Quantifying Literary Style

Grant Nelson

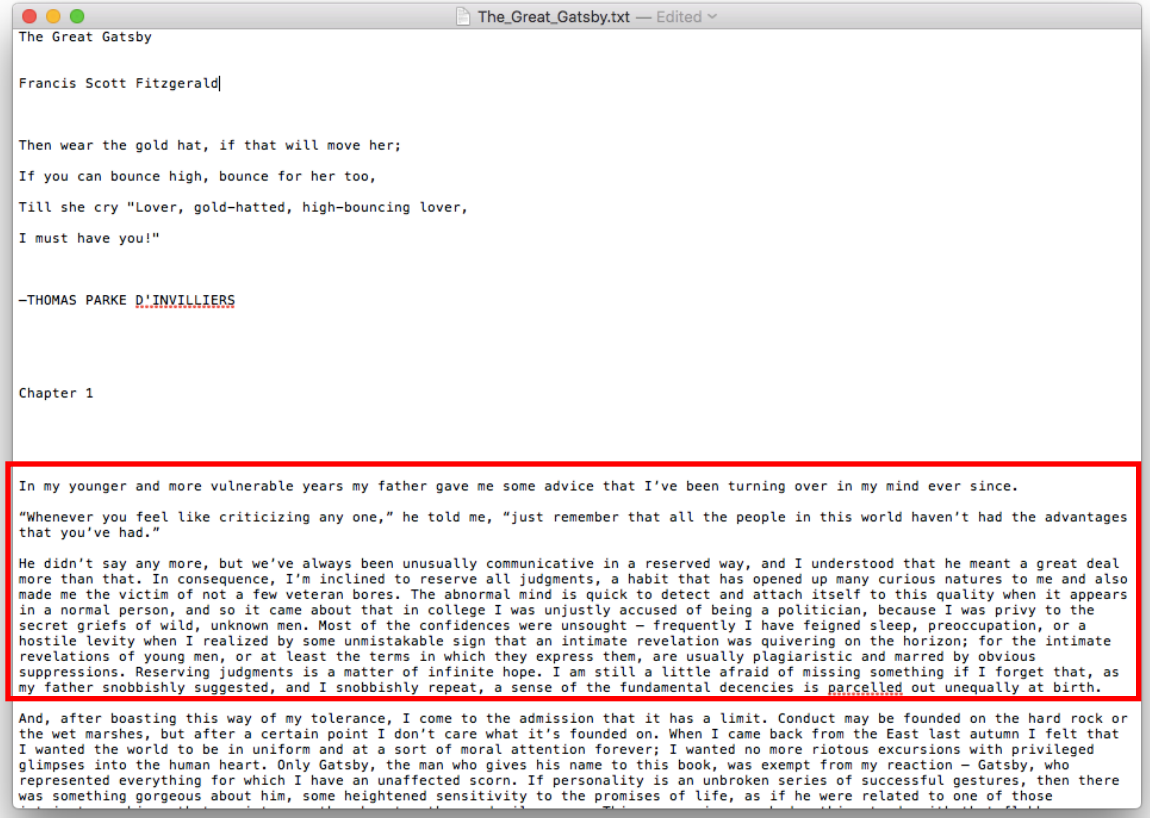
June 3, 2016

General Assembly

Introduction

- Style is a universally recognized aspect of literature, but often vaguely defined
- To what extent can style be quantified and reliably detected in an author's works?
- Given a chunk of text, predict which author wrote it
- Four contemporaneous authors from the 1920s and '30s:
  - William Faulkner
  - F. Scott Fitzgerald
  - Earnest Hemingway
  - John Steinbeck

Processing Data


The_Great_Gatsby.txt — Edited

The Great Gatsby

Francis Scott Fitzgerald

Then wear the gold hat, if that will move her;
If you can bounce high, bounce for her too,
Till she cry "Lover, gold-hatted, high-bouncing lover,
I must have you!"

—THOMAS PARKE D'INVILLIERS

Chapter 1

In my younger and more vulnerable years my father gave me some advice that I've been turning over in my mind ever since.

"Whenever you feel like criticizing any one," he told me, "just remember that all the people in this world haven't had the advantages that you've had."

