mlc.course: Team 2

\(*°▽°*)/

ABROR SHOPULATOV
NAVFALBEK MAKHFUZULLAEV
SAMIR IRGASHEV

Problem Statement

Efficiency

Traditional Q&A sessions can be inefficient, especially if there are many questions and limited time.

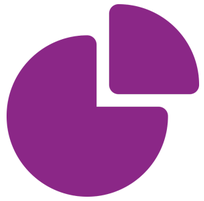
Limited Availability

Human labour is limited to a particular Q&A chats at a time

Pricey

Long term, AI chat-bot can offer financial flexibility

Problem Statement



Large Data

Data can be so large that normal person simply cannot answer the question correctly.

Facts

AI provides factual information contrary to the habit of making mistakes which applies to humans

How it works

Question: **Kitob nechta ertakni o'z ichiga olgan?**

Context: **Har bir ertak taxminan 20 bet, demak 6 ta atrofida.**

QA model

Prediction | Prediction | Prediction

Aggregation

Final Prediction

Answer: **6**

Image Source: https://www.coastdigital.co.uk/2019/10/28/seo-scoop-google-introduces-bert-to-rankings/

QA Model

- Reads and understands the retrieved documents to identify relevant passages that contain the answer to the question.
- Uses deep learning models such as BERT or RoBERTa to encode and analyze the text.
- Can be trained using supervised learning on a large dataset of question-answer pairs to improve accuracy and robustness.

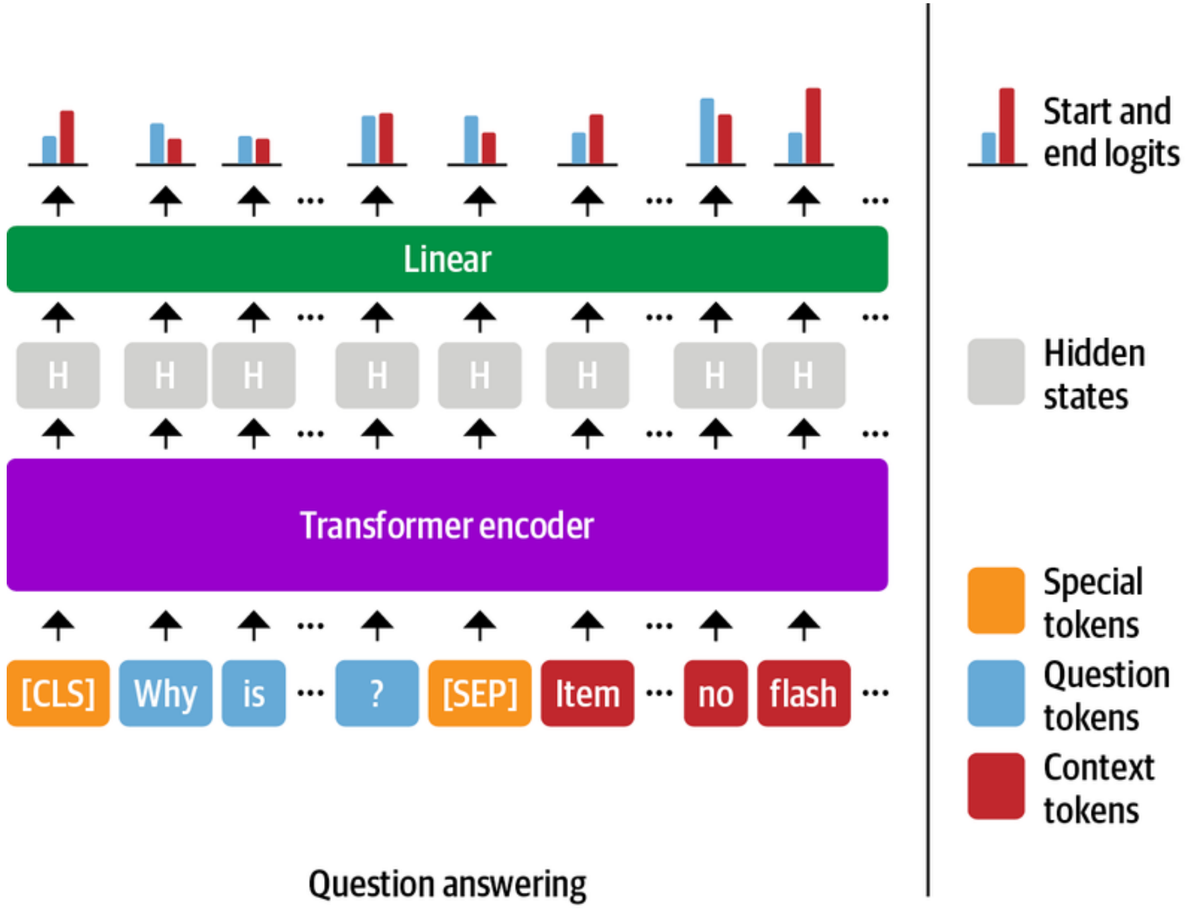
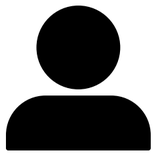


