PoCS What's The Story?

Principles of Complex Systems, Vols. 1, 2, & 3D
CSYS/MATH 300, 303, & 394
University of Vermont, Fall 2022
Assignment 23

In the beginning there was nothing, which exploded

Due: Weekly
https://pdodds.w3.uvm.edu/teaching/courses/2022-2023pocsverse/assignments/23/
Some useful reminders:
Deliverator: Prof. Peter Sheridan Dodds (contact through Teams)
Assistant Deliverator: Dylan Casey (contact through Teams)
Office: The Ether
Office hours: See Teams calendar
Course website: https://pdodds.w3.uvm.edu/teaching/courses/2022-2023pocsverse
Overleaf: LaTeX templates and settings for all assignments are available at
https://www.overleaf.com/project/631238b0281a33de67fc1c2b.

All parts are worth 3 points unless marked otherwise. Please show all your workingses clearly and list the names of others with whom you ~~conspired~~ collaborated.

For coding, we recommend you improve your skills with Python, R, and/or Julia. The (evil) Deliverator uses (evil) Matlab.

Graduate students are requested to use LaTeX (or related TeX variant). If you are new to LaTeX, please endeavor to submit at least n questions per assignment in LaTeX, where n is the assignment number.

Assignment submission:

Via Blackboard.

Let's write some papers.

This assignment will be built out over the last part of the semester. It will change in unpredictable ways.

Gitlab repository for ousiometry and telegnomics:

https://gitlab.com/petersheridandodds/ousiometry/

Please connect on Gitlab.

Python shifterator package:

https://github.com/ryanjgallagher/shifterator

Notes:

- We will perform analyses using with the original NRC VAD lexicon (about 20,000 words) as well as potentially one augmented with conjugations and plurals (about 32,000 words).
- Hopefully, the augmented lexicon—which simply covers more types—will perform well and we can just use that in our papers.
- In the foundational paper [1], and as discussed in lectures, we have determined that power-danger-structure is the framework of essential meaning.
- In the main data set, words are provided with scores in the VAD, GES, and PDS frameworks. Some of the Matlab scripts render VAD, GES, and PDS. Going forward, we will use the PDS framework only. Going forward, we will use the PDS framework only.
- We will still use the GES scores but relabel the main dimensions as Dangerous-Power (for Energy) and Safe-Power (SP, for Goodness).

  As a reminder, in terms of Danger and Power, we have: $E = \frac{1}{\sqrt{2}}(D + P)$ and $G = \frac{1}{\sqrt{2}}(-D + P)$.

  But again, we will not speak of Energy and Goodness going forward.
- We are using the compass of essential meaning with Danger as North, Power as East, Dangerous-Power at Northeast, etc.
- Some of this may be repeated below.

Notes on figures and captions and the Dr. Seuss method for writing a paper.

- Insert the figures you make into your paper and prepare a caption and discussion for the main text.
- Captions should be informative and reasonably self-contained. Some people will only look at figures in a paper, others will do so before reading the whole paper. Paper reading is nonlinear.
- For captions, see the ousiometrics paper as a recent example: https://storylab.w3.uvm.edu/ousiometrics/
- For the main text, prepare one or more paragraphs. The caption and main text should be distinct text.

1. Download the paper template thing here:

   https://github.com/petersheridandodds/universal-paper-template

   - Follow instructions as outlined in class and in the README to make a local version of the paper

     See Lecture #46 in stories here: https://pdodds.w3.uvm.edu/teaching/courses/2021-2022principles-of-complex-systems/stories/.
   - Renaming "paper-template" to something meaningful is important.

     Example:

     ...
     telegnomics-of-pratchett-discworld-series.bib
     telegnomics-of-pratchett-discworld-series.biblio.tex
     telegnomics-of-pratchett-discworld-series.body.tex
     telegnomics-of-pratchett-discworld-series.contributions.tex
     telegnomics-of-pratchett-discworld-series.keywords.tex
     ...
   - Run make-name-match-settingsfile.pl to adjust an internal reference to the base name.
   - Run the make-zip-file-for-overleaf.sh script to create overleaf.zip.
   - Upload to your Overleaf account.
   - Rename the project on Overleaf to match the naming convention. Prepend with YYYY-MM.

