Bez względu na branżę, w której pracujesz: IT, moda, żywność czy finanse, nie ma wątpliwości, że dane wpływają na Twoje życie i pracę. W pewnym momencie w tym tygodniu odbędziesz lub usłyszysz rozmowę na temat danych. Serwisy informacyjne relacjonują coraz więcej historii o wyciekach danych, cyberprzestępczości oraz o tym, jak dane mogą dać nam wgląd w nasze życie. Ale dlaczego teraz? Co sprawia, że ta era jest tak siedliskiem branż związanych z danymi? W XIX wieku świat był w szponach epoki przemysłowej. Ludzkość badała swoje miejsce w przemyśle obok gigantycznych wynalazków mechanicznych. Kapitanowie przemysłu, tacy jak Henry Ford, dostrzegli duże możliwości rynkowe tkwiące w tych maszynach i byli w stanie osiągnąć niewyobrażalne wcześniej zyski. Oczywiście epoka przemysłowa miała swoje plusy i minusy. Podczas gdy masowa produkcja umieszczała towary w rękach większej liczby konsumentów, nasza walka z zanieczyszczeniem również zaczęła się w tym czasie. W XX wieku byliśmy już dość biegli w tworzeniu ogromnych maszyn; teraz celem było uczynienie ich mniejszymi i szybszymi. Epoka przemysłowa się skończyła i została zastąpiona przez to, co nazywamy epoką informacji. Zaczęliśmy używać maszyn do gromadzenia i przechowywania informacji (danych) o nas samych i naszym środowisku w celu zrozumienia naszego wszechświata. Począwszy od lat 40. maszyny takie jak ENIAC (uważane za jeden z, jeśli nie pierwszy, komputer) obliczały równania matematyczne oraz uruchamiały modele i symulacje jak nigdy dotąd. W końcu mieliśmy przyzwoitego asystenta laboratoryjnego, który potrafił liczyć lepiej niż my! Podobnie jak w epoce przemysłowej, epoka informacji przyniosła nam zarówno dobre, jak i złe. Dobra była niezwykła technologia, w tym telefony komórkowe i telewizory. Zło w tym przypadku nie było tak złe, jak zanieczyszczenie na całym świecie, ale nadal pozostawiło nam problem w XXI wieku, tak wiele danych. Zgadza się, era informacji, w swoim dążeniu do pozyskiwania danych, eksplodowała produkcją danych elektronicznych. Szacunki pokazują, że w 2011 roku utworzyliśmy około 1,8 biliona gigabajtów danych (poświęć chwilę, aby pomyśleć, ile to jest). Zaledwie rok później, w 2012 roku, stworzyliśmy ponad 2,8 biliona gigabajtów danych! Ta liczba będzie rosła jeszcze dalej, aby osiągnąć około 40 bilionów gigabajtów danych w ciągu zaledwie jednego roku do 2020 roku. Ludzie przyczyniają się do tego za każdym razem, gdy tweetują, publikują na Facebooku, zapisują nowe CV w Microsoft Word lub po prostu wysyłają swojej mamie zdjęcie za pośrednictwem wiadomości tekstowej. Nie tylko tworzymy dane w niespotykanym dotąd tempie, ale także zużywamy je w przyspieszonym tempie. Zaledwie w 2013 roku, przeciętny użytkownik telefonu komórkowego zużył mniej niż 1 GB danych miesięcznie. Obecnie szacuje się, że liczba ta wynosi grubo ponad 2 GB miesięcznie. Nie szukamy tylko następnego quizu osobowości, szukamy wglądu. Wszystkie te dane tam są, niektóre z nich muszą być dla mnie przydatne! I może być! Więc my, w XXI wieku, mamy problem. Mamy tak dużo danych i wciąż robimy więcej. Zbudowaliśmy szalenie malutkie maszyny, które zbierają dane 24 godziny na dobę, 7 dni w tygodniu, a naszym zadaniem jest nadanie im sensu. Wprowadź wiek danych. To jest wiek, kiedy bierzemy maszyny wymyślone przez naszych XIX-wiecznych przodków i dane stworzone przez naszych odpowiedników z XX wieku i tworzymy spostrzeżenia i źródła wiedzy, z których każdy człowiek na Ziemi może skorzystać. Stany Zjednoczone stworzyły dla głównego naukowca danych zupełnie nową rolę w rządzie. Firmy technologiczne, takie jak Reddit, które do tej pory nie miały w swoim zespole analityka danych, teraz zatrudniają ich na prawo i lewo. Korzyść jest dość oczywista – wykorzystanie danych do tworzenia dokładnych prognoz i symulacji daje nam wgląd w nasz świat jak nigdy dotąd. Brzmi świetnie, ale w czym tkwi haczyk? W tej części poznamy terminologię i słownictwo współczesnego naukowca danych. Zobaczymy kluczowe słowa i wyrażenia, które są kluczowe w naszej dyskusji na temat nauki o danych. Zanim zaczniemy przyglądać się kodowi w Pythonie, podstawowym języku używanym przez nas, przyjrzymy się również, dlaczego korzystamy z nauki o danych i trzech kluczowych domenach, z których wywodzi się nauka o danych:

- Podstawowa terminologia nauki o danych
- Trzy dziedziny nauki o danych

• Podstawowa składnia Pythona

**Co to jest nauka o danych?**

Zanim przejdziemy dalej, spójrzmy na kilka podstawowych definicji, których będziemy używać. Wielką/okropną rzeczą w tej dziedzinie jest to, że jest ona tak młoda, że definicje te mogą się różnić w zależności od podręcznika, gazety i białej księgi.

**Podstawowa terminologia**

Poniższe definicje są na tyle ogólne, że można je stosować w codziennych rozmowach i pracy, przy wprowadzeniu do zasad data science. Zacznijmy od zdefiniowania, czym są dane. Może się to wydawać głupią pierwszą definicją, ale jest to bardzo ważne. Ilekroć używamy słowa „dane", odnosimy się do zbierania informacji w zorganizowanym lub niezorganizowanym formacie:

• Dane uporządkowane: Odnosi się to do danych posortowanych w strukturę wiersz/kolumna, gdzie każdy wiersz reprezentuje pojedynczą obserwację, a kolumny reprezentują cechy tej obserwacji.

