SINGAPORE UNIVERSITY
OF SOCIAL SCIENCES

ANL312
Text Mining and Applied Project Formulation

Group-based Assignment

July 2024 Semester

GROUP-BASED ASSIGNMENT

This assignment is worth 20% of the final mark for ANL312 Text Mining and Applied Project Formulation.

The cut-off date for this assignment is **18 October 2024, 2355hrs.**

This is a group-based assignment. You should form a group of **4 members** from your seminar group. Each group is required to upload a single report via your respective seminar group site in Canvas. Please elect a group leader. The responsibility of the group leader is to upload the report on behalf of the group. Those submitting individually will be given a 10 marks deduction.

It is important for each group member to contribute substantially to the final submitted work. All group members are equally responsible for the entire submitted assignment. If you feel that the work distribution is inequitable to either yourself or your group mates, please highlight this to your instructor as soon as possible. Your instructor will then investigate and decide on any action that needs to be taken. It is not necessary for all group members to be awarded the same mark.

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.

Note to Students:

Compose your report using Microsoft Office Word, and save as **.docx**.

Submit your SPSS Modeler stream in **.str** format.

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

For this GBA, it is mandatory that questions are not divided among group members. Each member must independently address and work on a question before engaging in group discussions of the question. In the event of a peer evaluation, each member is expected to submit their own original answers along with justifications for their individual contributions.

All peer evaluation requests must be submitted to the school at least three working days before the GBA due date. Late requests will not be considered.

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.

Question 1

Your GBA team has been tasked by the Management of Crowne Plaza Changi Airport to analyse the free-format hotel reviews received from their guests. Prior to engaging your GBA team, the Operations Manager at Crowne Plaza Changi Airport has been eye-balling the hotel reviews and the Management at Crowne Plaza Changi Airport wants to have a systematic way to analyse the data. Your GBA team is required to use a sample hotel review data in 2016 to build a text mining model.

There are 142 entries contained in **CrownePlazaReview_Sample.xlsx**. Table 1 gives a description of the variables contained in the file.

Table 1: Summary of variables in CrownePlazaReview_Sample.xlsx

S/N	Variable	Definition	Data type
1	ID	Entry ID	Continuous from 1 to 142
2	Room_type	Room type of the stay	Categorical
3	Traveller_type	Traveller type of guests	Categorical
4	Month	Month of the review	Categorical
5	Day	Day of the review	Continuous from 1 to 31
6	Rating	Rating of the review	Continuous from 1 to 10
3	Comment	Comment of the review	Free format text

(a) Discuss the business problem(s) and the business analytics goal(s) that your *team* can address and achieve using the IBM SPSS Modeler. Maximum 100 words.

(6 marks)

(b) Use a *Table* node and visually examine the content of **CrownePlazaReview_Sample.xlsx** for all 142 entries especially the comment column as shown in Figure 1.

