00:00:00,640 --> 00:00:07,200

today we are going to talk about a.

00:00:04,560 --> 00:00:09,920

subdomain of question answering called.

00:00:07,200 --> 00:00:12,160

long form question answering now before.

00:00:09,920 --> 00:00:14,240

we get into the specifics let's just.

00:00:12,160 --> 00:00:18,160

talk very quickly about question.

00:00:14,240 --> 00:00:21,760

answering as a subdomain of nlp.

00:00:18,160 --> 00:00:24,400

question answering has i think exploded.

00:00:21,760 --> 00:00:26,560

as a subdomain of nlp.

00:00:24,400 --> 00:00:28,720

in the past few years.

00:00:26,560 --> 00:00:31,840

mainly because i think question.

00:00:28,720 --> 00:00:33,520

answering is an incredibly widely.

00:00:31,840 --> 00:00:35,520

applicable.

00:00:33,520 --> 00:00:37,760

use case for nlp.

00:00:35,520 --> 00:00:40,399

but it wasn't possible to do.

00:00:37,760 --> 00:00:41,920

question answering or not anything good.

00:00:40,399 --> 00:00:44,239

with question answering.

00:00:41,920 --> 00:00:45,280

until we had transformed models like.

00:00:44,239 --> 00:00:46,239

that.

00:00:45,280 --> 00:00:48,000

so.

00:00:46,239 --> 00:00:50,239

that means that as soon as we got.

00:00:48,000 --> 00:00:52,800

something like bert the question.

00:00:50,239 --> 00:00:54,160

answering became viable and with the.

00:00:52,800 --> 00:00:56,320

huge.

00:00:54,160 --> 00:00:59,440

number of use cases for question.

00:00:56,320 --> 00:01:01,760

answering it obviously kind of took off.

00:00:59,440 --> 00:01:04,879

now question answering is quite.

00:01:01,760 --> 00:01:05,840

complicated but at its core.

00:01:04,879 --> 00:01:07,840

it's.

00:01:05,840 --> 00:01:12,000

basically just the.

00:01:07,840 --> 00:01:14,640

retrieval of information in a more human.

00:01:12,000 --> 00:01:16,560

like way and when we.

00:01:14,640 --> 00:01:19,360

consider this.

00:01:16,560 --> 00:01:21,600

i think it makes it really clear how.

00:01:19,360 --> 00:01:23,200

broadly applicable question answering is.

00:01:21,600 --> 00:01:25,280

because.

00:01:23,200 --> 00:01:26,479

almost every organization in the world.

00:01:25,280 --> 00:01:29,200

if not all.

00:01:26,479 --> 00:01:31,200

are going to need to retrieve.

00:01:29,200 --> 00:01:32,240

information.

00:01:31,200 --> 00:01:34,240

right and.

00:01:32,240 --> 00:01:36,479

for a lot of companies and particularly.

00:01:34,240 --> 00:01:39,840

larger organizations.

00:01:36,479 --> 00:01:42,640

i think the act of information retrieval.

00:01:39,840 --> 00:01:46,720

is actually a big component of their.

00:01:42,640 --> 00:01:49,280

day-to-day operations now at the moment.

00:01:46,720 --> 00:01:52,399

most organizations.

00:01:49,280 --> 00:01:53,520

do information retrieval across a suite.

00:01:52,399 --> 00:01:54,479

of tools.

00:01:53,520 --> 00:01:56,960

so.

00:01:54,479 --> 00:01:59,439

they will have people.

00:01:56,960 --> 00:02:02,479

using some sort of internal search tools.

00:01:59,439 --> 00:02:04,479

which are typically keyword based which.

00:02:02,479 --> 00:02:05,520

is.

00:02:04,479 --> 00:02:07,759

generally.

00:02:05,520 --> 00:02:10,399

not always that helpful sometimes it's.

00:02:07,759 --> 00:02:12,400

useful but a lot of the time.

00:02:10,399 --> 00:02:14,080

it's not great.

00:02:12,400 --> 00:02:16,400

and then another.

00:02:14,080 --> 00:02:19,840

key form of information retrieval in.

00:02:16,400 --> 00:02:21,200

most organizations is literally person.

00:02:19,840 --> 00:02:23,760

to person.

00:02:21,200 --> 00:02:24,959

so you go and ask someone who you think.

00:02:23,760 --> 00:02:27,520

will probably know where some.

00:02:24,959 --> 00:02:29,280

information is like a document or so on.

00:02:27,520 --> 00:02:31,920

and obviously.

00:02:29,280 --> 00:02:33,920

you know this sort of patchwork of.

00:02:31,920 --> 00:02:37,040

information retrieval.

00:02:33,920 --> 00:02:37,920

to an extent sure it works but.

00:02:37,040 --> 00:02:41,040

it's.

00:02:37,920 --> 00:02:42,800

inefficient now if we consider that many.

00:02:41,040 --> 00:02:46,080

organizations.

00:02:42,800 --> 00:02:48,319

contain thousands of employees.

00:02:46,080 --> 00:02:52,160

each of those employees producing pages.

00:02:48,319 --> 00:02:54,400

up on pages of unstructured data eg.

00:02:52,160 --> 00:02:56,879

pages of documents and text that are.

00:02:54,400 --> 00:02:58,159

meant for human consumption.

00:02:56,879 --> 00:03:00,400

in most.

00:02:58,159 --> 00:03:02,800

cases all of that information is just.

00:03:00,400 --> 00:03:04,959

being lost in some sort of void.

00:03:02,800 --> 00:03:07,360

okay and.

00:03:04,959 --> 00:03:09,360

rather than that information being lost.

00:03:07,360 --> 00:03:11,040

in a void that we're never going to see.

00:03:09,360 --> 00:03:13,840

again it becomes useless to the.

00:03:11,040 --> 00:03:16,879

organization or the company.

00:03:13,840 --> 00:03:19,519

we can instead place it in.

00:03:16,879 --> 00:03:21,519

a database that a.

00:03:19,519 --> 00:03:22,560

question answering.

00:03:21,519 --> 00:03:25,040

agent.

00:03:22,560 --> 00:03:26,959

has access to and when we ask a question.

00:03:25,040 --> 00:03:28,720

to that q a agent which we ask in a.

00:03:26,959 --> 00:03:30,400

human-like way.

00:03:28,720 --> 00:03:32,480

it will go and retrieve relevant.

00:03:30,400 --> 00:03:34,560

information for us instantly well not.

00:03:32,480 --> 00:03:36,720

instantly but pretty close.

00:03:34,560 --> 00:03:39,840

the majority of.

00:03:36,720 --> 00:03:42,080

data in the world is unstructured and.

00:03:39,840 --> 00:03:45,200

there's a few different sources for this.

00:03:42,080 --> 00:03:48,640

but i think places like forbes.

00:03:45,200 --> 00:03:51,120

estimate that number to be around 90.

00:03:48,640 --> 00:03:53,840

of the world's data so in your.

00:03:51,120 --> 00:03:56,480

organization you probably have a number.

00:03:53,840 --> 00:03:58,000

similar to this so 90 of your data is.

00:03:56,480 --> 00:03:59,280

unstructured.

00:03:58,000 --> 00:04:02,080

that means it's meant for human.

00:03:59,280 --> 00:04:04,959

consumption not machines and it means.

00:04:02,080 --> 00:04:06,400

it's liable to get lost in that void.

00:04:04,959 --> 00:04:08,080

where we're just never going to see that.

00:04:06,400 --> 00:04:11,280

information ever again.

00:04:08,080 --> 00:04:14,480

now that's massively inefficient.

00:04:11,280 --> 00:04:16,160

question answering is an opportunity.

00:04:14,480 --> 00:04:18,160

to not lose that.

00:04:16,160 --> 00:04:20,160

and actually benefit.

00:04:18,160 --> 00:04:23,680

from that information.

00:04:20,160 --> 00:04:25,680

now in question answering there are two.

00:04:23,680 --> 00:04:26,639

main approaches.

00:04:25,680 --> 00:04:30,320

in.

00:04:26,639 --> 00:04:31,919

both cases of question answering we.

00:04:30,320 --> 00:04:35,600

saw those documents.

00:04:31,919 --> 00:04:38,320

in or usually we saw those documents in.

00:04:35,600 --> 00:04:40,560

a document store or vector database so.

00:04:38,320 --> 00:04:42,800

these documents are what we would call.

00:04:40,560 --> 00:04:45,040

sentences or paragraphs.

00:04:42,800 --> 00:04:48,080

extracted from your.

00:04:45,040 --> 00:04:50,639

for example pdfs or emails or whatever.

