Bigrammwahrscheinlichkeiten

Ziel dieses Dokumentes ist es, die Berechnung von Satzwahrscheinlichkeiten durch Bigramme sowie die zu deren Verständnis benötigten stochastischen Konzepte herzuleiten. Anhand des Satzes *Studenten sind gut im Adventskranzbinden* soll dabei Schrittweise erläutert werden, wie dessen Wahrscheinlichkeit im TIGER-Korpus berechnet werden kann.

1 Bedingte Wahrscheinlichkeit

1.1 Intuition

Zunächst sollte man sich vor Augen führen, wie man Wahrscheinlichkeiten durch *relative Häufigkeiten* ausdrücken kann:
Gegeben eine *Grundgesamtheit* Ω mit $|\Omega| = N$ und ein *Ereignis* A mit $\mathrm{Fr}(A)$ Elementen, dann ist

$$P(A) = \frac{\mathrm{Fr}(A)}{N} \tag{1}$$

also die Anzahl der Elemente von A geteilt durch die Anzahl aller Elemente ($|\Omega|$).
Als Venn-Diagamm lässt sich das Ganze wie folgt darstellen:

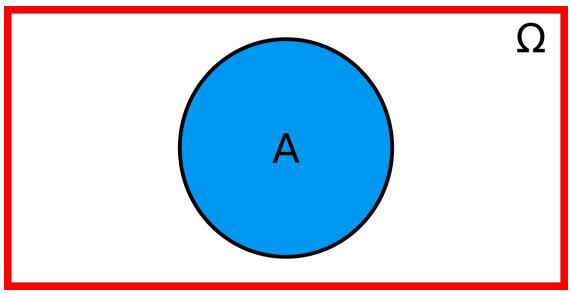

Abbildung 1: $P(A)$ als Venn-Diagramm

So ist zum Beispiel die Wahrscheinlichkeit des Wortes *Studenten*

$$P(\text{Studenten}) = \frac{\mathrm{Fr}(\text{Studenten})}{N} = \frac{71}{888238} = 7.99 \cdot 10^{-4}$$

Versucht man die Wahrscheinlichkeit des ganzen Satzes auf diese Weise zu ermitteln, wird man feststellen, dass der Satz als solches gar nicht im Korpus vorkommt. Um dieses Problem zu lösen, wird die Satzwahrscheinlichkeit auf ein Produkt von *bedingten Wahrscheinlichkeiten* reduziert, welche im Folgenden vorgestellt werden.

Die bedingte Wahrscheinlichkeit $P(B|A)$ ist die Wahrscheinlichkeit, dass das Ereignis B eintritt unter der Vorraussetzung, dass das Ereignis A eingetreten ist. Dies entspricht dem Verhältnis der Wahrscheinlichkeit, dass beide Ereignisse gemeinsam auftreten (in Zeichen: $P(A \cap B)$) zur Wahrscheinlichkeit von A als neue Bezugsgröße. Somit ergibt sich also

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \tag{2}$$

Das Ganze lässt sich wieder anschaulich in einem Venn-Diagramm darstellen:

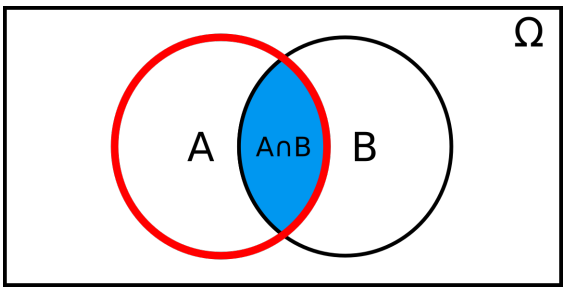

Abbildung 2: $P(B|A)$ als Venn-Diagramm

Setzt man nun die relativen Häufigkeiten nach (1) in (2) ein, so ergibt sich

$$\begin{aligned} P(B|A) = \frac{P(A \cap B)}{P(A)} &= \frac{\frac{\mathrm{Fr}(A \cap B)}{N}}{\frac{\mathrm{Fr}(A)}{N}} \\ &= \frac{\mathrm{Fr}(A \cap B)}{N} \cdot \frac{N}{\mathrm{Fr(A)}} \\ &= \frac{\mathrm{Fr}(A \cap B)}{\mathrm{Fr}(A)} \end{aligned} \tag{3}$$