• Niezorganizowane dane: Jest to rodzaj danych w dowolnej formie, zwykle tekstu lub nieprzetworzonego dźwięku/sygnałów, które muszą być dalej analizowane w celu uporządkowania. Za każdym razem, gdy otwierasz program Excel (lub dowolny inny program do obsługi arkuszy kalkulacyjnych), patrzysz na pustą strukturę wiersza/kolumny czekającą na uporządkowane dane. Te programy nie radzą sobie dobrze z niezorganizowanymi danymi. W większości będziemy mieli do czynienia z uporządkowanymi danymi, ponieważ najłatwiej jest z nich uzyskać wgląd, ale nie będziemy stronić od surowego tekstu i metod przetwarzania niezorganizowane formy danych.

Nauka o danych to sztuka i nauka zdobywania wiedzy poprzez dane.

Cóż za mała definicja tak dużego tematu i słusznie! Nauka o danych obejmuje tak wiele rzeczy, że zajęłoby strony, aby to wszystko wymienić.

Nauka o danych polega na tym, jak zbieramy dane, wykorzystujemy je do zdobywania wiedzy, a następnie wykorzystujemy tę wiedzę do wykonywania następujących czynności:

• Podejmować decyzje

• Przewidzieć przyszłość

• Zrozumieć przeszłość/teraźniejszość

• Twórz nowe branże/produkty

Ten tekst dotyczy metod nauki o danych, w tym sposobu przetwarzania danych, gromadzenia spostrzeżeń i wykorzystywania tych spostrzeżeń do podejmowania świadomych decyzji i prognoz. Nauka o danych polega na wykorzystywaniu danych w celu uzyskania nowych spostrzeżeń, których w przeciwnym razie byś nie zauważył. Jako przykład wyobraź sobie, że siedzisz przy stole z trzema innymi osobami. Wasza czwórka musi podjąć decyzję w oparciu o pewne dane. Należy wziąć pod uwagę cztery opinie. Użyłbyś nauki o danych, aby przedstawić piątą, szóstą, a nawet siódmą opinię. Dlatego nauka o danych nie zastąpi ludzkiego mózgu, ale go uzupełni, będzie z nim współpracować. Nauka o danych nie powinna być traktowana jako ostateczne rozwiązanie naszych problemów z danymi; jest to tylko opinia, bardzo poinformowana opinia, niemniej jednak opinia. Zasługuje na miejsce przy stole.

**Dlaczego nauka o danych?**

W erze danych jasne jest, że mamy nadwyżkę danych. Ale dlaczego miałoby to wymagać całego nowego zestawu słownictwa? Co było nie tak z naszymi poprzednimi formami analizy? Po pierwsze, sama ilość danych sprawia, że dosłownie niemożliwe jest przeanalizowanie ich przez człowieka w rozsądnym czasie. Dane zbierane są w różnych formach i z różnych źródeł, często bardzo niezorganizowane. Może brakować danych, mogą być niekompletne lub po prostu niepoprawne. Często dysponujemy danymi w bardzo różnych skalach, co utrudnia ich porównywanie. Weź pod uwagę, że patrzymy na dane w odniesieniu do wyceny używanych samochodów. Jedną cechą samochodu jest rok produkcji, a inną może być liczba mil na tym samochodzie. Kiedy oczyścimy nasze dane (któremu poświęciliśmy dużo czasu na przeglądanie w tej książce), relacje między danymi stają się bardziej oczywiste, a wiedza, która kiedyś była zakopana głęboko w milionach wierszy danych, po prostu wyskakuje. Jednym z głównych celów nauki o danych jest stworzenie wyraźnych praktyk i procedur w celu odkrycia i zastosowania tych relacji w danych. Wcześniej przyjrzeliśmy się nauce o danych w bardziej historycznej perspektywie, ale poświęćmy chwilę na omówienie jej roli w dzisiejszym biznesie na bardzo prostym przykładzie.

**Przykład - Sigma Technologies**

Ben Runkle, dyrektor generalny Sigma Technologies, próbuje rozwiązać ogromny problem. Firma konsekwentnie traci wieloletnich klientów. Nie wie, dlaczego odchodzą, ale musi coś szybko zrobić. Jest przekonany, że aby zmniejszyć ewakuację, musi tworzyć nowe produkty i funkcje oraz konsolidować istniejące technologie. Aby być bezpiecznym, wzywa swojego głównego analityka danych, dr Jessie Hughan. Nie jest jednak przekonany, że same nowe produkty i funkcje uratują firmę. Zamiast tego sięga do transkrypcji ostatnich biletów obsługi klienta. Pokazuje Runkle'owi najnowsze transkrypcje i znajduje coś zaskakującego:

• … Nie wiesz, jak to wyeksportować; prawda?"

• „Gdzie jest przycisk, który tworzy nową listę?"

• „Czekaj, czy wiesz, gdzie jest suwak?"

• „Jeśli nie mogę tego dzisiaj rozgryźć, to jest prawdziwy problem …"

Oczywiste jest, że klienci mieli problemy z istniejącym interfejsem użytkownika/UX i nie byli zmartwieni brakiem funkcji. Runkle i Hughan zorganizowali masową przebudowę UI/UX, a ich sprzedaż nigdy nie była lepsza. Oczywiście nauka zastosowana w ostatnim przykładzie była minimalna, ale ma sens. Zwykle nazywamy takich ludzi jak Runkle, kierowca. Dzisiejszy, typowy dyrektor generalny, który trzyma się przeczucia, chce szybko podejmować wszystkie decyzje i powtarzać rozwiązania, aż coś zadziała. Dr Haghun jest znacznie bardziej analityczny. Chce rozwiązać problem tak samo jak Runkle, ale szuka odpowiedzi na dane generowane przez użytkowników, a nie na instynktowne wyczucie. Nauka o danych polega na stosowaniu umiejętności analitycznego umysłu i używaniu ich tak, jak zrobiłby to kierowca. Obie te mentalności mają swoje miejsce we współczesnych przedsiębiorstwach; jednak to sposób myślenia Hagun dominuje w koncepcjach data science – wykorzystanie danych generowanych przez firmę jako źródła informacji, a nie tylko wybieranie rozwiązania i podążanie z nim.

