Audio file

2020-08-05_DS4DR_Text_as_data_34min.mp4

Transcript

00:00:02 Speaker 1

OK, we're going to cut this stuff and recording going on in.

00:00:13 Speaker 1

[Name redacted] here, trying out training videos. So this week we're going to talk about data science for disinformation response and start off the text analysis stuff.

00:00:24 Speaker 1

So let's get on with.

00:00:27 Speaker 1

These are going to be short videos. This is part of the series and they'll probably be three or four different videos for text analysis as part of the social Text Analysis section.

00:00:40 Speaker 1

And recaps data science is a process. Always remember that we're part of a process. We're collecting data, processing it, getting to the end of communicating the results. So this part is just looking at what is the text part of this text as the data recap again.

00:01:00 Speaker 1

Some of the things we're actually trying.

00:01:02 Speaker 1

To do.

00:01:02 Speaker 1

So tactical data tasks.

00:01:06 Speaker 1

Credibility verification. So we are looking at artifacts at the URL level. We're looking at text articles working out whether things contain or don't contain this information, rather misinformation. And we're looking at those sources, so whether.

00:01:26 Speaker 1

A publisher contains misinformation, disinformation. Whether a user is an account that's pushing it out. We're also looking at networks, so we're looking to see if there are networks of inauthentic activity. Tests can help us with that.

00:01:43 Speaker 1

And we're looking at looking for things at bot Nets again.

00:01:47 Speaker 1

Some of the text analysis we can do for looking for things like narratives can help us find those. Some just very, very straight. Hey, are they repeating the same text? Very, very zeroth level type of text work and then activity analysis. This is much more likely what we're going to be doing looking for things like those.

00:02:07 Speaker 1

Accounts looking for things like complicated computational applications. So and and looking at account patterns. So that recap stuff today.

00:02:20 Speaker 1

Going to talk about what we can do with text and specifically we're going to talk about what text actually is.

00:02:28 Speaker 1

So let's start this thing off. Text processing. Whole bunch of things we can do search. We can go look for things that are related to some text we're interested in.

00:02:40 Speaker 1

We do this all the time, so Google Bing searches named entity.

00:02:44 Speaker 1

Mission this is looking for people or places or organizations, objects that are interesting to us through the text. Quite often I I have several variants on my name. I'm actually quite hard to find.

00:03:05 Speaker 1

Because half the official documents refused to recognize the hyphen in my name because I have variants on my name because people misspell it because people forget the [Name redacted]. It's it's.

00:03:18 Speaker 1

A name is not always a name, and it doesn't always look the same. So named entity recognition is not that easy working out whether my name is the same person, so that that's part of information retrieval, that there's also a bunch of learning work we can do like classification.

00:03:37 Speaker 1

That, that part where we were talking about is this misinformation or not?

00:03:42 Speaker 1

So that's a classification problem. If you're feeding in a bunch of articles and saying, is this fake news is it's not fake news is fake news, is not fake news. If you've got stuff.

00:03:51 Speaker 1

That's marked up already.

00:03:53 Speaker 1

And you have classifications already marked up by humans.

00:03:58 Speaker 1

And it might not be. Is this miss facial or is this not misinformation that's actually hard? It might be. Is this targeting a specific group? Is this about a specific topic?

00:04:09 Speaker 1

And go look for more things on that topic that that's related to something called E Discovery. There's a whole legal field around doing that and clustering.

00:04:21 Speaker 1

So I've got this.

00:04:23 Speaker 1

Text or this article what else out there is related to this? What else is similar? And we we do a lot of this.

00:04:30 Speaker 1

By hand it it's like we've got this artifact. Someone's giving us this tweet. What else is talking about the same thing? Not. That's the same in the words, not necessarily connected directly, but what else is about the same thing?

00:04:43 Speaker 1

And we do this for text images. A lot of things. But today we're just talking about the text, so text clustering.

00:04:50 Speaker 1

As a biggie.

00:04:51 Speaker 1

Our topic identification topic following so this is about the narratives. So what are the themes we're seeing within the text? And for us, this is the narratives that we know and love, you know, COVID naive G and the the masks are all bad. 5G is going to kill us all.

