BERToid

By: Its GIF not GIF

Smart Lightweight Medical Query System (SLiMQ)

What are we solving

- A smart medical QA system to aid doctors that doesn’t hallucinate
- Lightweight enough to run on off-the shelf devices (Eg: mobile phones)

Approach

- Read the “Question Answering with Long Multiple-Span Answers by Ming Zhu et al.” which introduced the idea of comparing the context’s sentence similarity with the query as well as similarity with other sentences.
- Even though the paper claimed this method was suboptimal, this felt the most approachable for a hackathon.

Approach #1: Fine tune LLAMa on the MASHQA dataset

- The instruction prompt to LLAMA would be the query and the context and the target was the correct answers.

Approach #1: Fine tune LLAMa on the MASHQA dataset

- The instruction prompt to LLAMA would be the query and the context and the target was the correct answers.

- Challenge: Token limit of LLAMA

Approach #1: Fine tune LLAMa on the MASHQA dataset

- The instruction prompt to LLAMA would be the query and the context and the target was the correct answers.

- Challenge: Token limit of LLAMA

- Solution: Chunk the context and concatenate individual results using langchain or smtg similar.

Problem with the approach

Quantized LLAMA with less than a billion parameters is still unreasonably slow on a CPU.

Took 2 minutes to print one token locally (unacceptable)

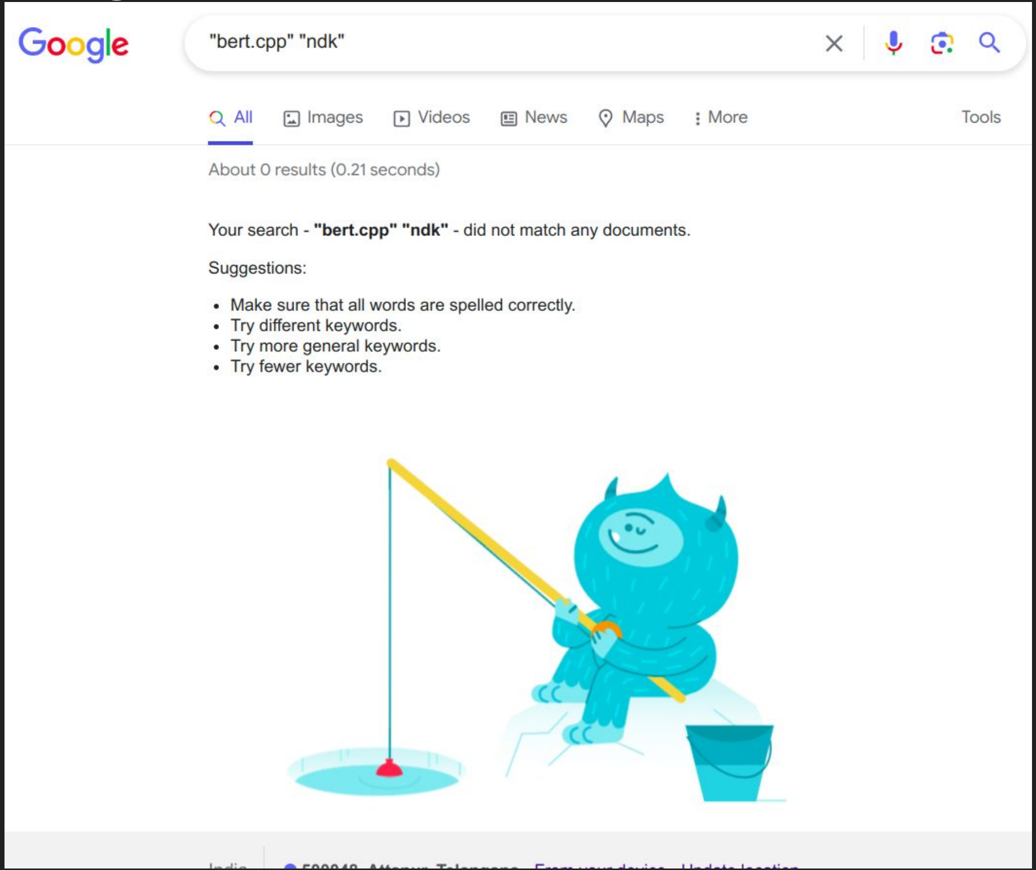
Approach #2: Use BERT

- Use BERT which is a relatively smaller model

- Get the similarity scores between each sentence of the context and the query.

