Werkzeuge für das wissenschaftliche Arbeiten

Python for Machine Learning and Data Science

Magnus Bender
bender@ifis.uni-luebeck.de
Wintersemester 2022/23

Inhaltsübersicht

1. Programmiersprache Python
   a) *Einführung, Erste Schritte*
   b) *Grundlagen*
   c) *Fortgeschritten*
2. Auszeichnungssprachen
   a) LaTeX, Markdown
3. Benutzeroberflächen und Entwicklungsumgebungen
   a) Jupyter Notebooks lokal und in der Cloud (Google Colab)
4. Versionsverwaltung
   a) Git, GitHub
5. Wissenschaftliches Rechnen
   a) NumPy, SciPy
6. Datenverarbeitung und -visualisierung
   a) Pandas, matplotlib, NLTK
7. Machine Learning (scikit-learn)
   a) Grundlegende Ansätze (Datensätze, Auswertung)
   b) Einfache Verfahren (Clustering, …)
8. DeepLearning
   a) TensorFlow, PyTorch, HuggingFace Transformers

Themen

I. Projektaufgabe 2

- Herangehensweise & Tipps

II. Auszeichnungssprachen

- Inhalt, Struktur & Form
- Semantische Auszeichnung
- Markdown
- LaTeX

Projektaufgabe 2

„Objektorientierung in Python“

- Klasse „Datensatz“ zur Verwaltung von Daten
  - Jedes Datum hat einen (eindeutigen) Schlüssel (Namen)
  - Oberklasse spezifiziert verschiedene Schnittstellen eines Datensatzes
  - Datensätze können erstellt, vereinigt, iteriert, … werden
- Daten werden „extern“ in einen Datensatz hinzugefügt
- Datensätze werden „extern“ genutzt.

dataset_usage.py

```python
from dataset import DataSetItem
from implementation import DataSet

data = DataSet([
  DataSetItem("Name 1", 11, "Inhalt 1“)
])
data += DataSetItem("Name 2", 12, "Inhalt 2")

del data["Name 2"]

for item in data:
  print(item)
```

Hier im Sinne von „außerhalb des zu schreibenden Programms“

Herangehensweise & Tipps

1. Anforderungen an Klasse
2. Spezifikation jeder Schnittstelle exakt umsetzen
3. Schrittweise lösen
   A. Schnittstellen einzeln nacheinander umsetzen
   B. Schnittstellen testen

- Inhalte des ersten Teils (erste drei Vorlesungen) beachten

Warum diese Art von Aufgabe?

Eigene Daten, die von einer Bibliothek verarbeitet werden sollen und dafür passenden *ansprechbar* sein müssen.

II.
Auszeichnungssprachen

Aber zuerst:

Fragen zur Aufgabe 2?

Inhalt, St

Inhalt

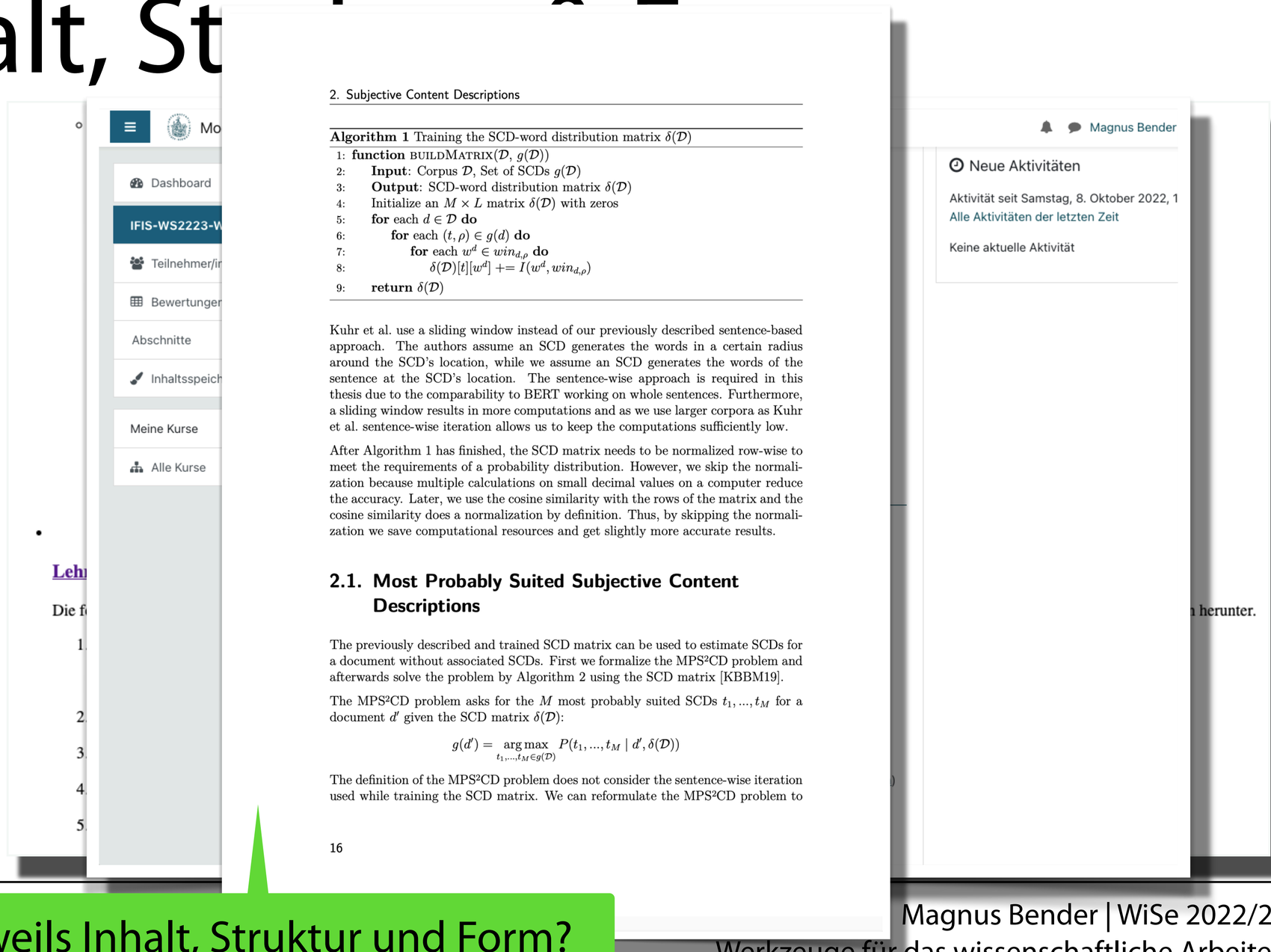
- Bedeutung eines Textes
- Die Wörter (und Sätze)

Struktur

- Aufbau des Dokuments
- Die Absätze, Kaptitel und Überschriften

Form

- Aussehen des Dokuments
- Die Darstellung, wie Farben, Schriftart, Markierungen und Kästen



2. Subjective Content Descriptions

Algorithm 1 Training the SCD-word distribution matrix $\delta(\mathcal{D})$

```
function BUILDMATRIX(D, g(D))
    Input: Corpus D, Set of SCDs g(D)
    Output: SCD-word distribution matrix δ(D)
    Initialize an M × L matrix δ(D) with zeros
    for each d ∈ D do
        for each (t, ρ) ∈ g(d) do
            for each w^d ∈ win_{d,ρ} do
                δ(D)[t][w^d] += I(w^d, win_{d,ρ})
    return δ(D)
```

Kuhr et al. use a sliding window instead of our previously described sentence-based approach. The authors assume an SCD generates the words in a certain radius around the SCD's location, while we assume an SCD generates the words of the sentence at the SCD's location. The sentence-wise approach is required in this thesis due to the comparability to BERT working on whole sentences. Furthermore, a sliding window results in more computations and as we use larger corpora as Kuhr et al. sentence-wise iteration allows us to keep the computations sufficiently low.

