# Informationsintegration

**Übung 5**

Gruppe 3 - Dettmer, Schwaß

17.7.2016

Programmaufruf mit `./match.sh`. Die trivialen Korrespondenzen werden quasi sofort ausgegeben, die restlichen brauchen u. U. ein paar Minuten.

## Vorgehen

### Daten vorbereiten

Die Daten in den `.owl`-Dateien wurden in folgende Form überführt:

```
...
MA_0000020|  back
MA_0000021|  abdomen pelvis perineum  |  lower body
...
```

Je Zeile werden also die Klasse, das Label und mögliche Synonyme angeführt.

Die Beschreibungen wurden dabei normalisiert:

- Nur Kleinbuchstaben
- Sonderzeichen entfernen (`- _ . , /`)
- Einige Stopwords entfernen (*and*, *at*, *for*, *of*, *or*, *the*, *to*)
- Unnötige Leerzeichen entfernen
- Am Anfang und am Ende jeweils zwei Leerzeichen setzen (wird später Relevant)

### Triviale Zuordnungen

Beim Einlesen dieser Dateien wird jeweils ein Mapping von Label auf Id und Synonyme erstellt. Da wir die Daten anhand der Label verknüpfen wollen, wird dieses als Schlüssel verwendet.

```
Label -> (Id, [Label])
```

Aus dem Schnitt der Schlüsselmengen dieser Mappings ergeben sich sofort die trivialen Zuordnungen, welche nach der Normalisierung völlig übereinstimmen.

### Restliche Zuordnungen

Für die restlichen Klassen wird eine Matrix mit der jeweiligen Ähnlichkeit erstellt. Die Ähnlichkeit zweier Klassen wird anhand der 3-Gramm-Überschneidung zwischen allen Bezeichnungen (Label und Synonyme) berechnet, wobei nur das Maximum weiterverwendet wird. Durch das Einfügen

von zusätzlichen Leerzeichen am Anfang der Begriffe stimmt das erste 3-Gramm nur mit einem anderen Begriffsanfang überein (analog für das letzte).

```
nGram 3 "  bar  "  = ["  b", " ba", "bar", "ar ", "r  "]
nGram 3 "  baz  "  = ["  b", " ba", "baz", "az ", "z  "]
sim bar baz = 2 / 8 = 0,25
```

Anschließend wird für jede Zelle der Matrix die Differenz zum Durchschnitt (Zeile und Spalte) berechnet. Anhand dieser Einträge werden in einem einfachen Greedy-Verfahren neue Zuordnungen gewählt. Klassen können dabei nur einmal an einer Zuordnung beteiligt sein. Stehen keine Zuordnungen mit ausreichender Ähnlichkeit ($> 0{,}53$) mehr zur Verfügung wird das Verfahren beendet.

## Ergebnis

| | |
|---|---|
| True positives | 1282 |
| False positives | 91 |
| False negatives | 234 |
| Precision | .9337 |
| Recall | .8456 |
| Recall+ | .5895 |
| F-Score | .8875 |