DENSE RETRIEVAL FOR LOCAL/LOW-RESOURCE LANGUAGES

November 2023

COURSE CODE: INFORMATION RETRIEVAL & EXTRACTION – CS4.406.M23

Advisor:
Prof. Rahul Mishra

Mentor:
Ankita Maity

Team Number:
4

Representatives:
Aadesh Ingle - 2022202017
Akhilesh Giriboyina - 2022202022
Samyak Jain – 2022201048

Academic year:
2023-2024

1. Problem Statement:

The problem statement involves developing an effective and efficient dense retrieval system for local or low-resource languages. Dense retrieval focuses on retrieving relevant passages or documents from a collection based on dense vector representations, typically obtained from pre-trained language models.

In the context of local or low-resource languages, there is a difficulty because of low data or no data available.
The challenge is to adapt existing retrieval techniques and models to manage limited data and linguistic resources while maintaining high retrieval performance.

Understanding of the Problem Statement: The problem entails addressing several key aspects:

- Creating high-quality dense embeddings for passages/documents.
- Designing an efficient indexing and retrieval mechanism for the stored embeddings.
- Adapting or fine-tuning pre-trained models to handle data scarcity of the target language.
- Evaluating the system's effectiveness in retrieving relevant information in the low-resource language.

3. Data Overview:

Here are the datasets which we have used:

- Wikipedia Dumps - DPR

A complete copy of all Wikimedia wikis, in the form of wikitext source and metadata embedded in XML. A number of raw database tables in SQL form are also available. These snapshots are provided at the very least monthly and usually twice a month.

- SQUAD - mT5

Stanford Question Answering Dataset (SQuAD) is a reading comprehension dataset, consisting of questions posed by crowdworkers on a set of Wikipedia articles, where the answer to every question is a segment of text, or span, from the corresponding reading passage, or the question might be unanswerable.

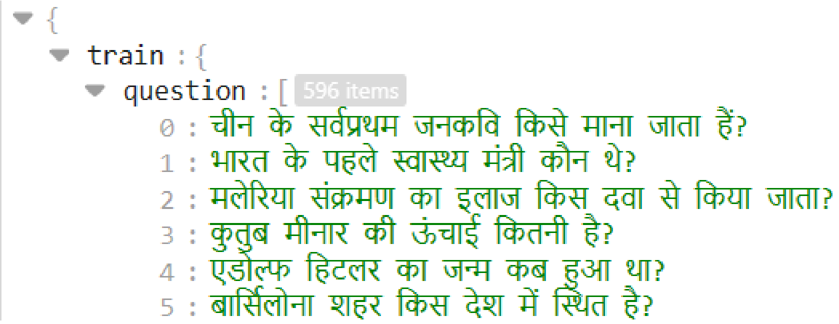
- TyDiQA - mT5

TyDi QA is a question answering dataset covering 11 typologically diverse languages with 204K question-answer pairs. The languages of TyDi QA are diverse with regard to their typology -- the set of linguistic features that each language expresses -- such that we expect models performing well on this set to generalize across a large number of the languages in the world. It contains language phenomena that would not be found in English-only corpora. To provide a realistic information-seeking task and avoid priming effects, questions are written by people who want to know the answer, but don't know the answer yet, (unlike SQuAD and its descendents) and the data is collected directly in each language without the use of translation (unlike MLQA and XQuAD).

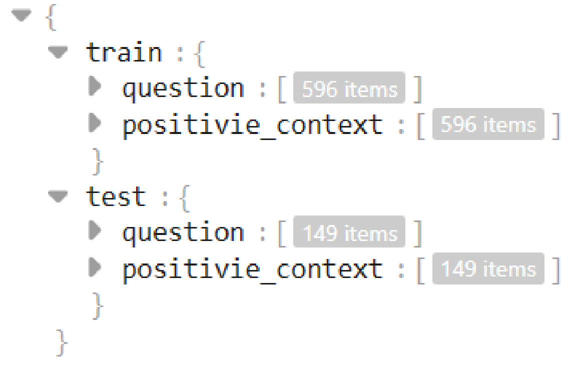
- mC4 - mT5

A multilingual colossal, cleaned version of Common Crawl's web crawl corpus. Based on Common Crawl dataset: "https://commoncrawl.org".

