

*Dual Bachelor in Data Science and Telecommunications Technologies Engineering 2020-2025*

***Final Project***

***Machine Learning Applications***



**Alejo González García (100454351)**

**Alonso Madroñal de Mesa (100454449)**

**Daniel Toribio Bruna (100454242)**

**Andrés Navarro Pedregal (100451730)**

# 1.Introduction

This project consists of a multiclass classification of previously preprocessed poems, that will be assigned to a period. In order to carry out this task, we make use of the tools explained in class regarding Natural Language Processing and Machine learning tools, such as feature extraction and selection.

# 2.Task 1: Text Preprocessing and vectorization

## 2.1. The dataset

In order to start the project, we first needed to obtain the dataset to work with. We could either download one available online, or create it ourselves. We opted for the second option, and we obtained the information from the web [Poetry Foundation](https://www.poetryfoundation.org/poems/browse#page=1&sort_by=recently_added).

For the creation we used a library called request to obtain the response of the get request and then we obtain the json response where we extract information like author, title, link to the poem, tags and snippet. With the link to the poem and using BeautifulSoup we get the html information. Poems were mainly in two formats, text or images. For the first one we just extracted the text that was inside the div with class 'o-poem', and for the second one, we used an OCR, in this case, pytesseract, to extract the text from the image that was in the div with class ‘c-assetStack-media’.

In the web [Poetry Foundation](https://www.poetryfoundation.org/poems/browse#page=1&sort_by=recently_added) only 5535 poems from the 47388 total poems have a period assigned, so we will have two different datasets, one composed with all the poems that will be used for the topic modeling task, see Figure 1, and another with only the poems that have a period and they can be used for the classification task, see Figure 2.

The resulting dataset contains the following information:

* An id
* The title of the poem
* The author
* A snippet of the poem
* The link to where the poem is (text or image)
* The categories that the poem has
* The period of the poem
* The text of the poem



Figure 1. Sample of the dataset obtained without periods



Figure 2. Sample of the dataset obtained with periods

## 2.2. Modifications and filtering of the dataset