* Shiyu Yuan: Hi, Professor.

1:52

* Carlo Lipizzi: Hello, she you! How are you?

1:54

* Shiyu Yuan: Thank you.

1:58

* Carlo Lipizzi: I am on campus.

1:59

* Shiyu Yuan: Oh. yeah, in your office.

2:01

* Carlo Lipizzi: Yeah. Yeah. Yeah, yeah. I mean I wanted to do the class from home, but then I had medium things to do, and I couldn't do it.

2:05

* Carlo Lipizzi: Hello, Christina!

2:22

* Carlo Lipizzi: So it's 6, 29. We will wait another

2:31

* Carlo Lipizzi: couple of minutes.

2:39

* Shiyu Yuan: No problem.

2:40

* Carlo Lipizzi: 6, 29. Let's wait for 6, 30 to officially start in the class.

3:00

* Carlo Lipizzi: Okay, so now is a 6, 30, and we can officially start. So it's a April the eleventh.

3:09

* Carlo Lipizzi: and it's a 6, 30.

3:19

* Carlo Lipizzi: This is going to be the last class for this courts, when I will present something next class will be you presenting your finals.

3:21

* Carlo Lipizzi: and and that will be the end of the course. So I really hope that it was interesting and useful for you.

3:41

* Carlo Lipizzi: But we will talk next week about that. So for up this week we will change things a little bit compared to what was originally scheduled, and in particular we will have a

3:49

* Carlo Lipizzi: couple of things. One will be she you presenting a a portion of the research she's doing with me for our Phd program.

4:08

* Carlo Lipizzi: Then I will. A couple of

4:25

* Carlo Lipizzi: research is applications that I developed.

4:33

* Carlo Lipizzi: Then I would talk briefly about a couple of scripts that they want to show to you to see what you could do with the

4:37

* Carlo Lipizzi: I mean that in practical uses all the writing scripts. Then we will talk about the final. I will present again the template for developing it. I will go through some examples of the final.

4:49

* Carlo Lipizzi: There are a couple in particular that I just want to show you. I will not share with you the actual presentation, because is one of the topics that you could eventually a peak for your final.

5:09

* Carlo Lipizzi: and then I will give you 1015 min to work on an in class assignment that could be a little bit different from the usual. So I will ask you to use a chat gpt to do some tasks.

5:25

* Carlo Lipizzi: so it will be fun.

5:45

* Okay. So let's start with the she you just because I don't want to keep her waiting for a portion for the entire class. So once

5:48

* Carlo Lipizzi: she will finish, she can go, or she can say, obviously, but she doesn't need to say. See you on the floor if you are.

6:00

* Shiyu Yuan: Hi, Professor, thank you, and

6:10

* Shiyu Yuan: I. I like your course very insightful. I I took 6, 24, and 6, 26 Highly recommend for you guys for this for yeah.

6:13

* Shiyu Yuan: and the

6:23

* Carlo Lipizzi: I I share my screen. Sure. Sure.

6:25

* Okay.

6:28

* Shiyu Yuan: Can you guys see my screen?

6:33

* Mina Shafik: Yup: okay. Good. Thank you.

6:38

* Shiyu Yuan: Okay.

6:47

* Shiyu Yuan: First of all, let me introduce myself. I'm. Sure, you pretty probably you guys have already familiar with me. My email address. Actually.

6:49

* Shiyu Yuan: I'm. Currently a Phd students working with Dr. Carlo and also the Ta. For this session. Today I will talk about one of the research we are doing now. It's about a sequence labeling task.

6:58

* Shiyu Yuan: Probably you guys are not familiar with what is sequence levy and task, but I believe after this presentation you can link this task to your normal life

7:14

* Shiyu Yuan: in case you are using Chij. Bt. Now, because this task is related to Ted Bt: the Chatbot. Okay. So I will present to this topic in the following.

7:25

* Shiyu Yuan: all the first I will explain what his sequence labeling, and what have the applications of second. Slowly, then, I will divide the sequence, label into 3 sub tasks and introduce how to do the what the method to do. Sequence labeling task.

7:38

* Shiyu Yuan: So first let's look at what is exactly sequence labeling

7:57

* Shiyu Yuan: according to a definition sequence, labeling task attempts to assign a categorical labor

8:02

* Shiyu Yuan: label in each member in the sequence. Actually, this is a very general definition for sequence labeling because we are doing the natural language processing research. So in natural language processing. The sequence is the sentence one

8:08

* Shiyu Yuan: and the member in the sequence. The tokens are the words: Probably you guys are not familiar with token that's totally fine.

8:23

* Shiyu Yuan: you can trade to the member, or as worse, the worst in the sentence and the sequence labeling task in an Rp. Is to assign each word a token in the sentence

8:31

* Shiyu Yuan: to a category based on their meaning or function in the sentence.

8:44

* Shiyu Yuan: The goal of sequence lovely is to understand the structure and meaning of the input sequence, which is the sentence by classifying the word and all the token according to a predefined set of categories.

8:49

* the predefined the set of category. I will. I will explain this part in the following slides, probably in the third section, the sequence lovely sub tasks.

9:04

* Shiyu Yuan: So let's look at the application of sequence level. And this is why we are doing this: An Rp. Research.

9:15

* Shiyu Yuan: The most common application of sequence labeling are the information extraction and the checkbox like a. G.

9:24

* Shiyu Yuan: So in information extraction we first need to know what is information extraction. It's a subfield of an Lp. Research that focuses on identifying and extracting structure the information from unstructured text that data

9:33

* Shiyu Yuan: as well know the machine or the computer cannot read the character of the words exactly. It can only read worse. It can only read numeric data. So

9:48

* Shiyu Yuan: we need to transform the raw text, the data. the row row text that data is the the character of the world you are seeing. Now

10:00

* Shiyu Yuan: we need to transfer the real text that data into a more structured or machine read by format and through this machine with a format to data. We extract to the head information from the real text that data

10:09

* Shiyu Yuan: so through sequence labeling tasks, we can do the name, the entity, recognition and the semantic role, liberty. Actually, these 2 tasks are this: 2 of the 3 sub tasks belongs to Sequence laby. You can treat the named, the entity and the semantic roles as a nun

10:23

* Shiyu Yuan: in the sentence all they

10:42

* Shiyu Yuan: significant information in the sentence

10:45

* Shiyu Yuan: sequence, labeling can provide information, extraction, tasks, the name, the entity and the semantic rows and

10:48

* Shiyu Yuan: event, extraction, knowledge based construction, and the P. On a 3 subtest of information, extraction, for example, through name to entity, we can. We can extract the named entity from the raw text.

10:57

* Shiyu Yuan: and we can do the event tracking through the extracted entities.

11:14

* Shiyu Yuan: Also we can use the extracted name to entity from the raw text data and based on their relationship to construct.

11:21

* Shiyu Yuan: to construct the knowledge base.

11:31

* Shiyu Yuan: And through this knowledge base we can do semantic search. I'll come all construct. I I partly to graph to do the recommendation: yeah, build a recommendation system.

11:33

* Shiyu Yuan: And for this third, the subtask of information, extraction, the question Answer. Actually, this is related to the next application of sequence. Lively. I will combine these 2 together.

11:46

* Shiyu Yuan: which is the chat box. The definition of Tbots is a software that designed to simulate a conversation with him. He my research, if you did not use the custom at the

12:01

* Shiyu Yuan: commercial check out to like that inserted in the website to deal with some of the some of your purchase, or some of the question you concern.

12:14

* If you did not use that

12:27

* Shiyu Yuan: kind of chat about you. My heart of Ch. I bt, and you may use Chi Gp to do some information retrieval out to just the entertainment.

12:30

* Shiyu Yuan: So this is the use usage of about to do information, retrieval, virtual assistant entertainment.

12:42

* Shiyu Yuan: Yeah, that's for the application for the usage of chat about. So how how sequence labeling tasks have with about

12:50

* Shiyu Yuan: also

13:00

* Shiyu Yuan: sequence labeling tasks that can provide the name to entity. And as part of speech to the Chatbot, for example, the talk about through the named entity extracted from your text.

