Selection models of language production support informed text partitioning: an intuitive and practical, bag-of-phrases framework for text analysis.

Jake Ryland Williams,1 James P. Bagrow,2 Andrew J. Reagan,2 Sharon E. Alajajian,1 Christopher M. Danforth,2 and Peter Sheridan Dodds2

1School of Information, University of California, Berkeley 102 South Hall #4600 Berkeley, CA 94720-4600.
2Department of Mathematics & Statistics, Vermont Complex Systems Center, Computational Story Lab, & the Vermont Advanced Computing Core, The University of Vermont, Burlington, VT 05401.

(Dated: February 1, 2016)

The task of text segmentation, or ‘chunking,’ may occur at many levels in text analysis, depending on whether it is most beneficial to break it down by paragraphs of a book, sentences of a paragraph, etc. Here, we focus on a fine-grained segmentation task, which we refer to as text partitioning, where we apply methodologies to segment sentences or clauses into phrases, or lexical constructions of one or more words. In the past, we have explored (uniform) stochastic text partitioning—a process on the gaps between words whereby each space assumes one from a binary state of fixed (word binding) or broken (word separating) by some probability. In that work, we narrowly explored perhaps the most naïve version of this process: random, or, uniform stochastic partitioning, where all word-word gaps are prescribed a uniformly-set breakage probability, \( q \). Under this framework, the breakage probability is a tunable parameter, and was set to be pure-uniform: \( q = 1/2 \). In this work, we explore phrase frequency distributions under variation of the parameter \( q \), and define non-uniform, or informed stochastic partitions, where \( q \) is a function of surrounding information. Using a crude but effective function for \( q \), we go on to apply informed partitions to over 20,000 English texts from the Project Gutenberg eBooks database. In these analyses, we connect selection models to generate a notion of fit goodness for the ‘bag-of-terms’ (words or phrases) representations of texts, and find informed (phrase) partitions to be an improvement over the \( q = 1 \) (word) and \( q = 1/2 \) (phrase) partitions in most cases. This, together with the scalability of the methods proposed, suggests that the bag-of-phrases model should more often than not be implemented in place of the bag-of-words model, setting the stage for a paradigm shift in feature selection, which lies at the foundation of text analysis methodology.

PACS numbers: 89.65.-s, 89.75.-k, 89.70.-a

In our previous work on text partitioning [1], and again in our study on the effects of text mixing [2], we have observed the relevance of the selection model proposed long ago by Simon [3]. We will use this work again, together with Zipf’s law [4, 5], as the basis for a model of frequency data of text.

The defining aspects of Simon’s model for language generation hold that as the words of a text appear, they do so independently, and in proportion to their historic frequencies of occurrence. The first assumption (of word-word independence) is simultaneously the basis for the ‘bag-of-words’ model that is pervasive in the computational text analysis community. It is clear that this assumption fails when one considers the clear dependence of words that are bound, like “New,” “York,” and “City.” As we have seen in [2], dependence can also be caused by mixing, where the rate of a word’s occurrence varies in a (global) mixed text as one transitions from document to document. These two types of dependence then indicate that for the Simon model to hold, and consequently, for the ‘bag-of-words’ model to hold, texts should be analyzed on the homogeneous (unmixed) scales of production, and the words of texts should be bound as phrases, according to their local dependence. The first criterion can be relatively easy to satisfy, and in our analyses will come for free by the curation of the Project Gutenberg eBooks [6] database. The second, however, poses a much greater challenge, as the meanings that bind words into idioms and other irreducible forms are only known a priori to the writers who generate text for their meanings—all else must interpret.

In our work on stochastic text partitioning [1] we proposed a simple mechanism for the fine-grained chunking of text into phrases. Under this framework, sentences were broken down stochastically, splitting on whitespace with a uniform bond-breakage probability, \( q \). While this (uniform) framework is clearly not the ideal mechanism to isolate phrases for, say, dictionary lookups, we did observe interesting improvements over word distributions (which too arise from stochastic text partitioning, setting \( q = 1 \)) in relation to Simon’s model.

Since we cannot ask all writers to resolve their texts with their intended partitions, the next best step may then be to explore informed stochastic partitions, defining \( q \)’s variation either empirically or through some model. While a strictly empirical approach might be best, it would likely require intensive data collection and surveying, and for this reason we will move forward in this
development with a hybrid approach. Here, we define \( q \) to be a function of two variables—the immediately adjacent (left and right) words framing a given whitespace (gap) in text. Considering the example clause

\[
\begin{array}{|c|c|c|}
\hline
q_0 & \text{Hot} & q_1 \\
q_2 & \text{dog} & q_3 \\
\hline
\end{array}
\]

we will define \( q_3 \) to be a function of the ordered pair, \((\text{Hot, dog})\), and the clause-edge partition probabilities \((q_0, \text{and \ } q_3)\) to be constant and equal to 1. For simplicity, we will assume in general that \( q_i \) are independent of one another, and approximate them with quasi-empirical breakage rates.

Since we possess no wealth of hand-partitioned text for training, we instead turn to bond-breakage information held latently inside of the Wiktionary \([7]\) by its trove of larger-than-word entries. We artificially construct a partitioned text by enumerating all \( N^2 \) ordered pairs of phrases listed under the dictionary of \( N \). Each pair of phrases is then considered a sentence, with one broken bond (between the two phrases), and potentially many unbroken bonds (internal to the phrases drawn). For an ordered pair \((w_l, w_r)\) of left and right-framing words, we then define \( q(w_l, w_r) \) to be the probability that \( w_l \) and \( w_r \) came from separate phrases, given their concurrent appearance in a sentence in our artificial text.

