Der Einsatz quantitativer Textanalyse in den Geisteswissenschaften: Bericht über den Stand der Forschung (R 5.2.3)

Version 06.10.2015
Cluster 5
Verantwortlicher Partner Uni Würzburg

DARIAH-DE
Aufbau von Forschungsinfrastrukturen für die e-Humanities

Dieses Forschungs- und Entwicklungsprojekt wird / wurde mit Mitteln des Bundesministeriums für Bildung und Forschung (BMBF), Förderkennzeichen 01UG1110A bis N, gefördert und vom Projektträger im Deutschen Zentrum für Luft- und Raumfahrt (PT-DLR) betreut.

GEFÖRDERT VOM

Bundesministerium für Bildung und Forschung

Projekt: DARIAH-DE: Aufbau von Forschungsinfrastrukturen für die e-Humanities
BMBF Förderkennzeichen: 01UG1110A bis N
Laufzeit: März 2011 bis Februar 2016

Dokumentstatus: Final
Verfügbarkeit: öffentlich
Autoren:

Frederik Baumgardt, Infai Leipzig
Sina Bock, UWÜ
Keli Du, UWÜ
Michael Huber, UWÜ
Matt Munson, Infai Leipzig
Stefan Pernes, UWÜ
Steffen Pielström, UWÜ
Michael Sünkel, MinfBA

Revisionsverlauf:

Datum	Autor	Kommentare
21.08.2015	Baumgardt, Bock, Du, Huber, Munson, Pernes, Pielström, Sünkel	Entwurf
06.10.2015	Pernes, Pielström	Einarbeitung von Kommentaren aus dem Konsortium

Inhaltsverzeichnis:

1. Einleitung

Eine Kehrseite der immer weiter zunehmenden Verfügbarkeit literarischer Texte in digitaler Form für die Forschung besteht darin, dass das verfügbare, zur Bearbeitung einer bestimmten Forschungsfrage idealerweise zu berücksichtigende Material sich mittlerweile oftmals durch seine schiere Menge der Möglichkeit einer intensiven Lektüre entzieht (Crane 2006). Dadurch gewinnt neben dem im 20. Jahrhundert als literarisches Analyseverfahren etablierten sog. *Close Reading*, das eine detaillierte und vollständige Analyse der untersuchten Texte vorraussetzt (Wenzel 2004), gewinnt damit das auf computergestützen Verfahren basierende *Distant Reading* immer mehr an Bedeutung (Moretti 2000). So lassen sich an einem digitalen Literaturkorpus sowohl sprachlich-stilistische als auch thematisch-inhaltliche Aspekte mit quantitativen Verfahren untersuchen.
Der Fokus bisheriger Forschung im sprachlich-stilistischen Bereich lag vor allem auf der Stilometrie, insbesondere mit dem Ziel der Autorenschaftszuschreibung. Ein mittlerweile weit verbreitetes Verfahren zur inhaltlichen Analyse auf Basis quantitativer Auswertungen ist das *Topic Modeling*, das die Untersuchung der Verteilung inhaltlicher Schwerpunkte in einem Korpus erlaubt. Von besonderer Bedeutung im Kontext großer Textbestände ist auch der Einsatz von *Text-Reuse* und Plagiatserkennungsverfahren, die es ermöglichen, die intertextuellen Eigenschaften analysierter Texte aufzudecken.

2. Stilanalyse

Steffen Pielström, Sina Bock, Michael Huber
Julius-Maximilians-Universität Würzburg

Eine Eigenschaft der menschlichen Wahrnehmung ist es die charakteristischen Merkmale einer Person oder einer Sache zu erfassen und einzuordnen - dabei ist Stil allgegenwärtig (Argamon et al. 2010). So wird ein literarisches Werk nicht nur durch den individuellen Stil seines Verfassers geprägt, sondern lässt sich auch mithilfe markanter Merkmale Gattungen und Epochen zuordnen.
Methoden computergestützter Stilometrie erlauben es stilistische Unterschiede zu quantifizieren und zu visualisieren. Somit lässt sich der Stil verschiedener Autoren vergleichen, anonyme oder undatierte Texte können einem Autor oder einer Epoche zugeordnet oder spezifische Eigenschaften innerhalb einer Gattung herausgestellt werden.

Hierfür stehen heutzutage verschiedene, frei verfügbare *Software Tools*, wie z.B. *Voyant*[1], *Stylo*[2] und *PyDelta*[3] zur Verfügung. Klassische Methoden in diesem Bereich sind die *Principal Component Analysis* und die Quantifizierung stilistischer Unterschiede durch Textabstandsmaße, neuere, teilweise darauf aufbauende stilometrische Verfahren bedienen sich Techniken aus dem Bereich der *Clustering*-Verfahren und des überwachten Maschienellen Lernens.

2.1 Principal Component Analysis

Will man nämlich die Unterschiedlichkeit zweier Texte modellieren, so kann man jeden Text als Datenpunkt in einem mehrdimensionalen Koordinatensystem betrachten. Die Achsen dieses Koordinatensystems repräsentieren messbare Eigenschaften der Texte, sog. ***features***. In der Autorenschaftsattribution sind das meistens die relativen Häufigkeiten der am häufigsten verwendeten Wörter, also überwiegend von Funktionswörtern wie “und”, “der” und “die”. Je nach Fragestellung kann auch die Verwendung anderer *features* sinnvoll sein, z.B. die Häufigkeiten von Wortgruppen, Grammatischen Konstruktionen oder selteneren Inhaltswörtern. In jedem Fall geht die Zahl der berücksichtigten *features*, oft sind es die Frequenzen von mindestens 50-100 der häufigsten Wörter, in der Regel über die drei visuell darstellbaren Dimensionen hinaus. Es gibt aber eine Reihe von Analysetechniken für derartige, hochdimensionale Datensätze.

Eines der ersten Verfahren, die in der quantitatitven Textanalyse eingesetzt wurden, ist die *Principal Component Analysis* (PCA), die lange vor der quantitativen Textanalyse von Karl Pearson (Pearson 1901) und von Harold Hotelling (1930) entwickelt wurde. Ziel der PCA ist, in einem hochdimensionalen Datensatz eine Betrachtungsebene zu finden, in der sich möglichst viel von der Varianz der Daten visuell erfassen lässt.

[1] http://voyant-tools.org/
[2] https://sites.google.com/site/computationalstylistics/stylo
[3] https://github.com/fotis007/pydelta

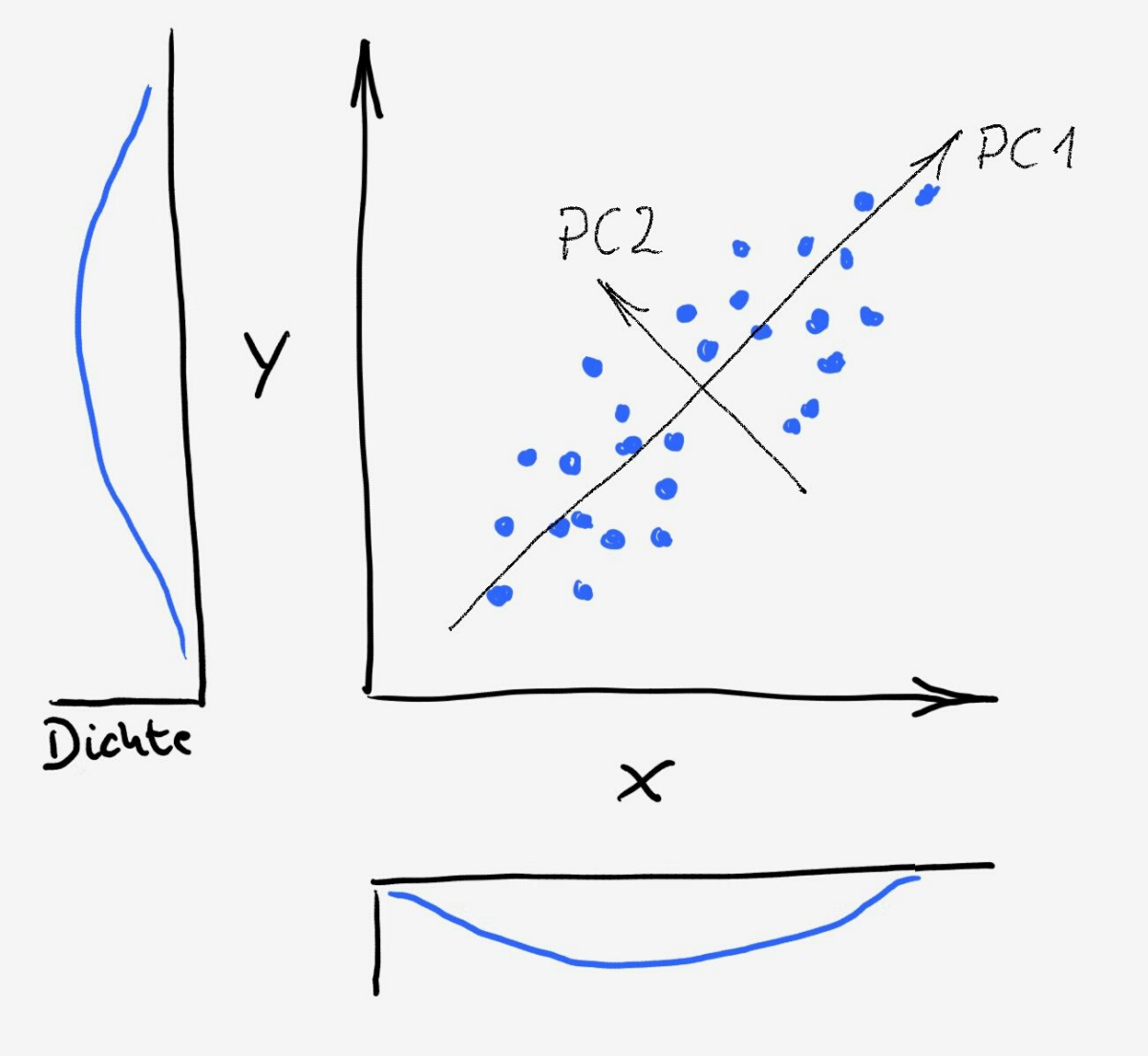

Abbildung 1: Vereinfachte Darstellung einer PCA auf nur zwei Dimensionen. Bei gleichzeitiger Betrachtung aller (zwei) Dimensionen sind hier deutlich zwei unterscheidbare Gruppen zu erkennen. Reduziert auf eine einzige Dimension, X oder Y, zeigt sich in den Daten aber keine bimodale Verteilung; die Gruppen lassen sich nicht mehr unterscheiden. Ebenso kann es in einem Datensatz mit 100 oder mehr Dimensionen schwierig werden, jene Dimensionen (oder Kombinationen von Dimensionen) auszumachen, in denen Unterschiede deutlich werden. Die Achsen der beiden Principal Components, die sich für diesen Datensatz berechnen lassen, sind hingegen an die Varianzverteilung der Datenpunkte angepasst. Aus DARIAH-DE Report 5.2.3: Stand der Forschung in der Textanalyse.

