NGEE ANN POLYTECHNIC

School of InfoComm Technology

Applied Analytics Assignment

Diploma in Cybersecurity & Digital Forensics
Diploma in Financial Informatics
Diploma in Information Technology
Year 2/3 (2022/2023), Semester 3/5

TEAM/INDIVIDUAL ASSIGNMENT

(40% of AA Module)

<u>Deadline for Submission:</u>

Presentation: 31th July 2022 (Sunday),23:59hrs
Report & Code: 14th August 2022 (Sunday),23:59hrs

Tutorial Group	:		
Team Number	:		
Tutor	:		
Members	:	Student No.	Student Name

Penalty for late submission:

10% of the marks will be deducted every day after the deadline.
NO submission will be accepted after 21st August 2022, 23:59 hrs.

1 Problem Statement

1.1 Objective

In this assignment, we will solve various Text Analysis problems using Python.

1.2 Dataset

The dataset (***bbc-text.csv***) consists of 2225 documents from the BBC news website corresponding to stories in five topical areas from 2004-2005. Detailed inform can be found from http://mlg.ucd.ie/datasets/bbc.html.

Column	**Description**
text	BBC news stories /Documents
category	The label for each document: business, entertainment, politics, sport, tech

Problem 1 (Individual) (70%)

Text analytics is the process of extracting meaningful information from natural language text. For example, in a news website, we may have news stories from different categories and text analysis techniques algorithm might be used to extract and associate various meaningful keywords in the stories/documents.

Using text analysis, you are required to extract keywords from each stories/document and analyze them using association rule mining.

1.3 Suggested Tasks

You are suggested to tackle this problem in *THREE* steps:

Step 1 – Text Data Preprocessing

- Download the datasets (***bbc-text.csv***) from POLITEMall.
- Cleanse the text data using proper python modules (e.g. NLTK, Regular Expressions).
- Transform the text data using Bag of Word and TF-IDF techniques.

Step 2 – Text Data Understanding

- Extract the Keywords for each document using TF-IDF matrix
- Analyze the extracted keywords using Association Rule Mining.
- Feel free to explore other suitable methods to analyze the text data, e.g. WordCloud.

Step 3 – Summarize the findings

- Summarize your work and provide suggestions for further improvements.

1.4 Suggested Report Format & Content Guidelines

Based on the above, write an **INDIVIDUAL** report with the following sections (see Table below). Sample content description is provided for each section. You are free to include other relevant information you deem necessary in the sections. You are strongly advised to use screenshots to capture details of work done.

*(Note: For a page with 1 inch margins, 11 point Calibri font, and minimal spacing elements, a good rule of thumb is **500 words** for a single spaced page)*

	Suggested Report Sections & Content Guidelines	**Word Count**
1.	Table of Contents	NA
2.	Introduction • Problem understanding and the approaches	Approx.: 750 words
3.	Text Data Preprocessing • Load and cleanse the text data • Transform the text data using Bag of Word and TF-IDF techniques	Approx.: 2000 words
4.	Text Data Understanding • Keywords extraction • Association rule mining on the extracted keywords • Other suitable methods	Approx.: 2000 words
5.	Summary and Further Improvements • Summarize your findings • Explain the possible further improvements	Approx.: 750 words

*Problem 2 (*Group：4-5 students per group*) (30%)*

Build a classification model to classify the BBC news stories (Documents) into different categories.

.

1.5 Suggested Tasks

You are suggested to tackle this problem in *TWO* steps:

Step 1 – Classification Modeling

- Sample the data into training data & testing data
- Build classification model(s) using training data to classify the BBC news stories (Documents) into different categories.
- Evaluate the model(s) performance (e.g. accuracy, confusion matrix and etc.) using testing data and see whether you can further improve the model performance through:
  - Tuning the model hyperparameters
  - Further cleanse or transform the text data
  - Other effective techniques

Step 2 – Summarize the findings

- Summarize your work and provide suggestions for further improvements.

1.6 Suggested Report Format & Content Guidelines

Based on the above, write an **Group** report with the following sections (see Table below). Sample content description is provided for each section. You are free to include other relevant information you deem necessary in the sections. You are strongly advised to use screenshots to capture details of work done.

*(Note: For a page with 1 inch margins, 11 point Calibri font, and minimal spacing elements, a good rule of thumb is **500 words** for a single spaced page)*

	Suggested Report Sections & Content Guidelines	**Word Count**
1.	Table of Contents	NA
2.	Classification Modelling • Build the model(s) • Evaluate and Improve the model(s)	Approx.: 2000 words
3.	Summary and Further Improvements • Summarize your findings • Explain the possible further improvements	Approx.: 750 words

2 Presentation and Demonstration

Each group will be allotted 40 minutes to present their Group & Individual work using slides. You are strongly suggested to add in screenshots, diagrams, images into your slides and practice the presentation in advance to make sure you can complete within 40 mins.

- Problem 1 (25 mins):
  - Individual Portion (20-25 mins): 5 mins per student
- Problem 2 (5mins):
  - Group Portion (5 mins)
- Q & A (10 mins)

The presentation will be conducted in Week 16/17 (**1st – 12th Aug 2022**) through Face-to-Face. Your tutor will provide detailed information later regarding to your presentation slot and etc.

3 Deliverables

For this assignment, you must submit all the following:

1. A set of **Presentation Slides** in POLITEMall
   - This is the set of presentation slides which you use to conduct your presentation.
   - The Final Slides Deck includes **One** Group Presentation Slides for Problem 1 and **Four/Five** Individual Presentation Slides for Problem 2.
   - Deadline for the slides submission is **Sunday 31st Jul 2022, 2359 hours**
2. A softcopy **Final Report** in POLITEMall.
   - The Final Report includes **One** Group Report for Problem 1 and **Four/Five** Individual Reports for Problem 2.
   - Deadline for report submission is **Sunday 14th Aug 2022, 2359 hours**
3. The **completed "AA_Assignment2_<Grp>_<studentname>.ipynb"** Jupyter Notebook File in POLITEMall.
   - The zip file includes **One** Group Jupyter Notebook for Problem 1 and **Four/Five** Individual Jupyter Notebooks for Problem 2.
   - Deadline for Jupyter Notebook submission is **Sunday 14th Aug 2022, 2359 hours**

Note: DO NOT PLAGIARIZE (please refer to Ngee Ann Polytechnic Plagiarism Policy webpage for more information)

4 Grading Criteria

	Problem 1	**Problem 2**	**Component Weightage**
	Individual	**Group**	
Presentation	**35%**	**15%**	**50%**
Final Report	**35%**	**15%**	**50%**

	Grading Criteria	**Component Weightage**
Presentation	a) Quality of work b) Flow of presentation based on content guidelines (see section 1.4) c) Quality of presentation slides d) Presentation and articulation skills	**50%**
Final Report	a) Quality of work b) Completeness of report based on suggested report sections and content guidelines (see section 1.4) c) Clarity of report, use of proper visual aids and use of proper grammar d) Quality of discussions and recommendations for further improvements	**50%**