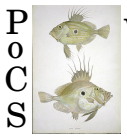


Principles of Complex Systems, Vols. 1, 2, & 3D
CSYS/MATH 300, 303, & 394
University of Vermont, Fall 2022
Solutions to Assignment 20
"Permission to yell in a bloodcurdling way, sir?"

Name: Patrick L. Harvey

Conspirators: Me, myself, and Irene (2000)

Bonus March facts:

1. The 0th day of March (or the last day of February) is always a Doomsday.
2. Pi Day is always a Doomsday.
3. For 2023: The Doomsday is Tuesday.

1. (3, 3, 3)

   Using the main text you chose at the start of the seasomester, plot happiness time series in the following ways using the labMT lexicon.

   The labMT word list was published with Ref. [**?**] in 2011, and has been occasionally upgraded to accommodate major changes in language use.

   See https://hedonometer.org and https://storywrangling.org for the current version.

   (a) Process (destroy) your text so that it is a simple text file with one 1-gram per line—a vector of 1-grams.

   To the extent possible, keep punctuation in as separate 1-grams. Periods, commas, semicolons, em dashes, ellipses, . . .

   You can submit this is a separate file, but okay to just say you've succeeded.

   (b) First use the full lexical lens provided by labMT.

   Make a single figure containing a stacked set of 7 plots with text windows of size $T = [10^z]$ for $z = 1, 1.5, 2, 2.5, 3, 3.5,$ and 4.0.

   Stacked here means separated and stacked vertically, as opposed to directly overlaid. See examples for Moby Dick at the end of this assignment.

   The notation $[\cdot]$ means round to the nearest integer.

   (c) Choose a 'good' text window from above, and repeat the analysis with lenses which exclude the central words around the neutral point.

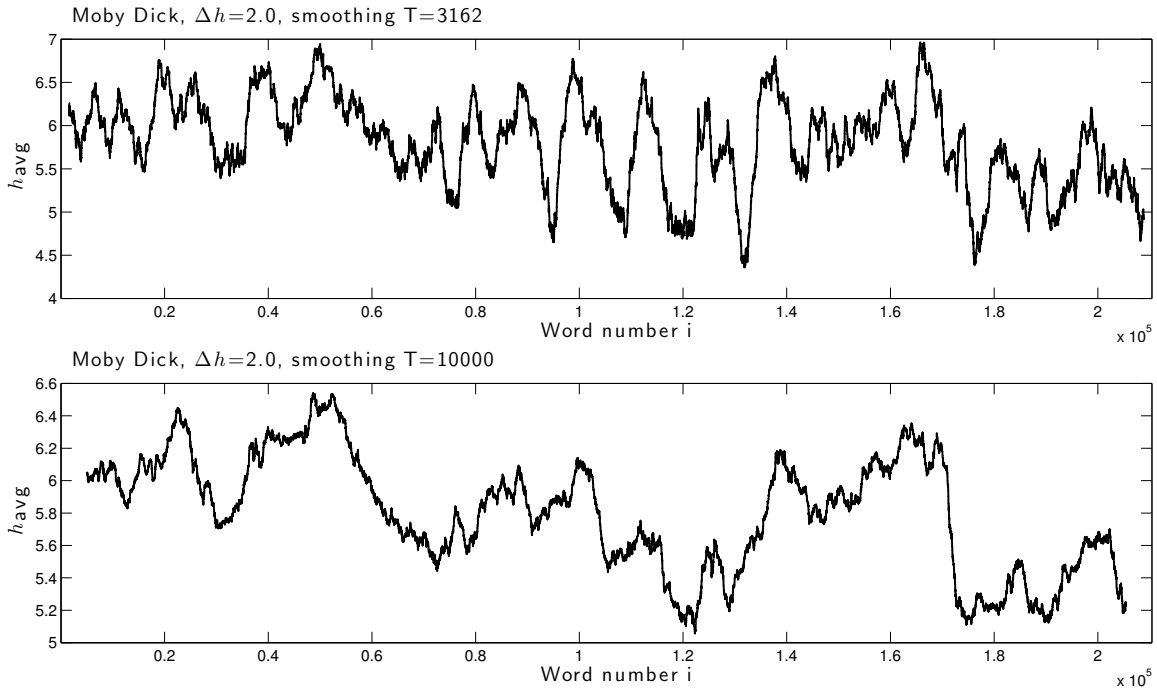
The blocked words are $h_{\rm avg} \pm \delta h_{\rm avg}$ where $\delta h_{\rm avg} = 0.5, 1.0, 1.5, 2.0, 2.5, 3.0,$ and 3.5.