He didn't say any more, but we've always been unusually communicative in a reserved way, and I understood that he meant a great deal more than that. In consequence, I'm inclined to reserve all judgments, a habit that has opened up many curious natures to me and also made me the victim of not a few veteran bores. The abnormal mind is quick to detect and attach itself to this quality when it appears in a normal person, and so it came about that in college I was unjustly accused of being a politician, because I was privy to the secret griefs of wild, unknown men. Most of the confidences were unsought — frequently I have feigned sleep, preoccupation, or a hostile levity when I realized by some unmistakable sign that an intimate revelation was quivering on the horizon; for the intimate revelations of young men, or at least the terms in which they express them, are usually plagiaristic and marred by obvious suppressions. Reserving judgments is a matter of infinite hope. I am still a little afraid of missing something if I forget that, as my father snobbishly suggested, and I snobbishly repeat, a sense of the fundamental decencies is parcelled out unequally at birth.

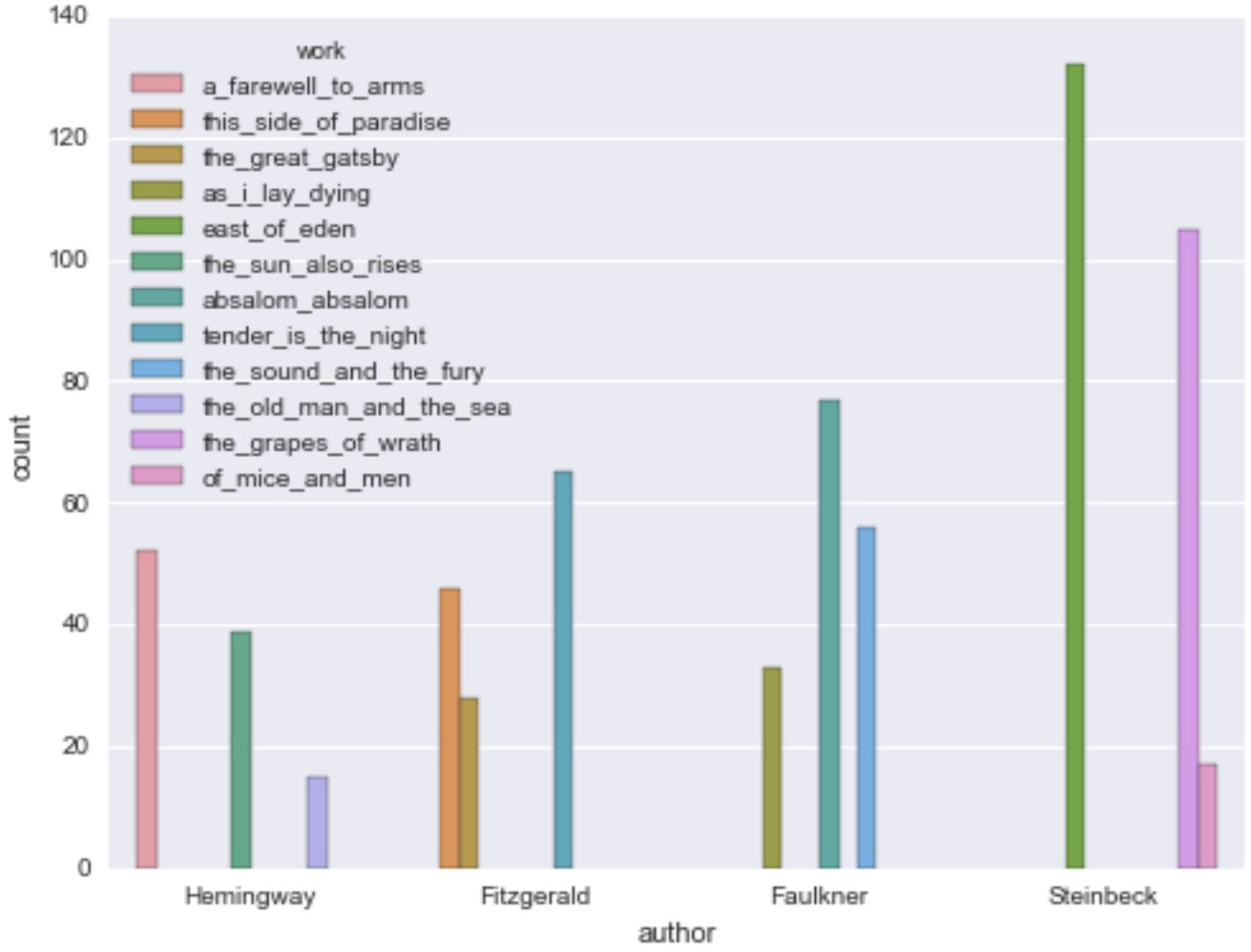
And, after boasting this way of my tolerance, I come to the admission that it has a limit. Conduct may be founded on the hard rock or the wet marshes, but after a certain point I don't care what it's founded on. When I came back from the East last autumn I felt that I wanted the world to be in uniform and at a sort of moral attention forever; I wanted no more riotous excursions with privileged glimpses into the human heart. Only Gatsby, the man who gives his name to this book, was exempt from my reaction — Gatsby, who represented everything for which I have an unaffected scorn. If personality is an unbroken series of successful gestures, then there was something gorgeous about him, some heightened sensitivity to the promises of life, as if he were related to one of those