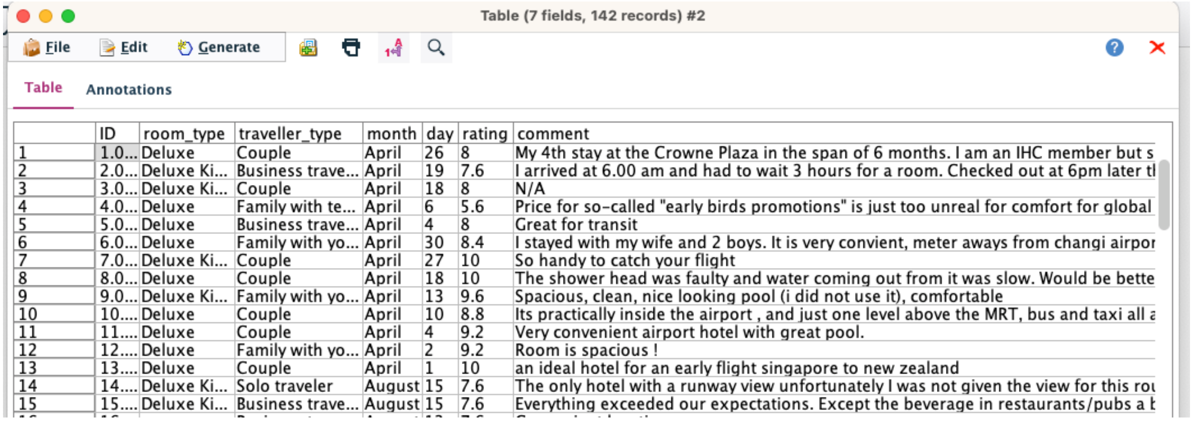


Table (7 fields, 142 records) #2

	ID	room_type	traveller_type	month	day	rating	comment
1	1.0...	Deluxe	Couple	April	26	8	My 4th stay at the Crowne Plaza in the span of 6 months. I am an IHC member but s
2	2.0...	Deluxe Ki...	Business trave...	April	19	7.6	I arrived at 6.00 am and had to wait 3 hours for a room. Checked out at 6pm later tl
3	3.0...	Deluxe Ki...	Couple	April	18	8	N/A
4	4.0...	Deluxe	Family with te...	April	6	5.6	Price for so-called "early birds promotions" is just too unreal for comfort for global
5	5.0...	Deluxe	Business trave...	April	4	8	Great for transit
6	6.0...	Deluxe	Family with yo...	April	30	8.4	I stayed with my wife and 2 boys. It is very convient, meter aways from changi airpor
7	7.0...	Deluxe Ki...	Couple	April	27	10	So handy to catch your flight
8	8.0...	Deluxe	Couple	April	18	10	The shower head was faulty and water coming out from it was slow. Would be bette
9	9.0...	Deluxe Ki...	Family with yo...	April	13	9.6	Spacious, clean, nice looking pool (i did not use it), comfortable
10	10....	Deluxe	Couple	April	10	8.8	Its practically inside the airport , and just one level above the MRT, bus and taxi all a
11	11....	Deluxe	Couple	April	4	9.2	Very convenient airport hotel with great pool.
12	12....	Deluxe	Family with yo...	April	2	9.2	Room is spacious !
13	13....	Deluxe	Couple	April	1	10	an ideal hotel for an early flight singapore to new zealand
14	14....	Deluxe Ki...	Solo traveler	August	15	7.6	The only hotel with a runway view unfortunately I was not given the view for this rou
15	15....	Deluxe Ki...	Business trave...	August	15	7.6	Everything exceeded our expectations. Except the beverage in restaurants/pubs a b

Figure 1: Screenshot of the Table Node

i) There are 3 types of comments in the raw dataset need to be addressed. Explain why the comments need to be addressed using the table below.

Table 2: Type of reviews to be addressed

S/N	**Type of comments to be addressed**	**Why the comments need to be addressed?**
1	Duplicate comments	
2	Useless comments like "nil", "N/A" and "."	
3	Special characters due to encoding issue like "&amp"	

ii) Implement the tasks needed in the IBM SPSS Modeler to address each type of comments.

In your answer:
1. Give the screenshots of your **stream** and key **settings** used in each node.
2. State the number of rows of text that remains after the data preparation tasks have been executed.

More credit will be given if the solution is applicable and efficient (in terms of number of nodes used and/or less conditions specified in a node) even if there is a million rows of text.

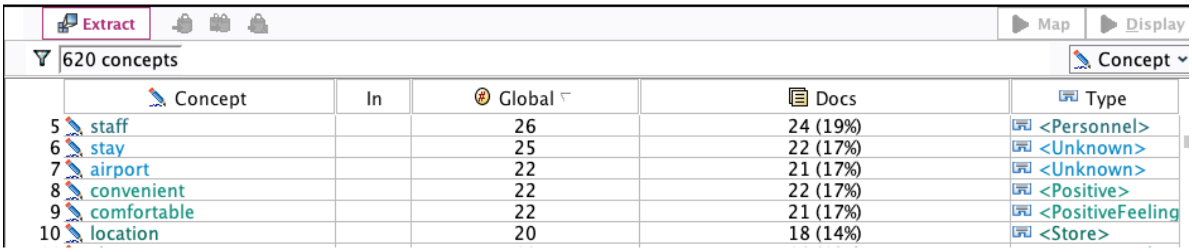
(18 marks)

(c) From the list of resource templates available, suggest three resource templates that would be most suitable for our current text collection. Run three separate *Text Mining* nodes with each using different resource templates based on your solution from part (b). In your answer, compare the libraries available and several statistics that are generated in the interactive workbench (e.g., number of concepts extracted) and justify which resource template you think is most suitable for the task at hand.

