Bauhaus-Universität Weimar
Fakultät Medien
Studiengang Mediensysteme

Datierung von Textdokumenten

Masterarbeit

Tsvetomira Boycheva Palakarska
Geboren am 02.07.1985 in Sofia

Matrikelnummer 30994

1. Gutachter: Prof. Dr. Benno Stein
Betreuer: Dr. Matthias Hagen und Dr. Martin Potthast

Datum der Abgabe: 3. Januar 2012

Danksagung

An dieser Stelle möchte ich allen danken, die durch ihre Unterstützung zum Gelingen dieser Masterarbeit beigetragen haben. Zunächst bedanke ich mich bei Herrn Prof. Dr. Benno Stein für die Überlassung des Themas dieser Masterarbeit. Ein besonders großer Dank geht auch an meine wissenschaftlichen Betreuer Herrn Dr. Matthias Hagen und Herrn Dr. Martin Potthast, die mir während der gesamten Zeit sehr motivierend, kompetent und hilfsbereit zur Seite standen.

Erklärung

Hiermit versichere ich, daß ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Weimar, den 3. Januar 2012

.....
Tsvetomira Boycheva Palakarska

Zusammenfassung

Die vorliegende Arbeit untersucht die Datierung von Textdokumenten. Dabei gilt es, das Jahr und somit implizit auch das Jahrzehnt und das Jahrhundert zu bestimmen, in dem ein Textdokument geschrieben wurde. Die automatische Datierung ist sowohl zur historischen Einordnung von Dokumenten, im Kontext der forensischen Analyse als auch in vielen Anwendungen des Information Retrieval von Bedeutung.

Zur Datierung von Textdokumenten werden Unigramm-Sprachmodelle auf Jahresbasis eingesetzt. Erstmals werden diese anhand des englischen Google-Books-N-Gramm-Korpus erstellt. Dieses geschieht für die Jahre 1800 bis 2008. Für ein zu datierendes Textdokument werden ebenfalls Unigramm-Sprachmodelle für dieselben Jahre gebildet. Die Sprachmodelle des Korpus und des undatierten Textdokuments werden mit zwei Rankingfunktionen, Produktionswahrscheinlichkeit und Kullback-Leibler-Divergenz, in Beziehung gesetzt. Auf Basis der Top-10 Rangordnungen kann das Jahr eines bestimmten Sprachmodells des Korpus schließlich als Datum für das zu datierende Textdokument bestimmt werden. Hierfür werden beispielsweise die Klassifizierer Top-1, Durchschnittliches-Jahr-aus-den-Top-10 und Zufälliges-Jahr-aus-den-Top-10 eingesetzt.

Bei den zu datierenden Textdokumenten handelt es sich um künstlich erzeugte Textdokumente, Zeitungsartikel und Belletristik verschiedener Textlängen. Auf den künstlich erzeugten Textdokumenten ist eine Datierungsgenauigkeit für Jahre zu 100% möglich. Diese Genauigkeit wird durch den Top-1-Klassifizierer bei dem Kullback-Leibler-Divergenz-Ranking ab einer Textlänge von 500 Wörtern erreicht. Für Zeitungsartikel und Belletristik bleibt diese Datierungsgenauigkeit unerreicht. Für Zeitungsartikel liegt die höchste Datierungsgenauigkeit bei 11.11% beim Produktionswahrscheinlichkeits-Ranking unter Verwendung des Klassifizierers Durchschnittliches-Jahr-aus-den-Top-10 für 1 000 Wörter. Für Belletristik beträgt die höchste Datierungsgenauigkeit 1.63%. Diese wird durch den Klassifizierer Zufälliges-Jahr-aus-den-Top-10 beim Produktionswahrscheinlichkeits-Ranking bei einer Textlänge von 5 000 Wörtern erzielt.

Inhaltsverzeichnis

1 Einleitung **3**

2 Grundlagen von Sprachmodellen **5**
2.1 N-Gramm-Modelle 5
2.2 Glättungstechniken 9
2.3 Distanz zwischen N-Gramm-Modellen 11

3 Verwandte Arbeiten **13**
3.1 Datierung durch Unigramm-Modelle 14
3.1.1 Datierung nach [de Jong u.a., 2005] 14
3.1.2 Datierung nach [Kanhabua und Nørvåg, 2008] 16
3.2 Datierung durch Heuristiken und Klassifikationsmethoden . . . 18
3.3 Zusammenfassung 21

4 Unsere Verfahren zur Datierung von Textdokumenten **23**
4.1 Google-Books-N-Gramm-Korpus 24
4.2 Rankingfunktionen 26
4.2.1 Produktionswahrscheinlichkeit 26
4.2.2 Kullback-Leibler-Divergenz 27
4.3 Klassifizierer 28

5 Evaluierung **31**
5.1 Korpora 31
5.2 Bewertungsmaße 33
5.3 Experimentaufbau 35
5.4 Experimente 38
5.5 Diskussion 48

6 Zusammenfassung und Ausblick **49**

A Tabellen zur Evaluierung **52**

Abbildungsverzeichnis **57**

Tabellenverzeichnis **58**

Literaturverzeichnis **59**

Kapitel 1

Einleitung

Ziel der automatischen Datierung von Textdokumenten ist die Bestimmung des Zeitraums, in dem ein Text geschrieben wurde. Mit dieser Information lassen sich Dokumente in ihren historischen Kontext einordnen und mit anderen Dokumenten in Beziehung setzen. Dies ist nicht nur für Historiker, Geisteswissenschaftler und Kriminologen von Bedeutung, sondern auch für viele Anwendungen des Information Retrieval. Im Bereich der Websuche zum Beispiel wird die Datierung dazu verwendet, um das Ranking der Suchergebnisse für den Nutzer aufzubereiten. Ziel ist es, dem Nutzer sowohl relevante als auch möglichst aktuelle Dokumente zu präsentieren.

Verfahren zur Datierung folgen dabei drei Paradigmen: (1) Verfahren, die Metainformationen wie das Publikationsdatum oder den Autor eines Textdokuments verarbeiten. Dies setzt allerdings voraus, dass diese Metainformationen verlässlich bestimmt werden können. (2) Stilbasierte Verfahren analysieren den im Textdokument verwendeten Schreibstil. Um den Schreibstil zu erfassen, werden Stilmerkmale verwendet. Darunter fallen unter anderem der Lesegrad oder die Verständlichkeit des Textes. (3) Inhaltsbasierte Verfahren verwenden zur Datierung zeitliche Ausdrücke, Schlüsselwörter oder alle Wörter des gesamten Textdokuments. Hierbei werden die Worthäufigkeiten über die Zeit betrachtet. Beispielsweise enthält ein Dokument aus der Zeit von Goethe im Gegensatz zu einem Text aus dem 21. Jahrhundert häufiger Wörter wie „gescheuter“, „Würckungskrafft“, „Tiergeripp“ und „Wissensqualm“. Ebenso ist es wahrscheinlicher, dass ein Dokument, welches nach dem 11. März 2011 veröffentlicht wurde, häufiger Wörter wie „Nuklearkatastrophe“ und „Fukushima“ enthält. Um solche Worthäufigkeiten für verschiedene Zeiträume zu erfassen, werden sogenannte Sprachmodelle eingesetzt. Diese werden häufig in der aktuellen Forschung zur Datierung verwendet.

In der vorliegenden Arbeit werden neue Verfahren zur Datierung von Textdokumenten auf Basis von Spachmodellen vorgestellt. Im Gegensatz zur beste-

henden Forschung werden erstmals Sprachmodelle auf Basis des Google-Books-N-Gramm-Korpus erstellt. Im Vergleich zu bisherigen Korpora, die ausschließlich auf Zeitungsartikeln basieren, bietet das Google-Books-N-Gramm-Korpus eine weitaus umfangreichere Korpusgröße, welche der Genauigkeit der Sprachmodelle zu Gute kommt. Des Weiteren grenzt sich diese Arbeit von aktueller Forschung ab, indem zwei Rankingfunktionen, Produktionswahrscheinlichkeit und Kullback-Leibler-Divergenz, und auch zahlreiche Klassifizierer zur Datierung von Textdokumenten untersucht werden. In den Experimenten werden nicht nur Zeitungsartikel, sondern erstmals auch Belletristik und künstlich erzeugte Dokumente berücksichtigt.

Die Arbeit ist wie folgt gegliedert: Das nachfolgende Kapitel 2 stellt die theoretischen Grundlagen von Sprachmodellen vor. Im Kapitel 3 werden verwandte Arbeiten zur Datierung in der aktuellen Forschung betrachtet. Im Anschluss werden die neu entwickelten Verfahren zur Datierung von Textdokumenten in Kapitel 4 vorgestellt. Experimente zur Güte der entsprechenden Verfahren werden in Kapitel 5 durchgeführt und deren Ergebnisse diskutiert. Eine Zusammenfassung und ein Ausblick schließen die Arbeit in Kapitel 6 ab.

Kapitel 2

Grundlagen von Sprachmodellen

Sprachmodelle werden in vielen Anwendungen der natürlichen Sprachverarbeitung eingesetzt, beispielsweise zur Spracherkennung [Jelinek u.a., 1991], Handschrifterkennung [Russell und Norvig, 2003], Rechtschreibkorrektur [Kukich, 1992], maschinellen Übersetzung [Brown u.a., 1990] und Part-of-Speech-Tagging [Schütze und Singer, 1994]. Ein Sprachmodell erfasst dabei die statistischen Gesetzmäßigkeiten einer Sprache, indem es eine Wahrscheinlichkeitsverteilung über das Vokabular dieser Sprache bildet. Jede Wortsequenz gehört somit mit einer bestimmten Wahrscheinlichkeit zu dieser Sprache.

Zur Erstellung eines Sprachmodells werden Beispieltexte benötigt, welche das sogenannte Trainingskorpus bilden. Zur Veranschaulichung: Gegeben sei ein Trainingskorpus, das aus Texten zur Geschichte deutscher Redewendungen besteht. Ein Sprachmodell, das darauf basiert, würde der Redewendung „Der frühe Vogel fängt den Wurm“ eine höhere Wahrscheinlichkeit zuweisen als der Redewendung „Der frühe Wurm fängt den Vogel“, da letztere keine Redewendung ist.

In der vorliegenden Arbeit werden Sprachmodelle auf Basis von N-Grammen, sogenannte N-Gramm-Modelle, eingesetzt. Hierzu werden in diesem Kapitel die theoretischen Grundlagen der N-Gramm-Modelle vorgestellt. Weiterhin werden Techniken zur Glättung von N-Gramm-Modellen sowie die Berechnung der Distanz für N-Gramm-Modelle betrachtet.

2.1 N-Gramm-Modelle

Ein Sprachmodell stellt eine Wahrscheinlichkeitsverteilung der Wörter w aus einem Vokabular V dar. Für ein Sprachmodell M über V gilt [Manning u.a., 2008]:

$$\sum_{w \in V} \Pr(w) = 1.$$

Dem Sprachmodell M liegt ein generatives Modell zugrunde. Wörter werden demnach iterativ „ausgewürfelt“: Jedes Wort w wird nach der Produktionswahrscheinlichkeit $\Pr(w \mid M)$ erzeugt. Diese Produktionswahrscheinlichkeit gibt an, wie wahrscheinlich w als zufällige Stichprobe aus M beobachtet wird (entsprechend auch für Wortsequenzen $w_1 \dots w_n$). Die Berechnung der Produktionswahrscheinlichkeit ist abhängig von der Wahl des Sprachmodells [Rosenfeld, 2000]. Die am häufigsten eingesetzten Sprachmodelle basieren auf N-Gramm-Modellen.

Ein N-Gramm ist eine geordnete Menge von N Wörtern. Zum Beispiel besteht ein Unigramm aus einem, ein Bigramm aus zwei, ein Trigramm aus drei Wörtern. Das N-Gramm-Modell bestimmt die Wahrscheinlichkeit eines Wortes w_k durch die vorhergehenden $N-1$ Wörter $w_{k-(N-1)} \dots w_{k-1}$. Diese bilden die Vergangenheit[1] des aktuellen Wortes. Beim Unigramm-Modell wird jedes Wort als unabhängig von den anderen betrachtet. Somit bezieht das Unigramm-Modell keine Vergangenheit bei der Berechnung der Wahrscheinlichkeit mit ein. Ein Bigramm-Modell bestimmt die Wahrscheinlichkeit eines Wortes w_k in Abhängigkeit vom vorhergehenden Wort w_{k-1}, wogegen ein Trigramm-Modell eine Geschichte von zwei Wörtern $w_{k-2}w_{k-1}$ bei der Berechnung der bedingten Wahrscheinlichkeit berücksichtigt. Diese Annahme ist für N-Gramm-Modelle höherer Ordnung verallgemeinerbar und wird als Markov-Annahme bezeichnet [Jurafsky und Martin, 2008]:

$$\Pr(w_k \mid w_1 \dots w_{k-1}) \approx \Pr\left(w_k \mid w_{k-(N-1)} \dots w_{k-1}\right).$$

Unter Berücksichtigung der Markov-Annahme wird die Produktionswahrscheinlichkeit mit einem N-Gramm-Modell wie folgt berechnet:

$$\Pr(w_1 \dots w_n \mid M) = \prod_{k=1}^{n} \Pr\left(w_k \mid w_{k-(N-1)} \dots w_{k-1}\right).$$

Um eine Vergangenheit für das erste Wort w_1 herzustellen, werden $N-1$ Zeichen <s> dem Wort w_1 vorangestellt [Jurafsky und Martin, 2008]. Diese Erweiterung entfällt bei der Berechnung der Produktionswahrscheinlichkeit mit einem Unigramm-Modell. Die Formel beim Unigramm-Modell vereinfacht sich zu:

$$\Pr(w_1 \dots w_n \mid M) = \prod_{k=1}^{n} \Pr(w_k).$$

Zur Veranschaulichung: Enthält ein Text nur die fünf Wörter „der“, „frühe“, „Vogel“, „fängt“, „den“ und „Wurm“, könnte ein mögliches Unigramm-Modell mit den Wahrscheinlichkeiten für jedes Wort wie folgt aussehen: (0.30, 0.15,

[1] Oft auch als Geschichte, Kontext oder Gedächtnis bezeichnet.

0.2, 0.18, 0.05, 0.12). Da beim Unigramm-Modell jedes Wort unabhängig vom Kontext betrachtet wird, kann die Produktionswahrscheinlichkeit jeder möglichen Wortkombination berechnet werden. Die Sequenz „der fängt Wurm“ hat in diesem Fall eine Produktionswahrscheinlichkeit von:

$$\begin{aligned} \Pr(\textit{der fängt Wurm} \mid M) &= \\ \Pr(\textit{der}) \Pr(\textit{fängt}) \Pr(\textit{Wurm}) &= \\ 0.30 \times 0.18 \times 0.12 &= \\ 0.00648\ . \end{aligned}$$

Zu beachten ist, dass es bei der Berechnung der Produktionswahrscheinlichkeit einer Wortsequenz als Produkt der Wahrscheinlichkeiten der Wörter zu einem numerischen Unterlauf kommen kann auf Grund von kleinen Werten. Anstatt der Multiplikation, wird daher die numerisch weniger „auffällige“ Addition der Logarithmen der Wahrscheinlichkeiten verwendet [Ney u.a., 1997].

Markov-Ketten

Prinzipiell gilt: Je höher die Ordnung des N-Gramm-Modells, desto genauer ist die Bestimmung der bedingten Wahrscheinlichkeiten und somit der gesamten Produktionswahrscheinlichkeit. Dies geht jedoch mit einem höheren Rechenaufwand einher, da N-Gramm-Modelle durch Markov-Ketten der Ordnung $N - 1$ repräsentiert werden.

Eine Markov-Kette ist ein gewichteter endlicher Automat, bei dem die Wörter einer Wortsequenz die Zustände darstellen. Die Übergänge zwischen den Zuständen werden durch die bedingten Wahrscheinlichkeiten der Zustände bestimmt und in einer Übergangsmatrix zusammengefasst [Jurafsky und Martin, 2008]. Da die Anzahl der Dimensionen der Übergangsmatrix der Ordnung N des N-Gramm-Modells entspricht, wächst die Komplexität exponentiell mit der Ordnung. Bei einer Vokabulargröße $|V|$ besitzt ein N-Gramm-Modell $|V|^N$ freie Parameter, welche die bedingten Wahrscheinlichkeiten (Übergängswahrscheinlichkeiten) darstellen. Beträgt die Vokabulargröße beispielsweise 20 000 Wörter und wird ein Bigramm-Modell verwendet, so ist die Anzahl der freien Parameter $20\,000^2 = 4 \times 10^8$; hingegen sind es bei einem Trigramm-Modell schon $20\,000^3 = 8 \times 10^{12}$ freie Parameter.

Beispiele für Markov-Ketten unterschiedlicher Ordnung sind in Abbildung 2.1 dargestellt. Eine Markov-Kette nullter Ordnung entspricht einem Graphen ohne Übergänge (siehe erste Zeile in der Abbildung), wobei bei der Markov-Kette erster Ordnung jeweils ein Übergang von einem Zustand zum nachfolgenden verläuft (siehe mittlere Zeile in der Abbildung). Bei der Markov-Kette zweiter Ordnung (siehe letzte Zeile in der Abbildung) sind es zwei

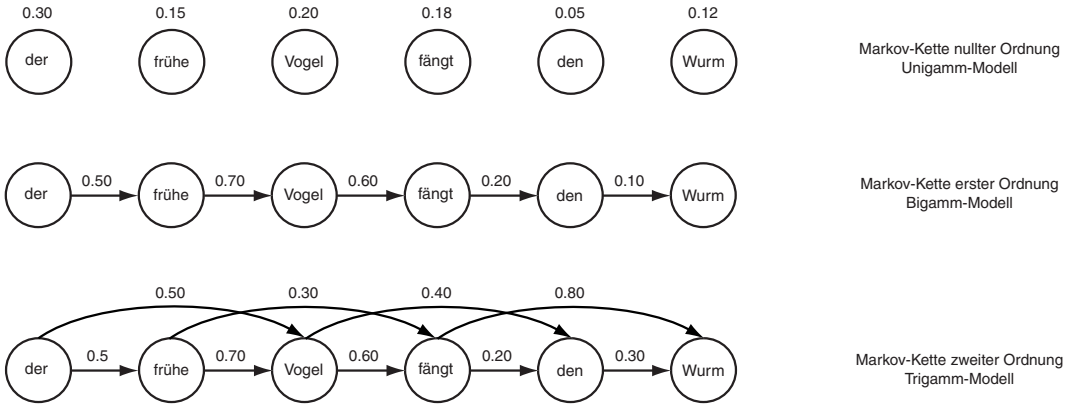


Abbildung 2.1: Markov-Ketten der Ordnungen Null bis drei für die Redewendung „der frühe Vogel fängt den Wurm“; adaptiert nach [Bishop, 2007].

Übergänge zu den zwei nachfolgenden Zuständen. Jeder solche Übergang bzw. jeder Wert eines solchen Übergangs stellt einen freien Parameter des N-Gramm-Modells dar.

Trainingskorpus und Testkorpus

Für die Schätzung der $|V|^N$ Parameter eines N-Gramm-Modells wird ein Trainingskorpus benötigt, welches eine Sammlung aus Texten desselben Genres darstellt, wie zum Beispiel Fachzeitschriften oder Science-Fiction-Literatur.

Die erstellten N-Gramm-Modelle können verwendet werden, um die Produktionswahrscheinlichkeiten von Texten eines Testkorpus zu berechnen oder die Distanz zu einem anderen N-Gramm-Modell zu berechnen. Das Trainingskorpus und das Testkorpus müssen sich dabei unabhängig auf dasselbe Themengebiet (Genre) beziehen. Weiterhin müssen beide Korpora disjunkt sein. Sind in beiden Korpora Texte vertreten, die aus unterschiedlichen Genres stammen oder sich in beiden Korpora wiederholen, so sind die berechneten N-Gramm-Modelle für gewöhnlich unbrauchbar [Jurafsky und Martin, 2008].

Die Texte beider Korpora werden einheitlich vorverarbeitet. Zu den Vorverarbeitungstechniken gehören üblicherweise das Entfernen der Stoppwörter und der Satzzeichen und die Rückführung der Wörter auf ihre Grundformen (Stemming). Des Weiteren werden zu den Sätzen nicht nur die Satzanfangszeichen <s>, sondern auch ein Satzendezeichen <\s> hinzugefügt. Das Satzendzeichen wird benötigt um aus einem N-Gramm-Modell eine „echte“ Wahrscheinlichkeitsverteilung zu machen [Chen und Goodman, 1996, Jurafsky und Martin, 2008].

Parameterschätzung

Die Parameterschätzung eines N-Gramm-Modells aus einem Trainingskorpus wird als Training bezeichnet. Dabei werden die absoluten Häufigkeiten der N-Gramme im Trainingskorpus verwendet. Die Schätzung erfolgt mit der Maximum-Likelihood-Schätzung[2] (ML-Schätzung). Diese gibt die bedingte Wahrscheinlichkeit eines Worts w_k unter einem N-Gramm-Modell wie folgt an:

$$\Pr_{ML}\left(w_k \mid w_{k-(N-1)} \dots w_{k-1}\right) = \frac{C\left(w_{k-(N-1)} \dots w_{k-1} w_k\right)}{\sum_{w \in V} C\left(w_{k-(N-1)} \dots w_{k-1} w\right)} \,, \qquad (2.1)$$

wobei C die absolute Häufigkeit einer Wortsequenz in einem gegebenen Trainingskorpus bezeichnet. Im Fall eines Unigramm-Modells wird die Wahrscheinlichkeit eines Unigramms durch seine relative Häufigkeit im Trainingskorpus ermittelt:

$$\Pr_{ML}\left(w_k\right) = \frac{C\left(w_k\right)}{\sum_{w \in V} C\left(w\right)} \,.$$

2.2 Glättungstechniken

Bei der ML-Schätzung besteht das Problem der knappen Daten in einem Trainingskorpus. Für Wortsequenzen $w_1 \dots w_n$ oder Wörter, die nicht im Trainingskorpus auftreten[3], sind die Häufigkeiten $C\left(w_1 \dots w_n\right)$ und damit die Wahrscheinlichkeiten $\Pr_{ML}\left(w_1 \dots w_n\right)$ gleich Null. Daher werden Glättungstechniken[4] eingesetzt, welche die Wahrscheinlichkeitsmasse der vorkommenden N-Gramme umverteilen, so dass die Wahrscheinlichkeiten auch für ungesehene Wörter definiert sind. Die ML-Schätzer werden wie folgt angepasst: Wörter, die im Trainingskorpus vorkommen, erhalten eine niedrigere Wahrscheinlichkeit und ungesehene Wörter erhalten eine Wahrscheinlichkeit größer Null.

