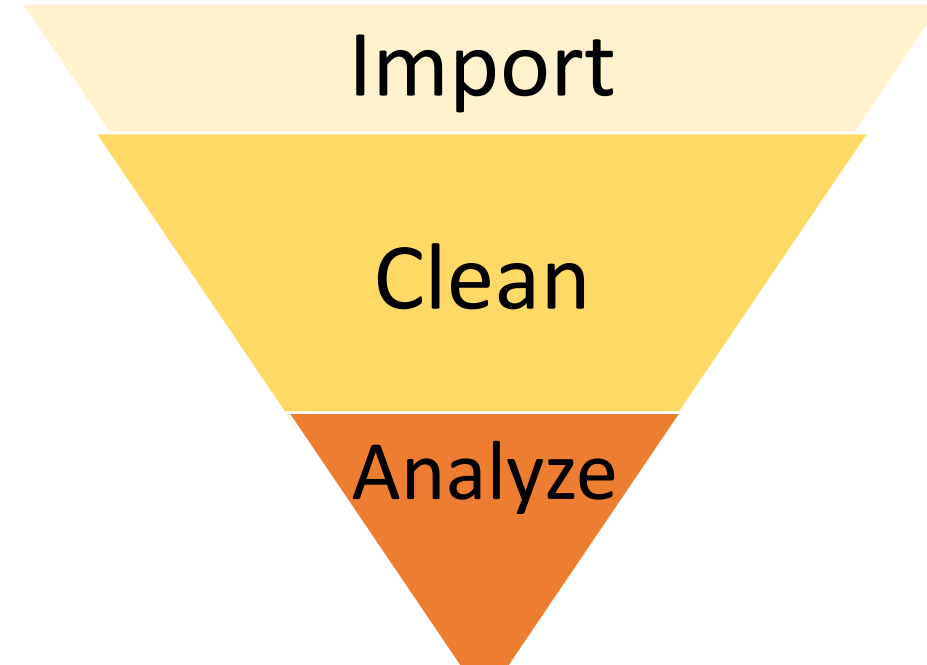
[Antony and Cleopatra](#)
[Coriolanus](#)
[Cymbeline](#)
[Hamlet](#)
[Julius Caesar](#)
[King Lear](#)
[Macbeth](#)
[Othello](#)
[Romeo and Juliet](#)
[Timon of Athens](#)
[Titus Andronicus](#)
[Troilus and Cressida](#)

