Text Summarization

**Author: Ioana-Maria Popescu**

**Abstract**

This is a template for preparing the papers for the Professional Internship classes. It is preferable that the papers be written in WORD, using the given template. You must convert your document also in Portable Document Format (.pdf). Please follow the instructions given in this template, for obtaining the final pdf form of your article.  We need both, the WORD source of your article and the pdf document containing the article. **The title mus****t be concise and suggestive. The abstract must clearly present the goal of the paper and the original contribution.** The paper must contain an Introduction section, where you present a motivation of your work, the existing similar research, the organization of the paper and underline the original part of your paper. You must also include in the end of the paper a Conclusion section, containing a comparison of your work with other existing ones, future directions of study, the importance of your work and conclude your main contribution. This section must be different from what you have written in the Abstract and Introduction.

Text summarization is the technique for generation a concise and precise summary of voluminous documents while focusing on the sections that convey useful information without losing the overall meaning.

This is a document whose intent is to present all the steps implemented to create a text summarization project.

# Introduction

Text summarization refers to the technique of shortening long pieces of text with the intention to create a coherent and fluent summary having only the main points outlined from the documents.

The Times New Roman font, is use in the article.

The setup settings for the article are:

Margins: left 2.5 cm

right 2.5 cm

top 3.2 cm

bottom 3.2 cm

On the first page, you must let an additional vertical space of 1 cm before the title

Paper format: A4, Alignment of paragraphs: 0.6cm

Title of the paper: Times New Roman, 18pt, Bold, Character spacing Extended by 0.2 pt

Author name: Times New Roman, 12 pt, Bold, Character spacing Extended by 0.2 pt

Abstract: Times New Roman, 10pt, indent from the article margins with 0.8cm left and right.

Title of section: Times New Roman, 16pt, Bold, Character spacing Extended by 0.4 pt, space before 16pt, after 10pt.

Normal text in the article: Times New Roman, 11pt, Character spacing Extended by 0.2 pt

# Implementation

## Format

The first step of the implementation is to specify a set of rules which all the words need to follow to have a unified and coherent set of words. These rules include eliminating punctuation, numbers, and white spaces.

### Extraction

Extraction represents an approach which involves picking up all the words contained in a document. To do that, the implemented project search key phrases presented in all given documents as in a pattern and extract them accordingly. As an example, a list of words is created by fetching all of the words contained in HTML tags such as “<text>” and “</text>”, respectively “<title>” and “</title>”. The implementation for getting the document’s topics is similar to the one explained above, but the searched tag is “metadata/codes”. Simulatively, the words and topics are counted in order to have an idea of the most and least words and topics in the documents.

Title of sub-subsection: Times New Roman, 12pt, Bold, Character spacing Extended by 0.2 pt, space before 16pt, after 6pt.

For entities like definitions, theorems, propositions, examples, lemmas, exercises, remarks, corollaries, use

- for the title of the entity: Times New Roman, 11pt, Bold, Character spacing Extended by 0.2 pt;

- numbering will be continuous in the article, without taking into account the number of section;

- for the body of the entity: Times New Roman, 11pt, Italic, Character spacing Extended by 0.2 pt;

- spacing: 6pt. before, 6pt. after the entity.

