Sorbonne Université Université

Ecole Doctorale Informatique, Télécommunications et Electronique - ED130

Inria de Paris / Équipe ALMAnaCH

Thèse de doctorat

Discipline : Informatique

Présentée par

Pedro Javier Ortiz Suárez

Dirigée par

Laurent Romary et **Benoît Sagot**

Pour obtenir le grade universitaire de

Docteur de Sorbonne Université

On Modern NLP Methods for Document Enriching

Présentée et soutenue publiquement le 32 octobre 2021 devant le jury composé de :

Alonzo Church	Princeton University	Examinateur
Margaret Hamilton	University of Michigan	Rapporteur
Emmy Noether	Georg-August-Universität Göttingen	Examinateur
Laurent Romary	Inria - ALMAnaCH	Directeur
Benoît Sagot	Inria - ALMAnaCH	Co-directeur
Claude Shannon	MIT	Examinateur
Alan Turing	Princeton University	Rapporteur

Abstract

Scientific documents often use LaTeX for typesetting. While numerous packages and templates exist, it makes sense to create a new one. Just because.

Contents

1 Introduction

> In which the reasons for doing this Ph.D. are laid bare for the whole world to see and we encounter some answers to questions in which, frankly, only an extremely small number of people were interested in the first place.

This package contains a minimal, modern template for writing your thesis. While originally meant to be used for a Ph. D. thesis, you can equally well use it for your honour thesis, bachelor thesis, and so on—some adjustments may be necessary, though.

1.1 Why?

I was not satisfied with the available templates for LaTeX and wanted to heed the style advice given by people such as Robert Bringhurst [1] or Edward R. Tufte [2, 3]. While there *are* some packages out there that attempt to emulate these styles, I found them to be either too bloated, too playful, or too constraining. This template attempts to produce a beautiful look without having to resort to any sort of hacks. I hope you like it.

1.2 How?

The package tries to be easy to use. If you are satisfied with the default settings, just add

```
\documentclass{mimosis}
```

at the beginning of your document. This is sufficient to use the class. It is possible to build your document using either LaTeX|, XeLaTeX, or LuaLaTeX. I personally prefer one of the latter two because they make it easier to select proper fonts.

Package	Purpose
`amsmath`	Basic mathematical typography
`amsthm`	Basic mathematical environments for proofs etc.
`booktabs`	Typographically light rules for tables
`bookmarks`	Bookmarks in the resulting PDF
`dsfont`	Double-stroke font for mathematical concepts
`graphicx`	Graphics
`hyperref`	Hyperlinks
`multirow`	Permits table content to span multiple rows or columns
`paralist`	Paragraph ('in-line') lists and compact enumerations
`scrlayer-scrpage`	Page headings
`setspace`	Line spacing
`siunitx`	Proper typesetting of units
`subcaption`	Proper sub-captions for figures

Table 1.1: A list of the most relevant packages required (and automatically imported) by this template.

1.3 Features

The template automatically imports numerous convenience packages that aid in your typesetting process. Table 1.1 lists the most important ones. Let's briefly discuss some examples below. Please refer to the source code for more demonstrations.

1.3.1 Typesetting mathematics

This template uses `amsmath` and `amssymb`, which are the de-facto standard for typesetting mathematics. Use numbered equations using the `equation` environment. If you want to show multiple equations and align them, use the `align` environment:

$$V := \{1, 2, \ldots\} \tag{1.1}$$

$$E := \big\{(u, v) \mid \text{dist}(p_u, p_v) \leq \epsilon\big\} \tag{1.2}$$

Define new mathematical operators using `\DeclareMathOperator`. Some operators are already pre-defined by the template, such as the distance between two objects. Please see the template for some examples. Moreover, this template contains a correct differential operator. Use `\diff` to typeset the differential of integrals:

$$f(u) := \int_{v \in \mathbb{D}} \text{dist}(u, v) \,\mathrm{d}v \tag{1.3}$$

You can see that, as a courtesy towards most mathematicians, this template gives you the possibility to refer to the real numbers $\mathbb{R}$ and the domain $\mathbb{D}$ of some function. Take a look at the source for more examples. By the way, the template comes with spacing fixes for the automated placement of brackets.

1.3.2 Typesetting text

Along with the standard environments, this template offers `paralist` for lists within paragraphs. Here's a quick example: The American constitution speaks, among others, of (i) life (ii) liberty (iii) the pursuit of happiness. These should be added in equal measure to your own conduct. To typeset units correctly, use the `siunitx` package. For example, you might want to restrict your daily intake of liberty to 750 mg.

Likewise, as a small pet peeve of mine, I offer specific operators for *ordinals*. Use `\th` to typeset things like July 4^{th} correctly. Or, if you are referring to the 2^{nd} edition of a book, please use `\nd`. Likewise, if you came in 3^{rd} in a marathon, use `\rd`. This is my 1^{st} rule.

1.4 Changing things

Since this class heavily relies on the `scrbook` class, you can use *their* styling commands in order to change the look of things. For example, if you want to change the text in sections to **bold** you can just use

```
\setkomafont{sectioning}{\normalfont\bfseries}
```

at the end of the document preamble—you don't have to modify the class file for this. Please consult the source code for more information.

Part I

Data

2 OSCAR

2.1 GOCLASSY

3 Modern French Data

3.1 LEM17

3.2 PRESTO MAX

3.3 PRESTO GOLD

4 Ancient/Medieval French Data

4.1 BERTrade Corpus

5 Other Data

Part II

Models

6 CamemBERT

7 FrELMo

8 D'Alembert

9 BERTrade

Part III

Downstream tasks

10 Parsing

11 POS Tagging

12 Named-Entity Recognition

13 Text Normalization

14 Document Structuration

Part IV

Real World Application

15 Le Petit Larousse

16 BASNUM

17 Scientific Papers

18 Modern French Treebank

19 Named-Entity Recognition

corpora

Bibliography

1. R. Bringhurst. *The Elements of Typographic Style*. 4th ed. Hartley & Marks Publishers, Vancouver, British Columbia, Canada, 2012.
2. E. R. Tufte. *Envisioning information*. Graphics Press, Cheshire, CT, USA, 1990.
3. E. R. Tufte. *The visual display of quantitative information*. 2nd ed. Graphics Press, Cheshire, CT, USA, 2001.