Corpus



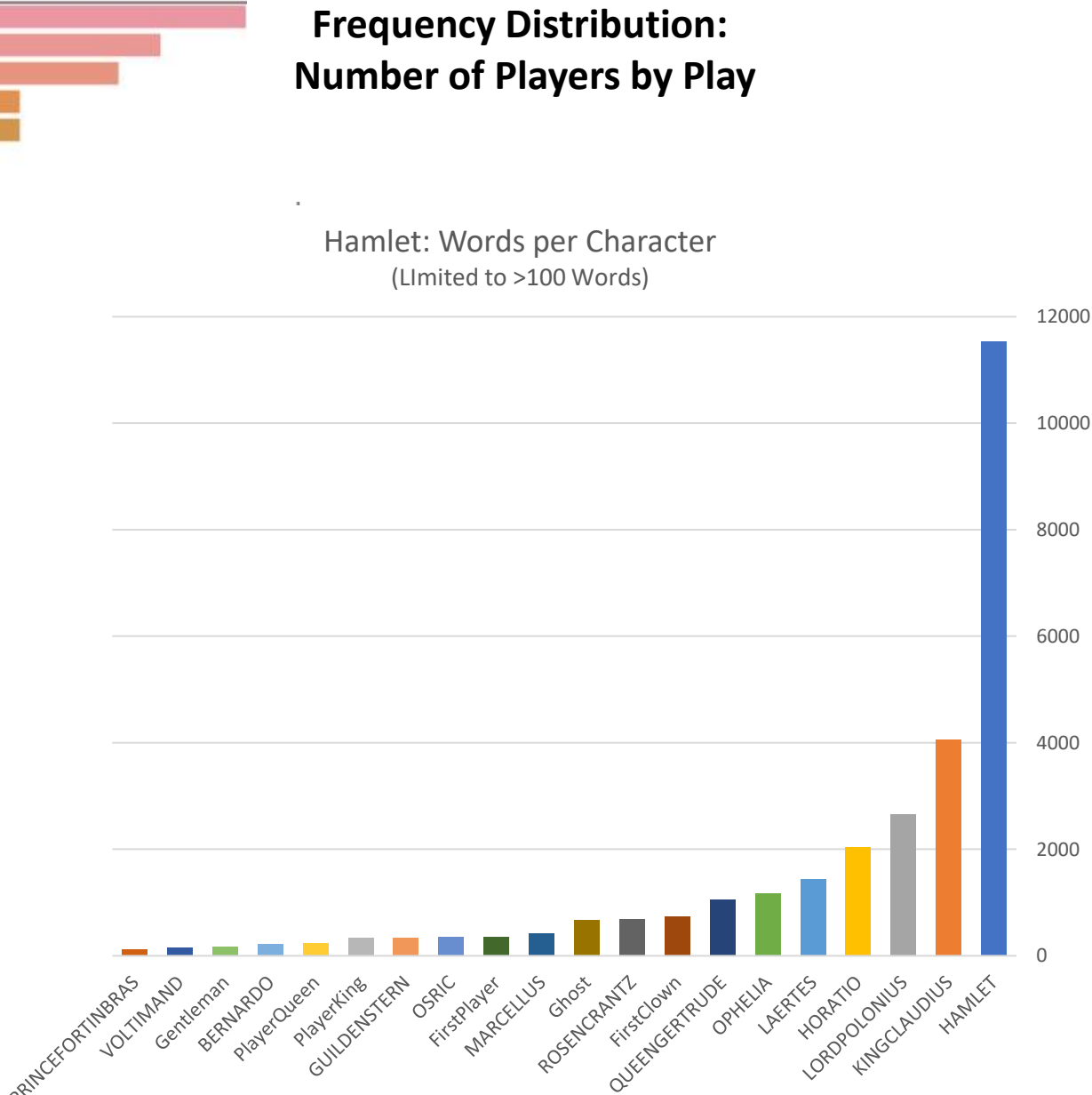
