   
Carlo Lipizzi: Hello.

2:30

 Elyse Spinelli: Hi! There! Evening, Professor.

2:34

 Carlo Lipizzi: How are you

2:37

 Elyse Spinelli: doing? Well? Happy to be out of work. We're doing our our release planning week this week. So it's

2:39

 Carlo Lipizzi: hard. Course.

2:46

 Carlo Lipizzi: What? What? What is the company you are working in?

2:49

 Elyse Spinelli: I work for? Lockheed Martin? I'm: a: software right now.

2:53

 Carlo Lipizzi: Okay, okay.

2:57

 Carlo Lipizzi: we have a quite a lot of business with Lo located Martin. We had more in the past, but we still have a quite some students from.

2:59

 Elyse Spinelli: We just happened to split my project into a dev, and then a set offs team, which

3:13

 Elyse Spinelli: kind of contrary to the like. I feel like that's set off. But you know. Oh, well, we'll see how it works out. I happen to be on the the set ups, too, so i'm excited to get

3:21

 Elyse Spinelli: it more into infrastructure and out of depth.

3:33

 Carlo Lipizzi: Good! Good good is this call so helping you in your activities as some also did contribution.

3:36

 Elyse Spinelli: it's helping. So I happen to have applied to a program within the company that, and even if I don't get it, i'm going to try to move into more role that python plays more of a rolling

3:46

 Elyse Spinelli: currently. I'm. I only use really like bash scripting yaml

3:59

 Elyse Spinelli: like pipeline sustainment.

4:06

 Elyse Spinelli: I use this necessarily in my day to day, but I from my capstone and undergrad. I actually did a AI project that I let. And so this is kind of bring me back to that. But I haven't used Python. Really, since then.

4:08

 Carlo Lipizzi: I mean that if you want to take the opportunity of the to do something that is more in line with what you are doing at work feel free to do so.

4:24

 Carlo Lipizzi: I mean that I will be happy to help you as much as I can.

4:35

 Elyse Spinelli: Yeah, I think only halfway through the degree. Also, there are classes that, as far as electives go.

4:41

 Elyse Spinelli: kind of inner working, the

4:50

 Elyse Spinelli: you know, trajectory of the with. Even some like systems work would be like, really, you know.

4:52

 Elyse Spinelli: Help? My.

5:00

 Carlo Lipizzi: Yeah. Yeah. Yeah. Yeah.

5:01

 Carlo Lipizzi: Okay. Good. So anyone else want to share how the courts it's helping or not helping what you do at work.

5:05

 Elyse Spinelli: I will say one other thing, I think.

5:19

 Elyse Spinelli: At least it's helped me gain some more

5:22

 Elyse Spinelli: planning credibility like within my team since i'm one of the youngest contributors, so like that's kind of been a nice knowledge. Slash confidence to have the you know the classes in my back pocket.

5:25

 Carlo Lipizzi: It's good. Good.

5:40

 Carlo Lipizzi: All right. Okay. So

5:42

 Carlo Lipizzi: let's move with the content for this class.

5:45

 Carlo Lipizzi: So we we basically have 2 more classes, and then they they will be the presentations of your final. So again, if you have any question on the finals, that's the moment for you to ask if

5:51

 Carlo Lipizzi: any. the again. The the final is going to be a a project.

6:11

 Carlo Lipizzi: You can develop the project in a team, as some of you already communicated to me.

6:20

 Carlo Lipizzi: You can pick one on the

6:29

 Carlo Lipizzi: suggestions that they provided the on canvas. You don't need to do one of those, but it's just for your convenience, somehow.

6:32

 Carlo Lipizzi: ere

6:44

 Carlo Lipizzi: or a possibly data that even question that you have, and you want to take the opportunity or the final to do it, feel free to do so.

6:57

 Carlo Lipizzi: Final will be pretty much similar to what you did in the last assignments.

7:06

 Carlo Lipizzi: the

7:16

 Carlo Lipizzi: keeping in mind that that is a final meaning. If for the assignments you do an interpretation of the results? That is a a few page for the final, it should be more. A substantial is not.

7:17

 Carlo Lipizzi: and 800 project is, is not a TV, so it's not a dissertation, meaning it is not something of 20 page or or a 50 page.

7:34

 Carlo Lipizzi: but it should be something around 1015 page, all included.

7:48

 Carlo Lipizzi: I I share with you the

7:55

 Carlo Lipizzi: structure that the final shouldn't have. So again, there are 4 phases defining the business defining the what the the data set and analyzing the data set.

7:59

 Carlo Lipizzi: to be sure to bet it somehow, to be sure that the the data set is able to address the the questions so that you play the in the first step.

8:14

 Carlo Lipizzi: then the the data preparation that is always a critical face.

8:28

 Carlo Lipizzi: And then you have the data exploration in a broad sense. So we are. You do all the

8:34

 Carlo Lipizzi: the visualization. So you calculate the metrics and things like that.

8:44

 Carlo Lipizzi: and it's where you actually will write the the narrative that will put together all those pieces in a a way that people I can read, it can understand, and even if they may not be experts.

8:53

 Carlo Lipizzi: all about how to do things. They may be expert or the domain, but they may not be expert in writing code or or extracting magics, 150,

9:12

 Carlo Lipizzi: so

9:24

 Carlo Lipizzi: a complexity.

9:26

 Carlo Lipizzi: The last assignments are up a little bit longer, so they are around the 100 lines. So, as I was mentioning last time, the number of lines is not necessarily a 100% through indication of the complexity of the problem.

9:30

 Carlo Lipizzi: But it's kind of difficult to have a a a problem

9:51

 and analyze the with the proper level of complexity with the 20 lines of code.

9:57

 Carlo Lipizzi: Yes, on not all the lines are created in well, but 20 lines, so it would be not on enough, probably not even 50, probably not not even 80.

10:03

 Carlo Lipizzi: So be sure that that you have a a, a, a a script that you would use to extract the the the metrics creating the scripts, the visualization. So

10:13

 Carlo Lipizzi: that is more or less, or at least

10:29

 Carlo Lipizzi: not shorter than the last assignments. So something around, I mean from 100 50, and a move.

10:34

 Carlo Lipizzi: So again, the length of of the script is not a a, a, a, through a complete indication of the complexity. But again, it is difficult to to do a complex analysis with just a few lines of code.

10:45

 That's

11:04

 Carlo Lipizzi: E. C.

11:06

 Carlo Lipizzi: Generally speaking, when you send emails to me at the she you, because this way you double the possibilities to get an answer soon, we will give you an answer.

11:14

 Carlo Lipizzi: But sometimes we are droning in emails, and in particular, when we are approaching at the end of the semester, and your email may get lost.

11:27

 Carlo Lipizzi: So just to be sure, add always she you in the email that you sent to me.

11:42

 Carlo Lipizzi: Okay. So this class and the following one: we'll be more on how to apply things. So there is nothing more that really we need to explore either in the informatics part or in the

11:50

 pied on paths.

12:11

 Carlo Lipizzi: The main reason is because obviously we could go into more details with Python. We didn't

12:14

 Carlo Lipizzi: erez agmoni work with the quite a lot of potentially useful libraries. We didn't spend too much time in applying some techniques. 101.

12:23

 Carlo Lipizzi: One of the things is what we we want to reduce loops using operations with Madrid sees, but that would be more on a spending time in a linear algebra, and this is not the goal of this course.

12:34

 Carlo Lipizzi: At this point I I really think that you have all the elements in terms of pied on that, and in terms of how to develop projects, to use a

12:51

 Carlo Lipizzi: coding, to create a story and to extract the meanings from a data set in the direction or questions that that you may have.

13:03

 Carlo Lipizzi: I will spend a a little bit of time on using a chat. Gpt.

13:17

 Carlo Lipizzi: A. Recently I mean Chat Gpt is a

13:26

 Carlo Lipizzi: representative of the family of large language models

13:35

 Carlo Lipizzi: ere

13:43

 Carlo Lipizzi: so Chat gpt is one is not the only one it's for sure. The one who created the the

13:51

 Carlo Lipizzi: the most of the hype that we have since the announcement, or the availability of this tool. That it was a you know November, but it's not the only one. There are some others that are coming.

14:00

 Carlo Lipizzi: I I've been working on a those.

14:18

 the

14:25

 Carlo Lipizzi: relatively a lot.

14:27

 Carlo Lipizzi: I've been interviewed on how to use those bots for education, and there is no answer, because no no one knows.

14:30

 Carlo Lipizzi: But I will share with you my opinion, and I will also share with you a little bit of insights on how to get the most out of them.

14:43

 Carlo Lipizzi: So it is a matter of fact, that is there a chat gpt like that, so that will come. It's definitely good to know what our

14:56

 Carlo Lipizzi: the the potential uses on one side and the limitations that they have, and also how to better use the the tool to get something done.

15:09

 Carlo Lipizzi: So we we will talk about that. It's not something that that was a scheduled, because I mean that we started the E. M. 6, 24 way before the so but I mean this area is evolving continuously

15:24

 Carlo Lipizzi: and I don't want to give you the impression that we are not following relevant changes that are happening in the sector in the area, having a a potential high impact in what you do.

15:43

 Carlo Lipizzi: So let's go down to what is more directly related to this courts, and let's go first on

16:03

 Carlo Lipizzi: the proposed solution for the current assignment. So the assignment was basically on downloading elements from a a web page and do some processing.

