

1. Definitionen

Bedeutung verwendeter Begriffe im Kontext der Thesis.

* 1. Merkmale
* Das Produkt ist kein Merkmal.
* Merkmal eines Produktes kann ein Bestandteil eines Produktes sein oder eine Eigenschaft.
* Nur explizite Merkmale, welche als Nomen vorkommen, z.B.: „Die Qualität des T-Shirts war schlecht.“
* Implizites Merkmal ist z.B.: „Das T-Shirt hat sich nach der ersten Wäsche verzogen.“
  + Schwer zu identifizieren für einen Tagger. Im Beispiel werden diese deshalb nicht bewertet.
  1. Meinungen
* Als Meinungen wird hier der Kontext bezeichnet, in welchem ein Merkmal auftaucht. Das kann positiv oder negativ sein.
* Pro Bewertung für jedes Merkmal zusammengefasst und anhand eines nummerischen Wertes dargestellt. Nur eine Positive +1 oder eine Negative -1 Tendenz der Meinung.
* Meinungsworte sind Worte, welche eine positive oder negative Meinung zu einem Merkmal ausdrücken.
* Manuell aufschreiben und Liste von Meinungsworten aus dem **SentiWS** (deutsch).

1. Datenbasis
   1. Externe Programme und Hilfsmittel

* **Scrapy** – Python Framework, welches schnelles scraping und crawling ermöglicht.
* **STTS** – deutsches, bekanntes Tagset.
* **Pyenchant** – Bibliothek zur Rechtschreibkorrektur für Python, basierend auf der **Enchant** Bibliothek.
* **NLTK** – Natural Language Toolkit ist eine Sammlung von Hilfsmitteln zur Verarbeitung in Python, darunter Segmentierung in Sätze und Worte, Part-of-Speech-Tagging.

1. Crawling

In diesem Beispiel werden mit den Kategorien Damenmode und Herrenmode gearbeitet und am Ende die merkmalsbasierte Stimmungsanalyse zusätzlich mit der Kategorie Multimedia getestet.

1. Analyse der Trainingsdaten

* Überblick über die Menge und Beschaffenheit der Bewertungen erhalten.

1. Vorverarbeitung

* Entfernung von sehr kurzen Bewertungen und Übertreibungen
  + Übertreibungen mit mehrfach auftretenden Satzzeichen (!!!!) oder (suuuuper) – max. 2 Zeichen außer bei Buchstabe s.
  + Bewertungen mit weniger als 10 Zeichen.
  + Optional: Alle nicht deutschen Rezensionen entfernen. [Fertig]
* Rechtschreibkorrektur
  + Pyenchant wird verwendet.
  + Deutsche Wörterbuch der Rechtsschreibung **Hunspell.**
* Stopwörter
  + Eigene Liste vorhanden. Diese verwenden.
* Wortwolke (ohne Tagger)
  + Ergebnis ansehen und dann entscheiden ob Tagger notwendig sein wird.
* Segmentierung in Sätze und Worte
  + NLTK wird verwendet.
  + Die Spezialisierung ist notwendig, um zu verhindern, dass nach Abkürzungen ein Satz fälschlicherweise getrennt wird.
* Part-of-Speech Tagging – Tokenisierung
  + Einteilung von Texten in einzelne linguistische Einheiten.
  + Zuerst in einzelne Sätze segmentieren. 🡪 **enchant**
  + Dann jeden Satz in Worte unterteilen.

1. Merkmalsextraktion durch häufigste Nomen

* Bewertungen mit Wortarten annotieren.
* Danach häufigste Nomen und Eigenworte betrachten.
  1. POS-Tagging
* Tagset STTS kann verwendet werden.
* Stanford Tagger 🡪 NLTK. Dewac Modell, für deutsches Web als Korpus. Dieses Modell wurde mit Negra Korpus trainiert.
  1. Häufigste Nomen
* Es wird nur ein Teil der Daten zum Taggen verwendet, weil es sonst zu lange dauert.
* Neben häufigsten Nomen werden auch die häufigsten Eigennamen in die Liste mitaufgenommen.
  1. Filterung
* Häufigsten Nomen/Eigennamen sind Marken oder Kategorienamen.
* Diese werden herausgefiltert.
* Stopwörter herausfiltern.