Exploratory analysis of word frequencies across corpus texts: towards a critical contrast of approaches

Andrew Hardie (Lancaster University, UK)

A recent trend in corpus linguistics is the adoption of *Latent Dirichlet Allocation* (LDA), already widely used by digital humanists (Blevins, 2010; Underwood, 2012) as a method for exploratory corpus analysis. LDA is a machine-learning approach to inducing structure in the content of a corpus based solely on word occurrence across texts or documents as data objects, one of a range of approaches usually if potentially misleadingly dubbed *topic modelling*. However, adopting this approach to the many-dimensional data of word frequency comes with a high price tag in terms of knowledge that the system ignores or makes nontransparent. The question this raises is whether that price tag is justified.

Various advantages have been asserted for LDA, albeit not without caveats (see Blei, 2012 for a selection of both). All such advantages notwithstanding, LDA has at least three substantive disadvantages. First, it is nondeterministic: randomisation is central to the algorithm. This is problematic from the perspective of scientific replicability for reasons too obvious to belabour. Second, its operation is opaque: the relationship between the underlying distribution data and the resulting statistical model is nontransparent to the analyst. Third, the theory of text generation underpinning the LDA algorithm is dubiously compatible with linguistic understandings of text, topic and discourse.

Moreover, although the lack of linguistic knowledge used in the construction of the model is presented as an advantage of LDA, this is equally characterisable as a disadvantage: the field of corpus analysis has invested much effort in the creation of precisely the knowledge resources which LDA is lauded for not requiring. What exactly does our acceptance of these disadvantages buy us? In examining this issue, we must venture comparisons to longer-established exploratory multivariate analysis approaches that are longer-established in corpus linguistics (cf. Biber, 1988, 1989).

Using example data drawn from the FLOB corpus, I will compare and contrast outcomes of different analytic procedures including LDA models and alternative approaches, with two questions in mind. First, to what extent are these outcomes *compatible with one another*? Second, to what extent are they *transparently interpretable* in linguistically meaningful terms?

References

Biber, D. (1988). *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, D. (1989). A typology of English texts. *Linguistics,* 27, 3–43.

Blei, D. (2012). Probabilistic Topic Models. *Communications of the ACM, 55*(4), 77-84.

Blevins, C. (1991). *Topic Modeling Martha Ballard's Diary*. Retrieved from http://historying.org/2010/04/01/topic-modeling-martha-ballards-diary/ [Accessed 2017-05-05].

Underwood, T. 2012. *Topic modelling made just simple enough*. Retrieved from http://tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough/ [Accessed 2017-05-05].