Toward Reproducibility in DH Experiments: A Case Study in Search of Edgar Allan Poe's First Published Work

Mark D. LeBlanc
mleblanc@wheatoncollege.edu
Wheaton College, United States of America

Summary

Reproducing experimental results is a hallmark of empirical investigation and serves both to verify and inspire. This paper is a call for more systematic documentation of computational stylistic experiments. Publishing only summaries of the methods and results of empirical work is an artifact of traditional print media. To facilitate experimental reproducibility and to help the growing community who wish to learn how to apply computational methods and subsequently teach the next generation of scholars, the publication of results must include (i) access to the digitized texts, (ii) a clear workflow and most essentially (iii) the source code that led to each and all of the experimental results. By way of example, we present the steps and process in a GitHub repository for computationally probing the unknown and contested authorship of an 1831 short story entitled "A Dream" as we seek evidence if this work is similar to other attributed works by Edgar Allan Poe. The entire framework is intended as a pedagogical jumpstart for others, especially those new to computational stylometry. If Poe did write the story, it would be his first published work.

Introduction

As the Digital Humanities gains access to a wide array of digitized corpora and matures to a discipline that creatively defines new methods for computationally close and distant readings, a growing gap has emerged between those who apply sophisticated programming, e.g., Stylo In R (Eder *et al.*, 2016) and those who are new to the game and need an introduction to the field. Typical of the community spirit in DH, significant efforts are underway to bridge this gap, including web-based tools for entry-level exploration including Voyant Tools (Sinclair and Rockwell, 2016) and Lexos (Kleinman *et al.*, 2016) and domain-specific introductions to programming, including Jockers' text (2014) and the Programming Historian (Crymble *et al.*, 2016). This paper attempts to narrow the gap by encouraging both sides to document their experimental methods more fully to embrace previous calls for the replication of experimental methods (Rudman, 2012 *et al.*) and thereby teach effective practices by "leaving a trail" of experimental methods that enable others to execute and extend.

A Good Mystery: Towards Reproducibility

A GitHub repository or "repo" offers a workflow that explores whether an 1831 story published under the attribution of only 'P' might have been written by Edgar Allan Poe. If so, it would be Poe's first published work. In addition to sharing a set of analytical methods applied in this experiment, the broader methodological-pedagogical goals are two-fold: (i) the dissemination of data and code should be championed as a cornerstone of DH research, thereby facilitating the replication of results and (ii) to share a workflow so that others may apply similar analyses to their texts of interest.

The workflow is stored as a set of numbered folders containing the texts *and* scripts (code) needed to complete each step. The workflow includes: collecting texts, the preprocessing, tokenization, and culling decisions made, unsupervised cluster analyses (k-means, hierarchical-agglomerative, bootstrap consensus tree), and supervised classification methods using Stylo in R's Delta, SVM, and NSC models. Each step represents scaffolding for a "teachable moment" with materials provided so faculty can more easily use them with students.

Scrubbing, Tokenization, Cutting, and Culling

Lexos, a web-based, open-source workflow of tools (Kleinman, *et al.*, 2016) was used to upload texts and "scrub" them by applying the following options: (i) convert words to lowercase, (ii) all punctuation was removed, (iii) however, a single word-internal hyphen and word-internal apostrophes were kept, and (iv) all digits were removed. Each individual word is considered as its own token. Larger stories were segmented ("cut") into pieces. We experimented with various culling options, e.g., keeping only the most frequent words that appear in each text at least once.

Cluster Analysis

As a set of initial probes, we compared the contested story "A Dream" to (i) other stories attributed to Poe and (ii) mixed in with stories by other contemporaries. In the repo, we share four variations using cluster analysis:

1. K-means clustering on only Poe's stories (using Lexos)
2. Hierarchical agglomerative clustering on only Poe's stories (uses a Python sklearn module and a script to convert the cluster to ETE and Newick formats)
3. K-means clustering when all stories by each author are concatenated together (Lexos)
4. Bootstrap Consensus Tree (using Stylo in R).

The result from the Bootstrap Consensus Tree is shown in Figure 1. Of interest is that each author's stories cluster consistently together (with the exception that Bird's initial section of "Sheppard Lee" and his "Calavar" are found in different clades, at six and eight o'clock). "A Dream" clusters with the smaller Poe texts. As you'll see, we couldn't resist tossing in the four stories sometimes attributed to Edgar's brother Henry ("Monte Video", "A Fragment", "The Pirate", and "Recollections"). These four stories are found within the cluster of Poe's known works (*c.f.* Collins, 2013).

A series of cluster analyses often serves well as a preliminary exploration, especially for scholars who are new to this game. Some of the file sizes are very small (*e.g.*, one-half of the Poe stories in this corpus have fewer than 2000 words) and when strict culling is enforced (top-N words that appear at least once in each segment), the available set of words is reduced to only 38 when dealing with "A Dream" and the other eighteen Poe stories. That noted, these exploratory investigations shed some light on why some scholars consider that Poe's "first published tale may have been 'A Dream'" (Silverman, 1991, p87).