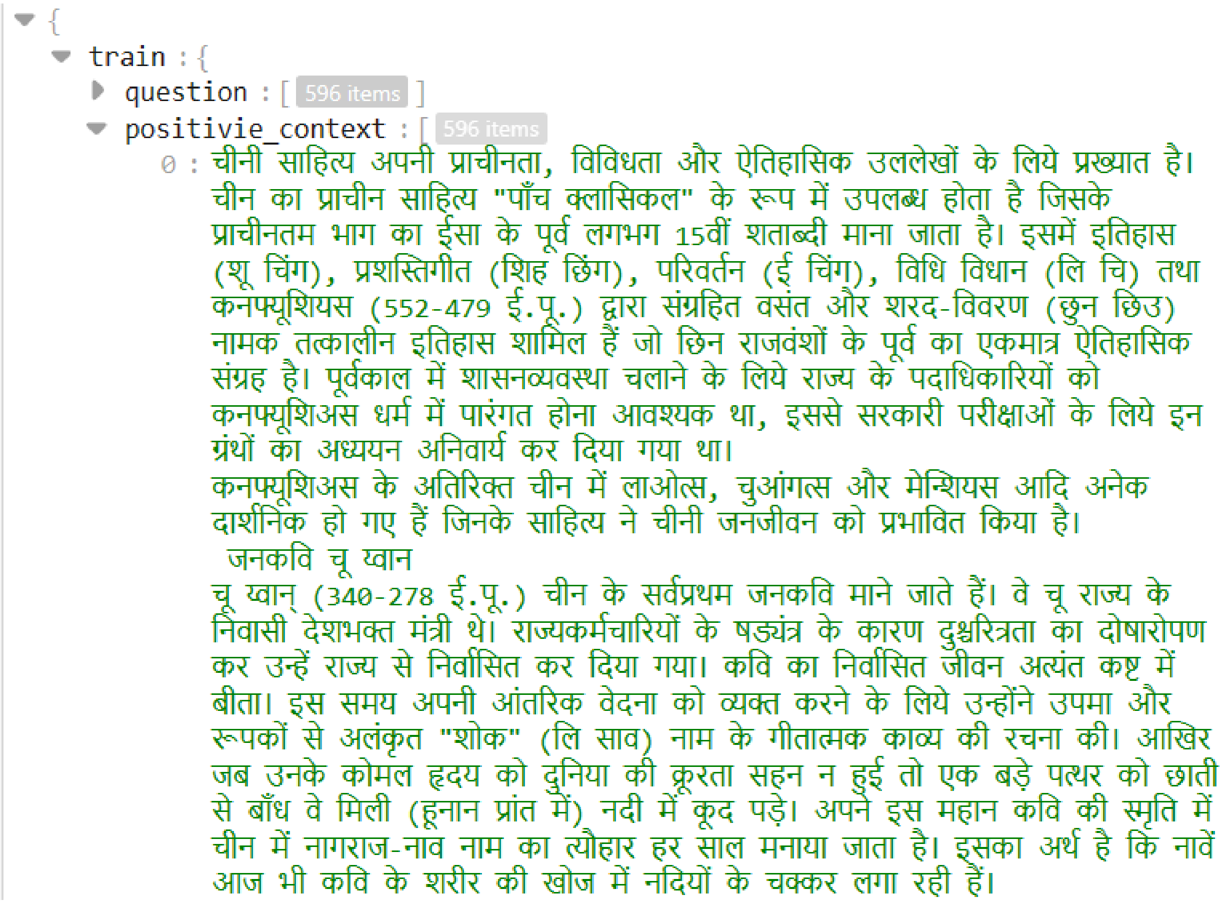
For the interim submission, we have decided to go with the **ChAII hindi dataset**. We have taken the questions, & positive contexts from the dataset. We have customized this dataset to our own needs. Following is the customized format for our dataset, which is different from the original dataset.



```
▼ {
  ▼ train : {
    ▼ question : [ 596 items
        0 : चीन के सर्वप्रथम जनकवि किसे माना जाता हैं?
        1 : भारत के पहले स्वास्थ्य मंत्री कौन थे?
        2 : मलेरिया संक्रमण का इलाज किस दवा से किया जाता?
        3 : कुतुब मीनार की ऊंचाई कितनी है?
        4 : एडोल्फ हिटलर का जन्म कब हुआ था?
        5 : बार्सिलोना शहर किस देश में स्थित है?
```



```
▼ {
  ▼ train : {
    ▶ question : [ 596 items ]
    ▶ positivie_context : [ 596 items ]
    }
  ▼ test : {
    ▶ question : [ 149 items ]
    ▶ positivie_context : [ 149 items ]
    }
  }
```



```
▼ {
  ▼ train : {
    ▶ question : [ 596 items ]
    ▼ positivie_context : [ 596 items
        0 : चीनी साहित्य अपनी प्राचीनता, विविधता और ऐतिहासिक उललेखों के लिये प्रख्यात है।
            चीन का प्राचीन साहित्य "पाँच क्लासिकल" के रूप में उपलब्ध होता है जिसके
            प्राचीनतम भाग का ईसा के पूर्व लगभग 15वीं शताब्दी माना जाता है। इसमें इतिहास
            (शू चिंग), प्रशस्तिगीत (शिह छिंग), परिवर्तन (ई चिंग), विधि विधान (लि चि) तथा
            कनफ्यूशियस (552-479 ई.पू.) द्वारा संग्रहित वसंत और शरद-विवरण (छुन छिउ)
            नामक तत्कालीन इतिहास शामिल हैं जो छिन राजवंशों के पूर्व का एकमात्र ऐतिहासिक
            संग्रह है। पूर्वकाल में शासनव्यवस्था चलाने के लिये राज्य के पदाधिकारियों को
            कनफ्यूशिअस धर्म में पारंगत होना आवश्यक था, इससे सरकारी परीक्षाओं के लिये इन
            ग्रंथों का अध्ययन अनिवार्य कर दिया गया था।
            कनफ्यूशिअस के अतिरिक्त चीन में लाओत्स, चुआंगत्स और मेन्शियस आदि अनेक
            दार्शनिक हो गए हैं जिनके साहित्य ने चीनी जनजीवन को प्रभावित किया है।
             जनकवि चू य्वान
            चू य्वान् (340-278 ई.पू.) चीन के सर्वप्रथम जनकवि माने जाते हैं। वे चू राज्य के
            निवासी देशभक्त मंत्री थे। राज्यकर्मचारियों के षड्यंत्र के कारण दुश्चरित्रता का दोषारोपण
            कर उन्हें राज्य से निर्वासित कर दिया गया। कवि का निर्वासित जीवन अत्यंत कष्ट में
            बीता। इस समय अपनी आंतरिक वेदना को व्यक्त करने के लिये उन्होंने उपमा और
            रूपकों से अलंकृत "शोक" (लि साव) नाम के गीतात्मक काव्य की रचना की। आखिर
            जब उनके कोमल हृदय को दुनिया की क्रूरता सहन न हुई तो एक बड़े पत्थर को छाती
            से बाँध वे मिली (हूनान प्रांत में) नदी में कूद पड़े। अपने इस महान कवि की स्मृति में
            चीन में नागराज-नाव नाम का त्यौहार हर साल मनाया जाता है। इसका अर्थ है कि नावें
            आज भी कवि के शरीर की खोज में नदियों के चक्कर लगा रही हैं।
```

Similar format is followed for test set.

