| | | | |
|---|---|---|---|
| **Course Code:** | COMP-546DL | **Document Title:** | Project |
| **Course Title:** | Deep Learning and Reinforcement Learning | **Lecturer:** | Michalis Agathocleous |

## Project Spam Filtering: Deep Learning (Deadline: 31/05/2024)

The main goal of this project is to apply Deep Learning methods in the context of a Spam Filter Application.

A Spam Filter is a program that is used to detect unsolicited and unwanted email and prevent those messages from getting to a user's inbox. Like other types of filtering programs, a spam filter looks for certain criteria on which it bases judgments. You have to investigate a Deep Learning method to automatically distinguish emails to spam or not spam classes.

**Project Tasks:**

1. Download the given dataset.
2. Create your own training and validation dataset.
3. Use Keras and/or tensorflow Neural Network API to build two different classifiers to handle this problem.
4. Use your own graphs to compare the classifiers and present your results.
5. Produce a report detailing what you tried, what the results are and what you would select as the final model. Also provide an accompanying python notebook and your final model. The notebook must contains the code of your experiments. The last cell must load your final model and apply it to a test set. The report can be within the notebook, similarly to the tutorials used in the course.

**Data:**

This is a csv file containing related information of 5172 randomly picked email files and their respective labels for spam or not-spam classification.

The csv file contains 5172 rows, each row for each email. There are 3002 columns. The first column indicates Email name. The name has been set with numbers and not recipients' name to protect privacy. The last column has the labels for prediction : 1 for spam, 0 for not spam. The remaining 3000 columns are the 3000 most common words in all the emails, *after excluding the non-alphabetical characters/words. For each row, the count of each word(column) in that email(row) is stored in the respective cells.* Thus, information regarding all 5172 emails are stored in a compact dataframe rather than as separate text files.

**Evaluation:**

Your method will be tested on an independent testing set. Your method will be evaluated based on F-measure:

F-measure(spam)= (2*Precision*Recall)/(Precision + Recall)

Final grade will be calculated based on the below table:

| Max Points | Task |
|---|---|
| 0.5 | Data Pre-processing |
| 2.0 | Methodology |
| 0.5 | Code based on Keras |
| 0.5 | Report |
| 0.5-\|1-r\|x0.5 | Where r is your method ranking based on F-measure |

**Deliverables:**

You have to send a .zip file with the name <your_id> .zip at ...@unic.ac.cy.

Include:

a. A detailed report for the above Tasks. Please explain in detail your:
   a. Model architectures
   b. Methodology
   c. Method evaluation
b. An accompanying python notebook containing your experiments and your final model. The last cell must load your final model and apply it to a test set.
c. Any other related code to this project.