### **Test des Assistenten**

Für den Test der Vorhersagefunktion des Assistenten werden wir folgendes Experiment durchführen. Zunächst erstellen wir ein Textkorpus aus Daten, die mit hoher Sicherheit dem Genre Märchen zuzuordnen sind. Hierfür nutzen wir die im TextGrid-Repositorium veröffentlichten deutschsprachigen Märchensammlungen.

Wir wählen die Märchentexte von Jacob und Wilhelm Grimm (215), Heinrich Pröhle (150), Karl Bartsch (52 + 47) und Ludwig Bechstein (81) aus und fügen sie in einem XML-TEI-Corpus zusammen. Zusätzlich werden 12 weitere Texte aus der Sammlung „Europäische Märchen“ aufgenommen. Aufgrund rechtlicher Gründe sind diese Texte in einer abgeleiteten Fassung dargestellt. Insgesamt umfasst unser Korpus somit 557 Texte.

Nun wählen wir ein Beispiel eines Märchentyp ATU 300 – Der Drachentöter aus der Kollektion von Karl Bartsch aus und annotieren den Text anhand des Markups für die inhaltliche Annotation des Märchens.

Der ausgewählte Text stellt ein komplexes Märchen dar, das aus vier Typteilen besteht. Nach den Anfangsepisoden aus den Typen ATU 567 (vom Essen des außergewöhnlichen Vogelherzes bis zur Flucht) und ATU 303 (Großziehen von zwei ähnlich aussehenden Brüdern bis zur Trennung der Brüder und Austausch der Lebensindikatoren) beginnt die Episode des Typs ATU 300 – Der Drachentöter. Diese umfasst den Inhalt von der Erkundigung der menschlichen Opfergabe bis zur Heirat der geretteten Prinzessin. Der Text wird weiter als Typ ATU 303 fortgesetzt. Dabei wird berichtet, wie der jüngste Bruder von der Hexe versteinert, aber dann mit Hilfe seines Bruders wieder erlöst wird.

Für das Experiment werden wir in diesem Märchen nur den Textteil des Typs ATU 300 erschließen, das heißt mit Hilfe des Markups labeln. Somit wird das Modell nur anhand positiv gelabelter Datensätze trainiert. Im Rahmen dieses Verfahrens werden aus allen gelabelten Datensätzen die Tokens extrahiert und ein Vokabular erstellt.

Dabei wird das Vokabular von sogenannten Stoppwörtern und anderen für den Vergleich unwesentlichen Wörtern bereinigt. Letztere gelten als seltene Wörter, die nur vereinzelt in den Datensätzen vorkommen und daher keine Rolle beim Vergleich dieser Datensätze mit anderen spielen. Für die Bereinigung des Vokabulars von vereinzelt stehenden Wörtern ist der Parameter min\_df=n zuständig. Angesichts der geringen Anzahl der Datensätze sowie der Kürze der Liste der in diesen Datensätzen anzutreffenden Tokens, wird der Wert des min\_df Parameters bei unserem Experiment auf den minimalen Wert (min\_df=2) gesetzt. Das bedeutet, dass jedes im Vokabular erfasste Wort wenigstens in zwei verschiedenen Datensätzen vorkommen muss.

Als nächstes werden die im Vokabular vorhandenen Wortformen nach ihren TF\*IDF-Werten skaliert. Die TF (Term Frequency) eines Tokens wird berechnet als:

Häufigkeit des Tokens in einem Datensatz / Gesamtzahl der Tokens im Datensatz

Die IDF (Inverse Document Frequency) eines Tokens wird berechnet als:

Loge (Gesamtzahl der Datensätze + 1 / Anzahl der Datensätze mit dem gesuchten Token + 1) + 1

Anschließend werden die TF\*IDF-Werte durch das L2-Normalisierungsverfahren umskaliert. Das somit berechnete Koeffizient bestimmt den Wert jedes einzelnen Tokens im Vokabular. Die Rangordnung der Tokens im Vokabular ermöglicht es, die Anzahl der Tokens im Vokabular effektiv zu korrigieren.

Für die Annäherung zwischen dem erstellten Vokabular und dem Modell einer konkreten Geschichte ist die Segmentierung der Trainingsdaten in natürliche Handlungsabschnitte von besonderer Bedeutung. Sie stellt sicher, dass Wörter, die in zwei aufeinanderfolgende Handlungsabschnitte (Offenkundigkeit der Krise – Auseinandersetzung mit der Krise) wiederholt vorkommen, mit getrennten Datensätzen dargestellt werden. Dadurch erhalten sie höhere TF\*IDF-Gewichte oder werden überhaupt im Vokabular erfasst.

Im nächsten Schritt wird aus den restlichen (nicht gelabelten) Datensätzen im Textkorpus ein Messdatenkorpus aggregiert. In diesen Messdatenkorpus wird auch der zuvor erstellte Modell-Datensatz eingefügt. Anschließend werden alle anderen Datensätze aus dem Messdatenkorpus mit dem Modell-Datensatz verglichen, um vorherzusagen, welche von ihnen diesem am nächsten stehen.

