

Ruprecht-Karls-Universität Heidelberg
Institut für Informatik
Sommersemester 2010
Seminar: Text Mining
Dozenten: Jannik Strötgen
Prof. Dr. Michael Gertz

Seminararbeit
Template für eine Seminararbeit

Name: Jannik Strötgen
Matrikelnummer: 1234567
Studiengang: Informatik (1. Fachsemester)
Email: stroetgen@uni-hd.de
Datum der Abgabe: 23. April 2010

Hiermit versichere ich **Jannik Strötgen**, dass ich die Hausarbeit mit dem Titel **Template für eine Seminararbeit** im Seminar **Text Mining** im **Sommersemester 2010** bei **Jannik Strötgen** und **Prof. Dr. Michael Gertz** selbstständig und nur mit den in der Arbeit angegebenen Hilfsmitteln verfasst habe. Zitate sowie der Gebrauch fremder Quellen, Texte und Hilfsmittel habe ich nach den Regeln wissenschaftlicher Praxis eindeutig gekennzeichnet. Mir ist bewusst, dass ich fremde Texte und Textpassagen nicht als meine eigenen ausgeben darf und dass ein Verstoß gegen diese Grundregel des wissenschaftlichen Arbeitens als Täuschungs- und Betrugsversuch gilt, der entsprechende Konsequenzen nach sich zieht. Diese bestehen in der Bewertung der Prüfungsleistung mit "nicht ausreichend"(5,0) sowie ggf. weiteren Maßnahmen.

Außerdem bestätige ich, dass diese Arbeit in gleicher oder ähnlicher Form noch in keinem anderen Seminar vorgelegt wurde.

Heidelberg, den 23. April 2010

Inhaltsverzeichnis

1	Einleitung	1
2	Allgemeine Informationen zum Template	1
3	Kapitel	2
3.1	Titelblatt	3
3.2	Antiplagiatserklärung	3
3.3	Inhaltsverzeichnis	4
3.4	Kapiteltiefe	4
4	Tabellen und Bilder	5
4.1	Tabellen	5
4.2	Bilder	6
5	Literaturliste	6
5.1	Bibliography Style	7
5.2	Arten von Literatur	7
5.2.1	Book	7
5.2.2	Article	8
5.2.3	Inproceedings	8

1 Einleitung

Dieses Dokument dient allen Teilnehmern des Seminars “Text Mining” als Vorlage zur Anfertigung Ihrer Hausarbeiten. In der Einleitung befindet sich normalerweise eine Hinführung zum Thema und eine Erläuterung der Problemstellung. Dabei ist es wichtig, dass in der Einleitung deutlich gemacht wird, was in den folgenden Kapiteln zu erwarten ist. In Ihrer Einleitung sollte außerdem bereits dargestellt werden, welche Ergebnisse Sie im Verlauf der Arbeit erzielt haben. In dieser Vorlage beispielsweise wird erläutert, wie Sie Tabellen oder Bilder verwenden können, und (ganz besonders wichtig!), wie Sie richtig zitieren. Und das Ergebnis dieser Vorlage ist hoffentlich, dass am Ende des Seminars Hausarbeiten abgegeben werden, die allen formalen Anforderungen entsprechen.

Allerdings handelt es sich bei diesem Template um ein *Starter Template*, d.h. es ist weder vollständig noch fehlerfrei (davon gehe ich zumindest aus). Eventuelle Fehler in Ihren Hausarbeiten dürfen nicht begründet werden mit:

In dem Template von der Kursseite war das aber auch so.

Es steht natürlich allen Teilnehmern frei dieses Template zu verwenden oder auch anzupassen. Selbstverständlich können Sie Änderungen vornehmen, um Verbesserungen zu erreichen. Sollten Ihnen irgendwelche Fehler auffallen, wäre ich Ihnen dankbar, wenn Sie diese nicht für sich behalten, sondern per Mail an mich schicken. Vielen Dank hierfür.

Am Ende der Einleitung sollten Sie beschreiben, wie der Rest der Arbeit aufgebaut ist:

Die Arbeit ist folgendermaßen strukturiert: Zunächst folgen in Kapitel 2 einige Informationen zu dem Template. Danach werden in Kapitel 3 einige Anmerkungen gemacht, wie die einzelnen Kapitel einer Seminararbeit aufgebaut werden sollen, bevor in Kapitel 4 ein paar Kleinigkeiten zu Tabellen und Bildern erläutert werden. Abschließend wird in Kapitel 5 erläutert, wie Sie richtig referenzieren. Im Vergleich zu Ihrer Arbeit enthält dieses Template keine Zusammenfassung, was natürlich nicht gut ist.