2. Baseline Method:

The problem is an open research topic and does not have many resources or implementations to tackle this problem.

So, as discussed with Professor, we have decided to keep the baseline as TFIDF.

We have used the same dataset that we will be using for our Advanced method.

$$TF(t,d) = \frac{number\ of\ times\ t\ appears\ in\ d}{total\ number\ of\ terms\ in\ d}$$

$$IDF(t) = log\frac{N}{1+df}$$

$$TF-IDF(t,d) = TF(t,d) * IDF(t)$$

For working on the this, we have chosen to approach this problem by creating a pipeline of 2 models: “Question Generation Model”, “Dense Passage Retrieval Model” which will be explained in detail in further sections.

3. Advanced Method:

We try to address the problem by using two different models - “Question Generation Model” and “Dense Passage Retrieval Model”.

- **Question Generation Model**
  The major challenge of the problem is to handle the unavailability of data for low resource language. For tackling this, we have chosen to create a model to generate low-resource language question.

  Assuming data augmentation of a low resource language could be the case where there is no data available. Hence, we tried to create synthetic data of target language on a Zero-Shot setting.

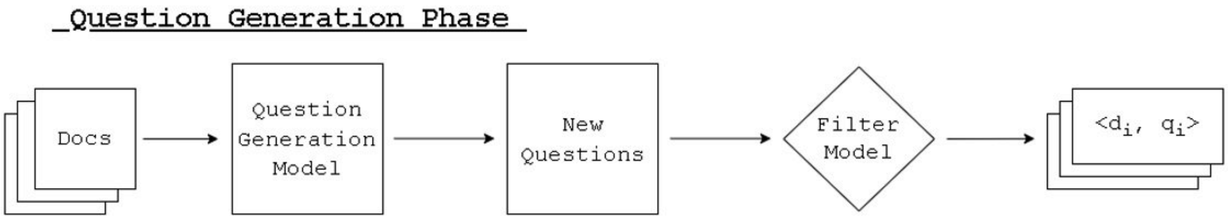
  We have taken a reference of Google’s mT5 model

Model Explanation
Base model used: mT5
Training Language: English
Dataset for training: SQUAD, mC4, TyDiQA
Target Language: Telugu

Aim: Create target question for the given target language Passage.

Approach:-
We are going in a zero-shot approach wherein we assume there is no Question data available for target language. So, we are fine-tuning our model with a highly available English language dataset SQUAD.

We have used the base model as mT5 and trained it on English data. Our assumption is that with model trained to generate questions on English, we can use it to generate synthetic data for our target Language.



Training

We have fine-tuned mT5 for the following tasks:-

- **SQUAD** English Data: To make the model learn to generate questions for a given passage, we have fine-tuned it on English SQUAD data.
  For this task we have created a **custom prefix token 'questiongeneration:'** which helps the model understand the prefix while fine-tuning.

- **Mask Language Modelling**: To avoid **catastrophic forgetting**, we are also continuing the MLM training of model for **mC4** and **TyDIQA** data.

We have chosen a ratio of 8:1 of SQUAD – MLM task while creating our dataset.

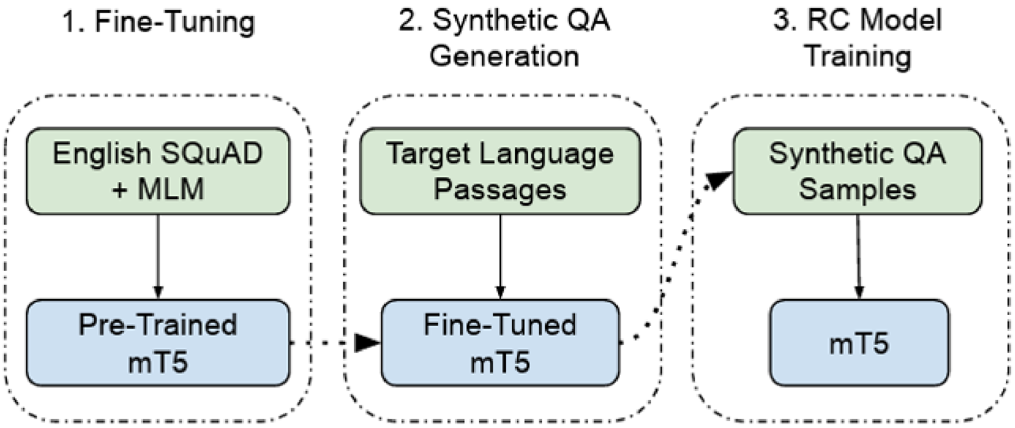
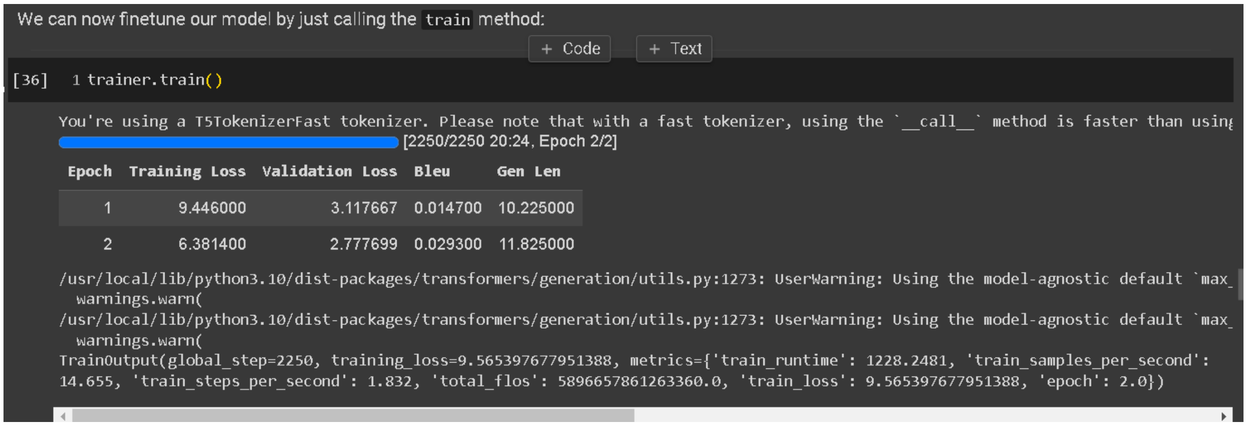


Figure 1: End-to-End pipeline: 1) Fine-tuning the generative model using SQuAD English samples and multilingual MLM. 2) Generating synthetic samples from Wikipedia passages of the target language using the fine-tuned generative model. 3) Training the downstream reading comprehension model using synthetic samples.

