ML Toolbox

Now, let me tell you about some of the tools that I work with to build machine learning models for reporting.

So one useful tool is Document Cloud, which is a completely free-to-use, open-source tool for helping you understand text documents. So let me give you a quick little demo of Document Cloud. So lots of different organizations upload popular documents PDF's to Document Cloud, and you can even upload your own. You can look at a documents that other people have uploaded. So I thought it'd be fun to take a look at ex-FBI Director James Comey's memos. And you can leave notes on different text within the tools, but another neat feature it has is something called "entity extraction.".

So an entity is something like a person, or a place, or an organization's name, and Document Cloud can identify organizations within Europe for your text. So, for example, we can see here that Trump Tower is not a person. OK. Obama is a person, and he is mentioned at these different locations within the document. Same thing for organizations. So Document Cloud recognizes that the White House is an organization, and here's where it comes up in the document. So Document Cloud is a pretty useful, free, open-source tool to analyze your documents.

There are lots of different ways to approach machine learning, and I've tried to break it into three different categories. On the left side, there are the DIY tools. These are free-to-use, mostly open-source tools. Like for example, libraries that you would use with Python or R, which are two different programing languages. This is sort of the path if you want to really learn about machine learning and data science, and get into the weeds. It's something that you probably need to be pretty competent at programing for.

And if you want to do some of the machine learning analysis I talked about earlier, you have to be willing to commit to learning a bit about data science, and what a model is and how it works, and how to train a model, and so on.

But the advantages that you can run all of these algorithms on your computer, locally, and you don't have to pay anyone or be pulled into any other service. So this is the DIY route. It allows you the most customization. You can build any type of model you want, but it is a bit of work. And you have to know programing and data science. This is the DIY route. I'm not going to talk about that one because I think it deserves probably a lot of time.

But I will talk about these two things on the right. So the Cloud AutoML and MLAPI's. So MLAPI is all the way on the right here. The idea here is that rather than training your own model to something very specific, like identify spy planes, maybe you want to do some of those more common machine learning tasks, like transcribe audio or identify common objects and images. And for that, you can use a model that somebody else has trained. And in this case, surprise, you can use a Google model. You could use some of Google's machine learning API's.

The advantage of this is you need to be able to call an API, so you have to know a little bit of programing. You will be working with the programmer. And it's really easy to use. Like, you really don't have to know a lot about data science to take advantage of these tools. The downsides, of course, are that you're working with the company. So you don't own

everything. You have to pay for the service, and the analysis does run in the Cloud. So depending on your application, this could be a no go. It depends.

And then this other category in the middle cloud, AutoML. This is also a Google tool. I'm really excited to introduce this to journalists because I think it really makes building custom models, that are models for very specific tasks, really approachable to anybody. You don't even really have to know a ton of programing, or even in some cases any programing, to use this tool to build your own custom model. Again, it is a Google tool. You have to pay for it. It runs in the Cloud, but otherwise, super neat.

So let's talk first about the API's. So Google has a bunch of pre-trained models that can help you accomplish common machine learning tasks: speech to text, text to speech. That's transcription and other way around. Text to speech. There's Cloud Vision. This allows you to do a lot of the features I talked about in the machine learning for photo section, like identify emotion, images, objects in images, and extract text. There's natural language. I'll talk about that in a second. Video intelligence helps you analyze videos so it can, for example, produce captions on videos. It can show you what objects are in videos. The translation API, of course, that allows you to translate between languages. And I'll just I'll show you a demo of one of these tools, the Cloud Natural Language Tool, so let's let's take a look at that.

So here I am on the natural language product page, where I can try out the API. Of course, you'd probably want to do this in code. So the first thing that this tool can do is entity analysis. So I actually talked about this a bit earlier because Document Cloud can also do entity analysis. But as you can see, you can identify things like organizations, consumer goods, locations, people, addresses, events, prices, numbers, all sorts of different stuff.

And this is really useful, for example, if you want, well first of all if you want to like parse forms and extract people's phone numbers and addresses. But also maybe you have a collection of articles about politics, and you want to organize the ones that talk about Obama and Trump and how these intersect. So that's where entity extraction can be pretty useful.

