XX ENMC
ENCONTRO NACIONAL DE MODELAGEM COMPUTACIONAL
VIII ECTM
ENCONTRO DE CIÊNCIA E TECNOLOGIA DE MATERIAIS
16 a 19 de Outubro de 2017
Instituto Politécnico - Universidade do Estado de Rio de Janeiro
Nova Friburgo - RJ

A SIMPLE MODEL FOR THE LITERARY ANALYSIS BY TEXT ANALYTICS: COMPARATIVE APPROACH AND EXAMPLE ON JAMES JOYCE AGAINST SHAKESPEARE AND THE BIBLE

Renato Fabbri[1] - renato.fabbri@gmail.com
Luis Henrique Garcia Ferreira[2] - henrique@hotmail.com
[1]University of São Paulo, Institute of Mathematical and Computer Sciences - São Carlos, SP, Brazil
[2]University of Campinas, Institute of Language Studies - Campinas, SP, Brazil

Abstract. *Literary analysis, criticism or studies is a largely valued field with dedicated journals and researchers which remains mostly within the humanities scope. Text analytics is the computer-aided process of deriving information from texts. In this article we describe a simple and generic model for performing literary analysis using text analytics. The method relies on statistical measures of: 1) token and sentence sizes and 2) Wordnet synset features. These measures are then used in Principal Component Analysis where the texts to be analyzed are observed against Shakespeare and the Bible, regarded as reference literature. The model is validated by analyzing selected works from James Joyce (1882-1941), one of the most important writers of the 20th century. We argue for the consistency of this approach, the reasons why we did not use other techniques (e.g. part-of-speech tagging) and ways by which the analysis might be adapted and enhanced.*

Keywords: *Text analytics, Literary criticism, Wordnet, Shakespeare, Bible*

1. INTRODUCTION

Literary criticism is performed by intellectuals using various techniques, including intuition and contextualization through erudition (Richards, 2003). Text analytics is usually considered a synonym of text mining, i.e. data mining applied to textual data, the extraction of meaningful information from texts by means of computer-aided analysis. A difference can be established nevertheless: text mining is more associated to earlier applications (e.g. dating to the 1980s) and to specific tasks, while the term text analytics is more frequent nowadays and might be related to a less purposeful processing of textual data. Accordingly, for example, a word cloud is more easily associated to text analytics while a search engine is more promptly associated to text mining (Wikipedia, 2017).

In this work we propose a very simple and generic model for literature analysis by means of statistical measures, Principal Component Analysis and comparison against reference literature. The uncomplicated methods favor the collaboration between researchers of different

background. For example: a computer science professional can understand, adapt and expand the techniques while a literature scholar can deepen the interpretation and the relevance of the conclusions.

Section 2 describes the corpus and methods. Section 3 is dedicated to presentation and discussion of results. Section 4 holds conclusions and further work considerations.

2. MATERIALS AND METHODS

2.1 Corpus

This work encompasses a comparison of the literature to be analyzed against reference literature. What is regarded as reference literature is arbitrary and we chose them, within this presentation and first formalization, to be the two utmost references of the English language (Norton, 2000; Boom, 1999):

- the complete works from William Shakespeare as given by the publication in the Gutenberg Project (Shakespeare, 1994): 36 plays (tragedies, comedies and historical) and poetry (2 batches). Shakespeare is often recognized as the greatest writer of the English language and is a universal reference of literature.

- The 80 books of the King James Bible, including Old Testament (39 books), Apocrypha (14 books) and New Testament (27 books). This is the most referenced English translation of the bible. These books are also universally accredited for their influence in English literature.

We should emphasize that changing this reference literature is very straightforward. One should only provide the corresponding text files and modify the scripts to read the intended records. If the works are well-known, the process should require only a quick search on the web (e.g. within Gutenberg or Archive.org projects), saving the text locally and then changing filenames in the scripts. Some possibilities include: other masters of English literature; a selection of poets; works from scientific literature of a specific field; works in other language, such as Machado de Assis and Clarice Linspector if analyzing works in Brazilian Portuguese. There is no reason why the corpus should not include data streaming or access to online resources, such as Wikipedia pages.

