   
Kevin Zeng: Oh, do we have class today?

4:41

 Yes, I believe so.

4:46

 Carlo Lipizzi: Hello! Hello, anybody.

5:04

 Carlo Lipizzi: It's

5:08

 Carlo Lipizzi: 6, 30.

5:09

 Carlo Lipizzi: So let's start the class.

5:11

 Carlo Lipizzi: First of all.

5:15

 Carlo Lipizzi: Is there any open issue, any something that you want to bring up and

5:16

 Carlo Lipizzi: review discuss?

5:25

 Scott Guetens: Yeah, Professor, I had a couple of questions about the quiz for this week.

5:27

 Carlo Lipizzi: Yeah. Question 5.

5:31

 Scott Guetens: I I was a little confused. What specifically was asking. Like it? It seemed like, so it says the answer was none of the above, but it it seemed like from like when I was researching.

5:34

 Scott Guetens: It was asking about the like

5:44

 Scott Guetens: classification models. Svm: naive, base rainforest, and none.

5:48

 Scott Guetens: And it said the answer was none, but

5:54

 Scott Guetens: from what I was reading it seemed like any of them could have been applicable.

5:57

 Carlo Lipizzi: Okay, what was the question just reminded me

6:00

 Scott Guetens: it was. Assume you've collected 5,000 textual social media posts, and your objective is to develop a classification model that it classifies into 3 categories, positive, negative and neutral.

6:06

 Scott Guetens: and then which model can be employed for the task.

6:19

 Scott Guetens: And the options were none of the above Svm. Naive bays and random forest. And I said, Svm.

6:23

 Carlo Lipizzi: Yeah, I mean, strictly speaking, that when you want to do a classification.

6:32

 Carlo Lipizzi: I mean either classification or class setting it, because in reality the the question was more on clustering than in classification. If you want to do class setting.

6:43

 Carlo Lipizzi: you may not want to apply models requiring a supervised approach.

6:56

 Carlo Lipizzi: So if you have your social media data set.

7:05

 Carlo Lipizzi: and

7:11

 Carlo Lipizzi: I mean that all those methods, the the the 3 options, are on the supervised, the approach.

7:13

 Carlo Lipizzi: But you don't have a supervision that in a. In that case, if you had a a supervision, and then you can apply pretty much any one of them.

7:25

 Carlo Lipizzi: because you don't the approach has to be unsupervised. and none of them. It's streetly unsupervised.

7:36

 Carlo Lipizzi: Then I I I mean you can stretch things so a a little bit.

7:48

 but the straight answer is none of the

7:54

 Scott Guetens: Okay. All right.

7:58

 Scott Guetens: I also had a question on Number 7. I'm: sorry I don't want to take up too much time.

8:00

 Scott Guetens: and the options are sentiment, analysis, speech, recognition, machine learning, translation, sorry machine translation and advertisement matching.

8:10

 Scott Guetens: So it says the answer is speech recognition. But I don't see how speech recognition is not an application of national language processing.

8:23

 Scott Guetens: Well, again, it's one of those cases, so that there is no real black and white, because it really depends which kind of speech recognition you do.

8:31

 Carlo Lipizzi: meaning that in strict terms it is not really processing text. When you do the process in text.

8:46

 Carlo Lipizzi: You do some of the things that we mentioned last time

8:58

 in the speech recognition. You basically are mentioning words

9:02

 Carlo Lipizzi: that you you here with words that are in in a sort of a cross reference table.

9:10

 Carlo Lipizzi: Then things are never black and white, meaning. There is no one to one

9:20

 Carlo Lipizzi: replacement for the world, for the for the sounds as the world

9:29

 Carlo Lipizzi: meaning. I would advise the number 7, and I will get more answers because

9:35

 Carlo Lipizzi: you are right. It's it's board the line. It really depends how you do the speech re recognition. If in a the speech recognition, you apply semantic rules.

9:45

 Carlo Lipizzi: then at the very end, meaning that that you do one step. That is a kind of a mechanical. But then the th that is a set that is more semantic, and then is more on the Nmp side.

10:00

 Carlo Lipizzi: So that that's a great points call. Yeah, I I appreciate that. Yeah, all right. Thank you very much. I appreciate you taking a look. Sure, Thank you.

10:16

 Carlo Lipizzi: All right. Other questions issues.

10:25

 Carlo Lipizzi: Okay. So let's move on. Now let me start sharing the screen

10:34

 Carlo Lipizzi: and let me go here, minimize this windows here and

10:44

 Carlo Lipizzi: okay. So

10:54

 what we have a

10:58

 Carlo Lipizzi: in the agenda is a reviewing the homeworker. We already spoke about the quizzes.

11:00

 Carlo Lipizzi: Then we will talk about what mining primarily, and

11:08

 Carlo Lipizzi: we will. We will talk a little bit about the final project. 250

11:14

 Carlo Lipizzi: and we will do an in-class exercise. We will talk about web mining, webcrolling, whatever you want to call it.

11:23

 Carlo Lipizzi: I mean the the topic is not as wide the for sure as a natural language processing.

11:37

 Carlo Lipizzi: but it can be pretty wide. It really depends on a specific cases. 200,

11:45

 Carlo Lipizzi: not all the websites are created. Well, I mean you may have a a a crawling that in some cases it's kind of a a straightforward, but in other cases we really require some sort of strategy

11:52

 Carlo Lipizzi: erez agmoni. We are not going to do that that way. We will mention it, but we will not do it. 250.

12:09

 Carlo Lipizzi: We Don't offer courses on web mining the school of business does.

12:17

 Carlo Lipizzi: I'm. Not totally sure that I mean that the topic would require an entire course. But this is just to say that there is a a wide topic while natural language processing it, it. It's really wide, and i'm creating a a courts on an Mp.

12:23

 But on web mining we will

12:43

 Carlo Lipizzi: use some of the approaches. We will do any class exercise, and you will do an assignment at home

12:47

 Carlo Lipizzi: on the same topic, and that could be pretty much it, and then you can use it for a a I don't know for the final or something that you may want to do

12:57

 Carlo Lipizzi: but anyway. So the previous assignment was about

13:09

 Carlo Lipizzi: analyzing the the website Proconn Org, and compare the 2 sides. the pros and the on a given topic.

13:19

 Carlo Lipizzi: So i'm using the same assignment since a while, because it it's pretty generic, and it really depends on what are the topics that you are picking.

13:32

 Carlo Lipizzi: So the example that you will see in my solution is not the most recent one, because I I don't even sure that the same topic is still available about the concept that will be exactly the same.

13:43

 Carlo Lipizzi: So you basically copy and paste the the 2 sides of the opinion. Then you do some cleaning. Then you you calculate the sentiment.

13:59

 Carlo Lipizzi: diagrams and the word cloud, and then you will write a brief record. That was the previous assignment.

