#CS657A

GROUP 9

MAHABHARATA

INFORMATION RETRIEVAL

Under **Prof. Arnab bhattacharya**

Team Members

Akanksha singh (21111005)

Arjun Singh (21111402)

Ayush Sahni (21111019)

Ayush Singh (21111020)

Neeraj Chouhan (21111044)

MOTIVATION 

- The Ramayana and the Mahabharata (Mahabharat in Hindi) can be considered as the greatest epics ever in the history of human civilization. They not only entertain the masses, but also enlighten them about duty, morality and salvation.
- These epics contain a lot of information about the ancient world and how people of that time used to live.
- Unfortunately, the two epics did not make it to the popular kids literature, school curriculum, or popular folklore in the West. Even the western film world chose to ignore them completely because their themes and spiritual values are alien to the western world.
- There should be easily and efficiently accessible sources of information from where people can learn about these epics. But they are very huge to read.

INTRODUCTION

- Question-answering is crucial to the Information Retrieval as it requires retrieval of relevant information from the corpus.
- Context is very important for retrieving relevant information for such applications.
- We are presenting a Question-Answering system on Mahabharata, which for a given question retrieves its answer from the whole mahabharata corpus.
- For this we have experimented on different settings of BM25 and different BERTs which focuses on bag-of-words and Self-Attention respectively.

DATASETS AND MODELS USED

Datasets

- Mahabharata Corpus.
- Manually created Question-Answer-Context set
- Extracting Question-Answers-Context from Internet

Models

- **BM25**
- BERT(for finding Relevance of the Passages)
  - **DistilRoberta-Bert**
  - **Bert-base-nli-mean-tokens**
  - **Pyserini**
- BERT (for Answering Questions) :
  - **Squad-Question-Answering Bert**

BM25

- BM25 is a bag-of-words retrieval function that ranks a set of documents based on the query terms appearing in each document.
- Given a query Q, containing keywords q_1....q_n, the BM25 score of a document D is:

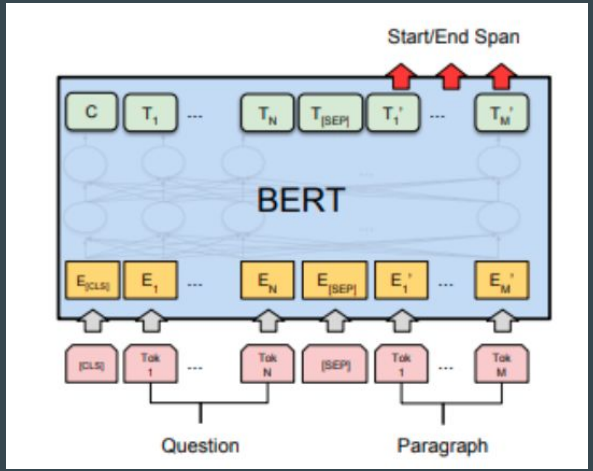
$$\text{score}(D,Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{\text{avgdl}}\right)}$$

Why to use BM25 rather than Tf-Idf:

- The simple TF-IDF rewards term frequency and penalizes document frequency whereas BM25 goes beyond this to account for document length and term frequency saturation.

