

# School of InfoComm Technology

**Deep Learning Assignment**

Diploma in CSF / FI / IT

Apr 2022 Semester

**ASSIGNMENT 2**

(40% of DL Module)

4th Jul 2022 – 12th Aug 2022

**Submission Deadline:**

**Presentation: 12th Aug 2022 (Week 17),**

**Report: 12th Aug 2022 (Friday), 11:59PM**

|  |  |  |
| --- | --- | --- |
| **Tutorial Group** | **:** | **P01 / P02 / P03 / P04** |
| **Student Name** | **:** |  |
| **Student Number** | **:** |  |

**Penalty for late submission:**

10% of the marks will be deducted every calendar day after the deadline.

**NO** submission will be accepted after 21st Aug 2022 (Sunday), 11:59PM.

# Assignment Specifications

# Problem 1 (Group：4-5 students per group) (60%)

## 1.1 Objective

Build a sentiment analysis model to predict the App review score based on the Google Play Store reviews.

## 1.2 Dataset

Each group need to collect the Google Play Store reviews by themselves.

* First, each student needs to select one google play store App and collect the review data using **google\_play\_scraper** 
  + the sample codes are provided in “**data\_collection.ipynb**”
  + the instructions on how to get the App ID is provided in Appendix

## 1.3 Suggested Tasks

You are recommended to tackle this problem by following the steps below. A Jupyter Notebook (**Assignment\_2\_p1.ipynb**) is provided for you to write the codes based on these steps.

**Step 1 – Data Loading and Processing (Group)**

* Data Collection
  + Each student needs to research online and select an App
  + Rank the selected App within the group
  + Based on the assigned App to your group, collect the review data from Google Play Store
  + Extract the content and score columns from the review data (min. 5000 data each)
* Data Processing
  + Convert the content and score into numeric tensors (X & y)
  + It may be helpful to cleanse the content before encode them into numbers
* Data Sampling
  + Split the dataset (X & y) into training set (X\_train & y\_train) and testing set (X\_test & y\_test).

**Step 2 – Develop a Sentiment Analysis Model using Training Data (Individual)**

Each student is required to develop **at least One model**, which should have:

* + Word Embedding Layer: You can use the pre-trained word embeddings (e.g. GloVe), or simply train the word embeddings using the provided dataset.
  + RNN Layers (e.g. LSTM, GRU and etc.) or 1D Conv Layer or other layer suitable for sentiment analysis task
* During training phase, please split the training data into training samples and validation samples. (**Hint**: use validation\_split in model fit() function)
* You are suggested to follow the universal machine learning workflow to develop the model and improve the model performance by tuning model hyperparameters, i.e.
  + Start with a baseline model
  + Scale up the model until it overfits
  + Regularize the model and tuning the model hyperparameters
* Analyze the Model Performance during training phase
  + **Remember** **to record all the model performance curves** (i.e. training and validation accuracy, training and validation loss scores) during training phase

**Step 3 – Evaluate the Models using Testing Data (Individual & Group)**

* Evaluate the Models Performance using X\_test & y\_test
* Compare and discuss the model performance
* Recommend the best model

**Step 4 – Use the best Model to make prediction (Group)**

* Use the input() function to record the user input
* Convert the user input into numeric tensor
* Feed the numeric tensor into your model and see whether the model can output a correct emoticon

## 1.4 Report Format & Content Guidelines

Write a report with the following sections and content guidelines. You are free to include other relevant information you deem necessary in the sections.