Hierfür werden die Dimensionen der Daten mit Hilfe der sog. **Singulärwertzerlegung** in ein neues Set von Variablen, die ***Principal Components***, transformiert. Diese *Principal Components* kann man als Achsen eines alternativen Koordinatensystems verstehen, in dem die selben Datenpunkte in der selben Anordnung aufgetragen sind. Die erste Achse

dieses neuen Bezugsystems (PC1) führt exakt durch die Datenpunkte in Richtung ihrer größten Ausdehnung, sie beschreibt also die größte Varianz der Daten (Abb. 1), die weiteren Achsen (PC2 bis PCn) repräsentieren die anderen neuen, orthogonal zur PC1 verlaufenden Achsen in Reihenfolge der Varianz, die der Datensatz in diesen Dimensionen jeweils hat. Folglich kann diese Technik eingesetzt werden, um aus einem Datensatz mit beliebig vielen Dimensionen eine zweidimensionale Darstellung (mit PC1 und PC2 als X- bzw. Y-Achse) zu erzeugen, die exakt diejenige Betrachtungsbene zeigt, in der der größte Teil der Datenvarianz zu sehen ist und Unterschiede zwischen Gruppen vermutlich am besten herausgestellt werden (Abb. 2, Smith 2002).

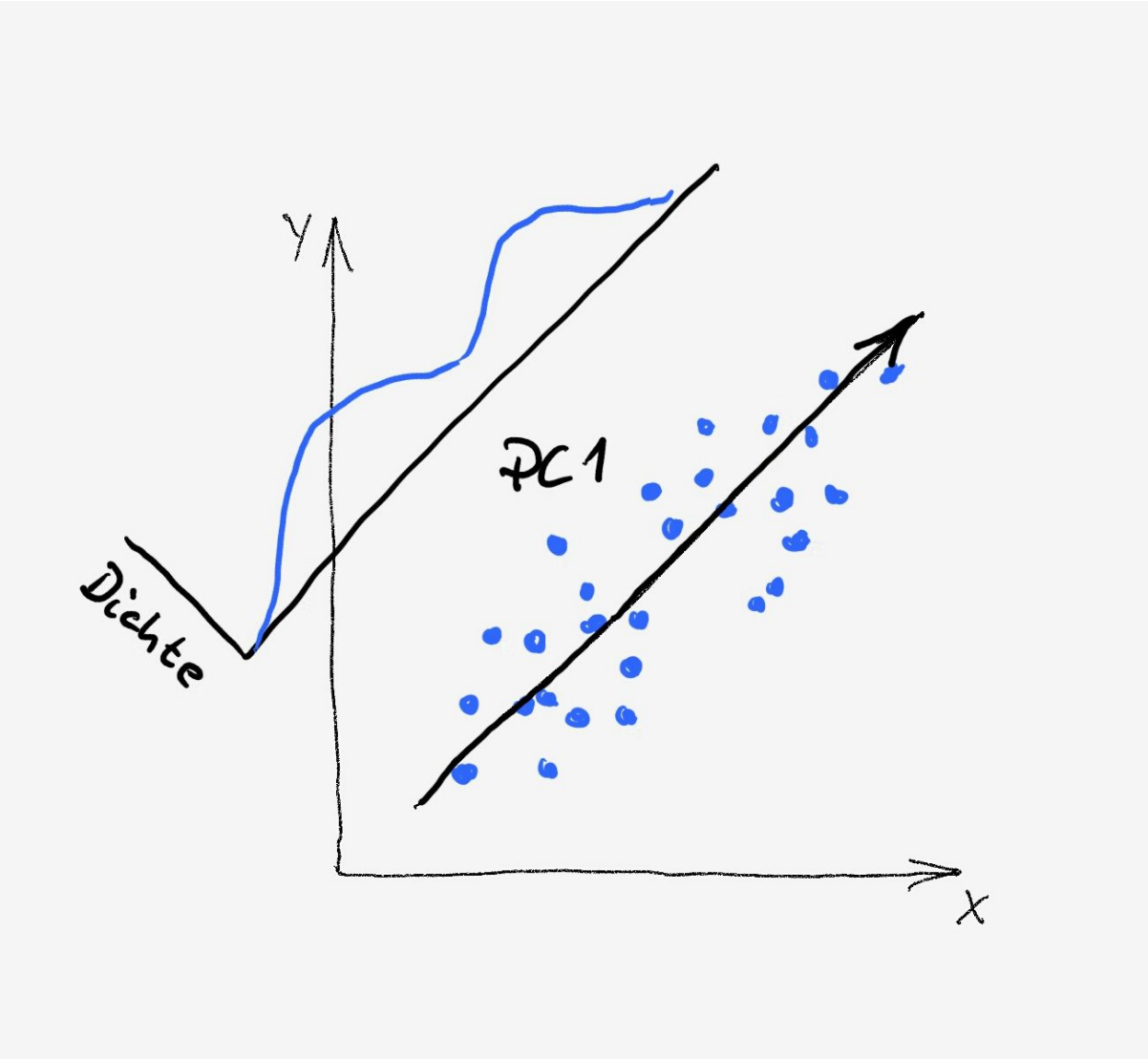

Abbildung 2: Entlang der neu berechneten Achse PC1 verläuft die Dichtekurve bimodal. Nun wird der Unterschied zwischen den beiden Gruppen schon in einer einzigen Dimension sichtbar. Aus DARIAH-DE Report 5.2.3: Stand der Forschung in der Textanalyse.

Dieses rechnerisch aufwändige Verfahren fand mit Aufkommen des Computers zunehmend mehr Berücksichtigung in unterschiedlichen Bereichen wie beispielsweise der Biologie, der Meterologie oder bei Bildkompressionsverfahren. Im Bereich der Textanalyse setzten Mosteller und Wallace (1964) die Methode zur Untersuchung der *Federalist* Papers erstmals im Zusammenhang mit Autorschaftsattributionen ein.

Vor allem, wenn es um die Zuordnung eines einzelnen Textes unbekannter Herkunft zu einem von zwei Autoren geht, für die jeweils mehrere sicher zugeordnete Vergleichstexte vorliegen, ist die PCA oftmals gut geeignet, die stilistische Ähnlichkeit zu einer der beiden Textgruppen visuell herauszustellen (Burrows 1989, Binongo und Smith 1999, Binongo 2003). Aber auch zur Analyse der zeitlichen Entwicklung von Schreibstilen (Brainerd 1980), oder der stilistischen Unterschiede zwischen Dialogen und narrativen Textpassagen (Burrows 1987), kann die PCA eingesetzt werden.

2.2 Die Quantifizierung stilistischer Unterschiede

Die Analyse stilistischer Unterschiede lässt sich noch weiter operationalisieren, indem man diese auch tatsächlich quantifiziert. Die aus der PCA bekannte Form der Modellierung von Texten als Datenpunkte in einem hochdimensionalen Koordinatensystem bietet hierbei die Möglichkeit, Abstände zwischen diesen Punkten direkt zu berechnen und als Maß für die stilistische Verschiedenheit zweier Texte zu verwenden. Als **Textabstandsmaße** bieten sich grundsätzlich die sog. ***Manhattan***-Distanz, d.h. die Summe der Abstände in den einzelnen Dimensionen[4], und die **Euklidische** Distanz[5] an. Das erste Verfahren dieser Art, daß in der Textanalyse erfolgreich war und bis heute in vielen Bereichen sehr erfolgreich eingesetzt wird, wurde von John Burrows (2002) vorgestellt.

Nachdem Burrows in seiner ersten Studie über die Figurenreden in Jane Austens Romanen mit statistischen Analysen erfolgreich hatte nachweisen können, dass die Sprechakte der Figuren sich systematisch unterscheiden (Burrows 1987) untersuchte er diese Methode im Zusammenhang mit Fragen bei Autorenschaftsattributionen und Epochenzugehörigkeit. In so genannten “geschlossenen” Spielen, also bei einer kleinen Anzahl potentieller Kandidaten, die als Autoren für einen anonymen Text in Frage kommen, können Autorenschaftsattributionen mit Hilfe der PCA sehr zuverlässig vorgenommen werden. Bei

[4] “Manhattan”-Distanz in Anspielung auf die Strecke, die man in einem rechtwinkligen Straßennetz zurücklegen muss, um von einem Punkt zu einem anderen zu gelangen.

[5] Hierbei handelt es sich um die kürzeste Verbindunglinie zwischen zwei Punkten in einem mehrdimensionalen Raum.

einem “offenen Spiel, in dem die möglichen Autoren kaum eingegrenzt werden können, reichen diese Methoden jedoch nicht mehr aus. In diesem Rahmen muss eine PCA für jeden Kandidaten, der dem Spiel hinzugefügt wird, erneut durchgeführt werden. Burrows Methode zielt darauf ab mit Hilfe statistischer Merkmale/Variablen aus einem offenen Spiel ein geschlossenes Spiel zu machen. 2002 stellte er dann ein Verfahren vor, das er Delta nannte. In seiner Studie verwendet er ein Referenzkorpus mit Texten von 25 Poeten des 17. Jahrhunderts. Zunächst wird dabei eine Liste der 150 häufigsten Wörter des Korpus (hierbei handelt es sich meist um Funktionswörter wie Artikel, Hilfsverben, Personalpronomen, Präpositionen, Konjunktionen). Da die Frequenzen in Worthäufigkeitslisten sehr schnell abfallen, alle Wörter aber gleich gewichtet werden sollten, damit die Analysen nicht von den häufigsten Wörtern dominiert werden, **standardisierte** er die Wortfrequenzen[6] und summierte die Beträge der Wertdifferenzen aller Dimensionen. **Burrows Delta** ist somit also ein Maß, das ausdrückt wie sehr sich die standardisierten, relativen Wortfrequenzen zweier Texte voneinander unterscheiden. Bei einem Text unbekannter Herkunft erlaubt Burrows Delta zu berechnen, welchen anderen Texten dieser am ähnlichsten ist. Burrows konnte mit diesem Verfahren bei einer Textlänge von über 2000 Wörtern 19 von 20 Gedichten dem richtigen Autor zuordnen. Bei einer geringeren Textlänge befand sich der entsprechende Autor in 85% aller Fälle noch unter den ersten fünf Kandidaten, so dass mit Delta aus einem offenen Spiel ein geschlossenen Spiel gemacht werden konnte.

Im Laufe der Zeit wurde Burrows Delta getestet und weiterentwickelt. So zeigte Hoover (2004a) dass Burrows Delta für Autorschaftsattributionen besonders gute Ergebnisse erzielt, wenn die 150 (oder mehr) häufigsten Wörtern berücksichtigt werden. Seine Untersuchungen zeigen außerdem, dass sich die Ergebnisse weiter optimieren lassen, wenn Personalpronomina und Wörter mit einer Frequenz von über 70% nicht in die Analyse miteinbezogen werden. Eine Auflösung von Kontraktionen in den von ihm untersuchten englischsprachigen Texten führt jedoch meist zu schlechteren Ergebnissen. In Hoovers Referenzkorpus konnten unter Berücksichtigung der 100 bis 300 häufigsten Wörter 18 von 20 Autoren treffend zugeordnet werden. Hoover (2004b) entwickelte und testete auch mehrere alternative Varianten von Delta, ohne aber eine wesentliche Verbesserung erreichen zu können.