Image Source: https://github.com/nlp-with-transformers/notebooks

Is that all?

All documents



Okay, read through and find the answer

Why don't you do it yourself?

Image Source: https://www.dbmgrow.com/googles-bert-looking-at-what-you-mean-not-what-you-say/

Chunking documents

What if we divide text into chunks!

Document	Document	Document	Document
Document	Document	Document	Document

✓ >1000x speedup

✓ Easy to scale

✗ Lower accuracy

How it works

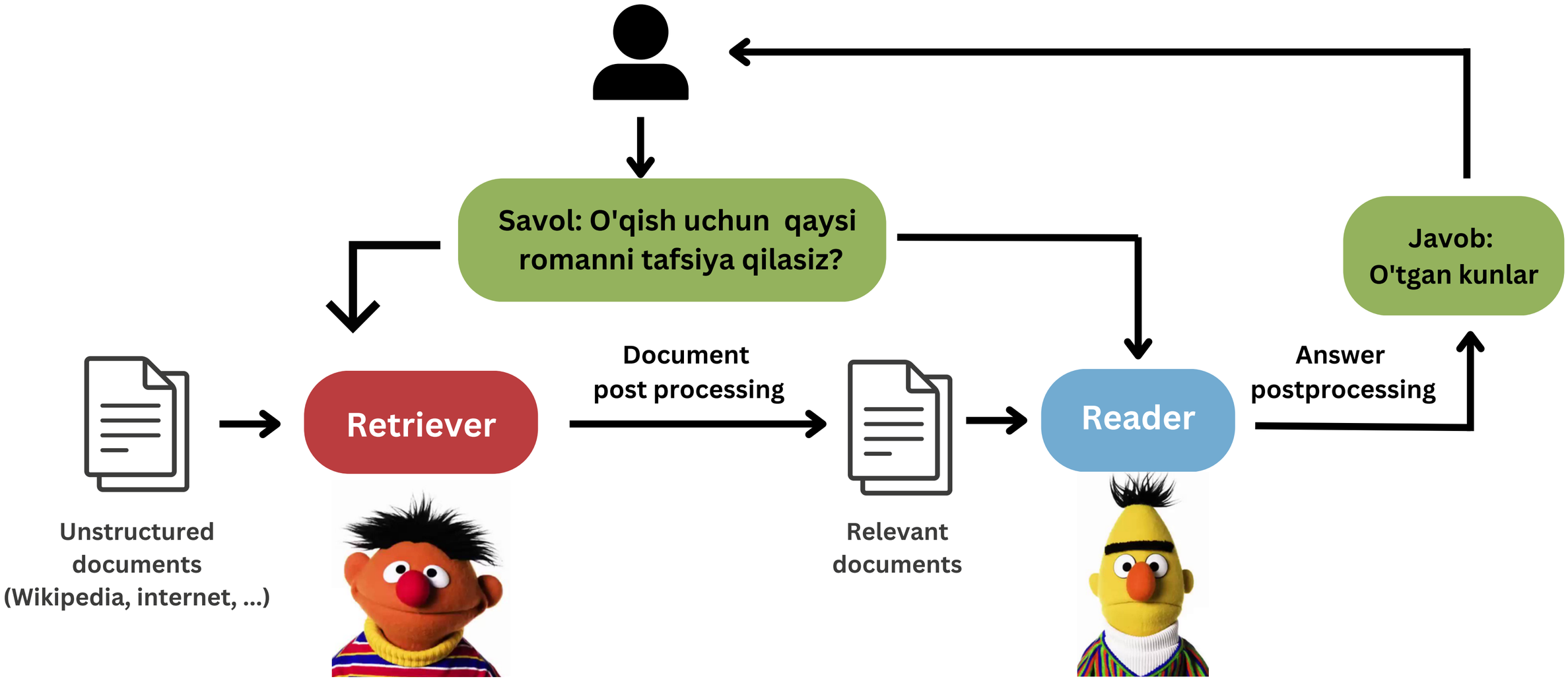
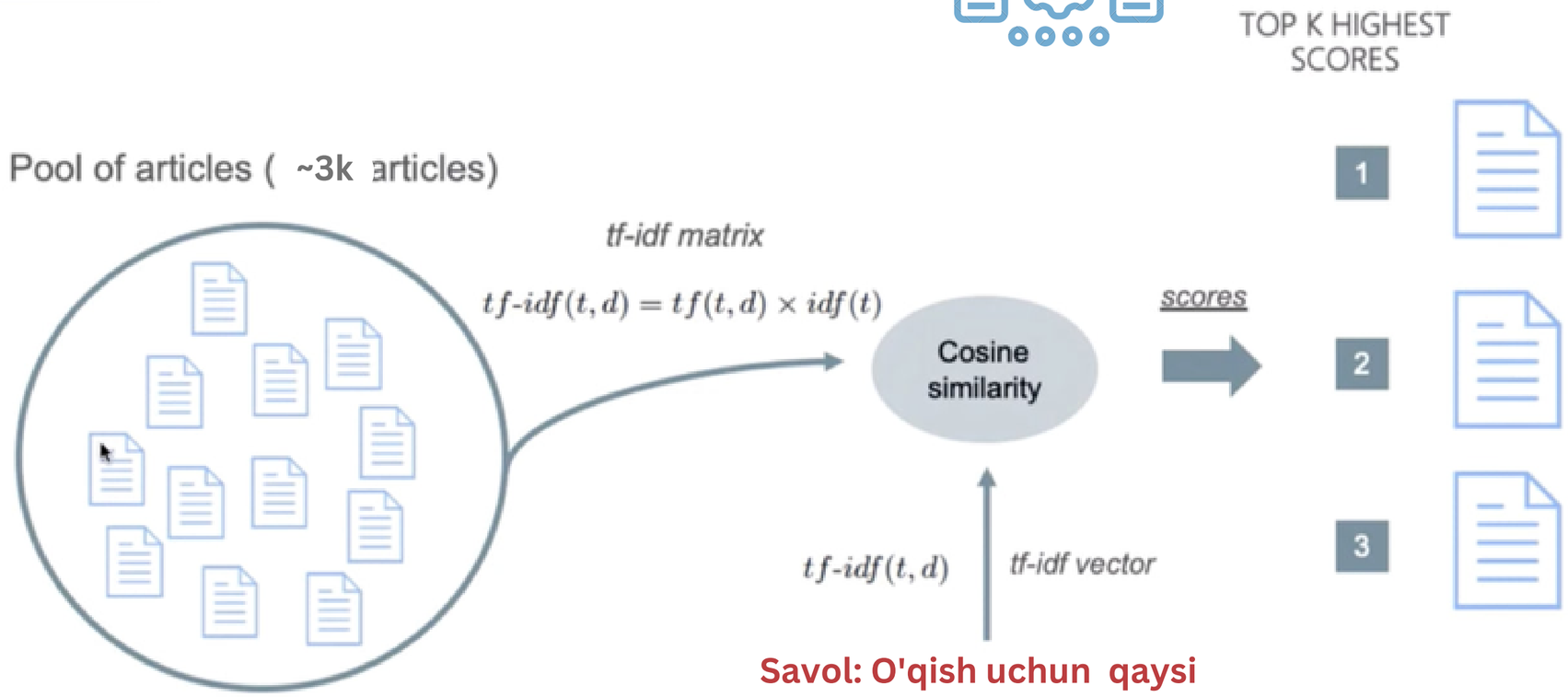


