Shakespeare Genre Analysis: Computational Signatures of Tragedy, Comedy, and History

IDS 570: Text as Data - Data Exploration Assignment

Shelly Cao
February 26, 2026

Research Question

To what extent do Shakespeare's genre categories (tragedy, comedy, history) correspond to measurable linguistic patterns in vocabulary, similarity structure, and syntactic complexity?

Corpus

This corpus consists of **21 Shakespeare plays** from the Folger Digital Texts edition:

Tragedies (8)
Hamlet, Macbeth, Othello, King Lear, Romeo and Juliet, Julius Caesar, Antony and Cleopatra, Coriolanus

Comedies (8)
A Midsummer Night's Dream, Much Ado About Nothing, Twelfth Night, As You Like It, The Merchant of Venice, The Taming of the Shrew, The Comedy of Errors, The Tempest

Histories (5)
Henry V, Richard III, Henry IV Part 1, Henry IV Part 2, Richard II

All texts are drawn from the same editorial source to ensure consistency in formatting and transcription.

Methods

This analysis employs three complementary quantitative approaches:

1. **TF–IDF (Term Frequency–Inverse Document Frequency)**
   Identifies lexically distinctive terms within each play relative to the full corpus.
2. **Pearson Correlation**
   Measures pairwise similarity between plays based on shared vocabulary patterns.
3. **Syntactic Complexity Analysis**
   Compares structural features (sentence length, clause density, subordination, coordination, and complex nominals) between selected plays using dependency parsing.

Analytical Goal

Rather than assuming genre categories reflect inherent linguistic structure, this project evaluates whether quantitative textual evidence supports, complicates, or challenges traditional genre classifications. This exploratory approach treats genre as a hypothesis to be tested rather than a given.

Setup & Imports

STEP 0: Load and Normalize Data

Load Texts

	play	genre	text	word_count
0	Hamlet	tragedy	Hamlet\nby William Shakespeare\nEdited by Barb...	32004
1	Macbeth	tragedy	Macbeth\nby William Shakespeare\nEdited by Bar...	18315
2	Othello	tragedy	Othello\nby William Shakespeare\nEdited by Bar...	27945
3	King Lear	tragedy	King Lear\nby William Shakespeare\nEdited by B...	27599
4	Romeo and Juliet	tragedy	Romeo and Juliet\nby William Shakespeare\nEdit...	25913

Text Normalization

Because the corpus consists of Early Modern English texts, normalization need to be applied to balance textual consistency with preservation of linguistic structure.

Normalization Choices

To improve comparability across documents while preserving rhetorical content, the following preprocessing steps were applied:

- Long s (ſ) normalization All instances of the long s character (ſ) were replaced with standard modern "s" to prevent tokenization errors and artificial feature splitting (e.g., *ſin* vs *sin*).
- Removed editorial metadata preceding the start of the play (before *ACT 1*).
- Removed separator lines and formatting artifacts.
- Removed stage directions enclosed in brackets (e.g., `[Enter Barnardo]`).
- Removed ACT and SCENE headers.
- Removed speaker labels (e.g., `HAMLET`, `BARNARDO`) both as standalone lines and when preceding dialogue.
- Converted all text to lowercase.
- Standardized whitespace.

```
Text Example After Normalization:

who's there?

nay, answer me. stand and unfold yourself.
long live the king!
barnardo?
he.

you come most carefully upon your hour.

'tis now struck twelve. get thee to bed, francisco.

for this relie
```

