Python and AI Power-Up Program Offline Class-20250826_113227-Meeting Recording

August 26, 2025, 6:02AM

1h 43m 34s

Hardip Patel started transcription

Hardip Patel 0:06

No, no.

Did you say?

I would.

Tirth 0:23

Are you waiting on anyone 12345611?

7.

Hmm.

Hardip Patel 0:31

OK, I think everyone is there now. Yeah, all right.

Tarun Jain 0:34

Oh, always.

Tirth 0:35

Yeah.

Hardip Patel 0:39

Yes.

Tarun Jain 0:39

Oh, OK, I'll share my screen.

Hardip Patel 0:43

Oh, I'm really.

Thanks.

Tarun Jain 0:48
So am I audible?


Hardip Patel 0:51
Yes. Can you do something?

Tarun Jain 0:52
Hello.
So I'm I'm audible, right?


Hardip Patel 0:58
Sir, I'm audible. Yeah.

Tarun Jain 1:02
OK, but my voice is getting repeated for some reason.
OK, now it's fine I guess.
Hello.


Tirth 1:12
Yeah.


Hardip Patel 1:12
Yeah, we can hear.


RamKrishna Bhatt 1:14
We can hear you.

Tarun Jain 1:16
OK, fine, fine. So probably what we will do is we'll just cover some of the essentials data center data science packages which we missed out in the last week, which was Numpy Pandas.
And matplotlib. So I'll just tell you what will be the usage of all three once one more time, and then we'll dive dive into how you can use all these libraries. So first we have Numpy. So Numpy is nothing but numerical Python.

And this is mainly used whenever we have to do any kind of data creation process, right? It can be in terms of creating some random vectors or if you want to play around with vectors, right? So in most of the cases, vectors are nothing but arrays, right? So if you want to.
Use any arrays. We will be using numpy. So can anyone tell me the difference between list and numpy?
I mean direct.
We actually covered this one of one example.
OK.

Tirth 2:36
Does it have something to relate with the data type that it stores or no?

TJ

Tarun Jain 2:40
Yeah, so when it comes to list it is mixed. In the sense you have 10, you have some other value and then you have true. So this is mixed data, right? So this is considered as list. But when it comes to array you only have to save the similar kind of data type which is 10/20/30.

Tirth 2:42
Right.
Babe.
Best right?

TJ

Tarun Jain 2:59
40 and so on. OK, so this is one difference. If you want to create and play around with the arrays that we have, we'll usually prefer numpy and most of the time when it comes to embeddings and.

Tirth 3:01
Right.

TJ

Tarun Jain 3:16
The token IDs. This will be in array.
So even though token ID is a tensor, but at the end of the day you can use the array

manipulation. Whatever methods are there right in array, you can also do that with tensors, right? Because at the end of the day tensors are nothing but arrays.
When you look at their data types and then we have pandas.
So pandas is basically used for data manipulation.
Let's suppose you want to perform some kind of filtering or grouping technique.
On CSV file, CSV or Excel.
During that time you will be working with Pandas. Now why is this important? Let's suppose you want to feed your data to our LLM, right? So basically whatever data you have here, right? This data, if it is CSV, most of the time chunking won't.
Work right? How do we usually pass the data to a level, at least in form of chunks?
But when your data is in CSV, let's suppose you have queries like what was the monthly sale?
Off.
April 2024 and you have some more queries like what is the average?
Off.
Average sales of last five to six months. Now if you look at this kind of queries, you can't handle it with just chunking process. You feed your CSV to vector database.
Vector database will save it and then fetch it. Most of the time what will happen is you will only have some information.
But you will need some kind of data manipulation to get this context. Now what is this kind of prompt? This is nothing but you filter. You filter based on the month data that you have, right? So if your data is in CSV and if you have time.
Period based queries during that time. Having understanding of pandas is very important, right? So using pandas.
You can.
Build anything related to. You can build anything related to data analysis.

Tirth 5:45
M.

Tarun Jain 5:46
Now what is this data analysis? You upload your CSV file, you ask any time period based question. You can also create graphs, graphs and plots right? And some of them is pie chart and then you have line chart right? So all these examples is something that you can create with pandas.

Now when it comes to visualization, this is where we'll be using matplotlib.
So these are the three libraries that are must that one needs to learn. Numpy, which is mainly for array creation and playing around with the numerical digits, and pandas is basically for data manipulation every time you're dealing with CSV and Excel and if you want to pass this.
Tabular data you can either call tabular.
Or structured.
Or you can sell time period based.
Right time period is also nothing but structured data.
Because when it comes to tabular structure and time period, you don't need vector databases. You can directly pass your CSV to LLM and you can build a pandas agent.
So now what this agent will do is every time you give a CSV file.
The the job of the agent is to run pandas command.
So I'll repeat what is happening. So you upload a CSV file. So the first step is upload a CSV file. Let me check if the tab is active.
Probably it is in sleep mode, but I'll show the demo of how data analysis agent looks like. But usually what will happen is your CSV file and what agent will do is it will not look into your data set. Based on your question it will run a pandas command.
And this pandas command is nothing but a code. Once it generates a code, that code is running in a interpreter. Code is running in a interpreter. So what is interpreter? The collab.
So if you look at the collab, right after every single line, it is getting you the output, right? So what is this line doing now? Let's suppose I do import numpy as NP. So once you have any Python command, it will run in a particular execution statement, right? And then it will give you the result. This is what you're also trying to do here.
Your interpreter is nothing but a tool. Once you give a CSV file, you ask any question, your interpreter will run a code and then once you get the response right, once you get the response from the interpreter.
This will go as a context to the LLM. So the result of the result from the interpreter is the context to the LLM and then LLM will generate the response.
Generate the final response.
Now this is only for the approach that you take from structured responses. Here this will not matter, right? So if you have anything related to this kind of questions which is monthly sales, it will give you the code. That code is executing in interpreter and then LLM will generate the response.

And you don't need any vector database to perform this task and it's sometime kind of time consuming, but it will give you better results compared to chunking process because it already has the actual data. Anyone has any question why we are using pandas or matplotlib?
1.

Tirth 9:18
Not as of now.

Tarun Jain 9:21
I'll show one demo and then probably we will start with the library usage.
OK, so here you're able to see your screen, right? Let me make this in a white theme.
Are you able to see this?

Tirth 9:44
Yes.

Ishan Chavda 9:45
Yes.

Tarun Jain 9:46
OK, so here if you see it is only accepting CSV data, I will upload ACSV data.
And this is the same data set which we have looked earlier, which is IMDb. If you see you have rank, you have year, you have rating, you have genre, you have certificate, you have runtime, you have budget and you have box office. So now these are the columns and.
It is structured data. Here I'll ask the question. The question is.
What is the?
Give me the movie name.

Tirth 10:25
With highest runtime.

Tarun Jain 10:26
With the with the highest budget.

Tirth 10:29
OK, but then.

Tarun Jain 10:31
Budget and runtime.

Tirth 10:33
OK.

Tarun Jain 10:42
So now if you see you have the movie with the is budget which is the professional and it's runtime is one hour 50 minutes. So it has considered a single movie, right? So here probably the question should be which movie.
Louise.
Have the Ayush.
Budget and.
The runtime.
Separately.
Do we have the Leon here?
OK, so now if you see the movie, Princess Mononoke has a budget of this particular thing and the longest runtime is Gone with the Wind, which is of three hours 58 minutes.

Tirth 11:43
But then previously it gave wrong answer.

Tarun Jain 11:46
So here what did it was it combined which has the IS budget and the runtime. So it was looking for one option with both the what you call possibilities.

Tirth 11:58
So in the previous one, did it just you know per minute? Did it calculate per minute basis and then it gives the Leon the professional?

Because if we see the highest budget, it is Princess.
Mononoke.

Tarun Jain 12:12
Uh, what is the budget of?
Movie.
So it did and operation here. Here what it did was it looked for specific whichever is the highest it gave you that name.
OK, why is it none?

Tirth 12:41
Maybe it's hallucinating. It's not calling actually the.

Tarun Jain 12:44
So here it is doing keyword search. So if you see here I'm giving the exact names that I'm supposed to give. In this particular app you're not targeting the particular columns.

Tirth 12:50
Mhm.
OK.

Tarun Jain 12:55
Give me the plot for the unique.
So whatever you see here, it's using pandas to run this particular comments. Now here it should use matplotlib. Matplotlib basically will give you the plot.

Tirth 13:18
OK.

Tarun Jain 13:25
Why is it not giving the plot?

Tirth 13:32
Yeah.

Tarun Jain 13:33
Count or OK, it's bar plotted. Should it be bar plot or is it?

Tirth 13:42
So.