*(Note: For a page with 1-inch margins, 12-point Arial font, and minimal spacing elements, a good rule of thumb is* ***500 words*** *for a single-spaced page)*

|  |  |  |
| --- | --- | --- |
|  | **Suggested Report Sections & Content Guidelines** | **Word Count** |
| 1. | Overview *(Group)*   * Describe the problem, the objective, and the approach. | Min: 500 words  Max: 1000 words |
| 2. | Data Collection and Processing *(Group)*   * Describe why you select this App and how you collect review data using google\_play\_scraper, including each student’s App choice and how you rank the 4-5 Apps within your group * Describe how you cleanse the data and convert the data into numeric tensors * Describe how you sample the data and why you need to sample the data | Min: 1000 words  Max: 2000 words |
| 3. | Develop the Sentiment Analysis Models using Training Data *(Individual: One Section per Student)*   * Describe how you build and train the models, tune the model hyperparameters to improve the model performance * Analyze the model performance during training phase | Per Student  Min: 500 words  Max: 1000 words |
| 4. | Model Evaluation using Testing Data *(Individual & Group)*   * Analyze the model performance during testing phase * Compare and Discuss the model’s performance * Recommend the best model and explain why | Min: 500 words  Max: 1000 words |
| 5. | Use the Best Model to Make Prediction *(Group)*   * How to apply the model on a real-life text input * Explain and Analyze the model prediction | Min: 500 words  Max: 1000 words |
| 6. | Summary *(Group)*   * Summarize your model performance and provide suggestions for further improvement. | Min: 500 words  Max: 1000 words |

# Problem 2 (Individual) (40%)

## 2.1 Objective

Implement a Recurrent Neural Network (RNN) to create an English language character generator capable of building semi-coherent English sentences from scratch, by building them up **character-by-character**. In particular, we will be using a complete version of J. K. Rowling’s famous book Harry Potter and the Philosophers Stone.

We can train a deep learning model to generate text automatically, character-by-character, by showing the model many training examples so that it can learn a pattern between text inputs and potential character outputs. With this type of text generation, each input is a string of valid characters like this one:

*“dogs are grea*”

The corresponding output is the next character in the sentence - which is 't' (since the complete sentence is 'dogs are great'). We need to show a model many such examples in order for it to make reasonable predictions.

## 2.2 Dataset

Download the dataset from POLITEMall:

* **Harry\_Potter\_Book1.txt**: the full text of Harry Potter and the Philosophers Stone

## 2.3 Suggested Tasks

You are recommended to tackle this problem by following the steps below. A Jupyter Notebook (**Assignment\_2\_p2.ipynb**) is provided for you to write the codes based on these steps.

**Step 1 – Data Loading and Processing**

* Data Loading
  + Open and read the Harry\_Potter\_Book1.txt file
  + Check the total number of characters in the original text
* Data Processing
  + Perform basic data cleansing by removing unnecessary characters
  + Identify a list of unique characters and punctuations in the clean text
  + Prepare data into training text and labels (X & y) using the “sliding window” method
  + Perform one-hot encoding on X & y and convert them into binary arrays

**Step 2 – Develop Sequence Generator Model**

* Each student is required to develop **at least ONE model**, which must consist of RNN Layers (e.g. LSTM, GRU and etc.)
* During training phase, please split the data into training samples and validation samples. (**Hint**: use validation\_split in model fit() function)
* You are suggested to follow the universal machine learning workflow to develop the model and improve the model accuracy, i.e.
  + Start with a baseline model
  + Scale up the model until it overfits
  + Regularize the model accordingly
* You are also suggested to use the trained models to generate new texts (e.g. 400 characters) applying proper sampling strategy and adjust the sampling function parameters (e.g. temperature) accordingly.
* Analyze the Model Performance during training phase
  + The model performance will be a combination of accuracy and the quality of generated texts. Feel free to develop your own metrics of evaluation on the quality of the generated texts.
  + **Remember** **to record all the model performance curves** (i.e. training and validation accuracy, training and validation loss scores) and **the generated texts** during training phase

**Step 3 – Use the developed Model to generate texts**

* Use the input() function to record new input from user
* Encode the input using one-hot encoding
* Feed the encoded input into your model and let the model generate 400 characters
* See whether the generated texts make sense to you

## 2.4 Report Format & Content Guidelines

Write a report with the following sections and content guidelines. You are free to include other relevant information you deem necessary in the sections.