16:18

 and some I mean, some cleaning, some processing, and some visualization.

16:36

 Carlo Lipizzi: So

16:44

 Carlo Lipizzi: again, like always, that's one of the medium possibilities for addressing this problem I create. I imported all the libraries one of the libraries I didn't use it. I don't know why I didn't delete it, but that's the way it is.

16:48

 Carlo Lipizzi: So. As I was mentioning last time it it it's a good way to proceed to create your own a cleaning function for a text. So in this case, again, may not be the most effective.

17:08

 But

17:28

 Carlo Lipizzi: look at it as a source of placeholder, so you can change it, you can make it yours. You can customize, but you want to have a your own, a text cleaning function.

17:32

 Carlo Lipizzi: So in this case this function is taking a a string to a process to clean the minimum length of all the words that we want to consider.

17:43

 Carlo Lipizzi: Assuming that that works, that they are shorter than a a given number of a characters, so are not semantically relevant, and then a list of so forth.

17:56

 Carlo Lipizzi: So basically the function is a transforming the string awards into a list, and then it's looping into the list 250.

18:07

 Carlo Lipizzi: It's a transforming in a lower case. It's checking, if it's Alpha medical or not, 250,

18:19

 Carlo Lipizzi: and then it's checking. If the length of the the word the it's bigger than the threshold that we set, and then checking if it's in the so, what not? If all good will append that the word to the list of clean words.

18:28

 Carlo Lipizzi: and then we loop up to the end of the list, and we'll return that, or of the initial string, and we'll return the the

18:50

 release of cleaning up in clean words.

19:01

 Carlo Lipizzi: So again, you have an explanation here on what is in and what is out.

19:04

 Then

19:11

 Carlo Lipizzi: it's about webpage. So I defined what did the what is the URL for the the web page? Do they want to get?

19:14

 Carlo Lipizzi: I'm getting the the content?

19:22

 Carlo Lipizzi: I'm moving the content into a 3 like structure, using a beautiful soup.

19:26

 Carlo Lipizzi: Then I initialize a string that we hold the the words in the headline.

19:34

 Carlo Lipizzi: i'm. Looping into the paragraphs Again: If New York Times will change the label for paragraphs from P to something else. My code will not work anymore, and

19:41

 Carlo Lipizzi: those are the limitations, or this one, and simulates are the limitations for a scraping content from a website.

19:58

 Carlo Lipizzi: and I appending the

20:11

 then that

20:18

 Carlo Lipizzi: initializing the listed will contain the so forth, reading the soap or file, extending the so so far, the the the so

20:22

 Carlo Lipizzi: setting the minimum length.

20:34

 Carlo Lipizzi: filtering the words, using the function that we saw before

20:37

 Carlo Lipizzi: extracting the diagrams

20:42

 Carlo Lipizzi: and getting the most, the 7 most common. I could be any number

20:46

 Carlo Lipizzi: then

20:54

 Carlo Lipizzi: creating a a a, at least of the most common

20:58

 Carlo Lipizzi: filtering, I mean extending the the list of words, so that I had the after the cleaning with the those common diagrams

21:06

 Carlo Lipizzi: transforming up the the list of awards into a string, because work cloud is taking a string solely.

21:18

 Carlo Lipizzi: and then pretty much printing it

21:29

 Carlo Lipizzi: and saving eventually as a Png file.

21:33

 Carlo Lipizzi: Then I analyze in the sentiment, using butter. positive, negative and neutral printing it and the process running it.

21:39

 Carlo Lipizzi: So you have here your work, Cloud.

21:52

 Carlo Lipizzi: You have the headlines.

21:58

 Carlo Lipizzi: I mean. Some more cleaning should be done.

22:01

 Carlo Lipizzi: and obviously is, a lot about. Trump is real time, and today it's pretty much trump day.

22:09

 I like it or not.

22:19

 Carlo Lipizzi: and that's that's it.

22:23

 Carlo Lipizzi: Sentiment

22:27

 Carlo Lipizzi: so again, like in most of the cases, a new drama is the vast majority.

22:30

 Carlo Lipizzi: It's more on the negative side now, so it it would be interesting to do it end of the day each day and see if and that is a trend somehow. It's negative.

22:38

 Carlo Lipizzi: anyway. So that was the

22:55

 Carlo Lipizzi: assignment that

23:02

 Carlo Lipizzi: I mean if you look at the length. Yes, there are quite a lot of spaces. There are a lot of comments the length it's 132, probably the real number would be more 100, even a little bit less than that.

23:04

 Carlo Lipizzi: So just to have an idea

23:23

 Carlo Lipizzi: that's a kind of a basic, simple, relatively simple, but that that there is a process in it, that I mean the the script, the a as a pipeline. So the cleaning, the downloading, the cleaning, and then you have the different analysis.

23:26

 Carlo Lipizzi: All right. So let me go here and let me

23:46

 Carlo Lipizzi: go into

23:52

 Carlo Lipizzi: the content.

23:56

 so that that was a

23:58

 Carlo Lipizzi: the the assignment that we just so the possible solution

24:01

 Carlo Lipizzi: they close this we do not need any more, so the content will change a little bit. So we spoke about the machine learning a little bit in the past. I I I want to fault you more on

24:11

 Carlo Lipizzi: how to extract metrics from text, and then we'll be talking about the the Ls.

24:27

 Carlo Lipizzi: So let me start with the and I will definitely give quite some time for the the in class the assignment.

24:35

 Carlo Lipizzi: So let me start with the

24:49

 Carlo Lipizzi: a couple of how we extract

24:55

 numbers from text.

25:05

 Carlo Lipizzi: So one of the issues let me stop for a second sharing this. So one of the issues that we always have when we deal with text

25:08

 Carlo Lipizzi: is that

25:18

 Carlo Lipizzi: text is. I mean it's text, and there is

25:21

 no number. I mean, not semantically relevant in the text, or may not be

25:27

 Carlo Lipizzi: how we get insights from the text in a numerical way, in a way that we can

25:34

 just take decisions based on that.

25:48

 Carlo Lipizzi: So decisions it could be of any kind. So we will solve this on 2 examples, that they did the in the past few years. So one is on social media. So when I started my Phd.

25:51

 The social media was kind of a beginning to be relevant.

26:10

 Carlo Lipizzi: But there were not many metrics to measure the

26:17

 Carlo Lipizzi: semantic, the meaning in the messenger that was in the social media.

26:24

 Carlo Lipizzi: So my idea was, we need to find a way that is beyond the the the, the the statistical analysis, the number of weeds, the number of bits per time interval, or the number of tweets from someone at the number of retweets. So those are loosely toppled with the semantic meaning.

26:29

 Carlo Lipizzi: Okay, obviously, you have more tweets from an individual that that individual is more relevant that someone I like myself with not many of those.

26:58

 Carlo Lipizzi: but it it it's really not a great indication, I mean. It is not in in the content itself, but it's more from like like looking from a a outside.

27:12

 Carlo Lipizzi: At that time they were not many solutions to go into the content, and extracting a meanings from the content.

27:24

 Carlo Lipizzi: and that was one.

27:37

 Carlo Lipizzi: After a few years I did a.

27:39

 Carlo Lipizzi: I was asked to be the principal investigator that is, the project manager, the one doing the the, the the designing, the solution for the problem in a large project for the the Department of Defense.

27:44

 Carlo Lipizzi: It was a the a 4 million dollar and change project over a 2 and a half years.

28:03

 Carlo Lipizzi: with the 20 between 20 and 25 people in the in in my team.

28:12

 and the question was

28:21

 Carlo Lipizzi: the

28:25

 Carlo Lipizzi: how we can do better the planning, the the the cycle of production or Webinars, so that

28:26

 Carlo Lipizzi: customer was the the Piccadilly at all. So

28:39

 Carlo Lipizzi: they plan a new weapon. So

28:44

 what in advance

28:49

 Carlo Lipizzi: and typically the way to do it is using the capabilities, meaning more powerful weapons and improving the effective mess all those weapons, so the usability and the effectiveness.

28:51

 Carlo Lipizzi: But that was basically it. There was nothing working more on now.

29:10

 Carlo Lipizzi: How they said now that

29:21

 Carlo Lipizzi: events may have an impact on the cycle, the life cycle, and eventually the the planning of new weapons.

29:25

 From now to I mean the duration of the life cycle.

29:38

 Carlo Lipizzi: So we we we started thinking, okay, we need to work on a a. We need to analyze what our the technologies that are making an impact on the competitive scenario of the

29:43

 development. And the so

30:06

 Carlo Lipizzi: we fall used on papers, pardons, and use giving a 3 different stage somehow, or the maturity, all the technology.

30:10

 Carlo Lipizzi: and we for use the on.

30:23

 Carlo Lipizzi: I mean, obviously for the Department of Defense. The for use was on countries.

30:27

 Carlo Lipizzi: and so the competitive scenario was with the competitors, being countries

30:36

 Carlo Lipizzi: in universities, and in particular in a graduate program. So the vast majority, meaning probably 85% or something around that are international students.

30:44

 Carlo Lipizzi: Those topics are are sensitive topics, and they need the at least

31:03

 Carlo Lipizzi: people lot.

