SINGAPORE UNIVERSITY OF SOCIAL SCIENCES

ANL312

End-of-Course Assessment - July Semester 2024

Text Mining and Applied Project Formulation

INSTRUCTIONS TO STUDENTS:

1. This End-of-Course Assessment paper comprises **9** pages (including the cover page).

2. You are to include the following particulars in your submission: Course Code, Title of the ECA, SUSS PI No., Your Name, and Submission Date.

3. Late submission will be subjected to the marks deduction scheme. Please refer to the Student Handbook for details.

IMPORTANT NOTE

ECA Submission Deadline: Friday, 01 November 2024 12:00 pm

ECA Submission Guidelines

Please follow the submission instructions stated below:

A - What Must Be Submitted

You are required to submit the following ONE (1) item for marking and grading:

- *A Report*

Please verify your submissions after you have submitted the above ONE (1) item.

B - Submission Deadline

- *The ONE (1) item of Report is to be submitted* ***by 12 noon*** *on the submission deadline.*
- *You are allowed multiple submissions till the cut-off date for each of the ONE (1) item.*
- *Late submission of any of the ONE (1) item* ***will be subjected to mark-deduction scheme*** *by the University. Please refer to Section 5.2 Para 2.4 of the Student Handbook.*

C - How the Item Should Be Submitted

- *The Report: submit online to Canvas via TurnItIn (for plagiarism detection)*
- *Avoid using a public WiFi connection for submitting large video files. If you are using public wireless (WiFi) connection (e.g. SG Wireless at public areas), you might encounter a break in the connection when sending large files.*

D - Additional guidelines on file formatting are given as follows:

1. Report	• *Please ensure that your Microsoft Word document is generated by Microsoft Word 2016 or higher.* • *The report must be saved in* ***.docx*** *format.*

E - Please be Aware of the Following:

Submission in hardcopy or any other means not given in the above guidelines will not

be accepted. You do not need to submit any other forms or cover sheets (e.g. form ET3) with your ECA.

You are reminded that electronic transmission is not immediate. The network traffic may be particularly heavy on the date of submission deadline and connections to the system cannot be guaranteed. Hence, you are advised to submit your work early. ***Canvas will allow you to submit your work late but your work will be subjected to the mark-deduction scheme****. You should therefore not jeopardise your course result by submitting your ECA at the last minute.*

It is your responsibility to check and ensure that your files are successfully submitted to Canvas.

F - Plagiarism and Collusion

Plagiarism and collusion are forms of cheating and are not acceptable in any form in a student's work, including this ECA. Plagiarism and collusion are taking work done by others or work done together with others respectively and passing it off as your own. You can avoid plagiarism by giving appropriate references when you use other people's ideas, words or pictures (including diagrams). Refer to the APA Manual if you need reminding about quoting and referencing. You can avoid collusion by ensuring that your submission is based on your own individual effort.

The electronic submission of your ECA will be screened by plagiarism detection software. For more information about plagiarism and collusion, you should refer to the Student Handbook (Section 5.2.1.3). You are reminded that SUSS takes a tough stance against plagiarism or collusion. Serious cases will normally result in the student being referred to SUSS's Student Disciplinary Group. For other cases, significant mark penalties or expulsion from the course will be imposed.

G - Use of Generative AI Tools (Allowed)

The use of generative AI tools is allowed for this assignment.

- *You are expected to provide proper attribution if you use generative AI tools while completing the assignment, including appropriate and discipline-specific citation, a table detailing the name of the AI tool used, the approach to using the tool (e.g. what prompts were used), the full output provided by the tool, and which part of the output was adapted for the assignment;*

- *To take note of section 3, paragraph 3.2 and section 5.2, paragraph 2A.1 (Viva Voce) of the Student Handbook;*

- *The University has the right to exercise the viva voce option to determine the authorship of a student's submission should there be reasonable grounds to suspect that the submission may not be fully the student's own work.*

- *For more details on academic integrity and guidance on responsible use of generative AI tools in assignments, please refer to the TLC website for more details;*

- *The University will continue to review the use of generative AI tools based on feedback and in light of developments in AI and related technologies.*

(Full marks: 100)

Section A (100 marks)
Answer all questions in this section.

Question 1

Guidelines for this End-of-Course Assessment report are as follows:

1. Draft a report on your proposed topic focusing on the practical application of text mining in a specific field or industry. The materials and sources used should substantially surpass the content covered in this course or any other ANL courses.
2. Construct a text mining project based on your selected topic.

3. You can choose either one of the following two options:

   - **Option 1:** Apply text categorisation using the IBM SPSS Modeler. You will need to find a dataset with minimum **100 rows** of text records for this option. Your response should include documentations of the effort put into improving the resource template and creating the categories from scratch (refer to the GBA steps). Provide screenshots of the text (5-8 samples) for each category, showing your effort to correctly categorise the text. **Do not use** the 'Build Categories' feature that automatically builds the categories as no credit will be given if this is used. Similarly, **do not use** the "Text Analysis Package".

