**Readme / Code user’s Manual (included in report)**

Title: SpringerLink Discipline Prediction

Authors:

* Daniel de las Cuevas Turel - 100406666
* Claudio Sotillos Peceroso - 100409401
* Pablo Reyes Martín - 100409333
* Bosco de Enrique Romeu - 100406718

This file contains three python notebooks and three folders plus the report.

\*\*Python notebooks brief summary:

* Web\_scraping: It contains the code and explanations of our web scraping procedure. ¡WARNING! If you are willing to execute the code of this notebook, be aware that it will take a long time in executing (it took us around 5 hours, even with "ray framework" parallelization).
* Task 1: In it, we fulfilled the first task. It consists of an initial preprocessing and analysis of our data set, the text processing pipeline, the LDA topic modeling and how we obtained the Nodes/Edges CSV files for our Graph Analysis.
* Task 2: For task 2 we have decided to compare the Abstract vectorization approaches, TF-IDF  and Word2Vec with different classifiers. ¡WARNING! If you are willing to execute the code of this notebook, it will take a lot of time. Concretely, the cells that take more time in executing are those related to the TSNE representation.

In order to run the code inside these notebooks you just have to change your Drive path.

Substitute it with the path where you have stored the final project file.

Just change this: |----------------------------------------------------|

v v

local\_folder = '/content/drive/MyDrive/ML Applications/Final Project/'

\*\*Description of the folders:

* Non-Clean Datasets: In it, we have stored the Data sets obtained with web scraping. The "firstphase.pickle" is the one which belongs to the outer web scraping (columns: [title,url, discipline name]). The "Dataset.pickle" is the one which belongs to the inner web scraping [titles,Abstract,KeyWords,Authors,Journal,PublicationDate,Accesses,urls,target]).

The "Dataset\_clean.pickle" is the same as the previous one, but with some cleaning modifications.

* Task 1 Checkpoints: It contains the definitive Data set after the initial cleaning of the initial web scraping Data set. It also contains the several checkpoints performed along the task 1 notebook.
* Task 2 Checkpoints: The same as task 1 checkpoints but non-structured.