After Algorithm 1 has finished, the SCD matrix needs to be normalized row-wise to meet the requirements of a probability distribution. However, we skip the normalization because multiple calculations on small decimal values on a computer reduce the accuracy. Later, we use the cosine similarity with the rows of the matrix and the cosine similarity does a normalization by definition. Thus, by skipping the normalization we save computational resources and get slightly more accurate results.

2.1. Most Probably Suited Subjective Content Descriptions

The previously described and trained SCD matrix can be used to estimate SCDs for a document without associated SCDs. First we formalize the MPS²CD problem and afterwards solve the problem by Algorithm 2 using the SCD matrix [KBBM19].

The MPS²CD problem asks for the M most probably suited SCDs $t_1, ..., t_M$ for a document d' given the SCD matrix $\delta(\mathcal{D})$:

$$g(d') = \underset{t_1,...,t_M \in g(\mathcal{D})}{\arg\max} P(t_1, ..., t_M \mid d', \delta(\mathcal{D}))$$

The definition of the MPS²CD problem does not consider the sentence-wise iteration used while training the SCD matrix. We can reformulate the MPS²CD problem to

16

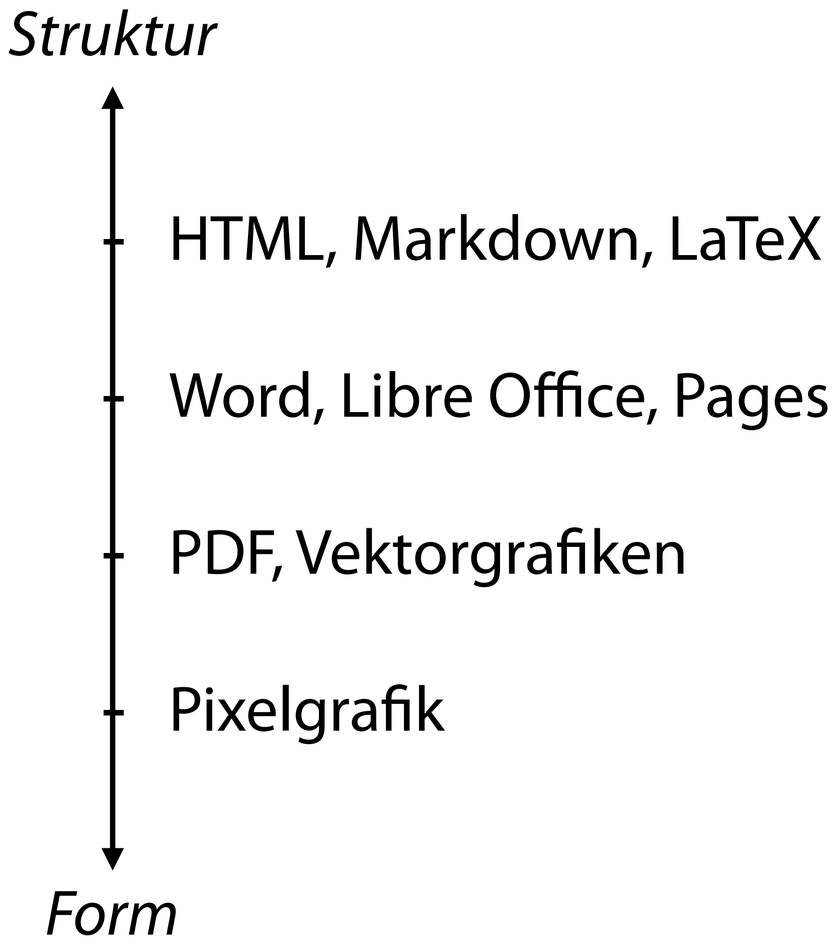
Was ist jeweils Inhalt, Struktur und Form?

Semantische Auszeichnung

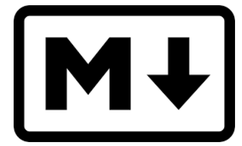
- Trennung zwischen Inhalt und Form
- Angabe des Inhalt mit Struktur

  Überschrift „Meine Übungsaufgabe“
- Anschließend Formatierung der Struktur

  Überschriften seien „**groß und fett**“



Markdown



```
# Markdown

> From [Wikipedia](https://en.wiki
wiki/Markdown), the free encyclope

## Article

Markdown is a lightweight markup l
creating formatted text using a pl
editor. John Gruber and Aaron Swar
Markdown in 2004 as a markup langu
appealing to human readers in its
form.

Paragraphs are separated by a blar

Two spaces at the end of a line
produce a line break.
Text can be sty ed _italic_, **bold**, or
`monospace`.
```

Markdown

> From Wikipedia, the free encyclopedia

Article

Markdown is a lightweight markup language for creating formatted text using a plain-text editor. John Gruber and Aaron Swartz created Markdown in 2004 as a markup language that is appealing to human readers in its source code form.

Paragraphs are separated by a blank line.

Two spaces at the end of a line
produce a line break. Text can be styled *italic*, **bold**, or `monospace`.

Inhalt und Struktur, aber keine Form

Verschiedene Formatierungen

Markdown

Es gibt auch Fußnoten[1].

A	B	C
1	2	3

```
print("Hallo ;)")
```

- [x] $\frac{1}{2}^2$
- [ ] $\frac{x_1^2+5x_2}{x_1}$

1. Tatsächlich ist das hier erweitertes Markdown ↩

Überschrift

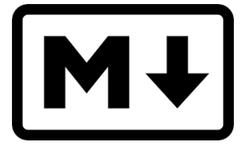
Unterüberschrift

> Quote

- **Fett**
- *kursiv*
- `Code`

1. Links
2. 

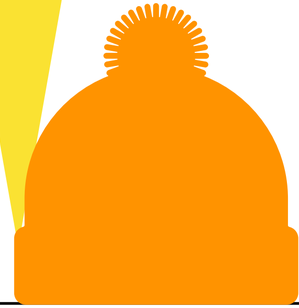
Markdown



Einmalig zu aktivieren: Einstellungen → Texteditor wählen → „Einfacher Text“
Unter den Eingabefeldern ist dann eine Auswahl „Markdown“ möglich.

