Stable Random Projection: Standardized universal dimensionality reduction for library-scale data

Benjamin Schmidt
bmschmidt@gmail.com
Northeastern University, United States of America

Summary

This paper describes a new method for dimensionality reduction, "stable random projection," (hereafter "SRP") distinctly suited for large textual corpora like those used in the digital humanities. The method is computationally efficient and easily parallelizable; scales to the largest digital libraries; and creates a standard dimensionality reduction space for all texts so that corpora and models can be easily exchanged. The resulting space makes a wide variety of applications suitable to bag-of-words data, such as nearest neighbor searches, classification, and semantic querying possible with data sets an order of magnitude smaller in size than traditional feature counts.

SRP is a minimal, universal dimensionality reduction with two distinctive features:

1. It makes *no distinction between in- and out-of-domain vocabularies.* In particular, unlike standard dimensionality reduction it creates a single space that can hold documents of *any language*.
2. It is *trivially parallelizable*, both on a local machine and through web-based architectures because it relies only on code that can be easily transferred across servers, rather than requiring large matrices or model parameters.

These two features allow dimensionality reduction to be conceived of as a piece of infrastructure for digital humanities work, rather than just an ad-hoc convention used in a particular project. This method is particularly useful for provisioners and users of text data on extremely large and/or multilingual corpora. This creates a number of new applications for dimensionality reduction, both in scale and in type. SRP features could usefully be distributed by libraries as a (much smaller and easier to work with) supplement to feature counts. After a description of the method, some novel uses for dimensionality reduction on such libraries are shown using a sharable dataset of approximately 4,500,000 books projected into SRP-space from the Hathi Trust.

Description of the method

The goal of SRP is to reduce of text of uncertain length to a much smaller fixed- length vector to which the many tools of textual analysis, machine learning, and linear algebra can be applied. The core technique used here for dimensionality reduction is *random projection.* Random matrix theory has emerged in the past few decades as an useful alternative to more computationally complex forms of dimensionality reduction.(Halko, Martinsson, and Tropp 2009) I make use here of the observation that it is possible to project into a space where points as determined purely by sampling randomly from the set [-1,1].(Achlioptas 2003) A true random number generator is not suitable for reproduction. The other core element of SRP, therefore, is a quasi-random projection for every individual word created using cryptographic hashes (specifically, SHA-1).

This allows the method to be defined algorithmically, making it easy to apply to any text. I have written short code libraries to implement the transformation in the three most important language for DH tool development: Python, R, and Javascript. These include a few necessary additional conventions such as minimal tokenization rules, a method for expanding beyond the 160 dimensions provided by SHA, and the byte-encoding of the Unicode character sets.

Comparison to existing methods

The gold standard for dimensionality reduction are techniques that make use of co-occurrences in the term-document matrix such as latent semantic indexing and independent components analysis. More recent techniques such as semantic hashing can be even faster and more efficient at optimally organizing documents in various types of vector spaces designed especially for particular documents.(Salakhutdinov and Hinton 2009) Another strategy finding recent use in the digital humanities is using an LDA topic model as dimensionality reduction, which produces neatly interpretable dimensions for analysis (Schöch, 2016; Fitzgerald,2016). In both the digital humanities and computer science, scholars frequently use "top-N"

words as a good enough approximation of the textual footprint, limiting the dimensions to a few hundred of the most common words in the corpus, producing what Maciej Eder has called "endless discussions of how many frequent words or n-grams should be taken into account" for stylometry.(Underwood 2014, Eder (2015))

These methods suffer two problems that make them problematic as a *general-use* feature reduction. First, the better ones are computationally complex, and quite difficult to perform on a very large corpus. Second, it is difficult or impossible to project *out-of-domain* documents into the space from a standard projection if they contain vocabulary different than the training corpus. This out-of-domain problem presents a particularly great problem for multilingual corpora, because texts that are missing or in sparsely-represented languages will behave erratically in the new environment.

Some other work in the digital humanities and computer science has used hashes, random projection, and other similar methods as an ad-hoc rather than infrastructural technique. SRP can be thought of as a particular species of *locality-sensitive hashing*, another version of which has been used by Douglas Duhaime to identify reuse in poetic texts based on three-letter phrases.(Duhaime 2016). Also related is the "hashing trick" in computer science(Weinberger et al. 2009), which is better than SRP in many ways for the short documents computer scientists frequently study, but takes significantly more memory to store for book-length documents (an edge case in the computer science literature, but among the most important for humanists).

