Nutzung maschinellen Lernens zur Extraktion von Paragraphen aus PDF-Dokumenten

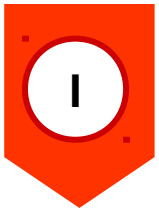
Albert-Ludwigs-Universität zu Freiburg

13.09.2016

Maximilian Dippel

max.dippel@tf.uni-freiburg.de

Überblick



Einführung

- Problemstellung und Motivation
- Ziele

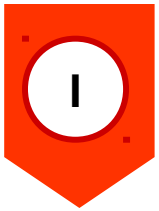
Ausführung

- Nutzung von maschinellem Lernen
- Erzeugung von Trainingsdaten

Analyse

- Laufzeit der Algorithmen
- Qualität der Ergebnisse

Überblick



Einführung

- Problemstellung und Motivation
- Ziele

Ausführung

- Nutzung von maschinellem Lernen
- Erzeugung von Trainingsdaten

Analyse

- Laufzeit der Algorithmen
- Qualität der Ergebnisse

Problemstellung und Motivation

Reducing quasi-ergodicity in a double well potential by Tsallis Monte Carlo simulation

Masao Iwamatsu[†]*and Yutaka Okabe[†]
*Department of Computer Engineering, Hiroshima City University
Hiroshima 731-3194, Japan
and
[†]Department of Physics, Tokyo Metropolitan University
Hachioji, Tokyo 192-0397, Japan

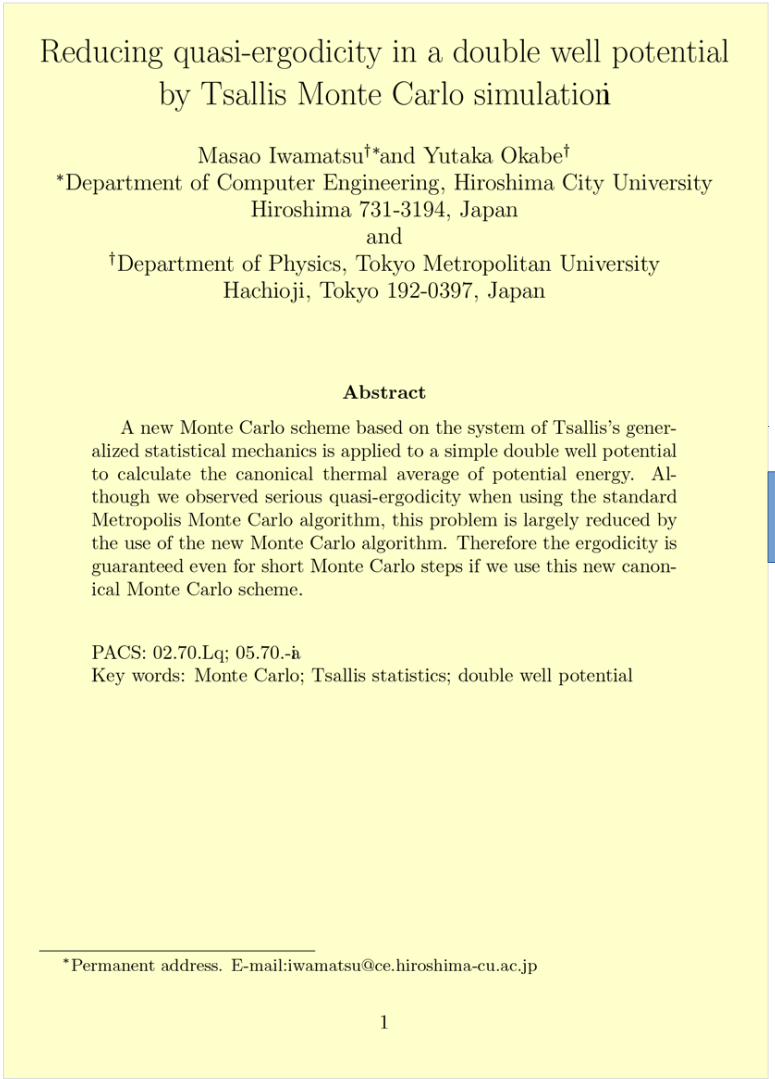
Abstract

A new Monte Carlo scheme based on the system of Tsallis's generalized statistical mechanics is applied to a simple double well potential to calculate the canonical thermal average of potential energy. Although we observed serious quasi-ergodicity when using the standard Metropolis Monte Carlo algorithm, this problem is largely reduced by the use of the new Monte Carlo algorithm. Therefore the ergodicity is guaranteed even for short Monte Carlo steps if we use this new canonical Monte Carlo scheme.

PACS: 02.70.Lq; 05.70.-a
Key words: Monte Carlo; Tsallis statistics; double well potential

*Permanent address. E-mail:iwamatsu@ce.hiroshima-cu.ac.jp

Problemstellung und Motivation



Reducing quasi-ergodicity in a double well potential by Tsallis Monte Carlo simulation

Masao Iwamatsu[†]*and Yutaka Okabe[†]
*Department of Computer Engineering, Hiroshima City University
Hiroshima 731-3194, Japan
and
[†]Department of Physics, Tokyo Metropolitan University
Hachioji, Tokyo 192-0397, Japan

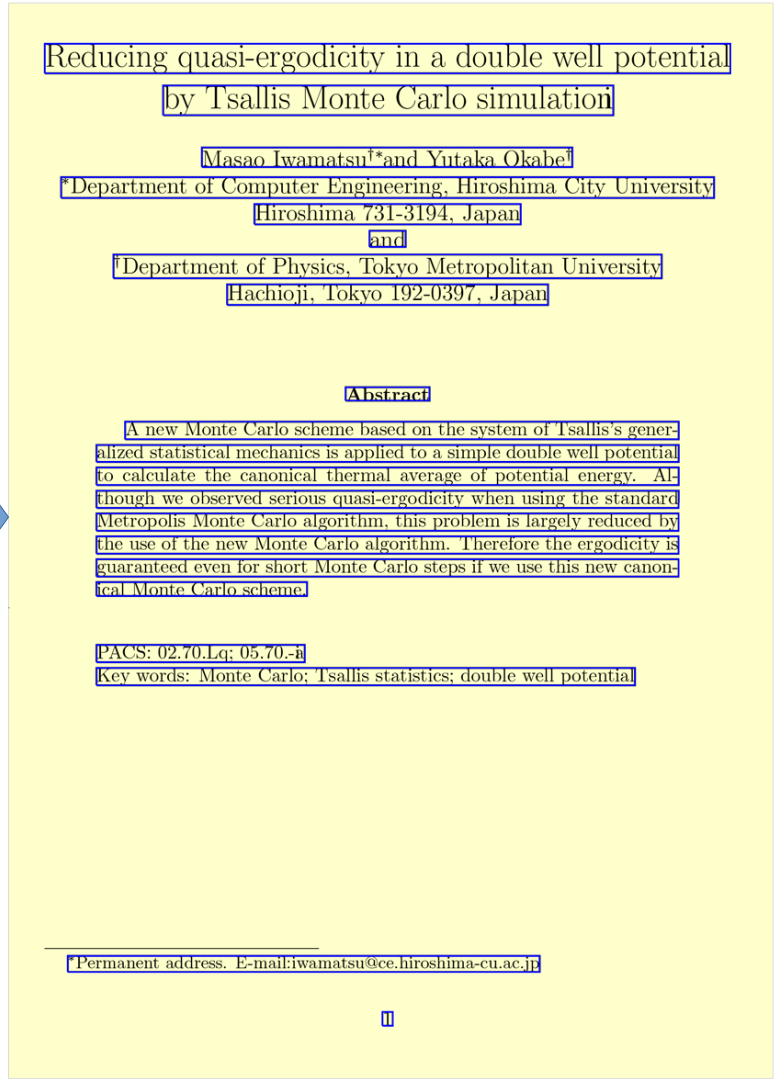
Abstract

A new Monte Carlo scheme based on the system of Tsallis's generalized statistical mechanics is applied to a simple double well potential to calculate the canonical thermal average of potential energy. Although we observed serious quasi-ergodicity when using the standard Metropolis Monte Carlo algorithm, this problem is largely reduced by the use of the new Monte Carlo algorithm. Therefore the ergodicity is guaranteed even for short Monte Carlo steps if we use this new canonical Monte Carlo scheme.

PACS: 02.70.Lq; 05.70.-a
Key words: Monte Carlo; Tsallis statistics; double well potential

*Permanent address. E-mail:iwamatsu@ce.hiroshima-cu.ac.jp

Converter



Reducing quasi-ergodicity in a double well potential by Tsallis Monte Carlo simulation

Masao Iwamatsu[†]*and Yutaka Okabe[†]
*Department of Computer Engineering, Hiroshima City University
Hiroshima 731-3194, Japan
and
[†]Department of Physics, Tokyo Metropolitan University
Hachioji, Tokyo 192-0397, Japan

Abstract

A new Monte Carlo scheme based on the system of Tsallis's generalized statistical mechanics is applied to a simple double well potential to calculate the canonical thermal average of potential energy. Although we observed serious quasi-ergodicity when using the standard Metropolis Monte Carlo algorithm, this problem is largely reduced by the use of the new Monte Carlo algorithm. Therefore the ergodicity is guaranteed even for short Monte Carlo steps if we use this new canonical Monte Carlo scheme.

PACS: 02.70.Lq; 05.70.-a
Key words: Monte Carlo; Tsallis statistics; double well potential

*Permanent address. E-mail:iwamatsu@ce.hiroshima-cu.ac.jp

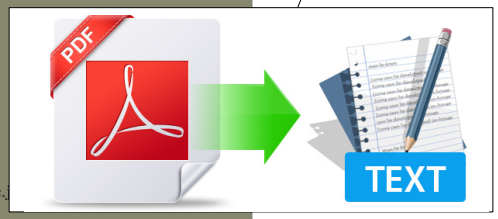
1

Problemstellung und Motivation

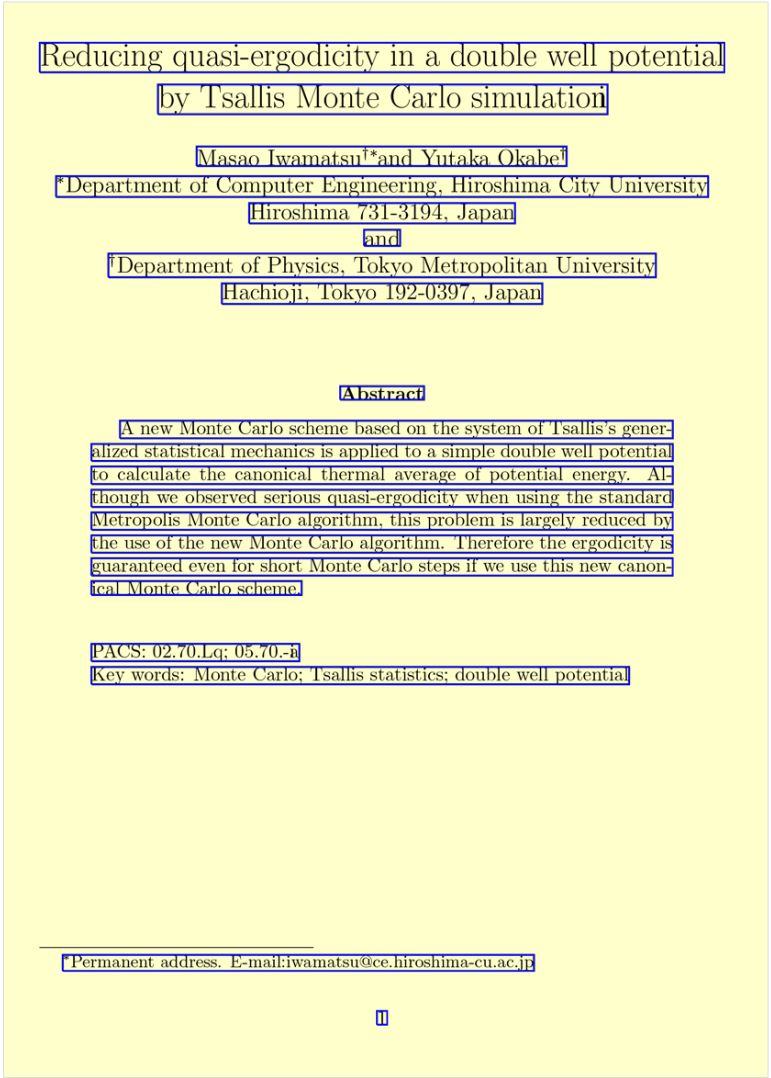
Input

Converter

Output



Problemstellung und Motivation



Reducing quasi-ergodicity in a double well potential by Tsallis Monte Carlo simulation

