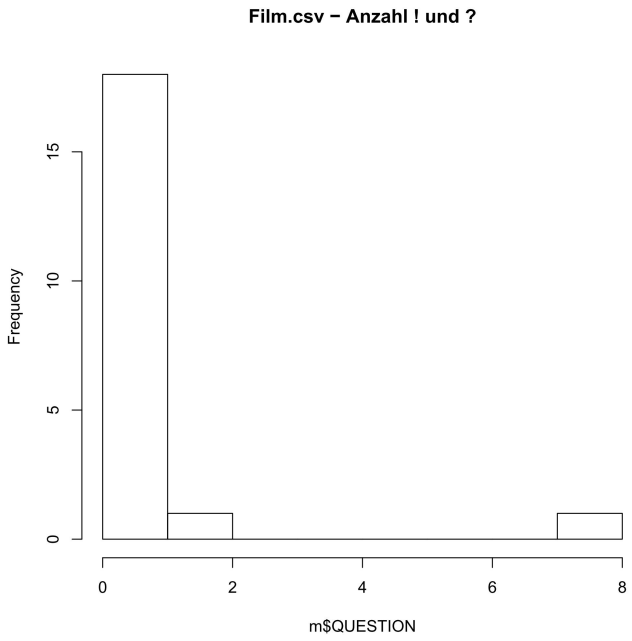
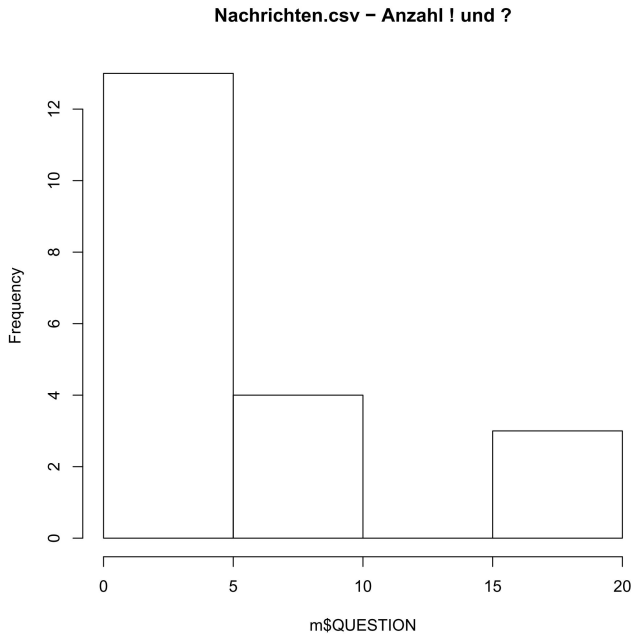
Textklassifikation

Herangehensweise:

Wir gehen nach dem Prinzip des beaufsichtigten Lernens vor. Aus Trainingsdaten extrahieren wir Merkmale, die wir an Testdaten kombinieren und zum Schluss an den Evaluationsdaten überprüfen.

- Wir verwenden als Auslese-Framework Antlr4
  Wir haben eine Grammatik definiert, die auf die Texte angewandt Kennzahlen(Merkmale) mit Hilfe der Klasse AnalysisListener ermittelt.
  Die Klasse Learning aus package learning arbeitet auf den Trainingsdaten und extrahiert die Merkmalsausprägungen in eine CSV-Datei.
- Mit einem R-Skript(drawPlots.R) haben wir für jedes Merkmal ein Histogramm, bzw. einen Boxplot, erstellt.
- Mit einem weiteren R-Skript erstellen wir eine xml-Datei mit dem "Brain". Dieses "Brain" enthält für die einzelnen Merkmale Bereiche mit zugehörigen Wahrscheinlichkeiten.
  Die Bereiche haben wir aus den Histogrammen, siehe exemplarisch unten, ermittelt. Wir haben versucht möglichst nicht zu detailiert den Kurvenverlauf abzubilden, aber Peeks zu erfassen.





(Bereiche für Film von 0-1 und 1-10000, sowie News von 0-5 und 5 bis 10000)
(Zugehörige Wahrscheinlichkeiten 0.9 und 0.1, sowie 0.55 und 0.45)

- Die Klasse Classifier lädt diese Brain.xml ein und kann dann auf die Test- und Evaluationsdaten angewendet werden. Sie addiert für jedes verwendete Merkmal den negativen Logarithmus der zugehörigen Wahrscheinlichkeit aus dem Bereich der Merkmalsausprägung.

Jetzt werden die unterschiedlichen Elemente vom Typ Classifier_Class verglichen, und der niedrigste Wert ist die passende Klasse.

Hiermit haben wir nun ein Model für die Klassifizierung der beiden Elemente vom Typ Classifier_Class, Film oder News, für die klassifiziert werden sollte.

Class Classifier:

Hier wird zuerst das Brain.xml eingeladen und deren Merkmale in die beiden Klassenattribute film und news übertragen. Danach werden alle Merkmale aller Test/Evaluations-Dateien der beiden Kategorien mit der Methode parseText extrahiert. Mit der Methode classify wird nun anhand dieser Merkmale die Kategorie der Datei ermittelt.