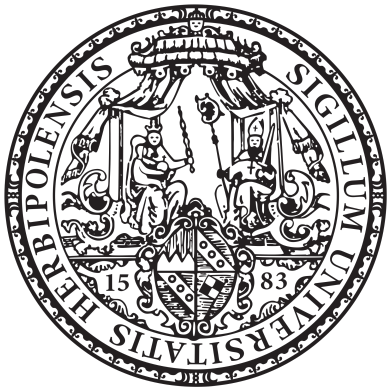
Autorschaftsattribution mithilfe von Machine Learning Verfahren und stilometrischen Verfahren (NUR DELTA). Eine vergleichende Analyse (TODO)

Jan Paulus



Seminararbeit

Institut für deutsche Philologie
Lehrstuhl für Computerphilologie und Neuere Deutsche Literaturgeschichte

Dozent: Thorsten Vitt

Würzburg, den 14.02.2020 (TODO)

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Zusammenfassung

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

1 Einleitung

Bevor Burrows 2001 sein Delta-Maß vorstellte, war Forschungsstand im Bereich der Stilometrie, eine begrenzte Anzahl an voraussichtlich ähnlichen Texten miteinander zu vergleichen (Evert u. a., 2017, S. ii5). Diese Vorgehensweise wird auch „closed games“ genannt (Burrows, 2002, S. 267). Bei „open games“ hingegen gibt es hinsichtlich eines unbekannten Textes zuvor keinen Anhaltspunkt, welchen Kandidaten-Texten der unbekannte Text ähnlich ist. Das Ziel von Burrows Verfahren war es deshalb, diese „open games“ in „closed games“ zu transformieren (Burrows, 2002, S. 268). Burrows verwendet dafür das Delta-Maß, welches die Ähnlichkeit über die Distanz von einem unbekannten Text zu einer Gruppe von Texten berechnet (Burrows, 2002). Dieses Maß hat die folgenden Jahren der Stilometrie-Forschung geprägt (Evert u. a., 2017, ii5f.). Als Alternative zum Delta-Maß wurden in den letzten Jahren auch vermehrt Machine Learning Verfahren wie K-Nearest Neighbors, Nearest Shrunken Neighbors oder Support Vector Machines verwendet (Jockers und Witten, 2010, S. 217). In dieser Arbeit soll Burrows Delta mithilfe von Nearest-Neighbor Verfahren als Klassifizierungsverfahren modelliert(AW?) werden. MEHR Daraufhin soll ein Vergleich von Burrows Delta mit weiteren Klassifizierungsverfahren wie Support Vector Machines, Multinomial Naive Bayes und Logistic Regression durchgeführt werden.

TODO:

- ZIEL DER ARBEIT: „Das Ziel dieser Arbeit ist es, Burrows Delta mit einigen ausgewählten Machine Learning Klassifizierungsverfahren sowie Deep Learning Methoden (BERT) zu vergleichen. Dabei sollen die zu untersuchenden Verfahren nicht nur auf Genauigkeit getestet werden, sondern auch die Komplexität der Anwendung und Implementierung soll ein Bewertungsfaktor sein. Für die Vergleiche werden verschiedene Versuchszenarien erstellt wie feature-anzahl-veränderung usw. Diese Ergebnisse sollen nach einer Analyse ebenfalls in die Vergleichsbewertung mit einfliessen. “
- hier werden quasi „open games“ behandelt, auch wenn Korpus etwas eingeschränkt ist
- IN DIESER ARBEIT: Autorschaftsattribution (Teilgebiet der Stilometrie: https://programminghistorian.org/en/lessons/introduction-to-stylometry-with-python)

TODO:

- ZETA auch als Maß nehmen (vllt noch weitere? und dann bert weglassen?) (ICH: eher nicht)
- RUDER angucken (siehe notizen)

- weiter literatur suchen mit begriff äuthorship autribution"
- NSC: https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.NearestCentroid.ht
-

