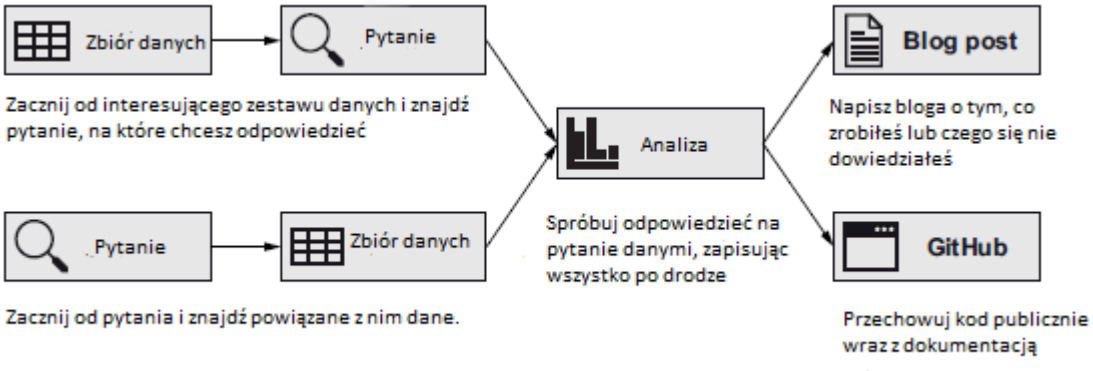
Budowanie portfolio

Właśnie ukończyłeś bootcamp, program studiów, zestaw kursów online lub serię projektów dotyczących danych w swojej obecnej pracy. Gratulacje — jesteś gotowy na pracę naukowca danych! Dobrze? Być może. Część Druga tej książki dotyczy tego, jak znaleźć, ubiegać się i uzyskać stanowisko w zakresie analityka danych, i z pewnością możesz rozpocząć ten proces już teraz. Ale kolejny krok może naprawdę pomóc Ci odnieść sukces: zbudowanie portfolio. Portfoliol to zestaw projektów związanych z analizą danych, które możesz pokazać ludziom, aby mogli zobaczyć, jakiego rodzaju prace związane z analizą danych możesz wykonać. Silne portfolio składa się z dwóch głównych części: repozytoriów GitHub (w skrócie repozytoriów) oraz bloga. Repozytorium GitHub przechowuje kod projektu, a blog prezentuje Twoje umiejętności komunikacyjne i niekodującą część pracy związanej z analizą danych. Większość ludzi nie chce czytać tysięcy linii kodu (twojego repozytorium); chcą szybkiego wyjaśnienia tego, co zrobiłeś i dlaczego jest to ważne (Twój blog). A kto wie – możesz nawet poprosić analityków danych z całego świata o czytanie Twojego bloga, w zależności od tematu. Jakx omówimy później nie musisz po prostu pisać na blogu o wykonanych analizach lub zbudowanych modelach; możesz również wyjaśnić technikę statystyczną, napisać samouczek dotyczący metody analizy tekstu, a nawet podzielić się radą dotyczącą kariery (np. Jak wybrałeś program studiów). Nie oznacza to, że musisz mieć bloga lub repozytorium GitHub wypełnione projektami, aby odnieść sukces jako data scientist. W rzeczywistości większość analityków danych tego nie robi, a ludzie cały czas otrzymują pracę bez portfolio. Ale tworzenie portfolio to świetny sposób, aby wyróżnić się, ćwiczyć swoje umiejętności w zakresie analizy danych i stawać się coraz lepszym. Mamy nadzieję, że to też świetna zabawa! Ta sekcja przeprowadzi Cię przez proces budowania dobrego portfolio. Pierwsza część dotyczy wykonania projektu analizy danych i zorganizowania go w serwisie GitHub. Druga część omawia najlepsze praktyki dotyczące zakładania i udostępniania bloga, aby uzyskać jak najwięcej korzyści z wykonanej pracy. Następnie przeprowadzimy Cię przez dwa prawdziwe projekty, które wykonaliśmy, abyś mógł zobaczyć cały proces od końca do końca.

Tworzenie projektu

Projekt związany z nauką o danych zaczyna się od dwóch rzeczy: interesującego zestawu danych i pytania, które należy zadać. Możesz na przykład wziąć dane ze spisu powszechnego i zapytać „Jak zmienia się w czasie sytuacja demograficzna w całym kraju?” Połączenie pytania i danych jest jądrem projektu, a dzięki tym dwóm rzeczom możesz rozpocząć naukę o danych.