00:04:48,080 --> 00:04:51,680

unstructured data you have out there.

00:04:50,639 --> 00:04:54,639

and.

00:04:51,680 --> 00:04:55,360

we retrieve data from that.

00:04:54,639 --> 00:04:57,759

and.

00:04:55,360 --> 00:04:59,759

then the next step is where we have the.

00:04:57,759 --> 00:05:02,320

two different forms of question.

00:04:59,759 --> 00:05:03,600

answering with that relevant information.

00:05:02,320 --> 00:05:06,320

that we have.

00:05:03,600 --> 00:05:08,320

from our sort of document store.

00:05:06,320 --> 00:05:10,639

based on a query that we've passed.

00:05:08,320 --> 00:05:12,560

through we either.

00:05:10,639 --> 00:05:14,960

generate an answer.

00:05:12,560 --> 00:05:16,560

or we extract an answer.

00:05:14,960 --> 00:05:18,240

so obviously when we're generating an.

00:05:16,560 --> 00:05:20,000

answer we look at the whole.

00:05:18,240 --> 00:05:20,960

all the context that we've retrieved and.

00:05:20,000 --> 00:05:24,639

we.

00:05:20,960 --> 00:05:27,919

use an nlp model to generate.

00:05:24,639 --> 00:05:31,039

some sort of human answer to.

00:05:27,919 --> 00:05:34,479

our query based on that information.

00:05:31,039 --> 00:05:36,479

otherwise we use an extractive model.

00:05:34,479 --> 00:05:39,840

which is literally going to take a.

00:05:36,479 --> 00:05:42,400

snippet of information from the data.

00:05:39,840 --> 00:05:43,520

that we have retrieved.

00:05:42,400 --> 00:05:45,360

so.

00:05:43,520 --> 00:05:47,440

there's a few.

00:05:45,360 --> 00:05:49,919

components that i just described there.

00:05:47,440 --> 00:05:51,840

there's that document store at the start.

00:05:49,919 --> 00:05:52,720

when we're using a document store which.

00:05:51,840 --> 00:05:55,759

we.

00:05:52,720 --> 00:05:57,600

will in most cases i'd imagine.

00:05:55,759 --> 00:05:59,440

we call that open book question.

00:05:57,600 --> 00:06:02,880

answering now the reason it's called.

00:05:59,440 --> 00:06:06,880

open book is it is like students in an.

00:06:02,880 --> 00:06:09,520

exam okay we have a typical exam you.

00:06:06,880 --> 00:06:11,680

don't have any outside materials to.

00:06:09,520 --> 00:06:13,600

refer to it's just you have to rely on.

00:06:11,680 --> 00:06:16,000

what is in your brain.

00:06:13,600 --> 00:06:18,800

that's very similar to.

00:06:16,000 --> 00:06:21,039

using for example generator model.

00:06:18,800 --> 00:06:24,000

that given a question it doesn't refer.

00:06:21,039 --> 00:06:26,400

to any document store it just refers to.

00:06:24,000 --> 00:06:27,600

what is within its own.

00:06:26,400 --> 00:06:29,840

memory.

00:06:27,600 --> 00:06:32,479

or its own model memory and that model.

00:06:29,840 --> 00:06:33,759

memory has been built during model.

00:06:32,479 --> 00:06:35,840

training.

00:06:33,759 --> 00:06:37,120

so that would be referred to as clothes.

00:06:35,840 --> 00:06:40,000

book.

00:06:37,120 --> 00:06:42,319

generative or abstractive q a on the.

00:06:40,000 --> 00:06:44,720

other hand we have a document store so.

00:06:42,319 --> 00:06:48,080

that document store is like we are in.

00:06:44,720 --> 00:06:50,000

our exam as students and we have a open.

00:06:48,080 --> 00:06:51,440

book that we can refer to for.

00:06:50,000 --> 00:06:53,680

information so we're not just relying on.

00:06:51,440 --> 00:06:56,160

what is in our head but we're looking at.

00:06:53,680 --> 00:06:58,720

the information in this book and we.

00:06:56,160 --> 00:07:01,440

still need to rely on the knowledge in.

00:06:58,720 --> 00:07:02,400

our head in order to apply what is in.

00:07:01,440 --> 00:07:06,000

that book.

00:07:02,400 --> 00:07:09,120

to the questions we're given in the exam.

00:07:06,000 --> 00:07:11,919

it's exactly the same for open book.

00:07:09,120 --> 00:07:14,160

abstractive question answering.

00:07:11,919 --> 00:07:16,560

in that you have the.

00:07:14,160 --> 00:07:18,479

generator model but we're not just.

00:07:16,560 --> 00:07:21,280

relying on a generator model to answer.

00:07:18,479 --> 00:07:23,360

our questions we are also relying on a.

00:07:21,280 --> 00:07:25,680

document saw which is our book.

00:07:23,360 --> 00:07:27,440

and what is called a retrieval model and.

00:07:25,680 --> 00:07:31,199

this retriever model.

00:07:27,440 --> 00:07:32,160

is going to take our question.

00:07:31,199 --> 00:07:36,880

it will.

00:07:32,160 --> 00:07:39,360

encode it into a vector embedding.

00:07:36,880 --> 00:07:41,120

takes it to that document store.

00:07:39,360 --> 00:07:43,919

which is.

00:07:41,120 --> 00:07:47,680

actually just a vector database in in.

00:07:43,919 --> 00:07:50,479

our scenario of what we're doing.

00:07:47,680 --> 00:07:52,479

and in a vector database what you have.

00:07:50,479 --> 00:07:54,800

is lots of other.

00:07:52,479 --> 00:07:57,199

vector embeddings which are essentially.

00:07:54,800 --> 00:07:59,759

numerical representations.

00:07:57,199 --> 00:08:01,520

of the documents that you stored in it.

00:07:59,759 --> 00:08:04,080

before so remember documents are those.

00:08:01,520 --> 00:08:05,840

chunks of paragraph or sentences.

00:08:04,080 --> 00:08:07,680

from different sources.

00:08:05,840 --> 00:08:08,879

that vector database has loads of these.

00:08:07,680 --> 00:08:09,919

um.

00:08:08,879 --> 00:08:13,919

these.

00:08:09,919 --> 00:08:17,280

what we call context vectors and we.

00:08:13,919 --> 00:08:19,840

pass our query vector into.

00:08:17,280 --> 00:08:23,280

that document store or vector database.

00:08:19,840 --> 00:08:25,520

and we retrieve the most similar.

00:08:23,280 --> 00:08:26,639

context vectors from there and pass them.

00:08:25,520 --> 00:08:30,720

back to.

00:08:26,639 --> 00:08:32,399

our sort of retrieval pipeline.

00:08:30,720 --> 00:08:33,599

then that is passed to our generator.

00:08:32,399 --> 00:08:35,360

model.

00:08:33,599 --> 00:08:36,800

our generator model is going to see the.

00:08:35,360 --> 00:08:39,440

query.

00:08:36,800 --> 00:08:41,120

followed by the set of retrieved.

00:08:39,440 --> 00:08:42,560

relevant hopefully.

00:08:41,120 --> 00:08:45,839

context.

00:08:42,560 --> 00:08:46,880

and it uses all of that to generate an.

00:08:45,839 --> 00:08:49,519

answer.

00:08:46,880 --> 00:08:50,959

okay so we can see with this open book.

00:08:49,519 --> 00:08:52,800

format.

00:08:50,959 --> 00:08:55,200

we are passing a lot more information.

00:08:52,800 --> 00:08:57,920

into the generator which allows the.

00:08:55,200 --> 00:08:59,120

generator to answer more specific.

00:08:57,920 --> 00:09:00,000

questions.

00:08:59,120 --> 00:09:02,240

now.

00:09:00,000 --> 00:09:04,720

long-form question answering which is.

00:09:02,240 --> 00:09:06,720

what we are going to go through.

00:09:04,720 --> 00:09:10,160

is one form.

00:09:06,720 --> 00:09:11,120

of this abstractive question answering.

00:09:10,160 --> 00:09:13,440

the.

00:09:11,120 --> 00:09:14,880

only difference with or the the one.

00:09:13,440 --> 00:09:17,519

thing that makes long-form question.

00:09:14,880 --> 00:09:19,760

answering long-form question answering.

00:09:17,519 --> 00:09:20,959

is that the generator model has been.

00:09:19,760 --> 00:09:23,760

trained.

00:09:20,959 --> 00:09:27,040

to produce a multi-sentence.

00:09:23,760 --> 00:09:29,279

output so rather than just outputting a.