2 Überblick über die Korpora

TODO: austauschen

Für die Experimente in dieser Arbeit sollen zwei Korpora benutzt werden, um einen aussagekräftigen Vergleich der Stilometrie- und Klassifizierungsverfahren zu gewährleisten. Zudem wurde auf Basis dieser beiden Korpora mithilfe von Segmentierungen und anderen Auswahlkriterien (TODO: habe ich das gemacht?) noch zwei weitere Versionen erstellt, die ebenfalls für die Vergleiche der stilometrischen Methoden und den Machine Learning Klassifizierungsverfahren herangezogen wurden. Das erste Korpus besteht aus Reden von deutschen Politikern aus dem 21. Jahrhundert (Barbaresi, 2018). Für die Benutzung in dieser Arbeit wurde es etwas bearbeitet. Zuerst wurden die Reden tokenisiert. Danach wurden die Namen der Redner aus den Reden entfernt, da dies eventuell die Klassifizierungen verfälschen könnte. Weiterhin wurde das Korpus auf 13 Redner mit jeweils 10 Reden gekürzt. Um eine Einheitlichkeit hinsichtlich der anderen Korpora (TODO: oder Korpus) zu gewährleisten, werden die Redner innerhalb dieser Arbeit als Äutoren"bezeichnet und ihre Reden als "Texte".

Das zweite Korpus stellt eine Variante des "Corpus of German-Language Fiction"von Frank Fischer und Jannik Strötgen dar (Fischer und Strötgen, 2017). Ihr Korpus ist eine extrahierte und konvertierte Version von Werken aus dem "Projekt Gutenberg-DE". Das Korpus ist zweigeteilt: Der Großteil besteht aus deutschen Werken von deutschsprachigen Autoren, der kleinere Teil aus Werken von nicht-deutschsprachigen Autoren, die ins Deutsche übersetzt wurden. Der zweite Teil wird in dieser Arbeit ignoriert, da er einige Probleme aufweist wie teilweise nicht übersetzte Texte und fehlende Erscheinungsjahre. Der Teil des Korpus mit Werken von deutschsprachigen Autoren besteht aus 2735 Prosa Werken von 549 verschiedenen Autoren. Die Erscheinungsjahre erstrecken sich dabei von 1510 bis 1940, wobei der größte Teil der Werke zwischen 1840 und 1930 erschienen ist.[1] Die Einteilung der Werke in die Gattung Prosa ist sehr vage, da die Prosagattungen sehr mannigfaltig sind (Bücher-Wiki, 2019) (TODO: Alternative?). Anhand des Korpus ist nicht erkennbar, zu welcher Prosagattung die einzelnen Werke gehören. Einige der Werke geben ihre Gattung zu Beginn des Textes an. Leider ist dies bei nur sehr wenigen Werken der Fall, eine einheitliche Angabe der Textgattung ist beim Original-Korpus nicht enthalten. Laut Angaben der Ersteller enthält das Korpus "mainly novels and short stories"(Fischer und Strötgen, 2017). Das Korpus von Fischer und Strötgen wird für die Untersuchungen in

[1] Dies wird auch durch Abbildung TODO bestätigt.

dieser Arbeit vorverarbeitet und reduziert, die genaue Vorgehensweise befindet sich in Appendix B.

TODO hier: Segmentierung und andere Auswahl!

- Segmentierung machen der Romane? damit bekommt man mehr romane, dafür sind die segmente nicht so lang (siehe SCHÖCH, zeta)
- experimentaufbau in kapitel 2 von SCHÖCH, zeta angucken

3 Stand der Forschung

Eines der meist benutzten und ältesten Methoden der Stilometrie ist die Autorschaftsattribution. Die ersten Herangehensweise des Autorschaftsattributions-Probleme wurden bereits im späten 19. Jahrhundert entwickelt und werden werden „einheitliche invarianten Ansätze“ (engl.: „Unitary Invariant Approach “) genannt (Argamon, Koppel und Schler, 2009, S. 10). 1964 nutzten Mosteller und Wallace multivariaten Analyse-Ansätze (engl.: „Multivariate Analysis Approach“) bei der Untersuchung der Federalist Papers (Argamon, Koppel und Schler, 2009, S. 10). Der grundlegende Gedanke hinter diesen Ansätzen ist es, dass durch eine Darstellung aller zu untersuchenden Dokumente in einem mehrdimensionalen Raum der Autor eines unbekannten Dokuments durch ein Distanzmaß ermittelt werden kann: Der Autor des Dokuments, welches am nächsten zum unbekannten Dokument in diesem mehrdimensionalen Raum liegt, ist wahrscheinlich auch der Autor des unbekannten Dokuments (Argamon, Koppel und Schler, 2009, S. 11). Eine der bekanntesten dieser Methoden ist Burrows Delta, welches nach seiner Einführung 2001 maßgeblich die Autorschaftsattributions-Forschung der folgenden Jahre prägte (Evert u. a., 2017, ii5f). (TODO ICH: hier burrows delta ausführlicher erklären? kann text strecken. ICH: einfaches Maß). Burrows Delta funktionierte in der Praxis sehr gut, jedoch waren die genauen Funktionsweisen (AW?) bis zur Erklärung durch Argamon ungeklärt. Der Ansatz von Argamon lieferte ein besseres Verständnis der grundlegenden Annahmen des Delta-Maßes, eine Einschränkung der Methoden sowie theoretisch fundierte Variationen und Erweiterungen des Maßes (Argamon, 2008). Eine dieser Variationen ist die Auffassung von Burrows Delta als achsengewichtetes Nearest Neighbor Klassifizierungsverfahren (Argamon, 2008, S. 132–135).