STEP 1: TF-IDF Analysis

Calculate TF-IDF

```
(21 plays × 5000 features)
```

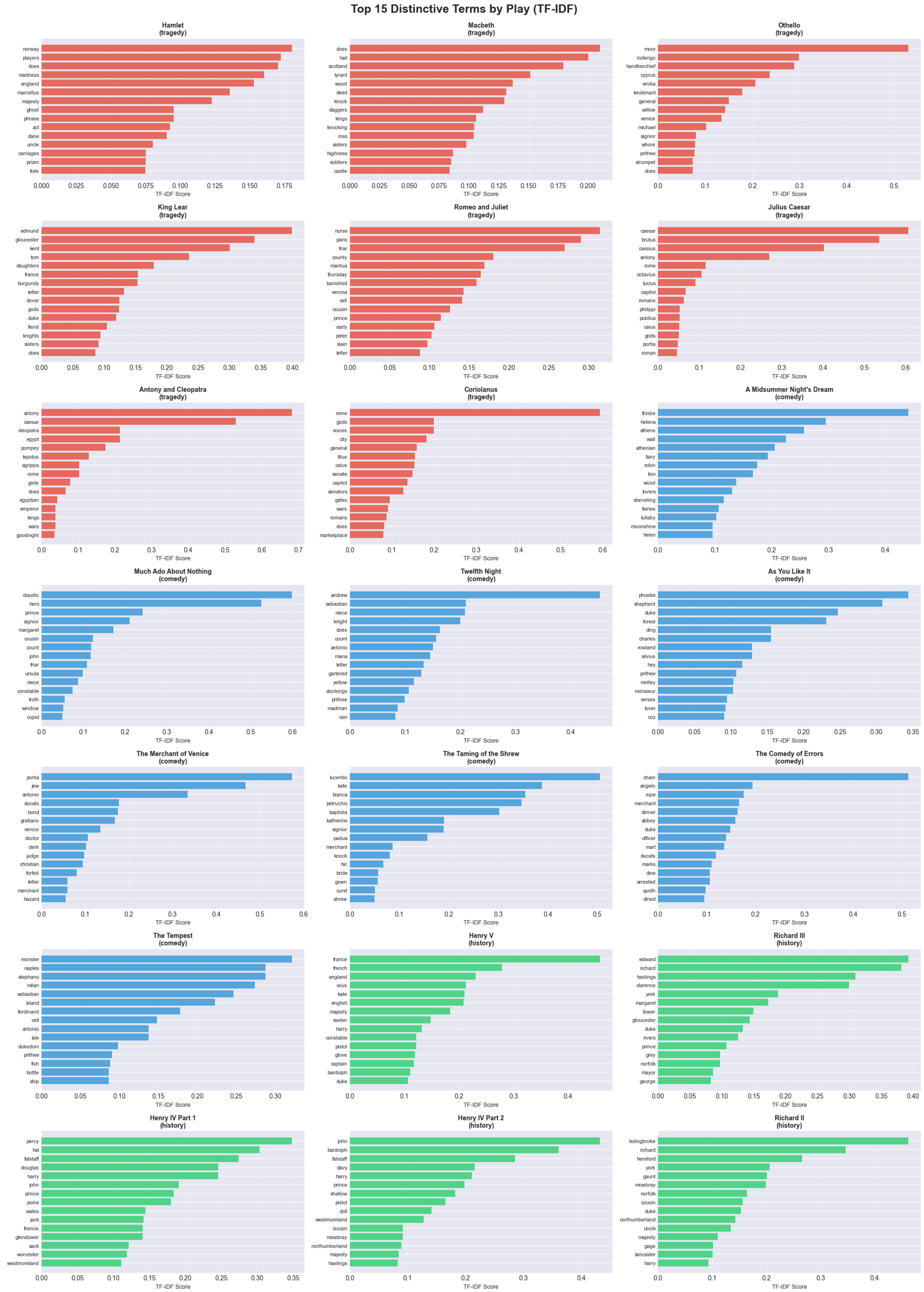
Extract Top TF-IDF Terms per Play

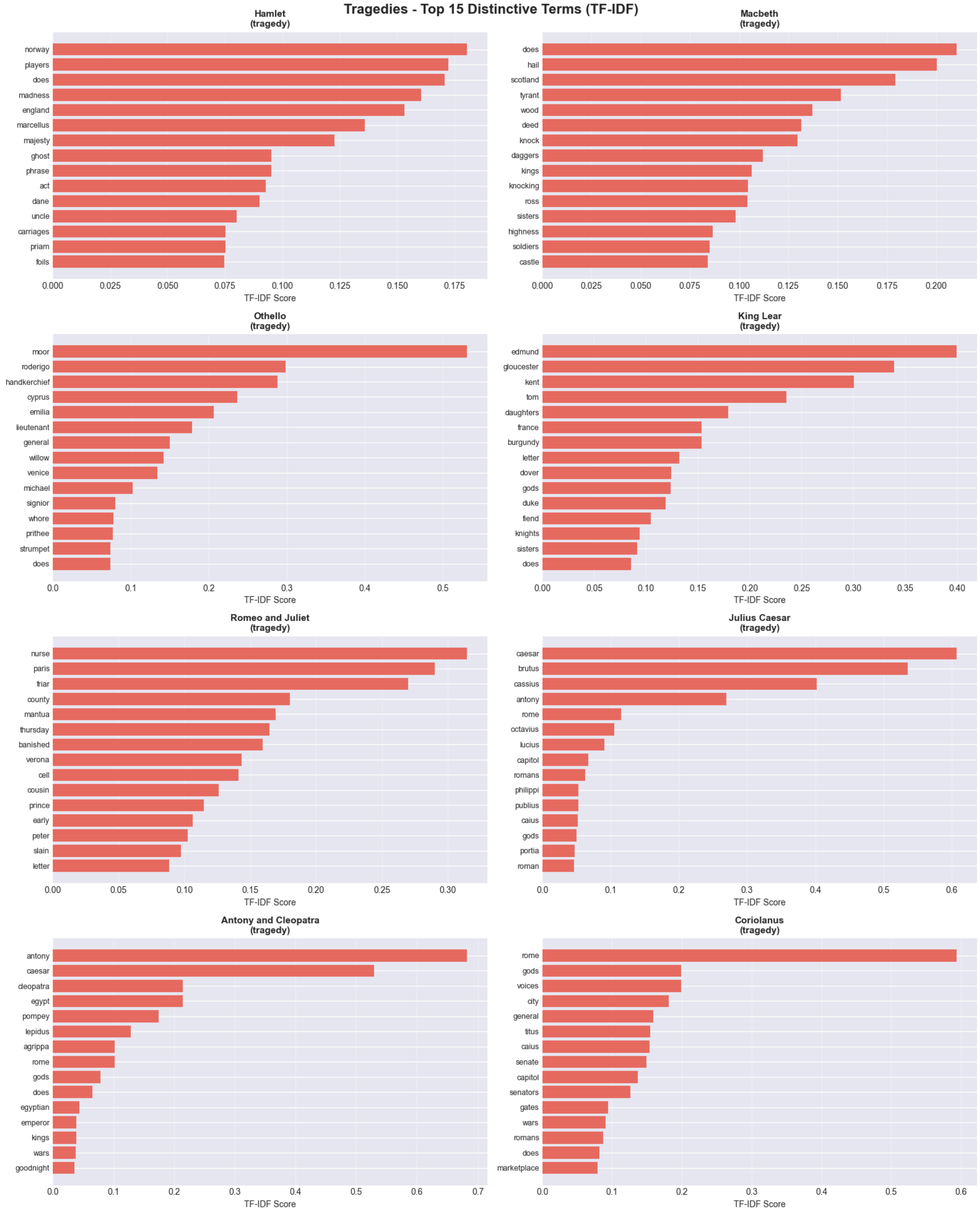
```
TOP 15 TF-IDF TERMS PER PLAY
================================================================================

Hamlet (TRAGEDY)
------------------------------------------------------------
norway(0.180), players(0.172), does(0.171), madness(0.160), england(0.153), marcellus(0.136), majesty(0.123), ghost
(0.095), phrase(0.095), act(0.093), dane(0.090), uncle(0.080), carriages(0.075), priam(0.075), foils(0.075)

Macbeth (TRAGEDY)
------------------------------------------------------------
does(0.210), hail(0.201), scotland(0.179), tyrant(0.152), wood(0.137), deed(0.131), knock(0.130), daggers(0.112), k
ings(0.106), knocking(0.105), ross(0.104), sisters(0.098), highness(0.087), soldiers(0.085), castle(0.084)

Othello (TRAGEDY)
------------------------------------------------------------
moor(0.531), roderigo(0.299), handkerchief(0.288), cyprus(0.237), emilia(0.206), lieutenant(0.179), general(0.150),
willow(0.142), venice(0.135), michael(0.102), signior(0.080), whore(0.078), prithee(0.077), strumpet(0.074), does
(0.074)

King Lear (TRAGEDY)
------------------------------------------------------------
edmund(0.399), gloucester(0.339), kent(0.301), tom(0.235), daughters(0.179), france(0.154), burgundy(0.154), letter
(0.132), dover(0.125), gods(0.124), duke(0.119), fiend(0.105), knights(0.094), sisters(0.091), does(0.086)

Romeo and Juliet (TRAGEDY)
------------------------------------------------------------
nurse(0.314), paris(0.290), friar(0.270), county(0.180), mantua(0.169), thursday(0.165), banished(0.159), verona(0.
143), cell(0.141), cousin(0.126), prince(0.115), early(0.106), peter(0.103), slain(0.098), letter(0.088)

Julius Caesar (TRAGEDY)
------------------------------------------------------------
caesar(0.607), brutus(0.535), cassius(0.402), antony(0.270), rome(0.115), octavius(0.105), lucius(0.091), capitol
(0.067), romans(0.063), philippi(0.053), publius(0.053), caius(0.052), gods(0.050), portia(0.048), roman(0.046)

Antony and Cleopatra (TRAGEDY)
------------------------------------------------------------
antony(0.682), caesar(0.529), cleopatra(0.214), egypt(0.214), pompey(0.175), lepidus(0.129), agrippa(0.102), rome
(0.102), gods(0.079), does(0.066), egyptian(0.044), emperor(0.039), kings(0.038), wars(0.038), goodnight(0.036)

Coriolanus (TRAGEDY)
------------------------------------------------------------
rome(0.593), gods(0.199), voices(0.199), city(0.182), general(0.159), titus(0.155), caius(0.154), senate(0.149), ca
pitol(0.137), senators(0.127), gates(0.095), wars(0.091), romans(0.087), does(0.082), marketplace(0.079)

A Midsummer Night's Dream (COMEDY)
------------------------------------------------------------
thisbe(0.441), helena(0.295), athens(0.257), wall(0.225), athenian(0.205), fairy(0.193), robin(0.175), lion(0.167),
wood(0.138), lovers(0.131), starveling(0.115), fairies(0.106), lullaby(0.103), moonshine(0.096), helen(0.096)

Much Ado About Nothing (COMEDY)
------------------------------------------------------------
claudio(0.599), hero(0.526), prince(0.242), signior(0.211), margaret(0.172), cousin(0.123), count(0.119), john(0.11
8), friar(0.109), ursula(0.098), niece(0.088), constable(0.075), troth(0.055), window(0.052), cupid(0.050)

Twelfth Night (COMEDY)
------------------------------------------------------------
andrew(0.453), sebastian(0.210), niece(0.209), knight(0.200), does(0.163), count(0.156), antonio(0.151), maria(0.14
6), letter(0.133), gartered(0.129), yellow(0.116), stockings(0.107), prithee(0.100), madman(0.087), rain(0.082)

As You Like It (COMEDY)
------------------------------------------------------------
phoebe(0.344), shepherd(0.308), duke(0.247), forest(0.231), ding(0.155), charles(0.155), rowland(0.129), silvius(0.
129), hey(0.116), prithee(0.107), motley(0.104), monsieur(0.103), verses(0.095), lover(0.093), coz(0.091)

The Merchant of Venice (COMEDY)
------------------------------------------------------------
portia(0.573), jew(0.467), antonio(0.335), ducats(0.178), bond(0.175), gratiano(0.167), venice(0.135), doctor(0.10
6), clerk(0.102), judge(0.098), christian(0.094), forfeit(0.080), letter(0.060), merchant(0.059), hazard(0.055)

The Taming of the Shrew (COMEDY)
------------------------------------------------------------
lucentio(0.506), kate(0.388), bianca(0.355), petruchio(0.347), baptista(0.302), katherine(0.191), signior(0.189), p
adua(0.157), merchant(0.086), knock(0.081), hic(0.068), bride(0.058), gown(0.057), curst(0.051), shrew(0.050)

The Comedy of Errors (COMEDY)
------------------------------------------------------------
chain(0.514), angelo(0.194), rope(0.175), merchant(0.166), dinner(0.163), abbey(0.159), duke(0.148), officer(0.14
0), mart(0.136), ducats(0.119), marks(0.110), dine(0.107), arrested(0.106), quoth(0.098), dined(0.095)

The Tempest (COMEDY)
------------------------------------------------------------
monster(0.321), naples(0.288), stephano(0.288), milan(0.274), sebastian(0.247), island(0.223), ferdinand(0.178), ce
ll(0.149), antonio(0.138), isle(0.138), dukedom(0.098), prithee(0.091), fish(0.088), bottle(0.087), ship(0.087)

Henry V (HISTORY)
------------------------------------------------------------
france(0.459), french(0.279), england(0.230), vous(0.213), kate(0.210), english(0.208), majesty(0.184), exeter(0.14
8), harry(0.132), constable(0.122), pistol(0.122), glove(0.120), captain(0.117), bardolph(0.111), duke(0.107)
```