00:05:10 Speaker 1

Because COVID doesn't exist, so looking for those topics looking topics we know looking for new ones starting to emerge, sentiment analysis.

00:05:21 Speaker 1

So there's different types of emotion.

00:05:25 Speaker 1

I mean, we, we humans, we have a huge range of emotions and the the kind of the zero thing here the, the the kind of real the, the the one-on-one on this is happy or sad.

00:05:38 Speaker 1

So if you think about it in terms of bank marketing, is are people happy or sad about your brand?

00:05:46 Speaker 1

And that that was kind of the first thing most people working on sentiment analysis. It's like you tag things as this is positive or negative about something feed that through machine learning and feed it something new and this is a positive and negative thing. So there's a bunch of ways to do that. We're going to talk.

00:06:04 Speaker 1

About a couple of those network and.

00:06:08 Speaker 1

Words entities, just like everything else. You can talk about words that that exist together. So Co occurrence is there. There are two words in the same.

00:06:16 Speaker 1

Space. You can talk about people as named entity recognitions. So when those those entities turn up in the same place, you can link them, make a network, do do the same network analysis we've already talked about.

00:06:29 Speaker 1

And then so we've talked about information retrieval, so finding stuff. We've talked about learning. So machine learning effectively as part of this is like making sense of piles of text and comprehension. So this is really starting to understand at a deeper level.

00:06:50 Speaker 1

So learning you can do just off off the words and you can think of text as just a bunch of symbols thrown together.

00:07:01 Speaker 1

But you also.

00:07:02 Speaker 1

Look at text as language and meaning.

00:07:07 Speaker 1

And communication and that that's where we get into natural language process, that's where we start talking about the meaning of sentences. We talk about what you're trying to communicate to another person with a similar mindset, a similar belief system or or not.

00:07:24 Speaker 1

So we also when we're talking about belief systems and mindsets, we talk about translation and we talk about things like localization because just speaking, the same language doesn't mean.

00:07:37 Speaker 1

That you understand in the same way, we still don't talk about truthiness as part of comprehension. So we we work on misinformation all the time. We kind of forget that that's also also part of how we communicate. People lie. People miss out pieces of truth, they they they have assumptions about.

00:07:57 Speaker 1

Each others world beliefs.

00:07:59 Speaker 1

And that gets us suggesting adjusting is some really cool stuff. It it's taking a large amount of text and reducing it down to a short summary.

00:08:13 Speaker 1

And this is one of my favorite pieces of text. Machine learning is take this huge chunk of stuff and just tell us what's important in the.

00:08:23 Speaker 1

And then there's creation. So you've probably seen some of the new stuff about GPT 2, GPT 3 being used on text, and one of the the cool things it can do is you you feed it stuff and it starts generating out, generating out credible text credible sentences, and that for us is interesting.

00:08:43 Speaker 1

Because the bad guys can use that. Actually the good guys can use it too as part of the response, but that that's another topic for another day. So where's this stuff coming from, sources?

00:08:55 Speaker 1

Oops, back to the sources. So so you get text anywhere there is text. It doesn't even have to have to be printed text. You you can do things like download off of radios and that there's a whole project to MIT that that just pulls and radio programs.

00:09:15 Speaker 1

Across across the US and and feeds that and transliterates it. So anywhere there's any form of human verbal communication or text or communication, you've got text.

00:09:27 Speaker 1

So social media, we work on this a lot now. Twitter, Facebook, Reddit, all the things.

00:09:34 Speaker 1

Including all of the text we keep seeing in bloody images. But you know, I just want we're good websites.

00:09:41 Speaker 1

So there are a lot of crawlers out there, so we we know and love the Internet archivearchive.org, but there are also sites like come and crawl which are really useful if we want to use text inputs because they they extract text, they say text in they have API's we can just go call call and pull stuff back.

00:10:01 Speaker 1

To us Wikipedia, Wikidata again put data in text form in ways that we can.

00:10:07 Speaker 1

Use documents. Big document stores, things like the Overview project are really useful for throwing a whole pile of reports and documents, getting OCR, pulling that data back out. OCR is fun.