Znajdowanie danych i zadawanie pytań

Kiedy myślisz o tym, jakich danych chcesz użyć, najważniejsze jest znalezienie interesujących Cię danych. Dlaczego chcesz wykorzystać te dane? Twój wybór danych to sposób na pokazanie swojej osobowości lub wiedzy dziedzinowej, którą posiadasz z poprzedniej kariery lub studiów. Jeśli na

przykład zajmujesz się modą, możesz przejrzeć artykuły o Tygodniu Mody i zobaczyć, jak zmieniły się style w ciągu ostatnich 20 lat. Jeśli jesteś zapalonym biegaczem, możesz pokazać, jak zmieniały się Twoje biegi z biegiem czasu, a także sprawdzić, czy Twój czas biegu jest związany z pogodą. Coś, czego nie powinieneś robić, to używać zestawu danych Titanic, MNIST lub innych popularnych początkowych zestawów danych. Nie chodzi o to, że te doświadczenia edukacyjne nie są dobre; mogą być, ale prawdopodobnie nie znajdziesz niczego nowego, co mogłoby zaskoczyć i zaintrygować pracodawców lub nauczyć ich więcej o Tobie. Czasami pozwalasz, aby pytanie zaprowadziło Cię do swojego zbioru danych. Możesz być ciekaw, na przykład, w jaki sposób rozkład płci na kierunkach studiów zmieniał się w czasie i czy ta zmiana jest związana z medianą zarobków po ukończeniu studiów. Następnie udałbyś się do Google i próbował znaleźć najlepsze źródło tych danych. Ale może nie masz palącego pytania, na które czekałeś tylko z umiejętnościami analizy danych. W takim przypadku możesz zacząć od przeglądania zbiorów danych i sprawdzenia, czy możesz zadać jakieś interesujące pytania. Oto kilka sugestii, od czego możesz zacząć:

* Kaggle.com - Kaggle rozpoczęło działalność jako strona internetowa dla konkursów data science. Firmy publikują zestaw danych oraz pytanie i zazwyczaj oferują nagrodę za najlepszą odpowiedź. Ponieważ pytania dotyczą modeli uczenia maszynowego, które próbują przewidzieć coś, na przykład to, czy ktoś nie spłaci pożyczki lub za jaką cenę sprzeda dom, użytkownicy mogą porównać modele na podstawie ich wydajności w zestawie testowym wstrzymania i uzyskać metrykę wydajności dla każdy. Kaggle ma również fora dyskusyjne i „jądra", w których ludzie dzielą się swoim kodem, dzięki czemu możesz dowiedzieć się, jak inni podeszli do zbioru danych. W rezultacie Kaggle ma tysiące zbiorów danych z towarzyszącymi pytaniami i przykładami analizowania ich przez inne osoby. Największą zaletą Kaggle jest również jego największa wada: przekazując (ogólnie oczyszczony) zestaw danych i problem, wykonał za Ciebie dużo pracy. Ty też masz tysiące ludzi borykających się z tym samym problemem, więc trudno jest wnieść wyjątkowy wkład. Jednym ze sposobów użycia Kaggle jest pobranie zestawu danych, ale zadanie innego pytania lub wykonanie analizy eksploracyjnej. Ale ogólnie uważamy, że Kaggle najlepiej nadaje się do uczenia się poprzez rozwiązywanie projektu, a następnie sprawdzanie, jak wypadłeś w porównaniu z innymi, a tym samym uczenie się na podstawie tego, co zrobiły ich modele, a nie jako fragmentu twojego portfolio.

* Zbiory danych w wiadomościach - Ostatnio wiele firm informacyjnych zaczęło upubliczniać swoje dane. Na przykład FiveThirtyEight.com, strona internetowa, która koncentruje się na analizie sondaży, polityce, ekonomii i blogowaniu sportowym, publikuje dane, których może użyć do artykułów, a nawet linki do surowych danych bezpośrednio ze strony z artykułami. Chociaż te zbiory danych często wymagają ręcznego czyszczenia, fakt, że są wiadomości oznaczają, że prawdopodobnie wiąże się z nimi oczywiste pytanie.

* API - API (interfejsy programowania aplikacji) to narzędzia programistyczne, które umożliwiają dostęp do danych bezpośrednio z firm. Wiesz, jak wpisać adres URL i dostać się na stronę internetową? Interfejsy API są jak adresy URL, ale zamiast strony internetowej otrzymujesz dane. Niektóre przykłady firm z przydatnymi interfejsami API to The New York Times i Yelp, które pozwalają odpowiednio pobrać ich artykuły i recenzje. Niektóre interfejsy API mają nawet pakiety R lub Python, które w szczególności ułatwiają pracę z nimi. Na przykład rtweet dla R pozwala szybko pobrać dane z Twittera, dzięki czemu można znaleźć tweety z określonym hashtagiem, jakie są popularne tematy w Kioto lub jakie tweety preferuje Stephen King. Pamiętaj, że istnieją ograniczenia i warunki korzystania z tych interfejsów API. Na przykład w tej chwili Yelp ogranicza Cię do 5000 połączeń dziennie, więc nie będziesz w stanie pobrać wszystkich recenzji. Interfejsy API doskonale nadają się do dostarczania niezwykle niezawodnych, uporządkowanych danych z wielu źródeł.

