

# Datierung von Textdokumenten

## Masterarbeit

Tsvetomira Palakarska

Bauhaus-Universität Weimar

3. Februar 2012

## Motivation

# Geisteswissenschaften

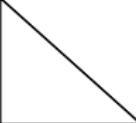
# IT-Forensik



# Internet



# Verfahren zur Datierung



Ihr naht euch wieder,  
schwankende Gestalten,  
Die früh sich einst dem  
trüben Blick gezeigt.  
Versuch ich wohl, euch  
diesmal festzuhalten?  
Fühl ich mein Herz noch  
jenem Wahn geneigt?

Gegeben: Undatierter Text  
Gesucht: Datum des Textes

# Verfahren zur Datierung

Ihr naht euch wieder,  
schwankende Gestalten,  
Die früh sich einst dem  
trüben Blick gezeigt.  
Versuch ich wohl, euch  
diesmal festzuhalten?  
Fühl ich mein Herz noch  
jenem Wahn geneigt?

11. JAN. 1808

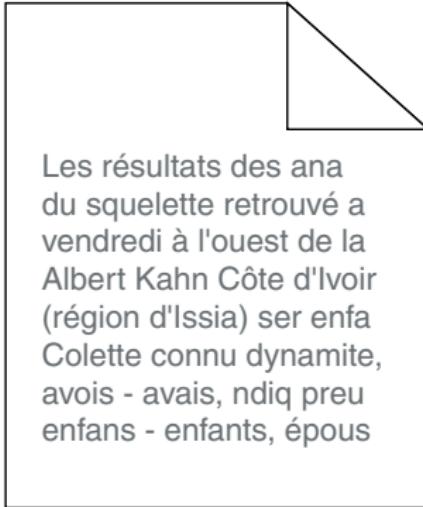


Gegeben: Undatierter Text

Gesucht: Datum des Textes

Paradigmen:

- Wissensbasiert
- Statistisch



Les résultats des ana  
du squelette retrouvé a  
vendredi à l'ouest de la  
Albert Kahn Côte d'Ivoir  
(région d'Issia) ser enfa  
Colette connu dynamite,  
avois - avais, ndiq preu  
enfans - enfants, épous

- Personennamen
- Neologismen
- Archaismen
- Rechtschreibreformen
  
- Ähnlichkeits-  
berechnung
- Klassifikation

Les résultats des ana  
du squelette retrouvé a  
vendredi à l'ouest de la  
**Albert Kahn** Côte d'Ivoir  
(région d'Issia) ser enfa  
**Colette** connu dynamite,  
avois - avais, ndiq preu  
enfans - enfants, épous

17. FEB. 1890



- Personennamen
- Neologismen
- Archaismen
- Rechtschreibreformen
  
- Ähnlichkeits-  
berechnung
- Klassifikation

Les résultats des ana  
du squelette retrouvé a  
vendredi à l'ouest de la  
Albert Kahn Côte d'Ivoir  
(région d'Issia) ser enfa  
Colette connu **dynamite**,  
avois - **avais**, ndiq preu  
enfans - **enfants**, épous

19. JUN. 1872



- Personennamen
- Neologismen
- Archaismen
- Rechtschreibreformen
  
- Ähnlichkeits-  
berechnung
- Klassifikation

Les résultats des ana  
du squelette retrouvé a  
vendredi à l'ouest de la  
Albert Kahn Côte d'Ivoir  
(région d'Issia) ser enfa  
Colette connu dynamite,  
**avois** - avais, ndiq preu  
**enfans** - enfants, épous

16. JUL. 1801



- Personennamen
- Neologismen
- Archaismen
- Rechtschreibreformen
  
- Ähnlichkeits-  
berechnung
- Klassifikation

Les résultats des ana  
du squelette retrouvé a  
vendredi à l'ouest de la  
Albert Kahn Côte d'Ivoir  
(région d'Issia) ser enfa  
Colette connu dynamite,  
**avoirs - avais**, ndiq preu  
**enfans - enfants**, épous

13. SEP. 1901



- Personennamen
- Neologismen
- Archaismen
- Rechtschreibreformen
  
- Ähnlichkeitsberechnung
- Klassifikation

Les résultats des ana  
du squelette retrouvé a  
vendredi à l'ouest de la  
Albert Kahn Côte d'Ivoir  
(région d'Issia) ser enfa  
Colette connu dynamite,  
avois - avais, ndiq preu  
enfans - enfants, épous

