2.3 Erwartungswert und Korpusperplexität

Eine Funktion $f: \mathcal{Y} \to \mathbb{R}$ wird *verallgemeinerte Häufigkeitsverteilung* genannt. Zu der Menge der verallgemeinerten Häufigkeitsverteilungen:

$$\mathcal{F}(\mathcal{Y}) = \{f: \mathcal{Y} \to \mathbb{R}\}$$

gehören insbesondere die natürlichen Häufigkeitsverteilungen. Im Gegensatz zu natürlichen Häufigkeitsverteilungen kann eine verallgemeinerte Häufigkeitsverteilung zum einen für unendlich viele $y \in \mathcal{Y}$ einen Wert $f(y) \neq 0$, und zum anderen sogar negative oder nicht-ganzzahlige Werte annehmen.

Sei $p(.)$ eine Wahrscheinlichkeitsverteilung. Der Wert:

$$p[f] = \sum_{y \in \mathcal{Y}} p(y) \cdot f(y)$$

wird *Erwartungswert von* f *bezüglich* p genannt.

Aus mathematischer Sicht stecken in dieser Definition zwei Schwierigkeiten. Erstens kann die Reihe $\sum_{y \in \mathcal{Y}} p(y) \cdot f(y)$ aus abzählbar unendlich vielen Termen bestehen, sodass man fordern sollte, dass die Reihe konvergiert. Zweitens muss das Summationsergebnis von der Reihenfolge der Summation unabhängig sein. Beide Schwierigkeiten führen dazu, dass definiert wird:

$$\begin{aligned} p[f] \text{ existiert} \quad &:\Leftrightarrow \sum_{y \in \mathcal{Y}} p(y) \cdot f(y) \text{ ist absolut konvergent} \\ &\Leftrightarrow \sum_{y \in \mathcal{Y}} p(y) \cdot |f(y)| < \infty \ . \end{aligned}$$

Die wichtigste Eigenschaft einer *Erwartung* $p[.]$ ist deren *Linearität*: für je zwei reelle Zahlen $\alpha, \beta \in \mathbb{R}$ und für je zwei verallgemeinerte Häufigkeitsverteilungen $f, g \in \mathcal{F}(\mathcal{Y})$, deren Erwartungswerte $p[f]$ und $p[g]$ existieren, existiert auch der Erwartungswert der verallgemeinerten Häufigkeitsverteilung:

$$\alpha f + \beta g: \mathcal{Y} \to \mathbb{R}, \quad y \mapsto \alpha \cdot f(y) + \beta \cdot g(y)$$

und es gilt:

$$p[\ \alpha f + \beta g\ ] = \alpha \cdot p[\ f\ ] + \beta \cdot p[\ g\ ] \ .$$

Diese Eigenschaft ist einfach zu beweisen und wird oft verwendet werden.

Leider wird in den Lehrbüchern zur Stochastik nie die schöne *zweite Linearitätseigenschaft* der Erwartung erwähnt: für je zwei reelle Zahlen $\alpha, \beta \geq 0$ mit $\alpha + \beta = 1$ und für je zwei Wahrscheinlichkeitsverteilungen $p(.)$ und $q(.)$, sowie für eine beliebige verallgemeinerte

Häufigkeitsverteilung $f \in \mathcal{F}(\mathcal{Y})$, deren Erwartungswerte $p[f]$ und $q[f]$ existieren, existiert auch der Erwartungswert $(\alpha p + \beta q)[f]$ und es gilt:

$$(\alpha p + \beta q)[f] = \alpha \cdot p[f] + \beta \cdot q[f] \ .$$

Natürlich ist auch die zweite Linearitätseigenschaft einfach zu beweisen, wenn man nur beachtet, dass

$$\alpha p + \beta q: \mathcal{Y} \to [0,1], \quad y \mapsto \alpha \cdot p(y) + \beta \cdot q(y)$$

wegen $\alpha + \beta = 1$, eine Wahrscheinlichkeitsverteilung ist.

