Universität Leipzig

WiSe 23/24

Modulname: Verfahren und Anwendungen in den Digital Humanities

Modulnummer: 10-DIH-0001

Dozent: Dr. Andreas Niekler

Sarah Szydlowski, Matrikel-Nr. 3720829

Jiacheng Lang, Matrikel-Nr. 3701369

**Genre-Specific Topics In Literature**

## **0. Author**

Die Autoren dieser Arbeit sind Sarah Szydlowski und Jiacheng Lang. Sarah war für die Einführung, *related works* und den Korpus, Jiacheng für die Methode und die Ausgabe der Ergebnisse verantwortlich.

## **1. Introduction**

„It is […] impossible to talk about texts without a concept of the conventions that bind and divide them, and it is only through the relay-point of genre that literary discourse is at all possible.“[1]

In den Digital Humanitiesspielt die Anwendung von *Topic Modeling* auf literaturwissenschaftliche Fragestellungen schon lange eine wichtige Rolle. Als analytisches Werkzeug im Methodenkasten der Literaturwissenschaftler kann sie versteckte Muster in großen literarischen Datensätzen offenlegen, die nur durch *close reading* nicht hätten erfasst werden können. Da Literatur nicht im „luftleeren Raum“ geschrieben wird und sie sich daher stets auf andere Werke bezieht, kann sie als ein untereinander vernetztes und sich gegenseitig beeinflussendes Polysystem[2] verstanden werden. Genre als System hat also die wichtige Funktion, sich gegenseitig referenzierende Werke zu Clustern zusammenzusetzen, um literaturwissenschaftliche Diskurse zu ermöglichen. Wendet man nun *Topic Modeling* auf die Analyse von literarischen Genres an, so können sich neue Ansätze und Perspektiven auf Literatur eröffnen. Dies soll der Gegenstand dieses Projektes sein.

**Was ist Genre?**

Es gibt viele Definitionen für Genre, was dafür spricht, dass es sich dabei um ein nicht unumstrittenes Konzept handelt. Eine mögliche Definition beschreibt Genre als eine Textsammlung, die gekennzeichnet ist durch „*institutional imperatives*, that is as sets of rules and properties of certain texts established through convention across a community.“. [3]

Die Texte innerhalb eines Genres besitzen mindestens ein ähnliches Charakteristikum, wie z.B. ihr Thema, Protagonist, Handlungsraum, Zielpublikum, kommunikativer Zweck oder ihre formale Gestaltung und Länge. Zudem wird in der rechnergestützten Genreanalyse davon ausgegangen, dass die Texte eines Genres Eigenschaften auf Textebene besitzen, die eine quantitative Analyse aufdecken kann.[4]

Eine solche Definition geht allerdings von einer idealen Literatur aus, die sich klar in Genres einordnen lässt, ohne sich dabei je diffus zwischen zwei Grenzen zu bewegen. In der Realität ist das jedoch nicht in allen Fällen so einfach, wodurch Genre eher ein subjektives gesellschaftliches Konstrukt ist. Dadurch wird besonders die Einordnung von Werken aus der Belletristik, die einen immensen gestalterischen Bewegungsspielraum besitzt, oft zur Ermessensentscheidung.

Es werden nun die drei in dieser Arbeit untersuchten Genres vorgestellt.

**Fantasy**

Das Genre Fantasy zeichnet sich in der Literatur dadurch aus, dass ihre Handlung in Welten spielt, die technologisch unterentwickelt sind und die bevölkert werden von magischen Dingen und Leuten.[5]

**Science-Fiction**

In der Science-Fiction ist der Grundbaustein der Handlung die (Natur-)Wissenschaft. In der Handlung wird eine Realität dargestellt, die durch die Veränderung der Wissenschaft, der gesellschaftlichen Strukturen und des Menschen geprägt ist. Science-Fiction gehört ähnlich wie Fantasy zu der fantastischen Literatur, da eine „Realität imaginativ entworfen wird, die dem empirischen Wahrheitsbegriff nicht unterliegt und daher auch pseudowissenschaftliche Zukunftsentwürfe einschließt.[6]

**Kinderliteratur**

Das entscheidende Merkmal des Genres Kinderliteratur ist die Zielgruppe, denn sie richtet sich an noch nicht erwachsene Rezipienten, angefangen vom Kleinkinderalter bis hin zu jungen Erwachsenen. Sie werden für gewöhnlich von Erwachsenen speziell für diese Zielgruppen verfasst und entweder vorgelesen oder von Kindern und Jugendlichen selbst konsumiert. Besonders an der Kinderliteratur ist, dass sie eine große Anzahl an Genres und Gattungen beinhaltet, wodurch es kaum ein konsistentes Repertoire an Merkmalen gibt, was die Inhalte und Formen, Handlungsräume oder Protagonisten betrifft.[7] Anhand von drei Merkmalen kann ein Werk jedoch zur Kinderliteratur hinzugezählt werden: 1. Wenn das Werk für Kinder wichtige Werte und Kenntnisse vermittelt (= Erziehungs- oder Sozialisationsliteratur), 2. Wenn es angepasst ist an die Bedürfnisse des kindlichen Sprachvermögens, Intellekts und der kognitiven Fähigkeiten (= kindgemäße Literatur), 3. Und zuletzt solche Werke, die Kindern literarische Regeln beibringen sollen mithilfe einer einfachen, redundanten und Mündlichkeit nachahmenden Erzählweise (= Anfänger- oder Einstiegsliteratur).[8]

**Related Work**

Die Analyse von literarischen Genres war in den vergangenen Jahren bereits das Ziel einiger Forschungsarbeiten. Vier solcher Arbeiten verfolgten einen ähnlichen Ansatz wie auch dieses Projekt und sollen an dieser Stelle als *related work* vorgestellt werden.