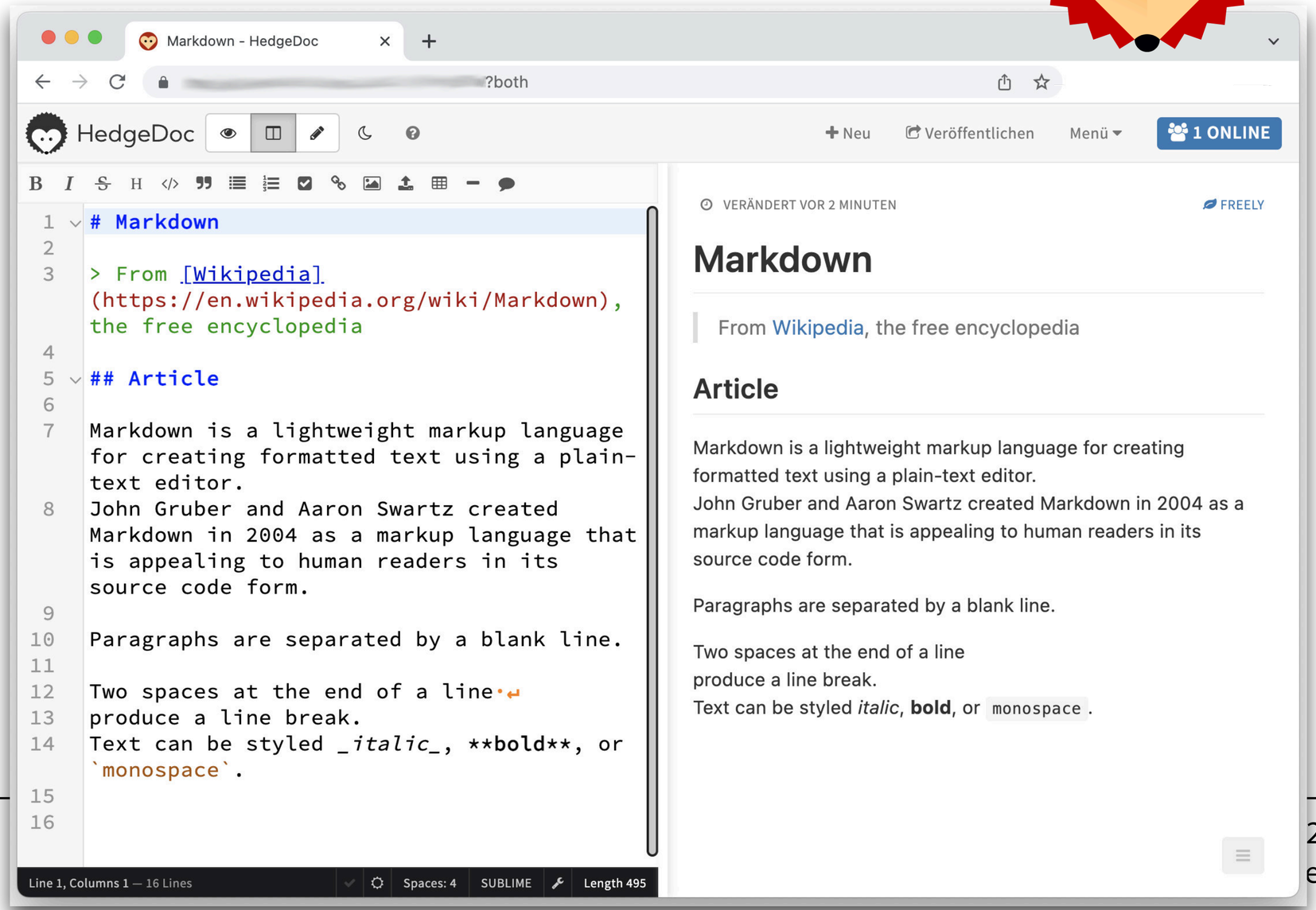
- Mittlerweile einer der Standards zum einfachen Formatieren von Texten
  - Kommunikation, z.B. *Moodle*, Discord, Slack
  - Code-Editoren, z.B. VS Code
  - Plattformen zur Versionsverwaltung, z.B. GitHub, GitLab
  - Plattformen zur Zusammenarbeit, z.B. HedgeDoc

„Oft lohnt es sich einfach mal Markdown-Syntax zu testen.“



Beispiel: HedgeDoc

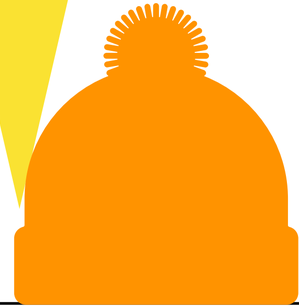
- Kollaborativ
- Markdown mit Erweiterungen
  - Gleichungen (LaTeX)
  - Diagramme
- Features & Demo



LaTeX

- Textsatzsystem
  - Geeignet für Übungszettel, Berichte, Zusammenfassungen, Abschlussarbeiten
  - Weniger für Notizen, Mitschriften (→ Markdown)
- Insbesondere Unterstützung von Gleichungen, Literaturverzeichnisse, Inhaltsverzeichnisse

„LaTe*ch* nicht LaTe*X*!“



Wir sehen auch hier Trennung von Inhalt (im „document“) und Form in der Präambel.

ste Dokument

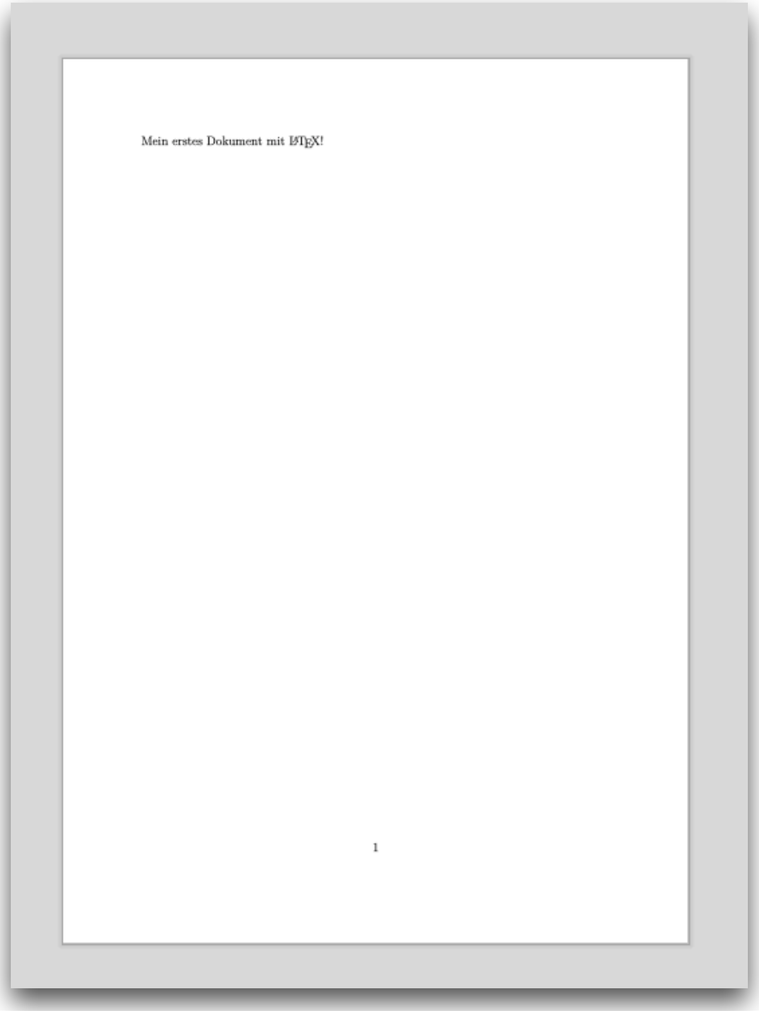
LaTeX

```
\documentclass[
  12pt, % Schriftgroesse
  a4paper, % Papier
  parskip=full % Absatzstil
]{scrartcl}

% Dateikodierung
\usepackage[utf8]{inputenc}
% Trennung "deutsch"
\usepackage[ngerman]{babel}
% Schriftart
\usepackage[T1]{fontenc}
\usepackage{lmodern}

\begin{document}

  Mein erstes Dokument mit \LaTeX!

\end{document}
```