Argamon (2008) wies in einer umfassenden mathematischen Analyse des Delta-Verfahrens auf mehrere Annahmen hin, die diese Methode implizit vorraussetzt, zeigte eine Reihe von Unzulänglichkeiten auf und entwickelte daraus eine Reihe von Verbesserungsvorschlägen. Zunächst stellte er heraus, daß bei der Berechnung Burrows Delta eine Standardisierung

[6] Ein Vorgang, der auch als z-Transformation bezeichnet wird.

oder z-Transformation mit der Manhattan-Distanz kombiniert wird. Während erstere normalerweise für normalverteilte Daten eingesetzt wird, eignet sich letztere eher für Daten, die einer Laplace-Verteilung folgen. Argamon schlägt daher vor, vorrausgesetzt die Wortfequenzen sind tatsächlich normalverteilt, statt der Manhattan-Distanz die Euklidische Distanz einzusetzen, ein Verfahren, das als **Argamons Delta** Eingang in einige gängige Stilometrie-Tools gefunden hat. Desweiteren weist Argamon darauf hin, daß die Wortfrequenzen eigentlich nicht alle voneinander unabhängig sind, d.h. viele Wörter korrelieren miteinander. Er schlägt vor, diese Korrelationen auf Basis einer Singulärwertzerlegung aus der *feature*-Matrix herauszurechnen, um so die statistische Unabhängigkeit der *features* herzustellen. Empirische Untersuchungen zeigen allerdings, daß Argamons Varianten der Delta-Methode in der Praxis keine Verbesserung der Erfolgsquote bei der Autorenschaftsattribution bringen (Jannidis et al. 2015).

Rybicki und Eder (2011) entwickelten eine Variante, die speziell an die Bedürfnisse stark flektierter Sprachen wie Polnisch und Latein angepasst ist. Im Vergleich zu einer weitgehend unflektierten Sprache, wie dem Englischen, ist bei Sprachen mit größerer morphologischer Formenvielfalt zu erwarten, daß die relative Häufigkeit der häufigen Wörter insgesamt weniger groß ist. Beim sog. **Eders Delta** werden die *features* nach ihrem Rang in der Liste der häufigsten Wörter gewichtet, um diesen Unterschied zu kompensieren.

Smith and Aldridge (2011) schlugen, in Anlehnung an etablierte Verfahren aus dem *Information Retrieval*, anstelle des bei Burrows verwendeten Manhattan-Distanzmaßes die Verwendung des Cosinus-Maßes vor. Die einzelnen, durch eine Reihe von numerischen *feature*-Werten repräsentierten Texte werden hierbei nicht als Datenpunkte in einem Koordinatensystem aufgefasst, sondern als Vektoren. Der Cosinus des Winkels zwischen diesen Vektoren dient hierbei als Maß für die Unterschiedlichkeit (Abb. 3). Empirische Tests konnten zeigen, dass **Cosinus Delta** für die Autorenschaftszuschreibung tatsächlich wesentlich bessere Ergebnisse ermöglicht als andere Varianten, und auch bei der Verwendung von mehr als 2000 Wörtern als *features* immer noch verlässlich ist, während die Performanz andere Maße in diesem Bereich beginnt wieder abzunehmen (Jannidis et al. 2015). Ein wesentlicher Grund dafür liegt vermutlich darin, dass in diesem Bereich der Wortliste zunehmend Worte auftreten, die nur in einzelnen Texten in hoher Frequenz vorkommen. Solche einzelnen Worte können die Abstände zwischen Texten, die vom selben Autor stammen, bei anderen Delta-Verfahren sehr groß werden lassen. Sie haben aber einen geringeren Effekt auf die Cosinus-Distanz, die Wirkung der Ausreißer wird hier in ähnlicher Weise gedämpft wie bei einer **Vektor-Normalisierung** (Jannidis et al. 2015).

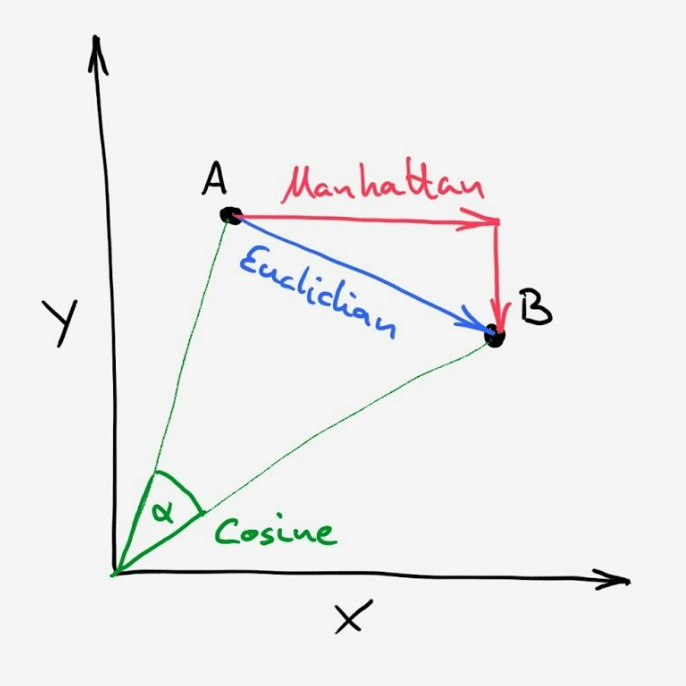

Abbildung 3: Der Abstand zweier Punkte A und B in einem Koordinatensystem: Manhattan-, Euklidische und Cosinus-Distanz. Aus Jannidis et al. 2015.

2.3 Clusteranalyse und überwachtes maschinelles Lernen

Bei der Untersuchung einer ganzen Gruppe von Texten kann die oben beschriebene Quantifizierung der Distanzen der einzelnen Texte zueinander die Grundlage für eine sog. Clusteranalyse bilden. Ziel der Clusteranalyse ist es, aus einer vollständigen quantifizierenden Beschreibung aller Zweierbeziehungen in einer Gruppe von Objekten Untergruppen, sog. ***Cluster***, zu identifizieren, in denen die Elemente eines Clusters sich einander möglichst ähnlich sind, während sich die Elemente aus verschiedenen Clustern möglichst unähnlich sind (Heyer et al. 2006). Solche Clusteringverfahren, die schon lange in den Naturwissenschaften eingesetzt werden, um Phänomene anhand messbarer Kriterien zu gruppieren[7], werden heutzutage auch zur Gruppierung von Texten eingesetzt und sind in den gängigen stilometrischen *Software Tools* integriert.

Ein weit verbreitetes Vorgehen hierbei ist, auf Basis der Delta-Abstände zwischen den Texten ein hierarchisches Clusteringverfahren durchzuführen und das Ergebnis in einem sog. Baumdiagramm oder Dendrogramm zu visualisieren. Solche Grafiken können dann interpretiert werden: Stilistisch ähnliche Autoren finden sich zum Beispiel auf benachbarten Ästen des Dendrogramms und ein Text unbekannter Herkunft sollte sich zwischen den anderen Texten des tatsächlichen Urhebers, oder zumindest in deren Nähe wiederfinden.

[7] z.B. in der biologischen Taxonomie oder bei der Klassifikation von Lebensräumen.

Für eine Klassifikationsaufgabe, wie die Zuschreibung eines einzelnen Textes unbekannter Urheberschaft zu einem von mehreren möglichen Autoren, kann aber auch ein Klassifikationsalgorithmus eingesetzt werden. Hierfür steht eine Reihe von Techniken aus dem Bereich des überwachten maschinellen Lernens (ML) zur Verfügung.

Maschinelles Lernen, unter anderem definiert als "...der Wissenserwerb eines künstlichen Systems."[8], kann als Begriff in diesem Zusammenhang recht irreführend sein. Einige der bereits weiter oben beschriebenen Analystechniken werden heute in maschinellen Lernverfahren eingesetzt und darum oft dem ML zugerechnet. Sowohl die PCA als auch die verschiedenen Formen der Clusteranalyse werden in gängigen Lehrbüchern als **unüberwachte** maschinelle Lernverfahren aufgeführt. "Unüberwacht" darum, weil sie helfen sollen, Muster und Strukturen zu erkennen, ohne dass irgendeine Form von Vorwissen bei der Analyse berücksichtigt wird: Die Autoren der Texte bekannter Urheberschaft spielen bei der Durchführung von PCA und Clusteranalyse keine Rolle, lediglich bei der anschließenden Interpretation.

Im Gegensatz dazu nutzen **überwachte** Verfahren vorhandene Metadaten. Sie werden vor allem für Klassifikationsaufgaben eingesetzt und benötigen stets ein sog. Test-Set, d.h. ein Set, anhand dessen sie Eigenschaften der einzelnen Klassen erlernen und neue Texte diesen Klassen zuordnen können.

Verfahren des ML werden in vielen alltäglichen Bereichen eingesetzt wie beispielsweise beim Filtern von Spam-Nachrichten. Hier wird ein Algorithmus regelbasiert, also mittels bestimmter Schlüsselwörter und/oder bereits vom Benutzer als Spam eingestufter E-Mail-Adressen so trainiert, dass er auch künftige unerwünschte E-Mail-Nachrichten mit hoher Wahrscheinlichkeit erkennen und aussortieren kann. Ein weiteres Anwendungsgebiet des Maschinellen Lernens ist der Bereich der Spracherkennung.

Auch für die Klassifikation literarischer Texte stellt z.B. die Software *Stylo*[9] eine Reihe etablierter Lernalgorithmen zur Verfügung, darunter die *Naive-Bayes*-Klassifikation, der *k-Nearest-Neighbour*-Algorithmus und *Support Vector Machines*. Als zuverlässigster Algorithmus für die Autorenschaftsattribution hat sich in empirischen Untersuchungen das ebenfalls in *Stylo* implementierte *Nearest-Shrunken-Centroids*-Verfahren erwiesen (Eder 2015).

[8] http://www.phonetik.uni-muenchen.de/~reichelu/kurse/machine_learning/machine_learning_1.pdf
[9] https://sites.google.com/site/computationalstylistics/stylo

Literatur:

Argamon, S. (2007) Interpreting Burrows's Delta: Geometric and Probabilistic Foundations. *Literary and Linguistic Computing*. [Online] 23 (2), 131–147.

Argamon, S. et al. (eds.) (2010) *The structure of style: algorithmic approaches to understanding manner and meaning*. Heidelberg ; New York: Springer.

Binongo, J. N. G. (2003) Who Wrote the 15th Book of Oz? An Application of Multivariate Analysis to Authorship Attribution. *CHANCE*. [Online] 16 (2), 9–17.

Binongo, J. & Smith, M. (1999) The application of principal component analysis to stylometry. *Literary and Linguistic Computing*. [Online] 14 (4), 445–466.

Brainerd, B. (1980) The Chronology of Shakespeare's Plays: A Statistical Study. *Computers and the Humanities*. 14 (4), pp. 221–230.

Burrows, J. (2002) 'Delta': a Measure of Stylistic Difference and a Guide to Likely Authorship. *Literary and Linguistic Computing*. [Online] 17 (3), 267–287.