Snapshot of Training question generation model:



We can now finetune our model by just calling the `train` method:

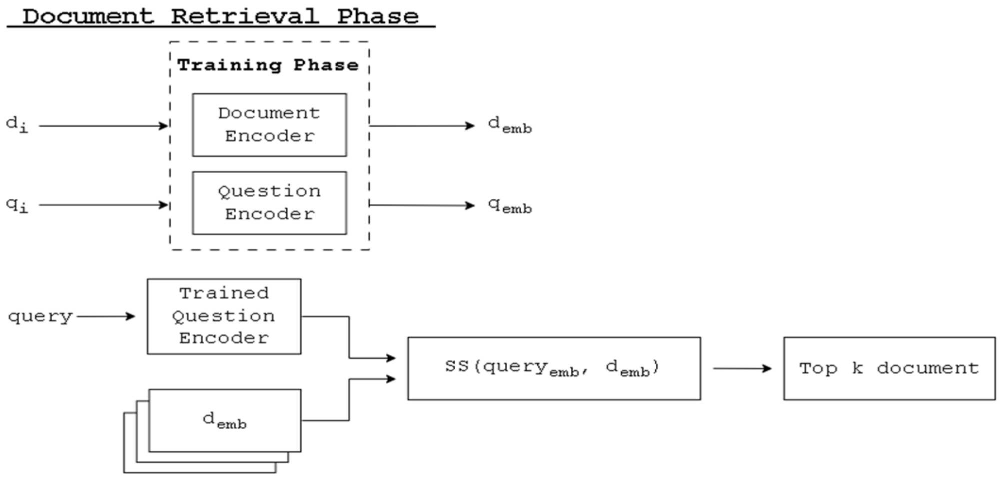
```
[36]   trainer.train()

You're using a T5TokenizerFast tokenizer. Please note that with a fast tokenizer, using the `__call__` method is faster than using
[2250/2250 20:24, Epoch 2/2]
 Epoch  Training Loss  Validation Loss  Bleu      Gen Len
     1       9.446000         3.117667  0.014700  10.225000
     2       6.381400         2.777699  0.029300  11.825000
/usr/local/lib/python3.10/dist-packages/transformers/generation/utils.py:1273: UserWarning: Using the model-agnostic default `max_
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/transformers/generation/utils.py:1273: UserWarning: Using the model-agnostic default `max_
  warnings.warn(
TrainOutput(global_step=2250, training_loss=9.565397677951388, metrics={'train_runtime': 1228.2481, 'train_samples_per_second':
14.655, 'train_steps_per_second': 1.832, 'total_flos': 5896657861263360.0, 'train_loss': 9.565397677951388, 'epoch': 2.0})
```

Inferencing

The trained mT5 model is fed with wiki passages data of our target language and generated synthetic data.
We could notice that the generated questions were not extremely well, this is because of the resource constraint we had to use “mt5-small”.
However, we could notice that our model is trying to generate the questions in target language even though it is not trained upon it.

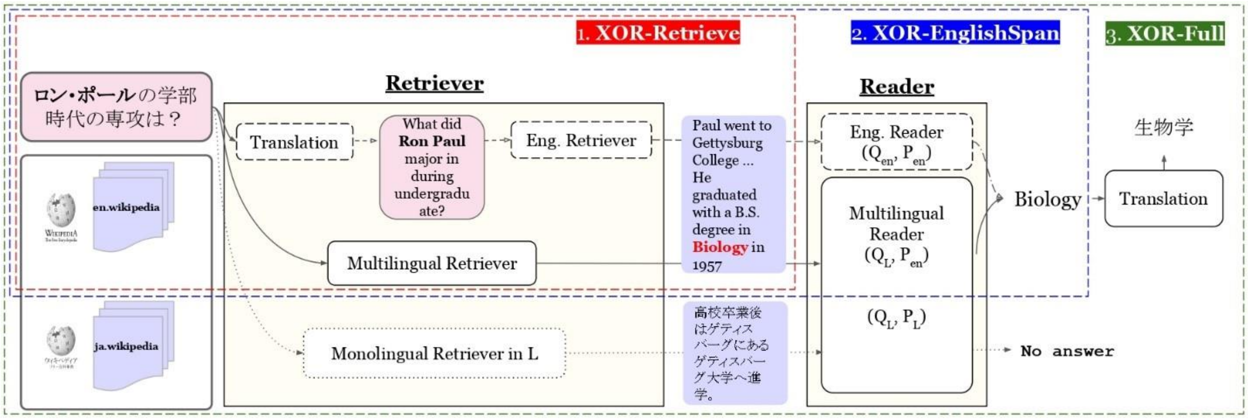
Note that in this approach we are not using target language’s question data i.e. **Zero Shot Approach**. Hence this approach can be extended to any of the target language. However, the restriction would be that the base model should have a representation of target language (mT5 in our case) meaning the data used in the model should be consisting of target language.

- **Our DPR approach:**



We have reviewed various papers and online resources for our project topic. We have tweaked the DPR paper along with some modifications as we do not have access to the Institute’s ADA cluster.