Für die bedingte Wahrscheinlichkeit $P(\mathrm{im}|\mathrm{gut})$, also die Wahrscheinlichkeit von *im* unter der Vorraussetzung, dass das vorherige Wort *gut* war, gilt

$$P(\mathrm{im}|\mathrm{gut}) = \frac{P(\mathrm{gut} \cap \mathrm{im})}{P(\mathrm{gut})} = \frac{\mathrm{Fr}(\mathrm{gut} \cap \mathrm{im})}{\mathrm{Fr}(\mathrm{gut})} = \frac{1}{372} = 2.67 \cdot 10^{-3}$$

1.2 Multiplikationssatz

Um die Satzwahrscheinlichkeit herleiten zu können, muss die Wahrscheinlichkeit $P(A \cap B)$ (für einen Satz mit 2 Wörtern) berechnet werden können.
Stellt man (2) nach $P(A \cap B)$ um, so erhält man den *Multiplikationssatz*:

$$P(A \cap B) = P(B|A) \cdot P(A) \tag{4}$$

Die Wahrscheinlichkeit für den Teilsatz *gut im* ist also

$$\begin{aligned} P(\text{gut} \cap \text{im}) &= P(\text{im}|\text{gut}) \cdot P(\text{gut}) = \frac{\text{Fr}(\text{gut} \cap \text{im})}{\text{Fr}(\text{gut})} \cdot P(\text{gut}) \\ &= \frac{\text{Fr}(\text{gut} \cap \text{im})}{\text{Fr}(\text{gut})} \cdot \frac{\text{Fr}(\text{gut})}{N} = \frac{\text{Fr}(\text{gut} \cap \text{im})}{N} \\ &= \frac{1}{888233} = 1.1 \cdot 10^{-6} \end{aligned}$$

2 Satzwahrscheinlichkeit

Somit lässt sich bereits die Wahrscheinlichkeit eines Satzes der Länge 2 berechnen. Damit man beliebig lange Sätze berechnen kann, muss der Multiplikationssatz aus (4) verallgemeinert werden, was zum *erweiterten Multiplikationssatz* (in den Folien als *Kettenregel* bezeichnet) führt.

2.1 Erweiterter Multiplikationssatz

Geht man zunächst von einem Satz mit 3 Wörtern aus, so erhält man $P(A \cap B \cap C)$, was man als $P((A \cap B) \cap C)$ klammern kann. Mit dem Multiplikationssatz ergibt sich daraus

$$P((A \cap B) \cap C) = P(C|A \cap B) \cdot P(A \cap B)$$

Der zweite Faktor $P(A \cap B)$ ist aber wiederum nach dem Multiplikationssatz $P(B|A) \cdot P(A)$, wodurch sich letztendlich Folgendes ergibt:

$$P(A \cap B \cap C) = P(C|A \cap B) \cdot P(B|A) \cdot P(A)$$

Letztendlich ist die Wahrscheinlichkeit des Satzes also die Wahrscheinlichkeit des erten Wortes multiplitziert mit der Wahrscheinlichkeit des zweiten Wortes gegeben sein Vorgängerwort multipliziert mit der Wahrscheinlichkeit des dritten Wortes gegeben die Wahrscheinlichkeit des Auftretens der beiden Vorgängerwörter.

Somit wäre zum Beispiel die Satzwahrscheinlichkeit für *Studenten sind gut*

$$P(\text{Studenten} \cap \text{sind} \cap \text{gut}) = P(\text{Studenten}) \cdot P(\text{sind}|\text{Studenten}) \cdot P(\text{gut}|\text{Studenten} \cap \text{sind})$$

2.2 Definition: Satzwahrscheinlichkeit

Verallgemeinert man dieses Konzept auf beliebig lange Sätze (Hier länge n), so erhält man

$$P(A_1 \cap A_2 \cap \ldots \cap A_n) = P(A_1) \cdot P(A_2|A_1) \cdot P(A_3|A_1 \cap A_2) \cdot \ldots \cdot P(A_n|A_1 \cap \ldots \cap A_{n-1}) \quad (5)$$

Für unseren Beispielsatz heißt das also

$$\begin{aligned} &P(\text{Studenten} \cap \text{sind} \cap \text{gut} \cap \text{im} \cap \text{Adventskranzbinden}) \\ &\quad = P(\text{Studenten}) \cdot P(\text{sind}|\text{Studenten}) \cdot P(\text{gut}|\text{Studenten} \cap \text{sind}) \\ &\quad \cdot P(\text{im}|\text{Studenten} \cap \text{sind} \cap \text{gut}) \\ &\quad \cdot P(\text{Adventskranzbinden}|\text{Studenten} \cap \text{sind} \cap \text{gut} \cap \text{im}) \end{aligned}$$

Da der Teilsatz *Studenten sind gut im* aber gar nicht im Korpus vorkommt[1], vereinfacht man die Abhängigkeiten, was zur *Bigrammwahrscheinlichkeit* führt.