31:14

 Carlo Lipizzi: I mean national. We have National is either a city zen, or a a green card holder.

31:18

 Carlo Lipizzi: We didn't have many of them so. And then, when you move to the real content, the the real text that you want to analyze that you need the the security guidance. So I got my my security guidance. But I was the only one in the team with the

31:30

 Carlo Lipizzi: because of that we switch the from the Via a scenario to and sort of an ultimate reality. So the alternate reality was about

31:50

 Carlo Lipizzi: the security industry, and we for use on the

32:07

 Carlo Lipizzi: operators in the security industry, publicly traded the companies working in the security industry. So at that point you have all the information you want, because they are a publicly rated.

32:16

 Carlo Lipizzi: All the information are public, because again, they are publicly traded companies. So and you have, but because if they are publicly traded.

32:30

 Carlo Lipizzi: chances are they are a relatively big. You have quite an other news, and there could be some pad and some papers that could affect the somehow. They've strategies in terms of technologies that they were using

32:41

 Carlo Lipizzi: Erez agmoni. But the problem was still there. So how you guessed a a competitive scenario, meaning doing a risk analysis. Risk analysis is based on numbers 250.

32:57

 Carlo Lipizzi: So you need to have numbers to work on a

33:11

 Carlo Lipizzi: risk panel and work on a what if analysis with the different scenarios that you can have in this type of competitive environment

33:18

 Carlo Lipizzi: decisions that could be what is going to happen in the competitive scenario If I invest more in one technology and less in another one.

33:34

 Carlo Lipizzi: What if I totally dismiss one technology? What if I go all the way with another technology. So all of those are elements that really

33:46

 Carlo Lipizzi: need to have a to evaluate.

34:01

 Carlo Lipizzi: So there there was a missing link social media

34:08

 Carlo Lipizzi: how you get so the metrics relevant I mean the semantic metrics relevant to to understand what's going on in the social media and on the other end how you get

34:13

 Carlo Lipizzi: the metrics that will allow you to work on a

34:30

 Carlo Lipizzi: management, risk management or a risk analysis in that the second case.

34:36

 Carlo Lipizzi: So let me

34:45

 Carlo Lipizzi: go into the 2 projects to

34:48

 Carlo Lipizzi: explain a little bit how we did it.

34:55

 Carlo Lipizzi: In both the cases I was a a force to create my own. I'll got it, because in both the cases there was a nothing that could help me at that point.

34:58

 Carlo Lipizzi: So

35:17

 Carlo Lipizzi: social media. Again, I I will go relatively fast, because I I really want to give you time for the in class exercise. So

35:20

 Carlo Lipizzi: so most of the analysis at that time, but even now are based on either a statistical analysis or a sentiment analysis. That is what it is.

35:32

 Carlo Lipizzi: So that

35:53

 Carlo Lipizzi: idea was to extract meaning somehow from a a text. Why, this is

35:54

 Carlo Lipizzi: social media at that time in particular. Now I don't know what it's going to be with a tweet or a

36:02

 Carlo Lipizzi: but we tweeted a social media was a what we call the a back channel for a

36:15

 Carlo Lipizzi: a real life event, meaning wise people. We are doing something in real life. They share the comments via the social media.

36:26

 Carlo Lipizzi: I still through

36:40

 Carlo Lipizzi: not sure that the Tweed, or is is still a the number one social media for those things. But still, when you are experiencing something that can be an event, or can be watching a TV show, or whatever

36:43

 Carlo Lipizzi: or commenting a political event. Again, you share your toes on that

37:07

 Carlo Lipizzi: this channel. So this channel is what is normally called back channel. So the front channel is real life that things that are happening. The back channel is is basically the social media giving you insights on what is happening.

37:20

 Carlo Lipizzi: We consider a 2 types of events events that are not evolving in time meaning. You have one single event, the presentation of a new product, or or a a movie, or something like that.

37:43

 Carlo Lipizzi: and you want to get some meaningful keeper 4 months indicators that that can give you a sense on what was going on, and then

38:04

 Carlo Lipizzi: events evolving in time. And you want to monitor how how things are going, and eventually change what you are doing, based on the feedback that you got.

38:19

 Carlo Lipizzi: So again

38:35

 Carlo Lipizzi: we use, you know, Twitter with most of the social media. You don't have a real conversation, probably the only social media. I mean that that there are very few that can really be. I consider a conversational where conversation is a You say something.

38:37

 Carlo Lipizzi: the other person is replying in a certain way, and there is a back and forth. So that's a compensation. There is no real conversation in most of the social media.

39:02

 Carlo Lipizzi: But when you have a a common ground, so the event that that you are tweeting about the then there is a sort of a convergency around this

39:15

 Carlo Lipizzi: common events, so this having a something in common, it's what is called the the common ground theory, and it's well documented in the the literature.

39:31

 So

39:44

 Carlo Lipizzi: the methodology that they use the was detecting the the event collecting the data. I use the a combination of the search Api

39:46

 Carlo Lipizzi: from a twitter to python python for all the cleaning Mongodb for getting the data.

40:00

 Carlo Lipizzi: I mean, when you download the the tweets, you get a a. Json file. A Json files are a really easy to be handled by a mongodb

40:10

 Carlo Lipizzi: Mongod. It' be it's a a non SQL or a database.

40:27

 Carlo Lipizzi: and the structure of the I mean, since to be kind of all the but the the the the engine of Mongodb is based on Jason structure, meaning. If you have a Jason, you dump the Json into Mongodb that's optimizing the tool You have.

40:36

 Carlo Lipizzi: What some other things I extract from one? Would it be some particular characteristics? And then those will go to a a Mysql database

40:56

 all the pre-processing

41:09

 Carlo Lipizzi: when you have a events that are happening in a given period of time. I created it back. It's all tweets just to compare

41:12

 Carlo Lipizzi: the semantics for each period of time.

41:26

 Carlo Lipizzi: Then I I combined the metrics in

41:33

 Carlo Lipizzi: sub events along this major again, so that i'm monitoring a and then create eventually classification model, and then visualize, evaluate the the capabilities of the model.

41:40

 Carlo Lipizzi: So the methodology was in in multiple stage. So the first stage is obviously collecting, as in the previous chart, the data

41:58

 Carlo Lipizzi: and then creating a 2 types of a networks, one that is the social network. That is the easy one.

42:08

 Notes are of the messages.

42:20

 Carlo Lipizzi: and 2 notes are connected. If one is citing another.

42:24

 Carlo Lipizzi: so doesn't need to be retweets. But if you have the name of a another, user meaning another screen name, then you have an edge from the 2 notes.

42:34

 and there's a social, and then I created what is called the bipartite network.

42:51

 Carlo Lipizzi: That is a so bypassed networks we are the the elements that are connected the are are these joint? So the TV gala example is when you have a I don't know people activities.

42:57

 Carlo Lipizzi: So you have a someone doing a a given activity. The same activity can be done by multiple things. So the same person I can do multiple activities.

43:16

 Carlo Lipizzi: So, looking from the outside. That is a network like all the other. Inside, the the 2

43:29

 Carlo Lipizzi: sets are semantically this joint, so you need to lay bowl up without the notes in one set, and what that the notes on the other set. But once you do that, then when you connect them.

43:37

 the network that you have is what is is called the bypass.

43:55

 Carlo Lipizzi: From the bipartite you can extract the it's called one mode where

43:59

 Carlo Lipizzi: there is no 2 sets, but it is just one. You can have a a network, all the people only a network of words only what's or a backgrounds.

44:07

 Carlo Lipizzi: who works are connected, if used by the same person.

44:23

 Carlo Lipizzi: because we then go with the unique words, the same word that can be used by multiple people like creating networks that that can be somehow complex, like the example that you have on the top right and and is a the other example.

44:28

 Carlo Lipizzi: Once you have this network, this network as a

44:47

 Carlo Lipizzi: words or engrams only

44:53

 Carlo Lipizzi: is a W. What is called a semantic network.

44:59

 Carlo Lipizzi: Someone, in some cases they call it a knowledge graph

45:04

 Carlo Lipizzi: I it

45:10

 Carlo Lipizzi: because you have all what's works in a broad sense. Again, it could be a single words or multiple words like diagram and gram in general.

45:11

 Carlo Lipizzi: Once you have that you can apply class setting. I'll go it so, and there are I'll go. It's for clustering in graphs, so the most popularly use the is the lobby and community, the detection method and basically you have a

45:25

 Carlo Lipizzi: classes awards. Those classes awards are potentially topics, 2

45:44

 Carlo Lipizzi: because those are worth, so that are highly interconnected. I had the student validating the topics created in in this way, and for using on the last topics the accuracy, I mean that

45:49

 Carlo Lipizzi: he did it manually on that 5,000 tweets, and then compared the results with my script, and the accuracy was a close to 90%.

46:09

 Carlo Lipizzi: So you have topics at that point.

46:22

 Carlo Lipizzi: If you analyze in time those topics, you can see how the conversation is evolving in time.

46:25

 Carlo Lipizzi: So we applied the to

46:32

 Carlo Lipizzi: quite a lot of different events. I. I share the scripts with people

46:35

 Carlo Lipizzi: in many places around the world, getting, I mean, and validation of the results.