24. AUG. 1913



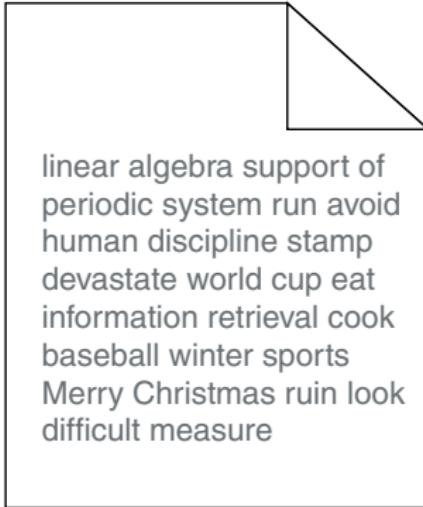
- Personennamen
- Neologismen
- Archaismen
- Rechtschreibreformen
  
- Ähnlichkeits-  
berechnung
- Klassifikation

Les résultats des ana  
du squelette retrouvé a  
vendredi à l'ouest de la  
**Albert Kahn** Côte d'Ivoir  
(région d'Issia) ser enfa  
**Colette** connu **dynamite**,  
**avois - avais**, ndiq preu  
**enfans - enfants**, épous

**27. JUL. 1884**



- Personennamen
- Neologismen
- Archaismen
- Rechtschreibreformen
  
- Ähnlichkeits-  
berechnung
- Klassifikation



linear algebra support of  
periodic system run avoid  
human discipline stamp  
devastate world cup eat  
information retrieval cook  
baseball winter sports  
Merry Christmas ruin look  
difficult measure

## Sprachmodelle für

- Kollokationen
- Konzepte
- Wortarten wie Verben

Datierung durch Vergleich  
von Sprachmodellen

linear algebra support of  
periodic system run avoid  
human discipline stamp  
devastate world cup eat  
information retrieval cook  
baseball winter sports  
**Merry Christmas** ruin look  
difficult measure

16. NOV. 1994



Sprachmodelle für

- Kollokationen
- Konzepte
- Wortarten wie Verben

Datierung durch Vergleich  
von Sprachmodellen

linear algebra support of  
periodic **system** run avoid  
human discipline stamp  
devastate world cup eat  
information retrieval cook  
**baseball winter sports**  
Merry **Christmas** ruin look  
difficult **measure**

**29. FEB. 1943**



## Sprachmodelle für

- Kollokationen
- Konzepte
- Wortarten wie Verben

Datierung durch Vergleich  
von Sprachmodellen

linear algebra support of  
periodic system run avoid  
human discipline stamp  
**devastate** world cup eat  
information retrieval cook  
baseball winter sports  
Merry Christmas ruin look  
difficult measure

12. APR. 1931



## Sprachmodelle für

- Kollokationen
- Konzepte
- Wortarten wie Verben

Datierung durch Vergleich  
von Sprachmodellen

linear algebra support of  
periodic system run avoid  
human discipline stamp  
devastate world cup eat  
information retrieval cook  
baseball winter sports  
Merry Christmas ruin look  
difficult measure

17. JUL. 2001

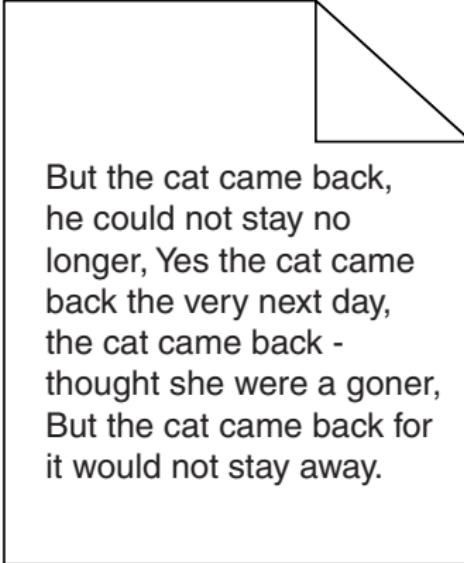


Sprachmodelle für

- Kollokationen
- Konzepte
- Wortarten wie Verben

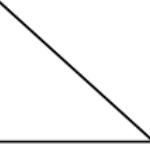
Datierung durch Vergleich  
von Sprachmodellen

# Sprachmodelle



But the cat came back,  
he could not stay no  
longer, Yes the cat came  
back the very next day,  
the cat came back -  
thought she were a goner,  
But the cat came back for  
it would not stay away.

# Sprachmodelle

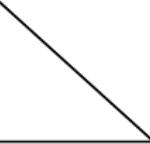


But the cat came back,  
he could not stay no  
longer, Yes the cat came  
back the very next day,  
the cat came back -  
thought she were a goner,  
But the cat came back for  
it would not stay away.

