IR
Assignment 4
Report

- Aditya Nangia
- 2020168

Custom Dataset

```
class GPT2ReviewDataset(Dataset):
    def __init__(self, tokenizer, reviews, max_len):
        self.max_len = max_len
        self.tokenizer = tokenizer
        self.eos = self.tokenizer.eos_token
        self.eos_id = self.tokenizer.eos_token_id
        self.reviews = reviews
        self.result = []

        for review in self.reviews:
            # Encode the text using tokenizer.encode(). We add EOS at the end
            tokenized = self.tokenizer.encode(review + self.eos)

            # Padding/truncating the encoded sequence to max_len
            padded = self.pad_truncate(tokenized)

            # Creating a tensor and adding to the result
            self.result.append(torch.tensor(padded))

    def __len__(self):
        return len(self.result)


    def __getitem__(self, item):
        return self.result[item]

    def pad_truncate(self, name):
        extra_length = len(tokenizer.encode(" TL;DR "))
        name_length = len(name) - extra_length
        if name_length < self.max_len:
            difference = self.max_len - name_length
            result = name + [self.eos_id] * difference
        elif name_length > self.max_len:
            result = name[:self.max_len + 3]+[self.eos_id]
        else:
            result = name
        return result
```

Rouge Scores

```
==================================================
TOTAL SCORES
==================================================
ROUGE 1
==================================================
Precision :0.15880294811960785
Recall :0.16046505461872948
F1 Score :0.13756710844493886
==================================================
==================================================
ROUGE 2
==================================================
Precision :0.02935442081887263
Recall :0.03153568777176795
F1 Score :0.025654143404486223
==================================================
==================================================
ROUGE L
==================================================
Precision :0.15541859141521178
Recall :0.1562466857567081
F1 Score :0.13450399596269583
==================================================
```