Christof Schöch untersuchte in seiner Arbeit „Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama“[9] das Genre Drama und grenzte die Daten sprachlich-geografisch und epochal ein. Auch fokussierte er sich auf Subgenres, die in der Gattung Drama vorkommen. Dabei war es unter anderem sein Ziel herauszufinden, welche semantischen *Topics* in der Sammlung von französischen Dramen vorkommen, ob die verschiedenen Subgenres des Dramas ebenfalls differenzierbare und dominante *Topics* haben und inwieweit *Topic Modeling* Ergebnisse liefern kann, die mit klassischer Genreunterscheidung in Einklang sind. Er fand heraus, dass die *Topics* gut interpretierbar waren, da beispielsweise thematische *Topics* sich deutlich unterscheiden ließen von solchen, die das Figureninventar oder Setting als Hauptbestandteil hatten. Ebenfalls bestätigte sich seine Hypothese, dass auch Subgenres klar differenzierte *Topics* besitzen.

Vom selben Autor stammt auch die Arbeit „Gattungen des Kriminalromans: Ein quantitativer, *Topic*-basierter Zugang“[10], in der er mit derselben Zielsetzung und ebenfalls unter Verwendung von *Topic Modeling* Kriminalromane analysierte. Eine seiner Haupterkenntnisse war, dass die Themen des Kriminalromans nicht so spezifisch sind, wie man durch die klassische Vorstellung von dieser Gattung annehmen könnte. Nur ein kleiner Teil der entstandenen *Topics* passten zu dem, was in der Literaturwissenschaft semantisch dem Kriminalroman zugeordnet werden kann. Doch Schöch bewies auch, dass die speziellen *Topics* des Kriminalromans deutlich zum Vorschein kommen, sobald man sie mit den Topics aus anderen Genres vergleicht.

Interviews gehören nicht zur Belletristik, doch die Forschung von Greene et al. „Topic modelling literary interviews from The Paris Review“[11] befasst sich mit dem Genre literarisches Interview, und untersucht, welche Hauptthemen in diesem Genre durch *Topic Modeling* zu finden sind. Auch hier konnte belegt werden, dass *Topic Modeling* eine effektive Methode ist, um bestimmte thematische Strukturen in spezifischen Genres zu untersuchen. Doch wiesen die Autoren auch darauf hin, dass *close reading* und Fachwissen benötigt werden, um die generierten *Topics* interpretieren zu können.

Als letztes *related paper* ist die Arbeit „Reading the *Quan Tang shi*: Literary History, Topic Modeling, Divergence Measures“[12] von Broadwell et al. anzuführen. In ihr wurden die Gedichte aus der Tang-Dynastie unter anderem mit *Topic Modeling* untersucht. Ziel war es, inhaltliche Überschneidungen in einem sehr großen Korpus aufzudecken. Dieses Paper ist ein relevantes Beispiel dafür, dass Topic Modeling ein nützliches Werkzeug ist, um Literaturgeschichte zu erforschen.

**Forschungsziel**

In diesem Projekt werden drei Genres mit jeweils drei Subgenres unter Verwendung von Topic Modeling analysiert und interpretiert. Betrachtet werden die Genres: 1. Science-Fiction mit den Subgenres Utopien/Dystopien, Roboter und Zeitreisen. 2. Kinderliteratur mit den Subgenres historische Kinderliteratur, Schulgeschichten, Märchen. 3. Fantasy und die Subgenres Folklore, Übernatürliches, historisches Fantasy.

Ziel ist es, herauszufinden, welche spezifischen *Topics* in diesen drei Genres vorkommen, ob und wie sich die Subgenres thematisch voneinander abgrenzen, obwohl sie zum selben übergeordneten Genre gehören, und schließlich sollen auch die drei verschiedenen Genres thematisch miteinander vergleichen werden, um herauszufinden, wie sie sich unterscheiden oder auch ähneln.

## **2. Data**

Die Daten für das Projekt wurden aus dem Korpus des „Project Gutenberg“ (PG) bezogen. PG ist eine kostenlose und der Öffentlichkeit frei zugängliche Online-Bibliothek. Sie umfasst aktuell über 70,000 Werke, die frei vom Urheberrecht sind, da die Urheber vor über 70 Jahren gestorben sind oder die Werke anderweitig für die Öffentlichkeit zugänglich gemacht wurden. Für das gegenwärtige Projektkorpus wurden englische Originalversionen verwendet. Im PG wird zum Finden von Literatur unter anderem mit einem *Tag*-System,

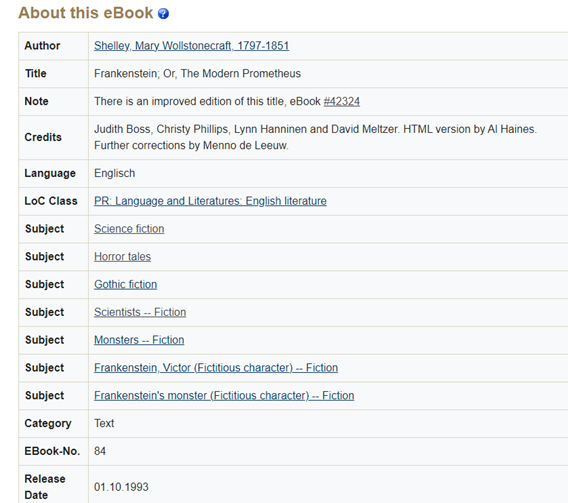


Abbildung 1: Beispiel des Tag-Systems

und „Bücherregalen“ gearbeitet.



Abbildung 2: Beispiel eines Bücherregals

Für die Datenbeschaffung wurden diese beiden Systeme als Wegweiser genutzt, um passende Belletristik für das Projektkorpus zu finden. Im ersten Schritt wurde entweder mithilfe eines *Tags* oder eines Bücherregals alle im PG vorhanden Werke eines bestimmten Genres gesucht. Im zweiten Schritt wurden die durch diese Suche eingegrenzten möglichen Bücher durch kursorisches *close reading* selektiert. Um entscheiden zu können, ob ein Werk tatsächlich zum Genre passt, wurden zusätzliche Informationen herangezogen, wie z.B. Inhaltszusammenfassungen. Die Auswahl erfolgte also auf Grundlage von subjektiver Entscheidung und Expertenwissen.