Umfassende Studien zu Glättungstechniken für N-Gramm-Modelle werden in [Chen, 1998], [Goodman, 2000] und [Zhai und Lafferty, 2004] vorgestellt. Im Folgenden wird auf die Add-Delta-Glättung sowie auf die Jelinek-Mercer-Glättung eingegangen. Die Parameter beider Glättungstechniken werden auf einem Held-Out-Korpus bestimmt [Jurafsky und Martin, 2008]. Dieses Korpus stammt wie das Trainings- und Testkorpus aus demselben Genre. Alle drei Korpora sind dabei disjunkt.

[2] engl. Maximum Likelihood Estimation

[3] Sogenannte ungesehene oder fehlende Wortsequenzen bzw. Wörter; engl. out of vocabulary words.

[4] engl. Smoothing

Add-Delta-Glättung [Lidstone, 1920, Johnson, 1932, Jeffreys, 1948]. Die Add-Delta-Glättung addiert zur absoluten Häufigkeit aller N-Gramme einen Wert δ, wobei $0 < \delta \leq 1$ gilt. Die bedingte Wahrscheinlichkeit aus Formel 2.1 wird entsprechend erweitert [Chen und Goodman, 1996]:

$$\Pr_{AD}\left(w_k \mid w_{k-(N-1)} \ldots w_{k-1}\right) = \frac{C\left(w_{k-(N-1)} \ldots w_{k-1} w_k\right) + \delta}{\sum_{w \in V} C\left(w_{k-(N-1)} \ldots w_{k-1} w\right) + \delta|V|} \ .$$

Für Unigramm-Modelle gilt demnach:

$$\Pr_{AD}\left(w_k\right) = \frac{C\left(w_k\right) + \delta}{\sum_{w \in V} C\left(w\right) + \delta|V|} \ .$$

Der Delta-Wert kann mit Hilfe des Expectation–Maximization-Algorithmus [Dempster u.a., 1977] auf einem Held-Out-Korpus oder durch Kreuzvalidierung auf dem Trainingskorpus bestimmt werden [Jurafsky und Martin, 2008].

Ein Nachteil der Add-Delta-Glättung ist, dass diese zuviel Wahrscheinlichkeitsmasse an ungesehene Wörter vergibt. Für eine Anwendung, die die Rechtschreibung überprüft, bedeutet dies, dass falsch geschriebene Wörter gegenüber korrekt geschriebenen Wörtern bevorzugt werden [Gale und Church, 1990]. [Church und Gale, 1991] zeigen, dass 46.50% der gesamten Wahrscheinlichkeitsmasse eines Bigramm-Trainingskorpus auf ungesehenen Bigramme verteilt werden. Hierbei ist $\delta = 0.000137$; die Trainingskorpusgröße beträgt 22 Millionen Bigrammen und die ungesehene Bigramme verlaufen sich auf 74.6 Milliarden. [Jurafsky und Martin, 2008] untersuchen die Add-Delta-Glättung, $\delta = 1$, auf einem Trainingskorpus mit $|V| = 1\,446$. Ein ungesehenes Bigramm, z.B. „food want“ bekommt eine Wahrscheinlichkeit von 0.00039, welches einer Häufigkeit von 0.43 entspricht. Dafür reduzieren sich die Wahrscheinlichkeiten der vorkommenden Bigramme stark. Zum Beispiel verändert sich die Wahrscheinlichkeit des Bigramms „chinese food“, $\Pr\left(\text{food} \mid \text{chinese}\right)$, um das Zehnfache – von 0.52 auf 0.052 nach der Glättung. Die entsprechende Häufigkeit $C\left(\text{chinese food}\right)$ reduziert sich dabei ebenfalls um das Zehnfach – von 82 auf 8.2.

Jelinek-Mercer-Glättung [Jelinek und Mercer, 1980, Brown u.a., 1992]. Bei der Jelinek-Mercer-Glättung (JM-Glättung) wird ein N-Gramm-Modell höherer Ordnung mit Modellen niedrigerer Ordnung linear interpoliert. Die Interpolation erfolgt dabei rekursiv [Chen und Goodman, 1996]:

$$\Pr_{JM}\left(w_k \mid w_{k-(N-1)}^{k-1}\right) = \lambda_i \Pr_{ML}\left(w_k \mid w_{k-[(N-1)-(i-1)]}^{k-1}\right) + (1 - \lambda_i) \Pr_{JM}\left(w_k \mid w_{k-[(N-1)-i]}^{k-1}\right),$$

wobei $1 \leq i \leq N$ gilt und $w_{k-(N-1)}^{k-1}$ als Abkürzung für $w_{k-(N-1)} \ldots w_{k-1}$ steht. Die Rekursion kann beendet werden, indem entweder die ML-Schätzung für

ein Unigramm-Modell, also

$$\Pr_{JM}(w_k \mid w_{k-1}) = \lambda_{N-1}\Pr_{ML}(w_k \mid w_{k-1}) + (1-\lambda_{N-1})\Pr_{ML}(w_k),$$

benutzt wird oder die Gleichverteilung für das N-Gramm-Modell nullter Ordnung eingesetzt wird:

$$\Pr_{JM}(w_k) = \lambda_N \Pr_{ML}(w_k) + (1-\lambda_N)\frac{1}{|V|}.$$

Die Gewichte λ_i können mit Hilfe des Expectation–Maximization-Algorithmus [Dempster u.a., 1977] oder des Baum-Welch-Algorithmus [Baum, 1972] auf einem Held-Out-Korpus ermittelt werden. Hierbei muss $\sum_i \lambda_i = 1$ gelten.

Die JM-Glättung in Verbindung mit Unigramm-Modellen wird häufig im Information Retrieval eingesetzt [Miller u.a., 1999, Lavrenko und Croft, 2001]. Dabei wird die Produktionswahrscheinlichkeit eines Unigramms w durch das Sprachmodell eines Dokuments M_d mit der Produktionswahrscheinlichkeit von w durch das Sprachmodell des gesamten Trainingskorpus[5] M_C interpoliert:

$$\Pr_{JM}(w \mid M_d) = \lambda \Pr_{ML}(w \mid M_d) + (1-\lambda)\Pr_{ML}(w \mid M_C). \tag{2.2}$$

Mittels linearer Interpolation kann einem häufigen Wort eine höhere Wahrscheinlichkeit zugewiesen werden, als es vor der Glättung hatte. In einem Beispiel von [Zhai, 2008] hat das Unigramm „for“ vor der Glättung eine Wahrscheinlichkeit von 0.025 in einem Text und 0.01 in einem anderen, so verändern sich diese Wahrscheinlichkeiten nach der Glättung zu 0.1825 im ersten Text und 0.181 im zweiten. Nachteilig ist, dass dadurch die ursprüngliche Differenz der Wahrscheinlichkeiten des Wortes aufgehoben wird. In der Folge wird die Produktionswahrscheinlichkeit eines Textes durch Beiträge von häufigen Wörtern dominiert. Des Weiteren ist die JM-Glättung von der Größe des Trainingskorpus abhängig: Für kleine Trainingsmengen (< 2.3 Millionen Wörter) werden bessere Ergebnisse für Bigramm- und Trigramm-Modelle erzielt im Vergleich zu den untersuchten Glättungstechniken; für größere Trainingsmengen (≥ 2.3 Millionen Wörter) ist die JM-Glättung für Bigramm- und Trigramm-Modelle ungeeignet [Chen und Goodman, 1996].

2.3 Distanz zwischen N-Gramm-Modellen

Die Distanz zwischen zwei N-Gramm-Modellen M und M' wird durch die Kullback-Leibler-Divergenz (KL-Divergenz) angegeben. Die KL-Divergenz

[5] Wird auch als Hintergrund-Sprachmodell und das Trainingskorpus als Hintergrundkorpus bezeichnet.

$KL(M \mid\mid M')$ ist ein Unähnlichkeitsmaß für zwei Wahrscheinlichkeitsverteilungen über dem gleichen Ereignisraum: In unserem Fall über das Vokabular beider Verteilungen V. Es wird gemessen, wie gut die Verteilung M' die Verteilung M approximiert. Die KL-Divergenz ist nichtnegativ und $KL(M \mid\mid M') = 0$ gdw. $M = M'$. Je größer die KL-Divergenz ist, desto unterschiedlicher sind beide N-Gramm-Modelle. Eine obere Schranke besitzt die KL-Divergenz nicht. Im Fall von Unigramm-Modellen wird die KL-Divergenz wie folgt berechnet:

$$KL(M \mid\mid M') = \sum_{w \in V} \Pr(w \mid M) \log \frac{\Pr(w \mid M)}{\Pr(w \mid M')} \ .$$

Die KL-Divergenz ist keine Metrik, da sie weder symmetrisch ist noch die Dreiecksungleichung[6] erfüllt. Trotzdem wird sie oft als Distanz zwischen zwei Wahrscheinlichkeitsverteilungen betrachtet. Vereinbart wird, dass $0 \log \frac{0}{\Pr(w|M')} = 0$ (unabhänging vom Wert $\Pr(w \mid M')$) und $\Pr(w \mid M) \log \frac{\Pr(w|M)}{0} = \infty$ gilt. Wenn ein Wort $w \in V$ existiert, so dass $\Pr(w \mid M) > 0$ und $\Pr(w \mid M') = 0$, dann ist $KL(M \mid\mid M') = \infty$ [Cover und Thomas, 1991].

Die KL-Divergenz wird häufig als Rankingfunktion in Anwendungen des Information Retrieval eingesetzt. [Lavrenko und Croft, 2001] stellen ein relevanzbasiertes Sprachmodell auf Basis einer Anfrage und gegebenen relevanten Dokumenten vor. Diese Dokumente stammen aus dem Ranking eines Retrieval-Modells oder durch Bewertungen von Benutzern. Die Anfrage und die dazugehörigen relevanten Dokumente werden dabei als zufällige Stichproben des relevanzbasierten Sprachmodells betrachtet. Die Berechnung dieses Modells erfolgt durch die Wahrscheinlichkeit des Auftretens der Wörter aus den relevanten Dokumenten zusammen mit den Anfragewörtern. Um weitere Dokumente auf Relevanz zur Anfrage zu untersuchen, wird die KL-Divergenz zwischen dem relevanzbasierten Sprachmodell und den Sprachmodellen der abzuschätzenden Dokumente bestimmt. [Xu und Croft, 1999] gruppieren Dokumente zur Bildung von thematischen Clustern. Um die Themenzugehörigkeit eines unbekannten Textes zu ermittelt, wird die KL-Divergenz zwischen dem Sprachmodell dieses Textes und den Sprachmodellen der thematischen Cluster gebildet.

[6] Für beliebige drei Punkte x,y und z im metrischen Raum (X, d) müsste gelten:
$d(x, y) \leq d(x, z) + d(z, y), d : X \times X \to \mathbb{R}$.

Kapitel 3

Verwandte Arbeiten

Bei der Datierung von Textdokumenten ist der Zeitraum, in dem ein gegebenes Textdokument verfasst wurde, zu bestimmen. Die bereits publizierten Verfahren zur Datierung eines Textdokuments verwenden Unigramm-Modelle, Heuristiken und Klassifikationsmethoden, welche aus einem Trainingskorpus bestimmt werden. Abbildung 3.1 stellt den entsprechenden Datierungsprozess dar.

Zur Evaluierung von Verfahren zur Textdatierung wird ein Testkorpus verwendet. Dieses enthält, ebenfalls wie das Trainingskorpus, Textdokumente mit bekannten Daten. Dabei müssen die Textdokumente beider Korpora aus demselben Genre stammen und nicht mehrfach vorkommen. Jedes Testdokument wird durch dasselbe Verfahren datiert. Das automatisch zugewiesene Datum wird anschließend mit dem tatsächlichen Datum verglichen. Als Bewertungsmaß dient häufig die Genauigkeit, welche den Anteil der richtig datierten Texte unter allen datierten Texten angibt.

In diesem Kapitel werden verwandte Arbeiten nach folgendem Schema vorgestellt. Zuerst wird das verwendete Trainingskorpus betrachtet. Anschließend

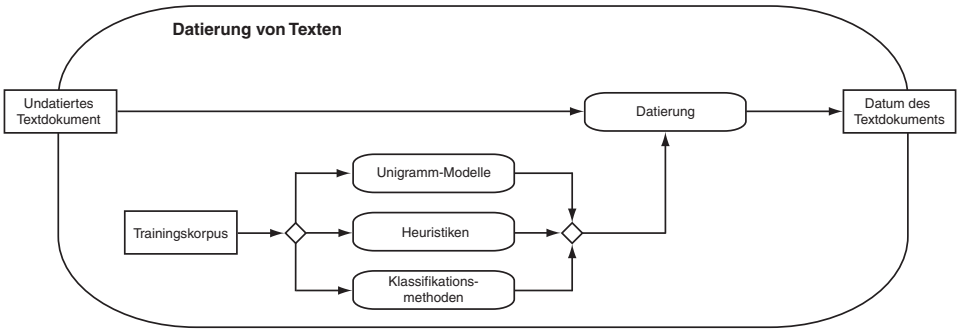


Abbildung 3.1: Bestimmung des Datums eines undatierten Textdokuments.

wird die jeweilige Dokumentvorverarbeitung besprochen, bevor Verfahren zur Datierung vorgestellt werden. Abschließend werden die erzielten Ergebnisse zusammen mit dem verwendeten Testkorpus vorgestellt und kritisch analysiert.

3.1 Datierung durch Unigramm-Modelle

Im Folgenden werden zwei Forschungsarbeiten vorgestellt, die sich mit Verfahren zur Datierung auf Basis von Unigramm-Modellen befassen.

3.1.1 Datierung nach [de Jong u.a., 2005]

Die Datierung von Textdokumenten auf Basis von Unigramm-Modellen wird erstmals in [de Jong u.a., 2005] eingesetzt.

Trainingskorpus. Das Trainingskorpus von [de Jong u.a., 2005] basiert auf einem Auszug des Twente-Korpus und enthält Zeitungsartikel aus den niederländischen Zeitungen *De Volkskrant* und *Algemeen Dagblad* von Januar 1999 bis Februar 2005 [Ordelman, 2003]. Die Dokumente im Trainingskorpus werden in konsekutive Zeiträume von vier Tagen, sieben Tagen, einem Monat und drei Monaten eingeteilt.

Unbekannt bleibt, wie viele Texte pro Zeitraum im Trainingskorpus vorhanden sind und welche Textlänge die Dokumente aufweisen, zumal das Trainingskorpus auch nicht frei verfügbar ist.

Dokumentvorverarbeitung. Alle Wörter mit einer Häufigkeit kleiner als zehn werden entfernt und Stemming für das Niederländische wird durchgeführt. Die Vokabulargröße wird dadurch von 1.3 Millionen Wörter auf 170 000 Wörter reduziert.

Verfahren zur Datierung. Für ein zu datierendes Textdokument d und das Trainingskorpus werden Unigramm-Modelle M_d bzw. M' mit Jelinek-Mercer-Glättung erstellt. Der Vergleich zwischen diesen Unigramm-Modellen erfolgt durch das normalisierte Log-Likelihood-Ratio-Maß (NLLR[1]). NLLR normalisiert M_d und M' durch das Unigramm-Modell M_C des Hintergrundkorpus. Das Hintergrundkorpus enthält dabei die Dokumente aus allen Zeiträumen. NLLR berechnet sich nach folgender Formel [Kraaij, 2004]:

$$NLLR\left(M_d, M'\right) = \sum_{w \in d} \Pr\left(w \mid M_d\right) \log\left(\frac{\Pr\left(w \mid M'\right)}{\Pr\left(w \mid M_C\right)}\right).$$

[1] engl. Normalized Log Likelihood Ratio

Zur Datierung werden von [de Jong u.a., 2005] zwei Methoden vorgestellt, die sich in der Art und Weise unterscheiden, wie die Unigramm-Modelle M' aus dem Trainingskorpus gebildet werden:

Vergleich auf Dokumentebene. Für alle Dokumente im Trainingskorpus werden Unigramm-Modelle M' mit Jelinek-Mercer-Glättung erstellt. Alle Trainingsdokumente werden nach dem NLLR-Maß $NLLR(M_d, M')$ sortiert. Der Zeitraum des Dokuments, der an Rang 1 der Sortierung steht oder der Zeitraum, dessen Dokumente den größten NLLR-Wert unter den Top-10 Rängen aufweisen, bestimmt das Datum des zu datierenden Textdokuments.

Vergleich auf Zeitraumebene. Je nach Einteilung des Trainingskorpus in Zeiträume vom vier Tagen, sieben Tagen, einem Monat und drei Monaten wird für jeweils alle Dokumente eines solchen Zeitraums ein Unigramm-Modell M' mit Jelinek-Mercer-Glättung erstellt. Der Zeitraum mit größtem NLLR-Wert bestimmt das Datum des zu datierenden Textdokuments.

Ergebnisse. Die Granularität des ausgegebenen Datums ist von der Granularität des eingesetzten Unigramm-Modells zu unterscheiden. Die Granularität des Datums ist abhängig von externen Faktoren, etwa ob das Publikationsjahr, die -woche oder der -tag angegeben werden soll. Beispielsweise könnten Unigramm-Modelle auf Wochenbasis erstellt werden, während die Anwendung das Publikationsjahr eines Zeitungsartikels ausgibt. Die Granularität des berechneten Datums in den folgenden Auswertungen von [de Jong u.a., 2005] beträgt ein Vierteljahr.

Das Testkorpus von [de Jong u.a., 2005], welches zur Evaluierung verwendet wird, enthält 500 Artikel aus den niederländischen Zeitungen *Trouw*, *Het Parool* und *NRC Handelsblad*.

Der Vergleich auf Dokumentebene ist dem Vergleich auf Zeitraumebene überlegen. Die besten Ergebnisse werden hierbei mit einer Genauigkeit von ca. 56% für den Zeitraum an Rang 1 des NLLR-Rankings und 47% für den Zeitraum mit größtem NLLR-Wert unter den Top-10 Rängen erzielt. Das liegt daran, dass Artikel, die über das gleiche Ereignis berichten, häufiger vorkommen und ihre Wortwahl ähnlicher ist. Somit ist die Ähnlichkeit zwischen den Unigramm-Modellen eines zu datierenden Dokuments und eines Trainingsdokuments größer als die zwischen den Unigramm-Modellen eines Dokuments und eines themen-ungebundenen Zeitraums.

Kritik. Die Autoren geben die Größe des Trainingskorpus nicht an, dadurch lässt sich nicht das Verhältnis zum Testkorpus abschätzen. Weiterhin ist das

Trainingskorpus auf Nachfrage nicht verfügbar, welches ein Nachvollziehen der Ergebisse unmöglich macht. Die Textlänge der betrachteten Dokumente beider Korpora wird ebenfalls nicht erwähnt. Das ist jedoch wichtig, um eine Abschätzung der minimal benötigten Länge für die zu datierenden Textdokumente abgeben zu können.

Obwohl der Vergleich auf Dokumentebene bessere Ergebnisse erzielt, ist diese Methode bei großem Trainingskorpus sehr rechenintensiv. Bei der Evaluierung wird eine Genauigkeit für Vierteljahre angegeben, offen dabei bleibt, ob das automatisch bestimmte Datum als korrektes Datum gilt, falls es sich innerhalb des Vierteljahres befindet, welches ab dem tatsächlichen Datum beginnt.

3.1.2 Datierung nach [Kanhabua und Nørvåg, 2008]

Aufbauend auf [de Jong u.a., 2005] verbessern [Kanhabua und Nørvåg, 2008] die Datierung von Dokumenten durch neue, angepasste NLLR-Maße und Vorverarbeitungsschritte.

Trainingskorpus. Das Trainingskorpus enthält Dokumente aus dem Internet Archive[2]. Bei den Dokumenten handelt es sich um zu verschiedenen Zeitpunkten aufgenommenen Versionen[3] von Webseiten, darunter größtenteils Zeitungsartikel, wie *ABC News*, *CNN* und *New York Post*. Insgesamt ergeben sich 9 000 Dokumente aus einem Zeitraum von acht Jahren (genauere Angaben hierzu fehlen). Das Trainingskorpus wird in konsekutive Zeiträume von einer Woche, einem Monat, drei Monaten, sechs Monaten und einem Jahr aufgeteilt.

Auch hier bleibt unbekannt, wie viele Texte pro Zeitraum im Trainingskorpus vorhanden sind und welche Textlänge die Dokumente aufweisen.

Dokumentvorverarbeitung. Die Dokumente werden folgenden Vorverarbeitungtechniken unterzogen:

- *POS-Tagging*[4]*:*
  Hierbei findet eine Zuweisung von Wortarten (wie Nomen, Verben und Adjektive) zu Wörtern, statt.

- *Extraktion von Kollokationen:*
  Kollokationen sind häufig gemeinsam auftretende Wörter wie zum Beispiel „Information Retrieval“ oder „Lineare Algebra“.

[2]http://www.archive.org/ (Letzter Zugriff: 21.12.2011.)
[3]Versionsgeschichten
[4]engl. Part of Speech

- *Auflösung sprachlicher Mehrdeutigkeiten*[5]*:*
  Hierbei werden mehrdeutige Begriffe und Ausdrücke aufgelöst. Zum Beispiel verweist der Begriff „Bank“ situationsbedingt entweder auf ein Geldinstitut oder eine Sitzgelegenheit.

- *Extraktion von Konzepten:*
  Ein Konzept ist ein Begriff oder eine Kombination von Wörtern. Beispiele hierfür sind Schlüsselwörter wie „Periodensystem“ oder „Geisteswissenschaft“.

