

Workshop Entwurf „Text Mining in R“

Jan R. Riebling

Umfang

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. Im Vorfeld des Workshops wird den Teilnehmern eine Installationsanleitung für die notwendigen Programmpakete sowie ein kurzes Tutorial zu R bereitgestellt werden.

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).