

# Self-Organizing Maps

## Project description

1st Semester of 2023-2024

Author

george-alexandru.dina@s.unibuc.ro

### 1 Introduction

- First of all, I chose this project because it is an interesting one and includes a topic that is increasingly widespread, that of neural networks.
  - The aim of this project is to present, implement and search for information to answer questions related to the topic of Self-Organizing Maps.
  - I started this project by making the PowerPoint presentation. During the presentation, we initially discussed what self-organizing maps mean, what the algorithm is, how it is implemented. Later I explained their most common application, that of Clustering. I have provided explanations and examples here. Starting from this step, I also made an implementation in Python for this situation. Later, during the presentation, we also discussed the other uses of Self-Organizing Maps and answered questions.
  - As an author, while carrying out this project I discovered what Self-Organizing Maps means, how it is implemented, how a neural network works and how to solve certain problems in everyday life.
- In the future, I would like to expand my knowledge and discover even more applications of this concept, also for others who are similar to him. In addition, I would like to deepen what I have already learned. For example, from the point of view of implementation, to make it recognize texts in different languages, with different punctuation and to become more interactive. (Possibly a graphic part, an input, so that you don't interact directly with the code when you want to find out what it looks like dividing your text into x clusters).

### 2 Approach

For the first time, I created the repository on github. It can be accessed by following this link: <https://github.com/alexdn7/archaeology-project>. There are all the files included in this project. This means the source code file (.py), the input file, the presentation and the pictures with the results.

After that, I started by making the PowerPoint presentation. In order to obtain the necessary information, I searched through several sources. All the sources necessary for the realization of this project are mentioned at the end of the PowerPoint presentation, and each cited section has a marker that refers you to the respective source.

After gathering the necessary information for the presentation part, I also realized an implementation of the Clustering problem using Self-Organizing Maps (which means unsupervised learning, trained using competitive learning ). This implementation is done in Python, using PyCharm IDE.

Within it, an English text is taken from a file, I divided it into words. After this step, it was necessary to transform this input consisting of words into a vector. For this I applied TfidfVectorizer from sklearn. Then I've applied MiniSom (function that implements Self-Organized Maps in Python) and I trained the model. I was able to group the words according to similarity.

After this part, I wrote a few more lines of code to display the result in the desired format. Practically, I knew for each word which cluster it corresponds to and based on this I grouped and made both the hitorigram and the table.

As an example, I obtained the following results (Figure 1, Figure 2).

The realization of the project lasted a few days, and the challenges that consumed quite a lot of time consisted of the lack of resources that specifically answered what was needed, the lack of complete

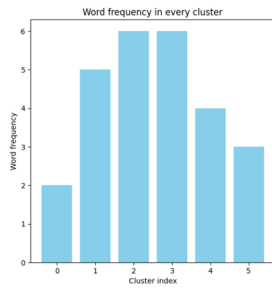


Figure 1: Word frequency for every cluster

of	are	there	many	passages	variations
of	issue	own	available	have	majority
the	but	suffered	from	by	
in	alteration	none	or		
randomised	injected	humour			
	words	which			

Figure 2: Words grouped by cluster

implementations that could be understood and the sources that bypassed the central topic, but they consumed a lot of time on introductions.

### 3 Limitations

A very big limitation was in terms of study resources. As I mentioned in section 2, it was quite difficult to find all the necessary information.

Another limitation would be from the point of view of the text provided as input. As I mentioned in the previous sections, I offered a text in English with spaces, without serious punctuation. When I offered a text that combines words from several languages, the answer was perhaps not the most desired.

There are also limitations in terms of the physical resources of the equipment. During the running of the algorithm, as the number of iterations/epochs increased, it became increasingly difficult to run.

### 4 Conclusions and Future Work

Carrying out this project I learned many new things. Personally, I didn't know anything about self-organizing maps and I didn't have much knowledge in this area of machine learning. But, as we explored the sources and realized the project, we learned many new things.

It is clear that there can be improvements, both in terms of documentation and implementation (Some of them were mentioned in Section 1).