14:12

 Carlo Lipizzi: So let me go to by charm.

14:24

 Carlo Lipizzi: and

14:30

 Carlo Lipizzi: that's the code. So i'm importing the different libraries. So the counter for counting elements. We will use a couple of times in this script.

14:32

 Carlo Lipizzi: but that's sentiment for the sentimental analysis work, cloud.

14:47

 Carlo Lipizzi: and not clock, but because it's required by water cloud.

14:51

 Carlo Lipizzi: I

14:56

 use the function we already mentioned the opportunity of doing so. But

14:58

 Carlo Lipizzi: for a a. When you do

15:03

 Carlo Lipizzi: any form of a natural language processing, you may want to have a a a you. You want to create your own function for a cleaning the text, because it's something that you will do over and over, and probably have it once you can reuse it.

15:10

 Carlo Lipizzi: This function, in particular is not the most sophisticated way.

15:28

 Carlo Lipizzi: I wouldn't use it alpha that much, because we know that it's not the most reliable or the detecting. If the content is no medical or not, but is an example of a function for a cleaning text.

15:34

 Carlo Lipizzi: then you have a little bit of explanation. The way is a

15:49

 Carlo Lipizzi: I mean the the comments sort of shaped in this way, because there there is a library in Python. They can go into the code and create the documentation out of the code, using the comments in the code

15:53

 Carlo Lipizzi: posted with this

16:12

 Carlo Lipizzi: specific format.

16:14

 Carlo Lipizzi: But I mean you Don't need to do that, but it's kind of a useful lot to have a a description of what what are the the parameters in, input and what is going to be the parameter for the output.

16:16

 Carlo Lipizzi: because down the road that you may not remember what you did. But you have a clearly as stated here.

16:31

 Carlo Lipizzi: So

16:37

 Carlo Lipizzi: that's function for a cleaning text

16:38

 Carlo Lipizzi: opening the files.

16:43

 Carlo Lipizzi: but I mean at the pro and corner that I previously created. Then, opening the stop or file.

16:46

 Carlo Lipizzi: initializing the the lease so that we like contain those elements. So the list of stopwords to the list of Pro Ancona

16:54

 Carlo Lipizzi: I

17:06

 Carlo Lipizzi: created 2 different variables for the entire line or for the single words

17:07

 Carlo Lipizzi: setting the parameter for the cleaning. One of the parameters is the minimum length for the word.

17:14

 Carlo Lipizzi: Assuming that words that they are shorter than 3 characters. They they

17:22

 Carlo Lipizzi: have no semantic meaning. So you send a minimum length.

17:30

 Carlo Lipizzi: Then, loading the stop word

17:35

 Carlo Lipizzi: the I added the some of the words, so that they know that they are going to be there because this was a pro corona marijuana.

17:41

 Carlo Lipizzi: and obviously cannabis. So marijuana drag. Legal legalization is going, are going to be popular, and

17:52

 Carlo Lipizzi: I mean they are. They may hide so that they want to stress more.

18:04

 Carlo Lipizzi: Then Obed, in the files

18:12

 Carlo Lipizzi: i'm cutting.

18:16

 Carlo Lipizzi: so that are not on the proper length. So in this case.

18:20

 Carlo Lipizzi: if it's not an empty one, and the len is more than 30, because it

18:27

 Carlo Lipizzi: i'm not considering a comments that are too short to get a a real opinion.

18:35

 Carlo Lipizzi: But I mean that

18:43

 I could have

18:44

 Carlo Lipizzi: parameter to make it more flexible, but the same concept. So this is pro and con.

18:46

 Carlo Lipizzi: then i'm doing the cleaning for both of them

18:54

 Carlo Lipizzi: using the function. Then I calculating the the common, the most common 10 for each one printing them.

19:00

 Carlo Lipizzi: Then, calculating the diagrams I created, the

19:13

 Carlo Lipizzi: that will

19:18

 hold the the values

19:21

 Carlo Lipizzi: a

19:24

 Carlo Lipizzi: well. Then doing i'm creating. I mean, this is definitely not the most optimized and the most rigorous way to calculate the diagrams. So I don't know if I mentioned it to you.

19:27

 Carlo Lipizzi: I brought a paper on how to Create diagrams. So it is. It's not exactly an easy task if you do it the right way. So in this case I'm. Just taking all the awards that are one next to the other, and then this this would be

19:43

 Carlo Lipizzi: largely, and then taking those that are most common.

20:02

 Carlo Lipizzi: So that's basically what I'm doing. And I'm adding, I mean, i'm joining the 2 words with an underscore.

20:08

 Carlo Lipizzi: And then taking the 5 most common, I mean

20:18

 Carlo Lipizzi: on those candidate playgrounds. and i'm printing them.

20:22

 Carlo Lipizzi: Then. Sentimentalities. sentiment, analysis.

20:30

 Carlo Lipizzi: The the library is taking a

20:36

 Carlo Lipizzi: strings, and not least I have a list and transforming the list into a string, calculating the the 3 components of the sentiment, positive, negative, and nutral.

20:40

 Carlo Lipizzi: Then for both.

20:53

 Carlo Lipizzi: I could have

20:57

 Carlo Lipizzi: use the function for that, but I mean just to repeat it. It's not elegant, but I mean it's just too, but

21:00

 could have been better

21:11

 Carlo Lipizzi: then printing it

21:13

 Carlo Lipizzi: then

21:17

 Carlo Lipizzi: generating the world cloud so same thing. I I joined the pro cones. I could use the the same that I did before, but that's fine.

21:21

 Carlo Lipizzi: And then

21:32

 Carlo Lipizzi: setting the param, the parameters, for the what cloud generally in the world cloud and printing it.

21:35

 Carlo Lipizzi: So if I run it.

21:44

 Carlo Lipizzi: I have a

21:49

 Carlo Lipizzi: the top words with the quotants. if I most common biograms.

21:53

 Carlo Lipizzi: that kind of make sense so

22:04

 Carlo Lipizzi: sentiment, analysis.

22:08

 Carlo Lipizzi: and then the word cloud here.

22:12

 Carlo Lipizzi: So now, in terms of the interpretation that I mean. Obviously everybody has a their own interpretation. So when you see something like this, you take out the new parallel because it's not giving you much of the the insights.

22:15

 Carlo Lipizzi: So

22:31

 Carlo Lipizzi: you have a the pro that is more on the negative, I mean, both are more on the negative

22:32

 Carlo Lipizzi: sentiment analysis. It's kind of a

22:40

 Carlo Lipizzi: a very row.

22:45

 Carlo Lipizzi: not very reliable measurement.

22:50

 Carlo Lipizzi: When I need to do something that is more, I mean a sensitive. I use different methods. I I use a more than the sentiment, the the emotions.

22:56

 Carlo Lipizzi: I I have a classification of a motion, and I calculate the the similarity, the words so, or the the phrases or the documents may have with those emotions.