- *Wortfilterung:*
  Wörter werden durch das *tf·idf*-Maß gewichtet, wobei nur die topgerankten Terme für die weitere Verarbeitung benutzt werden. Ab welchem Rang Wörter verwendet werden, wird von [Kanhabua und Nørvåg, 2008] nicht genannt.

Die Algorithmen, die zur Dokumentvorverarbeitung eingesetzt werden, sind nicht angegeben.

Verfahren zur Datierung. Unigramm-Modelle mit Jelinek-Mercer-Glättung werden für jedes zu datierende Textdokument und jeweils jeden Zeitraum des Trainingskorpus (eine Woche, einen Monat, drei Monate, sechs Monate, ein Jahr) erstellt, wobei ein Vergleich der Unigramm-Modelle auf Zeitraumebene stattfindet (wie dies bei [de Jong u.a., 2005] der Fall ist).

Des Weiteren werden Funktionen zum NLLR-Maß hinzugefügt, so dass zwei neue NLLR-Formeln entstehen. Zum einen wird jedes Wort durch seine Entropie[6] gewichtet und als Multiplikator benutzt, zum anderen werden Statistiken aus Google Zeitgeist[7] für jedes Wort benutzt und als additiver Faktor in die NLLR-Formel eingesetzt.

Ergebnisse. Das Testkorpus enthält, ebenfalls wie das Trainingskorpus, Dokumente aus dem Internet Archive. Hierbei handelt es sich um 1 000 Dokumente, die in die Zeitspanne des Trainingskorpus fallen. Das Testkorpus und das Trainingskorpus sind hierbei disjunkt.

Das Verfahren „Vergleich auf Zeitraumebene“ von [de Jong u.a., 2005] dient als Baseline. Allerdings nennen die Autoren nicht, ob zusätzlich zu den oben genannten Vorverarbeitungtechniken auch die Vorverarbeitung von [de Jong u.a., 2005] eingesetzt wird.

[5] engl. Word Sense Disambiguation

[6] Misst den Informationsgehalt eines Wortes: Je größer die Entropie eines Wortes, desto besser kann es ein Dokument von anderen trennen.

[7] http://www.google.com/intl/en/press/zeitgeist/index.html (Letzter Zugriff: 21.12.2011.)

Gezeigt wird, dass die Genauigkeit der modifizierten NLLR-Maße zusammen mit den Vorverarbeitungtechniken die Baseline übertreffen: Die Vorverarbeitungtechniken zusammen mit der entropiebasierten NLLR-Formel führen zur höchsten Genauigkeit von 71.56%; im Vergleich dazu beträgt die Genauigkeit der Baseline 53.62%.

Kritik. Die Autoren geben nicht an, welche Algorithmen zur Dokumentvorverarbeitung eingesetzt wurden. Diese können die Qualität der Unigramm-Modelle maßgeblich beeinflussen: Werden zum Beispiel Adverbien als Konzepte bestimmt, so ist ein Unigramm-Modell, welches darauf erstellt wurde, unbrauchbar, denn es gibt keine Konzepte, die nur durch Adverbien beschrieben werden. Weiterhin fehlen Angaben über die Anzahl der verwendeten POS-Taggs, Kollokationen, aufgelöster sprachlicher Mehrdeutigkeiten, Konzepten und Wörter bei der Erstellung der Unigramm-Modelle. Eine Einschätzung der minimal benötigten Länge für die zu datierenden Textdokumente ist somit nicht möglich.

Als Baseline wird das Verfahren von [de Jong u.a., 2005] „Vergleich auf Zeitraumebene“ herangezogen. Ob die Vorverarbeitung von [de Jong u.a., 2005] dabei mitberücksichtigt wurde, bleibt unbeantwortet. Die Baseline berechnet die Genauigkeit des ausgegebenen Datums für Vierteljahre. Allerdings verwenden [Kanhabua und Nørvåg, 2008] unter anderem Unigramm-Modelle für sechs Monate und ein Jahr, wobei hier nicht genannt wird, wie die Auswertung für Vierteljahre vorgenommen wird: Wir nehmen an, dass eine Angleichung der Genauigkeit des Datums an der Auflösung aller Unigramm-Modelle vorgenommen wurde, da ansonsten keine einheitliche Berechnung und Bewertung der Datierungsgenauigkeit gewährleistet ist.

3.2 Datierung durch Heuristiken und Klassifikationsmethoden

Im Gegensatz zur Datierung mit Unigramm-Modellen setzen [Garcia-Fernandez u.a., 2011] Heuristiken und Klassifikationsmethoden in Verbindung mit externem Wissen ein.

Trainingskorpus. Als Trainingskorpus dienen französische Zeitungsartikel aus sieben Zeitungen, die zwischen 1801 und 1944 veröffentlicht wurden. Diese Artikel sind in 2 396 Abschnitte zu je 300 und 500 Wörtern eingeteilt.

Dokumentvorverarbeitung. Es wird POS-Tagging und Stemming durchgeführt. Das Vokabular reduziert sich dadurch von 74 000 auf 52 000 Wörter.

Verfahren zur Datierung. Es werden sowohl chronologische Funktionen als auch Klassifikationsmethoden unterschieden, um das Jahr zu bestimmen, aus dem ein Textdokument stammt.

Chronologische Funktionen. Es werden drei chronologische Funktionen eingesetzt, die auf Heuristiken basieren. Dabei handelt es sich um Hinweise, die aus dem Inhalt eines zu datierenden Textdokuments stammen. Dazu gehören Personennamen, Neologismen[8], Archaismen[9] und Wörter, die vor den französischen Rechtschreibreformen von 1835 und 1878 verwendet werden. Die Funktionen bestimmen für jedes mögliche Jahr, das als Datum des zu datierenden Textdokuments in Frage kommt, einen Funktionswert, welcher die „Plausibilität“ des Datums angibt. Bei den ersten beiden chronologischen Funktionen wird der Funktionswert als Wahrscheinlichkeit angegeben. Die Funktionswerte der chronologischen Funktionen werden dabei durch Multiplikation und lineare Regression miteinander kombiniert.

- *Funktion auf Basis von Personennamen:*
  Mithilfe von Personennamen lässt sich schließen, ob ein zu datierendes Textdokument nach der Geburt der im Textdokument genannten Personen geschrieben wurde. Die Geburtstage der Personen werden mit Hilfe der französischen Wikipedia ermittelt. Abhängig von der Häufigkeit des Personennamens im Dokument wird jedem Jahr wie folgt eine Wahrscheinlichkeit zugewiesen: Ein Jahr, das vor dem Geburtstag einer Person liegt, bekommt eine Wahrscheinlichkeit von 0.3; für Jahre, die zwischen dem Geburtstag und 20 Jahre danach liegen, steigt die Wahrscheinlichkeit linear an, bis sie Eins erreicht.

- *Funktion auf Basis von Neologismen und Archaismen:*
  Ist das ungefähre Jahr des ersten Auftretens eines Wortes (Neologismus) bekannt, so wird eine niedrigere Wahrscheinlichkeit allen vorausgehenden Jahren und eine höhere Wahrscheinlichkeit allen nachfolgenden Jahren zugewiesen. Umgekehrt verhält es sich bei Archaismen.
  Zur Bestimmung der Neologismen und Archaismen wird die Häufigkeit der Wörter auf Grundlage der Google-Books-N-Gramme für die französische Sprache betrachtet: Liegt die Worthäufigkeit über gewissen Schwellwerten, so wird das Wort als Neologismus bzw. Archaismus betrachtet.

- *Funktion auf Basis von Rechtschreibreformen:*
  Zwischen 1801 und 1944 (der Zeitraum der betrachteten Texte) wurden in Frankreich zwei Rechtschreibreformen durchgeführt (1835 und 1878), die die Endung von Verben und des Plurals veränderten. Es werden Wörter

[8] ling. neugebildete Wörter
[9] ling. veraltete Wörter

aus dem zu datierenden Textdokument bestimmt, die aus der Zeit vor den beiden Rechtschreibreformen stammen. Auf Basis der Häufigkeit dieser Wörter wird jedem möglichen Datum ein Funktionswert zwischen 0 und 1 zugewiesen.

Klassifikationsmethoden. Die Kosinusähnlichkeit und eine One-vs-All-Support-Vector-Machine (SVM) werden als Klassifikationsmethoden für die Jahre 1801 bis 1944 eingesetzt. Hierbei werden die Abschnitte des Trainingskorpus in Jahre gruppiert, so dass für jedes Jahr ein Merkmalsvektor erstellt wird. Für ein zu datierendes Dokument wird ebenfalls ein Merkmalsvektor gebildet, welcher durch die Klassifikationsmethoden mit dem Merkmalsvektor eines Jahres in Beziehung gesetzt wird.

- *Kosinusähnlichkeit:*
  Die Kosinusähnlichkeit bestimmt als Datum eines Textdokuments das Jahr, dessen Merkmalsvektor am ähnlichsten zum Merkmalsvektor des Textdokuments ist. Beide Merkmalsvektoren werden für Buchstaben-1-bis-5-Gramme erstellt. Hierbei werden die Merkmalsvektoren zum einen mittels einer *tf·idf*-Gewichtung der Wörter berechnet, zum anderen mittels einer veränderten *tf·idf*-Gewichtung durch die Häufigkeiten der Wörter der Google-Books-N-Gramme. Im Fall der Google-Books-N-Gramme werden die Wörter nicht in ihrer Grundform verwendet. Zudem werden nur Wörter berücksichtigt, die mindestens zehn Mal im Jahr auftreten.

- *One-vs-All-SVM:*
  Für jedes Jahr wird eine SVM bestimmt, welche einen Text entweder zu diesem Jahr zuordnet oder zu den restlichen Jahren. Dadurch entstehen insgesamt 144 SVMs. Als Baseline-Merkmalsvektoren dienen Wort-Bigramme und Wortgrundformen. Zusätzlich zu diesen Merkmalsvektoren werden Geburtstage, Neologismen und Informationen zur Rechtschreibreform einzeln als Merkmale hinzugenommen.

Ergebnisse. Das Testkorpus von [Garcia-Fernandez u.a., 2011] enthält je 2 445 Abschnitte mit 300 und 500 Wörtern.

Als Auswertungsmetrik dient der Prozentsatz der korrekt datierten Jahre und Jahrzehnte. Für einen Abschnitt d zieht die Formel die Distanz vom vorhergesagten Jahr $\tau_p(d)$ zum tatsächlichen $\tau_r(d)$ Jahr in Betracht. Hierbei basiert die Auswertungsmetrik auf einer Gauß-Funktion und wird über die Anzahl der Testabschnitte K gemittelt:

$$S = \frac{1}{K} \sum_{i=1}^{K} e^{-\frac{\pi}{100}(\tau_p(d_i)-\tau_r(d_i))^2}.$$

Die chronologischen Funktionen erzielen bei der Kombination mittels linearer Regression bessere Ergebnisse als bei der Kombination mittels Multiplikation: Bei der Jahresbestimmung für Abschnitte mit 300 und 500 Wörtern ergibt sich eine Genauigkeit von 10% bzw. 14%, bei der Bestimmung des Jahrzehntes sind es 37% bzw. 42% korrekt datierter Dokumente.

Die besten Ergebnisse bei der Kosinusähnlichkeit werden für Buchstaben-Fünf-Gramme erzielt: Für eine Dokumentlänge von 300 und 500 Wörter werden 31% bzw. 36% angegeben. Ob diese Werte für Jahre oder Jahrzehnte erzielt werden, wird von den Autoren nicht angegeben.

Die Klassifikation mit einer One-vs-All-SVM zeigt, dass die Geburtstage als Merkmalsvektoren einen höchsten Wert von 24% im Vergleich zu den anderen Merkmalen erzielen.

Kritik. Bei der Auswertungsmetrik wird das vorhergesagte Jahr mit dem tatsächlichen Jahr verrechnet; wie die Berechnungsvorschrift für Jahrzehnte lautet, wird hingegen nicht angegeben. Wir vermuten, dass in diesem Fall die Genauigkeit für Jahrzehnte unabhängig von der Auswertungsmetrik bestimmt wird, indem geprüft wird, ob das ausgegebene Datum im Jahrzehnt des tatsächlichen Datums liegt. Daher können wir die erzielten Genauigkeitswerte für Jahrzehnte nicht bewerten. Weiterhin wird bei den Ergebnissen der Kosinusähnlichkeit und der One-vs-All-SVM nicht angegeben, ob es sich dabei um Genauigkeitswerte handelt oder ob dies die Ergebnisse der Auswertungsmetrik darstellen.

3.3 Zusammenfassung

Alle vorgestellten Publikationen verwenden zur Datierung Zeitungsartikel. Zeitungsartikel geben das aktuelle Tagesgeschehen wieder, so dass Unigramm-Modelle nicht nur auf Tagesbasis, sondern auch für unterschiedliche Zeiträume gebildet werden können. Hierbei wird die zeitliche Auflösung der Unigramm-Modelle von der Genauigkeit des ausgegebenen Datums unterschieden: Das Datum wird für Vierteljahre bestimmt [de Jong u.a., 2005]. Fraglich ist allerdings, ob dieser Zeitraum die beste Wahl darstellt. Wir schätzen ein, dass bei der Betrachtung von Zeitungsartikeln mit entsprechend großem Trainingskorpus (pro Monat über 10 000 Artikel) auch eine Genauigkeit für Monate sinnvoll ist. Zur Verbesserung der Datierungsgenauigkeit könnten Unigramm-Modelle verschiedener zeitlicher Auflösung miteinander linear kombiniert werden, indem jedes Modell unterschiedlich gewichtet wird. Eine weitere Möglichkeit besteht darin, das Datum durch eine Mehrheitsentscheidung von N-Gramm-Modellen unterschiedlicher Ordnung bestimmen zu lassen.

Die Datierung durch Heuristiken von [Garcia-Fernandez u.a., 2011] lässt sich ebenfalls durch weitere Heuristiken verbessern. Hier sind zeitliche Ausdrücke denkbar, die sowohl explizite Daten (1.11.2011 oder 11 Uhr), implizite Daten (heute oder letzte Woche) als auch Ereignisse (Parlamentswahlen oder Nationalfeiertag) beschreiben. Durch diese Angaben lässt sich eine Chronologie der Ereignisse im Text erstellen, so dass ein zeitlicher Ausdruck als Datum bestimmt werden könnte, der zu allen anderen Ausdrücken die geringste Entfernung aufweist. Weiterhin ließe sich das Datum eines zu datierenden Textdokuments eingrenzen, indem eine Überlappung der Zeitintervalle, die jede einzelne Heuristik bestimmt, gebildet wird.

Im nachfolgenden Kapitel 4 werden unsere Verfahren zur Datierung von Textdokumenten vorgestellt. Diese bauen auf keine der hier aufgeführten verwandten Arbeiten auf. Die einzige Gemeinsamkeit ist die Verwendung von N-Gramm-Modellen, denn wir sehen hierbei bessere Möglichkeiten, um Verfahren zur Datierung zu erstellen. Die vorgestellten Publikationen zur Datierung verwenden Dokumente aus sechs [de Jong u.a., 2005] und acht [Kanhabua und Nørvåg, 2008] Jahren zur Erstellung von N-Gramm-Modellen. Wir hingegen verwenden einen deutlich größeren Trainingskorpus, der Dokumente aus 208 Jahren umfasst, wobei pro Jahr 6 000 Bücher vorhanden sind. Dies bietet somit eine realistischere Grundlage zur Erstellung von N-Gramm-Modellen.

Kapitel 4

Unsere Verfahren zur Datierung von Textdokumenten

Ziel unserer Verfahren zur Datierung von Textdokumenten ist die Bestimmung des Zeitraums, in dem ein Text geschrieben wurde. Dem Datierungsprozess vorausgehend ist ein Vorverarbeitungsschritt, in dem zunächst die Dokumente des Trainingskorpus in konsekutive Jahre eingeteilt werden und als nächstes für diese Jahre und für ein undatiertes Textdokument N-Gramm-Modelle erstellt werden. Die anschließende Datierung des Textdokuments wird als ein Zwei-Schritt-Prozess durchgeführt: Eine Rankingfunktion stellt eine Ordnung zwischen den N-Gramm-Modellen der Jahre des Trainingskorpus in Bezug auf das N-Gramm-Modell des Textdokuments her. Ein Klassifizierer bestimmt abschließend das Jahr eines N-Gramm-Modells als Datum des Textdokuments. Dieser Prozess wird in Abbildung 4.1 dargestellt.

In diesem Kapitel wird auf die einzelnen Schritte detailliert eingegangen und das Diagramm aus Abbildung 4.1 wird entsprechend erweitert. Im nächsten

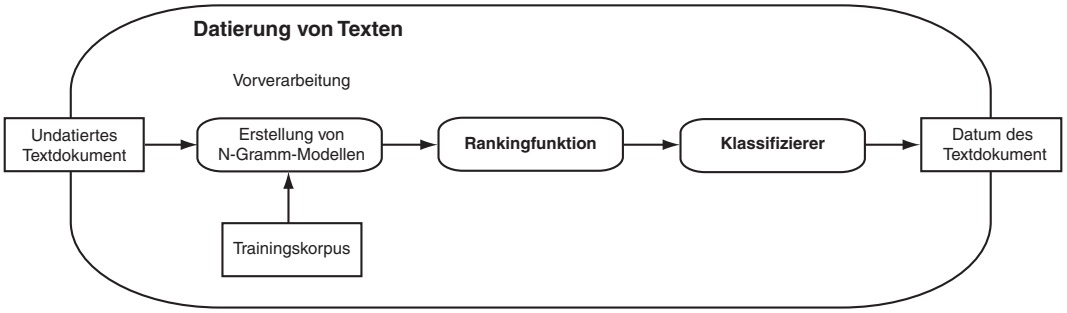


Abbildung 4.1: Die Datierung eines undatierten Textdokuments als ein Zwei-Schritt-Prozess mit Vorverarbeitung.

Abschnitt wird das Google-Books-N-Gramm-Korpus als Trainingskorpus vorgestellt. Auf Basis des Trainingskorpus werden N-Gramm-Modelle erstellt, die für die Rankingfunktionen von Bedeutung sind. Diese werden im Abschnitt 4.2 eingeführt. Im abschließenden Abschnitt 4.3 werden die verschiedenen Klassifizierer vorgestellt.

4.1 Google-Books-N-Gramm-Korpus

Als Trainingskorpus wird das Google-Books-N-Gramm-Korpus[1] verwendet. Dieses wurde 2009 erstellt und enthält an die 6 000 digitalisierte Bücher in englischer Sprache für jedes Jahr zwischen 1500 und 2008. Bücher mit niedriger OCR-Qualität wurden dabei entfernt. Die Bücher stammen aus Universitätsbibliotheken und Verlagen [Michel u.a., 2011]. Genauere Angaben zu den enthaltenen Büchern werden jedoch nicht gemacht: Die Genres der im Trainingskorpus enthaltenen Dokumente sind somit unbekannt. Dies stellt eine Herausforderung dar, denn die Genres des Trainingskorpus müssen mit den Genres der zu datierenden Dokumente übereinstimmen, um exakte N-Gramm-Modelle erstellen zu können und dadurch eine unverfälschte Datierungsgenauigkeit zu gewährleisten (vgl. Kapitel 2).

Das Trainingskorpus enthält sowohl Uni-, Bi-, Tri-, Vier- als auch Fünfgramme auf Wortbasis zusammen mit Metainformationen wie das Jahr des Auftretens und die zugehörige Häufigkeit. Diese Daten sind in insgesamt 1 510 Dateien organisiert, wobei sich N-Gramme unterschiedlicher Ordnung in getrennten Dateien befinden. Bei den N-Grammen wird zwischen Groß- und Kleinschreibung unterschieden. Ebenso sind sämtliche Satzzeichen vorhanden, welche auch als Wörter betrachtet werden. Jede Zeile einer Datei ist wie folgt aufgebaut:

```
N-Gramm<T>Jahr<T>Häufigkeit<T>Seitenanzahl<T>Bücheranzahl<CR>
```

Hierbei sind die Wörter des N-Gramms durch Leerzeichen voneinander getrennt; `<T>` ist ein Tabulator und `<CR>` ein Zeilenumbruch. Die N-Gramme einer Datei sind zuerst alphabetisch und anschließend chronologisch sortiert. Zwei Zeilen für das Trigramm „Museum of Moder sehen beispielsweise wie folgt aus:

```
Museum of Moder    1985    10    10    6
Museum of Moder    1986    16    13    6
```

[1] Das sogenannte *English One Million*, Version 20090715, welches unter folgender Adresse verfügbar ist: http://books.google.com/ngrams/datasets. (Letzter Zugriff: 21.12.2011.)

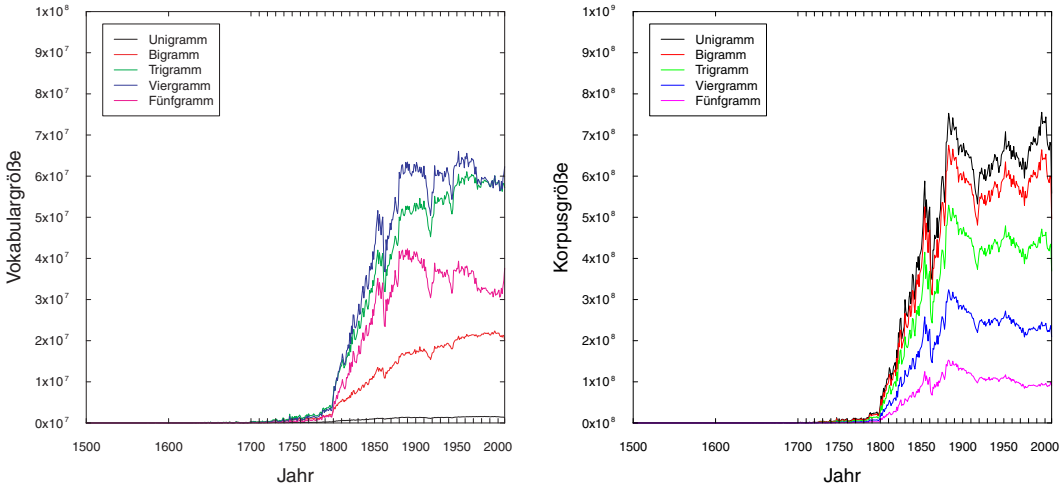


Abbildung 4.2: Vokabulargröße (links) und Korpusgröße (rechts) vom Trainingskorpus für die Jahre von 1500 bis 2008.