Die kleine Auflistung formaler Eigenschaften von Erwartungswerten soll mit der Bemerkung abgeschlossen werden, dass sich Erwartungswerte merkwürdig symmetrisch verhalten, wenn die verallgemeinerte Häufigkeitsverteilung eine Wahrscheinlichkeitsverteilung ist:

$$p[\ q\ ] = q[\ p\ ] \ .$$

Der Beweis ist trivial.

Erwartungswerte sind wichtige Grössen in der Stochastik, wie man sich leicht überzeugen kann, indem man den zentralen Begriff des ersten Abschnitts, d.h. den Begriff der Korpuswahrscheinlichkeit, etwas genauer untersucht:

$$p(f) = \prod_{y \in \mathcal{Y}} p(y)^{f(y)} \ .$$

Nach Logarithmierung (mit einer beliebigen logarithmischen Basis log-base $\in \{2, e, 10, \ldots\}$) wird aus dem Produkt eine Summe:

$$\begin{aligned} \log\ p(f) &= \sum_{y \in \mathcal{Y}} \log\ p(y)^{f(y)} \\ &= \sum_{y \in \mathcal{Y}} f(y) \cdot \log\ p(y) \ . \end{aligned}$$

Offensichtlich ist die Umformung formal richtig, doch tritt an dieser Stelle ein kleines Problem auf: wenn die Wahrscheinlichkeit $p(y)$ gleich 0 ist, ist der Logarithmus $\log\ p(y)$ gleich $-\infty$. Ist gleichzeitig die Häufigkeit $f(y)$ gleich 0, so ist zu klären, welchen Wert das Produkt $f(y) \cdot \log\ p(y)$, also $0 \cdot (-\infty)$ annehmen soll. Ein Blick auf die Korpuswahrscheinlichkeit $p(f)$ zeigt, dass das problematische y mit dem Wahrscheinlichkeitsfaktor $p(y)^{f(y)} = 0^0 = 1$ in die Korpuswahrscheinlichkeit eingeht. Derartige Datentypen werden also bei der Berechnung der Korpuswahrscheinlichkeit ganz einfach ignoriert. Dies spricht natürlich dafür, dass

$$0 \cdot (-\infty) := 0$$

definiert wird, sodass problematische Datentypen auch bei der Berechnung der logarithmierten Korpuswahrscheinlichkeit ignoriert werden. Eine mathematische Rechtfertigung dieser Definition ergibt sich aus dem Fakt, dass die Funktion $x \cdot \log\ x$ für $x \to 0$ gegen 0 konvergiert.

Nach Division mit der absoluten Häufigkeit erhält man die sogenannte *Average-Log-Likelihood*, die auch kurz *Log-Likelihood* genannt wird:

$$L(p; \tilde{p}) := |f|^{-1} \cdot \log\ p(f)\ .$$

Es folgt:

$$\begin{aligned} L(p;\tilde{p}) &= |f|^{-1} \cdot \sum_{y \in \mathcal{Y}} f(y) \cdot \log\ p(y) \\ &= \sum_{y \in \mathcal{Y}} |f|^{-1} \cdot f(y) \cdot \log\ p(y) \\ &= \sum_{y \in \mathcal{Y}} \tilde{p}(y) \cdot \log\ p(y) \\ &= \tilde{p}[\log\ p]\ . \end{aligned}$$

Wegen der durchgeführten Division mit der absoluten Häufigkeit $|f|$ ist die Log-Likelihood für zwei Korpora identisch, wenn diese dieselbe empirische Wahrscheinlichkeitsverteilung besitzen. Diese Eigenschaft ist an sich relativ bedeutungslos, hat aber doch für Korpora Konsequenzen, in denen dieselben Datentypen mit denselben empirischen Wahrscheinlichkeiten, jedoch mit unterschiedlichen Häufigkeiten vorkommen: Kommt zum Beispiel jedes Datentoken aus einem Korpus exakt x-mal, z.B. tausend- oder zehntausend-mal, in einem zweiten Korpus vor, so unterscheiden sich zwar die Korpuswahrscheinlichkeiten beider Korpora, doch die Log-Likelihood ist für beide Korpora identisch. Die Log-Likelihood sollte daher immer entweder auf der Grundlage des kleineren Korpus $f(.)$ oder der empirischen Wahrscheinlichkeitsverteilung $\tilde{p}(.)$ berechnet werden. Die Korpuswahrscheinlichkeit hingegen sollte auf der Basis der Log-Likelihood (oder einer äquivalenten Grösse), sowie der absoluten Häufigkeit berechnet werden.

