NGEE ANN POLYTECHNIC

School of InfoComm Technology

Applied Analytics Assignment

Diploma in Cybersecurity & Digital Forensics
Diploma in Data Science
Diploma in Information Technology
Year 2/3 (2023/2024), Semester 3/5

TEAM/INDIVIDUAL ASSIGNMENT

(40% of AA Module)

Deadline for Submission:

Presentation Slides: 13th August 2023 (Sunday),23:59hrs
Report & Code: 13th August 2023 (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 20th August 2023, 23:59 hrs.

1 Problem Statement

1.1 Objective

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

1.2 Dataset

EvolutionAI has collected a dataset of roughly 1M text Reddit posts, with 1013 distinct classes (1000 examples per class). The classes are based on the assumed 'topic' of the text post, the topics being a manually curated taxonomy based on subreddits.

For more information on the original dataset, please refer to the link: [The reddit self-post classification task | Kaggle]

In this assignment, we will be dealing with a subset of that data, a total of 2500 articles, split amongst 5 labels - each label contributing 500 articles to the pool.

The 5 labels are: 'soccer', 'snowboarding', 'triathlon', 'judo', and 'surfing'.

Column	Details
text	Reddit post
category	The label for each document/article: 'soccer', 'snowboarding', 'triathlon', 'judo', and 'surfing'.

Problem 1 (Individual) (70%)

1.3 Suggested Tasks

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

Step 1 – Text Data Preprocessing

- Download the dataset (***reddit_5.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.: 250 - 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.: 1000 - 2000 words
4.	Text Data Understanding • Keywords extraction • Association rule mining on the extracted keywords • Other suitable methods	Approx.: 1000 - 2000 words
5.	Summary and Further Improvements • Summarize your findings • Explain the possible further improvements	Approx.: 500 - 750 words
6.	Reflection • Suggest possible further improvement(s) to the current solution. • With reference to the module learning objectives stated, reflect on the skills learnt and the skills you could have learnt better.	Approx.: 500 - 1000 words

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

Step 1 – Classification Modeling

- Sample the data into training data & testing data
- Build classification model(s) using training data to classify the Reddit articles (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.

1.5 Suggested Report Format & Content Guidelines

Based on the above, write a **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 35 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 (5 mins)

The presentation will be conducted Face-to-Face in Week 18 (**14th – 20th Aug 2023**). Your tutor will provide detailed information later regarding your presentation slot and other arrangements. Please remember to **dress professionally and appropriately** for your presentation.

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.
   - Deadline for the slides submission is **Sunday 13th Aug 2023, 2359 hours**
2. A softcopy **Final Report** in POLITEMall.
   - Deadline for report submission is **Sunday 13th Aug 2023, 2359 hours**
3. The **completed "AA_Assignment_<Grp>_<studentname>.ipynb"** Jupyter Notebook File in POLITEMall.
   - Deadline for Jupyter Notebook submission is **Sunday 13th Aug 2023, 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	**20%**	**10%**	**30%**
Final Report	**50%**	**20%**	**70%**

	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	**30%**
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	**70%**