00:10:21 Speaker 1

It it can be difficult, there's a whole pile of stuff, and that's a separate thing we do. We we've been playing a lot with things like tesseract.

00:10:31 Speaker 1

But there's there's there's a there's work out there, there's stuff we can pull that there's a whole bunch of sources listed already in the big book. So go, go look in there. We'll keep adding to that common tools because most of this stuff is based on Python. So Python libraries, a lot of the machine learning stuff and some of the text.

00:10:51 Speaker 1

Traction stuff. Some of the text management stuff is sitting In Sync.

00:10:55 Speaker 1

Learn if you're getting into natural language processing so the semantic stuff, the grammar based stuff, the NLTK toolkit is really good. Gensim, Spacey. Also good for that. There is some things I'm going to talk about in a minute that are within those libraries that are useful. And if you're going to stand alone, GP.

00:11:15 Speaker 1

Two GPT 3 which I think is invite only at the moment, but you can get your hands on GPT 2 two and Weka is kind of the granddaddy of text munging tools, so thank you news.

00:11:29 Speaker 1

Right. Today, all that and we finally get to what we're talking about, which is that all text is data. And I've been talking about this as we go along. So let let's kind of go have a look at.

00:11:41 Speaker 1

Some of this, and as I'm talking, I'm going to show you bits of code and those bits of code we we've actually got a notebook.

00:11:49 Speaker 1

Sitting inside the GitHub for the team and we have a training folder there now, so you can just go pull that. Just run them use.

00:12:00 Speaker 1

So go, go get some raw text and the thing on the right is the text. So this is some of the COVID naive G tweets. We've been pulling those using the [Name redacted] derived data and it comes in a JSON file format.

00:12:20 Speaker 1

So we need to do a little bit of struggling to get it out of JSON into a pandas data frame because pandas data frame are just magic for machine learning.

00:12:29 Speaker 1

It's basically a row, column, spreadsheet type format, at which point you you can just run tools over it. It's it's it's, it's just easy. So over here we we kind of pull in pandas, we pull in Json.

00:12:47 Speaker 1

We read in that that data frame we clean it up a bit. Just all of that, you know, PD dot data frame, the reset index, the column is just the clean up on that data frame. You end up this row column with those those columns being the URL and the text. So it's the.

00:13:07 Speaker 1

URL for the tweet. So the address of the tweet and the text that's in that tweet.

00:13:13 Speaker 1

And then, because we could going to have a look at what's in it, we're we're just going to stick them all together into one big chunk of chunk of data. So we're just using the join function, which takes a Python list.

00:13:29 Speaker 1

And sticks it together. We're just going to stick it together with spaces. Messy. Horrible.

00:13:34 Speaker 1

Please, God, don't let any real Python coders look at this, but if you do, I apologize to the gods of Python, but we're doing this so we can kind of look at the code, look at the data. So sorry guys, make it cleaner so.

00:13:50 Speaker 1

Another way to look at text is well, so one way to look at text is the meanings in it. So I've talked about this a little, but this is the levels. So one way is to look at the structure of the sentences, the syntax.

00:14:09 Speaker 1

So you're looking at the grammar. You're looking at nouns and verbs.

00:14:15 Speaker 1

So you know the cat sat on the mat. You have nouns like cat and Mat. You have verbs like sat. You have like the fiddly words in the.

00:14:25 Speaker 1

Middle like the.

00:14:29 Speaker 1

But you have things like the the rules about which order they go in, like the British noun verb order is different to the German noun verb order, and you have different things like tense and case and gender agreements. So you have things like gendered words in French and German you don't have.

00:14:49 Speaker 1

In in British you have things like time, cases in verbs in things like English that you don't have in say.

00:15:01 Speaker 1

Mandarin Mandarin is great. You just have one verb. You just add extra words to tell you what when it happened, how it happened, who did it it it's different ways of working, so it's the structure, it's, it's how it's built.

00:15:18 Speaker 1

The sense mating the meaning is semantics, so the next layer up is one of the sentences we had was just light with radiation poisoning. Then it's like, what does, what does this hell does this mean?