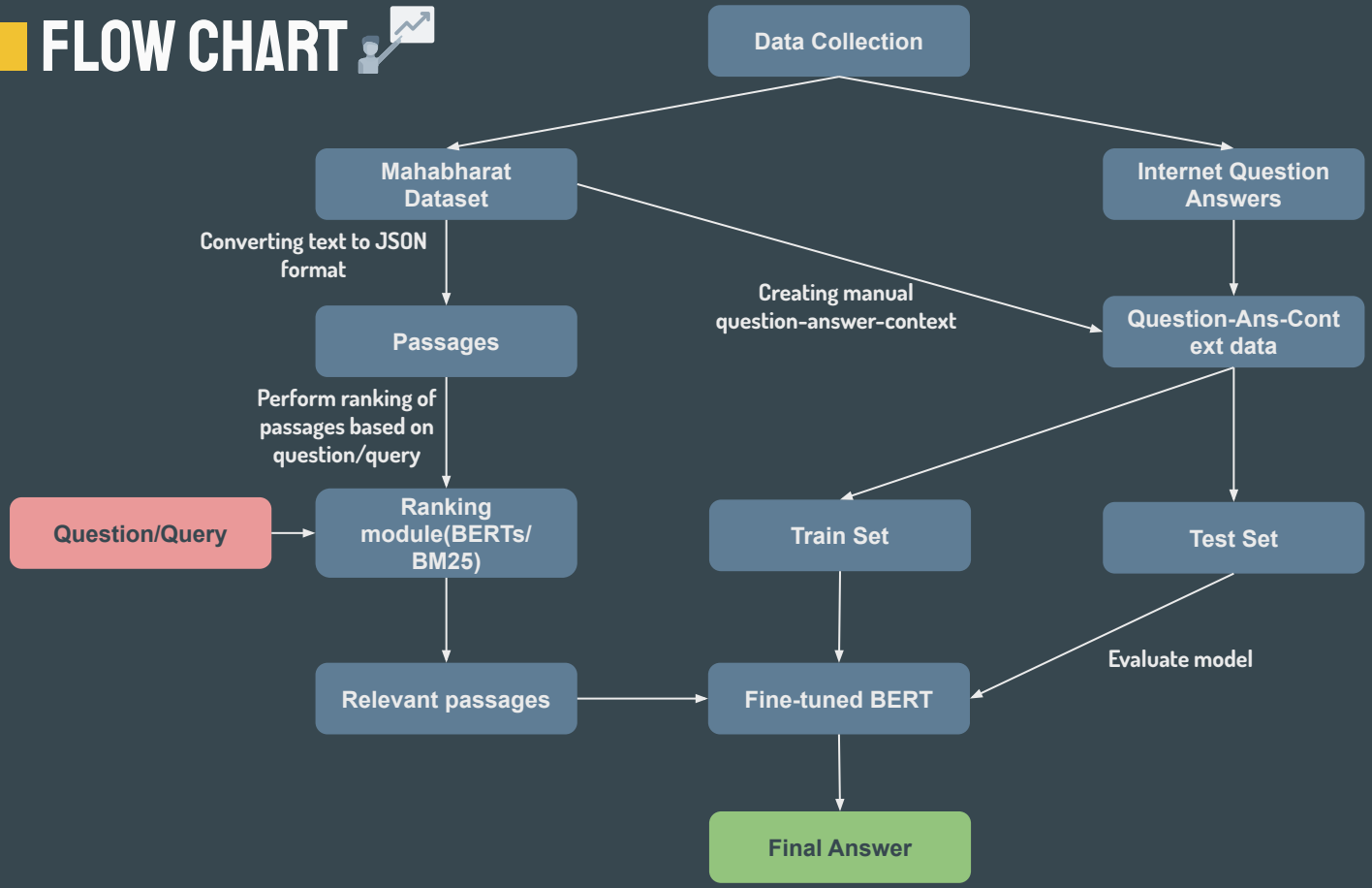
BERT

- Bidirectional Encoder Representations from Transformers (BERT) is a transformer-based machine learning technique for natural language processing (NLP) pre-training.
- BERT is at its core a transformer language model with a variable number of encoder layers and self-attention heads
- BERTs are highly tunable and can be used for numerous NLP Tasks, including :
    - Text classification
    - Text tagging
    - Question-Answering
    - Sentence-Pair Similarity
- In our project we are using BERTs for finding sentence-Pair Similarity and Question-Answering Tasks



Bert Structure for Question-Answering Task

FLOW CHART


Data Collection
Mahabharat Dataset
Internet Question Answers
Converting text to JSON format
Creating manual question-answer-context
Passages
Question-Ans-Context data
Perform ranking of passages based on question/query
Question/Query
Ranking module(BERTs/ BM25)
Train Set
Test Set
Evaluate model
Relevant passages
Fine-tuned BERT
Final Answer

QUESTION-ANSWER-CONTEXT DATASET CREATION

Question-Answer-Context Dataset is created in two phases:

- In First phase, We scraped through internet to find questions on Mahabharata along with their answers and context.
- In second, we dived into the Mahabharata corpus and created questions from it. We also extracted context and answers from the same.
- In total we were able to extract 270 questions.