13:02

* Shiyu Yuan: The tap out can detect or law or locate the entities, such as people, organization, location dates, or product

13:16

* Shiyu Yuan: from the raw text. and once the chat about

13:27

* Shiyu Yuan: detect the name, the entities in the real text, the row text that can be the question you asked

13:31

* Shiyu Yuan: in the

13:39

* Shiyu Yuan: the tabout can provide accurate and relevant information and provide you a customized response to your theory.

13:42

* Shiyu Yuan: Probably if you use travel before you may find that sometimes about can only provide us very channel, but and

13:50

* Shiyu Yuan: useless information

14:00

* Shiyu Yuan: through accurately locate the name. The entities in the question in your question or to in the customer's question, the checkout can locate.

14:03

* Shiyu Yuan: They

14:14

* Shiyu Yuan: salient information like these entities in this question, and to provide a more customized response. Also, part of speech can help the tap out to understand the grammar took place. Structure.

14:15

* Shiyu Yuan: and can help that about to generate more coherent and contextually accurate response. So this is what sequence labeling I have with Chatbot.

14:29

* Shiyu Yuan: So next I will introduce 3

14:43

* Shiyu Yuan: month subtas of sequence living.

14:46

* Shiyu Yuan: I've already mentioned them, or in the previous slides. They are the part of speech, hacking them to entity, recognition, and semantic, probably probably, for now you still feel a little bit confusion. So let's go step by step.

14:51

* Shiyu Yuan: Part of speech. Tagging is a a part of speech is a category to which a word is assigned to in accordance with it's syntactic fe a function. This definition is from the Oxford language

15:07

* Shiyu Yuan: in English. The main part of speech are now pronoun adjective, determined verb, etc., and these capitalize the character at the representation of part of speech in our research in I actually this representation are widely used in the research field.

15:21

* Shiyu Yuan: and let me, with a concrete example to illustrate what is part of speech. for example, in this sentence.

15:42

* Shiyu Yuan: But yeah, so, se shell on the shop. Our task is to assign each of the elements, which is the word in the sentence to its corresponding part of speech.

15:52

* Shiyu Yuan: So both of you announced so the verb se shell. Another now

16:06

* Shiyu Yuan: is interject interaction that is determined

16:12

* Shiyu Yuan: sure is another. Now.

16:17

* Shiyu Yuan: if you are, if you want to use data driven method, we want to train a model to learn the relationship between the part of speech and the world in our sequence.

16:19

* Shiyu Yuan: So this is part of speech tagging

16:32

* Shiyu Yuan: next the name to entity. Recognition

16:36

* Shiyu Yuan: is to identify and to characterize the named entities mentioned in the text and extract the entity span with levels.

16:39

* Shiyu Yuan: You may wonder why entity span, because sometimes the named entity is not a single word. It's a word freeze, such as Steve Jobs Steve is an is a verb is an entity belongs to person Entity Class

16:48

* Shiyu Yuan: jobs is another person is another entity that belongs to an person entity class. So Steve Jobs is actually the entity span

17:05

* Shiyu Yuan: the level format of named the entity are these: for actually different to data set has different to pre-defined the Ca category.

17:17

* Shiyu Yuan: Set. I just risk an example to illustrate the label format in in some larger

17:28

* Shiyu Yuan: named entity training data set. They I remember the biggest one has 37 named entities.

17:37

* Shiyu Yuan: So here is a just a example. It depends on what training data set you choose.

17:45

* Shiyu Yuan: So then. me illustrated this concept of through the same sentence

17:52

* Shiyu Yuan: in name to entity regulation. Our task ago has changed. We want to assign each of the element. Still each of the elements in the sentence, but 2 different function.

17:58

* Shiyu Yuan: The public yeah belongs to person

18:15

* Shiyu Yuan: like I said. I like, I explained, this is actually this to our entity. Span mobile. You is the beginner of the person. Entity UN is the inside of person entity

18:18

* Shiyu Yuan: another part, another entity is.

18:32

* Shiyu Yuan: it's plans to miss us, and sure.

18:36

* Shiyu Yuan: which belongs to the location.

18:39

* Shiyu Yuan: So the

18:42

* Shiyu Yuan: and now entities. So we tacked as now.

18:44

* Shiyu Yuan: So we want to build our we want to build a model to learn this relationship

18:48

* Shiyu Yuan: between the anti-tax entity type, all the not, and today, with their corresponding token in the sentence. So the last task is semantic, row labeling.

18:56

* Shiyu Yuan: This task is to like.

19:06

* Shiyu Yuan: and then labels towards our freezes in a sentence indicating this meant to grow. The semantic rules focus on who did what we use, subject, predicate, and object to represent, hold it? What? And we use a a capitalized as you be

19:09

* Shiyu Yuan: to represent a subject we capitalized, we to represent, to practice, predicate and capitalize the obj to represent the object.

19:28

* Shiyu Yuan: Actually, semantic role is related to the part of speech I just mentioned. Part of speech provided the statement of the syntax representation of each token, and a semantic role can indicate the relationship between the part of speeches of each world.

19:38

* Shiyu Yuan: So let me still use the same sentence to illustrate this concept.

19:56

* Shiyu Yuan: The for here we concern who did what? Who is for you

20:02

* Shiyu Yuan: which they are the subject Date? What about you? He sold his sell. So what se shell?

20:08

* Shiyu Yuan: So this is the semantic rose we care about in this sentence.

20:16

* Shiyu Yuan: mobile, and so shell. But you may wonder how about on the show, because we do not care who did what, to whom we'd only care about who did what? So on the show. Now, semantic crops we ignore that, but we still need to tech them as now.

20:22

* Shiyu Yuan: These 3 are more 3 sub tasks of semantic roles of sequence labeling tasks. And next, I will introduce the method to do this tasks

20:41

* Shiyu Yuan: we have classic sequence, label and the modern sequence label you can understand as a traditional sequence, labeling way and the modern a current, a more current sequence, leveling way

20:54

* Shiyu Yuan: for traditional sequence leveling, we have to. There are 2 branches, the header mark of model and conditional random field hidden. Markov model is

21:09

* Shiyu Yuan: the head of Markov is one.

21:20

* Shiyu Yuan: Oh.

21:22

* Shiyu Yuan: once it to the parents of one situation only depend on its immediate previous token

21:24

* Shiyu Yuan: a situation. So

21:31

* Shiyu Yuan: I will explain how he, the Markov model, use the impact of speech. Let's review what is part of speech a little bit. It's taking a sequence of words. Assign each word, a part of speech like not a verb.

21:35

* Shiyu Yuan: And in this section we need a 2 probability matrix. The first one is called the transition probability. It doesn't matter. We do not care what is called. We care what is calculated.

21:50

* Shiyu Yuan: We. We want to the proper the conditional probability of we be based on the occurrence of Md. And we calculate to this probability through the through counting the currents of Md. And they be divided by the current of Md.

22:03

* Shiyu Yuan: Actually this, for this calculation is from human engineer, and this whole match may measure the head, and Markov model is a feature engineering. It's belong to the traditional

22:23

* Shiyu Yuan: aspect. We actually we did not use this method anymore.

22:36

* Shiyu Yuan: Okay, so from the next, that is emission probability.

22:43

* Shiyu Yuan: Actually, this is a likelihood we want to pretty. We want to the likelihood, the conditional probability of the talking well based on it's

22:47

* Shiyu Yuan: based on it's part of speech. It's a Md.

23:00

* Shiyu Yuan: They for we formula we formulated this probability by counting the co-currence of Md. And will, by dividing by the current of Md. Actually this probability also from human engineered.

23:04

* Shiyu Yuan: Our goal is to get the probability of the it. We want to get the probability of each part of speech for this token, and we want to get the

23:19

* Shiyu Yuan: high is the probability.

23:33

* Shiyu Yuan: So this is. This is the example of the probability matrix.

23:36

* Shiyu Yuan: Like I said, these numbers, these probabilities are from human engineered, so in the traditional way. They use a lot of human labor to calculate the statistic and build the statistic model. But now we are not doing things in this way

23:41

* Shiyu Yuan: we're doing the

24:01

* Shiyu Yuan: your data driven method.

24:03

* Shiyu Yuan: This method has a to branch

24:05

* Shiyu Yuan: arm based model, and the transformer based model are is designed to sequence analysis, but

24:09

* Shiyu Yuan: it has some limitations. So the men's room is to use transformer based model.