* Otwarte dane rządowe - wiele danych rządowych jest dostępnych online. Możesz użyć danych ze spisu ludności, danych o zatrudnieniu, ogólnych badań społecznych i ton danych władz lokalnych, takich jak połączenia pod numer 911 lub liczniki ruchu. Czasami możesz pobrać te dane bezpośrednio jako plik CSV; innym razem musisz użyć API. Możesz nawet przesyłać wnioski z Ustawy o wolności informacji do agencji rządowych, aby uzyskać dane, które nie są publicznie wymienione. Informacje rządowe są świetne, ponieważ często są szczegółowe i dotyczą nietypowych tematów, takich jak dane dotyczące zarejestrowanych imion każdego zwierzęcia w Seattle. Wadą informacji rządowych jest to, że często nie są dobrze sformatowane, np. tabele przechowywane w plikach PDF.

* Twoje własne dane - Istnieje wiele miejsc, w których możesz pobrać dane o sobie; serwisy społecznościowe i usługi e-mail to dwa duże. Ale jeśli używasz aplikacji do śledzenia aktywności fizycznej, listy lektur, budżetu, snu lub czegokolwiek innego, zwykle możesz również pobrać te dane. Może mógłbyś zbudować chatbota na podstawie wiadomości e-mail ze współmałżonkiem. Możesz też przyjrzeć się najczęściej używanym słowom w tweetach i tym, jak te słowa zmieniały się w czasie. Być może mógłbyś monitorować spożycie kofeiny i ćwiczenia przez miesiąc, aby sprawdzić, czy możesz przewidzieć, ile i dobrze śpisz. Zaletą korzystania z własnych danych jest to, że Twój projekt ma gwarancję, że jest wyjątkowy: nikt inny wcześniej nie przeglądał tych danych!

* Web scraping - Web scraping to sposób na wyodrębnienie danych ze stron internetowych, które nie mają interfejsu API, zasadniczo poprzez zautomatyzowanie odwiedzania stron internetowych i kopiowania danych. Możesz stworzyć program do przeszukiwania witryny filmowej pod kątem listy 100 aktorów, ładowania profili aktorów, kopiowania list filmów, w których się znajdują, i umieszczania tych danych w arkuszu kalkulacyjnym. Musisz jednak zachować ostrożność: skrobanie witryny może być niezgodne z warunkami użytkowania witryny i możesz zostać zbanowany. Możesz sprawdzić plik robots.txt witryny, aby dowiedzieć się, co jest dozwolone. Chcesz być miły dla stron internetowych: jeśli odwiedzasz witrynę zbyt wiele razy, możesz ją wyłączyć. Ale zakładając, że warunki korzystania z usługi na to pozwalają i budujesz w czasie między trafieniami, scraping może być świetnym sposobem na uzyskanie unikalnych danych.

Co sprawia, że projekt poboczny jest interesujący? Naszą rekomendacją jest wybór analizy eksploracyjnej, w której każdy wynik prawdopodobnie nauczy czegoś czytelnika lub zademonstruje twoje umiejętności. Możesz utworzyć interaktywną mapę 311 połączeń w Seattle, oznaczonych kolorami według kategorii; ta mapa wyraźnie pokazuje twoje umiejętności wizualizacji i pokazuje, że możesz pisać o pojawiających się wzorcach. Z drugiej strony, jeśli spróbujesz przewidzieć rynek akcji, prawdopodobnie nie będziesz w stanie tego zrobić, a pracodawcy trudno jest ocenić twoje umiejętności, jeśli wynik jest negatywny. Kolejną wskazówką jest sprawdzenie, co pojawia się, gdy wyszukujesz swoje pytanie w Google. Jeśli pierwszymi wynikami są artykuły w gazetach lub posty na blogu, które odpowiadają dokładnie na pytanie, które zadałeś, możesz przemyśleć swoje podejście. Czasami możesz rozszerzyć analizę innej osoby lub wprowadzić inne dane, aby dodać kolejną warstwę do analizy, ale może być konieczne rozpoczęcie procesu od nowa.

Wybór kierunku

Budowanie portfolio nie musi być ogromnym zaangażowaniem czasowym. Doskonałość jest tu zdecydowanie wrogiem dobrego. Coś jest lepsze niż nic; Pracodawcy szukają przede wszystkim dowodów na to, że możesz kodować i komunikować się o danych. Możesz się martwić, że ludzie będą patrzeć i śmiać się z twojego kodu lub mówić: „Wow, pomyśleliśmy, że ta osoba może być w porządku, ale spójrz na jej okropny kod!” Jest bardzo mało prawdopodobne, że tak się stanie. Jednym z powodów jest to, że pracodawcy dostosowują swoje oczekiwania do poziomu stażu pracy: nie będziesz musiał kodować jak magister informatyki, jeśli jesteś początkującym naukowcem zajmującym się danymi.

