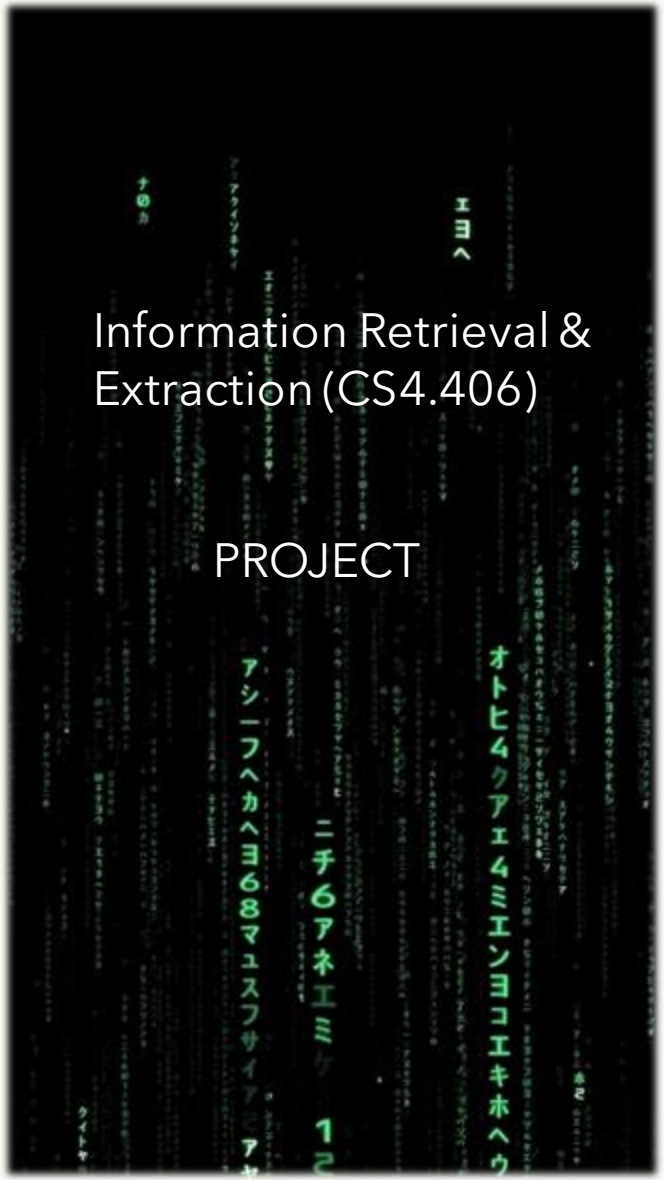
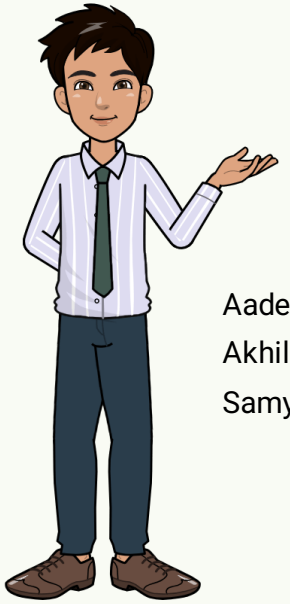
Dense Retrieval for low/local resource languages

Github Link



Information Retrieval & Extraction (CS4.406)

PROJECT



Advisor: Prof. Rahul Mishra

Mentor: Ankita Maity

TEAM 4

Aadesh Ingle	- 2022202017
Akhilesh Giriboyina	- 2022202022
Samyak Jain	- 2022201048

AGENDA

PROBLEM DESCRIPTION

DATASET

BASELINE IMPLEMENTATION

IMPLEMENTATION

EVALUATION & INFERENCING

PROBLEM DESCRIPTION

Creating high-quality dense embeddings for passages/documents in the **low-resource** language.

