Workshop „Text Mining in R“

Jan R. Riebling

Zusammenfassung

Der Workshop findet am 4.11. in München statt und ist auf ein sechs-stündiges Programm ausgelegt. Ziel des Workshops ist eine Einführung in das Text Mining mittels R. Unter Text Mining wird hier die Extraktion von strukturierten, numerischen Daten aus semi- oder unstrukturierten Textdaten sowie deren Analyse mittels quantitativen Verfahren verstanden. Im Rahmen des Workshops soll ein Überblick über die Breite der Thematik und mögliche Anwendungen gegeben werden. Material zur praktischen Übung sowie Hinweise zu weiterführenden Ressourcen werden im Rahmen der Veranstaltung bereitgestellt werden, sind aber aufgrund des zeitlichen Rahmens zum vertiefenden Selbststudium gedacht.

Thematischer Aufbau

1. Einführung in R, Text Mining und die Jupyter Notebook Arbeitsumgebung.
2. Strings, Textobjekte und Reguläre Ausdrücke.
3. Webscraping und HTML-Parsing.
4. Tokenisierung, Part-of-Speech Tagging und numerische Repräsentation von Texten.
5. Lesbarkeits, Komplexität und Modellierung sprachlicher Eigenschaften.
6. Machine Learning und semantische Analysen (Topic Models).

R installieren

Um einen möglichst schnellen Einstieg in das Thema zu finden sollte ein Laptop mit einer funktionierenden R Installation in den Workshop mitgebracht werden. Des Weiteren sollten die unten aufgeführten zusätzlichen Pakete des Tidyverse installiert werden. Zudem empfiehlt es sich das kurze R Introduction Tutorial bis zum Unterpunkt „Data Frame“ zu absolvieren, um ein Verständnis der grundlegenden Datentypen in R zu erlangen.

Basis R

Vorkompilierte R-Pakete für die gängigsten Betriebssysteme können auf der Webseite des Comprehensive R Archive Network (CRAN) heruntergeladen werden. Alternativ kann R auch über einen Paketmanager (z.B. Anaconda, Chocolatey, APT, etc.) Bei Problemen mit der Installation unter Windows finden sich hier weiterführende Informationen.

Zusätzliche Pakete

Zusätzliche R Pakete können über den in `base` enthaltenen Paketmanager direkt von einem CRAN Repositoriumsserver heruntergeladen werden. Dies geschieht durch Ausführen des Befehls `install.packages()`. Beim ersten Mal kann ein naheliegender CRAN-Server ausgewählt werden.

Alle für den Workshop notwendigen Pakete sind im Metapaket `tidyverse` enthalten. Hierbei handelt es sich um eine Sammlung von Text- und Datamining Werkzeugen die zum Teil stark von den grundlegenden R Konzepten abweichen. Ein Überblick der enthaltenen Pakete und die entsprechende Dokumentation findet sich hier. Folgende Eingabe auf der R-Eingabeaufforderung installiert die `tidyverse`-Pakete:

```
install.packages("tidyverse")
```

Integrated Development Environment

Nach der Installation von R steht normalerweise nur der R Interpreter und der Kommandozeilenzugriff darauf zur Verfügung. Um eine effizientere Nutzung und komplexeren Code zu ermöglichen können entweder auf dem System vorhandene IDEs (z.B. VSStudio) oder Editoren (z.B. Vim, EMACS) verwendet werden.

Im Workshop wird das Jupyter Notebook als grafische Benutzeroberfläche und Entwicklungsumgebung genutzt, da sich hiermit am besten eine einfache Live-Präsentation erstellen lässt. Die Unterlagen werden als Jupyter Notebook aber auch als Plaintext (Rmarkdown) und PDF Dateien zur Verfügung gestellt werden. Eine Installation von Jupyter Notebook kann auf ähnliche Arten und Weisen geschehen wie die Installation von R, z.B:

- Mittels des Python-Paketmanagers pip.
- Als Teil der Anaconda Python Distribution.

Um die Verbindung zwischen R und dem Notebook Server herzustellen, muss der IRkernel installiert werden. Dies geschieht durch Ausführen der entsprechenden Instruktionen in R:

```
install.packages("devtools")

devtools::install_github("IRkernel/IRkernel")
```

```
IRkernel::installspec()
```

Für die Teilnehmer wird wegen des leichteren Einstiegs und des besseren Anschlusses an die Community zusätzlich die Verwendung von RStudio als IDE empfohlen.

Ressourcen

Dokumentationen und Tutorials

- The R Manuals. Hier insbesondere „An Introduction to R“ und „R Installation and Administration“.
- R PROGRAMMING TUTORIAL – BEGINNERS , INTERMEDIATE AND ADVANCED.
- RStudio Dokumentation.

Beispiele für spezifische Pakete/Techniken

- R API Tutorial.
- Scraping HTML Tables and Downloading Files with R.

Frei verfügbare Bücher

- Text Mining with R: A Tidy Approach.
- R for Data Science.
- R Graphics Cookbook.
- Allgemeines zur Textanalyse: Methode und Methodologie quantitativer Textanalyse.