Präambel

Inhalt

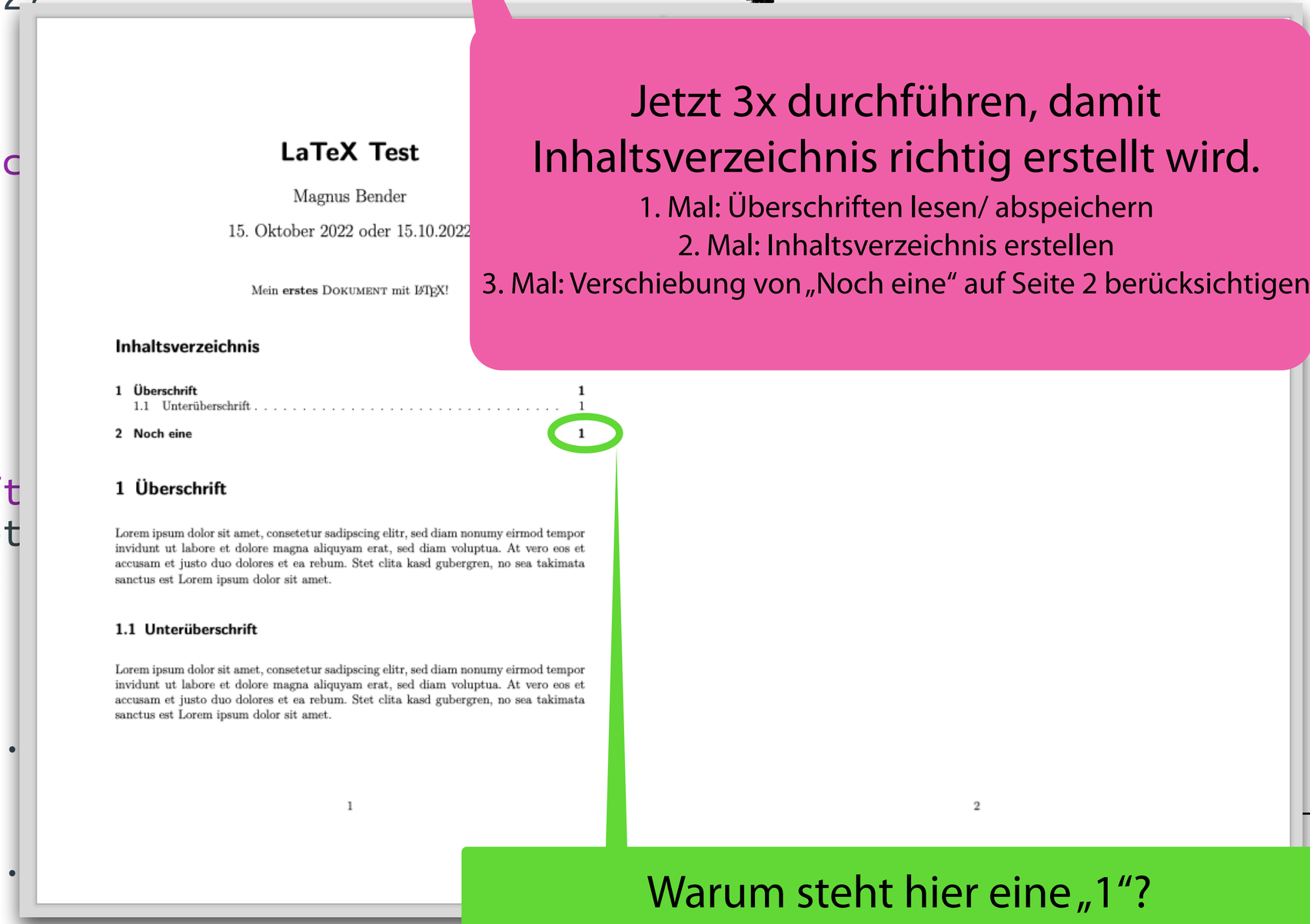
```
pdflatex ./document.tex
```



```
\begin{document}
  \title{LaTeX Test}
  \author{Magnus Bender}
  \date{\today{} oder 15.10.2022}
  \maketitle

  \begin{center}
    Mein \textbf{erstes} \textsc
  \end{center}

  \tableofcontents

  \section{Überschrift}
    Lorem ipsum dolor sit amet,

    \subsection{Unterüberschrift
      Lorem ipsum dolor sit amet
      At vero eos et accusam et

  \section{Noch eine}
    Lorem ipsum dolor sit amet,
    At vero eos et accusam et ..

    Lorem ipsum dolor sit amet,
    At vero eos et accusam et ..
\end{document}
```

3x ~~2x~~ `pdflatex ./document.tex`

LaTeX



Jetzt 3x durchführen, damit Inhaltsverzeichnis richtig erstellt wird.

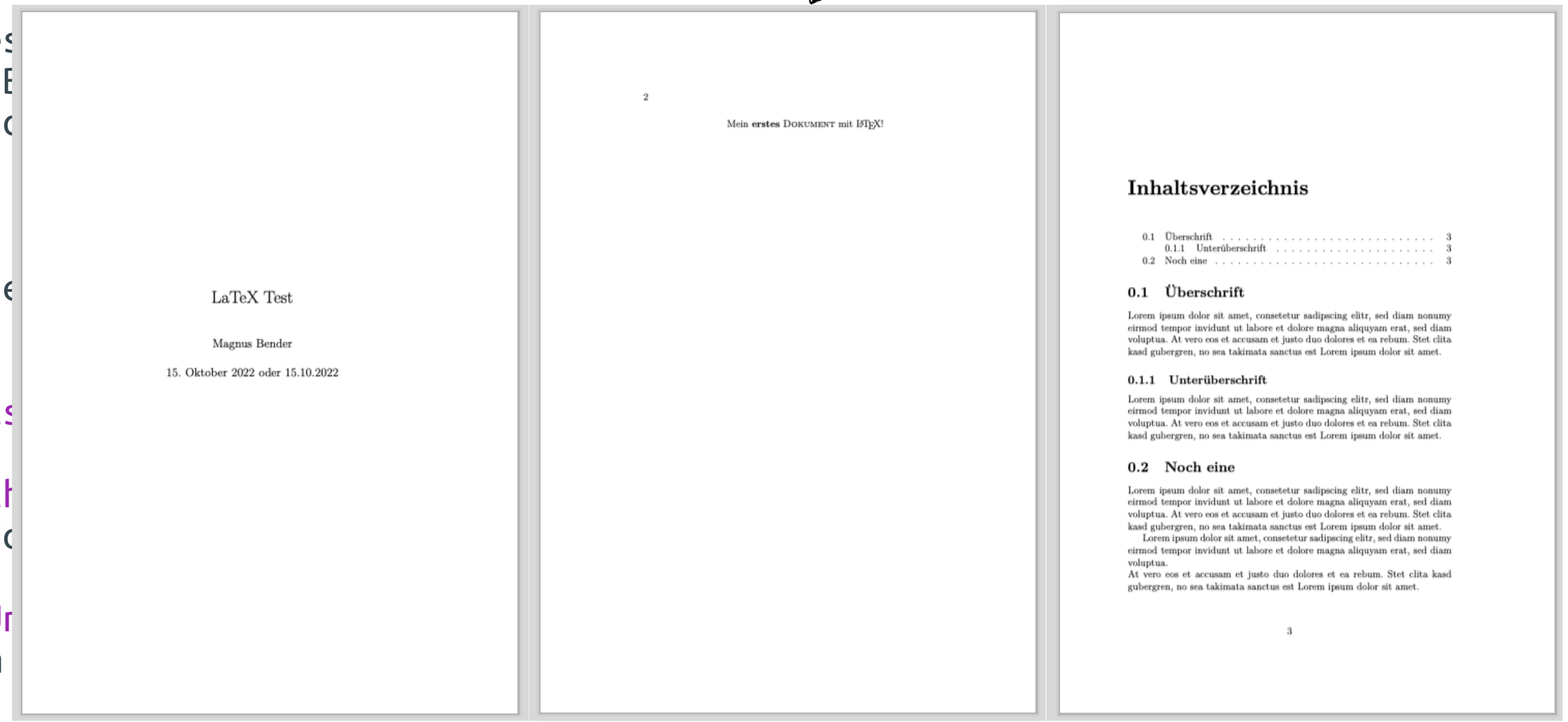
1. Mal: Überschriften lesen/ abspeichern
2. Mal: Inhaltsverzeichnis erstellen
3. Mal: Verschiebung von „Noch eine“ auf Seite 2 berücksichtigen

Warum steht hier eine „1“?

```
\documentclass[]{book}

% ...

\begin{document}
  \title{LaTeX Tes
  \author{Magnus E
  \date{\today{} c
  \maketitle

  \begin{center}
    Mein \textbf{e
  \end{center}

  \tableofcontents

  \section{Übersch
    Lorem ipsum do

    \subsection{Ur
      Lorem ipsum
      At vero eos

% ...
```

2x `pdflatex ./document.tex`

Gleicher Inhalt, aber „Buch“, somit andere Form!