24:17

* Shiyu Yuan: and I will introduce this very classic Fontaine approach. So first we divide the sentence into each tokens you can treat the token as word. So this sentence has an

24:22

* Shiyu Yuan: numbers of words, and we vectorize this because the machine cannot to read word or tokens. After vector after vectorization we put the token vector into the model, and we fix the previous several layers to maintain the

24:35

* Shiyu Yuan: parameter, and which one, the last several layers in the pretend model and the build a dense layer, because we are because we are doing the multi class classification. So we use soft Max.

24:54

* Shiyu Yuan: So actually, this is the end to end. Approach. We assign each token to one entity. If the second token is not a entity, we assign it as a

25:10

* Shiyu Yuan: 0. So this is the

25:24

* Shiyu Yuan: transformer based all the data driven method to do, and you are one of the sequence labeling tasks.

25:28

* Shiyu Yuan: So this is this: what I want to share with you guys, I don't know if I explain clearly. If you guys have any questions, please let me know.

25:38

* Shiyu Yuan: Thank you

25:49

* Carlo Lipizzi: all right. Thanks a lot to you.

25:55

* Shiyu Yuan: True, Professor.

25:58

* Carlo Lipizzi: Should I stop? Share all I I don't know if you can stop sharing unless we have some questions on on one particular slide, and I mentioned you. You can just retrieve it.

25:59

* Carlo Lipizzi: Okay. So one of the goals, all the the presentation that she you gave to us I mean, the most obvious is to give you an example of of the research that we are doing, but the second is to frame the research in a broader scope.

26:15

* Carlo Lipizzi: So the sequence labeling seems to be a very specific, deep niche research topic.

26:40

* Carlo Lipizzi: But when you do complex things as more details. We'd really make the difference in the old research.

26:54

* I think. Last class. I mentioned the fact that

27:05

* Carlo Lipizzi: that one of the things that that we do in a natural language processing basic things is a a programming. So you want to put together words with the same semantic meaning system, engineering, project management school of business, a Webinar mass destruction. All of those

27:08

* Carlo Lipizzi: are the same semantic concept spanning over multiple worlds.

27:29

* Carlo Lipizzi: So we we can do it the the way we do. We did the in the past, using a sort of brute force. You take the the words so one next to the other, and then you take only the most common, and those are the end grams.

27:36

* Carlo Lipizzi: That's a possibility. But if you want to do things the right way, then you may want to work on a the so forth. You want to remove the stopports, but if you remove the stop for so before doing the programming school of business, we never exist.

27:54

* Carlo Lipizzi: So you want to remove the so forth. So. But after you do the programming, and only if the stop for the is either at the beginning or at the end of the end. So that's an example. We wrote a paper on that

28:14

* Carlo Lipizzi: I I wrote a script that there is probably 200 lines just for that, just for a and gramming.

28:32

* Carlo Lipizzi: and seems to be a very niche problem.

28:41

* Carlo Lipizzi: But if you don't solve that, then Theola interpretation is not going to be right, because you will never know what is going on about the school of business or about the systems engineering. So you want to have that 250.

28:45

* Carlo Lipizzi: Another example is done. You You want to get some insights on on a a document.

29:02

* Carlo Lipizzi: but documents can be big. So you have a document or several 100 page, and the same concept can be scattered in different places in the document.

29:12

* Carlo Lipizzi: How do you do that. So the medal that I created was basically

29:25

* Carlo Lipizzi: deconstruct the document and reconstructed the in a visual paragraphs that are semantically coherent. So what I did was basically victorize the document 250,

29:32

* Carlo Lipizzi: and then create clusters based on the vectors. So at that point, and then going back to the actual world. So I recreated the the document 2

29:50

* Carlo Lipizzi: with a completely different structure.

30:03

* Carlo Lipizzi: Again. It

30:06

* Carlo Lipizzi: it's a very niche problem, teeny tiny, but without that you don't get results.

30:08

* Carlo Lipizzi: So again, we just wanted to give you an example of a thinking big, so a chatboard, but then a digging deep

30:17

* Carlo Lipizzi: going into one detail, and

30:32

* Carlo Lipizzi: so with.

30:37

* Carlo Lipizzi: So if you don't do the right labeling, you will never get the information you want, that you will never understand. If you are talking about a person or you. You are talking about the CD.

30:39

* Carlo Lipizzi: So that's the name, the entity, re recognition, and then the part of speech, the part of speech. It's particularly tree, because the role that the words are playing

30:54

* Carlo Lipizzi: is changing in time. If you consider Google Uber, so they are nouns. But we use them as verbs

31:12

* Carlo Lipizzi: meaning over the years language changed. So if you use a

31:23

* Carlo Lipizzi: methods that are static, so if you use that, we use the natural language, toolkit, natural language toolkit as a module, doing it the part of speech tagging. But it's based on a classification of the names and the other parts that is 20 years old.

31:32

* Carlo Lipizzi: and Uber may not be there, meaning, if you use it, Weber as a verb, we not get it.

31:52

* Carlo Lipizzi: and language is changing constantly. And then there are a jargons. So in everyday life we are using words with.

32:01

* Carlo Lipizzi: That is a not. The fish are one. But if you want to understand what people is saying, you need to recreate that so.

32:11

* And

32:23

* Carlo Lipizzi: the way you extract those it's really gridy gala. So

32:25

* Carlo Lipizzi: erez agmoni, we are exploring, creating a network and then applying network metrics to recreate these causality in the sequences of awards meaning in phrases 142.

32:33

* Carlo Lipizzi: If you don't, do that.

32:51

* Carlo Lipizzi: you Don't understand the text. So for human beings, so life is easy because you live in the present time. You use the words that people around you are using.

32:53

* Carlo Lipizzi: You are using them. It's probably in the same way, and you understand each other.

33:06

* Carlo Lipizzi: But machines do not have this common sense, and that's why we need to dig deep and create those algorithms, those methods that can really help us fulfill the the the task that that we have.

33:12

* Carlo Lipizzi: So again.

33:30

* Carlo Lipizzi: sometimes you can go high level. But sometimes you need to go deep. So probably is not even when not is definitely not the deepest

33:32

* Carlo Lipizzi: point that you can reach in a natural language processing. There are many other things that that are more I mean smaller targets, but pretty complex.

33:46

* But that's an example.

34:03

* Carlo Lipizzi: See you? What what do you think of this type on an analysis?

34:06

* Shiyu Yuan: Yes, Professor, actually, we think we have another kind of sequence labeling task in the recent days that is prompt to learning.

34:11

* Shiyu Yuan: That is also we design different part of a sequence. And this another story. Yes, i'll go with you, Professor, and I don't think our research is trivial. It's build a. So it's a foundation to the top tasks.

34:22

* Carlo Lipizzi: Yep.

34:42

* Carlo Lipizzi: Okay. Great thanks again to you.

34:43

* Shiyu Yuan: Okay.

34:51

* Carlo Lipizzi: All right. Okay. So let's move on with the the other topics we have, so

34:52

* Carlo Lipizzi: I want to spend as much time as possible on the final and then sometime on some other example. So let's let's start with the easy portion of examples, all the applications.

35:04

* Carlo Lipizzi: So I want to share with you

35:23

* Carlo Lipizzi: couple of researches that we did in recent times.

35:28

* Carlo Lipizzi: So let me share the screen and let me go here.

35:32

* Carlo Lipizzi: So this project that we call the Greek connect is a project that we started.

35:38

* Carlo Lipizzi: and yet enough to go pretty much and let me go into a presentation model.

35:50

* Carlo Lipizzi: It's it's it's it's, it's, it's, it's it's it's it's it's a

35:58

* Carlo Lipizzi: yeah

36:07

* Carlo Lipizzi: on 7 different tasks with the 1,400, the people participating all over the world. One of the 7 tasks was a financing, a sustainable future.

36:07

* Carlo Lipizzi: So we

36:24

* Carlo Lipizzi: created the a team. So the the point of contact, Martin Powell was a former a student. All the George this, who was was the Provost Stevens 2 problems. Ago

36:25

* Carlo Lipizzi: E. And Martin encouraged Georgia to participate. I I work with Georgia on a a couple of Dod projects, 250,

36:47

* Carlo Lipizzi: and he knew that they had some background in in natural language processing. Can we put together a the thing that I going to show to you?

