Part-of-Speech Profiling and Stylistic Textual Assessment Using Probabilistic Matrix Factorization

Jonathan Armoza
December 19, 2016

Introduction

When Emily Dickinson died in 1886, she left behind a body of unpublished writing that presents an appropriate case study for probabilistic modeling techniques. Of the roughly 1,800 individual poems, over 800 were handsewn into forty manuscript books called "fascicles." The page orderings of those books were lost as a result of a family feud over ownership rights and differing ideas about how to organize and publish such a large collection of poems. However, those page orderings were recovered by the forensic analysis of R.W. Franklin in *The Manuscript Books of Emily Dickinson* (Dickinson 1981). Still, the size of the collection has made comparative analysis of the poems a difficult task. The fascicle orderings of Dickinson's poems have become for some scholars a structuring basis for uncovering thematics or narratives in her writings.

Computational perspectives on the linguistic relationships within and among human-authored texts are the subject of an increasing amount of interest in humanities-based fields of study. Part of the inquiry into those perspectives has been gauging if and how they can demonstrate meaningful outputs for the bibliographic, historical, and cultural interests of those fields. For computational textual stylistics (also now called "digital humanities") simple indicators like proportional word frequency have remained useful means of gathering evidence for tasks like authorship attribution. In the past they have relied on an assortment of such methods with mixed success, but have been turning to the more statistically advanced techniques of machine learning and probabilistic modeling, especially as the size of their data sets and scope of their questions about those data sets increase. Topic modeling in particular has provided an expressive demonstration of the potential for probabilistic modeling techniques to uncover meaningful patterns in textual corpora as determined by latent factors. And it has been recently used over Dickinson's poems as a means of uncovering shared thematics within her fascicle books (Armoza 2016a).

As automatic part-of-speech (POS) tagging has become more reliable, it too has become part of the digital humanities toolkit. The project described in this paper demonstrates how combining POS tagging and probabilistic modeling can create informative profiles of texts that can complement more semantic-oriented modeling techniques. Quantifying texts in this way often produces vectorized feature data over a finite vocabulary that is employed over a set of finite syntactical patterns. In the case of topic models we get topic distributions over individual texts and word distributions over those topics. However, the categories of POS are far fewer in comparison to the vocabulary of one text let alone thousands. One potential method for understanding the relationships between vectorized POS data is collaborative filtering. It makes use of latent factor models like nonnegative matrix factorization (NMF) that can meaningfully cluster such data by feature correlation. Matrix factorization (MF) decomposes input data matrices, approximating factor matrices of reduced dimension. This project speculates that if dimension-reducing methods like topic modeling can produce latent, heterogeneous patterns of semantic cohesiveness of words, matrix factorization of POS data may similarly produce latent, heterogeneous patterns in syntactic-functional cohesiveness.

Netflix's 2006 collaborative filtering contest asked researchers to see if they could produce a more accurate movie recommendation system than its own (Koren et al. 2009). The data set included explicit user data in the form of dated ratings of movies, with varying degrees of sparseness. Using this metaphor of users and movies, the premise is that given a matrix V containing known user-rating data (and perhaps other implicit user attributes) we can derive approximate matrix $V^* = U^T W$ where U contains the contribution of latent factors of users and W contains the contribution of latent factors of the movies being rated. This factorization attempts to account for the biases of users and movies. It also enables a predictive capability where factors can act as a model for future, unknown ratings. To prevent overfitting the model based on the observed features, a regularization variable λ is employed as a "penalization of the Frobenius norm of the feature matrices" while minimizing the sum-of-squared-errors objective function (Salakhutdinov and Mnih 2008, p.4). This regularization is enhanced in probabilistic matrix factorization (PMF) which considers the conditional distribution over user ratings:

$$p(V|U, W, \sigma^2) = \prod_{i=1}^{N} \prod_{j=1}^{M} [\mathcal{N}(V_{ij}|U_i^T W_j, \sigma^2)]^{I_{ij}}$$

where "$N(x|\mu, \sigma^2)$ is the probability density function of the Gaussian distribution with mean μ and variance σ^2, and I_{ij} is the indicator function that is equal to 1 if user i rated movie j and equal to 0 otherwise" (p.2-3). Gaussian priors are placed on the user/movie feature vectors, and hyperparameters for this model include the variances of those priors as well as the variance of observation noise. Maximizing the log posterior over the user and movie features is equivalent to minimizing the objective function: the regularized sum-of-square errors which includes a λ_U and λ_W that reflect the separate variances of those Gaussian priors for each matrix, U and W (Salakhutdinov and Mnih 2008, p.3). [1] Koren et al. (2006) and Salakhutdinov and Mnih (2008) describe two methods for minimizing this objective function: gradient descent and alternating least squares.