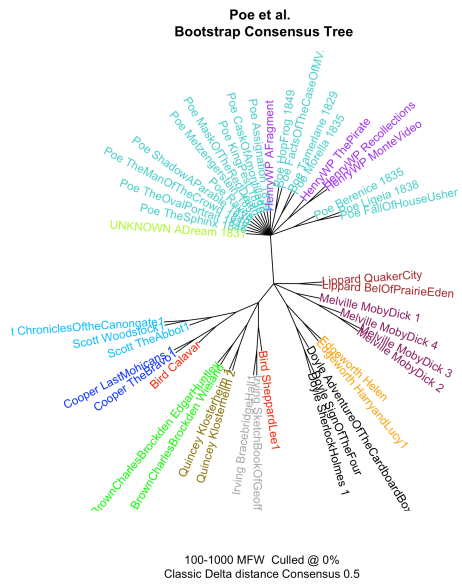

Figure 1. Using Stylo in R Bootstrap Consensus Tree (BCT) showing "A Dream" consistently clustering with other Poe stories. The BCT aggregates results over multiple cluster analyses and shows those texts that satisfy a consensus number of the individual trials. Using 12 different authors and at least two texts by each author for a total of 46 stories, Stylo formed clusters of the texts for the following frequency bands when using the most-frequent words: 100 to 1000 MFW.

Classification

Three classification models differentiated authorial writing style as implemented in Stylo in R. We scripted in R alongside Stylo to test "A Dream" over N-trials (N=10, 100) using a random selection of files for training sets in each trial. At least one text from each author is also included in the test set for each trial. A follow-up Python script parses the collected results to build confusion matrices for each author to provide metrics on how well the models predict each author's works. The most-frequently occurring, top-40 words (MFW, 1-grams) that appear in all the texts at least once were used.

		Confusion Matrix values for all Poe Stories			
Model	**Attributions of "A Dream" to Poe**	**True+**	**True-**	**False+**	**False-**
Delta	9	**13**	**200**	**0**	**7**
NSC	10	**16**	**170**	**30**	**4**
SVM	7	**14**	**198**	**2**	**6**

Table 1: Attributions of the contested story "A Dream" over ten (10) trials with "A Dream" and another randomly selected Poe story in the test set in every trial. Confusion matrix values for results of testing Poe texts over all trials provide overall measures of model effectiveness. In the three cases

where “A Dream” was attributed to a different author, Poe was ranked second.

Summary

We offer a start to an exploration to collect evidence as to whether Poe may have written the 1831 story “A Dream” (*c.f.,* Schöberlein (2016) who used the most frequent character 3-grams and attributed the story to Poe using Delta, but not so when using NSC nor SVM models). Evidence and methods aside, a GitHub repo provides a framework to share experimental workflows in a spirit similar to Jupyter notebooks, as well as one that facilitates both reproducible results and opportunities for subsequent contributions.

Notes

Forming an appropriate corpus is hard: thanks to Sam Coale, Ryan Cordell, Cary Gouldin, David Hoover, Shirrel Rhoades, and Ted Underwood. Four undergraduates: Weiqi Feng, Alec Horwitz, Jingxian Liu, and Khaled Sharafaddin worked with us on this problem. Thanks to Maciej Eder for his help with Stylo in R.

Bibliography

Crymble, A., Gibbs, F., Hegel, A., McDaniel, C., Milligan, I., Taparata, E., Visconti, A., and Wieringa, J., eds. (2016). *The Programming Historian*. 2nd ed.. Web: http://programminghistorian.org/.

Eder, M., Kestemont, M. and Rybicki, J. (2016). Stylometry with R: A package for computational text analysis. *R Journal*, 16(1): 107-121.

GitHub repository: A Good Mystery. . https://github.com/WheatonCS/aGoodMystery

Jockers, M. (2014). *Text Analysis with R for Students of Literature.* Springer, New York.

Kleinman, S., LeBlanc, M.D., Drout, M., and Zhang, C. (2016). Lexos v3.0. Web: http://lexos.wheatoncollege.edu.

Rudman, J. (2012). The State of Non-Traditional Authorship Attribution Studies -- 2012: Some Problems and Solutions. *English Studies*, v93(3), 259-274.

Schöberlein, S. (2016). Poe or Not Poe? A Stylometric Analysis of Edgar Allan Poe’s Disputed Writings. *Digital Scholarship in the Humanities*, July 24, 2016.

Silverman, K. (1991). *Edgar A. Poe: Mournful and Never-Ending Remembrance*. HarperCollins, New York.

Sinclair, S. and Rockwell, G., (2016). *Voyant Tools*. Web: http://voyant-tools.org/.