Die erste Zeile bedeutet, dass im Jahr 1985 das Trigramm „Museum of Moder“ 10 Mal auf 10 Seiten und in 6 Büchern vorkommt.

Zur Erstellung der N-Gramm-Modelle für die einzelnen Jahre sind nur die N-Gramme zusammen mit ihren Häufigkeiten von Interesse. Abbildung 4.2 stellt die Vokabulargröße[2] und die Korpusgröße[3] von 1500 bis 2008 dar. Da vor 1800 sowohl die Vokabulargröße als auch die Korpusgröße viel zu klein sind, werden in den später durchgeführten Experimenten nur Jahre ab 1800 betrachtet. Obwohl pro Jahr 6 000 Bücher vorhanden sein sollten, nehmen wir an, dass vor 1800 viele Bücher aufgrund OCR-Fehlern entfernt wurden.

Das Google-Books-N-Gramm-Korpus hat eine Größe von 959.4 GB. Eine Datenstruktur, die damit umgehen kann und eine konstante Zugriffszeit garantiert, ist der invertierte Index. Der von uns eingesetzte invertierte Index beinhaltet Paare bestehend aus einem Schlüssel (N-Gramm) und einer dazugehörigen sortierten Liste von Werten (Häufigkeiten von 1800 bis 2008). Für ein gegebenes N-Gramm lassen sich dadurch die zugehörigen Häufigkeiten mittels der Position des entsprechenden Jahres in der Liste erfragen. Zur Erstellung des invertierten Index aus dem Google-Books-N-Gramm-Korpus verwenden wir Komponenten der AItools-Software-Suite [Webis Group, 2011].

Bei der Erstellung des invertierten Index hat sich herausgestellt, dass das Google-Books-N-Gramm-Korpus nicht den Angaben entsprechend erstellt wurde. Dabei geht es um zwei Fälle, mit denen der invertierte Index nicht umge-

[2]Die Vokabulargröße wird durch die Anzahl der Wortformen im Korpus bestimmt.
[3]Die Korpusgröße wird durch die Anzahl der einzelnen Wörter im Korpus bestimmt.

hen kann. (1) Unter den Bi-, Tri-, Vier- und Fünfgrammen des Trainingskorpus gibt es auch N-Gramme, die nicht der angegeben Ordnung entsprechen. Beispielsweise befinden sich Tri- und Viergramme in Fünfgramm-Dateien. In den meisten Fällen handelt es sich dabei um N-Gramme, die Satzzeichen enthalten, welche wiederum nicht korrekt durch Leerzeichen von den Wörtern getrennt worden sind. Dies betrifft 0.059% (583.3 MB) der gesamten Korpusgröße. Während der Erstellung des invertierten Index wurden diese falsch einsortierten N-Gramme ignoriert. (2) Es gibt gleiche N-Gramme, die zwar in der „richtigen“ N-Gramm-Datei sind, allerdings für dieselben Jahre eine unterschiedliche Häufigkeit aufweisen. Zudem kommen diese Fälle in den Dateien nicht nacheinander, sondern sind verstreut. Daher nehmen wir als Häufigkeit eines N-Gramms für ein Jahr das erste Auftreten in einer Datei.

Die Erstellung des invertierten Index ist ein zeit- und rechenintensiver Prozess: Die Dauer umfasst drei Wochen bei voller Auslastung einer Maschine mit 8 GB Arbeitsspeicher und 4 Prozessoren mit 2.83 GHz Taktfrequenz. Der erstellte invertierte Index hat eine Größe von 248.1 GB.

4.2 Rankingfunktionen

Als Vorverarbeitungsschritt werden N-Gramm-Modelle M_t für jedes der 208 Jahre erstellt, $1800 \leq t \leq 2008$. Die N-Gramm-Modelle M_t werden dadurch erstellt, indem die Wahrscheinlichkeiten für jedes Wort im Vokabular aus diesem Jahr berechnet werden. Dies erfolgt durch Bestimmung des N-Gramm-Modells und der zugehörigen Glättungstechnik.

Für einen zu datierenden Text werden die N-Gramm-Modelle der möglichen Jahre, die für das Datum in Frage kommen, geordnet. Dabei erfolgt das Ranking einerseits nach der Produktionswahrscheinlichkeit dieses Textes durch die N-Gramm-Modelle der verschiedenen Jahre, andererseits wird der Vergleich durch die Kullback-Leibler-Divergenz zwischen den N-Gramm-Modellen des zu datierenden Textes und den N-Gramm-Modellen der Jahre herangezogen. Beide Rankingfunktionen werden nachfolgend vorgestellt.

4.2.1 Produktionswahrscheinlichkeit

Wir nehmen an, dass ein Jahr t, in dem Texte verfasst wurden, durch die Menge aller in diesem Jahr geschriebenen Wörter repräsentiert wird. Deckt dieses Vokabular viele Wörter aus einem zu datierenden Text d ab, so ist es wahrscheinlich, dass d aus diesem Jahr stammt. Das N-Gramm-Modell M_t dieses Jahres erzeugt d mit der Produktionswahrscheinlichkeit $Pr(d \mid M_t)$. Die Produktionswahrscheinlichkeit gibt an, wie wahrscheinlich d als zufällige Stichpro-

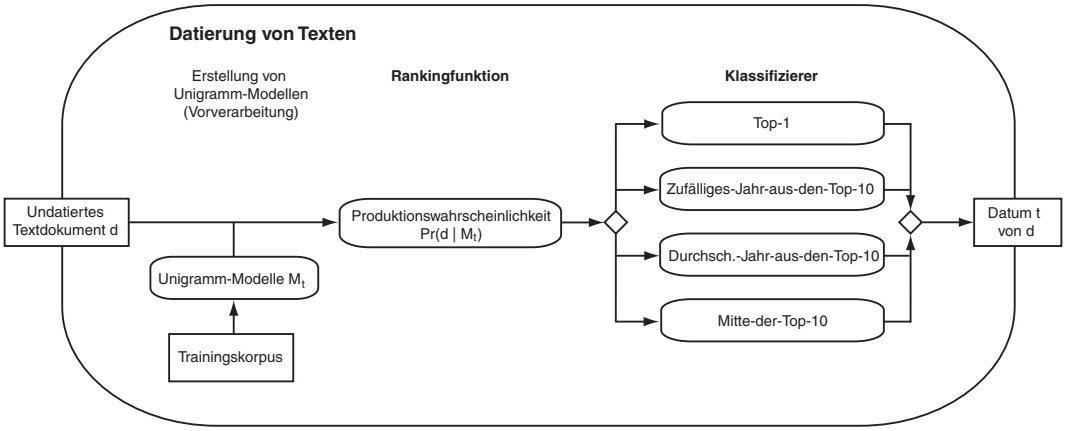


Abbildung 4.3: Die Datierung eines undatierten Textdokuments als ein Zwei-Schritt-Prozess mittels der Produktionswahrscheinlichkeit als Rankingfunktion.

be des entsprechenden Jahres beobachtet wird. Das Jahr, dessen Produktionswahrscheinlichkeit am größten ist, erzeugt das Textdokument auch am wahrscheinlichsten. Für d erfolgt das Ranking der Jahre nach folgenden Schritten (vgl. Abbildung 4.3):

1. Schätzung der Produktionswahrscheinlichkeiten $\Pr(d \mid M_t)$. Für d wird die Produktionswahrscheinlichkeit für jedes N-Gramm-Modell M_t berechnet.

2. Ranking nach absteigenden Wahrscheinlichkeiten $\Pr(d \mid M_t)$.

4.2.2 Kullback-Leibler-Divergenz

Für jedes zu datierende Textdokument d und jedes Jahr t des Trainingskorpus wird ein N-Gramm-Modell M_d bzw. M_t erstellt. Zu beachten ist, dass M_d anhand der Häufigkeiten im Text selbst erstellt wird; für M_t hingegen werden die entsprechenden Häufigkeiten im Trainingskorpus berücksichtigt. Zum Vergleich zwischen M_d und M_t wird die Kullback-Leibler-Divergenz (KL-Divergenz) verwendet. Das Jahr t, welches die geringste KL-Divergenz zum Textdokument aufweist, ist dem Textdokument auch am ähnlichsten. Für d erfolgt das Ranking der Jahre nach folgenden Schritten (vgl. Abbildung 4.4):

1. Berechnung der KL-Divergenzen $\mathrm{KL}(M_d \mid\mid M_t)$.

2. Ranking nach aufsteigenden $\mathrm{KL}(M_d \mid\mid M_t)$-Werten.

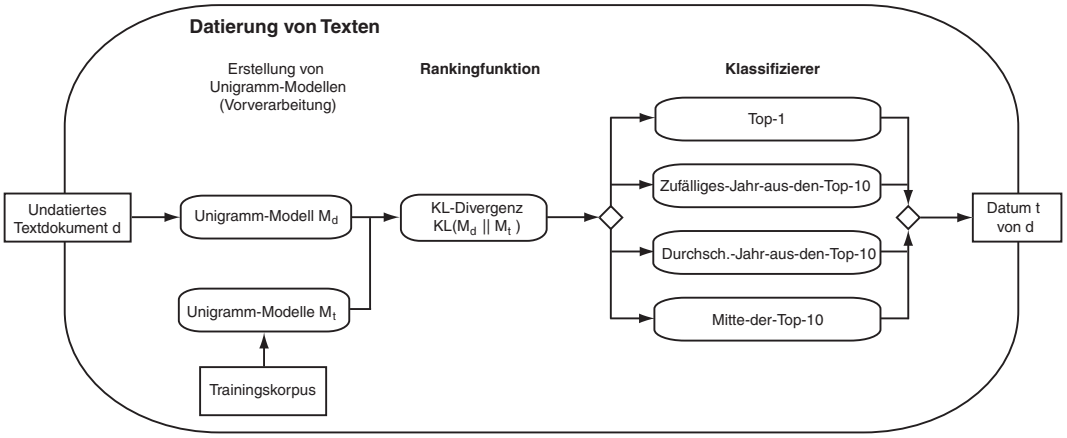


Abbildung 4.4: Die Datierung eines undatierten Textdokuments als ein Zwei-Schritt-Prozess mittels der KL-Divergenz als Rankingfunktion.

4.3 Klassifizierer

Die Berechnung der Produktionswahrscheinlichkeiten und KL-Divergenzen erzeugt ein Ranking der Jahre für ein zu datierendes Textdokument d. An erster Position der beiden Rankings sollte sich das tatsächliche Jahr von d befindet, da das N-Gramm-Modell dieses Jahres mit höchster Produktionswahrscheinlichkeit d erzeugt bzw. die kleinste Distanz zum N-Gramm-Modell von d aufweist. Steht das tatsächliche Jahr nicht an erster Position, sollte sich dieses unter den Top-10 befinden. Weiterhin sollten sich unter den Top-10 auch Jahre befinden, die nahe am tatsächlichen Jahr liegen. Auf Basis dieser Erwartungen werden verschiedene Klassifizierer eingesetzt, die das Datum von d bestimmen. Abbildung 4.5 stellt ein Beispiel zur Datierung eines Textdokuments dar. Für jeden Klassifizierer ist das automatisch bestimmte Datum angegeben. Folgende alternative Klassifizierer werden betrachtet:

Top-1

Das Datum wird durch das Jahr an erster Position des Rankings festgelegt. Bei dem Ranking anhand der Produktionswahrscheinlichkeit bestimmt das Jahr mit dem größten Wert das Datum: $\underset{t}{\operatorname{argmax}} \Pr(d \mid M_t)$. Im Gegensatz dazu wird als das Datum das Jahr mit dem kleinsten Wert beim KL-Divergenz-Ranking festgelegt: $\underset{t}{\operatorname{argmin}} \operatorname{KL}(M_d \mid\mid M_t)$.

Bei der Auswahl des Datums durch den Top-1-Klassifizierer nehmen wir an, dass das Jahr am Rang 1 beider Rankings jeweils dem tatsächlichen Jahr des Textdokuments entspricht.

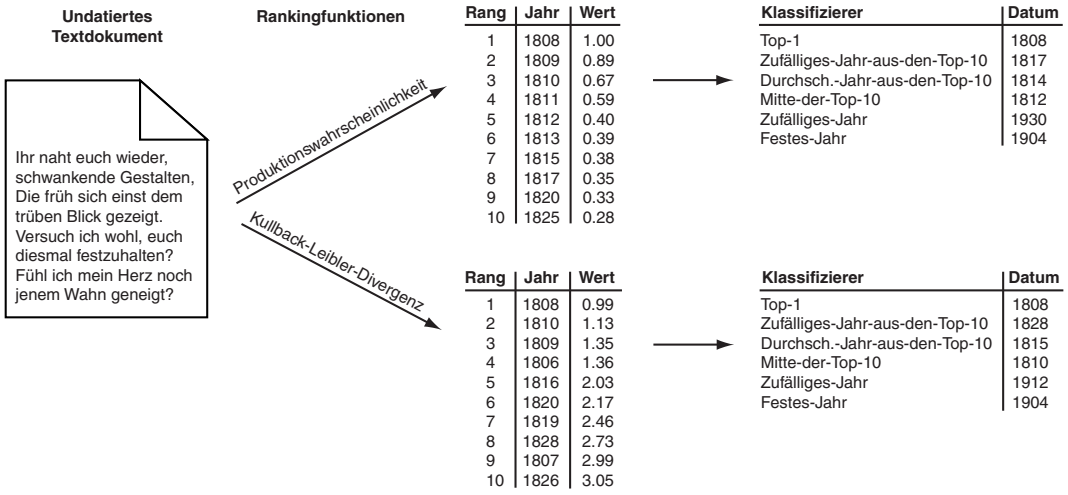


Abbildung 4.5: Beispiel zur Datierung eines Textdokuments, welches ursprünglich 1808 geschrieben wurde.

Zufälliges-Jahr-aus-den-Top-10

Ein Jahr, das an den Positionen 1 bis 10 im Ranking steht, wird zufällig als Datum von d ausgewählt.

Wir nehmen an, dass dieser Klassifizierer das tatsächliche oder das Jahr am nächsten hierzu bestimmen kann.

Durchschnittliches-Jahr-aus-den-Top-10

Aus den Jahren an den Rängen 1 bis 10 wird der Durchschnitt gebildet, welcher kaufmännisch gerundet wird: Liegt die Zahl an der ersten wegfallenden Dezimalstelle zwischen 5 und 9, dann wird aufgerundet, ansonsten wird abgerundet (DIN 1333 beschreibt diese Rundungsregel). Das gerundete Jahr bestimmt das Datum von d.

Durch die Durchschnittsbildung der Top-10 Jahren sollte dieser Klassifizierer das Jahr finden, welches sich am nächsten zum tatsächlichen Jahr befindet.

Mitte-der-Top-10

Aus den Rankings werden jeweils die Top-10 Jahre absteigend sortiert. Anschließend wird das Jahr an Position 5 als Datum ausgewählt.

Dadurch, dass wir Jahre mit großen Abweichungen zur Mitte ausschließen, werden Ausreißer vermieden. Wir nehmen an, dass das Jahr in der Mitte am geringsten vom tatsächlichen Jahr abweicht.

Unabhängig von den Rankings werden zwei weitere Klassifizierer als Baseline verwendet:

Zufälliges-Jahr

Ein zufälliges Jahr im betrachteten Zeitraum von 1800 bis 2008 wird als Datum festgelegt.

Festes-Jahr

Das Jahr 1904, also die Mitte zwischen 1800 und 2008, bestimmt das Datum von d. Wir nehmen an, dass das Jahr 1904 im Schnitt den geringsten Abstand zum tatsächlichen Jahr aufweist.

Kapitel 5

Evaluierung

In diesem Kapitel werden die vorgestellten Verfahren zur Datierung von Dokumenten experimentell untersucht. Zunächst werden Korpora vorgestellt, die zur Evaluierung dieser Verfahren erforderlich sind. Im Anschluss werden Bewertungsmaße zur Evaluierung der Datierung eingeführt, bevor auf den allgemeinen Experimentaufbau eingegangen wird. Abschließend werden die durchgeführten Experimente diskutiert.

5.1 Korpora

Für die Evaluierung der vorgestellten Verfahren wurden drei Testkorpora für Belletristik, Zeitungsartikel und künstlich erzeugte Texte erstellt. Jedes Testkorpus enthält Texte aus dem Zeitraum zwischen 1800 und 2008, wobei pro Jahr drei Texte zur Verfügung stehen. Die Texte werden in Längen von 100, 500, 1 000, 5 000, 10 000 und 100 000 Wörtern eingeteilt.

Projekt-Gutenberg-Testkorpus

Zur Erstellung des ersten Testkorpus werden 879 englische Büchern der belletristischen Literatur verwendet. Diese stammen aus dem Projekt Gutenberg[1], welches frei verfügbare Bücher zur Verfügung stellt. Die Bücher stammen aus dem Zeitraum zwischen 1806 und 1970. Eine Verteilung der Bücher zeigt Tabelle 5.1. Aus den 879 Büchern werden für den Aufbau des Testkorpus für jedes Jahr je drei Bücher zufällig ausgewählt. Aus den so bestimmten Büchern werden zusammenhängende Abschnitte mit 100, 500, 1 000, 5 000, 10 000 und 100 000 Wörtern zufällig gewählt. Falls ein Jahr nur durch ein Buch vertreten ist, werden mehrere Abschnitte davon zufällig ausgewählt. Auf diese Weise er-

[1] http://gutenberg.net.au/ (Letzter Zugriff: 21.12.2011.)

Jahr	Anzahl Bücher	Durchsch. Textlänge	Jahr	Anzahl Bücher	Durchsch. Textlänge	Jahr	Anzahl Bücher	Durchsch. Textlänge	Jahr	Anzahl Bücher	Durchsch. Textlänge
1806	1	67 234	1851	3	134 687	1887	4	113 797	1923	18	100 050
1808	1	98 111	1852	4	136 631	1888	2	77 555	1924	20	111 020
1809	1	106 726	1853	3	200 353	1889	6	86 520	1925	20	104 458
1810	1	81 925	1854	4	119 843	1890	4	126 290	1926	25	96 825
1813	1	124 153	1855	2	102 466	1891	5	111 754	1927	28	115 389
1814	1	163 285	1856	1	118 364	1892	5	100 406	1928	23	108 083
1816	2	111 728	1857	3	214 985	1893	3	104 609	1929	22	113 983
1817	3	98 891	1858	1	224 979	1894	6	81 907	1930	19	119 006
1819	3	155 198	1859	5	132 184	1895	8	109 550	1931	20	99 525
1820	1	187 249	1860	5	187 358	1896	6	82 467	1932	22	99 479
1821	2	172 260	1861	1	76 276	1897	3	11 3493	1933	22	92 751
1823	3	132 702	1862	2	297 081	1898	6	63 296	1934	9	101 836
1824	1	158 466	1863	2	162 964	1899	5	69 836	1935	16	103 805
1825	1	131 805	1864	3	178 551	1900	11	93 886	1936	11	96 038
1826	3	174 381	1865	4	194 948	1901	9	82 751	1937	8	87 289
1828	3	132 265	1866	1	309 641	1902	12	78 041	1938	7	130 696
1829	1	149 116	1867	3	176 280	1903	10	102 268	1939	6	116 171
1830	1	110 731	1868	2	198 589	1904	5	121 958	1940	12	91 670
1831	4	137 340	1869	7	283 587	1905	10	81 743	1941	7	98 262
1833	2	123 531	1870	6	146 987	1906	13	73 339	1942	3	143 173
1834	1	25 933	1871	5	156 567	1907	13	90 171	1943	3	101 910
1835	1	106 807	1872	1	167 534	1908	20	84 524	1944	7	91 865
1836	1	313 615	1873	2	206 739	1909	16	93 617	1945	3	124 765
1837	3	157 224	1874	2	149 118	1910	12	85 632	1946	2	75 421
1838	1	49 907	1875	4	208 047	1911	11	91 220	1947	5	94 272
1839	1	208 028	1876	5	184 588	1912	7	73 404	1948	2	94 011
1840	2	141 792	1877	2	213 328	1913	11	11 3751	1949	4	115 682
1841	1	218 848	1878	3	98 834	1914	10	97 197	1950	3	74 817
1843	4	170 558	1879	2	61 583	1915	10	120 609	1951	2	104 615
1844	4	217 139	1880	8	175 669	1916	7	78 445	1952	5	109 167
1845	3	159 530	1881	3	118 619	1917	5	120 425	1953	2	114 744
1846	2	85 473	1882	5	106 777	1918	6	90 834	1954	2	97 036
1847	6	158 979	1883	2	123 456	1919	19	96 267	1956	2	72 411
1848	6	145 176	1884	2	127 915	1920	14	101 623	1965	1	65 615
1849	2	243 990	1885	2	110 532	1921	8	103 628	1970	1	10 0908
1850	4	100 515	1886	6	101 938	1922	21	99 658			

Tabelle 5.1: Verteilung der Bücher aus dem Projekt Gutenberg pro Jahr mit ihrer durchschnittlichen Textlänge laut kaufmännischer Rundung.

geben sich 429 Bücher für jede Textlänge, außer für 10 000 und 100 000: Hierfür gibt es 428 bzw. 301 Bücher. Insgesamt enthält dieses Korpus 2 572 Texte.

Robust04-Testkorpus

Dieses Testkorpus enthält Artikel aus der Zeitung *Los Angeles Times* der Jahre 1989 und 1990, sowie Artikel der *Financial Times* aus dem Zeitraum zwischen 1991 und 1994. Diese Zeitungsartikel stammen aus dem Testkorpus des TREC 2004 Robust Track[2]. Tabelle 5.2 zeigt eine Zusammenfassung der vorhandenen Zeitungsartikel. Daraus werden Abschnitte der Länge 100, 500, 1 000, 5 000 und 10 000 zufällig ausgewählt. Für jedes Jahr gibt es pro Textlänge drei Zeitungsartikel, außer für Textlänge 5 000 und 10 000: Hier gibt es keine Zeitungsartikel aus 1991 bzw. es gibt nur einen Zeitungsartikel aus 1990 und zwei aus 1989. Das Testkorpus enthält insgesamt 81 Abschnitte.