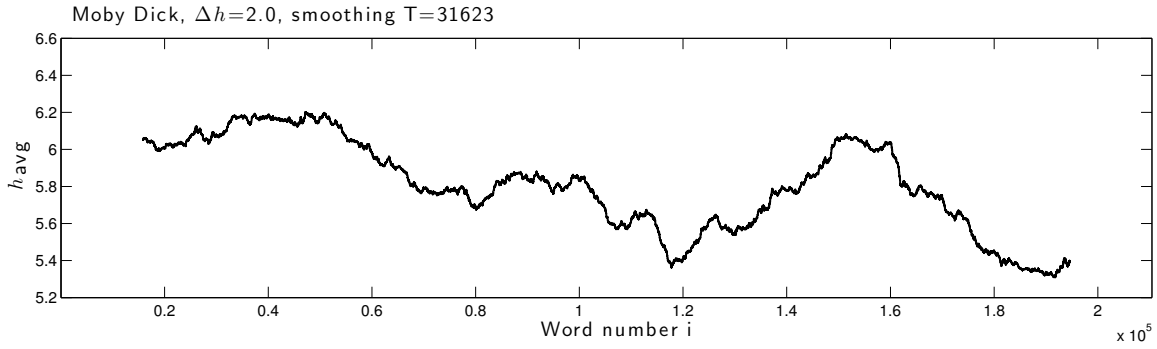
By 'good', we mean one that seems to you to produce a reasonable smoothing. Not too choppy, not too washed out.

Notes:

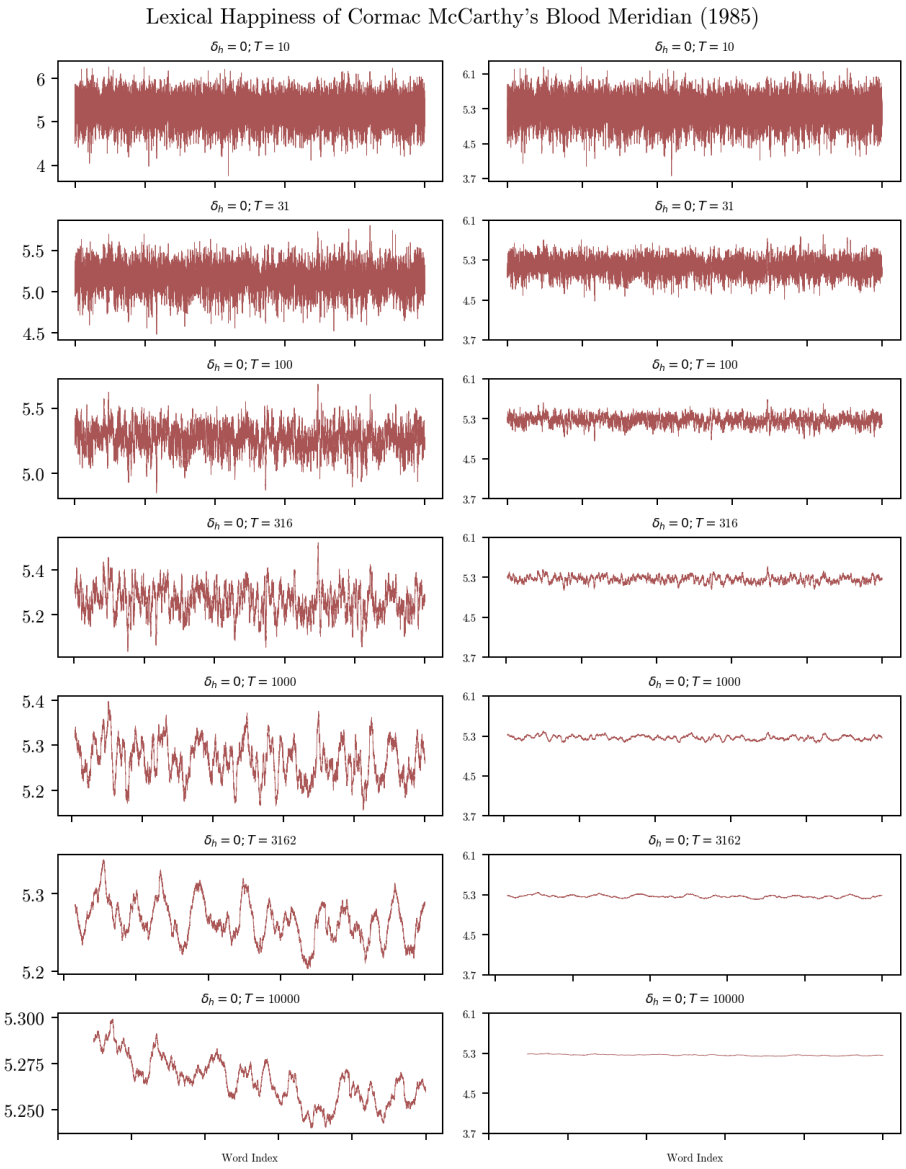
- The horizontal axis is "reading-experience time" corresponding to 1-grams in the text, running from 1 to N.
- The windows should overlap, sliding one word ahead each time. This is a simple averaging filter.
- Points should be located above the center of each window.
- So the point for the window running from n to $n + T - 1$ (T words) will be located at $n + (T - 1)/2$.
- Do not pre-filter the text for any given lens. Windows will contain variable numbers of words with and without happiness scores.