While our construction for the partition function \( q \) is highly artificial, it has substantial benefits in not requiring surveys, or having bias toward content types. However, if one wanted to implement these methods for analysis of a particular medium, e.g., Twitter, it might be best to survey that particular medium for partition probabilities, as the lay of the land for Twitter may look different than for Moby Dick. We also note here that \( q \)-defining partitions may be constructed by entirely different means and for different purposes. For example, one could bias \( q \) by pairwise word similarity, e.g., by edit distances \([5]\), and make repetitions like “la la” or “ha ha”, etc. more or less often partitioned.

Armed with the dictionary-informed partition function defined above, we now turn to our measure of partition quality. This descends directly from what we refer to as the Zipf/Simon, single-parameter power-law model for rank-frequency distributions. Zipf’s law succinctly formulates a relationship between the frequencies of occurrence of words, and their ranks by frequency (descending):

\[
f = C \cdot r^{-\theta},
\]

and results when languages are generated by Simon’s model.

The main caveat for the compatibility of Zipf’s empirical law with Simon’s theoretical model arises from the fact that Simon’s model is only able to generate single-scaling rank-frequency distributions with scaling parameter \( \theta \) in \([0, 1]\). While many texts exhibit multiple scalings and values of \( \theta \) larger than 1, we have seen that these anomalies are often the result word dependence. Consequently, we view the degree to which a particular Zipf/Simon model aligns with empirical rank-frequency data as a measure of the goodness of fit for the ‘bag-of-terms’ representation of a text (tokenization) that generated the rank-frequency data (regardless of how the terms are defined, i.e., words or phrases).

The Zipf/Simon fit for a text of \( N \) unique and \( M \) total words is defined as follows. Assuming the text was generated precisely according to Simon’s model, the constant word introduction rate is known by the text-wide average: \( \alpha = N/M \), from which it can be shown \([3]\) that the scaling parameter emerges to be \( \theta_{\text{mod}} = 1 - \alpha \). The exact form of the model’s fit is then obtained by computing the constant of proportionality: \( C_{\text{mod}} = N^\theta_{\text{mod}} \), whereupon we may take the coefficient of determination, or \( R^2 \), as a
measure of goodness of fit for the Zipf/Simon model.

Since the connection of Simon’s model to the parameter $\theta$ occurs naturally in the complimentary cumulative distribution function (CCDF) of term frequencies, we measure and regress all quantities along CCDFs, while we present all results in the intuitive and familiar rank-frequency (Zipf) representations.

In Fig. 1, we plot an example informed partition for Herman Melville’s well-known book “Moby Dick” (black line, main axes), and behind it, the spectrum of uniform stochastic partitions (y-axis). The discriminating line (red, dashed, $R_{\inf}^2 = R_{q=1}^2$) helps divide the collection into texts that are word- and phrase-based.

FIG. 2: We plot the goodness of fit ($R^2$) of the Zipf/Simon model, applied to the texts of the English Project Gutenberg eBooks database, where the dictionary-informed one-off partitions (x-axis) are plotted against the $q = 1$ uniform stochastic partitions (y-axis). The discriminating line (red, dashed, $R_{\inf}^2 = R_{q=1}^2$) helps divide the collection into texts that are word-and phrase-based.

To include dictionaries, spelling books, and books of extremely small size (often as placeholders for other media), indicating that low $R^2$ values appropriately identify texts unfit for analyses.

We have shown that informed stochastic text partitioning and fit-goodness tests are capable of improving the basic methodologies for feature selection in text analysis,
Here we present the percent of books in the Project Gutenberg English eBooks database that are phrase-based (solid green line, filled squares), word-based (solid blue line, filled circles), and poorly fit (red line, stars), when a cutoff in $R^2$ is applied (right). A clear dropoff in $R^2$ occurs when $R^2 \approx 0.7$, which is noted by the vertical, black dotted line. We also present curves that indicate the percents of books remaining above the cutoff when blanked usage of either words (blue dashed line, open circles) or informed phrases (green dashed line, open squares) are applied.

allowing for better adherence to assumptions (specifically, term-term independence) that are present in a vast collection of algorithms currently in practice throughout industry and academia. Additionally, phrase-based text analysis improves the independent interpretability of features—in the soft sense, at the level of user experience. With phrase-based text analysis, end-point users (e.g., policy makers or product users interpreting a list of phrase-feature topics from a topic model readout) who may not understand the machinery of an algorithm will be better able to interpret results, as phrases provide critical context for interpretation.

To make these tools both explorable and available for the community, we have with this letter released the first version of a Python package for text partitioning (for detailed information, see https://github.com/jakerylandwilliams/partitioner) as well as an explorable online appendix (http://jakerylandwilliams.github.io/partitioner/) that also aims to gather empirical partition data from the public, using the model for scientific research proposed by Volunteer Science [9] (https://volunteerscience.com/) that will enable us to improve informed partitions for future work.

---

* Electronic address: jakerylandwilliams@berkeley.edu
† Electronic address: james.bagrow@uvm.edu
‡ Electronic address: Andrew.Reagan@uvm.edu
§ Electronic address: sharon@ischool.berkeley.edu
¶ Electronic address: chris.danforth@uvm.edu
** Electronic address: peter.dodds@uvm.edu