Für die drei Genres und ihre drei Subgenres wurden jeweils zehn Bücher ausgewählt, insgesamt also 120 Bücher. Bei der Auswahl der Daten wurde darauf geachtet, möglichst repräsentative Werke auszuwählen. Als repräsentativ gelten solche, die möglichst ‚typisch‘ für ihr Genre sind, z.B. sehr bekannte Bücher oder solche von renommierten Autoren. Jedoch war das aufgrund der Datenlage im PG nicht immer möglich, da es sich um einen relativ alten Bestand handelt.

## 3. Methode

**Text Preprocessing**

Die Originaltexte der Werke, die im Korpus enthalten sind, stammen aus SPGC-Tokens 2018-07-18. Diese Texte wurden bereits mit drei Funktionen verarbeitet: isalpha function, lower function sowie einem Satz- und Wort-Tokenizer von NLTK.

Für die Textvorverarbeitung wurde zuerst der tokenisierte Text in Listen zerlegt und die englischen Stoppwörter aus den Texten entfernt. Als nächstes wurden die Wörter in den Listen mit POS-Tags versehen. Die Sprache der Texte im Korpus ist Englisch. Die Wörter in den Texten werden POS-tagged als Adjektiv, Verb, Substantiv und Adverb. In der Arbeit werden die Wörter, die zu den Kategorien Adjektiv, Verb und Adverb gehören, aus den Texten entfernt. Nur Wörter der Kategorie Substantiv werden beibehalten.

Da das Englische eine flektierende Sprache ist, wird eine Methode benötigt, um diese Wörter in ihre Stammformen umzuwandeln. Beim Stemming wird eine feste Regel verwendet, um Suffixe abzuschneiden. Die Lemmatisierung hingegen ist eine kontextuelle Analyse der Wörter. Aus diesem Grund ist die Lemmatisierung eine bessere Lösung für das Problem als das Stemming. Beibehaltene Wörter werden von WordNetLemmatizer lemmatisiert.

Es gibt viele Figuren in den Erzählungen, und in einigen von ihnen kommen die Figurennamen sehr häufig vor. Es ist wahrscheinlich, dass die Namen dieser Charaktere in den Topics vorkommen, allerdings sind sie keine sinnvollen Topic-Konstituenten. Wir versuchen, das NER-Paket von NLTK und Spacy zu verwenden, um das Problem zu lösen, aber diese Methoden können unsere Bedingungen nicht perfekt erfüllen. Einige Namen können nicht korrekt als Personennamen erkannt werden, einige andere Wörter, die keine Personennamen sind, werden als Namen erkannt, so dass zu viele irrelevante Wörter durch diese NER-Methoden aus dem Text entfernt werden. Letztendlich haben wir uns für eine Kombination aus manueller Beurteilung und NER-Methoden entschieden, um die häufigsten Charakternamen der Romane in einer Liste zu speichern und sie aus den Texten zu entfernen.

**Directory Preprocessing**

Der Zweck des Directory Processings ist es, die Berechnung der Topic-Verteilung von Subgenres zu ermöglichen. Durch das Processing erhalten wir den Index der Texte eines bestimmten Subgenres. Der Index gibt die Position der Texte eines bestimmten Subgenres, z.B. Roboter, im Directory des Genres, z.B. Science Fiction, an.

**LDA Model Training und Auswahl**

In unserem Projekt verwenden wir models.ldamodel von Gensim, um unser LDA-Modell zu trainieren. Der core estimation code basiert auf einem online variational Bayes Algorithm for Latent Dirichlet Allocation [13]. Bevor wir Topic-Modelle für die Textanalyse verwenden, müssen wir erst ein LDA-Topic-Modell trainieren. Zunächst müssen die vorverarbeiteten Texte in einen "Bag of Words" umgewandelt werden, der als Input für das Modell dient, und ein Textverzeichnis erstellen. Während des Trainings des LDA-Modells passen wir die Parameter des Modells an und evaluieren es mit Hilfe einer quantitativen Methode wie der Berechnung der Kohärenz von Topics und der menschlichen subjektiven Beurteilung durch das Chart von pyLDAvis. Die beiden Hauptparameter wurden angepasst: die Anzahl der Topics und die Anzahl der hochfrequenten Wörter, die aus dem Dictionary entfernt werden müssen.

**Quantifizierung der Qualität des Modells durch Topic Coherence**

Ein Topic setzt sich aus bestimmten Begriffen zusammen. Wenn diese Begriffe sich gegenseitig ergänzen, können wir sie als kohärent bezeichnen. Ein coherent fact set kann in einem Kontext interpretiert werden, der alle oder die meisten der facts umfasst [14]. Es gibt verschiedene Arten von Kohärenzmaßen, z. B. C\_v, C\_umass usw. Wir haben uns für die Verwendung des C\_v-Maßes für das CoherenceModel entschieden. Das C\_v measure basiert auf einem sliding window, einer One-Set-Segmentierung der Top-Wörter und einem indirekten Bestätigungsmaß, das normalisierte pointwise mutual information (NPMI) und die Cosinus-Ähnlichkeit verwendet [15].

Wir trainieren die LDA-Modelle mit verschiedenen Parametern und berechnen dann den Kohärenzwert dieser Modelle. Je höher der Score, desto kohärenter sind die Topics. Wir können die Modelle objektiv vergleichen und mehrere Modelle mit höheren Werten für weitere Tests beibehalten.