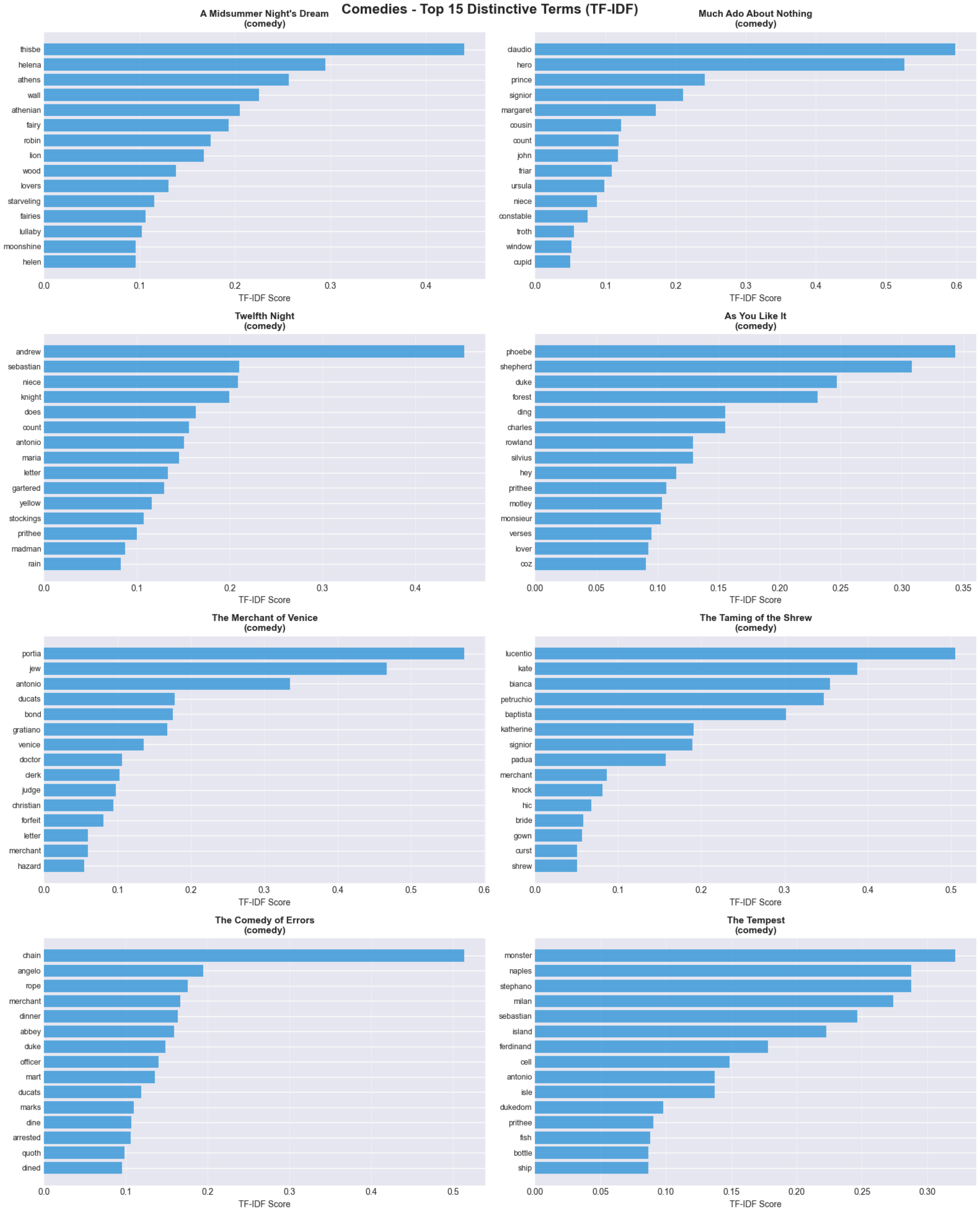
```
Richard III (HISTORY)
------------------------------------------------------------
edward(0.393), richard(0.382), hastings(0.309), clarence(0.299), york(0.188), margaret(0.173), tower(0.149), gloucester(0.144), duke(0.133), rivers(0.125), prince(0.108), grey(0.098), norfolk(0.098), mayor(0.086), george(0.083)

Henry IV Part 1 (HISTORY)
------------------------------------------------------------
percy(0.348), hal(0.303), falstaff(0.274), douglas(0.246), harry(0.245), john(0.191), prince(0.184), poins(0.180), wales(0.145), jack(0.142), francis(0.141), glendower(0.141), sack(0.121), worcester(0.118), westmoreland(0.111)

Henry IV Part 2 (HISTORY)
------------------------------------------------------------
john(0.432), bardolph(0.361), falstaff(0.286), davy(0.216), harry(0.211), prince(0.198), shallow(0.182), pistol(0.165), doll(0.141), westmoreland(0.128), cousin(0.092), mowbray(0.091), northumberland(0.089), majesty(0.085), hastings(0.083)

Richard II (HISTORY)
------------------------------------------------------------
bolingbroke(0.462), richard(0.346), hereford(0.266), york(0.206), gaunt(0.201), mowbray(0.199), norfolk(0.164), cousin(0.157), duke(0.153), northumberland(0.143), uncle(0.135), majesty(0.110), gage(0.102), lancaster(0.101), harry(0.093)
```

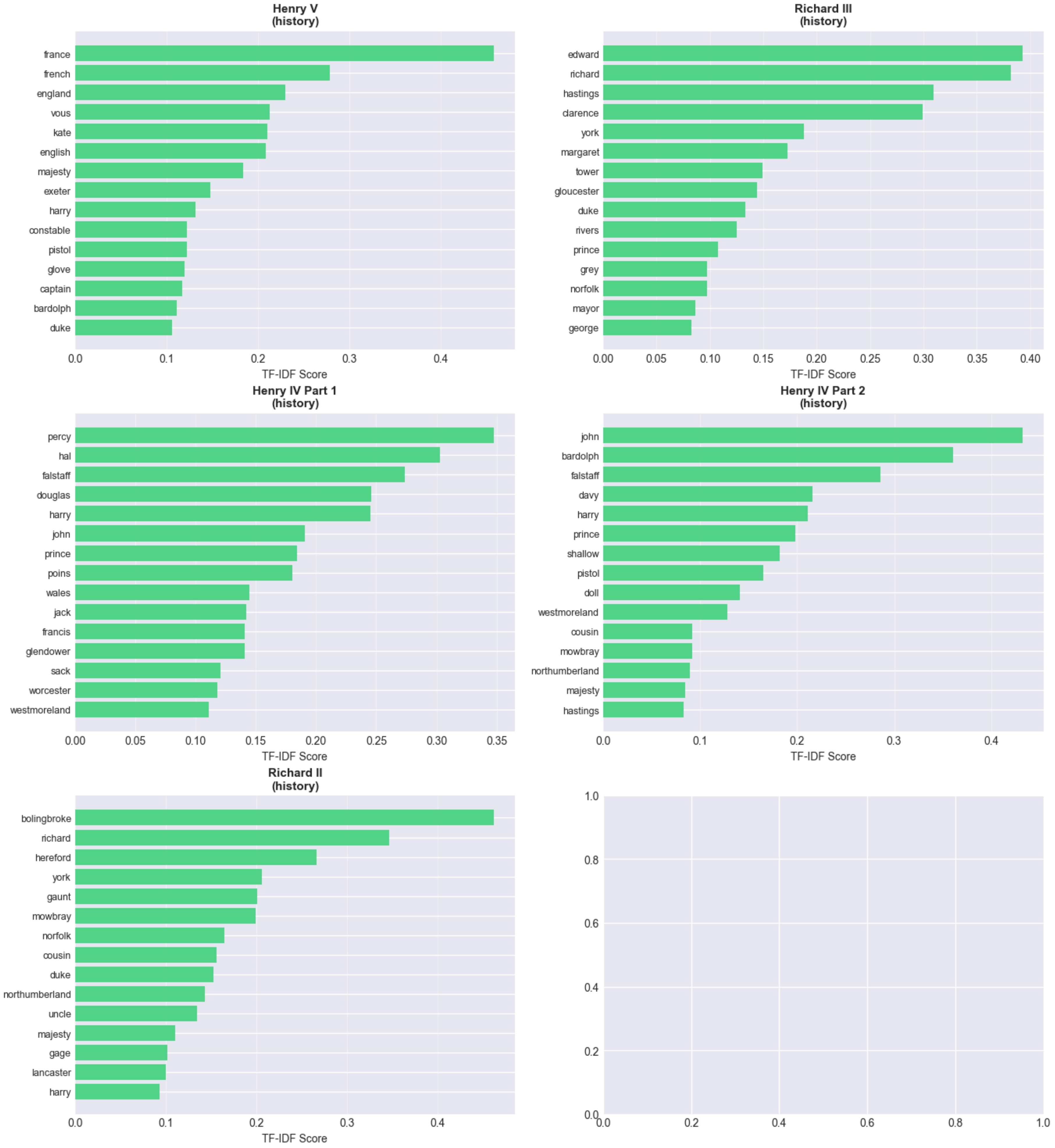
Create TF-IDF Summary Table

```
                     Play   Genre                                                                     Top 10 Distinctive Terms
                   Hamlet tragedy             norway, players, does, madness, england, marcellus, majesty, ghost, phrase, act
                  Macbeth tragedy                   does, hail, scotland, tyrant, wood, deed, knock, daggers, kings, knocking
                  Othello tragedy  moor, roderigo, handkerchief, cyprus, emilia, lieutenant, general, willow, venice, michael
                King Lear tragedy             edmund, gloucester, kent, tom, daughters, france, burgundy, letter, dover, gods
         Romeo and Juliet tragedy               nurse, paris, friar, county, mantua, thursday, banished, verona, cell, cousin
            Julius Caesar tragedy          caesar, brutus, cassius, antony, rome, octavius, lucius, capitol, romans, philippi
     Antony and Cleopatra tragedy                antony, caesar, cleopatra, egypt, pompey, lepidus, agrippa, rome, gods, does
               Coriolanus tragedy                  rome, gods, voices, city, general, titus, caius, senate, capitol, senators
A Midsummer Night's Dream  comedy                    thisbe, helena, athens, wall, athenian, fairy, robin, lion, wood, lovers
   Much Ado About Nothing  comedy                claudio, hero, prince, signior, margaret, cousin, count, john, friar, ursula
            Twelfth Night  comedy             andrew, sebastian, niece, knight, does, count, antonio, maria, letter, gartered
           As You Like It  comedy               phoebe, shepherd, duke, forest, ding, charles, rowland, silvius, hey, prithee
   The Merchant of Venice  comedy                  portia, jew, antonio, ducats, bond, gratiano, venice, doctor, clerk, judge
  The Taming of the Shrew  comedy     lucentio, kate, bianca, petruchio, baptista, katherine, signior, padua, merchant, knock
     The Comedy of Errors  comedy                   chain, angelo, rope, merchant, dinner, abbey, duke, officer, mart, ducats
              The Tempest  comedy         monster, naples, stephano, milan, sebastian, island, ferdinand, cell, antonio, isle
                  Henry V history             france, french, england, vous, kate, english, majesty, exeter, harry, constable
              Richard III history        edward, richard, hastings, clarence, york, margaret, tower, gloucester, duke, rivers
          Henry IV Part 1 history                      percy, hal, falstaff, douglas, harry, john, prince, poins, wales, jack
          Henry IV Part 2 history          john, bardolph, falstaff, davy, harry, prince, shallow, pistol, doll, westmoreland
               Richard II history bolingbroke, richard, hereford, york, gaunt, mowbray, norfolk, cousin, duke, northumberland
```