[2]http://trec.nist.gov/data/robust/04.guidelines.html (Letzter Zugriff: 21.12.2011.)

Jahr	Anzahl Artikel	Durchsch. Textlänge
1989	23 015	975
1990	24 954	967
1991	1 387	884
1992	16 327	872
1993	16 095	852
1994	17 002	846

Tabelle 5.2: Verteilung der Zeitungsartikel der *Los Angeles Times* und *Financial Times* pro Jahr mit ihrer durchschnittlichen Textlänge laut kaufmännischer Rundung.

Künstliches-Testkorpus

Dieses Korpus besteht aus Texten, deren Wörter nach den Verteilungen des Trainingskorpus für die Jahre 1800 bis 2008 gezogen wurden. Um Texte nach der Verteilung eines Jahres zu erzeugen, wird ein „Roulettekessel“ eingesetzt. Hierbei entsprechen die Kammern des Roulettekessels den Häufigkeiten der Wörter aus einem Jahr. Eine Kugel wird solange geworfen, bis eine Textlänge von 100, 500, 1 000, 5 000, 10 000 und 100 000 Wörtern erreicht wird. Wenn die Kugel in eine Häufigkeiten fällt, wird das entsprechende Wort zum Abschnitt hinzugefügt. Falls mehrere Wörter in Frage kommen, wird eins zufällig gewählt. Auf diese Weise werden pro Jahr drei Abschnitte erzeugt: Dies entspricht also 627 Abschnitte pro Textlänge. Das Testkorpus besteht somit aus insgesamt 3 762 Abschnitten.

5.2 Bewertungsmaße

Ziel der Verfahren zur Datierung ist es, das Jahr zu ermitteln, aus dem ein Textdokument stammt. Zur Bewertung der Datierungsgenauigkeit werden folgende Maße eingesetzt.

Durchschnittliche absolute Entfernung vom tatsächlichen Jahr

Der Betrag der Entfernung des tatsächlichen Jahres zum vorhergesagten Jahr wird berechnet. Für alle zu datierenden Textdokumente werden die Entfernungen aufsummiert und anschließend gemittelt.

Durchschnittliche Entfernung vom tatsächlichen Jahr

Die Entfernung vom tatsächlichen Jahr zum vorhergesagten Jahr wird bestimmt. Eine negative Distanz bedeutet, dass die Vorhersage zeitlich weiter in der Zukunft liegt als das tatsächliche Jahr. Eine positive Distanz sagt aus, dass das vorhergesagte Jahr zeitlich vor dem tatsächlichen Jahr liegt. Die Summe

der Entfernungen aller Textdokumente wird zur Bildung des Durchschnitts verwendet.

Standardabweichung

Die Standardabweichungen der durchschnittlichen absoluten Entfernungen und der durchschnittlichen Entfernungen wird berechnet.

Genauigkeit für korrekt datierte Zeiträume

Ermittelt wird die Genauigkeit für korrekt datierte Zeiträume bezüglich des

Jahres: Geprüft wird eine exakte Übereinstimmung des vorhergesagten mit dem tatsächlichen Jahr.

Jahrzehnts: Geprüft wird, ob das vorhergesagte Jahr im Jahrzehnt des tatsächlichen Jahres liegt. Hierbei werden als Jahrzehnt fünf Jahre vor dem tatsächlichen Jahr und vier Jahre danach aufgefasst.

Jahrhunderts: Hierbei wird bestimmt, ob das vorhergesagte Jahr im Jahrhundert des tatsächlichen Jahres liegt. Als Jahrhundert werden 50 Jahre vor dem tatsächlichen Jahr und 49 Jahre danach betrachtet.

Die Geltungsbereiche zur Berechnung der Genauigkeit für korrekt datierte Zeiträume fasst Abbildung 5.1 zusammen. Die Genauigkeit für Jahrhunderte ist immer größer gleich der Genauigkeit für Jahrzehnte, die wiederum größer gleich der Genauigkeit für Jahre ist.

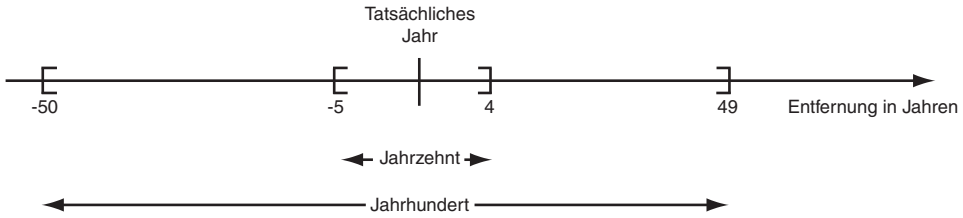


Abbildung 5.1: Geltungsbereiche zur Berechnung der Genauigkeit für korrekt datierte Jahre, Jahrzehnte und Jahrhunderte.

Ranking-Statistiken

Zusätzlich zu den genannten Maßen werden Statistiken für das tatsächliche Jahr in den Rankings der Produktionswahrscheinlichkeit und der KL-Divergenz angegeben. Durch diese Angaben wird überprüft, auf welchem Rang sich das tatsächliche Jahr befindet. Hierfür werden folgende Maße verwendet:

Durchschnittlicher Rang: Die Ränge aller tatsächlichen Jahre der Textdokumente werden summiert und anschließend wird der Durchschnitt gebildet.

Standardabweichung des durchschnittlichen Rangs.

Mean Reciprocal Rank (MRR): MRR wird als Bewertungsmaß für Ranking-Listen verwendet, indem der Kehrwert des Rangs des korrekten Jahres gebildet wird. MRR wird dabei über die Größe der Testmenge gemittelt [Jurafsky und Martin, 2008].

5.3 Experimentaufbau

In diesem Abschnitt wird der allgemeine Aufbau der Experimente vorgestellt. Um das Datum eines undatierten Textdokuments zu bestimmen, werden zunächst zwei Rankings aufgestellt. Zum einen werden die Jahre nach aufsteigender logarithmierter Produktionswahrscheinlichkeit gerankt, zum anderen nach absteigender KL-Divergenz. Als nächstes werden die in Kapitel 4 vorgestellten Klassifizierer auf diese Rankings angewandt. Für jedes Experiment werden die folgenden Parameter eingestellt:

- Rankingfunktion,
- Klassifizierer,
- Testkorpus und
- Textlänge.

Vor der Ausführung der Experimente wird zur Erstellung der Rankings die Ordnung des N-Gramm-Modells und die eingesetzte Glättungstechnik festgelegt.

Bestimmung des N-Gramm-Modells zur Datierung

Aufgrund der Komplexität, die N-Gramm-Modelle höherer Ordnung besitzen, werden ausschließlich Unigramm-Modelle auf Wortbasis in den Experimenten verwendet. Weiterhin werden alle Buchstaben in Kleinbuchstaben umgewandelt und sämtliche Formatierungen und Satzzeichen entfernt. Die Unigramm-Modelle werden auf Jahresbasis erstellt.

Hierfür wir ein neuer invertierter Index erstellt. Dieser basiert auf den 10 Unigramm-Dateien des Google-Books-N-Gramm-Trainingskorpus. Die Größe dieser Dateien ist 4.9 GB: Das sind 0.51% der Gesamtgröße des Trainingskorpus. Bei der Erstellung des invertierten Index wurde der zuvor gebaute

Wort	Jahr	Häufigkeit	ML-Schätzer	Delta-Glättung	JM-Glättung
waymar	1800	0	0	2.3119629E-12	2.7632553E-6
	1900	0	0	1.5116777E-13	7.9378057E-7
	2000	0	0	1.3708721E-13	6.6292379E-7
folklore	1800	1	2.3119648E-8	2.3121940E-8	3.4611583E-7
	1900	454	6.8630185E-7	6.8630187E-7	9.4297982E-7
	2000	3605	4.9419949E-6	4.9419940E-6	4.7731036E-6
legends	1800	139	3.2136311E-6	3.2136307E-6	3.6721360E-6
	1900	5579	8.4336520E-6	8.4336505E-6	8.3701548E-6
	2000	4235	5.8056445E-6	5.8056435E-6	6.0049481E-6
happiness	1800	2261	5.2273524E-5	5.2273483E-5	5.2001631E-5
	1900	44116	6.6689190E-5	6.6689177E-5	6.4975730E-5
	2000	16204	2.2213616E-5	2.2213611E-5	2.4947713E-5
childish	1800	86	1.9882897E-6	1.9882904E-6	2.3702784E-6
	1900	6276	9.4872916E-6	9.4872900E-6	9.1193801E-6
	2000	1794	2.4593450E-6	2.4593446E-6	2.7942282E-6
hearts	1800	2248	5.1972969E-5	5.1972928E-5	5.0686105E-5
	1900	33084	5.0012357E-5	5.0012348E-5	4.8921554E-5
	2000	12471	1.7096149E-5	1.7096146E-5	1.9296967E-5

Tabelle 5.3: Gegenüberstellung der Häufigkeit, ML-Schätzung, Add-Delta-Glättung ($\delta = 0.0001$) und JM-Glättung ($\lambda = 0.9$) für ausgewählte Wörter aus den Jahren 1800, 1900 und 2000. Die Vokabulargröße ist für 361 892 Wörter für 1800, 1 259 794 Wörter für 1900 und 1 508 469 Wörter für 2000.

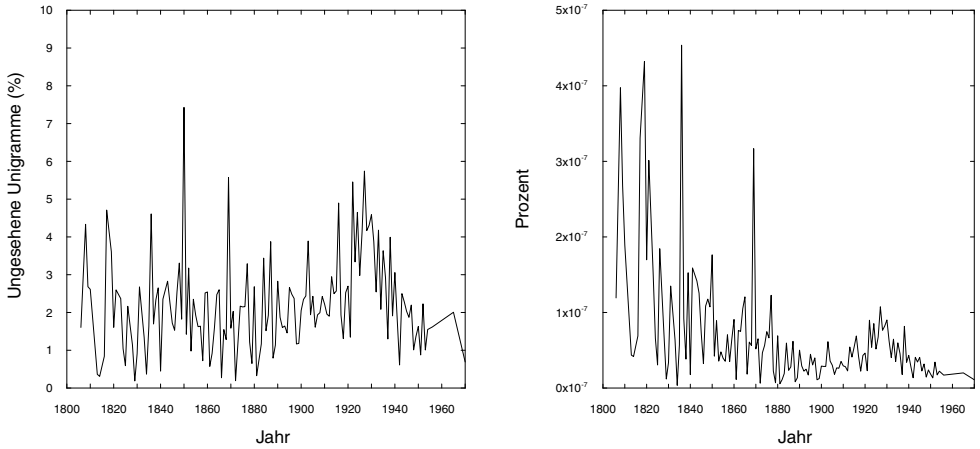
Index, siehe Abschnitt 4.1, verwendet. Der neue invertierte Index ist 9.4 GB groß.

Bestimmung der Glättungstechnik für das Unigramm-Modell

Im Folgenden werden sowohl die Add-Delta-Glättung als auch die Jelinek-Mercer-Glättung (JM-Glättung) für Unigramm-Modelle untersucht. Die Parameter beider Glättungstechniken wurden experimentell auf $\delta = 0.0001$ bzw. $\lambda = 0.9$ festgelegt.

Obwohl die Add-Delta-Glättung einen schlechten Ruf in der Fachwelt genießt [Church und Gale, 1991, Gale und Church, 1994, Jurafsky und Martin, 2008], zeigt Tabelle 5.3, dass die Wahrscheinlichkeiten, die die Add-Delta-Glättung berechnet, sich erst ab der fünften Nachkommastelle von den Wahrscheinlichkeiten des Maximum-Likelihood-Schätzers unterscheiden. Die Wahrscheinlichkeit für ungesehene Wörter ist zudem kleiner als für bekannte Wörter. Aus diesen beiden Gründen ist die Add-Delta-Glättung eine geeignete Glättungstechnik.

Untersuchungen der ungesehenen Wörter aus den Testkorpora Projekt-Gutenberg und Robust04 im Trainingskorpus zeigen, dass es wenige ungesehene Wörter gibt, denen ein geringer Prozentsatz der gesamten Wahrscheinlichkeitsmasse des Trainingskorpus zugewiesen wird. Die Anzahl der Wörter vom Projekt-Gutenberg-Testkorpus, die im Trainingskorpus fehlen, ist in Abbildung 5.2a dargestellt: Im Durchschnitt fehlen 2.2% der Wörter pro Jahr.



(a) Ungesehene Wörter des Projekt-Gutenberg-Testkorpus im Trainingskorpus.

(b) Prozent der gesamten Wahrscheinlichkeitsmasse des Trainingskorpus, die den ungesehenen Wörtern des Projekt-Gutenberg-Testkorpus zugewiesen wird.

Abbildung 5.2: Untersuchung des Projekt-Gutenberg-Testkorpus bzgl. ungesehene Unigramme.

Jahr	Ungesehene Unigramme	Wahrscheinlichkeitsmasse
1989	6.445622795	0.000000206
1990	6.540584748	0.000000213
1992	16.981031638	0.000000624
1993	12.894963241	0.000000423
1994	11.450457819	0.000000377

Tabelle 5.4: Prozent der ungesehenen Wörter des Robust04-Testkorpus im Trainingskorpus und der gesamten Wahrscheinlichkeitsmasse die diesem Testkorpus zugewiesen wird.

An diesen Wörtern werden durchschnittlich etwa $6.88 \times 10^{-8}\%$ der gesamten Wahrscheinlichkeitsmasse zugewiesen. Die Prozentwerte für alle Jahre sind in Abbildung 5.2b zu sehen. Für den Robust04-Testkorpus hingegen belaufen sich die ungesehenen Wörter auf durchschnittlich 10.86% der gesamten Größe des Robust04-Testkorpus, wie aus Tabelle 5.4 zu entnehmen ist. Hierfür werden durchschnittlich etwa $2.08 \times 10^{-7}\%$ der gesamten Wahrscheinlichkeitsmasse vergeben.

Bei der Jelinek-Mercer-Glättung (JM-Glättung) werden Unigramm-Modelle des zu datierenden Textes und eines Hintergrundkorpus linear interpoliert, siehe Formel 2.2 aus Kapitel 2. Die JM-Glättung weist im Gegensatz zur Add-Delta-Glättung eine höhere Wahrscheinlichkeit an ungesehene Wörter als an bekannte Wörter zu, wie aus Tabelle 5.3 zu entnehmen ist. Dies führt zu einer Überbewertung der ungesehenen Wörter, welches unerwünscht ist. Aus

diesem Grund wird in den weiteren Experimente die Add-Delta-Glättung mit $\delta = 0.0001$ eingesetzt.

5.4 Experimente

Die in diesem Abschnitt durchgeführten Experimente dienen der Beantwortung folgender Fragestellungen:

1. *Eignet sich der Google-Books-N-Gramm-Korpus zur Datierung von Texten?*
   Hierbei wird die Vokabularüberdeckung und Häufigkeitsverteilung der einzelnen Jahre im Trainingskorpus untersucht. Es wird erwartet, dass die Vokabulare und die Häufigkeitsverteilungen speziell genug sind, so dass die N-Gramm-Modelle für die verschiedenen Jahre voneinander zu unterscheiden sind.

2. *Sind die in dieser Arbeit vorgestellten Verfahren zur Datierung in der Lage, Textdokumente mit hoher Genauigkeit korrekt zu datieren?*
   Die Grundlage zur Beantwortung dieser Frage bildet die Experimentbeschreibung aus Abschnitt 5.3, wobei das Künstliche-Testkorpus eingesetzt wird. Es wird eine nahezu perfekte Datierung erwartet.

3. *An welchem Rang befindet sich das tatsächliche Jahr?*
   Hierbei wird auf Basis beider Rankingfunktionen untersucht, an welchem Rang sich das tatsächliche Jahr befindet. Es wird angenommen, dass sich an erster Position der beiden Rankings das tatsächliche Jahr eines zu datierenden Textes befindet, da das N-Gramm-Modell dieses Jahres mit höchster Produktionswahrscheinlichkeit den Text erzeugt bzw. die kürzeste Distanz zum N-Gramm-Modell des Textes aufweist. Steht das tatsächliche Jahr nicht an erster Position, sollte sich dieses unter den Top-10 befinden.

4. *Gibt es eine maximale Obergrenze für die Datierungsgenauigkeit?*
   Untersucht wird, welche Datierungsgenauigkeit die Klassifizierer maximal erreichen können. Hierfür wird ein „idealer“ Klassifizierer für die Top-10 verwendet, der Kenntnis über das tatsächliche Jahr und dessen Rang hat; somit kann dieser Klassifizierer das Jahr aus den Top-10 als Datum auswählen, welches sich am nächsten zum tatsächlichen Jahr befindet. Hierbei wird das Experiment aus Abschnitt 5.3 erneut für alle drei Testkorpora durchgeführt.

5. *Welche Kombination aus Rankingfunktion und Klassifizierer liefert die beste Datierung?*

Das Experiment aus Abschnitt 5.3 wird auf Basis des Projekt-Gutenberg-Testkorpus und des Robust04-Testkorpus durchgeführt.

Zudem werden die statistisch signifikanten Unterschiede in der Datierungsgenauigkeit der Klassifizierer im Vergleich zur Datierungsgenauigkeit des Baseline-Klassifizierers Zufälliges-Jahr für beide Ranking-Funktionen angegeben. Als Signifikanztest dient der Wilcoxon-Vorzeichen-Rang-Test mit einem Konfidenzintervall von 95% [Walpole und Myers, 1989]. Die Ergebnisse der durchgeführten Wilcoxon-Vorzeichen-Rang-Tests[3] gelten für alle Textlängen in den betrachteten Zeiträumen (Jahre, Jahrzehnte und Jahrhunderte).

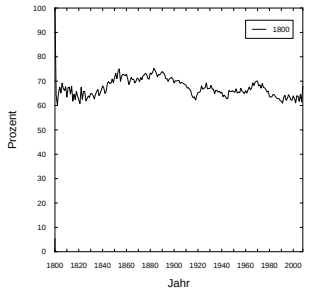
Eignet sich der Google-Books-N-Gramm-Korpus zur Datierung von Texten?

Um diese Frage zu beantworten, werden die Vokabulare und die Häufigkeitsverteilungen der einzelnen Jahre aus dem Trainingskorpus untersucht. Zum einen wird überprüft inwieweit sich die Vokabulare überlappen, zum anderen wird überprüft, ob sich diese anhand ihrer Häufigkeitsverteilungen trennen lassen.

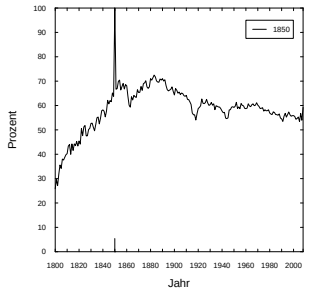
Im ersten Fall wird überprüft, wie viel Prozent des Vokabulars eines bestimmten Jahres in dem Vokabular eines anderen Jahres enthalten sind. Abbildung 5.3 stellt die Ergebnisse für sechs ausgesuchte Jahre dar; für alle anderen Jahre sehen die Kurvenverläufe ähnlich aus. Es zeigt sich, dass ein großer Teil des Vokabulars aus 1800 in allen anderen Jahren vorkommt. Je weiter die betrachteten Jahre jedoch voranschreiten, desto weniger Vokabular ist in den vorhergehenden Jahren zu beobachten. Der Prozentsatz des enthaltenen Vokabulars liegt in den nachfolgenden Jahren über 50%. Die Tendenz folgt einer Dreiecksfunktion mit dem Maximum beim betrachteten Jahr. Weiterhin ist zu sehen, dass knapp 30% des Vokabulars eines jeden Jahres in keinem anderen Jahr vorkommt. Hierbei handelt es sich in den meisten Fällen um Artefakte wie Kombinationen von englischen und nicht-englische Wörtern mit Zahlen und Sonderzeichen.

Um zu bestimmen, ob die Vokabulare voneinander zu trennen sind, wird die KL-Divergenz zwischen dem Unigramm-Modell eines jeden Jahres zu den Unigramm-Modellen aller Jahre berechnet. Abbildung 5.4 stellt die Ergebnisse ebenfalls für sechs ausgewählte Jahre dar; für alle anderen Jahre sehen die Kurvenverläufe ähnlich aus. Die KL-Divergenz zu vorhergehenden und nachfolgenden Jahren hat eine steigende Tendenz: Die Form folgt einer umgekehrten Dreiecksfunktion mit dem Minimum beim betrachteten Jahr. Die Vokabulare sind somit speziell genug und voneinander trennbar.

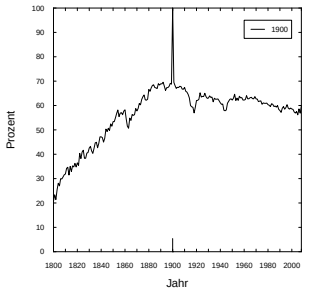
[3]Hierfür wurde die Statistik-Software SPSS 20.0 verwendet.



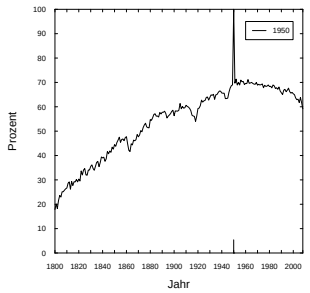
(a) Vokabularüberdeckung bzgl. des Jahres 1800.



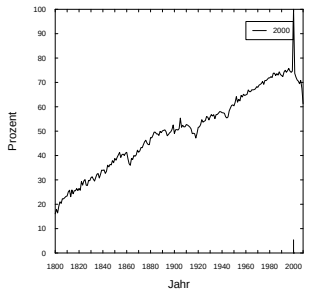
(b) Vokabularüberdeckung bzgl. des Jahres 1850.



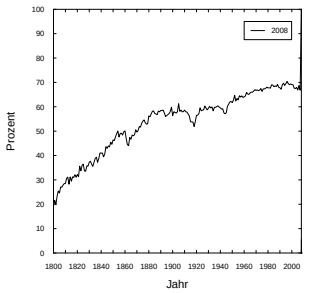
(c) Vokabularüberdeckung bzgl. des Jahres 1900.