Note: No need to retype the concepts in this step. No need to create any category in this step. The resource template required in part (d) is for consistency, may not reflect the answer to part (c).

(14 marks)

(d) Apply the Hotel Satisfaction (English) resource template on your cleaned text collection based on your solution from Part (b). Check the option for Accommodate Spelling for a minimum root character limit of 5 under the Expert tab. Fine-tune the resource template by examining the concepts to see if there needs to be a change in the type (e.g. location is typed as <Store> should be retyped as <HotelLocation>) and retyping the Unknown type (e.g. service is typed as <Unknown> should be retyped as <Service>) to some more relevant type as shown in Figure 3 below:



620 concepts

	Concept	In	Global	Docs	Type
5	staff		26	24 (19%)	<Personnel>
6	stay		25	22 (17%)	<Unknown>
7	airport		22	21 (17%)	<Unknown>
8	convenient		22	22 (17%)	<Positive>
9	comfortable		22	21 (17%)	<PositiveFeeling
10	location		20	18 (14%)	<Store>

Figure 3: Screenshot of the Extract concepts pane

(i) Retype at least 30 existing concepts using the format given in Table 3 below (excluding the examples given in the question above). Use an editable table in your answer, **do not** use an image.

Table 3: List of existing concepts that are to be retyped

S/N	Concept to be retyped	Default Type	More Appropriate Type	Library
e.g.	location	Store	HotelLocation	Local Library
1				
2				

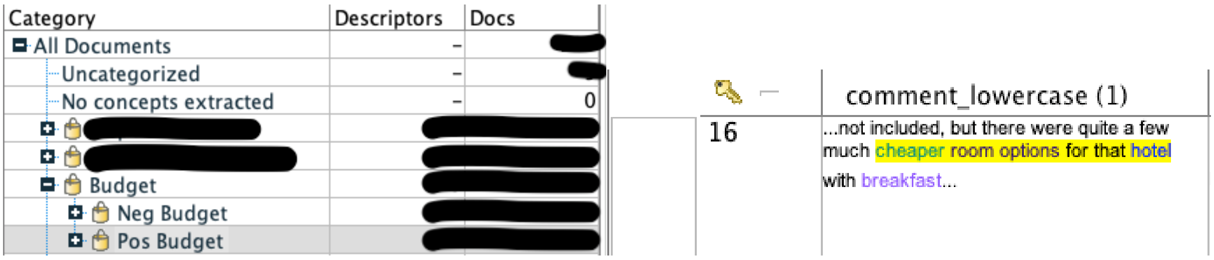
(ii) Define at least 10 synonyms (10 different targets) for spelling errors, short forms and other unstandardised forms of a concept using the format in Table 4 below. The synonyms should come from the dataset. Use an editable table in your answer, **do not** use an image.

Table 4: List of synonyms to be added

S/N	Targets	Synonyms	Library
1			
2			

(18 marks)

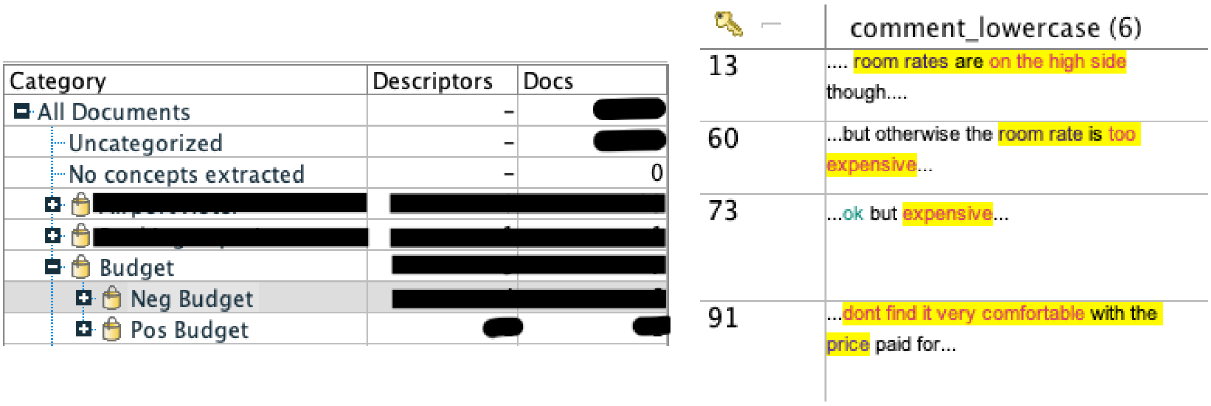
(e) Create a category model with at least 5 main categories. Develop subcategories and category rule(s) to ensure that each category is actionable, meaningful, and distinct. Samples of what your categories should look like are found in Figures 4 and 5 below for the "Pos Budget" category and "Neg Budget" category.