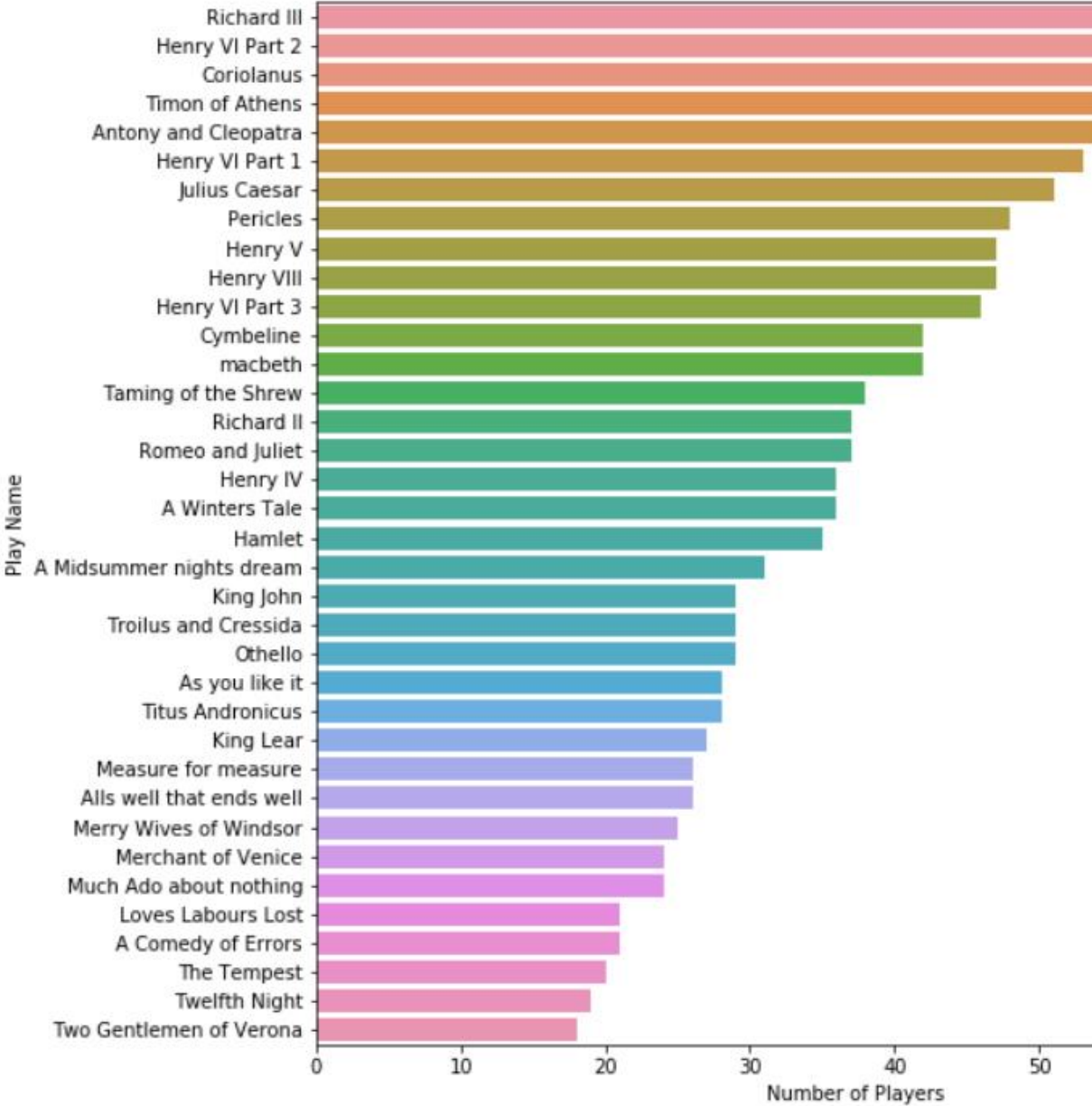
Data Preparation & Distributed Team Tasks