jockers 2010: zusammenfassung bei eder 2015 (169f.)

eder 2015: svm + n-grams angeblich beste kombo (koppel 2009, stratamos 2009); andere meinung ist, dass delta gleich gut wie svm (jockers 2010); n-gram nicht immer gut (eder 2015)

auch erwähnen: neben der Verwendung von Machine Learning Klassifizierungsverfahren auch Weiterentwicklung von Delta von Hoover oder Eder and Rybicki, 2013 (siehe EVERT, ii5) und SMITH, improving authorship ...

Die stilometrischen Verfahren wie Burrows Delta ähneln Machine Learning Verfahren (TODO: mehr). Der größte Unterschied zwischen den stilometrischen und den Machine Learning Verfahren

TODO

- EDER, 2015 (does size...)
  - S. 169 r-u: SVM + ngrams angeblich bester AA-Ansatz (ICH: hier Literatur der beiden genannten angucken) (notiz auch im nächsten kapitel)
  - S. 170 l-o: ngrams nicht für alle Sprachen sehr effektiv (siehe EDER 2011)
- TODO HIER: Jockers2010 (einleitung) Grundstein: Burrows Delta (brachte Forschung nach vorne)
- Argamon festigte Burrows Delta Methode. Zeigte hier auch schon Vorschläge für Klassifizierung (KNN)
- jockers zeigte machine learning verfahren
- ICH: hier an ebert sehr stark orientieren
- ICH: wenn NN, dann auch hier erzählen

NOTIZ: typischer vorgang stylo

- document term matrix mit häufigkeiten
- matrix normalisieren
- dann irgendwie distanz berechnen

4 Aufbau der Experimente

TODO

- ICH: Wie von Argamon bereits in seinem Paper zu ... vorgeschlagen wird Burrows Delta für die Experimente als Nearest Neighbors Classifier aufgefasst. Dies hilft vor allem bei dem Vergleich zwischen anderen Classifiern, da ein gleicher Aufbau gewährleistet werden kann (AF?!).
- JOCKERs text angucken, der beschreibt einige wichtige Maßnahmen, die stylom-Klassifizierung von normaler Dokument-Klassifizierung unterscheidet.
- 80/20 split (Pareto principle) erklären
- stratify erklären

- HOOVER, s. 470 (und bei burrows delta irgendwo): sagt, dass Erhöhung der Features die Accuracy erhöhen: das beweisen!? (notiz auch bei nächstem kapitel)
  - hier aufbau erklären:
  - welche verfahren?
  - was waren die ideen?
  - was waren probleme?
  - "Problematik bei Klassifizierung: viele Klassen, wenig Beispiele. Dies macht die Klassifizierung schwierig, auch im Hinblick der Evaluation, da cv der Cross Validation nicht größer als 2 sein darf (s.o.)"
- ...

5 Experimente

TODO

- HOOVER, s. 470 (und bei burrows delta irgendwo): sagt, dass Erhöhung der Features die Accuracy erhöhen: das beweisen!?
- EDER, 2015 (does size...), S. 169 rechts unten: SVM + ngrams angeblich bester AA-Ansatz (ICH: hier Literatur der beiden genannten angucken)
- ICH: siehe hier Notizen von hackmd für vers. Experimente
- ICH: viele Grafiken

UMFANG (TODO weg):

- min 9 Seiten
- einleitende worte mit theorie

6 Analyse der Experimente

TODO

- hier BURROWS/HOOVERS (S. 470) These, dass mehr features gleich bessere accuracy bei delta, analysieren

7 Schlussbetrachtung

TODO

-

8 DELETE ME

So gestaltet man eine Tabelle:

Tabelle 1: Beispielstabelle

A	B	C
D	per gram	11.65
	each	1.01
E	stuffed	32.54
F	stuffed	73.23
G	frozen	8.39

Literatur

Argamon, Shlomo (2008). „Interpreting Burrows's Delta: Geometric and Probabilistic Foundations“. In: *Literary and Linguistic Computing*, 23.2, S. 131–147.