Applications

This reduced space can be put to many of the same uses as a standard bag-of- words model in considerably less space and with the potential for building web facing tools. Among those to be described are:

1. **Duplicate detection**. SRP is quite accurate at identifying duplicate books in a computationally tractable space using cosine similarity, both inside a corpus and across disparate corpora.
2. **Similarity Search**. A prototype web page allows any user to paste in any text; it will hashed on the client side into the standard space, and a server can return in a few seconds the most similar documents. The top entries can function for duplicate detection; the lower ones presenting interesting opportunities for exploratory analysis. A search for *Huckleberry Finn*, for example, finds a large number of other American adventure novels about boys in the American west.
3. **Classification**
   - SRP features perform approximately as well as top-n words (~77%) on a pre-existing task described by Ted Underwood, separating high- from low-prestige poetry.(Underwood 2015)
   - A single hidden layer neural network trained with 640-dimensional SRP features can accurately classify a held-out sample of books into one of 225 Library of Congress Classification subclasses (for example, whether a work is PR: British Literature or PS: American Literature) with ~78% accuracy based on about 1 million training examples. A single classifier works in multiple languages simultaneously; its determinations on arbitrary pasted text are accessible for inspection through a web site.
   - A different single hidden layer neural network trained with SRP features and a novel encoding scheme for years using Google's TensorFlow framework can accurately predict the years for withheld books with a median error of four years from the true publication date.

SRP as Access

SRP fits in the DH2017's theme of "Access" in two ways.

First, it makes many forms of text analysis on huge digital libraries far more feasible for scholars without access to high performance computing resources. On large corpora, data storage and dimensionality reduction can be more resource- intensive than the actual analysis. The dimensionality-reduced dataset for the full Hathi Trust corpus can fit into 10 GB, easily storable on most computers; subsets are suitable for use in classroom or workshop settings.

Second, the ease with which it works with distributed web architectures, and its language agnosticism, can create new routes into neglected portions of large archives, particularly those with insufficient metadata.

Bibliography

Achlioptas, D. (2003). "Database-Friendly Random Projections: Johnson- Lindenstrauss with Binary Coins." *Journal*

of Computer and System Sciences, Special issue on PODS 2001, 66 (4): 671–87. doi:10.1016/S0022-0000(03)00025-4.

Duhaime, D. (2016). "Plagiary Poets. Plagiary Poets." http://plagiarypoets. io/.

Eder, M. (2015). "Visualization in Stylometry: Cluster Analysis Using Networks." *Digital Scholarship in the Humanities,* November, fqv061. doi:10.1093/llc/fqv061.

Fitzgerald, J. D. (2016) "What Made the Front Page in the 19th Century?: Computationally Classifying Genre in 'Viral Texts". July 13 2016 http://jonathandfitzgerald.com/blog/2016/07/13/keystone-paper.html

Halko, N., Martinsson,P.-G., and Tropp, J. A. (2009). "Finding Structure with Randomness: Probabilistic Algorithms for Constructing Ap- proximate Matrix Decompositions." arXiv:0909.4061 [Math], August. http://arxiv.org/abs/0909.4061.

Salakhutdinov, R., and Hinton, G. (2009). "Semantic Hashing*." International Journal of Approximate Reasoning,* Special section on graphical models and information retrieval, 50 (7): 969–78. doi:10.1016/j.ijar.2008.11.006.

Schöch, C. (2016, pre-publication) "Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama". *Digital Humanities Quarterly.*

Underwood, T. (2014). "Understanding Genre in a Collection of a Million Volumes, Interim Report." http://figshare.com/articles/Understanding_Genre_in_a_*Collection_of_a_Million_Volumes_Interim_Report*/1281251.

Underwood, T.. (2015). "The Literary Uses of High-Dimensional Space." *Big Data & Society* 2 (2): 2053951715602494. doi:10.1177/2053951715602494.

Weinberger, K., Dasgupta, A., Langford, J., Smola, A., and Attenberg, J. (2009). "Feature Hashing for Large Scale Multitask Learning." In *Proceedings of the 26th Annual International Conference on Machine Learning*, 1113–20. ICML '09. New York, NY, USA: ACM. doi:10.1145/1553374.1553516.