46:42

 Carlo Lipizzi: So this case, I mean, I consider that we are talking about. A few years ago I collect it. So this was a 2,015, but the concept is pretty much the same, was the announcement of what at the time was the new apple watch.

46:51

 Carlo Lipizzi: I collected the 700,000 tweets. There was a quite wide, less coverage for

47:13

 Carlo Lipizzi: everything that was presented in each particular moment of the event.

47:23

 Carlo Lipizzi: So I did the what I presented before, so I created the the relational network, then the semantic network.

47:31

 then I I They do the preparation.

47:43

 Carlo Lipizzi: And then I created the visualizations. So, instead of a timeline with the time I use the the single events

47:49

 Carlo Lipizzi: meaning the introduction on the apple TV, the iphone, the new functions for Macbook, and then finally the apple Watch.

48:00

 So.

48:12

 Carlo Lipizzi: and then I try to find a way to connect those. But this is not the the one that I want to spend more time. I want to spend time

48:14

 Carlo Lipizzi: right here.

48:24

 So

48:26

 Carlo Lipizzi: I was mentioning that there are 2 networks, the relational network and the semantic network. One of the metrics did they extract it from the graphs is the clustering, coefficient.

48:27

 Carlo Lipizzi: The clustering coefficient is a a way to measure the homophily of the network in a given time, meaning how connected the are the notes?

48:43

 Carlo Lipizzi: So if you look at this spy. Here you have a a high value for the clustering coefficient for both the relational and and the semantic networks.

48:55

 Carlo Lipizzi: That was the moment when they introduced the the new Macbook. So W. What is the interpretation? The interpretation is a. At that point. People where

49:08

 Carlo Lipizzi: talking about the the the same things. So the new keyboard that that was a total of failure, but was a the new keyboard, the new, the new monitor, so few topics

49:23

 Carlo Lipizzi: gathering, collecting the attention of the the Macbook enthusias. So it it was a people gathering together and talking about a few things.

49:42

 Carlo Lipizzi: This is when the the apple watch was introduced. So you have a relatively high, clustering, coefficient for the relational network, meaning the

49:59

 Carlo Lipizzi: people gathering together, talking about something. But the something it was all about the places, because they didn't know the product. So it was a.

50:11

 Carlo Lipizzi: It's good. I like it. It's cool. Not only knows how much it's going to cost. So things like that it's more on the as more to.

50:23

 Carlo Lipizzi: We use the same approach, to calculate or to evaluate the radicalization of a groups of people.

50:32

 Carlo Lipizzi: So we measure the in time, the clustering, coefficient, and the semantic, I mean the costing coefficient for both the semantic network and the relational network, and we created the a a composite index

50:44

 that is kind of a the summation. They normalize the summation of the 2.

51:00

 Carlo Lipizzi: If you analyze in time, when you see that there is a growth on this composite index.

51:07

 Carlo Lipizzi: There is a tendency to the radicalization, because you have a more people gathering together and talking about few things. So that's a radicalization. We use that to analyze

51:18

 Carlo Lipizzi: violence related to political elections in Kenya was a few years ago.

51:37

 Carlo Lipizzi: Okay, so this is an example. I don't want to spend too much time on that, but I just wanted to give you

51:47

 very briefly. This was a. On

51:56

 Carlo Lipizzi: predicting, so I collected data about 2 million tweets, 22 movies over a period of a few months

52:01

 Carlo Lipizzi: a. And then I I wanted to predict the either one on those 3. So

52:14

 Carlo Lipizzi: the box office revenues 3D score audience score. So we ended up not having much on both of the critics courts and the audience for a little bit on on the audience core, but definitely nothing with the critics score.

52:20

 Carlo Lipizzi: So we collected the let me go very fast. This. So we collected the all the data on the

52:40

 Carlo Lipizzi: then we created some prediction modeling. So in this case is a decision 3. But we also tried a different models.

52:51

 Carlo Lipizzi: Then we put together all the models in one chat. So we created that for categories of metrics, metrics on the sentiment, medics on the traffic that was not the number of tweets, but number of tweets per time unit

53:02

 Carlo Lipizzi: social meaning, the clustering coefficient for the social, and then some semantic medics.

53:20

 Carlo Lipizzi: So and we I played with the combination of all of those.

53:29

 Carlo Lipizzi: So the very end. One thing that is interesting, that that sentiment is is not a good predictor.

53:35

 Carlo Lipizzi: meaning using sentiment that you cannot really, I mean, in this case, was predicting the the box office revenues after week. You cannot really predict the the box office revenues based on sentiment. What people say is not what people do

53:43

 Carlo Lipizzi: so, but traffic, social and semantic. A combination of those is definitely what it it's more over an indicate or what people is going to do.

54:01

 Carlo Lipizzi: and then we applied to other cases. I don't want to spend too much time on that.

54:17

 Carlo Lipizzi: That's another application we try to predict.

54:26

 Carlo Lipizzi: to calculate for the the emotion generated by artists. I mean

54:31

 the

54:41

 Carlo Lipizzi: target was a music. So artist's, albums, and songs, using the Bluchnik emotional wheels. So there are those major emotions.

54:42

 Carlo Lipizzi: 3 shades, and we try to analyze the proximity of all the the the the

54:59

 Carlo Lipizzi: I mean. We can see that the comments that people left. We had a huge data set of comments, the

55:07

 Carlo Lipizzi: correlation somehow, between comments and emotions, because in this way you could eventually create a sort of a emotional DNA. All the the artists songs and albums, and then eventually

55:17

 Carlo Lipizzi: see if there are correlation between the number of all booms, a number of songs sold, and the emotional partners.

55:38

 Carlo Lipizzi: So there was another application. We use the

55:51

 so we for use on

55:57

 a fan of those 3. That's why we selected them.

56:05

 Carlo Lipizzi: and those are the numbers of the a number of artists, lyrics, rooms, and members on this, on this website that is no more active.

56:09

 Carlo Lipizzi: So again.

56:23

 Carlo Lipizzi: in this case we use the natural language toolkit. That was not a a a great way to do things so natural language toolkit. We mentioned that has a

56:24

 Carlo Lipizzi: whatnet where you have this taxonomy, and you measure the distance between what's and we measure the distance between the emotions. So the what to represent in the motions and the words in the comments like tools that you see here.

56:39

 Carlo Lipizzi: But the problem is a word net is relatively old.

56:58

 Carlo Lipizzi: and a comments are younger, and the match was not good. Do we revise it in that in a different way, and we didn't use a natural language toolkit. And I think we're definitely better.

57:03

 Carlo Lipizzi: Okay. So let me jump up to the other project that I was mentioning.

57:20

 Carlo Lipizzi: That is, this 1010. Not that was the internal lacking code for this project, and a again, the project

57:28

 Carlo Lipizzi: was a about getting metrics out of a a text.

57:47

 Carlo Lipizzi: So let me skip all of it, and let me go here. So the main issue, as I was mentioning before, was on the how to calculate the semantic matrix out of the

57:57

 Carlo Lipizzi: a text.

58:17

 Carlo Lipizzi: So that's why we I created the this

58:19

 Carlo Lipizzi: room theory.

58:24

 Carlo Lipizzi: So

58:25

 Carlo Lipizzi: let me

58:30

 Carlo Lipizzi: skip some of those and let me go here. So room theory. It's based on the framework theory that was created by Marvin Minsky in a mid or seventies.

58:31

 Carlo Lipizzi: The example that that means he use the was a. When you enter

58:48

 Carlo Lipizzi: into a room, you know, right away.

58:56

 Carlo Lipizzi: If a is a a bedroom, a bathroom, a a kitchen, not because there is a label saying a path from kitchen or bedroom, but because there are things in the room that it's kind of a a resonate with a classification that that

59:00

 Carlo Lipizzi: meaning a framework that you have in mind. There was a no computational component on it at that time.

59:20

 A.

59:30

 Carlo Lipizzi: But I like the idea of this framework approach. I added the a. A computational layer to it, and I

59:32

 Carlo Lipizzi: named it the room theory, using the example that Marvin Minsky was using.

59:46

 Carlo Lipizzi: So those rooms are a representation of the knowledge of the individual or the specific knowledge of the individual

59:52

 Carlo Lipizzi: entering the room or doing something on a given domain.

1:00:05

 Carlo Lipizzi: I use the at the time what to back is a form of a victorization. It is one of the for not the first, but one of the first of the new generation of a victorization of text. So you basically have

1:00:12

 Carlo Lipizzi: an application of the Meta for a of the so to what so are related.

1:00:33

 Carlo Lipizzi: If they up here somehow together.

1:00:44

 Carlo Lipizzi: What to back is calculating the conditional probability of one word, the appealing because of the other is using a a shallow neural network for doing that

1:00:48

 Carlo Lipizzi: we are. Charlotte means

1:01:02

 Carlo Lipizzi: one hidden layer. So you have input. Layer output layer, and he delay it in this case is one in the layer.

1:01:09

 Carlo Lipizzi: So

1:01:21

 Carlo Lipizzi: you have this conditional probability, meaning that you start with the matrix. That is a and we're in the unique words in the text that you are using by N. And you have numbers. Then you apply a reduction in the dimension.

1:01:23

 Carlo Lipizzi: like principal component analysis. then you have.

1:01:43

 Carlo Lipizzi: that is, in by whatever it is, the number of

1:01:48

 Carlo Lipizzi: typically we use a 200 or 300, meaning that each word or anram will become a a, vector with the 200 or 300 components.