Blocks of 1700 words

	author	work	block	text	lexical_diversity	grade_level	difficult_words	pct_dialogue
0	Hemingway	a_farewell_to_arms	0	In the late summer of that year we lived in a ...	0.329218	6.5	198	0.106992
1	Hemingway	a_farewell_to_arms	1	They will love you like a son." "He should go...	0.337639	5.5	214	0.678752
2	Hemingway	a_farewell_to_arms	2	two stopped talking and the captain shouted, ...	0.320115	5.5	208	0.487259
3	Hemingway	a_farewell_to_arms	3	encore," said Miss Ferguson. "Not really?" "N...	0.308266	5.5	193	0.429604

Final dataset

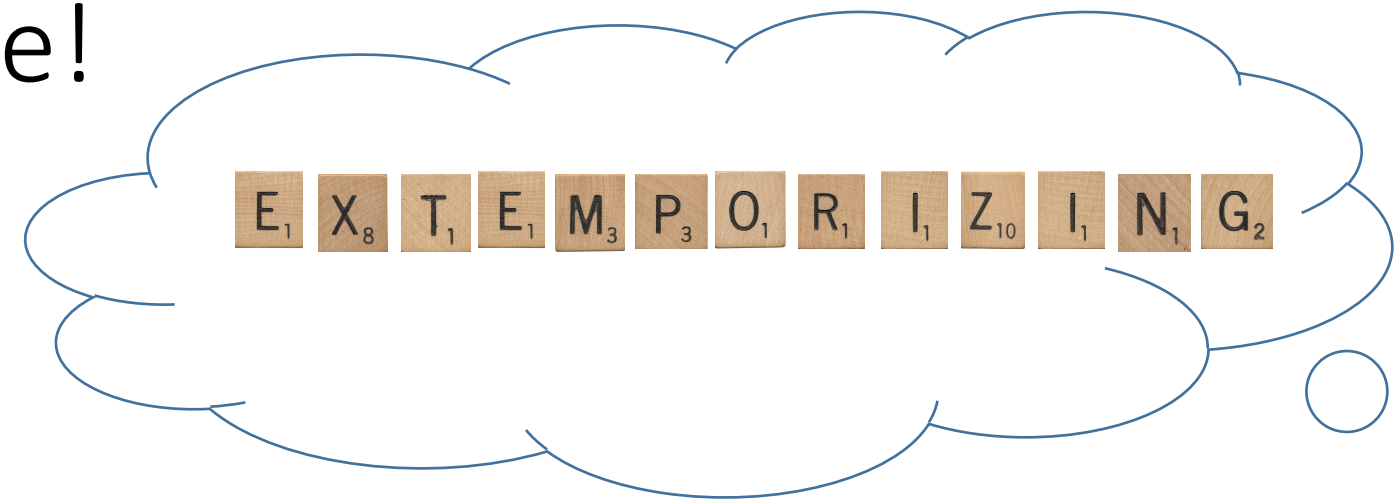
- 665 blocks of text, 33 features
- Class distribution:
  - Steinbeck 254
  - Faulkner 166
  - Fitzgerald 139
  - Hemingway 106
- Null accuracy rate: 38%



