Ferramentas para aprendizado de máquina

Agora eu vou falar sobre algumas das ferramentas com as quais eu trabalho para construir modelos de aprendizado de máquina para reportagens. Uma ferramenta útil é o Document Cloud, que é uma ferramenta completamente grátis, de código aberto, para ajudar você a entender documentos de texto.

Vou fazer uma demonstração rápida da Document Cloud. Muitas organizações fazem upload de documentos populares em PDF para a Document Cloud, e você pode também fazer isso. Você pode ver os documentos que outras pessoas subiram na ferramenta. Então eu pensei que seria divertido dar uma olhada nos memorandos do ex-diretor do FBI James Comey.

E você pode deixar comentários em vários textos dentro da ferramenta, mas outro recurso legal é algo chamado "entity extraction" (extração de entidade). Uma entidade é algo como uma pessoa ou um lugar, ou o nome de uma organização, e Document Cloud pode identificar organizações dentro do seu PDF ou do seu texto. Assim, por exemplo, podemos ver aqui que... Trump Tower não é uma pessoa. Ok. Obama é uma pessoa, e ele é mencionado nestes diferentes lugares dentro do documento. A mesma coisa para organizações. Document Cloud reconhece que a Casa Branca é uma organização, e é aqui que ela aparece no documento. Então Document Cloud é uma ferramenta muito útil, livre, de código aberto para analisar seus documentos.

Há muitas maneiras diferentes de abordar o aprendizado de máquina, e eu tentei dividi-las em três diferentes categorias. No lado esquerdo, há as ferramentas faça-você-mesmo. Estas são ferramentas grátis, a maioria de código aberto. Como por exemplo, as bibliotecas que você usaria com Python ou R, que são duas linguagens de programação. Este pode ser um caminho se você realmente quer aprender sobre aprendizado de máquina e ciência de dados, e entrar nesse campo. É algo para o qual você provavelmente precisa ser bem competente na programação. E se você quiser fazer algumas das análises com aprendizado de máquina que eu mencionei anteriormente, você tem que estar disposto a se comprometer a aprender um pouco sobre ciência de dados, e o que é um modelo e como ele funciona, e como treinar uma modelo, e assim por diante. Mas a vantagem é que você pode executar todos esses algoritmos no seu computador, localmente, e você não tem que pagar alguém ou ser levado a qualquer outro serviço. Então, esta é a rota faça-você-mesmo. Ela permite a maior personalização. Você pode construir qualquer tipo de modelo que quiser, mas dá um pouco de trabalho. E você precisa saber programação e ciência de dados. Esta é a rota faça-você-mesmo. Não vou falar sobre ela porque acho que ela provavelmente merece muito tempo. Mas vou falar sobre essas duas coisas à direita.

Então, temos o Cloud AutoML e as APIs de aprendizado de máquina. As APIs de aprendizado de máquina estão no canto à direita aqui. A ideia aqui é que em vez de treinar seu próprio modelo para algo muito específico, como identificar aviões espiões, talvez você queira fazer algumas dessas tarefas mais comuns de aprendizado de máquina, como transcrever áudio ou identificar objetos e imagens comuns. E para isso, você pode usar um modelo que alguém já treinou. E, neste caso, surpresa, você pode usar um modelo do

Google. Você pode usar algumas das APIs de aprendizado de máquina do Google. A vantagem disso é que você precisa ser capaz de chamar uma API, então você tem que saber um pouco de programação. Você estará trabalhando com um programador. E é realmente fácil de usar. Você realmente não tem que saber muito sobre ciência de dados para tirar proveito dessas ferramentas. A desvantagem, claro, é que você está trabalhando com uma empresa. Então você não é o proprietário de tudo. Você tem que pagar pelo serviço, e a análise é executada na nuvem. Então, dependendo de sua aplicação, este poderia ser um impeditivo. Depende.

