

---

# Anfrage-Sprachen

## Teil 3

Diese Folien wurden mit freundlicher Genehmigung von Peter Becker übernommen (<http://www2.inf.fh-rhein-sieg.de/~pbecke2m/retrieval/>).

# Das Soundex-Verfahren

---

**Problem:** Suche nach phonetisch ähnlichen Wörtern.

- Wörter werden hierzu auf interne Codes abgebildet.
- Phonetisch ähnliche Wörter sollen auf möglichst gleiche Codes abgebildet werden.
- Das bekannteste Verfahren ist der SOUNDEX-Algorithmus.

# Das Soundex-Verfahren

---

## SOUNDEX-Algorithmus

Der SOUNDEX-Algorithmus besteht aus den folgenden Schritten:

1. Nimm den ersten Buchstaben des Wortes und transformiere die restlichen Buchstaben (unabhängig von Groß- und Kleinschreibung) nach der folgenden Tabelle.

a e i o u h w y	→ 0
b f p v	→ 1
c g j k q s x z	→ 2
d t	→ 3
l	→ 4
m n	→ 5
r	→ 6

2. Streiche alle Nullen.
3. Reduziere alle hintereinander vorkommenden gleichen Zahlen auf eine Zahl.
4. Beschränke den ganzen Code auf maximal vier Stellen.

# Beispiel

---

Die Namen *Neumann* und *Newman* werden als gleich erkannt:

- Neumann  $\rightarrow$  N005055  $\rightarrow$  N555  $\rightarrow$  N5
- Newman  $\rightarrow$  N00505  $\rightarrow$  N55  $\rightarrow$  N5

# Bemerkungen

---

- Der SOUNDEX-Algorithmus dient insbesondere der Suche nach Namen.
- Der Code ist nicht immer erfolgreich, denn ähnlich geschriebene Wörter werden oft auf unterschiedliche Codes abgebildet:
  - Beispiel: Rodgers und Rogers.
  - Rodgers  $\rightarrow$  R032062  $\rightarrow$  R3262  $\rightarrow$  R326
  - Rogers  $\rightarrow$  R02062  $\rightarrow$  R2262  $\rightarrow$  R262
- In der Praxis wird der SOUNDEX-Algorithmus z.T. in leicht abgeänderter Form eingesetzt.