The Mixing Law and Experiments in Document Malformation

Jake Ryland Williams and Diana Solano-Oropeza

Departments of Information Science and Physics, Drexel University

Objectives

We investigate a potential explanatory mechanism behind the scaling breaks seen in Zipfian rank-frequency data. Our goals are:

- Replace the defunct explanatory core/non-core vocabulary (CNC) hypothesis with a mechanism that reproduces, explains, and predicts observed behavior.
- Introduce an ansatz solution: the Mixing Law.
- Test the Mixing Law and its feasibility in a Project Gutenberg corpus using natural and regressed parameters.
- Investigate a data malformation hypothesis that would explain model divergence at extreme document-vocabulary sizes by creating a synthetic data set from the corpus.

Defining and Interpreting the ML

In the abstract, we defined $P(r_w \mid \langle N\rangle)$ as the ML and proposed it interact with Zipf's law as a factor:

$$\overline{f}\,(r_w \mid \mathcal{D}) \propto \underbrace{r_w^{-\theta}}_{\text{Zipf's Law}} \cdot \underbrace{\left[1-\left(1+\frac{\langle r\rangle}{r_w}\right)^{-\frac{\langle N\rangle}{\langle r\rangle}}\right]}_{\text{Mixing Law}}$$

In Sec. 1, we outline the mathematics that asserts the need for the ML in determining mixture frequencies. But that math just asserts $P(r_w \mid \langle N\rangle)$'s existence and not what its form should be—in this report, the above is left as an *ansatz*.

Critical insight into the ML can be obtained by interpreting its—very natural—parameters. Understanding $\langle r\rangle = \frac{N}{H_N}$ as the harmonic average of mixture ranks, it becomes clear $\frac{\langle r\rangle}{r_w}$ is an odds ratio for w against the mixture vocabulary's centrality. Going deeper, $\frac{\langle N\rangle}{\langle r\rangle} = H_N\frac{\langle N\rangle}{N} \approx (\gamma + \log N)\frac{\langle N\rangle}{N}$ approximates the type-entropy from $\mathcal{D}$ covered by a mid-sized document. Currently, we're deriving a generative model that explains the ML and these phenomena through a semantic 'momentum' mechanism that operates a latent vocabulary, which we call the *Theory of Harmonic Resonance.*

Estimating Mixture Frequencies

Define text mixing mathematically for a set of k documents $\mathcal{D} = \{d_i\}_{i=1}^k$ by assuming each upholds ZL as $f(r_w \mid d_i) \propto r_{w,i}^{-\theta_i}$ over a vocabulary of $N_i = |d_i|$ words via a scaling exponent θ_i. The goal is to simplify the *mixture frequencies*: $\overline{f}(r_w \mid \mathcal{D}) = \sum_{i=1}^k f(r_w \mid d_i)$. This is challenged by variation in each word's ranks *across* the documents, i.e., the distribution of $\{r_{w,i}\}_{i=1}^k$ (local ranks). Now, ZL constrains the variation in exponents $\{\theta_i\}_{i=1}^k$ so that there exists θ with $\theta_i \approx \theta \approx 1$ across $\mathcal{D}$. Next, convert the scaling into a frequency model for each $w \in d_i$ via $\hat{f}(r_w \mid d_i) = \left(\frac{r_{w,i}}{N_i}\right)^{-\theta}$. To naïvely continue, we assume w appears in all k documents, so:

$$\overline{f}(r_w \mid \mathcal{D}) \approx \sum_{i=1}^k \hat{f}(r_w \mid d_i) = \sum_{i=1}^k \left(\frac{r_{w,i}}{N_i}\right)^{-\theta}$$

We then define θ-harmonic mixture frequencies:

$$\overline{\varpi}(r_w \mid \mathcal{D}) = \frac{\overline{f}(r_w \mid \mathcal{D})}{\sum_{i=1}^k N_i^\theta} \approx \sum_{i=1}^k r_{w,i}^{-\theta}\frac{N_i^\theta}{\sum_{i=1}^k N_i^\theta}.$$

Naturally appearing, the document-mass probabilities: $P_i = N_i^\theta / \sum_{i=1}^k N_i^\theta$ allow us to characterize $\overline{\varpi}$ via a P_i-weighted power mean of w's ranks:

$$\langle \vec{r}_w\rangle_{-\theta,\vec{P}} = \left(\sum_{i=1}^k r_{w,i}^{-\theta}P_i\right)^{-1/\theta} = \overline{\varpi}(r_w \mid \mathcal{D})^{-1/\theta}$$