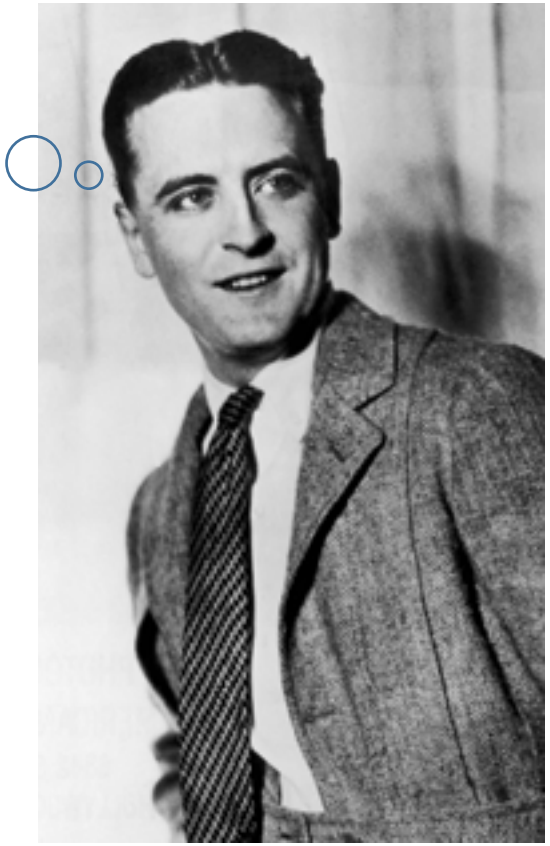
Features

- Word level
  - Length of word
  - Syllables per word
  - Year of most common use
- Sentence level
  - Words per sentence
  - Punctuation use (commas, semicolons, hyphens, etc.)
- Block level
  - Pct. dialogue
  - Reading difficulty
- Lower level features aggregated to block level

Scrabble!

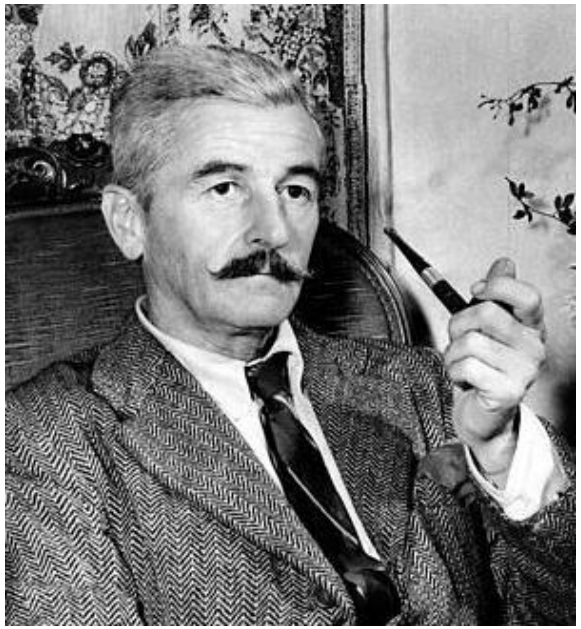


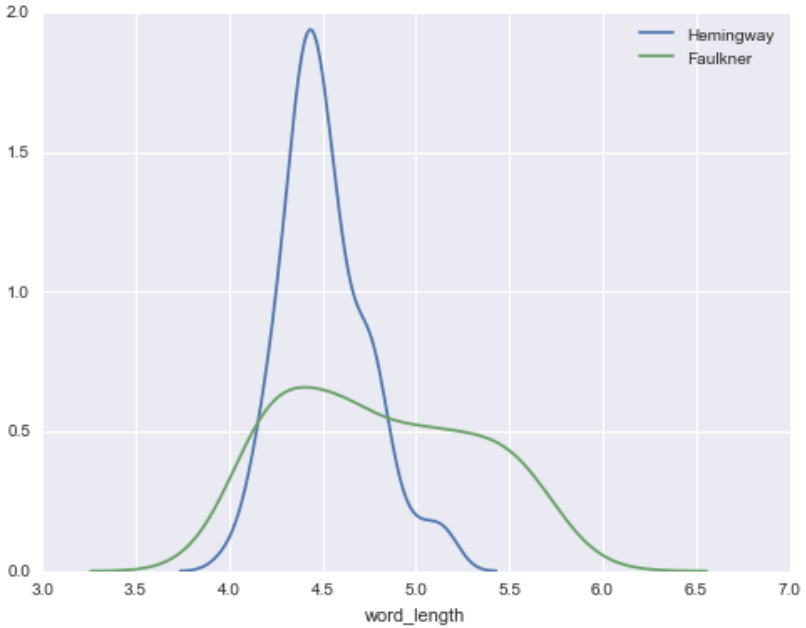
Author	Mean Scrabble Score
Fitzgerald	9.11
Faulkner	8.54
Steinbeck	8.15
Hemingway	8.09