Masao Iwamatsu[†]*and Yutaka Okabe[†]
*Department of Computer Engineering, Hiroshima City University
Hiroshima 731-3194, Japan
and
[†]Department of Physics, Tokyo Metropolitan University
Hachioji, Tokyo 192-0397, Japan

Abstract

A new Monte Carlo scheme based on the system of Tsallis's generalized statistical mechanics is applied to a simple double well potential to calculate the canonical thermal average of potential energy. Although we observed serious quasi-ergodicity when using the standard Metropolis Monte Carlo algorithm, this problem is largely reduced by the use of the new Monte Carlo algorithm. Therefore the ergodicity is guaranteed even for short Monte Carlo steps if we use this new canonical Monte Carlo scheme.

PACS: 02.70.Lq; 05.70.-a
Key words: Monte Carlo; Tsallis statistics; double well potential

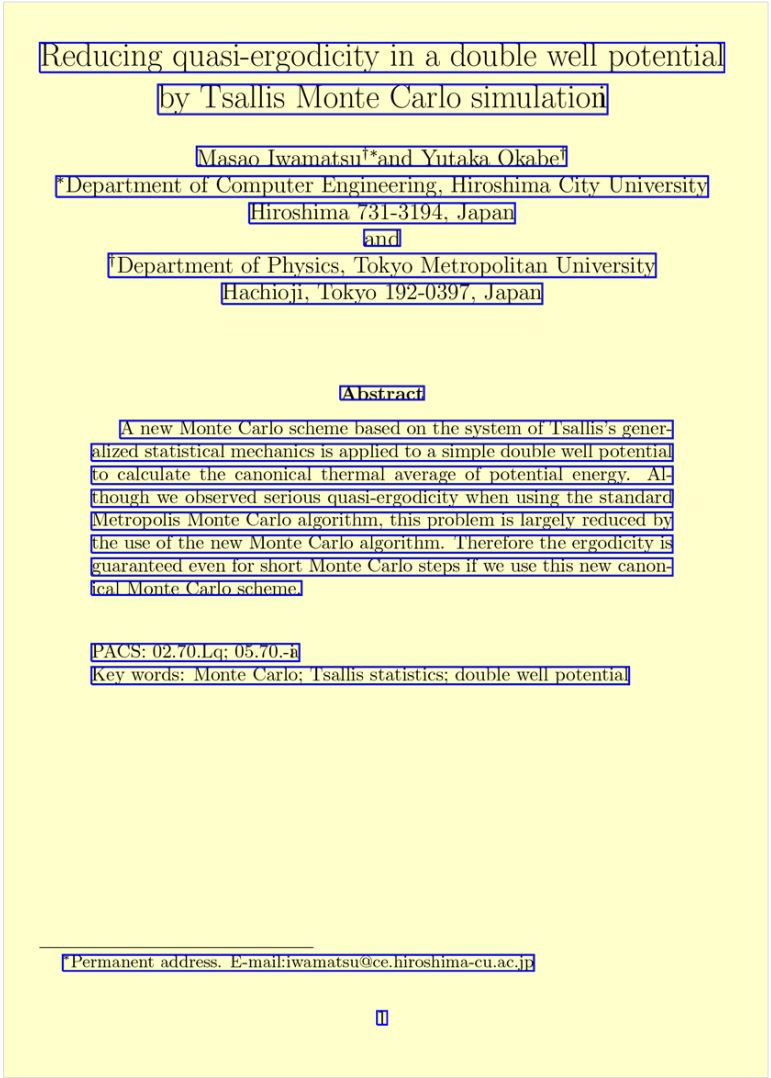
*Permanent address. E-mail:iwamatsu@ce.hiroshima-cu.ac.jp

1

Wir haben den reinen Text …

… aber sonst keine Informationen

Problemstellung und Motivation



Reducing quasi-ergodicity in a double well potential by Tsallis Monte Carlo simulation

Masao Iwamatsu[†]*and Yutaka Okabe[†]
[*]Department of Computer Engineering, Hiroshima City University
Hiroshima 731-3194, Japan
and
[†]Department of Physics, Tokyo Metropolitan University
Hachioji, Tokyo 192-0397, Japan

Abstract

A new Monte Carlo scheme based on the system of Tsallis's generalized statistical mechanics is applied to a simple double well potential to calculate the canonical thermal average of potential energy. Although we observed serious quasi-ergodicity when using the standard Metropolis Monte Carlo algorithm, this problem is largely reduced by the use of the new Monte Carlo algorithm. Therefore the ergodicity is guaranteed even for short Monte Carlo steps if we use this new canonical Monte Carlo scheme.

PACS: 02.70.Lq; 05.70.-a
Key words: Monte Carlo; Tsallis statistics; double well potential

*Permanent address. E-mail:iwamatsu@ce.hiroshima-cu.ac.jp

1

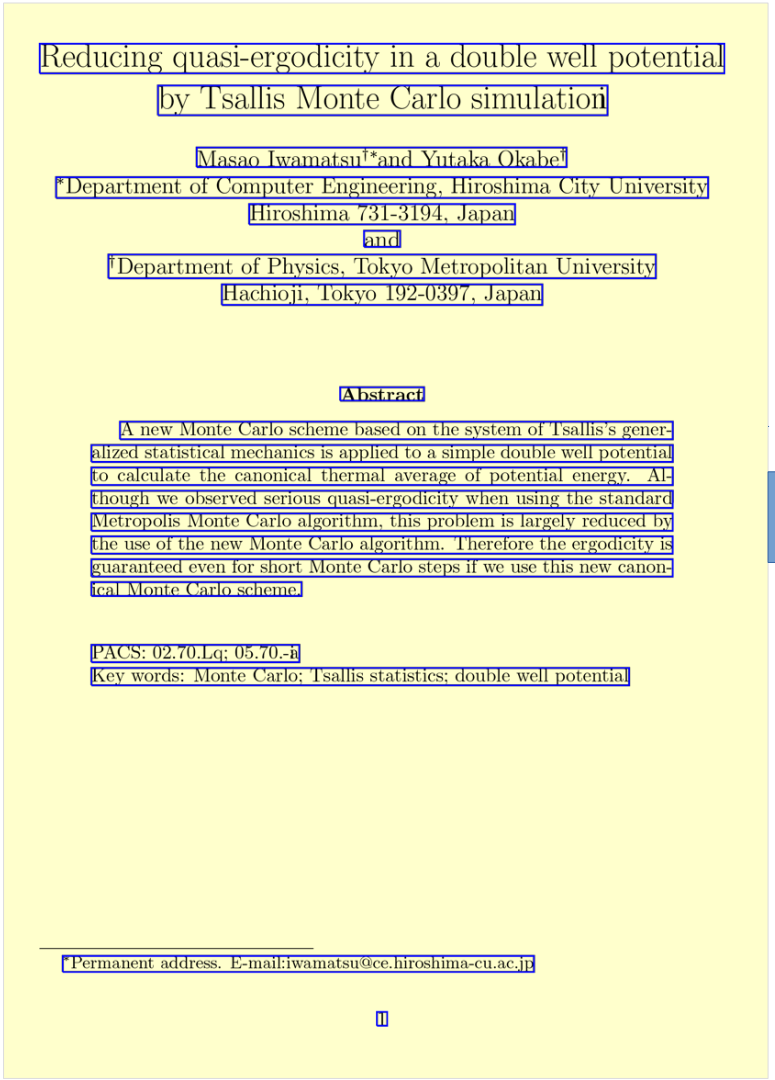
Gedankenspiel – Literaturrecherche

Vielleicht wollen wir ...

- … den ganzen Absatz bei Treffern
- … Bildunterschriften filtern
- … den Abstract extrahieren
- … alle Titel auflisten

→ **Strukturierung des reinen Textes**

Problemstellung und Motivation



Reducing quasi-ergodicity in a double well potential by Tsallis Monte Carlo simulation

Masao Iwamatsu[†]*and Yutaka Okabe[†]
*Department of Computer Engineering, Hiroshima City University
Hiroshima 731-3194, Japan
and
[†]Department of Physics, Tokyo Metropolitan University
Hachioji, Tokyo 192-0397, Japan

Abstract

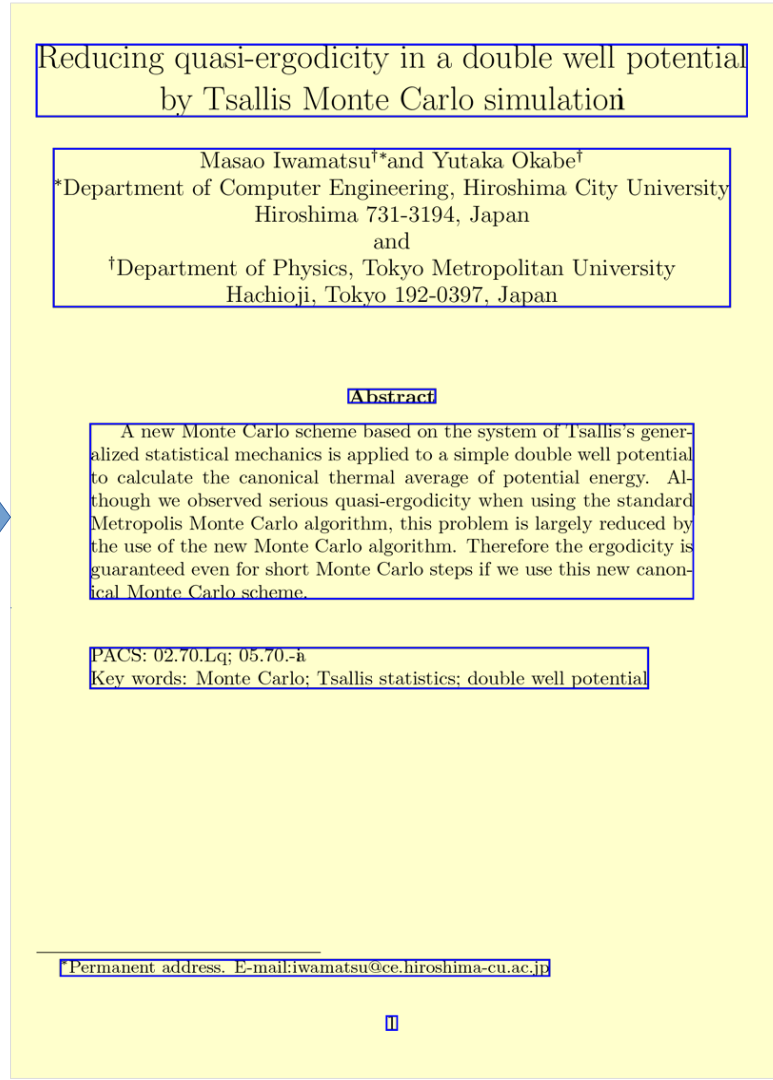
A new Monte Carlo scheme based on the system of Tsallis's generalized statistical mechanics is applied to a simple double well potential to calculate the canonical thermal average of potential energy. Although we observed serious quasi-ergodicity when using the standard Metropolis Monte Carlo algorithm, this problem is largely reduced by the use of the new Monte Carlo algorithm. Therefore the ergodicity is guaranteed even for short Monte Carlo steps if we use this new canonical Monte Carlo scheme.