Category	Descriptors	Docs
All Documents	-	[redacted]
Uncategorized	-	[redacted]
No concepts extracted	-	0
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
Budget	[redacted]	[redacted]
Neg Budget	[redacted]	[redacted]
Pos Budget	[redacted]	[redacted]

	comment_lowercase (1)
16	...not included, but there were quite a few much cheaper room options for that hotel with breakfast...

Figure 4: Pos Budget category – Positive comments related to price



Category	Descriptors	Docs
All Documents	-	[redacted]
Uncategorized	-	[redacted]
No concepts extracted	-	0
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
Budget	[redacted]	[redacted]
Neg Budget	[redacted]	[redacted]
Pos Budget	[redacted]	[redacted]

	comment_lowercase (6)
13	.... room rates are on the high side though....
60	...but otherwise the room rate is too expensive...
73	...ok but expensive...
91	...dont find it very comfortable with the price paid for...

Figure 5: Neg Budget category – Negative comments related to price

Each subcategory should have at least 1 category rule. Your final text mining model should have at most **10** documents uncategorised and **0** document with no concept extracted as shown in Figure 6.

In your answer:

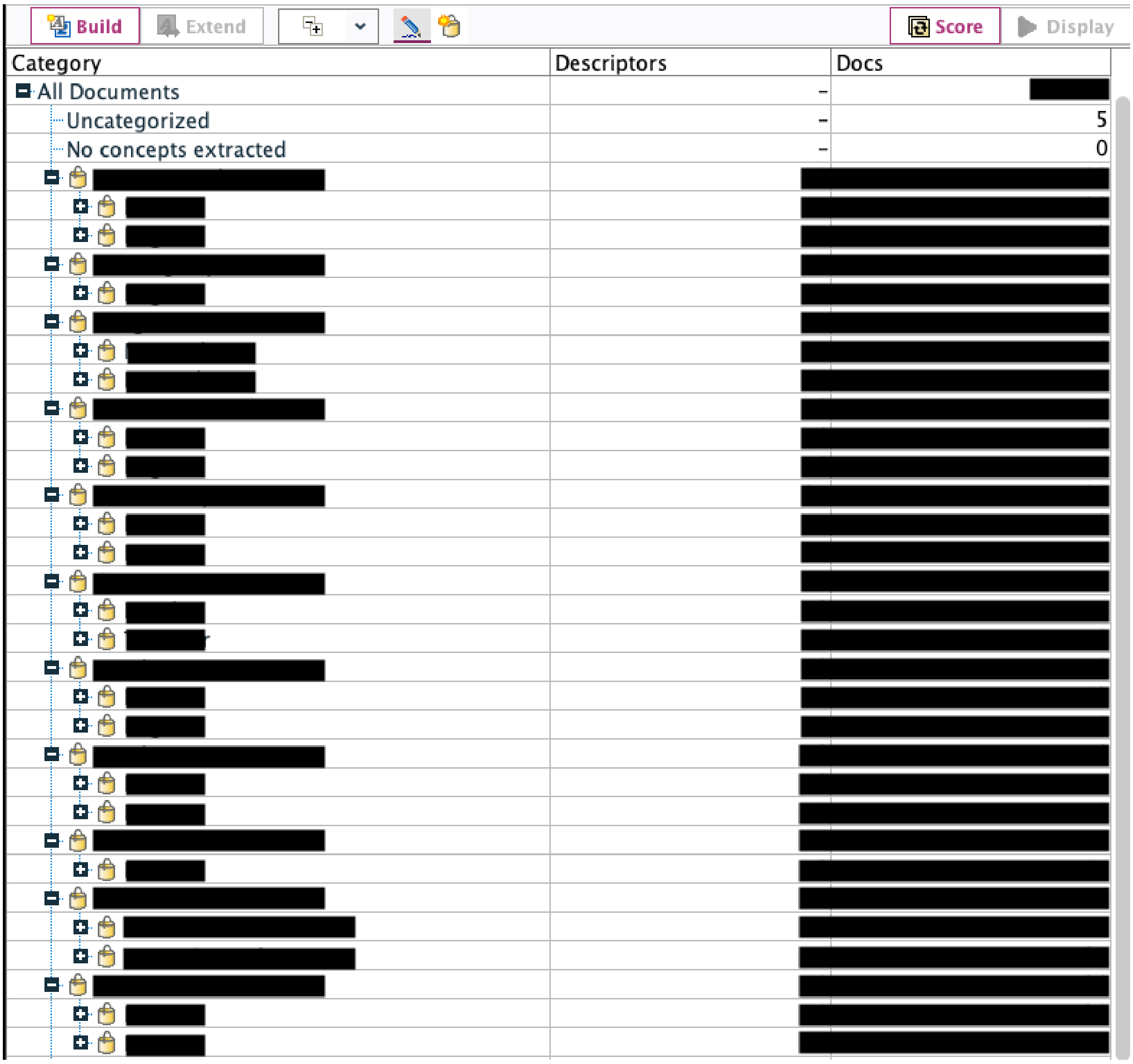
i) Provide a screenshot of your final text mining model showing **main categories** as well as **subcategories**.