**Diagram Venna**

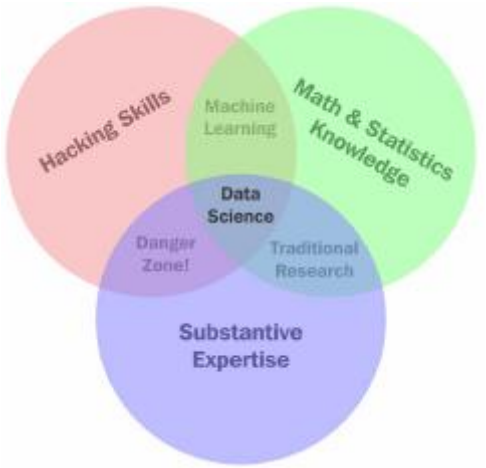
Powszechnym błędem jest przekonanie, że tylko osoby z doktoratem lub geniusze mogą zrozumieć matematykę/programowanie stojące za nauką o danych. To jest całkowicie fałszywe. Zrozumienie nauki o danych zaczyna się od trzech podstawowych obszarów:

• Matematyka/statystyka: jest to użycie równań i wzorów do przeprowadzenia analizy

• Programowanie komputerowe: jest to umiejętność używania kodu do tworzenia wyników na komputerze

• Wiedza dziedzinowa: odnosi się do zrozumienia problematycznej domeny (medycyna, finanse, nauki społeczne itd.)

Poniższy diagram Venna przedstawia wizualną reprezentację tego, jak przecinają się trzy obszary nauki o danych:




Osoby z umiejętnościami hakerskimi mogą konceptualizować i programować skomplikowane algorytmy za pomocą języków komputerowych. Posiadanie bazy wiedzy matematycznej i statystycznej pozwala teoretyzować i oceniać algorytmy oraz dostosowywać istniejące procedury do konkretnych sytuacji. Posiadanie wiedzy merytorycznej (ekspertyzy dziedzinowej) pozwala stosować koncepcje i wyniki w znaczący i skuteczny sposób.

Chociaż posiadanie tylko dwóch z tych trzech cech może uczynić cię inteligentnym, pozostawi również lukę. Weź pod uwagę, że jesteś bardzo biegły w kodowaniu i masz formalne szkolenie w zakresie daytradingu. Możesz stworzyć zautomatyzowany system do handlu na swoim miejscu, ale nie masz umiejętności matematycznych, aby ocenić swoje algorytmy, a zatem stracisz pieniądze w końcu. Tylko wtedy, gdy możesz pochwalić się umiejętnościami w zakresie kodowania, matematyki i znajomości domeny, możesz naprawdę wykonywać analizę danych.

To, co prawdopodobnie była dla Ciebie niespodzianką, to Domain Knowledge. To naprawdę tylko wiedza o obszarze, w którym pracujesz. Jeśli analityk finansowy zaczął analizować dane dotyczące zawałów serca, może potrzebować pomocy kardiologa, aby zrozumieć wiele liczb. Data Science to skrzyżowanie trzech kluczowych obszarów wspomnianych wcześniej. Aby uzyskać wiedzę z danych, musimy być w stanie wykorzystać programowanie komputerowe w celu uzyskania dostępu do danych, zrozumieć matematykę kryjącą się za modelami, które wyprowadzamy, oraz przede wszystkim zrozumieć miejsce naszych analiz w domenie, w której się znajdujemy. Obejmuje to prezentację danych. Jeśli tworzymy model do przewidywania zawałów serca u pacjentów, czy lepiej stworzyć plik PDF z informacjami lub aplikację, w której można wpisywać liczby i uzyskać szybką prognozę? Wszystkie te decyzje muszą podjąć badacz danych.

Pamiętaj też, że przecięciem matematyki i kodowania jest uczenie maszynowe. Szczegółowo przyjrzymy się uczeniu maszynowemu później, ale ważne jest, aby pamiętać, że bez wyraźnej możliwości uogólnienia jakichkolwiek modeli lub wyników na domenę, algorytmy uczenia maszynowego pozostają tylko algorytmami znajdującymi się na twoim komputerze. Być może masz najlepszy algorytm do przewidywania raka. Możesz być w stanie przewidzieć raka z ponad 99%

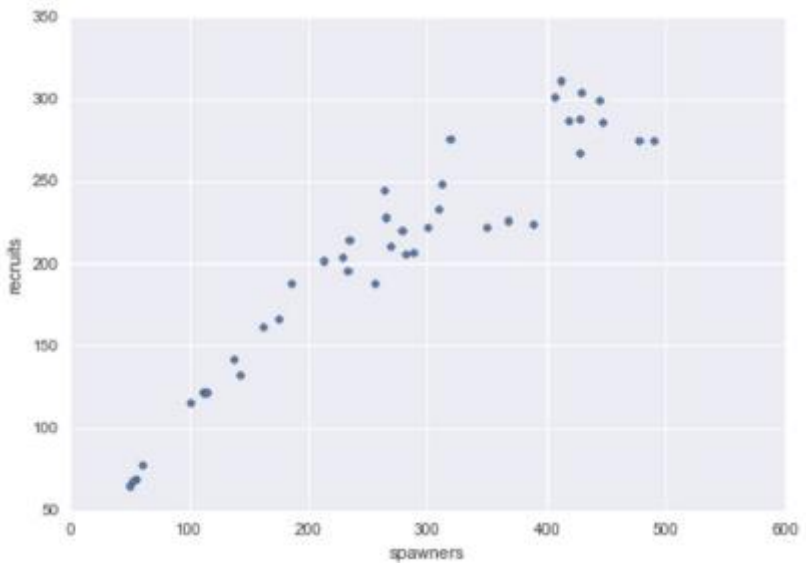
dokładnością na podstawie wcześniejszych danych pacjentów z rakiem, ale jeśli nie rozumiesz, jak zastosować ten model w praktycznym sensie, tak aby lekarze i pielęgniarki mogli z niego łatwo korzystać, Twój model może być bezużyteczny.

**Matematyka**

Większość ludzi przestaje słuchać, gdy ktoś wypowie słowo „matematyka". Kiwają głową, próbując ukryć całkowitą pogardę dla tematu. Te wpisy poprowadzą Cię przez matematykę potrzebną do nauki o danych, w szczególności statystyki i prawdopodobieństwa. Wykorzystamy te poddziedziny matematyki do stworzenia tak zwanych modeli. Model danych odnosi się do zorganizowanej i formalnej relacji między elementami danych, zwykle przeznaczonej do symulowania zjawiska w świecie rzeczywistym. Zasadniczo użyjemy matematyki w celu sformalizowania relacji między zmiennymi. Wiem, jak trudne może to być. Zrobię co w mojej mocy, aby wyjaśnić wszystko tak jasno, jak tylko potrafię. Pomiędzy trzema obszarami nauki o danych, matematyka jest tym, co pozwala nam przechodzić od domeny do domeny. Zrozumienie tej teorii pozwala nam zastosować model, który zbudowaliśmy dla branży modowej, do modelu finansowego. Matematyka omawiana tu obejmuje zakres od podstawowej algebry do zaawansowanego modelowania probabilistycznego i statystycznego. Nie pomijaj tych części, nawet jeśli już je znasz lub się ich boisz. Każde pojęcie matematyczne, które wprowadzam, robię z rozwagą, przykładami i celem. Matematyka jest niezbędna dla naukowców zajmujących się danymi.