- Print the sentences which are above some threshold of similarity

Major challenge


Google
"bert.cpp" "ndk"
All
Images
Videos
News
Maps
More
Tools

About 0 results (0.21 seconds)

Your search - **"bert.cpp" "ndk"** - did not match any documents.

Suggestions:

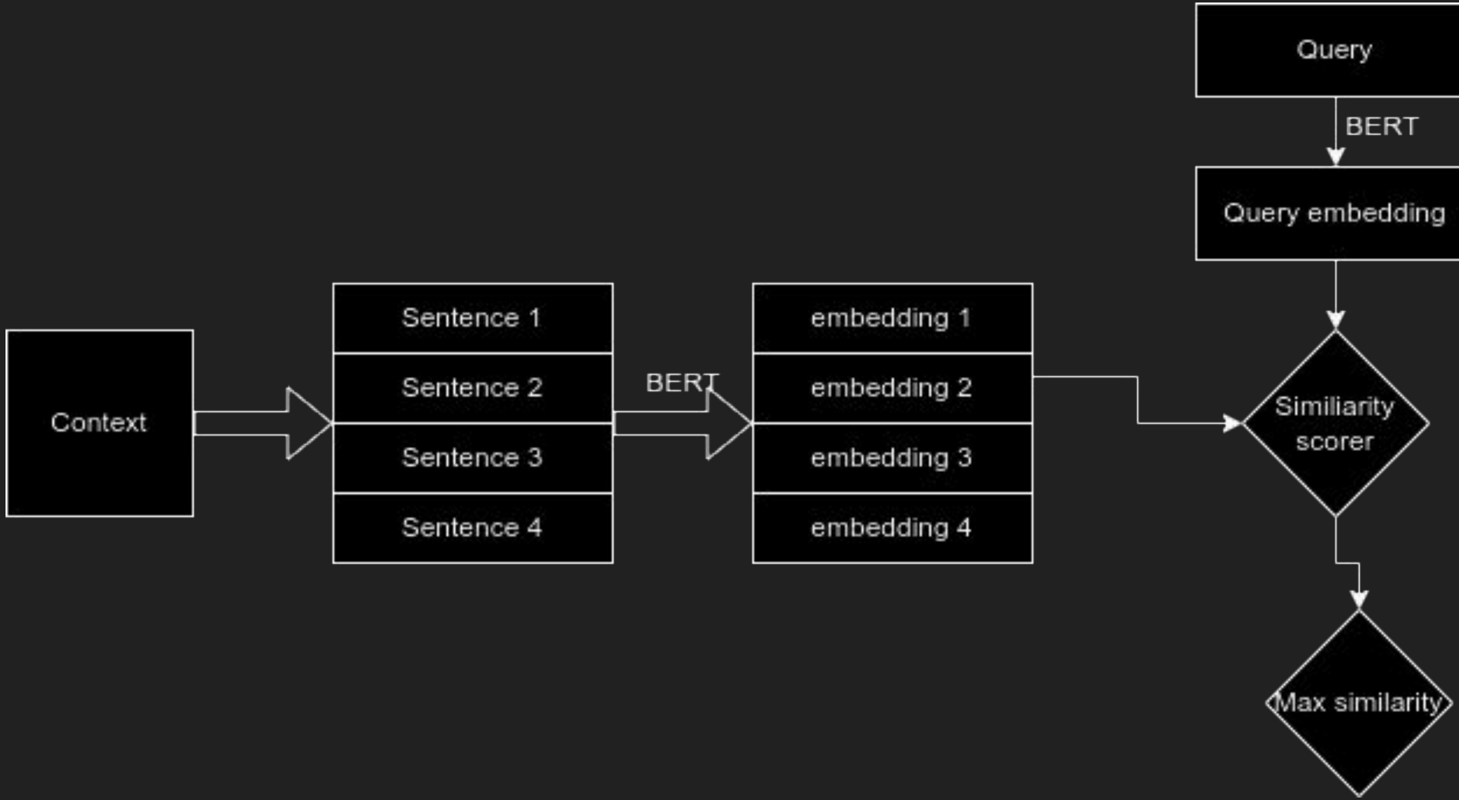
- Make sure that all words are spelled correctly.
- Try different keywords.
- Try more general keywords.
- Try fewer keywords.