*(Note: For a page with 1-inch margins, 12-point Arial font, and minimal spacing elements, a good rule of thumb is* ***500 words*** *for a single-spaced page)*

|  |  |  |
| --- | --- | --- |
|  | **Suggested Report Sections & Content Guidelines** | **Word Count** |
| 1. | Overview   * Describe the problem, the objective, and the approach. | Min: 500 words  Max: 1000 words |
| 2. | Data Loading and Processing   * Describe how you load and cleanse the data * Describe how you prepare the data and perform one-hot encoding | Min: 500 words  Max: 1000 words |
| 3. | Develop the Sequence Generator Model(s)   * Describe how you build and train the model(s), tuning model hyperparameters * Analyze the model performance, including both accuracy and the quality of generated texts | Min: 1000 words  Max: 2000 words |
| 4. | Use the developed Model to Generate Texts   * How to apply the model on a real-life text input * Analyze the generated texts and see whether it makes sense for you | Min: 500 words  Max: 1000 words |
| 5. | Summary   * Summarize your model performance and provide suggestions for further improvement. | Min: 500 words  Max: 1000 words |

# Recorded Presentation

Each group will be allotted a total of 45 minutes to present their Group & Individual work using slides. You are strongly suggested to add in screenshots, diagrams, images into your slides and practice the presentation in advance to make sure you can complete within 45 mins.

* + Problem 1 (20 mins):
    - Individual Portion (12-15 mins): 3 mins per student
    - Group Portion (5 mins)
  + Problem 2 (16-20 mins): 4 mins per student

The recorded presentation must be through MS Teams with all student’s camera on. The submission will be made by submitting the link in the comment section of the Assignment 2 submission.

# Deliverables

For this assignment, you must submit all the following:

Each Group is required to submit the following:

1. A Final Presentation Slides Deck

* **One** final deck of slides includes the Group Presentation Slides for Problem 1 and all the Individual Presentation Slides for Problem 2.

1. A Final Softcopy Report

* The Final Report includes **One** Group Report for Problem 1 and **Five** Individual Reports for Problem 2.

1. A zip file with Jupyter Notebook Files (.ipynb)

* The zip file includes **One** Group Jupyter Notebook for Problem 1 and **Four or Five** Individual Jupyter Notebooks for Problem 2.

1. Create a folder name **[Class]\_[Group]** and include items 1 to 3 above into it and Zip up the folder for submission in POLITEMall.
   * Deadline for the zip file submission is **Friday 12th Aug 2022, 2359 hours**
2. The **link to your recorded presentation**

* Submit the link of your recorded presentation in the **Comments** section of your Assignment 2 Submission
* Extend your Teams recording expiration date to **No expiration**
  + - Open your video recording in a browser and select the **i** icon at the upper right of the screen to launch the details pane and choose **No expiration** by clicking the calendar icon
    - Details are outlined in the following URL:

<https://support.microsoft.com/en-us/office/record-a-meeting-in-teams-34dfbe7f-b07d-4a27-b4c6-de62f1348c24#bkmk_view_change_expiration_date>

* **Edit the sharing of the recorded video** to be viewable by the tutor as outlined in this URL:

<https://docs.microsoft.com/en-us/stream/portal-edit-video>

* **Note**: if tutors cannot access the recorded video, and students need to be reminded by tutors to reset permission, students will be penalized for non-conformance.

# Grading Criteria

|  |  |  |  |
| --- | --- | --- | --- |
| A Total of 100% | **Problem 1** | | **Problem 2** |
| Group | Individual | Individual |
| **Report** | 15% | 15% | 20% |
| **Presentation** | 15% | 15% | 20% |

|  |  |
| --- | --- |
|  | **Grading Criteria** |
| **Presentation** | 1. Quality of work 2. Flow of presentation based on content guidelines 3. Quality of presentation slides 4. Presentation and articulation skills |
| **Final Report & Jupyter Notebooks** | 1. Quality of work 2. Completeness of report based on suggested report sections and content guidelines 3. Clarity of report, Quality of analysis and discussions 4. Use of proper visual aids and Use of proper grammar |

# Appendix

How to get App ID from Google Play Store?

Step 1: Go to <https://play.google.com/store/apps>

Step 2: Select the App you are interested and you can find App ID highlighted per below

