# IRE Assignment 3

## Detecting Well-Formed Search Queries

Deadline: 16th November 2021, Tuesday, 10:00 AM (IST)

## About the Assignment

In this assignment, you will read, understand and improve the following paper published at ECIR 2019:
[Inductive Transfer Learning for Detection of Well-formed Natural Language Search Queries]

In this paper, the authors study the well-formedness of a search query. Through this paper, you will understand the importance of well-formed queries and how NLP can be used to understand the queries and later be used in different downstream IR tasks.

Link to the dataset used in the paper:
[Query-wellformedness Dataset]

The assignment has two parts:

1. Theory Part - To summarize and evaluate your understanding of the paper.
2. Coding Part - To try hands-on a pretrained transformer model to take a shot at implementing the task.

Following is the distribution of the assignment questions along with their weights:

| Question No. | Nature of the question | Weight |
|---|---|---|
| 1 | Theoretical | 8 |
| 2 | Theoretical | 8 |
| 3 | Theoretical | 20 |
| 4 | Theoretical | 8 |
| 5 | Theoretical | 8 |
| 6 | Coding | 40 |
| *6 - Bonus* | *Coding* | *50* |

| | | |
|---|---|---|
| 7 | Theoretical | 8 |
| **Total** | **--** | **100 *(150 w/bonus)*** |

The weights have been indicated in bold square brackets below their respective questions.

# Theoretical Understanding Questions

Answer the following questions. Use diagram(s) wherever required.

**Q1)** What is the problem that the authors are trying to solve and why is it significant? Answer the question in not more than a paragraph.
**[8]**

**Q2)** What is transfer learning? Answer the question in not more than a paragraph.
**[8]**

**Q3)** Explain the architecture of the paper in your own words. Answer the question in not more than three paragraphs.
**[20]**

**Q4)** Discuss and describe the dataset in not more than one paragraph.
**[8]**

**Q5)** Describe the results and takeaways from the paper. Answer the question in not more than a paragraph.
**[8]**

# BERT-based Implementation

The recent storm of transformer-based pretrained language models has achieved SoTA in different NLP fields. In this question, you will get a hands-on experience on one of the most popular language models, BERT, and how it can be finetuned for any task.
Before starting this assignment you are expected to read about the basics of transformers (which is discussed in class) and BERT (also discussed in class, and a lot of good blogs are out there).

## Coding Task

**Q6)**
In the paper, the authors pretrained the AWD-LSTM model and finetuned it for their task. For this question, you have to take a publicly available BERT model and finetune it to predict for query well-formedness. More precisely, there is no need to pretrain the BERT model, instead, take the publicly available pretrained model and *simply finetune* it for the given problem statement. The choice of whether to lower case the text (BERT-uncased) or not (BERT-cased) is up to you. Also, you are advised to pay special attention to hyperparameter tuning for good results.
**[40]**

Please note that finetuning the BERT model might sound like a very difficult task, but in reality it is very easy yet highly effective. There are various blogs/tutorials that explain it, so you can freely use them for reference. The task can be easily done on Google Colab, so there would be no major computational issues as well.

## Bonus Task

This task is slightly tough and requires more a bit more effort compared to the above question. So not everyone has to do it, but those who do would be rewarded with a bonus.

**Q6) - Bonus**
As discussed in the paper, instead of directly finetuning a language model, they further pretrain it on the target dataset before finetuning it. Similarly, for this question, you have to *first further pretrain the BERT model* on the queries (most recommended using the Masked Language Modeling task) and then finetune it to predict for query well formedness.
**[50]**

Read more about this idea in the paper:
Don't Stop Pretraining: Adapt Language Models to Domains and Tasks.

## Theory question: to be done after the coding task

**Q7)**
In this paper, we understood the task of binary prediction for query well-formedness and finetuned a BERT model on the same task. Now, how can the same architecture be *extended to predict the rating of the query*? The rating has to be predicted in steps of 0.2 from 0 to 1, with 0 indicating a bad query and 1 indicating a perfect query. (i.e. the BERT model has to predict query well formedness rating instead of classifying for the same in a binary fashion).

Answer the question in not more than a paragraph.
**[8]**

The students have to submit a pdf answering the questions for the theory part. The code can be submitted either as a notebook or Python script file(s), as convenient to you.

# Submission Guidelines

1. Your submission folder is expected to include ***at least two files***:
    a. Answers.pdf - A PDF file containing answers to the theoretical questions. For the coding question(s), brief your results obtained.
    b. One or more files containing the solution code for the coding question(s). You are free to submit it as a *notebook* or *python script* at your convenience. Please note that the code should be reproducible for your results claimed in Answers.pdf. In case there exist multiple files, explain their functions in a readme file.
2. Zip the folder, name it in the format **<Roll_Number>.zip**. We hope the convention is self-explanatory, so please stick to the same while uploading on Moodle.
3. Please maintain academic discipline while completing the assignment, any detected violations can result in ***zero credit*** for the submission.
4. ***No extension*** would be possible beyond the announced deadline, so please plan accordingly.