00:09:27,040 --> 00:09:31,920

maybe a answer of.

00:09:29,279 --> 00:09:33,600

three or four words or one sentence it.

00:09:31,920 --> 00:09:34,720

is going to try and output a full.

00:09:33,600 --> 00:09:35,519

paragraph.

00:09:34,720 --> 00:09:38,320

um.

00:09:35,519 --> 00:09:41,519

answer to you okay so that's long form.

00:09:38,320 --> 00:09:43,440

question answering or lfqa so.

00:09:41,519 --> 00:09:48,160

we are going to implement.

00:09:43,440 --> 00:09:51,360

lfqa in haystack and haystack is a very.

00:09:48,160 --> 00:09:53,760

popular nlp library.

00:09:51,360 --> 00:09:54,800

mainly for question answering.

00:09:53,760 --> 00:09:56,560

now to.

00:09:54,800 --> 00:09:58,399

install haystack and the other libraries.

00:09:56,560 --> 00:10:00,240

that we need.

00:09:58,399 --> 00:10:02,560

today we.

00:10:00,240 --> 00:10:05,040

do this so we have a pip insole.

00:10:02,560 --> 00:10:07,440

we need the pancake compliance farm.

00:10:05,040 --> 00:10:10,160

haystack specified pine cone in there.

00:10:07,440 --> 00:10:11,440

datasets and pandas.

00:10:10,160 --> 00:10:14,079

actually.

00:10:11,440 --> 00:10:16,480

i think you can ignore pandas.

00:10:14,079 --> 00:10:18,720

let's remove that.

00:10:16,480 --> 00:10:21,519

so just these three here.

00:10:18,720 --> 00:10:23,600

uh with farm haystack.

00:10:21,519 --> 00:10:25,120

we are going to be using.

00:10:23,600 --> 00:10:26,320

something called the pinecone document.

00:10:25,120 --> 00:10:27,519

store so.

00:10:26,320 --> 00:10:32,160

for that.

00:10:27,519 --> 00:10:34,959

you need either version 1.3 or above now.

00:10:32,160 --> 00:10:36,560

to initialize that pinecone document.

00:10:34,959 --> 00:10:38,720

source so remember the document store is.

00:10:36,560 --> 00:10:41,440

that thing that you saw on the right.

00:10:38,720 --> 00:10:43,040

before so where we're storing all of our.

00:10:41,440 --> 00:10:43,760

context vectors.

00:10:43,040 --> 00:10:44,800

we.

00:10:43,760 --> 00:10:47,760

will.

00:10:44,800 --> 00:10:50,240

do this so we first need an api key.

00:10:47,760 --> 00:10:52,399

from pinecone so there's a link here.

00:10:50,240 --> 00:10:54,800

i'll just open it and show you quickly.

00:10:52,399 --> 00:10:57,279

and that will bring you to this page.

00:10:54,800 --> 00:10:58,959

here now you can sign up for free you.

00:10:57,279 --> 00:11:00,240

don't have to you don't need to pay for.

00:10:58,959 --> 00:11:01,839

anything and we don't need to pay for.

00:11:00,240 --> 00:11:04,160

anything to do what we're doing here.

00:11:01,839 --> 00:11:05,920

either it's all completely free so you.

00:11:04,160 --> 00:11:09,120

just sign up.

00:11:05,920 --> 00:11:11,120

and once you've signed up you will see.

00:11:09,120 --> 00:11:12,880

it should just be one project on your.

00:11:11,120 --> 00:11:16,160

home page.

00:11:12,880 --> 00:11:18,160

so for me it is the default project.

00:11:16,160 --> 00:11:19,120

james's default project so you can go.

00:11:18,160 --> 00:11:20,240

into.

00:11:19,120 --> 00:11:21,600

that.

00:11:20,240 --> 00:11:24,480

and then on the left over here we have.

00:11:21,600 --> 00:11:25,360

an api key so we open that.

00:11:24,480 --> 00:11:26,480

and.

00:11:25,360 --> 00:11:28,800

we.

00:11:26,480 --> 00:11:30,560

get our default api key we can just copy.

00:11:28,800 --> 00:11:32,399

it so we come over here.

00:11:30,560 --> 00:11:33,600

and we use that.

00:11:32,399 --> 00:11:36,560

to.

00:11:33,600 --> 00:11:39,440

authenticate our pinecone document store.

00:11:36,560 --> 00:11:40,399

back in our code so i would paste that.

00:11:39,440 --> 00:11:41,360

here.

00:11:40,399 --> 00:11:43,440

and.

00:11:41,360 --> 00:11:45,600

with that we just run this so we are.

00:11:43,440 --> 00:11:48,000

initializing our document store we are.

00:11:45,600 --> 00:11:50,800

calling out index so.

00:11:48,000 --> 00:11:52,480

remember document saw is actually vector.

00:11:50,800 --> 00:11:55,600

database in this case.

00:11:52,480 --> 00:11:57,839

and inside that vector database we have.

00:11:55,600 --> 00:11:59,760

what's called an index the index is.

00:11:57,839 --> 00:12:02,639

basically the list of all the context.

00:11:59,760 --> 00:12:05,440

vectors that we have we call that index.

00:12:02,639 --> 00:12:07,440

haystack lfqa now you can call it.

00:12:05,440 --> 00:12:10,560

whatever you want and just.

00:12:07,440 --> 00:12:12,480

when you are wanting to load this.

00:12:10,560 --> 00:12:14,880

document store again.

00:12:12,480 --> 00:12:16,800

you need to specify the correct index.

00:12:14,880 --> 00:12:18,160

that's all that's all that's the only.

00:12:16,800 --> 00:12:20,160

difference it makes.

00:12:18,160 --> 00:12:23,839

similarity we're using cosine similarity.

00:12:20,160 --> 00:12:24,639

and we're using embedding dimensions 768.

00:12:23,839 --> 00:12:28,800

now.

00:12:24,639 --> 00:12:30,079

it's important to align this to whatever.

00:12:28,800 --> 00:12:32,000

the.

00:12:30,079 --> 00:12:34,480

similarity metric and embedding.

00:12:32,000 --> 00:12:38,240

dimension of your retrieval model is.

00:12:34,480 --> 00:12:42,079

in our case cosine and 768 these are.

00:12:38,240 --> 00:12:45,279

pretty typical retriever model.

00:12:42,079 --> 00:12:47,040

metrics and dimensionalities.

00:12:45,279 --> 00:12:50,480

now we can go down we can check our.

00:12:47,040 --> 00:12:52,240

metric type we can also see the number.

00:12:50,480 --> 00:12:53,360

of documents and the embeddings that we.

00:12:52,240 --> 00:12:54,320

have in there.

00:12:53,360 --> 00:12:56,639

now.

00:12:54,320 --> 00:12:59,360

we don't have any at the moment because.

00:12:56,639 --> 00:13:00,639

we haven't pushed anything to our.

00:12:59,360 --> 00:13:03,120

document store.

00:13:00,639 --> 00:13:05,120

we don't have any data so we need to get.

00:13:03,120 --> 00:13:08,560

some data.

00:13:05,120 --> 00:13:11,920

for that we are going to use hogging.

00:13:08,560 --> 00:13:13,200

phase data sets so over here.

00:13:11,920 --> 00:13:18,320

we're going to use.

00:13:13,200 --> 00:13:20,639

this data set here which is a set of.

00:13:18,320 --> 00:13:23,120

snippets from wikipedia.

00:13:20,639 --> 00:13:25,200

there are a lot of them in folder states.

00:13:23,120 --> 00:13:26,320

that is nine gigabytes.

00:13:25,200 --> 00:13:29,200

now.

00:13:26,320 --> 00:13:31,680

to avoid downloading this full data set.

00:13:29,200 --> 00:13:33,920

what we do is set streaming equal to.

00:13:31,680 --> 00:13:37,440

true and what this will do is allow us.

00:13:33,920 --> 00:13:39,440

to iteratively load one record at a time.

00:13:37,440 --> 00:13:41,279

from this data set.

00:13:39,440 --> 00:13:44,320

and we can check what we have inside.

00:13:41,279 --> 00:13:48,079

that data set by running this so next.

00:13:44,320 --> 00:13:50,560

and create a iterable from our data set.

00:13:48,079 --> 00:13:53,839

and we see this so.

00:13:50,560 --> 00:13:55,920

the main things to take note of here.

00:13:53,839 --> 00:13:58,399

are section title and passage text.

00:13:55,920 --> 00:14:02,480

passage check text is going to create.

00:13:58,399 --> 00:14:04,880

our context or that sort of document.