Burrows, J. F. (1989) 'An ocean where each kind. . .': Statistical analysis and some major determinants of literary style. *Computers and the Humanities*. [Online] 23 (4-5), 309–321.

Burrows, J. F. (1987) *Computation into criticism : a study of Jane Austen's novels and an experiment in method*. Oxford: Clarendon Press.

Burrows, J. F. (1987) Word-Patterns and Story-Shapes: The Statistical Analysis of Narrative Style. *Literary and Linguistic Computing*. [Online] 2 (2), 61–70.

Crane, G. (2006) What Do You Do with a Million Books? *D-Lib Magazine*. [Online] 12 (3), . [online]. Available from: http://www.dlib.org/dlib/march06/crane/03crane.html (Accessed 20 August 2015).

Heyer, G. et al. (2008) *Text Mining: Wissensrohstoff Text: Konzepte, Algorithmen, Ergebnisse*. IT lernen. 1., korr. Nachdr. Herdecke ; Bochum: W3L-Verl.

Hoover, D. L. (2004a) Delta Prime? *Literary and Linguistic Computing*. [Online] 19 (4), 477–495.

Hoover, D. L. (2004b) Testing Burrows's Delta. *Literary and Linguistic Computing*. [Online] 19 (4), 453–475.

Moretti, F. (2000) Conjectures on world literature. *New Left Review*. 1. [online]. Available from: http://newleftreview.org/II/1/franco-moretti-conjectures-on-world-literature.

Pearson, K. (1901) LIII. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine Series 6*. [Online] 2 (11), 559–572.

Jannidis, Fotis et al. (2015) *Towards a better understanding of Burrows's Delta in literary authorship attribution*. [Online] [online]. Available from: http://dx.doi.org/10.5281/zenodo.18177 (Accessed 20 August 2015).

Rybicki, J. & Eder, M. (2011) Deeper Delta across genres and languages: do we really need the most frequent words? *Literary and Linguistic Computing*. [Online] 26 (3), 315–321.

Schöch, C. (2014) 'Corneille, Molière et les autres. Stilometrische Analysen zu Autorschaft und Gattungszugehörigkeit im französischen Theater der Klassik', in Christof Schöch; Lars Schneider (ed.) *Literaturwissenschaft im digitalen Medienwandel*. Beihefte zu Philologie im Netz, 7. PhiN. pp. 130–157. [online]. Available from: https://hal.archives-ouvertes.fr/hal-00957091.

Smith, L. I. (2002) *A tutorial on principal components analysis*. [online]. Available from: http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf. [online]. Available from: http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf.

Smith, P. W. H. & Aldridge, W. (2011) Improving Authorship Attribution: Optimizing Burrows' Delta Method*. *Journal of Quantitative Linguistics*. [Online] 18 (1), 63–88.

Wenzel, P. (2004) 'New Criticism', in *Grundbegriffe der Literaturtheorie*. Stuttgart und Weimar: Metzler Verlag.

Weiterführende Literatur

Eder, M. et al. (n.d.) 'Stylometry with R: a suite of tools', in *Digital Humanities 2013 Conference Paper*.

Eder, M. & Rybicki, J. (2013) Do birds of a feather really flock together, or how to choose training samples for authorship attribution. *Literary and Linguistic Computing*. [Online] 28 (2), 229–236.

Jannidis, F. & Lauer, G. (n.d.) Burrows's Delta and Its Use in German Literary History Matt Erlin & Lynne Tatlock (eds.). *Journal of Literary Theory*.

Murphy, K. P. (2012) *Machine learning: a probabilistic perspective*. Adaptive computation and machine learning series. Cambridge, MA: MIT Press.

Nünning, V. & Nünning, A. (2010) *Methoden der literatur- und kulturwissenschaftlichen Textanalyse. Ansätze, Grundlagen, Modellanalysen*. Stuttgart [u.a.]: Metzler.

Popescu, M. & Dinu, L. P. (2009) 'Comparing Statistical Similarity Measures for Stylistic Multivariate Analysis.', in Galia Angelova et al. (eds.) *RANLP*. 2009 RANLP 2009 Organising Committee / ACL. pp. 349–354. [online]. Available from: http://dblp.uni-trier.de/db/conf/ranlp/ranlp2009.html#PopescuD09.

Rajaraman, A. & Ullman, J. D. (2012) *Mining of massive datasets*. New York, N.Y. ; Cambridge: Cambridge University Press.

Ramsay, S. (2011) *Reading machines: toward an algorithmic criticism*. Topics in the digital humanities. Urbana: University of Illinois Press.

Underwood, T. (2015) ‘Plot arcs’ in the novel. The Stone and the Shell [online]. Available from: http://tedunderwood.com/2015/01/03/plot-arcs-in-the-novel/ (Accessed 20 August 2015).

3. Inhaltsanalyse

Stefan Pernes, Keli Du, Michael Huber
Julius-Maximilians-Universität Würzburg

Die in zunehmendem Umfang verfügbaren digitalen Korpora verlangen nach inhaltlicher Aufbereitung, die jedoch manuell nur schwer zu bewerkstelligen ist. Desweiteren bieten die Datenmengen interessante Möglichkeiten, semantische Felder zu modellieren, Inhalte auffindbar zu machen, Hypothesen über große Textsammlungen zu prüfen und unbekannte Texte nach ihren Inhalten zu klassifizieren. Für die Verfahren, die im Folgenden vorgestellt werden, gilt, dass nicht auf zuvor manuell definierte Wissensbestände oder Annotationen in den Texten zurückgegriffen werden muss, sondern der Ablauf komplett *ungesteuert* und anhand von *unstrukturiertem Text* stattfinden kann.

Key Words in Context (KWIC), bekannt vor allem aus der Korpuslinguistik, gilt als ein klassisches Werkzeug, das es ermöglicht, größere Textmengen im Hinblick auf bestimmte Schlüsselworte zu erschließen. KWICs werden genutzt, um eine Abfolge von Zeichen, meist Wörter in ihrem syntaktischen Zusammenhang oder dem Kontext ihrer Äußerung darzustellen, wobei die gesuchte Zeichenfolge zentriert untereinander dargestellt wird. Solche Abfolgen von Zeichen beziehungsweise Worten werden auch als *N-Gramme* bezeichnet.

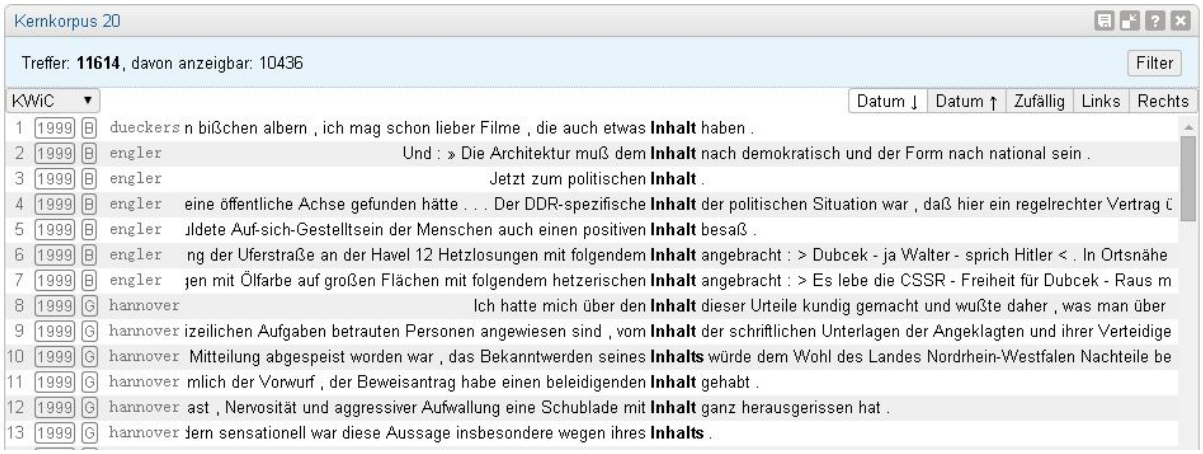

Abbildung 4: *Key Words in Context* im Digitalen Wörterbuch der deutschen Sprache (http://www.dwds.de)

Das Konzept der N-Gramme ist maßgeblich durch den Google N-Gram Viewer bekannt. Hierbei wird eine Suchanfrage in Form einer beliebigen Anzahl von *n* Zeichenfolgen an das Google-Books-Korpus gestellt. Das Ergebnis ist ein Graph, der die Antreffwahrscheinlichkeit besagter N-Gramme in einem chronologischen Kontext darstellt. Ein anderer Ansatz zur semantischen Erschließung von Texten ist das *Topic Modeling* (TM). Im Gegensatz zum reinen Auffinden linguistischer Muster und Trends in einem Korpus, können anhand von *Topic Modeling* auch umfassendere sprachliche Kontexte beschrieben werden.

3.1 Topic Modeling

Topic Modeling bezeichnet eine Gruppe von Verfahren, die es ermöglichen, anhand einer statistischen Analyse des lexikalischen Inventars Rückschlüsse auf die zugrunde liegende thematische Struktur einer Sammlung von Texten zu ziehen (Blei 2012). Ausgangspunkt ist meist eine *Term-Document* Matrix - jede Zeile in dieser Matrix steht für ein Wort bzw. eine Wortform und jede Spalte steht für eine zusammenhängende Textpassage, wie z.B. ein Absatz, Kapitel oder ein Dokument. Die Zellen enthalten die Worthäufigkeit für jedes Dokument. Eine solche Darstellung wird auch als **bag-of-words** Modell bezeichnet - eine Vereinfachung, die nicht Syntax und Wortfolge, sondern nur die Häufigkeit des Auftretens von Worten abbildet. Für die Bearbeitung einiger korpusanalytischer Fragestellungen ist diese Matrix bereits ausreichend. Um jedoch anhand von statistischen Methoden die Semantik, also den Bedeutungsgehalt eines Textes, besser erschließen zu können, wurde eine Reihe von Transformationen entwickelt, die es ermöglichen, Aussagen über die inhaltlichen Zusammenhänge von Worten und Dokumenten zu treffen. Da es sich dabei um einen rein mathematischen Ansatz handelt, funktionieren die Verfahren sprachunabhängig und können sogar eingesetzt werden, um thematische Strukturen in mehrsprachigen Korpora zu verfolgen. Ein wichtiger Einsatzbereich ist auch die Klassifikation und das Auffinden verwandter Dokumente (*Information Retrieval*). Desweiteren wurden Parallelen zum menschlichen Konzepterwerb festgestellt. Zurzeit ist *Latent Dirichlet Allocation* (LDA) ein weit verbreitetes Verfahren im Bereich des Topic Modeling. Als bekannte Vorgänger der LDA sollen im Folgenden auch zwei weitere Methoden skizziert werden: *Latent Semantic Analysis* (LSA) und *Probabilistic Latent Semantic Analysis* (pLSA). Synonym für beide wird auch die Bezeichnung *Latent Semantic Indexing* (LSI) verwendet. Beide Methoden zielen darauf ab, inhaltlich zusammenhängende Worte zu gruppieren und dabei auf die Bedeutung dieser Gruppierungen (beziehungsweise Dokumente, Passagen,...) zu schließen.

