Big Data Praktikum: Attribute extraction from eCommerce product descriptions

Gregor Pfänder Thilo Brummerloh

18. Mai 2021

1 Fragen für Freitag

- Datenvorverarbeitung genau beschreiben?
- Machine Learning/Neuronale Netzwerke genauer?
- Was wird in Testat gefragt/Ablauf?
- Welche Entscheidungen sollen fest stehen?
- Weitere Attribute labeln/trainieren? Weitere Schritte wie labeln beschreiben?

2 Thema

Produktseiten von Onlineshops enthalten oft viele unzureichend strukturierte Produktbeschreibungen oder sogar gar keine Beschreibungen. Zum Preisvergleich des gleichen Produkts auf verschiedenen Webseiten muss allerdings bekannt sein um welche Ausprägung eines Produkt es sich handelt. So können Preise nicht nur von Produkten, sondern auch ihren Unterausprägungen, wie dem Speicherplatz oder der Farbe, verglichen werden.

3 Fragestellung

Es sollte möglich sein mithilfe von einem Computerprogramm diese Arbeit automatisch durchzuführen. Wenn Produktspezifikationen innerhalb eines Freitextes vorliegen, sollte es möglich sein daraus Wörter und Wortgruppen zu erkennen und einer Spezifikation zuzuweisen. Es soll eine Methode zur Named Entity Recognition ausgewählt werden um Marke und Produktnummer herauszufinden.

4 Methodenüberblick

Alle Methoden lassen sich unter dem Begriff Named Entity Recognition zusammenfassen. Nachfolgend werden ausgewählte Methoden erklärt. Bei allen handelt es sich um Machine Learning Ansätze.

4.1 Naive Bayes

Naive Bayes Ansätze waren sehr beliebt[1] Eine Möglichkeit des NER liegt darin das Klassifizierungsproblem mit einem Naive Bayes Klassifikators zu lösen. Bei Naive Bayes wird sich ‚supervised learning' zu nutzen gemacht. Es werden also gelabelte Trainingsdaten benötigt. Anhand der Trainingsdaten werden für alle Werte die Wahrscheinlichkeiten bestimmt, dass der Wert einer bestimmten Klasse zugehört (bzw. eine Hypothese erfüllt). In unserem Fall könnten die Werte die Wörter einer Sequenz und mögliche Klassen ‚Attribut', ‚Attributwert' und ‚Keins von beiden' sein. Nachdem die Wahrscheinlichkeiten bekannt sind, können anhand des Naive Bayes Klassifikators die ungelabelten Daten immer der Klasse mit der höchsten Wahrscheinlichkeit zugeordnet werden. Um eine erfolgreiche NER Methode auf Basis von Naive Bayes entwickeln zu können benötigt es noch weitere Schritte wie beispielsweise das Verlinken zwischen Attributen und Attributwerten. Bewertung: Der Naive Bayes Klassifikator ist ein weiteverbreitetes Mittel zur Textklassifizierung und wird beispielsweise bei Spam Filtern erfolgreich eingesetzt um Wörter in die Klassen ‚Spam' (Schlecht) oder ‚Ham' (gut) einzuordnen. Allerdings ist der Ansatz wie bereits im Namen steht naiv. Er geht davon aus, dass keine Beziehungen zwischen den Werten bestehen. Diese sich also nicht gegenseitig beeinflussen. Diese Annahme ist kritisch bei der Textverarbeitung. In unserem Anwendungsfall könnte sich das unter anderem negativ auf die Extraktion von Attributwerten welche aus mehreren Wörtern zusammengesetzt sind auswirken. Bei solchen Attributwerten bestehen Abhängigkeiten zwischen mehreren Wörtern, die das Modell beachten sollte. [2]

4.2 CNNs

Neueste Ansätze verwenden CNNs zur Klassifikation. Zhu u. a. 2018 haben erfolgreich ein CNN verwendet um 2017 die genaueste Methode zur Named Entity Recognition von verschiedenen Biologieliteraturkorpora vorzustellen. Die GRAM-CNN genannte Methode hat laut den Autoren für das labeling von Biologieliteratur einen Genauigkeitsvorteil indem, nicht wie in klassischen LSTMs der ganze Satz betrachtet wurde, sondern nur die um ein Wort liegenden Nachbarworte. Die Worte, dieser N-Gram genannten Wortketten, werden in einem ersten Schritt zu einer Darstellung umgewandelt, die nicht das Wort selber enthält, sondern das Wort durch seine einzelnen Buchstaben als Vektor darstellt. Dazu wird dem Wort noch ein Part-of-Speech tag zugewiesen und der Character des Worts als Vektor dargestellt. Die Vektorisierung erfolgt durch Abgleich der Worte mit vorgefertigte Bibliotheken die auf ähnlichen Korpora Vektoren berechnet haben. Das Part-of-Speech tag enthält Informationen zu Abhängigkeiten eines Wortes zu anderen. Eine Eingabe in das

[1] Ghani u. a. 2006.

[2] (Quelle: Text Mining for Product Attribute Extraction) (Quelle: https://course.elementsofai.com/de/3/3)

GRAM-CNN ist also ein n-gram, das wie beschrieben durch seine Buchstaben mit Zusatzinformationen dargestellt wird. Innerhalb des CNNs wird diese Information versteckt verarbeitet und es liefert die features der Worte an ein CRF in dem die Verbindung von mehreren Worten erkannt werden kann. Damit sind auch auseinander geschriebene zusammengehörige Worte als solche erkennbar.
CNNs wurden auch in Lee, Chung u. a. 2019 und Lee, Park und Cho 2020 verwendet um Produkteigenschaften aus Benutzerbewertungen zu extrahieren. Mit den extrahierten Eigenschaften werden von den Autoren letztlich Sentiment-Scores der Benutzerbewertungen erstellt, die nicht nur die Erwähnung von bewertbaren Wörtern zählen, sondern auch die relative Gewichtung von verschiedenen Eigenschaften des bewerteten Produkts herausfinden.