Wie man sich leicht überlegt, ist der negative Erwartungswert $-\tilde{p}[\log\ p]$ eine nicht-negative reelle Zahl aus dem Intervall $[0, \infty]$. Sie wird *Cross-Entropie* genannt:

$$\begin{aligned} H_{\text{cross}}(p;\tilde{p}) &:= -\sum_{y \in \mathcal{Y}} \tilde{p}(y) \cdot \log\ p(y) \\ &= -\tilde{p}[\log\ p]\ . \end{aligned}$$

Es gilt natürlich einfach:

$$H_{\text{cross}}(p;\tilde{p}) = -L(p;\tilde{p})\ ,$$

sodass Cross-Entropie und Average-Log-Likelihood alternativ zu gebrauchende Begriffe sind, je nachdem, ob man negative oder positive Zahlen bevorzugt. Nach Umkehrung von Division und Logarithmierung erhält man aus der Log-Likelihood wieder die Korpuswahrscheinlichkeit. Es folgt:

$$\begin{aligned} p(f) &= \text{log-base}^{\,|f| \cdot L(p;\tilde{p})} \\ &= \text{log-base}^{\,|f| \cdot \tilde{p}[\log\ p]} \\ &= \text{log-base}^{\,-\,|f| \cdot H_{\text{cross}}(p;\tilde{p})}\ . \end{aligned}$$

Rein formal gesehen, lässt sich die Korpuswahrscheinlichkeit also durch einen Erwartungswert bezüglich der empirischen Wahrscheinlichkeitsverteilung berechnen, was die Bedeutung des Begriffs des Erwartungswertes an sich, der Log-Likelihood oder der Cross-Entropie im besonderen, sowie der empirischen Wahrscheinlichkeitsverteilung in das rechte Licht rückt.

Besonders erfreulich an dieser Formel ist, dass die Cross-Entropie und die Log-Likelihood sofern sie ihrer Natur nach überhaupt von einem Korpus unabhängig sein können, lediglich von der empirischen Wahrscheinlichkeiten der Datentypen im Korpus, nicht aber von deren absoluten Häufigkeiten oder der absoluten Häufigkeit aller Datentypen im Korpus abhängen. Deshalb ist es besonders störend, dass die Cross-Entropie und die Log-Likelihood von einem recht arbiträren Wert, wie der logarithmischen Basis log-base , abhängig sind. Durch die Einführung der sogenannten *Korpusperplexität* kann erreicht werden, dass die Korpuswahrscheinlichkeit auf eine noch natürlichere Grösse zurückgeführt wird:

$$\text{perplexity}(p;\tilde{p}) := \text{log-base}^{\,H_{\text{cross}}(p;\tilde{p})}\ .$$

In der Regel wird die Korpusperplexität für die logarithmische Basis log-base $= 2$ definiert; man kann sich allerdings leicht davon überzeugen, dass die Korpusperplexität von der speziellen Wahl der logarithmischen Basis log-base unabhängig ist: etwas später wird gezeigt werden, dass sich die Korpusperplexität auf die zwei einzelnen Werte der Korpuswahrscheinlichkeit und der Korpuslänge zurückführen lässt. Da diese beiden Werte von der speziellen Wahl der logarithmischen Basis unabhängig sind, gilt dies auch für die Korpusperplexität.

Andererseits ist die Korpusperplexität allein durch die logarithmische Basis log-base und die Cross-Entropie definiert. Somit ist die Korpusperplexität auch von der Korpuslänge unabhängig.