PACS: 02.70.Lq; 05.70.-a
Key words: Monte Carlo; Tsallis statistics; double well potential

*Permanent address. E-mail:iwamatsu@ce.hiroshima-cu.ac.jp

1

Converter



Reducing quasi-ergodicity in a double well potential by Tsallis Monte Carlo simulation

Masao Iwamatsu[†]*and Yutaka Okabe[†]
*Department of Computer Engineering, Hiroshima City University
Hiroshima 731-3194, Japan
and
[†]Department of Physics, Tokyo Metropolitan University
Hachioji, Tokyo 192-0397, Japan

Abstract

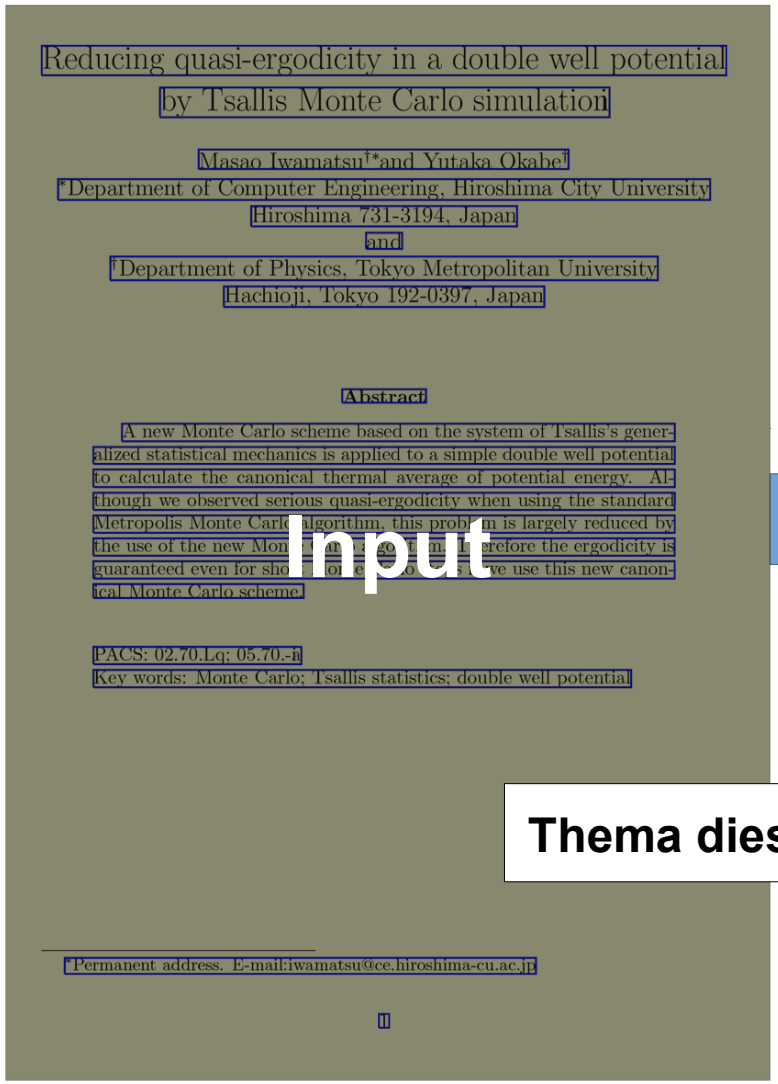
A new Monte Carlo scheme based on the system of Tsallis's generalized statistical mechanics is applied to a simple double well potential to calculate the canonical thermal average of potential energy. Although we observed serious quasi-ergodicity when using the standard Metropolis Monte Carlo algorithm, this problem is largely reduced by the use of the new Monte Carlo algorithm. Therefore the ergodicity is guaranteed even for short Monte Carlo steps if we use this new canonical Monte Carlo scheme.

PACS: 02.70.Lq; 05.70.-a
Key words: Monte Carlo; Tsallis statistics; double well potential

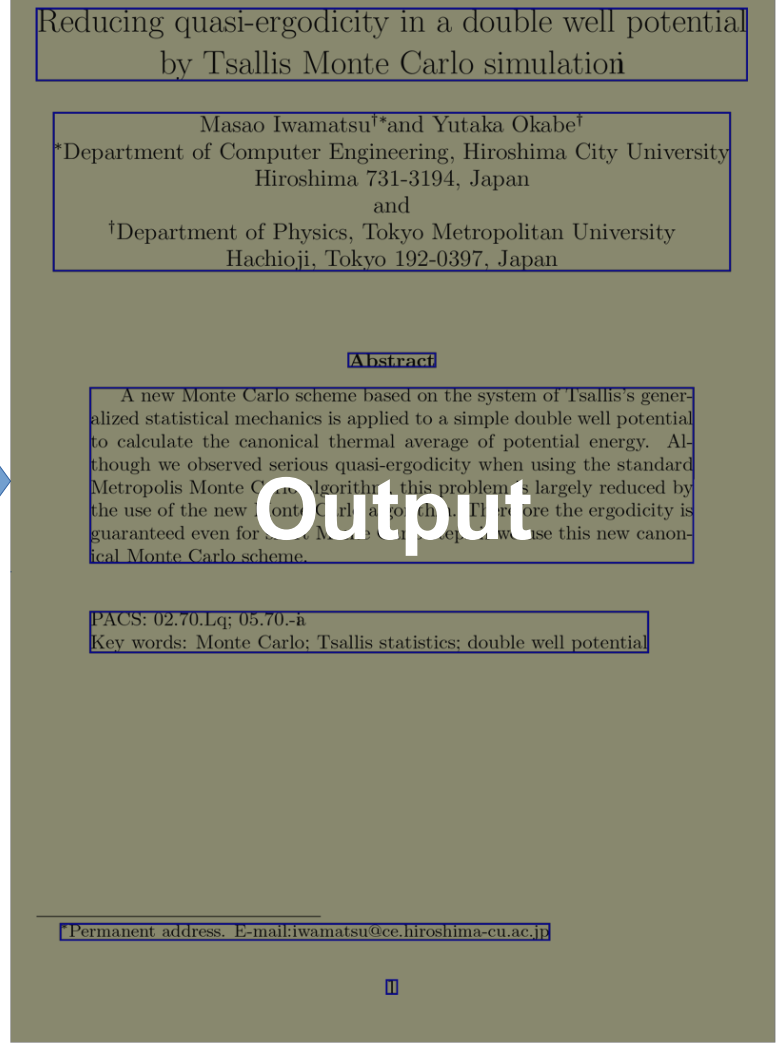
*Permanent address. E-mail:iwamatsu@ce.hiroshima-cu.ac.jp

1

Problemstellung und Motivation



Converter

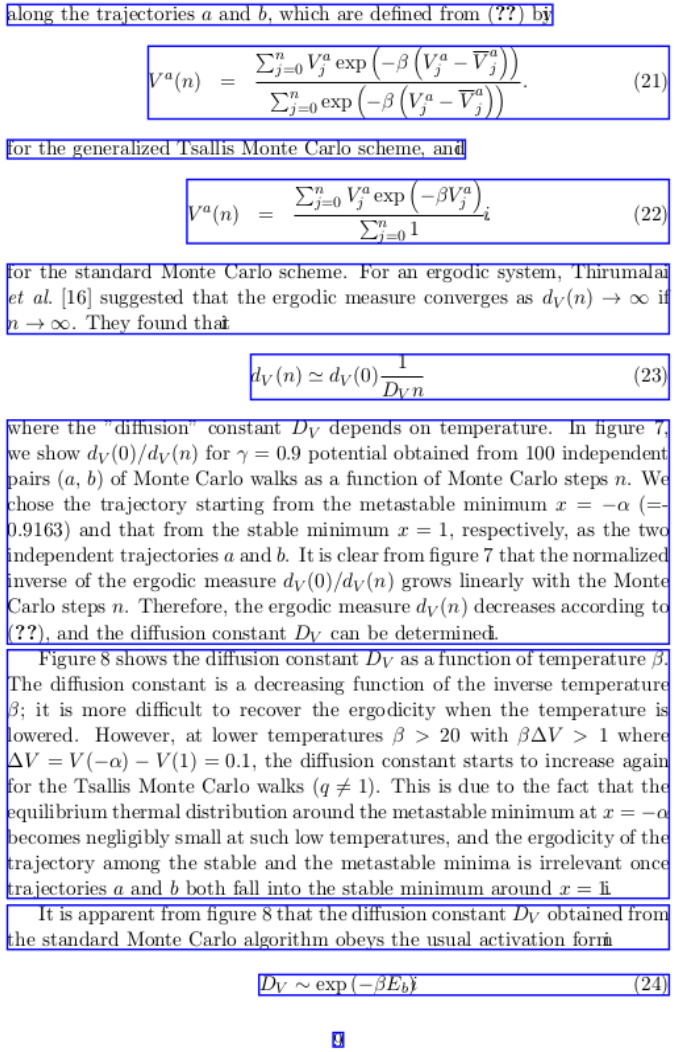


Thema dieser Arbeit

Ziele

Was sind Paragraphen?

- Text-Abschnitte
- Überschriften und Titeln
- Referenzen
- ...



along the trajectories a and b, which are defined from (??) by

$$V^a(n) = \frac{\sum_{j=0}^{n} V_j^a \exp\left(-\beta\left(V_j^a - \overline{V}_j^a\right)\right)}{\sum_{j=0}^{n} \exp\left(-\beta\left(V_j^a - \overline{V}_j^a\right)\right)}. \quad (21)$$

for the generalized Tsallis Monte Carlo scheme, and

$$V^a(n) = \frac{\sum_{j=0}^{n} V_j^a \exp\left(-\beta V_j^a\right)}{\sum_{j=0}^{n} 1}, \quad (22)$$

for the standard Monte Carlo scheme. For an ergodic system, Thirumalai *et al.* [16] suggested that the ergodic measure converges as $d_V(n) \to \infty$ if $n \to \infty$. They found that

$$d_V(n) \simeq d_V(0) \frac{1}{D_V n} \quad (23)$$

where the "diffusion" constant D_V depends on temperature. In figure 7, we show $d_V(0)/d_V(n)$ for $\gamma = 0.9$ potential obtained from 100 independent pairs (a, b) of Monte Carlo walks as a function of Monte Carlo steps n. We chose the trajectory starting from the metastable minimum $x = -\alpha$ (=-0.9163) and that from the stable minimum $x = 1$, respectively, as the two independent trajectories a and b. It is clear from figure 7 that the normalized inverse of the ergodic measure $d_V(0)/d_V(n)$ grows linearly with the Monte Carlo steps n. Therefore, the ergodic measure $d_V(n)$ decreases according to (??), and the diffusion constant D_V can be determined.