Low-Resource Dense Retrieval for Open-Domain Question Answering


1. XOR-Retrieve
2. XOR-EnglishSpan
3. XOR-Full
Retriever
Reader
ロン・ポールの学部時代の専攻は？
en.wikipedia
ja.wikipedia
Translation
What did Ron Paul major in during undergraduate?
Eng. Retriever
Multilingual Retriever
Monolingual Retriever in L
Paul went to Gettysburg College ... He graduated with a B.S. degree in Biology in 1957
高校卒業後はゲティスバーグにあるゲティスバーグ大学へ進学。
Eng. Reader (Q_{en}, P_{en})
Multilingual Reader (Q_L, P_{en})
(Q_L, P_L)
Biology
Translation
生物学
No answer

- It provides a comprehensive overview of the state-of-the-art methods for low-resource dense retrieval (DR) for both the document collection and a set of question-answer pairs to learn a dense retriever. This setting is the least challenging, as the retriever can learn from the question-answer pairs to directly predict the answer to the question.

Major Motivation from the above paper: We have decided to go for the documents-only dataset that makes the question generation model task the most important for the Retrieval model.

Our proposed implementation:

Given a collection of M text passages, the goal of our dense passage retriever (DPR) is to index all the passages in a low-dimensional and continuous space such that it can efficiently retrieve the top k passages relevant to the input question for the reader at run-time.

- Our dense passage retriever (DPR) uses a **dense encoder EP (·),** which maps any text passage to a d-dimensional real-valued vector and builds an index for all the M passages that we will use for retrieval.
- At run-time, DPR applies a different **encoder EQ (·)** that maps the input question to a d-dimensional vector and retrieves k passages of which vectors are the closest to the question vector.
- $$\operatorname{sim}(q,p) = E_Q(q)^{\intercal} E_P(p)$$ We define the similarity between the question and the passage using the dot product of their vectors:
- **ENCODER:** Although, in principle, the question and passage encoders can be implemented by any neural network, in this work, we use two independent BERT networks.
- **INFERENCING:** Given a question q at run-time, we derive its embedding $v_q = E_Q(q)$ and retrieve the top k passages with embeddings closest to v_q.
- **TRAINING:**

- The goal is to create a vector space such that relevant pairs of questions and passages will have a smaller distance (i.e., higher similarity) than the irrelevant ones by learning a better embedding function.

- **CUSTOM LOSS FUNCTION:**

$$L(q_i, p_i^+, p_{i,1}^-, \cdots, p_{i,n}^-) = -\log \frac{e^{\mathrm{sim}(q_i, p_i^+)}}{e^{\mathrm{sim}(q_i, p_i^+)} + \sum_{j=1}^{n} e^{\mathrm{sim}(q_i, p_{i,j}^-)}}$$

  - Let $\mathcal{D} = \{\langle q_i, p_i^+, p_{i,1}^-, \cdots, p_{i,n}^- \rangle\}_{i=1}^{m}$ be the training data that consists of m instances. Each instance contains one question q_i and one relevant (positive) passage p_i^+ along with n irrelevant (negative) passages $p_{i,j}^-$. We optimize the loss function as the negative log-likelihood of the positive passage:

- **POSITIVE AND NEGATIVE PASSAGES:**
  - For retrieval problems, it is often the case that positive examples are available explicitly, while negative examples need to be selected from an extremely large pool.
  - For instance, passages relevant to a question may be given in a QA dataset or can be found using the answer. All other passages in the collection, while not specified explicitly, can be viewed as irrelevant by default.
  - We have decided to do in **BATCH NEGATIVES**.

- **IN BATCH NEGATIVES:**
  - Assume that we have B questions in a mini-batch, and each one is associated with a relevant passage. Let **Q** and **P** be the (B×d) matrix of question and passage embeddings in a batch of size B.
  - **$S = QP^T$** is a (B ×B) matrix of similarity scores, where each row of which corresponds to a question paired with B passages.
  - In this way, we **reuse computation** and effectively train on B^2 (q_i, p_j) question/passage pairs in each batch.
  - Any (q_i, p_j) pair is a positive example when i == j and negative otherwise.

- This creates B training instances in each batch, where there are **B – 1 negative passages** for each question.

Code Structure of DPR:

a. **DataLoader**

We are sending question and it's corresponding context pairs in each batch. We have used in batch negative passages as mentioned earlier to speed up the process and save compute resources.

b. **Encoder**

We will be using class for generating the [CLS] embeddings of question/ query and paragraph, which we will be using further.

c. **Bidirectional Encoder:**

We will use this class as the model class that we will train. Here, we are using the above Encoder to get the [CLS] embeddings of both question/ query and paragraph.

d. **Train Bidirectional Encoder:**

We are running the train loop here and finetuning the BERT model to make embeddings of question and positive paragraphs as close as possible and vice versa for question and negative paragraphs.

Custom Loss Function: We are using NLL Loss(Negative Log-Likelihood), which is generally used in tasks like classification.

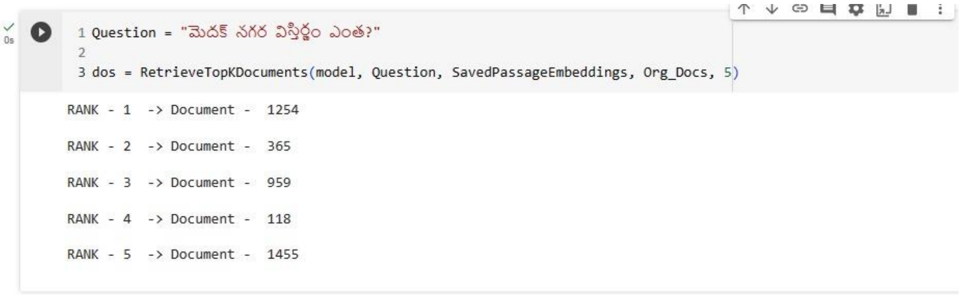
e. **LoadOrTrainModel:**

Finally, in this loop, we are running training the above class for a certain no of epochs, saving the state of the model if the loss is minimum at that epoch, and finally plotting the change in loss over the epoch.

