Text mining in practice: A discussion on user-applied text mining techniques in historical research.

Language: English, Duration: 60 minutes

In this panel we look at the application of text mining techniques in historical research. In recent years, text mining has come within reach of any vaguely computer-literate scholar. The growing availability of large digital text collections leads to growing abilities to apply digital and quantitative approaches to the study of historical texts. Commonly used languages and statistical environments such as Python and R, offer applicable software solutions for free. This has liberated historians and other humanities scholars from the shackles of time-consuming and often expensive programming work by hired external programmers.

Techniques like topic modelling, word embeddings, sentiment and emotion mining are increasingly being used in the humanities and social sciences. Historians, political scientists, sociologists and others now have the opportunity to use advanced text mining techniques on large datasets from their desktops. Although still mostly experimental, the potential gains now appear enormous.

It is often claimed that this enables researchers to study concepts and developments in longitudinal, systematic and quantitative ways that were impossible before. But what do these digital techniques really add to more traditional approaches? How can traditional approaches and innovative digital methodologies be paired in a meaningful and enriching manner? Does quantitative text analysis primarily provide context to existing knowledge, or is it a radical departure from what went before?

We believe that quantitative text analysis could well prove to be a dramatic, agenda-setting change. As yet, however, several problems need to be addressed. First, most of the techniques involved are less than a decade old, researchers are scattered among departments and disciplines, and there is as yet no overarching discussion about best practices, pitfalls and problems with methodology, or even a shared platform to discuss basic technical problems has been established. There is a distinct need for a better exchange of information and sharing of experience, both inside and outside the world of digital humanities.

A second problem that needs to be addressed is the slow advancement of new techniques in published research outside the narrow digital humanities world. Anecdotal evidence suggests that leading journals in the humanities, political and social sciences are not particularly keen on papers using text-mining methodologies. This unwillingness is at least in part inspired by the problem mentioned above. There are few established norms to evaluate the validity of new techniques. On the other hand, conservatism may also play a role.

A third problem, which also impacts publication opportunities, is that the bulk of publications sing text-mining techniques are still primarily *about* text mining. The corpora used, and the research questions asked, in many cases still seem peripheral to technological glitz. It is of course useful to investigate the technical opportunities that new techniques have to offer, but for the wider dissemination of these techniques it will probably prove necessary to tackle existing research problems in various fields and show that this particular field of the digital humanities has something to offer to the study of history.

We propose to discuss these problems with a mixed panel of experienced text mining researchers from different (sub-) disciplines. Our central goal is to discuss practices for validation of techniques and methodologies. We want to come up with a proposal for integrating text mining techniques in historical research practice in a meaningful, substantive, and contributive way, and pave the way for the move of text mining into common research practice, beyond the current hype.