Top 15 Distinctive Terms by Play (TF-IDF)
Hamlet (tragedy)
norway
players
does
madness
england
marcellus
majesty
ghost
phrase
act
dane
uncle
carriages
priam
foils
0.000 0.025 0.050 0.075 0.100 0.125 0.150 0.175
TF-IDF Score
Macbeth (tragedy)
does
hail
scotland
tyrant
wood
deed
knock
daggers
kings
knocking
ross
sisters
highness
soldiers
castle
0.000 0.025 0.050 0.075 0.100 0.125 0.150 0.175 0.200
TF-IDF Score
Othello (tragedy)
moor
roderigo
handkerchief
cyprus
emilia
lieutenant
general
willow
venice
michael
signior
whore
prithee
strumpet
does
0.0 0.1 0.2 0.3 0.4 0.5
TF-IDF Score
King Lear (tragedy)
edmund
gloucester
kent
tom
daughters
france
burgundy
letter
dover
gods
duke
fiend
knights
sisters
does
0.00 0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.40
TF-IDF Score
Romeo and Juliet (tragedy)
nurse
paris
friar
county
mantua
thursday
banished
verona
cell
cousin
prince
early
peter
slain
letter
0.00 0.05 0.10 0.15 0.20 0.25 0.30
TF-IDF Score
Julius Caesar (tragedy)
caesar
brutus
cassius
antony
rome
octavius
lucius
capitol
romans
philippi
publius
caius
gods
portia
roman
0.0 0.1 0.2 0.3 0.4 0.5 0.6
TF-IDF Score
Antony and Cleopatra (tragedy)
antony
caesar
cleopatra
egypt
pompey
lepidus
agrippa
rome
gods
does
egyptian
emperor
kings
wars
goodnight
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7
TF-IDF Score
Coriolanus (tragedy)
rome
gods
voices
city
general
titus
caius
senate
capitol
senators
gates
wars
romans
does
marketplace
0.0 0.1 0.2 0.3 0.4 0.5 0.6
TF-IDF Score
A Midsummer Night's Dream (comedy)
thisbe
helena
athens
wall
athenian
fairy
robin
lion
wood
lovers
starveling
fairies
lullaby
moonshine
helen
0.0 0.1 0.2 0.3 0.4
TF-IDF Score
Much Ado About Nothing (comedy)
claudio
hero
prince
signior
margaret
cousin
count
john
friar
ursula
niece
constable
troth
window
cupid
0.0 0.1 0.2 0.3 0.4 0.5 0.6
TF-IDF Score
Twelfth Night (comedy)
andrew
sebastian
niece
knight
does
count
antonio
maria
letter
gartered
yellow
stockings
prithee
madman
rain
0.0 0.1 0.2 0.3 0.4
TF-IDF Score
As You Like It (comedy)
phoebe
shepherd
duke
forest
ding
charles
rowland
silvius
hey
prithee
motley
monsieur
verses
lover
coz
0.00 0.05 0.10 0.15 0.20 0.25 0.30 0.35
TF-IDF Score
The Merchant of Venice (comedy)
portia
jew
antonio
ducats
bond
gratiano
venice
doctor
clerk
judge
christian
forfeit
letter
merchant
hazard
0.0 0.1 0.2 0.3 0.4 0.5 0.6
TF-IDF Score
The Taming of the Shrew (comedy)
lucentio
kate
bianca
petruchio
baptista
katherine
signior
padua
merchant
knock
hic
bride
gown
curst
shrew
0.0 0.1 0.2 0.3 0.4 0.5
TF-IDF Score
The Comedy of Errors (comedy)
chain
angelo
rope
merchant
dinner
abbey
duke
officer
mart
ducats
marks
dine
arrested
quoth
dined
0.0 0.1 0.2 0.3 0.4 0.5
TF-IDF Score
The Tempest (comedy)
monster
naples
stephano
milan
sebastian
island
ferdinand
cell
antonio
isle
dukedom
prithee
fish
bottle
ship
0.00 0.05 0.10 0.15 0.20 0.25 0.30
TF-IDF Score
Henry V (history)
france
french
england
vous
kate
english
majesty
exeter
harry
constable
pistol
glove
captain
bardolph
duke
0.0 0.1 0.2 0.3 0.4
TF-IDF Score
Richard III (history)
edward
richard
hastings
clarence
york
margaret
tower
gloucester
duke
rivers
prince
grey
norfolk
mayor
george
0.00 0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.40
TF-IDF Score
Henry IV Part 1 (history)
percy
hal
falstaff
douglas
harry
john
prince
poins
wales
jack
francis
glendower
sack
worcester
westmoreland
0.00 0.05 0.10 0.15 0.20 0.25 0.30 0.35
TF-IDF Score
Henry IV Part 2 (history)
john
bardolph
falstaff
davy
harry
prince
shallow
pistol
doll
westmoreland
cousin
mowbray
northumberland
majesty
hastings
0.0 0.1 0.2 0.3 0.4
TF-IDF Score
Richard II (history)
bolingbroke
richard
hereford
york
gaunt
mowbray
norfolk
cousin
duke
northumberland
uncle
majesty
gage
lancaster
harry
0.0 0.1 0.2 0.3 0.4
TF-IDF Score


Tragedies - Top 15 Distinctive Terms (TF-IDF)
Hamlet
(tragedy)
norway
players
does
madness
england
marcellus
majesty
ghost
phrase
act
dane
uncle
carriages
priam
foils
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
TF-IDF Score
Macbeth
(tragedy)
does
hail
scotland
tyrant
wood
deed
knock
daggers
kings
knocking
ross
sisters
highness
soldiers
castle
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200
TF-IDF Score
Othello
(tragedy)
moor
roderigo
handkerchief
cyprus
emilia
lieutenant
general
willow
venice
michael
signior
whore
prithee
strumpet
does
0.0
0.1
0.2
0.3
0.4
0.5
TF-IDF Score
King Lear
(tragedy)
edmund
gloucester
kent
tom
daughters
france
burgundy
letter
dover
gods
duke
fiend
knights
sisters
does
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
TF-IDF Score
Romeo and Juliet
(tragedy)
nurse
paris
friar
county
mantua
thursday
banished
verona
cell
cousin
prince
early
peter
slain
letter
0.00
0.05
0.10
0.15
0.20
0.25
0.30
TF-IDF Score
Julius Caesar
(tragedy)
caesar
brutus
cassius
antony
rome
octavius
lucius
capitol
romans
philippi
publius
caius
gods
portia
roman
0
0.1
0.2
0.3
0.4
0.5
0.6
TF-IDF Score
Antony and Cleopatra
(tragedy)
antony
caesar
cleopatra
egypt
pompey
lepidus
agrippa
rome
gods
does
egyptian
emperor
kings
wars
goodnight
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
TF-IDF Score
Coriolanus
(tragedy)
rome
gods
voices
city
general
titus
caius
senate
capitol
senators
gates
wars
romans
does
marketplace
0
0.1
0.2
0.3
0.4
0.5
0.6
TF-IDF Score