4. **Results:**

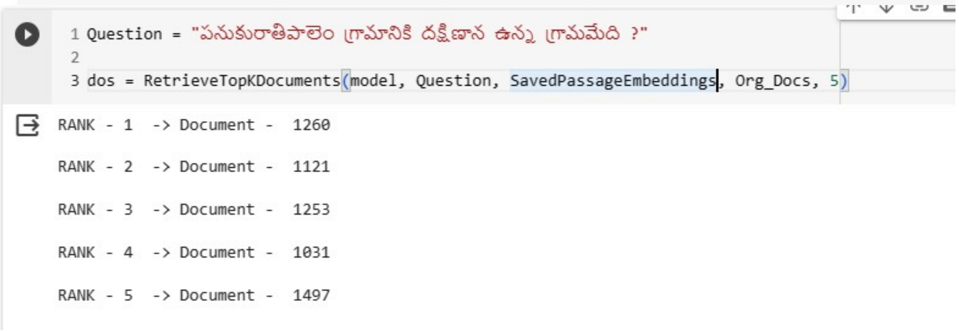
Metric	TF-IDF	Our Model
Accuracy	16.66%	81%
Mean Reciprocal Rank	16.66%	57.83%

OUR MODEL



```
Question = "మెదక్ నగర విస్తీర్ణం ఎంత?"

dos = RetrieveTopKDocuments(model, Question, SavedPassageEmbeddings, Org_Docs, 5)
```

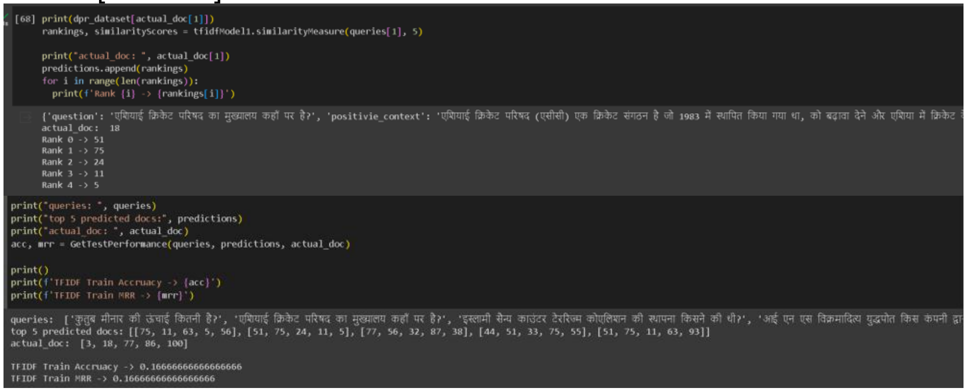
```
RANK - 1  -> Document -  1254

RANK - 2  -> Document -  365

RANK - 3  -> Document -  959

RANK - 4  -> Document -  118

RANK - 5  -> Document -  1455
```



```
Question = "పెనుకురాతిపాలెం గ్రామానికి దక్షిణాన ఉన్న గ్రామమేది ?"

dos = RetrieveTopKDocuments(model, Question, SavedPassageEmbeddings, Org_Docs, 5)
```

```
RANK - 1  -> Document -  1260

RANK - 2  -> Document -  1121

RANK - 3  -> Document -  1253

RANK - 4  -> Document -  1031

RANK - 5  -> Document -  1497
```

TF-IDF[baseline]:



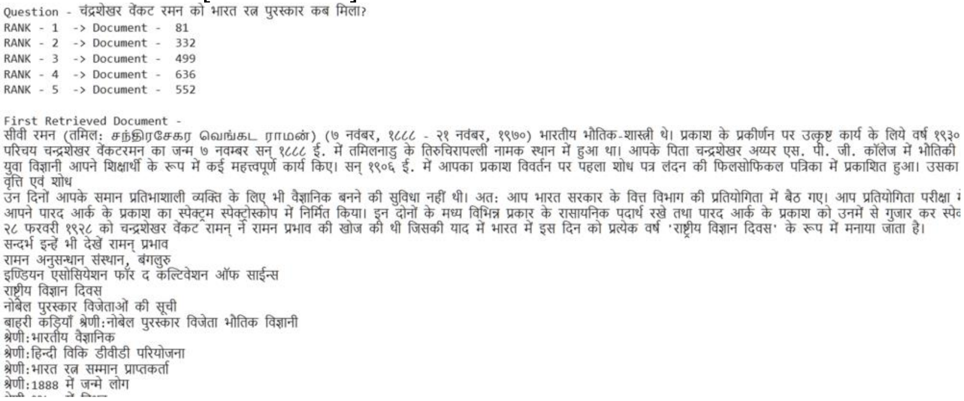
```
[68] print(dpr_dataset[actual_doc[1]])
     rankings, similarityScores = tfidfModel1.similarityMeasure(queries[1], 5)

     print("actual_doc: ", actual_doc[1])
     predictions.append(rankings)
     for i in range(len(rankings)):
       print(f'Rank {i} -> {rankings[i]}')
```

```
{'question': 'एशियाई क्रिकेट परिषद का मुख्यालय कहाँ पर है?', 'positivie_context': 'एशियाई क्रिकेट परिषद (एसीसी) एक क्रिकेट संगठन है जो 1983 में स्थापित किया गया था, को बढ़ावा देने और एशिया में क्रिकेट
actual_doc:  18
Rank 0 -> 51
Rank 1 -> 75
Rank 2 -> 24
Rank 3 -> 11
Rank 4 -> 5
```

```
print("queries: ", queries)
print("top 5 predicted docs:", predictions)
print("actual_doc: ", actual_doc)
acc, mrr = GetTestPerformance(queries, predictions, actual_doc)

print()
print(f'TFIDF Train Accruacy -> {acc}')
print(f'TFIDF Train MRR -> {mrr}')
```

```
queries:  ['कुतुब मीनार की ऊंचाई कितनी है?', 'एशियाई क्रिकेट परिषद का मुख्यालय कहाँ पर है?', 'इस्लामी सैन्य काउंटर टेररिज्म कोएलिशन की स्थापना किसने की थी?', 'आई एन एस विक्रमादित्य युद्धपोत किस कंपनी द्वा
top 5 predicted docs: [[75, 11, 63, 5, 56], [51, 75, 24, 11, 5], [77, 56, 32, 87, 38], [44, 51, 33, 75, 55], [51, 75, 11, 63, 93]]
actual_doc:  [3, 18, 77, 86, 100]

TFIDF Train Accruacy -> 0.16666666666666666
TFIDF Train MRR -> 0.16666666666666666
```