There's also a Sentiment Analysis API, where you can see if people are talking about entities positively or negatively. So, for example, if you wanted to see how people are feeling about a political candidate on Twitter, you could apply this API to Tweets. Of course, again, machine learning is probabilistic, so it might not be perfect and might not detect sarcasm. So use with some wisdom. And the tool can also do syntax parsing, identifying parts of speech. And it can also categorize blocks of text.

So that's using a pre-train model. Again, this is a really easy way to get started with machine learning if you don't want to dive into a lot of the dirty details, and if your task is sort of generic, not very specific.

But now let's say you do have one of these specific tasks like, for example, you want to spot illegal amber mining. For this, there's no out-of-the-box tool that identifies satellite photos of illegal mining, so you would have to build your own vision model. And there are many ways you could do it, and I told you that you could do it with the DIY path. But think Cloud AutoML, which is this Google Tool for building custom models, is a really easy way to get started.

The way it works is that you still have to provide a bunch of training data, so you still have to have satellite images of amber and what's considered illegal and what's not. You have to upload all this train data to the Cloud, and then this tool AutoML builds a model for you. And ultimately, it hosts it for you. So you can either make predictions through an API, or you can make predictions within the tool.

So let me just give you an example, a walkthrough of this. So, this is Google Cloud Platform. It's the interface to our Cloud Tools, and I'm in this tool called AutoML Vision. That's our tool for building vision-related custom models. And I'm going to build an object detection model. This is the type of model that not only identifies what's in an image, but actually puts a little box over the location of the thing within the image. And I'm going to try to build a model that identifies planes in satellite images. Not spy planes, just regular planes. So I have all this satellite imagery, and you can see it's actually already been annotated with planes. So, here you can see that there are all these little bounding boxes around planes. And if there is one that hasn't been labeled, which there is right here, I can just go like this and then I've added a new plane.

Now, you need a bunch of labeled data to make this work. So in this case, I have 161 labeled images. And the number of images you need to build a model that's labeled, really depends on the complexity of the task. So, maybe, identifying airplanes is relatively easy, so you need less examples than identifying pneumonia. I don't really know. It's a little bit of trial and error, but you need at least probably a couple hundred images per category.

Now, training a custom model is the easy part. You just click "train new model." Then rename the model. You can optimize it for having faster predictions or higher accuracy. Since we're journalists and we're probably not building real-time apps, we probably want to optimize for accuracy. We don't really care how long the predictions take to make. Then you can set a budget for how long you want the model to run.

So again, this is a paid for thing, but I should also add that you can get a bunch of Google Cloud credits for free if you want to play around with this. And also the types of projects that you would be doing as a journalist are not the ones that are going to cost a lot. You're probably training a model once and making a couple of predictions. So you don't have this high-bandwidth prediction making, so this really shouldn't be a bottleneck cost.

So you hit "train," and it takes about three or four hours to build the model. Then you can evaluate how well the model has done in this evaluate tab, here. You can see "precision recall." Recall is how good was the model at identifying all of the planes and not missing anything. And precision is how good was the model at not mistaking things for planes, so not mislabeling things that weren't planes as planes.

And then you can make predictions on new satellite data right here from within the UI. So I just have this picture on my desktop of Princeton Airport. I just took it from Google Maps, and it has these unlabeled planes. And let's see how well the model can do.

There you go. You can see the model identified lots of different planes in this photo. It also missed. Well, let's see. I honestly can't tell if there are planes here, but it missed a couple of planes. But I think that's good for you to see because, again, the remodels are usually pretty good, but not perfect. So it's important to keep in mind what you're going to do when they make an error.

So I just showed you Cloud AutoML vision. This helps you make custom photos, images, models. But you could also make a custom model like this on tabular data, like BuzzFeed planes on text. You could even classify videos or improve translation models. So there are lots of different offerings.

But anyway, this is my summary of the different tools that I think are the easiest to get started with machine learning. And definitely let me know if you try any of them and find they're good one way or the other.