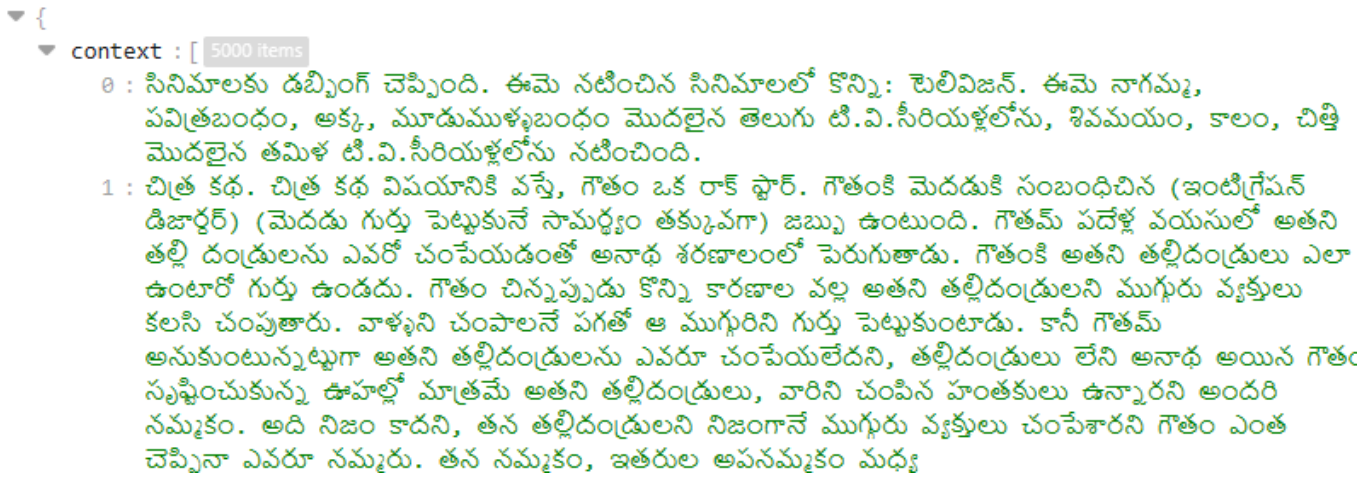
Designing an efficient retrieval mechanism for these embeddings.

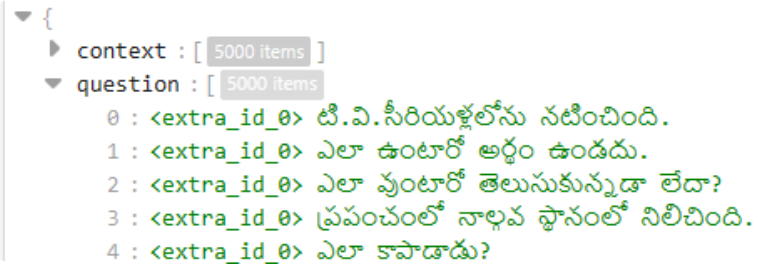
Adapting or fine-tuning pre-trained models to handle data scarcity of the target language.

Evaluating the system's effectiveness in retrieving relevant information in the low-resource language.

DATASETS

- SQuAD
- TyDiQA
- mC4
- Wikipedia Dumps
- ChAII





Discussion of Our Approach

- There were two approaches to choose from:
  - Zero Shot Approach
  - Few Shot Approach

- Architecture for the model

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

BASELINE IMPLEMENTATION

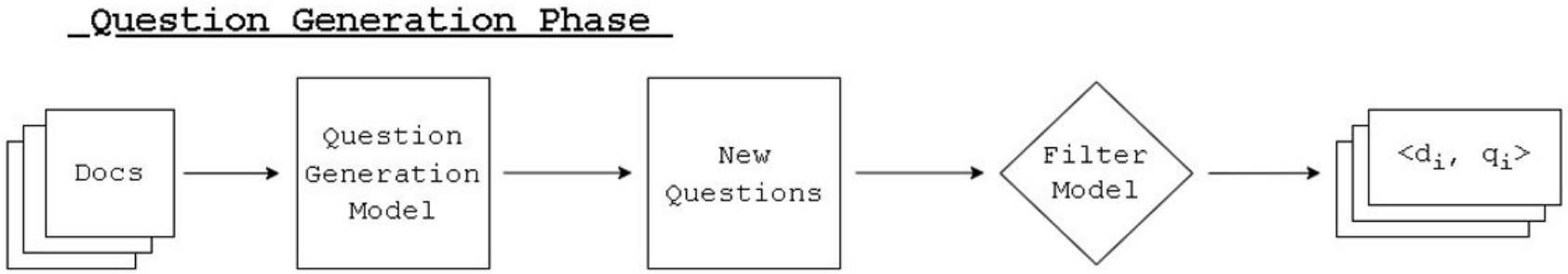
- TF-IDF Approach
- TF-IDF, or Term Frequency-Inverse Document Frequency, is a numerical statistic used in natural language processing and information retrieval to evaluate the importance of a word in a document relative to a collection of documents (corpus).
- It is calculated by multiplying the term frequency (how often a word appears in a document) by the inverse document frequency (reciprocal of the proportion of documents containing the word).