Ogólnie rzecz biorąc, większym zmartwieniem jest to, że w ogóle nie możesz kodować. Chcesz specjalizować się w wizualizacji? Stwórz interaktywny wykres za pomocą D3. Czy chcesz przetwarzać język naturalny? Użyj danych tekstowych. Nauczanie maszynowe? Przewiduj coś. Wykorzystaj swój projekt, aby zmusić się do nauczenia się czegoś nowego. Przeprowadzenie tego rodzaju praktycznej analizy pokaże ci luki w twojej wiedzy. Kiedy dane, którymi naprawdę jesteś zainteresowany, znajdują się w sieci, nauczysz się web scrapingu. Jeśli uważasz, że dany wykres wygląda brzydko, dowiesz się, jak tworzyć lepsze wizualizacje. Jeśli uczysz się samodzielnie, wykonanie projektu to dobry sposób na przezwyciężenie paraliżu związanego z niewiedzą, czego się dalej uczyć. Częstym problemem związanym z projektami z własnej motywacji jest przekroczenie zakresu. Overscoping to chęć zrobienia wszystkiego lub dodawania kolejnych rzeczy w miarę postępu. Zawsze możesz ulepszać/edytować/uzupełniać, ale wtedy nigdy nie kończysz. Jedną ze strategii jest myślenie jak Hollywood i tworzenie sequeli. Powinieneś zadać sobie pytanie i odpowiedzieć na nie, ale jeśli uważasz, że możesz chcieć wrócić do niego później, możesz zakończyć badania pytaniem lub tematem do dalszego zbadania (lub nawet „Kontynuować …?", jeśli tak musieć). Innym problemem jest brak możliwości obrotu. Czasami żądane dane nie są dostępne. Albo jest tego za mało. Albo nie jesteś w stanie go wyczyścić. Takie sytuacje są frustrujące i w tym momencie łatwo można się poddać. Ale warto spróbować dowiedzieć się, jak możesz uratować projekt. Czy wykonałeś już wystarczająco dużo pracy, aby napisać samouczek na bloga, być może o tym, jak zebrałeś dane? Pracodawcy poszukują osób, które uczą się na swoich błędach i nie boją się do nich przyznać. Samo pokazanie, co poszło nie tak, aby inni mogli uniknąć tego samego losu, nadal jest cenne.

Wypełnianie GitHub README

Może jesteś na bootcampie lub na studiach, w których już robisz własne projekty. Zadeklarowałeś nawet swój kod na GitHub. Czy to wystarczy? Nie! Minimalnym wymogiem dla użytecznego repozytorium GitHub jest wypełnienie pliku README. Masz kilka pytań, na które musisz odpowiedzieć:

* Czym jest projekt? Jakie dane wykorzystuje? Na jakie pytanie odpowiada? Jaki był wynik: model, system uczenia maszynowego, pulpit nawigacyjny czy raport?

* Jak zorganizowane jest repozytorium? To pytanie sugeruje oczywiście, że repozytorium jest w rzeczywistości zorganizowane w jakiś sposób! Istnieje wiele różnych systemów, ale podstawowym jest podzielenie skryptu na części: pobranie (jeśli dotyczy) danych, czyszczenie ich, eksploracja i ostateczna analiza. W ten sposób ludzie wiedzą, dokąd się udać, w zależności od tego, co ich interesuje. Sugeruje to również, że zachowasz swoją pracę zorganizowaną, gdy idziesz do pracy w firmie. Firma nie chce ryzykować, że cię zatrudni, a potem, gdy nadejdzie czas, aby przekazać projekt, dajesz komuś niekomentowany, 5000-wierszowy skrypt, którego rozgryzienie i użycie może być dla niego niemożliwe. Dobre zarządzanie projektami pomaga również w przyszłości: jeśli chcesz ponownie wykorzystać część kodu później, będziesz wiedział, dokąd się udać.

Ale chociaż robienie projektu i udostępnianie go publicznie w udokumentowanym repozytorium GitHub jest dobre, bardzo trudno jest spojrzeć na kod i zrozumieć, dlaczego jest to ważne. Po wykonaniu projektu następnym krokiem jest napisanie posta na blogu, który pozwoli ludziom dowiedzieć się, dlaczego to, co zrobiłeś, było fajne i interesujące. Nikogo nie obchodzi pet_name_analysis.R, ale wszystkich obchodzi „Użyłem R, aby znaleźć najgłupsze imiona zwierząt!"