00:14:02,480 --> 00:14:06,240

and there are a couple other things so.

00:14:04,880 --> 00:14:08,079

history.

00:14:06,240 --> 00:14:10,000

is going to be what we are going to.

00:14:08,079 --> 00:14:12,000

filter for in our data set this is a.

00:14:10,000 --> 00:14:14,959

very big data set and i don't want to.

00:14:12,000 --> 00:14:17,440

process all of it so i'm restricting our.

00:14:14,959 --> 00:14:20,480

scope to just history and we're going to.

00:14:17,440 --> 00:14:22,320

only return a certain number of records.

00:14:20,480 --> 00:14:24,079

from that section.

00:14:22,320 --> 00:14:26,560

that's important to us purely for that.

00:14:24,079 --> 00:14:28,320

filtering out of other.

00:14:26,560 --> 00:14:32,000

sections or.

00:14:28,320 --> 00:14:35,040

um titles in there section titles.

00:14:32,000 --> 00:14:36,639

and we'll we will include article title.

00:14:35,040 --> 00:14:38,320

as metadata.

00:14:36,639 --> 00:14:39,360

in our documents although it's not.

00:14:38,320 --> 00:14:41,360

really important because we're not.

00:14:39,360 --> 00:14:42,880

actually going to use it it's just so.

00:14:41,360 --> 00:14:45,040

you can see how you would include.

00:14:42,880 --> 00:14:47,360

metadata in there.

00:14:45,040 --> 00:14:48,240

in case you did want to use it.

00:14:47,360 --> 00:14:50,480

so.

00:14:48,240 --> 00:14:52,000

here what we're doing is filtering only.

00:14:50,480 --> 00:14:55,120

for documents.

00:14:52,000 --> 00:14:56,639

that have the section title history.

00:14:55,120 --> 00:14:59,120

okay.

00:14:56,639 --> 00:15:01,199

and we just get this iterable object.

00:14:59,120 --> 00:15:03,519

because we're streaming so.

00:15:01,199 --> 00:15:04,959

it just knows now when we're streaming.

00:15:03,519 --> 00:15:06,800

one by one.

00:15:04,959 --> 00:15:07,839

when it's pulling an object it's going.

00:15:06,800 --> 00:15:10,720

to check.

00:15:07,839 --> 00:15:12,959

if that object section title starts with.

00:15:10,720 --> 00:15:14,639

history if it does it will it will pull.

00:15:12,959 --> 00:15:16,320

it if not it will move on to the next.

00:15:14,639 --> 00:15:18,399

one okay so we're just going to pull.

00:15:16,320 --> 00:15:20,880

those with history.

00:15:18,399 --> 00:15:24,959

now what we need to do is process those.

00:15:20,880 --> 00:15:27,519

and add them to our document store.

00:15:24,959 --> 00:15:29,600

now what i've done here is said okay we.

00:15:27,519 --> 00:15:32,480

are only going to.

00:15:29,600 --> 00:15:34,240

pull 50 000 of those and no more at that.

00:15:32,480 --> 00:15:35,920

point we cut off.

00:15:34,240 --> 00:15:37,680

and it's actually it cuts off just.

00:15:35,920 --> 00:15:39,360

before 50 000..

00:15:37,680 --> 00:15:42,320

and what we're going to do is we're.

00:15:39,360 --> 00:15:44,079

going to add in a single batch so we're.

00:15:42,320 --> 00:15:46,399

going to loop through all of we're going.

00:15:44,079 --> 00:15:48,560

to pull all of these records we're going.

00:15:46,399 --> 00:15:51,440

to collect 10 000 of them and then we're.

00:15:48,560 --> 00:15:52,959

going to add them to our document store.

00:15:51,440 --> 00:15:53,759

and.

00:15:52,959 --> 00:15:54,639

this.

00:15:53,759 --> 00:15:57,120

is.

00:15:54,639 --> 00:15:59,920

a haystack document object.

00:15:57,120 --> 00:16:02,240

so we have a content the content is the.

00:15:59,920 --> 00:16:03,519

document text that big paragraph you saw.

00:16:02,240 --> 00:16:06,079

before.

00:16:03,519 --> 00:16:08,560

meta is any metadata that we'd like to.

00:16:06,079 --> 00:16:10,160

add in there now with the pinecone.

00:16:08,560 --> 00:16:12,079

document store we can use metadata.

00:16:10,160 --> 00:16:13,519

filtering although i won't show you how.

00:16:12,079 --> 00:16:15,920

to do that here.

00:16:13,519 --> 00:16:18,480

but that can be really useful if it's.

00:16:15,920 --> 00:16:20,160

something you're interested in.

00:16:18,480 --> 00:16:22,880

so that's how you'd add metadata to your.

00:16:20,160 --> 00:16:25,199

document as well and all i'm doing is.

00:16:22,880 --> 00:16:26,000

create adding that doc.

00:16:25,199 --> 00:16:27,839

to.

00:16:26,000 --> 00:16:30,959

a dots list.

00:16:27,839 --> 00:16:33,440

and we increase the counter.

00:16:30,959 --> 00:16:36,800

and once the counter.

00:16:33,440 --> 00:16:39,680

hits the batch size which is the 10 000.

00:16:36,800 --> 00:16:40,639

we write those documents to our document.

00:16:39,680 --> 00:16:41,680

store.

00:16:40,639 --> 00:16:42,399

now.

00:16:41,680 --> 00:16:44,480

you.

00:16:42,399 --> 00:16:46,160

will remember i said document store is a.

00:16:44,480 --> 00:16:47,839

vector database and inside the vector.

00:16:46,160 --> 00:16:49,839

database we have vectors.

00:16:47,839 --> 00:16:51,199

at the moment when we write those.

00:16:49,839 --> 00:16:52,880

documents we're not actually creating.

00:16:51,199 --> 00:16:56,000

those vectors because we haven't.

00:16:52,880 --> 00:16:57,759

specified the retriever model yet.

00:16:56,000 --> 00:16:59,680

we're going to do that later so at the.

00:16:57,759 --> 00:17:02,880

moment what we're doing is kind of.

00:16:59,680 --> 00:17:04,400

adding the documents as just plain text.

00:17:02,880 --> 00:17:07,600

to.

00:17:04,400 --> 00:17:10,160

almost be ready to be processed into.

00:17:07,600 --> 00:17:12,319

vectors to put into that vector database.

00:17:10,160 --> 00:17:14,720

so it's almost like they're in limbo.

00:17:12,319 --> 00:17:18,240

waiting to be added um.

00:17:14,720 --> 00:17:20,400

to our to our database.

00:17:18,240 --> 00:17:22,079

so we add all of those that can take a.

00:17:20,400 --> 00:17:23,760

it can take a little bit of time not not.

00:17:22,079 --> 00:17:25,120

too long though.

00:17:23,760 --> 00:17:28,000

and then once we.

00:17:25,120 --> 00:17:30,559

hit or get close to 50 000 we.

00:17:28,000 --> 00:17:33,440

break so we stop loop and.

00:17:30,559 --> 00:17:36,799

then we can see if we get the document.

00:17:33,440 --> 00:17:38,640

count we see that we have the almost 50.

00:17:36,799 --> 00:17:40,559

000 documents in there but then when we.

00:17:38,640 --> 00:17:41,440

look at the embedding count.

00:17:40,559 --> 00:17:42,960

zero.

00:17:41,440 --> 00:17:45,200

and that's because.

00:17:42,960 --> 00:17:47,679

we you know they're waiting to be added.

00:17:45,200 --> 00:17:50,160

into the vector database those um the.

00:17:47,679 --> 00:17:51,600

text documents so they are existing.

00:17:50,160 --> 00:17:53,360

documents they just don't exist as.

00:17:51,600 --> 00:17:54,480

embeddings yet.

00:17:53,360 --> 00:17:56,240

okay so.

00:17:54,480 --> 00:18:00,320

what we now need to do.

00:17:56,240 --> 00:18:02,799

is convert those documents into vector.

00:18:00,320 --> 00:18:03,760

embeddings.

00:18:02,799 --> 00:18:06,960

now.

00:18:03,760 --> 00:18:08,480

to do that we need a retriever model.

00:18:06,960 --> 00:18:11,039

now.

00:18:08,480 --> 00:18:12,320

at this point it's probably best to.

00:18:11,039 --> 00:18:15,679

check.

00:18:12,320 --> 00:18:18,559

if you have a gpu that.

00:18:15,679 --> 00:18:20,720

is available like a cuda enabled gpu if.

00:18:18,559 --> 00:18:22,160

you don't this set will take longer.

