Mid Semester Report

On

# The development of Data extraction-based NLP project


**BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI**

**June 2022**

Mufaddal Jiruwala
2020A7PS1720H

Mid-semester Report
on
Practice School-1
at

# National Informatics Centre, Hyderabad (Remote)

Prepared in partial fulfillment
ofPractice School -1
For
Dr. Pradheep Kumar
Faculty Instructor, Practice School - 1
National Informatics Centre (NIC)


Prepared by
Mufaddal Jiruwala
2020A7PS1720H

B.E. Electronics and Electrical
Birla Institute of Technology and Science, Pilani

Hyderabad Campus

# Acknowledgement

Sincere thanks to all those people who have helped me immensely through the course of my project until now. I appreciate my institution, Birla Institute of Technology and Sciences, Pilani, for running the Practice School program, which exposes us to the industry and gives us experience thatallows us to work on this project.

My special thanks to Dr. Pradheep Kumar, PS instructor at BITS Pilani, for his constant guidance and supervision as well as for providing beneficial suggestions on the project.

I am very grateful to the staff of NIC for conducting various programs for industry exposure and giving a platform for the project.

**TABLE OF CONTENTS**

# Introduction

- The project aims to generate a squad2.0 dataset containing questions and their answers from the text having information about various policies and schemes of the government
- It uses various NLP techniques like lemmatization, regular expression and various python libraries like pandas, nltk.
- Our main task is to fine tune a pre-trained Bert question answer model

# About the PS station-NIC

- National Informatics Centre (NIC) was established in 1976, and has since emerged as a "prime builder" of e-Government / e-Governance applications up to the grassroots level as well as a promoter of digital opportunities for sustainable development.
- NIC, through its ICT Network, "NICNET", has institutional linkages with all the Ministries /Departments of the Central Government, 35 State Governments/ Union Territories, and about 718 District administrations of India.
- NIC has been instrumental in steering e-Government/e-Governance applications in government ministries/departments at the Center, States, Districts and Blocks, facilitating improvement in government services, wider transparency, promoting decentralized planning and management, resulting in better efficiency and accountability to the people of India.

- **Software/Applications Used**
    1. Python
    2. NLP Techniques
    3. Pre-Trained Bert Model

- **Tasks performed to date**

| | |
|---|---|
| Task 1 | ➢ Told about Data extraction-based project<br>➢ Tried out data extraction from a sample pdf using python libraries |
| Task 2 | ➢ Learned about different NLP data cleaning techniques<br>1. Lemmatization<br>2. Regular expression<br>3. Tokenization<br>4. Removing stopwords<br>➢ Cleaned text of a sample pdf using these techniques |

| | |
|---|---|
| Task 3 | ➢ Learned about Bert model and its applications<br>➢ Tried pre-trained Bert model to generate answers of manually typed questions from sample text |
| Task 4 | ➢ Modified python scripts given by Nic mentors to generate question-answers fromcontent about various policies and schemes of various sectors undertaken by government.<br>➢ Generated a json file from the text file containing content and question answers based on it. |

- **Other works**
    1. Learned about the working and business model of NIC and participated in quiz
    2. Had a group discussion whose topic was “Work from home vs Work from Office” where everyone had their unique viewpoints and final conclusion supported hybrid model depending on type of work and conditions
    3. Maintained a weekly diary where all the work and information regarding the project was noted down

Mufaddal Jiruwala

2020A7PS1720H

# Glossary

**NLP**: Natural language processing (NLP) refers to the branch ofcomputer science—and more specifically, the branch of artificial intelligence or AI—concerned with giving computers the ability tounderstand text and spoken words in much the same way human beings can.

**Bert**: Bidirectional Encoder Representations from Transformers(BERT) is a transformer-based machine learning technique for natural language processing (NLP) pre-training developed by Google.

**Lemmatization:** Lemmatisation (or lemmatization) in linguisticsis the process of grouping together the inflected forms of a wordso they can be analyzed as a single item, identified by the word's lemma, or dictionary form.

**Stopwords**: A stop word is a commonly used word (such as “the”, “a”, “an”, “in”) that a search engine has been programmedto ignore, both when indexing entries for searching and when retrieving them as the result of a search query.

**Transformers**: A transformer is a deep learning model that adopts the mechanism of self-attention, differentially weighting the significance of each part of the input data. It is used primarily in the fields of natural language processing (NLP) and computer vision (CV).