“Reading DT Leaves”: A Digital Analysis of the “Lyrical” Novel

Tim Schott
University of Virginia ‘19

Presentation Outline

Motivation

Goal

Methods

Results

Future Work

References

Motivation: **“Lyrical” is everywhere**

Goodreads reviews: “This was the first novel I've read by McCarthy and I'm still sorting out my opinion. **His style is adroit, sometimes lyrical**.” - Adam Meade’s four-star review of *Blood Meridian* by Cormac McCarthy

But how can we make sense of a term this amorphous?

Motivation: **Critical foundation**

Virginia Jackson and *lyricization*: Jackson's work, *Becoming Lyric*, importantly highlights the instability of the lyric label. She writes, "to be lyric is to be read as lyric—and to be read as lyric is to be printed and framed as a lyric." I repurposed her conjectures regarding poetry to fit into this similarly unforgiving terrain

Franco Moretti and "distant reading": Devolves a digital-slant to analysis as he demonstrates important, novel DH work can take place with simple metrics like word frequencies — wherein "distance … is a condition of knowledge"

Martin Heidegger and action: "... action must apply theoretical cognition if it is not to remain blind. Rather, observation is a kind of taking care just as primordially as action has its own kind of seeing. Theoretical behavior is just looking, noncircumspectly. Because it is not circumspect, looking is not without rules; its canon takes shape in method." — *Being and Time* (69)

Goal: ***Define what makes fiction "lyrical" using traditional critique and digital methods***

Digital Pipeline

1. Corpus Creation
    a. clean texts
    b. store data
2. Feature Engineering
    a. calculate potentially useful markers of lyrical novels
3. Feature Reduction
    a. distill the information using Random Forests
4. Supervised Learning
    a. SVM classification

Methods: **Corpus**

Detective Corpus

	Author	Title
1	Agatha Christie	The Secret Adversary
2	Anna Katherine Green	The Leavenworth Case
3	Arthur Conan Doyle	A Study in Scarlet
4	Arthur Conan Doye	The Sign of Four
5	Arthur J. Rees	The Shrieking Pit
6	Arthur J. Rees	The Moon Rock
7	Arthur J. Rees	The Hand in the Dark
8	Carolyn Wells	The Maxwell Mystery
9	Edgar Wallace	The Angel Of Terror
10	Edgar Wallace	The Daffodil Mystery
11	Emmuska Orczy	The Old Man in the Corner
12	Ethel Lina White	The Spiral Staircase
13	Fred Merrick White	The Lady in Blue
14	Fred Merrick White	The Mystery of Room 75
15	G.K. Chesterton	The Innocence of Father Brown
16	Harrington Strong	The Brand of Silence
17	J.S. Fletcher	The Paradise Mystery
18	J.S. Fletcher	The Rayner Slade Amalgamation
19	J.S. Fletcher	The Scarhaven Keep
20	Mary Roberts Rinehart	The Circular Staircase
21	Mrs. Charles Bryce	The Ashiel Mystery
22	R. Austin Freedman	The Red Thumb Mark
23	Raymond Chandler	The Big Sleep
24	Wilkie Collins	The Moonstone

Lyrical Corpus

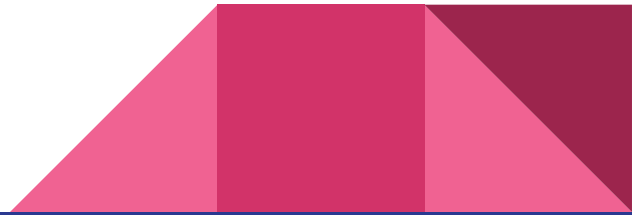
	Author	Title
1	Cormac McCarthy	Blood Meridian
2	Cormac McCarthy	The Road
3	D.H. Lawrence	The Rainbow
4	D.H. Lawrence	The Plumed Serpent
5	D.H. Lawrence	Women in Love
6	Edgar Allen Poe	The Narrative of Arthur Gordon Pym of Nantucket
7	Edgar Allen Poe	Eureka: A Prose Poem
8	F. Scott Fitzgerald	The Great Gatsby
9	Herman Melville	Billy Budd
10	Herman Melville	Moby Dick
11	James Joyce	Portrait of the Artist as a Young Man
12	Jean Rhys	Wide Sargasso Sea
13	Joseph Conrad	Heart of Darkness
14	Joseph Heller	Something Happened
15	J. M. Coetzee	Life & Times of Michael K
16	Malcolm Lawry	Under the Volcano
17	Oscar Wilde	The Picture of Dorian Gray
18	Thomas Pynchon	Gravity's Rainbow
19	Virginia Woolf	Mrs. Dalloway
20	Virginia Woolf	Orlando
21	Virginia Woolf	To The Lighthouse
22	Vladimir Nabokov	Pale Fire
23	Vladimir Nabokov	Lolita
24	William Faulkner	Absalom Absalom
25	William Faulkner	The Sound and the Fury
26	William H. Gass	In The Heart of the Heart of the Country