- 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>

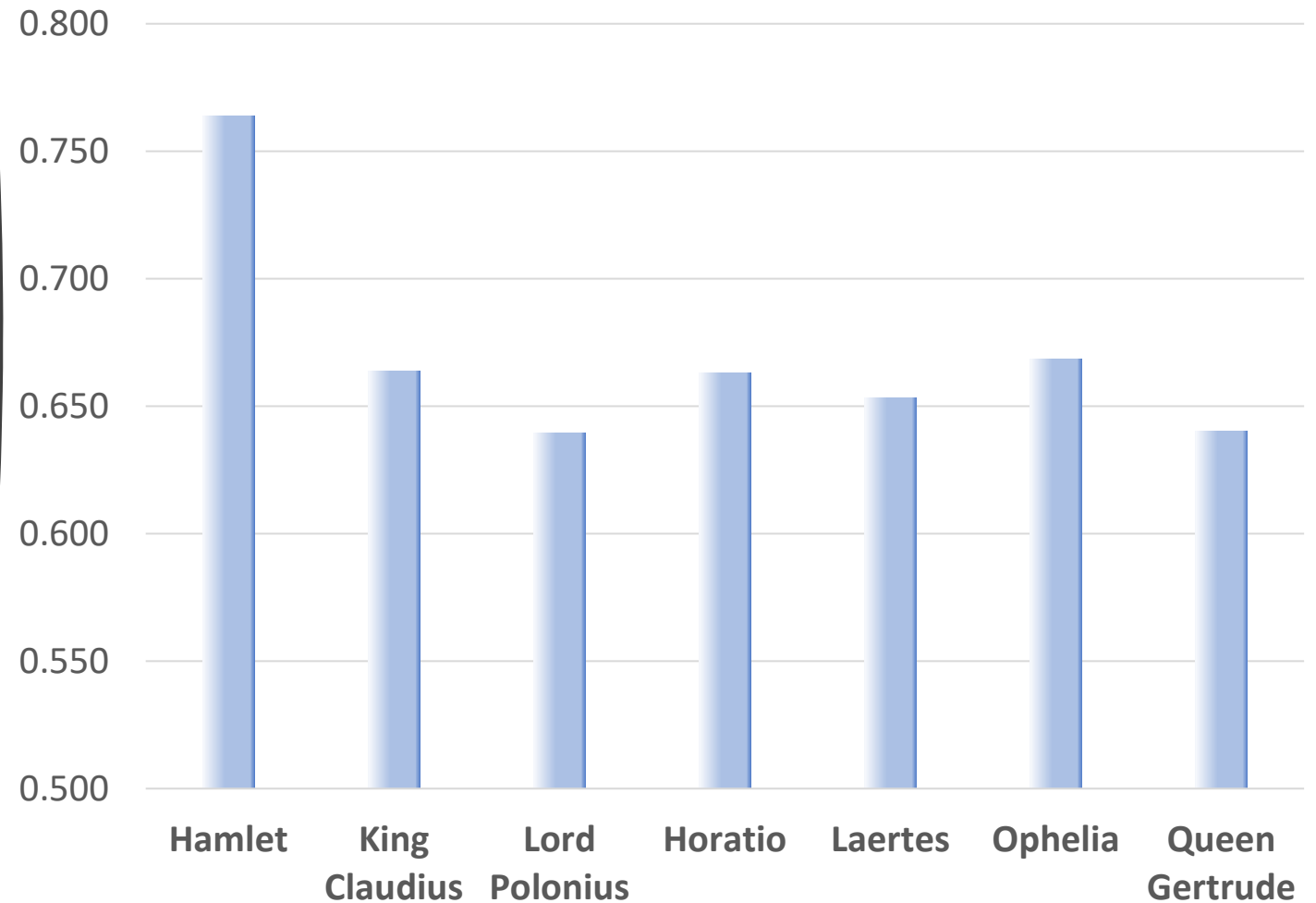
Player and Dialogue Visualizations



Only
Hamlet's
word choices