(d) Vokabularüberdeckung bzgl. des Jahres 1950.



(e) Vokabularüberdeckung bzgl. des Jahres 2000.



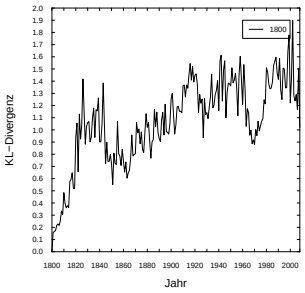
(f) Vokabularüberdeckung bzgl. des Jahres 2008.

Abbildung 5.3: Prozentsatz der Schnittmenge des Vokabulars eines ausgewählten Jahres mit allen anderen Jahren.

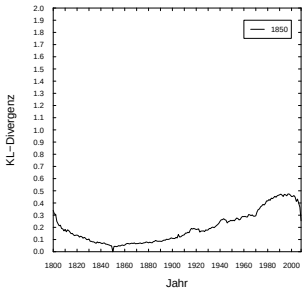
Sind die in dieser Arbeit vorgestellten Verfahren zur Datierung in der Lage, Textdokumente mit hoher Genauigkeit korrekt zu datieren?

Beispielhaft für drei künstliche Texte aus den Jahren 1811, 1880 und 1920 zeigt Abbildung 5.5 ihre Produktionswahrscheinlichkeit und KL-Divergenz, abgetragen über den Zeitraum zwischen 1800 und 2008. Es ist zu erkennen, dass die KL-Divergenz bei dem tatsächlichen Jahr den kleinsten Wert aufweist, welches für alle untersuchten Textlängen zutrifft. Somit ist das Jahr an Rang 1 des KL-Divergenz-Rankings korrekt. Der Top-1-Klassifizierer datiert in diesem Fall perfekt. Bei der Produktionswahrscheinlichkeit ist dies jedoch nicht der Fall.

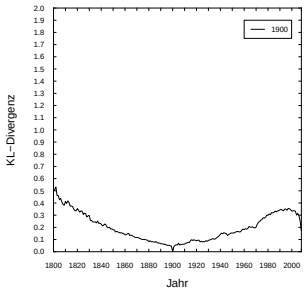
Die Experimentergebnisse für das Künstliche-Testkorpus bzgl. beider Rankingfunktionen zeigen, dass der Top-1-Klassifizierer mit Abstand die höchste Genauigkeit erreicht (vgl. Tabelle A.1 von Seite 53 und Tabelle 5.7 von Seite 46). Der Unterschied in der Genauigkeit des Top-1-Klassifizierers zur Base-



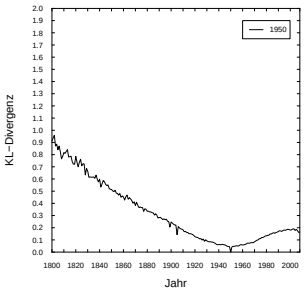
(5.4.1) $\mathrm{KL}\left(M_{1800} \mid\mid M_{t_i}\right)$.



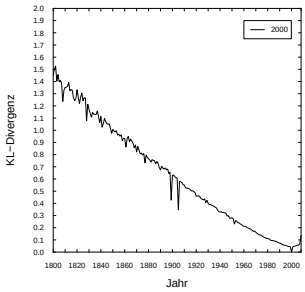
(5.4.2) $\mathrm{KL}\left(M_{1850} \mid\mid M_{t_i}\right)$.



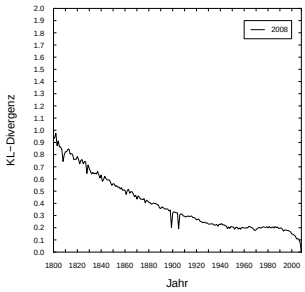
(5.4.3) $\mathrm{KL}\left(M_{1900} \mid\mid M_{t_i}\right)$.



(5.4.4) $\mathrm{KL}\left(M_{1950} \mid\mid M_{t_i}\right)$.



(5.4.5) $\mathrm{KL}\left(M_{2000} \mid\mid M_{t_i}\right)$.



(5.4.6) $\mathrm{KL}\left(M_{2008} \mid\mid M_{t_i}\right)$.

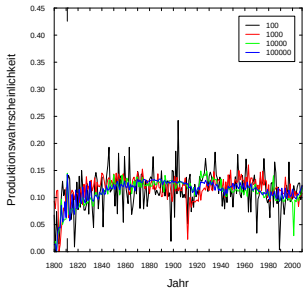
Abbildung 5.4: Gegenüberstellung der Kullback-Leibler-Divergenz $\mathrm{KL}\left(M_t \mid\mid M_{t_i}\right)$ für Unigramm-Modelle M_t von ausgewählten Jahren des Trainingskorpus zu den Unigramm-Modellen M_{t_i} des Trainingskorpus, $1800 \leq t_i \leq 2008$.

line für Jahre, Jahrzehnte und Jahrhunderte ist zudem statistisch signifikant.

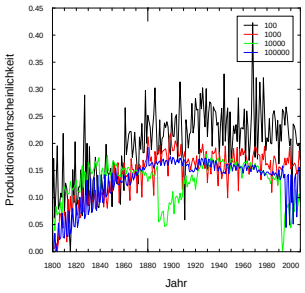
Bei dem Ranking anhand der Produktionswahrscheinlichkeit liegt die Genauigkeit des Top-1-Klassifizierers für Jahre, Jahrzehnte und Jahrhunderte über 70% ab einer Textlänge von 1 000 Wörtern und erreicht die höchste Genauigkeit mit 95.37% korrekt datierter Jahre, Jahrzehnte und Jahrhunderte bei einer Textlänge von 5 000 Wörtern.

Bei dem KL-Divergenz-Ranking liegt die Genauigkeit des Top-1-Klassifizierers für korrekt datierte Jahre, Jahrzehnte und Jahrhunderte bei einer Textlänge von 100 Wörtern bei 76.08%, 76.80% bzw. 79.90%. Ab 500 Wörter ist die Genauigkeit perfekt: Für alle datierten Zeiträume beträgt diese 100%.

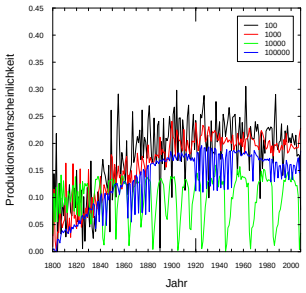
Unter den Klassifizierern, die auf den Top-10-Rankings basieren, gilt für beide Rankingfunktionen, dass der Klassifizierer Zufälliges-Jahr-aus-den-Top-10 die höchste Genauigkeit für Jahre und Jahrzehnte liefert: Beim KL-Divergenz-Ranking beträgt die höchste Genauigkeit für Jahre und Jahrzehnte 10.69% für 100 000 Wörter bzw. 32.06% für 10 000 Wörter. Ebenfalls für beide



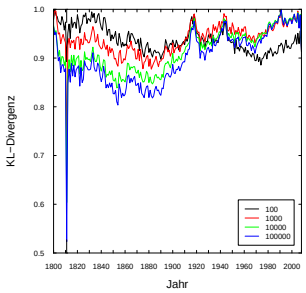
(5.5.1) Künstliches Dokument (1811)



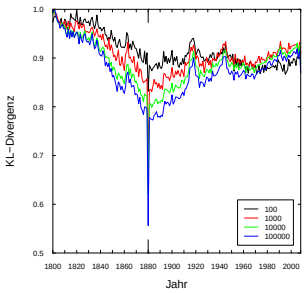
(5.5.2) Künstliches Dokument (1880)



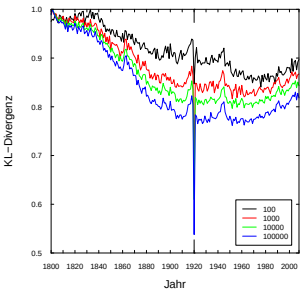
(5.5.3) Künstliches Dokument (1920)



(5.5.4) Künstliches Dokument (1811)



(5.5.5) Künstliches Dokument (1880)



(5.5.6) Künstliches Dokument (1920)

Abbildung 5.5: Produktionswahrscheinlichkeit (oben) und Kullback-Leibler-Divergenz (KL-Divergenz) (unten) für Texte unterschiedlicher Textlänge aus dem Künstlichen-Testkorpus.

Rankingfunktionen trifft zu, dass die höchste Genauigkeit für Jahrhunderte durch den Klassifizierer Durchschnittliches-Jahr-aus-den-Top-10 erzielt wird: Die Genauigkeit beträgt hierbei 96.65% für 10 000 Wörter.

Die Genauigkeiten der Baseline-Klassifizierer Zufälliges-Jahr und Festes-Jahr für Jahre, Jahrzehnte und Jahrhunderte werden von allen Klassifizierern übertroffen. Eine Ausnahme hierbei bildet der Klassifizierer Durchschnittliches-Jahr-aus-den-Top-10 für korrekt datierte Jahre für die Textlängen 100 und 500 beim Produktionswahrscheinlichkeits-Ranking und für 100 Wörter beim KL-Divergenz-Ranking.

Fazit. Es konnte gezeigt werden, dass die in dieser Arbeit vorgestellten Verfahren zur Datierung in der Lage sind, Textdokumente zu 100% korrekt zu datieren: Der Top-1-Klassifizierer kann erfolgreich zur Datierung eingesetzt werden. Dies gilt allerdings nur, falls die zu datierenden Textdokumente aus der Verteilung des Trainingskorpus stammen. Im Vergleich zum Top-1-Klassifizierer

sind die anderen Klassifizierer nur für eine Bestimmung des Jahrhunderts geeignet. Des Weiteren ist zu sehen, dass mit wachsender Textlänge weder die Genauigkeit steigt, noch die durchschnittliche Entfernung zum tatsächlichen Jahr reduziert wird. Entgegen unserer Erwartung gibt es keinen linearen Zusammenhang zwischen diesen Größen. Dies könnte daran liegen, dass manche Texte Wörter enthalten, die über mehrere Jahre hinweg eine ähnliche Häufigkeit aufweisen, so dass beispielsweise die Produktionswahrscheinlichkeit für diese Texte ähnliche Werte liefert, die wiederum für das tatsächliche Jahr nicht charakteristisch sind. Möglichkeiten dies zu unterbinden, bestehen in der Entfernung von Stoppwörtern, dem Einsatz von Stemming oder auch dem Filtern von Wörtern, die weniger als zehn Mal vorkommen. Hierbei ist allerdings ein Informationsverlust in Kauf zu nehmen.

An welchem Rang befindet sich das tatsächliche Jahr?

Von besonderem Interesse ist die Position des tatsächlichen Jahres in den Rankings der beiden Rankingfunktionen. Wir erwarten, dass sich das tatsächliche Jahr an erster Position des jeweiligen Rankings befindet. Trifft dies nicht zu, sollte das tatsächliche Jahr zumindest in der Top-10 liegen. Dadurch lässt sich untersuchen, inwieweit die Top-10-Klassifizierer zur Datierung geeignet sind.

Aus Tabelle 5.5 ist zu entnehmen, dass für beide Rankingfunktionen die Ränge der tatsächlichen Jahre von Büchern aus dem Projekt-Gutenberg-Testkorpus und der Zeitungsartikel aus dem Robust04-Testkorpus weit außerhalb der Top-10 liegen. Die Datierung sowohl mit dem Top-1-Klassifizierer als auch mit den Top-10-Klassifizierern erweist sich wider Erwarten als ungeeignet: Eine hohe Datierungsgenauigkeit für Jahre ist nicht zu erwarten.

Im Gegensatz dazu befinden sich die tatsächlichen Jahre für Texte aus dem Künstlichen-Testkorpus für die Textlängen 1 000, 5 000 und 10 000 in dem Top-10-Ranking der Produktionswahrscheinlichkeit. Für die Textlänge 5 000 ist sogar der durchschnittliche Rang 1.28 bei einem MRR von 0.97. Bei der KL-Divergenz hingegen ist der Top-Treffer für nahezu alle Textlängen für eine korrekte Datierung ausreichend. Damit lässt sich sowohl die perfekte Datierungsgenauigkeit des Top-1-Klassifizierer als auch die mangelhafte Datierungsgenauigkeit der Top-10-Klassifizierer für das Künstliche-Testkorpus erklären.

Entgegen unserer Erwartung verbessert sich der durchschnittliche Rang des tatsächlichen Jahres nicht mit steigender Textlänge. Das trifft für beide Rankingfunktionen auf allen Testkorpora zu. Eine Ausnahme bildet die KL-Divergenz bei dem Künstlichen-Testkorpus: Hier ist der Rang des tatsächlichen Jahres ab einer Textlänge von 500 Wörtern immer 1.0.

Rankingfunktion		Produktionswahrscheinlichkeit			KL-Divergenz		
Bewertungsmaß	Länge	PG	KN	R04	PG	KN	R04
Durchsch. Rang	100	76.46	47.98	68.67	124.93	4.78	40.5
	500	80.56	31.11	34.61	123.61	1.00	43.72
	1000	74.29	3.17	42.56	115.41	1.00	32.61
	5000	74.69	1.28	50.40	89.48	1.00	36.40
	10000	82.02	6.89	53.08	102.56	1.00	36.25
	100000	76.35	11.31		69.01	1.00	
Standard-abweichung	100	46.78	42.54	56.05	46.15	11.14	29.03
	500	50.83	54.51	42.47	45.91	0.00	36.49
	1000	43.99	7.55	39.25	52.16	0.00	23.63
	5000	44.64	2.16	31.16	59.21	0.00	22.26
	10000	52.49	23.90	30.17	59.63	0.00	28.07
	100000	47.84	38.00		40.60	0.00	
MRR	100	0.03	0.08	0.09	0.01	0.81	0.05
	500	0.04	0.46	0.07	0.01	1.00	0.06
	1000	0.04	0.78	0.11	0.02	1.00	0.05
	5000	0.04	0.97	0.03	0.04	1.00	0.04
	10000	0.04	0.85	0.03	0.03	1.00	0.05
	100000	0.04	0.89		0.04	1.00	

Tabelle 5.5: Ranking-Statistiken für das tatsächliche Jahr für die drei Testkorpora: Projekt-Gutenberg-Testkorpus (PG), Künstlicher-Testkorpus (KN) und Robust04-Testkorpus (R04).

Fazit. Gezeigt wurde, dass sich das tatsächliche Jahr bei Texten aus dem Künstlichen-Testkorpus immer am Rang 1 befindet und zwar ab einer Textlänge von 500 Wörtern. Für das Projekt-Gutenberg-Testkorpus und das Robust04-Testkorpus hingegen zeigt sich, dass die Rankingfunktionen das tatsächliche Jahr weit außerhalb der Top-10 einsortieren, so dass eine hohe Datierungsgenauigkeit für Jahre mit den Top-10-Klassifizierern für keines der beiden Testkorpora zu erreichen ist. Eine maximale Obergrenze für die Datierungsgenauigkeit mit Hilfe der Top-10-Ergebnissen lässt sich durch einen „idealen“ Top-10-Klassifizierer angeben.

Gibt es eine maximale Obergrenze für die Datierungsgenauigkeit?

Um eine maximale Obergrenze für die Datierungsgenauigkeit mit der Top-10 zu bestimmen, nehmen wir an, es gäbe einen „idealen“ Top-10-Klassifizierer, der Kenntnis über das tatsächliche Jahr und dessen Rang hat. Der „ideale“ Klassifizierer bestimmt aus den Top-10-Ergebnissen das Jahr als Datum, welches sich am nächsten zum tatsächlichen Jahr befindet. Falls sich das tatsächliche Jahr innerhalb der Top-10 befindet, wird dieses als Datum ausgewählt.

Für die drei Testkorpora sind die besten Ergebnisse des idealen Klassifizierers für die Rankingfunktionen in Tabelle 5.6 angegeben. Die Ergebnisse für das Künstliche-Testkorpus bzgl. beider Rankingfunktionen überraschen nicht: Eine nahezu perfekte Datierungsgenauigkeit ist gegeben, zumal das tatsächliche

Testkorpus		Projekt-Gutenberg		Künstliches-Testkorpus		Robust04-Testkorpus	
	Genauigkeit (%)	PW	KL	PW	KL	PW	KL
	Jahr	**6.99**	6.98	**99.04**	**100.00**	**22.22**	11.11
	Jahrzehnt	**38.54**	47.84	**99.04**	**100.00**	**66.67**	**50.00**
	Jahrhundert	**99.00**	99.34	**99.04**	**100.00**	**100.00**	**100.00**

Tabelle 5.6: Auflistung der höchsten Genauigkeiten des idealen Klassifizierers, die für die Rankingfunktionen Produktionswahrscheinlichkeit (PW) und Kullback-Leibler-Divergenz (KL) auf den drei Testkorpora erzielt werden. Hervorgehoben sind statistisch signifikante Verbesserungen in der Datierungsgenauigkeit des Klassifizierers im Vergleich zum Baseline-Klassifizierer Zufälliges-Jahr. Diese Ergebnisse sind ein Auszug aus Tabelle A.2 von Seite 54.

Jahr unter den Top-10-Ergbissen liegt. Wie bereits gezeigt, liegt das tatsächliche Jahr für die anderen beiden Testkorpora nicht unter den Top-10. Aus diesem Grund erwarten wir keine hohe Datierungsgenauigkeit für Jahre; für die Datierungsgenauigkeit für Jahrzehnte und Jahrhunderte allerdings erwarten wir Genauigkeiten über 80%.

Aus Tabelle 5.6 ist jedoch zu entnehmen, dass die Ergebnisse für Jahrzehnte deutlich schlechter als erwartet ausfallen: Für das Projekt-Gutenberg-Testkorpus liegt die höchste Genauigkeit für Jahrzehnte bei 47.84%, wogegen für das Robust04-Testkorpus diese 66.67% beträgt.

Fazit. Der „ideale“ Klassifizierer bestimmt für die Datierungsgenauigkeit für Jahre und Jahrzehnte eine maximale Obergrenze, die für das Projekt-Gutenberg-Testkorpus und das Robust04-Testkorpus sehr niedrig ist und dadurch unzureichend ist, um erfolgreich zur Datierung eingesetzt zu werden. Hingegen ist die Obergrenze der Datierungsgenauigkeit für das Künstliche-Testkorpus zu fast 100% korrekt.

Wir können hieraus bereits schließen, dass eine erfolgreiche Datierung nur für das Künstliche-Testkorpus gegeben ist, da dieses Testkorpus aus der Verteilung des Trainingskorpus stammt, welches nicht für die anderen beiden Testkorpora gegeben ist. Die charakteristischen Wörter der Belletristik (Projekt-Gutenberg-Testkorpus) und der Zeitungsartikel (Robust04-Testkorpus) sind über die Jahre entweder kaum im Trainingskorpus vertreten oder ihre Häufigkeiten sind nicht repräsentativ, so dass keine „exakteren“ Unigramm-Modelle erstellt werden können.

Testkorpus		Projekt-Gutenberg		Künstliches-Testkorpus		Robust04-Testkorpus	
Bewertungsmaß		PW	KL	PW	KL	PW	KL
Genauigkeit (%)	Jahr	1.63	1.00	**95.37**	**100.00**	11.11	6.67
	Klassifizierer	ZT10	FJ	T1	T1	DT10	ZJ
	Jahrzehnt	**8.64**	**8.31**	**95.37**	**100.00**	38.89	22.22
	Klassifizierer	ZT10 MT10	FJ	T1	T1	DT10	DT10
	Jahrhundert	**76.08**	**76.08**	**95.37**	**100.00**	**94.44**	**100.00**
	Klassifizierer	FJ	FJ	T1	T1	T1	T1 ZT10 DT10 MT10

Tabelle 5.7: Auflistung der Klassifizierer, die für die Rankingfunktionen Produktionswahrscheinlichkeit (PW) und Kullback-Leibler-Divergenz (KL), die beste Datierungsgenauigkeit auf den drei Testkorpora erzielen. Hervorgehoben sind statistisch signifikante Verbesserungen in der Datierungsgenauigkeit der Klassifizierer im Vergleich zum Baseline-Klassifizierer Zufälliges-Jahr. Die Abkürzungen stehen für folgende Klassifizierer ZT10: Zufälliges-Jahr-aus-den-Top-10; MT10: Mitte-der-Top-10; T1: Top-1; DT10: Durchschnittliches-Jahr-aus-den-Top-10; FJ: Festes-Jahr; ZJ: Zufälliges-Jahr. Diese Ergebnisse sind ein Auszug aus Tabelle A.1, Tabelle A.3 und Tabelle A.4 ab Seite 53.

Welche Kombination aus Rankingfunktion und Klassifizierer liefert die beste Datierung?

Obwohl gezeigt wurde, dass die bestmögliche Datierungsgenauigkeit für Jahre und Jahrzehnte selbst mit dem „idealen“ Klassifizierer niedrig ist, wollen wir die Ergebnisse für die Kombinationen aus Rankingfunktion und Klassifizierer vorstellen, die die höchsten Datierungsgenauigkeiten für Jahre, Jahrzehnte und Jahrhunderte im Hinblick auf die Testkorpora Projekt-Gutenberg und Robust04 erzielen (vgl. Tabelle 5.7).