**Przykład - modele rekrutacji ikry**

W biologii używamy, między innymi, modelu znanego jako model rekrutacji ikry, aby ocenić zdrowie biologiczne gatunku. Jest to podstawowa zależność między liczbą zdrowych jednostek rodzicielskich gatunku a liczbą nowych jednostek w grupie zwierząt. W publicznym zbiorze danych dotyczących liczby tarlaków i ikry łososia utworzono poniższy wykres, aby zobrazować związek między nimi.




Widzimy, że na pewno istnieje jakaś pozytywna relacja (w miarę, jak jeden idzie w górę, drugi też). Ale jak możemy sformalizować ten związek? Na przykład, gdybyśmy znali liczbę ikry w populacji, czy moglibyśmy przewidzieć liczbę rekrutów, które pozyska ta grupa i na odwrót? Zasadniczo modele pozwalają nam podłączyć jedną zmienną, aby uzyskać drugą. Rozważmy następujący przykład:

Rekruci = 0,5 * Ikra + 60

W tym przykładzie załóżmy, że wiedzieliśmy, że grupa łososi miała 1,15 (w tysiącach) tarła. Wtedy mielibyśmy:

Rekruci = 0,5 * 1,15 + 60

Rekruci = 60,575 (w tysiącach)

Ten wynik może być bardzo korzystny przy oszacowaniu, jak zmienia się stan zdrowia populacji. Jeśli uda nam się stworzyć te modele, możemy wizualnie zaobserwować, jak mogą zmieniać się relacje między dwiema zmiennymi. Istnieje wiele typów modeli danych, w tym modele probabilistyczne i statystyczne. Oba są podzbiorami większego paradygmatu, zwanego uczeniem maszynowym. Zasadniczą ideą stojącą za tymi trzema tematami jest to, że wykorzystujemy dane w celu opracowania najlepszego możliwego modelu. Nie polegamy już na ludzkich instynktach, raczej polegamy na danych. Celem tego przykładu jest pokazanie, w jaki sposób możemy zdefiniować relacje między elementami danych za pomocą równań matematycznych. Fakt, że wykorzystałem dane dotyczące zdrowia łososia, był nieistotny! Przyjrzymy się relacjom obejmującym dolary marketingowe, dane dotyczące nastrojów, recenzje restauracji i wiele innych. Głównym tego powodem jest to, że chciałbym, abyście (czytelnicy) mieli dostęp do jak największej liczby domen. Matematyka i kodowanie to narzędzia, które pozwalają analitykom danych cofnąć się i zastosować swoje umiejętności praktycznie w dowolnym miejscu

**Programowanie komputerowe**

Bądźmy szczerzy. Prawdopodobnie myślisz, że informatyka jest o wiele fajniejsza niż matematyka. W porządku, nie obwiniam cię. Wiadomości nie są wypełnione wiadomościami matematycznymi, tak jak wiadomościami na froncie technologicznym. Nie włączasz telewizora, aby zobaczyć nową teorię na temat liczb pierwszych, raczej zobaczysz raporty śledcze na temat tego, jak najnowszy smartfon może lepiej robić zdjęcia kotom czy coś takiego. Języki komputerowe to sposób, w jaki komunikujemy się z maszyną i każemy jej wykonywać nasze polecenia. Komputer mówi wieloma językami i, podobnie jak książka, może być napisany w wielu językach; podobnie, nauka o danych może być również wykonywana w wielu językach. Python, Julia i R to tylko niektóre z wielu dostępnych dla nas języków. My skupimy się wyłącznie na używaniu Pythona.

**Dlaczego Python?**

Pythona będziemy używać z różnych powodów:

• Python jest niezwykle prostym językiem do czytania i pisania, nawet jeśli nigdy wcześniej nie programowałeś, co sprawi, że przyszłe przykłady będą łatwe do przyswojenia i przeczytania później.

• Jest to jeden z najpopularniejszych języków, zarówno w środowisku produkcyjnym, jak i akademickim (w rzeczywistości jeden z najszybciej rozwijających się)

• Społeczność internetowa języka jest rozległa i przyjazna. Oznacza to, że szybkie wyszukiwanie w Google powinno dać wiele wyników osób, które napotkały i rozwiązały podobne (jeśli nie dokładnie te same) sytuacje

• Python ma gotowe moduły analizy danych, z których mogą korzystać zarówno nowicjusze, jak i doświadczeni naukowcy zajmujący się danymi

Ten ostatni jest prawdopodobnie największym powodem, dla którego skupimy się na Pythonie. Te gotowe moduły są nie tylko potężne, ale także łatwe do pobrania. Pod koniec pierwszych kilku rozdziałów te moduły będą bardzo wygodne. Niektóre z tych modułów to:

• pandas

• sci-kit learn

• seaborn

• numpy/scipy

• requests (to mine data from the Web)

• BeautifulSoup (do analizowania stron WWW-HTML)

**Praktyczny Python**

Zanim przejdziemy dalej, ważne jest sformalizowanie wielu wymaganych umiejętności kodowania w Pythonie. W Pythonie mamy zmienne, które są symbolami zastępczymi dla obiektów. Na początku skupimy się tylko na kilku typach podstawowych obiektów:

• int (liczba całkowita)

- Przykłady: 3, 6, 99, -34, 34, 11111111

• liczba zmiennoprzecinkowa (dziesiętna):

- Przykłady: 3,14159, 2,71, -0,34567

• wartość logiczna (prawda lub fałsz)

- Stwierdzenie, że niedziela jest weekendem, jest prawdziwe

- Stwierdzenie, że piątek jest weekendem, jest fałszywe

- Stwierdzenie, pi jest dokładnie stosunkiem obwodu koła do jego średnica, to prawda (szalony, prawda?)

