How to die in Shakespeare

A predictive tool for aspiring authors, by Ben Mathis

Play in, Lines out

Steps needed:

- Too much regex
- Cleaning stage directions
- Splitting lines by newlines to remove bias toward long-winded characters

```
MURDERER  He's a traitor.

SON
Thou liest, thou shag-eared villain!

MURDERER  What, you egg?
[Stabbing him.] Young fry of treachery!

SON  He has killed
me, mother.
Run away, I pray you.
[Lady Macduff exits, crying "Murder!" followed by the
Murderers bearing the Son's body.]
```
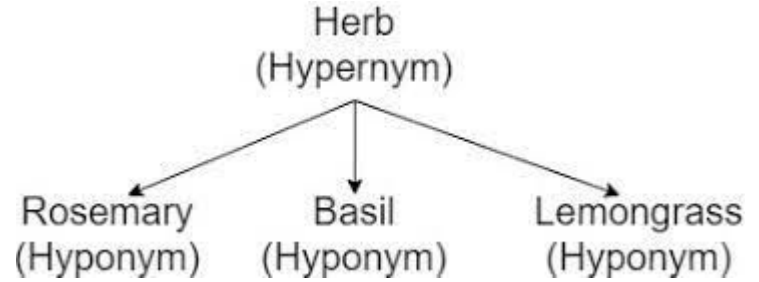
```
(r'(((([A-Z]{2,}(,*\s+and\s+|[,\.]{0,1}\s+|\/))+)(?!((([A-Z]{2,}(,*\s+and\s+|[,\.]{0,1}\s+|\/))+)))', line)
```

Words to Models

Part Of Speech Tagging

N-grams and other context-driven preprocessing steps were discarded due to Shakespeare's *poetic* grammar.

Dictionary Extravaganza

Problem words:

- Many, many contractions and uses of the poetic apostrophe
- "Thou [verb]st"
- Getting over 'ofer'
- Words that despite not being Middle English certainly are not in modern vernacular
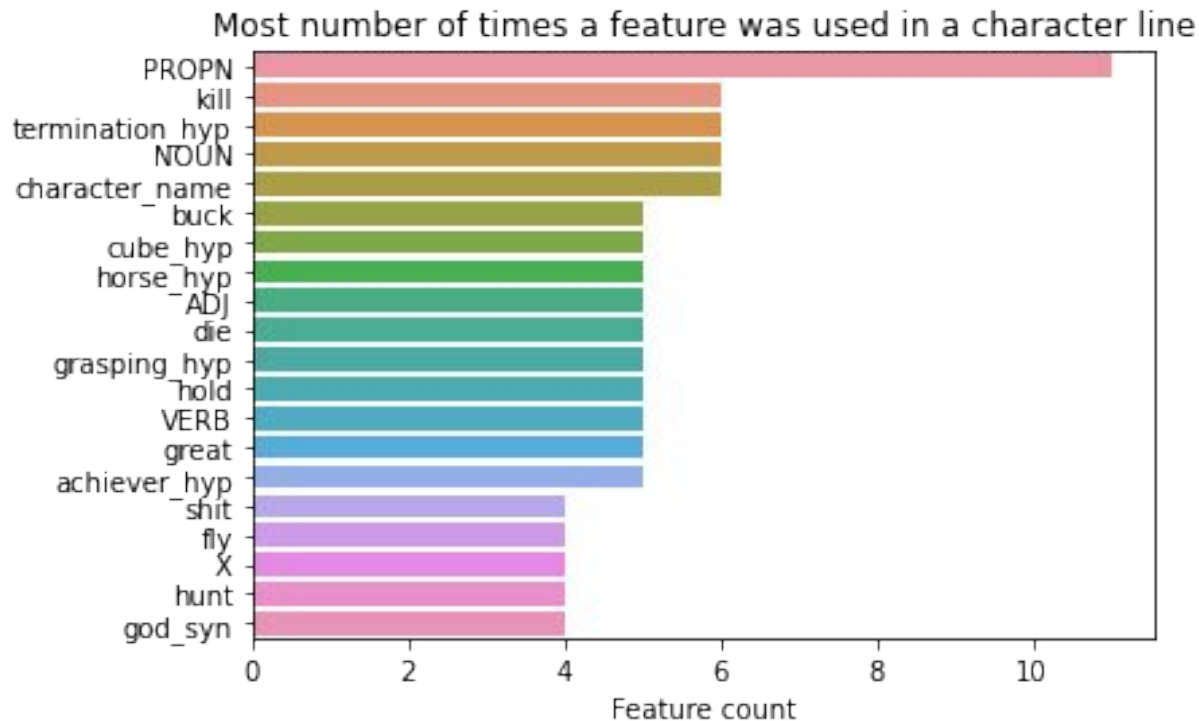
```
'affy': 'engage',
'invitis': 'in spite of clouds',
'gallowglasses': 'axe wielding Irish soldier',
'uncurable': 'incurable',
'alderliefest': 'most beloved',
'callet': 'scold',
'tainture': 'tainting',
'fraudful': 'fraudulent',
'deathful': 'mortal',
'sophister': 'sophist',
'customed': 'customary',
'behoof': 'benefit',
'charneco': 'sweet Portuguese wine',
'shearman': ' one who shears excess material from woollen cloth',
'sicil': 'sicily',
'pursuivant': ' royal messenger',
'unmeet': 'unready',
'clerkly': 'scholarly',
'traiterous': 'traitorous',
'woefull': 'sorrowful',
'occupat': 'business',
```
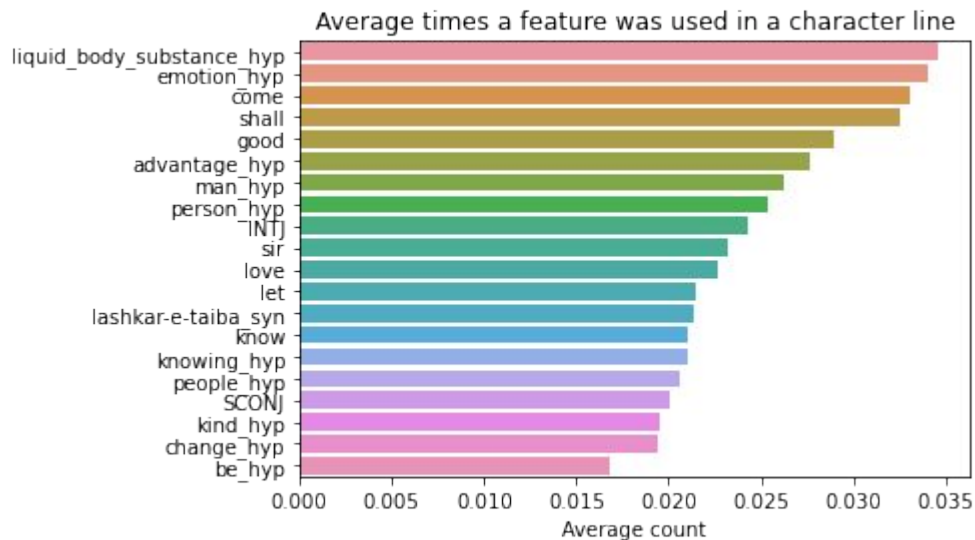
Full Process

- Start with a line of character dialogue
- Remove punctuation
- Run through at least five dictionaries
- Vectorize the word and its lemma
- Vectorize the words' hypernym if available, synonym if not

Exploration of the scrubbed data, part 1
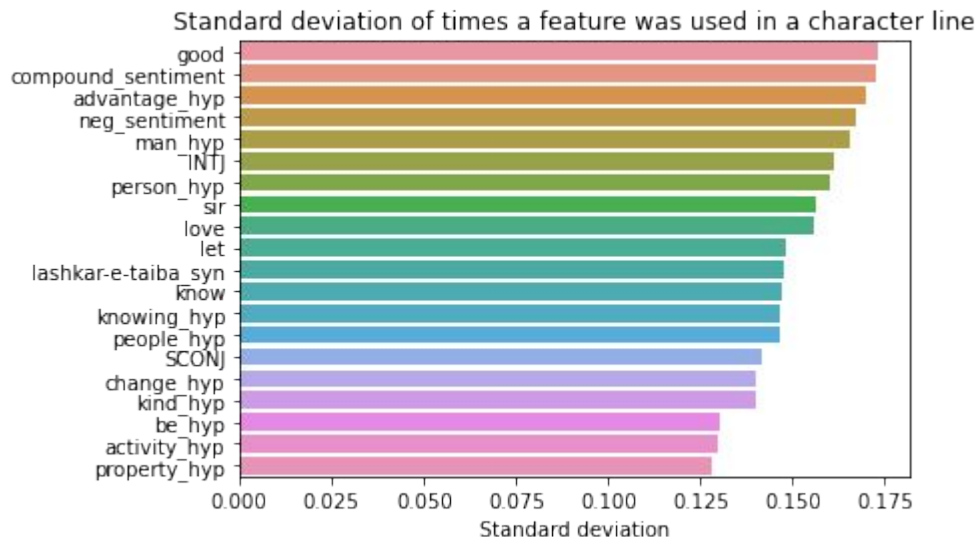

