Stylometry Experiments

Szakács Béla Benedek[1], Mészáros Tamás[1]

[1] *Budapest University of Technology and Economics, Faculty of Electrical Engineering and Informatics (VIK), H-*1111 Budapest*,* Hungary

Stylometry is a discipline that has been existed for a very long time yet it only gained popularity very recently with the sudden increase of computing power. The term itself covers a wide variety of methods used in analyzing texts of natural language, although it has a potential for using it in other fields. The most prominent employment of it is authorship attribution, which is the theme of my research as well.
The main problem concerning the use of stylometry is that it comprehends fields of sciences which lies between several different research areas, namely statistics, linguistics and IT. This means that in order to utilize it, some knowledge is required in all three disciplines. That is quite a rarity, so the main goal of my project was to make the application of this method easier for researchers with less knowledge in statistics ant IT.
Stylo is an R package developed by Maciej Eder, Jan Rybicki and Mike Kestemont specificaly for stylometric analysis. Throughout the project I have expanded on an already existing software called shtylo, developed by Dobi Jan Sándor, that is a web UI and a server encompassing stylo. I set up a built-in help to facilitate its use. I also created a wizard that, besides guiding the user through a step-by-step parameter-setting, makes some initial pre-analysis. I also started experimenting with a fully automated analyzer that, utilizing a searching method called simulated annealing, tries to find the optimal parameters for the analysis through using texts written by known authors.
Together with Margit Kiss from the Institute for Literary Studies of Hungarian Academy of Sciences I collected a large corpus of Hungarian texts that will serve as the basis for a comprehensive experiment aiming to provide statistics on several types of texts from several eras to help future research.

[1] Maciej Eder, Jan Rybicki and Mike Kestemont: “Stylometry with R: A Package for Computational Text Analysis”, The R Journal Vol. 8/1, Aug. 2016
[2] The stylo project: https://github.com/computationalstylistics/stylo
[3] The shtylo project: https://github.com/dobijan/shtylo/wiki

The research has been supported by the European Union, co-financed by the European Social Fund (EFOP-3.6.2-16-2017-00013).