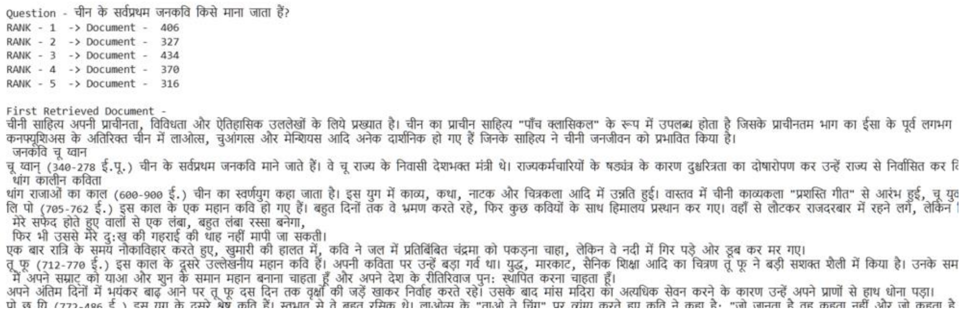
INFERENCING[TEST SET]:



```
Question - चंद्रशेखर वेंकट रमन को भारत रत्न पुरस्कार कब मिला?
RANK - 1  -> Document -  81
RANK - 2  -> Document -  332
RANK - 3  -> Document -  499
RANK - 4  -> Document -  636
RANK - 5  -> Document -  552

First Retrieved Document -
```

सीवी रमन (तमिल: சந்திரசேகர வெங்கட ராமன்) (७ नवंबर, १८८८ - २१ नवंबर, १९७०) भारतीय भौतिक-शास्त्री थे। प्रकाश के प्रकीर्णन पर उत्कृष्ट कार्य के लिये वर्ष १९३०
परिचय चन्द्रशेखर वेंकटरमन का जन्म ७ नवम्बर सन् १८८८ ई. में तमिलनाडु के तिरुचिरापल्ली नामक स्थान में हुआ था। आपके पिता चन्द्रशेखर अय्यर एस. पी. जी. कॉलेज में भौतिकी
युवा विज्ञानी आपने शिक्षार्थी के रूप में कई महत्त्वपूर्ण कार्य किए। सन् १९०६ ई. में आपका प्रकाश विवर्तन पर पहला शोध पत्र लंदन की फिलसोफिकल पत्रिका में प्रकाशित हुआ। उसका
वृत्ति एवं शोध
उन दिनों आपके समान प्रतिभाशाली व्यक्ति के लिए भी वैज्ञानिक बनने की सुविधा नहीं थी। अत: आप भारत सरकार के वित्त विभाग की प्रतियोगिता में बैठ गए। आप प्रतियोगिता परीक्षा
आपने पारद आर्क के प्रकाश का स्पेक्ट्रम स्पेक्ट्रोस्कोप में निर्मित किया। इन दोनों के मध्य विभिन्न प्रकार के रासायनिक पदार्थ रखे तथा पारद आर्क के प्रकाश को उनमें से गुजार कर स्पे
२८ फरवरी १९२८ को चन्द्रशेखर वेंकट रामन् ने रामन प्रभाव की खोज की थी जिसकी याद में भारत में इस दिन को प्रत्येक वर्ष 'राष्ट्रीय विज्ञान दिवस' के रूप में मनाया जाता है।
सन्दर्भ इन्हें भी देखें रामन् प्रभाव
रामन अनुसन्धान संस्थान, बंगलुरु
इण्डियन एसोसियेशन फॉर द कल्टिवेशन ऑफ साईन्स
राष्ट्रीय विज्ञान दिवस
नोबेल पुरस्कार विजेताओं की सूची
बाहरी कड़ियाँ श्रेणी:नोबेल पुरस्कार विजेता भौतिक विज्ञानी
श्रेणी:भारतीय वैज्ञानिक
श्रेणी:हिन्दी विकि डीवीडी परियोजना
श्रेणी:भारत रत्न सम्मान प्राप्तकर्ता
श्रेणी:1888 में जन्मे लोग