Image Source: https://twitter.com/bertsesame/status/1611411710895603725

Retriever



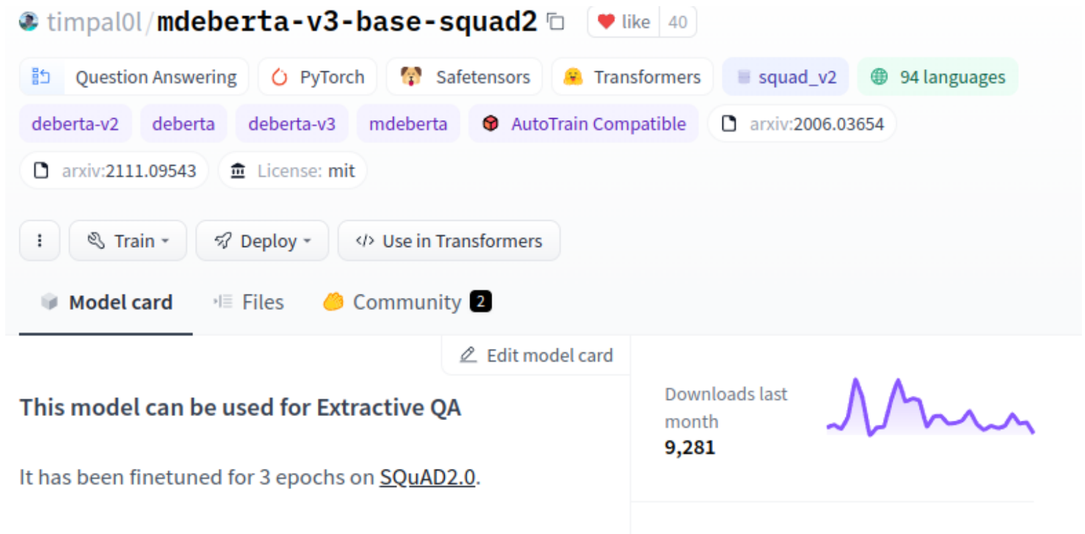
Savol: O'qish uchun qaysi romanni tafsiya qilasiz?

- **Retrieves relevant documents from a large corpus using technique called BM25.**
- **Can filter out irrelevant documents to reduce noise and improve accuracy.**
- **Can be fine-tuned using supervised learning to improve retrieval performance on a specific task.**

Image Source: https://github.com/cdqa-suite/cdQA/issues/344

Putting it all together

QA Model



https://hf.co/timpal0l/mdeberta-v3-base-squad2

Retriever

BM25 (Recommended)

Use BM25 if you are looking for a retrieval method that doesn't need a neural network for indexing. BM25 is a variant of TF-IDF. It improves upon its predecessor in two main aspects:

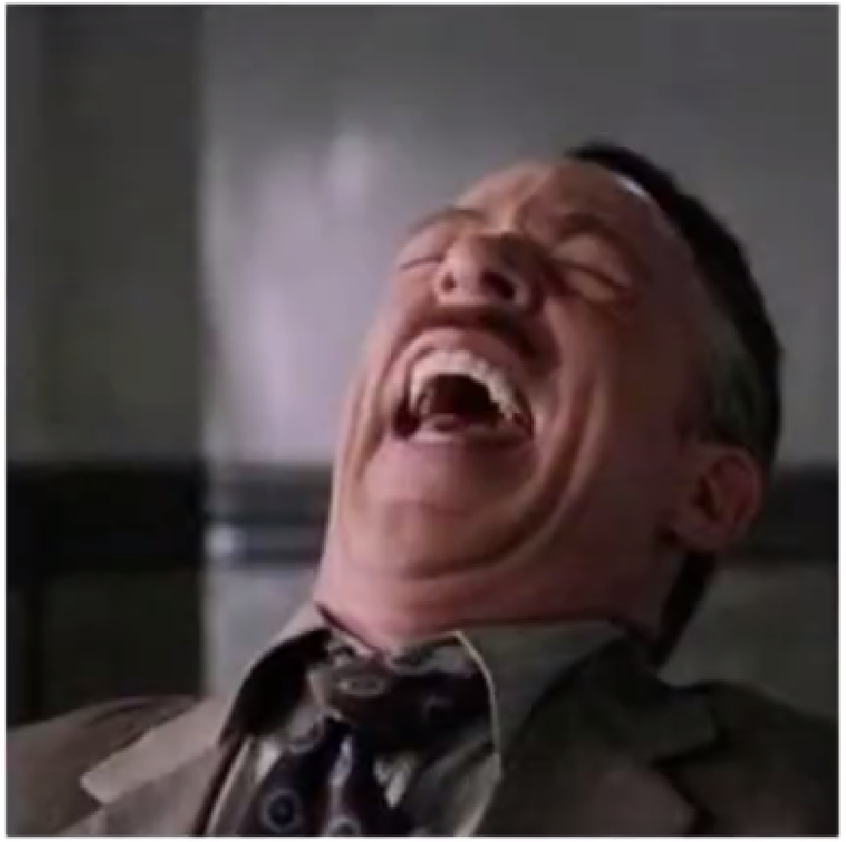
- It saturates `tf` after a set number of occurrences of the given term in the document
- It normalises by document length so that short documents are favoured over long documents if they have the same amount of word overlap with the query

Python

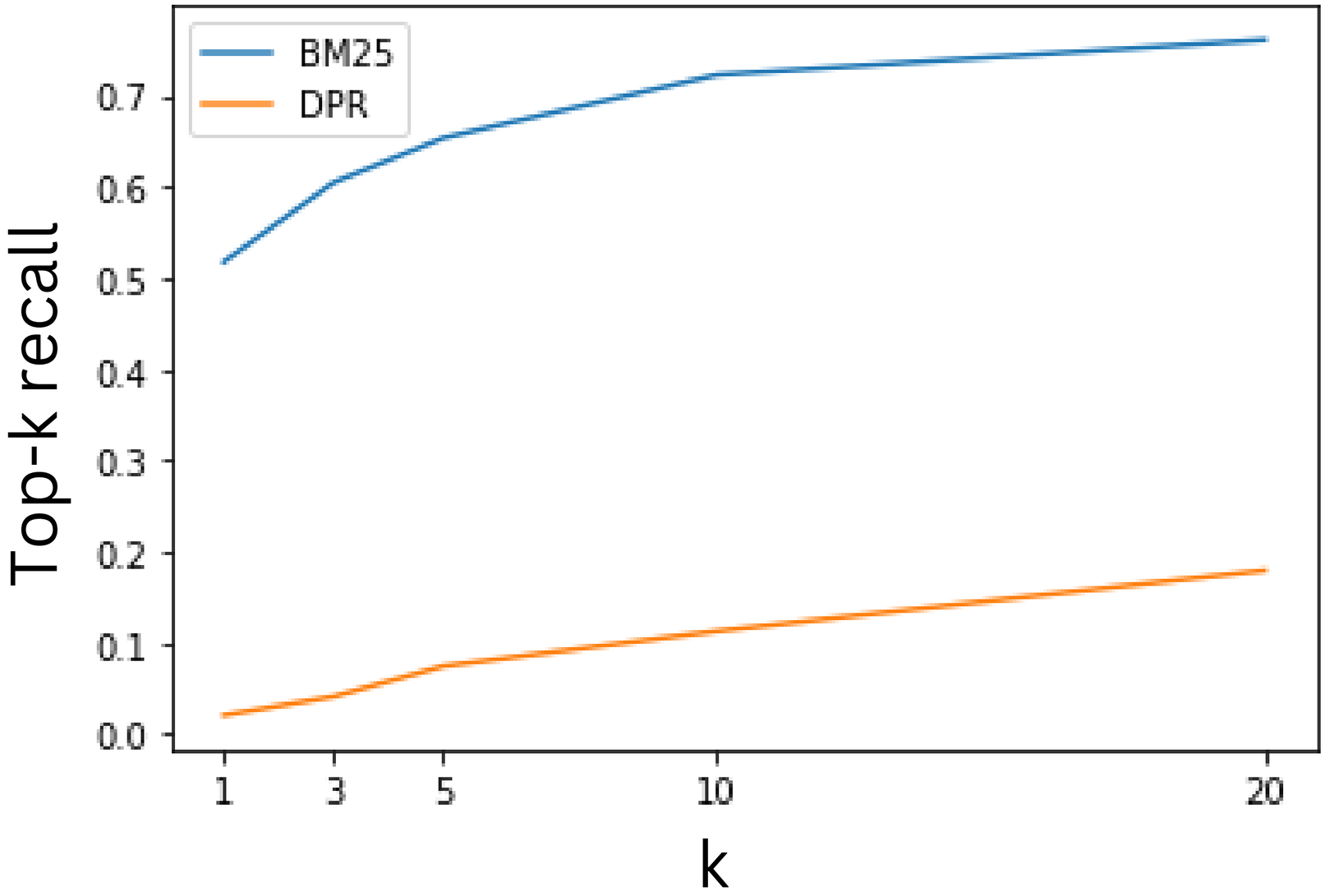
```
from haystack.document_stores import ElasticsearchDocumentStore
from haystack.nodes import BM25Retriever
from haystack.pipelines import ExtractiveQAPipeline

document_store = ElasticsearchDocumentStore()
... retriever = BM25Retriever(document_store)
... p = ExtractiveQAPipeline(reader, retriever)
```

