Building a Curriculum for Robo-Poets

W. Yandell

University of Denver

Abstract

[The abstract should be one paragraph of between 150 and 250 words. It is not indented. Section titles, such as the word Abstract above, are not considered headings so they don’t use bold heading format. Instead, use the Section Title style. This style automatically starts your section on a new page, so you don’t have to add page breaks. Note that all of the styles for this template are available on the Home tab of the ribbon, in the Styles gallery.]

Keywords: [Click here to add keywords.]

Building a Curriculum for Robo-Poets

In this first period of research into the Poetry Foundation dataset consisting of 15,000+ poems, the primary goals have been as follows: clean (carefully), vectorize (expansively), and explore (poems in their entirety and individual lines). These map onto the larger goal of constructing text generators capable of producing poems which are coherent and similar to those in the dataset. An outline of the pipeline under construction:

1. In – Data as DataFrame
2. Pre-process Data
3. Doc2Vec Fit-Transform
4. LSTM 1: Line Generator
5. LSTM 2: Line Predicter
6. Seed/Vocab Generator
7. Out – No-Prompt-Needed Document Generator

# Data

The data of interest comes from the Poetry Foundation website by way of John Hallman on Kaggle (<https://www.kaggle.com/johnhallman/complete-poetryfoundationorg-dataset>). This dataset retains 15,638 unique instances of strings labeled ‘content’ as well as columns denoting the ‘author’ (3310 different authors represented), ‘title’, along with the Poetry Foundation ID and an index for each document.

The variety of the data cannot be understated. While a standard corpus may have some number of unique words and stylistic grammars, for poems, these are regularities or commonalities in the dataset that must be carefully engaged. Many common transformations such as lemmatization, stemming, or even certain types of filtering could incidentally remove characteristic elements which may be confounding in another problem but necessary to one like this.

In its preprocessed state, the dataset’s relevant features are unstructured and thus ill-suited for the modeling required to create a robust and coherent generator. Thus new features must be created and the strings must be cleaned before being vectorized to manage the space complexity while retaining as much information as possible about the latent structural and semantic spaces of the documents.

# Preparation v. Analysis

Because the relevant machine learning algorithms require numeric features, the content of each poem must be parsed, cleaned, and vectorized. The processing pipeline goes as follows:

1. Split content into lines
2. Clean aberrations from lines
3. Remove empty lines
4. Filter poems/lines with outlier lengths
5. Tokenize lines and poems
6. Vectorize lines and poems

Cleaning too aggressively could remove relevant structures, such as line breaks and cleaning insufficiently could result in spurious results from distinctions without difference. Thus multiple options were explored with a final result that balances processing concerns and project goals such that each poem is split into lines prior to heavy-duty text cleansing applying regular expressions, natural python expressions, and tokenization by the Natural Language Toolkit. A heavy duty processor run directly on the texts would have erased the notion of ‘line’ which is a rather important one so this is the highest priority step in processing. Of course, other relevant and arguably as or more consequential information will almost certainly be lost as is the nature of loss in processing. Retaining just this much without lemmatization or stemming exhausted the 24 cores of my system for some time.

In early explorations, filtering was comparatively minimal to avoid incidences of false identity, however this introduced merely an alternate form of false identity, as in situations like the use of a contraction being converted to a non-relevant signifier merely by the brute force but restriction on punctuation/non-Unicode alone. Though this exploration of punctuation should have to occur at some point, in order to produce coherent and similar works to the underlying texts behind any given text transformation result. Further, important liter-analytical constructs can be complicated by the nuances of parsing such that lexical analysis in particular becomes wrought with error and thus a trade-off for each of the particular goals.

A sequential ordering between preparation analysis is tricky when they have such a recursive relationship. Analysis provides insight which feeds back to processing which feeds forward to new analysis. For instance, there were some texts in this dataset which had only one line of a thousand words and others with thousands of lines caused by scraping error. Some poems, after origin look-up, were revealed to be missing the entirety of their content save the epigraph. If you tell any learner that any object is an exemplar which is actually a bug a sufficient number of times, then you can reasonably expect that learner would recognize instances of bug as instances of the true class of that object and when asked to construct an instance of the object: bug. Not that ‘bug’ or ‘glitch’ aesthetics are not also interesting and valuable paths of exploration, just that they are outside the scope of the current task. As are explicit considerations or featurizing of elements of sound in a given text, though these are not outside the scope of a full view and exhaustive exploitation/exploration of the dataset in the reinforcement learning sense.

Explicitly, we may look at summary analytical values and distributions to get an idea of what the inputs are like at each stage of processing (at the author, document, and line level) and glimpses of the overall latent structures in play. Stylometric features of a text with respect to line or poem can include a wide variety of measures and representatives such as lexical diversity, length, extracted entities, part-of-speech distributions,

References

Last Name, F. M. (Year). Article Title. *Journal Title*, Pages From - To.

Last Name, F. M. (Year). *Book Title.* City Name: Publisher Name.

Footnotes

1[Add footnotes, if any, on their own page following references. For APA formatting requirements, it’s easy to just type your own footnote references and notes. To format a footnote reference, select the number and then, on the Home tab, in the Styles gallery, click Footnote Reference. The body of a footnote, such as this example, uses the Normal text style. (Note: If you delete this sample footnote, don’t forget to delete its in-text reference as well. That’s at the end of the sample Heading 2 paragraph on the first page of body content in this template.)]

Tables

Table 1

[Table Title]

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Column Head | Column Head | Column Head | Column Head | Column Head |
| Row Head | 123 | 123 | 123 | 123 |
| Row Head | 456 | 456 | 456 | 456 |
| Row Head | 789 | 789 | 789 | 789 |
| Row Head | 123 | 123 | 123 | 123 |
| Row Head | 456 | 456 | 456 | 456 |
| Row Head | 789 | 789 | 789 | 789 |

Note: [Place all tables for your paper in a tables section, following references (and, if applicable, footnotes). Start a new page for each table, include a table number and table title for each, as shown on this page. All explanatory text appears in a table note that follows the table, such as this one. Use the Table/Figure style, available on the Home tab, in the Styles gallery, to get the spacing between table and note. Tables in APA format can use single or 1.5 line spacing. Include a heading for every row and column, even if the content seems obvious. A default table style has been setup for this template that fits APA guidelines. To insert a table, on the Insert tab, click Table.]

Figures title:

Figure 1. [Include all figures in their own section, following references (and footnotes and tables, if applicable). Include a numbered caption for each figure. Use the Table/Figure style for easy spacing between figure and caption.]

For more information about all elements of APA formatting, please consult the APA Style Manual, 6th Edition.