## Wort

But  
the  
cat  
came  
back  
he  
could  
not  
stay  
no  
longer  
Yes

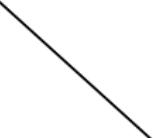
# Sprachmodelle



But the cat came back,  
he could not stay no  
longer, Yes the cat came  
back the very next day,  
the cat came back -  
thought she were a goner,  
But the cat came back for  
it would not stay away.

Wort	Häufigkeit
But	2
the	5
cat	4
came	4
back	4
he	1
could	1
not	2
stay	2
no	1
longer	1
Yes	1

# Sprachmodelle



But the cat came back,  
he could not stay no  
longer, Yes the cat came  
back the very next day,  
the cat came back -  
thought she were a goner,  
But the cat came back for  
it would not stay away.

Wort	Häufigkeit	Wahrscheinlichkeit
But	2	0.05
the	5	0.125
cat	4	0.1
came	4	0.1
back	4	0.1
he	1	0.025
could	1	0.025
not	2	0.05
stay	2	0.05
no	1	0.025
longer	1	0.025
Yes	1	0.025

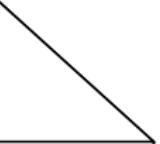
# Sprachmodelle

But the cat came back,  
he could not stay no  
longer, Yes the cat came  
back the very next day,  
the cat came back -  
thought she were a goner,  
But the cat came back for  
it would not stay away.

Wort	Häufigkeit	Wahrscheinlichkeit
But	2	0.05
the	5	0.125
cat	4	0.1
came	4	0.1
back	4	0.1
he	1	0.025
could	1	0.025
not	2	0.05
stay	2	0.05
no	1	0.025
longer	1	0.025
Yes	1	0.025

M

# Sprachmodelle



But the cat came back,  
he could not stay no  
longer, Yes the cat came  
back the very next day,  
the cat came back -  
thought she were a goner,  
But the cat came back for  
it would not stay away.

Wort	Häufigkeit	Wahrscheinlichkeit
But	2	0.05
the	5	0.125
cat	4	0.1
came	4	0.1
back	4	0.1
he	1	0.025
could	1	0.025
not	2	0.05
stay	2	0.05
no	1	0.025
longer	1	0.025
Yes	1	0.025

M

$$\Pr(\text{But the cat} \mid M) = \Pr(\text{But}) \Pr(\text{the}) \Pr(\text{cat}) = 0.05 \times 0.125 \times 0.1 = 0.000625$$

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

# Unsere Verfahren zur Datierung

# Google-Books-N-Gramm-Korpus

## ■ Charakteristik

- 209 Jahre: 1800 bis 2008
- 6 000 Bücher pro Jahr
- Häufigkeiten von Wort-N-Grammen
- Unbekannte Genres

## ■ Sprachmodelle $M_t$ ,

$$1800 \leq t \leq 2009$$



N-Gramm	Jahr	Häufigkeit
Christmasday	1800	1
Christmasday	1801	1
Christmasday	1802	2
Christmasday	1804	5
Christmasday	1805	3
Christmasday	1806	4
Christmasday	1807	14
Christmasday	1808	8
Christmasday	1809	2
Christmasday	1810	10
Christmasday	1811	6
Christmasday	1812	5
Christmasday	1813	9
Christmasday	1814	7
Christmasday	1815	3
Christmasday	1816	9
Christmasday	1817	11
Christmasday	1818	13
Christmasday	1819	10
Christmasday	1820	7
Christmasday	1821	10
Christmasday	1822	8
Christmasday	1823	18
Christmasday	1824	24
Christmasday	1825	4

# Rankingfunktionen

- Produktionswahrscheinlichkeit: Erzeugen von Wortsequenzen
  - 1 Schätzung der Produktionswahrscheinlichkeiten
  - 2 Absteigendes Ranking
- Kullback-Leibler-Divergenz: Vergleich von Sprachmodellen
  - 1 Berechnung der KL-Divergenzen
  - 2 Aufsteigendes Ranking

# Rankingfunktionen

Produktionswahrscheinlichkeit

Faust 1808

Ihr naht euch wieder,  
schwankende Gestalten,  
Die früh sich einst dem  
trüben Blick gezeigt.  
Versuch ich wohl, euch  
diesmal festzuhalten?  
Fühl ich mein Herz noch  
jenem Wahn geneigt?

