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 patterns within a book, which can highlight a particular genre or style.

2. Methodology

When building character networks from a textual narrative, the most widespread method consists in dividing the studied work into n textual units $u_1, \dots, u_n$, 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 methods which takes into consideration the text contained in the unit, helping to refine understanding of characters and their relationships.

2.1. Data organization

A textual narrative divided in n textual units $u_1, \dots, u_n$ is represented in this article by two tables. The first one is well known in the field of textual analysis, and consists in the $n \times v$ contingency table $\mathbf{N}$, as represented by Table 2.1, where v is the vocabulary size. In this table, each row is an unit, and each columns contain a variable V_i, which counts the number of times the word i appears in the unit. Using this table typically denotes a *Bag-of-Words* approach to our analyses.

The second table is a $n \times (p + p^2)$ table $\mathbf{C}$, as shown in Table 2.1, where p is the number of characters found in the narrative. Each row represents a textual unit, the first column is the text composing this unit, and the next p columns contains variables $O_1, \dots, O_p$ counting occurrences of each character in the unit. The $p \times p$ remaining columns contains variables C_{ij}, which count characters co-occurrences and can be defined with $C_{ij} = \min(O_i, O_j)$. Note that higher order co-occurrences could also be considered (e.g. variables

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)

denoting the co-occurrences of three characters), but it remains outside the scope of this article.

	a	alarme	alerte	...	est	...
u_{4756}	1	0	0	...	0	...
u_{4757}	0	1	1	...	1	...
u_{4758}	0	1	0	...	0	...
u_{4759}	0	2	2	...	2	...

Table 1
A snippet of the contingency table $\mathbf{N}$ extracted from *les Misérables*. Rows are paragraphs, columns are words in the vocabulary, and cells count the number of time each word appear in the unit.

	Azelma	Babet	Brujon	...	Babet-Brujon	...
u_{4756}	1	0	0	...	0	...
u_{4757}	0	1	1	...	1	...
u_{4758}	0	1	0	...	0	...
u_{4759}	0	2	2	...	2	...

Table 2
A snippet of the character table $\mathbf{C}$ extracted from *les Misérables*. Rows are paragraphs, the first p columns counts character occurrences, and the last $p \times p$ columns character co-occurrences.

This data organization can seem trivial, but it already gives an orientation to the subsequent analyses and should be kept in mind of the practitioner. Textual unit are now considered as *individuals* (in the statistical terminology), defined by their *variables* contained in the different columns of both tables. Moreover, subsequent analyses are oriented in searching how the character table $\mathbf{C}$ have an influence over the contingency table $\mathbf{N}$, i.e. searching which words are over-represented or under-represented knowing the characters composition in the unit. While authors use characters in order to build the narrative, we, to a certain extent, work backward: we are searching how character appearances in the textual unit influence her or his choice of words. The choice to include character co-occurrences along with occurrences is motivated by the fact that the interaction between two (or more) characters should be considered as an object of its own: this interaction is not necessarily the sum of its parts and gives a particular flavor to the unit.

The use of a contingency table to represent the textual resource present in the units denotes a *Bag-of-Words* approach. Using this is approach loses the information relative to the order of words in the units, but permits to transform a chain of characters, improper to statistical analyses, into a contingency table, an well studied mathematical object which permits the use of various kind of statistical methods.

This data organization also highlight the importance of chose a proper size for the units. These units should be large enough to contain enough words in order the properly capture the textual specificity of each unit, but not too large, as each unit should ideally capture a particular kind of state of characters in the narrative, or a particular interaction between them. Unfortunately, it is impossible to define an ideal size for all kind of textual resource, and this size should be balanced regarding knowledge of the studied work.

2.2. Methods

Various methods can be performed using the previously stated data organization, such as Sentiment Analysis (REF), Textometry (REF), or even Deep Learning methods. Here, we make to choice to extract a lower dimensional, numeric representation of each units, in other words, an *embedding*. Following the guideline stated in the previous section, the approach here is to use the contingency table N in order to embedded the textual units, which will help to see how resulting vectors depends on character occurrences and co-occurrences. Various kind embeddings can be performed, and we will focus here on two methods. Section (REF) describes *Correspondence Analysis (CA)*, and section (REF) focuses on *Pre-trained, static, word vectors (WV)*. In section (REF), we show two different way to use the table $\mathcal{C}$ in order to further embed characters and character pairs into the same spaces.

2.2.1. Correspondence Analysis

Using Correspondence Analysis in order to analyse textual resources has a long tradition (REF). It has the advantage to provide an embedding space, the factorial map, where units are placed alongside words, allowing to analyse or represent the different units in terms of word specificities. Moreover, units and words placement in the embedding space have a direct interpretation in terms of chi2 distance and profiles, which is desirable when interpreting results.

By performing the Correspondence Analysis, we get n vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ corresponding to units (rows) of table $\mathbf{N}$ and v vectors $\mathbf{w}_1, \ldots, \mathbf{w}_v$ corresponding to words (columns). Each of these vectors has a size of $\min(n, v) - 1$, which will generally be $n - 1$. (MORE EXPLANATION ON CA ? SEE MQ IV).

Note that unit vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ obtained from CA reflect their profile (in terms of words) regarding the mean profile. This mean that this analysis is in fact *contrastive*: we highlight unit variations among the studied text. This analysis can therefore becomes problematic when used on multiple texts: units belonging to the same text have a great chance to be close from each other, and the displayed variance might be largely composed by differences of style between text.

2.2.2. Pre-trained word vector

Pre-trained word vectors, based on Word2Vec (REF), Glove (REF), or Fasttext (REF), have recently received great attention from various fields (REFS). The generally originate from a training of very large corpora, such as Wikipedia (REF) or Common Crawl (REF) and results in a embedding containing a large quantities of word vectors.

2.2.3. Character and character pairs embeddings

3. Results

4. Conclusion

References