Es ist zu sehen, dass die Unterschiede der höchsten Genauigkeit für Jahre in Bezug auf die Baseline Zufälliges-Jahr weder für das Projekt-Gutenberg-Testkorpus noch für das Robust04-Testkorpus statistisch signifikant sind. Zumal diese Genauigkeiten deutlich unter den Genauigkeiten des idealen Klassifizierers liegen. Die höchste Genauigkeit auf Jahresbasis für das Projekt-Gutenberg-Testkorpus liegt bei 1.63% und wird vom Klassifizierer Zufälliges-Jahr-aus-den-Top-10 erzielt; für das Robust04-Testkorpus hingegen liegt diese bei 11.11% und wird vom Klassifizierer Durchschnittliches-Jahr-aus-den-Top-10 erreicht. Diese Genauigkeiten werden in Kombination mit dem Produktionswahrscheinlichkeits-Ranking erzielt. Die Datierung anhand des KL-Divergenz-Rankings ist zumal für das Projekt-Gutenberg-Testkorpus ungeeignet, da hier die höchsten Genauigkeiten für Jahre, Jahrzehnte und Jahrhunderte durch die Baseline Festes-Jahr erreicht werden (vgl. Tabelle 5.8). Für das Robust04-

Klassifizierer			Zufälliges Jahr		Festes Jahr (1904)		Top-1		Zufälliges Jahr aus den Top-10		Durchsch. Jahr aus den Top-10		Mitte der Top-10	
Bewertungsmaß		Länge	PW	KL	PW	KL	PW	KL	PW	KL	PW	KL	PW	KL
Genauigkeit (%)	Jahr	100		R04		PG	R04,**KN**	**KN**			PG		**KN**	
		500		PG			**KN**	**KN**					PG	
		1000			PG	PG	PG, **KN**	**KN**	PG		PG,R04	PG	PG	
		5000		PG,R04			**KN**	**KN**	PG					
		10000	PG	PG		PG	PG,**KN**	PG,**KN**	PG		R04	PG	PG	PG
		100000			PG	PG	**KN**	PG,**KN**		PG	PG			
	Jahrzehnt	100				**PG**	**KN**	**KN**			**PG**	R04	R04	
		500				**PG**	**KN**	R04,**KN**			R04			PG
		1000			**PG**	**PG**	PG,**KN**	**KN**	R04		**PG**	R04	**PG**	
		5000		R04			R04,**KN**	**KN**	**PG**	PG		R04		R04
		10000	R04	R04		**PG**	R04,**KN**	PG,**KN**	**PG**	R04	**PG**	PG		PG
		100000					**KN**	**KN**	**PG**	PG			**PG**	
	Jahrhundert	100			**PG**	**PG**		**R04**,**KN**			R04		**KN**	
		500				**PG**	**R04**	**KN**	R04		PG,R04	**R04**	R04	**R04**
		1000			**PG**	**PG**	PG,**R04**,**KN**	**KN**		**R04**		**R04**	PG	**R04**
		5000			**PG**	**PG**	PG,**R04**,**KN**	PG,**R04**,**KN**		**R04**	PG	**R04**	PG	PG,**R04**
		10000			**PG**	**PG**	**R04**, **KN**	**R04**,**KN**				**R04**	PG	**R04**
		100000			**PG**		**KN**	**KN**				PG		

Tabelle 5.8: Auflistung der Klassifizierer, die für die Rankingfunktionen Produktionswahrscheinlichkeit (PW) und Kullback-Leibler-Divergenz (KL), die beste Datierung auf den drei Trainingskorpora erzielen: Projekt-Gutenberg-Testkorpus (PG), Künstlicher-Testkorpus (KN) und Robust04-Testkorpus (R04). Mehrere Einträge eines Testkorpus für eine bestimmte Textlänge bei unterschiedlichen Klassifizierern bedeuten, dass diese Klassifizierer für den Testkorpus die gleiche Genauigkeit erzielen und somit austauschbar sind. Hervorgehoben sind die Testkorpora, bei denen der entsprechende Klassifizierer eine statistisch signifikante Verbesserung in der Datierungsgenauigkeit im Vergleich zum Baseline-Klassifizierer Zufälliges-Jahr für diesen Testkorpus liefert. Diese Ergebnisse basieren auf Tabelle A.1, Tabelle A.3 und Tabelle A.4 ab Seite 53.

Testkorpus erreicht der Klassifizierer Durchschnittliches-Jahr-aus-den-Top-10 für beide Rankingfunktionen eine höhere Datierungsgenauigkeit als beim Projekt-Gutenberg-Testkorpus, allerdings sind diese Werte für Jahre und Jahrzehnte zu niedrig.

Fazit. Es zeigt sich, dass keine Kombination aus Rankingfunktion und Klassifizierer eine hohe Datierungsgenauigkeit für Jahre uns Jahrzehnte liefert: Sowohl für das Projekt-Gutenberg-Testkorpus als auch für das Robust04-Testkorpus sind diese Datierungsgenauigkeiten unzureichend. Für Jahre geht die Datierungsgenauigkeit nicht über 11.11%, für Jahrzehnte nicht über 38.89% hinaus. Die Datierungsgenauigkeiten für Jahrhunderte sind hingegen für das Projekt-Gutenberg-Testkorpus mit 76.08% akzeptabel, für das Robust04-Testkorpus mit 100% perfekt. Beim Robust04-Testkorpus sind allerdings nur 81 Texte aus sechs Jahren vorhanden, so dass die höhe Genauigkeit für Jahrhunderte nicht überrascht. Dies lässt uns schließen, dass das Google-Books-N-

Gramm-Korpus als Trainingskorpus zur Erstellung von Unigramm-Modellen für Belletristik und Zeitungsartikel keine geeignete Wahl zur Datierung darstellt.

5.5 Diskussion

Die Ergebnisse für das Künstliche-Testkorpus zeigen, dass eine Datierung zu 100% erreicht werden kann: Der Top-1-Klassifizierer datiert nahezu perfekt. Im Vergleich zum Top-1-Klassifizierer sind die anderen Klassifizierer jedoch nur für eine Bestimmung des Jahrhunderts geeignet. Für die anderen beiden Testkorpora Projekt-Gutenberg und Robust04 ist eine hohe Datierungsgenauigkeit nur für Jahrhunderte gegeben.

Dies lässt uns schließen, dass sowohl das Projekt-Gutenberg-Testkorpus als auch das Robust04-Testkorpus eine andere Charakteristik als der Trainingskorpus haben: Die beiden Testkorpora enthalten Wörter, deren Häufigkeiten im Trainingskorpus entweder zu gering sind oder nicht hoch genug sind, um diskriminierende Unigramm-Modelle zu erstellen. Eine weiterer Grund für die niedrigen Datierungsgenauigkeiten auf beiden Testkorpora ist die Auflösung der Unigramm-Modelle: Für Texte des Projekt-Gutenberg-Testkorpus sind Jahres-Modelle zu fein, da belletristische Texte über mehrere Jahre geschrieben werden, wohingegen für Zeitungsartikel des Robust04-Testkorpus Jahres-Modelle zu grob sind, da Zeitungsartikel innerhalb wenigen Stunden geschrieben werden.

Kapitel 6

Zusammenfassung und Ausblick

In dieser Arbeit werden Verfahren zur Datierung von Textdokumeten untersucht. Ziel ist die Bestimmung des Jahres, in dem ein Textdokument verfasst wurde. Die Datierung der Textdokumente erfolgt dabei ausschließlich auf Basis von Unigramm-Modellen. Diese werden mittels des englischen Google-Books-N-Gramm-Korpus für die Jahre zwischen 1800 und 2008 erstellt.

Für ein zu datierendes Textdokument werden die Unigramm-Modelle aller Jahre nach den Rankingfunktionen Produktionswahrscheinlichkeit und Kullback-Leibler-Divergenz absteigend bzw. aufsteigend sortiert. Die Produktionswahrscheinlichkeit bestimmt dabei die Wahrscheinlichkeit, mit der das Unigramm-Modell eines bestimmten Jahres das zu datierende Textdokument erzeugt. Im Gegensatz dazu misst die Kullback-Leibler-Divergenz die Distanz zwischen dem Unigramm-Modell des zu datierenden Textdokuments und dem Unigramm-Modell eines Jahres.

Auf den erstellten Rankings werden sechs verschiedene Klassifizierer zur Bestimmung des Datums verwendet. Zwei davon, Zufälliges-Jahr und Festes-Jahr, sind Baseline-Klassifizierer; die restlichen vier werden auf den Top-10 Rängen der Rankings eingesetzt. Annahme ist, dass sich sowohl das tatsächliche Jahr eines Textes als auch Jahre mit geringer Entfernung zum tatsächlichen Jahr in den Top-10 befinden. Die vier Top-10-Klassifizierer heißen Top-1, Zufälliges-Jahr-aus-den-Top-10, Durchschnittliches-Jahr-aus-den-Top-10 und Mitte-der-Top-10.

Zur Evaluierung der Verfahren wurden drei Testkorpora auf Basis von Belletristik, Zeitungsartikeln und künstlich erzeugten Texten neu erstellt. Jedes Testkorpus enthält Texte aus dem Zeitraum zwischen 1800 und 2008, wobei pro Jahr drei Texte zur Verfügung stehen.

Es zeigt sich, dass die in dieser Arbeit vorgestellten Verfahren prinzipiell in der Lage sind, Textdokumente mit hoher Genauigkeit korrekt zu datieren. Experimentell wird das anhand der Datierung von Textdokumenten aus dem

künstlich erzeugten Testkorpus belegt. Bei dem Kullback-Leibler-Divergenz-Ranking beträgt die Datierungsgenauigkeit 100% für korrekt datierte Jahre, Jahrzehnte und Jahrhunderte unter Verwendung des Top-1-Klassifizierers ab einer Textlänge von 500 Wörtern. Somit ist eine nahezu perfekte Datierung gegeben, wenn die Testdokumente aus der Verteilung des Google Books N-Gramm-Korpus stammen.

Bezüglich der Testkorpora für Zeitungsartikel und Belletristik zeigt sich allerdings, dass eine solche Datierungsgenauigkeit unerreichbar ist: Für Zeitungsartikel wird die höchste Datierungsgenauigkeit mit 11.11% beim Produktionswahrscheinlichkeits-Ranking mittels des Klassifizierers Durchschnittliches-Jahr-aus-den-Top-10 bei einer Textlänge von 1 000 Wörtern erzielt. Für den belletristischen Testkorpus beträgt die höchste Datierungsgenauigkeit 1.63%. Diese wird durch den Klassifizierer Zufälliges-Jahr-aus-den-Top-10 bei dem Produktionswahrscheinlichkeits-Ranking bei einer Textlänge von 5 000 Wörtern erreicht.

Ein Grund für die schlechten Ergebnisse könnten die fehlerhaften Publikationsdaten des Google-Books-N-Gramm-Korpus sein [Nunberg, 2009]. Diese haben Auswirkungen auf die Genauigkeit der N-Gramm-Modelle, denn das Vokabular und die Häufigkeiten der jeweiligen Jahre sind somit nicht repräsentativ. Die schlechten Ergebnisse lassen sich auch damit erklären, dass belletristische Bücher über einen Zeitraum von mehreren Jahren geschrieben werden und somit N-Gramm-Modelle auf Jahresbasis zu fein sind. Bei Zeitungsartikeln trifft das Gegenteil zu: Diese geben das aktuelle Tagesgeschehen wieder, so dass N-Gramm-Modelle auf Jahresbasis zu grob sind. Des Weiteren lässt sich schließen, dass die Genres des Google-Books-N-Gramm-Korpus nicht mit den Genres der verwendeten Belletristik und Zeitungsartikel übereinstimmen. Wir vermuten, dass das Google-Books-N-Gramm-Korpus auf Basis von Fachliteratur erstellt wurde.

Zur Verbesserung der Datierungsverfahren in zukünftiger Forschung sind folgende Ansätze denkbar.

Bei der Berechnung der Genauigkeit für Jahre fordern wir aktuell eine exakte Übereinstimmung von vorhergesagtem und tatsächlichem Jahr. Falls das vorhergesagte Jahr allerdings geringfügig vom tatsächlichen Jahr abweicht, stellt das vorhergesagte Jahr trotzdem eine ausreichende Annäherung dar. Um dies bei der Berechnung der Genauigkeit zu berücksichtigen, könnte eine Gauß-Funktion benutzt werden, deren Erwartungswert beim tatsächlichen Jahr liegt und mit steigender Entfernung vom Erwartungswert abfällt.

Anstatt Unigramm-Modelle auf Basis des gesamten Vokabulars aus einem Jahr zu erstellen, könnte das Vokabular zunächst gefiltert werden. Dazu schlagen wir vor, N-Gramme zu entfernen, die nicht charakteristisch für das jeweilige Jahr sind: Beispielsweise N-Gramme, die weniger als zehn Mal im Jahr vor-

kommen. Weiterhin sollte untersucht werden, ob N-Gramm-Modelle auf Basis von Schlüsselwörtern, deren Häufigkeiten über einen bestimmten Schwellwert liegen, zur Verbesserung der Datierungsgenauigkeit beitragen.

Durch den Einsatz von N-Gramm-Modellen höherer Ordnung sollte ebenfalls eine Verbesserung der Datierungsgenauigkeit möglich sein, da diese die Reihenfolge der Wörter im Gegensatz zum Unigramm-Modell berücksichtigen.

Neben der Bildung von N-Gramm-Modellen auf Wortbasis könnten N-Gramm-Modelle auf Buchstabenbasis erstellt werden. Diese haben sich für die Spracherkennung im Vergleich zu N-Gramm-Modellen auf Wortbasis als besser erwiesen.

Um eine effiziente Datierung zu gewährleisten, könnten N-Gramm-Modelle mit unterschiedlicher zeitlicher Auflösung hierarchisch erstellt werden. In der Folge finden - im Gegensatz zu derzeitigen Ansätzen - für ein zu datierendes Textdokument weniger Vergleiche pro Auflösungsstufe statt. Zur Realisierung einer solchen Hierarchie würden zunächst für verschiedene Jahrhunderte N-Gramm-Modelle erstellt werden. Für die so bestimmten Jahrhundert-Modelle könnten folgend Jahrzehnte-Modelle erstellt werden. Anschließend werden für diese Jahrzehnte-Modelle Jahres-Modelle als feinste Auslösungsstufe erzeugt.

Anhang A

Tabellen zur Evaluierung

Nachfolgende Tabellen stellen die Experimentergebnisse für die Testkorpora Projekt-Gutenberg, Künstliches-Testkorpus und Robust04-Testkorpus dar.

Klassifizierer		Zufälliges Jahr		Festes Jahr (1904)		Top-1		Zufälliges Jahr aus den Top-10		Durchsch. Jahr aus den Top-10		Mitte der Top-10	
Bewertungsmaß	Länge	PW	KL	PW	KL	PW	KL	PW	KL	PW	KL	PW	KL
Durchsch. abs. Entfernung	100	68.40	71.07	52.25	52.25	53.02	26.74	54.28	65.65	46.67	66.49	45.52	72.01
	500	70.13	71.94	52.25	52.25	37.99	0.00	48.56	24.45	42.12	24.34	41.80	25.34
	1000	68.04	66.67	52.25	52.25	18.38	0.00	39.03	22.30	36.66	21.09	37.11	21.67
	5000	71.47	69.28	52.25	52.25	3.70	0.00	37.63	19.63	36.38	18.26	38.73	18.83
	10000	68.55	71.13	52.25	52.25	12.19	0.00	38.41	17.55	37.52	17.59	38.19	17.73
	100000	68.57	68.82	52.25	52.25	7.15	0.00	40.21	17.78	38.57	17.20	40.67	17.38
Standardabweichung	100	47.72	49.51	30.17	30.17	40.73	52.45	42.25	55.65	30.32	48.41	31.34	53.31
	500	48.11	51.10	30.17	30.17	41.29	0.00	41.88	23.46	29.47	18.60	31.56	21.45
	1000	50.72	47.25	30.17	30.17	31.86	0.00	31.57	20.44	25.14	16.86	26.53	19.43
	5000	49.85	47.82	30.17	30.17	17.15	0.00	29.04	18.74	23.71	15.27	25.55	17.18
	10000	46.99	49.97	30.17	30.17	28.35	0.00	31.41	18.39	25.70	14.77	26.22	16.55
	100000	49.19	48.04	30.17	30.17	24.83	0.00	30.67	18.67	24.03	14.72	25.80	16.04
Durchsch. Entfernung	100	-1.09	3.06	0.00	0.00	-9.68	-26.67	-14.07	-63.24	-13.56	-64.18	-7.95	-69.44
	500	1.94	-1.29	0.00	0.00	-13.71	0.00	-18.92	-21.77	-17.39	-22.53	-13.04	-23.05
	1000	-2.09	1.90	0.00	0.00	-11.84	0.00	-17.67	-19.30	-17.97	-19.21	-14.94	-19.45
	5000	1.56	2.48	0.00	0.00	-3.70	0.00	-15.55	-16.95	-15.89	-16.41	-13.57	-16.86
	10000	-1.93	-2.54	0.00	0.00	-10.07	0.00	-20.02	-15.18	-19.65	-15.76	-16.04	-15.76
	100000	-0.86	-0.60	0.00	0.00	-2.21	0.00	-14.78	-15.33	-14.59	-15.47	-12.94	-15.47
Standardabweichung	100	83.39	86.56	60.33	60.33	66.16	52.48	67.33	58.37	53.98	51.43	54.69	56.62
	500	85.02	88.24	60.33	60.33	54.40	0.00	61.27	25.96	48.37	20.76	50.73	23.89
	1000	84.84	81.69	60.33	60.33	34.82	0.00	46.99	23.29	40.66	18.98	43.10	21.65
	5000	87.12	84.14	60.33	60.33	17.15	0.00	44.92	21.19	40.41	17.24	44.37	19.12
	10000	83.09	86.89	60.33	60.33	29.17	0.00	45.40	20.39	41.01	16.71	43.46	18.44
	100000	84.38	83.93	60.33	60.33	25.74	0.00	48.36	20.73	43.04	16.53	46.39	17.89
Genauigkeit (%) Jahr	100	0.48	0.96	0.48	0.48	1.75	76.08	0.96	10.53	0.32	0.32	1.75	1.12
	500	0.48	0.16	0.48	0.48	34.13	**100.00**	5.58	10.21	0.16	1.75	2.07	2.87
	1000	0.32	0.48	0.48	0.48	70.33	**100.00**	8.61	8.13	0.64	**2.55**	**3.51**	**5.26**
	5000	0.32	0.64	0.48	0.48	**95.37**	**100.00**	8.61	7.50	0.48	2.07	0.80	3.83
	10000	0.80	0.80	0.48	0.48	80.70	**100.00**	**8.77**	10.05	0.64	2.39	2.23	3.35
	100000	0.80	0.32	0.48	0.48	88.52	**100.00**	7.97	**10.69**	0.96	1.91	0.96	3.35
Jahrzehnt	100	4.15	3.51	4.78	4.78	8.29	76.08	5.74	16.27	5.74	6.86	7.34	7.81
	500	4.15	4.78	4.78	4.78	35.73	**100.00**	12.76	24.24	5.42	17.07	9.41	22.17
	1000	6.22	3.67	4.78	4.78	70.49	**100.00**	**14.83**	24.56	**8.93**	19.62	**10.69**	23.76
	5000	4.94	4.94	4.78	4.78	**95.37**	**100.00**	14.67	25.68	5.42	23.44	7.81	27.75
	10000	4.63	3.99	4.78	4.78	81.34	**100.00**	14.51	**32.06**	7.81	23.29	8.13	**28.07**
	100000	5.26	4.31	4.78	4.78	89.31	**100.00**	**14.83**	31.10	6.06	**25.36**	6.86	27.91
Jahrhundert	100	41.95	39.55	**47.85**	**47.85**	55.50	79.90	55.34	47.85	55.82	42.26	59.17	42.26
	500	41.47	41.15	**47.85**	**47.85**	66.51	**100.00**	58.69	84.53	63.80	88.04	63.32	81.66
	1000	45.61	42.42	**47.85**	**47.85**	79.27	**100.00**	**66.19**	86.76	70.33	91.07	**67.94**	87.88
	5000	41.95	42.26	**47.85**	**47.85**	**95.37**	**100.00**	65.55	90.11	70.18	95.69	65.39	93.78
	10000	42.26	40.99	**47.85**	**47.85**	86.44	**100.00**	66.03	**92.03**	**70.49**	**96.65**	66.51	95.06
	100000	44.18	42.90	**47.85**	**47.85**	94.10	**100.00**	64.75	91.71	68.74	96.17	64.27	**95.53**

Tabelle A.1: Datierung der Textdokumente aus dem Künstlichen-Testkorpus anhand der Rankingfunktionen Produktionswahrscheinlichkeit (PW) und Kullback-Leibler-Divergenz (KL). Hervorgehoben sind statistisch signifikante Verbesserungen in der Datierungsgenauigkeit der Klassifizierer im Vergleich zum Baseline-Klassifizierer Zufälliges-Jahr.

Testkorpus			Projekt-Gutenberg		Künstliches-Testkorpus		Robust04-Testkorpus	
Bewertungsmaß		Länge	PW	KL	PW	KL	PW	KL
Genauigkeit (%)	**Durchsch. abs. Entfernung**	100	17.50	83.52	10.64	10.97	19.06	12.61
		500	19.36	83.52	8.00	0.00	8.22	12.72
		1000	23.10	66.82	3.53	0.00	6.56	6.94
		5000	26.69	25.39	0.74	0.00	31.00	10.07
		10000	22.57	29.45	2.94	0.00	25.00	14.33
		100000	15.38	10.32	1.80	0.00		
	Standard-abweichung	100	17.35	42.59	13.80	33.88	30.66	15.93
		500	17.90	42.59	17.64	0.00	18.71	15.42
		1000	18.83	42.09	13.82	0.00	6.68	6.08
		5000	23.99	24.46	7.50	0.00	36.98	9.35
		10000	22.39	27.93	12.92	0.00	26.43	15.56
		100000	15.15	11.61	9.54	0.00		
	Durchsch. Entfernung	100	-5.33	-83.51	-3.23	-10.97	8.94	9.50
		500	1.46	-83.52	-3.93	0.00	3.44	8.06
		1000	1.03	-66.73	-2.87	0.00	0.67	0.28
		5000	-17.57	-18.96	-0.74	0.00	25.80	3.13
		10000	-14.90	-25.86	-2.69	0.00	19.33	11.50
		100000	3.99	-4.31	-0.92	0.00		
	Standard-abweichung	100	24.06	42.62	17.13	33.88	34.98	17.96
		500	26.33	42.59	18.97	0.00	20.15	18.30
		1000	29.78	42.23	13.97	0.00	9.33	9.22
		5000	31.29	29.72	7.50	0.00	40.78	13.38
		10000	28.08	31.28	12.98	0.00	30.82	17.76
		100000	21.22	14.93	9.66	0.00		
	Jahr	100	5.59	0.00	15.79	88.68	11.11	11.11
		500	**6.99**	0.00	65.23	**100.00**	**22.22**	11.11
		1000	6.29	2.10	93.30	**100.00**	**22.22**	11.11
		5000	6.29	6.29	**99.04**	**100.00**	0.00	0.00
		10000	6.08	4.91	91.55	**100.00**	8.33	0.00
		100000	6.31	6.98	90.59	**100.00**		
	Jahr-zehnt	100	28.67	1.40	48.49	88.68	44.44	44.44
		500	31.00	1.40	71.93	**100.00**	**66.67**	38.89
		1000	23.08	6.99	93.30	**100.00**	61.11	**50.00**
		5000	20.28	24.48	**99.04**	**100.00**	20.00	26.67
		10000	34.11	26.64	93.94	**100.00**	25.00	33.33
		100000	**38.54**	47.84	92.82	**100.00**		
	Jahr-hundert	100	93.47	26.57	97.77	90.75	88.89	94.44
		500	93.01	26.57	94.26	**100.00**	94.44	94.44
		1000	89.51	38.46	96.17	**100.00**	100.00	**100.00**
		5000	80.42	80.42	**99.04**	**100.00**	73.33	**100.00**
		10000	84.58	74.30	97.13	**100.00**	75.00	**100.00**
		100000	**99.00**	99.34	98.72	**100.00**		

Tabelle A.2: Datierung der Textdokumente der drei Testkorpora mit einem idealen Klassifizierer in Verbindung mit den Rankingfunktionen Produktionswahrscheinlichkeit (PW) und Kullback-Leibler-Divergenz (KL). Hervorgehoben sind statistisch signifikante Verbesserungen in der Datierungsgenauigkeit des Klassifizierers im Vergleich zum Baseline-Klassifizierer Zufälliges-Jahr.

