|  |  |  |
| --- | --- | --- |
| BUE final logo  Informatics and Computer Science | **Coursework Brief Proofing & Printing Confirmation Sheet** | |
| Module Title: **Introduction to Natural Language Processing (NLP)** | | Module Code  22CSAI05H |
| Module Leader **Prof. Pester** | | Semester  **Two** |
| Proofed by **Prof. Barakat** | |  |

**I hereby confirm that:**

x

|  |  |
| --- | --- |
| * This coursework brief has been proof-read (spelling and grammar) * This coursework brief assesses the ILOs for the module * This coursework brief follows the approved template * All questions (and sub questions) have their marks specified | x |
| x |
| x |
|  |

**Signed (Proof Reader): Nahla Barakat**

**Signed (Module Leader) Andreas Pester**

|  |  |  |
| --- | --- | --- |
| BUE final logoInformatics and Computer Science | **22CSAI05H**  **Project and Report**  **2022/2023** | |
| Module Title: **Introduction to Natural Language Processing (NLP)** | | |
| Module Leader **Prof. Pester** | | Semester  **Two** |
| Assessment Weight  **35 % of the total course mark** | Due Date  **Week 6** | |

**Project Specification:**

You are required to deliver a group assignment (2 students).

**Project Deliverables**

|  |  |
| --- | --- |
| **Date** | **Deliverable** |
| Week 3 | Report Proposal and Specification |
| Week 6 | Final Report |
| Week 7 | Discussion and Presentation submission |

Specific details of each of this report deliverables will be given in due time.

**Assessment and Feedback:**

- Marking criteria will be attached to each phase specification

- Discussion feedback will be given to the report proposal and specification

- Written specific feedback to the final report will be sent to each person and discussed during sessions dedicated to submissions review

**Marking and Feedback Schedule:**

|  |  |  |  |
| --- | --- | --- | --- |
| Due Day | Deliverable | Marked by | Submission review |
| Week 3 | Report proposal and specifications | Week 3 | Week 3 |
| Week 6 | Final Report | Week 8 | Week 8 |
| Week 7 | Presentation and Discussion | Week 7 | Week 7 |

Along with the submitted assignment, you need to submit: a fully completed and signed *Coursework submission form* *and* a *Statement of Academic Honesty Form*. You can only submit your own work. Any student suspected of plagiarism will be subject to the procedures set out in the GAR.

**Guidelines:**

* This is a group assignment.

You are required to form a group of three students.

* Submission: The submission of the report + code is only on E-Learning server. A presentation and discussion will be implemented as a 2-hour in-lab session, if the teaching is offline or as a virtual presentation in a synchronous learning environment, of the learning is online.
* Assessment: Assessment will be based on the materials submitted to E-Learning and the presentation in the class or online. You should present the most important findings and results in a 10 min presentation and answer to the questions during the discussion.

The proposal and the report should contain information about the individual contribution of the authors. During the discussion every group member will be assessed also individually.

* Feedback: Feedback will be provided on the E-Learning module site two weeks after the submission.
* Along with the submitted assignment, you need to submit: a fully completed and signed *Coursework submission form* *and* a *Statement of Academic Honesty Form*. You can only submit your own work. Any student suspected of plagiarism will be subject to the procedures set out in the GAR.

The project for CW I consist of the following parts:

* Choice of the NLP problem to be handled
* Project plan
* Literature review (min 7 sources)
* Data Collection
* Data cleaning
* Data pre-processing
* Developing an own trained word embedding, with both methods COWB and skip-gram and compare the results by the weights according to the distance, speed, infrequent words.

It is up to the author to choose an appropriate outline for the report and the used literature. If you use information from the internet, GitHub, GitLab etc., specify, where and when accessed last time. Justify your choices in the project’s initial plan. The project report must be of a size of 1200-1400 words.

- you should present your report in 10 min presentation and answer to the questions

- Templates must be used to present your work. We will use the IEEE templates for this report.

- The report is a project report including a literature review (minimum 5 different sources) about the chosen NLP topic and its specifics.

- The coding should be done as a Colab python notebook using libraries like sci-learn, TensorFlow, Keras or others. Always the last version of the libraries should be used.