     Example: "2022-03: Telegnomics-of-Pratchett-Discworld-Series"
   - Share the project with Computational Story Lab (pdodds+compstorylab@uvm.edu).
   - If desired, connect with the Overleaf version using Overleaf's git option.

2. For the paper's introduction, write up a description of the corpus you're studying.

   - Basics: Title, time frame, kind of corpus, number of authors (could be one, could be many).
   - Place the corpus within a larger context of stories/media to which it belongs.
   - Discuss to the extent to which this corpus has been studied by others. Use Google Scholar to find papers. Add bibtex entries from Google Scholar to the .bib file in yoru paper's repository.

   For the paper's data section:

- Describe how you obtained the corpus itself (note online locations explicitly) and whatever data cleaning you and to perform.
- Record scale (number of 1-grams), units and number thereof (e.g., chapters in a book, books in a series, television episodes and seasons), and other basic overall quantifications.

3. Organize your corpora's derived data and contribute it to the Gitlab repository for ousiometry.

   Possible: Filtering a text through the ousiometric lexicon may be enough to make it non-reconstructible.

   Important!: Do not share any text that is under copyright. Hmmm.

4. Figure for paper + caption + discussion:

   Ousiometric time series for your text corpora.

   Plot five main time series that cover the compass of power-danger and then the third dimension of structure.

   - Danger.
   - Dangerous-Power (initially called Energy).
   - Power.
   - Safe-Power (initially called Goodness).
   - Structure

   Notes:

   - One book = five time series, one figure.
   - If you have a small collection of books, plot panels each time series within one overall figure.

     For example, for the 41 novels of the Discworld, a 7×6 format would work.
   - We will adjust depending on how the time series overlap.

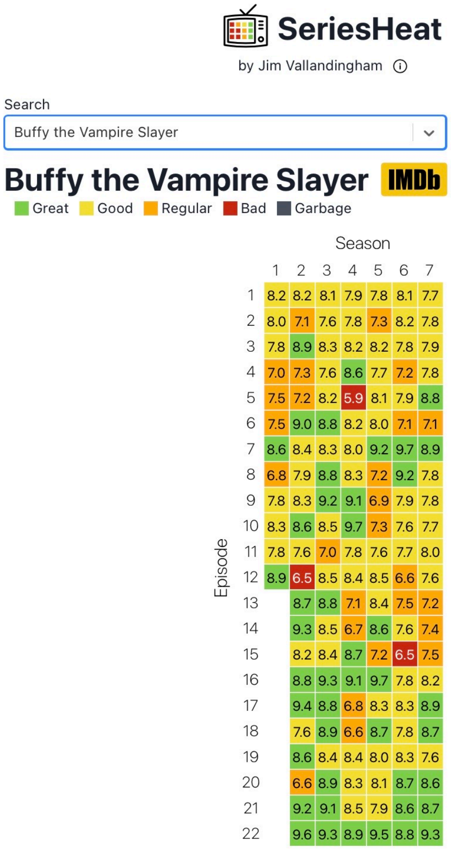
5. Optional figure for paper + caption + discussion:

   For large corpora, repeat the above at a natural intermediate scale, if possible.

   For example, for Buffy the Vampire Slayer, there are 144 episodes over seven seasons.

   We could plot Power, Danger, Dangerous-Power, and Safe-Power scores (and maybe structure) as one long time series.

   An alternate version—which shows the season structure—could be similar to that of the Series Heat visualization, based on IMDB scores:



SeriesHeat

by Jim Vallandingham

Search

Buffy the Vampire Slayer

Buffy the Vampire Slayer IMDb

🟩 Great 🟨 Good 🟧 Regular 🟥 Bad ⬛ Garbage

Season

Episode	1	2	3	4	5	6	7
1	8.2	8.2	8.1	7.9	7.8	8.1	7.7
2	8.0	7.1	7.6	7.8	7.3	8.2	7.8
3	7.8	8.9	8.3	8.2	8.2	7.8	7.9
4	7.0	7.3	7.6	8.6	7.7	7.2	7.8
5	7.5	7.2	8.2	5.9	8.1	7.9	8.8
6	7.5	9.0	8.8	8.2	8.0	7.1	7.1
7	8.6	8.4	8.3	8.0	9.2	9.7	8.9
8	6.8	7.9	8.8	8.3	7.2	9.2	7.8
9	7.8	8.3	9.2	9.1	6.9	7.9	7.8
10	8.3	8.6	8.5	9.7	7.3	7.6	7.7
11	7.8	7.6	7.0	7.8	7.6	7.7	8.0
12	8.9	6.5	8.5	8.4	8.5	6.6	7.6
13		8.7	8.8	7.1	8.4	7.5	7.2
14		9.3	8.5	6.7	8.6	7.6	7.4
15		8.2	8.4	8.7	7.2	6.5	7.5
16		8.8	9.3	9.1	9.7	7.8	8.2
17		9.4	8.8	6.8	8.3	8.3	8.9
18		7.6	8.9	6.6	8.7	7.8	8.7
19		8.6	8.4	8.4	8.0	8.3	7.6
20		6.6	8.9	8.3	8.1	8.7	8.6
21		9.2	9.1	8.5	7.9	8.6	8.7
22		9.6	9.3	8.9	9.5	8.8	9.3

https://vallandingham.me/seriesheat/#/?id=tt0118276

For discussion:

- What variation do you see? Does one dimension vary more than another?

6. Optional analysis and figure for paper + caption + discussion:

   If you have ratings/sales or any kind for your corpus, connect Power, Danger, Dangerous-Power, Safe-Power, STructure scores with those ratings.

   Run some simple analyses to see if there any correlations. Report what you find.

   Two possible figures for Buffy: Power or danger on the horizontal axis and IMDB rating on the vertical.

   It's important to report that there's no correlation.

7. Figure(s) for paper + caption + discussion:

   Time for some lexical calculus.

   Using word shifts, make comparisons of ousiometric scores for parts of your corpora.

   Possibilities:

- Any comparisons that call out for investigation.
- Word shifts comparing each episode or chapter to the overall corpus.
- Comparison of major highs or lows with a reasonable reference text.

- Incorporate figures into your paper.
- For large sets of figures, organize them as supplementary material (e.g., separate PDF booklet).

Some inspiration for books:

https://hedonometer.org/books/v1/?book=Pride%20and%20Prejudice&lens=[3,7]

8. Figure for paper + caption + discussion:

   Basic data: You will need Zipf distributions of word counts for your corpus.

   Put together two kinds: Case sensitive and case insensitive.

   For your main text of interest, plot ousiograms in the following ways:

   (a) Three views: Power-Danger, Power-Structure, Structure-Danger.

   (b) Types (lexicon only—ignore frequency of usage).

   (c) Tokens (full text—incorporate frequency of usage).

   (d) Use the ousiometric lexicon. See below for download.

   (e) Choose a reasonable bin size, perhaps 0.1 or $0.1/k$ where $k = 2$, 3.

   (f) Caption: Report the percentage of the ousiometric lexicon's words that appear in your main text's lexicon.

   (g) Caption and main text: Comment on the basic form of the distributions you find (symmetry, skew, and so on).

   (h) For Matlab, use the script figousiometer9000.m for large ousiograms and figousiometer9400.m for small ones.

9. A good thing to do in general: Test which words in a corpus are missing from the lenses you use.

   (a) Type level: What fraction of unique words (the lexicon) in a corpus are covered by the lens.

   (b) Token level: What fraction of all words (tokens) in a corpus are covered by the lens.

(c) Token level: For the words not in the lens, what does their count-rank distribution look like? Are the missing words problematic?

10. To contemplate:

    Read through the following paper, and start to think about how you might be able to extract character networks for your text:

    "Extraction and Analysis of Fictional Character Networks: A Survey"
    https://arxiv.org/abs/1907.02704

References

[1] P. S. Dodds, T. Alshaabi, M. I. Fudolig, J. W. Zimmerman, J. Lovato, S. Beaulieu, J. R. Minot, M. V. Arnold, A. J. Reagan, and C. M. Danforth. Ousiometrics and Telegnomics: The essence of meaning conforms to a two-dimensional powerful-weak and dangerous-safe framework with diverse corpora presenting a safety bias, 2021. Available online at https://arxiv.org/abs/2110.06847. pdf