Argamon, Shlomo, Moshe Koppel und Jonathan Schler (2009). „Computational Methods in Authorship Attribution“. In: *Journal of the American Society for Information Science and Technology* 60.1, S. 9–26.

Barbaresi, Adrien (2018). *A corpus of German political speeches from the 21st century*.

Bücher-Wiki (2019). *Prosa*. URL: `https://www.buecher-wiki.de/index.php/BuecherWiki/Prosa` (besucht am 03. 09. 2019).

Burrows, John (2002). „'Delta': a Measure of Stylistic Difference and a Guide to Likely Authorship“. In: *Literary and Linguistic Computing* 17.3, S. 267–287.

Evert, Stefan u. a. (2017). „Understanding and explaining Delta measures for authorship attribution“. In: *Digital Scholarship in the Humanities* 32.2 (Supplement), S. ii4–ii16.

Fischer, Frank und Jannik Strötgen (6. Jan. 2017). *Corpus of German-Language Fiction*. URL: `https://figshare.com/articles/Corpus_of_German-Language_Fiction_txt_/4524680/1` (besucht am 03. 09. 2019).

Jockers, Matthew L. und Daniela M. Witten (2010). „A comparative study of machine learning methods for authorship attribution“. In: *Literary and Linguistic Computing* 25.2, S. 215–223.

Appendix

A Daten und Code

Im Ordner „app“ befindet sich die Python-Datei „utils.py“, welche Nachbearbeitungs- und Reduktionsfunktionen für die Korpora beinhaltet. Die Dateien „texts_to_csv.py“ und „classification.py“ sind CLI-Tools. Mithilfe von „texts_to_csv.py“ können csv-Dateien der Korpora erstellt werden. Mit dem Tool „classification.py“ wurden die verschiedenen Stilometrie-Experimente in dieser Arbeit durchgeführt. Mit „visualization.py“ werden die Abbildungen anhand der Klassifikations-Ergebnisse visualisiert. Zudem befinden sich im Ordner „app“ auch noch einige Jupyter-Notebooks, die zur Nachbearbeitung und Reduktion der Korpora genutzt und in der die Abbildungen, die sich in dieser Arbeit befinden, erzeugt wurden.

analysis.py noch erklären

in figures corpora analysis abbildungen für reduktion und so in figures results klassifizierungsergebnisse