The marks for assignment I are distributed as follows:

1. Report **[60 marks]**
2. Presentation **[15 marks]**
3. Discussion **[25 marks]**

**[Assignment Total: 100 marks]**

Marking Scheme:

|  |  |  |  |
| --- | --- | --- | --- |
| Section | Evaluated | Marks | Total Marks |
| (a) | Report   * Content * Data collection (corpus) * Code (cleaning, pre-processing, training CBOW and skip-gram) * Literature | 20  10  20  10 | 60 |
| (b) | Slides  Time  Engagement with the audience | 7  5  3 | 15 |
| (c) | Discussion | 25 | 25 |

|  |  |  |
| --- | --- | --- |
| BUE final logoInformatics and Computer Science | **22CSAI08H**  **Project and Report**  **2022/2023** | |
| Module Title: **Introduction to Natural Language Processing (NLP)** | | |
| Module Leader **Prof. Pester** | | Semester  **Two** |
| Assessment Weight  **35 % of the total course mark** | Due Date  **Week 11** | |

**Instructions to students:**

* This is a group assignment.

You are required to form a group of three students.

* Submission: The submission of the report + code is only on E-Learning server. A presentation and discussion will be implemented as a 2-hour in-lab session, if the teaching is offline or as a virtual presentation in a synchronous learning environment, of the learning is online.
* Assessment: Assessment will be based on the materials submitted to E-Learning and the presentation in the class or online. You should present the most important findings and results in a 10 min presentation and answer to the questions during the discussion.

The proposal and the report should contain information about the individual contribution of the authors. During the discussion every group member will be assessed also individually.

* Feedback: Feedback will be provided on the E-Learning module site two weeks after the submission.
* Along with the submitted assignment, you need to submit: a fully completed and signed *Coursework submission form* *and* a *Statement of Academic Honesty Form*. You can only submit your own work. Any student suspected of plagiarism will be subject to the procedures set out in the GAR.

This assignment consists of the following parts:

* Setup the model for your chosen topic
* Train the model with the chosen word embedding.
* Evaluate the model the word embedding.
* Improve the model.

The Assignment: NLP model

The aim of this assignment to gain insight into the development and evaluation of a NLP model. The model should solve one of the typical tasks of NLP: Autocorrecting, Part of Speech Tagging, Autocompleting, Machine Translating, Text Summarizing etc. The objective is to work on the design and analysis of a NLP model using the developed word embedding from CW I or an existing word embedding and probabilistic, sequential, or other algorithms. The use of attention or transformers is optional, but not mandatory

It is up to the group to choose an appropriate outline for the report. If you use information from the internet, GitHub, GitLab etc., specify, where and when accessed last time. Justify your choices in the project’s initial plan. The report must be of a size of 1200-1400 words.

1. - you should present your report in 10 min presentation and answer to the questions
2. - Templates must be used to present your work. We will use the IEEE templates for this report.
3. - The coding should be as a Colab python notebook using libraries like sci-learn, TensorFlow, Keras, or others. Always the last version of the libraries should be used.

**Marking and Feedback Schedule:**

|  |  |  |  |
| --- | --- | --- | --- |
| Due Day | Deliverable | Marked by | Submission review |
| Week 8 | Project proposal and specifications | Week 8 | Week 8 |
| Week 11 | Final Report | Week 11 | Week 11 |
| Week 12 | Presentation and discussion, presentation submission | Week 12 | Week 12 |

You are required to submit your system as a jpyter Notebook running in Colab.

The requirements are as follows:

* Investigate an NLP problem and develop a model, use your word embedding from assignment 1 or an existing word embedding.
* You should test at least two variations of your model (explore the results of changing hyperparameters like size of vocabulary, learning rate etc.).
* Choose different metrics and evaluate your models with learning curves. Use TensorBoard, if applicable
* In the report justify your choices and give a critical analysis if your findings.
* The presentation (10 min – not more than 10 slides) should include your main findings. It is not repetition of the theory, but a presentation of your own work.
* Report+Code+Slides should be submitted to the E-Learning server. The presentation slides should be submitted after the discussion.

The marks for assignment two are distributed as follows:

1. Model setup **[20 marks]**
2. Implementation **[30 marks]**
3. Model evaluation and improvement **[30 marks]**
4. Presentation and discussion **[20 marks]**

**[Assignment Total: 100 marks]**

Marking Scheme:

|  |  |  |  |
| --- | --- | --- | --- |
| Section | Evaluated | Marks | Total Marks |
| (a) | Choice of Algorithms  Use of libraries  Test and Training model | 6  7  7 | 20 |
| (b) | Code  Programme structure  Commenting | 11  11  8 | 30 |
| (c) | Learning curves  Metrics  TensorBoard | 13  15  2 | 30 |
| (d) | Slides  Time  Discussion | 7  6  7 | 20 |

The Discussion is aimed at ensuring the students understand the code and the code presented is the student’s own work.