Methods: **Data mining and processing**

Created database using SQLite containing every word, sentence and paragraph from all 50 works.

Used R and Python for text cleaning - without corpus-creation packages like tm, coRpus or nltk.

Most works were sourced from Project Gutenberg or Archive.org, so their encoding was fairly uniform.



Methods: **Feature engineering**

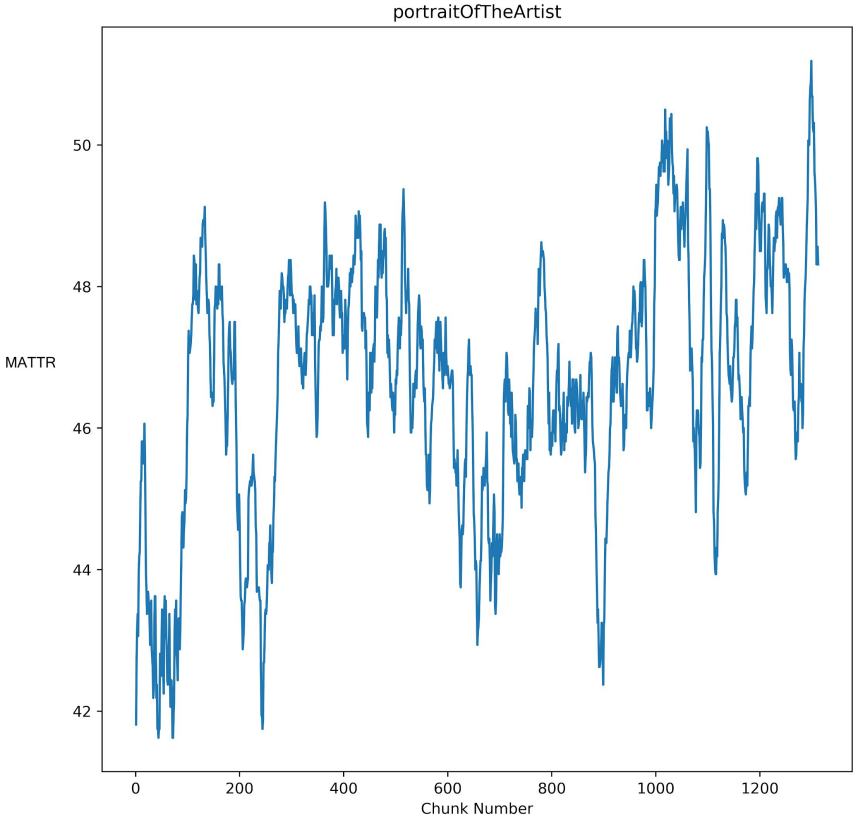
	lolita	mobyDick	mrsDalloway	orlando	paleFire	portraitOfTheArtist	pym
labels_vec	lyrical	lyrical	lyrical	lyrical	lyrical	lyrical	lyrical
word_counts_vec	112193	214183	64267	79547	68132	84927	100954
sent_counts_vec	5169	7381	3405	3292	2546	4452	3680
para_counts_vec	1189	2434	754	451	971	2208	625
commas_vec	8783	16135	6098	6047	4387	4258	8674
sent_comma_freq_vec	1.6991681	2.1860182	1.7908957	1.8368773	1.7230951	0.9564241	2.3570652
para_comma_freq_vec	7.3868797	6.6290058	8.0875332	13.4079823	4.5180227	1.9284420	13.8784000
words_per_sentence_vec	21.704972	29.018155	18.874302	24.163730	26.760408	19.076146	27.433152
words_per_paragraph_vec	94.35913	87.99630	85.23475	176.37916	70.16684	38.46332	161.52640
sents_per_paragraph_vec	4.347351	3.032457	4.515915	7.299335	2.622039	2.016304	5.888000
consecutive_counts_vec	347	262	265	243	150	421	284
consecutive_repeat_freq_vec	0.06713097	0.03549654	0.07782672	0.07381531	0.05891595	0.09456424	0.07717391
syll_and_word_freq_vec	1.419224	1.374087	1.383183	1.353024	1.449994	1.345332	1.454187
polsyll_and_word_freq_vec	0.09382938	0.08199997	0.07803383	0.07512540	0.10607350	0.06494990	0.10922796
syll_and_sent_freq_vec	30.804217	39.873459	26.106608	32.694107	38.802435	25.663747	39.892935
polsyll_and_sent_freq_vec	2.0365641	2.3794879	1.4728341	1.8153098	2.8385703	1.2389937	2.9964674
unique_counts_vec	14113	16747	7088	9397	11456	9190	9170
type_token_ratio_vec	0.12579216	0.07819015	0.11028988	0.11813142	0.16814419	0.10821058	0.09083345
mean_usage_frequency_vec	7.949621	12.789335	9.067015	8.465148	5.947277	9.241240	11.009160
median_MATTR	51.25	50.12	48.75	49.19	51.00	45.94	50.00
object_freq	0.03149929	0.03419506	0.02724571	0.03012056	0.02972172	0.02664641	0.02808210
relationship_freq	0.01655184	0.01973079	0.01588685	0.01544999	0.01646803	0.01523662	0.01571012
time_freq	0.04527912	0.05097510	0.04204335	0.04242775	0.04262314	0.04193013	0.04627850
self_freq	0.048077866	0.016028350	0.003112017	0.003142796	0.024790113	0.010185218	0.030033481
perceive_freq	0.01029476	0.01213915	0.01177899	0.01087407	0.01108143	0.01254018	0.01107435
i_freq	0.026579198	0.009818706	0.001571569	0.001684539	0.013826102	0.006358402	0.015888424
top_ten_freq	0.2466999	0.2344257	0.2447913	0.2578098	0.2466536	0.2763197	0.2614755
dialogue_freq	0.2380151	0.2002876	0.1942971	0.1873614	0.1266735	0.2504529	0.1200000
question_vec	414	870	361	218	161	556	120
exclamation_vec	262	1493	346	199	103	434	260
sentiment_vec	0.008131482	-0.000804251	0.010923513	0.006561174	0.025613752	-0.010018552	0.000531653