Of course, this user-item metaphor is not the only context for NMF. Those interested in tracing latent factors in genetic samples from populations of patients have also employed the method. In this case, the latent factors reflect "hidden genes" and "hidden patients." Zitnik et al. have developed a Python library for NMF for these purposes called "Nimfa" (Zitnik et al. 2012). Nimfa employs alternating least squares NMF with projected gradient and includes support for multiple MF variants including PMF. Koren et al. (2006) describes alternating least squares as more practical for training data that is less sparse. This makes Nimfa's choice ideal for the less sparse (~60% filled) POS data for this project, the 814 poems found in Emily Dickinson's fascicle books. The POS tagger used for this project, "spaCy," is an "averaged perceptron tagger...with Brown cluster features" that utilizes ~1GB of English language training data and performs with a claimed 93% accuracy (Honnibal and Johnson 2016a). [2] Below I describe a method for profiling the POS of texts via PMF as a means of comparative stylistic assessment.

PMF-POS Profiling Method

Taking Nimfa's patient-gene metaphor for PMF we substitute patients and genes with texts and POS, respectively. Texts are treated like unordered bags of words, each word being tagged with

[1] See Salakhutdinov and Mnih 2008 p.3 for log posterior and objective functions.

[2] More info on the spaCy text parser can be found at (Hannibal and Johnson 2016b).

a POS. Counts of those tags are vectorized for each text. Texts of varying size thus have summed counts for each POS category across the syntactical patterns of POS they contain. Using spaCy, we account for proper nouns, punctuation, determinants, adjectives, nouns, adverbs, whitespace, conjugations, verbs, participles, adpositions, numbers, pronouns, interjections, and symbolic characters. All remaining untaggable words are accounted for in a category, X. Each POS vector is normalized and preprocessed to ensure entries $V_{ij} >= 0$.

PMF is then used to identify "hidden texts," focusing on the latent factors of matrix U.[3] A consensus matrix of the PMF results over multiple iterations and runs of factorization is considered. Ward linkage is performed over this matrix to determine a hierarchical clustering which is then flattened to identify groups of texts correlated by those consensus factors. The POS vectors of each individual text in these clusters are then used to identify a centroid mean μ and within-cluster variance σ^2 (Halkidi et al. 2001).

These "hidden texts" qualify a unique "profile" of POS usage that reflect individual texts as well as the overall collection being modeled. Whereas a simple averaging of POS counts of a collection of texts would eliminate this individualized statistical information, describing and grouping the POS of these texts via PMF presents a more accurate characterization of them. The centroid mean μ and the within-cluster variance σ^2 for each cluster then accounts for the latent POS factors used by the text of each bibliographic container (book, manuscript, etc.) and the degree to which it varies from those factors. This profile may then be used as an evidentiary basis for stylistic comparisons (tasks like authorship attribution or semantic analysis) that implicitly or explicitly utilize POS. The latent factors themselves may also be further qualified by external information such as known authorship, known time period or locale of authorship, etc.

Usage Example: Emily Dickinson's "Fascicle" Manuscript Books

Using the method described above on the 814 poems of Emily Dickinson's forty fascicle books produces several perspectives of the latent POS speech usage within each poem and within each book. Below are descriptions of four such views, which refer to corresponding figures in the appendix of this paper.

Figure 1. *POS-PMF Profiles (by Poem) of Emily Dickinson's Fascicles*

In Figure 1, we take a look at the composition of each fascicle book by the centroid mean μ POS "profile" vector associated (via the PMF-POS profiling method) with each poem. These POS profiles stand in for the actual POS vector of each poem, each profile being given a unique color. Moving through each book from the top to the bottom of the graph we see a proportional representation (out of 100%) of their latent POS "genes." In total there are 247 clusters/profiles in this model. (Unfortunately, there is not enough room for displaying the full cluster to color legend in this space.) POS profiles exist in higher proportion in some books. Look at the blue "Cluster 7," for instance. However, the overall POS heterogeneity of the books adds credibility to an initial assumption behind topic modeling: that texts and collections of texts, particularly those containing more figurative language, are not semantically homogeneous.

[3]The problem of ideal factorization rank for PMF is left for the subject of future study, but the given example uses a rank of 2 dimensions.