23:10

 Carlo Lipizzi: Emotions are are better defined. So we know what an anger is. We know what joy is a sentiment is kind of a

23:23

 Carlo Lipizzi: it's combination of things. So a positive sentiment makes by a joy. But it really depends who you are.

23:36

 Carlo Lipizzi: So, and that's the point. So you have both with more negative

23:47

 Carlo Lipizzi: meaning. The people supporting the legal use of marijuana are stressing more the negative aspects that doesn't mean that the pro

23:53

 Carlo Lipizzi: is actually corner. But is, they are using more negative arguments to support the the pro marijuana

24:08

 Carlo Lipizzi: that could be. If we do not legalize marijuana, there will be more illegal news or more criminality.

24:17

 Carlo Lipizzi: So those are a negative concepts

24:29

 Carlo Lipizzi: on the corner.

24:32

 Carlo Lipizzi: Yeah, and

24:35

 Carlo Lipizzi: meaning against marijuana. They may have used the

24:36

 Carlo Lipizzi: arguments like it would be a sort of endemic user Everybody will use it. There will be no control things like that.

24:43

 Carlo Lipizzi: So what is interesting is the fact that the against marijuana they use more of positive arguments than the the pro marijuana.

24:55

 Carlo Lipizzi: Then that the what the cloud, the I mean it's probably

25:10

 Carlo Lipizzi: I mean I don't know how to interpret the State. That is a so big. But probably the argument was, the legislation should be at the State level.

25:16

 Carlo Lipizzi: The being so big in Washington could be Washington should take a format position in one way or in the other.

25:27

 Carlo Lipizzi: Then alcohol. It's probably saying

25:39

 Carlo Lipizzi: the user of marijuana is not going to be that much different from the user alcohol alcohol. It's legalized. We don't see why we should use we shouldn't we the marijuana, or vice versa, someone could say the number of

25:45

 Carlo Lipizzi: diseases or casualties related to alcohol is pretty high. We do not want to do the same with marijuana, so this could be a reason

26:04

 Carlo Lipizzi: legalizing should go. Then you have Colorado. That was one of the first seats legalizing a marijuana crime could be related to the fact that on the pro and the call the the pro legalization could be. It will reduce the crime.

26:16

 Carlo Lipizzi: The code now could be. People would be less in control of themselves, and there will be there could be more crime, whatever you read them, I mean in this case, i'm. Using just one of the 2

26:35

 Carlo Lipizzi: I don't remember which one it was. It was a and the pro. Okay.

26:53

 Carlo Lipizzi: So that's basically the all assignment that you saw the a python. We verbally analyze the results.

27:00

 Carlo Lipizzi: So if you go in analyzing the diagrams, I mean the

27:12

 Carlo Lipizzi: by, Grams are not telling a lot. Probably you may have wanted to remove the United States

27:21

 Carlo Lipizzi: because it is so popular. So

27:29

 Carlo Lipizzi: the pro the fact that it is economic activity could be related to the fact that then, having a legalized use of marijuana, would be good for the economy creating Jobs

27:33

 Carlo Lipizzi: emergency room. I could be in the call, and I could be. There will be more people in our emergency rooms. Heavy users, so it could be, could point to our individual. So.

27:50

 Carlo Lipizzi: having and heavy use of the

28:05

 Carlo Lipizzi: Madijuana Black Market. I don't know how interpret it, but I would expect more on the pro saying, this will reduce the Black Market.

28:11

 Carlo Lipizzi: but probably could be related to the fact that some people I can think that the

28:25

 Carlo Lipizzi: people may become a a more prone to the use of or recreational drugs, and may get a additional drugs from a black market for drugs that are not very well.

28:32

 Carlo Lipizzi: So I mean that those are all examples of how to interpret the data

28:48

 Carlo Lipizzi: on at the Topwards states Washington I mean, those are

28:56

 Carlo Lipizzi: the same that you saw in the the what cloud the world cloud the the sizes of the words are based on the frequency on the awards. So in the document.

29:02

 Carlo Lipizzi: so. And we have the same thing that we had in the what cloud? So States. Washington alcohol. Those are united is probably part of the United States that, should it be removed.

29:19

 Carlo Lipizzi: meaning

29:34

 Carlo Lipizzi: we should have removed the

29:37

 Carlo Lipizzi: If you go all the way here

29:42

 Carlo Lipizzi: we should have a standard that the so forth. So with the United that because probably

29:46

 Carlo Lipizzi: I mean not States, because United States, we want to remove it. We do not want to remove

29:54

 Carlo Lipizzi: so when they are by itself.

30:03

 Carlo Lipizzi: anyway. So that's all about the the assignment questions issues

30:08

 Carlo Lipizzi: All right. So let's move on

30:22

 Carlo Lipizzi: and let's go to

30:27

 Carlo Lipizzi: the

30:31

 Carlo Lipizzi: web mining there.

30:34

 Carlo Lipizzi: So the reason why you saw

30:37

 Carlo Lipizzi: this presentation for the engineering management program is because we are introducing the program in the Chinese market with the format, that it is a one, a semester in China to some assets in the Us. And and

30:39

 Carlo Lipizzi: Erez Agmoni, and we were reviewing the material. One of the things that was kind of interesting in the conversation we had today was the fact that the school of systems and enterprises as the highest one

30:59

 Carlo Lipizzi: employment rate across Stevens is 98, so the average on stevens is 95. After graduation from a master degree

31:15

 Carlo Lipizzi: so the average is 95. We have 98 in in terms of salary. We have a one or what program so that is a software engineering with $105,000 as a for salary. That is the highest in Stevens

31:29

 we are. The average is around 95 97.

31:49

 Carlo Lipizzi: So I mean just

31:53

 Carlo Lipizzi: give you some

31:56

 Carlo Lipizzi: background information.

31:58

 all right. So what mining using Python?

32:00

 Carlo Lipizzi: We we talk a little bit about the what web binding is, and then we will go in a python to see how to do it

32:05

 Carlo Lipizzi: in practical cases. So what is web binding web? Mining is basically going into into webpage and download the

32:16

 scrape content

32:28

 Carlo Lipizzi: if you think for a second, this is what.

32:32

 Carlo Lipizzi: in a sense, the different web such sites are doing, what Google is doing, what being is doing. So they go into the page. So what they actually do is to read the content.

32:35

 Carlo Lipizzi: tag the content, and then match the tags with the requests that you may have.

32:52

 Carlo Lipizzi: we may want to do something different, because I mean the goal of a so changing is just to let people

33:02

 Carlo Lipizzi: go to the page. This the specific page that is matching the the specific request.

33:12

 Carlo Lipizzi: We may want to do something more. We want to extract the the content to do some semantic analysis. We want to extract some metrics, so to do, some combined the calculation.