Pictured is the feature subset for: *Lolita, Moby Dick, Mrs. Dalloway, Orlando, Pale Fire, Portrait of The Artist* and *The Narrative of Arthur Gordon Pym*

A broad array of data-points that could potentially lead to a more comprehensive understanding of the formal elements that generate a“lyrical” work

Calculated with R and Python, committed to spreadsheet (later normalized).

Methods: **MATTR (Moving-Average-Type-Token-Ratio)**

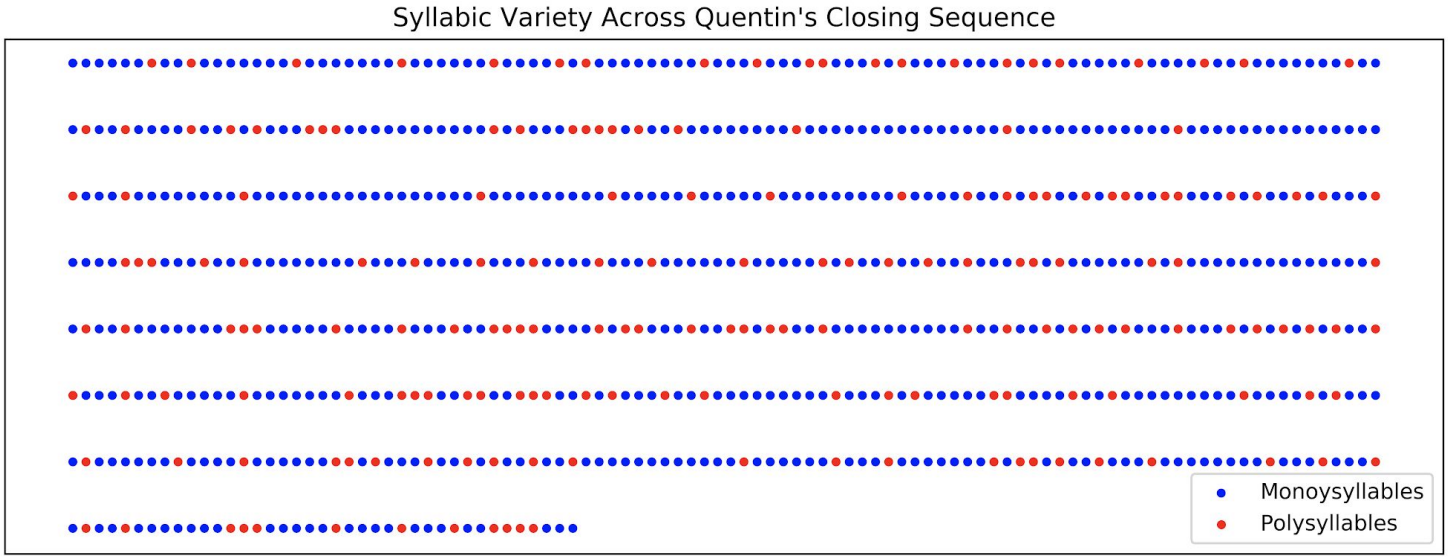


An improvement upon a commonly adopted metric.