Rang	$M_t$	Wert
1	1812	1.00
2	1809	0.89
3	1810	0.67
4	1811	0.59
5	1808	0.40
6	1813	0.39
7	1815	0.38
8	1817	0.35
9	1820	0.33
10	1825	0.28

Kullback-Leibler-Divergenz

Rang	$M_t$	Wert
1	1810	0.99
2	1808	1.13
3	1812	1.35
4	1806	1.36
5	1816	2.03
6	1820	2.17
7	1819	2.46
8	1828	2.73
9	1807	2.99
10	1826	3.05

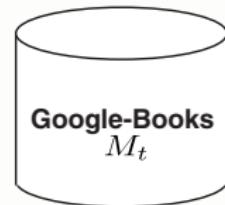
# Rankingfunktionen



Rang	$M_t$	Wert
1	1812	1.00
2	1809	0.89
3	1810	0.67
4	1811	0.59
5	1808	0.40
6	1813	0.39
7	1815	0.38
8	1817	0.35
9	1820	0.33
10	1825	0.28

$$\Pr(w_1 \dots w_n | M_t) = \prod_{k=1}^n \Pr(w_k | M_t)$$

Rang	$M_t$	Wert
1	1810	0.99
2	1808	1.13
3	1812	1.35
4	1806	1.36
5	1816	2.03
6	1820	2.17
7	1819	2.46
8	1828	2.73
9	1807	2.99
10	1826	3.05

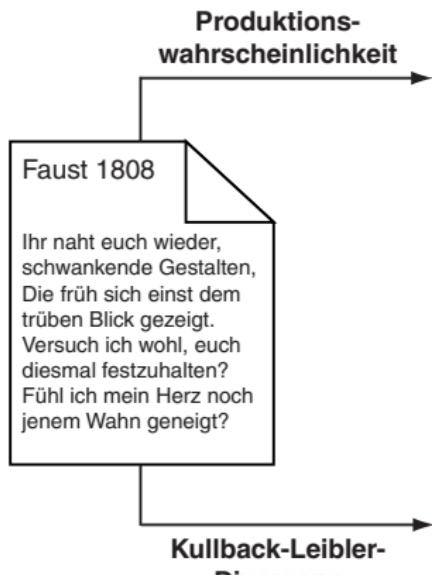


$$KL(M_d || M_t) = \sum_{w \in V} \Pr(w | M_d) \log \frac{\Pr(w | M_d)}{\Pr(w | M_t)}$$

# Klassifizierer

- Klassifizierer auf Basis des Top-10-Rankings
  - Top-1
  - Zufall Top-10
  - Durchschnitt Top-10
  - Mitte Top-10
- Baseline
  - Zufälliges Jahr
  - Festes Jahr

# Klassifizierer



Rang	$M_t$	Wert
1	1812	1.00
2	1809	0.89
3	1810	0.67
4	1811	0.59
5	1808	0.40
6	1813	0.39
7	1815	0.38
8	1817	0.35
9	1820	0.33
10	1825	0.28

Klassifizierer	Datum
Top-1	1812
Zufall Top-10	1817
Durchschnitt Top-10	1814
Mitte Top-10	1812
Zufälliges Jahr	1930
Festes Jahr	1904

Rang	$M_t$	Wert
1	1810	0.99
2	1808	1.13
3	1812	1.35
4	1806	1.36
5	1816	2.03
6	1820	2.17
7	1819	2.46
8	1828	2.73
9	1807	2.99
10	1826	3.05

Klassifizierer	Datum
Top-1	1810
Zufall Top-10	1828
Durchschnitt Top-10	1815
Mitte Top-10	1816
Zufälliges Jahr	1912
Festes Jahr	1904

# Evaluierung

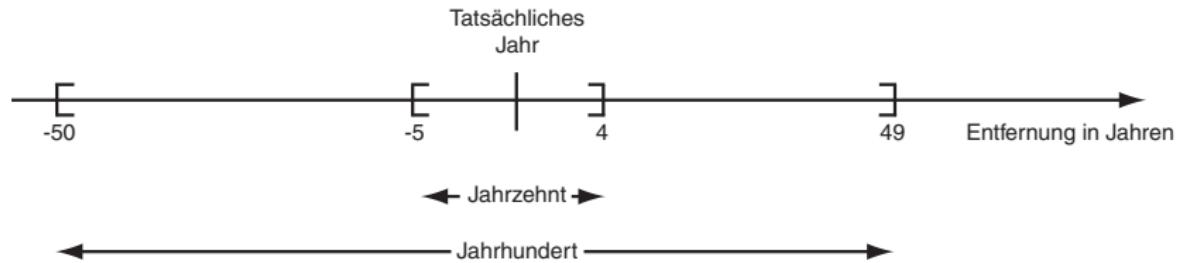
# Testkorpora

- Bücher Projekt Gutenberg (1808–1970)
  - Artikel aus Los Angeles Times und Financial Times (1989–94)
  - Künstliche Dokumente aus Sprachmodellen der Google-Books
- 
- Pro Jahr 3 Texte mit je 100, 500, 1 000, 5 000, 10 000, 100 000 Wörtern

