Woche 02 Daten: Textvorverarbeitung

Skript

Erarbeitet von
Ann-Kathrin Selker

Lernziele

- Vorverarbeitungsschritte in der Textverarbeitung kurz erklären können
- Vorverarbeitung beispielhaft zeigen können

Inhalt

Welche Vorverarbeitungsschritte sind notwendig, damit ein Machine-Learning-Modell mit Texten umgehen kann?

Projekt „KI für alle"
HeiCAD –Heine Center for Artificial Intelligence and Data Science

Einstieg

Texte liegen im Computer anhand von Strings vor. Strings sind Zeichenketten, die als Zahlen kodiert gespeichert werden. Eine Kodierung ist eine eindeutige und feste Zuordnung von einem Zeichen zu einer Zahl.
Die Verarbeitung von Text hängt sehr stark von der späteren Anwendung ab. Die hier vorgestellten Schritte sollen nur als Beispiel dienen, womit man sich bei der Textverarbeitung befassen kann. Insbesondere spielt die Sprache des Textes eine Rolle dabei, welche Schritte wie ausgeführt werden müssen.

Tokenisierung

Damit eine Maschine unseren Text verstehen kann, unterteilen wir ihn zuerst in Einzelteile, sogenannte Tokens. Es kommt auf die Anwendung an, wie die Aufteilung stattfindet, aber häufig wird der Text erst in Einzelsätze und dann in seine Einzelwörter aufgeteilt. Es ist also notwendig, den Text von vorne bis hinten durchzugehen und an den Leerzeichen, Zeilenumbrüchen usw. zu trennen. Je nach Sprache kann es aber auch nötig sein, einen Text in kleinere Einheiten als Wörter aufzuteilen. Jeder kennt bestimmt das berüchtigte Rindfleischetikettierungsüberwachungsaufgabenübertragungsgesetz, zumindest vom Namen her. Hier ist es selbst für Menschen noch nötig, das Wort in seine Einzelwortbestandteile aufzuteilen, um es zu verstehen. In diesem Schritt findet manchmal auch eine Auflösung von Verkürzungen statt, wie z.B. das Auflösen von „auf's“ zu „auf das“.

Normalisierung

Nach der Aufteilung in Tokens folgt häufig eine Normalisierung. Für den Computer sind die Wörter „Aber“ und „aber“ unterschiedliche Wörter. Das kannst du dir deutlich machen, indem du dir die ASCII-Kodierung dieser beiden Wörter anguckst: 65 98 101 114 und 97 98 101 114. Die Großschreibung von Wörtern erfolgt in vielen Sprachen vor allem am Satzanfang, sodass sich die Bedeutung dieser Wörter nicht ändert, nur weil sie großgeschrieben werden. Daher werden häufig alle Tokens einmal durchgegangen und Großbuchstaben in Kleinbuchstaben umgewandelt. Außerdem werden in diesem Schritt häufig Satzzeichen entfernt. Auch die Wortart eines Wortes kann wichtig sein, daher kann die Wortart mit einem sogenannten "Part-of-Speech-Tagger" bestimmt und gespeichert werden.
Allerdings gibt es auch Sprachen wie z. B. Deutsch, bei der die Großschreibung eine große grammatikalische Rolle spielt. Bei den Wörtern „Arm“ und „arm“ ändert die Großschreibung die Bedeutung stark. Daher müssen auch hier eventuell sprachspezifische Normalisierungsschritte durchgeführt werden. Die vorher bestimmte Wortart hilft in diesem Fall. Je nach Herkunft des Textes kann es auch nötig sein, Tokens zu korrigieren, also zum Beispiel Rechtschreibfehler zu entfernen.

Stoppwörter

In Texten gibt es häufig Wörter, die irrelevant oder so gut wie irrelevant für die Bedeutung des Textes sind. Ein Beispiel dafür sind Artikel (also z. B. „der“, „die“, „das“ im Deutschen) und Bindewörter (also z. B. „und“ im Deutschen). Diese Wörter können daher aus dem Text gestrichen werden, ohne dass Inhalt verloren geht. Du kennst solche Wörter bestimmt aus Suchmaschinen: Du findest deine Resultate, egal ob du nur die wichtigsten Stichpunkte angibst oder Suchanfragen in ganzen Sätzen schreibst. Der Fachbegriff lautet Stoppwörter, im englischen stop words. Die Liste der Stoppwörter kann je nach Anwendung unterschiedlich ausfallen.

Lemmatisierung/Stemming

Je nach Wortart und Stellung im Satz können Wörter verschiedene Formen haben und werden daher als unterschiedliche Wörter wahrgenommen, obwohl die grundlegende Bedeutung dieselbe ist. Bei der Lemmatisierung werden Wörter in ihre Grundform umgewandelt, also z. B. die Wörter „geht“, „ging“ und „gegangen“ in „gehen“. Beim Stemming werden Wörter auf ihren Stamm zurückgeführt, also z. B. „geh“ für gehen. Lemmatisierung und Stemming kennst du sicher ebenfalls von Suchmaschinen: Du musst nicht sämtliche grammatikalischen Fälle etc. deiner Suchbegriffe durchgehen, um auch Resultate zu finden, bei denen dein Suchbegriff zum Beispiel nur im Genitiv oder nur im Plural verwendet wird.

Abschluss

Alle genannten Vorverarbeitungsschritte benötigen eigene Algorithmen, in die immer noch viel Forschungsarbeit einfließt. Je nach Sprache gibt es für Python Funktionen, die dir einen Großteil der Vorverarbeitung abnehmen.

In diesem Video hast du einen kleinen Überblick über die verschiedenen Vorverarbeitungsschritte bei der Textverarbeitung bekommen.

Weiterführendes Material

https://www.kaggle.com/code/sudalairajkumar/getting-started-with-text-preprocessing

Disclaimer

die verwendeten Logos, alle in den Quellen ausgewiesenen Fremdmaterialien sowie alle als Quellen gekennzeichneten Elemente.