DIPLOMA IN DATA ANALYTICS CO-OP

Working with Data Using SQL and Python

Final Group Project

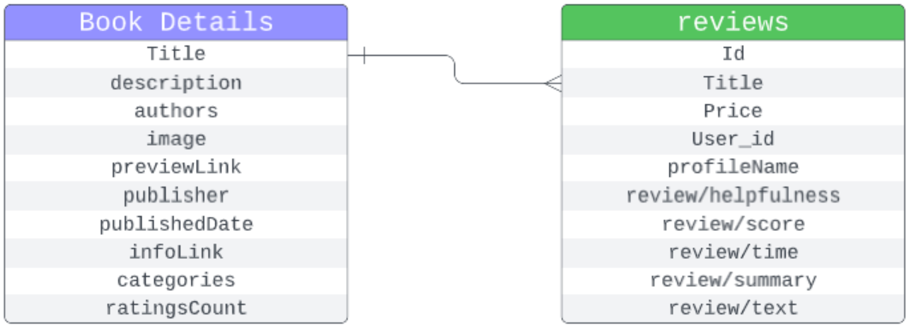
Instructions

- This is a group assignment worth 20% of the final grade.
- Read each section carefully and provide answers.
- The following are the expected deliverables:
    - **Deliverable 1 - Final Presentation (10%):** This is a technical presentation; Therefore, it must be professionally presented with adequate technical details. You may use a suitable presentation structure to address all tasks given. Each group must record their presentation using any collaboration tool and submit it along with the PPT file. Both the PPT file and the video must be submitted before the deadline. Please refer to the attached rubrics for the score breakdown.
    - **Deliverable 2 – Python Script Package (7%):** Each group needs to submit a Python code package (zip file). The file/s must be original and professionally arranged, and each code segment must be clearly explained by using comments. Please refer to the attached rubrics for the score breakdown.
    - **Deliverable 3 – Individual Reflection (3%):** Each “individual” needs to submit their own professional reflection as per the guidelines provided in this project assignment file.
- Time allocated for the presentation: ***15-20 minutes*** for the recorded group presentation.
- Toronto School of Management (TSoM) requires students to maintain high standards of academic integrity. Students are responsible for conducting themselves honestly and ethically in all aspects of their academic career and for becoming familiar with this policy and abiding by all aspects of it. To support academic honesty at TSoM, all work submitted by students may be reviewed for authenticity. In submitting their own work to TSoM, students consent to their submissions undergoing such a review and being retained in a database for comparison with other work submitted by fellow students.

Project Requirements (Deliverable 1 & 2)

Goal: Demonstrate proficiency in text mining (sentiment analysis) using Python.

Synopsis:



Book Details
Title
description
authors
image
previewLink
publisher
publishedDate
infoLink
categories
ratingsCount

reviews
Id
Title
Price
User_id
profileName
review/helpfulness
review/score
review/time
review/summary
review/text

The data package contains two files each corresponding to 212404 unique books that were available on Amazon between 1996 and 2014. 142.8 million reviews spanning May 1996 - July 2014 have also been collected. The two data files maintain the following structure:

File: Reviews	
Features	**Description**
id	The Id of Book
Title	Book Title
Price	The price of Book
User_id	Id of the user who rates the book
profileName	Name of the user who rates the book
review/helpfulness	helpfulness rating of the review, e.g. 2/3
review/score	rating from 0 to 5 for the book
review/time	time of given the review
review/summary	the summary of a text review
review/text	the full text of a review

File: Book Details	
Features	**Description**
Title	Book Title
Descripe	decription of book
authors	Neme of book authors
image	url for book cover
previewLink	link to access this book on google Books
publisher	Name of the publisheer
publishedDate	the date of publish
infoLink	link to get more information about the book on google books
categories	genres of books
ratingsCount	averaging rating for book

Your Tasks:

Task 1: Offer a brief explanation of the following terms:

1. ***[5 Marks]*** Text Mining
2. ***[5 Marks]*** Text Classification
3. ***[5 Marks]*** Text Clustering
4. ***[5 Marks]*** Sentiment Analysis

Task 2: Data Cleansing, Classification, and Word Clouds.

1. ***[5 Marks]*** Load the data into Python and perform the initial sanity checks.
2. ***[10 Marks]*** Generate initial analytical visualizations to understand the reviews (i.e., histogram) including a word cloud (using the wordcloud package). Discuss your findings.
3. ***[5 Marks]*** Classify reviews into two 'sentiment' categories called positive and negative.
4. ***[10 Marks]*** Generate positive and negative word clouds. Discuss your findings while comparing the positive and negative summaries (you may include other graphs if needed).