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

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

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

QUESTION GENERATION MODEL



- **Base Model :** mT5
- **Fine Tuning Datasets:** SQuAD, mC4, TyDiQA
- **Inferencing:** Wiki Passages for Target Language

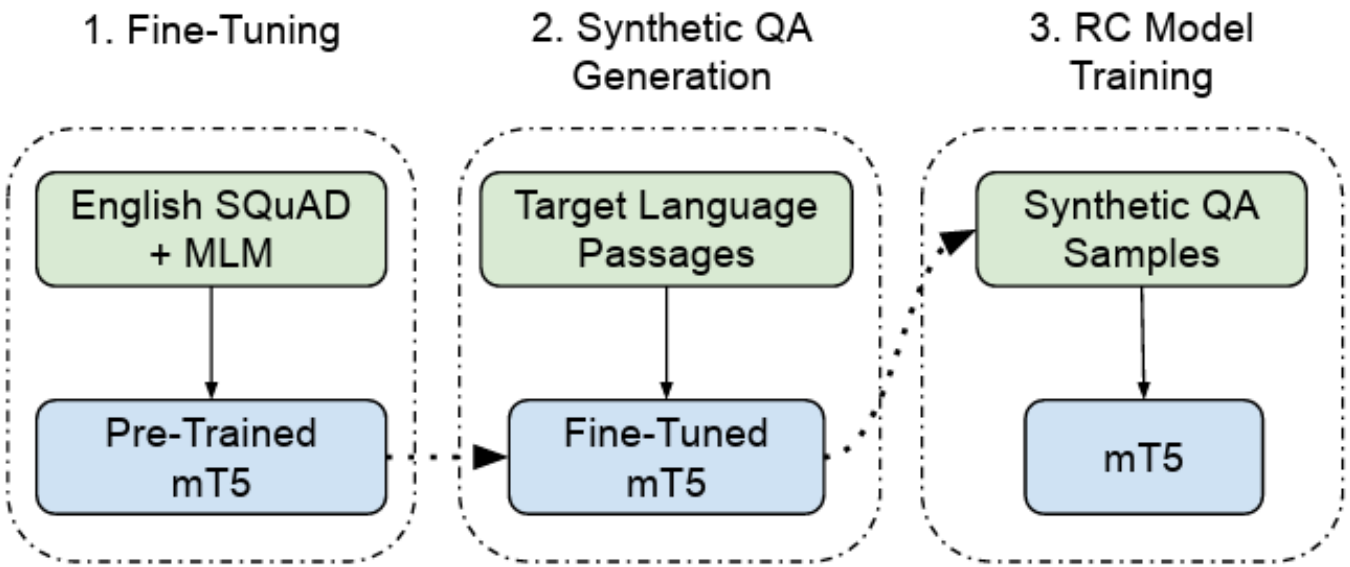
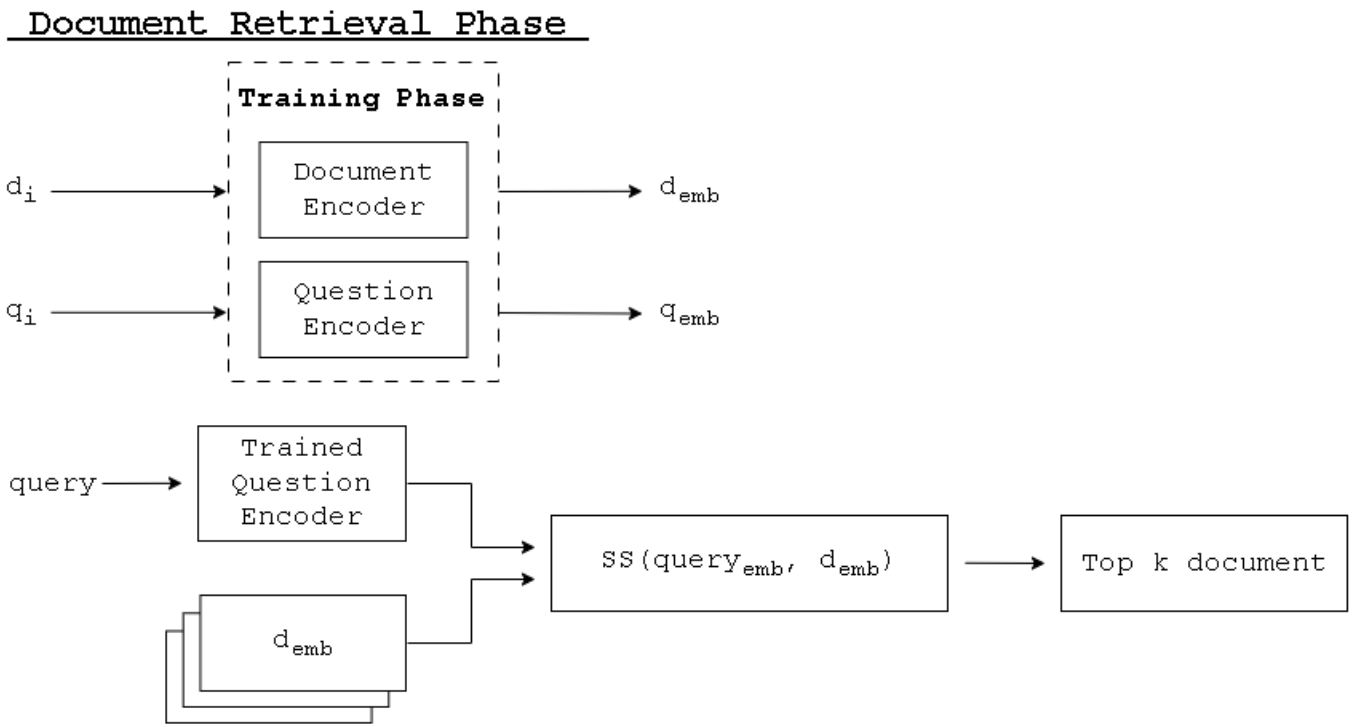