33:19

 Carlo Lipizzi: Erez agmoni. There are sites that are collecting information from different sites, aggregating and adding a a layer of integration and some additional analysis.

33:35

 Carlo Lipizzi: You may want to do web mining for a

33:47

 Carlo Lipizzi: optimizing the performances of your own website for seeing what is the distribution of content to analyze if there is any problem going. So for

33:53

 Carlo Lipizzi: several of those examples, web mining is useful that

34:10

 Carlo Lipizzi: even if mining mining, what mining data mining that those things are really different. So data mining is is more on that working on the

34:17

 Carlo Lipizzi: even if sometimes they'd be not so much the being text or or a similar, but it's more confined. So you know pretty much what you are going to get

34:33

 Carlo Lipizzi: with web mining. You don't know what you get, because you don't have control of the content. You may not even know what the structure is, because the owner or the website to change the the the structure overnight.

34:48

 And your beautiful script that was running yesterday is not running today anymore.

35:04

 Carlo Lipizzi: So when you do web mining, you really have quite a lot of issues, and obviously

35:10

 Carlo Lipizzi: the owners of the websites, and in particular, if they offer the content as a as a service for money they want to avoid that. You just go there and scrape the content.

35:20

 Carlo Lipizzi: This is the case for news. This is the case for a sites like a linkedin, so meaning that they are really doing their best to to make the life on the web scrab, or as much difficult as as possible.

35:36

 Carlo Lipizzi: Nevertheless, there are cases where this could be useful anyway. So there are sites that are open. I think about all the sides that are from the government.

35:54

 Carlo Lipizzi: You want to download the all the facts or the official filing of companies. Those are those. Those are information that are publicly available. You want to have the list of patents.

36:08

 Carlo Lipizzi: Things like that that they are, I mean, not saying it is easy to get them, but those are public, and you are not infringing any

36:24

 Carlo Lipizzi: or a limitation just downloading it. You're just doing it in an an automatic way.

36:36

 Carlo Lipizzi: What is the web that we want to? Mine is a a a a a huge number of page, so so no one can really count it. So 63, Leona, it could be 100 threed on, and we wouldn't notice the difference

36:45

 Carlo Lipizzi: that there is a lot of duplication, and some of the page are not indexed. If they are not indexed, they they are part of the so called dark Web.

37:03

 Carlo Lipizzi: Not necessarily. The dark web has to do with the illegal activities and Doesn't do.

37:17

 Carlo Lipizzi: It only means that you cannot. Google them. So Sometimes people do not want to be googled because you

37:25

 Carlo Lipizzi: is, is a pager that is restricted to the people. I know people. I will ask the the URL at the address by myself. And that's it.

37:34

 Carlo Lipizzi: So that's a possibility, meaning that those are not indexed by by Google. So again, not all the page.

37:47

 Carlo Lipizzi: Someone was saying that.

37:57

 Carlo Lipizzi: using the metaphor of the people, the iceberg. So we see the tip of the iceberg. But there is the rest of the iceberg. That is the rest of the web.

37:59

 Carlo Lipizzi: So when we want to do a web mining.

38:13

 Carlo Lipizzi: A few years ago I was teaching this topic to an undergraduate class

38:17

 Carlo Lipizzi: and a certain point. The students were asking me 2 questions. Why do you want to do it? First question and say on that? Is that legal?

38:24

 Carlo Lipizzi: So on the first one. Why you want to do it when there is Google, Google is not giving you the content, but is giving you the point to the content.

38:35

 Carlo Lipizzi: So if you want the actual content, because you want to do something with the content, then you need to go into the specific page. If there is more than one page, then more than one site, then I mean, if you do it manually, we'll take a

38:45

 Carlo Lipizzi: quite some time.

39:05

 Carlo Lipizzi: and then, if the content is changing in time, you want to create a system that will automatically go there and get the content for you instead of doing it each time manually.

39:07

 Carlo Lipizzi: So

39:20

 Carlo Lipizzi: that's why we want to do it. The E. Is it legal? Well, I mean companies like Google made a business out of it.

39:22

 Carlo Lipizzi: So if they are doing it, the the planetary level.

39:35

 Carlo Lipizzi: why shouldn't be able to do it? The at a Mini school level? Why, we can do with the our own web mining.

39:40

 Carlo Lipizzi: So it is legal. Up to a certain point there was a case a a few years ago between a linkedin and a small company.

39:53

 Carlo Lipizzi: So this is a small company.

40:05

 Carlo Lipizzi: was a scraping data from a linkedin, analyzing people up with the sort of

40:07

 Carlo Lipizzi: abnormal user of job search in a linkedin. Assuming that if someone is a changing the path of looking for jobs most likely is looking for a job.

40:19

 Carlo Lipizzi: and the employer may be interested to know, just to either make a counter offer or just let them have the employee, or, if

40:37

 Carlo Lipizzi: so.

40:50

 a

40:51

 Carlo Lipizzi: This small company was collecting the data, adding value every selling the the result. Linkedin.

40:52

 send

41:02

 Carlo Lipizzi: a legal note, saying.

41:04

 Carlo Lipizzi: Erez agmoni, what you are doing is an infringement of the terms of use, because you are supposed just to look at the results, not to use 150

41:08

 to create a business.

41:17

 Carlo Lipizzi: So it the the company was teeny, tiny, and compared to linkedin it

41:19

 Carlo Lipizzi: to take the case for the small company pro bono, just because he or she didn't like the fact of the big giant linkedin suing the teeny, tiny company.

41:29

 Carlo Lipizzi: So the litigation that went on for a quite long time.

41:46

 Carlo Lipizzi: couple of years. So in the meantime a and and there's more company one.

41:52

 Carlo Lipizzi: but in the meantime they lost clients. They lost key employees. They couldn't do Dave job in the proper way, and they went out of business. So the model of the story is, you really need to check. What is the thermal use of the

41:58

 Carlo Lipizzi: web mining that you are doing? May not be illegal, but maybe some restrictions.

42:18

 Carlo Lipizzi: Some companies created the their own business out of a web mining. So, indeed, that now is one of the 2 giants in a job search along with the Linkedin.

42:28

 Carlo Lipizzi: Initially.

42:42

 Carlo Lipizzi: the only thing that we are doing was to collect job postings the

42:46

 Carlo Lipizzi: from a different websites and integrate them in the website, adding a layer of all the analysis I mean filtering and analysis.

42:53

 Carlo Lipizzi: So that was the the initial web model or a business model that they have in the business.

43:06

 Carlo Lipizzi: Then, at the certain point they became pretty big. They started the accepting the accepting a sponsorship for a specific job. So a and some people play their jobs in a

43:15

 Carlo Lipizzi: a sort of an exclusive way into indeed, meaning that they didn't have to scrape the the content. But they have. They had their own, and they still have.