Tarun Jain 13:44
OK, we are counting it. For counting it should be probably histogram.
Wait, it says CPT.
All the unique plot.
Which is the right graph to use in matplotlib?
Habar plot.
So here it is using matplotlib.

Tirth 14:29
OK.

Tarun Jain 14:29
So now what we will try to do is we will try to see how to use these libraries 1st and then how once we start learning agent we will have something similar to this. But here we are using very smaller LLM model when it comes to.
Matlotlib It's not just for graph sometimes. Let's suppose you want to create report right report generation.

Tirth 14:54
Hmm.

Tarun Jain 14:54
For example, let's suppose you have logs of Cloudflare, right? And you need to create summaries like what is the unique IPS and other details, other important details. So during this time Matplotlib will also give you a entire report and this report it will help you in.
What are the different columns and what is the distribution of it, right? So if you want to get the report and summary in the distribution, Matplotlib will also help you

in generating that reports. So this is very important, all these three libraries, Numpy, Pandas and Matplotlib. So we'll start with Numpy.
Uh, do you have this notebook where we covered NLTK?
I'll reshare this notebook.
So.
So we'll start with numpy, then we'll cover pandas. And once pandas is done, we have matplotlib, right? And what is the best exercise to do after this? Let's suppose you complete pandas, you complete numpy and you complete matplotlib.

Tirth 16:08
Mhm.

TJ **Tarun Jain** 16:12
What's next?
So here basically we have to do something called as EDA. Anyone remembers what EDA is?
Have we have we covered this keyword before?

Tirth 16:40
No, I yeah, I'm. I'm not sure if I missed in some of the sessions, yeah.

TJ **Tarun Jain** 16:41
Hello.
OK, so basically EDA is nothing but exploratory data analysis. So what usually happens in exploratory data analysis is you are given a data, you want to see if the data has any outliers or not.

Tirth 16:51
OK.

TJ **Tarun Jain** 17:02
Outliers or empty data.

Tirth 17:03
M.

Tarun Jain 17:08

If you remember when we covered recommendation app rate recommendation app, there were two commands that we used. One is drop NA.
Drop and one more was fill in it. Do you remember these two comments?
So I'll open the recommendation up.
OK, so if you look at this recommendation app, we had something called as fill a fill in a command.
Where is that film?
This one is drop. Now what will drop do? Let's suppose you have total 10 columns.
Out of the 10 columns you're removing one column and when you are removing one column you have to give axis equals to one. So this column was not necessary, right?
You use that based on your analysis, right? Just like that you will have multiple.
Columns. Let's suppose you have 5 columns.
Let's take an real time example. So the real time example is the Titanic. So what were the features that we need to predict whether the given person will survive or not?
So here what we are supposed to do is you have a data set and the goal is to predict.

Tirth 18:41

EBay.
We here.

Tarun Jain 18:48

Whether the given person will survive or not.

Tirth 18:51

We had the gender.

Tarun Jain 18:53

So you have gender.

Tirth 18:54

We had the age.

Tarun Jain 18:56

Hello.

H then class.

Tirth 18:59

Uh, the deck. Yeah, so the glass.

Tarun Jain 19:03

Nick.

Tirth 19:04

And there was one more. What was it?

Nobody else remembers gender, age, class, ****.

Tarun Jain 19:22

Uh, we can also have current position, right? Like if the person is already uh.

Tirth 19:28

Near the boat.

Tarun Jain 19:29

Then probably will be survived. So here if you see you have total 5 columns that is already available to you. Now what EDA will do is let's suppose you want to perform EDA. You want to perform some kind of analysis like what is the distribution of data.

Tirth 19:37

Mhm.

Tarun Jain 19:46

What is the?

Tirth 19:48

Uh, there is an echo. Yeah, yeah, it is getting echoed from your side.

Tarun Jain 19:48

My voice is getting.

Hardip Patel 19:50
OK, thanks I can.

Tirth 20:18
Mr.

Mitesh Rathod 20:19
That's it.

Tarun Jain 20:19
Good column and sorry, not gender will be male or female, right?
And let's suppose there are few entries which is empty. So you have male, female and you have NAN. So NAN is nothing but not a number. So if you see in our data analysis also when I asked what is the budget of Leon the professional, it showed NAN.

Tirth 20:37
Hmm.

Tarun Jain 20:45
That means it's a null value, right? So now when you do EDA, you try to understand what is happening in your entire columns and what is the distribution of it. And if some distribution doesn't make sense, you have to cut that down, right? And in our case, if there is this particular column called gender.

Tirth 20:48
Hmm.

Tarun Jain 21:05
And you have male, female and NN. So do you think this will matter this rose?
No, right. So what you will do, you will just drop them.
And then what is the next column? You have age. Now age distribution is from 18 to 60 and if in case again if you find any what you call invalid invalid age, let's suppose. Some people, when they do data entry for this particular data set, you find uneven

data like you have -3 or you have more than 100. So basically this will cause an issue when you build a model. So if there is any distribution right if it is outside certain quartiles, so we have different.
And quartile, we have first quartile which is at 25%. Then you have second quartile which is at 50%. And in simple words, this is nothing but median, right? And then you have third quartile which is nothing but 75%.
So this is threshold. If it is less than 50%, that is nothing, but there is a middle value which is your median and the starting 25% of the data. Let's suppose you have the distribution 18 to 618 to 60.
So what is the 50% of this? What will be the median?

Tirth 22:36
We have it into 16 so.

Tarun Jain 22:37
18 60 / 2.
So 78 divided by two.

Tirth 22:45
Mm.

Hardip Patel 22:45
Uh.
3039.

Tarun Jain 22:53
39 That means your median is 39. So just like that if you divide this into half somewhere, not half.
Somewhere your first quarter will be around 24 or 25.
And here your 75 quartile will be somewhere around 49 or 51.
So here what we're trying to do is we're just trying to understand the distribution, what kind of data we have. You're just playing around with the data. You try to visualize the data, what is the unique values in it. So once you arrive at that particular solution, it will give you some kind of understanding.
Whether this column is required or not, if it is not required you can drop it. So that

decision you only take once you perform EDA, right? So here we will work on two different EDA. So one is we will work on video game analysis which is we will check what is the video game sales.
And then you will have one unknown data which will be given to you for an assignment, right? So now when I do video game sales, I will play around with different columns, try to understand if that column is making sense or not. If that column doesn't make sense either I will use that column.
Create a new column. If not, I will just drop that column. Is this clear?

Tirth 24:16
No.

TJ **Tarun Jain** 24:17
So this will be one assignment that we will do when we take the break on 27th or 28th and then what we will do is we will directly jump into Langchen and Lama index.
So once again I'm telling you every time we encounter CSV or Excel file we will not be working on drag.
Because RAG can't be performed on CSV and Excel data, but if you want to build an agent using your CSV data, that is when you can build something called as Panda's agent, right? So this is very important when we work with RAG and agent.
So in short, you will never perform brag on structured data.
Not performed on structure data.

Tirth 24:55
Mm.

TJ **Tarun Jain** 25:02
So the best option for structured data is text to SQL.
Right, because what is SQL? SQL is used for structured data. RAG is mainly used for unstructured data like PDFs, PPT and then images, videos, audios. That is where you can use RAG, but when it comes to structured data.

Tirth 25:19
Um.

Tarun Jain 25:23
Which is in tabular format. It's better to use text to SQL or you can use text to pandas.
Is this clear why we are learning this?

Tirth 25:34
Mhm.

Tarun Jain 25:35
EDI is very important. In simple words, if anyone is still confused in what EDI is, you can just write in simple terms data for some job.
Like understand what is your data.
OK, so let's begin with the Numpy. I hope the notebook is available.

Tirth 25:59
Yes, it is.

Tarun Jain 26:00
OK, so we will start off with a simple command which is the import statement and I hope this particular keyword everyone knows. So import numpy as is nothing but alias. Every single time when I have to refer to numpy I will be using NP right? NP is nothing but shop.
Short form of numpy and this can be any name. You can also keep this as just N right? Next time if you want to refer to numpy instead of NP it should be just N dot arrange or whatever functions you want to use.
Oh, it's trying to connect. Just one minute.
OK, why is this taking time?
Is collab working for you guys? Are you able to run?

Tirth 26:59
I'm running in on VS code, so I'm not sure.

Hardip Patel 27:03
Yes.

Tarun Jain 27:04
OK, now it's connected. Now it's connected. So now if I do NP dot a range, it will give me an error. Why? Because NP is not defined, it's a name error. So now if I use N, that means I'm using a function that belongs to numpy.
But it's better to use what you call very notable meaningful digits, meaningful characters. When it means numpy, N is nothing but starting numerical and then P is nothing but Python. So we usually use numpy as NP whenever we call it in input. Statement. Now I'll start with NP. So what will range do?