Beide Eigenschaften zusammen (Unabhängigkeit von der Korpuslänge und Unabhängigkeit von der logarithmischer Basis) machen die Korpusperplexität zu einer besseren Messgrösse für die Korpuswahrscheinlichkeit als die Log-Likelihood oder die Cross-Entropie, die zwar beide von der Korpuslänge nicht aber von der logarithmischen Basis unabhängig sind.

Paradoxerweise ist die Korpusperplexität in einem gewissen Sinn sogar ein besserer Wert als die Korpuswahrscheinlichkeit selbst: Da die Korpuswahrscheinlichkeit von der Korpuslänge abhängig ist, ist sie im Gegensatz zur Korpusperplexität leider nicht zum Vergleich von Korpora unterschiedlicher Länge geeignet (dies gilt sogar auch, wenn diese dieselbe empirische Wahrscheinlichkeitsverteilung aufweisen).

Die Korpusperplexität ist eine Funktion der Wahrscheinlichkeitsverteilung $p(.)$ der Datentypen im Korpus, sowie des Korpus $\tilde{p}(.)$. Daher kann die Korpusperplexität auf zwei Weisen interpretiert werden:

Die plausiblere Interpretation der Korpusperplexität ist: die durchschnittliche Anzahl (oder besser die erwartete Anzahl) von Datentoken, die an einer festen (aber beliebigen) Position im Korpus vorkommen werden. Hierbei wird die Wahrscheinlichkeitsverteilung $p(.)$ in schon gewohnter Weise als Generator von Datentypen angesehen. Beispielsweise bedeutet eine Korpusperplexität von 39, dass der Generator durchschnittlich 39 verschiedene Datentoken an den Positionen im Korpus erzeugen wird. Oder anders formuliert: Jedes Datentoken wurde an seiner Position im Korpus von dem Generator mit einer durchschnittlichen Wahrscheinlichkeit $\frac{1}{39}$ erzeugt. Je kleiner die Korpusperplexität ist, desto besser ist ein Generator an den vorgegebenen Korpus angepasst.

Um diese Interpretation zu verifizieren, muss lediglich der Zusammenhang zwischen Korpuswahrscheinlichkeit und Korpusperplexität untersucht werden. Offenbar gilt:

$$\begin{aligned} p(f) &= \text{log-base}^{\,-|f|\cdot H_{\text{cross}}(p;\tilde{p})} \\ &= \left(\text{log-base}^{\,H_{\text{cross}}(p;\tilde{p})}\right)^{-|f|} \\ &= \text{perplexity}(p;\tilde{p})^{-|f|} \\ &= \left(\frac{1}{\text{perplexity}(p;\tilde{p})}\right)^{|f|} . \end{aligned}$$

Wenn vorausgesetzt wird, dass jedes Datentoken im Korpus mit der Gleichverteilung

$$\pi : \mathcal{Y} \to [0,1], \quad y \mapsto |\mathcal{Y}|^{-1}$$

erzeugt wird, kann die Korpuswahrscheinlichkeit explizit berechnet werden:

$$\begin{aligned} \pi(f) &= \prod_{y\in\mathcal{Y}} \pi(y)^{f(y)} \\ &= \prod_{y\in\mathcal{Y}} \pi^{f(y)} \\ &= \pi^{\sum_{y\in\mathcal{Y}} f(y)} \\ &= \pi^{|f|} \\ &= \left(\frac{1}{|\mathcal{Y}|}\right)^{|f|} . \end{aligned}$$

Ein Vergleich der letzten beiden Resultate zeigt, dass die Korpusperplexität in dieser Situation gleich der Anzahl $|\mathcal{Y}|$ aller Datentypen ist:

$$\text{perplexity}(\pi) = |\mathcal{Y}| .$$

Natürlich kann dasselbe Ergebnis mit den Definitionen von Cross-Entropie und Korpusperplexität erzielt werden:

$$\begin{aligned} H_{\text{cross}}(\pi) &:= -\tilde{p}[\log \pi] \\ &= -\log \pi \quad (\text{denn } \pi \text{ ist konstant}), \end{aligned}$$

$$\begin{aligned} \text{perplexity}(\pi) &:= \text{log-base}^{\,H_{\text{cross}}(\pi)} \\ &= \text{log-base}^{\,-\log \pi} \\ &= \frac{1}{\pi} \\ &= |\mathcal{Y}| . \end{aligned}$$

Dieses Ergebnis korreliert sehr schön mit der oben angegebenen Interpretation der Korpusperplexität.