Das Verfahren umfasst folgende Schritte:

1. Erstellung und Bereinigung des Vokabulars: Das Vokabular wird von Stoppwörtern und anderen für den Vergleich unwesentlichen Wörtern bereinigt.

2. Berechnung der TF\*IDF-Werte: Für jedes Wort im Vokabular werden die TF\*IDF-Werte berechnet.

3. Erstellung dünnbesetzter Matrizen: Für jeden Datensatz wird eine dünnbesetzte Matrix erstellt, die die TF\*IDF-Werte der enthaltenen Tokens repräsentiert.

4. Vergleich der Matrizen: Jede dieser Matrizen wird mit der Matrix des Modell-Datensatzes verglichen.

Die Berechnung der Ähnlichkeit zwischen den Datensätzen beruht auf der Formel der Cosinus-Ähnlichkeit, die wie folgt definiert ist:

Dabei entspricht ***ai*** einem, in einem TF\*IDF Wert skalierte Merkmal (Token) aus dem Modell-Datensatz, ein ***bi*** entspricht hingegen einem in einem TF\*IDF Wert skalierte Merkmal (Token) aus dem beliebigen Mess-Datensatz. Je näher der Wert der berechneten Cosinus-Ähnlichkeit bei Eins liegt, desto plausibler ist die vorhergesagte inhaltliche Ähnlichkeit zwischen ***a*** und ***b*** Datensätzen. Diese Methode ermöglicht es, die Ähnlichkeit der Datensätze zu quantifizieren und diejenigen zu identifizieren, die dem Modell-Datensatz am nächsten stehen.

Die Datensätze mit den höchsten Koeffizienten, also die wahrscheinlichsten Vorhersagen, stehen an der Spitze der Liste. Anhand des Nachschlagewerks von Hans-Jörg Uther, dem „Deutschen Märchenkatalog“, lässt sich mühelos überprüfen, ob der Algorithmus alle im Textkorpus enthaltenen Texte bzw. ihre Teile gefunden hat, die zur gesuchten Märchenkategorie gehören (gemeint ist der Typ ATU 300 – Der Drachentöter).

Die Recherche im Katalog zeigt, dass in der Tat alle entsprechenden Texte durch den Algorithmus sicher erfasst wurden. Das sind wie folgt:

|  |  |
| --- | --- |
| Text im Katalog | Text ID im XML-Textkorpus |
| Bartsch 1879f. I, 474ff. | @n=“dummy20“ |
| Bechstein/Uther 1997 I, Nr.49 | @n=“dummy122“ |
| Grimm/KHM (1857) Nr. 60 | @n=“dummy484“ |
| Pröhle 1853, Nr. 4 | @n=“dummy234“ |
| Pröhle 1853, Nr. 5[[1]](#footnote-1) | @n=“dummy256“ |

In diesem Experiment gezeigte Verfahren ist insbesondere bei der Erstellung der Trainingsdaten einzusetzen. Mit den so gewonnenen Trainingsdaten ist es möglich, die Vorhersage nicht nur zu präzisieren, sondern auch auf kleinere Inhaltselemente zu richten. Um diese Eigenschaft zu demonstrieren, führen wir ein weiteres Experiment durch.

Wir trainieren das Modell anhand der zuvor gelabelten Datensätze, darunter die oben erwähnten Texte des Typs ATU 551 „Das Wasser des Lebens“.[[2]](#footnote-2) Die ID-Zeichen dieser Texte im Textkorpus beginnen mit dem Suffix „zyx\_“. Nach der Aggregation eines provisorischen Textkorpus aus diesen und den im ersten Experiment gelabelten Datensätzen (Datensätze mit dem ID-Suffix „dummy“) können wir alle in diesem Korpus vorhandenen Markupelemente abrufen und betrachten.

Als nächstes wählen wir aus dieser Liste ein oder mehrere Markupelemente aus und erstellen eine entsprechende Abfrage. Angenommen, wir suchen nach dem Motiv der Offenkundigkeit der Krankheit des Vaters. Die entsprechende Abfrage besteht aus den Attributen „a551“ und „Krankheit:“. Der Doppelpunkt nach dem Wort „Krankheit“ stellt sicher, dass ein weiterer Wortlaut wie „Krankheit\_beheben“ in der Suche ignoriert wird.

Das Ergebnis der anschließenden maschinellen Analyse bestätigt die hohe Präzision der eingesetzten Vorhersagefunktion. Der erste Treffer entspricht dem Textteil mit dem gesuchten Motiv. Außerdem ist der Unterschied zwischen den Koeffizienten des ersten und der folgenden Datensätze so groß, dass sie zweifelsohne zwei kontrastiv unterschiedliche Vorhersagen darstellen müssen. Tatsächlich, wie der oben erwähnte Katalog bestätigt, sind im Textkorpus keine anderen Texte mit gleichem Inhalt zu erwarten.

1. Der Text stellt eine Kombination von Typen 303 und 300 dar. Im Katalog von Uther nur bei der Erschließung des Typs ATU 303 erfasst. [↑](#footnote-ref-1)
2. Europäische Märchen. CD-Ausgabe [↑](#footnote-ref-2)