Pull Out All the Stops: Textual Analysis via Punctuation Sequences

Mason A. Porter (@masonporter), Department of Mathematics, UCLA

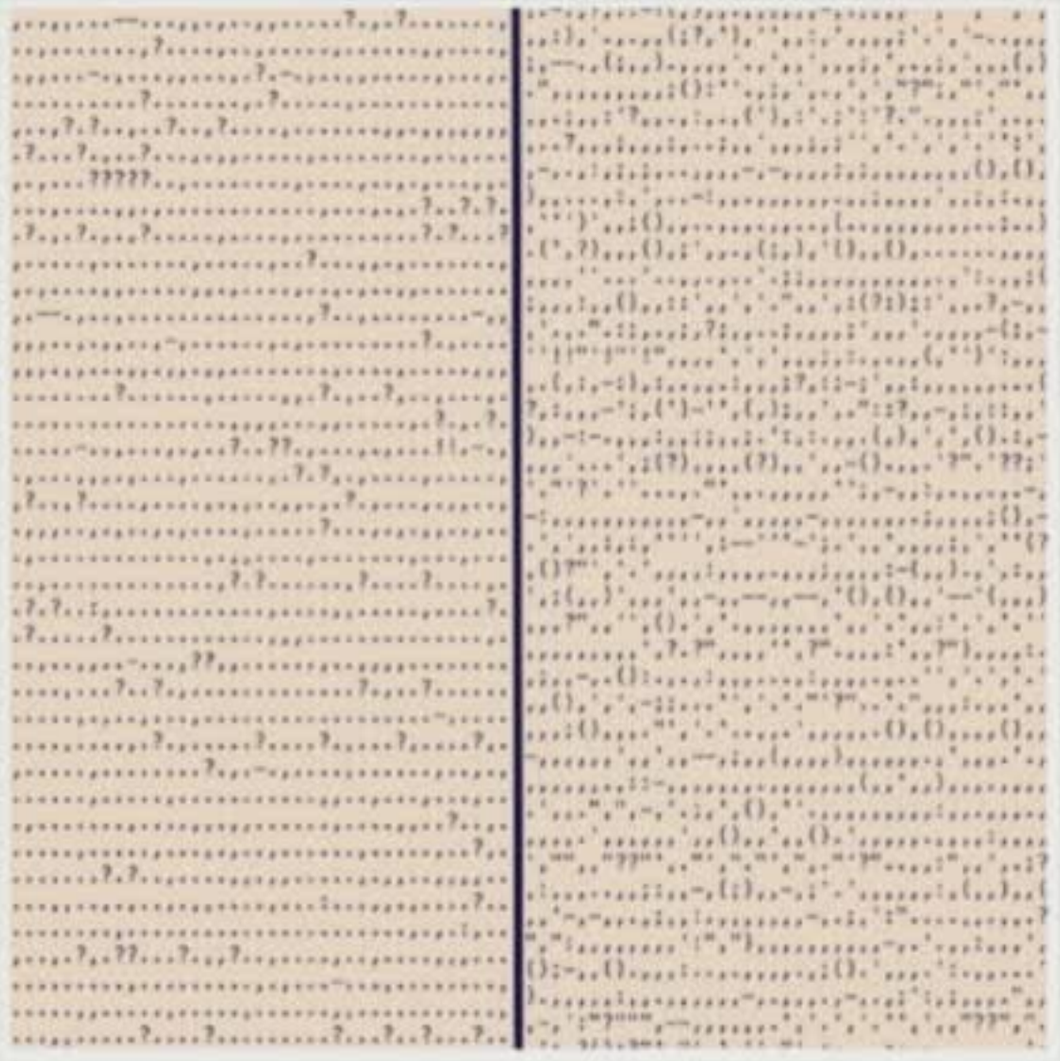
Collaborators: Alexandra Darmon, Marya Bazzi, Sam Howison

ArXiv:1901.00519; SocArxiv: https://osf.io/preprints/socarxiv/2rzsg

A Striking Blog Entry

- Adam J. Calhoun, "Punctuation in Novels"
  - https://medium.com/@neuroecology/punctuation-in-novels-8f316d542ec4

Cormac McCarthy,
Blood Meridian



William Faulkner,
Absalom, Absalom!