Im allgemeinen Fall gilt offenbar:

$$\text{perplexity}(p;\tilde{p}) = \sqrt[|f|]{\frac{1}{p(f)}} .$$

Die zweite Interpretation der Korpusperplexität betrifft den Fall, für den die Wahrscheinlichkeitsverteilung $p(.)$ als gegeben angesehen wird. Für eine fixierte Wahrscheinlichkeitsverteilung $p(.)$ schätzt die Korpusperplexität die Entropie eines beliebigen Korpus $\tilde{p}(.)$, welche ihrerseits als die Komplexität der im Korpus manifestierten Sprache interpretiert werden kann. Je grösser die Korpusperplexität ist, desto grösser ist die Korpus-Entropie.

Die *Korpus-Entropie* $H(\tilde{p})$ und die *relative Entropie* $D(\tilde{p}\|p)$, besser als *Kullback-Leibler-Distanz* bekannt, sind zwei weitere, sehr bekannte Erwartungswerte. Letztere ist definiert als:

$$D(\tilde{p}\|p) = \sum_{y\in\mathcal{Y}} \tilde{p}(y) \cdot \log \frac{\tilde{p}(y)}{p(y)} = \tilde{p}\left[\log \frac{\tilde{p}}{p}\right] ,$$

erstere als:

$$H(\tilde{p}) = -\sum_{y\in\mathcal{Y}} \tilde{p}(y) \cdot \log \tilde{p}(y) = -\tilde{p}[\log \tilde{p}] .$$

Für problematische $y \in \mathcal{Y}$ werden die folgenden Konventionen benutzt, die sich aus Plausibilitäts- und Stetigkeits-Argumenten ergeben:

$$0 \cdot \log \frac{0}{p(y)} := 0, \qquad \text{und} \qquad \tilde{p}(y) \cdot \log \frac{\tilde{p}(y)}{0} := \infty .$$

Es ist leicht einzusehen, dass die Korpus-Entropie wie die Cross-Entropie eine nicht-negative reelle Zahl ist.

Die sogenannte *Informations-Ungleichung* (siehe z.B. Cover und Thomas (1991)) ist von fundamentaler Bedeutung für die Informationstheorie und zu ihrem Beweis sind tiefe Ergebnisse nötig. Sie besagt, dass auch die Kullback-Leibler-Distanz nicht-negativ ist (und dann genau Null, wenn $\tilde{p} = p$):

$$D(\tilde{p} \| p) \geq 0 \ .$$

In Abschnitt 2.4 wird die Informationsungleichung im Rahmen der Fragestellung, welche Wahrscheinlichkeitsverteilung einen gegebenen Korpus am besten repräsentiert, bewiesen werden. Die Antwort wurde schon in Abschnitt 2.1 nahegelegt: es ist natürlich die empirische Wahrscheinlichkeitsverteilung. Die folgenden Ergebnisse sind auch in diesem Zusammenhang sehr interessant.

Unter Benutzung der bekannten Rechengesetze für den Logarithmus und der Linearität des Erwartungswertes rechnet man leicht nach, dass:

$$\begin{aligned} D(\tilde{p} \| p) &= \tilde{p} \left[\log \frac{\tilde{p}}{p} \right] \\ &= \tilde{p} \left[\log \tilde{p} - \log p \right] \\ &= \tilde{p} \left[\log \tilde{p} \right] - \tilde{p} \left[\log p \right] \\ &= -H(\tilde{p}) + H_{\text{cross}}(p; \tilde{p}) \ . \end{aligned}$$