1:01:52

 Carlo Lipizzi: meaning the words in the text that will become a points in a n dimensional space. Now, if you go down from n dimensions to 2 dimensions. You have a a a Cartesian space, and you have a points on the Cartesian space, representing the words, the more the words are closer.

1:02:05

 Carlo Lipizzi: the more they are logically related. So that's basically what is behind the the vectorization, then I mean

1:02:33

 Carlo Lipizzi: with the attention it's all you need, the they use the potential mechanism to do better, that just the conditional probability awards. And

1:02:44

 Carlo Lipizzi: but the concept, I mean the result. It's creating a a matrix representing a the text in that, no matter of terms.

1:03:02

 Carlo Lipizzi: So you have a text transform into numbers.

1:03:15

 Carlo Lipizzi: And that's basically how the room theory would work. So you have a

1:03:22

 Carlo Lipizzi: your room. That is a a numerical representation of your knowledge base meaning. You collect as many documents as as possible. You victorize the documents, and this is basically a Madrid

1:03:29

 Carlo Lipizzi: representing the knowledge on the specific domain.

1:03:47

 Carlo Lipizzi: Then you have words or engram so defining your your

1:03:51

 Carlo Lipizzi: You are a elements all the attention, because I I mean you as an individual, I can do different things.

1:03:59

 Carlo Lipizzi: So with the same knowledge you can do multiple things if your fog is the I don't know, in the financial aspects your fog is will will be on. I think so. The that are on the financial side. If you are in project management you are more on a those words.

1:04:11

 Carlo Lipizzi: So the the the benchmarks are our collections of of keywords, and eventually wait the all the different keywords. Then you have a document that you want to analyze. You scan at the document what by word?

1:04:28

 Carlo Lipizzi: And you calculate the proximity, all the words with the the the words in the benchmark, doing a look up getting the vectors from the the madrics. That is the room, and then the the result will be the

1:04:45

 Carlo Lipizzi: the distance between the document and each one of the benchmarks.

1:05:03

 Carlo Lipizzi: So in a symbolic terms, you basically have your corpus

1:05:08

 Carlo Lipizzi: that many documents representing the knowledge that that will be a victorized. Then you have a 2 elements that you have the keywords in the benchmark, and you have the the document that you want to analyze. So

1:05:16

 Carlo Lipizzi: what you do is basically what? By what? That you take one word from the text to evaluate one word from the benchmark, and you calculate the distance

1:05:34

 Carlo Lipizzi: that most simple example of calculating the distance is the cosine a similarity.

1:05:45

 Carlo Lipizzi: So you calculate the proximity that is a number, and this will tell you how much each word is a far from each one of the benchmarks.

1:05:52

 Carlo Lipizzi: and then you add all of them.

1:06:04

 Carlo Lipizzi: You normalize the the results, and you have the similarity between the document and the

1:06:08

 Carlo Lipizzi: and and the in each benchmark and the cumulative value for all the benchmarks.

1:06:18

 Carlo Lipizzi: meaning that at this point I can tell that one document is more on a given technology or another technology. So meaning at this point, i'm really getting the values.

1:06:26

 Carlo Lipizzi: So that was a the way we use the for all of this. Then

1:06:43

 Carlo Lipizzi: we created the a. A. A process, a pipeline. So you have a gathering the data meaning You have a news patents, papers doing a little bit of a preparation, going into a Mongodb.

1:06:52

 Carlo Lipizzi: creating a the rooms or or just analyzing it. And then the results, meaning the the actual metrics will go in in a relational database we use.

1:07:09

 Carlo Lipizzi: but it could be Mysql or or any other, and those will go with the use, the by, the 2 systems, this panel, and then another system for monitoring technologies.

1:07:26

 Carlo Lipizzi: We will go back to the applications. But the goal for today is just on how to extract metrics from text.

1:07:40

 Carlo Lipizzi: So I I would stop here, and I will go back to examples. So

1:07:53

 Carlo Lipizzi: next week.

1:08:01

 Carlo Lipizzi: But I just want to jump now into a chat, Gpt.

1:08:03

 I'm. Using a slides that I presented.

1:08:12

 Carlo Lipizzi: Here we go.

1:08:16

 Carlo Lipizzi: that I presented the to a workshop that they gave 2, 3 weeks ago

1:08:18

 Carlo Lipizzi: so large language models. So we mentioned a a chat gpt that is, a Gpt based on a transform that's so the transform that again are sort of the next generation to what? To back

1:08:29

 Carlo Lipizzi: ere this is an essential step. You cannot really do much if you do not do the transformation.

1:08:49

 Carlo Lipizzi: But anyway, so erez agmoni, I mentioned also the fact that what to back was a shallow neural network, meaning you have one hidden layer one

1:09:09

 Carlo Lipizzi: when you work with a neural network. So, and you already know that you have parameters meaning weights that you give to each one of the inputs. So those weights are are normally called parameters. So when you see those numbers.

1:09:19

 Carlo Lipizzi: Gp: 3 as 175 billiona parameter, so those are the weights in the the network. That is an indication of how complex is the model.

1:09:38

 Carlo Lipizzi: So we are going into several 1 billionparameters.

1:09:50

 Carlo Lipizzi: So, Bertha, that was a one on the first using transform. That was a 340 millionNow we have more than the same number in billions.

1:09:56

 Carlo Lipizzi: So when you train those models like Chat

1:10:07

 Carlo Lipizzi: Gpt, the training? It's a really an intensive process and an intense I mean a a very time consuming a resource, consuming a process.

1:10:15

 Carlo Lipizzi: There are a few months in the loop for cleaning the data, meaning eliminating all the in appropriate content, but also to start the tagging. So we will go back eventually on that. How J. G Gt. Was trained

1:10:29

 Carlo Lipizzi: if you look at the number. So at the bottom, on the page 4.6 million.

1:10:48

 Carlo Lipizzi: It was the cost to train the the board.

1:10:56

 Carlo Lipizzi: the energy that was used because I I mean it's a lot of computing time. They use the

1:11:01

 Carlo Lipizzi: quite a lot of N. Media Gpu Graphical processing units a. A. And they use a lot of energy. So

1:11:09

 Carlo Lipizzi: the amount of energy is enough to power a more than 30,000 American households for a day.

1:11:20

 Carlo Lipizzi: So that's how

1:11:28

 Carlo Lipizzi: I mean that energy, intensive is the training keeping in mind. The human mind is using a fraction of a fraction of that

1:11:31

 Carlo Lipizzi: much energy for the entire life. meaning that those systems are highly inefficient. So we are applying brute force to an algorithm that

1:11:42

 most likely it's not really representing the way we reason.

1:11:54

 Carlo Lipizzi: So those are the sources. So there are

1:11:59

 an estimated 45 data bytes of text

1:12:06

 Carlo Lipizzi: That's the distribution. So if you look at the distribution, you have a quite a lot of English, and that's an intrinsic buyer, so that you have in Chat G. G Gpt. Meaning that other languages are are under represented.

1:12:10

 Carlo Lipizzi: then that Yes, the web is pretty much in English, but that doesn't mean that that you don't have a good representation of the other languages.

1:12:27

 Carlo Lipizzi: How we evaluate the how they it's very difficult to say so in the 2,016.

1:12:41

 Carlo Lipizzi: This lambda the data set. That was a

1:12:51

 Carlo Lipizzi: use the as a benchmark for a task oriented the

1:12:58

 Carlo Lipizzi: language understanding so based on that.

1:13:04

 Carlo Lipizzi: I mean, they are pretty good compared to humans.

1:13:09

 Carlo Lipizzi: It it's just one of the ways to measure it. When you see 0 short, few short, the 0 shot models with no training data for that particular task.

1:13:13

 Carlo Lipizzi: few short meaning. You have some examples. So 0, short and not very effective

1:13:31

 Carlo Lipizzi: limitations. We we really need to understand how they work. So they work as a a matching patterns.

1:13:39

 Carlo Lipizzi: So you have the training generating

1:13:51

 Carlo Lipizzi: with the some human intervention, and then you have your quidy.

1:13:56

 Carlo Lipizzi: and what the B is doing is matching the the pattern in your query with the pattern, so that

1:14:04

 he does.

1:14:14

 Carlo Lipizzi: and then, once it does, the match is adding a a presentation, a a a conversational layer. To present the results.

1:14:15

 Carlo Lipizzi: The more data you provide in terms of you are quitting.

1:14:27

 Carlo Lipizzi: the more reach would be the answer, because there will be more elements to match. So that's a a key point.

1:14:35

 Carlo Lipizzi: So first of all.

1:14:44

 Carlo Lipizzi: those so far

1:14:49

 Carlo Lipizzi: are pretty much the same as a Google with the difference that you have a compilation, or or the answers, and a a a, a, a, a a presentation of the results in a plane, English or a plane, whatever is the language that you are using.

1:14:52

 Carlo Lipizzi: and that's valuable. But you are losing the reference to the sorts.

1:15:12

 Carlo Lipizzi: meaning that if you are using a chat gpt, or another Llm. Or something that that we go public.

1:15:18

 Carlo Lipizzi: There are chances that someone can sue you for a.