INFERENCING[TRAIN SET]



```
Question - चीन के सर्वप्रथम जनकवि किसे माना जाता हैं?
RANK - 1  -> Document -  406
RANK - 2  -> Document -  327
RANK - 3  -> Document -  434
RANK - 4  -> Document -  370
RANK - 5  -> Document -  316

First Retrieved Document -
चीनी साहित्य अपनी प्राचीनता, विविधता और ऐतिहासिक उललेखों के लिये प्रख्यात है। चीन का प्राचीन साहित्य "पाँच क्लासिकल" के रूप में उपलब्ध होता है जिसके प्राचीनतम भाग का ईसा के पूर्व लगभग
कनफ्यूशिअस के अतिरिक्त चीन में लाओत्स, चुआंगत्स और मेन्शियस आदि अनेक दार्शनिक हो गए हैं जिनके साहित्य ने चीनी जनजीवन को प्रभावित किया है।
 जनकवि चू य्वान
चू य्वान् (340-278 ई.पू.) चीन के सर्वप्रथम जनकवि माने जाते हैं। वे चू राज्य के निवासी देशभक्त मंत्री थे। राज्यकर्मचारियों के षड्यंत्र के कारण दुश्चरित्रता का दोषारोपण कर उन्हें राज्य से निर्वासित कर वि
 थांग कालीन कविता
थांग राजाओं का काल (600-900 ई.) चीन का स्वर्णयुग कहा जाता है। इस युग में काव्य, कथा, नाटक और चित्रकला आदि में उन्नति हुई। वास्तव में चीनी काव्यकला "प्रशस्ति गीत" से आरंभ हुई, चू यु
लि पो (705-762 ई.) इस काल के एक महान कवि हो गए हैं। बहुत दिनों तक वे भ्रमण करते रहे, फिर कुछ कवियों के साथ हिमालय प्रस्थान कर गए। वहाँ से लौटकर राजदरबार में रहने लगे, लेकिन
 मेरे सफेद होते हुए वालों से एक लंबा, बहुत लंबा रस्सा बनेगा,
 फिर भी उससे मेरे दु:ख की गहराई की थाह नहीं मापी जा सकती।
एक बार रात्रि के समय नौकाविहार करते हुए, खुमारी की हालत में, कवि ने जल में प्रतिबिंबित चंद्रमा को पकड़ना चाहा, लेकिन वे नदी में गिर पड़े ओर डूब कर मर गए।
तू फू (712-770 ई.) इस काल के दूसरे उल्लेखनीय महान कवि हैं। अपनी कविता पर उन्हें बड़ा गर्व था। युद्ध, मारकाट, सैनिक शिक्षा आदि का चित्रण तू फू ने बड़ी सशक्त शैली में किया है। उनके सम
 में अपने सम्राट् को याआ और शुन के समान महान बनाना चाहता हूँ और अपने देश के रीतिरिवाज पुनः स्थापित करना चाहता हूँ।
अपने अंतिम दिनों में भयंकर बाढ़ आने पर तू फू दस दिन तक वृक्षों की जड़ें खाकर निर्वाह करते रहे। उसके बाद मांस मदिरा का अत्यधिक सेवन करने के कारण उन्हें अपने प्राणों से हाथ धोना पड़ा।
पो छ यि (772-486 ई.) इस युग के दूसरे श्रेष्ठ कवि हैं। स्वभाव से वे बहुत रसिक थे। लाओत्स के "ताओ ते चिंग" पर व्यंग्य करते हुए कवि ने कहा है: "जो जानता है वह कहता नहीं और जो कहता है
```

5. **Discussion and Conclusions**:

Discussions/ Challenges:

- We are using Zero shot approach because we want to build the model which can be used on the mulitple languages (that are used to train mT5). We could have also used few shots approach as suggested by Prof, but we found the Zero shot approach more interesting/ exciting.
- We had to train the "mt5-small" of the mT5 model because of resource constraints.
- Explanation of dataset to be used for finetuned mT5 model was not mentioned clearly in the paper. We had to many iterations of experimentation to arrive at our final approach, which is 8 : 1 ratio for question generation and MLM task. Further we had decided to have 1 : 1 mixture of mC4 and TyDiQA dataset for MLM task.
- We also decided to go with DPR approach because it is possible to add on the passages in the runtime after the model has been freezed, we just had to create the embeddings and to our data.
- Also we have used FAISS(Facebook AI Semantic Search) for indexing the document encoding which helps in faster retrieval of documents. We have used both cosine similarity & the nearest neighbour search provide by FAISS for scoring the similarity of top 'k' documents.

Google Drive Link: https://drive.google.com/drive/folders/11PJv7jlOru8E7exs4Pl_yG1i_JIL7yIk?usp=share_link

References:

- Shakeri, S., Constant, N., Kale, M. S., & Xue, L. (2020). Towards Zero-Shot Multilingual Synthetic Question and Answer Generation for Cross-Lingual Reading Comprehension. ArXiv. /abs/2010.12008
- Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W. (2020). Dense Passage Retrieval for Open-Domain Question Answering. *ArXiv*. /abs/2004.04906
- Hedderich, M. A., Lange, L., Adel, H., Strötgen, J., & Klakow, D. (2020). A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios. *ArXiv*. /abs/2010.12309
- Hambarde, K. A., & Proenca, H. (2023). Information Retrieval: Recent Advances and Beyond. *ArXiv*. /abs/2301.08801
- Kulshreshtha, D., Belfer, R., Serban, I. V., & Reddy, S. (2021). Back-Training excels Self-Training at Unsupervised Domain Adaptation of Question Generation and Passage Retrieval. *ArXiv*. /abs/2104.08801
- Shen, X., Vakulenko, S., Del Tredici, M., Barlacchi, G., Byrne, B., & De Gispert, A. (2022). Low-Resource Dense Retrieval for Open-Domain Question Answering: A Comprehensive Survey. *ArXiv*. /abs/2208.03197
- Wang, K., Thakur, N., Reimers, N., & Gurevych, I. (2021). GPL: Generative Pseudo Labeling for Unsupervised Domain Adaptation of Dense Retrieval. *ArXiv*. /abs/2112.07577
- Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., & Gurevych, I. (2021). BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models. *ArXiv*. /abs/2104.08663
- Khattab, O., & Zaharia, M. (2020). ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. *ArXiv*. /abs/2004.12832
- Lu, W., Jiao, J., & Zhang, R. (2020). TwinBERT: Distilling Knowledge to Twin-Structured BERT Models for Efficient Retrieval. *ArXiv*. /abs/2002.06275

Presentation Link: 4-FinalPresentation.pptx

Github Link: **Dense-Retrieval-for-Low-Resource-Language**