To illustrate and validate the method, we performed and herein report an analysis of a selection of works written by James Joyce:

- Stephen Hero: written around 1905 and published posthumously in 1944, an autobiographical novel of which part is lost (Joyce threw it on fire after a number of rejections by publishers).

- Dubliners: published in 1914, it is a collection of 15 short stories about Dublin's middle class.

- A Portrait of the Artist as a Young Man: published in 1916, a condensed and reworked version of Stephen Hero.

- Ulysses: published in 1922, considered one of the most important works of the modernist literature.

- Finnegans Wake: published in 1939, often considered one of the most difficult fictional works of the English language.

2.2 Pre-processing

The reference literature (Shakespeare and Bible books) were separated and cleaned into individual files. As both collections do not hold well defined paragraph structures, these were discarded. These routines can be inspected through reading the scripts in Table 1.

Table 1: Files related to the analysis method proposed in this article. All files are found in a public git repository (Fabbri & Ferreira, 2017).

File	Description
`scripts/analysis.py`	Python script that makes the initial quantification of the books
`scripts/analysis2.py`	Python script that performs PCA and renders the figures with scatter plots
`scripts/bibleSeparation.py`	Python scripts that separates the King James Bible text into files with individual books
`scripts/shakespeareSeparation.py`	Python scripts that separates the text with the complete works of William Shakespeare into files with individual books
`corpus/*`	text files corresponding to individual books from Shakespeare, Joyce and the bible
`latex/*`	the PDF of this article and the files necessary to render it. It is the main documentation of the proposed analysis.

2.3 Analysis routine

As modeled until the moment, the analysis is performed by: the achievement of meaningful sets of textual elements, quantifying their incidences, taking overall measurements of these quantifications in each of the books, performing Principal Component Analysis (PCA) of the books in the measurements space, plotting the books within principal components and most meaningful measures, interpreting the results. We should look at each of these phases:

- Achievement of meaningful sets of textual elements: the original texts were separated into sets of: sentences, tokens, stopwords, known words (which are not stopwords), punctuations, tokens which are not stopwords or punctuation, Wordnet synsets of each known word.

- Quantization: each of the sets above were quantified by: means of their sizes in number of characters of each element, or by means of the number of elements they contain, or by means of synset characteristics (only depth was used in the example analysis).

- For the PCA, all the books were considered together. The z-score of each dimension (measure type) was performed to avoid meaningless prevalence of some measures over others (z-score of measures x_i is $x'_i = \frac{x_i - \mu(x)}{\sigma(x)}$ where $\mu(x)$ is the mean of all x_i and $\sigma(x)$

is the standard deviation). Then PCA was performed as usual: performing the eigendecomposition of the covariance matrix (where entry m_{ij} if the covariance of measures i with measures j) and observing the eigenvectors associated to the greatest eigenvalues.

- For visual inspection of the resulting structures, we used scatter plots of principal components and of measures which were relevant to our analysis of James Joyce's works.
- The interpretation of the results, we made discussions about literary analysis and the analysis of James Joyce before performing the quantitative analysis described above. When the final figures and measures were done, we had another round of considerations about what they revealed.

We discarded using other techniques mainly because of three reasons: 1) other methods involve greater complexity and would not favor the communication between interested parties; 2) e.g. part-of-speech tagging relies heavily on the vocabulary and the syntactic structure which are used with deep innovations by literary authors, especially from the last century and thereon; 3) using only the measures mentioned above, we already reached 20 dimensions. Nevertheless, we encourage adapting the method by inclusion of other measures and of other analysis procedures beyond PCA, and we will probably do so in further considerations of this efforts.

3. RESULTS AND DISCUSSION

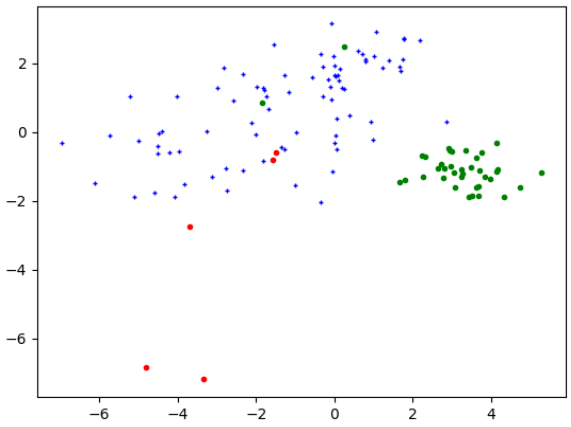

Figure 1: First two principal components.