43:30

 Carlo Lipizzi: But the initial business model was a no data that we are, they own, just integrated the content

43:44

 Carlo Lipizzi: it's great from

43:55

 and other websites.

43:58

 Carlo Lipizzi: So when we talk about web mining, there are a few things that we want to consider.

44:01

 So

44:11

 Carlo Lipizzi: first of all, how a web search is is working. So we mentioned Google, we mentioned being so, you basically had the web with the all the paid. So, and they usual links

44:13

 Carlo Lipizzi: that we haven't on on the web. You have a

44:30

 Carlo Lipizzi: this piece of software, the web that is going into the different page and indexing them indexing means tagging by given keywords.

44:33

 Carlo Lipizzi: So each page will have a

44:50

 Carlo Lipizzi: several keywords that are defining, based on the taxonomy that was created by the Google people or the people, whoever they are, to classify somehow the

44:53

 Carlo Lipizzi: So once you have the the page index in the, then you have a a a data set with the all the page and all the the indexes. I mean that the keywords, the tags.

45:12

 Carlo Lipizzi: Then you have the user placing up the equity.

45:27

 Carlo Lipizzi: and then what the search engine is doing is basically matching the keywords. So, and he's a pretty much an exact match between the the keywords. So in the queue and the the tags from the page.

45:32

 Carlo Lipizzi: and then the result will be presented.

45:52

 Carlo Lipizzi: Then things can be more sophisticated, so I can

45:56

 Carlo Lipizzi: automatically correct misspelling, just to be sure, rather than getting what the user really needs, I can have a synonym.

46:03

 Carlo Lipizzi: So

46:13

 Carlo Lipizzi: that's an additional step. So when the user is placing a query, then there is a an intermediate piece, analyzing that kind of natural language processing I mean

46:14

 Carlo Lipizzi: low level, but sort of analyzing the towards the keywords, eventually a little bit of context.

46:26

 Carlo Lipizzi: So that's the basic functionality.

46:37

 Carlo Lipizzi: Google being another. They also have a the placement meaning.

46:40

 Carlo Lipizzi: We mentioned that you have the tags for the page is tagging the query from the user and basically what the add the index

46:50

 Carlo Lipizzi: is doing is okay. I have one tag from or in tans from the greedy. I will match those stunts with the the list of advertisers with the same tang.

47:02

 Carlo Lipizzi: and then, based on who is paying the most. I will place a that piece of add on top of the list.

47:20

 Carlo Lipizzi: So the basic search engine, as no AD management component, but pretty much all of the search engines do have one.

47:30

 Carlo Lipizzi: So

47:45

 Carlo Lipizzi: search engine as a a a crawler that will create the the corpus of those tagging. Obviously, again, you need to define what are the tags.

47:46

 Carlo Lipizzi: so meaning that you need to create a sort of taxonomy of the page before classifying by the specific tags.

48:01

 Carlo Lipizzi: Then you have something that is the indexing. So the spider is just collecting. But then the indexer is creating the cross reference. So you have a tags and page, and you need to be able to go back and forth from one another.

48:13

 Carlo Lipizzi: and then you have a query processor that is getting the

48:30

 Carlo Lipizzi: quidies from the user and matching with what is in the indexer

48:36

 Carlo Lipizzi: again. I I already mentioned that

48:45

 Carlo Lipizzi: you may have a a query process, or that is a little bit of a intelligence in it, being able to work on like a visualization

48:49

 Carlo Lipizzi: on using a bul. So operator. So so in Google you can do, and or you can have a

49:01

 Carlo Lipizzi: must include the with the quotation. So all of these are additional functionalities to the query process. So, but not

49:12

 the the core ones.

49:21

 Carlo Lipizzi: So the data on the Internet, as we know, are based on the concept of a hyperlink. So page with

49:24

 Carlo Lipizzi: plain, basic content that dumb content. And then you have content to where you can click on it, and you will go to a different page. So those are the hyperlinks, and the hybrid links are linking a page one to the other.

49:34

 Carlo Lipizzi: That's an example of a how our we is going back and forth.

49:53

 Carlo Lipizzi: So the

50:00

 Carlo Lipizzi: they are based on a language that is called the HTML hypertext markup language that is a 3 like structure with the

50:04

 Carlo Lipizzi: I mean header and a food, or that will be the same HTML. And then you have the different components, like chapters like paragraphs of a document, and each one as a tanks that are on a in brackets with either the tag

50:18

 or the content, and all the content.

50:38

 Carlo Lipizzi: When you send the the a request that could be just open a page, or or doing something a more elaborated, that you use a a a protocol for a

50:41

 Carlo Lipizzi: requesting the page. So when you go to I don't know Cnn [and.com](http://and.com/) you write http

50:55

 Carlo Lipizzi: call on

51:03

 Carlo Lipizzi: So this Http is basically telling the browser, and then the the server to send your request in a format. That is Http.

51:07

 so and that

51:25

 Carlo Lipizzi: in the Hdp. There will be a Ww. There will be a name Dot, and a

51:26

 domain family.

51:35

 Carlo Lipizzi: and that's the standard for Http, and it's pretty much the same if you are sending a different types of request.

51:37

 Carlo Lipizzi: The request it's that will go from your browser to the server as more information that you may expect. So the server will see

51:47

 Carlo Lipizzi: in this sort of back at the request for for the page. But we'll also see information about the

52:06

 Carlo Lipizzi: who you are. So what is the the browser you you are using? What is the type of encoding that you are using? If there are cougis, what are the cookies that can be used.

52:16

 Carlo Lipizzi: What is the type of connection you have eventually? What is the the IP address that you are using

52:30

 Carlo Lipizzi: the response? The

52:39

 Carlo Lipizzi: you see, as a a page meaning an HTML file, that again, as a more features in it. So you also have a what is a

52:42

 the opening system on the server?

52:55

 Carlo Lipizzi: What is the the type of software that this is in? It is being used. What is the type of the release of the version of a Http. So those are additional information that 99% of the time, so we do not use.

52:57

 Carlo Lipizzi: But if we want to have more information about either the sender or the receiver. Then those are are available in the back and forth on the request

53:14

 Oops!

53:30

 Carlo Lipizzi: So that's another information. I will skip that.

53:31

 Carlo Lipizzi: So the spiders, the crawlers. Again. We can bite them. We need to be sure that we are using the the right protocol to go to the server, and we have the right level, so of the authorization to do it.

53:35

 Carlo Lipizzi: Keeping in mind that that when we work on a on a web page I can change meaning. We are not the owner on the page, and the owner on the page can change it the way they want.

53:58

 Carlo Lipizzi: So if we go into the HTML, and we are looking for one particular tag and take the content. After that tag.

54:14

 Carlo Lipizzi: Yeah, well, the owner can change it if they change it. Our roller is not going to work anymore. Some other times information on on multiple page meaning you need to click on something

54:24

 Carlo Lipizzi: to go to the next page, and then you need to stop when you reach again on the page for that particular information.