00:15:32 Speaker 1

But we can we can kind of radiation poisoning. We know what that is. It's a thing. And we we understand that fruit flies like an apple. So this is kind of a classic sentence because fruit flies like an apple and fruit flies like an arrow are very different sentences, have very different meanings.

00:15:51 Speaker 1

So it's that sense of what you're actually talking about, what you're transmitting to another person. Hey, narratives. Hi and.

00:16:02 Speaker 1

If I wasn't confusing enough there.

00:16:05 Speaker 1

You end up with the pragmatic layer, so quite often you get confusing sentence. It's like, just like radiation poisoning. Then it's like, what the hell was that about?

00:16:15 Speaker 1

And the reason?

00:16:16 Speaker 1

That said, what the hell was that?

00:16:17 Speaker 1

About is because you don't have the context for that sentence.

00:16:21 Speaker 1

You don't know what the just like was about. You don't know what's being compared to radiation poison.

00:16:27 Speaker 1

Think you don't know what the context is and pragmatics? End adds that context is like, what does the meaning around this? What are the the the mental models of the people who are talking about this? What is the verbiage? What are? What are the the sentences around this built up those mental models?

00:16:47 Speaker 1

So it's it's adding that history around that sentence.

00:16:51 Speaker 1

An aphra an aphra. Words like this and that, and their that they're they're placeholders for a thing that you're assumed to know what it is.

00:17:03 Speaker 1

So it's it's like missing information you you fill it in it it it's kind of, you know stuff we do but it's.

00:17:09 Speaker 1

The language version.

00:17:11 Speaker 1

So this is the stuff that natural language processing deals with, you know, syntax semantics from pragmatics. It's it's all that beautiful structural stuff.

00:17:20 Speaker 1

The good news is.

00:17:22 Speaker 1

You don't need to do that **** half the time.

00:17:24 Speaker 1

So most text processing, and certainly most machine learning text processing is handling text as bags of words. So bags of words basically means take all this text, chop it up into words.

00:17:39 Speaker 1

Do stuff with those words. Just count them up so.

00:17:45 Speaker 1

Bags of words. You've got sentences like. Oh, really? And just like with radiation poisoning then?

00:17:52 Speaker 1

Ah, just like with radiation poisoning. Then I think they were talking about 5G exposure. Single words, just like with.

00:18:02 Speaker 1

Then we get things like Tri Grams and bigrams. Remember I said that people can't spell my name.

00:18:07 Speaker 1

So one of the ways we get around that.

00:18:10 Speaker 1

Is instead of matching whole words, we match small pieces of those words.

00:18:15 Speaker 1

And the the the.

00:18:16 Speaker 1

The piece that works best is triples of letters.

00:18:20 Speaker 1

So J US UST St. space T space L, space L.

00:18:26 Speaker 1

I so just like width turns into these triples here and you use that to compare, say two different tweets to see if they were similar to each other enough to be.

00:18:41 Speaker 1

Connected together as a group.

00:18:44 Speaker 1

That's trigrams.

00:18:46 Speaker 1

Now, confusingly, you also get bigrams.

00:18:50 Speaker 1

And try as in three buy as in two.

00:18:54 Speaker 1

Quad as in four so bigrams you'd think would be two letters, but actually it's about two words. So sometimes you're interested in pairs of words instead of 6.

00:19:05 Speaker 1

Words. This gives you a little bit more information like COVID-19 is 2 words quite often and sometimes you want to know everything about COVID-19. Sometimes you want to see things like radiation poisoning together so bigrams just like like with with radiation. Again, you're doing those searches. It just gives you a richer.

00:19:26 Speaker 1

Way of searching and using those bags of words. Quite often you'll add the bigrams to your single words. It'll give you a better search.

00:19:36 Speaker 1

Stop what?

00:19:37 Speaker 1

Stop words as though little ****** words that you really don't need when when you're doing bags of words, when you're searching for.

00:19:45 Speaker 1

Stuff it's like.

00:19:48 Speaker 1

If you're looking for something, if you're doing a search, how often do you type the word? the IT it's just not adding any information.