TODO: erklären, wo tools, code –> github project

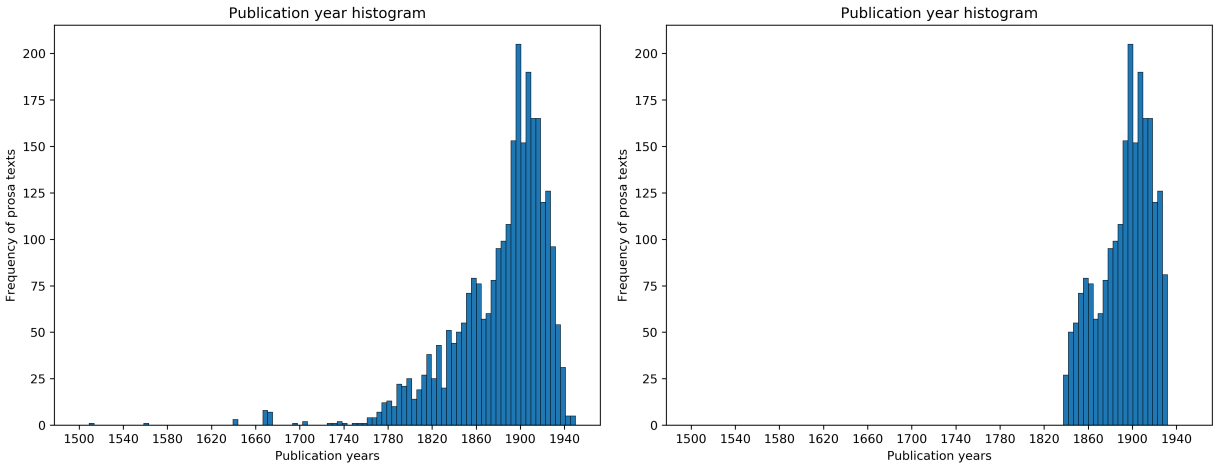


Abbildung 1: Das linke Histogramm zeigt die Verteilung der Häufigkeit der Erscheinungsjahre des unbearbeiteten Korpus. Wie von Strötgen und Fischer angemerkt, wurden die meisten Werke des Korpus im Zeitraum zwischen 1840 und 1930 erstellt. Das rechte Histogramm zeigt die Verteilung nach der Entfernung der Erscheinungsjahre außerhalb des Zeitraums 1840 bis 1930.

B Reduktion des Prosa Korpus

Das benutzte Korpus hat einige Probleme, weshalb es für die Untersuchungen in dieser Arbeit vorverarbeitet werden soll. Die Herangehensweise stützt sich dabei nicht auf eine literaturwissenschaftliche Wissensdomäne, sie soll eher auf schlichten Annahmen basieren. Die erste Veränderung soll eine Reduzierung des Korpus auf Werke sein, die zwischen 1840 und 1930 erschienen sind. Diese Aufteilung folgt der Aussage der Ersteller(AW?) des Korpus, dass der größte Teil der Werke des Korpus in diesem Zeitraum erschienen sei (Fischer und Strötgen, 2017). Dies wird durch Abbildung 1 (TODO) bestätigt, die eine Verteilung der Werke nach Erscheinungsjahren darstellt. Der Grund für die Reduzierung ist die Komplexität bei der Klassifizierung, welche durch ein kleineres Korpus verringert wird. Das reduzierte Korpus enthält nach der Reduzierung nur noch 2212 Werke, etwa zwanzig Prozent der Werke wurde entfernt. Die Anzahl der verschiedenen Autoren hat sich von 549 auf 439 verringert.

Die Ersteller(?) des Korpus merken einige bekannte Probleme mit dem Korpus an (Fischer und Strötgen, 2017). Fünf von neun dieser Probleme sind für eine Reduzierung uninteressant, da sie zum Beispiel außerhalb des Erscheinungszeitraums von 1840 und 1930 erschienen sind. Die anderen vier Probleme, bei denen es um Duplikate ging, wurden behoben, indem die problematischen Werke entfernt wurden. Nach der Entfernung bestand das Korpus noch aus 2208 Texten.

Die verschiedenen Prosagattungen der einzelnen Texte sind nicht bekannt. Dies ist problematisch, da die Menge an verschiedenen Textgattungen bei der Klassifikation

