Seminar Statistische Lernverfahren

Klassifikation von Rezensionstypen

Till Gräfenberg, Matthias Häußler, Alexander Kohlscheen, Michael Lau,
Tanja Niklas, Jonathan Schmitz

12. Dezember 2019

Problemstellung

- Ziel: Klassifizierung von Reviews in folgende Typen:

Texttyp	introvertiert	extrovertiert
emotional	stetig	initiativ
rational	gewissenhaft	dominant

- Gegeben: 439 bereits klassifizierte Reviews

Erstellen von Prädiktoren

- Klassifikation sollte durch verwendete Wörter geschehen
- Zurückführung auf Grundwörter notwendig
- Benutzung verschiedener Packages in `R` bzw. `Python` ermöglichte verschiedene Verfahren

Erstellen von Prädiktoren

Was ist Stemming?

- Verfahren, mithilfe dessen man verschiedene Varianten eines Wortes auf ihren gemeinsamen Wortstamm zurückführt
- Durch Abschneiden von Prä-/In- und Suffixen und Ersetzen von Umlauten, Diphtongen etc. Erzeugen von Wortstämmen
- Beispiele: gelernt $\rightarrow$ lernen; Wohnungen $\rightarrow$ Wohnung

Erstellen von Prädiktoren

Stemming

- Eigene Implementierung nach Vorgabe von COMPEON in `R`
- Für Englische Sprache bereits vorgefertigte Tools z.B.
  - `porterstemmer` von `nltk` in Python
  - `snowballstemmer` von `nltk` in Python

Erstellen von Prädiktoren

Was ist Lemmatisierung?

- Das Lemma ist im Bereich der Linguistik die Grundform eines Wortes
  $\rightarrow$ Wortform z.B. in einem Nachschlagewerk
- Zurückführung auf grammatikalische Grundformen
- Lemmatisierung als ein lexikonbasiertes Stemmingverfahren
- Auftretende Probleme des Vorgangs:
  - Ambiguitäten
  - Wahl des Lemmas eines Wortes (Verbinfinitiv vs. Nomen)
  - Kompositazerlegung nicht eindeutig (Beispiel: Wachstube)
  - Simplizia (Beispiel: Kreuzer, Tangente)
  - Unregelmäßigkeit von Verben im Deutschen

Erstellen von Prädiktoren

Lemmatisierung

- Erfordert vorgefertigte Packages z.B.
  - `SpaCy` in Python
  - `nltk` in Python
- Diese lieferten zusätzlich Informationen über die Wortart
- Auch hier für Englische Sprache ausgereifter als die deutsche Alternative

Erstellen von Prädiktoren

Wortliste

- Ausgangssituation: 439 Reviews über die Firma COMPEON
- 8792 Wörter in reviews_preprocessed (mitunter mehrfach)

cleaned_text	preprocessed_text
richtigen	richtig
darlehen	darleh
gewünschte	wunsch
taggenuae	taggenua
gegenüber	genub

Tabelle: Wortliste Beispiele