Figure 2. *Percent of POS Use in PMF-POS Profiles in Emily Dickinson's Fascicles*

In Figure 2, we turn to POS itself by examining the average POS categories of the POS profiles utilized in each fascicle book (those profiles given for individual poems in Figure 1). The results are again given a proportional representation (out of 100%). Though we could separate out each POS category in its own graph for a more visually exacting dynamic over the forty books, this view is given to get an initial perception of those dynamics. Emerging trends of correlated POS speech usage in this figure give the user a sense of how those "hidden poems" have informed POS speech usage instead of looking to raw counts. One observation of note is how punctuation features in relation to nouns. In a previous study of Dickinson's exaggerated use of em-dashes, her nouns emerge as negatively correlated with her increased use of the em-dash, particularly in her later books, 11-40 (Armoza 2016b). In Figure 2 that effect is less pronounced.

Figure 3. *Average Poem Variance from their PMF-POS Profiles by Fascicle*

In Figure 3, we pay attention to how the POS vectors of individual poems vary from their PMF produced mean POS profiles. Within-cluster variance σ^2 of each poem/profile is averaged and the dynamic of those averaged variances are tracked in the forty fascicle books, from left to right. Though no clear pattern emerges at first glance, the outlier Fascicle 29 stands out in how widely its poems vary from their PMF-POS profiles.

Figure 4. *PMF-POS Similarity for an Individual Poem*

In Figure 4, we look at how these PMF-POS profiles can act as a stylistic similarity indicator for individual texts. Here, Poem 6 from Fascicle 21, "They shut me up in Prose" is listed alongside those poems found within its cluster produced by PMF consensus. In addition to further close POS analysis of the individual poems, we can also examine semantic similarities. Just by reading the poems we can see that "They shut me up in Prose" speaks of form limiting creative expression. "The Brain, within its Groove", a poem linked by this PMF-POS profiling method, speaks of creative expression being difficult to tame once it is freed from form.

Conclusion

Though further refinement and testing of the PMF-POS profiling method is needed, MF in general offers a promising set of models and tests to understand the complicated nature of high dimensional metadata of large textual corpora. POS profiling via PMF in particular provides a probabilistic foundation for undertaking the use of these methods, where quantifiable assumptions about data sets can be included via priors. While topic modeling has aided the comprehensive study of Emily Dickinson's poetry within the author-informed contexts of her manuscript books (Armoza 2016a), the comprehensive POS usage of Emily Dickinson remains an area open for exploration. As methods of correlating large amounts of this kind of feature data will be necessary for such work, the PMF-POS profiling method of this project/paper offers a sound probabilistic alternative for comparison with the results of other decompositional and dimension-reducing methods like PCA. In addition, stylistic comparison via latent factors presents a way of comparing sections and entire texts of authors, eras, cultures and beyond. Though model overfitting based on observations remains a hazard, the predictive capability of MF also allows for more concrete speculation for particularly sparse textual metadata.

References

Armoza, Jonathan. 2016a. Topic Words in Context: Dickinson, The Fascicle, and the Topic Model. [Montreal, QC Canada]: McGill University. (Available upon request.)

Armoza, Jonathan. 2016b. Model as Archive: Computational Perspectives of Emily Dickinson. [New York, NY]: New York University. (Available upon request.)

Dickinson, Emily. 1981. The Manuscript Books of Emily Dickinson. R.W. Franklin, editor. Cambridge, MA: Harvard University Press.

Dickinson, Emily. 2013. Emily Dickinson Archive. Cambridge, MA: Harvard University Press. http://www.edickinson.org/.

Halkidi M, Batistakis Y, Vazirgiannis M. 2001. On Clustering Validation Techniques. Journal of Intelligent Information Systems, 17:2/3, 107-145.

Honnibal M, Johnson M. 2016a. Citation Information #272. https://github.com/explosion/spaCy/issues/272.

Honnibal M, Johnson M. 2016b. spaCy. https://github.com/explosion/spaCy.

Koren Y, Bell R, Volinsky C. 2009. Matrix Factorization Techniques for Recommender Systems. IEEE Computer Society. August 2009:42-49.

Salakhutdinov R, Mnih A. 2008. Probabilistic Matrix Factorization. [Toronto, ON Canada]: University of Toronto.

Zitnik, Marinka and Zupan, Blaz. 2012. Nimfa: A Python Library for Nonnegative Matrix Factorization. Journal of Machine Learning Research. 13:849-853. http://nimfa.biolab.si/.