**Die Anzahl der Topics**

Die Wahl der Anzahl der Topics hat erhebliche Auswirkungen auf die Qualität eines Topic Models. Es beeinflusst die Granularität der Topics. Eine höhere Anzahl von Topics kann eine komplexere Struktur der Daten darstellen und Nuancen und Unterschiede innerhalb des Korpus erfassen. Umgekehrt kann die Auswahl von weniger Themen das Modell zu sehr vereinfachen, was zum Verlust wichtiger thematischer Details führt. Der Vorteil ist jedoch, dass es sich besser an verschiedene Datensätze anpassen lässt. Im Hinblick auf Kohärenz und Interpretierbarkeit stellt eine optimale Anzahl von Topics ein Gleichgewicht dar, das umfassend genug ist, um verschiedene Themen in den Daten zu erfassen und gleichzeitig Redundanz und Störungen zu vermeiden.

| number | 5 | 10 | 15 | 20 | 25 |
| --- | --- | --- | --- | --- | --- |
| coherence | 0,69 | 0,68 | 0,66 | 0,66 | 0,66 |

Tabelle 1: Coherence score des Topic Models mit verschiedenen num\_topics

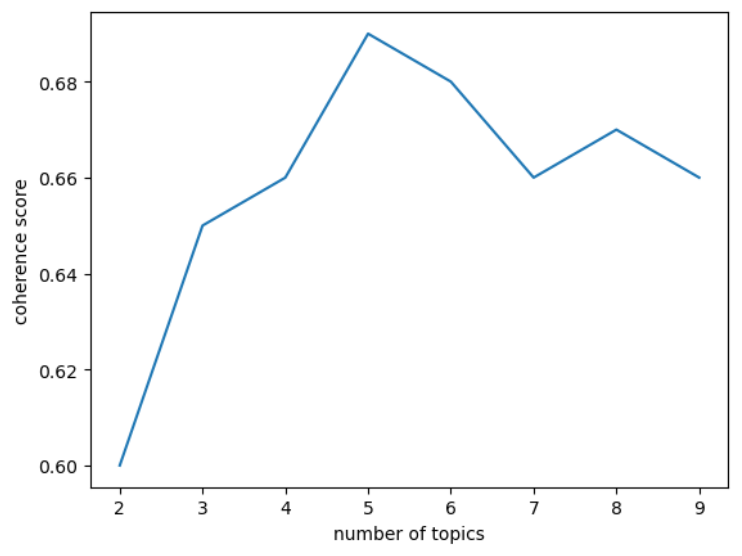
****

Abbildung 3: Coherence score von Topic Modellen mit unterschiedlichen num\_topics

Verwenden wir den Science-Fiction-Datensatz, so zeigt Tabelle 1 zwischen 5 und 25 Topics einen Abwärtstrend des Koherenzwertes. Danach gibt es weitere Untersuchungen zur Situation zwischen 2 und 9 Topics. Abbildung 3 zeigt, dass von 2 bis 5 Topics der Kohärenzwert des Modells einen Aufwärtstrend aufweist, von 5 bis 9 Topics zeigt der Kohärenzwert einen umgekehrten Trend. Somit bestätigen wir die Anzahl der Topics als fünf für unser LDA-Modell.

**Entfernen der Frequency Terms**

Obwohl einige default Stoppwörter aus dem Korpus entfernt wurden, gibt es immer noch viele Wörter mit hoher Frequenz im Verzeichnis, die relativ allgemeine Bedeutungen haben, was der Leistung des Modells nicht zuträglich ist. Wir hoffen, dass die vom Modell generierten Topics spezifischere und bedeutungsvollere Wörter enthalten können. Wie viele Wörter mit der höchsten Frequenz aus dem Verzeichnis entfernt werden sollten, ist der erste Parameter, der getestet werden muss. Wir testen drei Situationen für diesen Parameter: Wir behalten das ursprüngliche Verzeichnis als Kontrollgruppe bei, entfernen die ersten 100 Wörter und die ersten 200 Wörter als zwei weitere Situationen. Tabelle 2 zeigt die Kohärenzwerte des Modells in den drei Situationen:

| Entfernte Wörter | Original | 100 Wörter | 200 Wörter |
| --- | --- | --- | --- |
| Kohärenz | 0,26 | 0,63 | 0,69 |

Tabelle 2: Kohärenzwert von Topic-Modellen mit unterschiedlichen Wortlisten

Die Kohärenz des Modells erhöht sich deutlich, wenn wir die häufigen Wörter aus dem Dictionary streichen. Für das Training des Modells haben wir uns entschieden, die ersten 200 Wörter aus dem Dictionary zu entfernen.

**Subjektive Beurteilung mit pyLDAvis**

Wir hoffen, dass sich die vom Modell generierten Topics so wenig wie möglich wiederholen, aber die Berechnung des Kohärenzscores kann dies nicht gewährleisten. Im pyLDAvis-Chart gibt es einige Blasen, die den Topics des LDA-Modells entsprechen. Je größer die Blase ist, desto größer ist der Anteil der Texte im Korpus, die zu diesem Topic gehören.

Bei der subjektiven Beurteilung des LDA-Modells wollen wir vor allem feststellen, ob es einige Topics gibt, die den anderen zu ähnlich sind. Wir wollen nicht, dass die Topics eine Menge sich wiederholender Begriffe haben. Wir hoffen, dass diese Topics verschiedene Aspekte der gesamten Dokumente im Korpus repräsentieren. Im pyLDAvis-Chart bedeutet das, je weiter die Blasen voneinander entfernt sind und sich nicht überschneiden, dass diese Topics unterschiedlichere Bedeutungen haben.

Zum Beispiel haben das LDA-Modell mit 5 Topics und das mit 10 Topics beide einen hohen Kohärenzwert. In dieser Situation müssen wir ihre pyLDAvis-Charts vergleichen, um festzustellen, welche Anzahl von Topics des Modells wir wählen sollten. In Abbildung 4 ist zu sehen, dass sich diese 5 Topics nicht überschneiden und weiter voneinander entfernt sind. In Abbildung 4 überschneiden sich einige Topics. Daher wählen wir schließlich manuell die Anzahl der Topics als 5 für unseren nächsten Schritt der Arbeit aus.

