PROJECT TITLE: Scientific Research Papers have 2-3 tables defining various quantitative and non-quantitative aspects of the paper. In this project we are going to be finding out summaries defining these tables to create a dataset of table summarization.

PROJECT DESCRIPTION:

Section 1: Download and Preprocessing:

1. Download 1000 unique scientific papers in pdf format.
2. Write a program in python which converts the papers from pdf to text and XML formats.
3. The papers should be divided into folders by types for eg. Summarization, IOT, Sentiment analysis, translation etc.
4. Rename the text and the xml files in a uniform format and keep them in well defined folders.

Section 2: System Design:

Summary can be of two types, abstractive and extractive. For each table in each scientific paper you are going to be executing these steps:

1. For each table in each paper, the abstractive summary will be the caption of the table and the extractive summary will be the text in the paper mentioning the table by its number.
2. Resulting Summary have to be written after that.

But If the table is mentioned multiple times then there will be multiple extractive summary for one table.

Section 3: Output File preparation:

We have to prepare a single text file for 1000 papers. The format of the text file will be like this:
<Paper ID =1> <Table ID =1> <Abstractive Summary> =Table 1: Correlations between input features and average system performance for multi-document inputs of DUC 2001-2003, 2004G (generic task), 2004B (biographical task), All data (2002-2004) - UNnormalized and Normalized coverage scores</Abstractive Summary> <Extractive Summary> = NULL</Extractive Summary> </Paper ID =1>

Section 4: Submission checklist for all groups:

1. The well defined folders containing the 1000 scientific papers in PDF, TXT and XML formats .
2. All the python codes used in all sections with clear cut naming conventions.
3. The output .txt file in the above mentioned format.