are distinctive

Hamlet: Character Distinction

Average Pairwise Accuracy

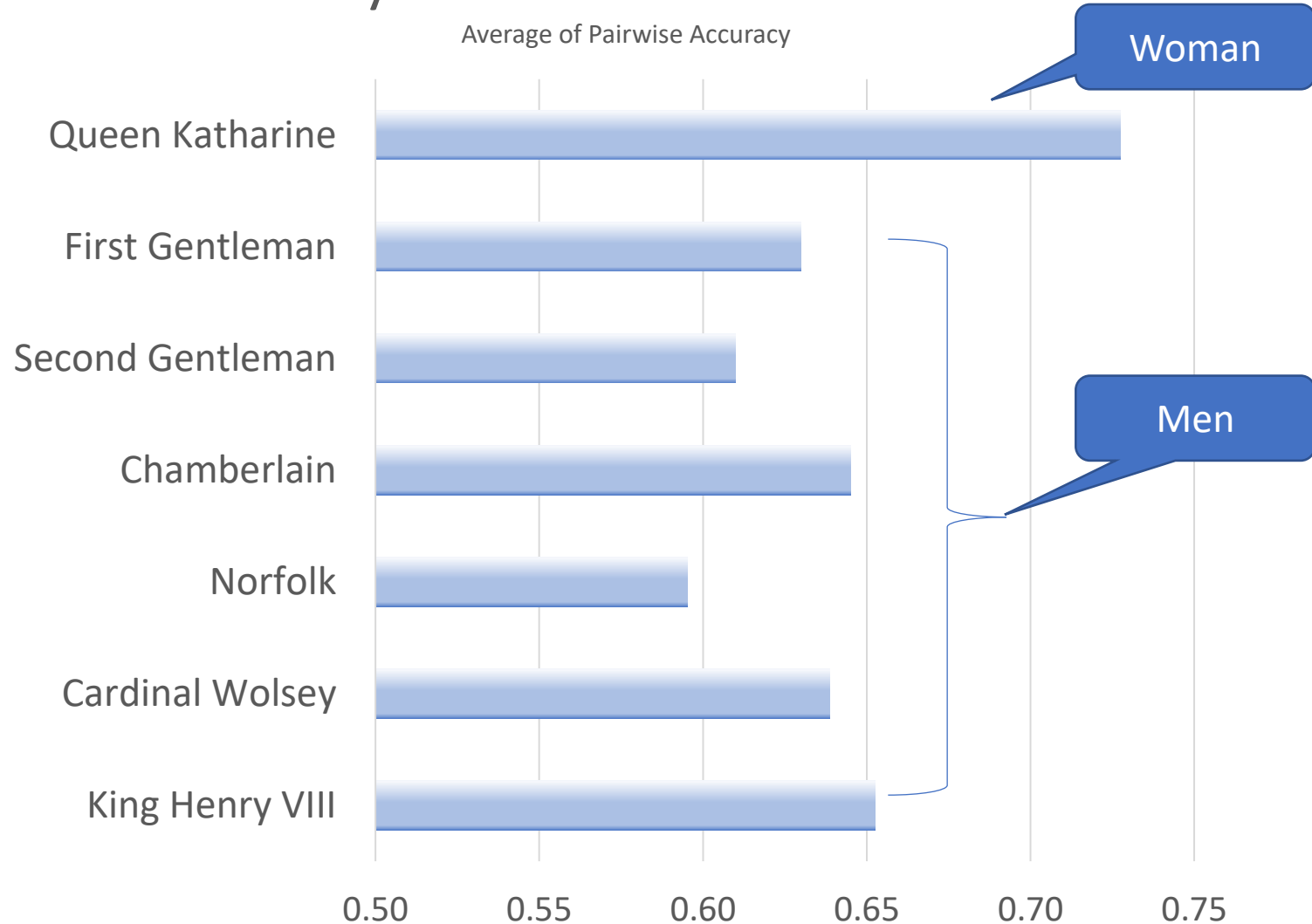


Henry VIII

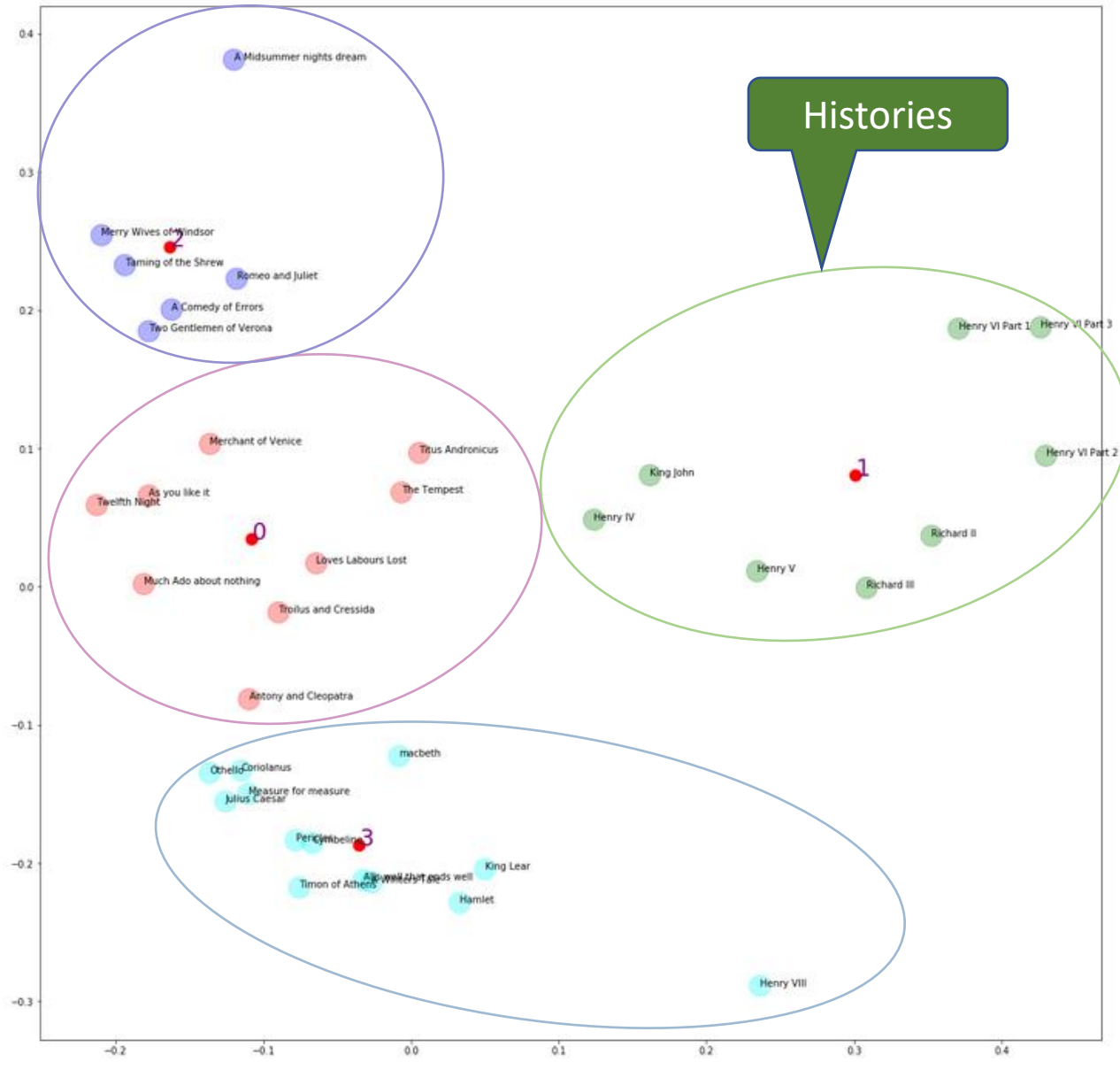
Only the female character stood out

Henry VIII: Gender Distinction

Average of Pairwise Accuracy



K-Means Clustering



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)