Three example averaging windows for Moby Dick with $\delta h_{\rm avg} = 2.0$:

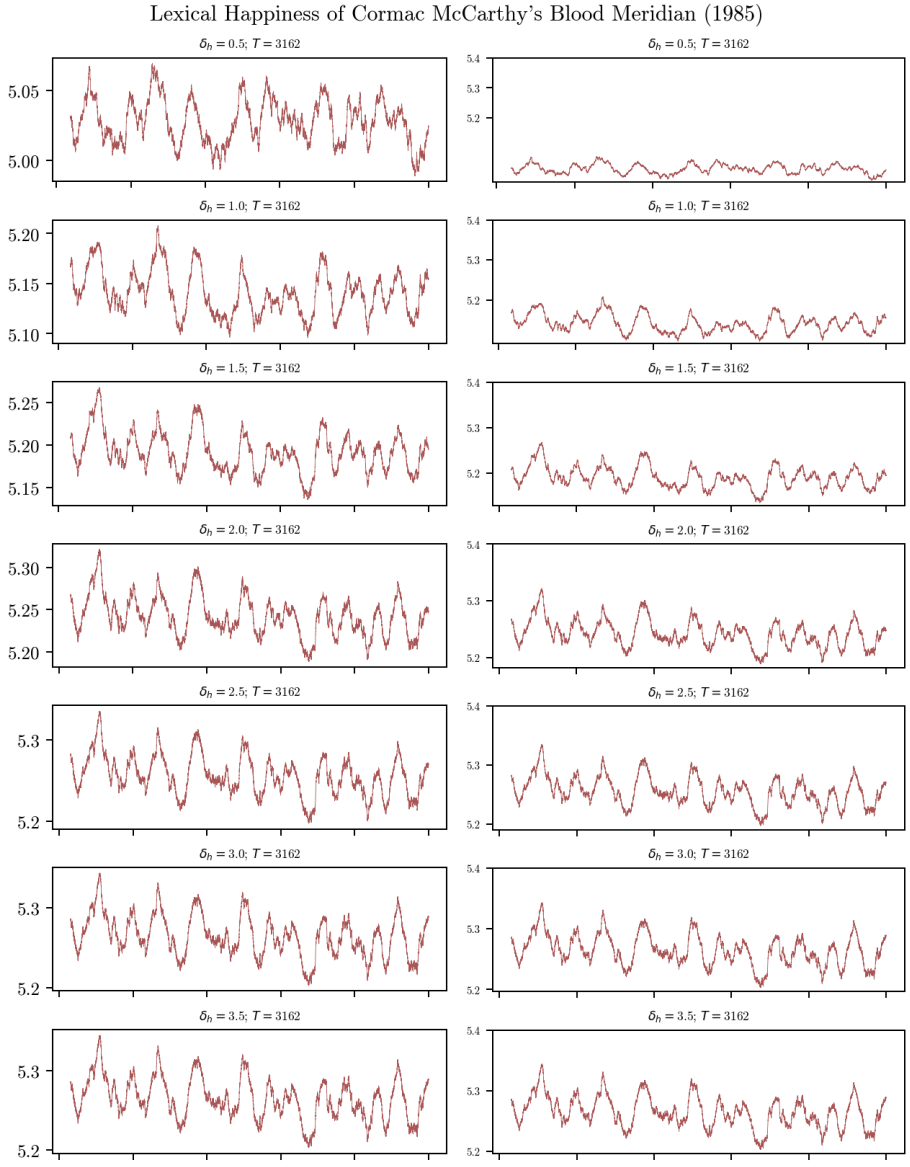


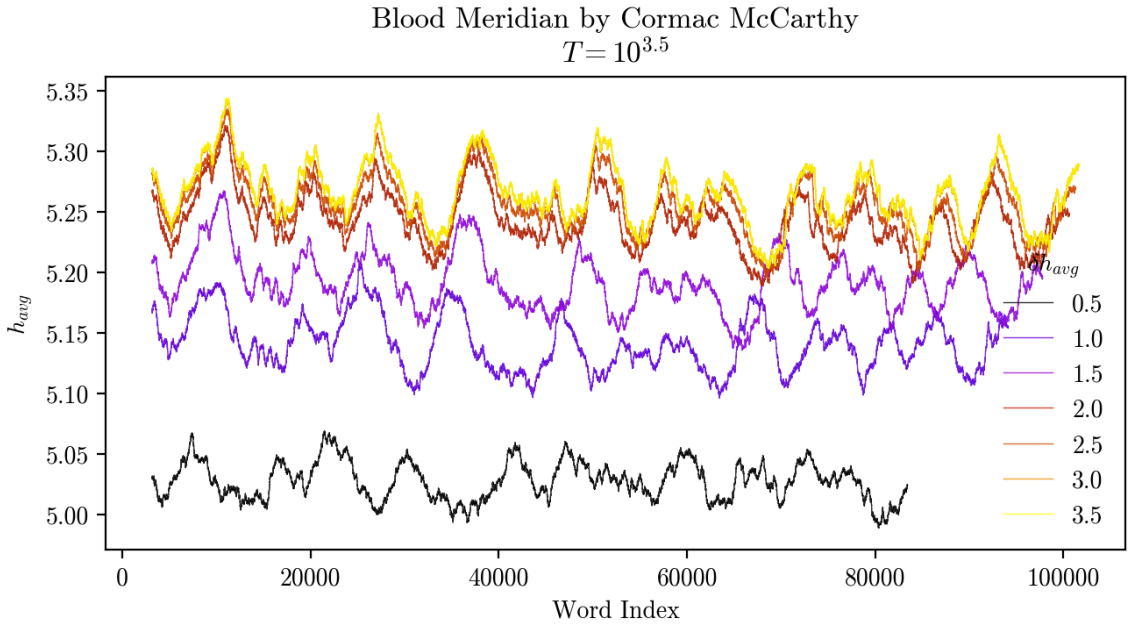

Moby Dick, Δh=2.0, smoothing T=31623
h_{avg}
6.6
6.4
6.2
6
5.8
5.6
5.4
5.2
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
Word number i
x 10^5

Solution: Part a can be viewed here if required. The figure below illustrates how the dynamic window size (T) affects the clarity of the lexical happiness in the text. That is, when we increase T, we are looking at (and averaging over) more words in the text. The result is essentially a smoothing ($+|-$) of the turbulence for happiness. However, we lose an amount of information proportional to T. Looking at the right hand side of the figure, we can clearly see this smoothing/noise reduction as $T \rightarrow \text{larger}$. Additionally, we see the loss of information at the left hand side of the lines for each increase in T. The only difference between the plots on the left hand side, versus the plots on the right hand side, are whether the y-axis is fixed (right) or variable (left).