We then merged these two sets of questions which are later split into training and testing sets. These sets are used to evaluate our model and also to fine tune SQuAD Question-Answering BERT

DATA PREPROCESSING

For Approach 1, 2 and 3:

- We have divided the complete corpus into passages. These passages are stored in dictionary format. The passage dictionary contains passage_id as key and its content as the value.
- Cleaning of text such as removal of non-ascii and non-alphanumeric is performed.
- The passages are the tokenized word level.

DATA PREPROCESSING

For Approach 4:

- Python Dictionary Developed for Approach 1 is then converted to json file. This json is fed as input to luceneSearcher.

```
{"id": "doc1", "contents": "contents of doc one."}
{"id": "doc2", "contents": "contents of document two."}
{"id": "doc3", "contents": "here's some text in document three."}
```

- We created another json file from corpus which is needed to be given as input to faissSearcher. The format of this json file is shown below:-

```
{
  "id": "CACM-2636",
  "contents": "Generation of Random Correlated Normal ... \n"
}
```

PYSERINI 

- Python toolkit for reproducible information retrieval research with sparse, dense and hybrid representations
- Sparse retrieval uses bag-of-words representations
- Dense retrieval uses transformer-encoded representations
- Hybrid retrieval integrates both sparse and dense approaches via linear combination of scores.

FINDING THE MOST RELEVANT PASSAGE FOR A QUESTION

- In Approach 1 :
  - Question is tokenized and then used as a Query for Finding Relevant Passages Using BM25
- In Approach 2 and 3 :
  - Every Passage is Converted to Embeddings using BERT(DistilRoberta Bert and Bert-base-nli-mean-tokens Bert for approach 2 and 3 respectively)
  - Question is converted to Embeddings using BERT and then these embeddings are used to find relevant Passages from their embeddings using Cosine Similarity.

$$similarity(A,B) = \frac{A \cdot B}{\|A\| \times \|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}}$$

FINDING THE MOST RELEVANT PASSAGE FOR A QUESTION

- In Approach 4 :
  - LuceneSearcher ranks the documents using following formula :

$$\text{score(q,d)} = \text{coord(q,d)} \cdot \text{queryNorm(q)} \cdot \sum_{t \text{ in } q}(tf(t \text{ in } d) \cdot idf(t)^2 \cdot norm(t,d))$$

Where,

- score(q,d) is the score assigned to document d for a given query q.
- queryNorm(q) is a normalizing factor used to make scores between queries comparable.
- tf(t in d) and idf(t) are term-frequency and inverse document frequency respectively
- coord(q,d) is a score factor based on how many of the query terms are found in the specified document. Typically, a document which contains more of the query terms will receive higher score.

RETRIEVING ANSWER FOR A QUESTION 

- The passage having highest score as obtained in previous step along with a question is given to BERT fine tuned on SQUAD(Stanford Question Answering Dataset). This model finds answer from the passage and gives it as output.
- If correct passage is given to this model then it is almost always accurate in fetching the correct answer.
- Situation is problematic only when an incorrect passage is given to this model by the previous stage.
- The previous stage sometimes can give incorrect passage to this model in situations such as ambiguous questions, the relevancy score of passages are very close to each other, etc…

FINE TUNING SQUAD QUESTION ANSWERING BERT

- To improve SQuAD Question-Answering Bert's Accuracy on our dataset, we further fine-tuned it.
- Since our dataset it quite small for such a complex Model, therefore we need to be careful while finetuning.
- We used AdamW as our Optimizer
- Learning Rate is set to 0.00001. Small Learning Rate would prevent noisy updates.
- We Fine tuned the model for 5 epochs.

```
Epoch 0: 100%|| 63/63 [01:04<00:00,  1.03s/it, loss=0.0196]
Epoch 1: 100%|| 63/63 [01:04<00:00,  1.03s/it, loss=0.008]
Epoch 2: 100%|| 63/63 [01:04<00:00,  1.03s/it, loss=0.0124]
Epoch 3: 100%|| 63/63 [01:04<00:00,  1.03s/it, loss=0.000733]
Epoch 4: 100%|| 63/63 [01:04<00:00,  1.03s/it, loss=0.00418]
```

RESULTS

Approach	Top-1 Accuracy	Top-5 Accuracy
Approach-1 (BM25 based)	15%	33%
Approach-2 (DistilRoberta based encoder)	31%	68%
Approach-3 (Bert-base-nli-mean-tokens based encoder)	25%	43%
Approach-4 (Pyserini-Toolkit based)	**79%**	**84%**