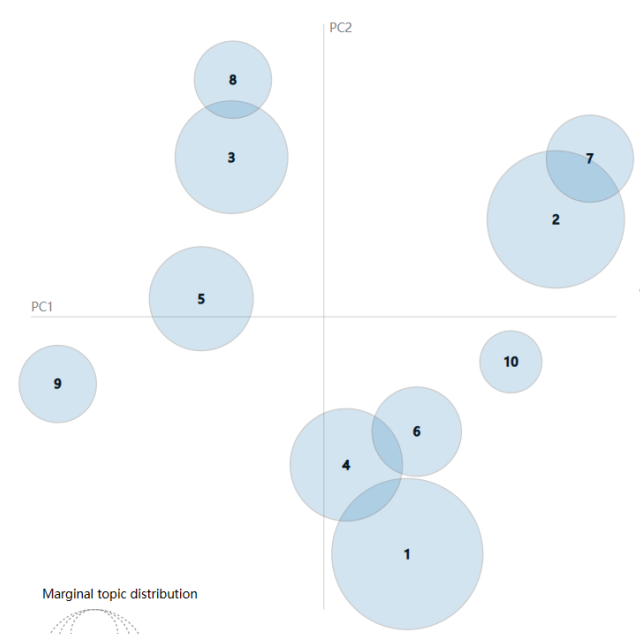
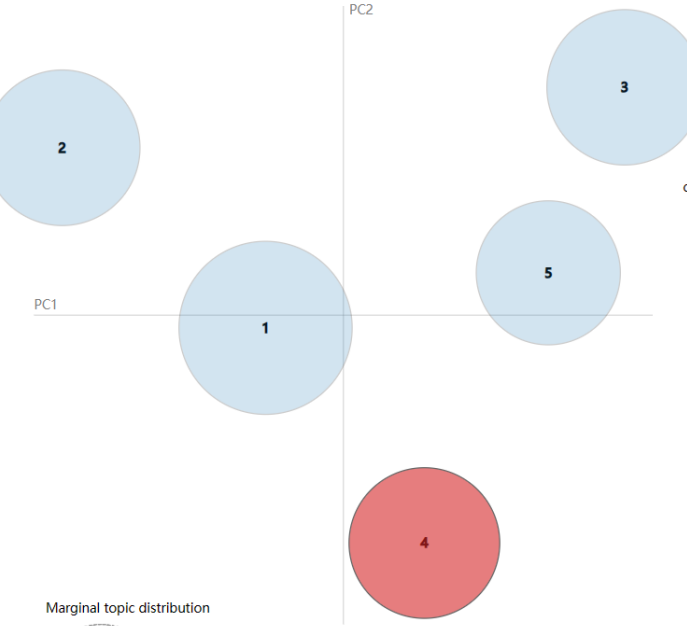


Abbildung 4: pyLDAvis zeigt die Topics des Modells num\_topics 5 & 10

**Die Methode der Topic-Verteilung**

Am Ende setzen wir num\_topics auf fünf, passes gleich zehn, alpha und beta auf automatisch und random\_state auf eine feste ganze Zahl, um das Ergebnis des Modells stabil zu halten. Wir trainieren ein LDA-Modell für jedes Level-1-Genre: Science Fiction, Kinderliteratur und Fantasy. Mit der eingebauten Funktion get\_document\_topics from lda\_model in Gensim können wir die Topic-Verteilung für das gegebene Dokument ermitteln. Jedem Topic wird eine ID und die entsprechende Wahrscheinlichkeit zugewiesen. Über das Directory Processing erhalten wir eine Liste der Topic-Verteilungen aller Dokumente desselben Subgenres. Schließlich können wir die durchschnittlichen Wahrscheinlichkeitswerte der Liste der Topic-Verteilungen als die Topic-Verteilungen der Subgenres zuordnen.

## 4. Result

**Wordclouds der Topics von Level-1-Genres**

Aus unserem Ergebnis ergeben sich einige typische Topics für verschiedene Level-1-Genres. Von den fünf Themen, die das Modell generiert, ist nicht jedes Thema mit dem Genre verbunden. Wir zeigen diese Themen im Zusammenhang mit Genre unten in einer Wordcloud.

In Abbildung 3 sehen wir, dass die Hauptwörter in Topic 1 mit Begriffen aus der Natur zu tun haben, insbesondere mit Begriffen, die mit dem Meer zusammenhängen, z.B. Insel, Schiff, Meer, Kapitän, Fels und Stein. Topic 2 bezieht sich auf das Konzept von Land und Gesellschaft. Topic 4 scheint sich auf die industrielle Produktion zu beziehen.

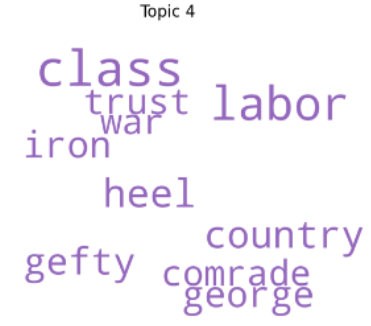
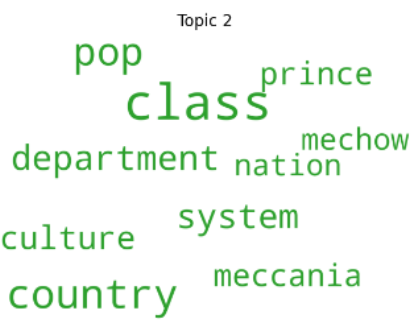
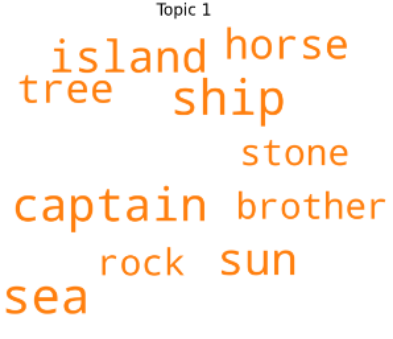


Abbildung 5: Die typischen Topics für das Genre Science-Fiction

In Abbildung 5 sehen wir, dass Topic 2 Schlüsselwörter wie „Fee“, „Vogel“, „Beere“, „Prinz“, „König“ und „Wald“ enthält, die immer in Märchenbüchern vorkommen. Topic 4 enthält das Wort „Schule“ als Kennwort.