54:43

 Carlo Lipizzi: So as you may not, we can do that because we see, we understand, and we act

54:56

 Carlo Lipizzi: a

55:04

 Carlo Lipizzi: for playing dumb cooler. So that's not possible. You cannot go to the page and get it the content if the part, the content is on multiple page.

55:06

 Carlo Lipizzi: so there are ways of doing it. So one of the ways is to have a

55:21

 Carlo Lipizzi: a system that can sort of simulate the the user so there are

55:26

 Carlo Lipizzi: the

55:37

 Carlo Lipizzi: Erez Agmoni, I mean, there are libraries in Python where you can download the an executable version of a browser, and your python 3

55:38

 Carlo Lipizzi: will run this digital browser, and it will act as it was a human.

55:50

 Carlo Lipizzi: So those are a little bit more complex so. But it's what we do when we have a a file with the higher level or complex it

55:59

 Carlo Lipizzi: erez agmoni. When you do a search, it really depends. What is the strategy on the search? So you can do, Bradford, meaning you go 150

56:12

 Carlo Lipizzi: from the root node to the first node. Then you move up horizontally until you reach the end.

56:23

 or you can go

56:30

 Carlo Lipizzi: all the way down and then go up so bathes. If we use a a fi f, or a alli, a full meaning

56:32

 Carlo Lipizzi: per scene for south or or last team for south approach there are pros and cones. So if you use this approach, for example, you can get lost, meaning there could be quite a lot of all the levels.

56:44

 Carlo Lipizzi: and you will never jump to the next one if you use this one while you are completing one level.

57:03

 Carlo Lipizzi: things can change in the meantime, and kind of will lose the initial relationship they had with the the higher level node.

57:12

 Carlo Lipizzi: So again, it really depends on the type of search that you are doing.

57:24

 Carlo Lipizzi: When you get the data from a server you use a protocol that again one side is a HTML in terms of a page that that exchange. But then you use a a protocol to get the page. So

57:31

 Carlo Lipizzi: in Python we use a there are several options, I mean in Python there are always a several options. You can just

57:50

 Carlo Lipizzi: ask Python to go to a certain page and get the HTML. You can use a libraries that are specific for one category of page. So there is a library or a newspaper that is called the newspaper.

58:05

 Carlo Lipizzi: and if you want to download it.

58:23

 Carlo Lipizzi: you need to install as a newspaper 3 K. I don't know why they're using this notation. Probably there is a newspaper that is a different one, and they do not want to be confused. Whatever is there is.

58:28

 So

58:40

 Carlo Lipizzi: one of the approaches is this one? So let me stop here for a second on.

58:42

 Carlo Lipizzi: and

58:49

 Carlo Lipizzi: the

58:53

 let me go

58:54

 Carlo Lipizzi: your.

58:56

 And then, when we go to

58:58

 Carlo Lipizzi: I, Sharma, with an example so beautiful Super, that is, from Ellis in the Wonderland

59:01

 Carlo Lipizzi: is a the Python library that is taking a a 3 like structure like a HTML like Jason.

59:08

 Carlo Lipizzi: ex Amanda.

59:20

 Carlo Lipizzi: and parts of the structure in a, that it's easier to navigate. So, generally speaking, you download the the HTML. You pass the the HTML to beautiful super, and then you start processing it.

59:21

 Carlo Lipizzi: So that's basically what we normally do.

59:41

 Carlo Lipizzi: So before going to the actual code, let me spend 3 min on the Apis.

59:44

 Carlo Lipizzi: I think we already mentioned what an Api is. The Api application program interface is a a protocol for exchanging information between the service.

59:54

 Carlo Lipizzi: So when you

1:00:08

 Carlo Lipizzi: want to download the tweets.

1:00:11

 Carlo Lipizzi: and you use a a a protocol created by Twitter to get or do something on the

1:00:14

 Carlo Lipizzi: the tweets that are in date service.

1:00:27

 Carlo Lipizzi: Either they are to the website, the the the web server, or a store in the in the server.

1:00:32

 Carlo Lipizzi: The Api is a better option compared to

1:00:43

 Carlo Lipizzi: web coding, because

1:00:50

 Carlo Lipizzi: it's a like a contract that you have with the owner of the website. So you are saying, okay, I will give you my credentials. So, using the credentials, so I can do several things. So some of the Apis are free, meaning

1:00:54

 you just need to

1:01:14

 Carlo Lipizzi: create the the the the credentials for the authentication process. But there is no cost associated with that. In some other cases you pay for

1:01:17

 Carlo Lipizzi: you getting the data using the Api. If you have a subscription to a news service, you can download the a certain number of

1:01:33

 Carlo Lipizzi: news based on the contract you have through Apis, so you will specify what are the

1:01:48

 Carlo Lipizzi: I mean, the the the page you want to kind of 3, 6 based on the contract you have, and you will get

1:02:00

 Carlo Lipizzi: everything that you requested. subject to the contract you have.

1:02:08

 Carlo Lipizzi: So if you have a an Api, either a with free or paid access is definitely a better option that just scraping the content. Sometimes you don't have the Api, and

1:02:15

 Carlo Lipizzi: you need to do

1:02:31

 the the scraping

1:02:34

 Carlo Lipizzi: we twitter. You have a a limitation in terms of a number of quidies you can do, because the Api it's free in that case, but with the limitations

1:02:36

 Carlo Lipizzi: they don't offer directly a

1:02:50

 Carlo Lipizzi: paid Api. Probably, you know, mass will change something.

1:02:54

 Yeah. But

1:03:00

 Carlo Lipizzi: in the past it was not possible, and I think it is still not possible.

1:03:01

 Carlo Lipizzi: My dissertation was on analyzing Tweets meaning. I spend quite a lot of time on the

1:03:07

 Carlo Lipizzi: and in the past it was free with limitations in terms of a number of tweets that you can download in an interval of time

1:03:17

 Carlo Lipizzi: unless you do things in different ways.

1:03:28

 Carlo Lipizzi: then that I mean, the way out is to create a time based script, meaning you download the the maximum number of tweets for the time you have. Then you place

1:03:31

 Carlo Lipizzi: your system in a holding mode for

1:03:46

 Carlo Lipizzi: a few seconds, and then you start again.

1:03:54

 In

1:03:57

 Carlo Lipizzi: In that case you will never reach the the maximum number that is allowed a, and you will continue. I mean. We take longer, obviously, but you will get the tweets you want.

1:03:57

 Carlo Lipizzi: So again, that's an example.

1:04:12

 Carlo Lipizzi: There is an Api. You have a a.

1:04:18

 Carlo Lipizzi: So those for our the that you are asking. So the the credentials. So there are 4 credentials in a Tweed.