Hardip Patel 27:50
Um, it will, uh, start with the below, make a list of numbers and from.

Tarun Jain 27:51
So let's suppose if I give zero to 10.
Zero to 9, no. So range will never give you the list. So if I do range of zero to 10, it is a function of range. This is sequence right? Since this is sequence, you can convert this into list.

Tirth 28:00
Today.

Hardip Patel 28:14
OK, OK.

Tarun Jain 28:16
Right and here also in numpy if you use this particular command A and range it will give you an array right now if you print the type of.
Here and it will be numpy's object and we saw this earlier as well when we were working with embeddings.

Tirth 28:34
Mhm.

Tarun Jain 28:37

So you have numpy N dimension array.
OK, and now if you see, let's try to understand the difference on how to how to check the.

Tirth 28:42
Hmm, OK.

Tarun Jain 28:51
Length.
Of the array, right? So there are two ways. One probably you might need to count.
The number of elements.
Within the array.
And one more can be to count the total dimensions of an array.
So you have numpy dot array and then you start with range one to 10. And now if you print this either you can use a range or you can use array. So this particular command is the right. This is similar to list of range.

Tirth 29:41
Hmm.

Tarun Jain 29:44
Zero to 10. Here it's like you're doing type conversion right here. What you're trying to do is you're directly using that particular command which is numpy dot a range and this is type conversion. Now what I'm trying to do is I'm printing 1 to 10 which will basically print 1 to 9. So the length of an array is also same and then.
Size is also same, but if I convert my 1D to 2D that is when it will be different. So I'll show you the difference now.
So if I print array.
What is this?
It's in one dimension, correct?

Tirth 30:22
Hmm hmm.

Tarun Jain 30:23

It starts with one and it ends with 9. But now if I run the same thing here if you see this is my first dimension and then this is my second dimension. Now if I run this. Total How many elements do I have within this array?
Elements. Elements is 18 and how much of length do I have? The length is 2 and length is nothing but the number of dimension. Either you can do N dim.

Tirth 30:44
18.
Cool.

TJ

Tarun Jain 31:02
Why is it giving 3?
OK, this is 1 dimension and then you have two dimension and this is 3 dimension.

Tirth 31:19
Why is it 3 dimension?

TJ

Tarun Jain 31:20
So if you see here here this is 1 column and then here you have one more column two and then there is three. There are total 3 square bracket.

Tirth 31:33
In the in the above one you have.

Hardip Patel 31:33
So in the last one in the NP dot array we have added a range 110 under already one, yeah.

Tirth 31:43
Right, right.

TJ

Tarun Jain 31:49
Now this is 2 DRA. If you see this is 1 and this is 2.

Tirth 32:01
M.

Tarun Jain 32:02
So if I have to do visual representation, let's suppose.
I'll come to xcalidra. You have 123. This is 1D array.

Tirth 32:14
One day.

Tarun Jain 32:21
Now this is 2D array. So 2D array will be like what is the rank of this?

Tirth 32:30
The length of this is.

Tarun Jain 32:31
Ranking the sun shave.

Ishan Chavda 32:33
Cool again.

Tarun Jain 32:35
Row cross column. So how many rows do I have? Two cross three. OK, so now if I want to convert this into a 3D array, I just have to put one more X this thing.

Hardip Patel 32:35
2.

Tirth 32:36
3.
Yeah, 2 comma 322 rules.

Hardip Patel 32:43
Oh, sorry.

Tirth 32:44
3.

Tarun Jain 32:52
What do you call square bracket and just close this square bracket. So now that we have three it's it is nothing but AN dimension. So how do we calculate the shape of this? So you just have to calculate the total.
Number of.
2D pairs you have then row and column of that particular 2D. So how many total number of 2D pairs do you have? It's just one, right?

Tirth 33:25
Mm.

Tarun Jain 33:26
So it will be 1 cross.
Two cross 3. But if I have like this, let's suppose I have 5.
6/7.
Then I have 8-9 and 10.
Now how many total number of 2D pairs I have? This is 1.

Hardip Patel 33:53
Right.

Tarun Jain 33:54
It starts here, it ends here. This is 1 2D pair and then this is 2nd 2D pair and then this is for the.

Hardip Patel 33:54
04.
OK.

Tirth 33:59
Mm.

Tarun Jain 34:05
This last one is for the 3D, so this will be.
Two cross 2 cross three. Why? Because both the 2D pairs that I have, it only has two rows and it has three columns.

Tirth 34:14
Mm.

Tarun Jain 34:20
Is this clear?

Tirth 34:20
Hmm.
Yes.

Tarun Jain 34:25
So here I just have two square brackets.
So let's try with that example.
Array for 2D123.
Then 456. Now if I do print of array 2D dot NDM.

Tirth 34:47
No, you have to do with N pine, yeah.

Tarun Jain 34:52
So this is list. List doesn't have end function.

Tirth 34:55
Mhm.

Tarun Jain 34:56

It is 2 and then if I try one more square bracket outside this, it should be 3.
But if I want to print the shape of it array 2D dot shape it is 1 comma 2 comma three.
But if I remove this brackets it will be just two comma 3.

Tirth 35:18
Mm.

Tarun Jain 35:19
And outside this I'm creating one more 2D which is 567 then 8910. What will be the shape of this?

Tirth 35:34
Mm.

Tarun Jain 35:35
I hope this is clear, right? You just have to count the number of 2D pairs you have within your 3D array. So how many pairs do I have? This is my one pair.

Tirth 35:37
It.
Mm.

Tarun Jain 35:47
I'll just bring this to the new line.

Tirth 35:50
What happens if you remove 10? Just 10 like it take the Max one.

Tarun Jain 35:53
It is another way.
Because usually it should match.

Tirth 36:00
That's her.

Tarun Jain 36:01
So whenever it comes to rank matrix, your rows and columns should match with your previous segment.


Tirth 36:09
M.
OK.

Tarun Jain 36:13
Now because let's suppose if you have 2 + 3.


Tirth 36:18
Hmm.

Tarun Jain 36:18
Just a 2D array which is 123234. But if you just remove this, there is no rank for your second row.


Tirth 36:22
Mhm.
Hmm.

Tarun Jain 36:29
Because it's on value error, you have a missing number.
So far, what did we cover? We covered.
NP dot array.
And for calculating the number of shapes, what is the function?


Tirth 36:54
And dim.

Tarun Jain 36:55
N dot N dim and then if you want to calculate the total number of elements that you have, let's suppose I have how many elements do I have here 3/6.

Hardip Patel 37:05
Size.

Tirth 37:05
Face.

Tarun Jain 37:09
912 So if I do size it should be 12.

Tirth 37:09
Tell me.

Tarun Jain 37:17
Now where is size useful? Let's suppose when you create visualization right. In visualization you need 3 images in one row, 3 images in one row and you have total 4 columns.

Tirth 37:23
Uh.

Tarun Jain 37:35
So it will be like image, image, image, image.
If you want to visualize all these things right the image. So when we come to matplotlib you will see how to add 4 to 5 images. During that time having the shape knowledge is very important because you can also create the transpose of it which we will come in a bit.
OK, so now the next command is what is the size of a particular array, the memory consumption. So let's suppose you want to calculate.
The memory usage of the given array.
Let's suppose you have the array. You have year range one.
And now if I print it just one single value which starts with 0. And now if I do this size array of size and then you have something called as array dot item size right? So this item size is for the every element that you have. Now if I print this you are getting 8. This eight is nothing but bits. Sorry, 8 bytes.

And if you want to directly print this, you have one more function called N bytes.
So this is for the memory consumption. Let's suppose I give one comma 10.
Now what will be the output of this?

Tirth 39:20
OK.

Ishan Chavda 39:21
80.

Tarun Jain 39:24
Why 72? I'm giving 10 year.

Hardip Patel 39:27
9.
Mm.

Tarun Jain 39:28
If I give 11.
Now it is.

Tirth 39:34
1 to 10.

Tarun Jain 39:36
Uh, it is 1010 into 880.

Tirth 39:38
Uh.

Tarun Jain 39:41
And if I directly print this, it is 80. Is this clear?

Hardip Patel 39:47
Yes.

Tirth 39:48

Yeah.

TJ **Tarun Jain** 39:48

So first was how do we create numpy and then endu and size is for the total number of elements. And if you want to calculate the size or memory, what is the memory of the given array? Either you can multiply with item size, if not you can directly print N bytes.

And now this we have already covered how to create 2D. It has to be same number of rows and columns.

And now this is very important.

Seed every single time whenever you are working with numpy.

Whenever you are working with numpy or.

Dodge.

Or anything related to ogging face.