• ciąg (tekst lub słowa złożone ze znaków)

- „Kocham hamburgery” (przy okazji, kto nie?)

- "Mat jest niesamowity"

- Tweet to ciąg

• lista (zbiór obiektów)

- Przykład: [1, 5.4, Prawda, "jabłko"]

Będziemy musieli również zrozumieć kilka podstawowych operatorów logistycznych. W przypadku tych operatorów należy pamiętać o typie danych logicznych. Każdy operator oceni True lub False. Rzućmy okiem na następujące ilustracje:

• == zwraca wartość Prawda, jeśli obie strony są równe; w przeciwnym razie zwraca się do Fałsz

- 3 + 4 == 7 (oznacza wartość Prawda)

- 3 – 2 == 7 (oznacza wartość Fałsz)

• < (mniej niż)

- 3 < 5 (prawda)

- 5 < 3 (fałsz)

• <= (mniejszy lub równy)

- 3 <= 3 (prawda)

- 5 <= 3 (fałsz)

• > (większe niż)

- 3 > 5 (fałsz)

- 5 > 3 (prawda)

• >= (większe lub równe)

- 3 >= 3 (prawda)

- 5 >= 3 (fałsz)

Podczas kodowania w Pythonie użyję znaku krzyżyka (#), aby utworzyć komentarz", który nie będzie przetwarzany jako kod, ale służy jedynie do komunikacji z czytelnikiem. Wszystko na prawo od znaku # jest komentarzem do wykonywany kod.

**Przykład podstawowego Pythona**

W Pythonie używamy spacji/tabulatorów do oznaczania operacji należących do innych linii kodu. Zwróć uwagę na użycie instrukcji if. Oznacza dokładnie to, co myślisz, że oznacza. Jeśli instrukcja po instrukcji if ma wartość True, część z kartami pod nią zostanie wykonana, jak pokazano w poniższym kodzie:

```
X = 5,8

Y = 9,5

X + Y == 15,3 # To prawda!

X - Y == 15,3 # To jest nieprawda!

5if x + y == 15.3: # Jeśli stwierdzenie jest prawdziwe

print "Prawda!" # wyświetl coś!
```

Instrukcja print „Prawda!" należy do linii if x + y == 15.3: linii poprzedzającej ją, ponieważ znajduje się tuż pod nią. Oznacza to, że instrukcja print zostanie wykonana wtedy i tylko wtedy, gdy x + y równa się 15.3.

Zauważ, że następująca zmienna list, my_list, może przechowywać wiele typów obiektów. Ma dane wejściowe typu int, float, boolean i string (w tej kolejności):

```
my_list = [1, 5.7, Prawda, "jabłka"]

len(my_list) == 4 # 4 obiekty na liście

my_list[0] == 1 # pierwszy obiekt

my_list[1] == 5.7 # drugi obiekt
```

W poprzednim kodzie:

• Użyłem polecenia len, aby uzyskać długość listy (która wynosiła cztery).

• Zwróć uwagę na indeksowanie zerowe w Pythonie. Większość języków komputerowych zaczyna liczyć od zera zamiast od jednego. Więc jeśli chcę pierwszy element, nazywam indeks zero, a jeśli chcę 95 element, nazywam indeks 94.

**Przykład - parsowanie pojedynczego tweeta**

Oto trochę kodu w Pythonie. W tym przykładzie przeanalizuję kilka tweetów na temat cen akcji (jednym z ważnych studiów przypadku będzie próba przewidzenia ruchów rynkowych w oparciu o popularne nastroje dotyczące akcji w mediach społecznościowych):

```
tweet = "RT @j_o_n_dnger: $TWTR now top holding for Andor, unseating $AAPL"

words_in_tweet = first_tweet.split(' ') # list of words in tweet

for word in words_in_tweet: # for each word in list

 if "$" in word: # if word has a "cashtag"

print "THIS TWEET IS ABOUT", word # alert the user
```

Wskażę kilka rzeczy na temat tego fragmentu kodu, linia po linii, w następujący sposób:

• Ustawiamy zmienną do przechowywania tekstu (znanego jako ciąg znaków w Pythonie). W tym przykładzie tweet, o którym mowa, to „RT @robdv: $TWTR teraz trzyma górę dla Andora, usuwa $AAPL”

• Zmienna words_in_tweet tokenizuje tweet (oddziela go słowem). Gdybyś miał wydrukować tę zmienną, zobaczyłbyś co następuje:

```
['RT',

'@robdv:',

'$TWTR',

'now',

'top',

'holding',

'for',

'Andor,',

'unseating',

'$AAPL']
```

• Powtarzamy tę listę słów. Nazywa się to pętlą for. Oznacza to po prostu, że przeglądamy listę jeden po drugim.

• Tutaj mamy kolejną instrukcję if. Dla każdego słowa w tym tweecie, jeśli słowo zawiera znak $ (w ten sposób ludzie odwołują się do notowań giełdowych na Twitterze).

• Jeśli poprzednie stwierdzenie if jest prawdziwe (tzn. jeśli tweet zawiera tag cashtag), wydrukuj je i pokaż użytkownikowi. Wynik tego kodu będzie następujący:

```
THIS TWEET IS ABOUT $TWTR
```

THIS TWEET IS ABOUT $AAPL

Otrzymujemy to wyjście, ponieważ są to jedyne słowa w tweecie, które używają cashtagu. Za każdym razem, gdy używam Pythona, upewnię się, że w każdym wierszu kodu mówię tak jasno, jak to tylko możliwe, o tym, co robię.

**Wiedza domenowa**

Jak wspomniałem wcześniej, ta kategoria skupia się głównie na posiadaniu wiedzy na konkretny temat, nad którym pracujesz. Na przykład, jeśli jesteś analitykiem finansowym pracującym na danych giełdowych, masz dużą wiedzę domenową. Jeśli jesteś dziennikarzem i przyglądasz się światowym wskaźnikom adopcji, możesz skorzystać z konsultacji z ekspertem w tej dziedzinie. Spróbujemy pokazać przykłady z kilku problematycznych dziedzin, w tym medycyny, marketingu, finansów, a nawet obserwacji UFO! Czy to oznacza, że jeśli nie jesteś lekarzem, nie możesz pracować z danymi medycznymi? Oczywiście nie! Świetni analitycy danych mogą zastosować swoje umiejętności w dowolnym obszarze, nawet jeśli nie są w tym biegli. Analitycy danych mogą dostosować się do dziedziny i wnieść znaczący wkład po zakończeniu analizy. Dużą częścią wiedzy domenowej jest prezentacja. W zależności od odbiorców, sposób prezentacji wyników może mieć duże znaczenie. Twoje wyniki są tak dobre, jak Twój środek komunikacji. Możesz przewidzieć ruch na rynku z 99,99% dokładnością, ale jeśli Twój program jest niemożliwy do wykonania, Twoje wyniki pozostaną niewykorzystane. Podobnie, jeśli twój pojazd jest nieodpowiedni do pola, twoje wyniki będą równie niewykorzystane.