1:15:28

 Carlo Lipizzi: because you don't know what the source is going to be. so I will show you in a moment the the way i'm using it.

1:15:37

 Carlo Lipizzi: So again.

1:15:46

 Carlo Lipizzi: It is not intelligence, but is a a a a nicer way to do what Google was doing, since more than a decade.

1:15:48

 Carlo Lipizzi: But it's great. I mean you have a a plain English interface.

1:16:02

 Carlo Lipizzi: i'm starting a project on having 2 projects. But before I go that let me go in another element that is relevant in in terms of limitation or or

1:16:09

 Carlo Lipizzi: and they are not domain specific

1:16:23

 Carlo Lipizzi: meaning, considering how they had been. Train the and those are the sources

1:16:28

 Carlo Lipizzi: you you cannot have a specific knowledge in one particular area. If you are in the defense industry.

1:16:36

 I mean that.

1:16:43

 Carlo Lipizzi: apart from the fact that some of the material is restricted, and for sure, it is not a public domain, but even in the in the public domain that are a several elements that are that could be potentially relevant if you want to do something in the defense industry. But in this case those elements there will be a drop in the ocean.

1:16:45

 Carlo Lipizzi: meaning the answer so that you will get that will be highly deluded, and you will not get the the knowledge from one specific domain.

1:17:06

 Carlo Lipizzi: So we we are working on 2 projects one to create an for the school of systems and enterprises, and I will start using 6, 24 as a an example.

1:17:19

 Carlo Lipizzi: the and we'll be sort of a Tudor out automatic tube, or for a 624, and I will use the the transcripts of my classes along with the articles, papers and textbooks.

1:17:36

 Carlo Lipizzi: and then eventually I will expand the to other courses. I will initially focus on 2 of the most popular courses that are 6, 24 and 6 12, but it is project management

1:17:55

 Carlo Lipizzi: the second thing that then doing a second project is to use the

1:18:10

 Carlo Lipizzi: the presentation. Uhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh

1:18:20

 Carlo Lipizzi: so it's like the next up blow instead of having visualization. So I have plain English.

1:18:30

 Carlo Lipizzi: Think about the the defense industry. You have people on the field, and they need to take decisions. They input the the values from the environment, and they have a scenarios.

1:18:38

 Carlo Lipizzi: But yes, it's a scenario. But at this point you have more either numbers or graphs. It would be great to have something in plain English. So once you have the conversational element that transform me, the conversational element from a written text to

1:18:55

 Carlo Lipizzi: oral text with the an automatic reader they can read. It is a no brain. And I mean that technology is there since 10 years.

1:19:18

 Carlo Lipizzi: So that's something I i'm working on.

1:19:29

 Carlo Lipizzi: I mentioned the mixed approach very briefly.

1:19:35

 So

1:19:40

 Carlo Lipizzi: I just want to go here for a moment.

1:19:45

 Yeah.

1:19:49

 Carlo Lipizzi: E: we are going a apparently to what the sort of so on one side the you have what is called the the prompt, the engineering again, a a a pattern matcher.

1:19:50

 Carlo Lipizzi: the better is your query. The more detailed, the more patents you are providing and the better output you are going to get.

1:20:04

 Carlo Lipizzi: There are jobs for prompt engineers. So people knowing how those lms are working, and on the other end the experts of the domain able to provide those input

1:20:16

 Carlo Lipizzi: it's basically like being a a super. User so that's a type of skill that is not

1:20:35

 Carlo Lipizzi: deep into the technology, but more on the use of the technology. And then you have on the opposite side what I call the the cognitive engineering, where you have those who will develop the next generation of of large language models.

1:20:45

 Carlo Lipizzi: I mean that the way they work now is is not efficient, and it's going to be changed. So that is no representation of the knowledge that is essential if you really want to do something that with an intelligent behavior. But we don't know how to do it.

1:21:05

 Carlo Lipizzi: So this discipline that will be at the crossroad between math, the cognitive science software, engineering, coding, the abstract thinking is what will generate jobs, not many, because it would be very challenging, but

1:21:22

 very critical jobs.

1:21:43

 Carlo Lipizzi: So briefly.

1:21:46

 Carlo Lipizzi: that's the way i'm using a chat Gpt: so for your information.

1:21:50

 Carlo Lipizzi: i'm writing a book that is called that I mean the name is still to be fully defined. Societal impacts of artificial intelligence and machine learning.

1:21:57

 Carlo Lipizzi: So th the idea is how artificial intelligence machine learning is is used right now, and what the impact could be to the society.

1:22:09

 Carlo Lipizzi: So the first question was, is this a revolution? So what is a revolution? How you measure a revolution. So

1:22:21

 Carlo Lipizzi: in that the the history of humankind that we had several social revolution, including the Industrial Revolution or the Dj. The the Revolution, are those potentially

1:22:30

 Carlo Lipizzi: the same impact that could the AI machine learning revolution be.

1:22:48

 Carlo Lipizzi: So I had to create somehow a a little bit of a

1:22:56

 Carlo Lipizzi: background. So I use the as a writing body Chat.

1:23:04

 Gpt.

1:23:13

 Carlo Lipizzi: I created the chats. So each chapter is a a

1:23:15

 Carlo Lipizzi: plenty of information. Each one is giving a pattern to match. So when I place a question

1:23:23

 Carlo Lipizzi: over here

1:23:35

 Carlo Lipizzi: over over here. I will get different answers

1:23:38

 Carlo Lipizzi: from another one.

1:23:43

 Carlo Lipizzi: So

1:23:45

 Carlo Lipizzi: that's

1:23:49

 the user that I encourage you to do

1:23:51

 Carlo Lipizzi: so. If you and place a query that is a a straightforward.

1:23:54

 Carlo Lipizzi: you will get an answer. That would be not much better that what you could get from Google.

1:24:02

 Carlo Lipizzi: But if you have something that is more complex, you will have answers that will be more complex. So keep in mind that. and

1:24:12

 Carlo Lipizzi: I get no problem. If you use a just gpt as a sort of a a in your development.

1:24:27

 Carlo Lipizzi: Don't, do copy and paste because you are doing your itself not a favor.

1:24:41

 Carlo Lipizzi: When you will get a job, I mean, down the road. The those boss will be so powerful that everybody and and so widespread that everybody will using it.

1:24:48

 Carlo Lipizzi: But if you use them now, let's say for coding.

1:25:00

 Carlo Lipizzi: keeping in mind that the chat G. G. Gpt was 3 in the on a github because Microsoft invested 10 billiondollars in open AI, and Microsoft is also your own of Github.

1:25:04

 Carlo Lipizzi: meaning all the code that is in Github, but somehow was used to train a chat. G.

1:25:19

 Carlo Lipizzi: If you use it for coding the you may or may not get a a a a good piece of code.

1:25:29

 Carlo Lipizzi: how you decide that if it's good, or if it's bad. you basically don't

1:25:39

 Carlo Lipizzi: unless you know how to code. So i'm introducing a a chat Gpt now, and not at the beginning of the course, because if you use it now, now, you can even using it for coding. But you know

1:25:47

 Carlo Lipizzi: what is right and what is wrong. So

1:26:05

 Carlo Lipizzi: also keep in mind that we will apply the same concept. The the assignments are individual. If any of you will do the same assignment in terms of coding. Each one will get the the to dollar divided by N.

1:26:10

 Carlo Lipizzi: Whatever is the tool that you are using. If you are just using someone else, or if you are using a chat G for it.

1:26:28

 Carlo Lipizzi: there is no way to detect, as today to detect if an output is generated by you, man, or from a boat.

1:26:39

 Carlo Lipizzi: So but again use it at this point. Use it to do better what you are doing.

1:26:49

 Carlo Lipizzi: Okay. So a again. There is not much time for it's a 7 min to 8,

1:27:00

 Carlo Lipizzi: and

1:27:13

 Carlo Lipizzi: erez agmoni there is not much time for the in-class exercise. But I want to introduce the the exercise and give you a little bit over 150 sense of what's going on, and then I will give you a few minutes just to

1:27:15

 Carlo Lipizzi: again familiar with it.

1:27:32

 So the

1:27:36

 Carlo Lipizzi: the data set

1:27:39

 Carlo Lipizzi: is one of the data sets, and I downloaded the from a tweed. So in particular, it's a

1:27:42

 Carlo Lipizzi: 20,000 rows, I mean. Tweets are more complex. But what I keep in the data set is a sender timestamp and tax.

1:27:54

 Carlo Lipizzi: and this is what you have here. So you have the sender. This is a

1:28:06

 a timestamp, but that is using what is called the epoch time.

1:28:12

 Carlo Lipizzi: That is a I mean some node that created a few years ago. That is, a number of seconds from the first introduction of unix.

1:28:17

 Carlo Lipizzi: But I mean there are functions, so to transform that number into a real time.

1:28:27

 Carlo Lipizzi: and then you have a in the text. So you have rt for retweet. You have the app for mentioning.

1:28:36

 Carlo Lipizzi: and so on.

1:28:47

 Carlo Lipizzi: So

1:28:49

 Carlo Lipizzi: you want to print the 5 most active sender, the 10 most retweeted, the tweets, the 5 most sided the screen name and the 10 most popular hashtags

1:28:53

 Carlo Lipizzi: so retweets Again, you cannot recognize them because there is Rt.