LaTeX Test

Magnus Bender

15. Oktober 2022 oder 15.10.2022

2

Mein **erstes** Dokument mit LaTeX!

Inhaltsverzeichnis

0.1 Überschrift

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

0.1.1 Unterüberschrift

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

0.2 Noch eine

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua.
At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

3

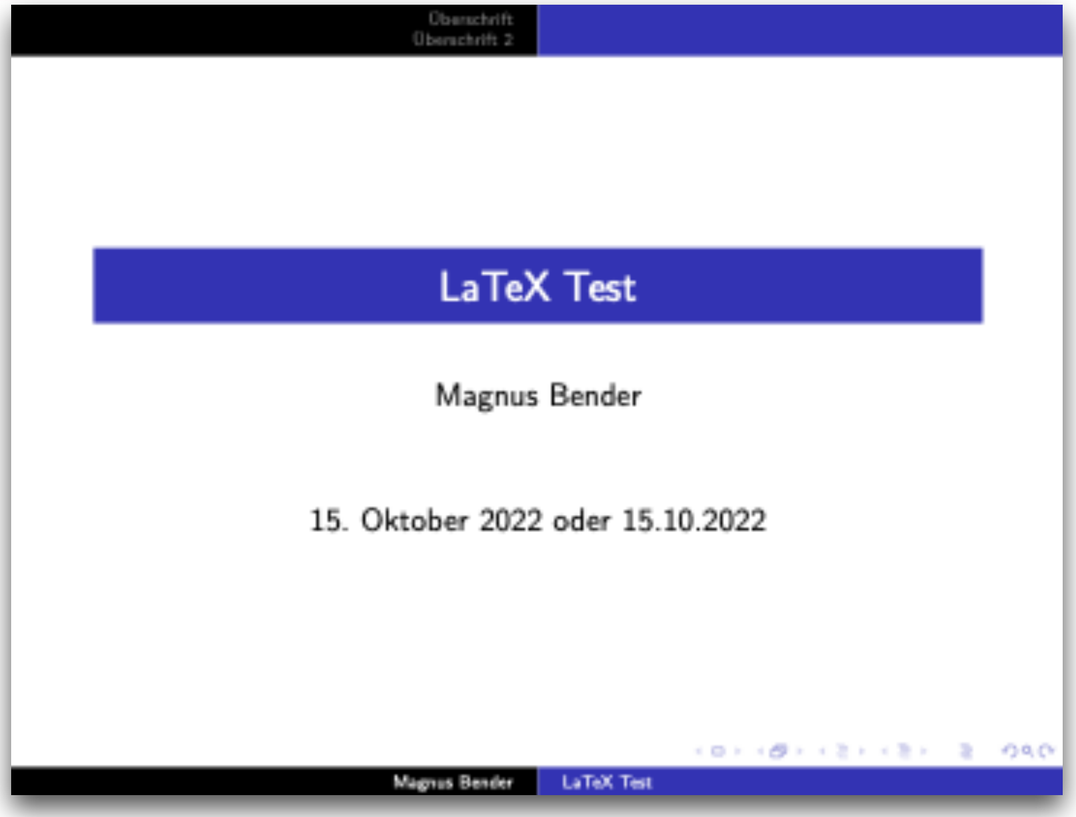
Umwandlung von LaTeX

- Wie erstelle ich die PDF?
  - „`pdflatex ./document.tex`“
  - „`pdflatex ./document.tex`“„`pdflatex ./document.tex`“
  - „`pdflatex ./document.tex`“ `...?`
- LaTeX Mk
  - „`latexmk -pdf document.tex`“

LaTeX & Beamer

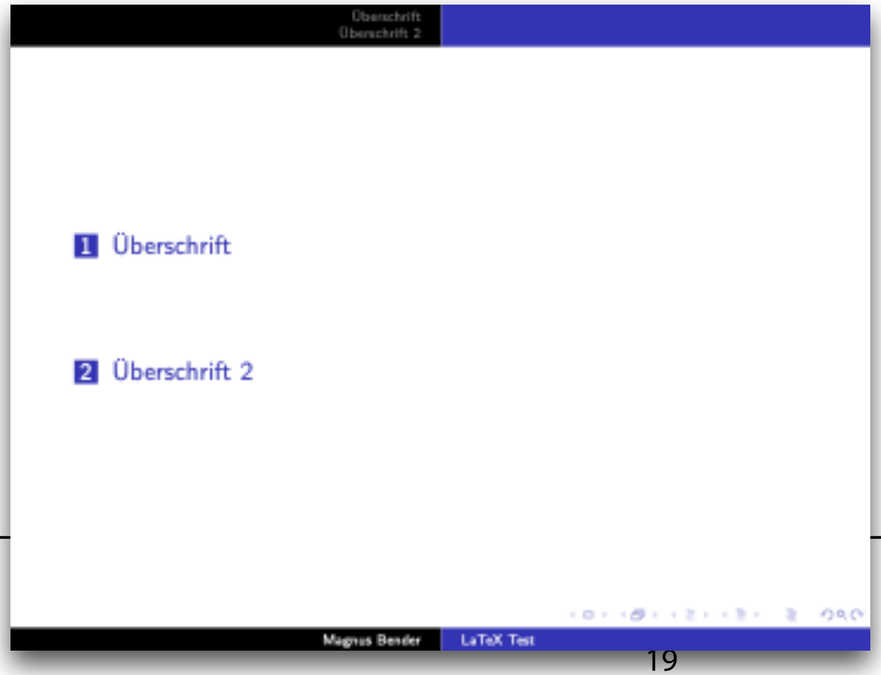
```
\documentclass{beamer}

% Dateikodierung
\usepackage[utf8]{inputenc}
% Trennung "deutsch"
\usepackage[ngerman]{babel}
% Schriftart
\usepackage[T1]{fontenc}
\usepackage{lmodern}

\title{LaTeX Test}
\author{Magnus Bender}
\date{\today{} oder 15.10.2022}

\usetheme{Luebeck}

\begin{document}

  \frame{\titlepage}

% ...
```

Präsentation, wieder andere Form, aber auch kleine Änderungen an der Struktur nötig!