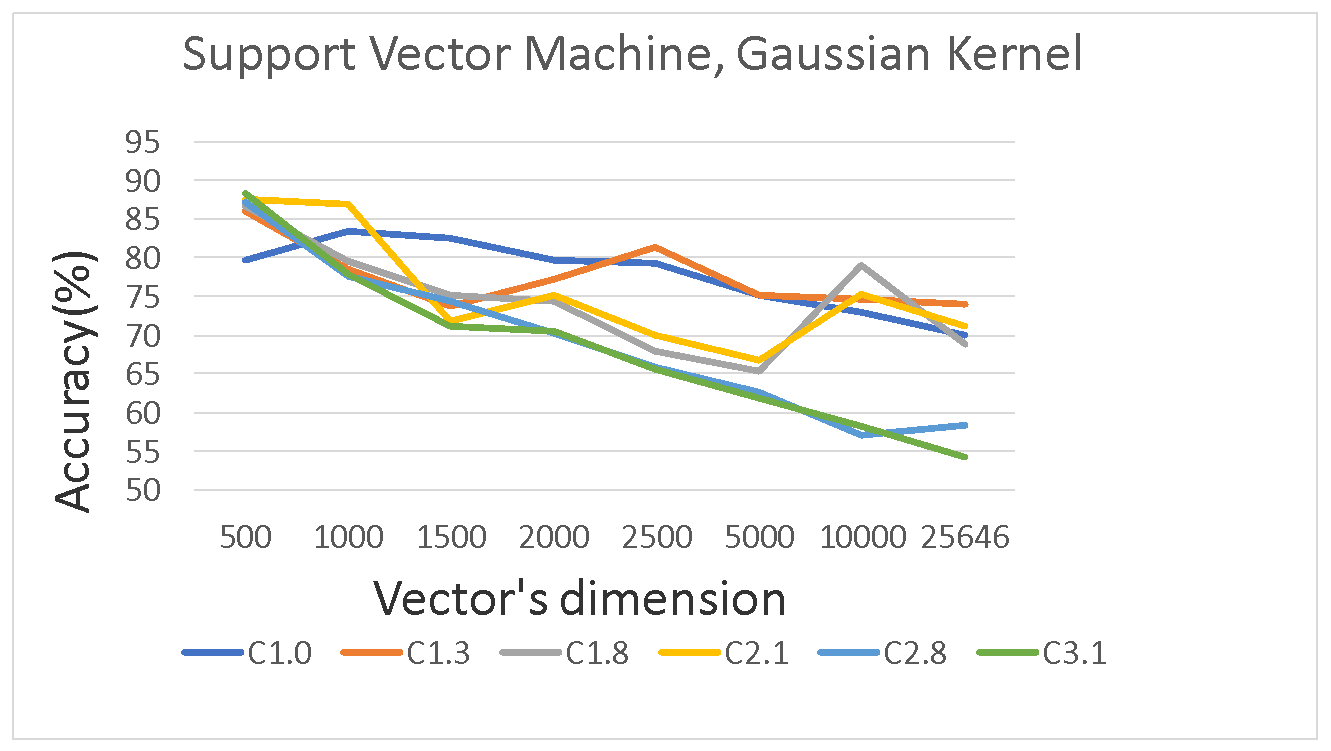
Example:

**Definition 1** *This is a definition.*

The equations will be centred, with the number align right, like in the next example

x=a+z (1)

All figures (Fig. 1:, Fig. 2:, …) and tables (Table 1:, Table 2:, …) should be continuous numbered with Arabic numerals. Table captions should be placed above the tables and Figure caption is always placed below the illustration. Short captions are centered, while long ones are justified. If the Table and Figure is taken from another part (web sites or other articles), the source of the Table/Figure is specified immediately after the caption.



**Figure 1**. Influence of vector dimension in classification accuracy (source: AMTM course)

**Table 1**. BoW document representation (source: AMTM course)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| *Document* | *classification* | *clustering* | *task* | *learning* | *supervised* | *unsupervised* | *technique* |
| D1 | 1 | 1 | 2 | 1 | 0 | 0 | 1 |
| D2 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| D3 | 0 | 1 | 1 | 1 | 0 | 1 | 1 |

It is not recommended to put code in the article.

## Acknowledgment

The Acknowledgement will be written before References, as normal text:

**Acknowledgement:** **The name of the professor who supervised the work or project (if any) should be written here. Use a sentence like** “This work was supervised by Professor *first\_ name last\_name*, from *affiliation*”.

**In the Acknowledgement section should also be noted the founder (if any) in the form like “**This paper is founded from the research Grant M12” or “This work benefits from founds given by xx”.

# References

For the references’ section:

- for the title of section References use the same settings as for other section, that is: Times New Roman, 16pt, Bold, Character spacing Extended by 0.4 pt, space before 16pt, after 10pt.

- for the items: Times New Roman, 10pt, Character spacing Extended by 0.2 pt. The name of the cited books and the names of the journals or proceedings where are published the cited articles, will be Italic. The references’ items were obtained as a numbered list with numbered position aligned left at 0cm. and the text position having the characteristics: tab space after 0.6cm, indent at 0.6cm.

1. Carl de Boor, *A practical guide to splines*, Springer-Verlag, New York Heidelberg Berlin, 1978.
2. R.A. DeVore and G.G. Lorentz. *Constructive Approximation*. Springer Verlag, Berlin Heildelberg New York, 1993.
3. T. Popoviciu. Introduction à la théorie des différences divisées. *Bull. Math. de la Soc. Roumaine des Sci.,* XLII: 65–78, 1940.

At the end of the article will be situated the details regarding the affiliation of the authors.

NAME of Author

Name of the Institution1

Name of department1

Address of the institution1

COUNTRY 1

E-mail: name1@server1