   - **Option 2:** Apply topic modelling using R programming. You have the option to include sentiment analysis, but this is not mandatory. You will need to find a dataset with minimum **500 rows** of text records for this option. Using other software, e.g., IBM SPSS Modeler, for necessary data preparation is allowed for this option.

   For **topic modelling**, please ensure the following:

   - Customize the stop words.
   - Try a few different values of *k* and document the process in your answer.
   - Evaluate the appropriateness of your final topic model. Include screenshots of the top 5 texts for each topic discovered; present them in your report, not in the appendix. To do this, sort the gamma values of each topic in the gammaDF table, and refer to the text in your original dataset using the doc id.
   - Name each topic appropriately.
   - More credits will be given if you improve the model by incorporating additional steps beyond the example given in the Study Guide.

   Regarding the **optional sentiment analysis** section, if you choose to include it, please follow these guidelines:

- Describe the R package you selected for this project, especially if it differs from the one used in the Study Guide.
- Evaluate the appropriateness of the sentiment scores generated by the selected R package based on the text you collected.
- Include appropriate screenshots of the texts after sentiment analysis along with their corresponding sentiment scores/polarity. Present the screenshots with the first 30 rows in your report, not in the appendix.

- Please note that sentiment analysis is optional. Including it may or may not earn additional credits, depending on the quality of the analysis.

4. Possible sources of references for your report:

    - Internet websites

    e.g.,
    http://videolectures.net/,http://www.kdnuggets.com/index.html,https://www.kaggle.com,http://www.dextra.sg,
    https://github.com/stepthom/text_mining_resources

    - Journal articles (Use SUSS library (https://library.suss.edu.sg/) or Google Scholar).
    - Conference papers especially those from the SAS Global Forum where they feature text mining applications (https://www.sas.com/en_us/events/sas-global-forum/program/proceedings.html)

Note: Wikipedia can be used as an initial source of information but should not be cited as a reference.

- Please write between 3,000 to 5,000 words (excluding cover page, table of content, reference and appendices). Marks will be deducted for those that are below 3000 words. For those reports that exceed 5000 words, only the part that is below 5000 words will be graded and the rest will be ignored. No word limit for individual section.

- Font size 12, Times New Roman, 1.5 lines spacing.

- Reference citation and reference list: Use American Psychological Association (APA) referencing style. Please refer to ANL312 Study Unit 1 for details.

- You **must** acknowledge and reference all sources used.

Up to 25 marks of penalties will be imposed for inappropriate or poor paraphrasing. For serious cases, they will be investigated by the examination department. More information on effective paraphrasing strategies can be found on
https://academicguides.waldenu.edu/writingcenter/evidence/paraphrase/effective

.

Topic Selection

1. You are required to select a topic of your choice. A list of topics is provided in Appendix A. You may propose a topic that is not on the list. However, your topic must be related to things you have learnt in this course.
2. Once you have selected the topic, conduct research to look for a dataset for your project and references for your Literature Review section.
3. Please note that any "hotel review" data is **not allowed** for this year's ECA.

The report must include the following sections, each carrying a specific weightage stated in the relevant sub-question.

Question 1a

Introduction: Based on your chosen topic, discuss the project background, the business analytics concepts/issue(s) involved, the project objective(s) and the role of text mining in achieving the objective(s). Define all relevant terms used (e.g., text mining, N-grams).

(12 marks)

Question 1b

Literature Review: Describe ***two (2)*** references (must be research articles from journal/conference/academic report/thesis) that apply text mining in a way relevant to your selected topic. Include the general background of the study, the dataset used, the details of the text mining process applied, as well as relevant findings and conclusions. Discuss the implications of the references to the current project.

(20 marks)

Question 1c

Body: Use the CRISP-DM framework to organize your report. You are required to find a small dataset and construct a text mining model to achieve your project objective(s).

(38 marks)

Question 1d

The last two sections of the report include:
Summary: Summarise the key findings, insights, and conclusions obtained from your text mining analysis.
References: List all the sources you cited in your report and follow the APA referencing style.

Additionally, the entire report will be evaluated based on the coherence and balance maintained across all sections. The Introduction should provide a clear motivation for the project. The Literature Review section should thoroughly review materials closely related to the project. In the CRISP-DM section, the steps should be logically presented, demonstrating a sound approach to text mining. The Summary section should effectively wrap up the project, highlighting key findings and insights. The references should be relevant and support the key ideas presented. The writing should be professional, with good and plain English, and adhere to all the instructions given.

(30 marks)

Appendix:

Appendix A: List of topics for selection.

No.	Topic
1	Application of text mining in education
2	Application of text mining in preventive maintenance
3	Application of text mining in healthcare/medical area
4	Application of text mining in human resource operations/workplace
5	Application of text mining in manufacturing
6	Application of text mining on market intelligence
7	Application of text mining on security/criminal detection
8	Application of text mining on product satisfaction and feedback (specify the product)
9	Application of text mining on social media (specify the specific social media and area eg. Twitter or OCBC Bank)
10	Application of text mining on customer satisfaction (specify the company or product)
11	Application of text mining on employee satisfaction

----- END OF ECA PAPER -----