37:00

* Carlo Lipizzi: So the

37:12

* Carlo Lipizzi: problem that we want to. That we want to to address. So was a

37:14

* Carlo Lipizzi: how to find projects that can be funded by Siemens financial services.

37:20

* Carlo Lipizzi: projects that are in the sustainability area. So that was the target that that that we had. So again we won a the hackathon

37:30

* Carlo Lipizzi: and let me go. Here we use the the room theory that I presented the last time. The room theory is basically a a knowledge driven approach

37:49

* Carlo Lipizzi: that is based on analyzing documents using a computational representation of the knowledge of the specific domain.

38:08

* Carlo Lipizzi: The architecture is basically you create the the corpus collecting documents in this case, but related to what Zoom's financial services was doing

38:25

* Carlo Lipizzi: so. All the documentation on all of past projects, the way you're doing things, what is the background of the people doing the same job?

38:41

* Carlo Lipizzi: And then we transform the this

38:52

* Carlo Lipizzi: ere

38:57

* Carlo Lipizzi: that we're computational representation of the of the domain.

39:04

* Carlo Lipizzi: then that with the same knowledge you can do multiple things. So we define the specific tasks on with the keywords and related weights

39:13

* Carlo Lipizzi: mit ctl. And so keywords are at basic keywords. And then all the the synonyms, all the misspelling. And again, weights, because not all the keywords are created equal. 101.

39:27

* Carlo Lipizzi: We don't do an exact match. But we basically measure the proximity of those keywords with the documents, so that we are analyzing.

39:39

* Carlo Lipizzi: So then, we built a a webcroller going into the web and getting those projects to be analyzed.

39:52

* Carlo Lipizzi: Then, using this room theory, we had the projects that are potentially of interest for Siemens financial services to be funded.

40:03

* Carlo Lipizzi: So then, you have those elements, and you do some visualization to present the the results.

40:19

* Carlo Lipizzi: So we developed a a basic user interface where you have a you can select the target industry.

40:30

* Carlo Lipizzi: You have the different documents with a caller code for the potential interest they may have, and a distribution

40:39

* Carlo Lipizzi: in terms of what are the keywords, I mean based on the keywords that you provide in the in the benchmark.

40:53

* Carlo Lipizzi: The proof of concept was a a simpler than that. So W. Was basically a room. We had a relatively small room, initially, I mean for the hackad on we had the 450 documents.

41:02

* Carlo Lipizzi: The benchmark was 300. The words of races, I mean, I mean, what's can be engrams again.

41:22

* Carlo Lipizzi: because

41:31

* Carlo Lipizzi: we wanted to provide that semantic concept, not just the work.

41:33

* Carlo Lipizzi: And a

41:38

* Carlo Lipizzi: for the hackad on. We didn't have the wet roller ready, and we collected manually the documents. Now, in the current version we do have a crawler

41:40

* Carlo Lipizzi: and the algorithm for matching. We are the same of the room theory, and we use the some sort of basic graphic visualizations.

41:53

* Carlo Lipizzi: So that's basically, what you have. You have all the different benchmarks or keywords, and you have the different projects, and you have a this sort of a correlation magic. So that is telling you what are the projects that are potentially more interesting.

42:07

* Carlo Lipizzi: We added that this

42:27

* Carlo Lipizzi: project, 0. There is a sort of a a control project that was not related with the topic to see if the system was doing his job. But not.

42:29

* Carlo Lipizzi: And then you couldn't have a a little bit more. That will be on a a view for

42:40

* Carlo Lipizzi: the different I mean the intersection between documents and and benchmark, and you can visually see what are the projects that are matching.

42:50

* Carlo Lipizzi: or better, what are the keywords that are matched? The the most by the different benchmarks.

43:08

* Carlo Lipizzi: So a. And that's basically it. We are after the hackaton. Siemens gave us a another. Here of a funding.

43:17

* Carlo Lipizzi: We are expanding the the system. The system again has a crawler. The roller is working fine. We expanded the the benchmark, the corpus.

43:32

* Carlo Lipizzi: the crawler, so far, is a dumb crawler. So it's just collecting a elements based on keywords that we provide. And then the system is actually doing the job

43:46

* Carlo Lipizzi: We are thinking about adding a layer all the you know common sense, because one of the things that seems to be easy, but it's it's kind of complicated. You have a document that is a white paper, so it's a commercial pro document.

44:01

* Carlo Lipizzi: It is not a request for a a proposal or an idea of a project to be founded for a human that would be easy to say. It's a commercial promotion is now the Vla request for funds

44:19

* Carlo Lipizzi: for a machine it is. It's more complicated.

44:36

* Carlo Lipizzi: Machines do not have the common sense, so we will use something like a chat gpt equivalent to to filter out those projects that are only

44:40

* Carlo Lipizzi: not projects, but they are just commercial products to commercial documents.

44:53

* Carlo Lipizzi: Same thing for the location. So sometimes you have a document.

45:00

* Carlo Lipizzi: and you have Cds. You have vitro it. But actually, the the project is going to happen in Nairobi. So because Siemens financial services is is for use in this case on North America, Nairobi would be or no interest for them.

45:05

* Carlo Lipizzi: So for a human it would be easy just to go through everything and say, okay, this is not of interest to us. Again. We will use a a a lunch language model to do this software pre screening.

45:26

* Carlo Lipizzi: That was the

45:43

* I mean the the 3 of us, the the founding members are still there.

45:57

* Carlo Lipizzi: We are thinking about the next stage, meaning the

46:05

* Carlo Lipizzi: Erez Agmoni. The system is now working, I mean. Obviously, there is always room for improvement, but it's delivering results. 150

46:11

* Carlo Lipizzi: Siemens wants to continue working with us, want to use the system.

46:22

* Carlo Lipizzi: So we are thinking about creating a company that could be with Stevens inside.

46:28

* Carlo Lipizzi: So I had conversations with the Provost, with the senior Vice provost for research and anthropologists with our general counselor, and we are going in the directions

46:36

* Carlo Lipizzi: when there is a sort of the organization in our point of contact move to a another company.

46:56

* Carlo Lipizzi: So right now the idea of the company is kind of on old. But before our point of contact

47:06

* left.

47:19

* Carlo Lipizzi: I mean. He involved the the Cfo, the the CEO, or Siemens financial services, North America. That is, 20,000 people. Then he seems to be on board. But

47:21

* Carlo Lipizzi: again. The direction is to spin off this activity, creating a company, and then we'd see what is going to happen. So that's an example. It's not end to end. But I mean it is a pretty, let's say mature example.

47:36

* and then

47:56

* Carlo Lipizzi: of application of something that we discussed.

47:58

* Carlo Lipizzi: something completely different. Let me share this research that we did about probably 2 years ago.

48:04

* and was a a join effort with unicef

48:17

* Carlo Lipizzi: during the pandemic

48:22

* A. There there was a a a a spike in domestic violence, and children abuse. So we we wanted to track that. We wanted to create some awareness on on the problem.

48:27

* So

48:43

* Carlo Lipizzi: we created the a.

48:45

* Carlo Lipizzi: A joint team with some people from Stevens, some people from Unicef. and that's the Unicef team.

48:49

* Carlo Lipizzi: That's the Stevens's team. So myself.

49:02

* Carlo Lipizzi: The 2 of us are 2 of the main point of contacts when is about analytics in a, in a broad sense, is more focus on on a resilience and more on data, analytics and natural language processing it's.

49:09

* Carlo Lipizzi: We are a 2 of our Phd students, both employed. Now.

49:32

* Carlo Lipizzi: Korea is in Bangladesh as a senior data scientist, and Fernando same thing in a data consulting company.

49:39

* Carlo Lipizzi: So

49:52

* Carlo Lipizzi: there is no need to to give you the background on the team. But I just want to go into the project.

49:53

* Carlo Lipizzi: So we we wanted to again create the some sort of a a awareness we for use the on a different countries for sure, on the Us. But we wanted to expand a a little bit.

50:04

* Carlo Lipizzi: and that we use the social media to get what people was

50:19

* Carlo Lipizzi: thinking was doing and trying to map somehow the facts with a story Again, it's all about the storytelling

50:29

* Carlo Lipizzi: 1 million issues. Some

50:42

* Carlo Lipizzi: data may be misleading. We are talking about a topic that is

50:46

* Carlo Lipizzi: very sensible and sensitive.