If you're training a model, you have to define a seed value. So this seed value is nothing but it will make sure your data is uniform.

So there are usually just three values that usually people give. Either seed value will be 0 or it will be 42 or it will be 123, right? And the standard one is 42.

And you can just search for it like random seed value ideal.

Random seed ideal value.

Open.

Usually you will have the common choices include zero or two because it's basically based on this Hitchhiker's Guide to Galaxy. Again, there is no specific reason why 42 was picked. It was barely on some assumption that was made, right? If not, people also choose between 123.

Hardip Patel 41:47

And it goes unset of the universe, I guess.

TJ **Tarun Jain** 41:51

Oh, what?

Hardip Patel 41:52

It is in Hijacker's Guide to Galaxy. It is known as the answer to the universe, so it is the most unique number 42.
Just a side fact.

Tarun Jain 42:05
Uh, maybe?
But if there is any choice or reason to pick 42, there are some reasons, but in in general there is no practical usage of it. Just because it started to work, many people just use 42 as a default value.
So we I also looked this when we covered random. So just make sure whenever you are training your model right during training process.
So during training process, multiple calculations will be done.
And now what are these calculations? It's nothing but NP dot between your weights and the actual data that you have. There will be multiple matrix multiplication. So when you do this, multiple arrays will be created since all the arrays that you create.
Initially the weights will be random.
In neural network.
The initial.
Weights are random after training.
The weights are updated.
Now what are these weights? When we covered open source and closed source, I said you every single LL model. If they expose their model on hugging phase, that means it is open weights, right? And if you have those open weights when you start training initially.
It is random and once you start with the training process, these weights are updated and during this process multiple calculation happens, right? And there might be chances you need to have uniform vectors that you are creating. So every time you have training process, just make sure you define your torch.
Numpy seed value to be 42 and most of the times if you are working in any Kaggle competition, right? Let's suppose after this curriculum you are joining some Kaggle competition.
You might see the starting file lines of code will have seed value.
Why? Because when you're training, you will have some set of weights. When you're doing inference, you need some kind of uniformity, right? So this is the reason why in most of the competition of Kaggle you will see the starting 5 lines of code remain

same, which is.
They set random seed of numpy to be 42 and you will also have torch. You will also have random and if you're using CUDA for CUDA also you will define your seed value to be 42. So is this clear what seed value is? It's just to have some uniformity whenever you are working with vectors.
So the command is numpy dot random dot seed 42 and let's create some random vectors. So far what we did was we used range and once we define range we give the starting sequence and we give ending sequence.
And then we convert it into an array.
Now what this will do is it will print the given range which starts with 0 and 10 and we know the sequence. What if I need to randomly generate a sequence right? So during this time you have a function called NP dot random and then you have a function called rand. So what this will do is it will print.
The random 10 digits. So if you see this particular function as random values in the given shape and now if I give 10 how many size of elements will it create?
But.
Size. What will be the size of it? It will be 10. If I give 129 here, what is the size of this?

Hardip Patel 45:53
And.

Tirth 45:53
10.
9129 is 9.

Hardip Patel 46:00
Date.

Tarun Jain 46:01
Correct. So here I'm I'm just giving 10. It will start with 0 and it will end till 10.

Hardip Patel 46:03
Hey.

Tarun Jain 46:09
But if you see the values is random.

Tirth 46:09
M.

Tarun Jain 46:14
And you never know what values we'll get again.

Tirth 46:19
Hmm.

Tarun Jain 46:21
But once you define N array and when you once you run this N array it is done. But if you rerun this particular cell it will be different values.

Tirth 46:30
But why is it always less than one? Like is that something expected or it can be greater than one as well?

Tarun Jain 46:36
No, this will be 0 to one as per the documentation. If you see the uniform distribution is it starts with 0 and here if you see it's a parenthesis, that means one is excluded.

Tirth 46:49
OK, OK.

Tarun Jain 46:50
Can you see this distribution? Here it is square, here it is parenthesis. That means one is excluded.

Tirth 46:53

Mm.
OK.

Tarun Jain 46:58
So it's like 0 to 0.99.

Tirth 47:00
Open.

Tarun Jain 47:08
And what is the N dim?

Tirth 47:13
And.

Tarun Jain 47:14
One. OK, now that we have covered how to create random strings, usually most of the common functions will be reshaped.
Now let me show you one of the formula WPX.
Y equals two.
Does anyone know transpose function linear regression?
Yeah.
Yeah, it's here. Does anyone know what is transpose?
It's a simple math. Uh, we had this in our mathematics as well.
Can anyone recall what is this T?

Tirth 48:17
So you know if you have one row, three column, it converts to three rows, one column.

Tarun Jain 48:23
Correct. So let's suppose I have this particular array itself 123234. So what is the shape of it? It is 2 + 3 and this is my array. Now if I do transpose of.

Tirth 48:32
Two by three, it becomes 3 cross two, yeah.

Tarun Jain 48:39
This array it will be 3 cross two. It's like you're swapping your rows into columns.

Tirth 48:46
Hmm.

Tarun Jain 48:47
Right, so this function is also very important in most of the calculations that we do, and it's not just with this formula that I'm showing even in some of the search techniques that we use, right? One of them is for neural sparse calculation. We have transpose.
Obviously we will not write the mathematics of it because we have frameworks that will handle it. But I'm just telling you most of the formulas that we have, you'll also have transpose arrays. So transpose is nothing, but if you have any original array, you're just swapping off it. So now let's suppose I'm multiplying this.
Array into trans array, which is the transfers of array. Can anyone just tell me what will be the rank if I multiply these two?

Tirth 49:37
2 cross 2.

Tarun Jain 49:39
It will be 2 cross 2. So this is 3 cross 2 cross three. This is 3 cross two. So this should be same right? This should be same and then your final output will be 2 cross 2.

Tirth 49:41
But.
Mm mm.
Hey.

Tarun Jain 49:55
Now what kind of multiplication is this? This is simple matrix multiplication.


Tirth 49:59
Yeah.

Tarun Jain 50:04
OK, so here what I'm trying to do is I have N array. Now N array is nothing but a single dimension.
Since it already has ten size, what can I do with this ten size? Either I can create 2 cross three, sorry 2 cross 5 or I can create 5 cross two. So what does this mean?


Tirth 50:45
And it will be 10.
With them.

Tarun Jain 50:48
It will be 10. It's the same thing right? So I can use reshape, but if I do 4 cross three it will give me an error because when I do reshape the size of whatever I have it should match.


Tirth 50:49
Hey.

Tarun Jain 51:02
Now here the size is 12 which is not an right way to reshape right? So here if you see N array has total 10 elements and if I do dot reshape whatever combinations that you try here the rows and columns combination.
It should have total same size of an array.


Tirth 51:29
M.

Tarun Jain 51:30

Now 2 comma 5 is nothing but two into five which is 10. So now if I do the shape of it, it's 2 comma 5 and if I print the DIM end DIM.

Tirth 51:44
Cool.

Tarun Jain 51:44
It is true. Why? Because now I have.
Two rows with five column each.

Tirth 51:54
Mm.

Tarun Jain 51:55
So the function is reshape.
And if I want to do the transpose of it.
OK.
I'll just do NRR dot T.
How many rows do I have now?

Tirth 52:10
5 rows, 2 columns.

Tarun Jain 52:12
Pyros and Oopalams.

52:13
OK.

Tarun Jain 52:17
And the function for transpose is just T.
Whatever array you have dot D.
You can try with a few more combinations. So now if you see I have 5 comma two.
Why? Because I did my existing array has only 10 elements, so even if I try the combination of five comma two it will work. But if I do 4 comma 3.

They should throw an error.
Because you cannot reshape an array of size 10 into a shape 4 comma 3.
Till transpose. Is it clear?

Tirth 52:59
Yes.

TJ **Tarun Jain** 53:00
So it's very simple when it comes to numpy. The only thing what you need to know is the simple matrix multiplication. I'll write what are the important functions that one must need to know. One is the dot operation. Now what is the dot operation?
NP dot dot Where do we use this?

Hardip Patel 53:18
No.

TJ **Tarun Jain** 53:21
In which formula did we encounter this?
Anyone because the uh technique algorithm where we used NP dot.
Cos of E star B divided by magnitude of A into magnitude of B. So which formula was this?

53:55
We'll see.

Tirth 53:55
This was for the frequency calculation for in vocab. We did this.

TJ **Tarun Jain** 54:03
Pairwise similarity.

Tirth 54:04
Pairwise similarity. Yeah, yeah, pairwise similarity, right? Right.

