Stylometry with R

A 3-Day Workshop by Maciej Eder and Jan Rybicki at the Göttingen Centre for Digital Humanities

Overall aims and significance of the workshop

Stylometry, or the study of measurable features of (literary) style, such as sentence length, vocabulary richness and various frequencies (of words, word lengths, word forms, etc.), has been practised since at least the middle of the 19th century, and has found numerous practical applications not only in authorship attribution research but also in other research areas. Patterns of stylometric similarity and difference also provide new insights into relationships between different books by the same author, between books by different authors, between authors differing in terms of chronology or gender, or between translations of the same author or group of authors. All of this helps, in turn, to find new ways of looking at works that seem to have been studied from all possible perspectives.

Stylometry belongs to the field of quantitative text analysis, which has been identified by the EADH (formerly the ALLC) as one of a number of key digital methods and is seen as an important component in the establishment of Digital Humanities (DH) at the Georg---August---University Göttingen. The Göttingen Centre for Digital Humanities (GCDH), operational at the University since 2011, has already made significant progress in this regard and is now looking to improve the local DH capabilities by organising workshops focusing on key areas with the help of external experts. The "Stylometry and R" workshop would make a vital contribution to this endeavour.

Specific goals

The workshop participants will receive a solid grounding in the theory behind, and the algorithms for, stylometric analysis as well as hands---on, practical work with real data sets. Each participant will apply the workshop's techniques to collections of texts within their own discipline and thus emerge from the workshop with a better understanding of their own data and concrete ideas on how to further apply stylometric, and more general linguistic, analytical methods within their field of expertise. They should also be able to edit existing R scripts, customizing them to their own specific needs.

Description of the workshop and its methodology

This three---day workshop, led by Maciej Eder and Jan Rybicki, is a theoretical and practical introduction to stylometric analysis in the R statistical programming environment, using the collection of Eder's and Rybicki's R scripts, which perform multivariate analyses of the frequencies of the most frequent words, the most frequent word n---grams, and the most frequent letter n---grams. The scripts perform a range of supervised and unsupervised textual analyses and will, therefore, give the participants an excellent overview of analytical methods and a suite of easy---to---use tools to apply them in their own research. The Stylometry and R Workshop 2 workshop's first day will be spent learning what R can do with

texts and why this is important. The second and third days will be spent using R to analyse texts, both those provided by the instructors and those brought by the participants. The participants will be working with R on their own computers at all stages of the workshop. The proper installation of the R software package will be ensured and supported by the GCDH Team.

Workshop plan and timetable

3 days, 3 x 1,5 hour sessions/day

Day 1:

1. A shock introduction to stylometry/A shock introduction to stylometry in R
2. What R can do (keywords, wordcloud, richness).
3. Just to make sure we're all on the same page: statistics in texts. Introduction to the scripts

Day 2:

1. Playing with the multivariate analysis script: authorship attribution and beyond
2. Playing with R (unsupervised analysis, supervised machine---learning analysis)
3. Playing with R (rolling Delta, Zeta)

Day 3:

1. TRanslation
2. ChRonology? GenRe? GendeR?
3. Shifting to Tags