DOCTORATE IN STUDY PROGRAM NAME

NOVA University Lisbon
month, year

DOCTORATE IN STUDY PROGRAM NAME
SPECIALIZATION IN SPECIALITY NAME

NOVA University Lisbon
month, year

Grammar of Time Time Series Data Mining in Steroids

This document was created with the (pdf/Xe/Lua)LaTeX processor and the NOVAthesis template (v6.9.14) [**novathesis-manual**].

Dedicatory lorem ipsum.

Acknowledgements

Acknowledgments are personal text and should be a free expression of the author.

However, without any intention of conditioning the form or content of this text, I would like to add that it usually starts with academic thanks (instructors, etc.); then institutional thanks (Research Center, Department, Faculty, University, FCT / MEC scholarships, etc.) and, finally, the personal ones (friends, family, etc.).

But I insist that there are no fixed rules for this text, and it must, above all, express what the author feels.

"You cannot teach a man anything; you can only help him discover it in himself." (Galileo)

Abstract

Regardless of the language in which the dissertation is written, a summary is required in the same language as the main text and another summary in another language. It is assumed that the two languages in question are Portuguese and English.

The abstracts should appear first in the language of the main text and then in the other language. For example, if the dissertation is written in Portuguese the abstract in Portuguese will appear first, then the abstract in English, followed by the main text in Portuguese. If the dissertation is written in English, the abstract in English will appear first, then the abstract in Portuguese, followed by the main text in English.

In the LaTeX version, the NOVAthesis template will automatically order the two abstracts taking into account the language of the main text. You may change this behaviour by adding

```
\abstractorder(<MAIN_LANG>):={<LANG_1>,...,<LANG_N>}
```

to the customization area in the document preamble, e.g.,

```
\abstractorder(de):={de,en,it}
```

The abstracts should not exceed one page and, in a generic way, should answer the following questions (it is essential to adapt to the usual practices of your scientific area):

1. What is the problem?
2. Why is this problem interesting/challenging?
3. What is the proposed approach/solution?
4. What results (implications/consequences) from the solution?

Keywords:

Resumo

Independentemente da língua em que a dissertação esteja redigida, é necessário um resumo na mesma língua do texto principal e outro resumo noutra língua. Pressupõe-se que as duas línguas em questão sejam o português e o inglês.

Os resumos devem aparecer primeiro na língua do texto principal e depois na outra língua. Por exemplo, se a dissertação for redigida em português, o resumo em português aparecerá primeiro, seguido do resumo em inglês (*abstract*), seguido do texto principal em português. Se a dissertação for redigida em inglês, o resumo em inglês (*abstract* aparecerá primeiro, seguido do resumo em português, seguido do texto principal em inglês.

Na versão LaTeX o template NOVAthesis irá ordenar automaticamente os dois resumos tendo em consideração a língua do texto principal. É possível alterar este comportamento adicionando

```
\abstractorder(<MAIN_LANG>):={<LANG_1>,...,<LANG_N>}
```

à zona de customização no preâmbulo do documento, e.g.,

```
\abstractorder(de):={de,en,it}
```

Os resumos não devem ultrapassar uma página e, de forma genérica, devem responder às seguintes questões (é essencial adaptá-los às práticas habituais da sua área científica):

1. Qual é o problema?
2. Porque é que é um problema interessante/desafiante?
3. Qual é a proposta de abordagem/solução?
4. Quais são as consequências/resultados da solução proposta?

Palavras-chave:

Contents

List of Figures

List of Tables

Glossary

This document is incomplete. The external file associated with the glossary 'main' (which should be called `template.gls`) hasn't been created.

Check the contents of the file `template.glo`. If it's empty, that means you haven't indexed any of your entries in this glossary (using commands like `\gls` or `\glsadd`) so this list can't be generated. If the file isn't empty, the document build process hasn't been completed.

If you don't want this glossary, add `nomain` to your package option list when you load `glossaries-extra.sty`. For example:

```
\usepackage[nomain]{glossaries-extra}
```

Try one of the following:

- Add `automake` to your package option list when you load `glossaries-extra.sty`. For example:

  ```
  \usepackage[automake]{glossaries-extra}
  ```

- Run the external (Lua) application:

  ```
  makeglossaries-lite.lua "template"
  ```

- Run the external (Perl) application:

  ```
  makeglossaries "template"
  ```

Then rerun LaTeX on this document.
This message will be removed once the problem has been fixed.

Acronyms

This document is incomplete. The external file associated with the glossary 'acronym' (which should be called `template.acr`) hasn't been created.

Check the contents of the file `template.acn`. If it's empty, that means you haven't indexed any of your entries in this glossary (using commands like `\gls` or `\glsadd`) so this list can't be generated. If the file isn't empty, the document build process hasn't been completed.