1:29:08

 Carlo Lipizzi: The

1:29:16

 Carlo Lipizzi: screen name. It's in the text is with the at sign.

1:29:18

 Carlo Lipizzi: and the hashtag as the

1:29:25

 Carlo Lipizzi: like over here and there are a few others

1:29:29

 here and there.

1:29:35

 Carlo Lipizzi: Okay. So I don't even create the the breakout rooms. So because it's a 7, 56. So

1:29:39

 Carlo Lipizzi: at this point I just want to be sure that you have everything published. So

1:29:56

 be sure

1:30:04

 Carlo Lipizzi: that you have what you need.

1:30:07

 Carlo Lipizzi: Yeah, you should have everything.

1:30:14

 Carlo Lipizzi: But I will post the solution for this in-class assignment in a moment.

1:30:21

 So we definitely don't have time for working on it

1:30:31

 Carlo Lipizzi: my apologies. but I hope it was useful.

1:30:35

 Carlo Lipizzi: So let me go into the next exercise.

1:30:39

 Carlo Lipizzi: That will be Exercise Number 11. So that's basically, what you have

1:30:47

 Carlo Lipizzi: when you click on it, and

1:30:56

 Carlo Lipizzi: instead of having a like that, that, you will have it

1:30:59

 Carlo Lipizzi: as a file.

1:31:07

 Carlo Lipizzi: So the file there is a spring 23 wn dot Docs

1:31:09

 Carlo Lipizzi: is a

1:31:15

 Carlo Lipizzi: is this one. So

1:31:19

 Carlo Lipizzi: I created that this assignment today. So if you see that there is something that it is not clear. Let me know, because I I

1:31:23

 Carlo Lipizzi: may have a a skip, something that that could be useful to you.

1:31:34

 Carlo Lipizzi: So one of the

1:31:40

 Carlo Lipizzi: considerations that some of the students in past semesters had on this course was, Why, Don't, we have more assignments that are more on how to use what what

1:31:44

 Carlo Lipizzi: we did

1:32:03

 Carlo Lipizzi: for managing situations

1:32:04

 Carlo Lipizzi: so. And that's why I created the this part. One of the things that I was managing in the past few years was a

1:32:09

 Carlo Lipizzi: the scheduler courses. So scheduling courts is in a university is not an easy task. Because you have a students. You have courses, you have instructors.

1:32:18

 Carlo Lipizzi: and you have classrooms.

1:32:32

 Carlo Lipizzi: So you need to match all those components in a way that you are serving the students and the best way possible. But you are using your faculty in the proper way 150,

1:32:34

 Carlo Lipizzi: meaning that each faculty as a teaching load, and you want to be sure that they fulfill that they load the they load. So in the proper way.

1:32:46

 Carlo Lipizzi: A. You need to be sure that the classrooms will be not too big, so there are many moving paths.

1:32:56

 Carlo Lipizzi: As you know it's even so. We have a a platform that is used for managing a the the community, somehow, that is called Work Day, with what they we can extract the the list of courses.

1:33:07

 Carlo Lipizzi: and the list would look pretty much similar to this one.

1:33:27

 Carlo Lipizzi: So you have a what is the academic units, the section, and a bunch of other information. distract or name or name, sir. did they

1:33:34

 Carlo Lipizzi: anonymized so instead of the name, you have numbers, so it's. Tractors are those numbers.

1:33:46

 Carlo Lipizzi: and then that the credit our what for each one and and and then you have a

1:33:55

 Carlo Lipizzi: some other information that they added.

1:34:03

 Carlo Lipizzi: actually just to be very open.

1:34:06

 Carlo Lipizzi: So the the actual file that we download the from what they is a 5, with all the courses from all the schools.

1:34:18

 Carlo Lipizzi: So I basically had to do a little bit of homework to cleaning it. To do this cleaning what I did that was creating a a a a a

1:34:28

 Carlo Lipizzi: configuration file.

1:34:41

 Carlo Lipizzi: That is this

1:34:44

 Carlo Lipizzi: where I have the name of the file. Then I have the

1:34:47

 Carlo Lipizzi: list of a names over a companies. We serve. What is the semester programs.

1:34:54

 Carlo Lipizzi: and then sort of a widely some black lease our courses that they want to keep or eliminate. And then the threshold for a

1:35:02

 Carlo Lipizzi: we consider as accountable courses with more than 8 students.

1:35:13

 Carlo Lipizzi: So, using those criteria, I

1:35:21

 Carlo Lipizzi: I created this script

1:35:26

 Carlo Lipizzi: to extract from everything, then what they is providing this script. That is a a reduced version. And it also added the level and program.

1:35:29

 Carlo Lipizzi: while i'm doing the calculation. So that's the data set that that you will use

1:35:43

 Carlo Lipizzi: on this data set. Those are it's 94 rose. This is related to to

1:35:50

 Carlo Lipizzi: last year spring 2,022. So those are the

1:35:59

 Carlo Lipizzi: the, the, the the rules, the the columns that you have meaning they attributes. And

1:36:06

 Carlo Lipizzi: you want to. The overall goal is to

1:36:15

 Carlo Lipizzi: get a sense of how things are going in that semester, eventually creating a dashboard to make sure that we are monitoring what's going on, and those are some of the elements that could that give an indication

1:36:23

 Carlo Lipizzi: printer the 5 courses with the highest number of students. So the number of students is in this column here named. The enrollment count.

1:36:41

 Carlo Lipizzi: Bring the 5 instructors from this column. Here instructors teaching assistant

1:36:53

 Carlo Lipizzi: with the highest number of students. Again, students are from enrollment count. Compare the total number of or students for a undergraduate graduate and a corporate, meaning that the corporate education

1:37:02

 Carlo Lipizzi: you will get that from

1:37:21

 Carlo Lipizzi: level.

1:37:24

 Carlo Lipizzi: So level is standing. Ug, undergraduate G Graduate Corp Corporation. You will skip the mixed ones.

1:37:25

 Carlo Lipizzi: Comparing that means, calculate the values so, and describe the results in a narrative in in the narrative path of the assignment.

1:37:37

 Carlo Lipizzi: Compare the number of courses that run at full capacity. So you have a this

1:37:48

 Carlo Lipizzi: that is a section status that will tell if it's open

1:37:57

 Carlo Lipizzi: or close the so if it's closed, meaning they are at capacity, open their top.

1:38:03

 Carlo Lipizzi: create a pie, chat with the distribution of students supper program a pie sh with the distribution of students per type of delivery.

1:38:11

 Carlo Lipizzi: You have a.

1:38:21

 Carlo Lipizzi: You have this.

1:38:24

 Carlo Lipizzi: this delivery mode in person or online.

1:38:27

 Carlo Lipizzi: and perform any other analysis that could make a sense to better monitor the the semester.

1:38:32

 Carlo Lipizzi: So you would read the file. You do the analysis. You will submit the the script and the interpretation of the So again, they need to be original, not describing the the process, but describing the the findings.

1:38:39

 Carlo Lipizzi: So what we are doing now is basically to use the script to get insights and eventually to take actions. So you have a

1:38:56

 Carlo Lipizzi: programs that are too small. You have a

1:39:08

 Carlo Lipizzi: classes that are too big. You have some instructors that are overloaded things like that. So all of those will give you somehow data points to take a decisions

1:39:14

 Carlo Lipizzi: all right. It's 805. That's pretty much the end of the class Questions.

1:39:32

 Kevin Zeng: Hey, Professor? It's me again. Yeah, Can you Did you see my email regarding the quiz for module 10, I think there are some issues or issues with question 5 and 6. It was marked wrong. But I think it's right

1:39:42

 Carlo Lipizzi: question. 10

1:39:55

 Kevin Zeng: Quiz: 10.

1:39:58

 Carlo Lipizzi: Yeah.

1:40:00

 Carlo Lipizzi: Okay, I will check it definitely. Check it.

1:40:05

 Kevin Zeng: Thank you for us, sir.

1:40:09

 Kevin Zeng: Yeah. When when when you send that email it's again add the she, you as well, just to be sure that you double your chances.

1:40:11

 Kevin Zeng: The the data set is quite, I I guess

1:40:28

 Kevin Zeng: you're reviewed and approved right, though the one I I sent you regarding the the airplane. Okay.

1:40:32

 Kevin Zeng: And for the

1:40:38

 Kevin Zeng: what do you call it? Analysis? Is it performed as a group? Or do we have to submit our in the visual analysis, like, what is your expectation of

1:40:40

 Kevin Zeng: of that portion?

1:40:48

 Carlo Lipizzi: Well, when you do a a a a group project. There is no need to to do things different. You want to do one single report, one single script that will contain the contribution of everybody

1:40:49

 Carlo Lipizzi: keeping in mind that that

1:41:08

 Carlo Lipizzi: if you have more people the an at least it should be more complex. because otherwise I I couldn't really compare

1:41:13

 Kevin Zeng: something that was generated by 3 people with something that was generated by one. and and the expectation for the graphs like, how many

1:41:23

 Kevin Zeng: is too little, or what's your expectation for that.

1:41:34

 Carlo Lipizzi: Well, graphs are just one indication of the complexity. So sometimes it could be metrics. So you you saw the presentation, the the the the for the tweets

1:41:37

 Carlo Lipizzi: A. I mean, if you consider a the the clustering coefficient is a number. But getting there, it took me several 100 of lines of code.