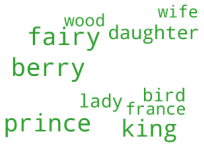


Abbildung 6: Die typischen Topics des Genres Kinderliteratur

In Abbildung 6 sehen wir, dass es in beiden Topics einige Fabelwesen/Kryptide oder Tiere gibt, z.B. Pigasus, Phönix, Kojote und Pferd. Fabelwesen/Kryptide [16] sind Kreaturen, von denen Kryptozoologen annehmen, dass sie in natürlichen Lebensräumen vorkommen könnten, deren Existenz derzeit jedoch weder wissenschaftlich bestätigt noch anerkannt ist.



Abbildung 7: Die typischen Topics des Genres Fantasy

**Topic-Verteilung von Subgenres**

Abbildung 8 zeigt, dass das Modell für das Subgenre Roboter eine höhere Wahrscheinlichkeit für Topic 4 anzeigt, aber eine geringere Wahrscheinlichkeit für Topic 3 als andere. Die Subgenres Zeitreise und Utopie können durch Topic 0 und Topic 2 unterschieden werden.

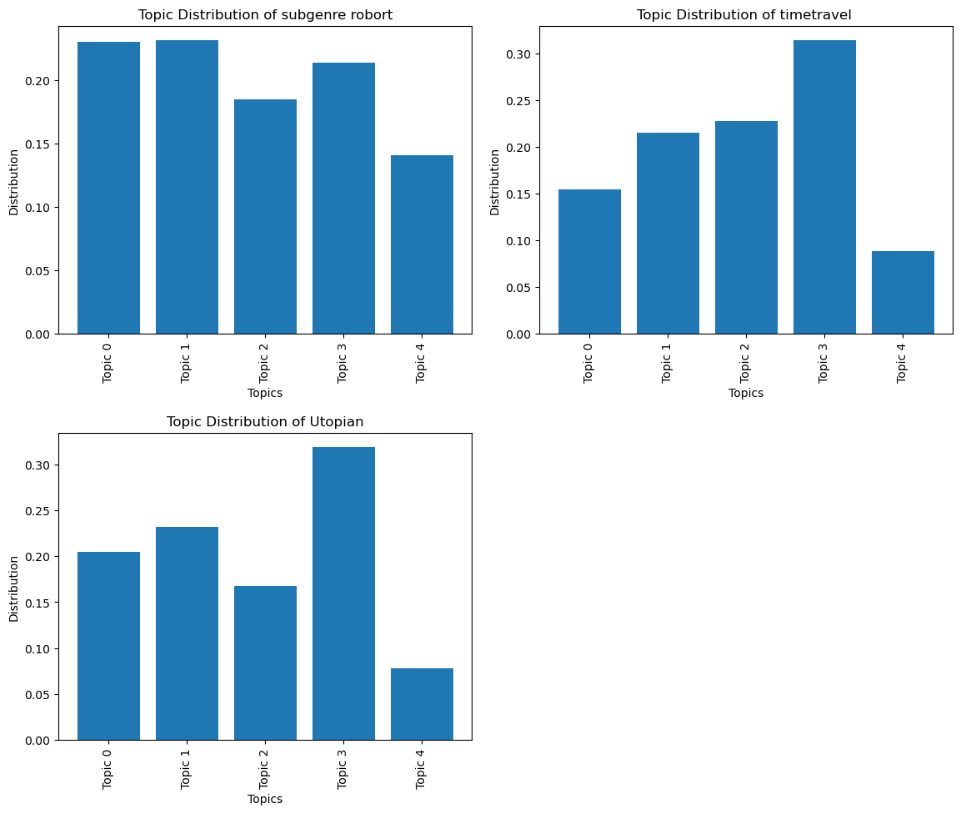
****

Abbildung 8: Topic-Verteilung der Subgenres Roboter, Zeitreise und Utopie

Abbildung 9 zeigt den Hauptunterschied in der Topic-Verteilung zwischen den Subgenres Märchen und Schulen bei Topic 0 und Topic 4.

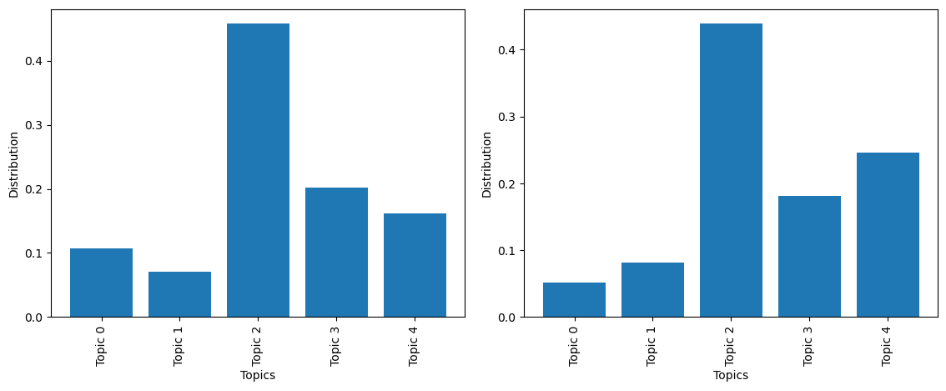
****

Abbildung 9: linkes Bild: Märchen-Subgenre, rechtes Bild: Schul-Subgenre

Abbildung 10 zeigt, dass die Topic-Verteilung des Fantasy-Subgenres „historisch“ und Folklore relativ ähnlich ist. Sie unterscheiden sich nur geringfügig in der zugeordneten Wahrscheinlichkeit von Topic 3.