Damit ist die folgende wichtige Aussage über das Verhältnis der drei Entropie-Grössen bewiesen: Die Cross-Entropie ist die Summe aus Korpus-Entropie und relativer Entropie:

$$H_{\text{cross}}(p; \tilde{p}) = H(\tilde{p}) + D(\tilde{p} \| p) \ .$$

Eine erste Konsequenz aus der Positivität der Kullback-Leibler-Distanz ist daher, dass die Cross-Entropie stets grösser ist als die Entropie:

$$H_{\text{cross}}(p; \tilde{p}) \geq H(\tilde{p}) \ .$$

Wenn man, analog zur Definition der Korpusperplexität zwei neue Grössen einführt:

$$\text{perplexity}_H(\tilde{p}) := \text{log-base}^{\ H(\tilde{p})} \ ,$$

sowie:

$$\text{perplexity}_D(p; \tilde{p}) := \text{log-base}^{\ D(\tilde{p} \| p)} \ ,$$

die man vielleicht *innere Korpusperplexität* und *relative Korpusperplexität* nennen könnte, so errechnet sich die Korpusperplexität elegant als das Produkt dieser zwei neuen Grössen:

$$\begin{aligned} \text{perplexity}(p; \tilde{p}) &= \text{log-base}^{\ H(\tilde{p}) + D(\tilde{p} \| p)} \\ &= \text{log-base}^{\ H(\tilde{p})} \cdot \text{log-base}^{\ D(\tilde{p} \| p)} \\ &= \text{perplexity}_H(\tilde{p}) \cdot \text{perplexity}_D(p; \tilde{p}) \ . \end{aligned}$$

Interessant ist, dass beide neu eingeführten Grössen, ebenso wie die Korpusperplexität , von der logarithmischen Basis (und natürlich von der Korpuslänge) unabhängig sind. Diese Eigenschaft haben ihre Basisgrössen Entropie und relative Entropie leider nicht. Wichtiger jedoch ist, dass sich die gute Interpretierbarkeit der Korpusperplexität , sowohl auf die innere Korpusperplexität als auch auf die relative Korpusperplexität überträgt. Hierzu dient:

$$\begin{aligned} p(f) &= \left(\frac{1}{\text{perplexity}(p; \tilde{p})} \right)^{|f|} \\ &= \left(\frac{1}{\text{perplexity}_H(\tilde{p}) \cdot \text{perplexity}_D(p; \tilde{p})} \right)^{|f|} \\ &= \left(\frac{1}{\text{perplexity}_H(\tilde{p})} \cdot \frac{1}{\text{perplexity}_D(p; \tilde{p})} \right)^{|f|} \ . \end{aligned}$$

Wenn die inverse Korpusperplexität, d.h. der Wert $\frac{1}{\text{perplexity}(p;\tilde{p})}$, als durchschnittliche Wahrscheinlichkeit interpretiert werden kann, mit der ein Prozess an einer festen (aber beliebigen) Position im Korpus ein Datentoken erzeugt, so kann, mit der eben hergeleiteten Formel für die Korpuswahrscheinlichkeit, dieser Prozess dahingehend interpretiert werden, dass er aus zwei stochastisch unabhängigen Modulen besteht:

Das erste Modul beschreibt nur den Korpus (oder wenn man so will, die Sprache, die in diesem speziellen Korpus codiert ist). Hierbei wird allein die innere Beschaffenheit des Korpus analysiert, ohne irgendwelche weitere, d.h. äussere Einflüsse zu berücksichtigen: mit der durchschnittlichen Wahrscheinlichkeit $\frac{1}{\text{perplexity}_H(\tilde{p})}$ liegt an einer bestimmten (aber beliebigen) Korpusposition ein bestimmtes Datentoken vor, wenn zur Bestimmung dieser Wahrscheinlichkeit nur der Korpus selbst, d.h. der Generator $\tilde{p}(.)$ benutzt wird. Wenn die Wahrscheinlichkeit $\frac{1}{\text{perplexity}_H(\tilde{p})}$ gross ist, dann ist der Korpus einfach, weil er "vorherberechenbar" ist. Ist die Wahrscheinlichkeit klein, dann ist der Korpus komplex, d.h. er ist "unvorherberechenbar".