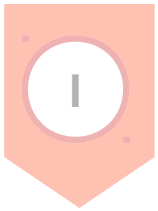
Figure 8 shows the diffusion constant D_V as a function of temperature β. The diffusion constant is a decreasing function of the inverse temperature β; it is more difficult to recover the ergodicity when the temperature is lowered. However, at lower temperatures $\beta > 20$ with $\beta\Delta V > 1$ where $\Delta V = V(-\alpha) - V(1) = 0.1$, the diffusion constant starts to increase again for the Tsallis Monte Carlo walks ($q \neq 1$). This is due to the fact that the equilibrium thermal distribution around the metastable minimum at $x = -\alpha$ becomes negligibly small at such low temperatures, and the ergodicity of the trajectory among the stable and the metastable minima is irrelevant once trajectories a and b both fall into the stable minimum around $x = 1$.

It is apparent from figure 8 that the diffusion constant D_V obtained from the standard Monte Carlo algorithm obeys the usual activation form

$$D_V \sim \exp\left(-\beta E_b\right) \quad (24)$$

9

Überblick



Einführung

- Problemstellung und Motivation
- Ziele

Ausführung

- Nutzung von maschinellem Lernen
- Erzeugung von Trainingsdaten

Analyse

- Laufzeit der Algorithmen
- Qualität der Ergebnisse

Nutzung von maschinellem Lernen

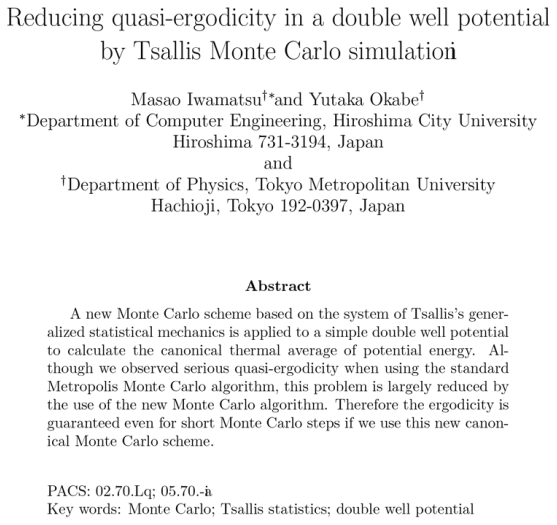
Ausgangspunkt

- Tabelle im TSV-Format
- Wörter und Zeilen
- Zusätzliche Informationen

Nutzung von maschinellem Lernen

Ausgangspunkt

- Tabelle im TSV-Format
- Wörter und Zeilen
- Zusätzliche Informationen



Reducing quasi-ergodicity in a double well potential
by Tsallis Monte Carlo simulation

Masao Iwamatsu[†*] and Yutaka Okabe[†]
*Department of Computer Engineering, Hiroshima City University
Hiroshima 731-3194, Japan
and
[†]Department of Physics, Tokyo Metropolitan University
Hachioji, Tokyo 192-0397, Japan

Abstract

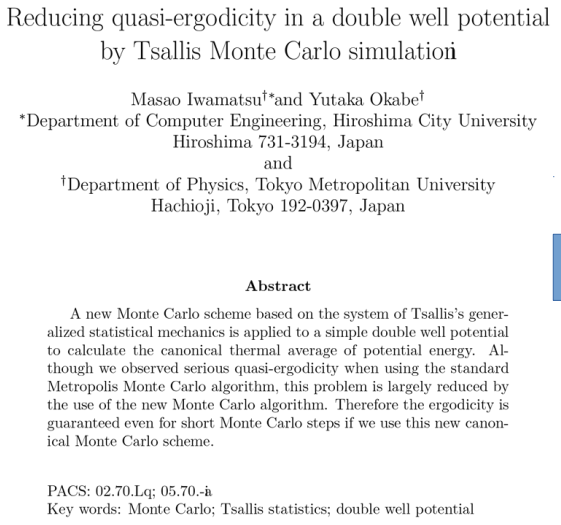
A new Monte Carlo scheme based on the system of Tsallis's generalized statistical mechanics is applied to a simple double well potential to calculate the canonical thermal average of potential energy. Although we observed serious quasi-ergodicity when using the standard Metropolis Monte Carlo algorithm, this problem is largely reduced by the use of the new Monte Carlo algorithm. Therefore the ergodicity is guaranteed even for short Monte Carlo steps if we use this new canonical Monte Carlo scheme.

PACS: 02.70.Lq; 05.70.-a
Key words: Monte Carlo; Tsallis statistics; double well potential

Nutzung von maschinellem Lernen

Ausgangspunkt

- Tabelle im TSV-Format
- Wörter und Zeilen
- Zusätzliche Informationen



Reducing quasi-ergodicity in a double well potential by Tsallis Monte Carlo simulation

Masao Iwamatsu[†]*and Yutaka Okabe[†]
*Department of Computer Engineering, Hiroshima City University
Hiroshima 731-3194, Japan
and
[†]Department of Physics, Tokyo Metropolitan University
Hachioji, Tokyo 192-0397, Japan

Abstract

A new Monte Carlo scheme based on the system of Tsallis's generalized statistical mechanics is applied to a simple double well potential to calculate the canonical thermal average of potential energy. Although we observed serious quasi-ergodicity when using the standard Metropolis Monte Carlo algorithm, this problem is largely reduced by the use of the new Monte Carlo algorithm. Therefore the ergodicity is guaranteed even for short Monte Carlo steps if we use this new canonical Monte Carlo scheme.

PACS: 02.70.Lq; 05.70.-a
Key words: Monte Carlo; Tsallis statistics; double well potential

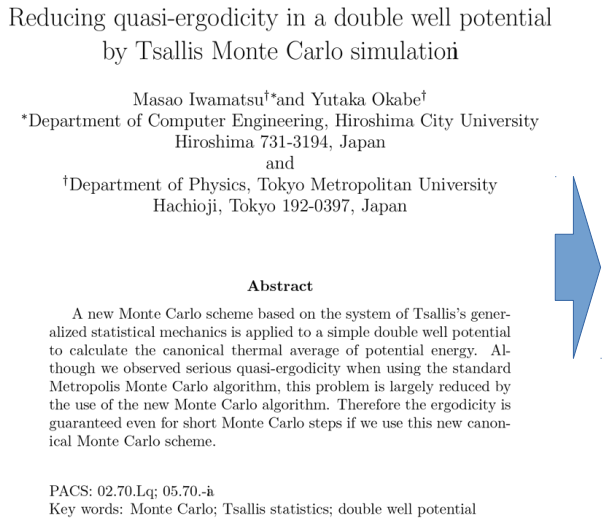
feature	text	page	minX	minX	maxY	minY	Font	...
line	Reducing quasi-ergodicity in a double	1	125	647	484	663	font-33	...
line	by Tsallis Monte Carlo simulationi	1	187	626	423	641	font-33	...
line	Masao Iwamatsu and Yutaka Okabe	1	208	598	401	609	font-32	...
line	*Department of Computer Engineerin	1	134	582	475	593	font-32	...
line	Hiroshima 731-3194, Japan	1	235	568	374	579	font-32	...
line	and	1	295	557	314	565	font-32	...
line	†Department of Physics, Tokyo Metro	1	161	540	448	553	font-32	...
line	Hachioji, Tokyo 192-0397, Japan	1	221	526	388	537	font-32	...
line	Abstract	1	283	476	327	483	font-4	...
line	A new Monte Carlo scheme based on	1	168	456	457	465	font-31	...

Nutzung von maschinellem Lernen

Ausgangspunkt

- Tabelle im TSV-Format
- Wörter und Zeilen
- Zusätzliche Informationen

FontSize, FontColor, Role



Reducing quasi-ergodicity in a double well potential by Tsallis Monte Carlo simulation

Masao Iwamatsu[†]*and Yutaka Okabe[†]
*Department of Computer Engineering, Hiroshima City University
Hiroshima 731-3194, Japan
and
[†]Department of Physics, Tokyo Metropolitan University
Hachioji, Tokyo 192-0397, Japan

Abstract

A new Monte Carlo scheme based on the system of Tsallis's generalized statistical mechanics is applied to a simple double well potential to calculate the canonical thermal average of potential energy. Although we observed serious quasi-ergodicity when using the standard Metropolis Monte Carlo algorithm, this problem is largely reduced by the use of the new Monte Carlo algorithm. Therefore the ergodicity is guaranteed even for short Monte Carlo steps if we use this new canonical Monte Carlo scheme.

PACS: 02.70.Lq; 05.70.-a
Key words: Monte Carlo; Tsallis statistics; double well potential

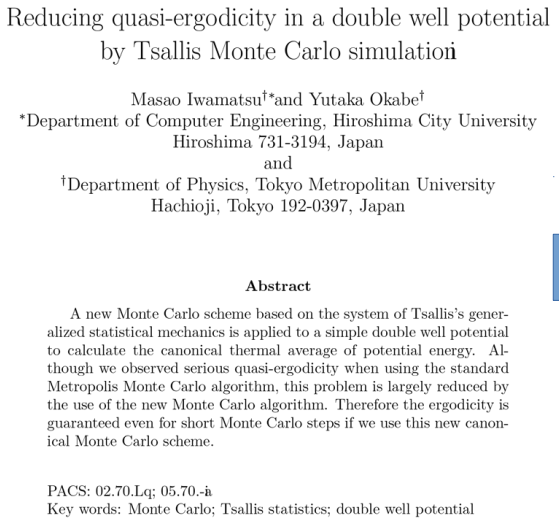
feature	text	page	minX	minX	maxY	minY	Font	...
line	Reducing quasi-ergodicity in a double	1	125	647	484	663	font-33	...
line	by Tsallis Monte Carlo simulationi	1	187	626	423	641	font-33	...
line	Masao Iwamatsu and Yutaka Okabe	1	208	598	401	609	font-32	...
line	*Department of Computer Engineerin	1	134	582	475	593	font-32	...
line	Hiroshima 731-3194, Japan	1	235	568	374	579	font-32	...
line	and	1	295	557	314	565	font-32	...
line	†Department of Physics, Tokyo Metro	1	161	540	448	553	font-32	...
line	Hachioji, Tokyo 192-0397, Japan	1	221	526	388	537	font-32	...
line	Abstract	1	283	476	327	483	font-4	...
line	A new Monte Carlo scheme based on	1	168	456	457	465	font-31	...

Nutzung von maschinellem Lernen

Ausgangspunkt

- Tabelle im TSV-Format
- Wörter und Zeilen
- Zusätzliche Informationen

FontSize, FontColor, Role



Reducing quasi-ergodicity in a double well potential by Tsallis Monte Carlo simulation

Masao Iwamatsu[†]*and Yutaka Okabe[†]
*Department of Computer Engineering, Hiroshima City University
Hiroshima 731-3194, Japan
and
[†]Department of Physics, Tokyo Metropolitan University
Hachioji, Tokyo 192-0397, Japan

Abstract

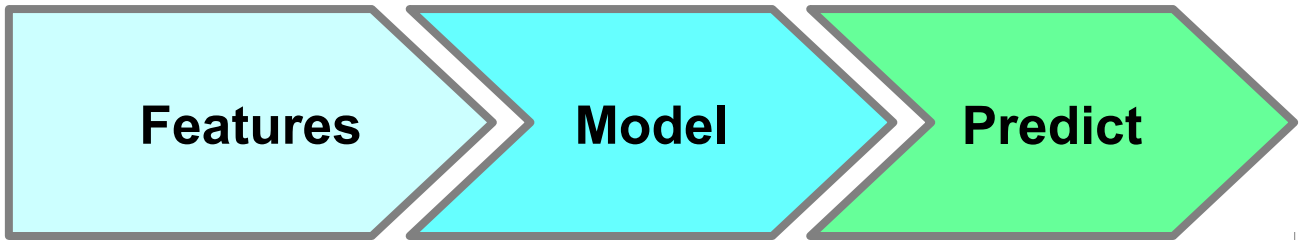
A new Monte Carlo scheme based on the system of Tsallis's generalized statistical mechanics is applied to a simple double well potential to calculate the canonical thermal average of potential energy. Although we observed serious quasi-ergodicity when using the standard Metropolis Monte Carlo algorithm, this problem is largely reduced by the use of the new Monte Carlo algorithm. Therefore the ergodicity is guaranteed even for short Monte Carlo steps if we use this new canonical Monte Carlo scheme.