Looking at the P_i denominators, we similarly define:

$$\langle \vec{N}\rangle_\theta = \left(\frac{1}{k}\sum_{i=1}^k N_i^\theta\right)^{1/\theta} = \left(\frac{1}{k}\frac{\overline{f}(r_w \mid \mathcal{D})}{\overline{\varpi}(r_w \mid \mathcal{D})}\right)^{1/\theta}$$

which is solved back for $\overline{f}$ in terms of averages:

$$\overline{f}(r_w \mid \mathcal{D}) \approx k\left(\frac{\langle \vec{r}_w\rangle_{-\theta,\vec{P}}}{\langle \vec{N}\rangle_\theta}\right)^{-\theta}$$

To generalize for words that *don't* appear in all documents, we scale by a probability, $P(r_w \mid \langle N\rangle)$ representing the chance that a document of average size $\langle N\rangle$ from $\mathcal{D}$ contains a word of mixture-rank r_w:

$$\overline{f}(r_w \mid \mathcal{D}) \approx k \cdot \left(\frac{\langle \vec{r}_w\rangle_{-\theta,\vec{P}}}{\langle \vec{N}\rangle_\theta}\right)^{-\theta} \cdot P(r_w \mid \langle N\rangle)$$

Supposing w's mixture rank and expected local rank scale: $\langle \vec{r}_w\rangle_{-\theta,\vec{P}} \propto r_w$, we recover the ML's role:

$$\overline{f}(r_w \mid \mathcal{D}) \propto r_w^{-\theta} \cdot P(r_w \mid \langle N\rangle). \quad \blacksquare$$

Experiments and Results

We test the Mixing Law by setting up an experiment, where we use a corpus (collection of texts) composed of Project Gutenberg texts, divided into decile bins ranging from the smallest to the largest by numbers of tokens. In general for each bin we perform 35 experiment-instances. In each, we sample some number of books from the bin, mix their frequencies, and store mixture metadata to compute:

$$\tilde{N} = \text{vocabulary size of entire mixture}$$
$$\langle \tilde{N}\rangle = \theta\text{-power mean of doc. vocabulary sizes}$$

as 'natural' parameters, i.e., determined by the metadata. We compares these with other, regressed parameters, $\hat{N}$ and $\langle \hat{N}\rangle$, treating N and $\langle N\rangle$ as learnable quantities (see Python's scipy.optimize.minimize Nelder-Mead implementation). We then store the ratio of the two models' perplexities and observe how they relate. We compare results from this procedure with the same from a synthetic corpus produced by smashing and sewing together various books in the middle deciles (4–7) to *simulate* extreme-decile books that exist as the result of malformation.