50:52

* Carlo Lipizzi: and some people it's not comfortable discussing those experiences

50:57

* with me.

51:06

* Carlo Lipizzi: I mean on public records. And then there is the issue, or a multi-languages.

51:07

* Carlo Lipizzi: So we collected the tweets from a 15 different countries. In those data frames we collected the data from red. It

51:14

* Carlo Lipizzi: tweets red. It's a they are very different. So tweets are are less for use, the shorter red. It's a more for use longer.

51:28

* Carlo Lipizzi: We'd read it.

51:46

* Carlo Lipizzi: You have communities with tweets, not so much

51:48

* so. We use the

51:53

* Carlo Lipizzi: modeling to represent all of that. So the first case study is a using tweed.

51:57

* Carlo Lipizzi: So again, there are some advantages. It's a easy to access was so easy to access before you know. Mask. It's widely used, based on text.

52:05

* Carlo Lipizzi: So I I mean Instagram. Will it be great, but it's way much more difficult to process images than text.

52:23

* Carlo Lipizzi: and because you have a the possibility to comment, I mean to

52:33

* to side. There is a sort of conversation that can and gone.

52:41

* Carlo Lipizzi: Okay. We collected the the data, we

52:48

* Carlo Lipizzi: and identify the conversation so that we are more related to the topic of our research. We mapped the results by geography, all 3

52:53

* Carlo Lipizzi: location, and we analyze the results.

53:07

* So

53:11

* Carlo Lipizzi: those are our hate speeches in different States before and after.

53:12

* Carlo Lipizzi: So

53:20

* Carlo Lipizzi: before Covid and after Covid. So that was the increase for this Rito Columbia, for example. So this is hate speech

53:22

* Carlo Lipizzi: around the world so

53:32

* abusive, non abusive.

53:36

* Carlo Lipizzi: There.

53:40

* Carlo Lipizzi: I mean that again, as usual.

53:41

* Carlo Lipizzi: Swedish are always the best in the world. But unfortunately, what this big and the number of people in Sweden, unfortunately, is not as much as the people in Indonesia or in Brazil, meaning

53:46

* Carlo Lipizzi: they're having so much of abuses, but compared to non abusive.

54:00

* Carlo Lipizzi: it's really

54:08

* warring up there.

54:10

* Carlo Lipizzi: So then

54:12

* Carlo Lipizzi: we created a sort of a an index to measure it.

54:16

* and that this is the distribution. So you have an average, and and countries that are above, over below that

54:22

* Carlo Lipizzi: similar thing, using a red. It

54:34

* Carlo Lipizzi: so again, red. It is very different from to either. It's a social media, but it's a different animal

54:38

* Carlo Lipizzi: you have Subreddits subreddits are very focus on a given topic. Mit, ctl and so we focus on on abuse-related, subreddits, abuse survivors, abuse survivors of abuse and a massive violence. 150.

54:50

* Carlo Lipizzi: We determine a control group just to have a a a way to create a context for the results that that we had.

55:07

* Carlo Lipizzi: And again, one of the limitations in this case is that the credit is used primarily by the Us. Meaning you, Don't have a

55:19

* Carlo Lipizzi: the same amount of data from other countries that you have with with them.

55:31

* Carlo Lipizzi: That's an example of a one on the post

55:36

* Carlo Lipizzi: collected data, analyzing in time expecting topics. That's what we did.

55:42

* Carlo Lipizzi: Number of newly active users on a abuse Re related the Reddit. So there is definitely a a a spike, even just considering the number

55:51

* Carlo Lipizzi: of users, and not the content or what they said. Then, using

56:03

* nda, we had that I mean. We

56:11

* Carlo Lipizzi: kind of manipulated a little bit, the nda adding labels to the different topics.

56:16

* Carlo Lipizzi: So that's basically where those were more

56:23

* Carlo Lipizzi: related.

56:29

* Carlo Lipizzi: So abuse related subred. It's a are among the topics with the highest growth during the lockdown. So

56:32

* Carlo Lipizzi: that was a something

56:42

* scaring somehow.

56:45

* Carlo Lipizzi: So let me skip that.

56:47

* Carlo Lipizzi: So those are

56:50

* the

56:54

* Carlo Lipizzi: topics.

56:55

* Carlo Lipizzi: So again

56:56

* we detected the 5 topics.

56:59

* Carlo Lipizzi: and we label them, and we extracted the keywords that are meaning the the different topics.

57:04

* Carlo Lipizzi: and that that's basically for each one of the topics, the distribution in time

57:13

* Carlo Lipizzi: that is giving you an idea of what was going on.

57:19

* So if you look at the first one, that's

57:24

* Carlo Lipizzi: really growing a lot a lot I mean it's it's it's carries out so 33 more exposure to abusive languages, language, and cyber bowling on twitter

57:28

* Carlo Lipizzi: 37 us states so a large and increase in tweets containing abusive language. 94 more child abuse on ready during the lockdown 88 more intimate partner abuse re but already during the lockdown.

57:40

* Carlo Lipizzi: So

58:00

* Carlo Lipizzi: those are the facts. So the paper that we published the following this presentation was a a highly sided

58:01

* Carlo Lipizzi: There are 2

58:16

* Carlo Lipizzi: more out to to the story. So the the first one, when you write something, stay on facts that can really drive attention.

58:19

* Carlo Lipizzi: because at that point you will get more exposure.

58:34

* Carlo Lipizzi: The sag on the most important.

58:40

* Carlo Lipizzi: What we do can really be useful to expose Some situations, create or increase awareness.

58:42

* Carlo Lipizzi: Some of those researches are not so easy meaning you can. You may not find them on the New York Times, not because they they are not good, but but because I mean it. It took us

58:54

* Carlo Lipizzi: 3 months to develop it.

59:09

* Carlo Lipizzi: Jordan is may not have the same, the the the actually I'll spend so much time. But again you can do a lot of good or a lot of working with data and leveraging on them and presenting them in in the proper way.

59:13

* Carlo Lipizzi: So that's something that they wanted to share with you. Another thing that they want to share with you is

59:34

* a few scripts.

59:46

* Carlo Lipizzi: So

59:48

* in particular.

59:50

* Carlo Lipizzi: 2 of them

59:55

* Carlo Lipizzi: let me share this screen again.

59:57

* Carlo Lipizzi: Let me go here. So

1:00:04

* Carlo Lipizzi: analyzing text is a pretty general need.

1:00:09

* Carlo Lipizzi: So she you was mentioning

1:00:16

* Carlo Lipizzi: retrieval in in text. There is a genetic term. It really depends what you want to do with the the text. What are the information that you want to extract? Some of the tasks kind of a repetitive. So we

1:00:21

* Carlo Lipizzi: spend some time, for for example, not analyzing text and extracting the most frequent words, the

1:00:41

* Carlo Lipizzi: diagrams.

1:00:51

* Carlo Lipizzi: By the way, let me stop here for a second and let me go into diagrams.

1:00:56

* Carlo Lipizzi: Yesterday I was teaching a a class on a a visualization course, of course 6, 22,

1:01:04

* Carlo Lipizzi: with Professor Ramirez

1:01:15

* Carlo Lipizzi: pretty much each semester. We do a sort of an exchange of of classes at a certain point, and at each one of these classes, using my my

1:01:21

* Carlo Lipizzi: experience in a natural language processing and applying that to visualizations and or to network analysis, or a combination of the 2,

1:01:36

* Carlo Lipizzi: one of the students, and asked me Why, she's also a student in my other 6 24 class. She asked me, why you are asking us to do diagrams. So what's the use of diagrams.

1:01:48

* Carlo Lipizzi: Well, diagrams are an essential component of the or engram, so are an essential component of of the conversation. I mentioned that half an hour ago.

1:02:08

* Carlo Lipizzi: so school of business project management information technology. So all of those are are in Grams diagrams, 3 grams or more, with the same semantic meaning.

1:02:26

* Carlo Lipizzi: So you don't want to leave the words as individual words, but you want to aggregate them, creating some sort or representing

1:02:43

* Carlo Lipizzi: one single semantic concept.

1:02:55

* Carlo Lipizzi: Once you have that that could be diagram. It could be engram again, a web on a mass destruction. There are quite some words, but we just one single semantic element.

