# Ghulam Ishaq Khan Institute (GIKI)

## Assignment # 2

| | |
|---|---|
| **Subject:** Natural Language Processing | **Course Code: AI** - 361 - A - Spring - 23 |
| **Class:** BS AI, **Batch:** Fall – 2020 | **Submission Deadline:** 02/May/2023 - Tuesday (11:59 - PM) |
| **Course Instructor:** M. Qasim Riaz - Lecturer - FCSE | **Total Marks:** 50 (Marks are divided problem wise) |
| **Course TA:** Mr. Ali Aftab | |

**Note (Read notes & instructions first)**

- First of all, read the instructions and statements of each exercise/question carefully then write the solution.
- It is written in front of each question that you have to upload it as handwritten or Python Notebook File.
- *For Handwritten:*
  - In case of multiple questions, give heading of each question's number or the exercise you are going to solve (don't write statement of question)
  - Mention page number and your roll number at corner of each page.
  - Take pictures using applications like camscanner on your phone.
  - Then select all pictures in the camscanner application and convert them into a pdf file using option in application.
  - The name of your pdf file should contain your assignment number and your roll number as shown in following example, For Example if your roll number is 2022532 and you have done assignment number 2 then the name of file should be as ---> 2022532_2.pdf
  - Then upload that pdf file at Microsoft teams. Remember the sequence of pages should be right.
  - Also keep the same original pages/hardcopy with you so that you can show/submit me later if required.
- *For Jupyter Notebook Code File:*
  - Create different file for each question & assignment
  - Name of each file should contain your roll number, assignment number & question number in a specific format.
  - For Example, if your roll number is 2022532 you are doing 2nd assignment and question no 5 then file name of your Python Notebook file should be written as ---> 2022532_2_5.py (Similarly, create for each question)
  - Now upload all of these files at Microsoft teams.

**CHEATING/COPY CASE or LATE SUBMISSION even 1 minute late will be graded as STRAIGHT ZERO MARKS.**

**So be on time make no excuse.**

## Machine Learning Based Classification of News Text

**Assignment Description:**

In this assignment, you will be working with two popular text classification datasets, namely, the 20 Newsgroups Dataset and the Reuters News Dataset. You will be implementing and comparing the performance of multiple machine learning models for text classification, including Naive Bayes, Logistic Regression, SVM, and Random Forest.

**Getting Dataset:**

1. ***20 Newsgroups Dataset:***
   - The dataset can be downloaded from the official scikit-learn library using the fetch_20newsgroups() function. Here's an example code to download the dataset:

```
from sklearn.datasets import fetch_20newsgroups
# Download the dataset
newsgroups_train = fetch_20newsgroups(subset='train')
newsgroups_test = fetch_20newsgroups(subset='test')
```

   - It can also be downloaded from the UCI Machine Learning Repository using this link:

     https://archive.ics.uci.edu/ml/datasets/Twenty+Newsgroups

2. ***Reuters News Dataset:***

- The dataset can be downloaded from the official NLTK library using the reuters corpus. Here's an example code to download the dataset:

```
import nltk
nltk.download('reuters')
from nltk.corpus import reuters

# Get the list of fileids for training and testing sets
train_docs = [d for d in reuters.fileids() if d.startswith("train")]
test_docs = [d for d in reuters.fileids() if d.startswith("test")]

# Load the dataset
train_data = [reuters.raw(doc_id) for doc_id in train_docs]
train_labels = [reuters.categories(doc_id)[0] for doc_id in train_docs]
test_data = [reuters.raw(doc_id) for doc_id in test_docs]
test_labels = [reuters.categories(doc_id)[0] for doc_id in test_docs]
```

- It can also be downloaded from the NLTK Data repository using this link:
  https://www.nltk.org/book/ch02.html#reuters-corpus

**Your Tasks:**

In this assignment, you will be working with above mentioned dataset. Your tasks are as following:

1. Dataset Preparation:
   1.1. Download and extract the 20 Newsgroups Dataset and the Reuters News Dataset.
   1.2. Preprocess the text data by removing stop words, stemming, and lemmatizing, as per your preference.
   1.3. Split the datasets into training and testing sets in a ratio of 80:20.
2. Feature Extraction:
   2.1. Convert the preprocessed text data into numerical vectors using one of the following feature extraction techniques: Bag of Words, TF-IDF, or Word Embeddings.
3. Model Implementation:
   3.1. Implement Naive Bayes, Logistic Regression, SVM, and Random Forest models for text classification using the scikit-learn library.
   3.2. Train each of the models on the training set and evaluate their performance on the testing set using evaluation metrics such as accuracy, precision, recall, and F1-score.
4. Model Comparison:
   4.1. Compare the performance of the four models based on the evaluation metrics and choose the best performing model.
   4.2. Show accuracy, precision, recall and f-measure of all four models in a graph using Matplotlib.

**Deliverables Your submission should include the following:**

- A Jupiter notebook with your code and analysis.
- Handwritten Report of a half page that mentions why results of one model is better than the other

**Note: Any student proved/caught with any kind of cheating/plagiarism/late submission will be subject to zero marks**

**o --- | --- Good Luck --- | --- o**