Topic one



Topic 2



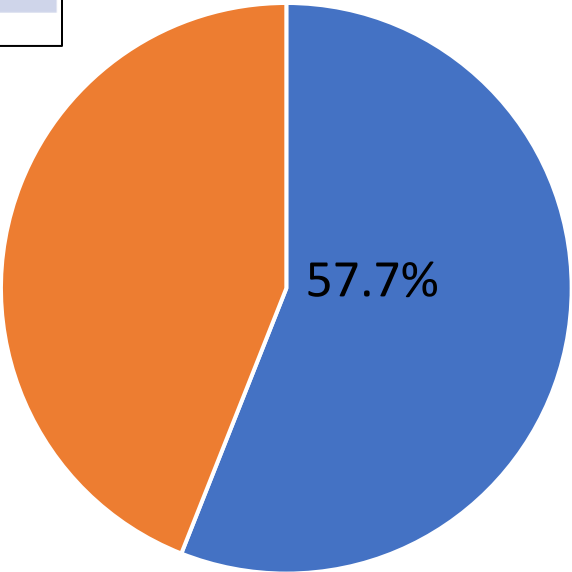
COMEDIES	TRAGEDIES
All's Well That Ends Well	Antony and Cleopatra
As You Like It	Coriolanus
Comedy of Errors	Cymbeline
Love's Labour's Lost	Hamlet
Measure for Measure	Julius Caesar
Merchant of Venice	King Lear
Merry Wives of Windsor	Macbeth
Midsummer Night's Dream	Othello
Much Ado about Nothing	Romeo and Juliet
Taming of the Shrew	Timon of Athens
Tempest	Titus Andronicus
Twelfth Night	Troilus and Cressida
Two Gentlemen of Verona	
Winter's Tale	

Comedy vs Tragedy : Topic Modeling (LDA)

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

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

Topic modeling Accuracy



■ Match ■ Mismatch

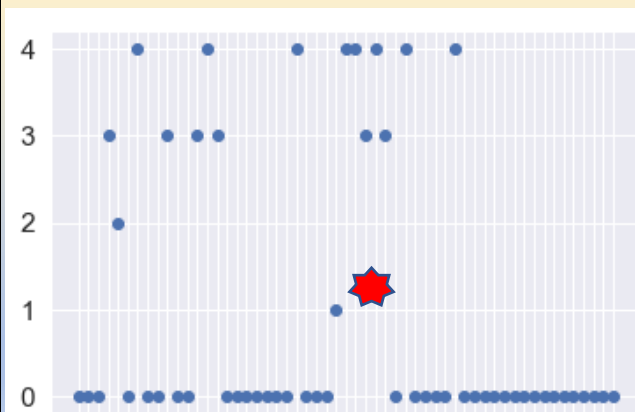
Much Ado About Stopwords with a k-means Assessment

- 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% (average 5 play samples).
- 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

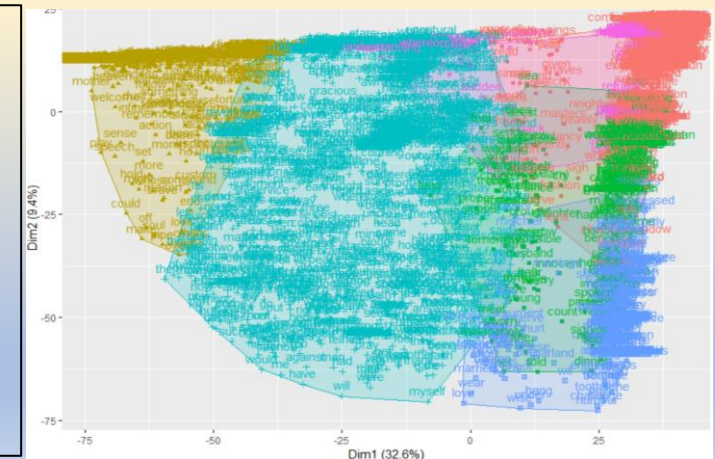
[8] Significant 2nd level cast (4)
[6] main family (3)
[1] Benedict – main character (2)
[1] Hamlet! – curious mistake distance math identified! (1)
[39] Ensemble cast (0)



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

Backup Slides

Backup slides of other methods

Total Words Across all Plays

Word count: 2,252,816

Frequency distribution of most common tokens:

(' ', 150763)

('''', 141184)

('the', 46815)

(''', 45151)

('.', 43273)

('I', 43251)

('and', 36292)

('to', 31590)

