ROBERT GORDON UNIVERSITY ABERDEEN

CM4603 – Coursework 1 (Group)

Academic Year	**2023-24**
Semester	**1**
Module Number	**CM4603**
Module Title	**Language Processing and Information Retrieval**
Assessment Method	**A group submission of code and report followed by a group presentation.**
Deadline (time and date)	**12 midnight on 9th December 2023**
Submission	**Assessment Dropbox in the Module Study Area in CampusMoodle.**
Word Limit	**1500 words**
Use of Generative Artificial Intelligence (AI) text	**IS authorised**
Module Co-ordinator	**Ruvan Weerasinghe**

What knowledge and/or skills will I develop by undertaking the assessment?

Students will be able to explore various natural language pre-processing tasks including tokenization and the get a feel for the different feature extraction methods in preparing textual data for modelling tasks. In doing so, they will be able to grasp the effects of these techniques on text processing tasks such as classification.

On successful completion of the assessment students will be able to achieve the following Learning Outcomes:

1. *Describe and critically review natural language processing techniques.*
2. *Select, analyse and apply NLP algorithms to reason with textual content.*
3. *Pre-process and transform textual content for algorithms to satisfy information retrieval needs using a range of similarity metrics.*

Please also refer to the Module Descriptor, available from the module Moodle study area.

What is expected of me in this assessment?

Task(s) - content

Task 1: Tokenization

You are required to use the WikiText-2 dataset[1] (https://blog.salesforceairesearch.com/the-wikitext-long-term-dependency-language-modeling-dataset/) as a source for obtaining a wide-coverage English vocabulary. Fully describe the dataset in terms of number of articles, the total and unique tokens in the whole dataset and a histogram of article lengths. What are the insights you gained about the data? Apply at least three (03) different tokenization/sub-word segmentation schemes such as white space, regex, stemming, lemmatizing, wordpiece and byte pair encoding with justification. Comment on the effectiveness of the different schemes using any suitable method.

Task 2: Feature Extraction

Describe the Reuters news dataset distributed with this coursework as in task 1, but with additional information about the number of categories and the number of articles in each category. Decide on the practical number of news categories for which text classification can be applied, justifying your reason to select such a number. Clean and pre-process the data as required to prepare the data for feature extraction. Apply the top three (03) tokenization schemes found in Task 1 and create two (02) sparse vector and two (02) dense vector (non-transformer) representations of the dataset to extract features from the text. Describe the resulting shapes of the data matrix in terms of the number of rows and columns, justifying their suitability for downstream classification tasks.

Task 3: Text classification

Use the feature extraction methods applied in Task 2 to evaluate the performance of three (03) significantly different non-deep learning algorithms for predicting categories of news in the dataset, justifying the rationale for selecting the three algorithms. Interpret the results of the performance of each combination above.

Task 4: Using pre-trained vectors

Compare the performance of the best tokenization and vectorization combinations explored in Task 1 and Task 2 with that of a pre-trained contextual embedding for the three (03) algorithms selected in Task 3. Finally compare the performance of these models with a deep learning model using contextual embedding.

[1] *You may use the Wikitext-103 dataset if needed.*

What is expected of me in this assessment?

Task 5: Improving performance

Discuss any considerations given by your group outside the direct specification of this coursework brief and/or how you would improve on the rigour of your exploration of the above data and models.

Task(s) - format

*You are required to **formulate solutions for each of tasks 1 through 4** above, clearly explaining your Python code and specifying the outputs produced by the code for the datasets used **in an iPython Notebook named Solution_Group#.ipynb** based on the template given (the # in the filename should be replaced with your group number number – 1 to 9). For each such part, a descriptive summary with an interpretation should be given for the output obtained after each executable cell. The notebook should be compressed as a .zip file. You also need to submit a PDF version of the notebook using a tool described at https://saturncloud.io/blog/how-to-export-jupyter-notebook-as-pdf. The name of the file should be the same as the notebook, except that the extension will be .pdf.*

*In addition, you should **write a comprehensive report of not more than 1500 words** on how you used generative AI in carrying out your coursework. You should NOT reproduce the detailed responses of the Generative AI, but provide a **detailed description** of the prompts you employed step by step in undertaking the coursework. The **PDF version of this report should be named, Report_Group#.pdf** where the # in the filename should be replaced with your group number number – 1 to 9.*

Your iPython Notebook files (the zipped notebook and the converted pdf) and your study report should be submitted as three separate files to Campus Moodle. Note that the PDF files should NOT be compressed. Submissions which do NOT adhere to these formatting and naming conventions would incur penalties.

