Metadaten-basierte Visualisierungen im Stilometrie-Paket „Stylo“

Pielström, Steffen
pielstroem@biozentrum.uni-wuerzburg.de
Julius-Maximilians-Universität Würzburg, Deutschland

Maciej, Eder
maciejeder@gmail.com
Pedagogical University of Kraków, Polen

Die Programmbiblothek *Stylo* (Eder et al. 2016) für die Programmiersprache *R* bietet ein breites Spektrum an Funktionen für die stilometrische Analyse von Textcorpora, darunter Clusteranalyse auf der Basis von Wort- und NGramm-Frequenzen, Textklassifikation und die Identifikation distinktiver Merkmale für eine bestimmte Textgruppe. Die Funktionen nehmen dabei ganze Ordner nicht vorverarbeiteter Textdateien und geben umfangreiche Analyseergebnisse zurück, in den meisten Fällen inklusive fertiger Visualisierungen. Zusätzlich können viele der wichtigen *High-Level*-Funktionen auch über ein graphisches *Userinterface* bedient werden, womit die Basisfunktionalitäten von *Stylo*, obwohl es sich um ein Programmbibliothek handelt, auch ohne Programmierkenntnisse genutzt werden können. Nicht zuletzt bedingt durch den Komfort und die Einsteigerfreundlichkeit dieser Zugänge ist die *Stylo*-Bibliothek eines der populärsten Werkzeuge für die stilometrische Forschung in den Digital Humanities.

Dabei ist *Stylo* ursprünglich aus einer Sammlung von Skripten und Funktionen entstanden, die von den Entwicklern selbst für ihre Forschung gebraucht wurden. Die schrittweise Weiterentwicklung und Funktionserweiterung spiegelt in vielen Fällen die Bedürfnisse und Forschungsinteressen des Entwicklerteams wieder, und auch die Art und Weise, wie bestimmte Probleme in *Stylo* gelöst werden, ist nicht zuletzt durch die Arbeitsgewohnheiten der Entwickler bestimmt.

Ein Aspekt, der immer wieder zu Nachfragen von Usern geführt hat, ist der Umgang mit Metadaten in der durch die Community wohl am häufigsten genutzte *High-Level*-Funktion *stylo()*. Diese Funktion nimmt ein Corpus in Form eines Ordners mit Textdateien und erzeugt daraus wahlweise eine Clusteranalyse in Form eines Baumdiagramms, oder eine Hauptkomponentenanalyse, dargestellt als *Scatterplot*, um die Ähnlichkeitsbeziehungen der Texte untereinander darzustellen. Texte, die aufgrund von Vorwissen einer bestimmten Gruppe zugeordnet werden, erscheinen in der Visualisierung in der gleichen Farbe. So werden zum Beispiel bei einem klassischen Autorenschaftsproblem alle Texte, von denen vorher bekannt ist, daß sie von der gleichen Autorin/vom gleichen Autor stammen, in der gleichen Farbe dargestellt. Dadurch lässt sich in der Graphik schnell erkennen, wie gut Texte einer Gruppe tatsächlich nach stilometrischen Kriterien zusammen clustern.

Die Informationen über die Gruppenzugehörigkeit eines Textes entnimmt *Stylo* traditionell dem Dateinamen. Dafür muss jede Textdatei nach der Konvention *Gruppe_Dokument.Endung* benannt sein. Das Drama "Hamlet" von Shakespeare wird also zum Beispiel mit dem Dateinamen Shakespeare_Hamlet.txt versehen, wenn alle Stücke von Shakespeare in der gleichen Farbe erscheinen sollen.

Bislang war die systematische Benennung der Textdateien der einzige Weg, solche Information zur Gruppenzugehörigkeit an die Funktion zu übermitteln. Von Nutzerweite wurde immer wieder der Wunsch nach zusätzlichen Möglichkeiten geäußert, Metadaten zur Gruppenzugehörigkeit der Texte an die Funktion zu übergeben.

In den neueren *Stylo*-Versionen haben wir nun eine flexiblere Möglichkeit implementiert. Die Funktion *stylo()* vefügt nun über einen Parameter *metadata*, dem die Information zur Gruppierung der Texte in Form einer Gruppierungsvariable übergeben werden kann. Im einfachsten Fall ist das ein Vektor, dessen Länge der Anzahl der Texte im Corpus entspricht, und der für jeden Text ein Gruppenlabel liefert.

```
authornames <- c("Goethe", "Goethe", "Goethe", "Rodan", "Rodan", ...)
stylo(metadata = authornames)
```

Die Funktion akzeptiert sowohl Faktor als auch einen Vektor von Strings als Gruppierungsvariable. Die andere Möglichkeit ist, die Information zur Gruppenzugehörigkeit der Texte in einer CSV-Datei zu hinterlegen und dem Parameter den Dataipfad als String zu übergeben. Die betreffende CSV-Datei enthält eine Spalte mit der Überschrift "filename", die alle Dateinamen des Corpus in alphabetischer Reihenfolge enthält, und mindestens eine weitere Spalte mit Gruppenlabels. Um die Spalte mit der gewünschten Gruppierungsvariable auszuwählen wird der Titel der gewünschten Spalte an den Funktionsparamter *grouping.column* übergeben.

```
stylo(metadata = "metadata.csv", grouping.column = "author")
```

Der Default-Wert ist "author". Wenn dem Paramter *grouping.column* kein Wert zugewiesen wird, muss die Datei eine Spalte mit dem Default-Wert "author" als Überschrift enthalten.

Dieser zusätzliche Parameter in der *stylo()*-Funktion erlaubt nun flexibel mit der Gruppenzugehörigkeit der Texte zu experimentieren, ohne daß dafür die Textdateien umbenannt werden müssen. Das Poster wird diese neuen Funktionalitäten vorstellen und durch Codebeispiele und Visualisierungen erläutern.

Bibliographie

Eder, Maciej / Rybicki, Jan / Kestemont, Mike (2016): „Stylometry with R: a package for computational text analysis“, in *R Journal*, **8**(1): 107-121, url: https://journal.r-project.org/archive/2016/RJ-2016-007/index.html