1:02:58

* You want to use them

1:03:10

* Carlo Lipizzi: erez agmoni, in pretty much all the representations you have. You want to use them for the word cloud. The most frequent words, even if They are not really words, but they are those chunks, those semantic elements, 150

1:03:13

* Carlo Lipizzi: you want to use. When you do some topic analysis.

1:03:32

* Carlo Lipizzi: you do, Lda, as an example of topic for a presentation, and you want to have a

1:03:39

* Carlo Lipizzi: business, a single word, or a business management as a diagram.

1:03:46

* Carlo Lipizzi: So that's the reason why you want to have a diagrams, because you want to extract from the text not only the words, but you want to extract the the

1:03:53

* Carlo Lipizzi: concepts, the the semantic elements in those who can span across multiple words. So that's something that I mean that you may have the same question that

1:04:06

* Carlo Lipizzi: this you have

1:04:21

* Carlo Lipizzi: all right. So

1:04:25

* Carlo Lipizzi: this script is kind of a a genetic script. It's relatively long, 240, and change lines of code.

1:04:27

* Carlo Lipizzi: and he is a doing a

1:04:39

* Carlo Lipizzi: kind of a repetitive tasks in N IP: so is a text cleaning

1:04:41

* Carlo Lipizzi: Most common words generated in the Engrams doing the core. And so let me stop here for a moment. Co-corance

1:04:50

* Carlo Lipizzi: A co-currence means we mentioned briefly last week means

1:05:01

* Carlo Lipizzi: appearing together. So you have a a large text, and you have a

1:05:12

* Carlo Lipizzi: some words, so that they are appearing together.

1:05:20

* Carlo Lipizzi: So the assumption is that if they are repeating together, that means they are related. So this being related as a value, because when you have more words that are related, then you have a

1:05:23

* Carlo Lipizzi: a set of related topic. What that will create a topic.

1:05:40

* Carlo Lipizzi: So, using this call quorance awards across the document is a way to analyze the document.

1:05:46

* Carlo Lipizzi: Eventually you can create a network out of it with what's the that are related because they are appealing one next to the other?

1:05:55

* Carlo Lipizzi: Then what is next to the other is to be defined, meaning, if they are one next to the other, I mean

1:06:07

* Carlo Lipizzi: one following the other, then he is a diagram. But if they are in works of distance, so degree of separation.

1:06:16

* Carlo Lipizzi: Then you can define the degrees of separation between the words. That will be that that threshold, Obviously the larger is the document, and the the smaller is going to be the number of degrees of separation, because otherwise you, you would have a a, a medium words.

1:06:28

* Carlo Lipizzi: and then you may want to count

1:06:47

* Carlo Lipizzi: the number of times the words are appearing, because if they appear together only once, they may not be relevant. So

1:06:52

* Carlo Lipizzi: with those parameters you can create a network based on this call quorance. So, and that's why in my script i'm calculating the coordinates.

1:07:02

* Carlo Lipizzi: Then what cloud some topic, modeling and some statistics.

1:07:16

* Carlo Lipizzi: So basically I imported the All the different libraries.

1:07:23

* Carlo Lipizzi: Some are for visualization, some are for calculation.

1:07:31

* This is a a

1:07:36

* Carlo Lipizzi: and

1:07:39

* very basic cleaning function.

1:07:40

* Carlo Lipizzi: This is to

1:07:44

* Carlo Lipizzi: calculate the the most common elements

1:07:48

* Carlo Lipizzi: generating the engram, sir. generating this matrix of cocaine.

1:07:53

* Carlo Lipizzi: then generating the work cloud

1:08:03

* Carlo Lipizzi: generating the the visualization for the nda model for representing topics.

1:08:09

* Carlo Lipizzi: And then you define whatever is the file you read the software file you, I mean, load the the file files into.

1:08:18

* Carlo Lipizzi: At least you said that the minimal length Sorry about that.

1:08:30

* Carlo Lipizzi: You set the window separation, and that I was mentioning to you

1:08:39

* the number of engrams that you want.

1:08:46

* Carlo Lipizzi: How many of

1:08:50

* Carlo Lipizzi: I mean, how big and as should be in Engrams, how many of them you want. and then that some naming a number of topics

1:08:52

* Carlo Lipizzi: here i'm generating from the Madrid. So the Co. Currents madrics what is in the network analysis? It's called the adjacency madrics.

1:09:04

* and

1:09:18

* Carlo Lipizzi: from that I'm. Generating the the graph

1:09:20

* Carlo Lipizzi: and then saving it for a future analysis

1:09:25

* Carlo Lipizzi: Then.

1:09:31

* Carlo Lipizzi: and

1:09:34

* Carlo Lipizzi: from the cleaning I I get the vocabulary that is the least of unique words

1:09:35

* Carlo Lipizzi: and

1:09:42

* creating the work cloud

1:09:44

* Carlo Lipizzi: topic, modeling topic, modeling a more that I think needs the vocabulary. That's why I generated it.

1:09:46

* Carlo Lipizzi: And then I basically generating some statistics. The number of words. the number of unique words. So the entropy meaning how diversified is the user awards. So a text with the higher entropy than another

1:09:55

* Carlo Lipizzi: means that there are a wider variety awards.

1:10:18

* Carlo Lipizzi: So when you run it, let me run it. And

1:10:25

* Carlo Lipizzi: so you are going to have a

1:10:33

* Carlo Lipizzi: the what cloud!

1:10:37

* Carlo Lipizzi: And that's Why not? You have a

1:10:40

* Carlo Lipizzi: the engrams. the topics

1:10:50

* Carlo Lipizzi: That's the this that I was mentioning. and there are some graphs. So this, for example.

1:10:54

* Carlo Lipizzi: is the graph

1:11:05

* i'm opening it

1:11:09

* Carlo Lipizzi: with the her FDA that is being generated. So when you go into the different, so we define a for topics. So when you go on each one on the right side, that you have the wars or and Graham.

1:11:11

* So you have a natural language processing senior business consultant.

1:11:30

* Carlo Lipizzi: So those are engrams.

1:11:35

* and

1:11:39

* Carlo Lipizzi: this is actually from something different. But the the the same concept.

1:11:42

* Carlo Lipizzi: You can eventually change the relevance. Madrid.

1:11:47

* Carlo Lipizzi: So this was for the resume. Okay, let me go to what has been generated in this case.

1:11:53

* Carlo Lipizzi: It is right here. This is what has been generated in this case. Sorry about that

1:12:00

* Carlo Lipizzi: was 5, not for

1:12:12

* topics. And again

1:12:16

* Carlo Lipizzi: you have the different topics. The relevance is all of them.

1:12:19

* Carlo Lipizzi: You have a

1:12:25

* another interesting representation that is a

1:12:27

* this one up.

1:12:34

* Carlo Lipizzi: So I was mentioning that that you can generate a network out of the awards, using the call quor, and so and that's basically what it's been generated.

1:12:39

* Carlo Lipizzi: So you have the different elements and how they are connected.

1:12:51

* So, going into it.

1:12:58

* Carlo Lipizzi: you can get a sense eventually from that you can do a clustering, using a a, an algorithm and it's called the lobbying community detection. And you can get more insights.

1:13:01

* Carlo Lipizzi: You can change what is called the gravity, meaning how much the notes are attracted, either one to the other or to the center. You can change the what it's called the solver, that is the visualization.

1:13:16

* Carlo Lipizzi: and then you have the work cloud. So you have the work cloud. You have the Mba. You have the network, and you have some statistics. I use that the when we were hiding

1:13:33

* new people at Ssc. So I wanted to compare those results

1:13:46

* Carlo Lipizzi: so, and it's pretty much the same as the previous one. I'm not

1:13:52

* Carlo Lipizzi: keeping in mind that when you have a a resume, most likely the resume is in a Pdf. So I had this.

1:13:58

* Carlo Lipizzi: There's a script that was a reading a Pdf. And transforming the the Pdf. Into a text is a relatively straightforward

1:14:11

* Carlo Lipizzi: script, using a library. There are several of them

1:14:25

* Carlo Lipizzi: is the one that I use in this case. So, using this one, it's basically where I got

1:14:34

* Carlo Lipizzi: this one. So

1:14:46

* Carlo Lipizzi: so

1:14:53

* Carlo Lipizzi: this Lda is related to my resume.