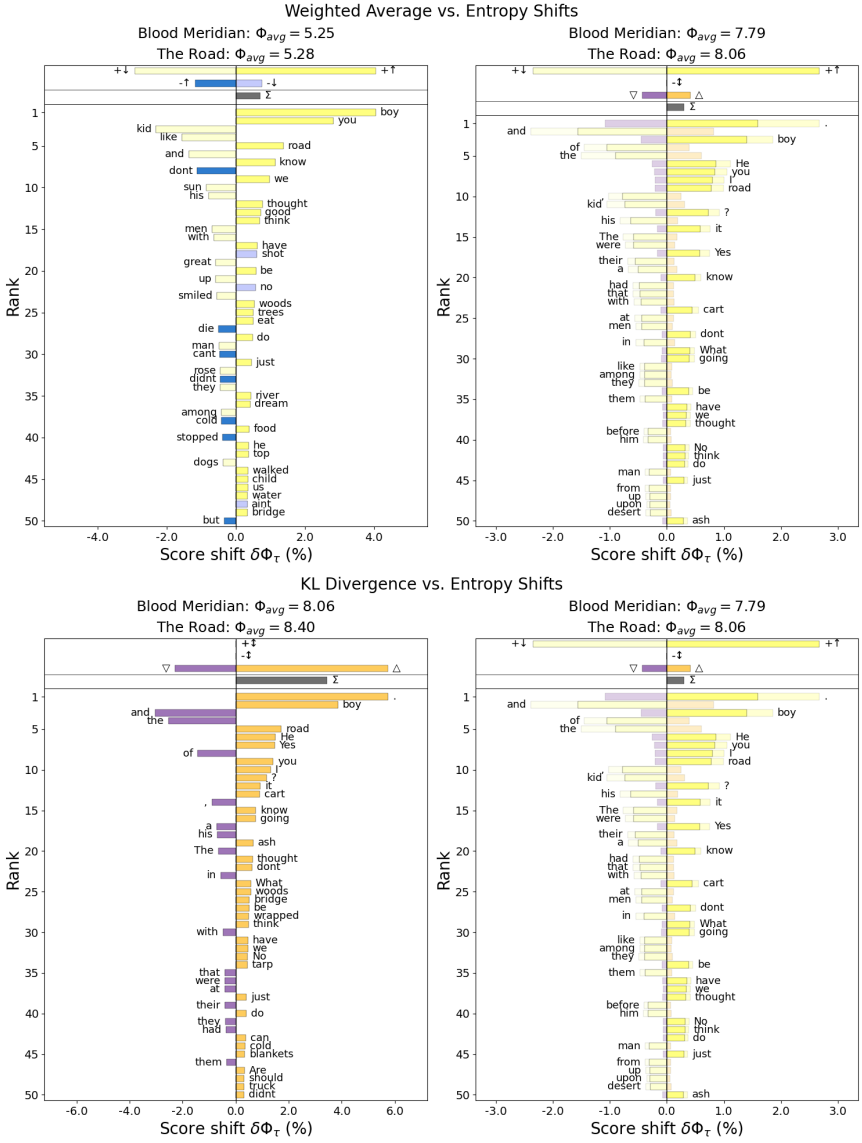
The figure below illustrates how the dynamic neutral point (δh_{avg}) affects the clarity of the lexical happiness in the text. That is, when we increase δh_{avg}, we are looking at fewer words around the mean happiness of our text window (T). The result is essentially a shifting ($\uparrow \mid \downarrow$)of the lexical happiness. However, the amount of shift is relative to h_{avg}. Looking at the right hand side of the figure, we can clearly see this shifting as $h_{\text{avg}} \rightarrow \text{larger}$. Additionally, we see the same loss of information as before for our fixed value of T ($10^{3.5} \approx 3162$). That is, the first/last $\frac{T}{2}$ words cannot be averaged against the non-existent words before/after them since T is constant. We could use alternative methods to address this issue, however we will simply use a linear scan of the text with fixed T for now. The only difference between the plots on the left, versus the plots on the right, are whether the y-axis is fixed (right-side) or variable (left-side).




Blood Meridian by Cormac McCarthy
$T = 10^{3.5}$
h_{avg}
5.35
5.30
5.25
5.20
5.15
5.10
5.05
5.00
0
20000
40000
60000
80000
100000
Word Index
αh_{avg}
0.5
1.0
1.5
2.0
2.5
3.0
3.5

Until next time, we can ponder over why Cormac McCarthy's The Road may be happier than Blood Meridian.

Spoiler: The latter may be a bit more gruesome.



□