00:19:59 Speaker 1

So think about this in terms of information. I mean, if just like with radiation poisoning, then you're probably going to search on radiation poisoning. So quite often you throw out all the stops.

00:20:11 Speaker 1

So there's a sentence somewhere down here with on. Then what? So they they they been out in the basically most of the sentence gets thrown out because most of it stop words.

00:20:23 Speaker 1

So we're going to talk about bits of this in a SEC and justice head on over so.

00:20:30 Speaker 1

Ah, stop it. More stop words. So this is Psychic learns default list of English stop words. Not every language has a stop word list yet. This is something I kind of worked on a few years back. There were ways to learn them.

00:20:47 Speaker 1

But it's not a huge list. It's got things like 6 and 60, and meanwhile in it that you can argue that some words should be on it and.

00:20:55 Speaker 1

Some words shouldn't.

00:20:57 Speaker 1

And there's actually I in the presenter notes and in the example code I've added a way to add your own stop words.

00:21:07 Speaker 1

Basically, if you're using these quite often, you have to tune.

00:21:12 Speaker 1

Especially if you're using a domain so.

00:21:15 Speaker 1

We will end.

00:21:16 Speaker 1

Up adding stop words that are specific to this information specific to the domains we're working on, it's just going to happen that way. There will be words that just going to be Internet words that just turn up everywhere and aren't adding information for us. We'll find them.

00:21:33 Speaker 1

OK.

00:21:34 Speaker 1

Named entity recognition.

00:21:37 Speaker 1

Mentioned this already.

00:21:39 Speaker 1

Names of people, organizations, locations and text so we can use these to create social graphs. And Spacey is pretty useful for this. Spacey is good at this.

00:21:51 Speaker 1

So this is easy as you load in. So here is a named entity recognition recognizer. This is the English core in Spacey and just throw it a sentence here. We've done Bill Gates selling 5G COVID-19 data to Microsoft and said go find us all the named entities.

00:22:13 Speaker 1

And we've just run this the NLP space, the NLP on it, and we pulled out that Bill Gates as a person.

00:22:20 Speaker 1

But five is a cardinal number, and that Microsoft is an organization.

00:22:25 Speaker 1

Cool. Great. Do this over big text text. Start building out the graphs. We can do things with this.

00:22:33 Speaker 1

All right, let's go grab some text.

00:22:37 Speaker 1

So This Is Us. Slinging some psychic learn around one of the useful things in machine learning is the feature extraction. So one of the things in here is the count vectorizer, which basically just pulls out all the words, count them all up.

00:22:56 Speaker 1

It's it's not any more exciting than that. Really. It's not. So we've just pulled the coat.

00:23:06 Speaker 1

Pulled in the library, the Count Vectorizer library run it. So we create an instance of it using the English stop words.

00:23:15 Speaker 1

We haven't extended that we can if we want. We've fitted that to so DF text was the all of the text from all of the tweets that.

00:23:26 Speaker 1

We had earlier.

00:23:27 Speaker 1

So we've just fitted it to the text that we had in those tweets and that's given us this array here.

00:23:36 Speaker 1

So the way to read this and go look at the code again is the first number is the tweet, the second number is the code for the word over on the right. I've written a bit of code so you can actually look at the words and the number of counts that you found in the whole old text and.

00:23:56 Speaker 1

So word 0.

00:24:00 Speaker 1

Sorry, sorry, tweaked #0 has one instance of word number 38301 instance of 117592 instances of 282983 tweets are pretty small. Pretty boring, but hell we're running running the tweets at the moment. If we've done this off articles, we'd get a lot.

00:24:19 Speaker 1

Larger camps usually and to the right, we've got the top words. So COVID turns up a lot, 5G turns up.

00:24:28 Speaker 1

A lot our T turns up a lot because this is text. There's a lot of retweets, 19 of course HTTPS because there's a lot of URLs we haven't scraped the URLs off this at Kerch, L and Dell. There was a lot of Spanish language for some reason in the COVID COVID naive G were data.

00:24:47 Speaker 1

So how? What the hell? Do we? No. How do we?

00:24:52 Speaker 1

Make sense of this?