Comedies - Top 15 Distinctive Terms (TF-IDF)
A Midsummer Night's Dream
(comedy)
thisbe
helena
athens
wall
athenian
fairy
robin
lion
wood
lovers
starveling
fairies
lullaby
moonshine
helen
0.0
0.1
0.2
0.3
0.4
TF-IDF Score
Much Ado About Nothing
(comedy)
claudio
hero
prince
signior
margaret
cousin
count
john
friar
ursula
niece
constable
troth
window
cupid
0.0
0.1
0.2
0.3
0.4
0.5
0.6
TF-IDF Score
Twelfth Night
(comedy)
andrew
sebastian
niece
knight
does
count
antonio
maria
letter
gartered
yellow
stockings
prithee
madman
rain
0.0
0.1
0.2
0.3
0.4
TF-IDF Score
As You Like It
(comedy)
phoebe
shepherd
duke
forest
ding
charles
rowland
silvius
hey
prithee
motley
monsieur
verses
lover
coz
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
TF-IDF Score
The Merchant of Venice
(comedy)
portia
jew
antonio
ducats
bond
gratiano
venice
doctor
clerk
judge
christian
forfeit
letter
merchant
hazard
0.0
0.1
0.2
0.3
0.4
0.5
0.6
TF-IDF Score
The Taming of the Shrew
(comedy)
lucentio
kate
bianca
petruchio
baptista
katherine
signior
padua
merchant
knock
hic
bride
gown
curst
shrew
0.0
0.1
0.2
0.3
0.4
0.5
TF-IDF Score
The Comedy of Errors
(comedy)
chain
angelo
rope
merchant
dinner
abbey
duke
officer
mart
ducats
marks
dine
arrested
quoth
dined
0.0
0.1
0.2
0.3
0.4
0.5
TF-IDF Score
The Tempest
(comedy)
monster
naples
stephano
milan
sebastian
island
ferdinand
cell
antonio
isle
dukedom
prithee
fish
bottle
ship
0.00
0.05
0.10
0.15
0.20
0.25
0.30
TF-IDF Score



TF-IDF Interpretation

1. Do some documents share distinctive vocabulary?

Yes. Several documents share distinctive vocabulary, particularly within the same genre and historical context.

Within Tragedies:

- Roman plays share political vocabulary: *Julius Caesar*, *Antony and Cleopatra*, and *Coriolanus* all feature "caesar," "rome," and political titles ("general," "senate," "capitol")
- English historical tragedies share geographic terms: *Hamlet* ("norway," "england"), *Macbeth* ("scotland")
- Domestic tragedies share family/relationship terms: *Othello* and *King Lear* include relationship markers ("lieutenant," "daughters," "kent")

Within Comedies:

- Romantic comedies share terms of courtship and social hierarchy: "duke," "count," "prince" appear across *Twelfth Night*, *As You Like It*, and *Much Ado About Nothing*
- Italian comedies share geographic vocabulary: *The Merchant of Venice* ("venice," "ducats"), *The Taming of the Shrew* ("padua," "signior"), *The Comedy of Errors* ("ducats," "merchant")
- Pastoral/magical comedies: *A Midsummer Night's Dream* ("fairy," "wood," "athens") and *The Tempest* ("island," "monster") share fantastical settings

Within Histories:

- All histories share aristocratic titles and English geography: "duke," "prince," "york," "john"
- The Henry IV plays are particularly close, both featuring "falstaff," "harry," "prince," "john," "bardolph"

Cross-genre patterns:

- Religious language appears inconsistently: "friar" (*Romeo and Juliet*, *Much Ado About Nothing*), "gods" (*King Lear*, *Antony and Cleopatra*, *Coriolanus*)
- The interrogative "does" appears in multiple tragedies (*Hamlet*, *Macbeth*, *Twelfth Night*), possibly reflecting rhetorical questioning

2. Are distinctive terms topical, rhetorical, or technical?

Predominantly topical, with some rhetorical elements:

The distinctive terms are mostly topical and character-driven, with some rhetorical and technical terms.

Topical (majority):

- Character names dominate: Nearly every play's distinctive vocabulary is character-driven (*Hamlet*: "marcellus," "ghost"; *Othello*: "roderigo," "emilia"; *Romeo and Juliet*: "nurse," "friar"). This reflects Shakespeare's character-centric dramaturgy.
- Geographic specificity: Place names establish setting (*Othello*: "cyprus," "venice"; *Macbeth*: "scotland"; *Julius Caesar*: "rome," "philippi")
- Plot-specific objects: Key props appear as distinctive terms (*Othello*: "handkerchief"; *Macbeth*: "daggers"; *The Comedy of Errors*: "chain," "rope"; *The Merchant of Venice*: "bond")

Rhetorical (limited):

- The appearance of "does" in *Hamlet*, *Macbeth*, and *Twelfth Night* may indicate rhetorical questioning patterns
- *Hamlet*'s "phrase" and "act" suggest metatheatrical or philosophical discourse
- Forms of address vary by genre: "signior" (Italian comedies), "majesty" (English histories/tragedies)

Technical:

- Military/political titles: "lieutenant," "general," "capitol," "senate"
- Legal/commercial terms: "ducats," "bond," "merchant," "clerk," "judge"

Insight: TF-IDF here primarily captures what makes each play's world unique (its cast of characters, setting, and central objects) rather than distinctive rhetorical strategies. This suggests Shakespeare's lexical variation operates more at the level of dramatic world-building than at stylistic differentiation.

3. Are there documents whose distinctiveness seems driven by noise or formatting?

There is little evidence of formatting noise in the results.

- Character labels appear to have been successfully removed.
- Stage directions do not dominate.
- No artifacts like "ACT" or "SCENE" appear.
- No OCR distortions or stray punctuation dominate.

Some generic verbs like *does* appear in multiple tragedies, but this is likely due to contextual emphasis rather than preprocessing error.

Overall, the TF-IDF output appears clean and substantively meaningful.

But for more refined lexical analysis, we might:

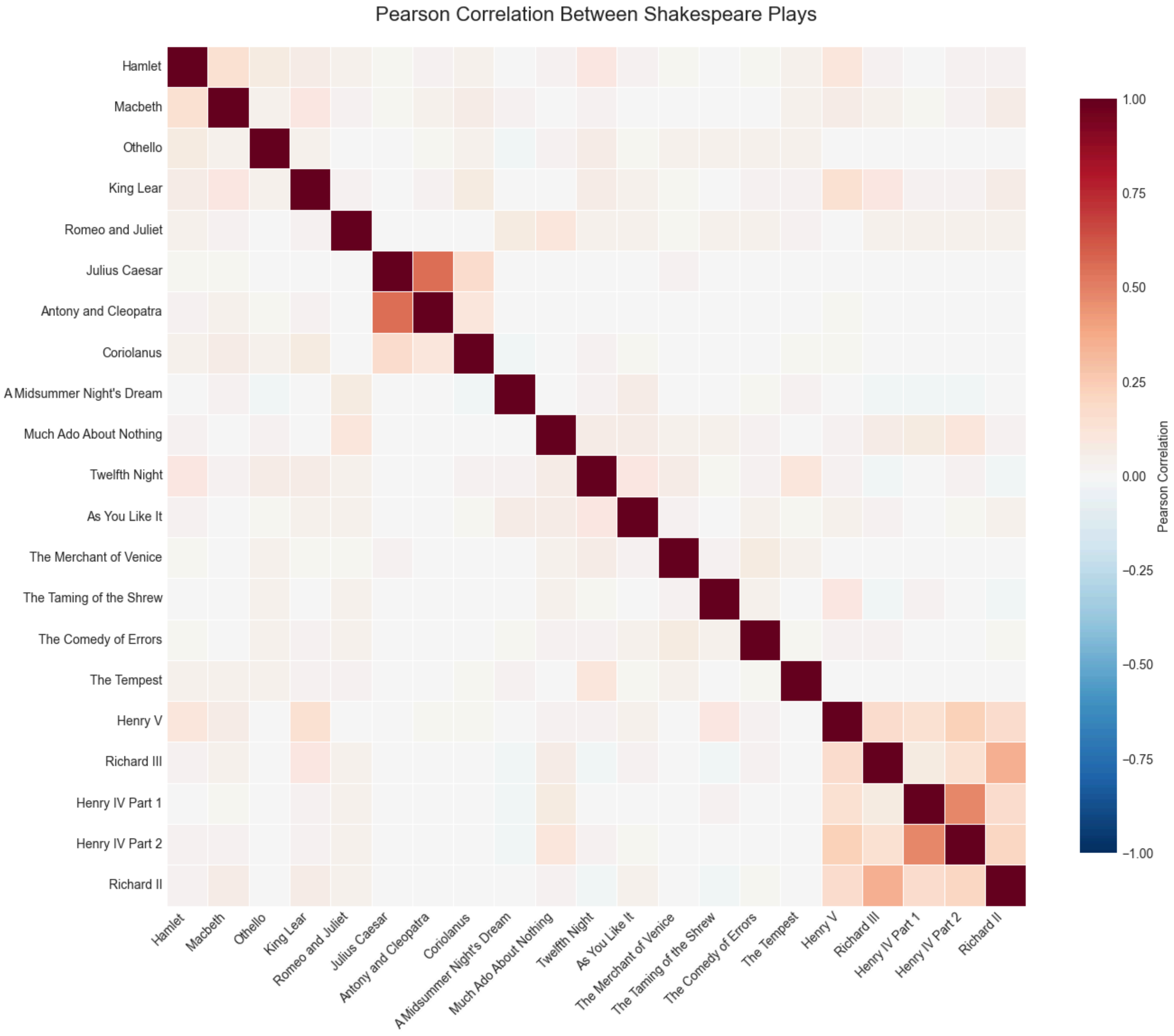
1. Remove character names systematically
2. Verify that stage directions are fully removed
3. Standardize archaic verb forms ("does"/"doth")
4. Consider lemmatization to group inflected forms