PACS: 02.70.Lq; 05.70.-a
Key words: Monte Carlo; Tsallis statistics; double well potential

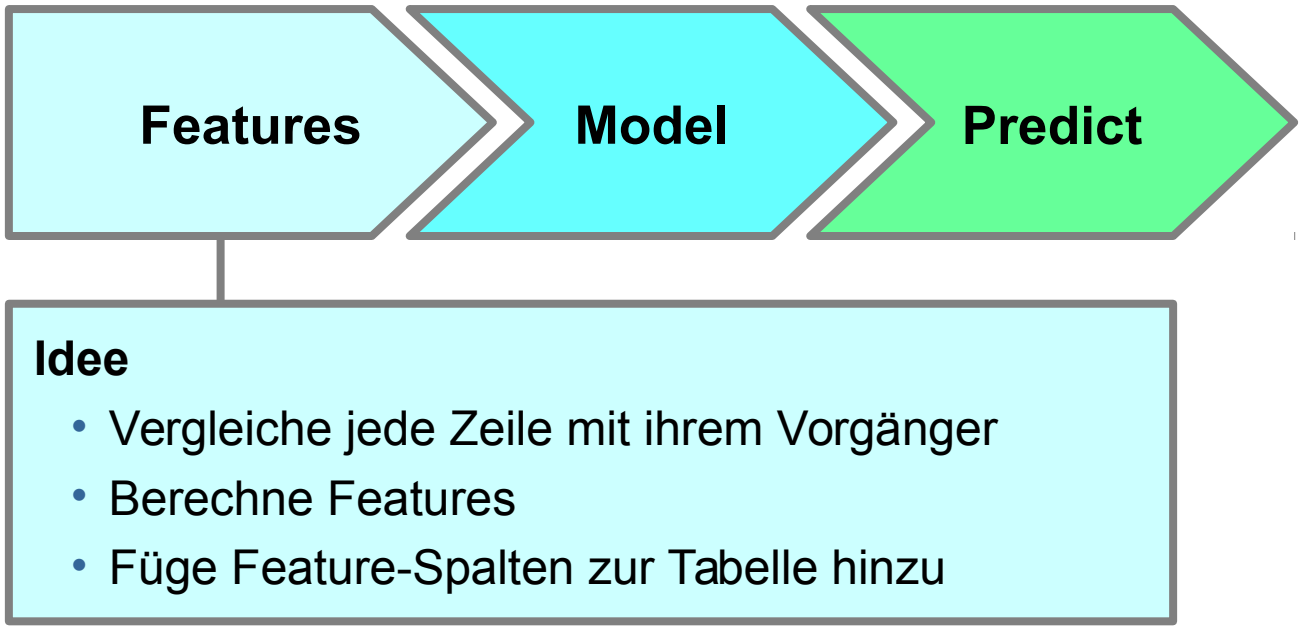
feature	text	page	minX	minX	maxY	minY	Font	...
line	Reducing quasi-ergodicity in a double	1	125	647	484	663	font-33	...
line	by Tsallis Monte Carlo simulationi	1	187	626	423	641	font-33	...
line	Masao Iwamatsu and Yutaka Okabe	1	208	598	401	609	font-32	...
line	*Department of Computer Engineerin	1	134	582	475	593	font-32	...
line	Hiroshima 731-3194, Japan	1	235	568	374	579	font-32	...
line	and	1	295	557	314	565	font-32	...
line	†Department of Physics, Tokyo Metro	1	161	540	448	553	font-32	...
line	Hachioji, Tokyo 192-0397, Japan	1	221	526	388	537	font-32	...
line	Abstract	1	283	476	327	483	font-4	...
line	A new Monte Carlo scheme based on	1	168	456	457	465	font-31	...

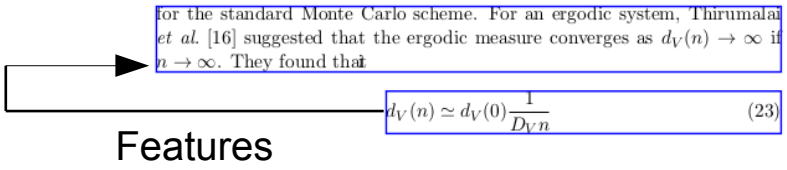
Wir interessieren uns für die *line*-Einträge und ihre Informationen

Nutzung von maschinellem Lernen

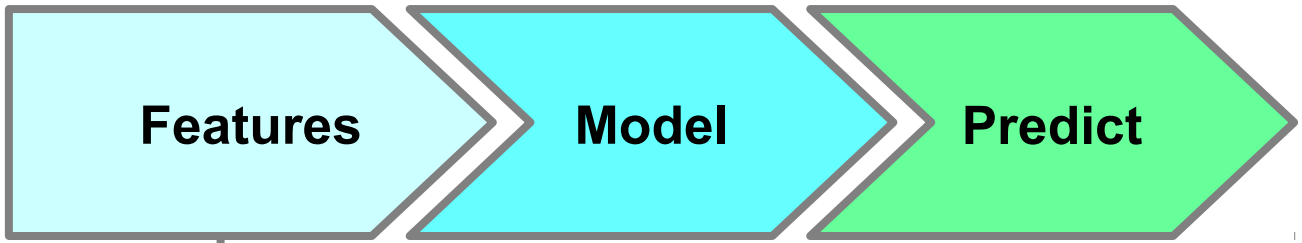


Nutzung von maschinellem Lernen





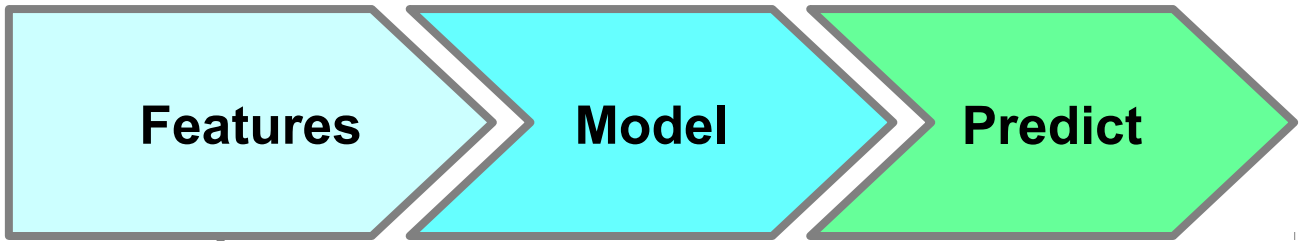
Nutzung von maschinellem Lernen



Auswahl 1/2

- DistanceToPreviousLine
- DifferenceInStartPositionToPreviousLine
- SameAverageFontAsPreviousLine
- SameConnectedFontAsPreviousLine
- SameAverageFontSizeAsPreviousLine
- SameConnectedFontSizeAsPreviousLine
- SameConnectedFontColorAsPreviousLine
- SpecialFontPreviousLine
- SameRoleAsPreviousLine

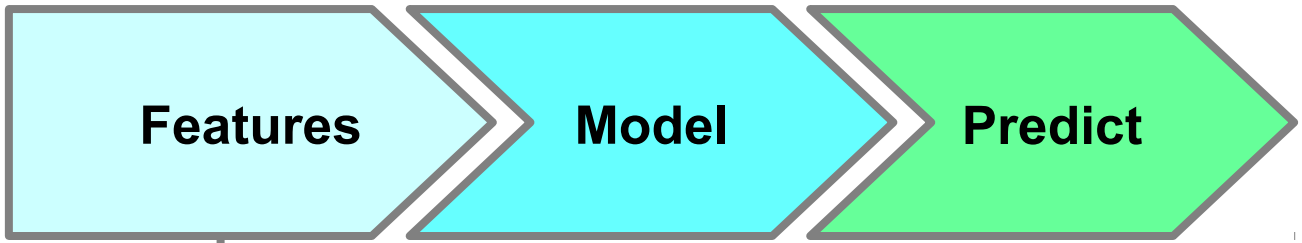
Nutzung von maschinellem Lernen



Auswahl 1/2

- **DistanceToPreviousLine**
- **DifferenceInStartPositionToPreviousLine**
- SameAverageFontAsPreviousLine
- SameConnectedFontAsPreviousLine
- SameAverageFontSizeAsPreviousLine
- SameConnectedFontSizeAsPreviousLine
- SameConnectedFontColorAsPreviousLine
- SpecialFontPreviousLine
- SameRoleAsPreviousLine

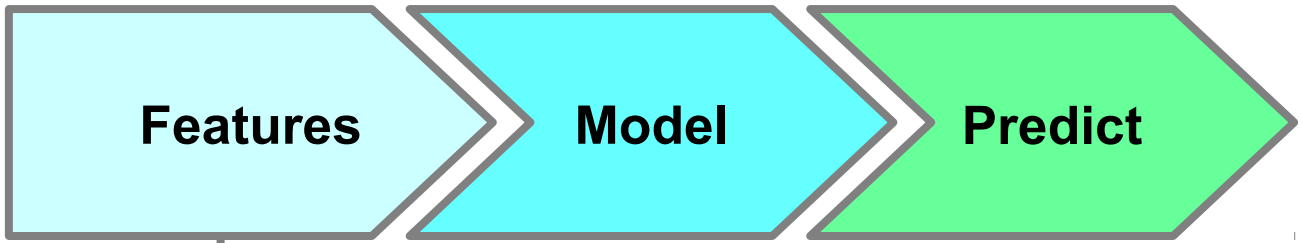
Nutzung von maschinellem Lernen



Auswahl 1/2

- DistanceToPreviousLine
- DifferenceInStartPositionToPreviousLine
- **SameAverageFontAsPreviousLine**
- **SameConnectedFontAsPreviousLine**
- **SameAverageFontSizeAsPreviousLine**
- **SameConnectedFontSizeAsPreviousLine**
- **SameConnectedFontColorAsPreviousLine**
- **SpecialFontPreviousLine**
- SameRoleAsPreviousLine

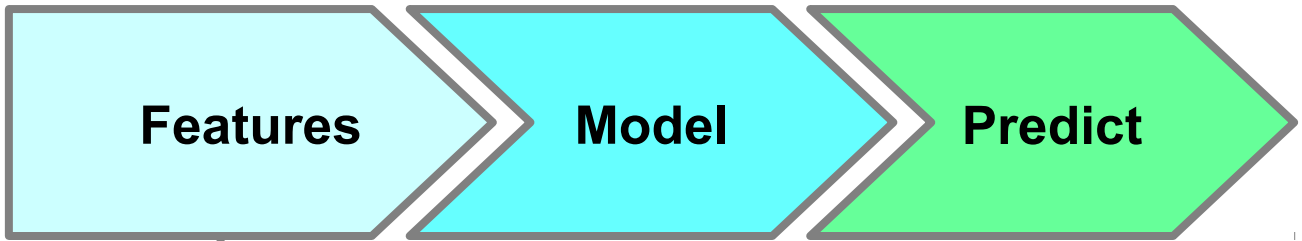
Nutzung von maschinellem Lernen



Auswahl 1/2

- DistanceToPreviousLine
- DifferenceInStartPositionToPreviousLine
- SameAverageFontAsPreviousLine
- SameConnectedFontAsPreviousLine
- SameAverageFontSizeAsPreviousLine
- SameConnectedFontSizeAsPreviousLine
- SameConnectedFontColorAsPreviousLine
- SpecialFontPreviousLine
- **SameRoleAsPreviousLine**

