Refining character relationships using embeddings of textual units

Guillaume Guex[1]

[1]Departement of Language and Information Sciences, University of Lausanne, Switzerland

Abstract

A clear and well-documented LaTeX document is presented as an article formatted for publication by CEUR-WS in a conference proceedings. Based on the "ceurart" document class, this article presents and explains many of the common variations, as well as many of the formatting elements an author may use in the preparation of the documentation of their work.

Keywords

LaTeX class, paper template, paper formatting, CEUR-WS

1. Introduction

Distant reading tools allow researchers, from various fields, to quickly gain knowledge on textual corpora without actually reading them. Purposes of these methods are various, but can be mainly categorized into two groups: in the first case, these methods are used in order to tag, classify, or summary large quantities of texts, in order to quickly structure information or to deliver a speech over the whole studied corpus. Methods in this case rely heavily on Big Data and make an extensive use of Machine Learning algorithms. In the second case, researchers use these methods to underline hidden structures in a particular text, helping them to refine their understanding of it and reinforce stated hypotheses. Methods in this setting can also rely on Machine Learning, but must typically be build with more caution and attention to details: corpus are smaller, analyses are closer to the work, and methods must be more transparent in order to appropriately interpret results.

Automatic extraction and analysis of *character networks* from literacy works typically belong in the latter group. These methods aim at representing various interactions occurring between fictional characters found in a textual narrative with a graph, thus showing explicitly the hidden structure of character relationships constructed by the author. This structure might allow to find hidden pattern within book, which can highlight a particular genre or author style and help to understand a part of the "flavor" given by the author to the text.

2. Methods

When building character networks from a textual narrative, the most widespread method consists in dividing the studied work into n textual units $u_1, \ldots, u_n$, which can be, e.g., sentences, paragraphs, or chapters, and counting characters co-occurrences in these units. Usually, the text constituting these units is discarded and the resulting network displays edges which roughly represent an aggregated number of interactions between characters. However, by doing so, the aggregation occurs on various type of interactions and will give little information about the type of relationship which exist between characters. In this paper, we propose a data organization, leading to various type of analyses, which permits the use of the text contained in the unit in order to characterize relationship

2.1. Data organization

A textual narrative divided in n textual units $u_1, \ldots, u_n$ can be represented in a $n \times (p + 1)$ table N, where p is the number of characters found in the narrative. Each line represent a textual unit, the first column is the text composing this unit, and the remaining p columns contains the number of character

3. Results

4. Conclusion

References

COMHUM 2022: Workshop on Computational Methods in the Humanities, June 9–10, 2022, Lausanne, Switzerland

guillaume.guex@unil.ch (G. Guex)

0000-0003-1001-9525 (G. Guex)