TJ **Tarun Jain** 54:08

Right now one is dot and after dot the seed value is very important only when you're training, only when you're training the model and then random is also very important.
Creating random shapes and then transpose.
And in general, most of the array manipulations. Let's suppose you want to do any type conversion into array. During that time we use Numpy because we don't have Numpy as a default data type in Python. If you want to play with arrays, you have to use Numpy, right? So if you want to.
Type cached into array. Basically you have to use number itself.
And here if you see N array of shape double equals to N array of dim this will always be same. So what is this N array dot shape?
This is nothing but five comma two. Now if I do the length of this.

Tirth 55:19
Is too.

TJ

Tarun Jain 55:20
This is nothing but NDM.
Let's suppose I have a three dimension. I have one cross 2 cross 3, so the shape is 1 cross 2 cross 3 and if I print the length of it.
It is 3 which is nothing but your NDM. So either you can print your N dimension using this logic. If not you can use N whatever array is there dot NDM.

Tirth 55:49
Mhm.

TJ

Tarun Jain 55:49
So let's see what will be the shape of this thing. OK, since the result is already here, what I will do is I'll change this a bit.
OK, so can anyone tell me the shape of this array?

Tirth 56:33
One comma 2 comma 41 cross 2 cross 4.

TJ

Tarun Jain 56:37

Uh, look at it again.
How many 2D arrays do I have?

Hardip Patel 56:44
1220

Tirth 56:45
02 you have to to.

Hardip Patel 56:47
One to two.

Tarun Jain 56:48
So cross.

Tirth 56:49
2 cross 2 cross 4.
Cross four. Yeah. Not one. Cross two. Cross four. 2 cross 2 cross four. Yeah, yeah.

Hardip Patel 56:56
Oh.

Tarun Jain 56:59
Is this clear how 3D dimension will be calculated?

Tirth 57:03
Yes, yes, yes.

Tarun Jain 57:04
So where is this useful? Can anyone recall? We showed this example when we covered tuple.
So let's suppose I have this here, but instead of two cross 2 cross four I have 512 cross 512 cross three. So which example is this?

Ishan Chavda 57:30
It is something related to image.

Tarun Jain 57:32
Image right? So now how will I decouple it? I will have width, I will have height and I will have channel equals to 512 cross 512 plus three. So arrays what is image? Image is nothing but pixels right? If I upload an image.

Hardip Patel 57:39
Hi.
You know.

Tarun Jain 57:52
For us, we are able to see OK, you have RGB channel, it is colorful image. But once you upload this image into Python framework, everything will be converted into pixels, right? So can anyone just recall, I mean this is basic math, what is the range of pixel values?
Probably if you have done CSS right, you might have seen the range of uh values.

Ishan Chavda 58:15
I need to.

Tarun Jain 58:16
So what is the range of colors?

Tirth 58:19
255255255 so zero to 255.

Ishan Chavda 58:22
You look.

Tarun Jain 58:24
02255 What is 0?

Tirth 58:25
Hey.
Uh, for black.

Hardip Patel 58:28
Length 255 by.

Tarun Jain 58:29
Black and 255.

Tirth 58:30
255 white.

Tarun Jain 58:32
Right, so whenever you're dealing with images, right, whether it is just for a simple image manipulation or using multimodal drag where you want to upload your images, feed that in a vector database and then do similarity.

Tirth 58:34
Yeah.

Tarun Jain 58:47
So you will be working with images. So the image pixels that you have, it will be in this shape right? Either it will be fight well, if not it will be very high like 1080 then. So here what will be the width? It will be 1080. What is the height? It is 1920. Channel will be same. Channel will be 3 only for images.

Hardip Patel 59:10
What is channel?

Tarun Jain 59:12
Uh, channel is nothing but RGB.

Tirth 59:15
OK.

Hardip Patel 59:18
OK.

Tarun Jain 59:20
So if you have something like 255 cross zero comma zero, it's nothing but red.

Tirth 59:29
Hmm.

Tarun Jain 59:31
And if you have something like 0 comma 255 comma zero, it's nothing but green.

Tirth 59:40
Yeah.

Tarun Jain 59:40
And if you have like this thing, it is blue. But if you have some values like where blue is dominated and let's suppose you have very less red 123 and here it is something like 185 one second.

Hardip Patel 59:43
You.

Tarun Jain 1:00:09
So here if you see if you have any shape like this right 123 comma 0185 that means you will have any color which blue is dominated but with some mix of red right? So this is what will be there when you will be working with image.

Tirth 1:00:19
Mm.

Tarun Jain 1:00:24
And the shape of images will be width, height and channel which will be in 3D array. Is this clear?


Tirth 1:00:36
Mhm.

Tarun Jain 1:00:38
OK, so let's try to also understand some of the mathematical.


Hardip Patel 1:00:40
But.

Tarun Jain 1:00:45
Operations that we can do mathematical operations using numpy. So far whatever we saw was just manipulation of or creating the arrays and what are the different functions that you have within that arrays. So if you see you define an array and then you're using the methods or.
Attributes of it. But what are the direct mathematical operations that you can directly use from numpy? So here I have an example of marks which is numpy array and you have a list. So now if I convert this list into an array, what is the data type of this?


Tirth 1:01:23
NN array NB array.

Tarun Jain 1:01:26
Will be numpy dot ND array.


Tirth 1:01:30
NBA.

Tarun Jain 1:01:31
Which is N dimension area.
So there are other mathematical operations like NP dot mean. So one that we

covered earlier was NP dot dot right? And here you give vector one and then here you give vector 2 and based on this what it will do element wise element it will multiply.
And then it will create the summation of it, right? It's like summation of.
Multiplication between element one comma element 2.
So here you have one more bracket.
Then you again have to do plus.
Multiply between.
Element one. So this should be element one of vector one, but everyone knows right what auto operator is.

Hardip Patel 1:02:33
Yes.

Tirth 1:02:34
The lot is multiplication.

Tarun Jain 1:02:35
OK.
dot is multiplication but with summation.

Tirth 1:02:39
OK.

Tarun Jain 1:02:42
So summation in the sense, let's suppose you have 123.
Then you have 456. What is the NP dot of this?
First we will do one into four, which is 4 plus this is 10.
And this is 18.

Tirth 1:03:07
It.

Tarun Jain 1:03:09
414 How much is this is 32?

Tirth 1:03:14
Yeah.

Tarun Jain 1:03:14
So this is your dot operator. First you multiply element wise and then you do multiple summation which is addition.
And now what we will do is we'll look at other commands. So you have a list now. Sorry, you have an array of marks. You can also check the mean of it. So what is the mean? You just have to create the total of all the marks and divided it by 1234567.
So do the sum of this and divide it by 7.
Sum of marks.
Divided by length of marks.
It's the same thing.

Tirth 1:03:59
Yeah.

Tarun Jain 1:04:00
P dot mean we also have median. Can anyone tell me the median of this?
What is the first rule of median?

Tirth 1:04:11
Min and Mac.

Tarun Jain 1:04:14
OK, you will take min and Max.
So in order to know min and Max, what are we supposed to do first?

Tirth 1:04:21
We'll sort it.

Tarun Jain 1:04:23
First will be sort. So when I sort what array will I get? It will be 65.

Tirth 1:04:31
65.

Tarun Jain 1:04:32
66.

Ishan Chavda 1:04:33
It.

Tirth 1:04:34
70, 78, 80.

Tarun Jain 1:04:34
But then I have 7.

RamKrishna Bhatt 1:04:35
Thank you.

Tarun Jain 1:04:39
And it's E 88 and 91 and 94.
1234567 So this three are here, this three is here. Eight is our median.
Everyone understood. So let's suppose I add two more here. So here I'll add 100 and then I'll add.

Tirth 1:05:05
Yeah.

Tarun Jain 1:05:13
97 Now what will be the median?

Tirth 1:05:19
4488 will be the median.

Tarun Jain 1:05:23

Hmm.
Then you also have variance. So variance is not that much useful, but when you're trying to create standard deviation during that time, you'll need variance, right? Standard deviation is also very important for one of the search technique. I'll cover this shortly once we try to learn.
Beyond 25 in detail, so standard deviation will come to normalize, right? Normalize. I'll remove this beyond 25. So standard deviation is mainly to normalize the given array.
And then we also have one more command line space. Let's suppose you have zero to 100 and between zero to 100 you need to create the equal distribution of total 5 elements.
I hope these commands are clear, right? Mean, median, variance and standard deviation. This is what we studied earlier in mathematics. We just have to use numpy and then dot that particular function and then give array.

Margi Varmora 1:06:26
Yeah.
OK.

Tirth 1:06:34
We'll have to just revise the definition of variance and deviation, but there is that we know that. Yeah, OK.

Tarun Jain 1:06:38
But mainly variance will not be good. But when you're trying to create standard deviation, mainly it's like you have variance. It's like square root of variance.