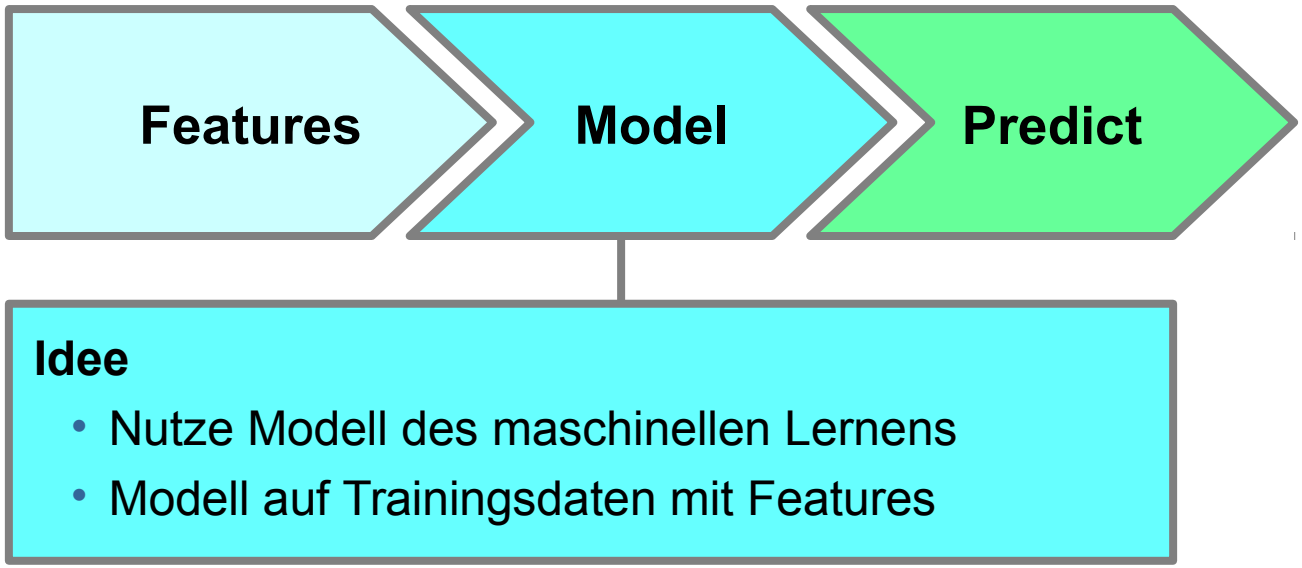
Nutzung von maschinellem Lernen



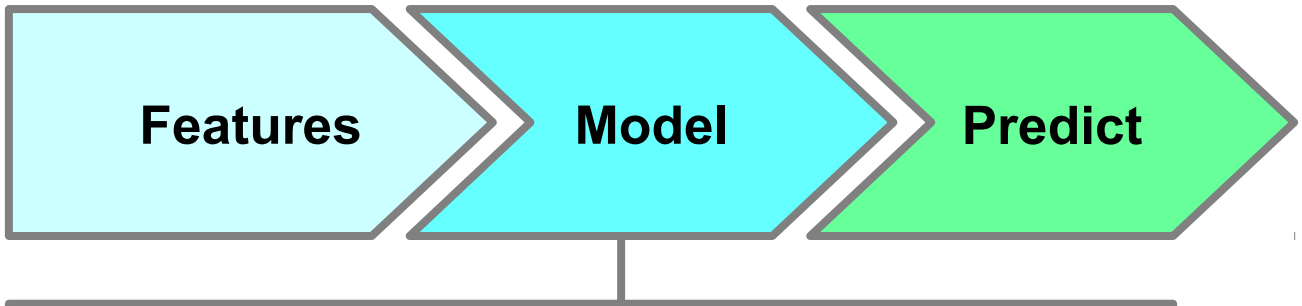
Auswahl 2/2

- FirstLineOfPage
- OverlapPreviousLine
- IsListing
- AfterListing
- IsTitle
- AfterTitle
- FormulaSwap
- IsPageNumber

Nutzung von maschinellem Lernen



Nutzung von maschinellem Lernen

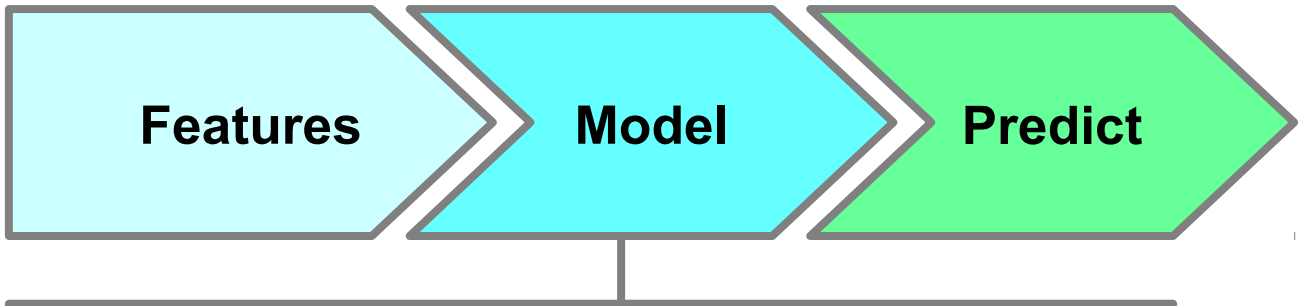


Idee

- Nutze Modell des maschinellen Lernens
- Modell auf Trainingsdaten mit Features

feature	text	page	minX	minY	maxX	maxY	mostCommonFont	...
line	Modeling	1	117	719	491	731	font-5	...
line	Andrew J	1	149	683	462	695	font-32	...
line	submitted	1	248	626	366	636	font-31	...
line	Received	1	121	568	361	578	font-32	...
line	1Departm	1	88	152	468	164	font-32	...
line	2LHEA, M	1	88	129	372	141	font-32	...
line	—23—	2	293	759	318	767	font-32	...
line	ABSTRA	2	268	721	343	730	font-5	...
line	We We m	2	101	683	483	694	font-32	...

Nutzung von maschinellem Lernen



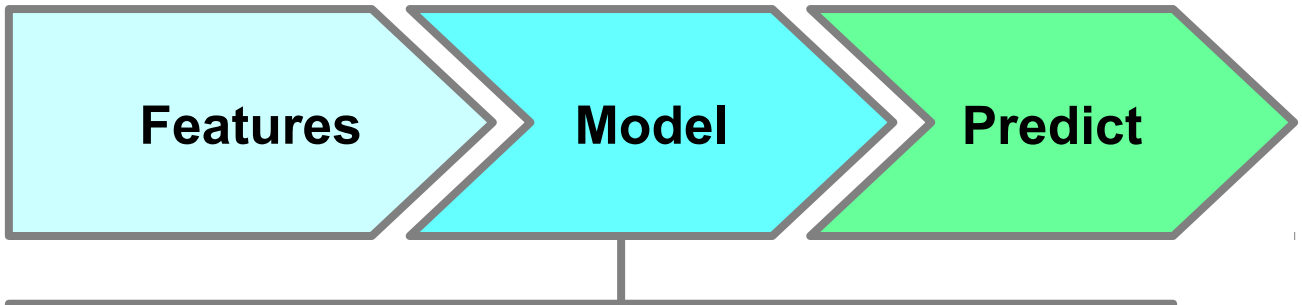
Idee

- Nutze Modell des maschinellen Lernens
- Modell auf Trainingsdaten mit Features

feature	text	page	minX	minY	maxX	maxY	mostCommonFont	...
line	Modeling	1	117	719	491	731	font-5	...
line	Andrew J	1	149	683	462	695	font-32	...
line	submitted	1	248	626	366	636	font-31	...
line	Received	1	121	568	361	578	font-32	...
line	1Departm	1	88	152	468	164	font-32	...
line	2LHEA, M	1	88	129	372	141	font-32	...
line	—23—	2	293	759	318	767	font-32	...
line	ABSTRA	2	268	721	343	730	font-5	...
line	We We m	2	101	683	483	694	font-32	...

FEATURES

Nutzung von maschinellem Lernen



Idee

- Nutze Modell des maschinellen Lernens
- Modell auf Trainingsdaten mit Features

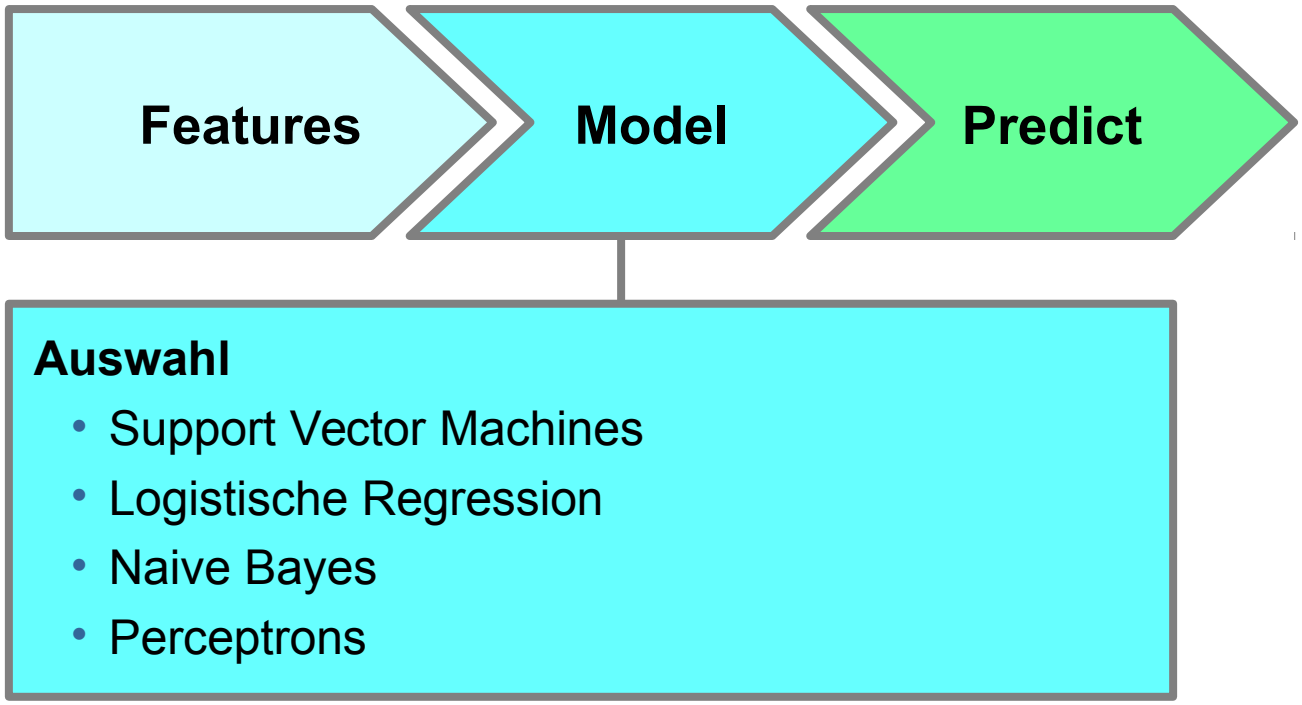
feature	text	page	minX	minY	maxX	maxY	mostCommonFont	...
line	Modeling	1	117	719	491	731	font-5	...
line	Andrew J	1	149	683	462	695	font-32	...
line	submitted	1	248	626	366	636	font-31	...
line	Received	1	121	568	361	578	font-32	...
line	1Departm	1	88	152	468	164	font-32	...
line	2LHEA, M	1	88	129	372	141	font-32	...
line	—23—	2	293	759	318	767	font-32	...
line	ABSTRA	2	268	721	343	730	font-5	...
line	We We m	2	101	683	483	694	font-32	...