3 Bigrammwahrscheinlichkeit

Das Problem bei der bisherigen Berechnung der Satzwahrscheinlichkeit liegt darin, dass man hierbei immer noch auf das Vorkommen von relativ langen Teilsätzen ($A_1 \cap \ldots \cap A_{b-1}$) angewiesen ist. Diese kommen in der Praxis allerdings viel zu selten vor ($\rightarrow$ Sparse-Data-Problem). Daher bedient man sich der *Markov-Annahme*, um die Berechnung zu Approximieren und damit zu vereinfachen.

3.1 Markov-Annahme

Im Kontext der Satzwahrscheinlichkeit sagt die Markov-Annahme aus, dass die Wahrscheinlichkeit eines Wortes nicht von all seinen Vorgängern, sondern nur von m Vorgängern abhängt. Im Allgemeinen heißt das

$$P(A_n|A_1 \cap \ldots \cap A_{n-1}) := P(A_n|A_{n-m} \cap A_{n-m+1} \cap \ldots \cap A_{n-1}) \tag{6}$$

Man spricht dann von einer $m+1$-Gramm-Wahrscheinlichkeit.

Zur Verdeutlichung hier nochmal ein konkretes Beispiel:

$$P(A_5|A_1 \cap A_2 \cap A_3 \cap A_4) := P(A_5|A_3 \cap A_4)$$

Statt von allen Vorgängerwörtern hängt die Wahrscheinlichkeit nur noch von den 2 letzten Vorgängern A_3 und A_4 ab. m ist also 2 und man spricht von einem $2+1$-Gramm, also einem *Trigramm*.

Für die letzte bedingte Wahrscheinlichkeit des Beispielsatzes heißt das also

$$P(\text{Adventskranzbinden}|\text{Studenten} \cap \text{sind} \cap \text{gut} \cap \text{im}) = P(\text{Adventskranzbinden}|\text{gut} \cap \text{im})$$

[1] Sollte uns das zu denken geben?

Allerdings kann es auch bei Trigrammen passieren, dass sie gar nicht im Korpus auftauchen, was in obigem Beispiel der Fall ist. Das führt letztlich zur Bigrammwahrscheinlihckeit.

3.2 Definition: Bigrammwahrscheinlichkeit

Die Bigrammwahrscheinlichkeit ergibt sich aus der Markov-Annahme mit $n = 1$. Es wird also angenommen, dass die Wahrscheinlichkeit eines Wortes nur von dessen Vorgängerwort abhängig ist.

$$P(A_n|A_1 \cap \ldots \cap A_{n-1}) := P(A_n|A_{n-1}) \tag{7}$$

Somit ergibt sich für die Satzwahrscheinlichkeit

$$P(A_1 \cap \ldots \cap A_n) = P(A_1) \cdot P(A_2|A_1) \cdot P(A_3|A_2) \cdot \ldots \cdot P(A_n|A_{n-1}) \tag{8}$$

4 Beispiel

Um die Wahrscheinlichkeit des Beispielsatzes *Studenten sind gut im Adventskranzbinden* zu berechnen geht man also wie folgt vor:
Aus (8) ergibt sich

$$\begin{aligned} &P(\text{Studenten} \cap \text{sind} \cap \text{gut} \cap \text{im} \cap \text{Adventskranzbinden}) \\ &\qquad = P(\text{Studenten}) \cdot P(\text{sind}|\text{Studenten}) \cdot P(\text{gut}|\text{sind}) \cdot P(\text{im}|\text{gut}) \\ &\qquad \cdot P(\text{Adventskranzbinden}|\text{im}) \end{aligned}$$

Daraus wiederum wird mit (3) und (1)

$$\begin{aligned} P(\ldots) = &\frac{\text{Fr}(\text{Studenten})}{N} \cdot \frac{\text{Fr}(\text{Studenten} \cap \text{sind})}{\text{Fr}(\text{Studenten})} \cdot \frac{\text{Fr}(\text{sind} \cap \text{gut})}{\text{Fr}(\text{sind})} \cdot \frac{\text{Fr}(\text{gut} \cap \text{im})}{\text{Fr}(\text{gut})} \\ &\cdot \frac{\text{Fr}(\text{im} \cap \text{Adventskranzbinden})}{\text{Fr}(\text{im})} \end{aligned}$$

Setzt man die entsprechenden Frequenzen ein, ergibt sich

$$P(\ldots) = \frac{71}{888233} \cdot \frac{1}{71} \cdot \frac{3}{2019} \cdot \frac{1}{372} \cdot \frac{1}{5138} = 8.75 \cdot 10^{-16}$$