00:24:54 Speaker 1

So, oh, by the way, bigrams, if you want bigrams, you can just modify this slightly. There's in the Count vectorizer after the stop words because English you say ngram range equals 2 comma two again. Also in the example and it gives you the by words as well.

00:25:10 Speaker 1

Backgrounds as well. OK, one thing that people will tell you about is they do A2 TFIDF and it sounds all big and exciting, but it's just another way of looking at that table back there.

00:25:26 Speaker 1

So TF term frequency.

00:25:29 Speaker 1

This is how important each word is to each document.

00:25:36 Speaker 1

And it's like not.

00:25:40 Speaker 1

Really. You know, not that exciting. It's just like.

00:25:43 Speaker 1

How many times does this word turn up versus how many words there are in this document?

00:25:49 Speaker 1

So it's a straight count again.

00:25:54 Speaker 1

So you've got TF is how important a word is to a document to a single document.

00:26:00 Speaker 1

IDF inverse document frequency. So how common is this worked?

00:26:06 Speaker 1

In this corpus, so corpus means all of the documents, so all of the tweets. So how common is this word in?

00:26:12 Speaker 1

All of.

00:26:12 Speaker 1

The tweets and.

00:26:16 Speaker 1

This together gets you a sense.

00:26:19 Speaker 1

Of how important a word is in a document across the whole corpus.

00:26:25 Speaker 1

So how significant is this worked in this place and you get that commonly by just multiplying with two numbers together.

00:26:36 Speaker 1

Is it important? The document? Is it rare in which case it's really important? Psychic learn.

00:26:43 Speaker 1

Again, it's like it learn has a library for that, like learn has a library for lot.

00:26:48 Speaker 1

So thanks. So TFIDF transformer, we've said use IDF because you can do it.

00:26:55 Speaker 1

Without and we just do a fit transform, so generally there's like a pattern to this fit transform transform says just basically do the thing and we do the thing.

00:27:09 Speaker 1

To those word counts. So back over here we created the word counts using count back.

00:27:15 Speaker 1

So these are the word counts here and over here we say.

00:27:21 Speaker 1

Now show us the TFIDF and now instead of having like count of one or two or three, we've got ourselves for the very first up at the top.

00:27:34 Speaker 1

Tweet zero. We've now got sort of .1 and point twos and .0, so you know in tweet zero word 98888881. Just really isn't that important, but we go down to the very last of the tweets we've got because we got, you know, 3800 of them.

00:27:54 Speaker 1

And we see down at the bottom word, 2849 is really quite important. It's like .93 that that's pretty everything rare and important to this document because it's.

00:28:10 Speaker 1

Quite a few of them. And and I went.

00:28:12 Speaker 1

And looked this up. It tells you.

00:28:15 Speaker 1

Which I've not actually seen before, so no idea what it is, but hey, it's rare and important to this to this particular document worth paying some attention to.

00:28:26 Speaker 1

This becomes important when we do machine learning. We're going to talk about that in a later thing.

00:28:33 Speaker 1

Sentiment the fields. Remember I talked about emotion and how we talk about, you know, positive and negative scores, but we can also talk about all the other emotions if you're talking about scoring. Is it this thing or not this thing?

00:28:49 Speaker 1

We can talk about happy. We can talk about sad. We can talk about angry.

00:28:55 Speaker 1

You name an emotion.

00:28:57 Speaker 1

If you're prepared to go through a bunch of texts or tweets and taggers, positive, negative or scoring in some way, we can do the learning of that. So ways to do that. Oh, by the way, yeah, the the thing over.

00:29:11 Speaker 1

To the right.

00:29:12 Speaker 1

That that's a that's a like at scale.

00:29:15 Speaker 1

So very positive down to very negative. You see these a lot. You see an awful lot on things like surveys. I mean, if you've done surveys, you usually see five things and that that's, you know, psychological, psychological thing. It's kind of like a sensible number of things to to have. And there's five things you can't sit.

00:29:33 Speaker 1

On the fence.

00:29:39 Speaker 1

Ways to do this.

00:29:41 Speaker 1

So word based sentiment.

00:29:45 Speaker 1

You can give words individual word skills.