Tirth 1:06:49
OK.

Tarun Jain 1:06:49
NP dot I don't know if square root is there, it's there. Here you just start to add variance. So this is nothing but your standard deviation.

Tirth 1:06:58
OK.

Margi Varmora 1:06:58
The.

Tarun Jain 1:07:03
12.1380 whatever formula I said.

Margi Varmora 1:07:05
2.

Tirth 1:07:07
Mm.

Tarun Jain 1:07:07
So this function at all is not required. You can just create variance and then you can use square root. But instead of directly calculating variance, we can directly calculate standard deviation.

Tirth 1:07:23
Open.

Tarun Jain 1:07:26
OK, so there are a few more comments which are useful in order to create some random strings, right? And if you need a random array, sorry not string, random array of equal distribution. Let's suppose you want to go to zero to 100.
In five attempts, how will you go? You will start with 0, then 25, then 50, then 75, then 100 if I make this 10.
Let's suppose if I make this nine, it is zero, then 12.525. If I print the length of this 25, right length of distribution.
It should be equal to whatever value you add at the end. So it's like you're starting with a particular value. You're ending at a particular Max value that you need, but you should take total 9 attempts to reach from zero to 100. So your zeroth index and

-1 index is fixed.
So in between that what values you have to fill that is the equal distribution and you have a logic. I mean the length of whatever distribution you have, it should be equivalent to the value that you have given here.
Is this clear? We just have to use line space. Now where is this useful? Let's suppose you create a histogram, right? In histogram you have X axis and Y axis.

Tirth 1:09:03
Mm-hmm. OK.

TJ **Tarun Jain** 1:09:08
So let's suppose this Y axis that you have, right? You want to plot something, but this Y axis should be equally distributed. Sorry, this X axis.

Tirth 1:09:16
M.

TJ **Tarun Jain** 1:09:19
Should be equally distributed.

Tirth 1:09:19
OK.

TJ **Tarun Jain** 1:09:25
So we use this function called NP dot line space to have a distribution.

Tirth 1:09:32
Mhm.

TJ **Tarun Jain** 1:09:34
A meaningful distribution. Whenever we are plotting anything, you can also call line space to be a placeholder.
Line space is a placeholder values.
For the X axis.
You can also do the same thing for Y axis. If in case you need equal distribution on Y

axis, you can also add line space to create the values. So this will just act as a value. So here, let's suppose you need to have zero. You want to go to London.
OK, 100 is a very short value. Let's probably use 10,000.
But you want to go in 15 attempts. If you directly plot without giving any line space, you will see some kind of very congested values, right? Every single value will be very congested. But if you need equal distribution, let's suppose from zero to 10,000, I need to have 15 or I need to have 20.

Tirth 1:10:28
Mm.

Tarun Jain 1:10:37
And during that time you'll have a very proper plots. So this can be used for line graphs or it can be used for histograms. This is just to ensure that you have a very valid or you have a very meaningful distribution when you are plotting. So this is just a placeholder.
Is this clear?

Tirth 1:10:58
Yeah.

Tarun Jain 1:11:02
And let's look at the final example which we already covered. I'm creating two different arrays, so the first array is 2 comma three. This is the formula that I showed earlier. I have WX, so W is nothing but weights.
X is nothing but my input and then you have. This is the formula that we use for neural network. You multiply your input with the weights.
So here I'm just taking my input to be random and the weights also to be random and the weights shape is 2 comma three and my X axis shape is 3 comma one. When I do matrix multiplication my.
2nd which is my N value of first array should match with the M value of the 2nd array and the array rank will be two comma one. This math you already saw. So now if I do matrix multiplication with these two values.

Tirth 1:11:56

Mm.
Yeah.

Tarun Jain 1:12:01
It is 2 comma one.
Is this clear?


Tirth 1:12:09
Yep.

Tarun Jain 1:12:11
And then you have dot operator for matrix multiplication it is matmul and for dot operator with summation it is just NP dot dot. And this is the formula you have by MC so Y and M. This is again one of the formula that we use for sigmoid, but I'll not cover sigmoid now.
We'll take this later. We'll take this example later.
For time being, let's just cover matmul which is nothing but matrix multiplication dot operator and log is already there log. If you want it in array you can use numpy. If not the best one is math dot log.


Tirth 1:13:07
Yeah.

Tarun Jain 1:13:09
Just to summarize, where will we use NumPy? Only when we are doing anything related to arrays, which is the type conversion into arrays. There are few important mathematical operations like dot, mean, median, matrix multiplication.
Because most of the operations that we are dealing when it comes to neural networks or AI, it is mainly related to vectors to vector matrix multiplication, right? So we'll be using dot and it will be numpy and when we talk about.
The vectors that we are using, it's mainly related to vectors of weights, right? So weights initially it is random.
When working with AI.
Weights are basically.
Randomly generated.

Initially.
And then we have seed value. Seed value is only used whenever we are training it. If in case you're not training anything, you don't have to use seed. Now whenever we are building with rag operation right for rag, basically you just need cosine similarity. You're not training anything. So when you're not training anything, there is no need to use seed value, but when you're fine tuning it.
Let's suppose you bring your own data. You're modifying the model weights. So during that time we will use it only when you're training and then when you're working with arrays, you also have a function called transpose. So we just have to ensure whenever we are using transpose.
And if you're using reshape, it should match.
It should match the size of the original array.
Is this clear all the basic operations that we covered in numpy?

Tirth 1:15:04
Mhm.

TJ

Tarun Jain 1:15:04
Any doubts before we proceed to Pandas?

Tirth 1:15:10
Not yet.

TJ

Tarun Jain 1:15:13
OK, so in most of the cases, if in case you want to know any comments, right, numpy is very freely available in the sense most of the operations that you want to use in numpy are very limited, like either it will be matrix multiplication or it will be something related to shapes.
Because some embeddings will be in different shapes, so you just have to check what is the shape of it and if you want to any modify the type conversion you will just use MP dot array if not list. So the limited operations is what you will be performing when it comes to numpy but ever since if you if you get stuck.
You can directly use their numpy.
Sorry, it's NP. So this will give you all the operations and then you can check if there is any operations that is making sense in my particular use case or not. And if you

use any function, hardly it will be just two or three parameters, right? So if you look at dot operator, what is the input variables for dot?
You need total two different arrays. Uh, where is dot?
So what is the syntax for dot? Basically you need 2 metrics so that you can multiply.
Same goes for matmul, but if you want to calculate for standard deviation or mean, you know that OK if I want to calculate a mean, you just give me the array.
So if you look at this, if you want to calculate mean, you just give me the array and only that particular input variable is fine when you have to calculate this. So in most of the functions you will never find more than two or three attributes or functions.
Is this clear? We'll probably proceed.

Hardip Patel 1:16:57
Isn't isn't matrix multiplication and dot product are are not same?

Tarun Jain 1:17:05
No, not most of the cases not same.

Tirth 1:17:06
It's.

Tarun Jain 1:17:09
So how is metric?

Tirth 1:17:09
With dot dot there is summation as well, no in matrix multiplication.

Tarun Jain 1:17:13
Oh, is the voice.
Matrix multiplication you have 34545681011.
And then you have one more metrics.
This is actually 2D. I'm not adding any square brackets here, so this is self understood.

Hardip Patel 1:17:34
Hmm.

Tirth 1:17:37
Yeah.

Tarun Jain 1:17:40
Here you have 91011101213.
OK, So what happens in dot product is you will check element wise element.
What happens in matrix multiplication is this particular thing will start with this one.

Tirth 1:18:02
What a multi level rules. Oh yeah.

Tarun Jain 1:18:06
Right. And then this particular thing will go with this one.
Then again, this particular thing will go with this one. So what is the shape of this? It is 3 cross 3. The shape of this it is 3 cross 3. So your output will also be 3 cross 3, but if it was.

Tirth 1:18:23
close.
Big Ghost tree.

Tarun Jain 1:18:29
4 cross 3 and if this is 3 cross 4, what will be your output?

Tirth 1:18:34
Or crossword.

Hardip Patel 1:18:36
Oh, this moment.

RamKrishna Bhatt 1:18:36
Waterproof Rd.

Tarun Jain 1:18:36

So if you look at the logic on how matrix multiplication works, it's a bit different. It's like you're doing first row with column and then you will proceed with the 2nd row, first column, then second row.
First column. So how are values added? So basically when you're going first, right, let's suppose.
You're calculating 345910910 and 13. It will come here.
3 into 9 plus.
4 into 10 plus.

Tirth 1:19:15
4 into 10.
I went to 30.

TJ

Tarun Jain 1:19:20
I into 13. So this is your first value. Then for second value you will have again 3 into 10, four into 12, five into 12. So this is a dot product. dot product is only this much, but the entire operation that you're doing it here is matrix multiplication.