E então esta outra categoria no meio, Cloud AutoML. Esta é também uma ferramenta do Google. Estou realmente animada para apresentar isso a jornalistas, porque acho que realmente torna a construção de modelos personalizados, que são modelos para tarefas muito específicas, realmente acessível a qualquer pessoa. Você nem precisa saber muito de programação, ou qualquer coisa de programação, em alguns casos, para usar esta ferramenta para construir seu próprio modelo personalizado. Novamente, é uma ferramenta do Google. Você tem que pagar para usá-la. Ela é executada na nuvem, mas fora isso, super legal.

Então vamos falar primeiro sobre as APIs. O Google tem um monte de modelos pré-treinados que podem ajudar a realizar tarefas comuns de aprendizado de máquina: speech-to-text (transcrição fala-para-texto), text-to-speech (texto-para-fala). Essas são a transcrição e o caminho oposto, transformar texto em fala. Tem a Cloud Vision. Ela permite que você faça várias das tarefas que eu comentei na parte sobre aprendizado de máquina para fotos, como identificar emoção, objetos em imagens, e extrair texto. Tem a linguagem natural. Vou falar sobre isso daqui a pouco. Inteligência de vídeo ajuda você a analisar vídeos e pode, por exemplo, produzir legendas em vídeos. Pode mostrar quais objetos estão nos vídeos. A API de tradução, claro, que lhe permite traduzir de um idioma para outro.

Vou fazer uma demonstração de uma destas ferramentas, a ferramenta Google Cloud Natural Language, então vamos dar uma olhada nisso. Aqui estou eu na página do produto da Natural Language, onde posso experimentar a API. Claro, você provavelmente vai querer fazer isso em código. Então a primeira coisa que esta ferramenta pode fazer é análise de entidade. Eu comentei sobre isso antes, porque a ferramenta Document Cloud também pode fazer uma análise de entidade. Mas como você pode ver, você pode identificar coisas como organizações, bens de consumo, localizações, pessoas, endereços, eventos, preços, números, várias coisas diferentes. E isso é realmente útil, por exemplo, se você quiser, bem, se você quiser analisar formulários e extrair números de telefone e endereços das pessoas. Mas também talvez você tenha uma coleção de artigos sobre política, e você quer organizar os que falam sobre Obama e Trump e como estes se cruzam. Então é aí que extração de entidade pode ser bastante útil. Há também uma API de análise de sentimento, em que você pode ver se as pessoas estão falando sobre entidades de maneira positiva ou negativa. Assim, por exemplo, se você quisesse ver como as pessoas estão se sentindo sobre um candidato político no Twitter, você poderia aplicar esta API aos tuítes. É claro que, mais uma vez, o aprendizado de máquina é probabilístico, por isso pode não ser perfeito e pode não detectar o sarcasmo. Portanto, use com sabedoria. E

a ferramenta também pode fazer análise sintática, identificação de partes de fala. E também pode categorizar blocos de texto. Então, isso é usar um modelo pré-treinado. Mais uma vez, esta é uma maneira muito fácil para começar com aprendizado de máquina se você não quiser mergulhar em um monte de detalhes sujos, e se a sua tarefa é genérica, não muito específica.

Mas agora vamos dizer que você tem uma dessas tarefas específicas como, por exemplo, você quer flagrar mineração ilegal de âmbar. Para isso, não há nenhuma ferramenta pronta que identifica fotos de satélite de mineração ilegal, então você teria que construir o seu próprio modelo de visão. E você poderia fazer isso de várias maneiras, e eu disse que você poderia fazer com o caminho faça-você-mesmo. Mas acho que Cloud AutoML, que é esta ferramenta do Google para construir modelos personalizados, é realmente uma maneira fácil de começar.

O modo como funciona é que você ainda tem que fornecer muitos dados de treinamento, então você ainda tem que ter imagens de satélite de mineração de âmbar e o que é considerado ilegal e o que não é. Você tem que fazer upload desses dados de treinamento para a nuvem, e então esta ferramenta AutoML constrói um modelo para você. E, finalmente, o hospeda para você. Então você pode fazer previsões por meio de uma API, ou você pode fazer previsões dentro da ferramenta. Vou dar um exemplo, um passo a passo deste processo.