**Trochę więcej terminologii**

To dobry moment, aby zdefiniować więcej słownictwa. W tym momencie prawdopodobnie podekscytowany przeglądasz wiele materiałów z zakresu analizy danych i widzisz słowa i wyrażenia, których jeszcze nie używałem. Oto kilka typowych terminologii, z którymi możesz się spotkać:

• Uczenie maszynowe: odnosi się do zapewnienia komputerom możliwości uczenia się na podstawie danych bez wyraźnych „reguł" narzucanych przez programistę. Widzieliśmy wcześniej koncepcję uczenia maszynowego jako połączenie kogoś, kto ma zarówno umiejętności kodowania, jak i matematyki. Tutaj próbujemy sformalizować tę definicję. Uczenie maszynowe łączy moc komputerów z inteligentnymi algorytmami uczenia w celu zautomatyzowania odkrywania relacji w danych i tworzenia potężnych modeli danych. Mówiąc o modelach danych, zajmiemy się następującymi dwoma podstawowymi typami modeli danych:

• Model probabilistyczny: Odnosi się do wykorzystania prawdopodobieństwa do znalezienia związku między elementami, który zawiera pewien stopień losowości.

• Model statystyczny: Odnosi się do wykorzystania twierdzeń statystycznych do sformalizowania relacji między elementami danych w (zwykle) prostym wzorze matematycznym.

Chociaż zarówno modele statystyczne, jak i probabilistyczne mogą być uruchamiane na komputerach i mogą być pod tym względem uważane za uczenie maszynowe, oddzielimy te definicje, ponieważ algorytmy uczenia maszynowego generalnie próbują uczyć się relacji na różne sposoby.

• Eksploracyjna analiza danych (EDA) odnosi się do przygotowania danych w celu standaryzowania wyników i szybko uzyskiwać wgląd. EDA zajmuje się wizualizacją i przygotowaniem danych. W tym miejscu zamieniamy niezorganizowane dane w uporządkowane dane, a także usuwamy brakujące/niepoprawne punkty danych. Podczas EDA stworzymy wiele rodzajów wykresów i użyjemy

tych wykresów do zidentyfikowania kluczowych cech i relacji do wykorzystania w naszych modelach danych.

• Eksploracja danych to proces znajdowania relacji między elementami danych. Eksploracja danych to część nauki o danych, w której staramy się znaleźć relacje między zmiennymi (pomyśl model spawn-recruit).

• Do tej pory bardzo starałem się nie używać terminu big data. To dlatego, że uważam, że ten termin jest często nadużywany. Chociaż definicja tego słowa różni się w zależności od osoby, duże zbiory danych. Big Data to dane, które są zbyt duże, aby mogły zostać przetworzone przez pojedynczą maszynę (jeśli Twój laptop uległ awarii, może to oznaczać przypadek big data).

**Studia przypadków związane z nauką o danych**

Połączenie matematyki, programowania komputerowego i wiedzy dziedzinowej sprawia, że nauka o danych jest tak potężna. Często jednej osobie trudno jest opanować wszystkie trzy z tych obszarów. Dlatego bardzo często firmy zatrudniają zespoły analityków danych zamiast jednej osoby. Przyjrzyjmy się kilku potężnym przykładom analizy danych w działaniu i ich wynikom.

**Studium przypadku - automatyzacja rządowej pracy kancelaryjnej**

Wiadomo, że roszczenia z tytułu ubezpieczenia społecznego są poważnym problemem zarówno dla agenta, który je czyta, jak i dla osoby, która je napisała. Niektóre roszczenia zajmują ponad 2 lata, aby zostać rozwiązane w całości, a to absurd! Co składa się na roszczenie? To głównie tekst. Wypełnij to, potem tamto, potem to i tak dalej. Widać, jak trudno byłoby agentowi czytać je przez cały dzień, formularz po formularzu. Musi być lepszy sposób! Cóż, jest. Elder Research Inc. przeanalizował te niezorganizowane dane i był w stanie zautomatyzować 20% wszystkich formularzy ubezpieczenia społecznego osób niepełnosprawnych. Oznacza to, że komputer mógłby przejrzeć 20% tych pisemnych formularzy i wydać opinię na temat zatwierdzenia. Co więcej, firma zewnętrzna, która jest wynajęta do oceny aprobat formularzy, faktycznie nadała formom maszynowym wyższą ocenę niż formy ludzkie. Tak więc komputer nie tylko poradził sobie z 20% obciążeniem, ale średnio radził sobie lepiej niż człowiek.

**Wystrzelaj wszystkich ludzi, prawda?**

Zanim dostanę mnóstwo wściekłych e-maili, w których twierdzą, że data science oznacza koniec ludzkiej pracy, pamiętaj, że komputer był w stanie obsłużyć tylko 20% obciążenia. Oznacza to, że prawdopodobnie spisał się fatalnie na 80% form! To dlatego, że komputer był chyba świetny w prostych formach. Obliczenie twierdzeń, które zajęłoby człowiekowi minuty, zajęło komputerowi sekundy. Ale te minuty sumują się i zanim się zorientujesz, każdy człowiek jest ratowany przez ponad godzinę dziennie! Formularze, które mogą być łatwe do odczytania przez człowieka, są również łatwe do odczytania przez komputer. Dopiero gdy forma staje się bardzo zwięzła lub gdy pisarz zaczyna odchodzić od zwykłej gramatyki, komputer zaczyna zawodzić. Ten model jest świetny, ponieważ pozwala ludziom poświęcić więcej czasu na te trudne roszczenia i poświęca im więcej uwagi bez rozpraszania się natłokiem papierów.