India
500048, Attapur, Telangana - From your device - Update location

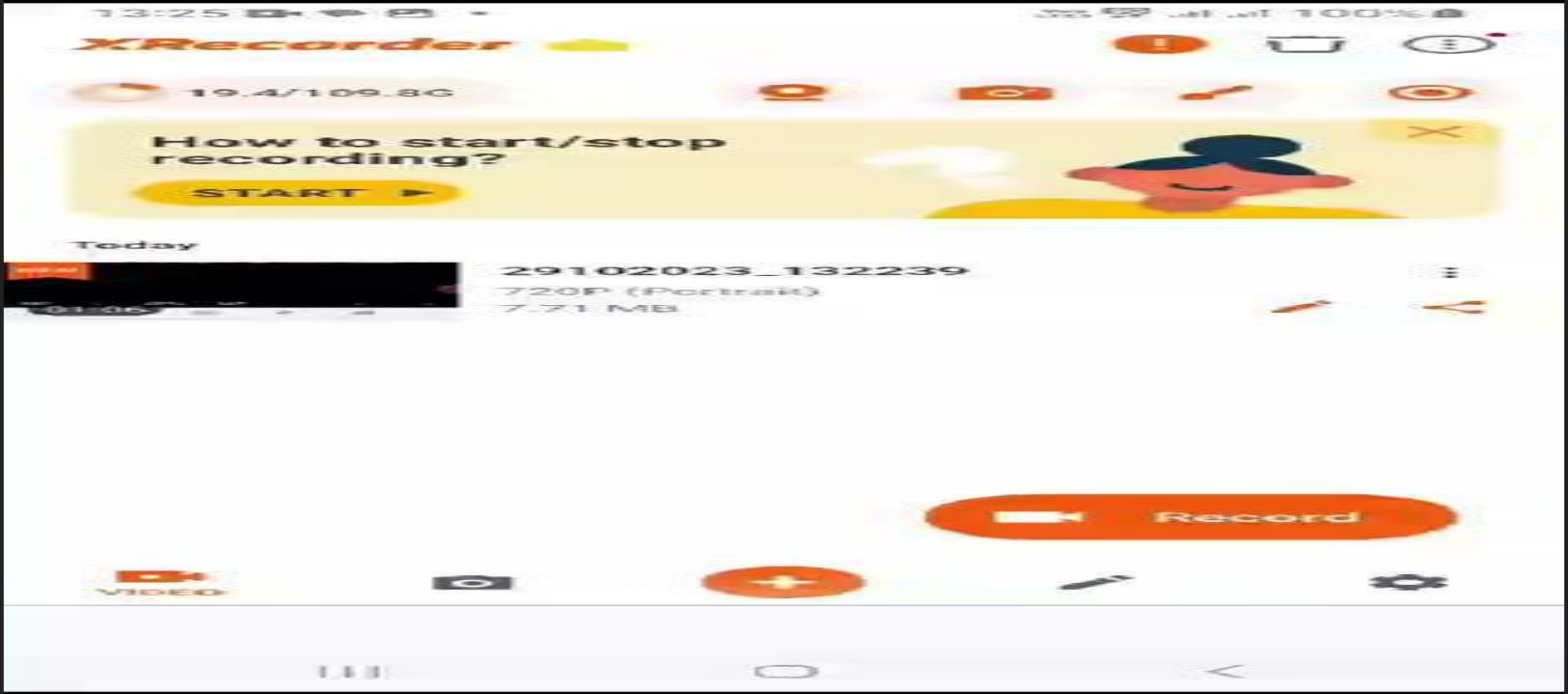
Ported bert.cpp to android

It can run using termux

Flow Chart


Query
BERT
Query embedding
Context
Sentence 1
Sentence 2
Sentence 3
Sentence 4
BERT
embedding 1
embedding 2
embedding 3
embedding 4
Similiarity scorer
Max similarity

Demo



How to start/stop recording?

Today

29102023_132239
720P (Portrait)
7.71 MB

thank you!!