Ausgehend von einem ausreichend großen Korpus kann mit Hilfe der **Latent Semantic Analysis** (Landauer und Dumais 2008) ein *semantischer Raum* berechnet werden, der Aussagen darüber zulässt, wie ähnlich die im Korpus enthaltenen Worte und Dokumente sind. Grundlage ist eine *Term-Document* Matrix und - wie für einige andere Verfahren auch (siehe: PCA) - die Singulärwertzerlegung (*Singular Value Decomposition*, SVD) der Matrix. Die SVD ermöglicht es, eine wechselseitige Einschränkung zu berücksichtigen, nämlich die Berechnung der Wortenbedeutungen als durchschnittlicher Effekt auf die Bedeutung von Dokumenten und umgekehrt, die Bedeutung von Dokumenten als durchschnittlicher Effekt auf die Bedeutung von Worten. Das anschließende Ergebnis dieser Operation sind zwei Matrizen, die Informationen zu den im Modell enthaltenen Worten, respektive Dokumenten, enthalten sowie eine dritte Matrix, die den *Eigenwerten* der Ursprungsmatrix entspricht. Im Fall der LSA gilt weiters, dass eine reduzierende Form der SVD angewandt wird, die eine bestmögliche *k*-dimensionale Approximation an die Ursprungsmatrix darstellt. Im Anschluss ist es aufgrund der einheitlichen Vektorform der berechneten Wort- und Dokument-Matrizen möglich, die Ähnlichkeiten von Wort-Wort, Dokument-Dokument, und Wort-Dokument, anhand eines Abstandsmaßes für Vektoren, wie z.B. der *Kosinus-Ähnlichkeit*, zu berechnen. Landauer und Dumais weisen darauf hin, dass die relative Nähe der Worte zueinander, die aus dem Modell abgelesen werden kann, nicht einer einfachen *Kookkurrenz* der Worte entsprechen, sondern einer “Ähnlichkeit der Effekte, die jene Worte auf die Textpassagen haben, in denen sie vorkommen” (Landauer und Dumais 2008).

Obwohl mit LSA erstmals ein Verfahren zur Berechnung von sogenannten *latenten semantischen Räumen* entwickelt wurde, handelt es sich dabei noch nicht um ein Topic Model im engeren Sinn. Das Konzept des *Topics* wird in der Regel als *latente Variable* umgesetzt und repräsentiert dabei ein Teilvokabular des Korpus, aus dem sich Dokumente zusammensetzen können. Im Gegensatz zu Worten und Dokumenten handelt sich dabei um eine unbeobachtbare Größe, die erst durch die Berechnung des Modells entsteht. LSA kommt ohne solche latenten Variablen aus, was in diesem Fall bedeutet, dass sich jedes Dokument aus einem einzigen Vokabular - nämlich auf Grundlage des gesamten Korpus - zusammensetzt.

Im Rahmen einer Erweiterung des Verfahrens, nämlich der **Probabilistic Latent Semantic Analysis** (pLSA) (Hofmann 1999), wird es möglich, Dokumente als Zusammensetzung mehrerer Topics zu beschreiben. Dazu wird für jedes Dokument eine Wahrscheinlichkeitsverteilung über eine vorher festgelegte Anzahl von Topics (latente

Variable) angenommen. Anhand dieser Verteilung werden schließlich die Worte eines Dokuments “gezogen”. Es handelt sich dabei um eine realistischere Annäherung an die inhaltliche Zusammensetzung eines Korpus beziehungsweise eines Textes. Ein weiterer Unterschied ist, dass keine Heuristik (“die ersten *k* Singulärwerte”) zur Dekomposition der Matrix eingesetzt, sondern anhand einer *Expectation-Maximization* eine optimale Approximation an die Ursprungsdaten erreicht wird. Obwohl das Verfahren dadurch robuster gegen *Overfitting* wird (eine übermäßig genaue Modellierung der Trainingsdaten, die sich nicht ausreichend auf neue, bisher unbekannte Fälle generalisieren lässt), können die damit generierten Modelle nicht ohne Qualitätsverlust um neue Dokumente erweitert werden. Gerade bei großen Korpora kann das rechnerisch aufwändig und unpraktikabel sein. Es handelt sich dabei um eine Eigenschaft, die in erster Linie bei *Document Classification* und *Retrieval*-Aufgaben zum Nachteil gerät - ein Nachteil der jedoch einen wesentlichen Grund für die Entwicklung der **Latent Dirichlet Allocation** (LDA) darstellt (Blei 2012).

Vom Grundprinzip her unterscheiden sich pLSA und LDA als *probabilistische Topic Models* nicht wesentlich voneinander - in beiden Fällen wird von beobachtbaren Variablen - den Worten und Dokumenten - auf die nicht-beobachtbare, latente Variable der Topics geschlossen. Folgende Punkte fassen die Beziehung zwischen Topics, Dokumenten und Worten noch einmal zusammen:

- Ein **Dokument** setzt sich mehreren Topics in jeweils unterschiedlicher Gewichtung zusammen.
- Ein **Topic** ist eine Wahrscheinlichkeitsverteilung über das gesamte Vokabular des Korpus. Die Wahrscheinlichkeit, dass ein Wort zu dem Thema gehört, ist unterschiedlich. Zum Beispiel gehört das Wort „Sprache“ mit hoher Wahrscheinlichkeit zu dem Thema „Sprachwissenschaft“. Im Vergleich dazu gehört das Wort „Brezel“ mit geringer Wahrscheinlichkeit zu dem gleichen Thema. Diese „kleinere Wahrscheinlichkeit“ kann sehr nah an 0% liegen, aber sie ist keinesfalls gleich 0%.
- Ein **Wort** kann mit hohen Wahrscheinlichkeiten mehreren Topics angehören. Zum Beispiel ist das Wort „Spiel“ ein wichtiger Begriff nicht nur für das Thema „Sport“, sondern auch für das Thema „Theater“.

Probabilistische Topic Models können am einfachsten als generativer Prozess beschrieben werden. Es wird ein zufälliger Prozess angenommen, der zur Entstehung der Dokumente geführt hat: Zu Beginn wird eine zufällige Zusammensetzung von Topics für das gesamte

Korpus generiert. Anschließend wird für jedes Wort in jedem Dokument zufällig ein Topic aus der Topic-Verteilung ausgewählt. Auf Grundlage dieses Topics wird schließlich zufällig ein Wort aus dem Gesamtvokabular des Korpus ausgewählt. So werden für jedes Dokument die Worte in einem zweistufigen Prozess generiert.

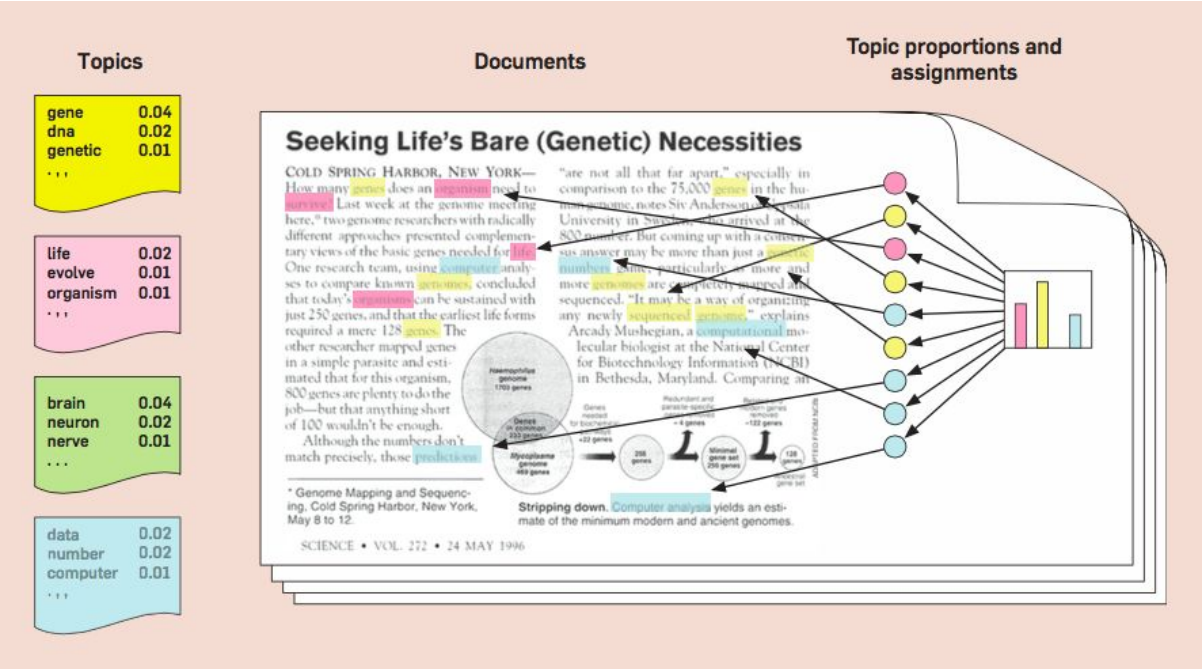

Abbildung 5: Intuition zu Topic Modeling: Aus der Gesamtmenge von Topics (links) wird für jedes Dokument eine Topic-Zusammensetzung gewählt (Histogramm-Skizze rechts). Anschließend wird für jedes Wort eine Topic-Zuordnung getroffen (farbige Kreise) und anhand dieses Topics ein Wort gewählt (Blei 2012).

In der praktischen Anwendung eines Topic Models auf ein bestehendes Korpus handelt es sich jedoch nicht um einen generativen Prozess, sondern um eine Umkehrung des eben beschriebenen Vorgangs. Während die Topics und ihre Verteilung für den generativen Prozess den Ausgangspunkt darstellen, sind sie in der Anwendung auf ein bestehendes Korpus die unbekannte, latente Variable. In diesem Fall sind die Dokumente bekannt und beobachtbar, die Topics, ihre Verteilungen für jedes Dokument sowie die Zuweisung zu den Worten sind jedoch unbekannt und müssen aus den beobachtbaren Informationen gewonnen werden. Die Frage lautet also: Wie sieht die versteckte Topic-Struktur aus, die am wahrscheinlichsten das zu untersuchende Korpus erzeugt hat? Dieser Vorgang beschreibt die zentrale technische Problemstellung probabilistischer Topic Models, denn hierbei ist es notwendig, alle denkbaren Topic-Zusammensetzungen zu berücksichtigen, um zu einem möglichst optimalen Ergebnis zu gelangen (auch: *Parameter Estimation*).