Esta é a plataforma do Google Cloud. É a interface para nossas ferramentas na nuvem, e eu estou nessa ferramenta chamada AutoML Vision. Essa é a nossa ferramenta para a construção de modelos personalizados de visão. Vou construir um modelo de detecção de objetos. Este é o tipo de modelo que não só identifica o que está em uma imagem, mas também coloca uma pequena caixa sobre a localização da coisa dentro da imagem. E eu vou tentar construir um modelo que identifica aviões em imagens de satélite. Não aviões espiões, apenas aviões normais. Então eu tenho essas imagens de satélite, e você pode ver que na verdade elas já foram anotadas com aviões. Aqui você pode ver que estas pequenas caixas estão ao redor dos aviões. E se algum não foi rotulado, como este bem aqui, eu posso apenas vir aqui assim e então eu adicionei um novo avião. Agora, você precisa de muitos dados rotulados para fazer isso funcionar. Neste caso, eu tenho 161 imagens rotuladas. E o número de imagens que você precisa para construir um modelo de rotulagem realmente depende da complexidade do tarefa. Então, talvez, identificar aviões é relativamente fácil, por isso você precisa de menos exemplos do que para identificar pneumonia. Eu realmente não sei. É um pouco de tentativa e erro, mas você precisa de pelo menos provavelmente algumas centenas de imagens por categoria.

Agora, treinar um modelo personalizado é a parte fácil. Você apenas clica em "train new model" (treinar novo modelo). Depois, renomeie o modelo. Você pode otimizá-lo para ter previsões mais rápidas ou maior precisão. Como somos jornalistas e provavelmente não estamos construindo aplicativos em tempo real, nós provavelmente queremos otimizar a precisão. Não nos importamos quanto tempo as previsões vão levar para serem feitas. Depois, você pode definir um orçamento para o tempo que você deseja que o modelo funcione.

Novamente, esta é uma ferramenta paga, mas também devo acrescentar que você pode obter muitos créditos para a Google Cloud de graça se você quiser brincar com isso. E também os tipos de projetos que você estaria fazendo como jornalista não vão custar muito. Você provavelmente está treinando um modelo uma vez e fazendo algumas previsões. Então você não tem essa enorme largura de banda de previsões a serem feitas, então esse realmente não deve ser um custo proibitivo.

Você clica "train" (treinar), e leva cerca de três ou quatro horas para construir o modelo. Então você pode avaliar quão bem o modelo se sai nesta guia de avaliação aqui. Você pode ver "precision recall". Recall é quão bem o modelo se saiu ao identificar todos os aviões e não deixar passar alguma coisa. E precisão é quão bom foi o modelo em não confundir as coisas com aviões, então não rotular erroneamente coisas que não eram aviões como aviões. Daí você pode fazer previsões sobre novos dados de satélite desde aqui, de dentro da interface do usuário. Então eu só tenho esta foto no meu desktop do Aeroporto de Princeton. Eu só peguei do Google Maps, e tem estes aviões sem rótulo. Vamos ver o quão bem o modelo pode se sair. Aí está. Você pode ver que o modelo identificou vários aviões diferentes nesta foto. Ele também deixou alguns para trás. Vamos ver. Eu honestamente não consigo dizer se há aviões aqui, mas ele deixou passar alguns aviões. Mas eu acho que isso é bom que você veja porque, novamente, os modelos são geralmente muito bons, mas não são perfeitos. Portanto, é importante ter em mente o que você vai fazer quando eles cometem um erro.

Então eu acabei de mostrar a ferramenta Cloud AutoML vision. Ela ajuda a fazer modelos personalizados de fotos, imagens. Mas você também pode fazer um modelo personalizado como este com dados tabulares, como os aviões do BuzzFeed, ou texto. Você pode até mesmo classificar vídeos ou melhorar modelos de tradução. Há várias possibilidades diferentes.

De todo modo, este é o meu resumo das diferentes ferramentas que eu considero as mais fáceis para começar com o aprendizado de máquina. E definitivamente me conte se você experimentar qualquer uma delas e achar que elas são boas de um jeito ou de outro.