Klassifizierer		Zufälliges Jahr		Festes Jahr (1904)		Top-1		Zufälliges Jahr aus den Top-10		Durchsch. Jahr aus den Top-10		Mitte der Top-10	
Bewertungsmaß	Länge	PG	R04	PG	R04	PG	R04	PG	R04	PG	R04	PG	R04
Durchsch. abs. Entfernung	100	64.87	108.11	38.85	87.50	85.06	38.56	47.37	45.67	39.78	36.17	37.62	36.67
	500	62.62	85.61	38.85	87.72	40.60	20.33	41.84	35.67	36.72	25.61	36.90	28.33
	1000	62.56	110.61	38.85	87.50	37.95	20.94	38.03	43.56	36.69	31.00	37.03	36.61
	5000	58.62	109.27	38.85	87.60	36.92	69.27	38.87	115.87	38.37	109.80	37.43	134.40
	10000	60.18	84.17	38.77	88.08	43.51	59.42	48.05	88.25	42.67	90.83	38.79	105.42
	100000	65.38		34.60		39.35		35.30		33.88		35.91	
Standardabweichung	100	44.82	54.64	25.49	1.71	53.86	43.87	37.46	52.08	27.02	39.57	24.14	42.56
	500	43.44	51.28	25.49	1.74	27.38	25.57	29.99	59.12	22.47	43.74	23.24	50.68
	1000	43.24	47.86	25.49	1.71	24.22	19.95	23.90	59.74	22.15	42.68	22.75	47.63
	5000	43.35	50.22	25.49	1.85	22.58	57.83	25.76	70.82	24.83	43.24	23.41	54.12
	10000	42.84	59.24	25.48	1.91	30.65	51.10	38.50	76.00	29.84	59.03	25.00	70.73
	100000	44.14		22.42		26.91		24.71		20.59		23.45	
Durchsch. Entfernung	100	-22.04	106.67	-18.51	87.50	-60.95	28.22	-18.36	41.78	-20.58	34.17	-12.13	35.22
	500	-17.94	84.17	-18.51	87.72	11.97	10.56	2.69	25.11	0.68	19.28	3.21	20.89
	1000	-15.32	109.61	-18.51	87.50	-14.51	11.17	-3.75	38.22	-4.51	27.11	-8.51	32.72
	5000	-19.49	109.27	-18.51	87.60	-7.51	66.87	-16.84	114.00	-16.51	109.80	-11.51	133.47
	10000	-20.68	83.67	-18.39	88.08	-31.67	54.58	-29.29	84.42	-29.50	88.50	-17.12	103.25
	100000	-17.04		-15.21		23.78		7.58		9.43		18.56	
Standardabweichung	100	75.70	57.40	42.62	1.71	80.14	51.14	57.53	55.25	43.46	41.31	43.02	43.77
	500	74.07	53.61	42.62	1.74	47.48	30.91	51.41	64.32	43.05	46.88	43.49	54.18
	1000	74.49	50.10	42.62	1.71	42.62	26.68	44.76	63.29	42.62	45.25	42.62	50.39
	5000	70.26	50.22	42.62	1.85	42.62	60.59	43.48	73.80	42.62	43.24	42.62	56.37
	10000	70.92	59.94	42.60	1.91	42.77	56.24	54.16	80.23	42.91	62.47	42.86	73.86
	100000	77.02		38.32		41.31		42.42		38.51		38.67	
Genauigkeit (%) Jahr	100	0.23	0.00	0.70	0.00	0.23	5.56	0.93	0.00	1.17	0.00	0.93	0.00
	500	0.23	0.00	0.70	0.00	0.70	0.00	0.93	0.00	0.70	0.00	1.40	0.00
	1000	0.23	0.00	0.70	0.00	0.70	5.56	0.70	0.00	0.70	11.11	0.70	0.00
	5000	0.47	0.00	0.70	0.00	0.70	0.00	1.63	0.00	0.70	0.00	0.70	0.00
	10000	0.93	0.00	0.70	0.00	0.93	0.00	0.93	0.00	0.47	8.33	0.93	0.00
	100000	0.66		1.00		0.66		0.66		1.00		0.66	
Jahrzehnt	100	4.20	0.00	6.99	0.00	3.50	5.56	6.99	11.11	7.69	22.22	6.76	27.78
	500	5.36	5.56	6.99	0.00	6.53	22.22	7.23	16.67	6.99	38.89	8.16	27.78
	1000	3.26	0.00	6.99	0.00	6.99	16.67	6.76	27.78	6.99	22.22	6.99	22.22
	5000	6.29	6.67	6.99	0.00	6.99	13.33	7.46	6.67	6.99	0.00	6.99	0.00
	10000	6.31	16.67	7.01	0.00	7.24	16.67	7.94	8.33	**7.94**	8.33	6.78	8.33
	100000	4.65		**8.31**		8.31		**8.64**		5.98		**8.64**	
Jahrhundert	100	45.69	16.67	69.93	0.00	30.54	72.22	59.21	66.67	66.43	77.78	69.46	72.22
	500	44.29	22.22	69.93	0.00	65.97	83.33	65.03	**83.33**	70.16	83.33	69.23	83.33
	1000	47.09	11.11	69.93	0.00	69.93	**94.44**	67.60	77.78	69.23	83.33	69.93	83.33
	5000	50.82	13.33	69.93	0.00	69.93	40.00	67.13	26.67	69.93	6.67	69.93	13.33
	10000	49.77	33.33	70.09	0.00	61.92	50.00	60.28	41.67	63.32	33.33	70.09	33.33
	100000	40.86		**76.08**		65.45		**73.75**		**74.75**		**70.76**	

Tabelle A.3: Datierung der Textdokumente aus dem Projekt-Gutenberg-Testkorpus (PG) und dem Robust04-Testkorpus (R04) anhand der Rankingfunktion Produktionswahrscheinlichkeit. Hervorgehoben sind statistisch signifikante Verbesserungen in der Datierungsgenauigkeit der Klassifizierer im Vergleich zum Baseline-Klassifizierer Zufälliges-Jahr.

Klassifizierer		Zufälliges Jahr		Festes Jahr (1904)		Top-1		Zufälliges Jahr aus den Top-10		Durchsch. Jahr aus den Top-10		Mitte der Top-10	
Bewertungsmaß	Länge	PG	R04	PG	R04	PG	R04	PG	R04	PG	R04	PG	R04
Durchsch. abs. Entfernung	100	61.81	80.83	38.85	87.50	97.51	20.94	95.16	21.33	94.51	20.11	92.51	21.28
	500	63.39	89.72	38.85	87.50	83.52	24.17	90.89	23.44	90.51	22.28	90.51	25.78
	1000	66.45	93.22	38.85	87.50	83.52	17.94	85.20	17.28	84.51	13.06	83.52	16.17
	5000	61.59	87.47	38.85	87.60	36.83	23.40	40.38	18.40	39.97	15.60	37.43	19.07
	10000	63.12	96.17	38.77	88.08	43.21	23.83	65.06	22.00	59.99	20.33	46.58	21.17
	100000	59.46		34.60		37.94		37.67		33.32		34.06	
Standardabweichung	100	42.81	56.51	25.49	1.71	42.62	17.88	43.30	19.76	42.62	18.26	42.62	18.04
	500	44.34	53.02	25.49	1.71	42.60	19.10	42.58	21.82	42.62	17.67	42.62	18.69
	1000	47.02	60.30	25.49	1.71	42.60	15.44	45.75	15.45	42.62	11.64	42.60	14.44
	5000	43.18	68.28	25.49	1.85	22.41	14.53	29.20	10.62	26.93	10.52	23.41	13.87
	10000	44.88	74.83	25.48	1.91	30.54	15.24	47.58	18.07	40.31	16.17	33.44	15.99
	100000	42.86		22.42		25.37		25.35		18.95		20.11	
Durchsch. Entfernung	100	-20.11	80.72	-18.51	87.50	-97.51	18.50	-95.16	18.00	-94.51	19.22	-92.51	19.17
	500	-18.95	89.72	-18.51	87.50	-83.51	20.61	-90.89	18.78	-90.51	21.39	-90.51	24.78
	1000	-18.76	92.44	-18.51	87.50	-83.51	8.50	-85.02	10.39	-84.51	9.39	-83.51	10.06
	5000	-20.33	86.27	-18.51	87.60	-6.51	17.40	-21.11	9.20	-22.51	11.73	-11.51	14.40
	10000	-18.57	93.83	-18.39	88.08	-31.39	12.00	-56.58	13.67	-58.39	12.83	-38.39	12.67
	100000	-12.99		-15.21		24.79		1.12		0.79		9.79	
Standardabweichung	100	72.45	56.67	42.62	1.71	42.62	20.40	43.30	22.84	42.62	19.20	42.62	20.27
	500	75.00	53.02	42.62	1.71	42.62	22.89	42.58	25.94	42.62	18.74	42.62	19.99
	1000	79.21	61.49	42.62	1.71	42.62	22.09	46.09	20.72	42.62	14.76	42.62	19.21
	5000	72.42	69.79	42.62	1.85	42.62	21.35	45.14	19.15	42.62	14.72	42.62	18.68
	10000	75.19	77.74	42.60	1.91	42.60	25.62	57.40	24.97	42.60	22.59	42.60	23.31
	100000	72.14		38.32		38.32		45.39		38.32		38.32	
Genauigkeit (%) Jahr	100	0.23	5.56	0.70	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	500	1.40	0.00	0.70	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	1000	0.00	0.00	0.70	0.00	0.00	0.00	0.00	0.00	0.70	0.00	0.00	0.00
	5000	1.40	6.67	0.70	0.00	0.70	0.00	0.93	0.00	0.70	0.00	0.70	0.00
	10000	0.70	0.00	0.70	0.00	0.70	0.00	0.23	0.00	0.70	0.00	0.70	0.00
	100000	0.66		1.00		1.00		1.00		0.00		0.00	
Jahrzehnt	100	5.13	16.67	6.99	0.00	0.00	16.67	0.47	16.67	0.00	22.22	0.00	11.11
	500	4.90	5.56	6.99	0.00	1.40	16.67	0.23	11.11	0.00	11.11	0.00	11.11
	1000	5.83	5.56	6.99	0.00	1.40	11.11	0.70	16.67	0.70	22.22	1.40	16.67
	5000	5.13	6.67	6.99	0.00	6.99	0.00	8.86	0.00	6.99	6.67	6.99	6.67
	10000	4.44	8.33	7.01	0.00	7.01	0.00	4.91	8.33	7.01	0.00	7.01	0.00
	100000	5.98		**8.31**		8.64		8.97		6.98		6.98	
Jahrhundert	100	43.36	27.78	69.93	0.00	16.78	94.44	17.95	83.33	18.88	88.89	20.28	88.89
	500	44.99	27.78	69.93	0.00	26.57	83.33	20.51	83.33	21.68	88.89	21.68	88.89
	1000	43.36	27.78	69.93	0.00	26.57	94.44	25.87	**100.00**	25.87	**100.00**	26.57	**100.00**
	5000	46.62	46.67	69.93	0.00	69.93	**100.00**	66.90	**100.00**	67.83	**100.00**	69.93	**100.00**
	10000	46.50	41.67	70.09	0.00	62.38	**100.00**	44.63	75.00	44.16	**100.00**	58.18	**100.00**
	100000	51.16		**76.08**		66.78		68.77		78.74		74.75	

Tabelle A.4: Datierung der Textdokumente aus dem Projekt-Gutenberg-Testkorpus (PG) und dem Robust04-Testkorpus (R04) anhand der Rankingfunktion Kullback-Leibler-Divergenz. Hervorgehoben sind statistisch signifikante Verbesserungen in der Datierungsgenauigkeit der Klassifizierer im Vergleich zum Baseline-Klassifizierer Zufälliges-Jahr.

Abbildungsverzeichnis

2.1 Beispiele zur Markov-Ketten der Ordnungen Null bis drei. . . . 8

3.1 Bestimmung des Datums eines Textdokuments. 13

4.1 Die Datierung eines Textdokuments als ein Zwei-Schritt-Prozess. 23
4.2 Vokabulargröße und Korpusgröße vom Trainingskorpus. 25
4.3 Die Datierung eines Textdokuments mittels der Produktionswahrscheinlichkeit. 27
4.4 Die Datierung eines Textdokuments mittels der Kullback-Leibler-Divergenz. 28
4.5 Beispiel zur Datierung eines Textdokuments. 29

5.1 Geltungsbereiche zur Berechnung der Genauigkeit für korrekt datierte Zeiträume. 34
5.2 Untersuchung des Projekt-Gutenberg-Testkorpus bzgl. ungesehene Unigramme. 37
5.3 Prozentsatz der Schnittmenge des Vokabulars eines ausgewählten Jahres mit allen anderen Jahren. 40
5.4 Gegenüberstellung der Kullback-Leibler-Divergenz für Unigramm-Modelle von ausgewählten Jahren. 41
5.5 Produktionswahrscheinlichkeit und Kullback-Leibler-Divergenz für Texte unterschiedlicher Textlänge aus dem Künstlichen-Testkorpus. 42

Tabellenverzeichnis

5.1 Verteilung der Bücher aus dem Projekt Gutenberg. 32
5.2 Verteilung der Zeitungsartikel des Robust04-Testkorpus. 33
5.3 Gegenüberstellung der Häufigkeit, ML-Schätzung, Add-Delta-Glättung und JM-Glättung für ausgewählte Wörter. 36
5.4 Prozent der ungesehenen Wörter des Robust04-Testkorpus und Wahrscheinlichkeitsmasse dieses Testkorpus. 37
5.5 Ranking-Statistiken für das tatsächliche Jahr für die drei Testkorpora. 44
5.6 Auflistung der höchsten Genauigkeiten des idealen Klassifizierers für die drei Testkorpora. 45
5.7 Auflistung der Klassifizierer mit höchster Genauigkeit für die drei Testkorpora. 46
5.8 Auflistung der besten Klassifizierer auf den drei Trainingskorpora. 47

A.1 Datierung der Textdokumente aus dem Künstlichen-Testkorpus anhand der Rankingfunktionen Produktionswahrscheinlichkeit und Kullback-Leibler-Divergenz 53
A.2 Datierung der Textdokumente der drei Testkorpora mit einem idealen Klassifizierer. 54
A.3 Datierung der Textdokumente aus dem Projekt-Gutenberg-Testkorpus und dem Robust04-Testkorpus anhand der Rankingfunktion Produktionswahrscheinlichkeit. 55
A.4 Datierung der Textdokumente aus dem Projekt-Gutenberg-Testkorpus und dem Robust04-Testkorpus anhand der Rankingfunktion Kullback-Leibler-Divergenz 56

Literaturverzeichnis

Leonard E. Baum. An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. *Inequalities*, 3:1–8, 1972.

Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 1st ed. 2006. corr. 2nd printing edition, Oktober 2007.

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, und Paul S. Roossin. A statistical approach to machine translation. *Computational Linguistics*, 16:79–85, Juni 1990.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, Jenifer C. Lai, und Robert L. Mercer. An estimate of an upper bound for the entropy of English. *Computational Linguistics*, 18(1):31–40, 1992.

Stanley F. Chen. An Empirical Study of Smoothing Techniques for Language Modeling. Technical Report TR-10-98, Computer Science Group, Harvard University, 1998.

Stanley F. Chen und Joshua Goodman. An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, ACL '96, Seiten 310–318, Stroudsburg, PA, USA, 1996. Association for Computational Linguistics.

Kenneth W. Church und William A. Gale. A comparison of the enhanced Good-Turing and deleted estimation methods for estimating probabilities of English bigrams. *Computer Speech and Language*, 5:19–54, 1991.

Thomas M. Cover und Joy A. Thomas. *Elements of information theory*. Wiley-Interscience, New York, NY, USA, 1991.

Franciska de Jong, Henning Rode, und Djoerd Hiemstra. Temporal Language Models for the Disclosure of Historical Text. In *Humanities, computers and*

cultural heritage: Proceedings of the XVIth International Conference of the Association for History and Computing (AHC 2005), Seiten 161–168. Royal Netherlands Academy of Arts and Sciences, 2005.

Arthur P. Dempster, Nan M. Laird, und Donald B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society, series B*, 39(1):1–38, 1977.

William A. Gale und Kenneth W. Church. Poor estimates of context are worse than none. In *Proceedings of the workshop on Speech and Natural Language*, HLT '90, Seiten 283–287, Stroudsburg, PA, USA, 1990. Association for Computational Linguistics.

William A. Gale und Kenneth W. Church. What's wrong with adding one. In *Corpus-Based Research into Language. Rodolpi*, 1994.

Anne Garcia-Fernandez, Anne L. Ligozat, Marco Dinarelli, und Delphine Bernhard. When was it written ? Automatically Determining Publication Dates. In *In proceedings of String Processing and Information Retrieval (SPIRE)*, Pisa, Italy, Oktober 2011.

Joshua Goodman. A Bit of Progress in Language Modeling. Technical Report MSR-TR-2001-72, Microsoft Research, 2000.

Harold Jeffreys. *Theory of Probability.* Clarendon Press, Oxford, 1948.

Frederick Jelinek und Robert L. Mercer. Interpolated estimation of Markov source parameters from sparse data. In *Proceedings, Workshop on Pattern Recognition in Practice*, Seiten 381–397. North Holland, Amsterdam, 1980.

Frederick Jelinek, Bernard Merialdo, Salim Roukos, und M. Strauss. A dynamic language model for speech recognition. In *Proceedings of the workshop on Speech and Natural Language*, HLT '91, Seiten 293–295, Stroudsburg, PA, USA, 1991. Association for Computational Linguistics.

W.E. Johnson. Probability: deductive and inductive problems. *Mind*, 41:421–423, 1932.

Daniel Jurafsky und James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition.* Prentice Hall, 2nd edition edition, Februar 2008.

Nattiya Kanhabua und Kjetil Nørvåg. Improving Temporal Language Models for Determining Time of Non-timestamped Documents. In *Proceedings of the 12th European conference on Research and Advanced Technology for Digital*

Libraries, ECDL '08, Seiten 358–370, Berlin, Heidelberg, 2008. Springer-Verlag.

Wessel Kraaij. *Variations on Language Modeling for Information Retrieval*. PhD thesis, University of Twente, Juni 2004.

Karen Kukich. Techniques for automatically correcting words in text. *ACM Comput. Surv.*, 24:377–439, Dezember 1992.

Victor Lavrenko und W. Bruce Croft. Relevance based language models. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '01, Seiten 120–127, New York, NY, USA, 2001. ACM.

George J. Lidstone. Note on the general case of the Bayes-Laplace formula for inductive or a posteriori probabilities. *Transactions of the Faculty of Actuaries*, 8:182–192, 1920.

Christopher D. Manning, Prabhakar Raghavan, und Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 1 edition, Juli 2008.

Jean-Baptiste Michel, Yuan K. Shen, Aviva P. Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, und Erez L. Aiden. Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*, 331(6014):176–182, Januar 2011.

David R. H. Miller, Tim Leek, und Richard M. Schwartz. A hidden Markov model information retrieval system. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '99, Seiten 214–221, New York, NY, USA, 1999. ACM.

Hermann Ney, S. Martin, und Frank Wessel. Statistical language modeling using leaving-one-out. *Corpus Based Methods in Language and Speech Processing*, Seiten 174–207, 1997.

Geoffrey Nunberg. Google's Book Search: A Disaster for Scholars, August 2009. URL `http://chronicle.com/article/Googles-Book-Search-A/48245/`.

Roeland J.F. Ordelman. *Dutch speech recognition in multimedia information retrieval*. PhD thesis, Enschede, 2003.

Ronald Rosenfeld. Two decades of statistical language modeling: Where do we go from here? In *Proceedings of the IEEE*, Seiten 1270–1278, 2000.

Stuart Russell und Peter Norvig. *Artificial Intelligence: A Modern Approach*. Prentice-Hall, Englewood Cliffs, NJ, 2nd edition edition, 2003.

Hinrich Schütze und Yoram Singer. Part-of-speech tagging using a Variable Memory Markov model. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, Seiten 181–187, Stroudsburg, PA, USA, 1994. Association for Computational Linguistics.

Ronald Walpole und Raymond Myers. *Probability and Statistics for Engineers and Scientists*. Macmillan Publishing Company, 1989.

Webis Group. AItools-Software-Suite, 2011. URL `http://aitools.de/`.

Jinxi Xu und W. Bruce Croft. Cluster-based language models for distributed retrieval. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '99, Seiten 254–261, New York, NY, USA, 1999. ACM.

ChengXiang Zhai. *Statistical Language Models for Information Retrieval*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2008.

ChengXiang Zhai und John Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems*, 22:179–214, April 2004.