# Bewertungsmaß

Genauigkeit für korrekt datierte

- Jahre
- Jahrzehnte
- Jahrhunderte

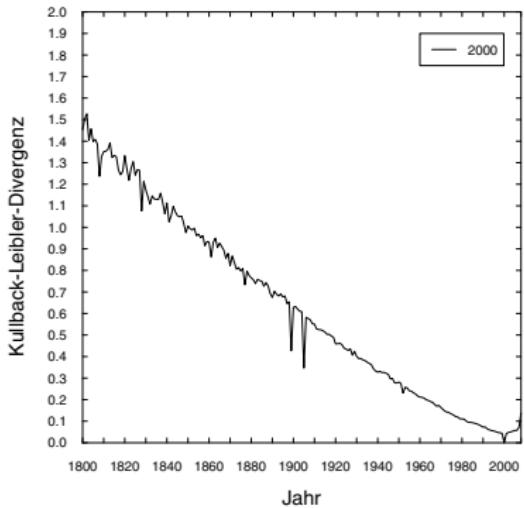
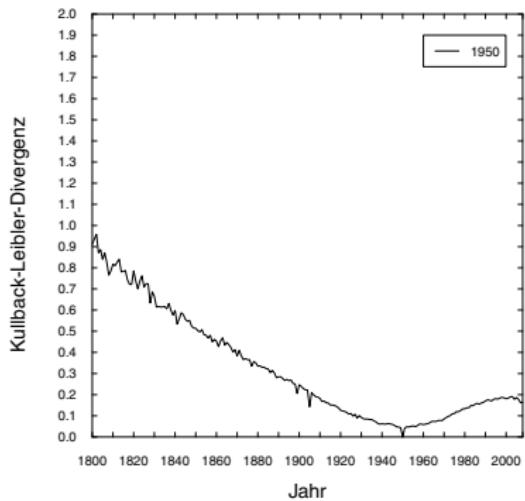


# Experimente

- Eignung von Google-Books
- Beste Datierung von Rankingfunktion und Klassifizierer
- Eignung unserer Verfahren
- Rang des tatsächliche Jahres
- Maximale Obergrenze der Datierungsgenauigkeit

# Eignung von Google-Books

## ■ Untersuchung der Sprachmodelle



# Beste Datierung von Rankingfunktion und Klassifizierer

Genauigkeit (%)	Testkorpus	Künstliche Texte		Zeitungsaufgaben		Bücher	
		PW	KL	PW	KL	PW	KL
Bewertungsmaß	Jahr	95.37*	100*	11.11	6.67	1.63	1
	Klassifizierer	T1	T1	DT10	BL	ZT10	FJ
	Jahrzehnt	95.37*	100*	38.89	22.22	8.64*	8.31*
	Klassifizierer	T1	T1	DT10	DT10	ZT10	FJ
	Jahrhundert	95.37*	100*	94.44*	100*	76.08*	76.08*
	Klassifizierer	T1	T1	T1	T1	FJ	FJ
				ZT10			
				DT10			
				MT10			

\* signifikante Verbesserungen in der Datierungsgenauigkeit bzgl. der Baseline ZJ

PW : Produktionswahrscheinlichkeit

KL : Kullback-Leibler-Divergenz

T1 : Top-1

ZT10: Zufälliges Jahr Top-10

DT10: Durchschnitt Top-10

MT10: Mitte Top-10

ZJ : Zufälliges Jahr

FJ: Festes-Jahr

# Zusammenfassung und Ausblick

- Eignung von Google-Books
- Erfolgreiche Datierung von künstlichen Texten
- Zukünftige Forschung
  - Trainingskorpora mit bekannten Genres
  - Evaluierungsmaß
  - Filtern des Vokabulars
  - Sprachmodelle
    - höherer Ordnung
    - auf Buchstabenbasis
    - unterschiedlicher zeitlicher Auflösung
  - Klassifizierer

# Zusammenfassung und Ausblick

- Eignung von Google-Books
- Erfolgreiche Datierung von künstlichen Texten
- Zukünftige Forschung
  - Trainingskorpora mit bekannten Genres
  - Evaluierungsmaß
  - Filtern des Vokabulars
  - Sprachmodelle
    - höherer Ordnung
    - auf Buchstabenbasis
    - unterschiedlicher zeitlicher Auflösung
  - Klassifizierer

Danke für die Aufmerksamkeit!