2 Allgemeine Informationen zum Template

In diesem Template wird davon ausgegangen, dass Sie bereits \LaTeX verwendet haben: Insbesondere wird nicht erklärt, wie ein Dokument kompiliert wird

usw. Sollten Sie noch nie L^AT_EX benutzt haben, sollten Sie davor eine Suchmaschine Ihrer Wahl nach einem “L^AT_EX Tutorial” fragen oder sich ein L^AT_EX Büchlein in der Bibliothek ausleihen und dieses dann ein bisschen durcharbeiten.

Das Template besteht aus fünf Dateien, wovon sich eine in einem Unterordner befindet.

- seminararbeit.tex (L^AT_EX Quelldatei)
- seminararbeit.pdf (pdf der kompilierten L^AT_EX Quelldatei)
- seminararbeit.bib (Bib_TE_X Quelldatei)
- plainnat-d.bst (Stylefile für die Bibliographie)
- pics/stopp_bild.png

In der L^AT_EX Quelldatei können Sie als allererstes die persönlichen Angaben ändern. Diese befinden sich in der Quelldatei relativ weit oben und beinhaltet folgende Zeilen:

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
Titel, Autor, Seminar, Semester, Dozent %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
\newcommand{\mytitle}{Template für eine Seminararbeit}
\newcommand{\myauthor}{Jannik Strötgen}
\newcommand{\myseminar}{Text Mining}
\newcommand{\mysemester}{Sommersemester 2010}
\newcommand{\mydozent}{Jannik Strötgen}
\newcommand{\mydozentTwo}{Prof. Dr. Michael Gertz}
```

Sobald Sie diese Definitionen angepasst haben, verändert sich bereits die Titelseite und die Antiplagiatserklärung.

3 Kapitel

Bevor in diesem Kapitel beschrieben wird, wie die Kapitel und Unterkapitel aufgebaut sein sollen, wird in Abschnitt 3.1 beschrieben, welche Informationen auf dem Titelblatt zu finden sein sollen. In Abschnitt 3.2 folgt die Beschreibung der Antiplagiatserklärung und in Abschnitt 3.3 wird beschrieben, wie das Inhaltsverzeichnis entstanden ist.

3.1 Titelblatt

Auf dem Titelblatt befinden sich folgende Informationen, die sich in drei Bereiche gliedern lassen(Kopf, Körper und Fuß). Im Kopfbereich sollen folgende Informationen zu finden sein:

- Name der Universität
- Name des Instituts
- Semester
- Name des Seminars
- Name des Dozenten

Im Körperbereich des Titelblatts müssen Sie folgende Angaben machen:

- Art der Hausarbeit
- Titel der Hausarbeit

Im untersten Teil, sollen die notwendigen persönlichen Informationen des Autors zu finden sein. Diese sind:

- Name des Autors
- Matrikelnummer
- Studiengang mit Fachsemesterzahl
- Ihre Emailadresse
- Datum der Abgabe

3.2 Antiplagiatserklärung

Nach dem Titelblatt folgt zunächst eine Antiplagiatserklärung. Diese enthält den Text

Hiermit versichere ich [**Name des Autors**], dass ich die Hausarbeit mit dem Titel [**Titel der Arbeit**] im Seminar [**Name des Seminars**] im [**Name des Semesters**] bei [**Name des Dozenten oder der Dozentin**] selbstständig und nur mit den

in der Arbeit angegebenen Hilfsmitteln verfasst habe. Zitate sowie der Gebrauch fremder Quellen, Texte und Hilfsmittel habe ich nach den Regeln wissenschaftlicher Praxis eindeutig gekennzeichnet. Mir ist bewusst, dass ich fremde Texte und Textpassagen nicht als meine eigenen ausgeben darf und dass ein Verstoß gegen diese Grundregel des wissenschaftlichen Arbeitens als Täuschungs- und Betrugsversuch gilt, der entsprechende Konsequenzen nach sich zieht. Diese bestehen in der Bewertung der Prüfungsleistung mit "nicht ausreichend"(5,0) sowie ggf. weiteren Maßnahmen.

Außerdem bestätige ich, dass diese Arbeit in gleicher oder ähnlicher Form noch in keinem anderen Seminar vorgelegt wurde.

Zusätzlich wird diese Seite mit Datum und Platz für die Unterschrift versehen. Die Antiplagiatserklärung steht auf einer eigenen Seite. Damit das so auch wirklich geschieht, wurde in diesem Template der Befehl