Założenie bloga

Blogi pozwalają pochwalić się swoimi myślami i projektami, ale mogą również oferować nietechniczne spojrzenie na twoją pracę. Wiemy, wiemy - właśnie nauczyłeś się wszystkich tych wspaniałych technicznych rzeczy! Chcesz się tym pochwalić! Jednak bycie naukowcem zajmującym się danymi

prawie zawsze wiąże się z przekazywaniem swoich wyników laikom, a blog da ci doświadczenie w tłumaczeniu procesu analizy danych na język biznesowy.

Potencjalne tematy

Załóżmy, że stworzyłeś bloga. Czy ludzie naprawdę będą zainteresowani Twoimi projektami? Nie masz jeszcze tytułu naukowca danych; jak możesz kogoś czegoś nauczyć? Warto pamiętać, że najlepiej jest uczyć ludzi kilka kroków za sobą. Zaraz po zapoznaniu się z koncepcją, taką jak używanie ciągłej integracji dla swojego pakietu lub tworzenie modelu TensorFlow, nadal rozumiesz nieporozumienia i frustracje, które miałeś. Wiele lat później trudno jest postawić się w myśleniu początkującego. Czy kiedykolwiek miałeś nauczyciela, który był wyraźnie bardzo inteligentny, a mimo to w ogóle nie potrafił przekazywać pojęć? Nie wątpiłeś, że znają temat, ale nie mogli go dla ciebie rozbić i wydawali się być sfrustrowani, że nie od razu go zrozumiałeś. Spróbuj myśleć o swoich odbiorcach jak o sobie sześć miesięcy temu. Czego się nauczyłeś od tego czasu? Jakie zasoby chcesz, aby były dostępne? To ćwiczenie jest również świetne do świętowania twoich postępów. Mając tak wiele do nauczenia się w nauce o danych, łatwo poczuć że nigdy nie zrobiłeś wystarczająco dużo; zatrzymanie się, aby zobaczyć, co osiągnąłeś, jest miłe. Posty na blogu poświęcone naukom o danych można pogrupować w cztery kategorie:

* Samouczki z dużą ilością kodu - samouczki pokazują czytelnikom, jak robić takie rzeczy jak web scraping lub głębokie uczenie się w Pythonie. Twoi czytelnicy będą na ogół innymi aspirującymi lub praktykującymi naukowcami danych. Chociaż samouczki nazywamy ciężkim kodem, zwykle nadal będziesz chciał, aby było tyle wierszy tekstu, co kod, jeśli nie więcej. Kod generalnie nie jest oczywisty; musisz przeprowadzić czytelnika przez to, co robi każda część, dlaczego chcesz to zrobić i jakie są wyniki.

* Samouczki z dużą ilością teorii - te samouczki uczą czytelników koncepcji statystycznej lub matematycznej, na przykład tego, czym jest empiryczna analiza Bayesa lub jak działa analiza głównych składowych. Mogą mieć jakieś równania lub symulacje. Podobnie jak w przypadku samouczków z dużą ilością kodu, twoją publicznością są zwykle inni naukowcy zajmujący się danymi, ale powinieneś pisać tak, aby każdy, kto ma trochę wiedzy matematycznej, mógł podążać za nimi. Samouczki z rozbudowaną teorią są szczególnie dobre do zademonstrowania umiejętności komunikacyjnych; panuje stereotyp, że wielu technicznych ludzi, zwłaszcza jeśli mają stopień doktora, nie potrafi dobrze wyjaśnić pojęć.

* Fajny projekt, który zrobiłeś - Mamy nadzieję, że Cię przekonaliśmy, nie musisz pracować tylko nad przełomowym rozpoznawaniem obrazów medycznych. Możesz również dowiedzieć się, który z filmów Zmierzch używa tylko słów z Burzy Szekspira. Na przykład Julia Silge użyła sieci neuronowych do wygenerowania tekstu, który brzmi jak Jane Austen. Te posty na blogu mogą bardziej skupiać się na wynikach lub procesie, w zależności od tego, jaka była najciekawsza część twojego projektu.

* Zapisywanie swoich doświadczeń - nie musisz tylko pisać na blogu o samouczkach lub swoich projektach z zakresu analizy danych. Możesz opowiedzieć o swoich doświadczeniach na spotkaniu lub konferencji poświęconej analizie danych: jakie prelekcje Cię zainteresowały, porady dla osób, które wybierają się na ich pierwsze spotkanie lub jakie zasoby udostępnione przez prelegentów. Ten rodzaj postu może być pomocny dla osób, które rozważają udział w tym samym wydarzeniu w następnym roku lub które nie mogą uczestniczyć w konferencjach z powodów logistycznych lub finansowych. Ponownie, tego typu posty na blogu dają potencjalnym pracodawcom wgląd w to, jak myślisz i komunikujesz się.

