Group Assignment (25% of the total module marks)

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

**Group:** Form a group comprises of **Four or Five** members to work on this assessment component.

Use the **Text Corpus.txt** given in the Group Assignment Data folder to answer Q1 – Q3 of this assessment component.

The file **Text Corpus.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. The unknown words should be treated as UNK.

Perform suitable pre-processing to perform the following tasks.

Q1. Form a unigram language model (25 marks)

Implement unsmoothed and smoothed unigram language models in python using any available techniques and report the output.

Q2. Form a bigram language model (25 marks)

Implement unsmoothed and smoothed bigram language models in python using any available techniques and report the output. (12 marks)

Q3. Work on sentence probabilities (25 marks)

Using the *smoothed* model values, carry out the following tasks.

1. Compute manually the sentences probabilities using the unigram model. (5 marks)
2. Compute manually the sentences probabilities using the bigram model. (5 marks)
3. Justify which language model is more suitable to calculate the sentence probabilities. (5 marks)
4. Implement and report the respective sentence probabilities in python using both unigram and bigram language models. (10 marks)

Q4. Supervised Text Classification (25 marks)

Use the **Musical\_Instruments\_Reviews.csv** data set available in the **Group Assignment Data** to perform the following tasks.

1. Export the resulting data set along with the predicted sentiments to a .csv format using python and report the relevant code used for this operation. *(Note: The resulting data set must be submitted via the given submission link in MOODLE).* (5 marks)
2. Use the data set exported in the Q4.2 to build a supervised sentiment classification model using *Naïve Bayes Classifier* from NLTK and report the following model performance measures:
3. Accuracy
4. Precision
5. Recall
6. F1 Score

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

Deliverable:

1. Complete & running Python code.
2. Output stating the above FOUR (04) performance measures.

(20 marks)

Deliverables:

**Report:**

Word count: 2000 words

The report must be prepared in a professional manner following the proper documentation formats.

**Codes:** Python code should run without any arguments. It should read files in the same directory. 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 or .pdf*) 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**.