FEATURES

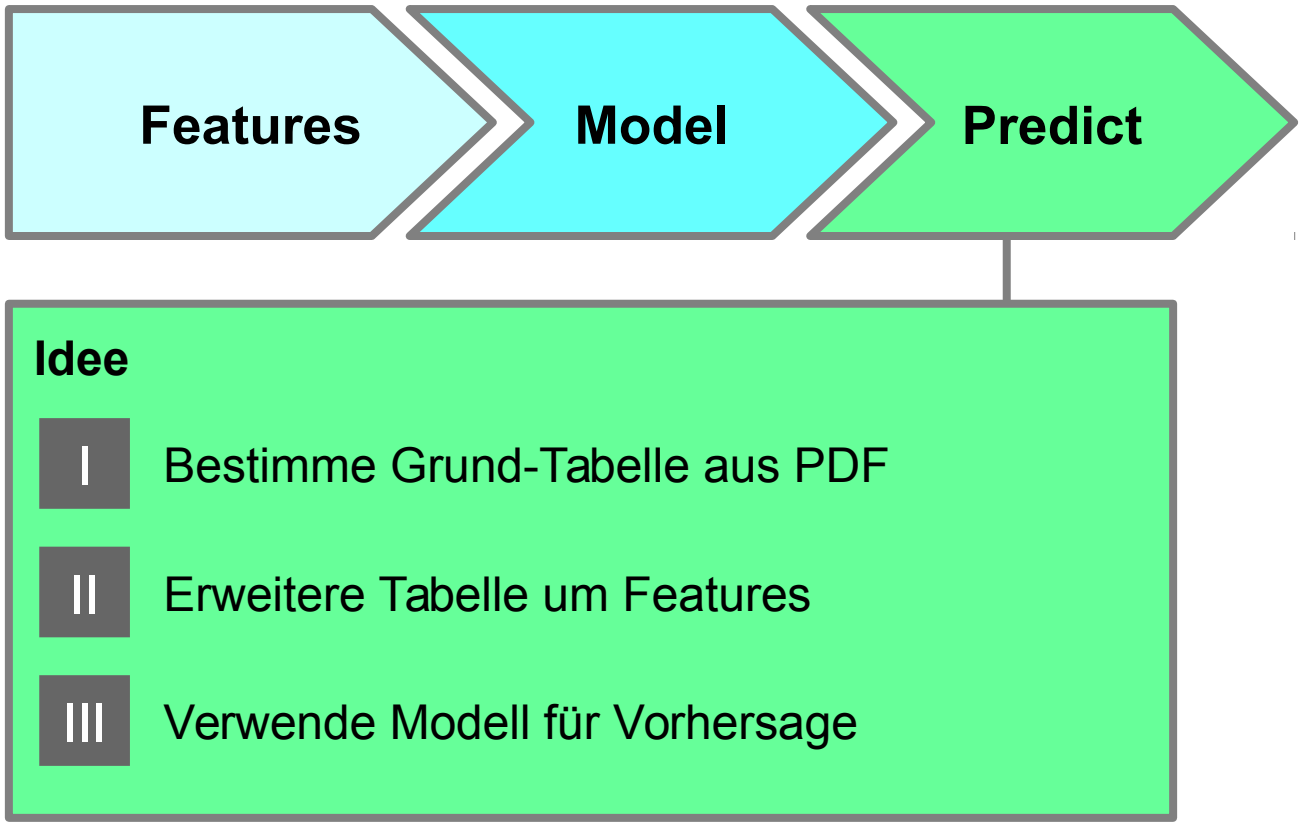
is_start_of_paragraph
1
1
0
0
0
0
1
1
1

LÖSUNG

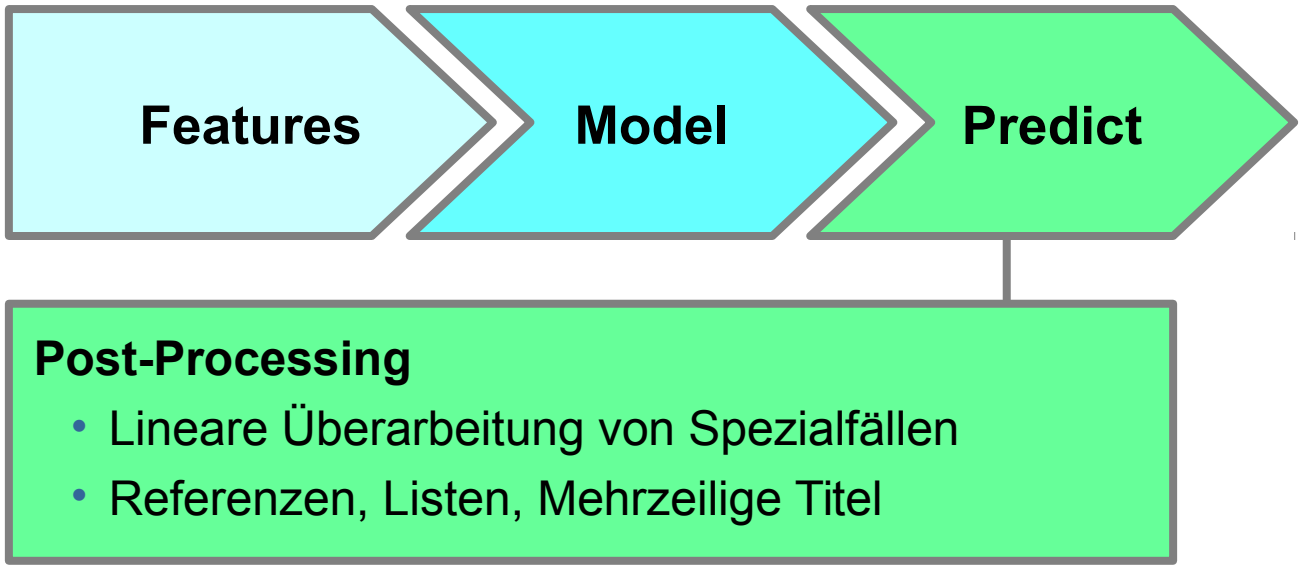
Nutzung von maschinellem Lernen



Nutzung von maschinellem Lernen



Nutzung von maschinellem Lernen



Vorführung

Erzeugung von Trainingsdaten

Vorraussetzungen

- Musterlösung
- Verbindung von Features und Lösung

Erzeugung von Trainingsdaten

Vorraussetzungen

- Musterlösung
- Verbindung von Features und Lösung

feature	text	page	minX	minY	maxX	maxY	mostCommonFont	...
line	Modeling	1	117	719	491	731	font-5	...
line	Andrew J	1	149	683	462	695	font-32	...
line	submitted	1	248	626	366	636	font-31	...
line	Received	1	121	568	361	578	font-32	...
line	1Departm	1	88	152	468	164	font-32	...
line	2LHEA, M	1	88	129	372	141	font-32	...
line	—23—	2	293	759	318	767	font-32	...
line	ABSTRA	2	268	721	343	730	font-5	...
line	We We m	2	101	683	483	694	font-32	...