Logistyka

Ale gdzie umieścić swoje ciekawe pisarstwo? W przypadku bloga masz dwie główne opcje:

* Stwórz własną stronę internetową. Jeśli pracujesz w języku R, sugerujemy skorzystanie z pakietu blogdown, który umożliwia stworzenie strony internetowej dla bloga przy użyciu kodu R (dzikiego, prawda?). Jeśli używasz Pythona, Hugo i Jekyll to dwie opcje, z których obie umożliwiają tworzenie statycznych witryn blogowych i dostarczanie wielu motywów, które zbudowali inni ludzie, umożliwiając pisanie postów na blogu w przecenach. Sugerujemy, abyś nie martwił się zbytnio swoim motywem i stylem; po prostu wybierz ten, który Ci się podoba. Nie ma nic gorszego niż niepisanie postów na blogu, ponieważ zbytnio się rozproszyłeś, zmieniając wygląd bloga. Prosty jest prawdopodobnie najlepszy; zmiana motywu może być uciążliwa, więc lepiej nie wybierać takiego, na który możesz mieć dość za sześć miesięcy.

* Korzystanie z Medium lub innej platformy blogowej. Medium to bezpłatna, internetowa platforma wydawnicza. Firma na ogół nie pisze treści; zamiast tego zawiera treści setek tysięcy autorów. Medium i podobne strony to dobre opcje, jeśli chcesz szybko zacząć, ponieważ nie musisz się martwić o hosting lub uruchomienie strony internetowej; wystarczy kliknąć „Nowy post", zacząć pisać i publikować. Możesz również uzyskać większy ruch, gdy ludzie przeszukują witrynę blogów w poszukiwaniu terminów takich jak data science lub Python. Ale jedną obawą jest to, że jesteś na łasce platformy. Jeśli firma zmieni model biznesowy i umieści wszystko za zaporą, na przykład, nie możesz nic zrobić, aby Twoje posty na blogu były bezpłatne. Nie możesz też tworzyć prawdziwej sekcji biograficznej ani dodawać innych treści, takich jak strona z linkami do przemówień, które wygłosiłeś.

Jednym z typowych pytań, jakie ludzie mają na temat blogowania, jest to, jak często muszą publikować i jak długie powinny być te posty. Te rzeczy są zdecydowanie osobistymi wyborami. Widzieliśmy ludzi, którzy prowadzą mikroblogi, publikując krótkie posty wiele razy w tygodniu. Inne osoby spędzają miesiące między postami i publikują dłuższe artykuły. Istnieją pewne ograniczenia; chcesz mieć pewność, że Twoje posty nie zaczną przypominać Ulissesa. Jeśli Twój post jest bardzo długi, możesz podzielić go na części. Ale chcesz pokazać, że potrafisz komunikować się zwięźle, ponieważ jest to jedna z podstawowych umiejętności w zakresie analizy danych. Kierownictwo, a nawet Twój menedżer, prawdopodobnie nie chcą lub nie muszą słyszeć wszystkich fałszywych startów, które miałeś, ani 20 różnych rzeczy, których próbowałeś. Chociaż możesz zdecydować się na krótkie podsumowanie swoich falstartów, musisz szybko dotrzeć do sedna i ostatecznej ścieżki. Jedynym wyjątkiem jest sytuacja, w której ostatnia metoda zaskoczy czytelników. Jeśli na przykład nie użyłeś najpopularniejszej biblioteki do rozwiązania problemu, możesz wyjaśnić, że nie skorzystałeś, ponieważ okazało się, że biblioteka nie działa. A co jeśli martwisz się, że nikt nie przeczyta Twojego bloga, a cała Twoja praca pójdzie na marne? Cóż, jednym z powodów posiadania bloga jest to, że pomaga on w podawaniu pracy. Możesz umieścić linki do swoich postów na blogu w swoim CV, gdy odwołujesz się do projektów z zakresu analizy danych, a nawet pokazujesz je osobom biorącym udział w wywiadach, zwłaszcza jeśli posty mają ładne interaktywne wizualizacje lub pulpity nawigacyjne. Nie jest ważne, aby mieć setki czy tysiące czytelników. Może być fajnie, jeśli otrzymasz komentarz na poziomie Medium lub jeśli pojawisz się w biuletynie firmy zajmującej się badaniem danych, ale ważniejsze jest posiadanie odbiorców, którzy będą czytać, cenić i angażować się w materiał, niż mieć wysokie wskaźniki. Nie oznacza to, że nie możesz nic zrobić, aby zbudować czytelnictwo. Po pierwsze, powinieneś się reklamować; chociaż to banał, posiadanie #marki jest przydatne do budowania sieci w dłuższej perspektywie. Nawet jeśli coś wydaje się proste, prawdopodobnie jest to nowość dla grupy praktykujących analityków danych tylko dlatego, że dziedzina jest tak duża. Ludzie w firmach, dla których chcesz pracować, mogą nawet czytać Twoje rzeczy! Twitter jest dobrym miejscem do rozpoczęcia pracy nad przykładowymi projektami; możesz podzielić się wiadomościami po opublikowaniu posta i użyć odpowiednich hashtagów, aby zdobyć szersze grono czytelników. Ale Twój blog jest cenny, nawet jeśli nikt (oprócz Twojego partnera