Figures 1, 2 and 3 exposes the works of Joyce, Shakespeare and bible books within the principal components. As the second and third component held very near spreads, and the first two components summed only $\approx 50\%$ of all dispersion, we chose to use the first three components. As can be noticed, Joyce's works are very distinct from Shakespeare, and some of them are also very distinct from bible books. Even so, some of them fall near bible books.

Figures 4 and 5 are direct plots of measures we idealized to probe the extent of the need of interpretation by the reader. The first of these plots is dedicated to the amount of unknown words, and the conclusion is that some of the works have an very distinctive amount of unknown

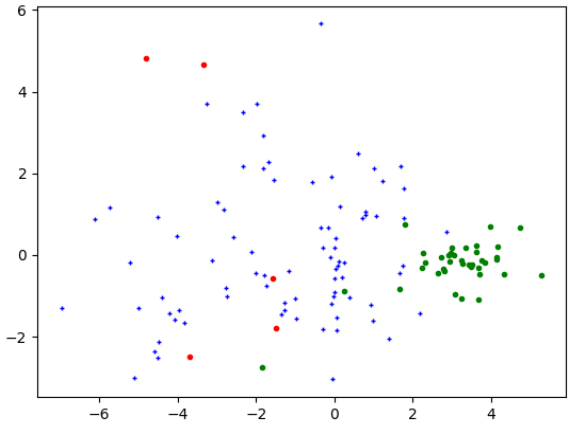

Figure 2: First and third principal components.

words, but all of them fall on the greater amount of unknown words among the most meaningful tokens when the same amount of unknown words is considered. The second plot is dedicated to synset depth. The lower the depth, the more abstract the concept is. In this plot, we conclude that three of the works by Joyce lie on the more abstract margin among the reference works, but two of them lie within the middle and the more concrete (less abstract) books.

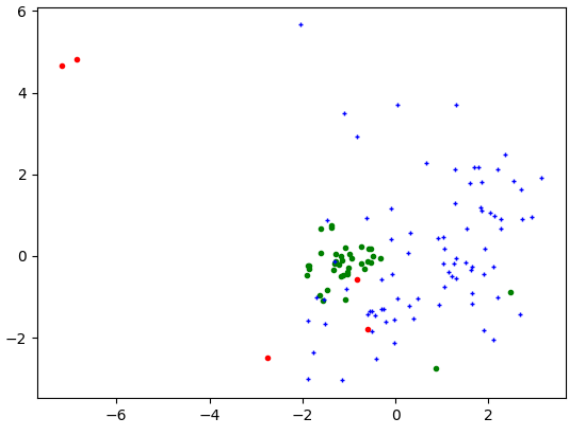

Figure 3: Second and third principal components.

We propose to validate and illustrate the method by analyzing works by Joyce, but, as this is the first work of the kind which analyzes Shakespeare and the bible, as far as the authors know, some considerations about them are also opportune. First, the works by Shakespeare lie in a notably more restricted domain when compared against the bible. Second, they are perfectly distinguishable with respect to the first two principal components: a simple Bayesian inference of neural network should be able to correctly classify a book from one group or the other. Third, Shakespeare uses a less abstract language at least in the sense captured by the depth of the synsets. This diversity is convenient for a reference literature to compare something against.