Hardip Patel 1:19:22
I think we're on top 30.
Yeah.
2.
Yeah.

TJ

Tarun Jain 1:19:42
The only thing in matrix multiplication you have to remember is this two has to be same.
Oh, is this clear?

Tirth 1:19:58
Yeah.

Hardip Patel 1:19:59
Mhm.

Tarun Jain 1:20:00
OK, so I'll share a GitHub repo for pandas. I did create it in a very detailed format.


Hardip Patel 1:20:01
Yeah.

Tarun Jain 1:20:11
So this will be fun because we have different kind of data set. We have simple data set, we have randomly generated data set. I have an anime data set and we also have some.
We have data, we have random and we have some Naruto data set. So what we will do is we will try to work with pandas. Are you able to open this GitHub repo?


Tirth 1:20:32
Yes.

Tarun Jain 1:20:34
This is a notebook. What we can do is first we can download this notebook.
And then we can upload it in Collab.
And why is pandas used? Pandas is basically used for data manipulation. Anything that is related to data analysis, right? Whatever operations you do in Excel, the operations of Excel that you do, it can be done using pandas as well.
So I'm downloading this and now I'll come back to collab.research.com and here on the left hand side if you see you have examples, you have recent, you have upload here in upload, I'll just upload this.
Let me know once it's done. So the starting three to four commands, right? We have already covered that when we worked with cosine similarity. If you remember, I added some documents in a dictionary and I convert that the dictionary into. Dataframe.
So this particular command you might have already seen.
PD dot data frame and then what is this? It is a dictionary dictionary. How many keys I have?

Tirth 1:22:22
You have two keys now.

Tarun Jain 1:22:25
So 2 keys in the sense I have total 2 columns and then how many values I have. I have total 4 values. These four values is my.

Tirth 1:22:27
Hmm.
Uh.

Tarun Jain 1:22:34
Rose.
So let me know once you have, uh, uploaded the notebook.

Tirth 1:22:42
Load it.

Tarun Jain 1:22:55
Everyone are done or?

Hardip Patel 1:23:00
Yes.

Tirth 1:23:01
Sorry, I'm using VH code.

Tarun Jain 1:23:03
Hi in VS code also we can run.

Tirth 1:23:05
Yeah.
How do we how do we print a data frame like just print DF or is there a?

Tarun Jain 1:23:10
So uh, this comes.
So either you can do DF dot head. So what DF will do is here. How many rows do I have? I just have four rows so it will display all. But if you directly print DF, let's suppose you have around. OK, this will give an error. I didn't run this.


Tirth 1:23:17
OK.
Mhm.
Mm.

Tarun Jain 1:23:37
Here now what I will do is I will have animate and I will just print.
List of.
Random sorry range. I'll make it 1000.
So now if I print DF, it will only print limited starting first pyros and then the last pyros.
But if I do DF dot head it will print total starting file and you can decide how much total head you need. If you need 10 you can give 10, but if you give around 100 it won't print everything. There will be this dash dash dash in between.


Tirth 1:24:31
Mhm.

Tarun Jain 1:24:33
So till 20 or 25 you can see it in collab.


Tirth 1:24:38
Mhm.

Tarun Jain 1:24:41
And if I give tail, it's from the bottom.

Tirth 1:24:45
Yeah.

Tarun Jain 1:24:45
If I want to randomly print, what was the function?

Tirth 1:24:56
And.

Tarun Jain 1:24:58
Of what?

Tirth 1:24:59
OK.

Tarun Jain 1:25:01
It starts with this.

Tirth 1:25:11
Temple.

Tarun Jain 1:25:14
Samples. So now if I do DF dot sample it is randomly generated some 20 rows.
Not generated, displayed. I'll revert back to the basic example.
So whenever we are working with pandas, usually the data that we have will be either in Jason, but the most popular is CSV and Excel, but it also supports Jason, it supports HTML and it also supports XML and other format as well, but the commonly 2.
Used data set with CSV or Excel. Why? Because pandas is very similar to what you have in.
OK, I'll just print DF because I have only four rows, right? Pandas is similar to what you have in Excel. So if there are any operations that you you can do in Excel, you can also perform that in pandas as well. So DF is 1. DF is nothing but data free.
And if you want to create data frame from scratch and if you have very limited data,

you just have to do PD dot data frame and then you need to have dictionary. So whatever keys that you have will be your column and whatever values you have will be your rows.

Margi Varmora 1:26:35
Yeah.

Tarun Jain 1:26:38
Is this clear? And if I do the type of DF, it is pandas core frame data frame. So data frame is the class and in simple terms data frame can be referred as.

Margi Varmora 1:26:52
OK.

Tarun Jain 1:26:57
Excel representation or you can say it rows and column representation.

Margi Varmora 1:26:57
Mhm.
M.

Tarun Jain 1:27:06
In pipe element.
Uh, hello.

Margi Varmora 1:27:15
Eka 3317.

Tirth 1:27:15
Yeah, yeah, I think Margi is just unmute.

Hardip Patel 1:27:19
Who's a?

Tarun Jain 1:27:20

OK, so now let's see how you can add new column. So if you remember we use this syntax in recommendation system where we had a new column called overview and then what we did over there is once you do overview, we just combine some of the columns.

Tirth 1:27:21
Yeah.

TJ **Tarun Jain** 1:27:38
But what if when you're creating a new column, you should have same number of values. So how many values do I have here? I have 123 and four. Let's suppose I ignore one of them.

Margi Varmora 1:27:43
OK.

TJ **Tarun Jain** 1:27:57
So I'm creating a new column but I'm only giving 3 values. But technically I need to have total 4 values. So now if I run this it's an error. Why? Because the length of the values is 3 and it doesn't matches the length of the index which is 4. If you want to bypass this you have to add.

Tirth 1:28:17
Mt value.

TJ **Tarun Jain** 1:28:18
NP dot NAN.
So what is this NP dot NN?

Tirth 1:28:24
Not a number.

TJ **Tarun Jain** 1:28:26
It's not a number.
So now my DF will have a empty value if I do is null.

dot sum.
Have one year.
Is this here?

Tirth 1:28:44
Yes.
I have a hard stop at 1:00, but I can extend till 5 more minutes. Do you think we can cover this today's session in next 5 minutes or we need more time?

Tarun Jain 1:28:54
We need some time, probably 15 more minutes. We can wind up then.

Tirth 1:28:58
Then I'll I'll have to drop then.

Tarun Jain 1:29:04
OK, we can cover two more functions, then we can probably drop off.

Tirth 1:29:04
OK. Yeah. Thank you. I'll I'll.
You can continue. I'll I'll just go through the recording if that's fine, yeah.

Tarun Jain 1:29:12
No, what we can do is we have to have EDA. So for EDA we can do pandas and matplotlib together.

Tirth 1:29:18
OK.

Tarun Jain 1:29:20
OK, so first command was we are creating data frame and 2nd what we are trying to do is we are creating a new column and whenever we create a new column we just have to ensure whatever values you're filling should match with the existing length. So what is the existing length? It's length of DF.
Which is 4. And if you're adding new values it should also match with four. And

earlier also we looked at this example which had overview and then you're combining the existing data itself which was rate. Then we have name. So those columns already had the same number of values which didn't give us any error.

Tirth 1:29:56
Hmm.

TJ **Tarun Jain** 1:30:02
But if you're adding your own values, it should be the same. I hope it's clear on how to add new column.

Tirth 1:30:08
Yes.

TJ **Tarun Jain** 1:30:09
And the command to check whether the given data is null or not. It is DF dot is null. If you just print this, it will show you all the names of the columns and then it will show false, false, false. But if you want to actually get the actual number you can use dot some function.

Hardip Patel 1:30:10
Yes.

TJ **Tarun Jain** 1:30:29
So if you see now in our given data, if I just print data.
I have no empty data, so empty data will be usually denoted by NAN or it will be none.

Hardip Patel 1:30:44
Uh, it it this. Uh is it OK to use fill any right?

TJ **Tarun Jain** 1:30:44
So these are the two values.
Uh, what will do is let's suppose uh.
This is an N.

Now this is 1.
You can do DF dot fill NA and then you can give a empty value for it. Now this is an empty value.

Tirth 1:31:13
OK.

Tarun Jain 1:31:14
So here you can also see no value.
No value. So it's like a placeholder what you want to give in that empty value space.

Hardip Patel 1:31:28
OK.

Tarun Jain 1:31:30
And there is also one more way you can add new column which is if you give any value. Let's suppose you're creating a new column which is check and if you just give one value it will repeat it for entire thing. This is like Excel.
In Excel, if you remember, if you give, let's suppose this is an Excel sheet, right? You start with check and you're giving one to this particular column which is one piece. And then if you just drag that to all the columns, what will happen in all the other columns?