LaTeX & Beamer

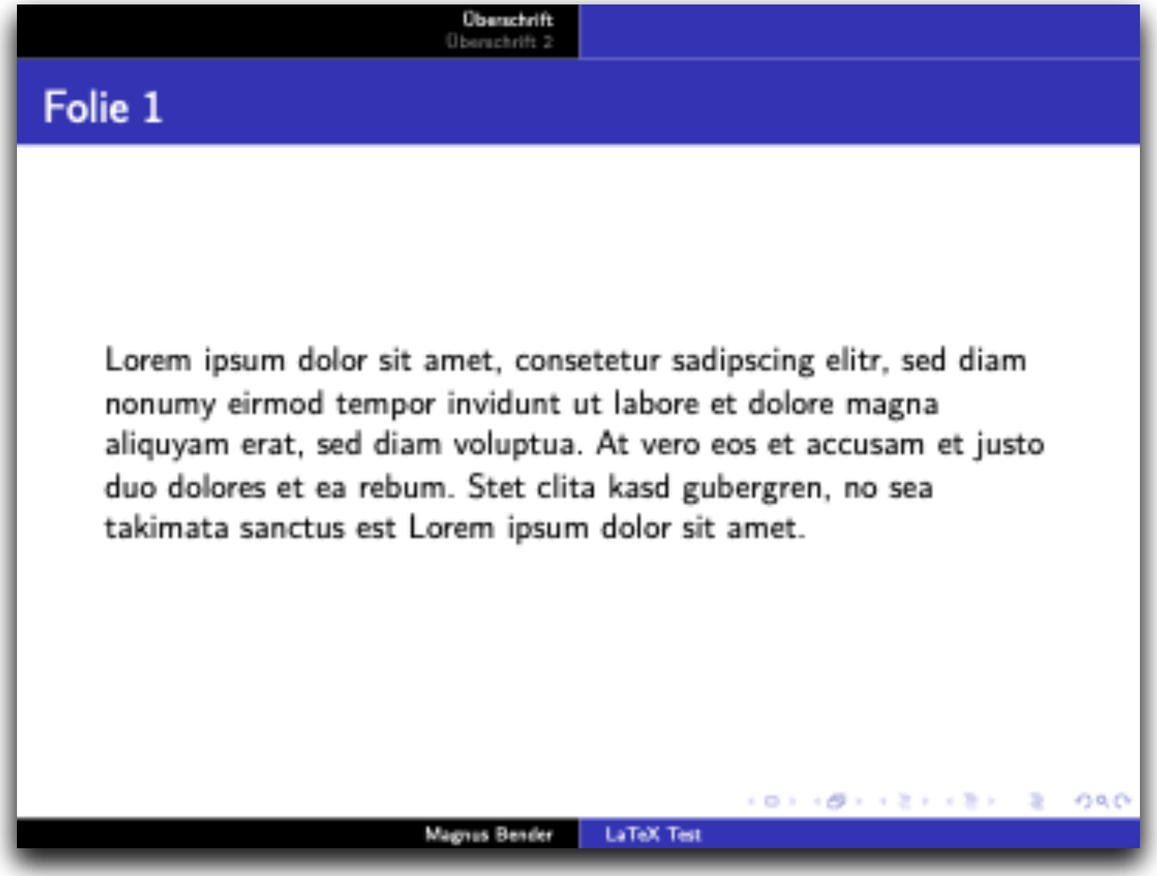
```
% ...

\begin{document}

  \frame{\titlepage}

  \begin{frame}
    \begin{center}
      Meine \textbf{erste} \alert{Präsentation} mit \LaTeX!
    \end{center}
  \end{frame}

  \begin{frame}
    \tableofcontents
  \end{frame}

% ...
```



LaTeX & Beamer

```
% …

  \section{Überschrift}

  \begin{frame}{Folie 1}
    Lorem ipsum dolor sit amet, … \pause
    At vero eos et accusam et …
  \end{frame}

% …
```

Folien mit \pause animieren, erzeugt zwei Folien, eine nur mit den Inhalten vor und auch mit den Inhalten nach \pause.



LaTeX & Beamer

```
% ...

    \section{Überschrift 2}

  \begin{frame}{Folie 2}
    \begin{block}{Hinweis}
      Ein Text
    \end{block}

    \begin{alertblock}{Wichtig}<2->
      Wieder ein Text
    \end{alertblock}
  \end{frame}

  \section{Überschrift 3}

\end{document}
```

Es gibt verschiedene Blöcke in Beamer.

Neben \pause gibt es auch die <x-y> Syntax.

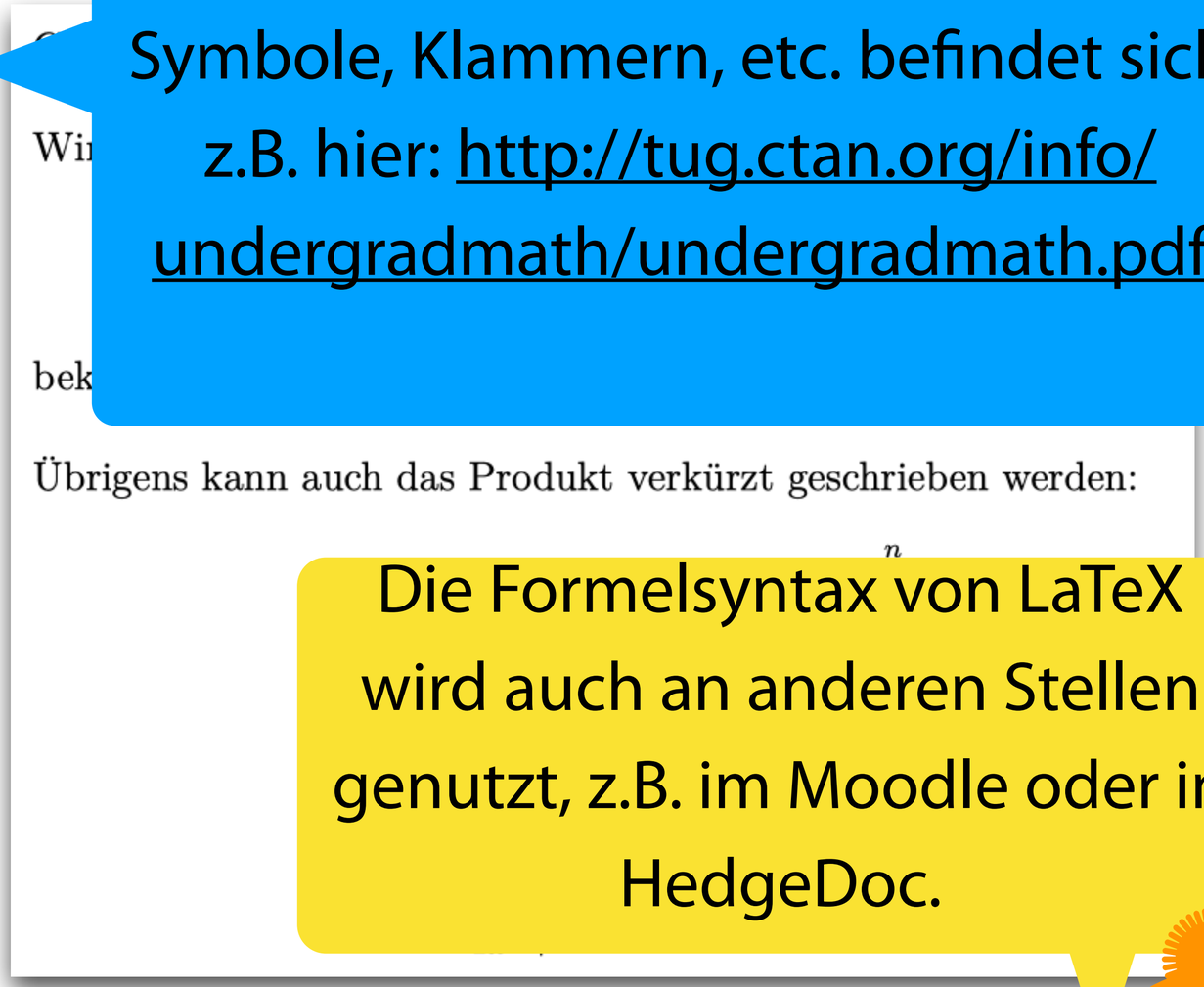
Fazit: Beamer

- *Beamer* erlaubt es Präsentationen mit LaTeX zu erstellen
  - Insbesondere ist die Umgebung `frame` neu, welche eine Folie darstellt
  - Verschiedene Themes für das Design der Folien
  - Weiterhin Trennung zwischen Form und Struktur