Most number of times a feature was used in a character line

Exploration of the scrubbed data, part 2
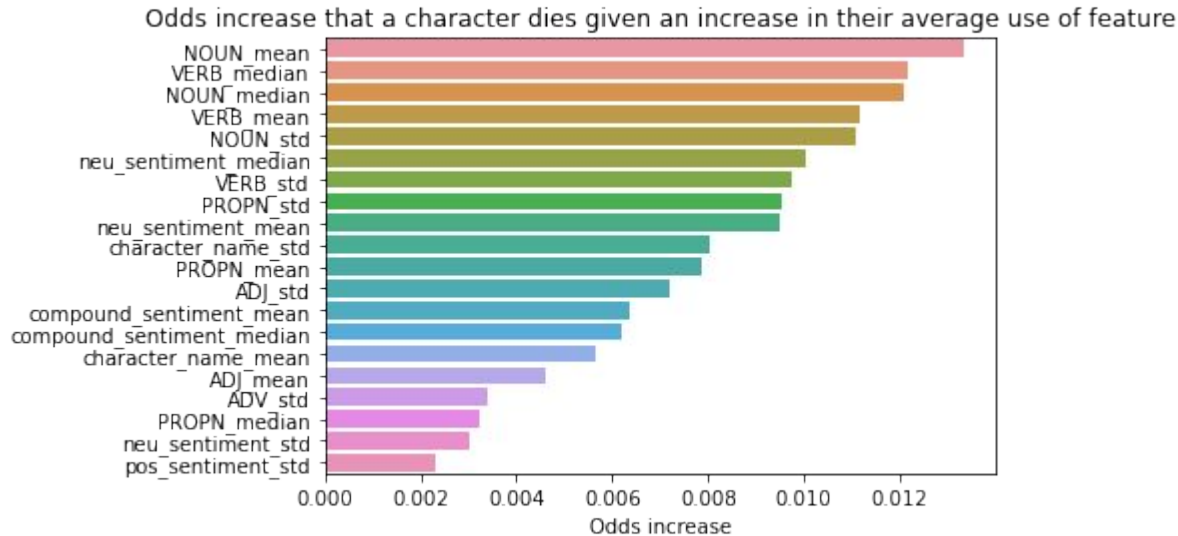

Average times a feature was used in a character line

Many of these features are important to the model for its decision making.

Exploration of the scrubbed data, part 3

Standard deviation of times a feature was used in a character line

Some of these features are as well.

Multinomial Naive Bayes' Opinions

Odds increase that a character dies given an increase in their average use of feature

Important note: this was not the final model used for predictive purposes. This model only has .84 ROC AUC score.

Multinomial Naive Bayes' Opinions, cont.

Odds increase that a character dies given an increase in their average use of feature

Feature	
liquid_body_substance_hyp_std	
emotion_hyp_std	
come_std	
shall_std	
good_std	
advantage_hyp_std	
change_hyp_std	
person_hyp_std	
let_std	
lashkar-e-taiba_syn_std	
people_hyp_std	
love_std	
know_std	
knowing_hyp_std	
kind_hyp_std	
be_hyp_std	
activity_hyp_std	
communicate_hyp_std	
speak_std	
like_std	

Odds increase: 0.0000 0.0005 0.0010 0.0015 0.0020

If you want to live don't talk about:

- Bodily fluids
- Emotions
- Goodness
- Change
- Talking.

Characters consigned to death by the model.

One liners:

SECOND AND THIRD KNIGHTS - The two noble kinsmen:

We'll follow cheerfully.

GLANSDALE - Henry VI, part 1:

And I, here, at the bulwark of the bridge.

Characters with actual dialogue:

SALISBURY - Henry V

SECOND LORD - Timon of Athens