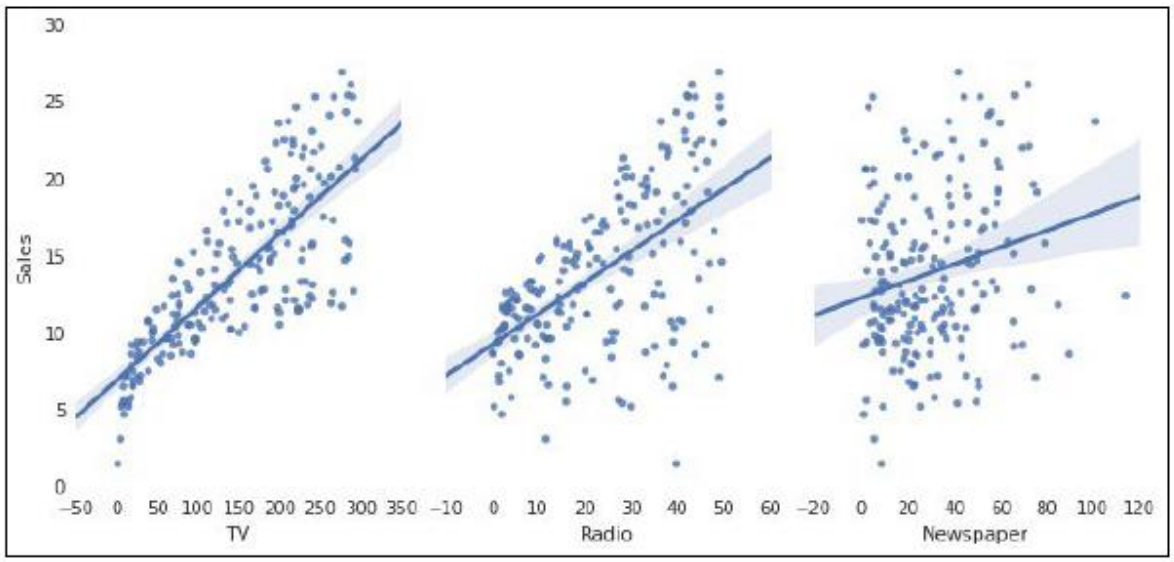
Zauważ, że użyłem słowa model. Pamiętaj, że model to relacja między elementami. W tym przypadku związek zachodzi między słowami pisemnymi a statusem zatwierdzenia roszczenia.

**Studium przypadku - dolary marketingowe**

Zestaw danych pokazuje związek między pieniędzmi wydanymi w kategoriach telewizji, radia i gazety. Celem jest analiza relacji między trzema różnymi mediami marketingowymi i ich wpływu na sprzedaż

produktu. Nasze dane mają postać struktury wierszowej i kolumnowej. Każdy wiersz reprezentuje region sprzedaży, a kolumny informują nas, ile pieniędzy wydano na każde medium i jaki zysk uzyskano w tym regionie.

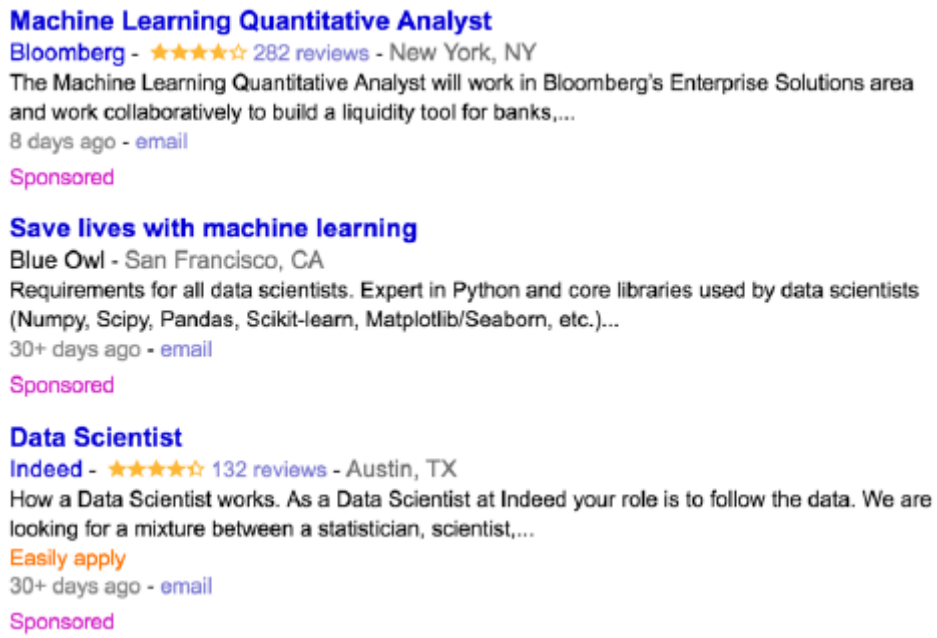
Zwykle badacz danych musi poprosić o jednostki i skalę. W tym przypadku powiem wam, że telewizja, radio i gazeta są mierzone w „tysiącach dolarów", a sprzedaż w „tysiącach sprzedanych gadżetów". Oznacza to, że w pierwszym regionie wydano 230 100 USD na reklamę telewizyjną, 37 800 USD na reklamę radiową i 69 200 USD na reklamę w prasie. W tym samym regionie sprzedano 22 100 sztuk. Na przykład w trzecim regionie wydaliśmy 17 200 USD na reklamę telewizyjną i sprzedaliśmy 9 300 widżetów. Jeśli wykreślimy każdą zmienną w funkcji sprzedaży, otrzymamy następujące wykresy:




Zwróć uwagę, że żadna z tych zmiennych nie tworzy bardzo silnej linii i dlatego może nie działać dobrze przy przewidywaniu sprzedaży (samodzielnie). Telewizja jest najbliższa w tworzeniu oczywistej relacji, ale nawet to nie jest wspaniałe. W takim przypadku będziemy musieli uformować bardziej złożony model niż ten, który zastosowaliśmy w modelu spawner-recruiter i połączyć wszystkie trzy zmienne w celu modelowania sprzedaży. Ten typ problemu jest bardzo powszechny w nauce o danych. W tym przykładzie próbujemy zidentyfikować kluczowe cechy związane ze sprzedażą produktu. Jeśli uda nam się wyizolować te kluczowe cechy, możemy wykorzystać te relacje i zmienić wysokość wydatków na reklamę w różnych miejscach z nadzieją na zwiększenie sprzedaży.