Participation in the final physical presentation is mandatory for all group members. The presentation should include a demonstration of the relevant iPython notebook and the reflections in the report. Viva questions will be directed at specific group members.

How will I be graded?	
A number of subgrades will be provided for each criterion on the feedback grid which is specific to the assessment. The overall grade for the assessment will be calculated using the algorithm below.	
A	At least 50% of the subgrades to be at Grade A, at least 75% of the subgrades to be at Grade B or better, and normally 100% of the subgrades to be at Grade C or better.
B	At least 50% of the subgrades to be at Grade B or better, at least 75% of the subgrades to be at Grade C or better, and normally 100% of the subgrades to be at Grade D or better.
C	At least 50% of the subgrades to be at Grade C or better, and at least 75% of the subgrades to be at Grade D or better.
D	At least 50% of the subgrades to be at Grade D or better, and at least 75% of the subgrades to be at Grade E or better.
E	At least 50% of the subgrades to be at Grade E or better.
F	Failing to achieve at least 50% of the subgrades to be at Grade E or better.
NS	Non-submission.
NB: Non-participation in the presentation will result in the next lower grade being awarded to the group member concerned.	

ROBERT GORDON UNIVERSITY ABERDEEN

Feedback grid

GRADE	A	B	C	D	E	F
DEFINITION / CRITERIA (WEIGHTING)	**EXCELLENT**	**COMMENDABLE/VERY GOOD**	**GOOD**	**SATISFACTORY**	**BORDERLINE FAIL**	**UNSATISFACTORY**
Task 1 **(2 subgrades)**	The group describes the dataset used comprehensively and explores 5 or more ways of tokenizing text, commenting on their merits using at least one metric.	The group describes the dataset used very well and explores at least 5 ways of tokenizing text, commenting on their merits using a justified metric.	The group describes the dataset used and explores multiple ways of tokenizing text, commenting on their merits using a metric.	The group describes the dataset used and explores some ways of tokenizing text, and use a metric to compare them.	The group does not describe the dataset used adequately and explores some ways of tokenizing text.	The group does not describe the dataset used adequately and uses some standard ways of tokenizing text.
Task 2 **(1 subgrade)**	The group uses multiple sensible sparse and dense vector representations to extract features of the dataset, specifies the exact resulting matrices and justifies them for downstream tasks.	The group uses two sensible sparse and two sensible dense vector representations to extract features of the dataset, specifies the shape of the resulting matrices and justifies them for downstream tasks.	The group uses two sparse and two dense vector representations to extract features of the dataset, specifies the shape of the resulting matrices and mentions their suitability for downstream tasks.	The group uses two sparse and two dense vector representations to extract features of the dataset, specifies the shape of the resulting matrices and interprets the numbers.	The group uses sparse and dense vector representations to extract features of the dataset, specifies the shape of the resulting matrices.	The group uses sparse and dense vector representations to extract features of the dataset, specifies the shape of the resulting matrices.
Task 3 **(1 subgrades)**	The group carefully reason out the choice of the 3 algorithms to use for modelling and interprets the results obtained by each in a comprehensive way.	The group justifies the choice of the 3 algorithms to use for modelling and interprets the results obtained by each in a detailed way.	The group makes a sensible choice of 3 algorithms to use for modelling and interprets the results obtained by each.	The group demonstrates the use of 3 significantly different algorithms to use for modelling and interprets the results obtained by each.	The group demonstrates the use of 3 different algorithms to use for modelling and comments on the results obtained by each.	The group demonstrates the use of 3 algorithms to use for modelling but fails to make any meaningfu comments on the results obtained by each.