Form

```
Gegeben sei $ x^2 + 5x - \alpha = 12 $.

Wir nehmen an, dass die Gauß-
Summe\footnote{siehe z.B. Wikipedia}
$$ \sum^n_{i=1} i = \frac{n(n+1)}2 $$
bekannt ist.

Übrigens kann auch das Produkt verkürzt
geschrieben werden:
\[
  y_1 \cdot ... \cdot y_n
  = \prod_{i=1}^n y_i
\]

\begin{align*}
  3x^2 &+  4x &= 0 \\
  2x^2 &+ 10x &= 0 \\
  4x^2 &+     &= 0 \\
\end{align*}
```

Eine Übersicht über die verschiedenen Symbole, Klammern, etc. befindet sich z.B. hier: http://tug.ctan.org/info/undergradmath/undergradmath.pdf


Wi

bek

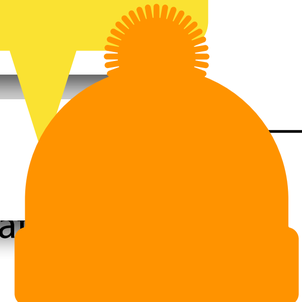
Übrigens kann auch das Produkt verkürzt geschrieben werden:

$$y_1 \cdot \ldots \cdot y_n = \prod_{i=1}^{n} y_i$$

$$\begin{aligned} 3x^2 + 4x &= 0 \\ 2x^2 + 10x &= 0 \\ 4x^2 &= 0 \end{aligned}$$

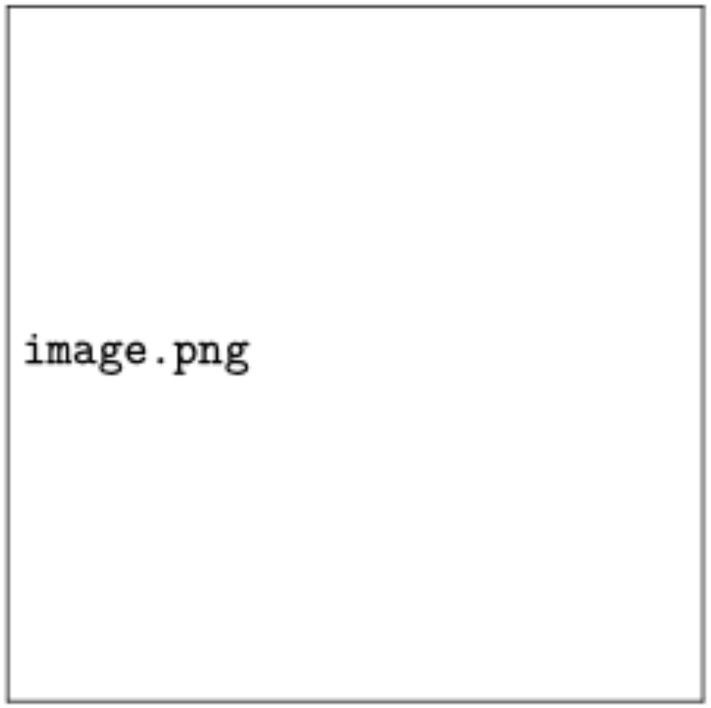
[1] siehe z.B. Wikipedia

Die Formelsyntax von LaTeX wird auch an anderen Stellen genutzt, z.B. im Moodle oder in HedgeDoc.



Weitere Elemente in LaTeX

Paket	Version	Anzahl
Numpy	1	12
Scipy	1.7	200
Gensim	2.4	30



i) Python

ii) Java

iii) LaTeX

- Auch Listen
- in Aufzählungen sind
- mit LaTeX
- möglich.

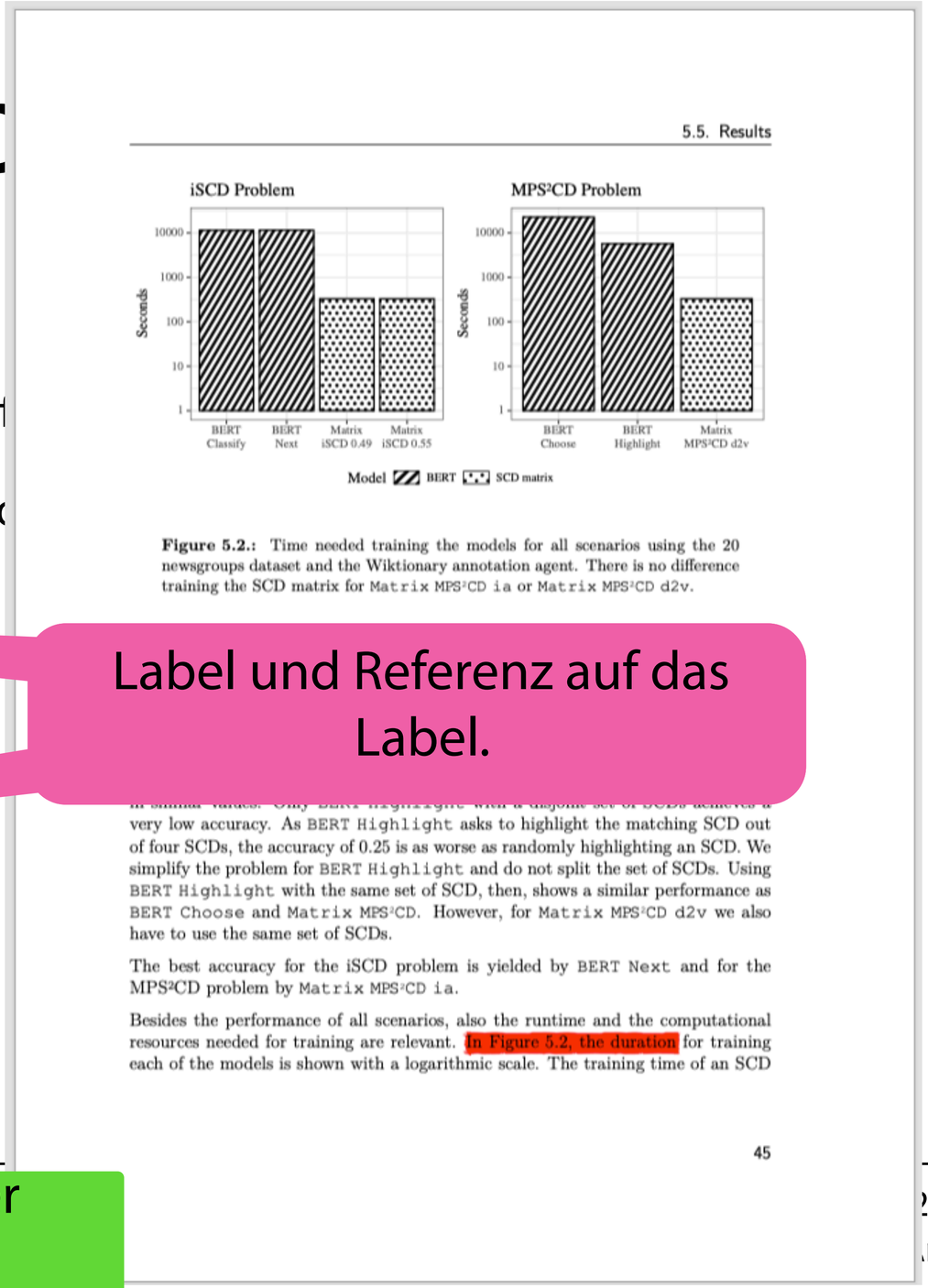
Ein Begriff der zu erklären ist.

Am Anfang wird der Begriff hervorgehoben und danach folgt der Text.

```
\begi
  \it
  \it
  \it
    \
    \
\end{
\begi
  \it
    c
  \it
    w
\end{
line
d danach folgt der Text.
```

Gleitumgeb

```
\begin{figure}
  \centering
  \includegraphics[width=\textwidth]{res/runtime_de
  \caption{
    Time needed training the models for all scenario
  }
  \label{fig:dur}
\end{figure}

Besides the performance of all scenarios,
also the runtime and the computational
resources needed for training are relevant.
In \ref{fig:dur}, the duration for
training each of the models is shown
with a logarithmic scale.
```



Label und Referenz auf das Label.