1:14:55

* Carlo Lipizzi: So you have a business intelligence. You have a natural language processing a step, and Cecil of technology.

1:14:59

* Carlo Lipizzi: So the things that you can expect. So I I did the something very straightforward. I went to Linkedin. I downloaded the I transformed the profile into a Pdf. And then I use the Pdf

1:15:06

* again.

1:15:23

* Carlo Lipizzi: It's just an example. you can see, actually, the user. All the

1:15:24

* the engrams.

1:15:32

* Carlo Lipizzi: So Stevens is the technology, business intelligence, natural language processing lab. In this case, senior business consultant, principal investigator

1:15:34

* Carlo Lipizzi: a developing solution. So you want to have a machine learning. So you want to have those and grabs

1:15:46

* Carlo Lipizzi: all right. So, very briefly, I want to show you

1:15:58

* couple of finals.

1:16:08

* Carlo Lipizzi: So in particular, I want to show you this one that is, from last year.

1:16:10

* Carlo Lipizzi: and my opinion is one of the

1:16:21

* best

1:16:25

* Carlo Lipizzi: in

1:16:27

* Carlo Lipizzi: recent semesters, and was for sure, the best

1:16:28

* Carlo Lipizzi: in that particular semester. So let me share the screen.

1:16:35

* Let me go here.

1:16:39

* Carlo Lipizzi: It's on one of the tasks of the problems that you have as options for this year. So global migration dynamics.

1:16:41

* Carlo Lipizzi: Joshua Johnson

1:16:53

* Carlo Lipizzi: a good, the table of content. I mean that that was a kind of massive 30 page, and you would see the quality of the visualizations and the analysis and the conclusions

1:16:57

* Carlo Lipizzi: in particular. So you have a introduction, research questions, data, description, preparation, representation, conclusions, references, appendixes, list of figures.

1:17:10

* Carlo Lipizzi: list of Tables abstract.

1:17:25

* Carlo Lipizzi: So why we are doing what we we are doing why this problem is relevant. Why, we are talking about the migration. So that's the description. I'm not going to spend much time, but I just want to give you a sense of the the way it was done.

1:17:29

* Carlo Lipizzi: Some terminology that is always good, because we assume that we are using sometimes. But everybody knows, but not necessarily. Everybody does

1:17:48

* Carlo Lipizzi: so. Factors push and pool.

1:18:00

* Carlo Lipizzi: so push job opportunities, hyp population.

1:18:04

* Carlo Lipizzi: cool availability, or regular work.

1:18:10

* Carlo Lipizzi: higher wages, education, opportunities, and so on. Results

1:18:16

* Carlo Lipizzi: questions.

1:18:23

* Carlo Lipizzi: So you want to have questions. What are the major global migration dynamics and so on.

1:18:24

* Carlo Lipizzi: Data, description.

1:18:33

* Carlo Lipizzi: exploratory data analysis.

1:18:36

* Carlo Lipizzi: preparation.

1:18:40

* Carlo Lipizzi: and then some visualization. So

1:18:45

* Carlo Lipizzi: with this one, you really see who is going where.

1:18:50

* Carlo Lipizzi: So you'll see. I don't know people from Europe, North America going

1:18:55

* to North Africa eventually, and so on.

1:19:03

* Carlo Lipizzi: So source, destination numbers

1:19:09

* Carlo Lipizzi: the range that some from the different regions

1:19:18

* Carlo Lipizzi: global migration in terms of immigration, immigration, self, immigration. meaning from

1:19:26

* I don't know Latin America to Latin America, different

1:19:36

* Carlo Lipizzi: countries eventually.

1:19:40

* Carlo Lipizzi: and so on. That's another interesting one.

1:19:45

* Carlo Lipizzi: So you have going Where? From? Where?

1:19:50

* Carlo Lipizzi: So this is a sun key diagram created for analyzing flows, and this is a flow of people. I mean it's not fluids, but it's the same concept.

1:19:56

* Carlo Lipizzi: same thing. So I none

1:20:12

* Carlo Lipizzi: want to go into each one

1:20:16

* Carlo Lipizzi: conclusions. So that's

1:20:19

* Carlo Lipizzi: I mean, you want to have something at that, because that's your take.

1:20:23

* Carlo Lipizzi: So

1:20:30

* Carlo Lipizzi: migration is ongoing and happens from all over to all over. So

1:20:32

* Carlo Lipizzi: they are working side by side. Somehow.

1:20:39

* Carlo Lipizzi: Obviously.

1:20:45

* Carlo Lipizzi: people in less developed the lower income region are more likely to migrate to more developed, and that's something

1:20:47

* Carlo Lipizzi: that it is not a surprise. A little bit more surprising is people in more developed region, so do not always migrate to more developed region, but may actually migrate back to less developed regions.

1:20:55

* Carlo Lipizzi: I think about the the countries that are growing I think about. I mean, the typical example is China.

1:21:10

* Carlo Lipizzi: I mean. 30 years ago China was a an underdeveloped country. Now it's definitely a a developed country people who migrated from China to us 150,

1:21:20

* Carlo Lipizzi: 30 years ago, or 4 years ago. Now they may want to go back. So it it. It's a an inverse migration that is happening

1:21:33

* Carlo Lipizzi: anyway, so that and then references and some appendix is so that's a way to use the the concept of appendix is

1:21:45

* Carlo Lipizzi: you don't need to have them another big fun of having a lot of tables, but I mean that it's an appendix. You can use it or you can out.

1:21:58

* Carlo Lipizzi: Okay, so I really wanted to share with you those very briefly. Another, I think that they want to share with you is again at something that they already presented. But I want to be sure

1:22:09

* Carlo Lipizzi: that you have it, and you consider it in the proper way is a a data exploration template. So be sure that you have a a structure that is somehow similar to this one. So you want to have a the goals you want to cover those faces that are

1:22:26

* Carlo Lipizzi: business understanding, date, understanding, data, preparation and data representation. So you want to cover all those phases those can be like in

1:22:51

* Carlo Lipizzi: in the document that they represented before as a table of content. You can call them in different ways. So you don't want to call like in this case a data understanding data description that's absolutely fine.

1:23:07

* but you need to have those steps

1:23:25

* Carlo Lipizzi: so, and that for each one you will add the your content. So you have this. Use it not necessarily to create a presentation that the but as a table of content for your a final document.

1:23:28

* Carlo Lipizzi: One more thing that I want to share with you, and then we will spend a few minutes playing with the the in class assignment.

1:23:46

* Carlo Lipizzi: I probably already

1:23:59

* Carlo Lipizzi: share with you a little bit.

1:24:05

* Carlo Lipizzi: Let me share the screen now.

1:24:15

* Carlo Lipizzi: So we are now talking about the chat, Gpt.

1:24:19

* Carlo Lipizzi: The way you present you prompt you

1:24:26

* Carlo Lipizzi: type. The query

1:24:33

* Carlo Lipizzi: will effect heavily the way they bought the

1:24:35

* will provide the answer.

1:24:41

* Carlo Lipizzi: So this case I created the several chats from for

1:24:43

* Carlo Lipizzi: different topics. So this is mixed, then I have a future AI technologies. So

1:24:50

* Carlo Lipizzi: when I place a a query in this chat, I will have an answer that is different from the answer that I could get from another chat.

1:24:59

* Carlo Lipizzi: So

1:25:11

* the more

1:25:12

* Carlo Lipizzi: context you provide for your query, the more accurate is going to be your answer.

1:25:15

* Carlo Lipizzi: So use Chat Gpt. Use it wisely. So keep in mind that that it is likely that those bots will take out the

1:25:22

* Carlo Lipizzi: some of the jobs.

1:25:37

* Carlo Lipizzi: but most likely the jobs that will be lost will be the bottom 10%.

1:25:39

* Carlo Lipizzi: So mit ctl, and you don't want to be in the bottom 10%. So you need to use the boats. But now it be replaced by the bolts 150.

1:25:46

* So the idea of the next assignment

1:25:57

* Carlo Lipizzi: in class you will have just 10 min, so no more than that I just want you to think in those terms

1:26:02

* Carlo Lipizzi: is about.