00:29:49 Speaker 1

So you can say that happy.

00:29:53 Speaker 1

Is a very positive thing.

00:29:56 Speaker 1

And bummer is a very negative thing. Or maybe it's just a bit negative, but you can score them.

00:30:03 Speaker 1

And then when you're giving a sentence, you sum up the scores of the words and sentence. So something's gonna be neutral, you know, and doesn't do anything.

00:30:12 Speaker 1

Banana probably doesn't do anything.

00:30:14 Speaker 1

Even though it's not a stop word and and there are whole dictionaries. I'm I'm going to show those in a second.

00:30:22 Speaker 1

Of scores for different different words. Again, lots of English ones, not so much. Other languages actually. So some some have got multiple languages, and so we we may get lucky there.

00:30:37 Speaker 1

No one thing you can do is to use those as a seed for machine learning. So you learn the other words that you learn the sentiment of other words that Co occur.

00:30:45 Speaker 1

With the words that you've got scores for.

00:30:48 Speaker 1

Another one is to score documents. You've got the words in the documents.

00:30:54 Speaker 1

You've got scores on documents you can machine learning to to score individual individual words, individual sentences and get.

00:31:03 Speaker 1

Scores for new sentences you haven't seen before.

00:31:06 Speaker 1

And the really, really hard stuff.

00:31:11 Speaker 1

And this really you're not using it a lot. Most stuff is word based or document based.

00:31:18 Speaker 1

It is thinking about natural language processing.

00:31:21 Speaker 1

Things like satire are really hard. We actually had site our detectors at one thing I worked on because it just shows up as positive even though it's it's completely negative, like slash S just, you know, doesn't show. There's like nice work, bro or, you know, the the usual satirical.

00:31:42 Speaker 1

Irony just doesn't doesn't work on machines and and also you have the language of emoticons. So there is a motion dictionary so sentiment dictionaries they they look like this over to the right is part of an emoticon sentiment dictionary that this just.

00:31:58 Speaker 1

Positive and negative plus 1 -, 1.

00:32:01 Speaker 1

Big Grand Daddy's here wordstat senti wordnet.

00:32:05 Speaker 1

There there, there's more. So you can.

00:32:06 Speaker 1

Go look those up.

00:32:08 Speaker 1

So we just had a little canter through representations.

00:32:12 Speaker 1

And and text and how to think about text?

00:32:21 Speaker 1

And text processing and.

00:32:24 Speaker 1

How it looks like as they.

00:32:27 Speaker 1

So we've looked at it as bags of words. We've talked about it a little in terms of.

00:32:34 Speaker 1

Syntax, semantics, pragmatics, and natural language.

00:32:39 Speaker 1

And we have a lot more to talk about.

00:32:41 Speaker 1

In terms of text, especially things we can do with it, we have a lot more to talk about before we really get to the disinformation.

00:32:49 Speaker 1

Response parts SO1 representation we haven't talked about is text as vectors, so so word vector representations.

00:32:58 Speaker 1

You you basically have a a multiple dimensional space and each word is a point in that space and you can do things like talk about how close.

00:33:08 Speaker 1

In space, words are to each other.

00:33:10 Speaker 1

Which makes things like clustering easier.

00:33:14 Speaker 1

We haven't gone into the machine learning algorithm, so we haven't talked about the classic algorithms like Latent Dirichlet analysis, which is used a lot. We we if you've looked at our our examples on.

00:33:27 Speaker 1

Text learning. You've seen some LDA or support vector machines. We haven't talked about deep learning on text. We haven't talked about some of the natural language processing, and then we'll keep NLTK libraries. We can do that.

00:33:39 Speaker 1

And we really need to talk about this info response application. So if you're a text expert.

00:33:47 Speaker 1

These lectures, these first lectures probably aren't for you. We'll we'll kind of yell when we're doing this info response stuff.

00:33:55 Speaker 1

We we've done.

00:33:56 Speaker 1

This we we are kind of starting on our journey of handing over some some.

00:34:02 Speaker 1

What are we called the videos? Sorry my words fail me on an lecture on text and now now it's over to you.

00:34:12 Speaker 1

Thank you. And we're done.