Some remarks on text data visualization and codec transparency

Bryan Jurish
Berlin-Brandenburgische Akademie der Wissenschaften
jurish@bbaw.de

By "visualization", I refer to a generic algorithmic procedure by which an underlying data source may transformed to graphical form for direct human consumption, e.g. as a network graph, tag cloud, motion chart, etc. A "text data visualization" is simply a visualization procedure using a (digital) text corpus as its underlying data source. In this talk, I submit the following propositions for discussion:

1. Visualization procedures – especially text data visualizations – cannot always be clearly distinguished from the preprocessing machinery which supplies their input parameters, since the formal model underlying a given visualization procedure imposes hard constraints on the structure of those parameters. There is little to be learned from a network graph visualization of a flat list of unweighted terms, for example.

2. Visualization tool-chains can be understood as *filters* in the sense of Shannon's (1948) model of communication. Text data visualizations tend overwhelmingly to be "lossy" filters, degrading messages passed through them. Such lossiness is due at least in part to an implicit demand for high compression rates on the tool-chains as a whole – we already have the flat serial text-encoding available to us.

3. In Shannon's terms, natural language is itself a lossy filter (Reddy, 1979). Moreover, the human users who are the final consumers of the visualization's output can be assumed to be equipped with a great many more integrated lossy filters, e.g. linguistic filters for parsing (minimal attachment) and interpretation (semantic priming), perceptual ones for motion detection, cognitive filters for object independence and causal relations, as well as cultural ones for shared experience and common knowledge. Adding another (lossy) filter to one's data intake process increases the informational "distance" in Moretti's (2013) sense, but does not change the fact that the communication channel between the information source (text, author, object) and the recipient (ourselves, subjects, minds) is already noisy (i.e. fallible).

4. The "intuitivity" often predicated of (text) data visualizations is nothing more or less than an exploitation of the human users' pre-existing perceptual/cognitive/cultural filters by use of color, motion, size, or shared conventional signs. Such exploitation can be considered successful to the extent that all and only the *relevant* data is passed through both the programmatic and user-integrated filters.

5. Assuming that the ultimate aim of the visualization pipeline is *communication* of the text-encoded message to the human user, Grice's (1975) well-known cooperative principle suggests that our task as builders of visualization tool-chains would be most effectively performed by maximizing our filter codecs' transparency, optimizing our tool-chains for the users' common research goals, analogous to the optimization of popular lossy audio codecs (e.g. mp3, ogg) for the human auditory perceptual apparatus.

References

P. Grice. Logic and conversation. In P. Cole and J. Morgan, editors, *Syntax and Semantics*, volume 3: Speech Acts, pages 41–58. Academic Press, 1975.

F. Moretti. *Distant reading*. Verso Books, 2013.

M. J. Reddy. The conduit metaphor: A case of frame conflict in our language about language. In A. Ortony, editor, *Metaphor and Thought*, pages 284–310. Cambridge University Press, 1979.

C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948.