Die Ausarbeitung sowie die Quelle der verwendeten Daten wurden in Github abgelegt und sind unter dem folgenden Link zu erreichen: palimpalim08/Phase2_Data_Analysis (github.com)

Zu Beginn wird der Datensatz in Python mithilfe der Bibliothek pandas eingelesen. Der Datensatz enthält Rezensionen, die später analysiert werden.

Verwendete Bibliotheken:

- pandas: zum Einlesen und Bearbeiten der Daten
- re: zur Textbearbeitung und Mustererkennung (z.B. Entfernen von Sonderzeichen)

Um Textdaten für maschinelle Lernmodelle nutzbar zu machen, müssen sie in numerische Form umgewandelt werden. Dies geschieht durch die **Vektorisierung**. Wie in Phase 1 beschrieben sind zwei gängige Techniken hierfür **Bag of Words (BoW)** und **TF-IDF**.

Bei der **Bag of Words (BoW)** Technik wird jedes Dokument in eine Repräsentation umgewandelt, bei der gezählt wird, wie oft jedes Wort im Text vorkommt. Anschließend wird eine Matrix wird erstellt, in der jede Zeile und jede Spalte ein Wort darstellt, gefüllt mit der Häufigkeit des jeweiligen Wortes. Die Anzahl der Wörter kann durch einen Parameter wie max_features begrenzt werden (z.B. 500 häufigste Wörter).

Bei der **TF-IDF (Term Frequency-Inverse Document Frequency)** Technik wird die Häufigkeit eines Wortes mit seiner Seltenheit in dem gesamten Dokument gewichtet. Häufige, aber wenig bedeutende Wörter (z.B. "der", "und") werden so niedriger bewertet als seltene, aber bedeutende Wörter. Analog zu BoW wird hier eine Matrix erstellt, aber mit den TF-IDF-Werten anstelle der bloßen Häufigkeit. Die Ergebnisse der BoW- und TF-IDF-Analysen können dann als CSV-Dateien gespeichert werden.

Der **Coherence Score** bewertet, den Zusammenhang der identifizierten Themen. Ein höherer Coherence Score deutet auf sinnvollere Themen hin.

- **Bibliotheken**:
  - gensim.models.coherencemodel.CoherenceModel: zur Berechnung des Coherence Scores.
  - gensim.matutils.Sparse2Corpus: zur Umwandlung der BoW-Matrix in ein Format, das gensim versteht.
  - gensim.corpora.Dictionary: zur Erstellung eines Wörterbuchs.

Zur Berechnung wird die BoW-Matrix in ein Gensim-kompatibles Format umgewandelt (Sparse2Corpus). Anschließend wird ein LDA-Modell für verschiedene Anzahlen von Themen (z.B. 2 bis 10) trainiert. Der Coherence Score wird für jede Themenanzahl berechnet, indem überprüft wird, wie gut die Wörter innerhalb der Themen zusammenpassen. Die Themenanzahl mit dem höchsten Coherence Score wird als optimal gewählt.