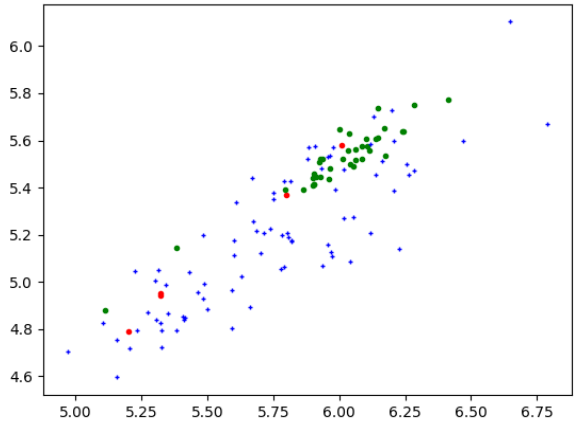

Figure 4: Synset depths. Lower depth is regarded here as evidence of abstraction.

Finally, we believe to have reached a good result in terms of the model proposed for the analysis. The analysis is very simple, which favors both elaboration of variants and the understanding by interested researchers which are potentially from diverse and multidisciplinary background. It is robust, in the sense that it does not rely in canonical vocabulary or syntactic structures. Furthermore, the method is very fast: pre-processing and then processing and rendering the figures can all be performed in dozens of seconds.

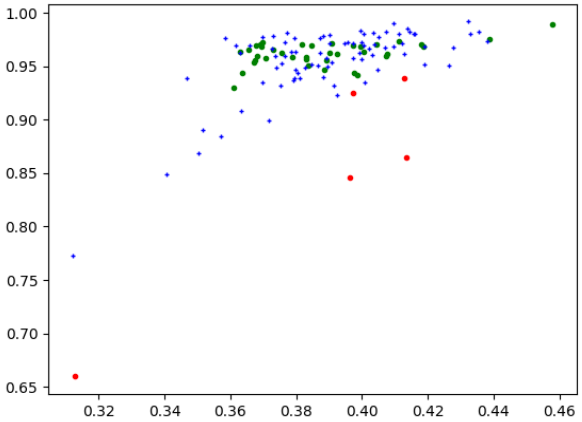

Figure 5: Fraction of known words among all tokens and significant words. Lower faction is regarded here as evidence of abstraction.

4. CONCLUSIONS AND FUTURE WORK

The analysis method proposed yields interesting results for literary criticism. It is robust, easily adaptable and fast. Also, the online availability of the scripts and the reference corpus,

all in public domain, facilitates reuse and the achievement of derivatives. The example analysis, revealed both distinctive traces of the set and can be used to argue quantitatively in favor of the thesis that the style of Joyce calls the reader to fill the meaning gaps generated by the abstraction.

In further efforts, we should:

- Deepen the analysis of the reference literature (book by Shakespeare and in the bible) to better contextualize any literature we consider against them.
- Expand the use of Wordnet to encompass synonymy, antonymy, meronymy, etc. Also to consider specific roots of nouns, adjectives, verbs and adverbs.
- Report this endeavor to the literary criticism scholars. This should be done at least in two ways: by describing the method and its relevance within the humanities background; and by exposing results from analyzing specific authors, such as Joyce, Ezra Pound, etc.
- Consider other measures of abstraction. Can we use length of words and sentences?
- Vary the methods and state reasonable generic bounds e.g. for splitting a work to obtain more data points.

Acknowledgements

The authors thank the open source software developers, especially those who enabled this work by developing the Python language and the NLTK; the open culture movement, especially the collaborators of the Gutenberg and the Archive.org projects, which enabled this work by making the literary works available; the IFSC/USP, ICMC/USP and IEL/UNICAMP researchers for their attentive collaboration whenever we required opinions and directions for learning and researching.

REFERENCES

Richards, I. A. (2003). Principles of literary criticism. Routledge.

Text mining. (2017, July 31). In Wikipedia, The Free Encyclopedia. Retrieved 02:02, August 10, 2017, from https://en.wikipedia.org/w/index.php?title=Text_mining&oldid=793168242

Norton, D. (2000). A history of the English Bible as literature. Cambridge University Press.

Bloom, H. (1998). Shakespeare: The invention of the human. Riverhead Trade (Paperbacks).

Shakespeare, W. (1994). The Complete Works of William Shakespeare. The Project Gutenberg.

Fabbri, R., Ferreira, L.H.G. (2017). Public git repository with the scripts and corpora for the quantitative analysis to support literary criticism. From https://github.com/ttm/joyce