Hierin besteht die mathematische Komplexität solcher Topic Models und gleichzeitig auch der Vorteil von LDA gegenüber pLSA: Da die Anzahl aller möglichen Topic-Zusammensetzungen exponentiell groß und nicht direkt berechenbar ist (da sich die Topic Struktur auch auf jedes einzelne Wort auswirkt), muss der Vorgang approximiert werden. Solche aufwändigen Berechnungen möglichst effizient zu gestalten ist ein aktiver Forschungsbereich, wobei sich besonders Ansätze der Bayes'schen Statistik dafür zu eignen scheinen. Die für LDA am häufigsten eingesetzte Methode ist das *Gibbs Sampling*, ein Verfahren der statistischen Physik, das in der Lage ist, sich über ein ausgeklügeltes Ziehen von Stichproben an hochdimensionale, multivariate Verteilungen anzunähern und dabei auch auf unbeobachtbare Variablen, die in dem multivariaten Modell enthalten sind, zu schließen. (Blei 2012)

3.2 Aktueller Forschungsstand im Bereich Topic Modeling

Der Anwendungsbereich von TM liegt vor allem in der Analyse von großen Korpora, wobei es konventionellen quantitativen Methoden überlegen ist (Evans 2014). TM ermöglicht sowohl einen Überblick über im Korpus behandelte Themen, was eine Diskursanalyse großer Korpora zulässt, als auch zur Evaluation von Forschungsergebnissen (Heuser 2012). Ein weiterer Anwendungsbereich für TM ist das *Clustering* von Texten. Hierbei werden die produzierten Themen für eine Klassifizierung von Texten benutzt, wie zum Beispiel Genre oder Epoche (Goldstone & Underwood, 2012).

Über den Forschungsstand im Bereich *Topic Modeling* hat David Mimno eine ausführliche Literaturliste erstellt[10]. Die Liste enthält über 100 Literaturangaben, angeordnet in 15 Teilbereiche. Im Hinblick auf eine Anwendung in der Literaturwissenschaft und anderen Philologien sind inbesondere die folgenden zu nennen:

- „Bibliometrics": Topic Modeling wird z. B. dafür eingesetzt, einflussreiche Texte im Korpus zu identifizieren (Gerrish und Blei 2010).
- „Cross-language": Anwendung von Topic Modeling für die Erzeugung der multilingualen Topics (Jagarlamudi und Daumé III 2010).
- „Evaluation": Methoden zur Evaluation von Topic Models (Wallach, Murray, Salakhutdinov und Mimno 2009).

[10] http://mimno.infosci.cornell.edu/topics.html

- „NLP“: Eine erweiterte Form von LDA kann auch für Natural Language Processing Aufgaben angewendet werden, wie z.B. für Part-of-Speech Tagging (Toutanova und Johnson 2007).
- „Networks“: Topic Models kann auch für die Klassifizierung von Netzwerken angewendet werden. Zum Beispiel die Entdeckung von Gruppen von Entitäten und ihre Attributten in einem Entity-Relationship Modell (Wang, Mohanty und McCallum 2005).
- „Non-parametric“: Erweiterungen, die in der Lage sind, optimale Parameter selbst zu finden. Dabei entstehen hierarchische Topics, ähnlich wie bei einem hierarchischen Clustering (Blei et al. 2003).

Literatur

Blei, D. M. et al. (2003) ‘Hierarchical topic models and the nested Chinese restaurant process’, in *NIPS*. 2003 [online]. Available from: http://books.nips.cc/papers/files/nips16/NIPS2003_AA03.pdf.

Blei, D. M. (2012) Probabilistic topic models. *Communications of the ACM*. [Online] 55 (4), 77.

Evans, M. S. (2014) A Computational Approach to Qualitative Analysis in Large Textual Datasets Daniele Fanelli (ed.). *PLoS ONE*. [Online] 9 (2), e87908.

Gerrish, S. & Blei, D. M. (2010) ‘A language-based approach to measuring scholarly impact’, in *ICML*. 2010 [online]. Available from: http://www.cs.princeton.edu/ blei/papers/GerrishBlei2010.pdf.

Heuser, R. et al. (2012) *A Quantitative literary history of 2,958 nineteenth-century British novels : the Semantic cohort method* [online]. Available from: http://litlab.stanford.edu/LiteraryLabPamphlet4.pdf.

Hofmann, T. (1999) ‘Probabilistic Latent Semantic Analysis’, in *Proc. of Uncertainty in Artificial Intelligence, UAI’99*. 1999 Stockholm: . [online]. Available from: http://citeseer.ist.psu.edu/hofmann99probabilistic.html.

Jagarlamudi, J. & III, H. D. (2010) ‘*Extracting Multilingual Topics from Unaligned Comparable Corpora*’, in 2010 pp. 444–456. [online]. Available from: http://dx.doi.org/10.1007/978-3-642-12275-0_39.

Landauer, T. K. & Dumais, S. (2008) Latent semantic analysis. *Scholarpedia*. 3 (11), 4356.

Toutanova, K. & Johnson, M. (2007) ‘A Bayesian LDA-based model for semi-supervised part-of-speech tagging’, in *NIPS*. 2007 pp. 1521–1528. [online]. Available from: http://books.nips.cc/papers/files/nips20/NIPS2007_0964.pdf.

Underwood, T. & Goldstone, A. (2012) What can topic models of PMLA teach us about the history of literary scholarship? The Stone and the Shell [online]. Available from:

http://tedunderwood.com/2012/12/14/what-can-topic-models-of-pmla-teach-us-about-the-history-of-literary-scholarship/ (Accessed 21 August 2015).

Wallach, H. et al. (2009) 'Evaluation Methods for Topic Models', in *ICML*. 2009 [online]. Available from: http://www.cs.umass.edu/ mimno/papers/wallach09evaluation.pdf.

Wang, X. et al. (2005) 'Group and Topic Discovery from Relations and Their Attributes', in *NIPS*. 2005 [online]. Available from: http://books.nips.cc/papers/files/nips18/NIPS2005_0819.pdf.

Weiterführende Literatur

Blei, D. M. et al. (2003) Latent Dirichlet Allocation. *J. Mach. Learn. Res.*. 3993–1022.

Blei, D. M. (2012) Probabilistic topic models. *Communications of the ACM*. [Online] 55 (4), 77.

Burton, M. (213AD) The Joy of Topic Modeling. Words [online]. Available from: http://mcburton.net/blog/joy-of-tm/ (Accessed 21 August 2015).

Gerrish, S. M. & Blei, D. M. (n.d.) *A Language-based Approach to Measuring Scholarly Impact*.

Jockers, M. L. (2013) *Macroanalysis: digital methods and literary history*. Topics in the digital humanities. Urbana: University of Illinois Press.

Li, S. et al. (2009) 'A Framework of Feature Selection Methods for Text Categorization', in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*. ACL '09. 2009 Stroudsburg, PA, USA: Association for Computational Linguistics. pp. 692–700. [online]. Available from: http://dl.acm.org/citation.cfm?id=1690219.1690243.

McCallum, A. et al. (2007) 'Joint Group and Topic Discovery from Relations and Text', in *Proceedings of the 2006 Conference on Statistical Network Analysis*. ICML'06. 2007 Berlin, Heidelberg: Springer-Verlag. pp. 28–44. [online]. Available from: http://dl.acm.org/citation.cfm?id=1768341.1768345.

Mehrotra, R. et al. (2013) 'Improving LDA Topic Models for Microblogs via Tweet Pooling and Automatic Labeling', in *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '13. [Online]. 2013 New York, NY, USA: ACM. pp. 889–892. [online]. Available from: http://doi.acm.org/10.1145/2484028.2484166.

Riddell, A. B. (2012) A Simple Topic Model (Mixture of Unigrams). Allen B. Riddell [online]. Available from: https://ariddell.org/simple-topic-model.html (Accessed 21 August 2015).

Steyvers, M. & Griffiths, T. (2007) '*Latent Semantic Analysis: A Road to Meaning*', in T. Landauer et al. (eds.) Laurence Erlbaum.

Underwood, T. (2012) Where to start with text mining. The Stone and the Shell [online]. Available from: http://tedunderwood.com/2012/08/14/where-to-start-with-text-mining/ (Accessed 21 August 2015).

Historical Text Reuse Detection.
An examination of the State of the Art.

Frederik Baumgardt, Tufts University
Matt Munson, Universität Leipzig

Introduction

Based on the idea that Text Reuse is evidence for a reception of preceding ideas, it is therefore of interest to research that attempts to reconstruct and understand the genesis of a historic text or the life of its author(s). Text Reuse can take many different forms, from verbatim copies to recognizable paraphrases to vague allusions; it crosses language barriers, time spans and geographic distances. This can make it challenging for humans and machines alike to discover instances of Text Reuse.
The following document recapitulates the current state of the research into automated detection of Historical Text Reuse in large collections of text.

Theory

From a mathematical perspective the result of a Text Reuse Detection is a subgraph in the complete bipartite graph between the individual passages (reuse units) in a pair of text bodies. The sets of reuse units in both bodies V1 and V2 potentially interconnect with another through a complete set of edges E to form a graph G = (V1, V2, E).
The process of Text Reuse Detection operates as a filter on the set of edges to first eliminate the impossible relationships and then select the most likely ones. The result is a subgraph H = (W1, W2, F) with W1≤V1, W2≤V2 and F≤E where the vertices in W1 and W2 are often projections of V1, V2 respectively into a space of reduced dimensionality and the edges in F are weighted according to a scoring scheme.

Methodology

Most published approaches to Historical Text Reuse Detection follow an architecture of three processing stages:

1. Dimensionality reduction in V

This stage is applied both to reduce the cardinality of the set of vertices to make it feasible to computationally analyze larger corpora and to expand its members to equivalence classes

that bridge between variations which occur during non-verbatim citations or other mutations that happen to a text during its transmission. It entails the projection of the original text passages into a range space that eliminates items which are not significant enough to signal reuse instances and that merges equivalent items.

2. Dimensionality reduction in E

In this stage a number of computationally simple mechanism are applied to filter out implausible, unlikely or unwanted connections between text passages. This functions as a preparation to allow for more complex methods in the third step.

3. Scoring in E

This stage involves a detailed and sophisticated comparison of the remaining candidate pairs of text passages. The test applied here range from simple check for exact matches to alignment algorithms with complex scoring matrices.

Notably, one of the approaches does not obviously follow this scheme and instead applies a more monolithic procedure. The newer work by Scheirer, Forstall and Coffee is applying a single algorithm — Latent Semantic Indexing — that simultaneously reduces the datasets dimensionality and scores the remaining data points.

Implementations

1. Dimensionality reduction in V

Frequency thresholds come in 3 kinds - max pruning, min pruning and significance thresholds. Max pruning or stop-word deletion is removing the most frequent and common terms that are present in a large number of text passages and do not add identifying features. Min pruning or unique term removal is cutting the long tail of the term frequency table, as singular terms don't provide a feature that can be linked to another text passage (except when mapped to an equivalence class). Significance thresholds are a similar idea to stop-word deletion with the addition of a metric other than absolute frequency to determine the term's (or feature's) significance.

Equivalence classes for words can be spelling variants, lemmata, synonyms, concepts or topics and are formed by groupings the terms and their common transformations during citations. In addition to reducing the number of types and improving computational efficiency, they are widely used to find citations that vary from their source texts. The examined works used alignment algorithms such as the Levensthein distance, Needleman-Wunsch and

FastSS to detect spelling variants (and sometimes - as in Smith - even larger deviations); lemmatizer (e.g. LemLat) to map to lemmata; WordNet for finding synonyms; and semantic analysis and topic modeling methods such as pLSA, LSI and LDA to identify associated concept and topics.

W-shingling is a set of the n-grams in a text passage or its projection into a space of equivalence classes. Generally the n-grams are overlapping, but in some cases they can be disjoint.