Appendix

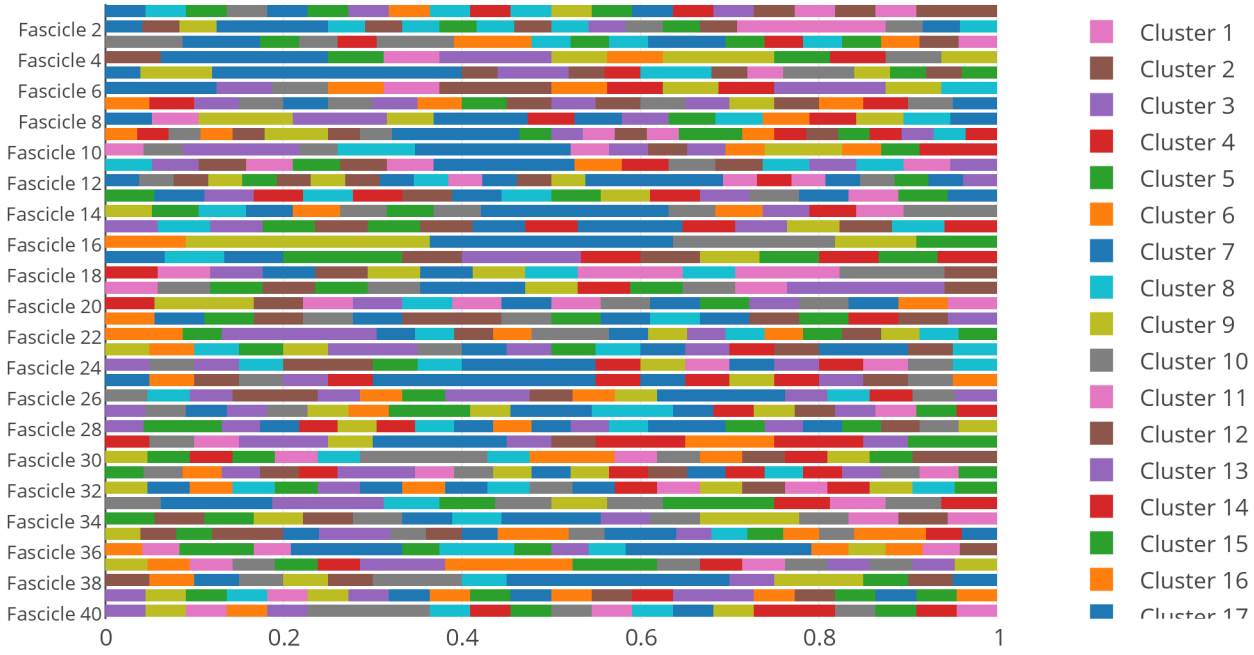

Figure 1: POS-PMF Profiles (by Poem) of Emily Dickinson's Fascicles

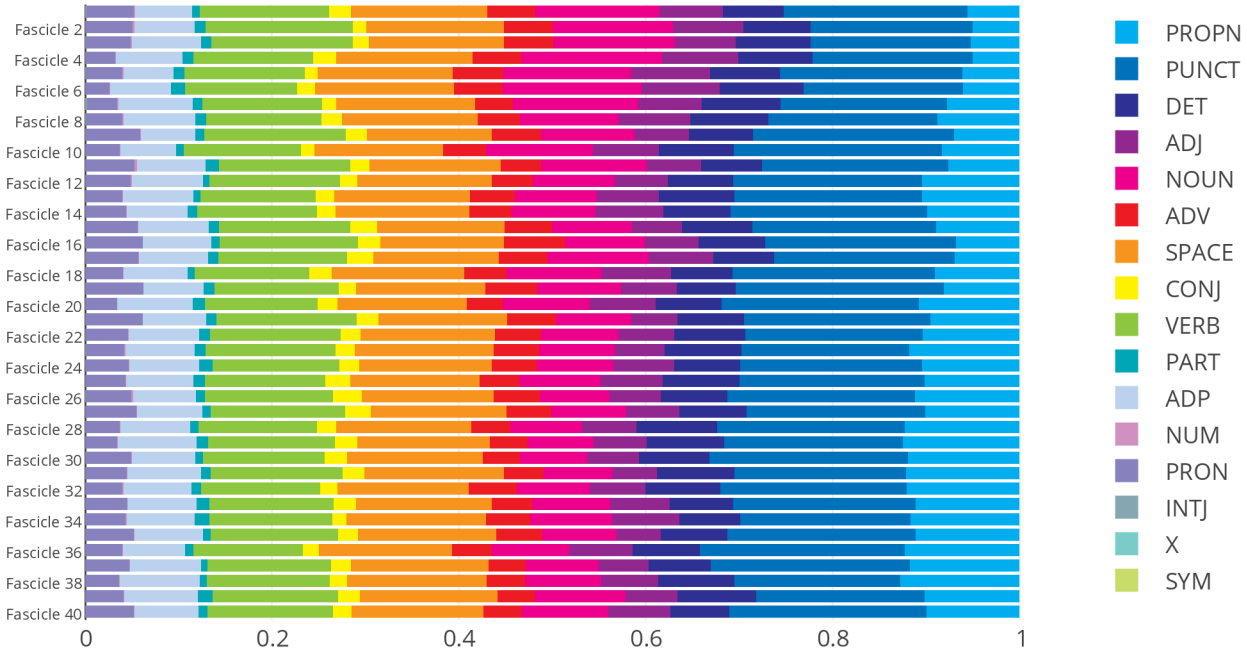

Figure 2: Percent of POS Use in PMF-POS Profiles in Emily Dickinson's Fascicles

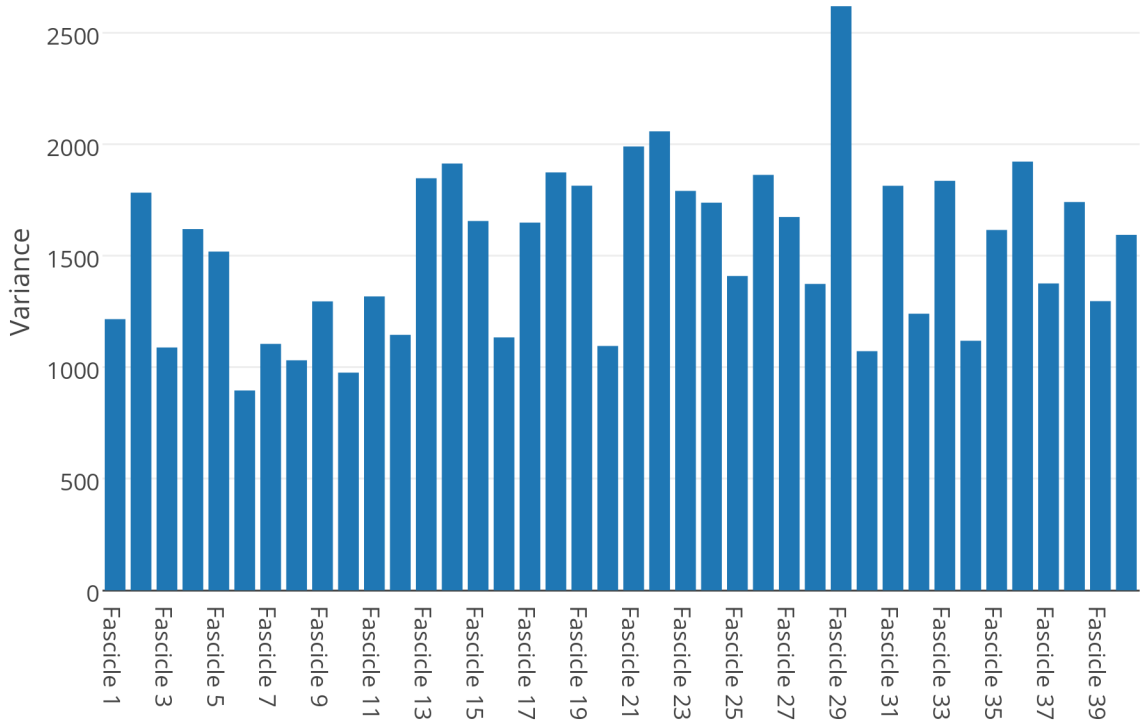

Figure 3: Average Poem Variance from their PMF-POS Profiles by Fascicle

Poem 6 of Fascicle 21: They shut me up in Prose
Part of Speech Cluster: 128
Similar Poems:
Poem 7 of Fascicle 12: "She sweeps with many-colored Brooms"
Poem 19 of Fascicle 25: "So glad we are - a Stranger'd deem"
Poem 3 of Fascicle 26: "I measure every Grief I meet"
Poem 16 of Fascicle 26: "The Brain, within its Groove"
Poem 9 of Fascicle 30: "I gained it so"
Poem 24 of Fascicle 34: "I many times thought Peace had come"

Figure 4: PMF-POS Similarity for an Individual Poem