Come2Data

# TRAINING: Python für die Datenanalyse in den Sozialwissenschaften

## Teil 2: Analyse verschiedener Datenarten und Web-Ressourcen

## *Einführung zu Natural Language Processing – Begriffe und Konzepte*

**SPEAKER:** Matthias Täschner

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

SACHSEN
Diese Maßnahme wird gefördert durch die Bundesregierung aufgrund eines Beschlusses des Deutschen Bundestages.
Diese Maßnahme wird mitfinanziert durch Steuermittel auf der Grundlage des von den Abgeordneten des Sächsischen Landtags beschlossenen Haushaltes.

# INHALT UND ZIELE

- **Crashkurs** Natural Language Processing - Begriffe und Kern-Konzepte für Einsteiger
- Natural Language Processing (NLP)
  - Kombiniert Linguistik und Informatik / Data Science
  - Ermöglicht maschinelle / automatisierte Verarbeitung menschlicher Sprache
  - Nutzung in Such-Maschinen, ChatBots, Übersetzung, …
- Ziel: Unstrukturierten Text / Sprache in Strukturierte Daten überführen
- Methoden und Konzepte
  - Tokenisierung, Stopwörter, Stammformen
  - Alternative Repräsentationsformen von Text
  - Ähnlichkeitsberechnung, Themenmodellierung, Sentimentanalyse, …
- Weiterführende Quellen z.B.
  - „Speech and Language Processing “ (Stanford), https://web.stanford.edu/~jurafsky/slp3/
  - „Natural Language Processing with Python“ (NLTK), https://www.nltk.org/book/

# Unser Ausgangsdokument – unstrukturierter Text

**The dog chased the cat across the yard.**

# Tokenisierung

- Text in einzelne Einheiten (Token) zerlegen
- Oftmals mit Normalisierung, z.b. Kleinschreibung, Satzzeichen entfernen, etc.

**The dog chased the cat across the yard.**

↓

**[“The”, “dog”, “chased”, “the”, “cat”, “across”, “the”, “yard”, “.”]**

↓

**[“the”, “dog”, “chased”, “the”, “cat”, “across”, “the”, “yard”]**

# Stopwörter

- Stopwords / Stopword-Removal
- Identifikation und Entfernung von Wörtern, die wenig zur Bedeutung beitragen

**The dog chased the cat across the yard.**

↓

**[“The”, “dog”, “chased”, “the”, “cat”, “across”, “the”, “yard”, “.”]**

↓

**[“the”, “dog”, “chased”, “the”, “cat”, “across”, “the”, “yard”]**

↓

**[“dog”, “chased”, “cat”, “across”, “yard”]**

# Stammformen

- Stemming / Lemmatisierung
- Wörter in Grundform überführen

**The dog chased the cat across the yard.**

↓

**[“The”, “dog”, “chased”, “the”, “cat”, “across”, “the”, “yard”, “.”]**

↓

**[“the”, “dog”, “chased”, “the”, “cat”, “across”, “the”, “yard”]**

↓

**[“dog”, “chased”, “cat”, “across”, “yard”]**

↓

**[“dog”, “chase”, “cat”, “across”, “yard”]**

# Repräsentationsformen

- Alternative Repräsentation von Wörtern / Dokumenten
- Numerische Darstellung von Text