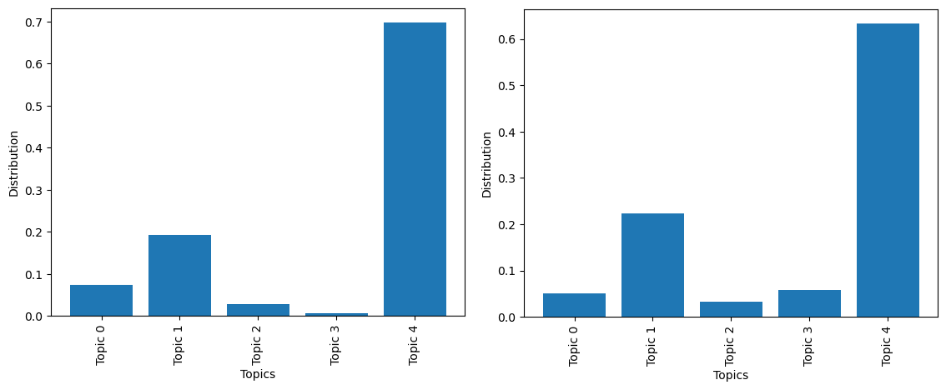


Abbildung 10: linkes Bild: f\_historical subgenre, rechtes Bild: Folklore subgenre

Es gibt andere nicht-signifikante Ergebnisse, die man in den Code-Skripten sehen kann.

## 5. Conclusion

Es fällt auf, dass sich die Begriffe in den generierten Topics sich in fünf Kategorien einsortieren lassen: Figureninventar, Schauplätze, Requisiten und Konzepte. Eine weitere Kategorie wären Begriffe, die ohne Kontext nicht verständlich sind (Sonstige).

| Figuren | Schauplätze | Requisiten | Konzepte | Sonstiges |
| --- | --- | --- | --- | --- |
| Captain  Prince  Comrade  Wife  Fairy  Daughter  Brother  Lady  King  Captain  Doctor  \*Miss  Aunt  God | Island  Sea  Department  Wood  France  City  Ship  School  Mountain  Country | Horse  Stone  Rock  \*System  \*Iron  Berry  Tree  Bird  \*Chapter  Dollar  Food  Pigasus  Clef  Horse  Sun  Ruby  Phoenix  Coyote  Carpet | \*Class  Culture  Trust  \*Heel  Minute  Everything  Nation  Soul  \*Youth  Labor | Pop  Meccania  Gefty  Robert  George  Mechow  Tressa |

Die mit einem Stern markierten Begriffe können mehrdeutig sein.

Die im Report vorgestellten Topics besitzen alle mindestens eine Figur, einen Schauplatz und eine Requisite. Einige haben zudem noch ein Konzept. Mit Hilfe von Topic Modeling kann es also möglich sein zu sehen, in welchem Genre was für Figuren typisch sind, wo sie sich aufhalten und mit “was” etwas gemacht wird. Dabei ist dieses “etwas” jedoch die Nullstelle, denn in diesen Ergebnissen gibt es keine Handlungsmarkierer. Es lässt sich jedoch so viel sagen, dass in manchen (Sub-)Genres besonders viele Figuren auftreten, wie im Topic 2 der Kinderliteratur deutlich wird. Dadurch könnte man die Hypothese aufstellen, das im Subgenre Märchen die Figuren ein wichtiger Bestandteil sind, welche die Handlung vorantreiben, das Genre ist wäre somit eher charaktermotiviert. Im Kontrast dazu besitzt Topic 2 des Genres Science Fiction überwiegend Konzepte, wodurch man annehmen kann, dass für dieses Subgenre eine Art “Kommentar der Gesellschaft” maßgeblich ist. Außerdem wird deutlich, dass je höher die Anzahl an “sonstigen” Begriffen ist, umso größer die Wahrscheinlichkeit, dass es sich um ein wenig mimetisches, von der Realität abweichendes Genre handelt, wie die Topics 2 und 4 von Science-Fiction suggerieren. In diesem Aspekt weichen sie zum Beispiel vom Subgenre Märchen ab, dass zwar ebenfalls ‘fantastisch’ ist, aber wohl doch stärker die Realität nachahmt.

Aus den Ergebnissen geht hervor, dass die Genres der Level-1-Ebene einige klare, typische Topics haben, welche die Genres unterscheiden. Bei den Subgenres können wir, wenn wir die Subgenres detaillierter festlegen, z. B. Roboter versus Zeitreise, deutlichere Unterscheidungsmerkmale finden. Wenn wir Subgenres wie Folklore oder Historical unterscheiden wollen, wird es schwieriger.

Um die typischen Topics für Genres und Subgenres besser zu finden, gibt es unserer Meinung nach einige Möglichkeiten: Erstens sollte der Korpus optimiert werden. Es kann mehr Originaltexte enthalten. Dreißig bis fünfzig Bücher für ein Genre sind nicht zu viel. Wir sollten relativ detaillierte Subgenres als Forschungsziel wählen und zu allgemeinen Subgenres vermeiden. Zweitens können die Namen von Personen, Orten und Organisationen das Ergebnis stark verwirren. In einem Werk gibt es einige Namen von Personen, die immer wieder erwähnt werden. Dies ist ein wesentlicher Unterschied zum Umgang mit Produktbewertungen oder kurzen Texten auf Twitter. Eine bessere Lösung wäre, diese speziellen Begriffe vor der Tokenisierung zu entfernen. Die Verwendung einer besseren NER-Lösung ist ebenfalls wichtig, da sie diese Begriffe präziser erkennen kann und die Wahrscheinlichkeit falscher Entfernungen verringert. Drittens können wir verschiedene Werte des Parameters alpha testen, der die markanten Topics hervorheben und den Einfluss der bedeutungslosen Topics abschwächen kann. Viertens kann die Suche nach einer geeigneten Methode und einem Modell zur Quantifizierung des Unterschieds zwischen den Themen der Genres der Level-1-Genres ebenfalls eine Richtung sein, an der gearbeitet werden kann.