00:18:20,720 --> 00:18:23,760

unfortunately.

00:18:22,160 --> 00:18:25,600

but if you do.

00:18:23,760 --> 00:18:26,840

that's great because this will be pretty.

00:18:25,600 --> 00:18:31,120

quick in most.

00:18:26,840 --> 00:18:32,440

cases depending on your gpu of course.

00:18:31,120 --> 00:18:34,080

so we.

00:18:32,440 --> 00:18:36,799

initialize.

00:18:34,080 --> 00:18:38,559

our retrieval model so we're using the.

00:18:36,799 --> 00:18:39,919

embedding retriever and this allows us.

00:18:38,559 --> 00:18:41,679

to use what are called sentence.

00:18:39,919 --> 00:18:45,200

transform models from the sentence.

00:18:41,679 --> 00:18:46,799

transformers library now i'm using this.

00:18:45,200 --> 00:18:48,400

model here.

00:18:46,799 --> 00:18:50,720

and we can find all the sentence.

00:18:48,400 --> 00:18:52,480

transform models over on the hooking.

00:18:50,720 --> 00:18:55,679

face model hub so let's have a quick.

00:18:52,480 --> 00:18:58,799

look at that so we are here.

00:18:55,679 --> 00:19:01,679

co models and i can paste.

00:18:58,799 --> 00:19:03,360

that model name.

00:19:01,679 --> 00:19:04,480

maybe i'll just do flight sentence.

00:19:03,360 --> 00:19:06,480

embeddings.

00:19:04,480 --> 00:19:07,919

now flight sentence embeddings are a set.

00:19:06,480 --> 00:19:10,559

of models that were.

00:19:07,919 --> 00:19:13,840

trained on a lot of data.

00:19:10,559 --> 00:19:15,840

using the flex library but there are a.

00:19:13,840 --> 00:19:17,840

lot of other sentence transform models.

00:19:15,840 --> 00:19:20,480

see the one we're using here.

00:19:17,840 --> 00:19:24,320

so for example if we go sentence.

00:19:20,480 --> 00:19:26,480

transformers you see all of the default.

00:19:24,320 --> 00:19:27,919

models used by the sentence transformers.

00:19:26,480 --> 00:19:31,919

library.

00:19:27,919 --> 00:19:33,440

so we are using this mpnet model.

00:19:31,919 --> 00:19:36,320

we also specify that we're using.

00:19:33,440 --> 00:19:38,640

sentence transformers model format and.

00:19:36,320 --> 00:19:40,720

when we initialize our retriever we also.

00:19:38,640 --> 00:19:43,840

need to add the document store that.

00:19:40,720 --> 00:19:45,520

we'll be retrieving documents from.

00:19:43,840 --> 00:19:46,960

so we've already initialized our.

00:19:45,520 --> 00:19:49,200

document source so we just add that in.

00:19:46,960 --> 00:19:49,200

there.

00:19:49,600 --> 00:19:54,480

and at this point it's time for us to.

00:19:53,360 --> 00:19:57,360

update.

00:19:54,480 --> 00:19:59,039

those embeddings so when we say update.

00:19:57,360 --> 00:20:02,159

embeddings what this is going to do is.

00:19:59,039 --> 00:20:03,200

look at any all the documents.

00:20:02,159 --> 00:20:05,840

that are.

00:20:03,200 --> 00:20:07,840

ready and with your document store and.

00:20:05,840 --> 00:20:11,520

it's going to use the retriever model.

00:20:07,840 --> 00:20:14,080

that you pass here and embed them into.

00:20:11,520 --> 00:20:16,640

vector representations of.

00:20:14,080 --> 00:20:20,080

those contents and then you can saw.

00:20:16,640 --> 00:20:21,600

those in your pinecone vector database.

00:20:20,080 --> 00:20:25,440

that will be processed.

00:20:21,600 --> 00:20:27,039

and at this point we could run.

00:20:25,440 --> 00:20:30,480

this get embedding count again and we.

00:20:27,039 --> 00:20:34,080

would get this 4995.

00:20:30,480 --> 00:20:36,080

value now another way that you can also.

00:20:34,080 --> 00:20:38,400

see this number is.

00:20:36,080 --> 00:20:39,760

if we go back.

00:20:38,400 --> 00:20:41,760

to.

00:20:39,760 --> 00:20:44,400

our pinecone dashboard.

00:20:41,760 --> 00:20:47,120

we can head over to our index so.

00:20:44,400 --> 00:20:49,360

haystack lfqa.

00:20:47,120 --> 00:20:52,159

we click on that.

00:20:49,360 --> 00:20:53,919

scroll down and we can click on index.

00:20:52,159 --> 00:20:56,159

info and then we can see the total.

00:20:53,919 --> 00:20:58,000

number of vectors which is the same okay.

00:20:56,159 --> 00:20:59,360

so that will that number will be.

00:20:58,000 --> 00:21:01,280

reflected.

00:20:59,360 --> 00:21:03,440

in your vector database once you have.

00:21:01,280 --> 00:21:05,120

updated the embeddings using a retriever.

00:21:03,440 --> 00:21:07,840

model.

00:21:05,120 --> 00:21:12,640

okay and at that point we can.

00:21:07,840 --> 00:21:14,480

just test the first part of our lfqa.

00:21:12,640 --> 00:21:16,720

pipeline which is just a document store.

00:21:14,480 --> 00:21:19,039

and a retriever so we.

00:21:16,720 --> 00:21:21,440

initialize this document search pipeline.

00:21:19,039 --> 00:21:23,280

with our retrieval model and we can ask.

00:21:21,440 --> 00:21:25,919

the question uh where is the first.

00:21:23,280 --> 00:21:29,440

electric power system belt okay and all.

00:21:25,919 --> 00:21:31,760

this is going to do is retrieve.

00:21:29,440 --> 00:21:33,840

the relevant context it's not going to.

00:21:31,760 --> 00:21:35,919

generate an answer yet it's just going.

00:21:33,840 --> 00:21:37,679

to retrieve what it thinks is the.

00:21:35,919 --> 00:21:38,640

relevant context.

00:21:37,679 --> 00:21:42,559

so.

00:21:38,640 --> 00:21:45,760

we have here electrical power system in.

00:21:42,559 --> 00:21:50,080

1881 two electricians built the world's.

00:21:45,760 --> 00:21:53,280

first power system in goldman in england.

00:21:50,080 --> 00:21:53,280

which is pretty good.

00:21:53,440 --> 00:21:58,320

so that's pretty cool.

00:21:55,200 --> 00:22:00,559

and what we now need to do is we have.

00:21:58,320 --> 00:22:02,480

our document store or vector database.

00:22:00,559 --> 00:22:04,559

and then we have our retriever model now.

00:22:02,480 --> 00:22:08,000

we need to initialize our generator.

00:22:04,559 --> 00:22:10,240

model to actually generate those answers.

00:22:08,000 --> 00:22:14,720

so we come down here we are going to be.

00:22:10,240 --> 00:22:14,720

using a sequence to sequence generator.

00:22:15,200 --> 00:22:20,240

and we are going to be using this model.

00:22:18,000 --> 00:22:22,080

here so this again you can find this on.

00:22:20,240 --> 00:22:23,919

the hooking face.

00:22:22,080 --> 00:22:24,880

model hub.

00:22:23,919 --> 00:22:27,440

and.

00:22:24,880 --> 00:22:30,240

there are different generator models you.

00:22:27,440 --> 00:22:32,000

can use here but you do want to find one.

00:22:30,240 --> 00:22:34,720

that has been trained for long-form.

00:22:32,000 --> 00:22:37,120

question answering so for example.

00:22:34,720 --> 00:22:39,600

we have the bar lfqa.

00:22:37,120 --> 00:22:40,960

that you can find here or you have the.

00:22:39,600 --> 00:22:44,400

bot.

00:22:40,960 --> 00:22:45,919

explain like m5 model that we can find.

00:22:44,400 --> 00:22:48,640

here.

00:22:45,919 --> 00:22:50,880

now i think the bar left ua model seems.

00:22:48,640 --> 00:22:53,600

to perform better so we have gone with.

00:22:50,880 --> 00:22:55,280

that also it's been trained with a newer.

00:22:53,600 --> 00:22:56,400

data set.

00:22:55,280 --> 00:22:58,720

and.

00:22:56,400 --> 00:23:00,559

yeah we just initialize it like that now.

00:22:58,720 --> 00:23:03,360

when we say sequence the sequence that's.

00:23:00,559 --> 00:23:05,520

because it is taking in a sequence of.

00:23:03,360 --> 00:23:07,600

characters or some some input it's going.