Tirth 1:31:46
Mm.
It will just copy.

Tarun Jain 1:32:02
In Excel, if you give one and if you just choose the drag for every columns also it will be one.

Tirth 1:32:04
Hello.
Yeah, it will copy.

Tarun Jain 1:32:08

This is the same logic. So if you give any static value, let's suppose if I give a list, it's an error.

So now what it is trying to do? Hey there are total 4 values but you only give one value. But if you give a static value this will add it for entire thing. This is like Excel drag option that we do.

And this is again indexing method indexing. There are two way you can do it. One either you can use DF then slash animate. If not you can use DF dot animate. It's the same thing.

Either you can use a dot operator or you can use square bracket with that particular column name and if you just give anyway and this E this particular key doesn't exist, you'll get key error.

Tirth 1:32:56

Um.

Tarun Jain 1:32:57

This is similar to how you do it in dictionary. In dictionary also you have it. Let's suppose you have anime as a key. You can directly use that particular dictionary name and anime, but here it is data frame and the column name. Either you can use square bracket or you can use dot operator.

Tirth 1:33:15

Oh.

Tarun Jain 1:33:16

But if you want to extract multiple column at once, even that is possible. So if you remember when we concatenate rate, when I created a new column overview, we used DF. We had a main square bracket and inside that we had.

Two columns, one is main character and one more is animate and then you had join operator. Here I'm not using join but if you want to extract 2 columns you just have to define a main which is your first square bracket and then inside that you can define any multiple columns that you need. So here I'm defining main character and animate and then it will.

Display those particular columns. So for multiple column selection you can define double square bracket.

Tirth 1:33:55
Hmm.

Tarun Jain 1:34:04
Probably from here we can continue in the next class of modifying index and then unloading it from CSV file.
It.

Mitesh Rathod 1:34:15
See you next time. See you next time.

Tarun Jain 1:34:19
Oh, whatever.
Hello.

Tirth 1:34:25
No, I think it was by mistake.

Hardip Patel 1:34:26
Read PD code HR.

Tarun Jain 1:34:28
OK, so let's do one thing. What we'll do is here. If you see I have something called as reading CSV file and we just have two more functions. So probably you can complete these two more function. Then tomorrow we can directly start with reading CSV file where we will work with.
The Excel data set.

Mitesh Rathod 1:34:48
OK.

Tarun Jain 1:34:49

There is just two more functions. OK, so now let's suppose you create 2 columns, but you also need to have a index column. So for our movie data set we had rank column. So for rank it was like you went to 0 to 1.
Right or zero to some particular number. Here what I'm trying to do is you have a dictionary of that particular number and if you want to index it, so I have index and then I'm giving ABCD. So now if I run this and if I check the data frame you have index as ABCD.
So you're not creating a new column. So whatever existing index is there, you're just changing it. So when you do filtering, it becomes easy for you to filter. Either you can give serial numbers, serial numbers in the sense you'll have actual serial numbers of this particular anime. Let's suppose bleach I give.

Tirth 1:35:24
M.

Tarun Jain 1:35:42
Some B123. So this your B123 will be your index. There are two ways you can use index. Either you can create a new column. If you create a new column it is a new feature.
New column equals to new feature, whereas index is not a feature.
Why? Because if I print.
DF dot columns.
I will only get animate episode. I won't get that particular column. But let's suppose when I add this index right, I add serial number which is DF dot serial number.
And I have a list of serial numbers like I have 0123 then N 123.
And then I have B123.
I now have G123. What am I doing it here? What is the syntax for?

Hardip Patel 1:36:58
Setting serial number new setting new column.

Tarun Jain 1:37:03
So this will add a new column. Now if I print DF.
So basically these are just serial numbers, right? But when I print the total number of columns, you have a new column. So basically when you add a new column, it's a

new feature. But if I give the same thing in the index.
And if I rerun this.
It will be added in the index and you don't have to use a new feature. So this is there is a new function called index if in case you want to modify your existing index. Usually it will be serial numbers. If not, it will be just increment values like 01234 till end by default.
And you can also define set index as episode. Let's suppose you had serial number, right? I'll go with serial number again.
If you go with this approach, let's imagine I have ABCD.
Now if I print DF you have ABCD, then you have serial number which is 0123 N 123B123. If I print number of columns, how many columns do I have?

Hardip Patel 1:38:15
8.

TJ

Tarun Jain 1:38:16
Have three, but if I do set index as serial numbers, now the data frame will be different. It is serial number, anime and episode. Now this serial number is my index and how did I define this by using set index?

Tirth 1:38:16
Thank you.

TJ

Tarun Jain 1:38:33
So this is only used when if you forget index here. Let's suppose you don't include this index. You can use a function called set index and again whatever values you have it should match with the existing list of the values. If you give just three serial number then it will be an error again.
Is this clear?

Tirth 1:38:53
Yeah.

TJ

Tarun Jain 1:38:55

OK, now all these things that you're seeing right anime, episode, serial number, this individual column that you have, the data type of individual column is series.

Hardip Patel 1:38:55
Yes.

Tarun Jain 1:39:10
Data type of individual.
Column is series whereas.
Data type of the entire data is data frame.
This is for Pandas.
So whatever anime you have read this particular column, this particular column just one is called series. So if you see a series is 1 dimension representation of the data, but data frame is a multi dimension. Why? Because it is in two dimension.
What is 2 dimension? You have rows, you have columns. So if you do DF dot shape. If you do DF dot shape, why is it 4 comma 3? Because you have total 4 rows and you have total 3 columns right? But pandas series is nothing but just a single dimension. So if you see you are creating PD dot series which is good, bad, neutral and then you are defining the index. You have ABCD good, bad, neutral. It is a single dimension. Same goes for this series as well DF2PD series. Then you have you are randomly creating 10 values.
And for each 10 values you are giving the index starting from 1:00 to 10:00, so every single one dimension.
One dimension data in pandas is called.
Series.
Is this clear?

Hardip Patel 1:40:56
Yes.

Tarun Jain 1:40:59
So this was just the basic uh.

Hardip Patel 1:40:59
Yes.

Tarun Jain 1:41:03

What you call basic function to get started with pandas, we define data frame. Then we have series which is 1 dimension. But if you want to index there are two ways we can index it. Either we can use DF dot that particular column or you can give it as. Yeah so there are two ways we can do it. And then what else did we cover? How do you add new column? You can just do DF dot the new column name and whatever values you give should match with the existing values.

And if you give any static value like one, it will add 1 to every single rows that you have. So this is the syntax for new column. So tomorrow what we will do is we will start with how you can load the existing CSE file.

And there are different platforms like Kaggle is there from where you can load the CSV file and you also have DeFi. DeFi basically is the AI planet itself. We also have some data set which can be used and in recent days we have seen hugging phase. Hugging phase can also be used.

You can convert your ugging phase data to CSV and then load it here, but which is not required. We will directly work with CSV files.

Uh, is this any questions in Ampai and Pandas?

Tirth 1:42:21

What I look now?

Tarun Jain 1:42:23

OK, so the major thing when it comes to numpy, pandas and marplotlib is EDA because there you will have an actual data. In our case we will take video game sales and we will just see via first month what was the sale, second month what was the sales and then we'll have some distribution.

Hardip Patel 1:42:25

No.

Tarun Jain 1:42:41

And once you are completed with EDA then we can go with RAG. So this was one pending topic. This was supposed to be covered before NLP and after this we were supposed to start NLP. But now we will complete this then go with Langchain and

RAMA index.
So hardly two more class will be just pandas, matplotlib and one class will be on EDA.
ETA is like try to understand how data works. That's it.

Tirth 1:43:05
8.

Ishan Chavda 1:43:10
Nothing.

Tirth 1:43:11
Thank you. Thank you, everyone. Thank you, Tarun.

Tarun Jain 1:43:14
Yeah.

RamKrishna Bhatt 1:43:14
Just small confirmation, do we have session tomorrow? As Tarun was saying that he's not.

Hardip Patel 1:43:15
Yes.

Tarun Jain 1:43:19
Yeah, tomorrow will be on leave. I'll probably do RSVP no.

RamKrishna Bhatt 1:43:24
OK, OK.

Tirth 1:43:25
OK. OK. Thanks.

Tarun Jain 1:43:26
Yeah.

Hardip Patel 1:43:28

Save. Thank you.

TJ

Tarun Jain 1:43:29

Yeah.

RamKrishna Bhatt 1:43:29

You. Thank you. Bye, bye.

Hardip Patel 1:43:30

Yeah.

Tirth stopped transcription