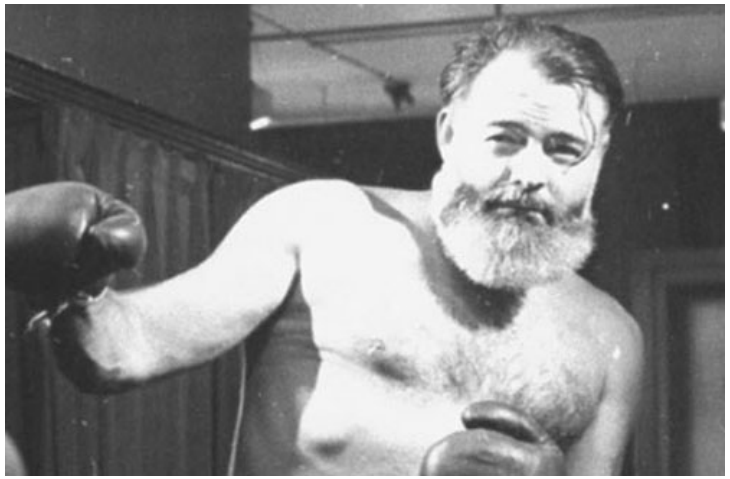
Faulkner vs. Hemingway

"He has never been known to use a word that might cause the reader to check with a dictionary to see if it is properly used."





"Poor Faulkner. Does he really think big emotions come from big words?"



Metric (75th percentile)	Faulkner	Hemingway
Difficult words / block	258	209
Grade level	7.5	5.5
Total punctuation / sentence	2.94	.56

One vs. Rest Classification

- Turn response variable (author) into dummies
- For each block of text, estimate probability it was written by each author separately
- Highest predicted probability wins
- Predicted probabilities across authors don’t have to add up to 1
  - Avoids assumption that authors’ styles are mutually exclusive

Naïve Bayes

- Overall accuracy: 76%
- Predicted probabilities:

Faulkner blocks:

Faulkner_prob	0.538855
Fitzgerald_prob	0.023598
Hemingway_prob	0.384323
Steinbeck_prob	0.376416

Fitzgerald blocks:

Faulkner_prob	0.097287
Fitzgerald_prob	0.984562
Hemingway_prob	0.025160
Steinbeck_prob	0.132055

Hemingway blocks:

Faulkner_prob	0.018931
Fitzgerald_prob	0.039987
Hemingway_prob	0.953860
Steinbeck_prob	0.898222

Steinbeck blocks:

Faulkner_prob	5.650974e-07
Fitzgerald_prob	0.1574381
Hemingway_prob	0.8430932
Steinbeck_prob	0.9482283

Decision Tree

- Overall accuracy: 82%
- Important features by author:

Faulkner:

word_count_std	0.440538
difficult_words	0.172554
hyphens_mean	0.097979
total_punct_mean	0.078675
sentiment	0.047124

Fitzgerald:

hyphens_mean	0.698249
syllable_count_std	0.085571
word_count_mean	0.059230
colons_mean	0.040841
lexical_diversity	0.034442

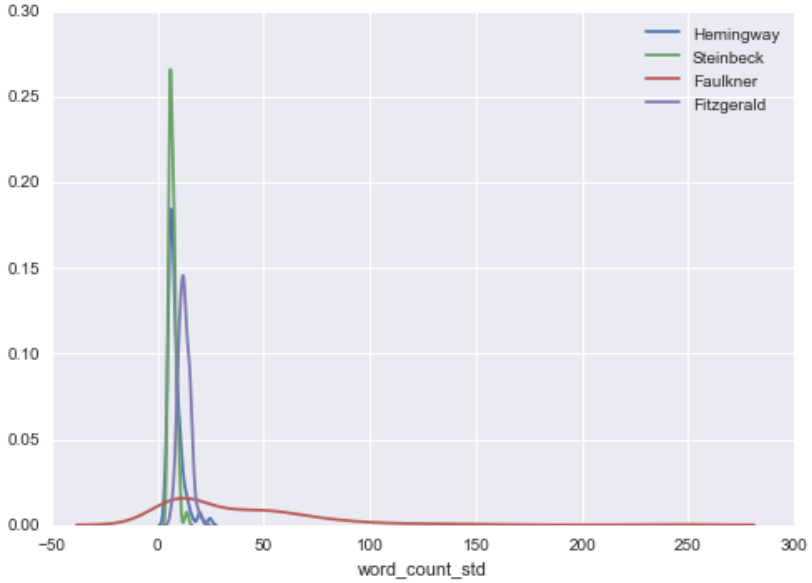
Hemingway:

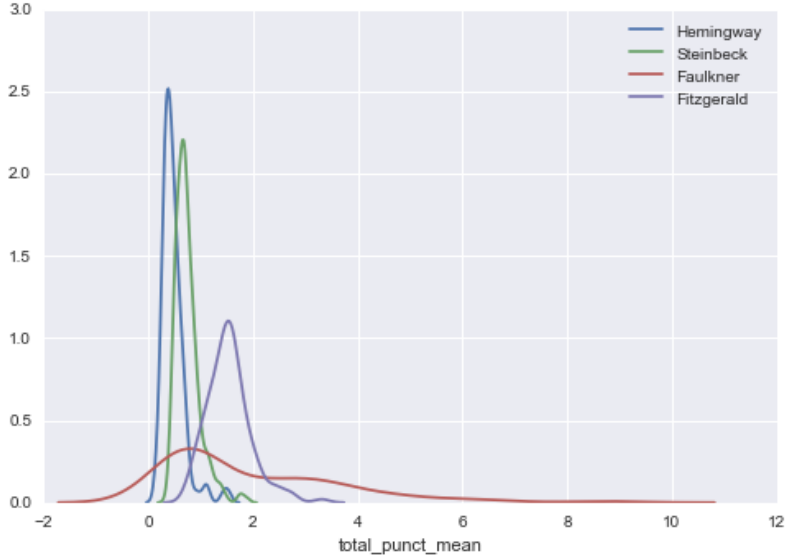
total_punct_mean	0.497074
most_freq_year_25%	0.100032
length	0.077185
lexical_diversity	0.058059
word_count_max	0.043742

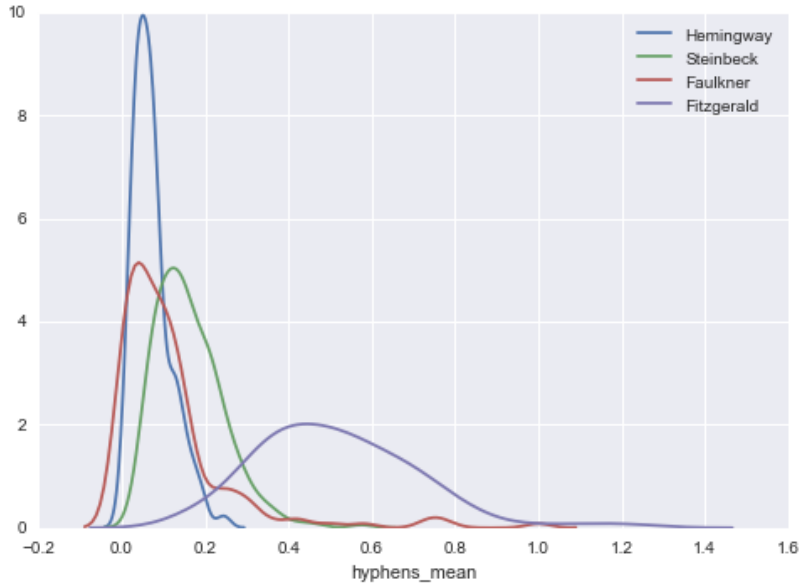
Steinbeck:

word_count_std	0.383177
total_punct_mean	0.213508
colons_mean	0.072063
difficult_words	0.060956
length	0.045470

Distribution of Important Features







Random Forest

- Number of estimators: 150
- Max features: 6
- OOB Score: 92%
- After removing less important features, OOB Score: 91%

	feature	importance
13	hyphens_mean	0.125702
5	word_count_std	0.118986
2	difficult_words	0.085739
0	lexical_diversity	0.075648
8	word_count_max	0.072013
14	total_punct_mean	0.067158
17	syllable_count_std	0.054101
10	commas_mean	0.048380
12	colons_mean	0.048214
24	most_freq_year_50%	0.036376

Conclusions

- Results really dependent on choices in data processing/feature extraction
  - Hyphen/n-dash/m-dash conflation
  - Choice of block size
  - Choice of aggregation for word- and sentence-level features
- Would be interesting to do more NLP-ey features
- Text is incredibly rich…
- BUT there's a lot more to literature than this

The End