**Module Specification – (Programme Specs Ver. 5.1) 2021/22**

|  |  |  |
| --- | --- | --- |
| **Module Code:**  21CSAI05H | **Title:**  Introduction to Natural Language Processing (NLP) | |
| **Level:** 6 | **Modular weight:** 10 | **Faculty/Dept:** ICS |
| **Pre-requisite modules:**  CSAI01I, CSCI05I, SCIB07P, SCIB03C, CSCI02C | | |
| **Reassessment:**  No restriction | | |
| **Module Leader:** | | |
| **Semester taught:**  Two | | |
| **Date of latest revision:** April 2019 | | |

**Aims**

The aim of this module is to introduce the main topics in theoretical linguistics, as well as the practical hands on experience for text processing. The student will learn how to make sense of the vast amounts of texts available online and other sources.

**Intended Learning Outcomes**

***On completion of this module students should be able to:***

**Knowledge and Understanding**

1. Describe and understand the core functions of corpus readers, stemmers, taggers and parsers. [A1, A3, A11**,** A12].
2. Understand different Natural Language Processing techniques [A11**,** A12].

**Intellectual Skills**

1. Design experiments with implemented Natural Language Processing techniques on data sets, and evaluate and reflect on the results [B11].
2. Evaluate the performance of Natural Language Processing systems [B11].
3. Recognize potential real-world applications of Natural Language Processing and evaluate the suitability of Natural Language Processing techniques; given previously unseen task. [B1, B5, B11].

**Practical and Professional Skills**

1. Use a systematic approach to conducting experiments, given a data sets from any domain. [C2, C4, C6, C11, C12]
2. Use appropriate software tools and apply Natural Language Processing (NLP) methods to text data to fulfil a specific objective. [C2, C4, C6, C11, C12].

**General and Transferable skills**

1. Develop interpersonal skills working, communicating and collaborating in a team on assigned projects. [D2, D6, D8].

**Employability**

***This module will provide opportunities for students to:***

1. Understand the importance of being self-motivated in order to progress the area of work. [A.1, A.5]
2. Understand how to build small practical NLP systems for a number of domains. [B.1.1]
3. Demonstrate effective time management to manage time effectively so as to prioritise tasks and to work to deadlines. [C.1.5]

**Indicative Content**

Introduction to NLP grammar, types and tokens, part-of-speech tagging, parsing, lexical semantics, information retrieval and extraction (document matching, named-entity recognition), text classification.

**Methods of Learning, Teaching and Assessment**

Total student effort for the module: 100 hours on average over 1 semester.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Type of session** | **ILOs**  **Assessed** | **Typical Student Effort** | | |
| **Typical number in the semester/s** | **Typical hours per week** | **Total hours** |
| Lecture | 1-6 | 12 | 2 | 24 |
| Tutorial | - | - | - | - |
| Laboratory | 4-8 | 12 | 2 | 24 |
| Private study | 1-8 |  |  | 52 |

**Assessment**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Assessment Type** | **Weight %** | **ILOs**  **Assessed** | **Exam Semester** | **Exam/**  **Written Coursework Length** |
| Two group projects , weight of each is 35% | 70 | 1-8 | 1 | N/A |
| Written class test | 30 | 1,2,5 | 1 | 2 Hours |

**Methods of Feedback**

***In response to assessed work:***

* Developmental feedback generated through teaching activities.
* Feedback will be provided for each assessed component in written form as appropriate.
* Generic exam feedback will be given on the e-learning system.

***Developmental feedback generated through teaching activities:***

* Dialogue between students and staff in workshops and labs

**Indicative Reading List**

* M. Raghavan, and Schutze. Introduction to Information Retrieval. Cambridge University Press, 2008.
* Jurafsky and Martin: Speech and Language Processing (2nd Edition). Pearson, 2009.
* D. Jurafsky & J. H. Martin: Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition, Prentice Hall, (2nd Edition), 2009.
* S. Bird, E. Klein and E. Loper, Natural Language Processing with Python, O'Reilly Media, 2009.