RESULTS

```
QUES :  For how much duration battle took place on the field of Kurukshetra?
TRUE ANS : three years
0 th ANSWER : a full three years
1 th ANSWER : seventeen s
2 th ANSWER : seventeen s
3 th ANSWER : so huge was the army of both parties
4 th ANSWER : fierce battle ensued
QUES :  Who did put the three princesses of Kashi kingdom on his chariot and left the Kashi kingdom?
TRUE ANS : Bhishma
0 th ANSWER : bhishma
1 th ANSWER : he put her on his chariot and rode toward his kingdom
2 th ANSWER : bhishma
3 th ANSWER : duhshasana
4 th ANSWER : bhishma
QUES :  Who did challenge Bhishma while he was proceeding toward Hastinapura with three princesses of Kashi kingdom?
TRUE ANS : King Salwa
0 th ANSWER : king salwa
1 th ANSWER : all opponents
2 th ANSWER : susharma
3 th ANSWER : the kings of avantipura
4 th ANSWER : susharma
QUES :  Whom did Ambika and Ambalika accept as their husband?
TRUE ANS : Vichitravirya
0 th ANSWER : draupadi
1 th ANSWER : bhishma
2 th ANSWER : a goddess
3 th ANSWER : king salwa
4 th ANSWER : vichitravirya
QUES :  What did King Salwa tell Amba when she requested him to accept her as his queen?
TRUE ANS : I no longer desire you for my queen, for you have been touched by another
0 th ANSWER : i no longer desire you for my queen
1 th ANSWER : i no longer desire you for my queen
2 th ANSWER : at heart i had chosen king salwa as my husband
3 th ANSWER : benediction
4 th ANSWER : pleaded with me to allow her to go to his kingdom
QUES :  Who was Amba's maternal grandfather?
TRUE ANS : Hotravahana
0 th ANSWER : hotravahana
1 th ANSWER : vasudeva
2 th ANSWER : vasudeva
3 th ANSWER : yadu
4 th ANSWER : janardana
```

RESULTS OF APPROACH 2 ON THE DATASET

CONCLUSION

- We successfully developed a Question-Answering system for Mahabharata.
- We can conclude that question answering system depends heavily on retrieving relevant documents.
- We Observed that Bert based approaches for retrieving the passage works a lot better than Bag-of-word based techniques like BM25 which shows that semantic meaning is very important for Question-Answering Systems, such as ours.
- Our best Model(Approach 4) obtained 79% Top-1-accuracy on the dataset.

FUTURE WORK 

We successfully created a Question-Answering system using BERT but there are still few areas which can be improved.

- We can incorporate various other Epics and create a single model to Answer Question based on different Epics.
- Currently our model reads English Mahabharata corpus and returns answers in English Language.
- Since Mahabharata was originally written in Sanskrit Language, and better translations are available in Hindi Language, work can be done to transfer Model to answer the questions in Hindi/Sanskrit Language using Hindi/Sanskrit Corpus.
- Currently Our model takes 3-4 secs to answer a single question, which is little slow of real time question answering. A faster and more efficient approach can be used to make our model faster.

REFERENCES

- Pyserini- An Easy-to-Use Python Toolkit to Support Replicable IR Research with Sparse and Dense Representations - https://arxiv.org/abs/2102.10073
- Faiss: A library for efficient similarity search - https://engineering.fb.com/2017/03/29/data-infrastructure/faiss-a-library-for-efficient-similarity-search/
- pyserini 0.16.0 - https://pypi.org/project/pyserini/
- Huggingface-Transformers - https://huggingface.co/docs/transformers/index
- Question Answer set - https://www.funtrivia.com/trivia-quiz/Religion/The-Great-Mahabharatha-366607.html
- Blog on BERTs - https://jalammar.github.io/illustrated-bert/
- BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding - https://arxiv.org/pdf/1810.04805.pdf
- CoQA: A Conversational Question Answering Challenge - https://arxiv.org/pdf/1808.07042.pdf
- Distilbert - https://huggingface.co/distilbert-base-uncased
- Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks - https://arxiv.org/abs/1908.10084

THANK YOU

ANY QUESTIONS?