Übung 3: Log-Likelihood-Ratio

Rechenübung

Gegeben sei die folgende Kontingenztabelle:

	München	~~München~~	
Bayern	23	?	211
~~Bayern~~	?	?	?
	315	?	15700

Berechnen Sie die fehlenden beobachteten Werte (?) in der Tabelle.

Berechnen Sie dann die erwarteten Werte:

$$E_{ij} = \frac{O_{i-}O_{-j}}{O_{--}}$$

Berechnen Sie schließlich den Loglikelihood Ratio:

$$LLR = 2\sum_{ij} O_{ij} \log_2 \frac{O_{ij}}{E_{ij}}$$

Implementierung

Laden Sie an der Adresse `www.cis.uni-muenchen.de/~schmid/lehre/StatNLP/data/word-pairs.txt.gz` eine komprimierte Datei mit Adjektiv-Nomen-Paaren und ihren Häufigkeiten herunter und dekomprimieren Sie die Datei. Jede Zeile der Datei enthält ein Wortpaar mit seiner Häufigkeit. Leerzeichen dienen als Trennzeichen. Beispiel:

1 Aachener Agentur
1 Aachener Altbauwohnung
1 Aachener Amt

Schreiben Sie ein Programm, welches die Wortpaare aus der Datei einliest und für jedes Wortpaar den **Log-Likelihood-Ratio** (LLR) berechnet. Geben Sie dann die Wortpaare und ihren LLR absteigend nach LLR sortiert aus.

Was ist zu tun, wenn ein O_{ij} in der Formel den Wert 0 besitzt?