Task 3: Prediction

1. ***[15 Marks]*** Build a simple logistic regression model to predict the sentiment category based on a text-based review. Discuss your findings.
2. ***[15 Marks]*** Build a multinomial logistic regression model to predict the rating of a book based on its text-based review. Discuss your findings.

Deliverable 3:

Individual Reflection and Peer Review. ***[3% of the final score | 30 points]***

1. Each student on your team must write their independent reflection on the project. Your reflection article should be no more than 1 page (excluding the peer evaluation) and must be uploaded separately. Your task:
    a. ***[10 Marks]*** Choose a topic from the SQL series that you found most interesting and explain it.
    b. ***[10 Marks]*** Explain how well the chosen topic was applied in the project.
    c. ***[5 Marks]*** Explain the strengths the team demonstrated when executing the project.
    d. ***[5 Marks]*** Discuss the areas to improve when working in this group.
2. Complete and include the peer evaluation.

Marking Rubric: Deliverable 1 – PowerPoint File and Live Presentation

Question		0-20%	20-50%	50-80%	80-100%	**Final Marks**
Sections		The answer does not meet the question’s requirements. Invalid or incorrect answer.	A poorly structured and presented answer that demonstrates the lack of understanding of the subject matter. The section addresses some of the topic’s objectives. The answer has major grammatical errors.	A well-structured and presented answer. The answer addresses most of the underlying subject matters of the question. The section addresses most of the topic’s objectives. The answer has limited grammatical errors.	An exceptional answer that covers all the underlying subject matters of the question. The answer exerts critical thinking and includes examples. The section addresses all of the topic’s objectives. The answer has no grammatical errors.	**Marks Out of 100**
Introduction		0-2	2-3	3-4	4-5	
Task 1	**1: Text Mining**	0-2	2-3	3-4	4-5	
	2: Text Classification	0-2	2-3	3-4	4-5	
	3: Text Clustering	0-2	2-3	3-4	4-5	
	4: Sentiment Analysis	0-2	2-3	3-4	4-5	
Task 2	**1: Data loading**	0-2	2-3	3-4	4-5	
	2: Initial Visualizations	0-2	2-5	5-8	8-10	
	3: Sentiment Classification	0-2	2-3	3-4	4-5	
	4: Sentiment Analysis	0-2	2-5	5-8	8-10	
Task 3	**1: Simple logistic regression**	0-4	4-8	8-12	12-15	
	2: Multinomial logistic regression	0-4	4-8	8-12	12-15	
PPT Presentation File		0-2: Poor quality	2-3: Structure is maintained. Include Notes.	3-4: Structure is maintained. Descriptive Notes. Accurate referencing.	4-5: Structure is maintained. Accurate referencing. Well-written Notes. High-quality presentation.	
Group Presentation (Audio/Visual)		0-2: Poor Quality Presentation	2-5: Satisfactory Presentation	5-8: Very Good Presentation	8-10: Exceptional Presentation.	
Total						

Marking Rubric: Deliverable 2 – Python Code

Question		0-20%	20-50%	50-80%	80-100%	**Final Marks**
Sections		Invalid or incorrect answer. The code doesn't run, and the steps are incorrect.	The code works but the intermediate steps are not compatible with the answer. A correct answer with a poor approach. Some of the steps are correct, although the code doesn't work.	The code works and most of the steps meet the question's expectations. Some comments are available. Most of the steps are correct, although the code doesn't work.	The code works and all the steps meet the question's expectations. The code is well commented and well-structured.	**Marks Out of 70**
Task 2	**1: Data loading**	0-2	2-3	3-4	4-5	
	2: Initial Visualizations	0-2	2-5	5-8	8-10	
	3: Sentiment Classification	0-2	2-3	3-4	4-5	
	4: Sentiment Analysis	0-2	2-5	5-8	8-10	
Task 3	**1: Simple logistic regression**	0-4	4-8	8-12	12-15	
	2: Multinomial logistic regression	0-4	4-8	8-12	12-15	
Python code file quality		0-2: Poor quality	2-5: Structure is maintained. Some code is explained using comments.	5-8: Structure is maintained. Most of the important code sections are explained using comments.	8-10: Structure is maintained. All important code sections are explained using comments.	
Total						