Downsampling is a quasi-random removal of features or terms in a text passage. A popular implementation is the following.

2. Dimensionality reduction in E

Frequency thresholds repeat the min pruning from the previous step on the dimension-reduced set of vertices — unique n-grams and possibly n-grams that are too all-too frequent are being removed.

Feature selection is a qualitative filter on the edges and only removes (or keeps) edges that match predetermined criteria. This can be applied to search for certain kinds of text reuse or to eliminate systemic errors.

3. Scoring of E

Containment is a simple and fast test as to whether V1 is a substring of V2 or W1 a subset of W2, and variants test for the longest substring or largest subset.

Alignment algorithms are a more complex and flexible approach and commonly identify instances of text reuse that involve a change of word order.

Semantic analysis-based significance scores can be reused as well or - if performance is sufficient - used as a stand-alone replacement for the dimensionality reduction architecture (LSI).

Conclusion

There are large similarities and minute differences between most of the approaches to Historical Text Reuse Detection. The architecture of dimensionality reduction on vertices and edges, often using stop-word deletion and w-shingling, and subsequent scoring with varying sequence alignment algorithms is widely being deployed. Differences occur in the final scoring of candidate pairs and in specialization on specific textual species.

New approaches based on semantic analyses are being tested recently, with some benefits at the cost of certain limitations. More general, the goal to detect with high recall short

citations while accounting for an array of possible variations from its source remains unsolved, as character-level irregularities as well as changed word-orders are being recognized but further alteration still pose significant problems in particular with decreasing citation lengths.

References

M. Buechler, et al. (2014). Scaling Historical Text Re-Use. Proceedings of the IEEE International Conference on Big Data 2014

P. Clough, et al. (2002). METER: MEasuring TExt Reuse. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)

C.W. Forstall, S. Jacobson & W.J. Scheirer (2011). Evidence of intertextuality: investigating Paul the Deacon's Angustae Vitae. Literary and Linguistic Computing

J.G. Ganascia, P. Glaudes & A. Del Lungo (2014). Automatic detection of reuses and citations in literary texts. Literary and Linguistic Computing

C. Lyon, J. Malcolm & B. Dickerson (2001). Detecting short passages of similar text in large document collections. Proceedings of Conference on Empirical Methods in Natural Language Processing

G.H. Roe, et al. (2012). Intertextuality and Influence in the Age of Enlightenment: Sequence Alignment Applications for Humanities Research. Digital Humanities 2012 Conference Abstract

W.J. Scheirer, W. Forstall & N. Coffee (2014). The Sense of a Connection: Automatic Tracing of Intertextuality by Meaning. Digital Scholarship in the Humanities (DSH)

J. Seo and B.W. Croft (2008). Local text reuse detection. Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval

D. Smith, et al. (2014) Detecting and modeling local text reuse. Proceedings of the ACM +IEEE-CS Joint Conference on Digital Libraries

Text Re-use Detection, eine praktische Betrachtung

Michael Sünkel, Otto-Friedrich-Universität Bamberg

Aufbauend auf den im vorherigen Kapitel dargelegten Grundlagen und konzeptionellen Überlegungen soll nun exemplarisch ein praktischer Ansatz vorgestellt werden. Der folgende Bericht ist dabei angelehnt an die Dissertation von Marco Büchler (Büchler, 2013).
Um die Einsatzmöglichkeiten derartiger Ansätze und das breite Spektrum der Aufgabenstellungen zum Text Re-use anzudeuten soll einleitend eine Übersicht über die Arten des Text Re-use gegeben werden. Einerseits kann ein eher wörtlicher Re-use vorliegen, der *Syntactic Text Re-use*. Andererseits gibt es den eher semantisch ähnlichen Re-use, den *Semantic Text Re-use*. Ist die *Source* eines Text Re-use nicht Teil der Digital Library, dann spricht man von einem *Incomplete Text Re-use*. Im Folgenden finden Sie eine detailliertere Auflistung der Arten des Text Re-use nach (Büchler, 2013).

- ➢ Syntactic Text Re-use
  - ■ Idiomatic Text Re-use
    - ● Idiom
    - ● Winged Word
  - ■ Quotation
    - ● Verbatim
    - ● Near Verbatim
  - ■ Ciphering
  - ■ Patch Writing
  - ■ Edition
- ➢ Semantic Text Re-use
  - ■ Allusion
  - ■ Paraphrasing
    - ● Paraphrase
    - ● Analogy
    - ● Translation
  - ■ Ghostwriting
  - ■ Summarizing
- ➢ Incomplete Text Re-use
  - ■ Fragmentary Authors

- Incomplete Digital Library

Ein Tool, das bei der Erstellung eines Zitations- bzw. Text Re-use-Graphen behilflich sein kann, ist TRACER (Büchler, 2013), welches im Kontext des vom BMBF geförderten Projektes eTraces[11] entstanden ist. Hinter diesem Namen verbirgt sich die Implementierung einer Information Retrieval Pipeline, die sprachunabhängig dabei unterstützt, historischen Text Re-use zu erkennen. Die zugrundeliegende Architektur besteht aus 7 Schichten (Segmentation, Preprocessing, Featuring, Selection, Linking, Scoring, Postprocessing), welche im Folgenden näher erläutert werden.

1. Segmentation (Digital Library → Re-use Units)

Bei der Segmentierung wird eine Digitale Bibliothek in einzelne Re-use Units zerlegt, zwischen denen die "Ähnlichkeit" hinsichtlich Text Re-use bestimmt werden soll. Dabei unterscheidet man zwischen *Disjoint* und *Overlapping* Segmentation.

Je nach Größe und Ausdehnung der zu untersuchenden Text Re-uses muss hier eine andere Entscheidung getroffen werden. Erstrecken sich die Text Re-uses über ganze Sätze, entscheidet man sich für eine disjunkte Zerlegung nach Sätzen. Erstrecken sich Text Re-uses hingegen über Paragraphen oder noch größere Teile eines Dokumentes, darf man nicht so feingranular separieren. *Disjunkte Segmentierung* kann auf Zeilen-, Satz-, Absatz- und Dokumentebene geschehen.

Bei kürzeren Text Re-uses verwendet man *Overlapping Segmentation*. Die Länge einer Re-use Unit kann hier über die Anzahl der enthaltenen Token (Wörter) angegeben werden. Um die Re-use Units zu bilden wird das Segmentierungsfenster dabei jeweils um ein Token verschoben, sodass bei einer Re-use Unit Länge von *w*, ein bestimmtes Token in *w* Re-use Units vorkommt. Dieses Verfahren nennt man *Fixed Size Moving Window*. Die Größe einer Re-use Unit muss nicht festgelegt werden und kann auch dynamisch bestimmt werden (*Dynamic Size Moving Window*), was den Rechenaufwand aber deutlich erhöht.

Die Qualität der Text Re-use Detection kann bei falscher Segmentierung leiden. Man kann beispielsweise einen konkreten Text Re-use durch Segmentierung auf zwei benachbarte Re-use Units aufteilen. Anschließend ist der Re-use in keiner der beiden Units hinreichend signifikant zum Original.

2. Preprocessing (Re-use Units → Cleaned Re-use Units)

[11] http://etraces.e-humanities.net/; letzter Abruf: 21.08.2015

Der zweite und wichtigste Schritt ist das Preprocessing, welches die Re-use Units für die weitere Verarbeitung aufbereiten soll. Auf Buchstabenebene können hier unnötiger “Whitespace” und diakritische Zeichen entfernt werden, welche vor allem im Altgriechischen Wörter in syntaktischer Hinsicht verschieden machen, was bei der Text Re-use Analyse aber oft nicht erwünscht ist. Ein weiterer Preprocessingschritt, der auf Buchstabenebene durchgeführt werden kann, ist die Konvertierung aller Groß- zu den entsprechenden Kleinbuchstaben. Viele antike Texte wurden ausschließlich in Großbuchstaben verfasst und Groß- / Kleinschreibung erst von Editoren eingeführt.

Die einzigen sprachabhängigen Bearbeitungsschritte finden im wortbasierten Preprocessing statt. Diese kann man jedoch frei konfigurieren, was das Tool an sich sprachunabhängig macht. Linguistische Ansätze, wie Lemmatisierung und Synonym- und Kohyponymbetrachtung finden hier lexikonbasiert statt, d. h. man gibt selbst ein Wörterbuch an, das flektierte Wortformen auf Stammformen abbildet bzw. Wörtern semantisch identische Wörter oder Oberbegriffe zuordnet. Besitzt man für die zu untersuchende Sprache beispielsweise nur einen regelbasierten bzw. heuristischen Stemmer, muss man sich zuerst mit Hilfe des Vokabulars seiner Digitalen Bibliothek und des Stemmers selbst eine lexikonbasierte Lemmatisierung erzeugen. Diese linguistischen Methoden kommen vor allem dann zum Einsatz, wenn man nach semantischem Text Re-use sucht. Bei wörtlichem Re-use würde es wenig Sinn machen, mit Lemmatisierung, Synonymen und Kohyponymen zu arbeiten.

Neben den sprachabhängigen gibt es noch die sprachunabhängigen Methoden der Zeichenkettenverarbeitung. Man kann beispielsweise mit *Word Length Replacement* ein Wort durch seine Wortlänge darstellen. Außerdem finden *T9-like Recoding*, *Length Reduced Words* und *String Similarity* Anwendung. Sinnvollen Einsatz haben diese Methoden bei Sprachevolution, Dialekten und unterschiedlicher Rechtschreibung.

3. Featuring (Cleaned Re-use Units → Digital Fingerprint)

Re-use Units können, wie sie bisher vorliegen, nur sehr schlecht verglichen werden. Um das Vergleichen performanter zu gestalten, werden die Re-use Units in sogenannte Features unterteilt. Danach repräsentiert man eine Re-use Unit durch einen Vektor, der für jedes Feature die Featurehäufigkeit enthält. Diese Repräsentation wird Digitaler Fingerabdruck genannt. Die Dokumentenkollektion kann auf diese Weise als Matrix dargestellt werden, in der die einzelnen Re-use Units den Zeilen und die Features den Spalten entsprechen.

Die am stärksten strukturierenden und differenzierenden Featuring Methoden nennt man *Syntactical Featuring*. Diese verwendet man vorzugsweise bei *Duplicate Detection*, *Near*

Duplicate Detection und *Plagiarism Detection*. Die Re-use Units werden hier in *n*-grams zerlegt. Überlappen sich diese und haben alle *n*-grams eine konstante Länge, so nennt man dieses Verfahren *Shingling*. Bei einer Zerlegung in *n*-grams mit einer dynamischen, von der Featurehäufigkeit abhängigen, Länge, spricht man vom *Longest Common Consecutive Words* Verfahren. Weitere überlappende *n*-gram Techniken sind *Sparse Orthogonal Bigrams* und *Distance-based Bigrams*. Diese beziehen die Distanz der beiden Wörter eines Bigrams mit ein und werden so robuster gegenüber Einschüben.