00:23:05,520 --> 00:23:10,080

to output a sequence of characters egb.

00:23:07,600 --> 00:23:13,600

output the answer.

00:23:10,080 --> 00:23:15,760

and if you are curious that the input.

00:23:13,600 --> 00:23:18,000

will look something like what you see.

00:23:15,760 --> 00:23:20,080

here okay so we have the question and.

00:23:18,000 --> 00:23:22,320

then we have the user's query.

00:23:20,080 --> 00:23:25,200

it's followed by context and then we.

00:23:22,320 --> 00:23:27,919

have this sort of p token here and that.

00:23:25,200 --> 00:23:29,760

p token um indicates in a model the.

00:23:27,919 --> 00:23:32,960

start of a new context that has been.

00:23:29,760 --> 00:23:35,360

retrieved from our document store.

00:23:32,960 --> 00:23:38,080

so in this case we've retrieved three.

00:23:35,360 --> 00:23:40,000

contexts and all of that is being passed.

00:23:38,080 --> 00:23:41,919

to the generator model.

00:23:40,000 --> 00:23:44,640

where it will then.

00:23:41,919 --> 00:23:46,880

generate an answer based on all of that.

00:23:44,640 --> 00:23:49,120

okay so.

00:23:46,880 --> 00:23:50,480

yeah that we just initialized up the.

00:23:49,120 --> 00:23:52,159

generator model.

00:23:50,480 --> 00:23:54,400

and then we.

00:23:52,159 --> 00:23:56,799

initialize the generative q and a.

00:23:54,400 --> 00:23:58,480

pipeline we pass in the generator and.

00:23:56,799 --> 00:24:00,080

the retrieval model we don't need to.

00:23:58,480 --> 00:24:02,640

include document saw here because the.

00:24:00,080 --> 00:24:04,320

document store has already been passed.

00:24:02,640 --> 00:24:05,840

to the retriever model when we're.

00:24:04,320 --> 00:24:08,559

initializing that so it's almost like.

00:24:05,840 --> 00:24:09,760

it's embedded within the retriever.

00:24:08,559 --> 00:24:11,440

so we don't need to worry about adding.

00:24:09,760 --> 00:24:13,039

that in there and then we can begin.

00:24:11,440 --> 00:24:15,919

asking questions and this is where it.

00:24:13,039 --> 00:24:18,720

starts to get i think more interesting.

00:24:15,919 --> 00:24:20,480

now one thing to make note of here is we.

00:24:18,720 --> 00:24:22,880

have this top k parameter and that's.

00:24:20,480 --> 00:24:25,200

just saying how many.

00:24:22,880 --> 00:24:26,720

uh contexts to retrieve in in the.

00:24:25,200 --> 00:24:28,320

context of our.

00:24:26,720 --> 00:24:30,840

retriever model and then for the.

00:24:28,320 --> 00:24:32,880

generator how many answers to.

00:24:30,840 --> 00:24:34,240

generate so in this case we're.

00:24:32,880 --> 00:24:36,480

retrieving.

00:24:34,240 --> 00:24:39,200

three contacts and then we are.

00:24:36,480 --> 00:24:41,360

generating one answer based on the query.

00:24:39,200 --> 00:24:43,600

and those three contexts like we saw in.

00:24:41,360 --> 00:24:45,440

the example.

00:24:43,600 --> 00:24:47,840

so.

00:24:45,440 --> 00:24:50,559

in this i'm asking what is a wall.

00:24:47,840 --> 00:24:53,360

of currents it's good to be specific um.

00:24:50,559 --> 00:24:54,240

to test this and if we have.

00:24:53,360 --> 00:24:55,279

the.

00:24:54,240 --> 00:24:57,840

data.

00:24:55,279 --> 00:24:59,520

within our data set it seems to be.

00:24:57,840 --> 00:25:03,200

pretty good at pulling that out and.

00:24:59,520 --> 00:25:05,760

producing a relatively accurate answer.

00:25:03,200 --> 00:25:08,159

so the war occurrence was a rivalry.

00:25:05,760 --> 00:25:10,320

between thomas edison and george.

00:25:08,159 --> 00:25:12,720

westinghouse's companies over which.

00:25:10,320 --> 00:25:14,720

former transmission dc or ac was.

00:25:12,720 --> 00:25:16,159

superior okay.

00:25:14,720 --> 00:25:17,840

that's the answer.

00:25:16,159 --> 00:25:19,600

which is pretty cool and we can see what.

00:25:17,840 --> 00:25:21,520

it's pulled that from so.

00:25:19,600 --> 00:25:24,240

it's pulled it from.

00:25:21,520 --> 00:25:26,559

this content.

00:25:24,240 --> 00:25:28,960

this content.

00:25:26,559 --> 00:25:29,840

and this content okay so there were.

00:25:28,960 --> 00:25:31,919

three.

00:25:29,840 --> 00:25:36,000

um parts that got.

00:25:31,919 --> 00:25:37,200

fed into the into the model and that's.

00:25:36,000 --> 00:25:38,640

that's good we can see a lot of.

00:25:37,200 --> 00:25:40,720

information there but maybe we can see a.

00:25:38,640 --> 00:25:43,679

little bit too much information so we.

00:25:40,720 --> 00:25:45,919

can actually use the print answers.

00:25:43,679 --> 00:25:48,320

utility to.

00:25:45,919 --> 00:25:49,200

minimize more outputting them and here.

00:25:48,320 --> 00:25:50,720

we get.

00:25:49,200 --> 00:25:52,720

just this which is obviously a lot.

00:25:50,720 --> 00:25:55,279

easier to read so we just pass our.

00:25:52,720 --> 00:25:57,520

result into print answers and specify.

00:25:55,279 --> 00:26:00,240

details of minimum the rest of that is.

00:25:57,520 --> 00:26:04,480

the same as what we asked before.

00:26:00,240 --> 00:26:05,360

okay so it's much more readable now.

00:26:04,480 --> 00:26:07,600

one.

00:26:05,360 --> 00:26:09,919

thing to point out here is.

00:26:07,600 --> 00:26:13,039

that this is actually a very good answer.

00:26:09,919 --> 00:26:14,880

but maybe there's not that much detail.

00:26:13,039 --> 00:26:17,840

now if we find that we're not getting.

00:26:14,880 --> 00:26:20,559

much detail in our answers or that the.

00:26:17,840 --> 00:26:22,559

answer is just wrong.

00:26:20,559 --> 00:26:23,360

what the issue might be.

00:26:22,559 --> 00:26:24,640

is.

00:26:23,360 --> 00:26:28,000

first.

00:26:24,640 --> 00:26:30,480

the retrieved context.

00:26:28,000 --> 00:26:32,799

may not contain any relevant information.

00:26:30,480 --> 00:26:34,559

for the model to actually.

00:26:32,799 --> 00:26:37,039

view and answer the.

00:26:34,559 --> 00:26:39,520

question correctly okay so it's not.

00:26:37,039 --> 00:26:42,240

retrieving relevant information from.

00:26:39,520 --> 00:26:43,679

that external um sort of open book.

00:26:42,240 --> 00:26:46,080

document source.

00:26:43,679 --> 00:26:47,840

and the second is.

00:26:46,080 --> 00:26:49,840

if it's not also not retrieving.

00:26:47,840 --> 00:26:52,320

information from there and it's also not.

00:26:49,840 --> 00:26:53,840

retrieving information from you remember.

00:26:52,320 --> 00:26:57,760

i mentioned that these models can have a.

00:26:53,840 --> 00:26:59,760

memory if it's not able to find any.

00:26:57,760 --> 00:27:02,640

relevant information within its memory.

00:26:59,760 --> 00:27:04,320

for your particular query.

00:27:02,640 --> 00:27:05,679

if both of those.

00:27:04,320 --> 00:27:07,279

conditions.

00:27:05,679 --> 00:27:08,720

are.

00:27:07,279 --> 00:27:10,480

not satisfied so we don't have.

00:27:08,720 --> 00:27:12,159

information or relevant information.

00:27:10,480 --> 00:27:13,600

coming from the external source and we.

00:27:12,159 --> 00:27:15,919

don't have relevant information coming.

00:27:13,600 --> 00:27:17,679

from the model memory the generator is.

00:27:15,919 --> 00:27:20,000

going to output.

00:27:17,679 --> 00:27:21,360

usually something nonsensical.

00:27:20,000 --> 00:27:24,799

okay.

00:27:21,360 --> 00:27:26,960

so in this scenario we have two options.

00:27:24,799 --> 00:27:28,080

really the generator model we can.