1:04:22

 Carlo Lipizzi: You can do it directly, using I mean creating the script. I mean a a bomb with the query, asking all the elements what you can use. A library in this case is that Twitter

1:04:37

 Carlo Lipizzi: That will do in this case the authorization for you passing the the the credentials that you created.

1:04:53

 Carlo Lipizzi: and

1:05:05

 Carlo Lipizzi: and then eventually you can use the same library to get the tweets you want.

1:05:07

 Carlo Lipizzi: So let me go now on

1:05:14

 Carlo Lipizzi: the example. So in this example

1:05:19

 Carlo Lipizzi: i'm working on the the New York Times. So if we go here, and we do

1:05:25

 Carlo Lipizzi: so. That's what we have right now on the the New York Times.

1:05:35

 Carlo Lipizzi: So if I go

1:05:40

 Carlo Lipizzi: here, so i'm importing a beautiful super and this library that is the one that we use to go to the website and get the the HTML.

1:05:42

 Carlo Lipizzi: So i'm

1:05:55

 Carlo Lipizzi: getting in body the HTML page. Then i'm using beautiful soup to parts it in a tree like structure.

1:05:58

 Carlo Lipizzi: and then I think, printing the the content.

1:06:10

 Carlo Lipizzi: So if I run it.

1:06:15

 Carlo Lipizzi: that's what I have, so I have

1:06:20

 Carlo Lipizzi: everything that is, the the ruling was a and so on, that this one

1:06:26

 Carlo Lipizzi: the shooter, and this this one.

1:06:34

 Carlo Lipizzi: and so on.

1:06:42

 Carlo Lipizzi: So

1:06:43

 Carlo Lipizzi: I mean, we also have things that we may not need that this 5 min rate 6 min speed that is, those

1:06:45

 Carlo Lipizzi: pieces of information is not something that we really want.

1:06:55

 Carlo Lipizzi: So, anyway.

1:07:00

 Carlo Lipizzi: he is a 7, 32

1:07:02

 Carlo Lipizzi: before we go to the in class exercise

1:07:07

 Carlo Lipizzi: I want to spend a few minutes on

1:07:16

 on the final project.

1:07:24

 Carlo Lipizzi: So

1:07:26

 Carlo Lipizzi: I will publish shortly

1:07:29

 Carlo Lipizzi: the That's it. There we go.

1:07:33

 Carlo Lipizzi: The ideas for the final project, so the ideas for the final project

1:07:39

 Carlo Lipizzi: will be

1:07:47

 Carlo Lipizzi: on this document that you will have online.

1:07:51

 Carlo Lipizzi: The final project is basically defining a problem

1:07:58

 Carlo Lipizzi: having a a data set to address the problem and then do all the data preparation for, and then using scripts, getting insights from the data set

1:08:04

 Carlo Lipizzi: mit Ctl. And that can address the questions you had in

1:08:20

 Carlo Lipizzi: the problem definition.

1:08:26

 Carlo Lipizzi: and then

1:08:28

 they can explain your findings.

1:08:30

 Carlo Lipizzi: That's basically it.

1:08:34

 Carlo Lipizzi: Now the level of complexity of the script has to be at the same level all the

1:08:36

 Carlo Lipizzi: I mean not less, let's say, than the level of complexity or the last assignments.

1:08:47

 Carlo Lipizzi: So if you go and see, there are

1:08:54

 Carlo Lipizzi: that right here we have a

1:08:58

 Carlo Lipizzi: about the

1:09:02

 200 lines. You want to have something that is at least

1:09:04

 Carlo Lipizzi: Erez agmoni, 200 lines, 100, 5,200, then obviously not all the lines are created equal. So there are a lot of comments, some blank lines, 152

1:09:11

 Carlo Lipizzi: additional comments. There is a sort of a redundancy here, the dimension before. So those 2 could be combined in one single creating a

1:09:21

 Carlo Lipizzi: a function

1:09:35

 Carlo Lipizzi: ere I mean that

1:09:39

 Carlo Lipizzi: yeah, If you optimize this code from under the 90, you can go easily to

1:09:44

 Carlo Lipizzi: a little bit more than 100. Let's say 120 something around that

1:09:53

 Carlo Lipizzi: so? You want to have something around the 200? Why, I'm measuring the

1:09:59

 Carlo Lipizzi: complexity by lines of code.

1:10:08

 Carlo Lipizzi: I mean, i'm assuming that there is a a direct correlation between the number of lines and the complexity. That's not always true.

1:10:11

 Carlo Lipizzi: but it's true most of the time. So you want to be sure that you have a code that is, with the I mean complexity enough to go into the data creating a enough insights.

1:10:23

 So

1:10:41

 Carlo Lipizzi: that's the goal that you have a problem. You have a data set. You clean the data. Either. W. Whatever is the way you want to do it, and then you create your your script to analyzing it. Then you use the outcome to write the paper that could be around 10 page 15 page, or included.

1:10:43

 Carlo Lipizzi: It's like some of the last assignments that we did think about the pro con, but expanded.

1:11:08

 Carlo Lipizzi: I'm. Suggesting 3 topics.

1:11:19

 Carlo Lipizzi: You don't need to to use those if you have a a better option if you have something in your mind that you want to analyze, keeping in mind that you need the problem. But you need also the data

1:11:24

 Carlo Lipizzi: and getting the data sometimes is not so easy because either they are not available. They are in a format that is not usable or is not enough data, so there may be quite a lot of issues.

1:11:36

 Carlo Lipizzi: but I leave it in open. So for this class I leave it open for the other classes. I'm teaching

1:11:53

 Carlo Lipizzi: on Wednesdays. I don't give

1:12:01

 Carlo Lipizzi: that students the option, because this class is more, and I can have a I can manage a a multiple

1:12:04

 Carlo Lipizzi: searches problems that you may have in a class with 55 people. It wouldn't be possible.

1:12:14

 Carlo Lipizzi: So we did in the past, but it was a complicated because you have a data set, then you didn't attach the data set, and I need to chase you. What is the data set? But it is this one. But this is not what you use.

1:12:21

 Carlo Lipizzi: I mean that if you multiply by 55 it will become a really difficult to handle

1:12:37

 Carlo Lipizzi: giving in mind that

1:12:44

 Carlo Lipizzi: that we need to 5 the final grade within 72 h after the end of the final

1:12:46

 Carlo Lipizzi: Erez agmoni, meaning we don't have time to all the back and forth on I mean, then there is flexibility. But the rule is 72 h, 2.

1:12:59

 Carlo Lipizzi: So

1:13:10

 Carlo Lipizzi: to shorten up on the time i'm giving you 3 options. So so analyze a people. Migration data is one that you would read it. There is no need for me to go through that analyze research projects.

1:13:11

 Carlo Lipizzi: analyzing COVID-19 data. So each one as a challenges. they are

1:13:26

 pretty much equivalent.