Doc1

**The dog chased the cat across the yard.**

Doc2

**A cat sits in the yard.**
**There is another cat.**

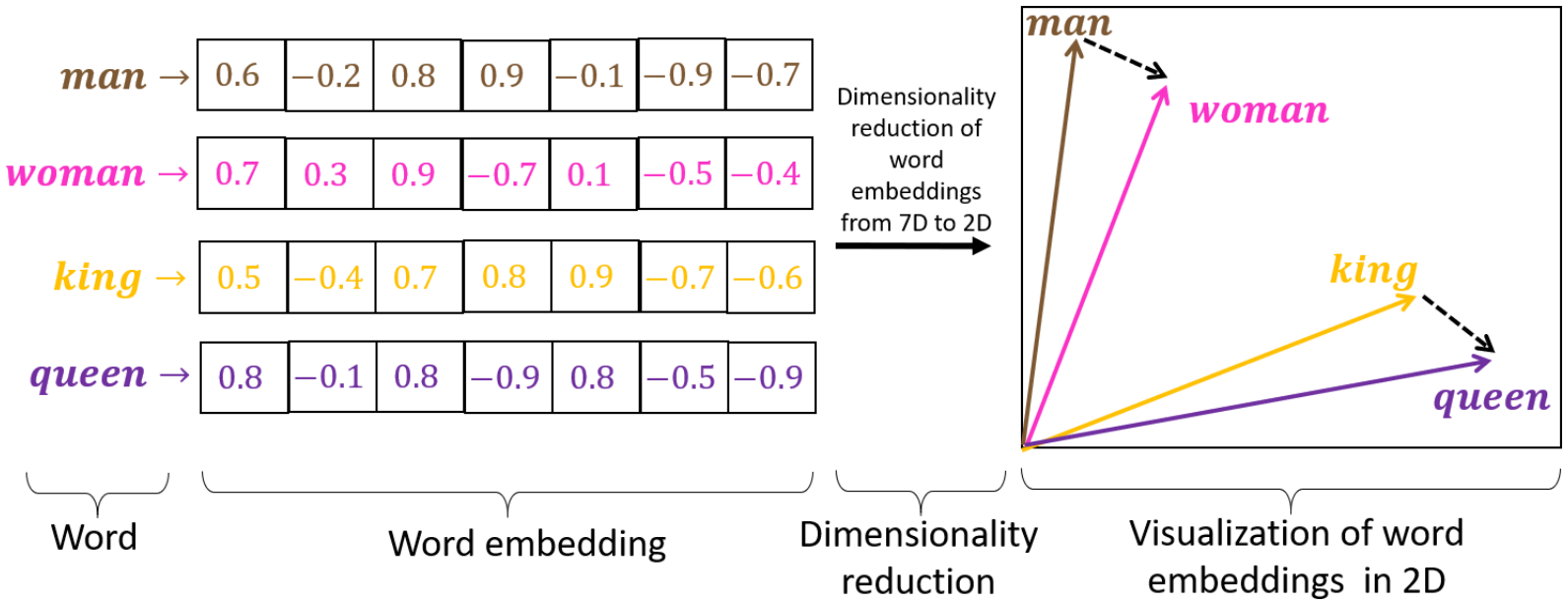
- **Bag of Words**
  - Repräsentation eines Dokuments bzgl. der Anzahl der im Korpus vorkommenden Wörter

| | dog | chase | cat | across | yard | sit | in | there | be | another | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Doc1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | → [ 1, 1, 1, 1, 1, 0, 0, 0, 0, 0 ] |
| Doc2 | 0 | 0 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | → [ 0, 0, 2, 1, 1, 1, 1, 1, 1, 1 ] |

- **TF-IDF (Term Frequency – Inverse Document Frequency)**
  - Gewichtete Bedeutung eines Worts bzgl. eines Dokuments und der Wörter aus Korpus
  - Seltene Wörter („dog“) erhalten höheres Gewicht als häufigere Wörter („cat“)
  - TF-IDF Vektor für Doc1 → [0.14, 0.14, 0, 0.14, 0, 0, 0, 0, 0, 0]
  - TF-IDF Vektor für Doc2 → [0, 0, 0, 0, 0, 0.09, 0.09, 0.09, 0.09, 0.09]

# Repräsentationsformen

- Alternative Repräsentation von Wörtern / Dokumenten
- Numerische Darstellung von Text
- **Embeddings**
  - Text wird in eine n-dimensionale Vektor-Darstellung überführt
  - Vektoren mit ähnlicher Richtung im n-dimensionalen Raum haben auch ähnliche Semantik




Rozado, David (2020). Word embeddings map words in a corpus of text to vector space. PLOS ONE. Figure. https://doi.org/10.1371/journal.pone.0231189.g008 (CC BY 4.0)