Für aussagekräftigere Ergebnisse sollte sehr genau kontrolliert werden, welche Werke in den Korpus aufgenommen werden, damit die Ergebnisse sich tatsächlich auf ein Genre verallgemeinern lassen. Bei einem kleinen Korpus ist eine Probe auf die Eignung noch mit kursorischem Close-Reading umsetzbar. Bei über 50 Werken ist es allerdings nicht möglich und auch nicht sinnvoll, so eine Prüfung durchzuführen. Will man das Korpus des PGs verwenden, kann man zwar durch das Tag-System die Literatur eines Genres finden, doch es ist nicht gewährleistet, dass das Werk: 1. ins richtige Genre einsortiert wurde, und 2. das es repräsentativ für das Genre ist. Das Tagging wird kollaborativ von Freiwilligen im Projekt durchgeführt, dabei ist es nicht klar, ob es sich um Literaturwissenschaftler handelt oder um Laien, die ein Genre-Tag rein nach ihrer eigenen subjektiven Einschätzung vergeben.

Doch auch wenn das Korpus von einem Experten zusammengestellt wird, ist die Realität, dass die meiste Belletristik zu individuell ist und daher nur selten der literaturwissenschaftlichen Theorie folgt. Da es in der Natur der Sache liegt, kann es also kaum einen perfekten Korpus geben, der nur aus typischen, mustergültigen Werken eines bestimmten Genres besteht. Daher stellt sich die Frage, ob durch computergestützte Genreanalyse überhaupt generalisierbare Aussagen über ein ganzes Genre getroffen werden können oder ob man vielmehr nur Aussagen über ein Fragment von Genre trifft, so wie es sich im Korpus des Projekts präsentiert.

**Literaturverzeichnis**

[1] Rayment, A. (2014). *Fantasy*. Amsterdam: Editions Rodopi B.V. S. 13

[2] Vegleiche dazu Even Zohars Polysystemtheorie

[3] Schöch, C. (2022): Computational Genre Analysis. The Dragonfly’s Gaze. Marseille: OpenEdition. URL: https://dragonfly.hypotheses.org/1219, DOI (archive copy): 10.5281/zenodo.7553854. Letzter Zugriff: 29.03.2024

[4] Ebd.

[5] Rayment, A. (2014). *Fantasy*. Amsterdam: Editions Rodopi B.V. S. 15-18

[6]

Göller, K. (2012). Literatur im Banne der Zukunft: Science Fiction. Regensburger Universitätszeitung. URL: https://epub.uni-regensburg.de/26745/1/ubr13495\_ocr.pdf ,DOI: 10.5283/epub.26745. S. 1, Letzter Zugriff: 29.03.2024

[7] Kümmerling-Meibauer, B. (2012). *Kinder- und Jugendliteratur: Eine Einführung*. Darmstadt: WBG (Wiss. Buchges.). S. 9

[8] Ebd. S. 17

[9] Schöch, C. (2021). Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama. http://digitalhumanities.org:8081/dhq/vol/11/2/000291/000291.html

[10] Schöch, C. (2017). Gattungen des Kriminalromans: Ein quantitativer, topic-basierter Zugang. In: Lang, S., Schöch, C. Gattungen des Kriminalromans. In: Dialogische Krimianalysen: Fachdidaktik und Literaturwissenschaft untersuchen aktuelle Krimiliteratur aus Belgien und Frankreich, 37-59. Frankfurt am Main: Peter Lang, 2017. – DOI: 10.5281/zenodo.8328336.

[11] Greene, D., O'Sullivan, J., O'Reilly, D.. (2024). Topic modelling literary interviews from The Paris Review. Digital Scholarship in the Humanities. 10.1093/llc/fqad098. Letzter Zugriff: 29.03.2024

[12] Broadwell, P., Chen, J., Shepard, D. (2019). Reading the Quan Tang shi: Literary History, Topic Modeling, Divergence Measures. http://digitalhumanities.org:8081/dhq/vol/13/4/000434/000434.html. Letzter Zugriff: 29.03.2024

[13] Hoffman, Matthew, et al. ‘Online Learning for Latent Dirichlet Allocation’. Advances in Neural Information Processing Systems, vol. 23, Curran Associates, Inc., 2010. Neural Information Processing Systems, https://papers.nips.cc/paper\_files/paper/2010/hash/71f6278d140af599e06ad9bf1ba03cb0-Abstract.html.

[14] Röder, Michael, et al. ‘Exploring the Space of Topic Coherence Measures’. Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, Association for Computing Machinery, 2015, pp. 399–408. ACM Digital Library, https://doi.org/10.1145/2684822.2685324.

[15] Kapadia, Shashank. ‘Evaluate Topic Models: Latent Dirichlet Allocation (LDA)’. Medium, 24 Dec. 20https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0.

[16] ‘List of Cryptids’. Wikipedia, 15 Feb. 2024. Wikipedia, <https://en.wikipedia.org/w/index.php?title=List_of_cryptids&oldid=1207495110>.

Abb. 1: https://www.gutenberg.org/ebooks/84, Screenshot aufgenommen am 28.03.2024

Abb. 2: https://www.gutenberg.org/ebooks/bookshelves/search/?query=children%7Cchristmas%7Cchild%7Cschool, Screenshot aufgenommen am 28.03.2024