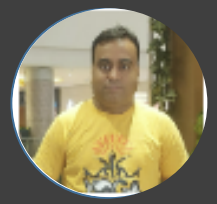
AI, ML | Dataset creation and classification of Unit test cases

Problem Statement

- Whenever developer writes a code we need to check the code covers all the necessary cases and don't crashes on production, so to avoid these issues tester writes test cases to check the quality of code before deploying it on production.
- Test cases specially Negative test cases are very important to avoid any crashes or serious bug in code.
- Objective of this work let is to create the dataset of Positive and Negative test cases(Python) and Automate the process of classifying test cases into Positive or Negative Using AI.
- Team would be required to create dataset from open source and build AI Algorithm to classify the test cases into Positive or Negative.

Objective

N Devaraj, Lead Engineer
devaraj.n@Samsung.com
+91-9663795469



Anandamay, Sr.Chief Engineer
anandamay.c@Samsung.com
+91-7411072266

Additional Documentation:

- https://brightsec.com/blog/unit-testing/
- https://www.geeksforgeeks.org/positive-vs-negative-vs-destructive-test-cases/
- https://www.softwaretestinghelp.com/positive-and-negative-test-scenarios/
- https://arxiv.org/pdf/1910.03474.pdf
- "Text Classification via Large Language Models" - https://arxiv.org/pdf/2305.08377.pdf

Work-let expected duration – 6 months

3-5 Members

Expectations

- Create dataset for Python Language which should consist :
  1. function code,
  2. corresponding test case and
  3. class of test case (Positive or Negative)
- Augment the dataset to increase the dataset and balance positive and negative class count
- Verification and Validation of dataset classes
- Find and Train Suitable Language model/Custom Generative model (BERT/GPT family LLMs) for classification task
- Tune the hyper-parameters to get Accuracy>60% and F1 Score>50%

Training/ Pre-requisites

- Hands-on Python
- Knowledge of writing test cases for python function
- Knowledge of Deep Learning
- Knowledge of NLP and Large Language models

Kick Off < 1st Month >

- Understanding of Function code writing and corresponding Test case classes
- Understand the NLP Basics like Text Embedding, Tokens, vocabulary, Encode-Decoder, Transformer

Milestone 1 < 2nd Month >

- Open source dataset preparation – Find any open source dataset , Data scraping
- Create dataset from Open source Language model like ChatGPT, BARD or LLaMA etc.
- Automating the labelling of test cases

Milestone 2 < 4th Month > Verification and validation of Dataset classes

Split The Dataset into Train, Validation and Test set keeping class balanced

Develop ML pipeline for text (Function + Test case) classification using Language model like BERT/GPT

Closure < 6th Month > Iterative Tuning of hyper parameters and

Iterative testing the models and get the best model among experiments

Develop prediction script to run in terminal- Given Function-Test case , It should classify Negative or Positive.