**Group:** Form a group comprises of **Strictly Four** members to work on this assessment component. Group members are expected to hold equal responsibility in completing the assignment.

Assignment - Group - Part A (50 Marks)

In this part of the assignment, you are expected to carry out **practical text pre-processing, basic processing, and text normalization** in Python using Natural Language Toolkit (NLTK), Regular Expression (re), and textblob.

Work on the following questions and write suitable answers along with Python codes. Share the relevant codes in the appropriate position to support each answer in the report.

Text Data

Use **Data\_1.txt** to answer Q1 & Q2 and use **Data\_2.txt** to answer Q3 given in Assignment Data.zip.

Q1. Form tokenization and Filter stop words & punctuation (10 marks)

1. Demonstrate word tokenisation using the *split function*, *Regular Expression and* *NLTK packages* separately and report the output. (2 marks)
2. Justify the most suitable tokenisation operation for text analytics. Support your answer using obtained outputs. (3 marks)
3. Demonstrate *stop words and punctuations* removal and report the output suitably along with the stop words found in the given text corpus. (2 marks)
4. Explain the importance of filtering the stop words and punctuations in text analytics. (3 marks)

Q2. Form word stemming (10 marks)

1. Explain the importance of stemming in text analytics (3 marks)
2. Demonstrate word stemming using *Regular Expression, Porter Stemmer* and *Lancaster Stemmer* and report the output. (3 marks)
3. Explain the differences among *Regular Expression stemmer, Porter Stemmer* and *Lancaster Stemmer* using the obtained output. (4 marks)

Q3. Form Parts of Speech (POS) taggers & Syntactic Analysers (10 marks)

1. Demonstrate POS tagging using *NLTK POS tagger, textblob POS tagger* and the R*egular Expression tagger* and report the output. (3 marks)
2. Explain the differences of the POS taggers using the output obtained in the above question. (3 marks)
3. Draw possible parse trees for the given sentence using suitable python codes and report the output along with the code. (4 marks)

Q4. Work on sentence probabilities (10 marks)

Use the **Data\_3.txt** given in the Assignment Data.zip to Carry out the following tasks.

The file **Data\_3.txt** is a ***text*** ***corpus*** which contains three sentences. Each sentence is limited by the sentence pads such as <s> and </s> as the starting and end of the sentence respectively. You are expected to use the appropriate equations to perform the respective tasks.

*Perform suitable pre-processing if required.*

1. Compute manually the sentences probabilities using the *unsmoothed* bigram model. (3 marks)
2. Compute manually the sentences probabilities using the *smoothed* bigram model. (3 marks)
3. Implement in python using both *unsmoothed* and *smoothed* bigram language models and report the respective sentence probabilities. (4 marks)

**Individual Component (10 Marks)**

**Q5. Alternative Approach Implementation**

Each student must independently implement an **alternative technique** for the task Q1 and compare it with the group’s approach.

* Implement tokenization using an alternative approach. (3 marks)
* Compare and contrast the alternative approach with the group’s approach. (2 marks)
* Explain why the alternative approach is better, worse, or just different. (5 marks)

Assignment - Group - Part B (50 Marks)

In this part of the assignment, you are required to work on the **implementation of text analytics techniques and methodologies** using Python.

**Objective:** The main objective of this assignment is to build an effective Supervised Text Classification (Sentiment Prediction/Spam Detection/Authorship Identification/Fake Review Detection/Cyberbully Detection/etc.) model using the suitable dataset.

Q1. Dataset and Exploratory Data Analysis (EDA) (12 marks)

1. Choose a medium size secondary dataset, from the suitable and possible resources. Define the classification problem and describe your selection and suitability of the dataset for the defined problem. You may produce a comprehensive literature review on some previous studies on the chosen dataset. Justify your selection of such dataset on the suitability for your work.

(7 marks)

1. Perform an Exploratory Data Analysis (EDA) on your dataset using the suitable methods. Critically analyse the dataset and suggest **n** (n = no. of members in the group) predictive models that may be suitable for your dataset and to discover the solution for the defined problem. All the data preparation activities are expected to take place and reported.

(5 marks)

Q2. Supervised Text Classification (14 marks)

Build Supervised Text Classification models using the suitable Python libraries such as NLTK, Sklearn, Spacy, etc and report the relevant model performance measures.

The python codes must be neat with clear output. Provide relevant comments in the code to explain the purpose of the code snippet.

(14 marks)

Q3. Hyper parameter selection (12 marks)

Perform the following tasks for the models built in Q2.1

1. List and explain the hyper parameters of all your models. (5 marks)
2. Using grid search or random search methods, report the hyper parameters of your models with results. (7 marks)

Q4. Evaluation & Discussion of the predictive models’ results (12 marks)

1. Select evaluation metrics and briefly describe them. (3 marks)
2. Perform evaluation and present your results. (3 marks)
3. Discuss and critically analyze the results your predictive models built to choose the best one. You may compare the obtained results with the previous works. (6 marks)

Deliverables:

**Report:**

Word count: 4000 words

You shall submit your work in **a report** format. Each section should consist of necessary introduction, brief description of the chosen methods together with a justification of your choice, clear presentation of the results with appropriate graphics and/or tables and critical analysis.

**Codes:** The output must be as specified in respective questions. Suitable comments must be inserted in proper places in the code.

**Softcopy:** The relevant softcopies such as the report *(.doc or .docx*) and the python code files *(.py or. ipynb)* must be uploaded via the specified submission links available in the MOODLE.

Academic Integrity

Copying or paraphrasing someone's work (code included) or permitting your own work to be copied or paraphrased, even if only in part, is not allowed, and will result in disciplinary action. Your grade should reflect your own work.

Basically, 'plagiarism' means representing someone else's work as if it is your own. This is a very serious academic offence for all students within the University regulations and is particularly reprehensible for a researcher. Please do not even consider it. Remember that accidental plagiarism (or the appearance of it) may be avoided by referencing your work properly. This gains you credit, not loses it! The simple rule is that you must not represent the ideas of other people (whether they are published works or the work of other students) as your own.

Carefully read APU guidelines and policies on the Webspace. The golden rule on plagiarism is **DO NOT DO IT**.