ii) Provide the top 2 positive and top 2 negative subcategories based on document counts from your final text mining model.

- Each positive/negative category is to be documented as in Figures 4 and 5 (**need to show category rule(s) in your answer**).
- Each category should have an appropriate name, description, category rule(s) and screenshots of all/sample documents categorised in that category

(if a category has more than 8 documents, show at least 8 sample documents in your report).

iii) Based on the top 2 negative subcategories, provide at least 2 feasible recommendations to the Management of Crowne Plaza. Maximum 150 words for part e (iii)



Category	Descriptors	Docs
All Documents	-	[redacted]
Uncategorized	-	5
No concepts extracted	-	0
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]

Figure 6: Screenshot of your final text mining model

(44 marks)

All streams created in IBM SPSS Modeller are also required to be submitted in soft copy to Canvas (after you have submitted your GBA). Please save the stream as Group X GBA solution.str where X is your GBA group number. Save all the updates to the chosen resource template by going to File -> Update Modeling Node -> Keep session

work only. Do this multiple times and use a different stream name each time so that you can recover your work easily.

Your report should be succinct but not at the expense of excluding relevant details. Highlight points that are relevant to your discussion. Use plain and simple language. Some questions may not come with absolutely right or wrong answers. For such questions you have the liberty to express your views about the problem. However, your points must be supported by evidence and good reasoning. It's the quality and not the length that counts. Make sure you follow the report guidelines and style specified in this assignment.

The topics in the main report should be presented in the order according to the sequence of the tasks/questions listed in the assignment; that is, in the order of (a), (b), ..., etc. You can have several sub-sections within a section if you deem appropriate.

It is important to include all relevant tables and figures in the report as evidence to support the answers given.

The follow are some details of report format:

- Length: should not exceed 20 pages (including the relevant graphs and tables, but excluding the cover page, table of content, and appendix)
- Font Style: Times New Roman
- Font size: 12
- Line spacing: 1.5
- Margins: 1 inch or 2.54cm for top, bottom, right and left
- Include the page number on each page

Further suggestions:

- Ensure minimal grammatical and typographical errors
- Write in plain English
- Write clearly and appropriately to the context
- Cite appropriate sources
- Provide a reference or bibliography at the end of the main report
- Include less relevant details in the Appendix
- Good overall presentation of the report

---- END OF ASSIGNMENT ----