00:27:26,960 --> 00:27:30,559

increase.

00:27:28,080 --> 00:27:32,320

its size so we can use a larger.

00:27:30,559 --> 00:27:34,399

generator model because larger generator.

00:27:32,320 --> 00:27:37,200

models have more.

00:27:34,399 --> 00:27:39,200

model parameters which means they have.

00:27:37,200 --> 00:27:41,200

basically more memory that they've.

00:27:39,200 --> 00:27:43,120

learned during training.

00:27:41,200 --> 00:27:44,559

or.

00:27:43,120 --> 00:27:46,880

we can.

00:27:44,559 --> 00:27:50,240

increase the amount of data that we are.

00:27:46,880 --> 00:27:53,120

pulling from the document store okay so.

00:27:50,240 --> 00:27:55,279

if we are just returning three.

00:27:53,120 --> 00:27:57,600

documents or contacts.

00:27:55,279 --> 00:28:00,240

we can increase it to 10 because then.

00:27:57,600 --> 00:28:02,799

the generator is being fed a lot more.

00:28:00,240 --> 00:28:03,919

information and it might be that the.

00:28:02,799 --> 00:28:05,039

correct.

00:28:03,919 --> 00:28:07,919

um.

00:28:05,039 --> 00:28:11,200

information that we need may come in.

00:28:07,919 --> 00:28:12,399

maybe context five or context six and.

00:28:11,200 --> 00:28:14,159

nine.

00:28:12,399 --> 00:28:16,320

and the generator will see that and be.

00:28:14,159 --> 00:28:19,679

like okay that's the answer i'm going to.

00:28:16,320 --> 00:28:21,200

you know reformat this into a into my.

00:28:19,679 --> 00:28:24,159

answer.

00:28:21,200 --> 00:28:25,520

okay so we can try that here now we.

00:28:24,159 --> 00:28:28,480

already got a good answer but we can.

00:28:25,520 --> 00:28:31,440

just see what we get if we increase the.

00:28:28,480 --> 00:28:33,840

retriever so audio retrieved number of.

00:28:31,440 --> 00:28:36,080

documents so it increased up to 10 and.

00:28:33,840 --> 00:28:37,840

we see that we get this much longer.

00:28:36,080 --> 00:28:39,600

chunk of text now.

00:28:37,840 --> 00:28:41,360

and.

00:28:39,600 --> 00:28:43,360

i think the first half of this is.

00:28:41,360 --> 00:28:44,799

relatively accurate.

00:28:43,360 --> 00:28:46,440

so we have.

00:28:44,799 --> 00:28:49,679

uh this in.

00:28:46,440 --> 00:28:51,360

1891 first power system it was installed.

00:28:49,679 --> 00:28:52,880

in the united states i think that's.

00:28:51,360 --> 00:28:55,600

relatively.

00:28:52,880 --> 00:28:57,520

relatively correct um.

00:28:55,600 --> 00:28:59,200

and then it starts to get a little bit.

00:28:57,520 --> 00:29:00,480

silly after that.

00:28:59,200 --> 00:29:04,480

because.

00:29:00,480 --> 00:29:06,559

you know we've pulled more contacts.

00:29:04,480 --> 00:29:09,039

from our documents store.

00:29:06,559 --> 00:29:10,320

but with that we have pulled in more.

00:29:09,039 --> 00:29:12,399

irrelevant.

00:29:10,320 --> 00:29:14,080

information because we're retrieving 10.

00:29:12,399 --> 00:29:16,480

now so there's a good chance that the.

00:29:14,080 --> 00:29:18,080

last few of those are not relevant so.

00:29:16,480 --> 00:29:20,320

we're feeding a lot of irrelevant.

00:29:18,080 --> 00:29:22,000

information into our generator model and.

00:29:20,320 --> 00:29:24,799

so it starts to get confused and then it.

00:29:22,000 --> 00:29:27,200

can start to ramble uh like we.

00:29:24,799 --> 00:29:29,039

like we see here.

00:29:27,200 --> 00:29:32,159

so.

00:29:29,039 --> 00:29:34,399

that's what we see happening um.

00:29:32,159 --> 00:29:35,600

another thing i want to point out is.

00:29:34,399 --> 00:29:38,159

that the.

00:29:35,600 --> 00:29:39,360

generator has this memory.

00:29:38,159 --> 00:29:40,559

so.

00:29:39,360 --> 00:29:42,559

a lot of people.

00:29:40,559 --> 00:29:44,000

always think when they hear okay the.

00:29:42,559 --> 00:29:46,000

generator has memory does that mean i.

00:29:44,000 --> 00:29:47,840

don't need the document store because we.

00:29:46,000 --> 00:29:49,919

we have this memory can't just fine-tune.

00:29:47,840 --> 00:29:53,039

the model so that it knows everything.

00:29:49,919 --> 00:29:55,200

within my particular use case.

00:29:53,039 --> 00:29:56,320

in some cases yes you might be able to.

00:29:55,200 --> 00:29:57,760

do that.

00:29:56,320 --> 00:30:00,240

but it.

00:29:57,760 --> 00:30:03,279

generally only works for more.

00:30:00,240 --> 00:30:05,039

general uh questions or general.

00:30:03,279 --> 00:30:07,279

knowledge if you start to get specific.

00:30:05,039 --> 00:30:09,279

it tends to fail with that sort of.

00:30:07,279 --> 00:30:11,840

memory part because the memory can only.

00:30:09,279 --> 00:30:12,640

source so much information.

00:30:11,840 --> 00:30:14,799

and.

00:30:12,640 --> 00:30:17,360

in the end what you will probably need.

00:30:14,799 --> 00:30:19,279

is you want a model with good memory so.

00:30:17,360 --> 00:30:20,240

it can kind of maybe pull out some facts.

00:30:19,279 --> 00:30:22,000

from there.

00:30:20,240 --> 00:30:24,880

but for anything specific it's probably.

00:30:22,000 --> 00:30:25,919

going to need to refer to its document.

00:30:24,880 --> 00:30:28,159

store.

00:30:25,919 --> 00:30:29,440

so what we have done here is we've asked.

00:30:28,159 --> 00:30:31,440

the same question.

00:30:29,440 --> 00:30:34,159

but this time i've replaced the retrieve.

00:30:31,440 --> 00:30:35,760

documents with just nothing and.

00:30:34,159 --> 00:30:36,880

we can see the result of that straight.

00:30:35,760 --> 00:30:38,480

away so.

00:30:36,880 --> 00:30:39,520

the answer is i'm not sure what you mean.

00:30:38,480 --> 00:30:41,120

by war.

00:30:39,520 --> 00:30:43,279

so it's.

00:30:41,120 --> 00:30:45,440

it has no idea what the war of currents.

00:30:43,279 --> 00:30:47,039

is it doesn't have that information.

00:30:45,440 --> 00:30:49,679

within its memory.

00:30:47,039 --> 00:30:50,799

so without that external document.

00:30:49,679 --> 00:30:51,760

source.

00:30:50,799 --> 00:30:53,679

it.

00:30:51,760 --> 00:30:54,720

it doesn't know what to say it's just.

00:30:53,679 --> 00:30:57,039

okay.

00:30:54,720 --> 00:30:58,799

i don't even know what war is.

00:30:57,039 --> 00:31:01,200

but like i said.

00:30:58,799 --> 00:31:03,279

in some cases particularly when you're.

00:31:01,200 --> 00:31:04,880

asking more general knowledge query it.

00:31:03,279 --> 00:31:07,039

will be able to pull that out from its.

00:31:04,880 --> 00:31:09,919

memory so.

00:31:07,039 --> 00:31:12,240

who is the first person on the moon.

00:31:09,919 --> 00:31:14,240

it it knows this because it's it's such.

00:31:12,240 --> 00:31:16,640

a common thing to know it's probably.

00:31:14,240 --> 00:31:18,159

seen it in the training data that the.

00:31:16,640 --> 00:31:20,320

model has been trained on a million.

00:31:18,159 --> 00:31:22,399

times maybe not a million but a few.

00:31:20,320 --> 00:31:23,279

times at least.

00:31:22,399 --> 00:31:25,679

so.

00:31:23,279 --> 00:31:28,320

that is the first pers man to walk on.

00:31:25,679 --> 00:31:30,799

the moon was neil armstrong okay.

00:31:28,320 --> 00:31:32,880

cool so.

00:31:30,799 --> 00:31:34,880

i think that's pretty much it we can ask.

00:31:32,880 --> 00:31:36,960

a few more questions uh when was the.

00:31:34,880 --> 00:31:38,720

first electrical power system built so.