i zwierzaka) go nie czyta. Pisanie posta na blogu to dobra praktyka; zmusza cię do uporządkowania swoich myśli. Podobnie jak nauczanie osobiście, pomaga również uświadomić sobie, że nie wiesz czegoś tak dobrze, jak myślałeś, że wiesz.

Praca nad przykładowymi projektami

W tej sekcji przeprowadzimy Cię przez dwa przykładowe projekty, od początkowego pomysłu przez analizę do końcowego publicznego artefaktu. Wykorzystamy prawdziwe projekty : stworzenie aplikacji internetowej dla freelancerów zajmujących się nauką o danych, aby znaleźć najlepiej dopasowane zawody i uczenie się sieci neuronowych poprzez szkolenie jednego na zbiorze danych z zakazanymi tablicami rejestracyjnymi.

Freelancerzy zajmujący się badaniem danych

PYTANIE

Kiedy byłem początkującym naukowcem zajmującym się danymi, zainteresowałem się jednym ze sposobów, w jaki niektórzy badacze danych zarabiają dodatkowe pieniądze: freelancerem. Freelancing to robienie projektów dla kogoś, u kogo nie jesteś zatrudniony, niezależnie od tego, czy jest to inna osoba, czy duża firma. Projekty te obejmują od kilku godzin do miesięcy pracy w pełnym wymiarze godzin. Możesz znaleźć wiele ofert pracy dla freelancerów opublikowanych na stronach internetowych freelancerów, takich jak UpWork, ale ponieważ nauka o danych to bardzo szeroka dziedzina, stanowiska w tej kategorii mogą obejmować wszystko, od tworzenia stron internetowych, przez analizę w programie Excel, po przetwarzanie języka naturalnego na terabajtach danych. Postanowiłem zobaczyć czy mógłbym pomóc freelancerom przebrnąć przez tysiące ofert pracy, aby znaleźć te, które najlepiej do nich pasują.

ANALIZA

Aby zebrać dane, wykorzystałem API UpWork do pobrania aktualnie dostępnych ofert pracy i profili wszystkich w kategorii Data Science and Analytics. Skończyło się na 93 000 freelancerów i 3000 miejsc pracy. Chociaż API umożliwiło stosunkowo łatwy dostęp do danych (ponieważ nie musiałem robić web scrapingu), nadal musiałem tworzyć funkcje do wykonywania setek wywołań API, obsługi, gdy te wywołania API nie powiodły się, a następnie przekształcania danych tak Przydałby mi się. Ale zaletą tego procesu było to, że ponieważ dane nie były łatwo dostępne, nie było setek innych osób pracujących nad tym samym projektem, jak byłoby, gdybym użył danych z konkursu Kaggle. Po uzyskaniu danych w dobrym stanie przeprowadziłem analizę eksploracyjną. Przyjrzałem się, jak poziom wykształcenia i kraj wpływają na zarobki freelancerów. Stworzyłem również wykres korelacji umiejętności wymienianych przez freelancerów, który pokazał różne typy freelancerów: programiści stron internetowych (PHP, jQuery, HTML i CSS), finanse i księgowość (rachunkowość finansowa, księgowość i analiza finansowa) oraz dane gromadzenie (wprowadzanie danych, generowanie leadów, eksploracja danych i web scraping) wraz z „tradycyjnym" zestawem umiejętności data science (Python, uczenie maszynowe, statystyki i analiza danych). Na koniec utworzyłem wynik podobieństwa między tekstem profilu a tekstem stanowiska i połączyłem ten wynik z nakładaniem się umiejętności (zarówno freelancerów, jak i umiejętności z listy stanowisk), aby stworzyć pasujący wynik dla freelancera i pracy.

PRODUKT KOŃCOWY

W tym przypadku nie napisałem posta na blogu. Zamiast tego stworzyłem interaktywną aplikację internetową, w której ktoś mógł wprowadzić swój tekst profilu, umiejętności i wymagania dotyczące pracy (takie jak minimalny wynik opinii dla oferty pracy i czas trwania pracy), a dostępne oferty byłyby filtrowane aby spełnić te wymagania i posortowane według tego, jak dobrze pasują do użytkownika.