Example Quote from Ursula K. Le Guin
(from her 2014 collection, *Wave of the Mind*)

I don't have a gun and I don't have even one wife and my sentences tend to go on and on and on, with all this syntax in them. Ernest Hemingway would have died rather than have syntax. Or semicolons. I use a whole lot of half-assed semicolons; there was one of them just now; that was a semicolon after "semicolons," and another one after "now."

Sequences of Punctuation

Let $\Theta = \{\theta_1, \ldots, \theta_{10}\}$ denote the (unordered) set of ten punctuation marks (see section 2.1). Let n denote the total number of documents in our corpus; and let $D_k = \{\theta_1^k, \ldots, \theta_{n_k}^k\}$, with $k \in \{1, \ldots, n\}$, denote the ordered set of n_k punctuation marks in document k. As an example, consider the following quote by Ursula K. Le Guin (from

The set D_k for this quote is $\{, , . , . , . , ; , ; , “ , , , ” , “ , . , ”\}$, with $n_k = 12$. From D_k, we can calculate $\boldsymbol{f}^{1,k}$, $\boldsymbol{f}^{2,k}$, and $\boldsymbol{f}^{3,k}$.

f^4, f^5, f^6: considering the numbers of words between punctuation marks

tween punctuation marks. Let $D_k^w = \{w_0^k, w_1^k, \ldots, w_{n_k-1}^k\}$ denote the number of words that occur between successive punctuation marks in D_k, with w_0^k equal to the number of words before the first punctuation mark. Therefore, w_1^k is the number of words between punctuation marks θ_1^k and θ_2^k, and so on. The set D_k^w for Le Guin's comment is $\{25, 6, 9, 2, 9, 7, 5, 1, 0, 4, 1, 0\}$, where we count "don't" as two words and we also count "half-assed" as two words. The minimum number of words that can occur between suc-

```
I don't have a gun and I don't have even one wife and my sentences tend
to go on and on and on, with all this syntax in them. Ernest Hemingway would
have died rather than have syntax. Or semicolons. I use a whole lot of half-
assed semicolons; there was one of them just now; that was a semicolon after
"semicolons," and another one after "now."
```

f^5: quantifies frequencies of the numbers of words between successive punctuation marks

The entries of the feature $\boldsymbol{f}^{5,k} \in [0,1]^{n_s \times 1}$, which quantifies the frequency of the number of words between successive punctuation marks, are

$$f_i^{5,k} = \frac{|\{w_l^k \in D_k^w \,|\, w_l^k = i\}|}{n_k} \,. \tag{2.5}$$

In the Le Guin quote, recall that $D_k^w = \{25, 6, 9, 2, 9, 7, 5, 1, 0, 4, 1, 0\}$ (which includes 9 unique integers), so the $n_s \times 1$ vector (with $n_s = 40$, as mentioned above) $\boldsymbol{f}^5$ has 9 nonzero entries. For example, $\boldsymbol{f}_1^5 = 2/12$ (because 0 occurs twice out of $n_k = 12$ total punctuation marks) and $\boldsymbol{f}_4^5 = 0$ (because 3 never occurs out of $n_k = 12$ possible times).

The feature $f^{5,k}$ tracks word-count frequency between successive punctuation marks, without distinguishing between different punctuation marks.

Comparing Feature Vectors: Kullback-Leibler (KL) Divergence

- Need a measure of similarity

- KL divergence (aka: relative entropy)
  - Discrete probability distributions p, q
  - Measures how much one probability distribution deviates from a second
  - Smaller means less deviation

$$KL(p \mid q) = \sum_{i=1}^{n} p_i \log (p_i / q_i)$$