Figure 1: End-to-End pipeline: 1) Fine-tuning the generative model using SQuAD English samples and multilingual MLM. 2) Generating synthetic samples from Wikipedia passages of the target language using the fine-tuned generative model. 3) Training the downstream reading comprehension model using synthetic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

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

Passage Retrieval Model



- **Base Model:** mBERT
- **Fine Tuning Dataset:** Dataset from Question Generation Model
- **Inferencing:** Question -> Top ‘k’ Documents

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

Evaluation & Inferencing

```
[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(dpr_dataset[actual_doc[0]])
rankings, similarityScores = tfidfModel1.similarityMeasure(queries[0], 5)

print("actual_doc: ", actual_doc[0])
predictions.append(rankings)
for i in range(len(rankings)):
  print(f'Rank {i} -> {rankings[i]}')
```

```
{'question': 'कुतुब मीनार की ऊंचाई कितनी है?', 'positivie_context': 'कुतुब समूह के अन्य उल्लेखनीय स्थलों एवं निर्माणों हेतु देखें मुख्य लेख\nकुतुब मीनार भारत में दक्षिण दिल्ली शहर के महरौली भाग में स्थित, ईंट से
actual_doc:  3
Rank 0 -> 75
Rank 1 -> 11
Rank 2 -> 63
Rank 3 -> 5
Rank 4 -> 56
```

Inferencing (From Train data)

```
Question - इंडोनेशिया की राजधानी क्या है
RANK - 1  -> Document -  464
RANK - 2  -> Document -  30
RANK - 3  -> Document -  147
RANK - 4  -> Document -  576
RANK - 5  -> Document -  260

First Retrieved Document -
```