00:31:36,960 --> 00:31:40,960

we asked his name to start and it will.

00:31:38,720 --> 00:31:42,640

give us this answer.

00:31:40,960 --> 00:31:44,720

um.

00:31:42,640 --> 00:31:46,720

if we want to confirm.

00:31:44,720 --> 00:31:48,080

that this is correct so this is what i.

00:31:46,720 --> 00:31:49,519

did with this i was a bit confused.

00:31:48,080 --> 00:31:51,519

because google is telling me something.

00:31:49,519 --> 00:31:55,200

different.

00:31:51,519 --> 00:31:56,720

you can print out the contents.

00:31:55,200 --> 00:31:59,919

using this so we loop through the.

00:31:56,720 --> 00:32:01,519

results documents and we just print dot.

00:31:59,919 --> 00:32:03,279

content.

00:32:01,519 --> 00:32:05,600

and this okay so.

00:32:03,279 --> 00:32:07,840

two electricians built first uh power.

00:32:05,600 --> 00:32:09,679

system at gold damming in england so.

00:32:07,840 --> 00:32:12,000

that information is actually coming from.

00:32:09,679 --> 00:32:13,120

somewhere it's not just making it up.

00:32:12,000 --> 00:32:15,039

so.

00:32:13,120 --> 00:32:17,279

that can be really useful another thing.

00:32:15,039 --> 00:32:19,160

uh just to be aware of with generators.

00:32:17,279 --> 00:32:20,559

is that they can.

00:32:19,160 --> 00:32:21,760

[Music].

00:32:20,559 --> 00:32:24,480

generate.

00:32:21,760 --> 00:32:26,240

misleading uh information.

00:32:24,480 --> 00:32:28,640

um so you need to be careful with that.

00:32:26,240 --> 00:32:31,039

so for example in this one i asked where.

00:32:28,640 --> 00:32:32,880

did covert 19 originate now this is.

00:32:31,039 --> 00:32:34,720

pretty unfair because the generator.

00:32:32,880 --> 00:32:36,399

probably hasn't seen anything about.

00:32:34,720 --> 00:32:37,519

curving 19.

00:32:36,399 --> 00:32:39,840

and.

00:32:37,519 --> 00:32:41,760

at the same time.

00:32:39,840 --> 00:32:45,120

it doesn't have any covert 19.

00:32:41,760 --> 00:32:47,039

information within its uh document store.

00:32:45,120 --> 00:32:48,320

because we looked at history not not.

00:32:47,039 --> 00:32:51,760

anything else.

00:32:48,320 --> 00:32:54,480

um so it just says covered 19 isn't a.

00:32:51,760 --> 00:32:57,039

virus which it is it's a bacterium.

00:32:54,480 --> 00:32:58,559

okay so straight away it's pretty pretty.

00:32:57,039 --> 00:33:00,080

wrong.

00:32:58,559 --> 00:33:03,600

so.

00:33:00,080 --> 00:33:05,519

it's just one example of where.

00:33:03,600 --> 00:33:06,880

you need to just be cautious with this.

00:33:05,519 --> 00:33:08,480

sort of thing because it can just give.

00:33:06,880 --> 00:33:10,799

completely wrong.

00:33:08,480 --> 00:33:13,840

answers if it doesn't have the relevant.

00:33:10,799 --> 00:33:15,519

information available to it so with that.

00:33:13,840 --> 00:33:17,519

uh there's a couple of things you could.

00:33:15,519 --> 00:33:19,440

do to mitigate that you can.

00:33:17,519 --> 00:33:20,880

one just include the sources of.

00:33:19,440 --> 00:33:22,399

information if you.

00:33:20,880 --> 00:33:24,240

if you build some sort of search.

00:33:22,399 --> 00:33:26,559

interface make sure you include those so.

00:33:24,240 --> 00:33:29,200

users can look at that and see where.

00:33:26,559 --> 00:33:30,720

this information is coming from.

00:33:29,200 --> 00:33:32,159

and two.

00:33:30,720 --> 00:33:34,399

there are.

00:33:32,159 --> 00:33:37,519

sort of confidence scores that are.

00:33:34,399 --> 00:33:40,240

given to these answers so you could put.

00:33:37,519 --> 00:33:43,600

a threshold like you say anything be.

00:33:40,240 --> 00:33:46,080

below 0.2 confidence we just don't show.

00:33:43,600 --> 00:33:48,320

or we show um.

00:33:46,080 --> 00:33:50,960

i'm not confident in this answer.

00:33:48,320 --> 00:33:53,120

but it might be this or something along.

00:33:50,960 --> 00:33:54,399

those lines.

00:33:53,120 --> 00:33:57,360

okay.

00:33:54,399 --> 00:33:59,200

so that's just one drawback um let's.

00:33:57,360 --> 00:34:01,120

we'll just go through a few final.

00:33:59,200 --> 00:34:03,360

questions so what was nasa's most.

00:34:01,120 --> 00:34:05,039

expensive project i would say the space.

00:34:03,360 --> 00:34:07,600

shuttle project.

00:34:05,039 --> 00:34:09,280

um that's correct tell me something.

00:34:07,600 --> 00:34:12,000

interesting about history of the earth.

00:34:09,280 --> 00:34:14,159

in this case it really it's nothing it's.

00:34:12,000 --> 00:34:15,280

not already history i don't think.

00:34:14,159 --> 00:34:17,440

um.

00:34:15,280 --> 00:34:19,359

but it does give us an interesting fact.

00:34:17,440 --> 00:34:21,040

about the magnetic field being weak.

00:34:19,359 --> 00:34:22,560

compared to the rest of the solar system.

00:34:21,040 --> 00:34:24,000

i don't know if that's true or not it.

00:34:22,560 --> 00:34:25,760

seems like it.

00:34:24,000 --> 00:34:27,359

might not be.

00:34:25,760 --> 00:34:28,800

when it says compared to the rest of so.

00:34:27,359 --> 00:34:31,599

this is someone thinking is it weak.

00:34:28,800 --> 00:34:34,000

compared to mars i don't think so so.

00:34:31,599 --> 00:34:35,599

um that might not be true another thing.

00:34:34,000 --> 00:34:38,000

to be wary of.

00:34:35,599 --> 00:34:40,639

who created the nobel prize and why.

00:34:38,000 --> 00:34:44,079

so this one is correct and i think quite.

00:34:40,639 --> 00:34:45,760

interesting and how is no prize funded.

00:34:44,079 --> 00:34:47,520

uh we kind of see it down here so i know.

00:34:45,760 --> 00:34:49,679

the information is in there hence why.

00:34:47,520 --> 00:34:52,000

i've asked the question.

00:34:49,679 --> 00:34:54,800

and it it tells you that as well with a.

00:34:52,000 --> 00:34:56,480

little bit more information.

00:34:54,800 --> 00:34:58,480

so.

00:34:56,480 --> 00:34:59,280

that is it.

00:34:58,480 --> 00:35:02,640

for.

00:34:59,280 --> 00:35:04,960

long form question answering with haysak.

00:35:02,640 --> 00:35:07,040

as i said at the start i think.

00:35:04,960 --> 00:35:11,040

question answering is one of the most.

00:35:07,040 --> 00:35:13,920

widely applicable forms of nlp or use.

00:35:11,040 --> 00:35:16,640

cases of nlp.

00:35:13,920 --> 00:35:19,359

it it can be applied almost everywhere.

00:35:16,640 --> 00:35:21,119

so it's a really good one to just sort.

00:35:19,359 --> 00:35:25,040

of go away and.

00:35:21,119 --> 00:35:28,240

see you know maybe i can implement like.

00:35:25,040 --> 00:35:29,280

document search in my organization or i.

00:35:28,240 --> 00:35:31,520

can.

00:35:29,280 --> 00:35:35,839

create some sort of internal search.

00:35:31,520 --> 00:35:38,079

engine that helps people in some way.

00:35:35,839 --> 00:35:39,760

and i think in a lot of organizations.

00:35:38,079 --> 00:35:42,560

it's very.

00:35:39,760 --> 00:35:45,520

possible to do this and add a lot of.

00:35:42,560 --> 00:35:46,320

benefit and reduce a lot of friction.

00:35:45,520 --> 00:35:49,280

in.

00:35:46,320 --> 00:35:51,440

uh day-to-day processes.

00:35:49,280 --> 00:35:54,400

of most companies so.

00:35:51,440 --> 00:35:55,680

that's it for this video i hope it's.

00:35:54,400 --> 00:36:00,160

been useful.

00:35:55,680 --> 00:36:00,160

and i will see you in the next one.