**Studium przypadku - co zawiera opis stanowiska?**

Szukasz pracy w data science? Świetnie, pozwól mi pomóc. W tym studium przypadku „wydobyłem" (zaczerpnięte z Internetu) 1000 opisów stanowisk dla firm aktywnie zatrudniających analityków danych (stan na styczeń 2016 r.). Celem jest przyjrzenie się niektórym z najczęstszych słów kluczowych, których ludzie używają w opisach stanowisk.




```
import requests

# used to grab data from the web

from BeautifulSoup import BeautifulSoup

# used to parse HTML

from sklearn.feature_extraction.text import CountVectorizer

# used to count number of words and phrases (we will be using this module a lot)
```

Pierwsze dwa importy służą do pobierania danych internetowych z witryny Indeed.com, a trzeci import ma po prostu zliczać, ile razy pojawia się słowo lub fraza.

```
texts = []

# hold our job descriptions in this list

for index in range(0,1000,10): # go through 100 pages of indeed

page = 'indeed.com/jobs?q=data+scientist&start='+str(index)

# identify the url of the job listings

web_result = requests.get(page).text

# use requests to actually visit the url

soup BeautifulSoup(web_result)

# parse the html of the resulting page

for listing in soup.findAll('span', {'class':'summary'}:

# for each listing on the page

texts.append(listing.text)
```

# append the text of the listing to our list

Okej, zanim cię stracę, wszystko, co robi ta pętla, to przeglądanie 100 stron opisów stanowisk i dla każdej strony chwytanie każdego opisu stanowiska. Ważną zmienną są tutaj teksty, czyli lista ponad 1000 opisów stanowisk:

type(texts) # == list

vect = CountVectorizer(ngram_range=(1,2), stop_words='english')

# Get basic counts of one and two word phrases

matrix = vect.fit_transform(texts)

# fit and learn to the vocabulary in the corpus

print len(vect.get_feature_names()) # how many features are there

# There are 11,293 total one and two words phrases in my case!!

Pominąłem tutaj trochę kodu. Wyniki są następujące (reprezentowane jako fraza, a następnie liczba wystąpień):

experience 320

machine 306

learning 305

machine learning 294

techniques 266

statistical 215

team 197

analytics 173

business 167

statistics 159

algorithms 152

datamining 149

software 144

applied 141

programming 132

understanding 127

world 127

research 125

datascience 123

methods 122

join 122

quantitative 122

group 121

real 120

large 120

**Godne uwagi rzeczy:**

• Uczenie maszynowe i doświadczenie znajdują się na szczycie listy. Doświadczenie przychodzi z praktyką.

• Po tych słowach znajdują się słowa statystyczne, które sugerują znajomość matematyki i teorii.

• Słowo zespół jest bardzo wysoko, co oznacza, że będziesz musiał pracować z zespołem analityków danych; nie będziesz samotnym wilkiem.

• Słowa informatyki, takie jak algorytmy i programowanie, są rozpowszechnione.

• Słowa techniki, rozumienie i metody oznaczają więcej podejścia teoretycznego, ambiwalentne dla dowolnej dziedziny.

• Słowo biznes implikuje określoną dziedzinę problemu.

Jest wiele interesujących rzeczy, o których warto wspomnieć w tym studium przypadku, ale największym wnioskiem jest to, że istnieje wiele kluczowych słów i fraz, które składają się na rolę nauki o danych. To nie tylko matematyka, kodowanie czy wiedza domenowa; to naprawdę połączenie tych trzech pomysłów (zilustrowanych na przykładzie jednej osoby lub w zespole wieloosobowym) sprawia, że nauka o danych jest możliwa i skuteczna.

**PODSUMOWANIE**

Na początku zadałem proste pytanie, jaki jest haczyk nauki o danych? Cóż, jest jeden. To nie tylko zabawa, gry i modelowanie. Nasze poszukiwania coraz inteligentniejszych maszyn i algorytmów muszą mieć swoją cenę. Gdy szukamy nowych i innowacyjnych sposobów odkrywania trendów w danych, bestia czai się w cieniu. Nie mówię o krzywej uczenia się matematyki czy programowania, ani o nadmiarze danych. Epoka przemysłowa pozostawiła nas w ciągłej walce z zanieczyszczeniem. Kolejna era informacji pozostawiła po sobie ślad big data. Jakie więc niebezpieczeństwa może nam przynieść wiek danych? Epoka danych może prowadzić do czegoś znacznie bardziej złowrogiego - odczłowieczenia jednostki przez masowe dane. Coraz więcej osób wskakuje w dziedzinę nauki o danych, większość bez wcześniejszego doświadczenia w matematyce lub CS, co na pierwszy rzut oka jest świetne. Przeciętni analitycy danych mają dostęp do milionów danych profili randkowych, tweetów, recenzji online i wielu innych, aby przyspieszyć swoją edukację. Jeśli jednak wskoczysz do nauki o danych bez odpowiedniego kontaktu z teorią lub praktykami kodowania i bez poszanowania domeny, w której pracujesz, ryzykujesz nadmierne uproszczenie samego zjawiska, które próbujesz modelować. Załóżmy na przykład, że chcesz zautomatyzować lejek sprzedaży, tworząc uproszczony program, który szuka w LinkedIn bardzo konkretnych słów kluczowych w profilu LinkedIn danej osoby.

keywords = ["Saas", "Sprzedaż", "Przedsiębiorstwo"]

Świetnie, teraz możesz szybko przeskanować LinkedIn, aby znaleźć osoby spełniające Twoje kryteria. Ale co z osobą, która pisze „Software as a Service” zamiast „Saas” lub błędnie pisze „enterprise”. W jaki sposób Twój model zorientuje się, że ci ludzie również dobrze do siebie pasują? Nie należy ich zostawiać tylko dlatego, że data scientist na skróty w tak łatwy sposób nadmiernie uogólnił ludzi. Programista zdecydował się uprościć wyszukiwanie drugiego człowieka, szukając trzech podstawowych słów kluczowych i skończył z wieloma niewykorzystanymi szansami na stole. W następnym rozdziale przyjrzymy się różnym typom danych, które istnieją na świecie, od dowolnego tekstu po wysoce ustrukturyzowane pliki wierszy/kolumn. Przyjrzymy się również operacjom matematycznym, które są dozwolone dla różnych typów danych, a także wywnioskujemy spostrzeżenia na podstawie formy danych, które przychodzą.