इंडोनेशिया गणराज्य (दीपान्तर गणराज्य) दक्षिण पूर्व एशिया और ओशिनिया में स्थित एक देश है। १७५०८ द्वीपों वाले इस देश की जनसंख्या लगभग 26 करोड़ है, यह दुनिया का तीसरा सबसे अधिक आबादी

इतिहास

इसा पूर्व ४थी शताब्दी से ही इंडोनेशिया द्वीपसमूह एक महत्वपूर्ण व्यापारिक क्षेत्र रहा है। बुनी अथवा मुनि सभ्यता इंडोनेशिया की सबसे पुरानी सभ्यता है। ४थी शताब्दी इस्सा पूर्व तक ये सभ्यता काफी उन्नति कर

नामोत्पत्ति

इसका और साथ के अन्य द्वीप देशों का नाम भारत के पुराणों में दीपान्तर भारत (अर्थात सागर पार भारत) है। यूरोप के लेखकों ने १५० वर्ष पूर्व इसे इंडोनेशिया (इंद= भारत + नेसोस = द्वीप के लिये) दिया

अर्थव्यवस्था

इंडोनेशिया एक मिश्रित अर्थव्यवस्था है, जिसमे निजी क्षेत्र एवं सरकारी क्षेत्र दोनों की भूमिका है। इंडोनेशिया दक्षिण-पूर्वी एशिया की सबसे बड़ी अर्थव्यवस्था है और जी-२० अर्थव्यवस्थाओं में से एक है। सन २०१०
विश्व व्यापार संगठन के अनुसार 2020 में चीन को पीछे छोड़ कर इंडोनेशिया विश्व का सबसे बड़ा निर्यातक बन जाएगा। तेल और गैस, इलेक्ट्रिकल उपकरण, प्लाय-वुड, रबड़ एवं वस्त्र मुख्य निर्यात रहेंगे। रसा

भाषा

यहाँ की मुख्य भाषा-भाषा इंडोनेशिया है। अन्य भाषाओं में जावा, बाली, भाषा सुंडा, भाषा मदुरा आदि भी हैं। प्राचीन भाषा का नाम कावी था जिसमें देश के प्रमुख साहित्यिक ग्रन्थ हैं।

चुनौतियां

लेकिन इसके बाद से इंडोनेशिया का इतिहास उथलपुथल भरा रहा है, चाहे वह प्राकृतिक आपदाओं की वजह से हो, भ्रष्टाचार की वजह से, अलगाववाद या फिर लोकतंत्रीकरण की प्रक्रिया से उत्पन्न चुनौतियां ह

प्राचीन राजवंश
श्रीविजय राजवंश
शैलेन्द्र राजवंश
संजय राजवंश
माताराम राजवंश
केदिरि राजवंश
सिंहश्री
```
```

Inferencing (From Train data)

```
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 ख्वान
चू ख्वान् (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 दि
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 युव
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 पुन: 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 ये 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

Inferencing (From Testdata)

```
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 में 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 स्पेक्ट्र 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

Evaluation : Accuracy & Mean Reciprocal Rank

TF IDF

```
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
```

OUR MODEL

```
Train Accuracy of Retrieved Documents : 0.8157453936348409
Train MRR of Retrieved Documents : 0.5783919597989943

Train + Test Accuracy of Retrieved Documents : 0.7466487935656837
Train + Test MRR of Retrieved Documents : 0.5149016979445923
```

Our Novel Ideas

- Data Augumentation of Target Language.
- Zero Shot Approach.
- Question Generation -> DPR
- FAISS Usage for Retrieval using Indexing.

Note:-
For the implementation, we have researched and came through the discussed pipeline which was not yet followed or implemented.

CHALLENGES FACED

There were not many references to start with for Dense Retrieval for Low resource languages especially for completely Zero Shot Approach.

Lack of resource: In the paper it was suggested to use ”mt5” but we could only use "mt5-small".

The code of the paper was not provided by the authors and some things were unclear in the paper.

Data Prep: loading of complete mc4 dataset in RAM

THANK YOU!
धन्यवाद!
ధన్యవాదాలు!