Try one of the following:

- Add `automake` to your package option list when you load `glossaries-extra.sty`. For example:

  ```
  \usepackage[automake]{glossaries-extra}
  ```

- Run the external (Lua) application:

  ```
  makeglossaries-lite.lua "template"
  ```

- Run the external (Perl) application:

  ```
  makeglossaries "template"
  ```

Then rerun LaTeX on this document.

This message will be removed once the problem has been fixed.

SYMBOLS

This document is incomplete. The external file associated with the glossary 'symbols' (which should be called `template.sls`) hasn't been created.

Check the contents of the file `template.slo`. If it's empty, that means you haven't indexed any of your entries in this glossary (using commands like `\gls` or `\glsadd`) so this list can't be generated. If the file isn't empty, the document build process hasn't been completed.

Try one of the following:

- Add `automake` to your package option list when you load `glossaries-extra.sty`. For example:

  ```
  \usepackage[automake]{glossaries-extra}
  ```

- Run the external (Lua) application:

  ```
  makeglossaries-lite.lua "template"
  ```

- Run the external (Perl) application:

  ```
  makeglossaries "template"
  ```

Then rerun LaTeX on this document.
This message will be removed once the problem has been fixed.

Chemical Symbols

This document is incomplete. The external file associated with the glossary 'chemical' (which should be called `template.chs`) hasn't been created.

Check the contents of the file `template.cho`. If it's empty, that means you haven't indexed any of your entries in this glossary (using commands like `\gls` or `\glsadd`) so this list can't be generated. If the file isn't empty, the document build process hasn't been completed.

Try one of the following:

- Add `automake` to your package option list when you load `glossaries-extra.sty`. For example:

  ```
  \usepackage[automake]{glossaries-extra}
  ```

- Run the external (Lua) application:

  ```
  makeglossaries-lite.lua "template"
  ```

- Run the external (Perl) application:

  ```
  makeglossaries "template"
  ```

Then rerun LaTeX on this document.
This message will be removed once the problem has been fixed.

1

Introduction

How to make Grammar of Time, make sense of Event Detection, etc..: - time dependencies between subsequences, such as word dependencies and association

- summarization of subsequences with letters

- annotation

1.1 Motivation

- In recent years, many wearables, many data is being acquired in shape of time series

- It becomes difficult to be able to analyze this amount of data, for this,

In recent years, the continuous increase in accessible wearable technology has contributed to a significant amount of data available. The continuous production of data from wearable devices through the usage of mobile phones, smartwatches, hearables, wristbands and other non-invasive wearable sensors has provided a valuable quantity of information. As reported in *Tankovska et al.*, the wearable devices usage has more than doubled in the interval between 2016 and 2019, reaching 722 million [**tankovska_23_2020**] [**novathesis-manual**], leading to a large volume of time series data being gathered in all possible scenarios, by monitoring patients in healthcare institutions, tracking everyday activities of humans, recording machines in industrial processes or workers motion while performing their tasks. It has never been so easy to gather data about any aspect of our life, work, education, society or industry. Of course, having relevant information about a subject is beneficial, but the overwhelming amount of data brings tremendous challenges in the ability to save, process, analyze and retrieve interpretable and meaningful information from which we can act upon[**bigdata**]. The usage of machine learning, with supervised methods, has brought ways of easing some of these tasks, but these challenges are all tied togheter. For instance, we can only use machine learning if we have good quality and labelled data (*Garbage-in, Garbage-out*, which is very hard to do with the amount of data being acquired, often unstructured and unlabeled.

Data preparation is essential for data analysis and the development of ?? applications.

After data acquisition, the effort in preparing it implies several challenges and is an active research subject. One of the critical issues in data preparation is reported to be the lack of labeled data. This process is sensitive and time consuming, which complexity increases with data quantity. Nevertheless, accurately labeled data is particularly relevant for data analysis and even fundamental to train supervised and semi-supervised ?? models [**roh2019survey**]. In the work of *Roh et al.* is mentioned that data scientists only rely on a small portion of the available datasets because it is too expensive to label all the data [**roh2019survey**]. This shows how paramount it is to have solutions that can improve the existing strategies in data labeling and reduce the required effort in doing so.

In addition, being time seres one of the most common data type in nature [**puttinghuman**], problems arise in any domain. In that sense, more tools should be developed with the intent of being more intuitive and letting the human be in the time series analytical loop [**puttinghuman**]. This would democratize time series analysis to non-experts and increase productivity of experienced data analysts.

Therefore, there is an increasing need for tools that can help accelerate information retrieval in data, summarize the data to find patterns of interest or filter irrelevant information and intuitively and expressively use search mechanisms for specific events or patterns.

This work proposes several solutions to time series data mining that contribute to ease the interpretation of the data, how to find relevant events and patterns, as well as expressive mechanisms for pattern search based on language.

