Vorab

Dieses Manual vermittelt grundlegende Kenntnisse und Fähigkeiten im Umgang mit dem Formalismus der regulären Ausdrücke. Es ist abgestimmt auf die Arbeit mit Textdaten und eignet sich auch für Neueinsteiger.

Stellen wir die Definitionsfragen vorerst einmal zurück und widmen uns zum Einstieg einem Beispiel, das dem fortgeschritteneren Gebrauch zuzuordnen ist und Anfänger sicherlich daran zweifeln lässt, jemals mit regulären Ausdrücken umgehen zu können. Doch es gibt auch gute Nachrichten: Erstens muss der Ausdruck zu diesem Moment noch lange nicht verstanden werden, zweitens wird es in diesem Manual nicht mehr viel komplizierter. Und so kann ein regulärer Ausdruck aussehen:

[A-Za-z]{4} [A-Za-z]*([A-Za-zü]{4}){2}\.

Das Beispiel wirkt undurchschaubar kompliziert und in der Tat mögen reguläre Ausdrücke zunächst abschrecken, doch das Einarbeiten lohnt sich, zumal sie nicht nur ein systematisches sondern bald auch ein zeitsparendes Durchsuchen von Textdaten erlauben.

Sie erweisen sich als überaus nützlich, will man ein abstraktes Muster im Text finden, ohne zu wissen, wie dieses durch Zeichen und Zeichenketten besetzt wird. Mit anderen Worten: Reguläre Ausdrücke werden herangezogen, wenn man eine Vorstellung davon hat, was es in vorliegenden Daten zu finden gilt, ohne zu wissen, wie das Gesuchte im Detail auszusehen hat. Doch wie hat man sich so etwas vorstellen? Wie lassen sich Zeichen und Zeichenfolgen umschreiben? Welche Kriterien können herangezogen werden, um Zeichen und abstrakte Muster zu beschreiben? Und wie lassen sich sich aus konkreten Textdaten überhaupt Muster abstrahieren? Wie lassen sie sich mithilfe regulärer Ausdrücke formalisieren?

Genug verwirrt. Bevor der Einsatz regulärer Ausdrücke näher erläutert werden kann, zunächst eine kleine Übung, die verdeutlichen soll, inwiefern Textdaten Variablen und abstrakte Muster zugrunde liegen. Die Fragestellung: Welche formalen Eigenschaften haben Datensatz A und Datensatz B in Abgrenzung zu Datensatz C gemeinsam? (Der Schwierigkeitsgrad steigt vom ersten bis zum siebten Beispiel. Es lohnt sich wirklich, die Beispiele vor dem Durchlesen der Lösung zu durchdenken, da sie so für das Verständnis regulärer Ausdrücke einen großen Vorteil bringen.)

	Datensatz A	Datensatz B	Datensatz C
1)	G	A	1
2)	A	B	a

Die Gemeinsamkeiten in 1 und 2 sind schnell erklärt. Während es sich in 1) bei Datensatz A (*G*) und Datensatz B (*A*) um Buchstaben handelt, besteht Datensatz C aus einer Ziffer (*1*).
Im zweiten Beispiel liegt der Unterschied zwischen den Datensätzen A (*A*) und B (*B*) auf der einen und C (*a*) auf der anderen Seite in der Groß- und Kleinschreibung der Buchstaben.

	Datensatz A	Datensatz B	Datensatz C
3)	aub	avb	cud
4)	abc 1	abd 1	efc1

Ein wenig komplizierter wird es in den Beispielen 3) und 4). Hier spielt neben dem Zeichentypus auch die Abfolge der Zeichen ebenfalls eine Rolle. Die Datensätze A und B im dritten Beispiel zeichnen sich in Abgrenzung zu C dadurch aus, dass beide dreistellig sind, mit einem *a* beginnen und mit einem *b* enden (natürlich handelt es sich auch um Buchstaben, Kleinbuchstaben, aber darin unterscheiden sie sich nicht von C). Das mittlere Zeichen unterscheidet sich in A und B, ist also für das gemeinsame Muster egal. Umschreiben ließen sich die Folgen folglich durch "*a* egal *b*".
Beispiel 4) zeigt zudem, dass auch Leerstellen von Bedeutung sind. Das Muster, das ausschließlich den Datensätzen A und B gemeinsam ist, besteht hier nicht nur aus der Folge "a b egal", sondern lässt sich auf "*a b* egal *Leerstelle 1*" erweitern.

	Datensatz A	Datensatz B	Datensatz C
5)	Der Hund schläft im Haus.	Niemand war im Haus.	Im Haus war niemand.

In Beispiel 5) drängt sich das gemeinsame *im Haus* auf. Dies erscheint zwar auch in Datensatz C, tritt aber anders als in A und B nicht am Ende des Satzes auf. Hier spielt also die Position des Syntagmas *im Haus* eine Rolle.

	Datensatz A	Datensatz B	Datensatz C
6)	ABC DEF GHI.	JKL MNO GH1.	TUV WXY IGH.

Ähnliches lässt sich auch für Beispiel 6) sagen. Mit Ausnahme von *G* und *H* liegen keine gemeinsamen Buchstaben vor. Um die Datensätze A und B von C abgrenzen zu können, muss auch die Position von *GH* beachtet werden. Während auf die Buchstabenkombination in A und B ein beliebiges Zeichen und schließlich ein Punkt folgt, ist der Punkt in Datensatz C direkt hinter *GH* platziert. Folglich könnte eine Umschreibung der entscheidenden Sequenz in A und B "*G H* egal *Punkt*" lauten.

	Datensatz A	**Datensatz B**	**Datensatz C**
7)	Alle Auen sind grün.	Auch Autobahnen sind grau.	Mein Hund kann Sitz.

Die Datensätze des siebten Beispiels lassen möglicherweise mehrere Lösungen zu und es dient lediglich dazu, den Blick für abstrakte Muster im Text zu schärfen. Alle drei Datensätze bestehen aus einem durch einen Punkt beendeten Satz, der sich aus vier Wörtern zusammensetzt, welche mit Ausnahme von *Autobahnen* stets aus vier Buchstaben bestehen. Bezüglich der Gemeinsamkeiten von Datensatz A und B in Abgrenzung zu C ließe sich anführen, dass die ersten beiden Wörter in beiden Fällen mit einem *A* beginnen, das dritte Wort *sind* ist und das vierte Wort mit *gr* beginnt. Anders gesagt: "*A* egal egal egal *Leerstelle A* unbestimmt oft egal *Leerstelle sind gr* egal egal *Punkt*".

Will man ein Muster entwickeln, das auf alle drei Datensätze in Beispiel 7) passt, kann man dazu folgende Beschreibung heranziehen: Der Satz besteht aus vier Wörtern, von denen sich das erste, das dritte und das vierte aus je vier Buchstaben zusammensetzen. Das zweite ist beliebig lang. Um nun den Bogen zum Anfang zu spannen: Genau das sagt das Anfangsbeispiel *[A-Za-z]{4} [A-Za-z]*([A-Za-zü]{4}){2}\.* aus.

Übrigens lassen sich im Internet Seiten finden, auf denen anhand von Beispieltexten reguläre Ausdrücke erklärt werden und ausprobiert werden können. Zwei Beispiele hierfür:

http://www.regexr.com/

https://regex101.com/

Up Reguläre Ausdrücke in der Arbeit mit Textdaten Reg-was? Next