1:41:53

 Kevin Zeng: Okay, good.

1:42:08

 Carlo Lipizzi: And that's complex city. So

1:42:09

 Carlo Lipizzi: I for use Mmm: more on the complexity of discrete than the complexity of the All Research. In a broad sense. Then the number of either a tables or

1:42:13

 Kevin Zeng: okay. Thank you, Professor.

1:42:28

 Carlo Lipizzi: Sure

1:42:31

 Carlo Lipizzi: other questions.

1:42:33

 Scott Guetens: Alright, I just yeah real quick. I just wanted to reiterate what Kevin said. I also had an issue with 5 and 6. So I'm: yeah, I have it. Yeah, yeah, I I i'm sorry if I didn't do much again. We are.

1:42:36

 Carlo Lipizzi: But yeah, no worries. I didn't even email, but I figured I would just mention it because he had also

1:42:54

 Scott Guetens: same time. Thank you.

1:42:59

 Carlo Lipizzi: Sure other questions tissues.

1:43:01

 Carlo Lipizzi: Okay. So it's 8 or 9. That's the end of the class. I hope you enjoyed the at least the part on that chat, Gpt, and I hope

1:43:09

 Carlo Lipizzi: that could be useful again in the future. We will use those tools more and more. E-pola is thinking about replacing with the

1:43:21

 Carlo Lipizzi: so we will have a sort of a

1:43:35

 Carlo Lipizzi: personal assistance that would be hopefully more smart than studio or or Alexa so, or whatever my likes just started.

1:43:38

 Kevin Zeng: Professor, Why are your thoughts on on Chat Gbt, and the future of jobs? Do you think it will always be always remain as a supplement or a tool for developers, or

1:43:49

 Kevin Zeng: do you think it will replace

1:43:59

 Kevin Zeng: people in the future?

1:44:02

 Carlo Lipizzi: It's a great question, no one as a real answer. But there are some indications. So

1:44:04

 Carlo Lipizzi: I mean, if you look at the history A.

1:44:15

 When you have a new tool, you have some jobs that will go away

1:44:20

 Carlo Lipizzi: typically are the jobs that are more repetitive jobs that are more easy to Peter Blaze.

1:44:26

 and that's what is going to happen with the chat.

1:44:35

 Carlo Lipizzi: So

1:44:41

 Carlo Lipizzi: you need to add value. So companies will pay you for adding value If you have something that is already providing value, you cannot just

1:44:43

 Carlo Lipizzi: use it and

1:44:56

 Carlo Lipizzi: pass the through. So you need to do better. and that's why I, for example, presented you how i'm using it

1:44:58

 Carlo Lipizzi: so by productivity. In writing the book improved

1:45:09

 Carlo Lipizzi: for coding would be the same, so you can use a chat gpt just to save the time that you would use the on a stack overflow.

1:45:18

 Carlo Lipizzi: But we not replace you when you have a problem that is really complex.

1:45:33

 Carlo Lipizzi: I mean, right now, we also have the limitation that the chat gpt is not taking fines, meaning that if you want to do a processing of a fine, not like the one in the current assignment. You cannot pass the file and have it process.

1:45:40

 Carlo Lipizzi: but you can have some help.

1:45:59

 Carlo Lipizzi: So you want to use that to do things better

1:46:02

 Carlo Lipizzi: if you think for a

1:46:07

 Carlo Lipizzi: You probably saw the movie. I don't remember the name, but the movie on NASA with the human calculators.

1:46:10

 Carlo Lipizzi: So an army of people doing manually the calculations for the satellites.

1:46:22

 Carlo Lipizzi: Then they introduced the computers and their job was gone. But the job of a coding started.

1:46:33

 Carlo Lipizzi: So once we will have Lms that that that that could be more powerful.

1:46:44

 Carlo Lipizzi: that will have an impact, so we need to be ready

1:46:54

 Carlo Lipizzi: to move up to be more expert of the domain if you want to go in the prompt engineering side more expert on how to improve the the quality of those things.

1:46:57

 Carlo Lipizzi: One of the things that we are working on is on that, using those tools to provide the the call. More knowledge

1:47:10

 Carlo Lipizzi: representing common knowledge is a pain in the neck, because we don't really know what it is. So we know things, because just we know it.

1:47:20

 Carlo Lipizzi: how you represent it.

1:47:32

 Carlo Lipizzi: So one of the ideas is to use those tools based on a such a large data set to provide the common knowledge, and then use the common knowledge coupled with the domain that specific knowledge that we create.

1:47:35

 Carlo Lipizzi: So you have a layer that is more complex.

1:47:53

 Carlo Lipizzi: Another thing that that we are working on is to use those Lms as a sort of dispatcher for a specific knowledge, so it is said, over getting the results.

1:47:58

 Carlo Lipizzi: They will tell us what is the the island of knowledge that we want to activate?

1:48:13

 Carlo Lipizzi: Then they dial and could be another model based on something similar to an Lm. Or something similar to my room theory, and you will get the answer from that.

1:48:20

 Carlo Lipizzi: So again that

1:48:36

 those are the examples on how to build on top of it.

1:48:37

 Carlo Lipizzi: But

1:48:43

 there will be an impact so lower level jobs will go.

1:48:44

 Carlo Lipizzi: But that is what happened in the Industrial revolution.

1:48:51

 Carlo Lipizzi: So who is using the faxes anymore? So people manufacturing taxes, they are gone.

1:48:59

 Carlo Lipizzi: Who is using the the analog photo

1:49:08

 Carlo Lipizzi: with all the chain or products that are related to that. Think about digital music and streaming. So we have a less movies less here that we have a

1:49:15

 Carlo Lipizzi: less Cds or similar things. So that's the way it is.

1:49:34

 Carlo Lipizzi: That would be an impact for sure. And we need to be prepared. We need to know how to use them. Is this an answer, Kevin.

1:49:40

 Kevin Zeng: Yeah, Absolutely. Thank you for us, sir. And and I also read that it's it. It's really bad math. Is there a reason why

1:49:53

 Kevin Zeng: I I heard it can't do calculations correctly.

1:50:02

 Kevin Zeng: like simple. Well, I mean that is a domain chat. Gpt is not good in any domain, but it's good in all domains and like.

1:50:05

 Kevin Zeng: Okay.

1:50:19

 Carlo Lipizzi: And there's the reason, I mean, if you ask questions on the the defense industry, if you ask questions on something

1:50:21

 Carlo Lipizzi: that is not genetic. You will not get much, because I mean, if you remember the sources that is being used by a chat

1:50:32

 Carlo Lipizzi: gpt. They are very generic sources

1:50:43

 Carlo Lipizzi: that could be good to represent the common knowledge, the there are some paypal. So on a. I mean.

1:50:47

 Carlo Lipizzi: what is the common knowledge? So the common knowledge is what we have in our mind, because we are using more and more digit.

1:50:58

 Carlo Lipizzi: We could assume that then everything that is available line now

1:51:06

 Carlo Lipizzi: is the equivalent of the common knowledge.

1:51:12

 Carlo Lipizzi: There's an assumption. Now. There are papers saying that is what

1:51:16

 Carlo Lipizzi: is a available as open source, this formation, or everything. Assuming that this could be possible to do on the summation of everything that is available as an open source. Is there a presentation of the common knowledge of people

1:51:21

 Carlo Lipizzi: I

1:51:40

 Carlo Lipizzi: we don't know for sure, because we don't know we can really quantify what what is available, and we can not quantify the common knowledge. Common knowledge is very depending on the culture and and your specific culture meaning is really difficult.

1:51:41

 Carlo Lipizzi: But we can assume that what is available in terms of common knowledge is what is available as common knowledge to to an average.

1:51:57

 Carlo Lipizzi: So for those things. those tools are good and will be even better. So we have a. A. Gpt. For our chat. G. G Gpt. Is based on a Gtt. Too.

1:52:09

 Carlo Lipizzi: By the way, keep in mind that the chat gpt is based on a data that that stops at 2,021, meaning whatever it is, after 2,021 is not there

1:52:22

 Carlo Lipizzi: meaning If you are asking, I don't know something that is happening that happened after that.

1:52:35

 Carlo Lipizzi: It's not like Google. It is pretty much real time, but it's bad

1:52:43

 Carlo Lipizzi: because of the training it's. I don't know how many millions, but it's not something that you do on a regular basis

1:52:48

 Carlo Lipizzi: into one week.

1:52:57

 Carlo Lipizzi: so it's a batch process. That's another challenge. So we need to move. We need to have a way to have more real time.

1:52:58

 Carlo Lipizzi: Human mind. It's real time. We learn

1:53:09

 Carlo Lipizzi: every moment that every every sale on the from whatever we do, and what we learn now will increase our our our knowledge. This is not the way those tools are working.

1:53:14

 Carlo Lipizzi: Anyway it it it it's a very large topic, and

1:53:28

 Carlo Lipizzi: I would be happy to address other questions You may have down the road.

1:53:35

 Kevin Zeng: That that will be all. Thank you.

1:53:42

 Carlo Lipizzi: All right. Okay. So thank you. All. Sorry for keeping you till 8, 20,

1:53:45

 Carlo Lipizzi: and I hope I address your questions. Feel free to send us an email life to anything else.