GRADE	A	B	C	D	E	F
DEFINITION / CRITERIA (WEIGHTING)	**EXCELLENT**	**COMMENDABLE/VERY GOOD**	**GOOD**	**SATISFACTORY**	**BORDERLINE FAIL**	**UNSATISFACTORY**
Task 4 **(2 subgrades)**	The group successfully implements multiple pre-trained word and character embedding schemes for feature extraction and compares and comments on the performance of the algorithms with own tokenization and feature extraction schemes.	The group successfully implements multiple pre-trained word embedding schemes for feature extraction and compares and comments on the performance of the algorithms with own tokenization and feature extraction schemes.	The group successfully implements a pre-trained word embedding scheme for feature extraction and compares and comments on the performance of the algorithms with own feature extraction schemes.	The group successfully implements a pre-trained word embedding scheme for feature extraction and compares the performance of the algorithms with own feature extraction schemes.	The group implements a pre-trained word embedding scheme for feature extraction and reports the performance of the algorithms with own feature extraction scheme.	The group fails to successfully implement a pre-trained word embedding scheme for feature extraction and reports the performance of the algorithms with own feature extraction scheme.
Task 5 **(1 subgrade)**	A comprehensive reflection on using generative AI in executing the coursework for tokenization, pre-processing, modelling and diagnostics with several considerations unspecified in the coursework brief.	A detailed reflection on using generative AI in dealing with tokenization, class imbalance, overfitting and other important data pre-processing, modelling and diagnostics with some considerations unspecified in the coursework brief.	A detailed reflection on the use of generative AI for executing the tokenization, data pre-processing, modelling and diagnostics steps in the coursework.	The report comments on various tokenization, data pre-processing and model diagnostics steps of the coursework.	The report comments only on some selective pre-processing steps, modelling choices and diagnostics.	The report fails to comment on the choices made with respect to pre-processing, modelling or diagnostics.
Presentation **(1 subgrades)**	A high level of group cohesion is evident in the presentation and the answers to questions to the point. Excellent class engagement with all weekly activities submitted.	A good flow is evident in the presentation and the answers to questions to the point. High level of class engagement with all weekly activities submitted.	A good flow is evident in the presentation and all questions are answered by some member(s) of the group. Good class engagement with most weekly activities submitted.	The presentation is understandable and most questions are answered by some member(s) of the group. Satisfactory class engagement with most weekly activities submitted.	The presentation is not fully clear and only some questions are answered adequately by the group. Some class engagement with some weekly activities carried out.	The presentation is confusing and most questions are not answered adequately by the group. Poor class engagement with several weekly activities not carried out.

Coursework received late will be regarded as a non-submission (NS) and one of your assessment opportunities will be lost.

What else is important to my assessment?

What is the Assessment Word Limit Statement?

It is important that you adhere to the Word Limit specified above. The Assessment Word Limit Statement can be found in Appendix 2 of the RGU Assessment Policy. It provides detail on the purpose, setting and implementation of wordage limits; lists what is included and excluded from the word count; and the penalty for exceeding the word count.

What's included in the word count?

The table below lists the constituent parts which are included and excluded from the word limit of a Coursework; more detail can be found in the full Assessment Word Limit Statement. Images will not be allowed as a mechanism to circumvent the word count.

Excluded	Included
Cover or Title Page	Main Text e.g. Introduction, Literature Review, Methodology, Results, Discussion, Analysis, Conclusions, and Recommendations
Executive Summary (Reports) or Abstract	Headings and subheadings
Contents Page	In-text citations
List of Abbreviations and/or List of Acronyms	Footnotes (relating to in-text footnote numbers)
List of Tables and/or List of Figures	Quotes and quotations written within "…"
Tables – mainly numeric content	Tables – mainly text content
Figures	
Reference List and/or Bibliography	
Appendices	
Glossary	

What are the penalties?

The grade for the submission will be reduced to the next lowest grade if:

- The word count of submitted work is above the specified word limit by more than 10%.
- The submission contains an excessive use of text within Tables or Footnotes.

What else is important to my assessment?

What is plagiarism?

Plagiarism is "the practice of presenting the thoughts, writings or other output of another or others as original, without acknowledgement of their source(s) at the point of their use in the student's work. All materials including text, data, diagrams or other illustrations used to support a piece of work, whether from a printed publication or from electronic media, should be appropriately identified and referenced and should not normally be copied directly unless as an acknowledged quotation. Text, opinions or ideas translated into the words of the individual student should in all cases acknowledge the original source" (RGU 2022).

What is collusion?

"Collusion is defined as two or more people working together with the intention of deceiving another. Within the academic environment this can occur when students work with others on an assignment, or part of an assignment, that is intended to be completed separately" (RGU 2022).

For further information please see Academic Integrity.

What if I'm unable to submit?

- The University operates a Fit to Sit Policy which means that if you undertake an assessment then you are declaring yourself well enough to do so.
- If you require an extension, you should complete and submit a Coursework Extension Form. This form is available on the RGU Student and Applicant Forms page.
- Further support is available from your Course Leader.

What additional support is available?

- RGU Study Skills provide advice and guidance on academic writing, study skills, maths and statistics and basic IT.
- RGU Library guidance on referencing and citing.
- The Inclusion Centre: Disability & Dyslexia.
- Your Module Coordinator, Course Leader and designated Personal Tutor can also provide support.

What are the University rules on assessment?

The University Regulation 'A4: Assessment and Recommendations of Assessment Boards' sets out important information about assessment and how it is conducted across the University.