Warum ist die Grafik im PDF an einer anderen Stelle?

BibTeX und TikZ

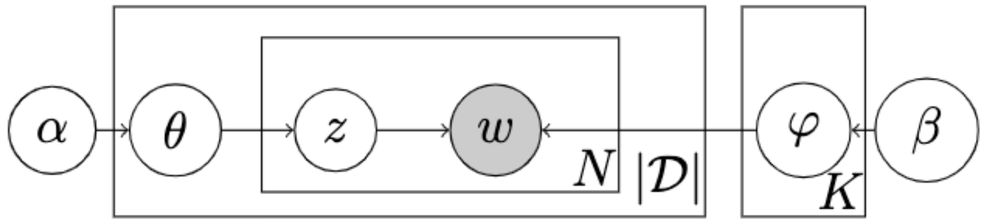
- BibTeX
  - Literaturverzeichnisse und -zitate mit LaTeX
- TikZ ist kein Zeichenprogramm
  - Grafiken und Abbildungen direkt in LaTeX erstellen

document.tex

```
Open Information Extraction
(OpenIE)~\cite{OpenIE}
extracts triples of
subject, predicate and
object.
```

literature.bib

```
@article{OpenIE,
  title = "Leveraging Linguistic Structu
  author = "Angeli, Gabor and Johnson Pr
  journal = "Proceedings of the Associat
  year = "2015",
  publisher = "Association for Computati
                                115/v1/P15-
```



Information Extraction. In: *Proceedings of the Association of Computational Linguistics (ACL)* (2015), 344–354. https://doi.org/10.3115/v1/P15-1034

[All83] Allen, James F.: Maintaining Knowledge about Temporal Intervals. In: *Commun. ACM* 26 (1983), November, Nr. 11, 832–843. https://doi.org/10.1145/182.358434

[BBG+21a] Bender, Magnus ; Braun, Tanya ; Gehrke, Marcel ; Kuhr, Felix ; Möller, Ralf ; Schiff, Simon: Identifying and Translating Subjective Content Descriptions Among Texts. In: *Int. J. Semantic Computing* 15 (2021). – Accepted for publication

[BBG+21b] Bender, Magnus ; Braun, Tanya ; Gehrke, Marcel ; Kuhr, Felix ; Möller, Ralf ; Schiff, Simon: Identifying Subjective Content Descriptions among Text. In: *Proceedings of the 15th IEEE International Conference on Semantic Computing (ICSC-21)* (2021). https://doi.org/10.1109/ICSC50631.2021.00008

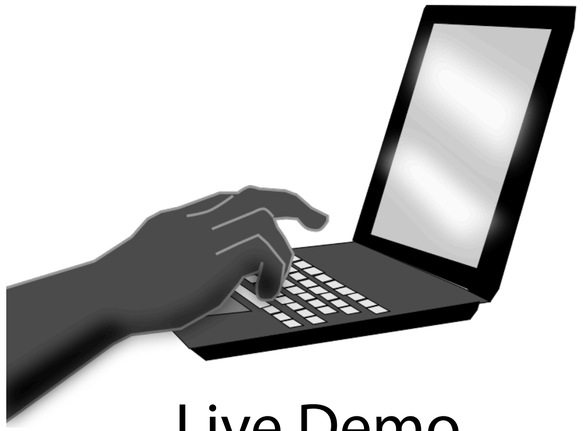
[BMR+20] Brown, Tom B. ; Mann, Benjamin ; Ryder, Nick ; Subbiah, Melanie ; Kaplan, Jared ; Dhariwal, Prafulla ; Neelakantan, Arvind ; Shyam, Pranav ; Sastry, Girish ; Askell, Amanda ; Agarwal, Sandhini ; Herbert-Voss, Ariel ; Krueger, Gretchen ; Henighan, Tom ; Child, Rewon ; Ramesh, Aditya ; Ziegler, Daniel M. ; Wu,

Links & Tools LaTeX

- https://www.mlte.de/latex/
  - Video und Unterlagen zu Vorlesungen über LaTeX, Beamer & TikZ
  - Klassen für Übungszettel
  - Vorlagen für Präsentationen und Abschlussarbeiten
- http://en.wikibooks.org/wiki/LaTeX
  - Wikibook
- http://detexify.kirelabs.org/
  - Sonderzeichen bestimmen
- https://www.overleaf.com/
  - Online und ohne Installation Dokumente in LaTeX verfassen
- https://app.diagrams.net/
  - Diagramme und Grafiken schnell und einfach erstellen
    (kein Bezug zu LaTeX!)

Übungszettel mit LaTeX

- Erstellung eines Übungszettels mit LaTeX
  - Wichtigste Pakete in der Präambel
  - Name, Seitenzahl, Gruppe auf jeder Seite
- Alternativ Nutzung einer vorgefertigten Klasse, z.B.
  - https://www.mlte.de/latex/exercise-class/



Live Demo

Zusammenfassung

- Auszeichnungssprachen
- Markdown
- LaTeX
  - Grundlagen
  - Beamer
  - Formeln
  - Beispiel: Übungszettel

Inhaltsübersicht

1. Programmiersprache Python
   a) *Einführung, Erste Schritte*
   b) *Grundlagen*
   c) *Fortgeschritten*
2. Auszeichnungssprachen
   a) *LaTeX, Markdown*
3. Benutzeroberflächen und Entwicklungsumgebungen
   a) Jupyter Notebooks lokal und in der Cloud (Google Colab)
4. Versionsverwaltung
   a) Git, GitHub
5. Wissenschaftliches Rechnen
   a) NumPy, SciPy
6. Datenverarbeitung und -visualisierung
   a) Pandas, matplotlib, NLTK
7. Machine Learning (scikit-learn)
   a) Grundlegende Ansätze (Datensätze, Auswertung)
   b) Einfache Verfahren (Clustering, …)
8. DeepLearning
   a) TensorFlow, PyTorch, HuggingFace Transformers

Anhang: Präambel Übungszettel

```
\documentclass[
  oneside,      % Einzelne Seiten
  12pt,      % Schriftgroesse
  a4paper,      % Papier
  parskip=full    % Absatzstil
]{scrartcl}

% Dateikodierung
\usepackage[utf8]{inputenc}

% Trennung "deutsch"
\usepackage[ngerman]{babel}

% Schriftart
\usepackage[T1]{fontenc}
\usepackage{lmodern}

% Anfuehrungszeichen
\usepackage[german=quotes]{csquotes}

% Mathematische Symbole, Umgebungen
\usepackage{amssymb, amsfonts, amsthm, amsmath}

% Auflistungen und Grafiken
\usepackage{paralist, graphicx}

% Fuss- und Kopfzeile
\usepackage[headsepline,footsepline]{scrlayer-scrpage}
\makeatletter
  \clearpairofpagestyles
  \ifoot{\@subtitle}
  \ofoot{\@title}
  \ihead{\@author}
  \ohead{\pagemark}
\makeatother
```

Anhang: Inhalt Übungszettel

```latex
% Angaben ueber Aufgabenblatt und Autor
\author{Magnus Bender (LaTeX Beispiel)}
\title{Einführung in Web and Data Science\\Übungszettel 1}
\subtitle{Gruppe 5, Übung Mo. 10 Uhr}

\begin{document}
  % Titel
  %  Kann man auch einfach auskommentieren, dann spart man
  %  den Platz auf der ersten Seite!
  \maketitle

  % Ueberschrift ohne Laufende Nummer
  \section*{Aufgabe 1}
    % Lösungen, …

  \section*{Aufgabe 2}
    % Lösungen, …

\end{document}
```