Nie pozwoliłem, żeby to, co doskonałe, było tutaj wrogiem dobrego; jest wiele sposobów na ulepszenie projektu. Wyciągnąłem zlecenia tylko raz, a ponieważ wykonałem ten projekt cztery lata temu, aplikacja nadal działa, ale żadna z tych ofert nie jest już dostępna. Aby aplikacja była wartościowa na dłuższą metę, musiałbym co wieczór pobierać zlecenia i aktualizować oferty. Mogłem też stworzyć bardziej wyrafinowany algorytm dopasowywania, przyspieszyć początkowy czas ładowania aplikacji i sprawić, by wygląd był bardziej wyszukany. Jednak pomimo tych ograniczeń projekt zrealizował kilka ważnych celów. Pokazał, że mogę wziąć projekt i pozwolić ludziom na interakcję z nim, a nie ograniczać się do statycznych analiz, które znajdowały się na moim laptopie. Miał rzeczywisty przypadek użycia: pomaganie freelancerom w znalezieniu pracy. Przeprowadziło mnie to przez cały cykl projektu z dziedziny nauki o danych: zbieranie danych, czyszczenie ich, przeprowadzanie analiz eksploracyjnych i generowanie ostatecznego wyniku.

Szkolenie sieci neuronowej na obraźliwych tablicach rejestracyjnych

PYTANIE

Gdy dorastałem jako naukowiec zajmujący się danymi, zawsze byłem sfrustrowany, gdy widziałem zabawne posty na blogach, w których ludzie trenowali sieci neuronowe, aby generować takie rzeczy, jak nowe nazwy zespołów, nowe Pokémony i dziwne przepisy kulinarne. Myślałem, że te projekty są świetne, ale sam nie wiedziałem, jak je wykonać! Pewnego dnia przypomniałem sobie, że słyszałem o zbiorze danych wszystkich niestandardowych tablic rejestracyjnych, które zostały odrzucone przez stan Arizona za zbyt obraźliwe. Gdybym mógł zdobyć ten zestaw danych, byłby to idealny sposób na nauczenie się tworzenia sieci neuronowych - mógłbym stworzyć własne obraźliwe tablice rejestracyjne.

ANALIZA

Po złożeniu wniosku o rejestrację publiczną w Departamencie Transportu Arizony otrzymałem listę tysięcy obraźliwych tablic rejestracyjnych. Nie wiedziałem nic o sieciach neuronowych, więc po otrzymaniu danych zacząłem przeszukiwać internet w poszukiwaniu wpisów na blogu opisujących, jak je stworzyć. Jako użytkownik głównie R, z przyjemnością znalazłem pakiet Keras od RStudio do tworzenia sieci neuronowych w R. Załadowałem dane do R, a następnie sprawdziłem przykład pakietu RStudio Keras do generowania tekstu za pomocą sieci neuronowych. Zmodyfikowałem kod, aby działał z danymi z tablic rejestracyjnych; przykład RStudio był do generowania sekwencji długiego tekstu, ale chciałem trenować na siedmioznakowych tablicach rejestracyjnych. Oznaczało to utworzenie wielu punktów danych treningowych dla mojego modelu z każdej tablicy rejestracyjnej (jeden punkt danych do przewidywania każdego znaku na tablicy rejestracyjnej). Następnie przeszkoliłem model sieci neuronowej, chociaż na początku nie działał. Po odłożeniu projektu na miesiąc wróciłem i zdałem sobie sprawę, że moje dane nie są przetwarzane poprawnie. Kiedy naprawiłem ten problem, wyniki wygenerowane przez sieć neuronową były fantastyczne. Ostatecznie, mimo że nie zmieniłem zbytnio przykładu RStudio, pod koniec czułem się znacznie pewniej w tworzeniu i korzystaniu z sieci neuronowych.

PRODUKT KOŃCOWY

Napisałem post na blogu o projekcie, który opisuje, w jaki sposób uzyskałem dane, jak je przetwarzałem, aby były gotowe do sieci neuronowej i jak zmodyfikowałem przykładowy kod RStudio, aby działał dla mnie. Wpis na blogu był w dużym stopniu w stylu „Jestem nowy w sieciach neuronowych, a oto czego się nauczyłem”; Nie udawałem, że już wiem, jak to wszystko działa. W ramach wpisu na blogu stworzyłem obraz, który pobrał tekst z mojego modelu neuronowego i nadał mu wygląd tablic rejestracyjnych Arizony. Kod umieściłem również na GitHubie. Odkąd napisałem ten

post na blogu i udostępniłem mój kod, wiele innych osób zmodyfikowało go, aby stworzyć własne zabawne sieci neuronowe. To, czego nauczyłem się z tego głupiego projektu, ostatecznie pomogło mi w stworzeniu efektywnych modeli uczenia maszynowego dla ważnych zadań konsultingowych. Tylko dlatego, że oryginalna praca nie jest poważna, nie oznacza to, że nie ma w niej wartości!