Calculated with Python and graphed with matplotlib.

We can watch Stephen steadily increasing the number of unique words he uses as he matures into an aesthete and cements his ideals

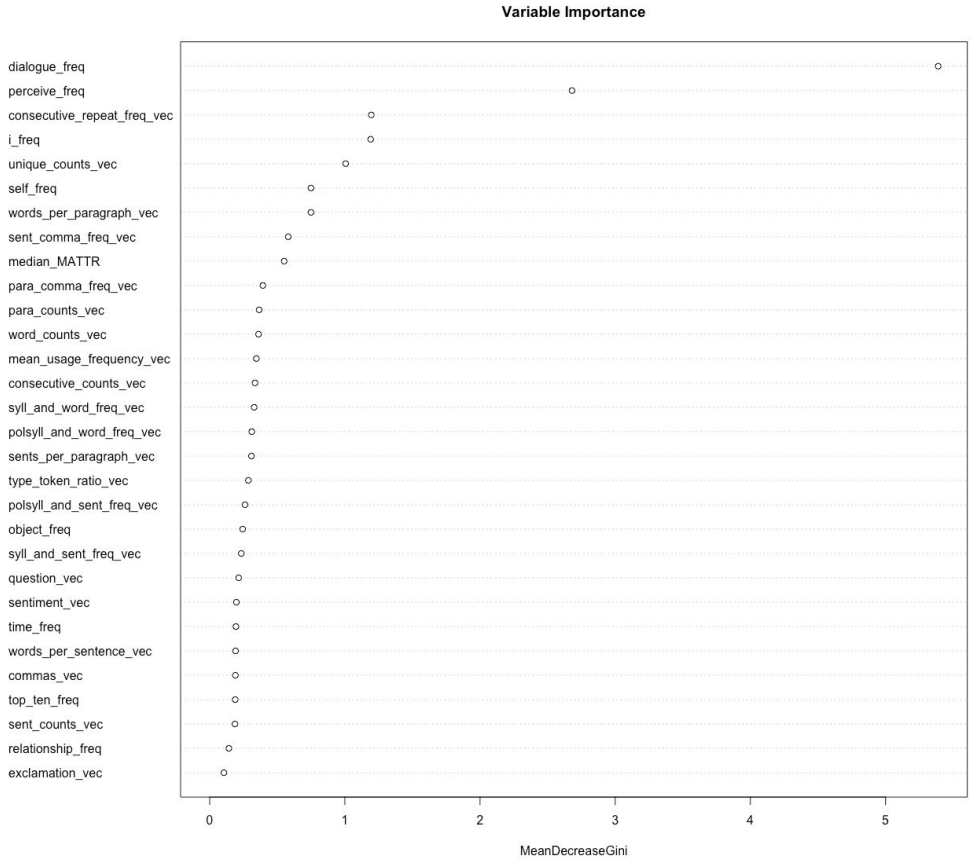
Methods: **Syllabic Richness in *The Sound and the Fury***



"...i was afraid to i was afraid she might and then it wouldnt have done any good but if i could tell you we did it would have been so and then the others wouldnt be so and then the world would roar away..."