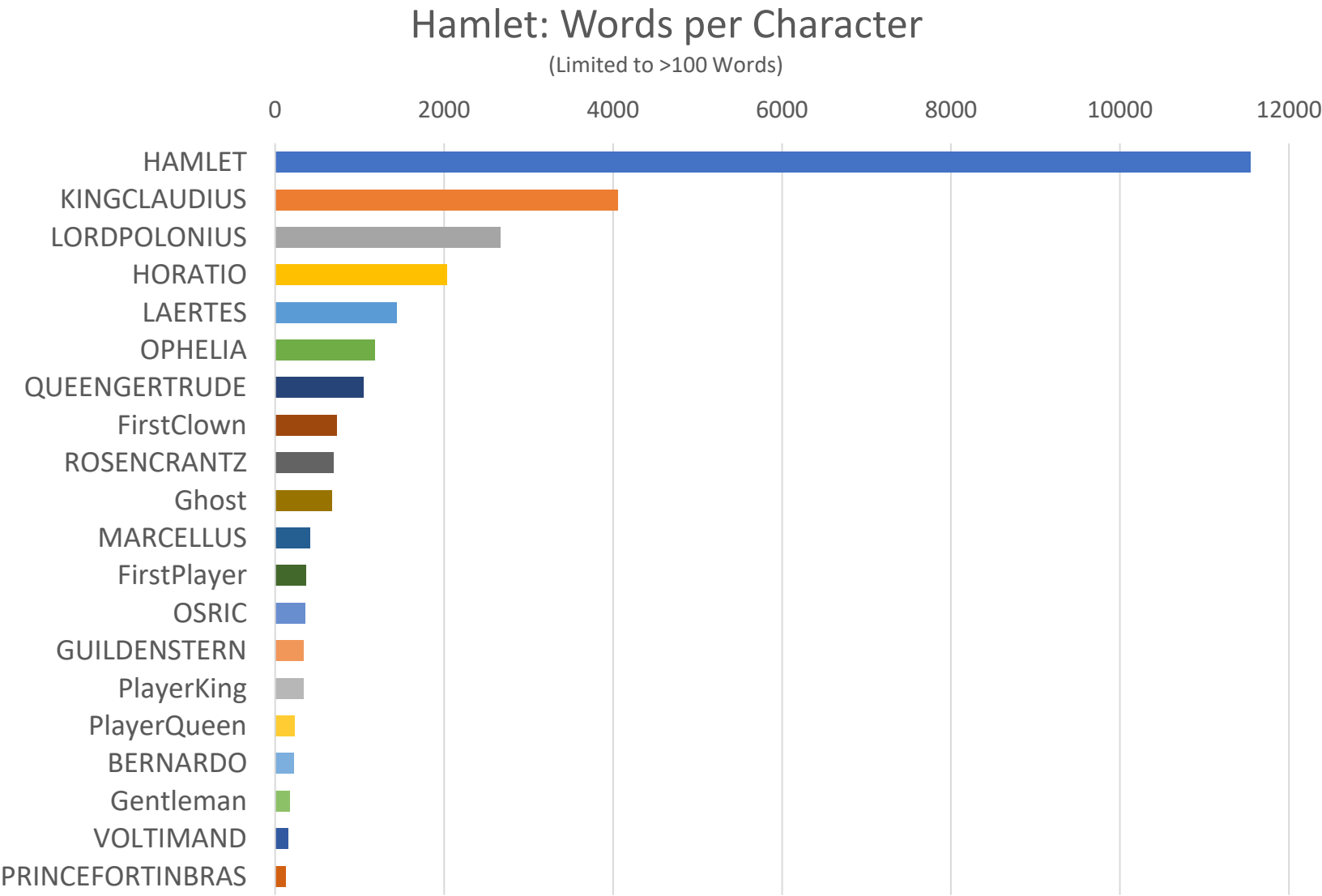
('of', 29857)

('.", 27740)



Remove stopwords and punctuation

Word Distribution: Hamlet



Play: Hamlet

Words per Character

Most: Hamlet
39% of all words

Wordcloud

Much Ado... w "Lord, good and shall" top 3 frequency