Zerlegt man eine Re-use Unit in *n*-grams ohne Überlappung, spricht man vom *Hashbreaking*. Da man ohne Überlappung deutlich weniger Features extrahiert, resultiert daraus in den meisten Fällen ein signifikanter Performance-Vorteil. Man unterscheidet zwischen *Local* und *Global Hashbreaking*. Das Adjektiv bezieht sich auf die Art der Information, die für die Zerlegung herangezogen wird. Beim *Local Hashbreaking* wird die Position eines Wortes innerhalb einer Re-use Unit verwendet. Beim *Global Hashbreaking* wird der Rang eines Features innerhalb der gesamten Dokumentenkollektion herangezogen. Hashbreaking ist wiederum gut für Duplikats- und Plagiatserkennung geeignet.

Neben dem Syntactical Featuring gibt es das *Semantic Featuring*. Zu den vertretenen Methoden zählen *Word Bigram* und *Co-occurrence*. Bei der Co-occurence Methode werden nicht nur benachbarte Wörter zu Bigrams gebildet, sondern auch entferntere. Dies ermöglicht eine bessere Erkennung semantischer Arten Text Re-use.

Als letzte Gruppe der *Featuring* Methoden müssen noch die *Non-statistic Approaches* genannt werden. Diese zeichnen sich dadurch aus, dass die Techniken dieser Klasse sehr gut streamingfähig sind. Zum Einsatz kommen *Pattern-based Approaches* wie *<ENTITY> <VERBUM DICENDI> Pattern*, *Canonical References* und *Surface Features*. Diese setzen voraus, dass der Editor eines Textes entsprechende Marker im Text gesetzt hat. Außerdem stehen noch *Signal Processing Techniques* zur Verfügung. Dazu zählen *Discrete Cosine Transformation*, *Fourier Transformation* und *Wavelets*. Dabei wird der Text als fortlaufendes Signal verstanden, ähnlich wie ein Audio- oder Videosignal.

4. Selection (Digital Fingerprint → Signature)

Uns liegen nun die in Features aufgeteilten Re-use Units vor. Nun stellt sich die Frage, welche der Features relevant und gut beschreibend sind und welche nicht. Im Selection Schritt geht es darum, irrelevante Features zu entfernen. Es gibt eine sehr große Anzahl an Selection Strategien, weswegen im Folgenden nur die vielversprechendsten genannt werden.

Die Selection Methoden lassen sich nach zwei Gesichtspunkten kategorisieren. Einerseits kann nach *Selection Knowledge* differenziert werden. Verwendet man nur *Local Selection Knowledge*, nutzt man lediglich Informationen, die beim Betrachten einer Re-use Unit zur Verfügung stehen, wie z. B. PoS-Tags oder Wortlängen. Beim *Global Selection Knowledge* nutzt man hingegen Informationen bezogen auf die gesamte Dokumentenkollektion. Wissen über Verteilungen und Abhängigkeiten zwischen Features kann man hierfür als Beispiele anführen. Andererseits kann man die Selektion auf zwei unterschiedlichen Ebenen durchführen, d. h. nach *Selection Usage* unterscheiden. Beschränkt man die Selektierung auf eine Re-use Unit, spricht man von *Local Selection Usage*. Übertragen auf die Feature-Matrix, die man nach dem Featuring erhält, werden hier Einträge durch Selektion auf 0 gesetzt, was das Entfernen eines Features aus einer Re-use Unit bedeutet. Beim *Global Selection Usage* hingegen, werden gesamte Spalten aus der Feature-Matrix entfernt, was im Gegensatz zur lokalen Methode zu einer deutlich reduzierten Datenmenge führt. Die resultierende Matrix wird *Signature-Matrix* genannt.

Die beiden einfachsten Selection Strategien sind das *min pruning* und *max pruning*. Dabei werden Features entfernt, die global gesehen eine Häufigkeit aufweisen, die unter bzw. über einem gewissen Schwellenwert liegen. Es ist beispielsweise empfehlenswert, alle Features mit einer Auftrittshäufigkeit von 1 zu entfernen, da diese auf keinen Fall für einen Text Re-use in Frage kommen. Außerdem kann man sehr häufige Features entfernen, da diese oft nur aus Funktions- bzw. Stoppwörtern bestehen. Allerdings muss man auf die zu erwartenden Arten von Text Re-use achten. Nach einem max pruning wird das Zitat von Shakespeare “Sein oder nicht sein” wohl nicht mehr erkannt werden.

Neben dem *Pruning* kommen noch

- *Depth Reduction* Strategien, wie *Local Random*, *Global Random*, *Global 0 mod p*, *Minimum Word Length* und *Winnowing Selection*,
- *Term Weighting* Strategien, wie *Inverted Feature Frequency*, *tf.idf* und *Frequency Class Selection*,
- *Information Theory* Strategien, wie *Self Information*, *Entropy* und *Redundancy Selection*,
- *Statistical Measure* Strategien, wie *Log Likelihood Ratio* und *Kullback Leibier Divergence Selection*,
- *Feature Dependency* Strategien, wie *Feature Correlation* und *Contrastive Semantics Selection*,
- *Global Word Class Selection und*
- *Inverted Category Index Selection*

zum Einsatz.

5. Linking (Signature → Candidate List)

Das Linking ist der zeitaufwändigste Schritt, mit einer Laufzeitkomplexität von $O(n^2)$. Für jedes Feature wird ein Link zu allen anderen Re-use Units gesetzt, die das gleiche Feature beinhalten. Dies kann man sich anhand der Link-Matrix veranschaulichen. Führt man eine Matrixmultiplikation der Signature-Matrix mit der transponierten Signature-Matrix durch, so erhält man die Link-Matrix, die für alle Re-use Paare, die Anzahl der gemeinsamen Features enthält. Die entstandenen Links werden auch Candidate List genannt, da sie potentielle Text Re-uses darstellen.

Die Linking Strategie, die im TRACER Tool verwendet wird, ist *Intra Digital Library Linking*, da die gesamte Berechnung innerhalb einer Dokumentenkollektion geschieht. Dem steht das *Inter Digital Library Linking* gegenüber, welches übergreifend über mehrere Digitale Bibliotheken arbeitet.

6. Scoring (Candidate List → Result List)

Das Scoring entscheidet, welche Kandidaten akzeptiert werden und in die tatsächliche *Result List* übernommen werden. Dafür muss die Ähnlichkeit der verlinkten Re-use Units bewertet werden, was entweder basierend auf Features oder Wörtern durchgeführt werden kann. Ausgehend von der Link-Matrix, lässt sich eine Scoring-Matrix berechnen, die anstelle der Anzahl der gemeinsamen Features einen Score führt. Nach dem Scoring kann auch der Re-use Graph berechnet werden.

Metriken wie *tf.idf* funktionieren gut, wenn man sie auf Dokumentebene einsetzt, jedoch sind Re-use Units meistens Sätze oder noch kleinere *Moving Windows*, was es schwierig macht, *Term Weighting*-Techniken zu verwenden. Daher setzt man auf mengenorientierte *Similarity & Distance Measure* Methoden. Zu diesen zählen *Overlap*, *Resemblance*, *Weighted Resemblance*, *Word Class Weighting*, *Containment* und *Weighted Containment*.

Neben den Similarity Measures können auch *Statistical Measures* verwendet werden. Hierzu zählen u. a. *Cross Validation*, *Mutual Information*, *Log-Likelihood-Ratio*, *Perplexity* und *Conditional Probability*. Diese finden im Bereich einer Text Re-use Analyse aber kaum Anwendung, da sie fast jeden potentiellen Text Re-use als statistisch signifikant einstufen würden, weil die Features einer *Power Law*-Verteilung unterliegen.

7. Postprocessing (Result List → Reduced Result List)

Eine Text Re-use Analyse produziert auf einer großen Digitalen Bibliothek ebenfalls eine große Menge an Text Re-use Daten, wodurch man sich schnell in der Datenmenge verlieren kann. Das *Postprocessing* kann daher als ein Nachbearbeitungsschritt verstanden werden, der für eine konkrete fachwissenschaftliche Fragestellung einen Re-use Graphen entsprechend nachbereitet und gegebenenfalls nicht relevante Daten entfernt.
Faktisch kann jede Graph Mining-Technik auf einen Re-use Graphen angewendet werden. Deswegen sind die folgenden Postprocessing Methoden nur ein Auszug aus vielen.

- *Bibliometry* (*PageRanking, h-Index, Impact Factor*)
- *Phonetic Postprocessing* (*Alliteration, Rhyming, Metric Analysis*)
- *Cluster Analysis* (*Author Disambiguation, Weak Ties Analysis, Re-use Cluster Analyis*)
- Further Graph Mining Approaches (*Temperature Analysis, Dotplot Analysis, Noisy Channel Mining, Text Decontamination, Author Dating*)

Querbezüge und weitere Quellen

Im vorliegenden Kapitel haben wir exemplarisch einen Workflow für die Erkennung und Untersuchung von Text Re-use nach Büchler (2013) beschrieben. Für die weitere Recherche bieten sich die Arbeit von Büchler selbst sowie die im vorherigen Kapitel angegebene Literatur an.
Mit der Aufgabenstellung der Erkennung von Text Re-use nah verwandt sind dabei -- wie oben bereits angerissen -- auch Fragestellungen der Plagiatserkennung (vergleiche z. B. Martins et al., 2014 oder Potthast et al., 2014) sowie Fragestellungen der Near Duplicate Detection -- z. B. zur Erkennung von Varianten oder Versionen eines Dokumentes -- (vergleiche z. B. Alonso et al., 2013 oder Potthast & Stein, 2008). Die Techniken, die in den Forschungsfeldern Text Re-use Erkennung, Plagiatserkennung sowie Near Duplicate Detection angewendet werden überlappen sich stark und auch die Forschungsfelder selbst sind nicht klar voneinander abzugrenzen, so dass man bei einer Recherche breit vorgehen sollte.

Literatur

Omar Alonso, Dennis Fetterly and Mark Manasse: *Duplicate News Story Detection Revisited.* Microsoft Research Technical Report MSR-TR-2013-60, May 2013, http://research.microsoft.com/apps/pubs/default.aspx?id=193265

Marco Büchler: I*nformationstechnische Aspekte des Historical Text Re-use*, Dissertation, 2013. http://nbn-resolving.de/urn:nbn:de:bsz:15-qucosa-108515

Martins, V. T., Fonte, D., Henriques, P. R., & da Cruz, D. (2014). Plagiarism detection: A tool survey and comparison. In OpenAccess Series in Informatics (OASIcs), 4566. http://drops.dagstuhl.de/opus/volltexte/2014/4566/

Martin Potthast, Matthias Hagen, Anna Beyer, Matthias Busse, Martin Tippmann, Paolo Rosso and Benno Stein: *Overview of the 6th International Competition on Plagiarism Detection*. In Working Notes for CLEF 2014 Conference, Sheffield, UK, September 2014, p. 845--876, http://ceur-ws.org/Vol-1180/CLEF2014wn-Pan-PotthastEt2014.pdf

Martin Potthast and Benno Stein: New Issues in Near-duplicate Detection. in Preisach, Burkhardt, Schmidt-Thieme, Decker (Eds.): Data Analysis, Machine Learning and Applications, Selected Papers from the 31th Conf. of the German Classification Society Berlin, ISBN 978-3-540-78239-1, pp. 601-609, Springer 2008.