Quentin Compson

Methods: **Feature set reduction through variable importance**



Top 4:

frequency of dialogue

frequency of perception words

frequency of anaphora

frequency of “I”

Methods: **Supervised Learning**

I implemented an SVM using the *e1071* package in R.

I built a final SVM using the two most extreme variables from an initial trial run. The levels *anaphora-frequency* and *perception-frequency* across my corpus were placed into this model.

I used "Leave-One-Out-Cross-Validation" (LOOCV) and a cost parameter of 2.5 and achieved 84% accuracy.

According to my classifier: novels with **high levels of anaphora** and **low levels of perception** belong in the "lyrical" class.

Misclassified Novels

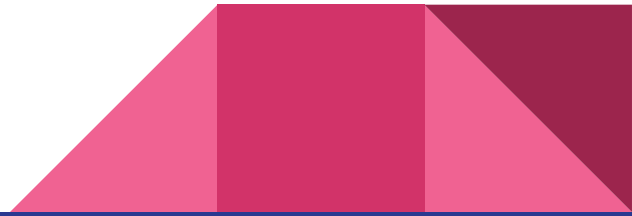
	Title	Author	Classified Label	Correct Label	Frequency of Anaphora	z-score	Frequency of Perception Words	z-score
1	Billy Budd	Herman Meville	detective	lyrical	0.03518	-1.69	0.013597	0.69
2	Eureka: A Prose Poem	Edgar Allan Poe	detective	lyrical	0.041651	-1.44	0.011179	-0.86
3	Heart of Darkness	Joseph Conrad	detective	lyrical	0.076887	-0.05	0.012584	0.04
4	The Sound and the Fury	William Faulkner	detective	lyrical	0.093796	0.61	0.01236	-0.1
5	Wide Sargasso Sea	Jean Rhys	detective	lyrical	0.069218	-0.35	0.013172	0.42
6	The Big Sleep	Raymond Chandler	lyrical	detective	0.103745	2.68	0.013809	-0.88
7	The Circular Staircase	Mary Roberts Rinehart	lyrical	detective	0.064207	0.44	0.013025	-1.33
8	The Mystery of Room 75	Fred Merrick White	lyrical	detective	0.073456	0.96	0.013804	-0.88

Results: **Technical and critical summary**

I did not formulate an unassailable blueprint for creating “lyrical” novels, but I think that’s okay.

Instead, I created a reusable database of cleaned literature for future study. I calculated remarkably minute yet consequential markers of syntax and style in the novel. I interfaced with the critical and theoretical authorities in an untraditional manner. And, using an intricate machine learning pipeline, I stake my claim:

The most salient feature of “lyrical” novels is their reliance on anaphora.



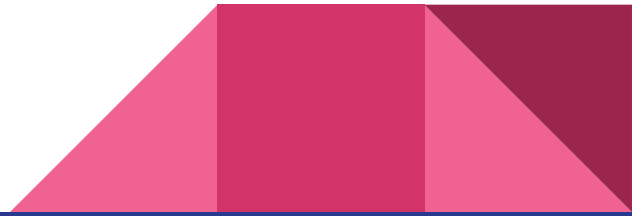
Future Work

Larger corpus (can never be too big)

Use unsupervised learning (Keras Attentive LSTM)

Branch into different mediums (music lyrics, tweets, etc.)

Delve into the sonic aspect of literature (what do lyrical novels *sound* like?)



References

contact email: tcs9pk@virginia.edu

github: github.com/timschott

full thesis: timschott.github.io/thesis

citations: github.com/timschott/dmp/references/11-09-dhcs.txt