4.3 RNNs (LSTMs)

RNNs sind die Standardmethode für Textklassifikation, dabei wird meist die Spezialform der LSTMs verwendet.[3] RNNs verwenden wie die vorherigen Methoden Trainingsdaten um einen Klassifikator zu trainieren. Eingaben sind Texte aus denen bereits die Interessanten Textstellen mit ihrer Beschreibung extrahiert wurden. Damit wird ein Neuronales Netzwerk trainiert. Das trainierte Netzwerk ist dann in der Lage beliebige Texte einzulesen und die darin enthaltenen Entitäten in Form von Markennamen zu erkennen und zuzuordnen. Mit diesen Informationen sollte es möglich sein größere Korpora von Produktbeschreibungen automatisch mit Informationen anzureichern.

5 Daten

5.1 Trainingsdaten

Je nach Methodenwahl werden unterschiedliche Daten benötigt. Zur Entity Resolution wäre nur ein Menge von ungeordneten Produktbeschreibungen nötig. Wenn allerdings zusätzlich ein Neural Network trainiert werden soll ist es auch nötig einen bereits gelabelten Datensatz zum training zu haben.

5.2 title

5.3 Beschreibung Datensatz

Zur Verfügung steht ein Datensatz mit 56005 Produkten in Form einer `.csv`-Datei. Bei den Produkten, die in den Daten erfasst wurden, handelt es sich um unterschiedliche Arten von Haushaltsgeräten. Die Produkte reichen dabei von Waschmaschinen und Wäschetrocknern bis zu Kühlschränken und Mikrowellen. Eine genaue Zusammenfassung welche Produkttypen in den Daten vorkommen, kann durch eine erste Datenbegutachtung nicht getroffen werden. Der Produkttyp wäre also ein weiteres Attribut, bei dem eine Extraktion sinnvoll wäre. Die Werte wurden aus den Produktseiten verschiedener eCommerce Seiten extrahiert. Die Aufteilung der Daten

[3] Majumder u. a. o. D.

nach Quelle ist wie folgt:

Quelle	Anzahl
shopee.my	24849
appliancesconnection.us	21434
sharafdg.ae	5443
productreview.au	2502
spencerstv.us	1777

Nachdem die .csv Datei mithilfe von Pandas in ein Dataframe überführt wurde, erhält man die Nachfolgende Datenstruktur:

id	source	name	productdescription	url	brand	modelnumber

Die Zeilen von Interesse sind ‚name' und ‚productdescription'. In diesen Feldern stehen die Sequenzen, aus denen die Attribute extrahiert werden sollen. Außerdem ist der Datensatz bereits für die Attribute ‚brand' und ‚modelnumber' gelabelt. (der Absatz könnte weggelassen werden. Um einen ersten Eindruck der Daten zu bekommen wird nachfolgen der erste Eintrag des Datensatzes aufgeführt:

```
id: 000266f5e7ab2344315290174dfb75f7
source: appliancesconnection.us
name: "Broan TEN136WW"
productdecription: "Broan TEN136WW Overview The Tenya 1 Series Under Cabinet
Range Hood by Broan offers 2-speed motor with up to 250 CFM of ventilation.
[...] Vertical/Horizontal Rectangular Duct 7 in. Vertical Round Duct 1
Year Limited Warranty"
url: https://www.appliancesconnection.com/broan-ten136ww.html?zipcode=20001
brand: Broan
modelnumber: TEN136WW)
```

Bei der ersten Begutachtung der Daten sind einige Problemstellen und mögliche Störfaktoren . . . :

1. "-Zeichen wird in der Zeile ‚productdescription' oft als Größen Bezeichnung benutzt: Dadurch denkt Pandas, dass der Satz vorbei ist und jedes weiter Komma wird als neue Zeile interpretiert. . .
2. Chinesiche Zeichen bei mind. einem Produkt
3. Andere Sonderzeichen: verarbeitbar oder Störfaktoren?
4. Verschiedene Quellen könnten zu verschiedenen Mustern führen: Beobachten wie gut unser Modell je nach Quelle funktioniert.
5. Marke und Modelnummer zu einfach zu extrahieren?: weitere Attribute festlegen und von Hand labeln?

Literaturverzeichnis

Ghani, Rayid u. a. (2006). "Text mining for product attribute extraction". In: *ACM SIGKDD Explorations Newsletter* 8.1, S. 41–48. ISSN: 1931-0145. DOI: 10.1145/1147234.1147241.

Lee, Younghoon, Minki Chung u. a. (2019). "Extraction of Product Evaluation Factors with a Convolutional Neural Network and Transfer Learning". In: *Neural Processing Letters* 50.1, S. 149–164. ISSN: 1370-4621. DOI: 10.1007/s11063-018-9964-8.

Lee, Younghoon, Jungmin Park und Sungzoon Cho (2020). "Extraction and prioritization of product attributes using an explainable neural network". In: *Pattern Analysis and Applications* 23.4, S. 1767–1777. ISSN: 1433-7541. DOI: 10.1007/s10044-020-00878-5.

Majumder, Bodhisattwa Prasad u. a. (o. D.). *Deep Recurrent Neural Networks for Product Attribute Extraction in eCommerce*. URL: http://arxiv.org/pdf/1803.11284v1.

Zhu, Qile u. a. (2018). "GRAM-CNN: a deep learning approach with local context for named entity recognition in biomedical text". In: *Bioinformatics (Oxford, England)* 34.9, S. 1547–1554. DOI: 10.1093/bioinformatics/btx815.