From Close to Distant and Back: How to Read with the Help of Machines

Rudi Bonfiglioli*, Federico Nanni**
* Textkernel, Netherlands
** University of Bologna, Italy

Digital Humanities (DH) is a variegate field of study that combines a humongous number of different interactions between humanist disciplines and the use of the computer. In recent years researchers have noticed a common trend among these different approaches characterized by the adoption of quantitative methods in the study of digital sources [1].

In particular, due to Franco Moretti's definition of "distant reading" as a group of text mining approaches proposed in opposition to traditional-hermeneutic analysis [2], the DH community is dividing itself in two opposite factions [3]. Central to this division is the idea that computational methods seem to move in the direction of making the work of the humanist irrelevant for the production of insights.

Starting with these assumptions, the purpose of our paper is twofold: first, we intend to stress how text mining methods will always need a strong support from the humanist, and second, we argue about the usefulness and necessity of advanced text mining approaches in the DH.

We would like to think of a humanist research involving computational techniques as a three steps process. The first step is a "close reading", which includes selecting a specific case study, crafting the initial features, and labeling of the training corpus. The second step is a "distant reading" since it involves performing a computational analysis. The third step is another "close reading", which consists of evaluation and interpretation of the results and the use of these results in a humanities research.

At the same time, we think that researchers should not renounce text mining approaches, but should instead experiment with advanced methods such as the ones belonging to the family of deep learning [5]. Deep learning techniques essentially perform representation learning, and therefore permit to automatically analyze text as a multilayered set of encoded features.

Reading

Traditional "close" readings of literary texts reach insights by considering a multitude of different factors, such as the choice of the vocabulary, the syntactical constructions employed, or knowledge of the author background or historical context. As humans, we discuss which combinations of values for those factors can signal "pathos" or "Victorian writing style" and then we reach insights by recognizing the patterns of those combinations of values in texts. In the domain of computational methods, we call those combinations of values "features", and texts are analyzed to expose patterns of such features. Ideally, we would like computational methods to work with the same features we use as humans, but there is no straightforward way to encode in them, for example, the idea of "Victorian writing style".

For the purposes of DH, we believe that a text has to be considered as a multi-layered medium with the various layers expressing increasing connotations of meaning. Simple word meaning is the most distant one, furthest away from the text; the next layer can be a syntactic layer, and then we can find layers that express information related to the historical context, author background and so on.

Designing the features implies choosing the right layer of meaning and then express features in digital terms, adapting the input representation if necessary. For such operations (known as "feature engineering") domain specific knowledge is known to be essential [4].

Therefore, a traditional "close reading" is an important prerequisite for running a computational analysis, the subsequent "distant reading". Afterwards, additional work must be done to translate the obtained patterns, or correlations, into insights for the humanities, which require at least a strong causation relationship. This work requires again a "close reading" of the text. Digital humanists seem to be seduced by the "Big Data" rhetoric of "making the data speak for itself", but while correlations alone may be enough to build a recommendation systems, they are not sufficient to build-up knowledge in the domain of humanities.

Deep reading

As we remarked before, assembling complex features that span over different "layers" and that capture with precision deep concepts of a text can be very difficult and time consuming. This is a known issue in the domain of AI [4]. Therefore, the effort spent in the field of "representation learning" has increased rapidly in recent years [5]. Representation learning aims to automatically discover explanatory factors, and thus features, decreasing the human labour spent in feature engineering. More specifically, deep learning methods that perform representation learning could be the most promising ones for the DH. This is because such approaches can automatically learn multi-leveled representations and generate hierarchical features capturing more (or less) abstract properties of the input, which matches the way texts are analyzed in the humanities. Thus, deep learning appears to be a suitable method for the distant-reading stage of a DH research, although, the advantages of its use still depend on a solid domain specific knowledge of the researcher.

Conclusion

Having observed the emerging factions in DH, we proposed a three-steps framework to conduct research using text mining techniques, and showed how the framework helps, reasoning at a deeper philosophical level, to blur the contrasts present in the field. We think that the use of advanced computational methods is an important area of research that must be pursued, and argued that deep learning could be beneficial. Moreover, we stressed the importance of understanding that qualitative knowledge rooted in the domain of humanities is essential and can not be ignored by works focused on computational methods. In this sense, we believe that, especially in the field of DH, exploiting complementarity between computational methods and humans will be the most advantageous research direction.

References

[1] D. Berry, “The computational turn: Thinking about the digital humanities”, *Culture Machine* 12(0), 2011.

[2] F. Moretti, *Distant reading*, Verso Books, 2013.

[3] T. Underwood, “Why digital humanities isn’t actually 'the next thing' in literary studies”, 2011, http://tedunderwood.com/2011/12/27/why-we-dont-actually-want-to-be-the-next-thing-in-literary-studies/

[4] P. Domingos, “A few useful things to know about machine learning”, *Communications of the ACM*, 55(10), 2012, pp. 78-87.

[5] Y. Bengio, “Deep learning of representations: Looking forward”, *Statistical language and speech processing*, 2013, pp. 1-37.