STEP 2: Pearson Correlation Analysis

Calculate Pairwise Correlations

```
Correlation matrix shape: (21, 21)

First 5x5 subset:
play             Hamlet  Macbeth  Othello  King Lear  Romeo and Juliet
play
Hamlet            1.000    0.150    0.073      0.064             0.032
Macbeth           0.150    1.000    0.037      0.087             0.022
Othello           0.073    0.037    1.000      0.037            -0.000
King Lear         0.064    0.087    0.037      1.000             0.022
Romeo and Juliet  0.032    0.022   -0.000      0.022             1.000
```

Create Similarity Heatmap



Find Most and Least Similar Pairs

```
================================================================================
TOP 5 MOST SIMILAR PLAY PAIRS
================================================================================
Julius Caesar (tragedy) <--> Antony and Cleopatra (tragedy)
  Correlation: 0.558

Henry IV Part 1 (history) <--> Henry IV Part 2 (history)
  Correlation: 0.483

Richard III (history) <--> Richard II (history)
  Correlation: 0.359

Henry V (history) <--> Henry IV Part 2 (history)
  Correlation: 0.229

Henry IV Part 2 (history) <--> Richard II (history)
  Correlation: 0.204

================================================================================
TOP 5 LEAST SIMILAR PLAY PAIRS
================================================================================
A Midsummer Night's Dream (comedy) <--> Henry IV Part 2 (history)
  Correlation: -0.023

The Taming of the Shrew (comedy) <--> Richard II (history)
  Correlation: -0.022

A Midsummer Night's Dream (comedy) <--> Henry IV Part 1 (history)
  Correlation: -0.021

Coriolanus (tragedy) <--> A Midsummer Night's Dream (comedy)
  Correlation: -0.021

Othello (tragedy) <--> A Midsummer Night's Dream (comedy)
  Correlation: -0.020
```

Pearson Correlation Interpretation

Most Similar Document Pairs

1. **Julius Caesar ↔ Antony and Cleopatra (r = 0.558)**

Both plays are Roman tragedies centered on overlapping political figures and institutions (Caesar, Antony, Rome, military leadership). The high correlation suggests that shared geopolitical setting and political vocabulary strongly drive lexical similarity.

2. **Henry IV Part 1 ↔ Henry IV Part 2 (r = 0.483)**

This is unsurprising, as the plays are direct continuations of one another. They share characters (Falstaff, Prince Hal), historical context, and dynastic themes. The strong correlation confirms that sequential historical narratives produce consistent lexical patterns.

Notably, the remaining highly similar pairs are also English histories, reinforcing that dynastic and national political discourse produces measurable lexical clustering.

Least Similar Document Pairs

1. **A Midsummer Night's Dream ↔ Henry IV Part 2 (r = -0.023)**

A Midsummer Night's Dream is a magical comedy set in Athens featuring fairies, lovers, and amateur actors, with vocabulary centered on enchantment, courtship, and theatrical performance ("fairy," "helena," "thisbe," "wall," "athens"). *Henry IV Part 2* is a political history focused on English nobility, rebellion, military campaigns, and royal succession, featuring vocabulary of statecraft and warfare ("bardolph," "falstaff," "harry," "john"). These plays represent opposite poles of Shakespeare's corpus—fantastical romance vs. political realism—and share virtually no lexical overlap.

2. **The Taming of the Shrew ↔ Richard II (r = -0.022)**

The Taming of the Shrew is an Italian domestic comedy about courtship, marriage, and social hierarchy within a merchant family ("lucentio," "kate," "bianca," "petruchio," "baptista," "padua"). *Richard II* is an English political history about royal deposition, divine right, and aristocratic rebellion ("bolingbroke," "richard," "york," "gaunt," "hereford"). The negative correlation reflects completely disjoint thematic and lexical domains—domestic/romantic vs. political/historical, Italian urban setting vs. English court and countryside.

These negative (near-zero) correlations indicate minimal overlap in vocabulary distribution.

Notably, *A Midsummer Night's Dream* appears in three of the five least similar pairs, suggesting it occupies an exceptionally distinct lexical space within the corpus, likely due to its unique magical/supernatural vocabulary.

This makes intuitive sense:

- *A Midsummer Night's Dream* centers on mythological, romantic, and pastoral themes (fairies, lovers, Athens).
- The histories focus on political conflict, warfare, monarchy, and national identity.
- Tragedies like *Othello* and *Coriolanus* revolve around military hierarchy and political authority.

Research Questions Emerging from Correlation Patterns

- Why do Roman tragedies show the strongest inter-play correlation despite not being a continuous narrative?
- Why are history plays' internal correlations surprisingly low despite shared characters and settings?
- Temporal scope creates variation (Richard II's medieval court vs. Henry V's Renaissance warfare)
- What makes *A Midsummer Night's Dream* such an extreme outlier? Would *The Tempest* (another magical play) show similar outlier status?
  - Do magical/supernatural plays form a distinct sub-genre with unique lexical profiles?

Conclusion:

The Pearson correlation analysis suggests that genre categories have measurable but moderate lexical coherence. Clear clustering among the English histories and the Roman political tragedies indicates that shared historical setting and character networks strongly influence vocabulary patterns. However, the relatively modest magnitude of even the highest correlation (r = 0.558) indicates that lexical overlap across plays remains limited.

Overall, Shakespeare's vocabulary appears to be shaped more by specific dramatic context—character ensembles, geopolitical setting, and thematic focus—than by broad genre conventions alone. These findings suggest that plot-level particularity exerts a stronger influence on lexical distribution than abstract genre classification, highlighting both the specificity of Shakespeare's dramatic worlds and his considerable lexical range.

STEP 3: Syntactic Complexity Analysis

Selection of Plays for Syntactic Comparison

Selected Plays:

- *Julius Caesar* (tragedy)
- *Antony and Cleopatra* (tragedy)

Reason for Selection

These two plays were selected because they show the strongest lexical similarity in the corpus (r = 0.558). Both the TF-IDF and Pearson correlation analyses indicate substantial shared vocabulary.

Shared distinctive terms:
"antony," "caesar," and "rome" appear in both plays, reflecting:

- Character continuity (Antony and Caesar/Octavius)
- Shared Roman political setting
- Common military and imperial vocabulary

At the same time, each play contains distinctive terms reflecting its specific focus:

- *Julius Caesar*: "brutus," "cassius," "capitol," "philippi": conspiracy and republican politics
- *Antony and Cleopatra*: "cleopatra," "egypt," "pompey," "agrippa": romance, empire, and the triumvirate

Thus, the plays are lexically similar but thematically distinct.

Despite their strong lexical similarity, do *Julius Caesar* and *Antony and Cleopatra* differ in syntactic complexity?