https://docs.haystack.deepset.ai/docs/retriever

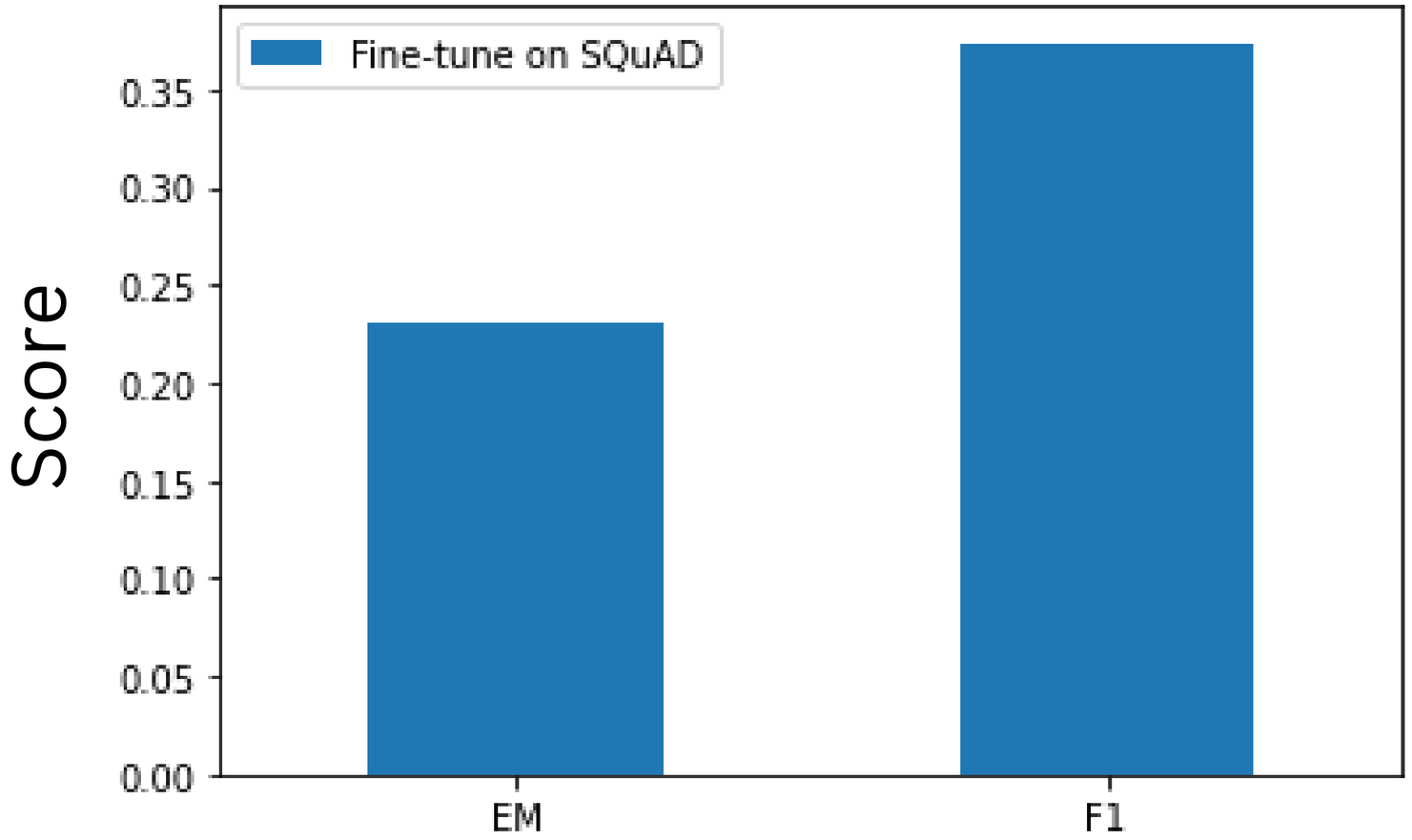
Now done



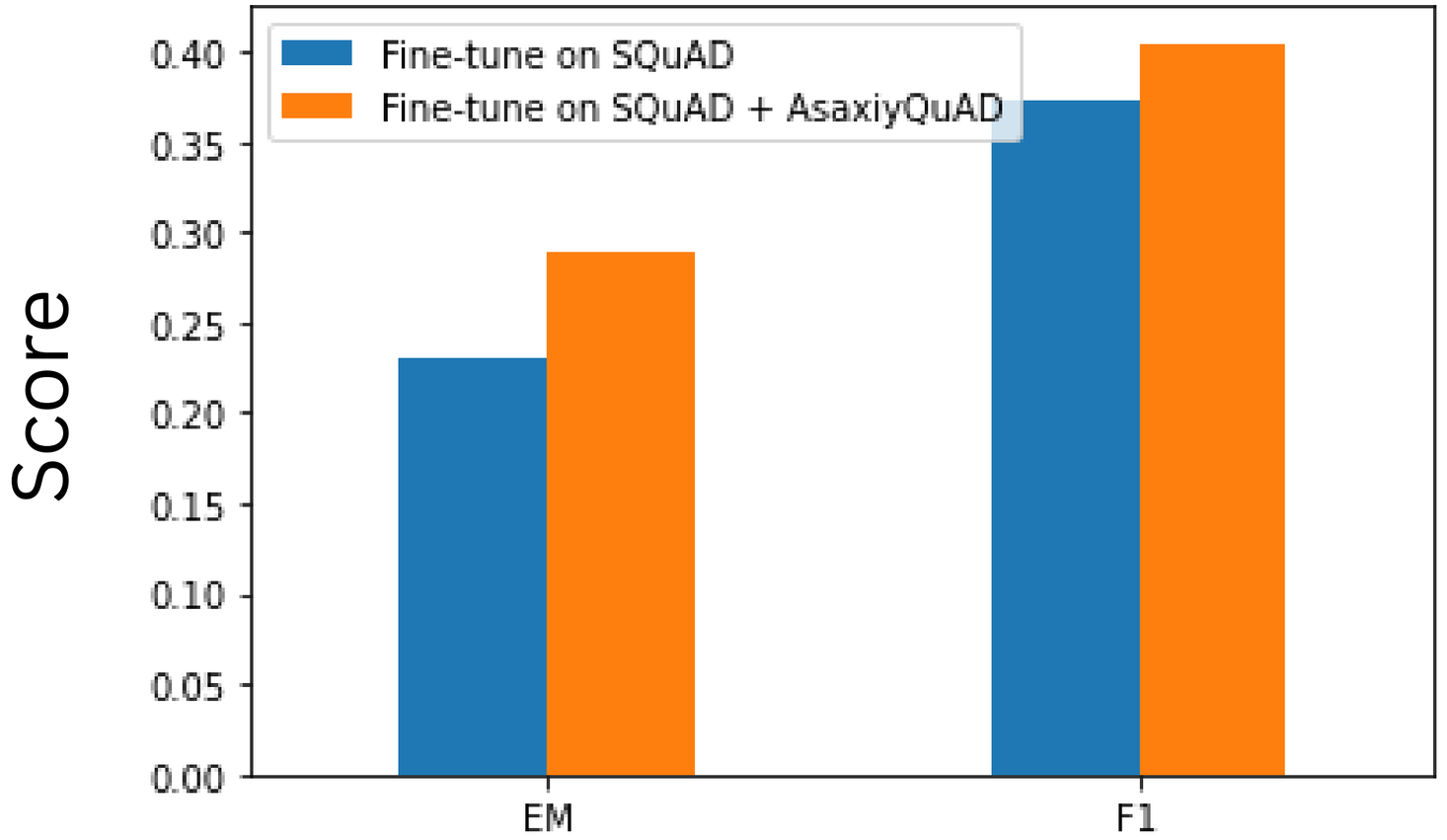
Evaluating the Retriever



Evaluating the Reader



Domain adaptation



Used tools

Data preparation

- haystack.deepset.ai
- An Natural Language Processing framework to the applications

  Haystack is an open-source framework for building search systems that work intelligently over large document collections.

- Haystack's annotation tool

  Annotation tool to label datasets for use with semantic search and question answering.

Haystack
annotation tool

navfalbek.uz@gmail.com

AsaxiyBooks | Documents | Questions | Import | Export Labels

AsaxiyBooks / Document 1461528

Amount of dataset that scrapped from web

MARK DOCUMENT AS DONE | SHOW LABELS OF ALL USERS

269 / 2948

This document is done, thus it cannot be changed anymore. Please open it again on the topleft if needed.

The scrapped data from asaxiy's web page

Questions

U

Dasturlashda yig'latadigan kitob bormi? 1

ADD CUSTOM QUESTION

Question that might be asked form used

Annotation Document

Search

Algoritmlash va dasturlash asoslari. C++ tili asoslari. Amaliy qo'llanma, nazariya, masalalar, mulohazalar, yechimlar, tavsiyalar
B. J. Beltayev
A. R. Azamatov
A. D. Rahimov
