Log-likelihood-ratio

Um zu ermitteln, ob die Frequenz k_1 einer Wortform *w* in einem Analysekorpus der Länge n1 statistisch signifikant über der Frequenz k_2 liegt, die aufgrund seiner Frequenz im Referenzkorpus der Länge n_2 zu erwarten gewesen wäre, wird bei der log-likelihood-ratio die relative Häufigkeit p_1 der Wortform *w* im Referenzkorpus mit ihrer relativen Häufigkeit p_2 im Analysekorpus verglichen. Als sog. *Nullhypothese* wird dabei angenommen, dass zwischen beiden kein Unterschied besteht, dass also $p_1 = p_2$.

Beim Vergleich zwischen dem Analyse- und Referenzkorpus wird zunächst für jede Wortform *w* ihre relative Häufigkeit im Analysekorpus und Referenzkorpus ermittelt:

$$p_1 = k_1/n_1 \text{ und } p_2 = k_2/n_2$$

Eine angemessene Modellierung der Wahrscheinlichkeit, dass *w* genau k-mal in einem Text der Länge n auftritt, ist nach Dunning (1993) die *Binomialverteilung*, d.h.

$$p(n, k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

Binomialverteilung von Wortformen

Für die Abschätzung der Wahrscheinlichkeit, dass eine Wortform w genau k-mal in einem Text der Länge n auftritt, nehmen wir an, dass wir jede der n Wortformen betrachten und entscheiden, ob die betrachtete Wortform x identisch ist mit w, also x= w, oder nicht. Da der Text insgesamt die Länge n hat, gibt es n solcher Entscheidungen, wovon k positiv sind, n-k negativ. Die Wahrscheinlichkeit, das w genau k-mal auftritt, ist p^k , die Wahrscheinlichkeit des gegenteiligen Ereignisses entsprechend $(1-p)^{n-k}$. Da jede der n Entscheidungen positiv sein kann, muss das Ergebnis noch multipliziert werden mit der Anzahl von Möglichkeiten, dass bei n Entscheidungen k positive enthalten sind $\binom{n}{k}$.

Wenn wir die Verteilung von Wortformen in einem Text mit der Binomialverteilung modellieren, setzen wir voraus, dass (1) jede Entscheidung *unabhängig* ist von allen vorangehenden und dass (2) das Wort *w* an jeder Stelle des Textes *gleich* wahrscheinlich ist. Im Hinblick darauf, dass die Wahrscheinlichkeit des Auftretens einer Wortform stark davon abhängt, welches Thema behandelt wird (vgl. LINK Topic Models) und dass zwischen Wortformen oft Abhängigkeiten bestehen, ist dies nicht ganz korrekt, kann jedoch für unsere Zwecke vernachlässigt werden.

Die *likelihood-Funktion* H ermöglicht einen Vergleich des Analyse- mit dem Referenzkorpus: H beschreibt die Wahrscheinlichkeit, dass ein Wort w in einem Analysekorpus der Länge n_1 k_1-mal und im Referenzkorpus der Länge n_2 k_2-mal gesehen wird unter der Voraussetzung, dass die Auftretenswahrscheinlichkeit im Analysekorpus durch p_1 und die im Referenzkorpus durch p_2 gegeben ist.

Die Nullhypothese lautet: $p_1 = p_2 = p$, d.h. w ist in beiden Texten gleichwahrscheinlich.

Die *likelihood-ratio* λ ist nun der Quotient zweier Maxima: dem durch die Nullhypothese gegebenen maximalen Wert der likelihood-Funktion H auf dem Teilraum Ω_0 geteilt durch das Maximum von H auf dem gesamten Ereignisraum Ω, also

$$\lambda = \max_{\omega \in \Omega_0} H(\omega, k) \,/\, \max_{\omega \in \Omega} H(\omega, k)$$

Die Maxima werden erreicht durch die maximum likelihood estimates $p_1 = k_1/n_1$, $p_2 = k_2/n_2$ und $p = (k_1 + k_2)/(n_1 + n_2)$.

Nach Einsetzen und Umformen erhält man die Prüfgröße $-2 \log \lambda$ (mit $L(p, k, n) = p^k(1-p)^{n-k}$):

$$\lambda = 2 [\log L(p_1, k_1, n_1) + \log (p_2, k_2, n_2) - \log L(p, k_1, n_1) - \log (p, k_2, n_2)]$$

Die Differenzanalyse besteht also darin, alle Wortformen, für die $-2 \log \lambda$ groß genug ist (und die mit einer gewissen Mindestfrequenz auftreten), herauszufiltern.

tf/idf (term frequency / inverse document frequency)

Das Maß *Term-Frequenz/Inverse-Dokument-Frequenz (tf/idf)* wurde von George Salton für die Indexierung von Wortformen im Information Retrieval eingeführt (1973). Es berücksichtigt folgenden naheliegenden Aspekt der Verteilung von Wortformen in einer Dokumentkollektion, der auf Karen Sparck-Jones zurückgeht (1972): Wortformen, die nur in wenigen Dokumenten, dort aber häufig vorkommen, sind spezifischer für diese Dokumente, als Wortformen, die insgesamt häufiger, dafür aber in sehr vielen Dokumenten vorkommen.

Die Termfrequenz f_{ik} mißt dabei, wie häufig eine Wortform k im Dokument i vorkommt, während die inverse Dokumentfrequenz IDF_k einer Wortform k berechnet wird als:

$$IDF_k = \log N/d_k + 1$$

N ist dabei die Gesamtzahl der Dokumente in der Kollektion und d_k die Anzahl der Dokumente, in denen die Wortform k auftritt. Die IDF wird also groß, wenn die Wortform k in wenigen Dokumenten auftritt. Salton verwendet das Produkt $w_{ij} = f_{ik} * IDF_k$ als Gewicht für die Spezifität einer Wortform k im Dokument i.

Eine Differenzanalyse von Dokumentkollektionen mit Hilfe des Maßes tf/IDF ermöglicht es, statistisch signifikante Unterschiede in der Verwendung von Wortformen zu ermitteln, indem solche Wortformen ausgewählt werden, die im Analysekorpus häufig, sonst aber selten auftreten. Dabei sichert eine hohe Termfrequenz im Analysekorpus die Repräsentativität einer Wortform für das Analysekorpus, die inverse Dokumentfrequenz bemisst die statistische Signifikanz in der unterschiedlichen Verwendung dieser Wortformen in Bezug auf das Referenzkorpus, da Wortformen bevorzugt werden, die im Referenzkorpus selten sind — so wie bei der IDF.

In der Praxis ist dieses Verfahren vor allem dann relevant, wenn die Texte im Analysekorpus meist sehr kurz sind, so wie beispielsweise Kurznachrichten in sozialen Medien. Wird als Referenzkorpus dann ein allgemeiner Wortschatz wie z.B. die Daten des Projekts Deutscher Wortschatz gewählt, so können auch für sehr kurze Texte statistisch auffällige Schlüsselwörter extrahiert werden.