In time series analysis, the data produced by such an amount of sensors can help the data science community develop tools to extract meaningful information from it, namely reporting, pattern recognition, event detection and segmentation of periodic signals, and classification, among other data mining tasks [**rodrigues2017noise**, **david_thesis**]. More importantly, having more data is even more beneficial when using ??, such as deep learning methods. There are numerous fields that would benefit from using these methods, namely healthcare, for the detection of physiological events, such as noise, sleep problems or epilepsy; on electrocardiogram (ECG) and/or electroencephalogram (EEG) [**cpd_medical_1**, **cpd_medical_2**, **cpd_medical_3**, **cpd_medical_4**, **dataset6**, **dataset7**]; climate change detection [**cpd_climate**]; speech recognition for audio segmentation [**cpd_audio**]; human activity analysis, in segmenting activities or changes in behaviour [**cpd_har_1**, **cpd_har_2**, **review_1**]; occupational health scenarios [**antonio**] and monitoring industrial processes for anomalies.

1) Achieving more with data 2) Make tools that ease the exploration of data 3) Make tools that help structure data into organized and labelled data 4) Improve the quality of the data used to train machine learning methods (Garbage in Garbage out) 5) Have visualization that make the understanding of the data more valuable

Many techniques exist to perform occupational health risk measures, but these rely on video inspection or other observational apparatus, from which specialists in occupational health domains are able to perform their evaluation. This evaluation typically relies on

existing screening tools, such as EAWS, OCRA, RULA, etc… but these have several limitations, namely the fact that these were designed for a "standard" subject, which can or not be the same for all workers.

Focus on this data and proxy datasets related with human motion data and human activity recognition

1.2 Context and Relevance

1.3 Research Questions

1.4 Thesis Structure

2

Theoretical Concepts

Tenho de fazer a associação entre os conceitos encontrados em séries temporais mas deveria falar em 1 - sensorização, 2 - o que é interessante fazer sobre os dados adquiridos e que seriam importantes no contexto do que é desenvolvido nessa tese, 3 -

Aqui, o leitor tem de perceber de que forma esta tese contribui para o domínio das séries temporais, com foco específico nos domínios abordados na tese, no âmbito da visualização, sumarização, anotação, segmentação

2.1 Time Series Fundamentals

2.2 Sensing Human Posture, Motion and Physiology

2.3 Linguistic Nature of Time Series

3

State of the Art

State of the art in the topics mentioned previously as well as what has been done considering the advances in time series data mining.

3.1 Information Retrieval from Time Series

Citing something online [**wiki:shuntingyard**, **flex**, **bison**].

3.2 Occupational Health Sensing and Problems

4

Data Description and Management

Explain our sources of data to explore the methods developed in this work. Show why you chose this type of data and which purpose it has been used.

4.1 Public Datasets

4.1.1 Classification Benchmark - UCR

4.1.2 UCI Machine Learning Repository

4.1.3 Physionet

4.1.4 CPD Benchmark

4.2 Acquired Datasets

4.2.1 Office Job Dataset

4.2.2 Industrial Job Dataset

5

Detection of Events and Summarization of Time Series

Have a small introduction about the problematic and inspiration in time series from the audio domain.

Será que poderia colocar isto sobre um tema maior? e depois ter tools desenvolvidas nesse ambito? Tools: Annotation, Segmentation, One-click segmentation of periodic events, summarization and profiling

5.1

5.1.1 Feature Representation

5.1.2 Self-Similarity

5.1.3 Novelty Search

5.1.4 Periodic Search

5.2 Time Series Profilling

5.2.1 Elements with Relevance

5.2.2 Minimalist Design

(Find better words to describe this)

5.2.3 Summarize Time Series

5.3 Further Developments

6

Text Mining Time Series

6.1 Synthatic Search on Time Series

We have made 2 approaches, one which uses a translation of the time series into the symbolic domain and another in which we associate extracted features with words.

I AM NOT HAPPY WITH THIS SEPARATION...IT FEELS LIKE IT COULD HAVE A MORE HOMOGENEOUS STRUCTURE

6.1.1 Time Series Representation

6.2 Towards Natural Language for Pattern Search

6.3 Classification of Time Series Documents

7

Data Description and Management

Explain our sources of data to explore the methods developed in this work. Show why you chose this type of data and which purpose it has been used.

7.1 Public Datasets

7.1.1 Classification Benchmark - UCR

7.1.2 UCI Machine Learning Repository

7.1.3 Physionet

7.1.4 CPD Benchmark

7.2 Acquired Datasets

7.2.1 Office Job Dataset

7.2.2 Industrial Job Dataset

A

NOVAthesis covers showcase

This Appendix shows examples of covers for some of the supported Schools. When the Schools have very similar covers (e.g., all the schools from Universidade do Minho), just one cover is shown. If the covers for MSc dissertations and PhD thesis are considerable different (e.g., for FCT-NOVA and UMinho), then both are shown.

B

Appendix 2 Lorem Ipsum

This is a test with citing something [**ecoop12-dias**] in the appendix.

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea

dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

I

Annex 1 Lorem Ipsum

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum

wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.