FEATURES

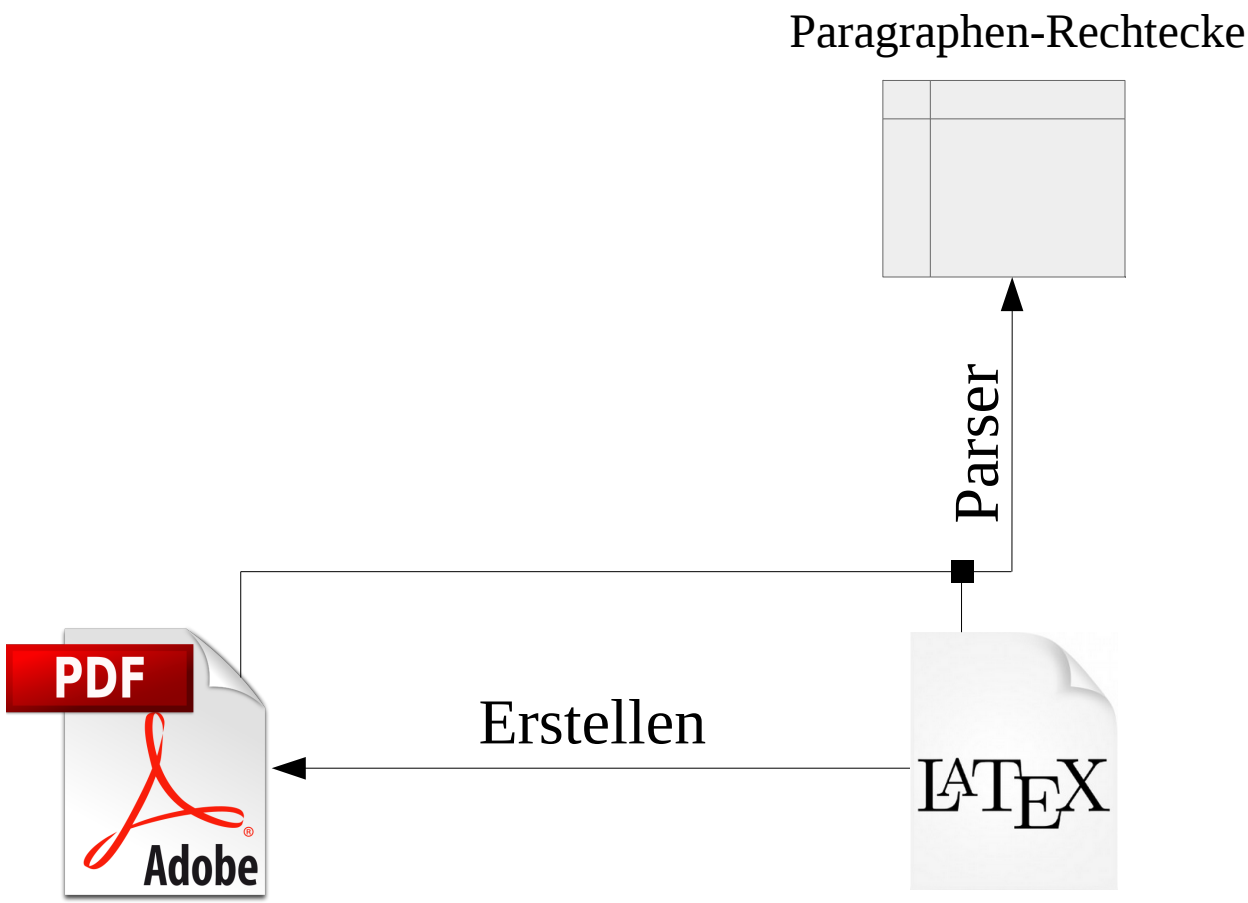
is_start_of_paragraph
1
1
0
0
0
0
1
1
1

LÖSUNG

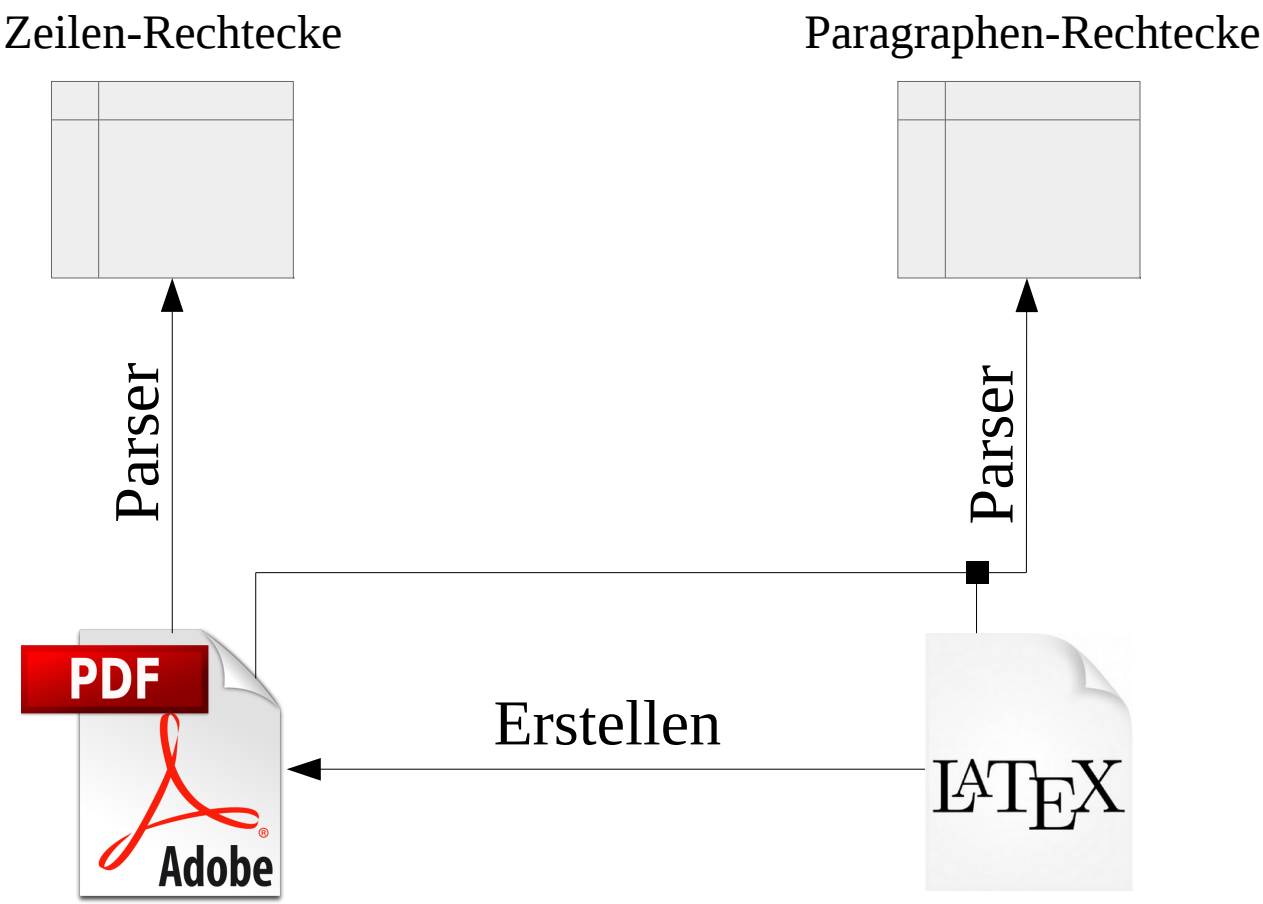
Erzeugung von Trainingsdaten

Erzeugung von Trainingsdaten

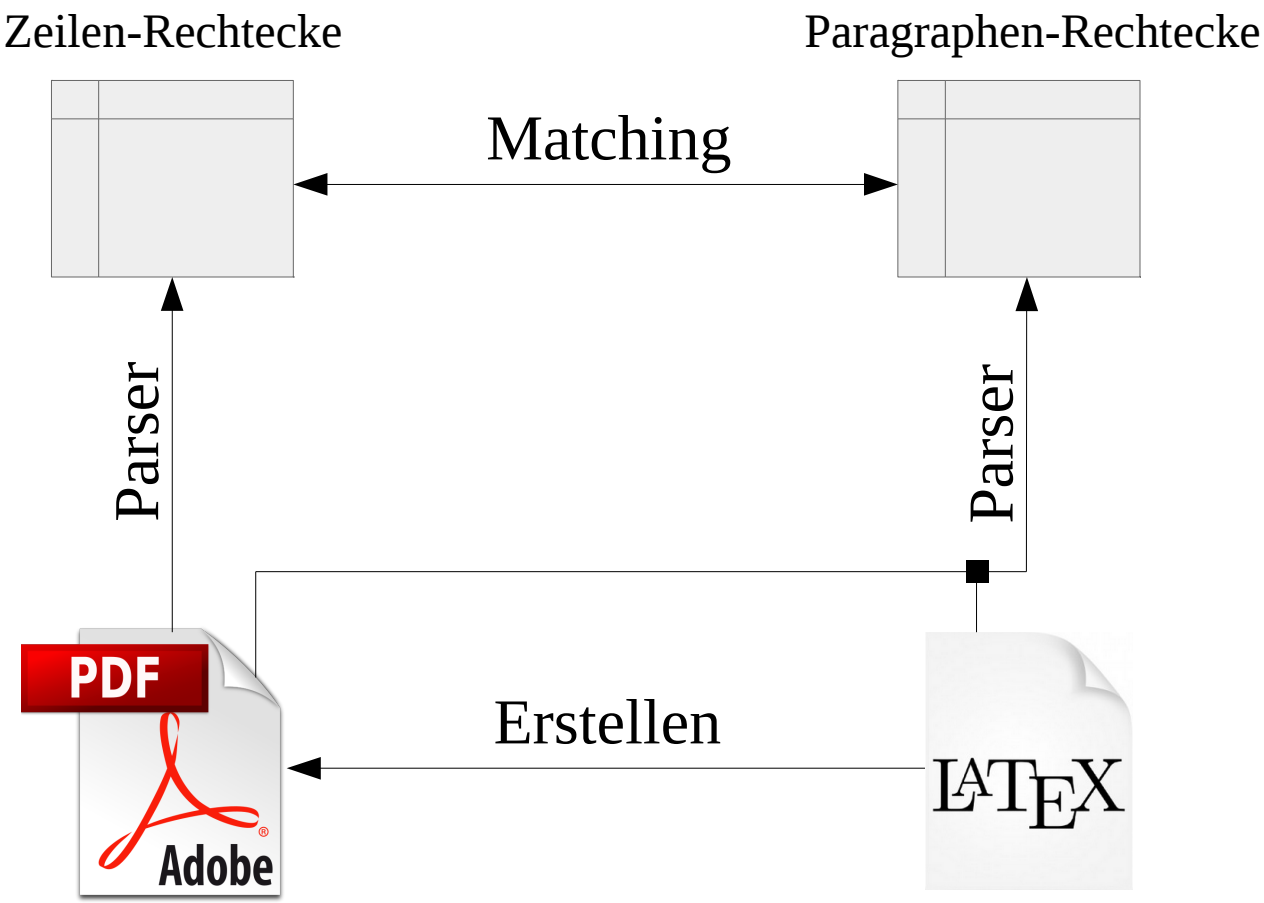
Erzeugung von Trainingsdaten



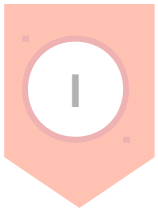
Erzeugung von Trainingsdaten



Erzeugung von Trainingsdaten



Überblick



Einführung

- Problemstellung und Motivation
- Ziele

Ausführung

- Nutzung von maschinellem Lernen
- Erzeugung von Trainingsdaten

Analyse

- Laufzeit der Algorithmen
- Qualität der Ergebnisse

Externe Algorithmen

Externe Algorithmen

Parsen eines PDFs

Input: PDF-Dokument

Output: Tabelle mit Informationen über jede Zeile des PDFs

Laufzeit: *O(A1)*

Laufzeit der Algorithmen

Externe Algorithmen

Parsen eines PDFs

Input: PDF-Dokument

Output: Tabelle mit Informationen über jede Zeile des PDFs

Laufzeit: *O(A1)*

Parsen eines TEXs

Input: TEX-Dokument

Output: Tabelle mit Paragraphen-Rechtecken

Laufzeit: *O(A2)*

Laufzeit der Algorithmen

Interne Algorithmen

N: Anzahl Zeilen M: Anzahl Features

Laufzeit der Algorithmen

Interne Algorithmen

N: Anzahl Zeilen M: Anzahl Features

Generieren von Trainingsdaten

Input: TEX-Dokument

Output: Trainingsdaten mit Lösung

Laufzeit: $O(N^2 + A1 + A2)$

Laufzeit der Algorithmen

Interne Algorithmen

N: Anzahl Zeilen M: Anzahl Features

Generieren von Trainingsdaten

Input: TEX-Dokument

Output: Trainingsdaten mit Lösung

Laufzeit: $O(N^2 + A1 + A2)$

Haupt-Algorithmus zur Paragraphen-Bestimmung

Input: PDF-Dokument

Output: Paragraphen-Lösung, Visualisierung

Laufzeit: $O(N \cdot M + A1)$

Laufzeit der Algorithmen

Interne Algorithmen

N: Anzahl Zeilen M: Anzahl Features

Evaluation eines Modells

Input: Verwendetes Modell, Trainings-Dokument

Output: Precision, Recall und F1-Score

Laufzeit: $O(N \cdot M)$

Qualität der Ergebnisse

Modell	Precision	Recall	F1
Logistische Regression	0.93	0.70	0.80
Support Vector Machines	0.98	0.98	0.98
Perceptrons	0.98	0.24	0.39
Naive Bayes	0.97	0.51	0.67

Analyse auf 1000 Datensätzen

Qualität der Ergebnisse

Anteil der berechneten Paragraphen-Anfänge, die korrekt sind

Modell	Precision	Recall	F1
Logistische Regression	0.93	0.70	0.80
Support Vector Machines	0.98	0.98	0.98
Perceptrons	0.98	0.24	0.39
Naive Bayes	0.97	0.51	0.67

Analyse auf 1000 Datensätzen

Qualität der Ergebnisse

Anteil der korrekten Paragraphen-Anfänge, die erkannt wurden

Modell	Precision	Recall	F1
Logistische Regression	0.93	0.70	0.80
Support Vector Machines	0.98	0.98	0.98
Perceptrons	0.98	0.24	0.39
Naive Bayes	0.97	0.51	0.67

Analyse auf 1000 Datensätzen

Qualität der Ergebnisse

Verrechnung von Precision und Recall

Modell	Precision	Recall	F1
Logistische Regression	0.93	0.70	0.80
Support Vector Machines	0.98	0.98	0.98
Perceptrons	0.98	0.24	0.39
Naive Bayes	0.97	0.51	0.67

Analyse auf 1000 Datensätzen

Qualität der Ergebnisse

Modell	Precision	Recall	F1
Logistische Regression	0.93	0.70	0.80
Support Vector Machines	**0.98**	**0.98**	**0.98**
Perceptrons	0.98	0.24	0.39
Naive Bayes	0.97	0.51	0.67

Analyse auf 1000 Datensätzen

Beobachtung 1: das beste Ergebnis

Qualität der Ergebnisse

Modell	Precision	Recall	F1
Logistische Regression	0.93	0.70	0.80
Support Vector Machines	0.98	0.98	0.98
Perceptrons	**0.98**	**0.24**	**0.39**
Naive Bayes	0.97	0.51	0.67

Analyse auf 1000 Datensätzen

Beobachtung 2: das schlechteste Ergebnis

Fazit

In Summe gute Ergebnisse

… auch wenn Spezialfälle Schwierigkeiten bereiten können

Mögliche Verbesserungen

… im Hinblick auf die verwendeten Features

… im Hinblick auf die verwendeten Algorithmen