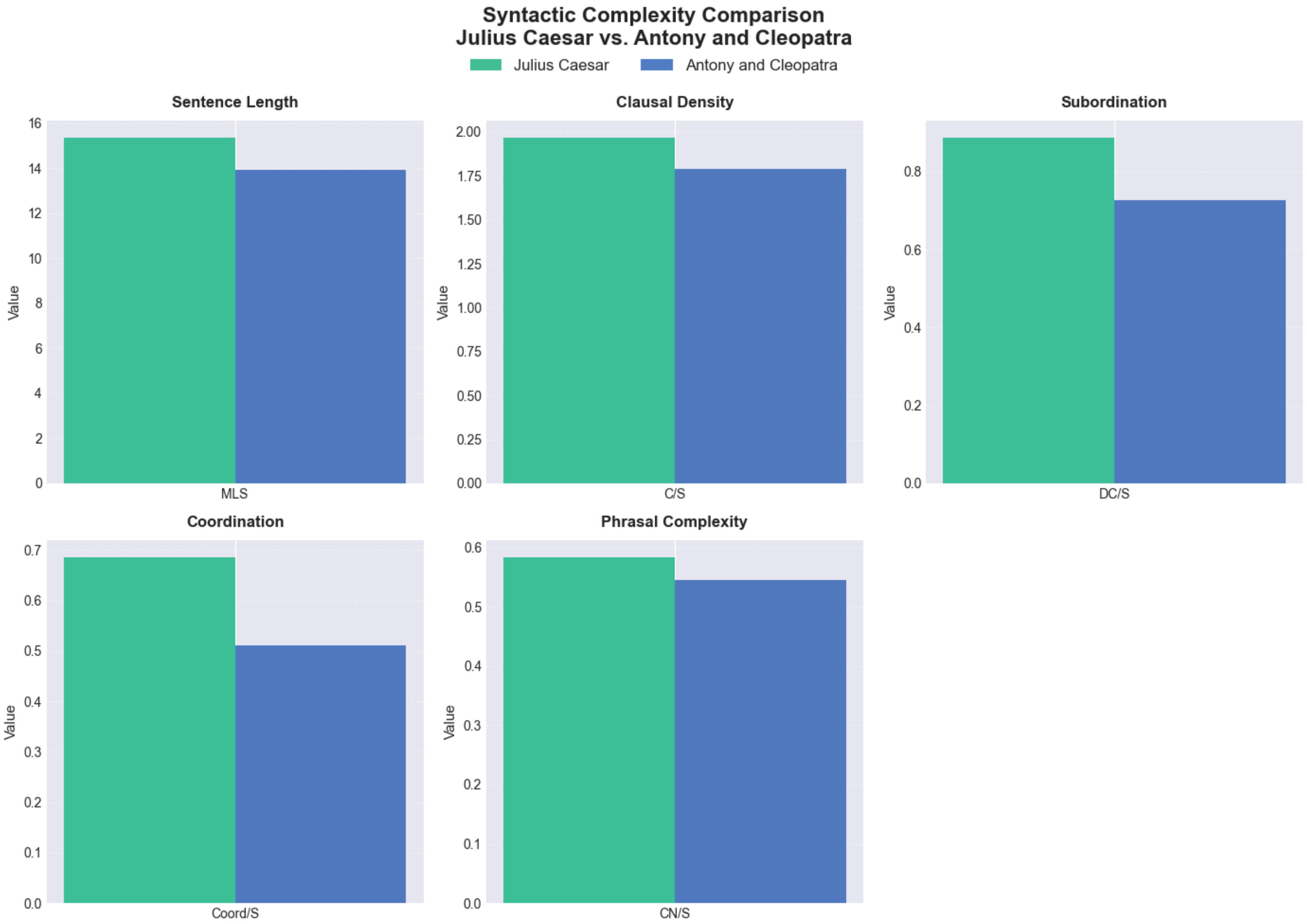
Define Syntactic Complexity Functions

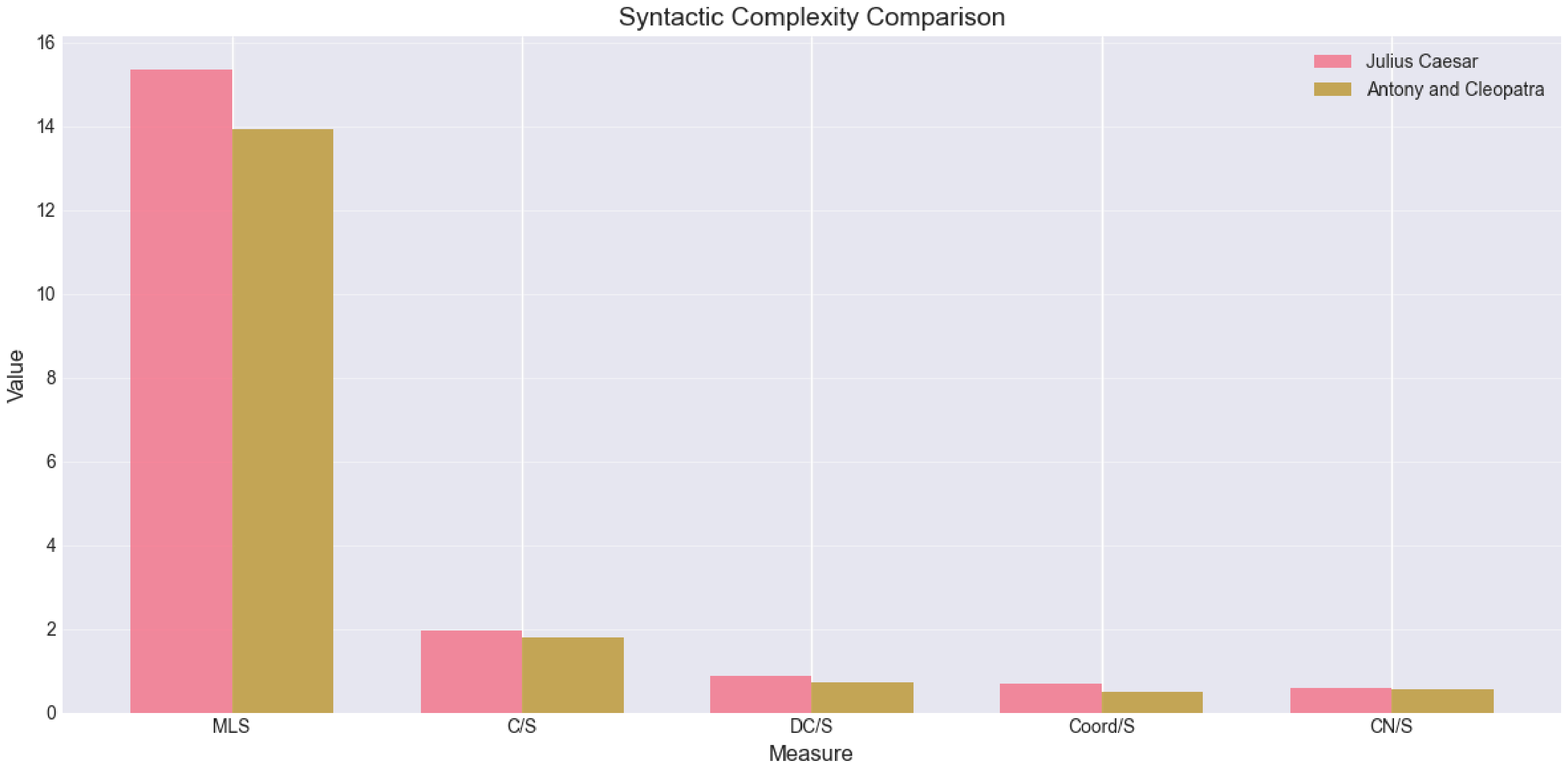
Calculate Complexity for Both Plays

Create Summary Table

```
SYNTACTIC COMPLEXITY COMPARISON
                      sentences    MLS   C/S  DC/S  Coord/S  CN/S
play
Julius Caesar              1440  15.37  1.97  0.89     0.69  0.58
Antony and Cleopatra       2024  13.92  1.79  0.73     0.51  0.54

Measures:
  MLS = Mean Length of Sentence
  C/S = Clauses per Sentence
  DC/S = Dependent Clauses per Sentence
  Coord/S = Coordination per Sentence
  CN/S = Complex Nominals per Sentence
```

Visualize Syntactic Differences





Extract Example Sentences for Interpretation

```
EXAMPLE SENTENCES FROM NORMALIZED TEXT:

Julius Caesar - Longest Sentence:
--------------------------------------------------------------------------------
[125 words, 9 dependent clauses, 13 coordination markers]

over thy wounds now do i prophesy
(which like dumb mouths do ope their ruby lips
to beg the voice and utterance of my tongue)
a curse shall light upon the limbs of men;
domestic fury and fierce civil strife
shall cumber all the parts of italy;
blood and destruction shall be so in use
and dreadful objects so familiar
that mothers shall but smile when they behold
their infants quartered with the hands of war,
all pity choked with custom of fell deeds;
and caesar's spirit, ranging for revenge,
with ate by his side come hot from hell,
shall in these confines with a monarch's voice
cry "havoc!"

Antony and Cleopatra - Longest Sentence:
--------------------------------------------------------------------------------
[86 words, 4 dependent clauses, 19 coordination markers]

nay, nay, octavia, not only that--
that were excusable, that and thousands more
of semblable import--but he hath waged
new wars 'gainst pompey; made his will and read it
to public ear;
spoke scantly of me; when perforce he could not
but pay me terms of honor, cold and sickly
he vented them, most narrow measure lent me;
when the best hint was given him, he not took 't,
or did it from his teeth.

Julius Caesar - Most Syntactically Complex (most dependent clauses):
--------------------------------------------------------------------------------
[105 words, 11 dependent clauses, 10 coordination markers]

and since you know you cannot see yourself
so well as by reflection, i, your glass,
will modestly discover to yourself
that of yourself which you yet know not of.
and be not jealous on me, gentle brutus.
were i a common laughter, or did use
to stale with ordinary oaths my love
to every new protester; if you know
that i do fawn on men and hug them hard
and after scandal them, or if you know
that i profess myself in banqueting
to all the rout, then hold me dangerous.

Antony and Cleopatra - Most Syntactically Complex (most dependent clauses):
--------------------------------------------------------------------------------
[41 words, 8 dependent clauses, 0 coordination markers]

when it pleaseth their deities to take the wife of a
man from him, it shows to man the tailors of the
earth; comforting therein, that when old robes are
worn out, there are members to make new.
```

Step 3: Syntactic Complexity Interpretation

How do the two texts differ in syntactic complexity?

Despite their high lexical similarity (r = 0.558), *Julius Caesar* and *Antony and Cleopatra* show meaningful differences in syntactic complexity:

Julius Caesar uses longer, more complex sentences:

- Longer sentences: MLS = 15.37 vs. 13.92 (about 10% longer)
- More clauses per sentence: C/S = 1.97 vs. 1.79
- More subordination: DC/S = 0.89 vs. 0.73 (22% more dependent clauses)
- More coordination: Coord/S = 0.69 vs. 0.51 (35% more "and," "but," "or")
- More complex noun phrases: CN/S = 0.58 vs. 0.54

In Summary: *Julius Caesar* consistently shows higher complexity across all five measures. The plays share similar vocabulary (same characters, Roman setting, political themes), but *Julius Caesar* builds longer, more intricate sentences with more subordinate clauses and coordinated phrases.

