ROBERT GORDON UNIVERSITY ABERDEEN

CM4603 – Coursework 1 (Group)

Academic Year	**2024-25**
Semester	**1**
Module Number	**CM4603**
Module Title	**Language Processing and Information Retrieval**
Assessment Method	**A group submission of code and report followed by individual viva**
Deadline (time and date)	**12 midnight on 29th November 2024**
Submission	**Assessment Dropbox in the Module Study Area in CampusMoodle**
Word Limit	**1500 words**
Use of Generative Artificial Intelligence (AI)	**Is authorised**
Module Co-ordinator	**Ruvan Weerasinghe**

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

Students will be able to collect and collate text data from online sources, explore various natural language pre-processing techniques on the data and perform diverse feature extraction methods in preparing this data for modelling tasks. They will then be able to grasp the effects of these techniques on text processing tasks such as classification and evaluate the resulting models. They will also be able to interpret the results of such modelling including explaining them.

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.*

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

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

What is expected of me in this assessment?

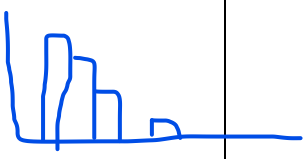
Task(s) - content

Sri Lanka is set on putting behind its past and regaining its position as one of the popular tourist destinations in Asia. You are tasked with achieving this by showcasing some of the best destinations it has to offer. To do this, you are required to access at least 5000 user reviews from at least 100 Sri Lankan hotels on tripadvisor.com and process this information using the following steps.

Task 1: Describe

After performing any cleaning steps on the raw data, fully describe the raw dataset and the cleaned dataset collected from TripAdvisor in terms of the number of hotels, the number of reviews for each hotel, the total number and the number of unique words in the whole dataset. You are also required to plot a histogram of article lengths (in words) of the full dataset. Perform any other exploratory data analysis (EDA) to describe the dataset for bonus points.

Skewed to the right the graph



Task 2: Establishing Ground Truth

The number of reviews is too large to be able to manually establish a 'gold standard' for the sentiment expressed in each review of this dataset. As an alternative, it is suggested that you use 3 different existing sentiment classifiers with at least one being a sentiment lexicon[1]. Use a majority vote scheme to label the reviews of the entire dataset as the way to establish the 'ground truth'.

can train a model

can use existing classifiers

Task 3: Feature Extraction

Create two (02) sparse vector and two (02) dense vector (non-transformer) representations of the dataset to extract features from the text dataset. Describe the resulting shapes of the data matrices in terms of the number of rows and columns, justifying your choice of vectorization methods used.

min df - used to remove typos
max df - used to remove higher occurence words like 'the' 'a'

Task 4: Text classification

Use each of the feature extraction methods applied in Task 3 to evaluate the performance of three (03) significantly different non-deep learning algorithms for predicting the sentiment of user reviews, justifying

[1] SentiWordNet, Bing Liu Lexicon and VADER are some examples.

What is expected of me in this assessment?

the rationale for selecting the three algorithms. Interpret the results of the performance of each combination above.

Task 5: Using pre-trained vectors

Compare the performance of the best combinations explored in Task 4 with that of a pre-trained ~~contextual~~ embedding used with the same three (03) algorithms selected in Task 4. Also compare the performance of the resulting models with a deep learning model using this same contextual embedding.

Task 6: Text Clustering

It is desirable to dig deeper to try to understand what the different aspects of the hotel that these sentiments are directed at. For this purpose, you are required to cluster the documents using topic modelling to arrive at a distinct set of aspects from the dataset. Justify your model by randomly taking 50 reviews from the test set and manually labelling the aspects they represent and then checking[2] how they would be clustered in your topic model.

grouping

Task(s) - format

*You are required to **formulate solutions for each of tasks 1 through 6** above, clearly explaining your Python code and specifying the outputs produced by the code for the dataset 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 10). 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 with the same filename as the notebook, except that the extension should be .pdf.*

*In addition, you should **write a comprehensive report of not more than 1500 words** in two parts**. Part A** should state any considerations given by your group outside the direct specification of this coursework brief. In addition, a description of what you would do differently to improve the rigour of your exploration if you had to do it again. **Part B** should contain a trace your learning journey as a group through the sources used and the kind of prompts you gave generative AI. Note that you **should NOT** reproduce the steps you carried out step by step in the coursework. The **PDF version of this report should be named,***

[2] A common way to compare clustering with ground truth include Rand Index and the Jaccard Coefficient.

What is expected of me in this assessment?

Report_Group#.pdf *where the # in the filename should be replaced with your group number number – 1 to 10.*

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 delays in grading and possibly result in penalties.

Participation in the individual physical viva is mandatory for all group members. Non-participation will result in a 'No Show' result which will be interpreted as a Non-Submission for the group member concerned.

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 **(1 subgrade)**	At least 5000 reviews from 100 different hotels are collected, cleaned appropriately, described using the specification. In addition, a meaningful EDA has been included.	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)**	How the ground truth for the reviews was arrived at is clearly documented. It is also justified and validated.	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)**	Two sparse vector and two dense vector feature extraction methods are clearly demonstrated. The choice of vectorization is well justified.	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 **(1 subgrade)**	Each of the vectorization techniques is used for training 3 significantly different non-DL algorithms, checking for and mitigating overfitting. The resulting predictive models are evaluated rigorously and interpreted accurately.	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 clear grasp of the use of contextual embedding and how it can be used with the algorithms used is demonstrated. A clear and valid comparison of these 3 classifiers with a deep learning classifier is presented, with adequate checks for validity.	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.
Task 6 **(1 subgrade)**	A clear comprehension of the value and purpose of topic modelling and its potential use in aspect-based sentiment analysis is evidenced. The validation of the clustering achieved with the ground truth has been properly carried out.	A clear application of topic modelling and its potential use in aspect-based sentiment analysis is evidenced. The validation of the clustering achieved with the ground truth has been properly carried out.	Topic modelling is applied to discover clusters in the reviews. The resulting clusters are validated.	The reviews are clustered to discover categories within them.	Some effort has been made to cluster the data using topic modelling which wasn't valid.	Little effort is made to cluster the data in order to understand the categories of the reviews.

GRADE	A	B	C	D	E	F
DEFINITION / CRITERIA (WEIGHTING)	**EXCELLENT**	**COMMENDABLE/VERY GOOD**	**GOOD**	**SATISFACTORY**	**BORDERLINE FAIL**	**UNSATISFACTORY**
Report & Viva **(2 subgrades)**	The group demonstrates understanding beyond the exact specification of the coursework in the way that other considerations have been addressed. In addition, the learning process is well tracked via the generative AI prompts that have been employed. This is evidenced in each member of the team.	The group appears to have worked well in addressing the coursework. Individual learning process is tracked via the generative AI prompts that have been employed. Most members of the team demonstrate this understanding.	Some group work is evidenced in addressing the coursework. The learning process is documented by each member. Members of the team demonstrate this understanding at different levels.	Individuals have contributed to the execution of the coursework. A very similar learning process is documented by all members. Members of the team provide some evidence of this learning.	Only some parts of the coursework have been executed and commented on. The learning process is not well captured. Members of the team don't have a clear idea of what was learnt.	Poor execution of the coursework and weak reflective report on the learning process by members. Poor understanding of the main learning by group members.

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.