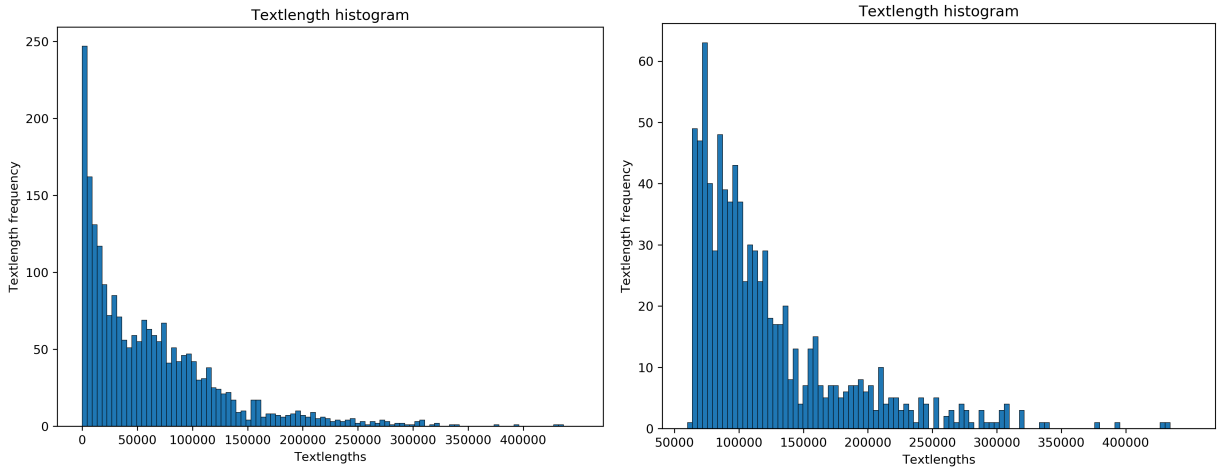


Abbildung 2: Die linke Grafik stellt die Häufigkeit von Texten anhand ihrer Textlänge dar. Sehr kurze Texte sind sehr häufig im Korpus und sehr lange Texte sind eher selten vertreten, die Verteilung ist also rechtsschief. Der Mittelwert der Gesamtheit aller Textlängen der Dokumente ist definiert als $\overline{x} = \frac{\sum L_D}{count(D_i)}$, wobei L_D die Textlängen der einzelnen Texte und $count(D_i)$ die Gesamtzahl aller Texte repräsentiert. Der Mittelwert $\overline{x}$ wird als Trennwert verwendet und alle Texte, deren Länge kürzer als der Mittelwert sind, werden aus dem Korpus entfernt. Die Verteilung der Texte anhand ihrer Textlängen nach dieser Reduzierung wird in der rechten Grafik dargestellt.

einen Bias[2] erzeugen kann. Durch eine Reduzierung der Prosagattungen kann auch der Bias verringert werden. Da die Prosagattungen der einzelnen Texte jedoch unbekannt sind, werden hier die Textlängen als loser Indikator für die verschiedenen Prosagattungen verwendet. Die Annahme ist, dass kürzere Texte eine erhöhte Wahrscheinlichkeit haben, Prosagattungen wie „Kurzgeschichte “, „Essay “ oder „Brief “ anzugehören. Da es jedoch keine eindeutigen Grenzen für die Einteilung von Prosagattungen gibt, wurde hier der Mittelwert ($\overline{x} = 63870$) als Trennwert verwendet.[3] Dieser ist aufgrund der großen Menge an kurzen Texten verzerrt. Nachdem die kürzeren Texte mithilfe des Mittelwerts entfernt wurden, sollen auch sehr lange Texte aus dem Korpus entfernt werden. Dies hat vor allem Performance-Gründe. Zuletzt wurden Autoren aus dem Korpus entfernt, von denen im Korpus weniger als drei Texte vorhanden waren.

Neben der Reduktion des Korpus wurde noch einige Nachbearbeitungsmaßnahmen getroffen. Innerhalb der ursprünglichen Texte wurden zu Beginn Titel, Autor und Erscheinungsjahr des Textes genannt. Diese Informationen wurden aus den Texten

[2] In dieser Arbeit wird der englische Begriff „Bias “ anstelle des deutschen Begriffs „systematischer Fehler “ benutzt.

[3] Siehe Abbildung 2. (TODO)

entfernt. [4] Die Befürchtung war es, dass bei einer Beibehaltung der Meta-Informationen im Text diese ein Einfluss auf die Ergebnisse der Klassifizierungen und stilometrischen Analysen haben und diese verfälschen könnten. Zuletzt wurden die Texte noch tokenisiert.

[4] Dies geschah in zwei Schritten: Zuerst wurde beim Einlesen der Texte ein Großteil der Meta-Informationen mithilfe eines regulären Ausdrucks entfernt. Dadurch konnten nicht alle Meta-Informationen beseitigt werden, da sich die Schreibweisen der Meta-Informationen von Titeln oder Autoren innerhalb des Textes teilweise geringfügig von den vorhandenen Meta-Informationen außerhalb des Textes unterschieden, wurde die Kosinus-Ähnlichkeit benutzt. Mit dieser wurde die Ähnlichkeit des Textanfangs jeweils mit Titel und Autor verglichen. Der Textanfang wurde dabei nach und nach vergrößert und iterativ mit dem Titel und Autor verglichen. Betrug die Ähnlichkeit gleich oder mehr als 60 Prozent, wurde der aktuelle Textanfang vom eigentlichen Text abgeschnitten und entfernt. Somit wurden auch einige andere Informationen als die Titel- und Autor-Informationen entfernt, die Menge ist jedoch im Vergleich der Textlänge so gering, dass dies keinen großen Einfluss auf die spätere stilometrische Analysen und Klassifizierungen haben sollte.

Eidesstattliche Erklärung

Hiermit versichere ich, die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt sowie die Zitate deutlich kenntlich gemacht zu haben.
Ich erkläre weiterhin, dass die vorliegende Arbeit in gleicher oder ähnlicher Form noch nicht im Rahmen eines anderen Prüfungsverfahrens eingereicht wurde.

Würzburg, den 6. Januar 2020 Autor