1:26:12

* Carlo Lipizzi: Let me share again the screen

1:26:16

* Carlo Lipizzi: and let me go here.

1:26:24

* So

1:26:26

* Carlo Lipizzi: this in classics that sizes, exercises on a chat gpt that I, calling the the B in the rest of the description, you want to perform. One of the following tasks, using the Board

1:26:27

* Carlo Lipizzi: called the documentation

1:26:42

* Carlo Lipizzi: input your code to the both and ask for a documentation

1:26:45

* creating a narrative, input your results and get a narrative from the boat

1:26:50

* Carlo Lipizzi: code, writing input, the the requirements. So from one of the first assignment and get the Python code.

1:26:56

* So

1:27:04

* Carlo Lipizzi: i'm going to share a

1:27:08

* Carlo Lipizzi: this in class assignment. You have it. Okay. So let's have a just 10 min, and then we will recombine that. So

1:27:14

* Carlo Lipizzi: I will create breakout rooms. But they are not really necessary, but I didn't do that.

1:27:31

* Carlo Lipizzi: So i'm creating a 3 breakout rooms with 3 participants each.

1:27:39

* Carlo Lipizzi: They are open to you in a less than 10 min.

1:27:45

* Carlo Lipizzi: Okay. So

1:33:00

* Carlo Lipizzi: so we have another 15 s.

1:33:33

* Carlo Lipizzi: 10 s.

1:33:37

* Carlo Lipizzi: All right. So

1:33:49

* Carlo Lipizzi: I want your input so what is your experience? What do you think? What did you do? What did you get

1:33:53

* Carlo Lipizzi: anyone?

1:34:08

* Elyse Spinelli: So we had noticed like it needed a login. So we kind of just my group and I were just. I was kind of telling a narrative about how my team at work actually use Chat Gpt because we were just doing release planning.

1:34:11

* Elyse Spinelli: and our chief architect actually wondered if it could write a description and a acceptance criteria for a story we wanted done for a feature in our Jira, and

1:34:26

* Elyse Spinelli: it was shockingly long, and like, quite accurate to what our goal was for writing the acceptance criteria.

1:34:37

* Carlo Lipizzi: and then, like talking, as in the English, was

1:34:47

* Elyse Spinelli: quite correct as well. So we actually ended up.

1:34:50

* Elyse Spinelli: you know, taking inspiration from it.

1:34:54

* Elyse Spinelli: And I thought, you know, in a professional setting. I thought that was pretty interesting.

1:34:57

* Carlo Lipizzi: but you use it as these at the very end, or you did some modifications.

1:35:01

* Elyse Spinelli: I think our chief architect was like for my own

1:35:11

* Carlo Lipizzi: pride. I'm going to change like, I'm going to add my own tidbits like. For the most part he did keep it

1:35:16

* Carlo Lipizzi: pretty much the same.

1:35:23

* Carlo Lipizzi: Wow: okay, that's great. Okay, Thank you.

1:35:25

* Carlo Lipizzi: Anyone else.

1:35:30

* Kevin Zeng: Yeah. So, Professor, this is Kevin at the At our organization level and in in Lockheed we we do block the use of it.

1:35:34

* Kevin Zeng: But there, there. Isn't really a policy of of using on assigned as a supplement to like.

1:35:40

* Kevin Zeng: you know, for for some some, you know, textual or verbiage, you know, assistance.

1:35:47

* Kevin Zeng: but I found that

1:35:54

* Kevin Zeng: it. It is really helpful if you provide the proper questions, and sometimes you just have to reformat the question. It is nice that it has a context of your previous submissions.

1:35:57

* Kevin Zeng: So that's really cool. I found that really interesting and powerful.

1:36:09

* Carlo Lipizzi: Yeah, I mean, at the very end. It's a sort of next level of a so changing them. So what the search engine is doing? It has a crawler going into the websites and tang.

1:36:14

* Carlo Lipizzi: and then you have a sort of database with the page and tags some. Then you have your query, and is is matching the tags in your query with the tags in the database and giving you the URL. So that's

1:36:29

* Carlo Lipizzi: a very straightforward the approach with the those bots. Instead of having a a tag, you have a pattern or a pattern so, and it is taking the patterns in your query and matching them with patterns

1:36:49

* Carlo Lipizzi: in the database it is, and then is adding a layer of a conversation.

1:37:06

* Carlo Lipizzi: The more patents you provide the in your query, the more matches will be in this sort of pattern, matching that that that you have, and that's why it's so important to to formulate the questions. So in the proper way.

1:37:13

* Carlo Lipizzi: It is becoming a a job, but it is called the the Prompt Engineering. We mentioned that probably last time up. But the way you formulate the query it's really important for the the quality of the answer.

1:37:32

* Carlo Lipizzi: So out of curiosity, what they did the just before the class was a to

1:37:48

* Carlo Lipizzi: fast. So

1:37:59

* Carlo Lipizzi: a script, this one, so saying, write a documentation for the following code.

1:38:04

* Carlo Lipizzi: And it was just a code.

1:38:11

* Carlo Lipizzi: And that's basically what I got.

1:38:13

* I mean, that

1:38:17

* Carlo Lipizzi: is not something that I can use directly.

1:38:18

* Carlo Lipizzi: but consider that this daily support is not a specific chat. If I

1:38:22

* Carlo Lipizzi: would

1:38:31

* Carlo Lipizzi: create something more on the documentation, I would probably get

1:38:31

* something better. But it's still a giving something.

1:38:38

* Carlo Lipizzi: I mean that writing documentation is a pain in the neck, and most of the people write the code, the I mean in theory. You should write code, the and documentation at the same time. The majority of developers don't do that.

1:38:43

* Carlo Lipizzi: So having something that will help you can be helpful.

1:39:02

* Anyway.

1:39:06

* Carlo Lipizzi: It's 806. Does the end of the class? Let me know. Let us know if you have any question on the final.

1:39:07

* Carlo Lipizzi: We we are definitely here to help. If you want to change at the last minute your project and you have a something that you really want to do it.

1:39:18

* Carlo Lipizzi: That's not a problem. But you need to send me not just the overall genetic idea, but more detail the the description all the project you want to do, and the data set you are going to use. Not

1:39:33

* Carlo Lipizzi: 2 page 2 paragraphs is absolutely fine, but they really need to understand what you have in mind. So that's

1:39:50

* Carlo Lipizzi: basically it. And again feel free to work in teams. So that's absolutely fine. Be sure that you are all working on the same project. I mean.

1:40:00

* Carlo Lipizzi: The main reason why I have no problem with this class is because you are professionals. So yes, you are here to get a a good grade. But you are here to learn now.

1:40:16

* Carlo Lipizzi: So

1:40:27

* Carlo Lipizzi: while in my other 6, 24 are things that different, the students may have different goals, they may be less mature. In this class I have no problem. I i'm pretty sure that you know how to do things in a broad way.

1:40:29

* Carlo Lipizzi: So again

1:40:44

* Carlo Lipizzi: thank you for being with me till 8 0 7, and again, if you have questions, so send me and see you an email, and we will get back to you. By the way just out of curiosity. We are

1:40:46

* Carlo Lipizzi: working on creating a sort of Ssc. Gpt. And I will use the transcripts of my classes as a sorts of data that that i'm going to use for training the bought

1:41:03

* Carlo Lipizzi: with those larger language models. You also have the possibility to do what is called fine-tuning.

1:41:24

* Carlo Lipizzi: providing a couple question answers. So those would be conceded before using the

1:41:31

* Carlo Lipizzi: matching of patterns.

1:41:43

* Carlo Lipizzi: So we are collecting the email from the past few semesters and creating a. A. Q. A. That we will feed the to the system to do that

1:41:46

* Carlo Lipizzi: for extracting a. Q. A. From the email. We are using a combination of our chat, Gpt. And New months.

1:42:01

* Carlo Lipizzi: so the initial Ssc. Gpt will be on 6, 24, and will be a sort of a tube or a for 6, 24, so just to let you know.

1:42:14

* Carlo Lipizzi: anyway. So thank you again. 809. Have a good night, and I see you next week again. If you have any question, let me know.

1:42:27

* Carlo Lipizzi: Next week you will present, so we will exchange her emails. In the meantime you will present your final projects, and we will start a sort of a a discussion on that.