1:13:37

 Carlo Lipizzi: So you will have all the

1:13:40

 Carlo Lipizzi: the data set here.

1:13:45

 Carlo Lipizzi: So that's basically on the final, I think, already published some of the examples. If not, I would do before next week.

1:13:47

 Carlo Lipizzi: Okay. So let me now go into

1:13:58

 Carlo Lipizzi: what is it? Yeah.

1:14:06

 Carlo Lipizzi: So the in-class assignment

1:14:11

 Carlo Lipizzi: you will write a program that will extract data from a web page and perform some analysis. So you can go to New York Times or anyone that you want to do it. You want to print the headlines, generate the word cloud, the for the what's in the headline.

1:14:18

 Carlo Lipizzi: so I will post. Why you will reaggregate in.

1:14:38

 I mean.

1:14:44

 Carlo Lipizzi: in the different rooms this

1:14:46

 Carlo Lipizzi: small script, to be sure that you can use it to. I mean write the code for the in-class assignment.

1:14:49

 and you will have a

1:14:59

 Carlo Lipizzi: let's say 15 min to work on it. So let me open that

1:15:02

 the breakout rooms.

1:15:09

 Carlo Lipizzi: so there are 3 breakout rooms with 2, 3 participants, but each room

1:15:15

 Carlo Lipizzi: I created it. I'm open it. I will pose the the recording, and I will make sure that all the content that you need that will be

1:15:22

 Carlo Lipizzi: available through a canvas. See you in 15 min

1:15:33

 Carlo Lipizzi: resuming the recording up.

1:15:45

 Carlo Lipizzi: So it's a 7, 58. All the rooms are closed.

1:15:48

 and people is coming back. So

1:15:56

 Carlo Lipizzi: questions, issues.

1:16:02

 Okay. So let me

1:16:11

 Carlo Lipizzi: share the screen and let me go through a possible solution.

1:16:13

 So for

1:16:19

 Carlo Lipizzi: this case I use the a library that I was mentioning before, that is newspaper. So when you import the newspaper, you need to make sure that that you import the the library as newspaper 3. K.

1:16:22

 Carlo Lipizzi: So even if, when you use it in a pydon, you will just write in port newspaper. If you try to install a newspaper, you' not going to get it.

1:16:38

 Carlo Lipizzi: but I mean once you install it up. It's it's a little bit more as smart than the

1:16:53

 Carlo Lipizzi: in a way, or doing it, using a request and a beautiful super one.

1:17:02

 Carlo Lipizzi: So, anyway, I imported this library.

1:17:11

 Carlo Lipizzi: I imported. What Cloud and Matt Plot, lib.

1:17:14

 Carlo Lipizzi: I specified what the the the

1:17:20

 is.

1:17:24

 Carlo Lipizzi: Then i'm extracting, using this library. Those are the parameters that that are

1:17:26

 required.

1:17:34

 Carlo Lipizzi: and

1:17:36

 I mean that

1:17:40

 Carlo Lipizzi: the library is giving a a little bit more feature so than the playing request. But you you are with size. You have the number of articles.

1:17:41

 Carlo Lipizzi: Then you have a

1:17:54

 Carlo Lipizzi: in that with

1:17:59

 Carlo Lipizzi: with paper you basically have the content. And then what you are doing here is a downloading the the Rt. Also, and passing it.

1:18:01

 and then

1:18:16

 Carlo Lipizzi: printing the

1:18:17

 Carlo Lipizzi: I mean, I created the some.

1:18:21

 Carlo Lipizzi: So so if the article is now in August, then

1:18:26

 some, I mean i'm reducing the the

1:18:33

 Carlo Lipizzi: I mean that

1:18:39

 title so that i'm getting you not me to do that.

1:18:41

 Carlo Lipizzi: So running it

1:18:45

 Carlo Lipizzi: 131 articles. Those are

1:18:52

 Carlo Lipizzi: the titles, and as the word cloud

1:18:58

 Carlo Lipizzi: and I will pass this, I I would do it right away.

1:19:05

 Carlo Lipizzi: Yeah. So

1:19:17

 Carlo Lipizzi: okay, I need to post it 7. Okay.

1:19:21

 Carlo Lipizzi: all right. So for next week exercise Number 10 mining webpage surprise. So you you will get

1:19:24

 Carlo Lipizzi: any news website. You will bring the headlines so generate the word cloud for the words and diagrams. Calculate the sentiment to write an interpretation of the results. Again, write an interpretation.

1:19:38

 Carlo Lipizzi: so you will import the libraries. You will remove what is not essentially using the the software file.

1:19:53

 Carlo Lipizzi: you will clean using a soap for fights and other the headlines.

1:20:05

 Carlo Lipizzi: You will extract the diagrams, just as in the previous assignment

1:20:12

 Carlo Lipizzi: you will merge the list of single words and diagrams. You will create what cloud

1:20:17

 Carlo Lipizzi: you can or cannot calculate the sentiment, keeping in mind

1:20:25

 Carlo Lipizzi: that when you have a diagram with the underscore, it will not match anything in the sentiment, because it's not a world that that is known by the

1:20:29

 Carlo Lipizzi: and then you will write a. What Pdf. Document 3 plus page with the interpretation.

1:20:42

 Carlo Lipizzi: So be sure that the interpretation is the original, that there are comments that

1:20:51

 I mean that

1:21:01

 Carlo Lipizzi: the document is important because of the very end.

1:21:02

 Carlo Lipizzi: We are not in computer science. So we do not write code. But we use code for doing something.

1:21:06

 Carlo Lipizzi: Okay, so that's basically it. It's 803. Again, my apologies to be a little bit late instead of 8. So let me stop sharing.

1:21:15

 Carlo Lipizzi: So if you don't have any other question, I will make sure that you will have a the in class except size. So the solution on the in class exercise on your canvas, just to use it eventually for a next assignment.

1:21:30

 Carlo Lipizzi: And as usual, if you have any question, send me and she you just to be sure that you have more chances to get an answer. We will get back to you

1:21:49

 Carlo Lipizzi: all right. Yes, Good question for the final. Is it a group assignment or individual?

1:22:01

 Carlo Lipizzi: Well, it's typically an individual assignment that, considering the type of class, we are, I would be okay with doing a a a group assignment.

1:22:08

 Carlo Lipizzi: a small groups meaning not all of you

1:22:23

 as a a single group.

1:22:28

 Carlo Lipizzi: So in the specs I wrote, 2 people could be 3 people. And that's absolutely fine.

1:22:31

 Kevin Zeng: Okay. thank you, Professor.

1:22:40

 Carlo Lipizzi: Do do we have to let me know through email if we form in my group. Yeah, I mean, we want to avoid the the the certain point. We will grade it. And then we would say, you did the an identical, the final and the other, and you cheated. So that's obviously not the case. Yeah.