ΥΣ19 **Artificial Intelligence II (Deep Learning for Natural Language Processing)**
Fall Semester 2022
Homework 1
25% of the course mark
Announced: November 8, 2022 Due: November 29, 2022 before 23:59)

In this homework you have to develop a sentiment classifier using logistic regression for the dataset `imdb-reviews.csv` which is provided. Each row in the dataset contains a URL for an IMDb movie (`https://www.imdb.com/`), a score between 0.0 and 10.0 and a review text. Your classifier should deal with 2 classes: *negative* sentiment (score between 0.0 and 4.0) and *positive* sentiment (score between 7.0 and 10.0). There are no reviews with intermediate scores in the dataset.

Before you do the homework, makes sure that you have studied the relevant slides of the course ("Introductory concepts of machine learning" and "Regression") and the relevant chapters 4 and 5 of the "Speech and Language Processing" book of Jurafsky and Martin (`http://web.stanford.edu/~jurafsky/slp3/`) or any other relevant literature you may find useful.

It is your responsibility to choose all the details of developing a good model (e.g., how to partition the given dataset into training and validation datasets, whether to do cross validation, whether to do regularization, which gradient-based training algorithm to use, how to choose the hyperparameters of the algorithm, how to make sure that your model does not underfit or overfit etc.).

You should plot learning curves that show that your models are not overfitting or underfitting.

You should use the toolkit Scikit-Learn (`https://scikit-learn.org/stable/`) and evaluate your classifier using *precision, recall* and *F-measure.*

Your code should be written in a way that your model can be evaluated on the test set by simply passing the path of the test file to a specific variable.

You should hand in:

1. A pdf document with a detailed explanation of your solution including citations to relevant literature that you might have used in developing your solutions. If you use LaTeX for your document, you will get a bonus 5%. In this case, you have to hand-in the LaTeX source files too.

2. One Colab notebook (`ipynb` file using `https://colab.research.google.com/`) containing your code.