Das zweite Modul beschreibt den Einfluss des Generators $p(.)$ auf den Generierungsprozess, wobei allerdings der Einfluss des ersten Moduls, d.h. die innere Beschaffenheit des Korpus , nicht mit berücksichtigt wird: ein bestimmtes Datentoken liegt dann an einer bestimmten (aber beliebigen) Korpusposition mit der durchschnittlichen Wahrscheinlichkeit $\frac{1}{\text{perplexity}_D(p;\tilde{p})}$ vor, wenn zur Bestimmung dieser Wahrscheinlichkeit zwar der Generator $p(.)$ benutzt wurde, jedoch zuvor die innere Beschaffenheit $\tilde{p}(.)$ des Korpus aus dem Generierungsprozess heraus dividiert wurde. Ist die Wahrscheinlichkeit $\frac{1}{\text{perplexity}_D(p;\tilde{p})}$ gross, dann ist der Korpus (modulo seiner inneren Beschaffenheit) einfach. Wenn die Wahrscheinlichkeit klein ist, dann ist der Korpus (in jedem Fall) komplex und unvorherberechenbar.

Ein Prozess, der stochastisch unabhängig aus diesen Einzelmodulen zusammengesetzt wird, erzeugt dann mit Wahrscheinlichkeit

$$\frac{1}{\text{perplexity}(p;\tilde{p})} \quad = \quad \frac{1}{\text{perplexity}_H(\tilde{p})} \cdot \frac{1}{\text{perplexity}_D(p;\tilde{p})}$$

an einer bestimmten (aber beliebigen) Korpusposition ein bestimmtes Datentoken.

Die hier angegebene Interpretation korreliert also sehr schön mit der oben angegebenen Formel zur Korpusperplexität.

Am Ende dieses Abschnitts sollen noch einmal alle bisher gefundenen Formeln für die Korpuswahrscheinlichkeit zusammengefasst werden:

$$\begin{aligned}
p(f) &= \prod_{y\in\mathcal{Y}} p(y)^{f(y)} \\
&= \text{log-base}^{\ |f|\cdot\tilde{p}[\log\ p]} \\
&= \text{log-base}^{\ |f|\cdot L(p;\tilde{p})} \\
&= \text{log-base}^{\ -\ |f|\cdot H_{\text{cross}}(p;\tilde{p})} \\
&= \text{log-base}^{\ -|f|\cdot(H(\tilde{p})+D(\tilde{p}\|p))} \\
&= \left(\frac{1}{\text{perplexity}(p;\tilde{p})}\right)^{|f|} \\
&= \left(\frac{1}{\text{perplexity}_H(\tilde{p})} \cdot \frac{1}{\text{perplexity}_D(p;\tilde{p})}\right)^{|f|} .
\end{aligned}$$

Für einen gegebenen Korpus $\tilde{p}(.)$ gelten daher die folgenden Gleichungen, die zwar zum Teil schon in diesem Abschnitt angesprochen wurden, auf die aber in den nächsten Abschnitten, insbesondere im Abschnitt 2.4 über Estimierungsverfahren, ausgiebig zurückgegriffen werden wird:

$$\begin{aligned}
\text{argmax}_p \quad p(f) &= \text{argmax}_p \quad \prod_{y\in\mathcal{Y}} p(y)^{f(y)} \\
&= \text{argmax}_p \quad \tilde{p}[\log\ p] \\
&= \text{argmax}_p \quad L(p;\tilde{p}) \\
&= \text{argmin}_p\ H_{\text{cross}}(p;\tilde{p}) \\
&= \text{argmin}_p\ D(\tilde{p}\|p) \\
&= \text{argmin}_p\ \text{perplexity}(p;\tilde{p}) \\
&= \text{argmin}_p\ \text{perplexity}_D(p;\tilde{p}) .
\end{aligned}$$