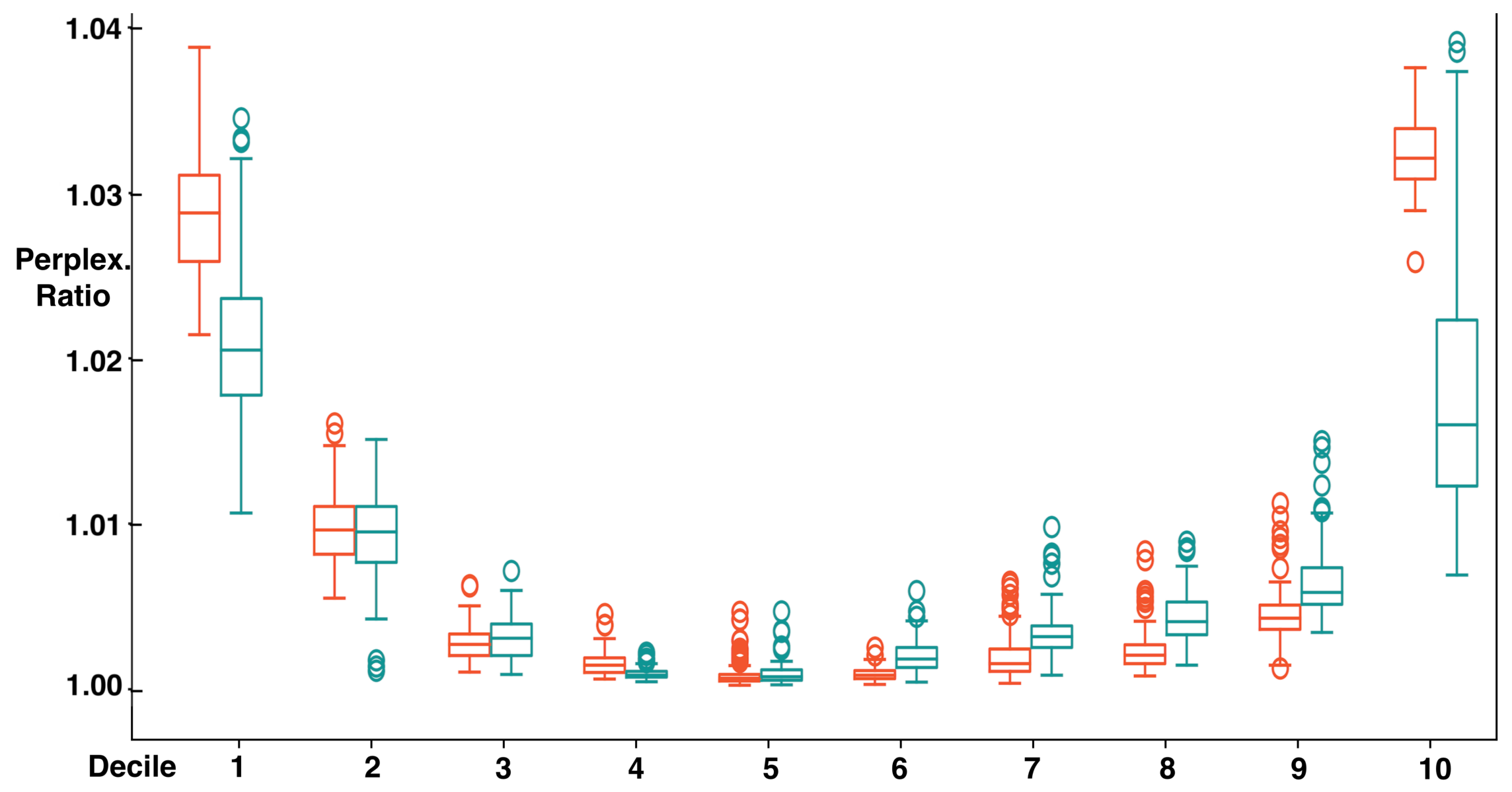

Figure 1: Model perplexity ratios for $\langle N\rangle$ & N measured by corpus and regression for samples of 100 documents (taken by decile). Large/small deciles (1–3 & 7–10) for synthetic documents (orange) are fragmentations/agglomerations of real documents (blue), taken from deciles 4–7.

Acknowledgements

This document is based on work supported by the National Science Foundation grant no. #1850014.

Interpreting ML Experiments

Comparing perplexity ratios across the synthetic and original data sets, we see that the data malformation hypothesis produces model divergence in anticipated directions, i.e., supporting the hypothesis that ML divergence is largely impacted by poor document 'resolution'. However, the amount of divergence we see goes well beyond that for the real data, indicating that other effects are likely going on, too, i.e., that govern generation of *truly* small and large 'documents,' like poetry and encyclopedias (respectively).

References

[1] G. K. Zipf.
The Psycho-Biology of Language.
Houghton-Mifflin, 1935.

[2] H. A. Simon.
On a class of skew distribution functions.
Biometrika, 42:425–440, 1955.

[3] Marin Gerlach and Eduardo G. Altmann.
Stochastic model for the vocabulary growth in natural languages.
Phys. Rev. X, 3:021006, 2013.

[4] J. R. Williams, J. P. Bagrow, C. M. Danforth, and P. S. Dodds.
Text mixing shapes the anatomy of rank-frequency distributions.
Physical Review E, 91:052811, 2015.

[5] Karen Spärck Jones.
A statistical interpretation of term specificity and its application in retrieval.
Journal of Documentation, 28:11–21, 1972.

[6] S. Robertson.
Understanding inverse document frequency: on theoretical arguments for IDF.
Journal of Documentation, 60:503–520, 2004.

Contact Information

- Web: http://jakerylandwilliams.github.io/
- Email: {jw3477,ds3367}@drexel.edu