B. A. Azamatov
D. T. Asrayeva
Sh. Z. Qambaraliyev

Label to the question

Qo'llanmada C++ dasturlash tili va elementlari, dastur tarkibi va tavsifi, unda dasturlash imkoniyatlari va usullari, tuzilmalar mohiyati erkin tarqatiladigan Codeblocks IDE asosida ochib berishga qaratilgan. Shu bilan birga, umumiy va nazariy ma'lumotlar, dastur namunlaridan tashqari takrorlash va mustaqil ishlash uchun ko'p sonli vazifalar keltirilgan. Keltirilgan vazifalarni hal etishda turli yondashuvlar va usullarning mohiyati cohib berishga harakat qilingan. Ma'lumotlar va vazifalar "oddiydan murakkabga" mantiqiy ketma-ketlikda bayon etilgan. Vazifalar qiyinlik darajasi bo'yicha A, B, C va D turkumlarga ajratilgan. Qo'llanmadan o'qituvchilar, o'quvchilar, talabalar, umuman dasturlash san'ati bilan qiziquvchilar foydalanishlari mumkin. Qo'llanma 11 va undan katta yoshli qiziquvchilar uchun mo'ljallangan.
Kitob 404 betdan iborat. 2021-yilda chop etilgan.
Ko'proq: https://asaxiy.uz/uz/product/algoritmlash-va-dasturlash-asoslari-s-tili-asoslari-amalij-kullanma-nazariya-masalalar-mulohazalar-echimlar-tavsiyalar-akademiknashr-2021

Resources

Codes(+model)

Dataset

Demo

github.com/shopulatov/UzBooksQA

hf.co/datasets/mlcourse-team2/asaxiy-quad-256

Thank you

ʃ ɔ ◕_◕ ʅ ɔ ʃ

mlc.course: Team #2