```
\newpage
```

verwendet. Außerdem soll auf dieser Seite keine Kopfzeile stehen. Deshalb wurde der Befehl

```
\thispagestyle{empty}
```

verwendet.

3.3 Inhaltsverzeichnis

Das Inhaltsverzeichnis kann mit dem Befehl

```
\tableofcontents
```

erstellt werden. Dabei werden alle Überschriften eingefügt, zusammen mit der Seitennummer der Seite, auf der das entsprechende Kapitel beginnt.

3.4 Kapiteltiefe

Im Allgemeinen soll sich die Gliederung auf drei Ebenen beschränken. Kapitelüberschriften, Abschnitts- und Zwischenüberschriften. Diese werden nach dem Schema „1; 1.1; 1.2; 1.2.1; 1.2.2“ usw. durchnummeriert. Hierfür dienen in diesem Template folgende Befehle:

Erste Zeile; erste Spalte	erste Zeile; zweite Spalte
Zweite Zeile; erste Spalte	zweite Zeile; zweite Spalte

Tabelle 1: So sieht eine Tabelle aus.

- `\section{Kapitel}`
- `\subsection{Abschnitt}`
- `\subsubsection{Unterabschnitt}`

4 Tabellen und Bilder

In diesem Kapitel wird beschrieben, wie Kapitel und Bilder eingefügt werden können. Wichtig ist hierbei, dass alle Bilder und Tabellen, die Sie eingefügt haben, eine Unterschrift haben und im Text referenziert werden. Dies funktioniert indem sie jeder Tabelle eine

`\caption{Ich bin eine Überschrift}`

und ein

`\label{tab:beispieltabelle}`

zuordnen. Im Text können Sie dann mit dem Befehl

`\ref{tab:beispieltabelle}`

auf die Nummer der Tabelle oder des Bildes zugreifen. Genauso greifen Sie übrigens auch auf die Kapitelnummern zu, wie Sie vielleicht bereits gemerkt haben. Also etwa so: In Kapitel 4.1 wird die Tabelle 1 gezeigt.

4.1 Tabellen

Manche Inhalte lassen sich sehr gut in einer Tabelle darstellen. Unnötige Informationen sind dagegen in Tabelle 1 dargestellt. Dafür wird die Tabelle im Text referenziert. Ist die Referenz nicht auf der gleichen Seite wie die Tabelle, kann mit dem Befehl

`\pageref{tab:tabelle}`

auch noch die Seitenzahl angegeben werden, auf der sich die Tabelle befindet. Also in etwa so: Auf Seite 5 wird die Tabelle 1 gezeigt.



Abbildung 1: Auch bei Bildern müssen Sie die Quelle angeben, wenn Sie sie nicht selbst erstellt haben (Quelle: http://www.chip.de/news/Endspurt-Letzter-Tag-der-Netzsperr-Petition_36913882.html).

4.2 Bilder

Dieses Kapitel dient ausschließlich dem Zweck, dass Sie sehen, wie sie eine Abbildung in ein L^AT_EX-Dokument einfügen können. Dafür dient die Abbildung 1 auf Seite 6.

5 Literaturliste

Wie sie bereits erfahren haben, soll Ihre Hausarbeit wissenschaftlichen Ansprüchen genügen. Dabei ist es besonders wichtig, dass Sie auch richtig und angemessen zitieren und referenzieren. Dass Sie dies tun, bestätigen Sie übrigens in Ihrer Antiplagiatserklärung.

Es gibt verschiedene Arten von wissenschaftlichen Artikeln, die Sie zitieren können. Bücher, wissenschaftliche Aufsätze aus Zeitschriften oder auch wissenschaftliche Artikel von Konferenzen, die nach Stattfinden der Konferenz veröffentlicht werden, in sogenannten *Proceedings* einer Konferenz. Da es sich um verschiedene wissenschaftliche Artikel handelt, werden auch alle etwas unterschiedlich behandelt. Wichtig ist auch, dass es Artikel mit einem, zwei oder mehreren Autoren gibt, was auch Auswirkungen hat. Da es sich bei dem vorliegenden Dokument jedoch lediglich um ein Starter Template handelt, können hier nicht alle Einzelheiten beschrieben werden.

Für Ihre Referenzen sollten Sie ein separates File anlegen. Sinnvoll ist oft auch Programme zu verwenden, mit denen man seine Bibliographie verwalten kann, z.B. JabRef¹. Bei dieser Gelegenheit haben wir auch gleich noch einmal Fußnoten kennengelernt. Und zusätzlich den Befehl:

```
\url{http://www.addresse.com}
```

5.1 Bibliography Style

In L^AT_EX besteht die Möglichkeit, verschiedene Styles zu verwenden, um auf die Bibliography zuzugreifen. In dem Paket dieses Templates ist ein angepasstes **natbib** File enthalten, das ein paar Anpassungen für die deutsche Sprache beinhaltet, sonst aber alle tollen Funktionen von natbib unterstützt. Zum Beispiel können damit auf die Autoren von bibliographischen Einträgen zugegriffen werden. Alle möglichen Funktionen sind auf der Webseite von natbib beschrieben².

Um das angepasste natbib File benutzen zu können, mussten in diesem Template folgende Sachen gemacht werden:

1. `\usepackage[numbers]{natbib}`
2. `\bibliographystyle{plainnat-d}`
3. Das File in dem die Literatur steht muss geladen werden:

```
\bibliography{seminararbeit}
```

5.2 Arten von Literatur

In diesem Kapitel werden einige Arten von Literatur vorgestellt. In der Literaturliste kann betrachtet werden, wie die verschiedenen Arten referenziert werden sollen.

5.2.1 Book

In diesem Kapitel wird ein Buch (book) zitiert. Wenn Sie sich bereits auf das Seminar vorbereitet haben, kennen Sie dieses Buch vielleicht schon. Bei Zitaten muss kenntlich gemacht werden, dass es sich um ein Zitat handelt. Dafür wird das Zitat in Anführungszeichen gestellt. Zitate müssen wörtlich

¹<http://jabref.sourceforge.net/>

²<http://merkel.zoneo.net/Latex/natbib.php>

übernommen werden, dürfen weder gekürzt, noch geändert werden, ohne dass dies mit eckigen Klammern deutlich gemacht wird. Sehen wir uns nun ein Beispiel an, in dem der Unterschied zwischen Text Mining und Data Mining erklärt wird. Text Mining ist ähnlich zu Data Mining, denn “analogous to data mining, text mining seeks to extract useful information” [1, S.1]. Bitte geben Sie vor allem bei Büchern immer auch die Seitenzahl an, wo sie das Zitat gefunden haben.

Statt wörtlicher Zitate ist es auch möglich zu paraphrasieren, z.B. so: In [1] haben Feldman und Sanger behauptet, dass Text Mining und Data Mining Ähnlichkeiten aufweisen.

5.2.2 Article

In diesem Kapitel wird auf ein System referenziert, das in einem wissenschaftlichen Artikel in einem Journal beschrieben wurde (article). Stellen Sie sich vor, wir schreiben etwas zu Ontologien und erzählen vorab, dass das schon andere gemacht haben. In etwa so:

Die in dieser Arbeit beschriebene Ontologie wurde rein zufällig aufgebaut und ergibt keinerlei Sinn. Anders die Ontologie YAGO, die mithilfe der Ressourcen Wikipedia und Wordnet aufgebaut wurde [3].

Auch bei solchen Publikationen können Sie direkt auf die Autoren zugreifen, also auf Suchanek et al. in diesem Fall. Dies geschieht mit

`\citeauthor{SuchanekEtAl2008}`

Dabei wird ein einzelner Autor direkt benannt, zwei Autoren als Autor1 und Autor2 und drei oder mehr Autoren als Autor1 et al.

5.2.3 Inproceedings

Ein drittes Beispiel für wissenschaftliche Publikationen sind Paper, die auf einer Konferenz eingereicht, angenommen, vorgestellt und anschließend in den Proceedings der Konferenz veröffentlicht wurden. Solche Artikel (sogenannte Inproceedings) haben eine etwas andere Darstellung in der Literaturliste.

Named Entity Recognition ist ein wichtiges Thema und kann für viele verschiedene Dinge benutzt werden, z.B. für die Verbesserung von Suchanfragen, indem versucht wird, Named Entities in Suchanfragen zu erkennen [2].

Literatur

- [1] Feldman, Ronan und James Sanger. *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press, New York, NY, USA, 2007.
- [2] Guo, Jiafeng, Gu Xu, Xueqi Cheng und Hang Li. Named entity recognition in query. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval (SIGIR '09)*, Seite 267–274, 2009.
- [3] Suchanek, Fabian M., Gjergji Kasneci und Gerhard Weikum. Yago: A large ontology from wikipedia and wordnet. *Web Semantics: Science, Services and Agents on the World Wide Web*, 6(3):203–217, 2008.