Do these differences align with or complicate your earlier lexical findings?

These syntactic differences complicate what we found earlier with TF-IDF and correlation analysis.

From the earlier analysis, we saw that *Julius Caesar* and *Antony and Cleopatra* have the highest correlation in the corpus (r = 0.558). They share distinctive vocabulary like "caesar," "antony," and "rome," which makes sense because they're both Roman political plays with overlapping characters.

But the syntactic analysis shows something different. Even though they use similar words, *Julius Caesar* builds more complex sentences. It has longer sentences, more subordinate clauses, and more coordination. *Antony and Cleopatra* uses the same Roman vocabulary but in shorter, simpler sentence structures.

This means that sharing vocabulary doesn't necessarily mean sharing sentence structure. The plays talk about the same things (Roman politics, Caesar, Antony) but say them differently. *Julius Caesar* uses complex syntax for public speeches and political debates. *Antony and Cleopatra* uses simpler syntax for private conversations between lovers.

So the findings don't completely align. The lexical analysis showed these plays are very similar, but the syntactic analysis reveals important differences in how they're written. Both findings are valuable—they just tell us different things about the plays.

What kinds of rhetorical or stylistic practices might these syntactic patterns reflect?

The syntactic differences make sense when you think about what's happening in each play.

Julius Caesar has higher complexity because it's full of public speeches. The play is about political conspiracy, senate debates, and persuading crowds. Brutus and Antony give long funeral speeches trying to convince the Roman people. Cassius argues with Brutus about strategy. These situations require complex sentences—you need subordinate clauses to build logical arguments, and you need coordination to list reasons and connect ideas.

For example, when Brutus speaks to the crowd, he uses complex syntax to make his case: "If there be any in this assembly, any dear friend of Caesar's, to him I say that Brutus' love to Caesar was no less than his..." This kind of sentence structure helps build a persuasive argument.

Antony and Cleopatra has lower complexity because it's more about personal relationships. Much of the play is private conversations between Antony and Cleopatra—lovers arguing, reconciling, expressing emotions. This kind of dialogue works better with shorter, more direct sentences. When Cleopatra is upset, she doesn't give long political speeches—she says things like "I am dying, Egypt, dying."

The patterns suggest:

- *Julius Caesar* uses complex syntax for public persuasion (convincing crowds, making political arguments)
- *Antony and Cleopatra* uses simpler syntax for private emotion (expressing feelings, having personal conversations)

Even though both plays are about Roman politics and share the same vocabulary, Shakespeare writes them differently based on the dramatic situation. Political speeches need one kind of sentence structure, and intimate conversations need another.

STEP 4: SYNTHESIS

Triangulating Evidence Across All Three Approaches

The Question:

Are Shakespeare's genre categories (tragedy, comedy, history) actually reflected in how the plays are written, or are they just based on plot and theme? And when two plays use similar vocabulary, does that mean they're written in the same style?

Evidence from TF-IDF:

The TF-IDF analysis showed that distinctive vocabulary is mostly about the specific content of each play, not really about genre.

What I found:

- Character names are the most distinctive terms in almost every play (like "brutus" and "cassius" in *Julius Caesar*, "thisbe" and "helena" in *A Midsummer Night's Dream*)
- Place names create clustering: Roman plays all have "rome" and "capitol"; Italian comedies have "venice" and "padua"
- There are some genre patterns (tragedies mention "blood" and "death" more; comedies mention "love" and "marry"), but they're not as strong as I expected
- The most distinctive words are usually specific to that play's story—character names, important objects (like Othello's handkerchief), and locations

What this means: The vocabulary that makes each play unique is more about its specific story, characters, and setting than about being a tragedy or comedy.

Evidence from Pearson Correlation:

The correlation analysis showed that plays do cluster somewhat by genre, but not as cleanly as I expected.

Key findings:

- The two most similar plays are *Julius Caesar* and *Antony and Cleopatra* (r = 0.558), these are both Roman political tragedies with overlapping characters
- Second most similar are *Henry IV Part 1* and *Part 2* (r = 0.483), this makes sense since they're basically one long play split in two
- The least similar plays are *A Midsummer Night's Dream* and various history plays (r ≈ -0.02), which means magical comedy vs. political history have almost nothing in common

Patterns I noticed:

- Roman plays (*Julius Caesar*, *Antony and Cleopatra*, *Coriolanus*) are more similar to each other than to other tragedies
- History plays are surprisingly different from each other—each one focuses on different political situations
- *A Midsummer Night's Dream* is really different from everything else (it showed up in three of the five least-similar pairs)

What this means: Plays cluster more by specific themes and settings (like "Roman politics" or "Italian city life") than by broad genres. Also, even the most similar plays only have moderate correlation (r = 0.558), which shows Shakespeare uses really varied vocabulary even when writing about similar topics.

Evidence from Syntactic Complexity:

The syntactic analysis of *Julius Caesar* and *Antony and Cleopatra* showed something surprising: even though these plays use similar vocabulary, they're written quite differently.

The comparison:

- *Julius Caesar* is more complex: longer sentences (MLS = 15.37 vs. 13.92), more subordinate clauses (DC/S = 0.89 vs. 0.73), more coordination (Coord/S = 0.69 vs. 0.51)
- *Antony and Cleopatra* is simpler across all measures

Why this matters:

- *Julius Caesar* has lots of public speeches—senate debates, funeral orations, political arguments. These need complex sentences with lots of subordinate clauses to build logical arguments
- *Antony and Cleopatra* has more private conversations between lovers. These use shorter, more direct sentences for emotional impact

What this means: You can use the same vocabulary (both plays talk about Caesar, Antony, and Roman politics) but write in different styles depending on what's happening in the scene. Political speeches need one kind of sentence structure, emotional conversations need another.

Synthesis and Conclusion:

After looking at all three methods together, here's what I learned about how Shakespeare writes:

The main findings:

1. **Vocabulary is about content, not genre**: The TF-IDF analysis showed that what makes each play's vocabulary distinctive is mostly its specific characters, settings, and plot. Character names and place names dominate the distinctive terms. Genre (tragedy vs. comedy) matters less than I thought.

2. **Plays cluster by specific themes, not just broad genres**: The correlation analysis showed that Roman plays cluster together, Italian comedies cluster together, and so on. But "tragedy" as a whole doesn't form a super tight cluster. The highest correlation in the whole corpus was only r = 0.558, which is moderate, not strong. This means even similar plays use pretty different vocabulary overall.

3. **Syntax works differently from vocabulary**: Even though *Julius Caesar* and *Antony and Cleopatra* have high lexical similarity (r = 0.558), they have different syntactic complexity. *Julius Caesar* uses longer, more complex sentences because it's full of formal speeches. *Antony and Cleopatra* uses simpler sentences because it's more about personal conversations. This shows that Shakespeare adjusts how he builds sentences based on the dramatic situation, not just on what words he's using.

What this all means:

Shakespeare's genre categories are real but they're not as strong as I expected. When I started, I thought tragedies would all cluster together and be clearly different from comedies. But what I found is more complicated:

- The strongest patterns are at a more specific level: Roman plays, Italian comedies, magical plays, etc.
- Vocabulary is driven mainly by the content of each play (who's in it, where it takes place, what happens)
- Syntax is driven by the dramatic mode (whether characters are giving speeches or having conversations, making arguments or expressing emotions)

The disconnect between lexical and syntactic findings is also important. *Julius Caesar* and *Antony and Cleopatra* share a lot of vocabulary (same characters, same Roman world), but they're structured differently because one is about public politics and one is about private passion. This suggests that Shakespeare thinks about word choice and sentence structure separately—he can use the same vocabulary but arrange it differently depending on what the scene needs.

Summary of Outputs

Files Generated:

TF-IDF Analysis:

1. `tfidf_all_plays.png` - Top 15 distinctive terms for all 21 plays (7×3 grid)
2. `tfidf_all_tragedies.png` - Top 15 distinctive terms for all 8 tragedies
3. `tfidf_all_comedies.png` - Top 15 distinctive terms for all 8 comedies
4. `tfidf_all_histories.png` - Top 15 distinctive terms for all 5 histories
5. `tfidf_by_genre_average.png` - Average distinctive terms by genre (3-panel comparison)
6. `tfidf_comparison.png` - Side-by-side comparison of 6 representative plays

Pearson Correlation Analysis: 7. `pearson_heatmap.png` - Correlation matrix for all 21 plays

Syntactic Complexity Analysis: 8. `syntactic_comparison_faceted.png` - Faceted comparison of *Julius Caesar* vs. *Antony and Cleopatra* across 5 measures

Data Tables: 9. `tfidf_results.csv` - Complete TF-IDF top terms for all plays 10. `syntactic_complexity.csv` - Syntactic complexity measures for selected plays